BACHELOR'S DEGREE THESIS

Degree in Data Science and Engineering

Title: Statistical analysis of the databases of the Institut Ramon Llull

Author: Anna Patrícia Orteu Irurre

Advisor: Esther Coll Caldas

Tutor: Francesc Rey Micolau

Institution: Institut Ramon Llull

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Facultat d'Informàtica de Barcelona

Facultat de Matemàtiques i Estadística

Escola Tècnica Superior d'Enginyeria de Telecomunicació de Barcelona

Resum

L'Institut Ramon Llull (IRL) és un organisme públic creat amb l'objectiu de promoure el català a l'exterior en un ampli ventall d'àmbits com l'acadèmic, les arts visuals o la traducció de textos; internacionalitzant la cultura catalana. L'Institut està constituït per tres àrees de negoci:

1. L'Àrea de Llengua i Universitats
2. L'Àrea de Creació
3. L'Àrea de Literatura i Pensament

Dins de totes tres es comparteix la voluntat d'estudiar l'atorgament de subvencions per a finançar un gran rang d'activitats que promocionen el català arreu del món. A més a més, dins l'Àrea de Llengua i Universitats també es té la voluntat d'analitzar:

1. L'organització d'estades lingüístiques en territoris de parla catalana
2. L'avaluació i la certificació de coneixements de la llengua catalana
3. La selecció de professorat de català a les universitats per a garantir la continuïtat de la docència
4. La justificació de les subvencions atorgades a les universitats, també anomenades: memòries

Totes les dades de les diferents àrees anteriorment comentades es troben guardades en diverses bases de dades internes de l'IRL. L'objectiu del projecte ha estat explorar-les per a què els usuaris interns de l'Institut puguin dur a terme accions basades en evidències. Aquesta anàlisi també els ha permès acceptar o rebutjar totes aquelles hipòtesis realitzades prenent com a base l'experiència.

Amb aquest propòsit s'ha dut a terme set guions, un per a cada conjunt de dades, que contenen tots els passos necessaris per a l'obtenció de les gràfiques desitjades. Aquestes visualitzacions han estat programades amb la llibreria interactiva de Python Altair. Tot el codi ha estat fet a través de l'entorn interactiu de l'aplicació Google Collaboratory.

Paraules clau: Institut Ramon Llull, cultura catalana, català, subvencions, estades lingüístiques, certificació, selecció, memòries, Àrea de Llengua i Universitats, Àrea de Creació, Àrea de Literatura i Pensament, gràfics, Python i Altair.

Resumen

El Instituto Ramon Llull (IRL) es un organismo público creado con el objetivo de promover el catalán en el exterior en un amplio abanico de ámbitos como el académico, las artes visuales o la traducción de textos; internacionalizando la cultura catalana. El Instituto está constituido por tres áreas de negocio:

1. El Área de Lengua y Universidades
2. El Área de Creación
3. El Área de Literatura y Pensamiento

Dentro de las tres se comparte la voluntad de estudiar el otorgamiento de subvenciones para financiar un gran rango de actividades que promocionan el catalán en todo el mundo. Además, dentro del Área de Lengua y Universidades también se tiene la voluntad de analizar:

1. La organización de estancias lingüísticas en territorios de habla catalana
2. La evaluación y la certificación de conocimientos de la lengua catalana
3. La selección de profesorado de catalán en las universidades para garantizar la continuidad de la docencia
4. La justificación de las subvenciones otorgadas a las universidades, también llamadas: memorias

Todos los datos de las diferentes áreas anteriormente comentadas se encuentran guardados en varias bases de datos internas del IRL. El objetivo del proyecto ha sido explorarlas para que los usuarios internos del Instituto puedan llevar a cabo acciones basadas en evidencias. Este análisis también les ha permitido aceptar o rechazar todas aquellas hipótesis realizadas tomando como base la experiencia.

Con este propósito se han realizado siete guiones, uno para cada conjunto de datos, que contienen todos los pasos necesarios para la obtención de las gráficas deseadas. Estas visualizaciones han sido programadas con la librería interactiva de Python Altair. Todo el código ha sido realizado a través del entorno interactivo de la aplicación Google Collaboratory.

Palabras clave: Instituto Ramon Llull, cultura catalana, catalán, subvenciones, estancias lingüísticas, certificación, selección, memorias, Área de Lengua y Universidades, Área de creación, Área de Literatura y Pensamiento, gráficos, Python y Altair.

Abstract

The Ramon Llull Institute (IRL) is a public entity created with the aim of promoting Catalan abroad in a wide range of fields such as academia, the visual arts and text translation; internationalizing Catalan culture. Specifically, it is composed of three major areas:

1. The Area of Language and Universities
2. The Creation Area
3. The Area of Literature and Thought

All three desire to study the concession of grants to finance a wide range of activities that promote Catalan around the world. In addition, in the Area of Language and Universities there is also the will to analyse:

1. The organization of language stays in Catalan-speaking territories
2. The evaluation and certification of knowledge of the Catalan language
3. The selection of Catalan teachers in universities to guarantee the continuity of teaching
4. The justification of the grants awarded to universities, also called: reports

All data from the distinct sections discussed above are stored in various internal IRL databases. The aim of the project has been to explore them so that internal users of the Institute could carry out evidence-based actions. This analysis has also allowed them to accept or reject all those hypotheses made based on experience.

To this end, seven scripts have been created, one for each dataset, containing all the steps required to obtain the desired graphs. These visualizations have been programmed with the Python Altair interactive library. All the code has been created through the interactive environment of the Google Collaboratory application.

Keywords: Ramon Llull Institute, Catalan culture, Catalan, grants, language stays, certification, selection, reports, Language and University Area, Creation Area, Literature and Thought Area, graphics, Python and Altair.

Agraïments

M'agradaria dedicar un moment a agrair a les persones que han fet possible aquest projecte. En primer lloc, a en Francesc Rey Micolau per haver acceptat la proposta per a ser el meu tutor del projecte. M'ha donat consells interessants i, sense les seves recomanacions i ajudes pràcticament instantànies, hauria sigut més difícil dur a terme aquest projecte.

En segon lloc, a Esther Coll Caldas, responsable de l'administració electrònica i gestió de la informació a l'Institut Ramon Llull, per a introduir-me en l'IRL i acompanyar-me en els primers passos.

En tercer lloc, a en Salvador Orteu, per ser incansable i respondre a les més de 10.000 preguntes fetes dia i nit.

També m'agradaria expressar el meu més sincer agraïment a altres treballadors de l'Institut Ramon Llull: Gemma Gil, Ignasi Massaguer i Victòria Oliva per les seves aportacions directes i assessorament en el camí.

En últim lloc, cal fer una menció especial a la meva família per la paciència que han tingut al llarg del desenvolupament del projecte i les desaparicions que han hagut d'aguantar.

És un error capital teoritzar abans de tenir dades.[1]

[1] Sherlock Holmes a "A scandal in Bohemia" – Sir Arthur Conan Doyle (1859 – 1930), escriptor.

Índex

1 Introducció

1.1 Introducció

Durant aquest projecte la Universitat Politècnica de Catalunya (UPC) [1] ha signat un conveni amb l'Institut Ramon Llull (IRL) [2] per a dur a terme una anàlisi estadística extensiva de diverses bases de dades d'aquest últim a través d'una estudiant del Grau en Ciència i Enginyeria de Dades (GCED) [3]. Aquesta anàlisi, explicada en les següents seccions amb més detall, té com a principal objectiu obtenir estadístiques rellevants que puguin donar lloc a la presa de decisió a partir de fets contrastats per part dels usuaris interns de l'Institut.

Il·lustració 1: Logotip de la UPC

Il·lustració 2: Logotip de l'IRL

Aquest conveni de cooperació entre l'IRL i la UPC va començar fa dos anys quan jo mateixa, Anna Patrícia Orteu, vaig realitzar unes pràctiques a l'Institut. Aquestes, em van fer veure que hi havia un gran potencial no analitzat referent a les dades amb què tractava cada dia, raó per la qual vaig proposar la realització d'aquest projecte, el qual va acabar prosperant.

L'Institut Ramon Llull [2] és un organisme públic creat l'any 2002 amb la finalitat de promoure el català a l'exterior en un ampli ventall d'àmbits com l'acadèmic, les arts visuals o la traducció de textos; internacionalitzant la cultura catalana. És degut a aquest objectiu que du a terme acords amb les universitats de l'exterior per tal de promoure-hi la docència d'estudis catalans i oferir suport. A més a més, també es relaciona amb associacions internacionals de catalanística per a estimular la recerca i els estudis avançats.

En paral·lel, d'acord amb el marc europeu comú de referència per a les llengües, realitza exàmens de català oficials a l'exterior per a poder acreditar el coneixement en diferents nivells. Finalment, l'Institut es preocupa de difondre la literatura catalana, no només assegurant la presència d'artistes de Catalunya i Balears en programes destacats de la creació contemporània internacional, sinó també fent-la present en festivals i fires de ressò mundial, entre d'altres.

Aquest gran ventall de seccions amb què treballa l'IRL va fer augmentar l'abast del projecte, el qual en un inici només pretenia analitzar les dades provinents de les subvencions que aquest atorga. El projecte es va augmentar doncs afegint les seccions d'inscripcions, certificació, selecció de professorat i memòries explicades més profundament en apartats posteriors. Sent-ne conscient d'aquest paper com a agent social i de servei públic l'IRL ha acceptat la realització d'aquesta anàlisi així com que aquesta es posi a mà de la societat. Aquestes dades han estat principalment analitzades i netejades amb Python [4] mentre s'ha utilitzat la llibreria Altair [5] per a realitzar les gràfiques a partir de les quals es duran a terme les decisions.

Tot seguit, cal comentar que el projecte es troba compost per 8 seccions. En la primera, la introducció, es podrà descobrir una petita descripció de sobre què tracta el projecte i quines han estat les raons per dur-lo a terme. Posteriorment, en la metodologia de treball es descriuen les

eines utilitzades i les tasques acomplertes. Tot seguit, es troba la part central del projecte, és a dir, la descripció de les dades emprades així com la neteja d'aquestes junt amb els gràfics creats. En últim lloc, es troben tres seccions complementàries: futurs passos que es podrien portar a cap amb l'objectiu de millorar l'abast del projecte o el resultat d'aquest, les referències usades per a realitzar-lo i els apèndixs.

Finalment, convé comentar breument que tot el codi, dades i resultats finals del projecte es poden visualitzar en un repositori de GitHub públic: https://github.com/AnnaPaty/Statistical-analysis-IRL.

1.2 Motivació

Com mencionat anteriorment, l'Institut Ramon Llull té l'objectiu principal d'internacionalitzar la cultura catalana promovent el català a l'exterior en un ampli ventall d'àmbits sent un dels més importants, l'acadèmic. Amb aquest propòsit, es va voler elaborar una anàlisi de les dades contingudes en els següents punts:

1. L'atorgament de les subvencions per a finançar activitats en tres sectors:
    a. L'àrea de llengua pretén reconèixer entitats acadèmiques que agrupen experts i estudiosos de la llengua i la cultura catalanes a diversos països. Aquesta té la meta de representar, fomentar i difondre la cultura catalana mitjançant la realització d'activitats diverses en els àmbits lingüístic, literari i cultural.
    b. L'àrea de creació pretén reconèixer la producció artística catalana i donar-la a conèixer a través de la difusió.
    c. L'àrea de literatura pretén reconèixer les obres així com les seves traduccions que esperen ser llegides pels lectors internacionals i donar-les a conèixer a través de la difusió.
2. L'organització d'estades lingüístiques en territoris de parla catalana.
3. L'avaluació i la certificació de coneixements de la llengua catalana.
4. La selecció de professorat de català a les universitats per a garantir la continuïtat de la docència.
5. La realització dels justificants de subvenció duts a terme per les universitats: les memòries.

Dins la primera secció, és a dir, de les subvencions, es té com a principal propòsit donar respostes a preguntes del tipus: "Quina és la subvenció que té més revocacions?" o "Quina és la subvenció més acreditada i quins elements podrien donar lloc a aquesta popularitat?". Mentre el ventall de preguntes s'estén de forma significativa pel cas de l'àrea de literatura en trobar-se aquestes lligades a les traduccions dels llibres i obres literàries, entre d'altres. Així doncs, els objectius es troben alineats amb els anteriorment mencionats afegint preguntes com "Quins són els llibres que es tradueixen més i a quins idiomes?".

A més a més, a l'hora de tractar amb les subvencions s'ha de tenir en compte que existeixes dos tipus molts diferenciats, que donarà lloc a fer diverses diferències entre aquestes. D'una banda, existeixen les subvencions de concurrència competitiva, que fan referència a aquelles on les persones físiques empresàries o persones jurídiques sol·liciten la subvenció dins d'una

convocatòria. Així doncs, per a poder ser beneficiàries d'aquestes subvencions han de complir els requisits establerts en les bases de la convocatòria i competir amb altres empreses o persones que també hagin sol·licitat l'ajuda econòmica. D'altra banda, les subvencions directes fan referència a aquelles sol·licituds que entitats importants realitzen a l'Institut Ramon Llull fora de cap convocatòria de manera independent. En aquest segon tipus els sol·licitants no han de competir amb altres persones o entitats per a l'obtenció de la subvenció.

Les altres 4 seccions es troben totes elles contingudes dins de l'àrea de llengua. Les dades de les anomenades estades lingüístiques en realitat contenen informació d'inscripcions, tant d'estades com de campus. En aquest cas interessarà saber elements com en quina universitat es duen a terme més inscripcions, si hi ha alguna raó perquè això sigui així o quina és l'activitat més ben puntuada i com es podria promocionar, si els interessa.

En tercer lloc, les dades de certificació són aquelles obtingudes a partir dels exàmens de català oferts per l'Institut. Dins d'aquest subtema, es voldrà analitzar, per exemple, si les notes assolides en aquests exàmens es poden relacionar amb les hores empleades en estudiar el català per part dels participants, la llengua materna o el lloc de realització de l'examen, entre d'altres.

En quart lloc, l'Institut en promoure l'ensenyament de la llengua, la literatura i la cultura catalanes a les universitats i altres centres d'estudi superior organitza anualment convocatòries de selecció de professorat d'estudis catalans a les universitats amb les quals col·labora per cobrir les vacants i garantir d'aquesta manera la continuïtat de la docència. Dins d'aquesta subsecció es voldrà saber, entre d'altres, quines són les universitats que necessiten cobrir vacants més repetidament o quines són les característiques que presenten més freqüentment els usuaris seleccionats.

Finalment, les memòries, que com dit anteriorment són justificants de subvenció, tenen com a mòbil principal demostrar que s'ha ensenyat català durant el període de la subvenció així com avaluar l'experiència. En aquest cas, l'anàlisi estarà més enfocada a saber quina és la universitat amb la qual el Llull té més relació i a descobrir, per exemple, si després de realitzar algun dels cursos els usuaris s'acaben presentant a alguna prova de certificació.

S'acabarà aquesta secció mencionant dos exemples molt clars que fan entendre la necessitat de dur a terme aquest projecte. D'una banda, pel que fa a l'àmbit de certificació, es volen portar a cap proves per tal que els estudiants de l'exterior puguin certificar el nivell de català. Però s'ha vist, tot i que no comprovat, que els últims anys ha augmentat molt l'afluència de persones catalanes que realitzen els exàmens, per culpa de les oposicions, que requereixen un títol. A més a més, es creu que a llocs de l'exterior els exàmens seran més fàcils i, per tant, els usuaris catalans se'n van a fora a fer els exàmens, traient els llocs als usuaris que el Llull realment vol examinar. Si totes aquestes hipòtesis es confirmessin i es poguessin mostrar als directius, s'hauria de considerar una manera de canviar el funcionament i redirigir el programa de certificació.

El segon exemple es troba relacionat amb les inscripcions als campus i estades. Durant l'any 2020 a causa de la COVID es van realitzar estades de forma virtual les quals van tenir una gran acollida. A més a més, durant aquestes es va comentar que alguns usuaris ja hi havien volgut assistir més d'un cop amb anterioritat, però que no ho havien pogut fer per temes de localització geogràfica. Arran d'aquest fet, es va mostrar un alt desig d'analitzar les procedències dels usuaris

que havien assistit a aquestes estades virtuals i mirar si realment valia la pena fer-ne més d'aquest tipus.

Aquests dos exemples formulats a partir d'hipòtesis que tenien els usuaris interns del Llull abans de començar el projecte són només una petitíssima part de tot el potencial que es podia extreure de les bases de dades d'aquest. Aquesta àmplia gamma de possibilitats que oferia aquest projecte va possibilitar que aquest es tirés endavant amb la supervisió i el vistiplau de tots els directius de l'Institut.

1.3 Objectiu

En aquesta secció es pretén deixar clar l'objectiu del projecte per a saber en tot moment cap on s'han de dirigir els esforços. Com mencionat adés el propòsit inicial és l'observació d'aquestes dades a partir de gràfics per a ajudar a treure estadístiques rellevants per tal que els usuaris interns puguin dur a terme accions basades en nombres comparables, repetibles i comprensibles; augmentant d'aquesta manera la importància de l'Institut.

Tot i que en un inici s'ha volgut analitzar tot allò que pogués ser rellevant, com el sexe, la procedència o el nombre d'usuaris apuntats a una activitat per any. Un cop s'han mostrat les primeres visualitzacions als usuaris interns de l'IRL, aquests han especificat allò que volien explorar de forma més concreta a partir de reunions realitzades al llarg del projecte. Perquè com John Tukey va dir, "El valor més gran d'una imatge és quan obliga a notar allò que mai esperàvem veure"[2]. A més a més, també s'ha tingut al llarg del projecte la voluntat de detectar elements que es podrien haver analitzat si s'haguessin recollit les dades per a comentar-ho amb els usuaris de l'IRL, la qual cosa milloraria qualsevol futura versió d'aquest projecte.

1.4 Glossari

Tot seguit es pot trobar una llista de tots aquells acrònims utilitzats durant el projecte per a poder-los consultar en cas de necessitat i facilitar l'enteniment d'aquest:

- IRL → Institut Ramon Llull
- BD → Base de dades
- FK → Foreign Key (clau forana)
- SQL → Structured Query Language (llenguatge d'interrogació estructurat)
- UPC → Universitat Politècnica de Catalunya
- GCED → Grau en Ciència i Enginyeria de Dades
- KB → Kilobyte

[2] John Tukey (1915 – 2000), estadístic.

2 Metodologia

Per a poder assolir els objectius del projecte dins del període requerit i amb la major qualitat possible ha estat necessari portar en tot moment un control de les diferents activitats i subactivitats que aquest engloba així com els recursos que necessitaven cadascuna d'elles. Per tal de fer-ho factible era aconsellable seguir algun tipus de metodologia. En aquest cas, s'ha decidit seguir una metodologia de tipus àgil [6] en permetre aquesta gestionar el projecte des d'una perspectiva incremental i iterativa on no és necessari que els propòsits es trobin 100% establerts des del primer moment.

Entre les múltiples metodologies de tipus àgil en gestió de projectes que existeixen, s'ha decidit seguir el mètode Scrum [7]. Aquest té com a principal finalitat aplicar de forma regular un conjunt de bones pràctiques per treballar de forma col·laborativa, és a dir, en equip i obtenir el millor resultat possible.

Tot i que és cert que en aquest projecte no es necessita una coordinació estreta entre diferents persones, un cop aclarits els objectius per part de l'IRL, la resta d'elements clau que conformen aquesta metodologia han ajudat a tirar endavant el projecte de forma organitzada. Així doncs, no només han assistit en obtenir uns resultats periòdics i prou immediats per a poder mostrar als usuaris de l'IRL, sinó també a adaptar-se d'una manera ràpida i organitzada als canvis i evolucions que el destí del projecte ha patit.

El procés s'ha iniciat amb la realització d'un calendari inicial on es van planificar totes les fites de les diferents activitats i les dependències que existeixen entre aquestes. La primera versió d'aquest va ser presentat dins la proposta de projecte i pla de treball inicial (Annex 1: Document "Fita Inicial") el qual va ser validat amb el tutor. Aquests primers passos van confirmar l'adequació de les fites a realitzar i van permetre identificar els riscos a tenir presents durant el cicle del projecte.

Finalment, abans de detallar amb més precisió els passos duts a terme per a la realització del treball, convé comentar que també s'ha usat la metodologia CRISP-DM per a la correcta creació d'aquests passos així com per ajudar a seguir unes bones pràctiques adequades. La metodologia CRISP-DM [18] [19] o procés estàndard de la indústria transversal per a la mineria de dades, és el model més emprat en l'anàlisi de dades que descriu els enfocaments utilitzats pels experts en mineria. Aquest és format per sis fases: comprensió empresarial, comprensió de dades, preparació de dades, modelatge, avaluació i disposició, no estrictament seqüencials. En ser aquest projecte un treball que no requeria la creació d'un model, no tots els passos han estat portats a terme, però sí que ha estat altament útil tant la comprensió com la preparació de dades del mètode.

2.1 Pla de treball

A continuació es mostrarà en detall quina ha estat l'organització d'aquest projecte. En concret es podrà observar quines han estat les activitats, les dependències entre aquestes i les fites. També es definiran les eines utilitzades per a realitzar la comunicació i la planificació de les reunions.

Pel que fa a les diverses tasques dutes a terme, convé comentar que el document de seguiment (Annex 2: Document "Informe de seguiment") fet a la meitat del projecte va ajudar molt a la planificació de la segona part, sobretot degut a la introducció de tota la secció de selecció de professorat en una fase ja avançada, com es pot llegir a l'annex.

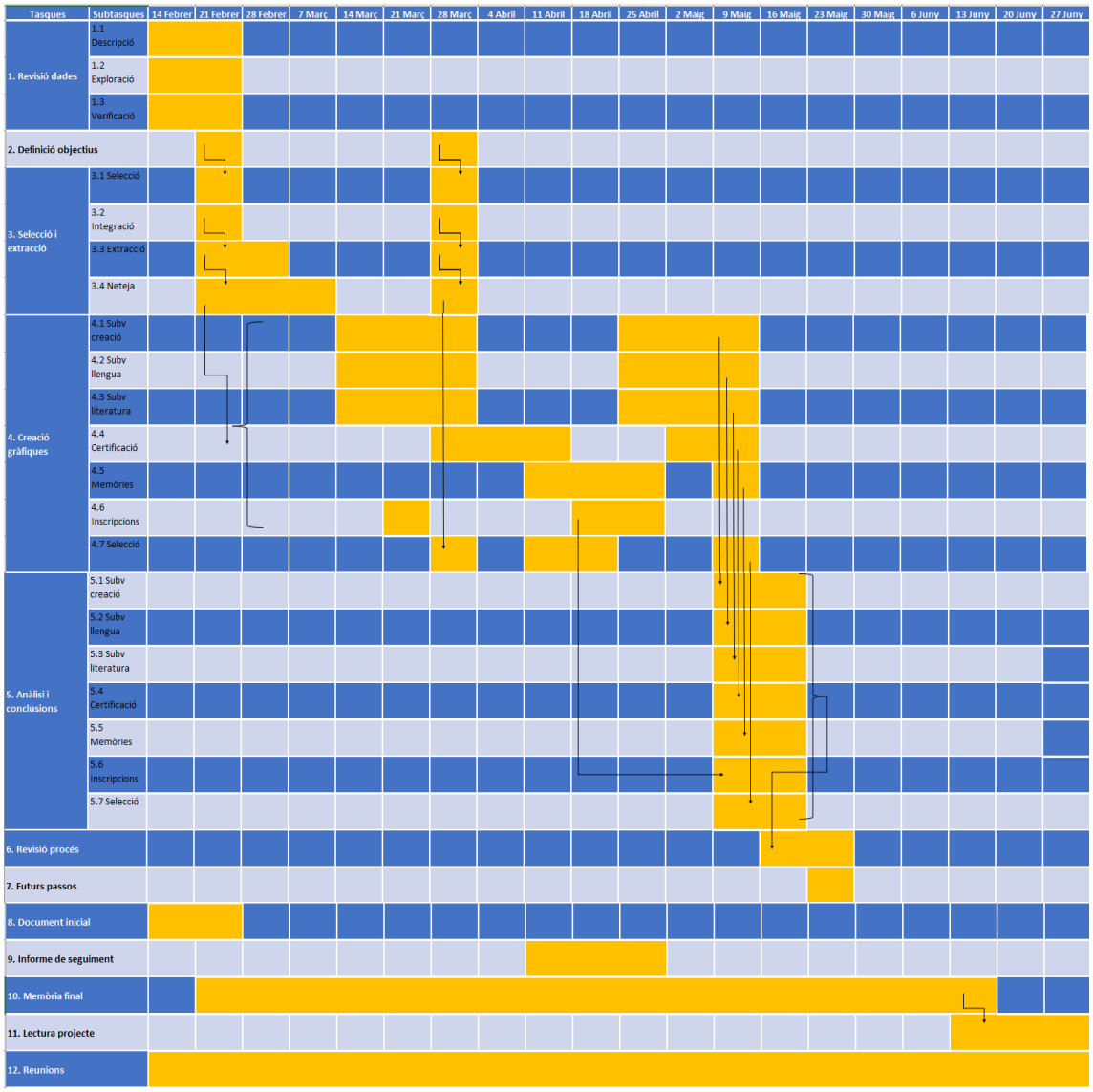
2.1.1 Tasques i fites. Diagrama de Gantt

El projecte es troba compost per 12 grans tasques i 3 fites o revisions del projecte. Aquestes 3 fites coincideixen amb l'entrega dels 3 documents importants que s'han hagut de lliurar al llarg d'aquest: la fita inicial i l'informe de seguiment, prèviament mencionats i, aquesta mateixa memòria. D'altra banda, cada tasca s'ha dividit en les subtasques necessàries per a poder tancar temes a un ritme acurat però no haver de realitzar un sobreesforç de seguiment del projecte. Aquestes es poden veure a continuació de forma visual en el digrama de Gantt (il·lustració 3).

En aquest es pot observar clarament les principals dependències del projecte:

1. La necessitat d'esperar a tenir els objectius clars per a seleccionar i extreure les dades amb què s'ha treballat al llarg de tot el projecte.
2. L'exigència d'efectivament extreure les dades abans de netejar-les.
3. No poder començar a dur a terme les gràfiques fins que les dades es trobessin completament netes.
4. Haver descrit les gràfiques per a saber quines visualitzacions podrien ser millorades, introduïdes o eliminades.

A més a més, gràcies a aquest diagrama també es pot afirmar que mentre portar a cap les visualitzacions de les dades de selecció i inscripcions va ser relativament senzill: només es van necessitar entre 3 i 4 setmanes de treball discontinu en el temps. Les dades de subvencions i certificació van donar lloc a més problemes i, per tant, a una despesa de temps major.



Il·lustració 3: Diagrama Gantt del projecte

2.1.2 Pla de reunions i comunicació

Per tal d'acomplir el projecte, s'ha comptat amb l'ajuda del personal de l'Institut Ramon Llull i de Francesc Rey Micolau, tutor d'aquest treball de fi de grau. D'una banda, la comunicació amb el tutor de la universitat no es trobava definida de forma clara. Simplement s'ha celebrat reunions en els moments indicats davant de qualsevol dubte que es pogués tenir.

D'altra banda, la comunicació amb l'IRL ha estat més constant. En un inici es va plantejar dur a terme una reunió cada dues o tres setmanes amb els caps de les àrees per a poder detectar elements ja prou analitzats i nous focus d'anàlisi. Tanmateix, a causa de la poca disponibilitat dels usuaris del Llull no va poder ser així. D'una banda, només es va poder parlar dues vegades amb l'àrea de literatura: una per a definir els objectius i l'altre per a detectar millores, les quals ja van ser suficients. Pel que fa a l'àrea de creació no s'ha pogut dur a terme cap reunió, raó per la qual totes les visualitzacions d'aquesta àrea s'han millorat a partir de les observacions fetes pels usuaris de les altres àrees.

L'àrea de llengua ha sigut amb la que ha existit una major comunicació duent a terme aproximadament una reunió setmanal un cop passat el primer mes del projecte. Com que pràcticament tots els focus d'aquest projecte pertanyien a aquesta àrea les reunions portades a cap van ser suficients i en tot moment útils. A més a més, al llarg de tot el projecte s'ha mantingut una comunicació pràcticament diària amb els desenvolupadors de les bases de dades utilitzades.

2.1.3 Eines emprades

D'una banda, per a la comunicació amb el tutor de la universitat s'ha usat el correu per a dubtes puntuals i el Google Meet per a solucionar temes més extensos. En segon lloc, per a la comunicació amb els desenvolupadors de la base de dades s'ha emprat Skype per a facilitar la constant i àgil comunicació. En últim lloc, s'ha fet ús del correu i el Microsoft Teams per a parlar amb els caps de les àrees de l'IRL, per adaptar-se a les seves necessitats.

3 Dades

Com explicat anteriorment, l'objectiu del projecte ha estat analitzar les dades de les subvencions, inscripcions, certificacions, seleccions de professorat i memòries de l'IRL. Per a fer-ho, ha estat necessari disposar d'uns conjunts de dades sobre els quals poder realitzar les consultes necessàries. Així doncs, en aquesta secció es detallarà quin ha estat el procés dut a terme per a obtenir aquests 7 conjunts de dades.

En concret es podrà trobar en un inici la fase de col·lecció de dades on es contarà de quines bases de dades s'ha extret la informació i com. S'hi inclou tant la selecció de característiques com la posterior integració i extracció. En segon lloc, es considerarà la descripció i exploració de les dades intermèdies. Aquesta és una fase de comprensió de les dades que permet familiaritzar-se amb aquestes, descobrir primers coneixements sobre les dades i detectar subconjunts interessants per formar hipòtesis sobre la informació oculta. Al llarg d'aquesta també s'ha entès alguns dels objectius de l'IRL així com s'ha detectat certes limitacions.

En tercer lloc, la neteja de dades serà duta a terme. En aquesta fase, a part d'identificar problemes de qualitat de les dades també es comprovarà qualsevol restricció inherent a una variable específica. En quart lloc, s'explicarà la construcció de noves variables. En últim lloc, es durà a terme una descripció de les dades finals sobre les quals es realitzaran les visualitzacions.

Cal tenir en compte que tot i haver plantejat aquests passos de forma seqüencial, varis d'ells s'han hagut de dur a terme diverses vegades a causa de la naturalesa iterativa del procés de selecció de dades i obtenció de conjunts completament analitzables.

3.1 Col·lecció de les dades inicials

L'Institut Ramon Llull disposa de 5 bases de dades (BD) principals. Dues d'aquestes, contenen informació no necessària per al projecte i, per tant, seran obviades. D'altra banda, en primer lloc, la base de dades Oracle conté les dades de les diferents subvencions i inscripcions. En segon lloc, la base de dades Ovirtllull conté les dades sobre les memòries i seleccions de professorat. Finalment, la de Certificació conté la informació sobre les certificacions.

Tanmateix, aquestes bases de dades prèviament mencionades contenen moltes altres taules sobre altres elements no analitzats. És doncs per aquesta raó que posteriorment s'explica quins han estat els passos per a seleccionar les taules i atributs possiblement interessants per a l'anàlisi, com s'han ajuntat i finalment extret.

3.1.1 Selecció de les dades

En aquest primer pas s'ha seleccionat la taula centre de cada tema. A continuació, s'ha mirat quines taules es trobaven relacionades amb aquesta que poguessin aportar informació rellevant i finalment s'ha seleccionat els atributs interessants de cadascuna d'aquestes.

Els criteris utilitzats per a la selecció inclouen la rellevància de les variables per a la posterior anàlisi, la qualitat i les limitacions tècniques, com ara el tipus de dades.

3.1.1.1 Subvencions

Tot i haver acabat formant 3 conjunts de dades diferents, un per a cada àrea, la primera selecció es va dur a terme de forma conjunta i, per tant, s'explicaran tots tres a la vegada. En primer lloc, se sabia que la taula *tsubvencions* de la BD Oracle era la base d'aquestes. A partir d'aquí, es va mirar quins eren els ID que apareixien a la taula *tsubvencions* que feien referència a alguna altra taula (FK) que pogués aportar informació rellevant. Aquest procés de mirar ID d'altres taules es va dur a terme de forma recursiva. Realitzant-lo, es va acabar obtenint relacions amb 12 taules diferents, que es poden visualitzar a l'Annex 3.1: Taules usades - Subvencions.

Tot seguit, es va anar seleccionant variable per variable de cada taula, comprovant si aportava informació rellevant i tenia suficients registres per a poder aportar estadístiques prou confiables. Tanmateix, després de dur a terme una primera extracció, es va notar que algunes variables inicialment cregudes interessants no eren necessàries raó per la qual es van eliminar permetent fins i tot reduir el nombre de taules a les quals s'havia de mirar. Aquesta segona anàlisi que va permetre eliminar variables ja es va fer de forma única per a cada àrea, raó per la qual la variable "*titelles*", per exemple, es troba en el conjunt de dades de creació però no en el de llengua.

Continuant amb les diferències entre les diferents àrees també cal comentar que moltes subvencions de literatura estan relacionades amb la traducció d'alguna obra o llibre en català a altres llengües. Això, va fer que aquestes subvencions en particular s'haguessin de vincular a 7 altres taules, que es poden observar altre cop en l'annex mencionat anteriorment.

D'altra banda, la necessitat d'haver d'introduir el nom dels diferents festivals així com els països on es van celebrar aquests en les subvencions de creació, va portar a la necessitat d'utilitzar la vista *v_nom_festival*.

Tot seguit, cal esmentar que els registres de les subvencions de les àrees es diferencien entre ells gràcies a una variable molt important anomenada "*seguretat*" que es troba a la taula *tsubvencions*. Si aquesta pren el valor 1 significa que és una subvenció de l'àrea de creació, si pren el valor 3, fa referència a una subvenció de l'àrea de llengua; mentre el 5 indica les de literatura. Les consultes finals d'aquestes extraccions es poden observar al GitHub, en concret als fitxers anomenats select_subv_llengua.sql, select_subv_creacio.sql i select_subv_lite.sql.

Abans d'acabar amb aquesta subsecció, convé comentar que l'àrea de llengua un cop analitzades les visualitzacions referents a les subvencions, també va voler saber les diferents relacions que havia tingut amb les entitats al llarg dels anys. Tanmateix, aquesta informació no es va poder extreure a partir del conjunt de dades explicat anteriorment perquè una relació no implica en tot moment un intercanvi de diners i, per tant, no totes les relacions es troben implícites a la taula base *tsubvencions*.

Encara que és cert que la informació no es considera doncs completament relacionada amb les subvencions, l'anàlisi d'aquesta nova petició afecta de forma directa a aquestes i, en conseqüència, es va voler mantenir dins d'aquest apartat. Aquesta petició va dur a la realització

d'una nova extracció de dades que en aquest cas té *tany_activitat* com a taula principal i 5 entitats més lligades a aquesta. El nom del fitxer que conté aquesta consulta és select_subv_evolucioUnis.sql.

3.1.1.2 Inscripcions

En segon lloc, com mencionat anteriorment les dades que fan referència a les inscripcions també es troben compreses a la BD Oracle. Tanmateix, en aquest cas la taula principal s'anomena *tinscripcions*. Aquesta al seu torn és complementada per dades contingudes a 7 taules observables a l'Annex 3.2: Taules usades – Inscripcions. En aquest cas, no s'ha eliminat cap taula a posterioritat, tot i que sí que s'ha eliminat diverses variables. La consulta exacta es pot observar altre cop al GitHub del projecte al fitxer anomenat select_inscripcions.sql.

3.1.1.3 Certificació

En tercer lloc, les dades de certificacions són emmagatzemades a la BD Certificació. En aquest cas, la taula origen és *clc_examens* i es troba relacionada amb 7 altres taules, explicades a l'Annex 3.3: Taules usades - Certificació. En algun moment també es va plantejar extreure més informació sobre les certificacions de la taula *clc_diplomes* i les relacionades amb aquesta. Tanmateix, després de parlar amb els desenvolupadors de la BD es va arribar a la conclusió que aquestes no aportarien més informació rellevant i, per tant, es van deixar de marge. La consulta final utilitzada per a aquest conjunt es pot visualitzar al document anomenat select_certificacio.sql.

3.1.1.4 Selecció de professorat

En quart lloc, les dades de selecció de professorat són emmagatzemades a la BD Ovirtllull. En aquest cas, la taula origen és *sp_formulari* i es troba relacionada amb 10 altres taules explicades a l'Annex 3.4: Taules usades – Selecció de professorat. Totes les consultes usades per a aconseguir el conjunt de dades desitjat es pot observar al document anomenat select_seleccio.sql.

3.1.1.5 Memòries

Finalment, les dades sobre les memòries es poden obtenir de la BD Ovirtllull de l'IRL, igual que en el cas de selecció. En aquest cas la taula principal s'anomena *xov_memoria* i al seu torn es relaciona amb 4 taules descrites a l'Annex 3.5: Taules usades – Memòries. Respecte a aquestes dades, mencionar que hi ha diverses columnes que podrien haver aportat informació rellevant a l'anàlisi, però que no s'han inclòs a la selecció, a causa de la dificultat que implicava al ser variables de text lliure. Com en els anteriors casos la consulta final executada es pot trobar al GitHub, en concret al fitxer anomenat select_memories.sql.

3.1.2 Integració de les dades

En aquesta secció s'explicarà el mètode mitjançant el qual la informació de diverses taules o registres s'ha combinat per a crear registres o valors nous. Fusionar taules fa referència a unir dues o més taules que tenen informació diferent sobre els mateixos objectes. Les dades combinades també cobreixen les agregacions, les quals fan referència a les operacions en què es calculen nous valors resumint la informació de diversos registres i/o taules.

És cert que si s'han observat les consultes mencionades al llarg de l'apartat 3.1.1 Selecció de dades ja s'haurà pogut comprovar com s'ha dut a terme aquesta integració, tanmateix, a continuació es realitzarà l'explicació teòrica d'aquesta secció. En tots els casos s'ha utilitzat una consulta SELECT d'SQL com a base. L'SQL [8] [9] és un llenguatge estàndard de comunicació amb bases de dades relacionals. La seva principal característica és l'alta simplicitat d'aquest, ja que amb pocs coneixements es poden portar a cap consultes bàsiques, mantenint en tot moment la completesa tant a nivell relacional com computacional. D'altra banda, una consulta SELECT és aquella que permet consultar les dades emmagatzemades en una BD, és la paraula clau que indica que la sentència que es vol executar és de selecció. Aquest tipus de consulta es troba normalment acompanyada pels següents elements:

- FROM: indica la taula des de la qual es vol recuperar les dades.
- JOIN: permet realitzar la consulta combinada de varies taules. Aquest, tot i que no és necessàriament sempre així, en aquest projecte és precedit per un INNER si se sap que totes les files de la selecció tenen indicat l'identificador pel qual es fa la relació entre les taules (ON) o per un LEFT, altrament.
- WHERE: permet especificar quelcom condició que les dades han de complir per a ser retornades per la consulta. Aquest admet els operadors lògics AND i OR.
- GROUP BY: especifica una manera d'agrupar de dades.
- ORDER BY: especifica l'ordre amb què es vol que es retornin les dades.

A part dels elements bàsics d'una consulta de selecció, també s'han utilitzat les següents funcions al llarg del projecte:

- Declaració CASE: itera per les diferents condicions de la consulta i retorna un valor quan es compleix la primera condició (com una sentència if-then-else). Si no hi ha condicions certes, retorna el valor de la clàusula ELSE. Si no hi ha ELSE, retorna NULL.
- SELECT DISTINCT: retorna només valors diferents.
- STRING_AGG: concatena els valors de les expressions de cadena i col·loca separadors entre ells. El separador no s'afegeix al final de la cadena.
- TO_CHAR: converteix una data o un número a una cadena donat un format.
- GROUP_CONCAT: retorna una cadena amb un valor no NULL concatenat d'un grup. Retorna NULL quan no hi ha valors que no siguin NULL.
- CAST: converteix un valor de qualsevol classe a un altre tipus de dades especificat. Per exemple, en el cas del projecte s'ha utilitzat en la consulta de les memòries per a convertir la variable "valor" a un nombre amb exactament 2 decimals.
- CREATE TABLE: crea una nova taula a la base de dades.

A més a més, per a facilitar les sentències se'ls ha donat un nou àlies a la majoria de taules i variables amb què s'ha tractat.

Tot seguit convé comentar el cas especial de la consulta de memòries. La majoria dels JOINs realitzats simplement han permès afegir informació de diverses taules a una mateixa fila. D'aquesta manera si existeixen, per exemple, 1000 files dins de la taula *tinscripcions*, després de tots els JOINs i la subseqüent introducció d'informació s'haurien obtingut el mateix nombre de files amb més columnes. Tanmateix, això no és així per a les memòries, la qual presenta dues peculiaritats.

D'una banda, una memòria es troba composta per n produccions, m activitats i p assignatures. No obstant això, la informació d'aquestes taules no s'ha concatenat com a noves columnes al final d'una mateixa tupla per culpa de la ineficiència de l'operació. Així doncs, la variable "*idmemoria*" (id de la taula *xov_memoria*) no identifica de manera única una tupla, sinó que aquesta es troba n*m*p cops en el conjunt de dades resultant en un augment considerable del nombre de tuples amb les quals tractar, dificultant l'anàlisi en general.

L'altra peculiaritat que presenta aquesta consulta fa referència a la informació provinent de la taula *xov_puntuacio*. Aquesta altre cop conté múltiples tuples per a una única memòria el qual donaria lloc a augmentar encara més la multiplicitat. Així i tot, en aquest cas sí que s'ha concatenat la informació de les diferents línies com a columnes en comptes de com a noves files. Per a fer-ho s'ha dut a terme una pivotació de la informació a força d'introduir tants JOINs com files hi ha a la taula *xov_puntuacions* referides a 1 única memòria. La pivotació ha estat facilitada per l'existència del camp "*concepte*", que indica la característica puntuada.

Finalment, cal especificar les peculiaritats de la consulta de selecció. Moltes de les taules d'aquesta van haver d'ajuntar-se en vistes per a obtenir una informació compacta i poder facilitar el SELECT final. També es va pivotar en tots els casos necessaris la informació de files a columnes per a tenir en tot cas una única fila que contingués tota la informació sobre una sol·licitud. Finalment, per a augmentar l'eficiència, per culpa de l'alt nombre de pivotacions que es van realitzar en el cas de les taules *sp_valor**, es va crear una taula amb un subconjunt de les dades finals que es volien extreure.

3.1.3 Extracció de les dades

Com mencionat al principi de tot d'aquest secció existeixen 3 bases de dades des de les quals s'ha extret informació, és a dir, 3 llocs diferents on s'han hagut d'executar les consultes explicades anteriorment per a obtenir els Excels desitjats sobre els quals poder treballar. La primera BD, Oracle, es troba configurada dins d'un programa anomenat MobaXterm [10], mentre tant Certificació com Ovirtllull es troben configurats al HeidiSQL [11].

3.1.3.1 MobaXterm

MobaXterm [10] és una aplicació de Windows que ofereix un munt de funcions adaptades per a programadors, administradors web, administradors d'IT i pràcticament tots els usuaris que necessiten gestionar els seus treballs remots d'una manera senzilla. L'aplicació consta principalment d'un terminal on es poden executar sentències Linux i SQL. Amb aquest s'ha realitzat l'extracció dels conjunts de dades dels 3 tipus de subvencions (subv_llengua.xlsx,

subv_creacio.xlsx i subv_lite.xlsx), la consulta adherida a les subvencions de llengua (subv_evolucioUnis.xlsx) i de les inscripcions (inscripcio.xlsx).

A continuació s'enumeraran els passos que cal portar a cap per a dur a terme les extraccions:

1. Primer de tot s'executa: `rlwrap -c sqlplus master@gestio.llull.intranet` per a entrar a Oracle i poder executar sentències SQL. Aquesta sentència requereix acompanyar-la de la contrasenya corresponent.
2. Per a redirigir la sortida del SELECT al document "result" i que no surti per la terminal s'introdueix en aquesta:

   ```
   set lines 32500;
   set pages 0;
   set trimspool on;
   spool result;
   ```

   Aquest conjunt de línies també milloren l'aparença de les dades seleccionades en el document obtingut.
3. S'executen les consultes (una cada cop).
4. Es para la redirecció del terminal al document executant: `spool off;`
5. S'obre el document "result" i es copien les files resultants a un Excel.
6. Es divideixen les files copiades únicament a la columna A en diverses columnes dient a l'Excel que divideixi en funció del caràcter "|". Aquest pas fa entendre el perquè la separació entre variables en aquests SELECT's és `||'|'||` i no una simple coma.
7. S'introdueix el nom de les columnes a la primera fila de l'Excel.

Al final s'aconsegueixen doncs els Excels .xlsx corresponents que es podran obrir amb Google Collaboratory utilitzant Python. La posterior exploració i neteja de dades s'ha dut a terme a partir d'aquests.

3.1.3.2 HeidiSQL

D'altra banda, el HeidiSQL [11] és un programa lliure que té l'objectiu de ser fàcil d'aprendre. Aquest programa, permet veure i editar dades i estructures des d'ordinadors que executen un dels sistemes de bases de dades MariaDB, MySQL, Microsoft SQL, PostgreSQL i SQLite. Dins d'aquest es troben configurades les bases de dades Ovirtllull i Certificació que han permès obtenir els Excels de memòries (memories.xslx), selecció de professorats (seleccio.xlsx) i certificació (certificacio.xlsx).

Tot seguit es poden veure els passos realitzats per a l'obtenció d'aquests:

1. S'executa la consulta (una per una) a la terminal del programa.
2. S'exporta el retorn en format .csv, incloent-hi els encapçalaments i indicant els espais buits amb els caràcters "\t".
3. S'obre l'Excel en format .csv, s'eliminen els "\t" i es converteix de .csv a .xlsx.

Al final s'aconsegueixen altre cop els Excel .xlsx corresponents que es podran obrir amb Google Collaboratory utilitzant Python.

3.2 Descripció i exploració de les dades intermèdies

En aquesta secció s'examinaran les propietats "superficials" de les dades adquirides. Es descriuran les primeres troballes o hipòtesis inicials i el seu impacte en la resta del projecte. Entre les característiques analitzades es trobaran el format de les dades i la quantitat.

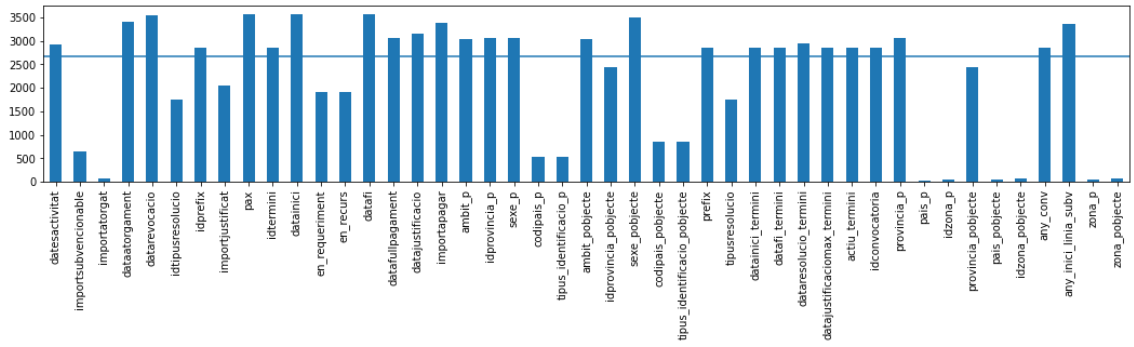
També serà el moment d'abordar el tema de si les dades adquirides compleixen els requisits pertinents mitjançant tècniques de consulta, visualització i informes, si fos necessari. Aquests requisits inclouen la distribució d'atributs clau, relacions entre parells de variables, propietats de subpoblacions significatives i anàlisis estadístiques simples. Aquestes anàlisis ajuden no només a saber les dades amb què s'està tractant sinó també a descobrir transformacions que s'ha de realitzar.

També, es detectaran variables que no és necessari analitzar o que no tenen prou informació per a ser rellevants, no detectades anteriorment. Aquestes han estat eliminades en la seva majoria de casos en els guions de Google Collaboratory (Python) que es poden observar al GitHub per a reflectir la feina feta. Tanmateix, en alguns casos s'han eliminat directament de les consultes inicials per a augmentar l'eficiència del procés en conjunt.

Finalment, cal comentar que tots i que tots els passos mencionats anteriorment s'han dut a terme en els guions de Python, a continuació només es podran trobar els elements més significatius d'aquesta exploració per a no exposar a l'usuari a informació massa detallada que podria fer pesada la lectura d'aquesta memòria.

3.2.1 Subvencions de llengua

Les dades d'aquest conjunt es troben compostes per 71 columnes i 3.578 entrades. Així mateix, hi ha 8 variables considerades importants: *idpersona*, *idestatsubvencio*, *idlinia_conv*, *idprefix*, *idtermini*, *pers_objecte*, *any_directe* i *concurrent*. 26 d'aquestes columnes superen el 75% de valors nuls per columna, com es pot veure amb la següent gràfica (on la línia marca el 75%):



Il·lustració 4: Nombre de valors nuls per columna en el conjunt de dades de subvencions de llengua

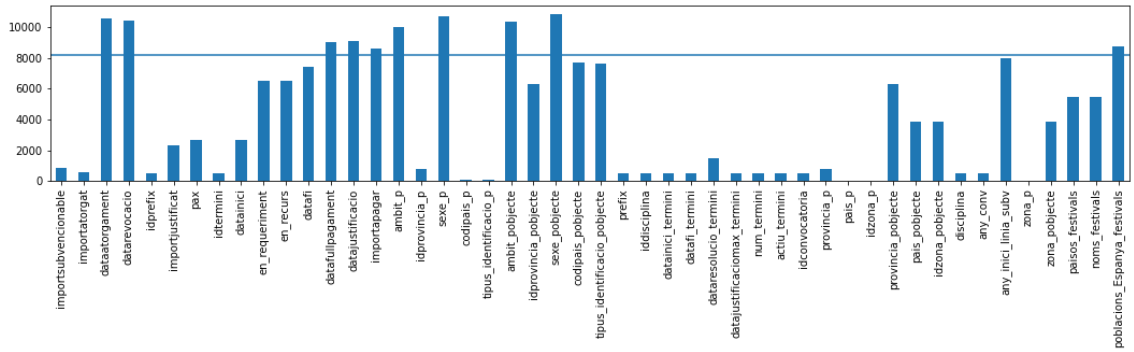
Tanmateix, només 14 files han estat eliminades en el primer apartat de l'anàlisi, en ser la resta de variables factors que precisament es volien analitzar encara que les estadístiques no fossin molt fiables. Aquestes eliminacions han dut a tenir 32 variables numèriques, 18 cadenes de text i 7 camps que contenen dates per a analitzar. A més a més les variables numèriques han estat dividides segons si eren categòriques (19), ID (8) o quantitatives (5). Les distribucions que

aquestes mostren, visualitzades en el cas de les variables categòriques amb histogrames, per als ID amb gràfics de punts i per a les quantitatives amb diagrames de caixes, es poden trobar al fitxer subv_llengua.html del GitHub, que conté el mateix que el fitxer subv_llengua.ipynb però de forma ja executada. Com a fet destacable d'aquesta exploració, es pot mencionar que els diagrames de caixes observats per a les variables quantitatives ja posaven de manifest que comprovar aquestes variables seria un dels passos més complicats d'aquest conjunt, com finalment ha succeït.

D'altra banda, cal recordar que per a aquesta àrea dins les subvencions s'havia extret un segon conjunt de dades per a comptar les relacions que havia tingut l'àrea amb les diferents entitats al llarg dels anys, encara que no comportessin un intercanvi de diners. Aquest és conformat per 8 columnes i 6.731 entrades. Dues d'aquestes variables són ID, una d'elles indica l'any i les altres cinc són cadenes de text. En aquest cas és important declarar que la coneixença que les dades són correctes ha evitat haver de realitzar un preprocessament sobre aquestes abans d'utilitzar-se per a crear visualitzacions, raó per la qual no seran mencionades en l'apartat 3.3.1.1 Neteja de dades i comprovació de restriccions – Subvencions de llengua.

3.2.2 Subvencions de creació

El conjunt de les subvencions de creació es troba compost per 78 columnes i 10.960 entrades. Les variables més importants tornen a ser: *idpersona, idestatsubvencio, idlinia_conv, idprefix, idtermini, pers_objecte, any_directe* i *concurrent*. Tanmateix, en aquest cas només 10 variables superen el 75% de valors nuls, com es pot veure en la il·lustració 5.



Il·lustració 5: Nombre de valors nuls per columna en el conjunt de dades de subvencions de creació

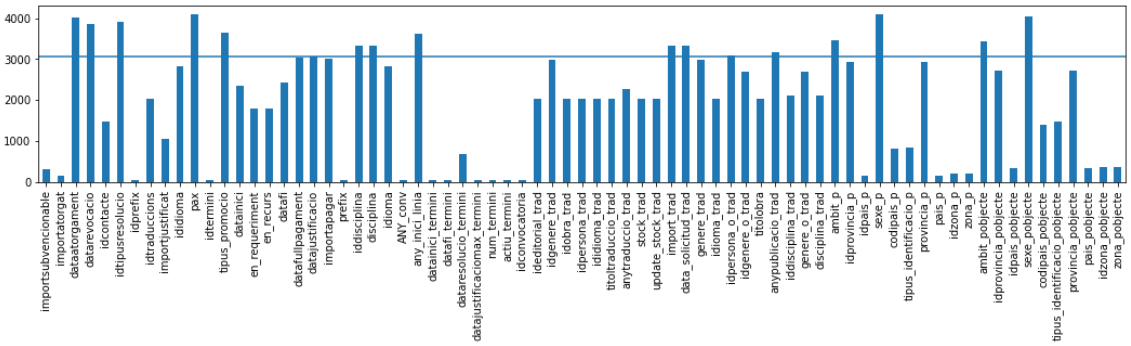
Després de dur a terme les eliminacions necessàries han acabat quedant 9 variables conformades per dates, 23 per cadenes de text i 39 per camps numèrics a analitzar. A més a més, per a dur a terme una primera visualització dels nombres aquests s'han dividit altre cop en 24 variables categòriques, 10 ID i 5 variables quantitatives, totes elles amb un gran nombre d'anomalies.

Com a observació especial en aquest cas dir que la variable poblacions_Espanya_festival, tot i ser una de les més importants per als usuaris interns de l'IRL, en aportar la informació d'on es realitzen les mobilitats per a les actuacions, semblava que seria una de les més complicades per a analitzar i així ha estat. Aquesta pot contenir diverses poblacions dins d'una mateixa tupla, a causa que una entitat només demana una subvenció per festival, encara que aquest es dugui a

terme en diferents localitzacions. A més a més, se sap que es vol saber quines han estat les poblacions més visitades per línies i anys de convocatòria, raó per la qual l'organització actual de les dades no és la més correcta, sent molt millor tenir una població per fila. Aquestes dues observacions han acabat duent a la necessitat de crear un nou conjunt de dades per a codificar la informació d'aquestes poblacions per línies i anys.

3.2.3 Subvencions de literatura

En tercer lloc pel que fa a les subvencions, el conjunt de dades de literatura al tenir informació sobre les obres traduïdes, augmenta el nombre de variables fins a 101 mentre el nombre d'entrades baixa respecte al cas anterior a 4.130. Les variables més importants tornen a ser les mateixes 8 que les mencionades en els dos casos anteriors, però en aquest cas *idtraduccions* s'afegeix a la llista. Així mateix, com s'observa a la figura 6 fins a 15 variables superen la línia horitzontal del 75% de nuls.

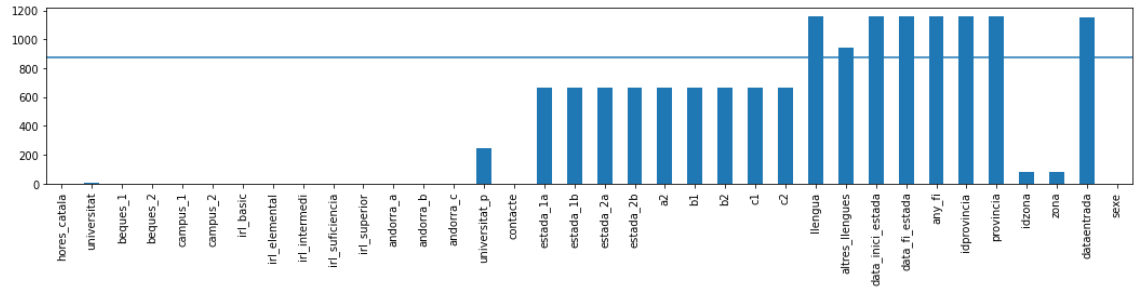


Il·lustració 6: Nombre de valors nuls per columna en el conjunt de dades de subvencions de literatura

Després dels processos d'eliminació s'ha tractat amb 13 variables de tipus data, 28 cadenes de text, 26 categòriques, 17 ID i 7 quantitatives. Un dels fets més destacables de l'exploració ha estat l'alt percentatge de valors incorrectes que s'han mostrat en les variables que expressaven una unitat temporal. Així mateix, una exploració més profunda dels imports també ha permès veure que els estats de les subvencions no es corresponien amb aquests, raó per la qual s'ha analitzat cas per cas. Aquesta última observació portada a cap s'ha notat en realitat per a tots els conjunts de subvencions.

3.2.4 Inscripcions

En aquest cas són 60 columnes i 1.167 files les que formen part de l'Excel, mentre el nombre de variables significatives són 3: *idpersona, idestada* i *any_estada*. A més a més, com es pot veure en la figura 7 el nombre de variables amb més del 75% de valors nuls s'ha reduït a 8, totes elles eliminables per la seva poca importància pel que fa a l'anàlisi general dels campus i les estades. Al final s'ha hagut d'analitzar doncs 12 cadenes de text, 1 data, 4 ID i 35 variables categòriques.

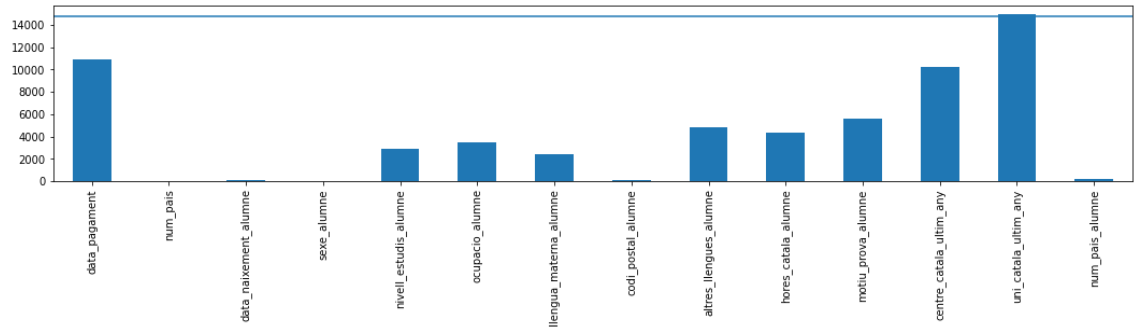


Il·lustració 7: Nombre de valors nuls per columna en el conjunt de dades d'inscripcions

Com a comentari rellevant, es pot dir que l'exploració de les variables que contenen noms d'universitats dins d'elles són les que van preocupar més al preveure que comportarien diversos problemes de cara a la uniformització, el qual s'ha complert. L'alta variabilitat d'aquestes variables ve donada perquè són totes dues cadenes de text agafades directament d'un formulari que es pot omplir de manera lliure de tal manera que la "Universitat de Barcelona", per exemple, es pot haver escrit com a "UB", "Universitat de Barcelona" o "Universitat de Barcelona (UB)", entre d'altres.

3.2.5 Certificació

El conjunt de dades referent a l'avaluació i la certificació de coneixements de la llengua catalana es troba compost en un inici per 62 columnes i 19.762 files. 8 d'aquestes columnes, totes ID i sense cap valor nul en cap fila, han estat les seleccionades com a importants: *id_examen, id_nivell, id_conv_seus, id_convocatoria, id_conv_seus_nivells, id_seus, id_pais* i *id_alumnes*. Tot seguit amb la visualització 8 es pot comprovar que és el conjunt de dades més complet al tenir només 1 columna que supera el 75% de valors nuls, tot i que no ha estat eliminada en ser un dels factors que es vol analitzar.

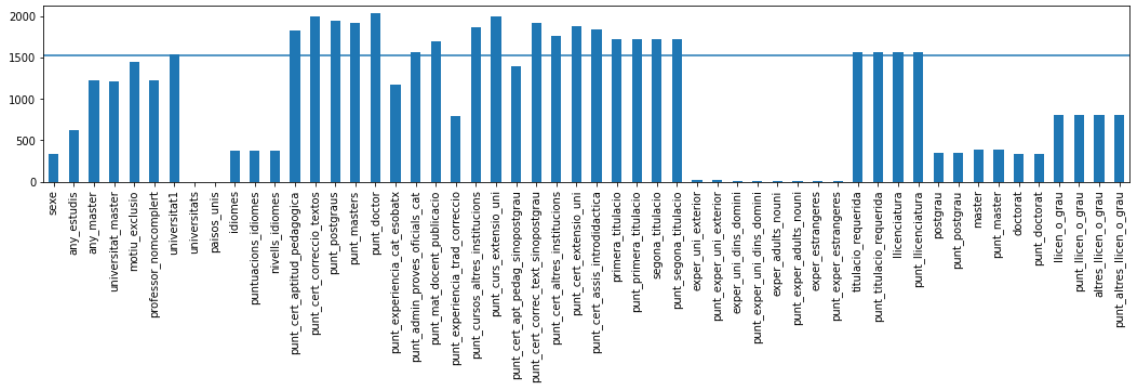


Il·lustració 8: Nombre de valors nuls per columna en el conjunt de dades de certificacions

En aquest cas la divisió en nombres, dates i cadenes de text resulta en 38, 8 i 20 variables, respectivament. D'altra banda, a l'hora d'analitzar els nombres en concret es tenen 19 variables categòriques, 11 ID i 9 quantitatives. És l'únic conjunt de dades que no ha mostrat cap anomalia pel que fa a les variables quantitatives, tanmateix, junt amb les memòries, semblava que seria un dels més difícils d'uniformitzar pel que fa a les variables de text, per culpa dels múltiples valors que prenien aquestes, i així ha estat.

3.2.6 Selecció de professorat

El conjunt de selecció de professors es troba compost per 81 columnes i 2.174 entrades, sent doncs el segon amb menys files inicials. També és un dels conjunts amb menys variables importants al ser només 3 d'elles essencials: *id_formulari, nom* i *id_convocatoria*. No obstant això, mostra un alt nombre de variables amb algun nul. En concret 21 d'aquestes superen el 75% de valors nuls per variable:



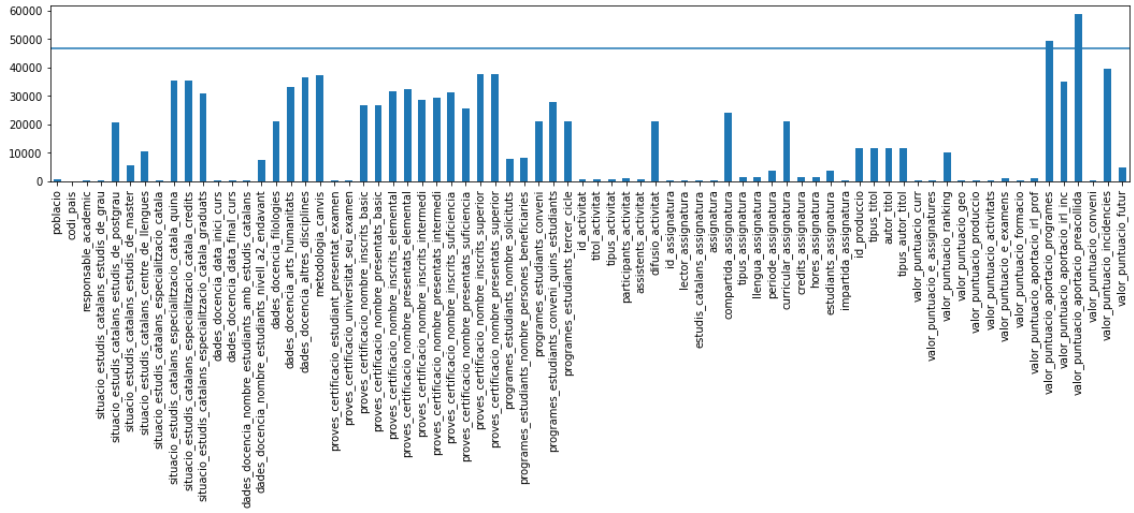
Il·lustració 9: Nombre de valors nuls per columna en el conjunt de dades de selecció de professorat

Pel que fa a les categories 3 d'aquestes variables són dates, 27 cadenes de text i 51 nombres. Així mateix, dins d'aquests nombres 2 són ID, 16 categòrics i 33 quantitatius. Com a observacions cal comentar que se sap que gràcies a la correcta extracció de les dades, les columnes que es jutgen relacionades concorden a la perfecció. Aquestes són les variables *idioma** amb *nivell_idioma** i *punt_idioma** i, d'altra banda, *uni_opcio** amb *pais_uni_opcio**. Aquest fet permet afirmar amb total seguretat que serà més fàcil tant netejar-les com visualitzar-les.

3.2.7 Memòries

En últim lloc el conjunt de dades de memòries a partir de l'exploració inicial es postula com el més difícil de netejar. Aquest és compost per 98 columnes i 62.370 entrades, el qual complica qualsevol operació que es vulgui realitzar sobre aquest al ser més costosa en l'àmbit computacional. D'altra banda, consta de 7 variables importants: *id, id_persona, universitat, any_academic, id_activitat, id_assignatura* i *id_produccio*. Així mateix, només 2 variables superen el 75% de valors nuls, com es pot veure a la figura 10.

S'ha acabat analitzant 2 variables compostes per dates, 28 per cadenes de text, 24 quantitatives, 5 ID i 35 categories numèriques. L'alt nombre de variables numèriques ha facilitat l'anàlisi al ser més fàcil comprovar valors numèrics que cadenes de text que poden prendre massa categories similars, a vegades no identificables. Les variables quantitatives han estat les que han mostrat més valors anòmals amb un gran marge en dependre d'activitats acomplertes de forma anual per les universitats, les quals varien dins d'un gran rang de valors, no només entre universitats sinó també al llarg dels anys. A part de la gran variabilitat pel que fa a les variables quantitatives, com mencionat anteriorment, les cadenes de text també han mostrat tenir diverses possibilitats per a una mateixa categoria. Un d'aquests exemples podria ser la variable *difusio_activitat* que a partir de text escrit de forma manual i lliure indica com ha estat difosa una activitat.



Il·lustració 10: Nombre de valors nuls per columna en el conjunt de dades de memòries

3.3 Neteja de dades i comprovació de restriccions

Durant aquesta secció de neteja s'ha examinat la qualitat de les dades, abordant preguntes com ara: Les dades són completes (cobreixen tots els casos requerits)? Són correctes, o contenen errors i, si hi ha errors, què tan freqüents són? Hi ha valors que falten a les dades? Si és així, com es representen, on es produeixen i amb quina freqüència es donen?

S'ha elevat doncs la qualitat de les dades al nivell requerit per les tècniques d'anàlisi seleccionades. Això ha implicat la selecció de subconjunts de dades, la inserció de valors predeterminats adequats o tècniques més ambicioses com l'estimació de dades que faltaven mitjançant modelització. S'ha descrit en tots els casos les decisions i accions que s'ha pres per abordar els problemes de qualitat de les dades i el perquè. Cal comentar que en tot moment s'ha tingut en compte les transformacions de les dades amb finalitats de neteja i el possible impacte en els resultats de l'anàlisi.

No obstant això, aquesta neteja no només ha implicat comprovar que les dades són correctes pel que fa al contingut, sinó també quant a format. Les transformacions de format es refereixen principalment a modificacions sintàctiques fetes a les dades que no canvien el seu significat, però que podrien ser requerides per l'eina de modelatge. En aquest cas, no s'ha dut a terme cap model, però són els canvis requerits per a poder visualitzar correctament les dades. Finalment, s'ha comprovat totes les restriccions que se sap que les variables han de complir a causa de la seva naturalesa. Aquestes han estat principalment comentades pels desenvolupadors de les BD d'on s'ha extret les dades.

Tot seguit cal comentar que al llarg d'aquest apartat es podrà observar una llista de totes les modificacions fetes així com una descripció més profunda d'aquelles considerades més importants. No obstant això, no s'inclouran totes i cadascuna de les decisions dutes a terme per a cada variable per la llargada en què això desembocaria. La resta d'informació es pot visualitzar al GitHub, tant als fitxers Python com als HTML, que permeten veure-ho més directament en ser la versió executada del primer. Finalment, cal declarar que aquest ha estat possiblement l'apartat més llarg i dur de realitzar de tot el projecte.

3.3.1 Subvencions de llengua

Per a aquest conjunt, en primer lloc, s'ha comprovat que cap tupla es trobés repetida, és a dir, que cap d'elles fes referència a una mateixa persona objecte i termini, tenint en compte que un usuari només pot presentar una sol·licitud dins d'un mateix termini. Aquesta restricció només s'ha comprovat per a les subvencions de concurrència competitiva, al ser les subvencions directes l'excepció de la regla. 11 files han trencat la restricció i per tant 6 d'elles (les que contenien menys informació) han estat eliminades. Aquesta restricció d'unicitat només s'ha validat a partir de l'any 2015, perquè antigament les normes eren diferents. Aquesta necessitat d'introduir la variable *any_directe* a la condició ha fet que s'hagués de trobar en tot moment informada, el qual en un inici no era així. Aquest segon problema ha estat solucionat imputant la variable a partir de la data de sol·licitud.

En tercer lloc s'ha eliminat 14 variables. 12 d'elles perquè tenien més del 75% de les files nul·les i a més a més no aportaven informació rellevant i, 2 perquè codificaven la mateixa informació que altres variables ja presents en el conjunt, sent les mantingudes més específiques. Tot seguit s'ha comprovat que totes les dates estiguessin entre els anys durant els quals l'IRL ha atorgat subvencions de llengua. Només 1 fila de la variable *datamaxjustificacio* trencava la restricció, tanmateix, després d'haver-la explorat amb l'àrea s'ha arribat a la conclusió que era correcte en fer referència a una subvenció plurianual.

S'ha progressat estudiant les variables que contenien text. Dins d'aquestes s'ha observat que les variables *tipusresolucio, pais_p* i *pais_pobjecte* tenien alguns valors que requerien una recodificació deguda principalment a faltes ortogràfiques que donaven lloc a la presència de categories duplicades. Pel que fa als nombres només s'ha extirpat 4 variables categòriques, totes elles per tenir només una categoria possible. L'import revocat també ha estat eliminat per a contenir valors estranys que no es podien confirmar.

Finalment, s'ha contrastat que totes les variables necessàries complissin els criteris inherents a elles mateixes. En primer lloc, s'ha revisat que totes les tuples referents a subvencions de concurrència tinguessin informades *idtermini, idconvocatoria* i *idprefix*, els ID principals per a aquest tipus de subvencions. Totes elles eren correctes i, per tant, s'ha passat a la segona comprovació d'aquesta subsecció.

Per a explicar aquest capítol s'ha de saber que existeixen tres tipus d'imports rellevants pel que fa a les subvencions. En primer lloc, l'import sol·licitat és aquell que l'usuari/entitat indica a l'hora d'omplir el formulari inicial. En segon lloc, l'import subvencionable és la part de l'import sol·licitat que els usuaris de l'IRL consideren que compleix els requisits de la convocatòria i, per tant, en tot moment ha de ser igual o inferior a l'import sol·licitat. En últim lloc, l'import atorgat és la quantitat de diners que finalment se li concedeixen als usuaris/entitats i, en conseqüència, ha de ser igual o menor a l'import subvencionable.

Pel que fa a aquesta clàusula, en primer lloc s'ha eliminat 60 tuples perquè contenien un import sol·licitat igual a 0, considerant-se subvencions no acabades. Posteriorment s'ha extirpat 7 files per a tenir un import subvencionable major que el sol·licitat i no poder confirmar quin dels dos era el correcte. Al fer aquesta declaració s'ha de tenir en compte que abans d'eliminar files s'ha fet que l'import subvencionable fos igual al sol·licitat si aquest també era igual a l'atorgat,

perquè s'ha assumit que l'usuari s'hauria equivocat a l'hora d'indicar l'import subvencionat. Finalment, s'ha extret 106 files per tenir un import atorgat major que el subvencionable.

Posteriorment, s'ha constatat que tots els estats de les subvencions concordessin amb els imports. Les modificacions fetes per a aquesta raó han estat múltiples. En penúltim lloc, s'ha validat que tota sol·licitud hagués estat feta dins dels terminis, temporalment parlant. En aquest cas cap element trencava la restricció i per això s'ha passat a mirar en últim lloc, que tota persona/entitat hagués rebut com a màxim 20.000 €/any ajuntant totes les subvencions de concurrència, al ser el màxim de diners que l'IRL pot atorgar a una mateixa persona/entitat per any. Dues files han trencat aquesta restricció, però l'àrea ha comentat que devien ser excepcions i, per tant, s'han continuat mantenint en el conjunt de dades.

3.3.2 Subvencions de creació

En segon lloc pel que fa a les subvencions, com en el cas anterior s'ha començat observant quantes tuples es trobaven duplicades. Tanmateix, en aquest cas s'ha afegit a la restricció les variables *noms_festivals* i *objecte*. Aquestes dues variables, de fet, han estat afegides en el conjunt de dades només per a comprovar aquesta restricció d'unicitat. Això és degut al fet que en aquesta àrea un usuari pot demanar més d'una subvenció dins d'un mateix termini si és per a acomplir diferents activitats, les quals es consideren esmentades en aquestes dues variables. En aquest cas 7 files han trencat la restricció, això no obstant, totes elles s'han mantingut dins del conjunt al ser falsos positius per generalment tenir un objecte no prou detallat. Cal comentar que també s'ha hagut d'imputar la variable *any_directe* a partir de *datasolicitud* per a completar aquest pas.

En segon lloc, s'han eliminat 7 variables. 5 d'elles perquè tenien més del 75% de les files nul·les i no aportaven informació rellevant i les altres 2 perquè codificaven la mateixa informació que altres variables ja presents en el conjunt, sent les mantingudes més específiques. Pel que fa a les dates aquest cop dos variables presentaven valors estranys. La fila que trencava la restricció per a cada cas ha estat posada a nul al no poder-se imputar.

Pel que fa a les variables compostes per text s'ha recodificat cadenes de *pais_p* i *paisos_festivals* altre cop per a tenir errors ortogràfics que donaven lloc a més categories de les correctes. Així mateix, la variable *poblacions_Espanya_festivals* ha estat utilitzada per a donar lloc a un nou conjunt de dades sencer. Aquest conjunt conté només les línies, l'any directe, les poblacions i una quarta variable que indica quants cops s'ha donat cada situació. A més a més, els noms de les poblacions també han hagut d'uniformitzar-se en contenir varis d'ells errors ortogràfics o utilitzar múltiples maneres per a referir-se a una població.

Posteriorment, pel que fa a les variables categories numèriques se n'ha extret tres per a tenir totes elles 1 sola categoria. També s'ha canviat el valor que prenia una fila de la variable *importsubvencionable* al detectar gràcies a un diagrama de caixes que era una anomalia. Altre cop s'ha eliminat la variable *importrevocat* degut a la impossibilitat de validar els valors.

Finalment, s'ha contrastat que totes les variables necessàries complissin els criteris inherents a elles mateixes. En primer lloc, s'ha repassat que totes les tuples referents a les subvencions de concurrència tinguessin informades les variables *idtermini, idconvocatoria* i *idprefix*. 5 tuples no

es trobaven correctament informades i han estat completament eliminades. També s'ha descartat 55 files per a tenir l'import sol·licitat igual a 0, 27 per a tenir un import subvencionable major que el sol·licitat, després d'haver recodificat tots els imports subvencionables incorrectes que es podien imputar i, 6 per a tenir un import atorgat major que el subvencionable. Cal tenir en compte que totes les files que trencaven restriccions d'aquest tipus s'han eliminat directament, en comptes de canviar a nul el valor de la variable, perquè són les columnes més importants del conjunt de dades i, si no són correctes, la probabilitat que les altres columnes ho siguin disminueix dràsticament. Així doncs, com que no es volia esbiaixar l'anàlisi i sent-ne conscient de les files amb què es tractava, s'ha considerat que era la millor opció.

Finalment, com en el cas anterior múltiples modificacions han estat dutes a terme perquè tots els estats de les subvencions concordessin amb els imports d'aquestes. No obstant això, en aquest cas cap fila ha trencat la restricció de sol·licituds entrades fora dels períodes ni de persones/entitats havent rebut més de 20.000 €/any mencionades en la secció anterior.

3.3.3 Subvencions de literatura

Per al cas de les subvencions de literatura, en primer lloc, s'ha eliminat una tupla per a no tenir indicada l'*idlinia*, una de les variables més importants del conjunt. Tot seguit, s'ha continuat imputant els valors no informats d'*any_directe* a partir de *datasolicitud* per a poder comprovar la restricció d'unicitat. A la restricció també se li ha hagut d'afegir la columna *objecte*, respecte a les variables usades per a l'àrea de llengua, a causa de que un usuari pot demanar diverses sol·licituds dins d'un mateix termini, si aquestes són per a traduir, per exemple, diferents obres; fet indicat a la variable *objecte*. 18 columnes han sortit com a repetides, però altre cop s'han mantingut totes en considerar-se correctes amb un objecte no prou detallat.

Pel que fa al pas d'eliminar variables en funció del percentatge de nuls per columna, 7 variables han estat extretes. També han estat extirpades *codipais_p* i *codipais_pobjecte* per a contenir exactament la mateixa informació que *pais_p* i *pais_pobjecte*, però aquestes darreres de forma més específica.

Posteriorment, al focalitzar l'atenció en les dates s'ha notat que fins a 5 variables mostraven valors fora dels intervals requerits, tanmateix, només els valors d'una d'aquestes s'ha pogut imputar a partir d'altres tuples. Les altres han estat recodificades a nul. Canviant a les cadenes de text, aquest cop s'ha extret la variable *disciplina_trad* en prendre només 1 valor i, per tant, no aportar informació rellevant de cara a l'anàlisi. Com en els casos anteriors s'ha hagut de recodificar categories de les variables *pais_p* i *pais_pobjecte* per problemes ortogràfics.

Tot seguit, dues columnes numèriques més han estat extretes per mostrar una única categoria. D'altra banda, s'ha advertit que algunes variables que codifiquen anys prenien valors incorrectes en ser massa petits. La variable *any_directe* s'ha pogut imputar a partir de la coneixença dels terminis de les subvencions mentre per a les variables *anytraduccio_trad* i *anypublicacio_trad* s'han hagut de canviar 7 i 9 files, respectivament, a nul. Altre cop s'ha eliminat l'import revocat per a no poder confirmar que els valors fossin correctes.

Finalment, s'ha contrastat que totes les variables necessàries complissin els criteris inherents a elles. En primer lloc, s'ha repassat que totes les tuples referents a subvencions de concurrència

tinguessin informades les variables *idtermini, idconvocatoria* i *idprefix*. 1 tupla no es trobava correctament informada i ha estat eliminada. També s'ha descartat 147 files per a tenir l'import sol·licitat igual a 0, 9 per a tenir l'import subvencionable major que el sol·licitat (després d'haver recodificat tots els imports subvencionables incorrectes que es podien imputar) i 4 per a tenir un import atorgat major que el subvencionable.

Finalment, altre cop múltiples modificacions han estat dutes a terme perquè tots els estats de les subvencions concordessin amb els imports d'aquestes. Tanmateix, aquest cop també s'ha hagut de fer alguna modificació pel que fa a les dates. Posteriorment s'ha notat que fins a 7 usuaris/entitats van rebre més de 20.000 €/any, però després de comprovar-ho amb l'àrea, tots ells s'han continuat mantenint en ser excepcions. Finalment, s'havia de comprovar que totes les subvencions referents a alguna traducció tinguessin informada la variable *idtraduccions*. No ha estat així per a 6 files i al ser aquesta una de les variables més importants del conjunt han estat directament eliminades.

3.3.4 Inscripcions

Pel que fa a les inscripcions com en els casos anteriors s'ha començat comprovant la restricció d'unicitat de les tuples. Aquest cop les variables implicades en el procés han estat *idpersona, idestada* i *any_estada*, perquè una persona no pot trobar-se físicament en una mateixa estada dos cops. En tot cas duplicat s'ha extret la tupla més antiga en ser la que més probablement tenia menys camps indicats.

Tot seguit totes les columnes que tenien més d'un 75% de nuls han estat eliminades. També s'ha suprimit les variables *tipuspersona* i *alfresco*; la primera per a ser una cadena de text amb només un valor possible que no aportava nova informació i la segona per a contenir la mateixa informació que *descripcio_ECCS*, aquesta segona de forma més específica.

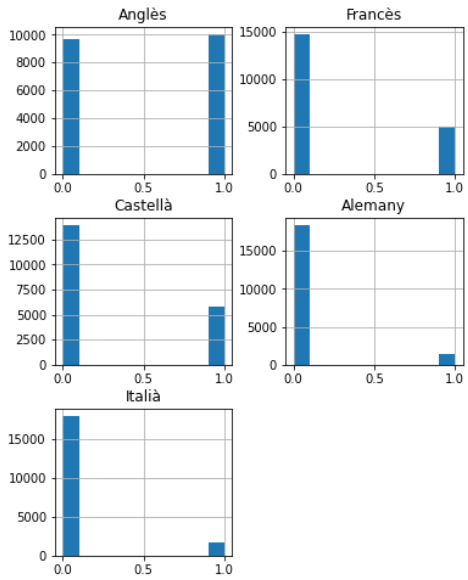
D'altra banda, pel que fa a la unificació de cadenes de text les variables *universitat* i *universitat_p* han estat les més difícils de tractar al mostrar múltiples opcions. Finalment, pel que fa als nombres, s'ha remogut 4 variables; 3 d'elles per a prendre una única categoria i l'altre per a tenir una distribució massa desequilibrada i no aportar informació rellevant.

3.3.5 Certificació

En la restricció d'unicitat que han de complir les dades de certificació es troben implicades les variables *id_alumnes, id_convocatoria* i *id_nivell*; perquè una mateixa persona es pot presentar múltiples cops a una convocatòria sempre que sigui a nivells diferents. Altre cop dins les duplicades s'ha eliminat les tuples més antigues. En segon lloc, s'ha hagut de modificar 323 tuples per a tenir una data de naixement de l'usuari incorrecte. S'ha considerat incorrecta tota aquella que fes referència a més de 95 anys o menys de 14 (límit legal) a l'hora de realitzar l'examen. Passant a les cadenes de text s'ha recodificat dues variables. En un dels casos aquesta recodificació ha fet augmentar de forma notòria els nuls en la variable, tanmateix, s'ha mantingut per sota del 50% de nuls i, per tant, no s'ha extirpat.

Tot seguit s'explicarà alguna modificació més complicada pel que fa a les variables que contenen cadenes de text. En primer lloc, s'enunciarà el cas de les variables *institucio_examen_conv_seus, poblacio_examen_conv_seus* i *tipus_institucio*. Pel que fa a les dues primeres variables, a l'hora d'explorar-les s'ha notat que moltes tuples contenien una 'a' evidentment incorrecte com a valor. En ser aquests camps responsabilitat directa dels usuaris interns de l'IRL s'ha comentat amb l'àrea. Aquesta ha enviat un Excel que engloba la correspondència entre cada convocatòria amb la població, institució i a més a més una nova columna que indica el tipus d'institució. S'ha modificat totes les tuples a partir d'aquest Excel per a assegurar que les dues variables inicialment mencionades fossin correctes, així com s'ha realitzat la inserció de la informació del tipus d'institució a *tipus_institucio*. Cal destacar que s'ha modificat totes les tuples, no només les files que contenien una 'a', perquè sobre aquest Excel també s'ha dut a terme un procés d'uniformització dels noms de les institucions al català.

Una altra de les modificacions complexes ha estat el cas de la variable *ocupacio_alumne*. Aquesta és agafada directament d'un camp de text lliure d'un formulari. Això dona lloc a múltiples variacions d'un mateix treball així com a múltiples ocupacions diferents. Com que és impossible uniformitzar-los un per un, s'ha unificat només les ocupacions més destacables. Una modificació semblant s'ha dut a terme per a la variable *altres_llengues_alumne*. En aquest cas s'ha creat 5 noves variables, cadascuna d'elles contenint informació d'una de les llengües més parlades pels usuaris: anglès, francès, castellà, alemany i italià.



Il·lustració 11: Distribucions de les llengües no maternes conegudes pels examinands

També s'ha hagut d'uniformitzar les variables *llengua_materna_alumne* i *uni_catala_ultim_any*, totes dues per a provenir de camps de texts lliures d'un formulari. D'altra banda, la neteja de dades dels nombres ha estat molt més fàcil en haver-se eliminat únicament dues variables per a contenir totes dues una sola categoria i, per tant, no aportar informació rellevant.

Pel que fa a la comprovació de les dades en funció de la seva naturalesa simplement s'ha hagut de comprovar que totes aquelles tuples que pertanyessin a algun país no exclòs de pagament i haguessin fet el nivell bàsic haguessin pagat 30 €. El preu esmentat s'ha augmentat en 5 € per als dos nivells intermedis i en 10 € per als dos superiors a l'hora de dur a terme la comprovació.

3.3.6 Selecció de professorat

Per al cas de selecció la restricció d'unicitat que han de complir les dades es troba composta per les columnes *id_convocatoria* i *nom*, al ser evident que un mateix usuari no pot presentar múltiples candidatures a un mateix lloc de treball. Altre cop s'ha mantingut en el conjunt les tuples més recents temporalment dins les 221 que han trencat la restricció. Així mateix, cal comentar que abans d'aquesta ha estat necessària una petita uniformització pel que fa als noms, per a assegurar que aquest era correcte. Aquesta mateixa modificació també ha estat feta per als noms dels professors, la qual es basa en canviar tots els noms a majúscules, per a assegurar que un mateix usuari/professor no estigui introduït a la base de dades en aquests dos formats.

La següent modificació duta a terme sobre les dades ja ha estat feta sobre cadenes de text, implicant doncs que no s'ha eliminat cap variable pel mig ni s'ha observat problemes en les dates. Diverses columnes que contenen cadenes de text han presentat la necessitat de recodificar la categoria "Seleccioneu una opció" que substituïa el nul. També han requerit una modificació de categories per a la correcta uniformització *universitat_master, universitat1, motiu_exclusio* i *pais*.

Un tema a part ha estat el procediment que s'ha hagut de dur a terme per a les variables *idiomes*, *nivells_idiomes, puntuacions_idiomes* i *universitats*. Totes elles s'han extret de la base de dades concatenant vàries opcions ordenades dins d'aquestes. Així doncs, s'ha separat la informació en el cas dels idiomes en 6 noves variables, en ser el màxim d'idiomes que algun usuari ha indicat i, en 3 per al cas de les universitats. Tanmateix, en el cas dels idiomes 3 d'aquestes columnes creades contenien més d'un 95% de nuls i, per tant, han estat eliminades.

Posteriorment, pel que fa als nombres s'ha hagut d'eliminar 3 variables per a tenir un nombre excessiu de nuls. Aquestes en un inici utilitzaven el 0 per a codificar el valor nul i per això no havien estat detectades. Aquest fet ha estat descobert gràcies a l'histograma fet per a explorar-les. L'histograma dels nombres categòrics també ha servit per a detectar que en el cas de la variable admès es feia servir tant el -1 com l'1 per a indicar que l'usuari havia estat admès.

3.3.7 Memòries

Les memòries com comentat diversos cops al llarg del projecte són justificants de subvenció dels diners atorgats a les universitats que imparteixen docència en català. Així doncs, la restricció d'unicitat és composta per *id_persona, universitat, any_academic, id_assignatura, id_activitat* i *id_produccio*. En realitat només les 3 primeres variables serien necessàries per a comprovar la restricció, però com que es té la informació de les diferents memòries multiplicada tants cops com activitats, assignatures i produccions s'han produït, ha estat necessari introduir aquests tres últims ID esmentats a la comprovació. Altre cop s'ha extret les més antigues dins les duplicades.

Tot seguit s'ha extret totes les variables que contenien més d'un 75% de nuls. Posteriorment, s'ha passat directament a considerar canvis sobre les variables que contenen cadenes de text, en ser les dates correctes. La primera modificació ha passat per a imputar la informació faltant de les variables *valor_puntuacio_geo* i *puntuacio_geo*. En segon lloc, s'ha recodificat els possibles valors que poden prendre 8 de les variables del conjunt de dades. Totes elles han

requerit aquest preprocessament per culpa de ser escrites de forma manual pels usuaris en camps de text lliure que indueixen a errors ortogràfics o a la utilització de múltiples maneres per a referir-se a un mateix fet. En tercer lloc, en analitzar les variables *valor_puntuacio_curr* i *valor_puntuacio_activitat* s'ha notat que totes dues contenien informació concatenada que hauria de pertànyer a diferents columnes. Aquest fet ha dut a la utilització de la funció "Split" de Python per a crear-les.

Finalment pel cas de les variables amb cadenes de text, cal comentar el cas de la variable *valor_puntuacio_aportacio_irl_prof*. Aquesta teòricament hauria de ser un nombre, no una cadena de text. Així doncs, s'ha buscat quins eren els valors que havien dut a classificar la variable com a text i s'han recodificat a nombres. Tot seguit, s'ha dut a terme un diagrama de caixes per a comprovar que els nombres prenguessin valors adequats. Existien dues anomalies molt clares que han estat explorades.

Tot seguit, cal destacar que un dels processos més tediosos d'aquest conjunt ha estat comprovar les anomalies que mostraven les variables quantitatives numèriques. No s'enumeraran una per una a l'haver realitzat fins a 12 comprovacions. Finalment, pel que fa a les restriccions inherents a la naturalesa de les dades només s'ha comprovat que si s'havien firmat convenis, la variable que indica aquest fet es trobés a "Sí".

3.4 Construcció o recodificació de variables per a les gràfiques

Aquesta tasca inclou principalment operacions constructives de preparació de dades, com ara la producció d'atributs derivats o registres nous complets. Els atributs derivats són atributs que es construeixen a partir d'un o més atributs existents al mateix registre. Tanmateix, al llarg d'aquesta secció també s'ha inclòs modificacions fetes en els formats de les dades o unions de categories demanades pels usuaris de les àrees. Inclou doncs totes les accions que es poden observar en les seccions "Exploració i neteja de dades" – "Pas 6: Construcció o recodificació de les variables per a realitzar els gràfics" dels guions continguts al GitHub.

3.4.1 Subvencions de llengua

Per a les subvencions de llengua, en primer lloc, s'ha creat la variable *dies_termini* per a poder respondre a la hipòtesi de si com més dies es troba oberta una convocatòria, més sol·licituds es reben, amb major facilitat. En segon lloc, s'ha recodificat les categories de 3 variables: *concurrent, identificacio_pobjecte* i *robinson_pobjecte*, perquè totes elles mostressin text en comptes de nombres i poder d'aquesta manera comprendre millor les dades. Finalment, per demanada de l'àrea, s'ha ajuntat algunes de les categories de la variable *idlinia_conv* i, en conseqüència, també de *descripcio_linia_subv*.

3.4.2 Subvencions de creació

En aquest segon cas de subvencions s'ha dut a terme exactament els mateixos passos que els mencionats anteriorment per a les subvencions de llengua. L'única acció que ha canviat ha estat la combinació de diverses categories de les variables *idlinia_conv* i *descripcio_linia_subv*, que s'ha substituït per a la reducció del text d'una de les etiquetes de *descripcio_linia_subv* perquè no ocupés tant d'espai en les gràfiques.

3.4.3 Subvencions de literatura

Per a aquest conjunt només dos tipus d'accions han estat necessàries. Com en els casos anteriors s'ha creat la columna *dies_termini* per a poder respondre a la hipòtesi de si com més dies es troba oberta una convocatòria, més sol·licituds es reben, amb major facilitat. D'altra banda, s'ha canviat de nombres a cadenes de text les categories de 5 columnes per a una millor comprensió.

3.4.4 Inscripcions

En aquest cas altre cop s'ha dut a terme dos tipus d'operacions. La primera ha estat la recodificació de 4 variables numèriques categòriques a cadenes de text per a entendre-les millor. La segona operació ha consistit en afegir informació de localització a les estades dutes a terme del 2013 al 2016, per petició de l'àrea, uniformitzant els noms de les estades.

3.4.5 Certificació

Aquest conjunt de dades és el més complex pel que fa a aquesta secció. En primer lloc, s'ha introduït dues noves variables: *data_inscripcio_convocatoria* (per a saber si com més dies es troba oberta una convocatòria, més alumnes s'apunten a aquesta) i *edat_alumnes*. Tot seguit 5 variables han estat recodificades per a tenir nivells que es poguessin comprendre més correctament. Pel que fa a la variable *pais_alumne*, per petició de l'àrea, s'han eliminat algunes de les possibilitats perquè aquestes no es mostressin en les visualitzacions finals, perdent doncs informació, però informació no rellevant de cara a l'IRL. Finalment, s'ha donat lloc a una nova variable anomenada *nivell_pais_hcat*. Aquesta codifica el nivell més comú donat un país i unes hores de català, necessària per a poder fer un gràfic molt concret: "Nivell obtingut en funció del país de naixement i les hores de català estudiades".

3.4.6 Selecció de professorat

Les dues primeres modificacions fetes pel cas de selecció tenen una alta relació amb les comentades en l'apartat anterior en haver introduït les variables *dies_convocatoria* i *edat_candidats* per a codificar els dies que les convocatòries s'han trobat obertes i les edats dels

candidats facilitant el codi dels posteriors gràfics. També s'ha dut a terme un procés similar amb la creació de les variables *fa_anys_estudis* i *fa_anys_master* que indiquen el temps que fa que un usuari ha assolit el grau/màster des que ha decidit presentar-se a la convocatòria. A més a més, la creació d'aquestes dues variables ha permès detectar nous valors incorrectes en les columnes *any_estudis* i *any_master*, que han estat canviats a nul. Finalment, s'ha recodificat diversos nombres a categories escrites per a poder comprendre-les millor.

3.4.7 Memòries

Pel que fa a aquesta secció, aquest conjunt és el més simple, en haver-se realitzat únicament dues modificacions molt senzilles. La primera d'elles ha estat retallar el nom d'una de les universitats evitant d'aquesta manera obtenir gràfiques massa grans per culpa de les etiquetes. La segona ha implicat indicar tots els valors d'una variable que inicialment eren nul a 0, perquè altrament la llibreria Altair no permetia utilitzar la funció "max".

3.5 Descripció de les dades finals

Al llarg d'aquesta secció es descriuran els conjunts produïts per la fase de preparació de dades que s'utilitzaran pel treball d'anàlisi principal del projecte. A més a més, també es descriuran els Excels no obtinguts a partir de les extraccions de les bases de dades, però usats al llarg de la secció de visualització, per a saber quines dades s'ha mostrat. Es pot accedir a aquests conjunts a les carpetes dades.zip del GitHub.

3.5.1 Subvencions de llengua

Pel que fa a les subvencions de llengua s'ha usat dos conjunts de dades: les que fan referència a les subvencions i les que contenen les relacions amb l'àrea de llengua i el tipus d'aquestes al llarg dels anys. Aquestes últimes també han estat extretes a partir d'una consulta SQL de la base de dades, però no han hagut de ser netejades.

En primer lloc, el conjunt de dades de subvencions pesa 722KB, està format per 3577 entrades i conté informació de 53 característiques. 21 d'aquestes són cadenes de text, 7 dates i 25 nombres. De fet, d'aquests 25 nombres 15 són racionals (float) mentre 10 són enters (integer). Les descripcions de cada variable es poden trobar a l'Annex 4.1 Descripció dels conjunts de dades - Subvencions de llengua, junt amb la indicació del seu tipus i el nombre de files no nul·les que contenen. Les variables més importants d'aquest són: *idpersona*, *idestatsubvencio*, *idlinia_conv*, *idprefix*, *idtermini*, *pers_objecte*, *any_directe* i *concurrent.*

D'altra banda, el conjunt que conté les relacions pesa 272KB, conté 6731 entrades i 8 columnes. Dins d'aquestes 3 contenen nombres, tots ells enters, mentre 5 són objectes. En aquest cas les variables més rellevants són *idpersona, anyacademic* i *tipuscentre*. Així mateix, les descripcions, tipus i nombre de files no nul·les de cada variable es poden observar en el mateix annex mencionat anteriorment.

3.5.2 Subvencions de creació

La informació usada per a crear les gràfiques referents a les subvencions de creació és altre cop emmagatzemada en dos conjunts de dades. En primer lloc, la informació sobre les subvencions és composta per 68 columnes, 10.811 files i pesa 3.595KB. 9 columnes contenen dates, 22 nombres racionals, 11 enters i 26 cadenes de text. Es pot obtenir informació més detallada sobre aquestes a l'Annex 4.2 Descripció dels conjunts de dades - Subvencions de creació. Les variables destacades aquest cop són: *idpersona, idestatsubvencio, idlinia_conv, idprefix, idtermini, pers_objecte, any_directe* i *concurrent.*

D'altra banda, l'Excel que conté informació sobre les poblacions d'Espanya agrupades per any i línia, redueix de forma significativa la mida respecte a l'anterior en 58KB. És compost per 2.241 files, 1 columna entera, 1 racional i 2 objectes. En aquest cas totes quatre variables són molt valuoses, però les indispensables serien: *any_directe, descripcio_linia_subv* i *poblacions_Espanya_festivals*. Altre cop, més informació sobre el nombre de valors nuls per columnes, entre d'altres, es pot aconseguir en l'annex mencionat anteriorment.

3.5.3 Subvencions de literatura

La informació de l'àrea de literatura es troba guardada en la seva totalitat en un Excel. Aquest pesa 1.656KB, conté 89 columnes i 3.925 files. 13 d'aquestes són dates, 37 nombres racionals, 7 enteres i 32 objectes. Les variables importants d'aquest són: *idpersona, idestatsubvencio, idlinia_conv, idprefix, idtermini, pers_objecte, any_directe, concurrent* i *idtraduccions*. Com en els casos anteriors es pot obtenir més informació del conjunt a l'Annex 4.3 Descripció dels conjunts de dades - Subvencions de literatura.

3.5.4 Inscripcions

Pel que fa als campus i les estades només s'ha necessitat 1 conjunt de dades per a poder dur a terme totes les gràfiques demanades. Aquest pesa 203KB, conté 1.162 files i és compost per 48 columnes. Una d'aquestes és una data, 11 són enteres, 21 racionals i 15 són objectes. Les variables destacables continuen sent: *idpersona, idestada* i *any_estada*. Altre cop es pot trobar més informació sobre aquest a l'Annex 4.4 Descripció dels conjunts de dades - Inscripcions.

3.5.5 Certificació

Pel cas de les proves d'avaluació i certificació de la llengua catalana només ha estat necessari l'ús d'un conjunt per a la creació de totes les visualitzacions. Aquest pesa 5.510KB, és compost per 71 columnes i conté 19.749 entrades. Així mateix, aquestes 71 columnes es divideixen en 8 que codifiquen dates, 22 objectes, 38 enters i 3 nombres racionals.

Les variables més importants continuen sent: *id_examen, id_nivell, id_conv_seus, id_convocatoria, id_conv_seus_nivells, id_seus, id_pais* i *id_alumnes*; que no han canviat

respecte a la secció de descripció i exploració de les dades. Finalment, es pot observar la descripció de cadascuna d'aquestes, el nombre de valors nuls per columna i el tipus a l'Annex 4.5 Descripció dels conjunts de dades – Certificació.

3.5.6 Selecció de professorat

Per a dur a terme les visualitzacions referents a selecció de professorat s'ha utilitzat dos conjunts de dades. El primer i més rellevant, en ser el que conté la majoria d'informació, pesa 903KB i s'ha obtingut a partir de l'extracció de les dades de la base de dades i la posterior neteja. Aquest inclou la informació de tots els formularis realitzats pels diferents candidats en les convocatòries així com si van ser finalment seleccionats i per a quin motiu. Està compost per 2.039 files i 95 columnes. 3 d'elles són dates, 24 contenen nombres racionals, 12 nombres enters i finalment la majoria, 56 columnes, són cadenes de text. Les variables més valuoses del conjunt són: *id_formulari, nom* i *id_convocatoria*. Si es vol aconseguir més informació d'aquest, es pot mirar l'Annex 4.6 Descripció dels conjunts de dades – Selecció de professorat.

En segon lloc, l'àrea va demanar que s'analitzés el nombre d'universitats a les quals donen suport (financen) per continent. Tanmateix, aquesta informació no es considerava a la base de dades d'on s'ha extret la informació. Aquest fet va dur a fer que els usuaris del Llull proporcionessin un Excel amb les dades que volien visualitzar. Aquest se sap que és completament correcte i per aquesta raó no s'ha dut a terme cap preprocès sobre aquest. No obstant això, cal comentar que sí que ha estat necessari dur a terme una modificació de l'organització de les dades per a poder ser carregades al Google Collaboratory. Aquestes són guardades al document anomenat seleccio_docencia_universitats.xlsx que pesa 43KB. Abans de començar la seva descripció, cal destacar que tot i que en un inici només volien mostrar les universitats a les quals havien finançat, l'Excel que van enviar també contenia la informació de les universitats a què podrien haver finançat a partir de l'any 2007 i, en conseqüència, la informació també ha estat lleugerament analitzada.

Acabar aquesta secció indicant que el conjunt de dades prèviament explicat es troba compost per 780 entrades i 5 columnes. 2 d'aquestes columnes són objectes mentre les altres 3 contenen nombres enters. Les variables més importants d'aquest són *continent* (en ser la raó de ser del conjunt), *any_academic* i *nombre_unis*. Finalment ressenyar que altre cop es pot obtenir la informació exacta de quants nuls presenten les variables, els noms, les descripcions i els tipus a l'annex mencionat en aquesta secció.

3.5.7 Memòries

Els Excels utilitzats per a realitzar els gràfics en el cas de les memòries han estat una mica especials. Com comentat al llarg del projecte la primera extracció de dades va portar a aconseguir la informació d'una mateixa memòria multiplicada tants cops com produccions, activitats i assignatures tinguessin associades. No obstant això, l'Excel resultant d'aquesta extracció i posterior neteja de dades pesa 25.552KB, conté 62.361 files i 111 columnes. Aquesta magnitud dur a la complicació del maneig de les dades a l'hora de crear els gràfics amb Altair.

Així doncs, com que se sabia que varies de les gràfiques usarien informació únicament d'una de les seccions en particular, a part de descarregar el conjunt de dades complet per a poder dur a terme alguna gràfica que requerís l'encreuament d'apartats, es van obtenir també 4 subconjunts del general.

El de les memòries pesa 270KB i conté només la informació d'aquestes de manera no duplicada. En segon lloc, es va separar la informació de les activitats i s'ha comprovat que pesa 222KB. El mateix s'ha fet tant per a produccions que ha resultat en un pes de 38KB, notant doncs que la majoria de memòries no tenen produccions associades i per a assignatures, donant lloc a 443KB de memòria. Aquests tot i no permetre l'encreuament d'informació han estat molt crucials per a la creació de les visualitzacions en ser molt menys pesats i no contenir informació repetida.

Posteriorment, s'explicaran les característiques de cadascun d'aquests subconjunts creats, tot i que en l'Annex 4.7 Descripció dels conjunts de dades – Memòries només es podrà trobar una taula que fa referència a les variables emmagatzemades en el conjunt de dades globals i, per tant, contindrà tota la informació. D'aquesta manera s'evitarà repetir text. Les variables més destacades d'aquest conjunt de dades globals *són universitat, any_academic, id_assignatura, id_activitat* i *id_produccio*, que difereix de les vistes a l'apartat 3.2.7 Descripció i exploració de les dades intermèdies - Memòries al no aparèixer *id_persona*. En un principi aquesta variable havia estat considerada important perquè els noms de les universitats no eren uniformitzats i, en conseqüència, no es podia confiar en la variable *universitat*, però sí en *id_persona*.

Tot seguit es durà a terme l'explicació de cadascun dels subconjunts extrets. En primer lloc, el subconjunt de les memòries és compost per 777 files i 82 columnes. En concret 2 d'aquestes columnes són dates, 20 contenen nombres enters, 43 racionals i 17 són cadenes de text. Les variables considerades crucials són *universitat* i *any_academic*.

En segon lloc, el subconjunt referent a les assignatures és constituït per 5.897 files i 15 columnes: 5 objectes, 5 enteres i 5 racionals. Aquest augment en el nombre de files respecte a les indicades per a les memòries permet concloure amb total seguretat que en mitjana s'ha inscrit entre 7 i 8 assignatures per memòria. Les variables importants en aquest cas són les dues mencionades anteriorment més *id_assignatura*.

Pel cas de les produccions el nombre de files disminueix dràsticament fins a 787, així com el nombre de columnes passa a ser 6. També es redueix la tipologia de les diferents variables en contenir només 4 cadenes de text i 2 variables formades per nombres enters. Les variables destacades tornen a ser *any_academic* i *universitat*, però en aquest cas es troben acompanyades per *id_produccio*.

Per acabar el conjunt d'activitats conté 2.609 files i 14 columnes, estadístiques mitjanes respecte a les mencionades en els dos subconjunts anteriors. 1 d'aquestes variables és integrada per nombres racionals, 9 per nombres enters i 4 per cadenes de text. Les variables notòries són *any_academic, universitat* i *id_activitat*.

4 Visualitzacions

Com descrit a la secció (1.3 Objectiu) el propòsit d'aquest projecte és analitzar les dades de l'Institut Ramon Llull perquè posteriorment els treballadors d'aquest puguin prendre decisions racionals basades en elements estadístics confiables. Aquesta anàlisi, es pot dur a terme de moltes maneres diferents, tanmateix, en aquest cas s'ha decidit realitzar diversos gràfics per mostrar la informació de forma visual.

Tot seguit, s'explicarà més concretament el perquè ha estat la tècnica escollida. En un inici es podrà observar un apartat dedicat a la teoria dels gràfics mentre més endavant s'enumeraran els diagrames duts a terme en cada secció d'anàlisi així com es posarà algun exemple per cadascuna d'aquestes.

Com a breu introducció d'aquesta secció afirmar que s'ha usat visualitzacions perquè aquestes no només ajuden a comprendre, sinó també a comunicar informació i coneixement de les dades d'una manera breu. Així mateix, un cop portades a cap estalvien temps a l'usuari evitant que hagi de mirar el conjunt de dades, que podria ser complicat. Per aquest motiu, quan es resol un problema de visualització és important saber des d'un inici què es vol visualitzar i per què.

4.1 Marc teòric de les visualitzacions

Com comentat anteriorment aquesta secció està focalitzada en l'àmbit teòric de les visualitzacions creades a l'apartat 4.2 Anàlisi de les visualitzacions. Així doncs, al llarg d'aquesta es podrà trobar una explicació més detallada del perquè ha estat la tècnica escollida. També es mencionaran els diferents tipus de gràfics usats en el projecte amb les seves característiques més rellevants i finalment es durà a terme una explicació de quina tècnica ha estat la usada per a construir-los.

4.1.1 Gràfics i les seves característiques

La visualització d'informació a través de gràfics es troba directament relacionada amb la comprensió de les dades subjacents. Aquests ajuden a l'ésser humà a realitzar un procés cognitiu per a comprendre les dades sent la principal raó per la qual seran l'eina emprada durant aquest projecte per a mostrar els resultats.

També, cal comentar que aquests ajuden a dur a terme tasques amb més eficàcia permetent saber de forma directa si els resultats observats són els que s'esperaven o si per contra mostren patrons inesperats. En el primer cas es continuaria amb els fluxos de treball existents, mentre el segon donaria lloc a una anàlisis completament nova, avaluant en tot moment la validesa del model estadístic.

Així mateix, a l'hora de portar-los a cap s'ha de tenir en compte diversos aspectes. En primer lloc, s'ha d'aconseguir una visualització expressiva que mostri la informació de les dades sense afegir ni evitar informació. En segon lloc, aquesta ha de ser efectiva, és a dir, ha de tenir en compte les capacitats cognitives del sistema visual humà. En últim lloc, ha de ser adequada el

qual fa referència a la compensació entre els esforços necessaris per crear la visualització i els beneficis que aquesta pot produir.

Dels punts anteriorment mencionats en destaquen les deficiències que pot tenir el sistema visual humà. Per aquest motiu, a continuació s'estudiaran diferents aspectes que poden afectar a la percepció de les visualitzacions i que en conseqüència, s'hauran de tenir en compte a l'hora de crear-les. Aquestes deficiències són causades pel procés que protagonitza el nostre ull en conjunt amb el cervell en rebre un estímul visual (processament preatent), les propietats que es detecten sense haver de prestar atenció (propietats preatentives) i els principis que descriuen com els humans agrupem elements similars, reconeixem patrons o simplifiquem imatges (principis de Gestalt).

4.1.1.1 Processament preatent

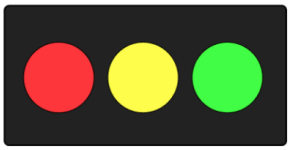
El processament de totes les visualitzacions és dividit en 3 etapes [12].

- **Etapa 1:** Processament ràpid: aquest és caracteritzat per extraure característiques de baix nivell. En ser una etapa que les neurones duen a terme en paral·lel, de forma simultània, no es pot controlar. Durant aquesta, es detecten la forma, els atributs espacials, l'orientació, el color, la textura i el moviment. Succeeix de forma automàtica independentment d'en què s'estigui centrant en aquell moment el focus cognitiu. La informació que es capta durant aquesta etapa és transitòria i, per tant, es guarda durant un temps curt en la memòria visual. Per les raons mencionades anteriorment, s'anomena tot sovint processament preatent.
- **Etapa 2:** Durant aquesta es reconeixen les estructures mitjançant una percepció dels patrons. Aquests patrons permeten identificar els objectes i saber si tenen, entre d'altres, contorns continus i regions o textures del mateix color. En ser un procés que es du a terme en sèrie és més lent que l'anterior.
- **Etapa 3:** Es troba caracteritzada per ser un procés seqüencial amb un objectiu dirigit constituint la base del pensament visual. La informació analitzada es redueix encara més concentrant-se en uns quants objectes conservats en la memòria de treball visual. S'utilitza principalment per a construir i respondre consultes visuals. Durant aquesta etapa es relaciona el visualitzat amb altres subsistemes com podrien ser la parla o el moviment.

4.1.1.2 Propietats preatentives

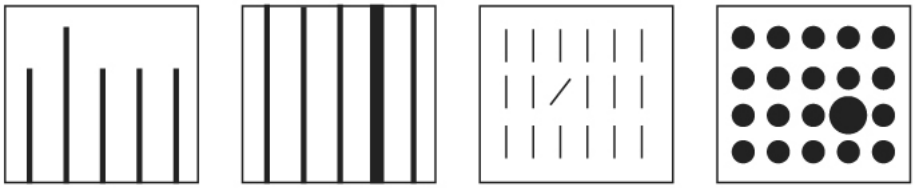
Existeixen un conjunt de propietats visuals molt limitades que poden ser processades de forma preatentiva, és a dir, durant la primera etapa del processament preatent. La informació codificada amb aquestes propietats ressalta inclús abans de ser-ne conscients, en menys de 500 mil·lisegons. Tenir en compte aquestes propietats pot ajudar molt a l'hora de dissenyar les visualitzacions al poder codificar amb aquestes les característiques que es vol que es percebin ràpidament. Existeixen quatre propietats preatents [12]: el color, la forma, el posicionament espacial i el moviment.

- **Color:** tot i escollir-se molts cops amb un objectiu estètic hi ha diverses normes a tenir en compte a l'hora de seleccionar-lo. En primer lloc, el color escollit per a l'objecte principal de la visualització ha de tenir un contrast adequat amb el fons. Tot seguit, les variables contínues s'haurien de representar amb un únic color i un gradient adequat. Una altra norma clara, per exemple, seria no representar el perill amb el color verd. La següent il·lustració ajuda a entendre el perquè:



Il·lustració 12: Semàfor de colors mostrant que el verd no pot ser escollit per a indicar perill

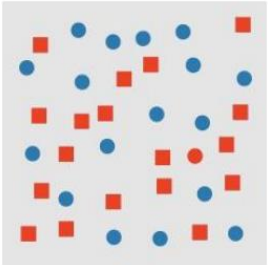
- **Forma:** l'ús d'aquesta propietat preatent inclou un ampli ventall d'atributs a utilitzar: la llargada, l'amplada i l'orientació de les línies, la mida, la curvatura, la forma i l'agrupació espacial. Les següents il·lustracions mostren l'efecte de les primeres quatre característiques:



Il·lustració 13: Visualitzacions que mostren que la llargada, l'amplada i l'orientació de les línies són propietats preatents, així com la mida

- **Posicionament espacial:** les visualitzacions en 1 sola dimensió són generalment pobres, consumint un espai no necessari. D'altra banda, les visualitzacions en 3D, tot i que a vegades són molt útils gràcies a la possibilitat de representar un major nombre de variables, poden ser confuses si es troben plasmades sobre una pantalla o paper amb el qual no es pot interaccionar. Per això, durant aquest projecte s'usaran en tot cas visualitzacions en dues dimensions, una manera completament adequada per a reconèixer les dades.
- **Moviment:** aquesta propietat tot i no considerar-se implementada durant el projecte, ja que totes les gràfiques mostrades són estàtiques, és molt útil en permetre cridar l'atenció d'algú amb èxit assegurat sempre que el moviment sigui adequat. Els moviments inadequats serien aquells massa ràpids que provoquen la pèrdua de l'usuari o massa constants que podrien provocar distraccions.

Tanmateix, s'ha de tenir en compte en tot moment que no es poden fer servir dues propietats preatentives al mateix temps, ja que la conjunció d'aquestes provoca la ineficàcia d'ambdues. Un exemple clar seria el mostrat a la següent figura:



Il·lustració 14: Cerca de conjunció per color i forma

Anteriorment, s'havia dit que tant el color com la forma són propietats preatentives, però en ajuntar-les en una mateixa visualització es nota que es requereix d'una cerca seqüencial per a captar totes les característiques d'aquesta (existeix 1 cercle de color vermell).

Posteriorment, cal comentar que qualsevol processament de propietats preatentives només funciona completament quan se sap què s'està buscant. En aquest cas, el cervell permet a les cèl·lules sensibles de l'element buscat tenir un paper més rellevant mentre les altres són parcialment silenciades. Finalment, cal destacar que aquestes són altament sensitives a qualsevol distractor i entrenar-se per a detectar-les més fàcilment no té cap influència.

4.1.1.3 Principis de Gestalt

Els principis de Gestalt [13][14] són lleis de la percepció visual humana que descriuen com s'agrupen elements similars, es reconeixen patrons o se simplifiquen imatges. Saber com funcionen pot ajudar a organitzar el contingut d'una visualització fent que aquesta sigui estèticament agradable i fàcil d'entendre. Aquests es regeixen pel fet que es tendeix a ordenar l'experiència d'una manera regular, ordenada i reconeixible.

L'ús d'aquests principis facilita a l'usuari percebre una sèrie d'elements individuals com un tot, per exemple, permetent-li entendre què veu a primera vista. A continuació s'explicaran alguns d'aquests principis, els considerats més importants de cara al projecte:

- **Principi de similitud:** estableix que quan dos objectes semblen semblants entre si, la ment humana els agrupa, pensant que tenen la mateixa funció.

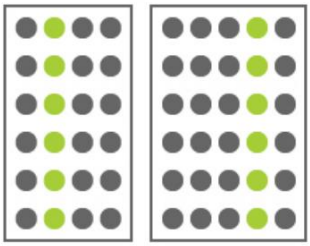
Il·lustració 15: Representació del principi de similitud

  En la imatge superior, tot i que totes les formes són iguals, s'observa clarament que cada color representa un grup diferent. De fet, es podria arribar a pensar que totes les columnes grises realitzen una funció separadora entre els altres grups.
- **Principi de proximitat:** estableix que els objectes que estan a prop semblen estar més relacionats que aquells més llunyans. És una propietat tan forta que anul·la la similitud

de color, forma i altres factors que podrien diferenciar un grup d'objectes, com afirmat en la il·lustració 16.

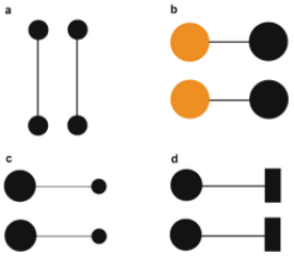
Il·lustració 16: Representació del principi de proximitat

- **Principi de contenció:** el principi de la regió comuna o contenció es troba molt relacionat amb el de proximitat. Afirma que quan els objectes es consideren dins la mateixa regió tancada, es perceben com agrupats. Altre cop anul·la altres principis com el de la forma o el color.



Il·lustració 17: Representació del principi de contenció

- **Principi de connectivitat:** estableix que els elements connectats visualment amb línies es perceben com un grup. Aquests semblen tenir una relació més forta que si estiguessin separats, com mostrat a la figura 18.



Il·lustració 18: Representació del principi de connectivitat

4.1.2 Tipus de gràfics

Al llarg d'aquest apartat s'explicarà de forma teòrica totes les tipologies dels gràfics que s'utilitzaran durant la secció 4.2 Anàlisi de les visualitzacions. Es comentaran les característiques, principals avantatges i desavantatges de cadascuna. Cal tenir en compte que tot i que s'avaluaran de forma separada, varies d'aquestes categories es poden ajuntar en una única visualització per a codificar informació a través dels màxims canals possibles.

4.1.2.1 Diagrama de línies

Tipus de diagrama que mostra la informació en una sèrie de punts connectats per segments de línies rectes. De fet, els punts que marquen cada valor concret no sempre són mostrats. L'ús més comú d'aquests és representar sèries temporals en permetre expressar molt bé com una variable canvia amb el temps dins d'una escala.

Facilita la cerca de patrons, anomalies i tendències en les dades. Tanmateix, quan diverses línies es creuen entre elles pot resultar difícil comprendre la informació. En aquest cas podria ser útil usar uns quants eixos paral·lels.

4.1.2.2 Diagrama de barres

Representació mitjançant barres horitzontals o verticals útil per a mostrar una variable categòrica (clau) en un eix i una variable discreta (valor) a l'altre. Normalment, la quantitat exacta de la variable discreta és la que s'expressa amb la longitud de la barra. L'ordenació de la categoria es pot dur a terme principalment per dos factors: per ordre alfabètic, el qual ajuda a la rapidesa de la cerca o pel valor de la variable discreta, que facilita comparar i analitzar els valors.

Una variació d'aquest tipus de visualització, també utilitzada durant el projecte, és el diagrama de barres apilades. Aquest permet representar no només una, sinó dues variables categòriques. La segona d'aquestes variables es codifica amb el color, dividint la barra en tants nivells com claus té la segona variable. No obstant això, cal tenir en compte que aquesta segona categoria afegida només pot prendre com a màxim una dotzena de valors, resultant altrament en un diagrama massa difícil d'analitzar. És usat per a representar la categoria dins del tot.

Finalment, existeix una tercera variació: els diagrames de barres apilats normalitzats. Aquests permeten codificar el mateix nombre de variables que les mencionades anteriorment. Tanmateix, la llargada de la barra (considerant tots els colors) sempre va de 0 a 100, indicant percentatges. Són útils per a concebre la mida de cada pila com a percentatge del conjunt.

4.1.2.3 Histograma

Representació gràfica associada a una variable contínua o discreta amb molts valors diferents que codifica la freqüència amb què cada valor d'aquesta es dona. Un dels eixos indica els rangs amb què es divideix la variable mentre amb l'altre es codifica la freqüència d'aquests rangs a partir de la llargada de la barra. Tot i ser altament emprats l'elecció del nombre de classes i l'amplada d'aquestes és sovint problemàtica.

4.1.2.4 Diagrama circular

Gràfic circular dividit en sectors on la longitud de l'arc i, per tant, l'angle central i l'àrea de cadascun, són proporcionals a la quantitat que representen. Cadascun d'aquests talls és

generalment codificat amb un color diferent. Pot ser molt útil si l'objectiu és comparar la mida d'un dels sectors en comparació a tot el gràfic, en comptes de talls entre ells. Altrament, en ser l'angle un dels canals que l'humà compara amb menys precisió la tasca pot esdevenir pràcticament impossible, sobretot si els angles de les seccions són molt iguals entre ells.

4.1.2.5 Mapa de calor

Codifiquen informació de dues claus (categòriques) i un valor (quantitativa). Són generalment emprats per a trobar grups o anomalies en les dades. Són formats per una matriu de dues dimensions (les claus de les dues variables categòriques formen els eixos) que conté a dins rectangles. La intensitat del color d'aquests rectangles codifica el valor. Poden presentar algun problema si s'intenta codificar un rang de valors massa gran on la majoria d'elements es concentren en uns valors molt iguals, al ser difícil distingir dues tonalitats de color molt iguals, com passava anteriorment amb l'angle.

4.1.2.6 Diagrama de dispersió

Codifica generalment la relació entre dues variables quantitatives, tot i que a vegades pot mostrar una variable categòrica en algun dels eixos. Generalment, els valors de la variable explicativa apareixen a l'eix X mentre els de la variable resposta es visualitzen a l'Y. Cada combinació d'aquestes dues apareix com un punt en el diagrama. Les distribucions que aquests formen permeten comparar i descobrir tendències així com anomalies.

Aquest tipus de visualització també mostra un subtipus utilitzat durant el projecte: els diagrames de bombolles. Aquests afegeixen una tercera dimensió al gràfic representant amb la mida del punt una nova variable.

4.1.3 Python i Altair

Per a dissenyar les gràfiques del projecte s'ha utilitzat el llenguatge de programació Python dins l'aplicació Google Colaboratory prenent Altair com el paquet principal. D'una banda, s'ha decidit usar Python [4] com a llenguatge de programació d'alt nivell degut a la seva gran llegibilitat del codi i sintaxi que permet als programadors expressar conceptes en menys línies de codi del què seria possible en llenguatges com C. A més, aquest presenta una gestió de la memòria automàtica i té una exhaustiva biblioteca estàndard. També, tot i que Python és utilitzat sovint com un llenguatge script, es pot fer servir en una àmplia gamma de contextos no-script a partir d'eines desenvolupades per tercers, com Google Colaboratory.

Google Colaboratory [15] és una aplicació que permet a través d'un entorn interactiu, no una pàgina web estàtica, escriure i executar codi Python al navegador. Els principals avantatges d'aquesta són que no es necessita realitzar cap configuració prèvia, que proporciona accés gratuït a GPU i que ofereix una alta facilitat per compartir el codi amb terceres persones. Gràcies a la seva facilitat a l'hora de fer-lo servir i el requeriment per part de l'IRL d'utilitzar una eina on

altres programadors poguessin visualitzar el codi, ha estat el sistema escollit on executar el projecte.

A més a més, amb aquesta eina es poden importar i emprar completament les biblioteques més populars de Python per analitzar i visualitzar dades com serien NumPy [16], Pandas [21] o Matplotlib [17]. Entre les biblioteques importables també s'hi troba Altair, l'escollida per a implementar els gràfics principals d'aquest projecte. Altair [5] és una biblioteca de visualització estadística declarativa per a Python, basada en Vega i Vega-Lite la font de la qual és disponible a GitHub. La interfície és senzilla, amigable i coherent, permetent produir visualitzacions atractives i efectives amb una quantitat de codi mínima, raó principal per la qual ha estat escollida.

Tot i que no s'explicarà en detall la programació de les gràfiques a causa de la llargada en què això desembocaria, tot seguit es descriuran els elements més destacables per a crear una visualització amb Altair. En primer lloc, s'hauria d'importar, com mencionat anteriorment, la llibreria. Aquesta acció seria duta a terme amb la següent línia de codi:

```
import altair as alt
```

Tot seguit, un cop importades les dades, abans d'escriure el codi de les gràfiques, seria necessari executar la següent instrucció per a poder visualitzar conjunts amb més de 5.000 files.

```
alt.data_transformers.disable_max_rows()
```

A partir d'aquí ja es podrien codificar elements com el següent fragment de codi, que declara una relació entre les columnes desitjades a plotejar del conjunt de dades i els canals:

```
alt.Chart(dd[['descripcio', 'any_estada']]).mark_bar().encode(
      x = alt.X('descripcio:N', title = "Campus i estades", sort =
      alt.EncodingSortField('any_estada')),
      y = alt.Y('count():Q', title = "Nombre de participants"),
      color = alt.Color('any_estada:O', scale=alt.Scale(scheme='cat
      egory10'), legend=alt.Legend(title=['Anys'])),
).properties(
      title = "Sol·licituds d'inscripció en les estades lingüístiqu
      es al llarg dels anys"
)
```

El codi en concret dona lloc a la il·lustració 22 observada a l'apartat 4.2.4 Anàlisi de les visualitzacions – Inscripcions. Com es pot veure l'objecte **Chart** agafa el conjunt de dades com a argument. Més concretament, agafa només les columnes necessàries per a la gràfica concreta, augmentant d'aquesta forma l'eficiència. Tot seguit s'usa l'atribut **mark** per a declarar com es presentaran les dades. Aquesta paraula sempre va seguida del camp que indica el tipus de visualització com per exemple circle, point, line, text, bar o tick. El "bar" fet servir en aquest cas indica que serà un gràfic de barres. Posteriorment, dins de l'**encode** es relaciona cada canal de la gràfica amb una columna del conjunt de dades. Els canals més emprats durant el projecte han estat: x, y, color, size, opacity i facet. Finalment, gràcies a l'objecte **properties** es poden modificar els aspectes més estètics del gràfic. Dins d'aquest es pot veure tot sovint la declaració del títol a

partir de title o els camps resolve_scale i width/height. El primer permet lligar o deslligar la interacció dels canals de diferents gràfiques mentre els segons permeten fixar una mida.

A part dels objectes vistos en aquest tros de codi, també han estat altament usats al llarg del projecte transform_joinaggregate, transform_calculate i transform_filter. El primer permet agregar dades i realitzar càlculs sobre els grups creats mentre el segon possibilita realitzar càlculs, principalment matemàtics, sobre les columnes. L'últim permet filtrar dades que no es volen mostrar en el gràfic. Finalment, altres objectes altament utilitzats han estat hconcat o "|", que permeten mostrar dues visualitzacions concatenades horitzontalment, vconcat o "&", que permeten la mateixa concatenació, però en aquest verticalment i, layer, que sobreposa dos gràfics inicialment codificats com a entitats independents.

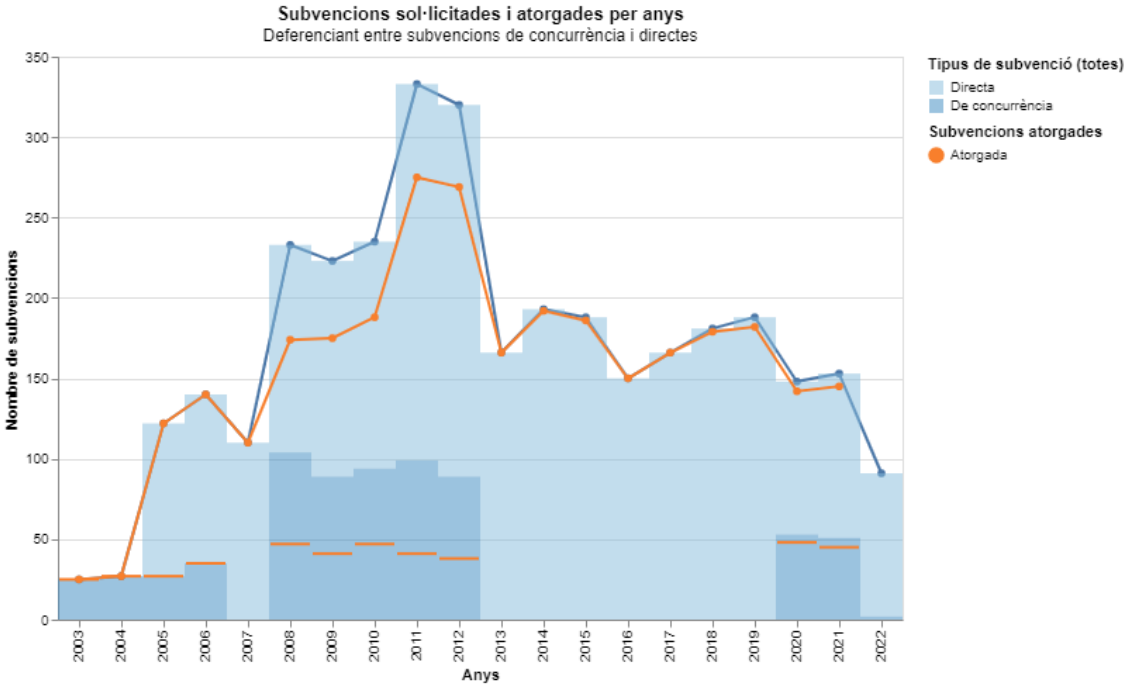
4.2 Anàlisi de les visualitzacions

Al llarg del projecte s'ha dut a terme múltiples visualitzacions i, per tant, no totes elles podran ser mostrades en aquesta secció. Tot seguit es podrà observar doncs una o dues visualitzacions, les considerades més rellevants, amb les seves explicacions, per cada apartat. La resta de gràfiques podran ser vistes junt amb les seves descripcions al GitHub del projecte. Tots els fitxers HTML les mostren, però els fitxers *_grafics.html han estat dissenyats expressament amb aquest propòsit mostrant únicament les gràfiques i les descripcions, evitant tant el codi com la neteja de dades prèvia. Finalment, també es comentarà el nombre exacte de gràfics codificats en cada apartat. Respecte a l'última afirmació cal tenir en compte que les visualitzacions que a partir del canal facet donen lloc a múltiples gràfiques s'han considerat com a una de sola.

4.2.1 Subvencions de llengua

Dins d'aquest primer apartat de subvencions s'ha creat 47 gràfics. Tot seguit, s'explicarà en detall el gràfic considerat més important pel que fa a les subvencions. Aquest també es mostrarà per a les altres àrees, d'aquesta manera no només es podrà comparar amb si mateix sinó també amb els valors de les altres àrees.

La il·lustració dona una informació general sobre quantes sol·licituds s'ha demanat a l'IRL al llarg dels anys i quantes se n'ha acabat atorgant dins de l'àrea de llengua. Així mateix, també informa sobre si aquestes eren directes o de concurrència gràcies a un gràfic de línies sobreposat amb un gràfic de barres.



Il·lustració 19: Subvencions sol·licitades i atorgades per anys en l'àrea de llengua

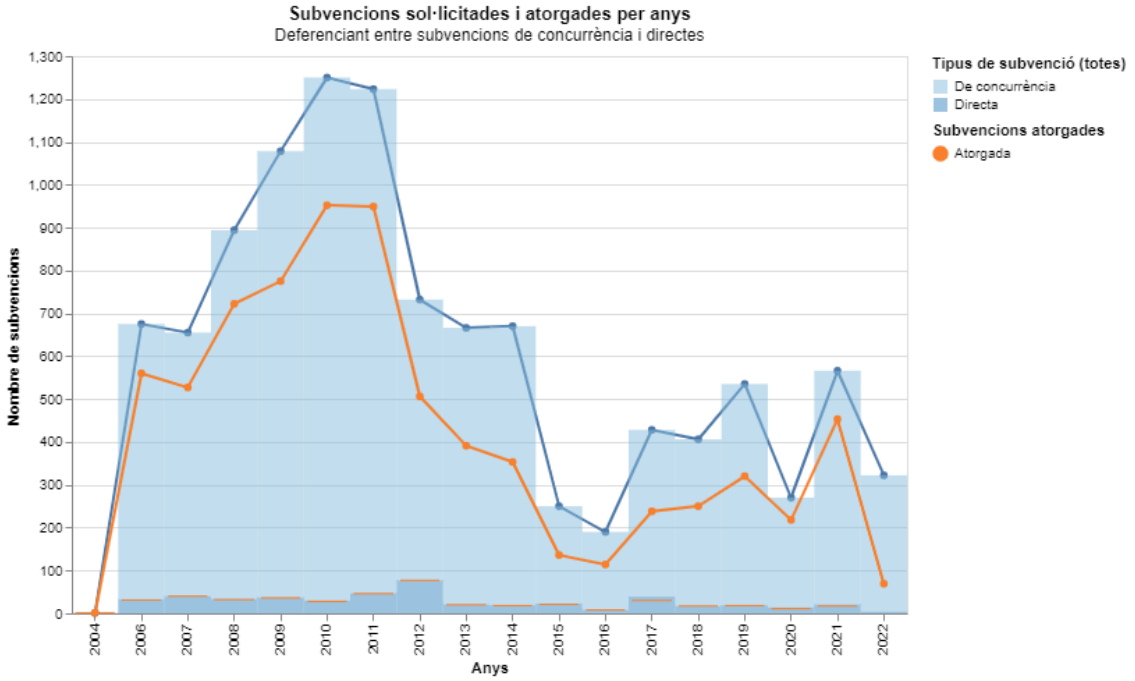
En concret, l'eix X mostra els anys durant els quals s'ha estat atorgant les subvencions (des del 2003 fins a l'actualitat, 2022), mentre l'eix Y codifica el nombre de sol·licituds. D'altra banda, la línia blava fa referència al nombre de sol·licituds demanades i la taronja al nombre de sol·licituds atorgades. Així mateix, les barres blau fosc indiquen les subvencions de concurrència mentre les blau clar les directes. Finalment, per ajudar a saber la proporció de subvencions atorgades/no atorgades de tipus directe/concurrent, s'ha incorporat un tic taronja a cada any. Aquest, mostra quantes subvencions de concurrència es van acabar atorgant respecte les sol·licitades.

En avaluar el gràfic es nota clarament com els anys de 2008 a 2012, que són els anys durant els quals hi va haver un major nombre de subvencions de competència competitiva, van ser els anys amb major nombre de sol·licituds demanades i al mateix temps atorgades. Tanmateix, aquests són també els anys amb una major diferència entre el nombre de sol·licituds atorgades i demanades, el qual es pot comprendre fàcilment pel fet que rarament les sol·licituds directes són denegades mentre les de concurrència ho són en diverses ocasions. A més a més, el fet que el tic taronja d'aquests anys es trobi aproximadament al mig de les barres de color blau fosc indica clarament que va ser culpa d'aquestes la diferència en nombre entre les dues categories.

D'altra banda, pel que fa als anys en què només s'ha estat atorgant subvencions directes, destaca l'any 2014 en ser el que mostra un major nombre de sol·licituds demanades i atorgades seguit molt de prop pel 2019. Finalment, cal comentar que l'any 2022 conté valors encara molt petits perquè aquest encara no s'ha acabat i, per tant, no s'han sol·licitat totes les subvencions. A més a més, cap de les demanades ha estat encara atorgada raó per la qual la línia taronja desapareix.

4.2.2 Subvencions de creació

En aquest segon cas de subvencions s'ha creat 46 visualitzacions. Posteriorment, com comentat anteriorment, s'analitzarà altre cop el gràfic de les subvencions demanades i atorgades per anys, però en aquest cas per a l'àrea de creació:



Il·lustració 20: Subvencions sol·licitades i atorgades per anys en l'àrea de creació

Com es pot observar aquest continua donant una informació general sobre quantes sol·licituds s'ha demanat a l'IRL al llarg dels anys, en aquest cas dins l'àrea de creació i, quantes se n'ha acabat atorgant. No s'explicarà tota la codificació perquè ja estat explorada anteriorment, però sí que s'advertirà que en aquest cas les barres blau fosc indiquen les subvencions directes mentre les blau clar les de concurrència, a diferència d'en el cas anterior, degut al major nombre de subvencions de concurrència que existeixen. Si s'hagués mantingut la codificació anterior, en ser la llargada de les barres blau fosc significativament superior, el blau clar hauria passat completament desapercebut. També, el tic mostra quantes subvencions directes es van acabar atorgant respectes les sol·licitades, no les concurrents. Finalment, l'eix X que mostra els anys durant els quals s'ha estat atorgant les subvencions, només va des del 2004 - només una subvenció atorgada - fins a l'actualitat, 2022.

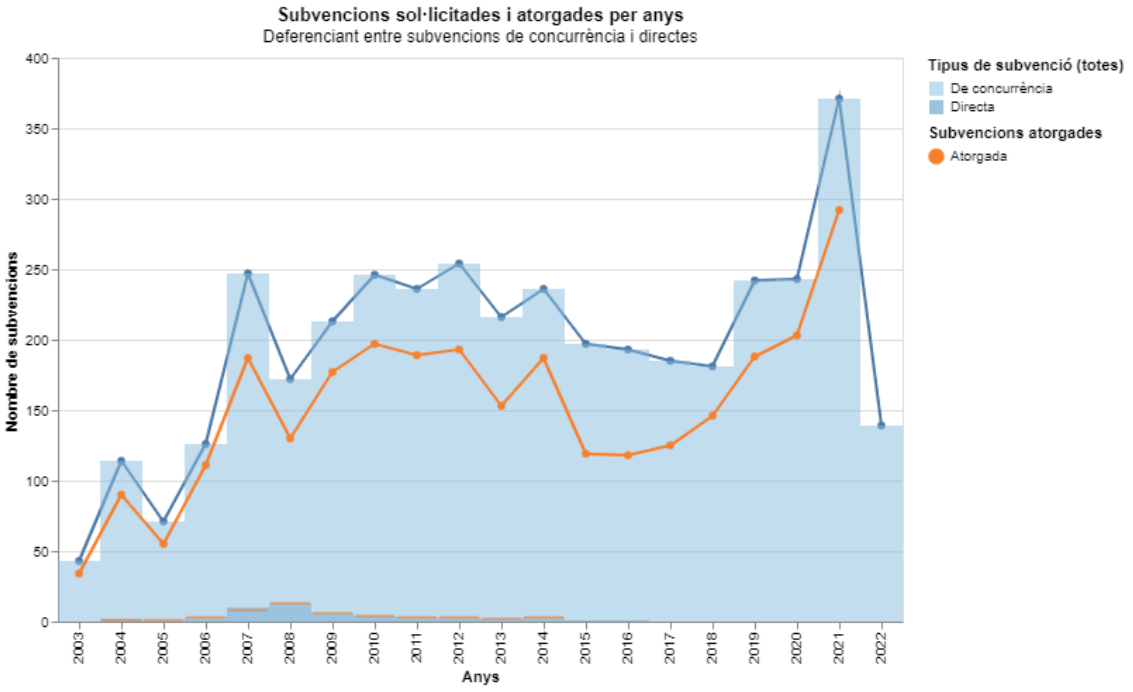
A l'hora d'analitzar-lo es comprova que així com les sol·licituds directes s'han mantingut aproximadament constants al llarg dels anys en uns valors d'entre 20 i 100, sent pràcticament totes atorgades, les de concurrència han patit més variacions. En aquest segon cas les sol·licituds van anar creixent de forma pràcticament lineal des del 2007 fins al 2010 el qual es pot assumir que va ser degut a l'augment de la popularitat d'aquestes. A partir d'aleshores, aquests valors van patir un dalt a baix fins al 2016, sobretot del 2011 al 2012 i del 2014 al 2015. Aquesta caiguda es pot explicar pel fet que l'any 2016 es van acabar de compactar els expedients de manera que els usuaris dins la línia de RETRO van poder demanar 1 única sol·licitud per a totes les activitats desitjades, en comptes d'haver-ne de demanar una per cada activitat. A partir d'aleshores els

valors es van recuperar en aparèixer moltes noves varietats (línies) i van tornar a augmentar a excepció de l'any 2020, degut a la COVID, on es van tornar a ajuntar algunes línies reduint d'aquesta manera el nombre de sol·licituds, però en general no el nombre de diners atorgats (com es pot observar en altres gràfiques).

Així mateix, si es comparen les subvencions sol·licitades amb les finalment atorgades, es pot advertir com aquestes últimes sempre s'han trobat entre 50 i 200 punts per sota. També s'analitza que els anys on hi ha una major distància entre aquestes dues categories són els anys on el nombre de sol·licituds demanades és més alt mostrant doncs que tot i que a més subvencions demanades, més atorgades, no s'acaba de poder fer front a totes elles. Convé acabar mencionant que l'any 2022 encara no mostra les dades de les subvencions directes perquè encara no s'han rebut així com tampoc mostra el nombre final de subvencions sol·licitades i atorgades perquè aquest encara no ha acabat.

4.2.3 Subvencions de literatura

Per a les subvencions de literatura 67 gràfics han estat necessaris per fer front a totes les preguntes i hipòtesis que s'havien plantejat els usuaris de l'IRL així com per a cobrir tota l'exploració de les dades que s'ha considerat necessària. Aquest nombre és molt més alt que els dos anteriors principalment degut a la profunda anàlisi que s'ha fet pel que fa a les subvencions de traducció. La gràfica visualitzada en el dos casos anteriors per a l'àrea de literatura es reconeix com:



Il·lustració 21: Subvencions sol·licitades i atorgades per anys en l'àrea de literatura

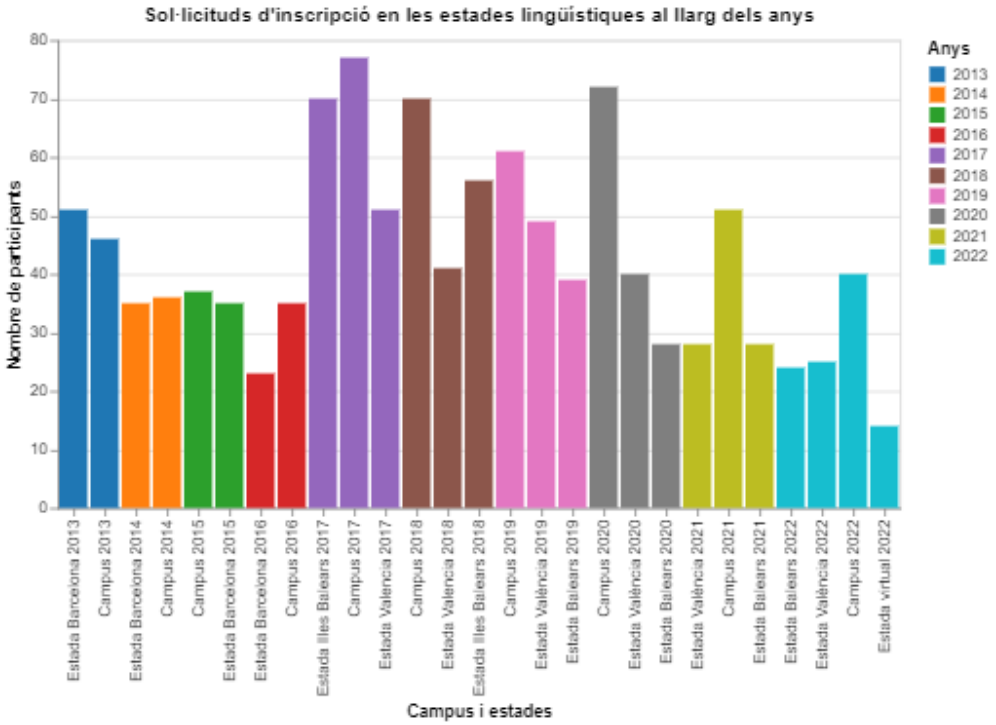
Com es pot observar la codificació d'aquesta gràfica és exactament la mateixa que es tenia per al gràfic de l'àrea de creació. En analitzar la visualització de forma individual es nota que així com durant els primers anys el nombre de sol·licituds rebudes era molt similar a les atorgades, al cap d'aquests s'ha anat diferenciant arribant a mostrar diferències de fins aproximadament 80

sol·licituds l'any 2015. Cal mencionar que l'any 2022 mostra un nombre molt petit d'instàncies demanades perquè quan les dades es van extreure encara no havien entrat totes al sistema, així com tampoc s'havien atorgat. En comparar-la amb les altres àrees s'arriba a la conclusió que és la que rep menys sol·licituds directes, de fet, aquestes han desaparegut per complet a partir de l'any 2014. També s'adverteix que és la segona amb més peticions, després de l'àrea de creació. Aquest últim fet es pot entendre perquè l'àrea de llengua és principalment composta per subvencions directes, que són moltes menys en nombre que les subvencions de concurrència que es reben en formalitzar convocatòries.

Tot seguit, cal comentar que se sap que la pujada del 2008 al 2009 va ser deguda a la presència de l'IRL a la Fira del Llibre de Frankfurt, Alemanya. Finalment, cal fer notar que el percentatge de sol·licituds directes atorgades sembla ser aproximadament del 100%, assumint doncs que totes les denegacions es donen en subvencions de concurrència. Aquest fet no causa estupefacció en ser una conseqüència inherent dels dos tipus de subvencions. Recordar que en les subvencions de concurrència els usuaris/entitats que les sol·liciten han d'aplicar dins d'un termini especificat per l'Institut i competir amb altres candidatures, mentre en el cas de les directes són les entitats (generalment importants) qui per voluntat pròpia detallen la seva situació al Llull i aquests els atorguen els diners sense haver d'entrar dins de cap competició.

4.2.4 Inscripcions

Pel que fa als campus i estades, l'anàlisi de les sol·licituds rebudes només ha portat a fer 21 visualitzacions. En aquest cas altre cop s'analitzarà únicament la primera visualització duta a terme per a aquest conjunt, referent a la secció "Estada o campus més realitzat", per a posar al lector en situació de les dades amb què es tracta:



Il·lustració 22: Sol·licituds d'inscripció en les estades lingüístiques al llarg dels anys

En aquesta s'ha usat l'eix horitzontal per a indicar el nom de l'estada o campus. Aquest nom indica tant l'any en què s'ha portat a cap (ordenat per aquest) com el tipus d'aquesta. En segon lloc, l'eix vertical indica el nombre de sol·licituds. Finalment, el color de les barres fa referència als anys en què aquests campus i estades han tingut lloc.

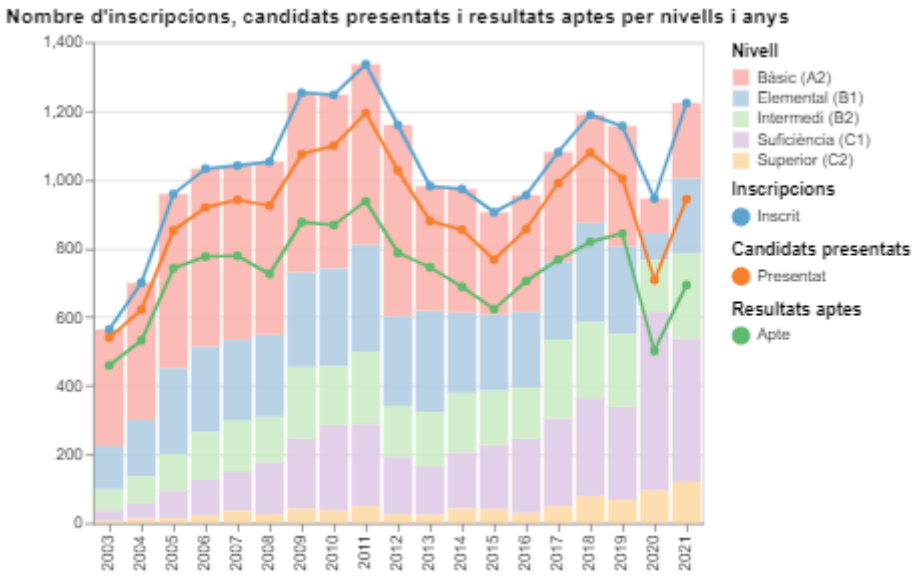
A l'hora d'examinar la gràfica s'adverteix primerament que fins l'any 2016 només es va realitzar 1 campus i 1 estada a Barcelona a l'any mentre a partir d'aquest l'estada es va dividir en dos: Illes Balears i València, desapareixent la de Barcelona. A més a més, l'any 2022 ha afegit una nova varietat d'estades, la virtual. Aquesta última estada fa referència a una activitat feta en comptes de presencialment de forma telemàtica perquè usuaris de l'exterior no hagin d'anar físicament a València o a les Illes Balears. Després de parlar amb els usuaris de l'àrea es va saber que es va pensar que seria una bona idea la introducció d'aquesta nova modalitat després de les experiències viscudes durant la COVID. Tanmateix, observant els resultats preliminars que es tenen del 2022 (tenint en compte que al moment d'extreure les dades el període d'inscripció no havia acabat), s'assenyala que possiblement no és tan bona idea tenint en compte el baix nombre de sol·licituds a aquesta.

Posteriorment, es veu que només 4 activitats superen els 65 usuaris apuntats, dues d'elles l'any 2017. Finalment, s'afirma que no es pot designar quina activitat és més popular, si el campus o alguna de les dues estades, en anar-se intercanviant el títol de "màxim nombre de participants" al llarg dels anys.

4.2.5 Certificació

Pel que fa a les activitats que avaluen i certifiquen els coneixements de la llengua catalana 40 diagrames han estat necessaris per a mostrar tota la informació sol·licitada pels usuaris de l'IRL. En aquest cas abans de posar l'exemple indicar que les visualitzacions s'han dut a terme en un script diferent d'on s'havia portat a cap la neteja de dades. Això és degut al fet que en un inici no se sabia que era millor seleccionar les columnes necessàries per a cada gràfica per a millorar l'eficiència i, per tant, en carregar-se totes les dades en cada gràfica el sistema deixava de funcionar per problemes de memòria. Això no passava en els altres casos a causa del fet que certificació té el conjunt de dades més complet amb què s'ha codificat gràfiques durant aquest projecte. Tot i que tal com està ara ja seria factible que les gràfiques es poguessin trobar junt amb la neteja, s'ha deixat en dos guions separats per a no induir a possibles errors i córrer riscs innecessaris.

Tot seguit es mostrarà altre cop la primera visualització d'aquest apartat per posar en situació a l'usuari, que en aquest cas mostra el nombre d'inscrits, presentats i aprovats als exàmens de certificació de forma anual. No obstant això, en aquest cas a continuació es mostrarà la mateixa informació dividida per nivells dels exàmens, per a mostrar la gran influència que tenen aquests sobre els valors indicats.



Il·lustració 23: Nombre d'inscripcions, candidats presentats i resultats aptes per nivells i anys

Aquesta primera visualització mostra el nombre (eix Y) de candidats inscrits (línia blava), finalment presentats (línia taronja) i aprovats (línia verda) per any (eix X). D'altra banda, el gràfic de barres sobreposat indica el nombre d'usuaris que es van inscriure a cada nivell al llarg dels anys. A l'hora d'analitzar-la, s'observa fàcilment com l'any 2011 va ser l'any en què es va realitzar més inscripcions i que també hi va haver més aprovats. En contraposició, els anys 2003 i 2004 són els que mostren un menor nombre d'inscrits, explicable per la no gran popularitat d'aquest servei als inicis, així com pel fet que aquests dos anys només hi va haver una convocatòria. Relacionat amb aquest tema, també es nota com l'any 2020 mostra un pic inferior. Aquest es pot explicar sabent que l'any de la COVID es va fer una excepció en portar a cap només 1 convocatòria.

Tot seguit, es pot advertir que en general com major és el nombre de persones presentades de forma anual, més diferència hi ha entre aquest i el nombre d'aprovats. Finalment, pel que fa als nivells es manifesta com a mesura que passen els anys el nivell més comunament dut a terme passa de ser el bàsic a ser el de suficiència, mentre el superior ha estat en tot moment el menys demanat.

Com comentat anteriorment, tot seguit es mostra la mateixa informació però codificada per als cinc nivells de forma separada:

Il·lustració 24: Nombre d'inscripcions als exàmens per anys i nivells

(indicant el nombre de presentats i aptes finals)

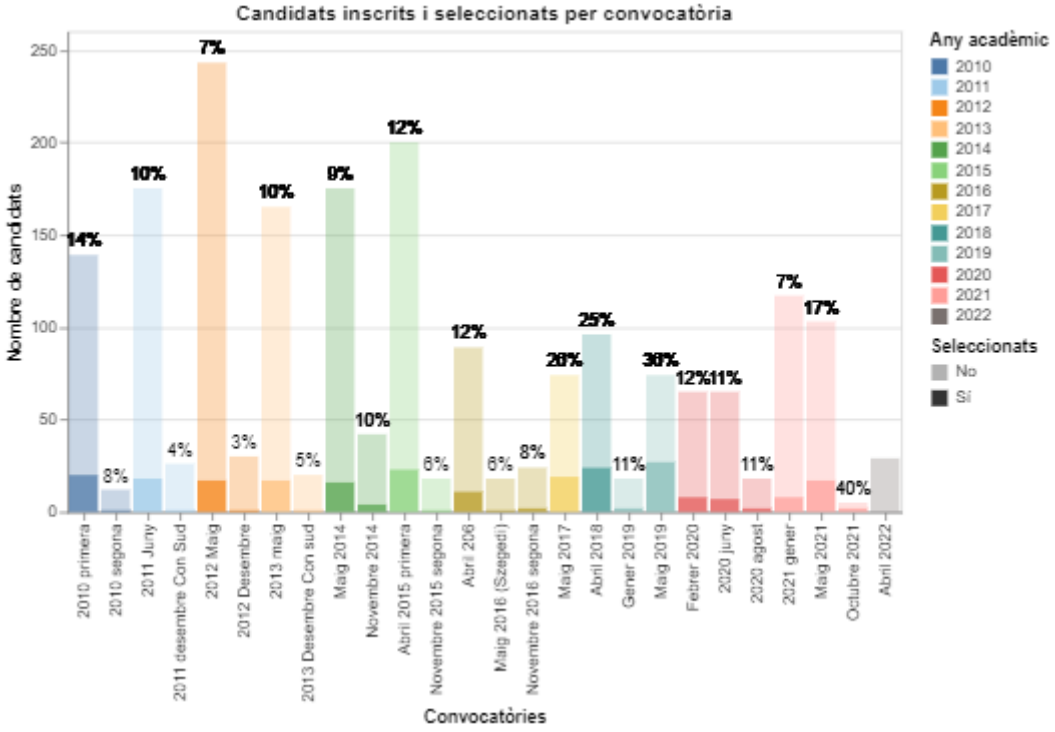
Si s'analitzen inicialment les gràfiques de forma conjunta, es pot comprovar com el nivell bàsic és el que mostra un valor major d'inscrits en 1 sol any sent aproximadament de 550 el 2012. Aquest màxim es troba seguit de ben a prop pel nivell suficiència que mostra un valor de 520 candidats inscrits aproximadament l'any 2020. De fet, si es pensa amb consciència, aquest màxim del nivell suficiència és encara més sorprenent en haver-lo aconseguit l'any 2020, un dels anys amb menys inscrits totals dels últims anys com comentat anteriorment degut a la COVID.

Subseqüentment, si s'examina gràfic per gràfic, es nota clarament la tendència decreixent en la primera visualització, la del nivell bàsic. Així mateix, tal com enunciat anteriorment es nota altre cop que com major és el nombre de presentats, major és el rang mostrat respecte al nombre d'aptes. Si es mou l'atenció al gràfic del nivell elemental, s'adverteix que el nombre d'inscrits ha estat molt estable al llarg dels anys movent-se la majoria de valors dins d'un rang de 85 punts, del 225 a 310, aproximadament (obviant els anys on només hi va haver 1 convocatòria).

Posteriorment, es manifesta que els tres gràfics de nivell superior mostren tots ells tendències creixents pel que fa al nombre d'inscrits (tot i que l'intermedi de forma més tímida). Finalment, cal assenyalar que el nivell suficiència és el que mostra una major diferència entre el nombre de persones presentades a un examen i el nombre de persones aptes de forma continuada al llarg dels anys. Aquest fet du a pensar que és un nivell normalment dut a terme perquè el demanen en alguna feina o oposició, però que generalment els examinands no estan preparats quan s'hi presenten.

4.2.6 Selecció de professorat

Tot seguit pel que fa a l'apartat de selecció de professors per a les universitats per mantenir la docència han estat necessàries 35 visualitzacions per a respondre totes les preguntes i hipòtesis plantejades pels usuaris de l'IRL. A més a més, també s'ha creat dos conjunts de gràfiques especials que mostren les puntuacions atorgades als diferents candidats. Totes aquestes gràfiques, tant les puntuacions com les 35 visualitzacions inicialment mencionades poden ser visualitzades a seleccio_grafics.html. Tot seguit cal comentar que la majoria de gràfiques en aquest cas no han estat dutes a terme tant per a explorar les dades sinó per a comparar la informació dels candidats que han estat seleccionats vs els que no. Posteriorment, es visualitza la primera gràfica d'aquest conjunt, que altre cop posa en context de les dades a analitzar:



Il·lustració 25: Candidats inscrits i seleccionats per convocatòria

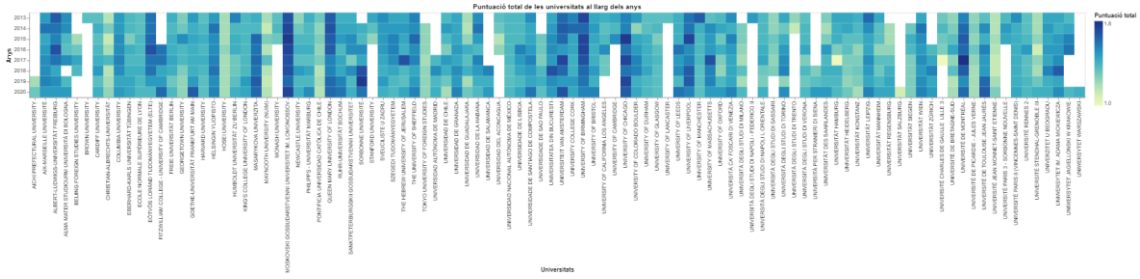
Amb aquesta s'ha volgut obtenir una visió general de quantes convocatòries s'ha dut a terme al llarg dels anys, quants usuaris han sol·licitat alguna plaça en cadascuna d'elles i quants han acabat sent seleccionats. Tenint en compte aquest propòsit s'ha creat un gràfic de barres on l'eix X mostra totes les convocatòries que hi ha hagut al llarg dels anys (color) de forma ordenada. D'altra banda, l'eix Y fa referència al nombre de candidats mentre amb l'opacitat s'indica si aquests han estat seleccionats (fort) o no (clar). Finalment, el text escrit s'ha emprat per a indicar el percentatge de seleccionats dins de cada convocatòria.

En analitzar la visualització es nota que la majoria d'anys hi ha hagut dues convocatòries: la primera gran i la segona petita. En les de major volum generalment el percentatge de candidats seleccionats és major que en les de menor volum tot i que en totes elles aquest és molt baix trobant-se entre el 5-15% aproximadament, a excepció de 4 convocatòries que superen el 25% de seleccionats. També s'adverteix com aquest percentatge no es mostra per a la convocatòria de l'any 2022 perquè en el moment d'extreure les dades s'havien fet les sol·licituds, però encara

no s'havia pres cap decisió. Finalment, es manifesta com amb el pas dels anys es tendeix a portar a cap més convocatòries per any (3) rebent menys sol·licituds dins d'aquestes.

4.2.7 Memòries

Finalment, per als justificants de les subvencions atorgades a les universitats, és a dir, les memòries, s'ha creat 100 diagrames. En aquesta secció en comptes de mostrar la primera gràfica del guió, com en tots els casos anteriors, es mostrarà les puntuacions atorgades per l'IRL a les diferents universitats que han portat a cap les justificacions al llarg dels anys. D'aquesta manera, es podran observar tant els anys durant els quals s'ha realitzat memòries, com el nombre d'universitats que n'han fet cada any. Així mateix, també es podrà saber quina puntuació li ha atorgat el Llull a aquestes de forma anual i, per tant, quina és la que universitat més ben considerada tenint en compte els seus estàndards.



Il·lustració 26: Puntuació total de les universitats al llarg dels anys

Com es pot veure en l'eix Y hi ha universitats que han estat fent justificants des del 2013, mentre només es tenen dades completes per a totes elles fins al 2020, perquè aquestes es duen a terme un any després de la finalització de l'any. Així mateix, també es nota gràcies al nombre d'universitats de l'eix X que aproximadament 80 universitats han participat en algun moment en la creació d'alguna memòria. D'altra banda, el rang de la intensitat del color groc – blau només va de l'1 a l'1,8 aproximadament, confirmant doncs que a totes les universitats se'ls a atorgat ponderacions dins d'aquest rang, tot i que se sap que es ponderen dins d'una escala que va del 0 al 2.

A primera vista es nota que la Universitat de Moskovski és la que ha obtingut millors resultats al llarg dels anys i, per tant, la considerada top 1 dins de l'IRL. Altres universitats amb bones puntuacions al llarg dels anys han estat la Queen Mary, Birmingham, Chicago, Liverpool i Mont-Real. Si es compara aquest rànquing de classificació amb el rànquing global portat a cap per usuaris externs de l'IRL (visualitzat en una altra gràfica d'aquest projecte) es nota que mentre per les entitats externes les universitats dels Estats Units d'Amèrica són més ben considerades, per a l'IRL les universitats del Regne Unit són millor.

Les 4 millors puntuacions aconseguides al llarg de tots els anys en què s'ha dut a terme memòries han estat:

- La Universitat de Mont-Real l'any 2017 amb una puntuació d'1,82
- La Universitat de Moskovski l'any 2015 amb una puntuació d'1,79
- La Universitat de Chicago l'any 2015 amb una puntuació d'1,79
- La Universitat de Sorbona l'any 2019 amb una puntuació d'1,79

5 Conclusions

El propòsit inicial del projecte era l'observació de set subconjunts de dades:

1. Les subvencions de l'àrea de llengua, creació i literatura
2. Les sol·licituds realitzades a les estades lingüístiques en territoris de parla catalana
3. Les inscripcions i candidats aptes en exàmens de certificació
4. Les sol·licituds per a la selecció de professorat a les universitats
5. La justificació de les subvencions atorgades a les universitats

Tots aquests es pretenien analitzar a partir de gràfics per a ajudar a treure estadístiques rellevants per tal que els usuaris interns de l'Institut Ramon Llull poguessin dur a terme accions basades en nombres comparables, repetibles i comprensibles. Basant-se en el mencionat a l'apartat anterior, es pot afirmar que aquest s'ha complert amb escreix. Un exemple concret d'aquest fet es podria donar mirant les visualitzacions creades per a l'àrea de selecció del professorat on gràcies a aquestes se sap quins són els factors més rellevants que fan que un usuari sigui seleccionat o no per a cobrir les vacants.

Tanmateix, com en tots els projectes diversos elements podrien ser millorats. En primer lloc, cal mencionar que si s'hagués dut a terme aquest projecte coneixent des d'un principi la totalitat de les bases de dades que s'anava a utilitzar i les taules que aquestes contenien, hauria estat no només ràpid sinó més efectiu computacionalment, en haver estalviat l'extracció de variables que no aportaven informació, per exemple. En segon lloc, la gran quantitat de temps que s'ha hagut d'invertir en la neteja d'aquestes ha estat determinant. Si aquest s'hagués pogut reduir encara que fos a la meitat, no només hauria estat possible la realització de més gràfiques, sinó la millora de les ja existents.

Tot seguit pel que fa al contingut de les visualitzacions cal declarar que si s'hagués conegut prèviament que aquest projecte s'anava a donar, l'IRL clarament hauria recollit algunes dades de diferent manera. De fet, diverses gràfiques desitjades pels usuaris interns de l'IRL, no s'han pogut crear perquè les dades necessàries per a fer-ho no es trobaven a la BD ni es podien imputar. S'ha advertit que cada cop que ells volien extreure aquesta informació la codificaven de forma manual, però això només es pot fer si consideres entre 20 i 30 sol·licituds, és a dir, una convocatòria particular, no si es miren les dades d'entre 8 i 18 anys enrere. Posteriorment, pel que fa a la planificació la realització de les reunions inicials 3 setmanes abans del que es van donar també hauria ajudat molt a concentrar els esforços en el desitjat des d'un primer moment.

No obstant això, cal expressar que el mètode d'anàlisi, és a dir, la utilització de la llibreria Altair per a crear els gràfics a partir del llenguatge de programació Python ha estat un gran encert, en permetre fer noves visualitzacions de forma molt automàtica i elegant un cop es tenien totes les dades ben formatades. Òbviament com tot llenguatge de programació té algunes restriccions, però la majoria de cops s'ha trobat solucions per a superar les dificultats.

Finalment, cal constatar altre cop que les més de 300 visualitzacions creades al llarg d'aquest projecte, amb les seves corresponents explicacions, satisfan les peticions mencionades pels usuaris interns de l'Institut Ramon Llull, l'organització beneficiària.

6 Passos futurs

Tot seguit es llistaran un conjunt de passos que si es duguessin a terme millorarien aquesta primera versió del projecte d'anàlisi de dades de l'Institut Ramon Llull. En primer lloc, es comentaran alguns elements de l'extracció de dades inicial que, si haguessin estat duts a terme de diferent manera, hauria desembocat en una millora en l'àmbit computacional del projecte en general. En segon lloc, com comentat anteriorment al llarg del projecte s'ha tingut present la voluntat de detectar elements que es podrien haver analitzat si s'haguessin recollit les dades per a comentar-ho amb els usuaris de l'IRL. Aquestes també s'enumeraran. Finalment, es llistaran algunes gràfiques no realitzades, però que si es creessin millorarien l'abast.

Començant doncs per les millores en l'extracció de les dades inicials, cal tenir en compte, com comentat anteriorment a la conclusió, que si el projecte s'hagués dut a terme coneixent des d'un principi la totalitat de les bases de dades que s'anava a utilitzar i les taules que aquestes contenien hauria estat més efectiu a escala computacional. Això es pot explicar amb el fet que diverses variables com "*cobrament*", "*ajut_viatge*" o "*ordre_pagament*", totes elles contingudes en els conjunts de subvencions, han estat extretes i tractades, però en cap cas visualitzades a causa del poc interès que aquestes tenen per part dels usuaris del Llull. En un futur projecte es podria doncs eliminar completament la seva presència des d'un principi. Un segon exemple que hauria millorat el projecte tant en l'àmbit computacional com en el de l'organització seria el diferent tractament dels conjunts de les dades de memòries. L'extracció d'aquestes dades ha estat feta de forma conjunta agafant no només les dades de les memòries, sinó de les produccions, assignatures i activitats relacionades amb aquestes. No obstant això, al final cap dels gràfics ha requerit l'encreuament d'aquests subconjunts i, per tant, una extracció dividida per a cada element hauria estat més senzilla i efectiva.

Tot seguit es mostrarà la llista d'aquells elements que s'haurien d'incloure a la base de dades per a millorar qualsevol futura versió d'aquest projecte. Així mateix, també s'inclouran aquells elements que ja estan sent recollits, però potser no de la millor manera.

1. En primer lloc, la variable "*importrevocat*" dels subconjunts de subvencions ja està sent recollida. No obstant això, en la majoria dels casos no es troba ben indicada, tot i ser un dels factors que l'IRL volia analitzar.
2. En segon lloc, la comunitat autònoma dels candidats de selecció de professorat s'hauria de recollir, al ser atractiu per l'IRL l'anàlisi del nombre de candidats presentats i finalment seleccionats a partir d'aquesta agrupació.
3. Tampoc s'ha considerat correctament codificada la variable "*exclos*" de selecció de professorat, tot i ser-hi present a la BD.
4. També per al mateix conjunt de dades, en comptes de voler conèixer les poblacions dels usuaris, que no interessen, s'hauria de dur a terme la següent divisió geogràfica: "València", "Balears", "Catalunya", "Resta de l'estat espanyol" i "Resta del món". De fet, un error comú de les bases de dades de l'IRL, no només de la de selecció, és la presència tant del país de naixement com del codi postal i població actuals de l'usuari en qüestió, mentre no es creen els grups mencionats anteriorment, quan és el que es vol analitzar. Un pensaria que si es té el codi postal i el país, ja es pot obtenir la categorització

desitjada, però això no és del tot correcte, en necessitar el codi postal un país associat (no el de naixement, sinó l'actual) per a poder crear els subconjunts correctament.

5. Continuar dient que el màxim de diners que es poden atorgar a una convocatòria, que teòricament és una informació ja recollida, s'hauria de codificar correctament. Aquest valor permetria analitzar si el fet d'oferir més diners dona lloc a un increment del nombre de sol·licituds o no.
6. Finalment, el sexe un dels elements que més s'ha demanat de conèixer, no es troba informat en diversos conjunts i si s'hi troba, no està omplert en la majoria dels casos. Aquest factor seria prou fàcil d'arreglar amb la simple introducció de la pregunta corresponent als formularis amb què tramiten cada element.

Finalment, cal comentar dos aspectes que també s'haurien pogut visualitzar i no s'ha fet, però en aquest cas no per la incorrecta codificació de les variables o la seva inexistència. En primer lloc, es podria comprovar si el fet que hi hagi un canvi en la metodologia dona lloc a una disminució o increment del nombre de persones apuntades a una activitat o assignatura, pel que fa al subconjunt de les memòries. La creació d'aquest gràfic podria requerir l'encreuament dels diferents subapartats de les memòries i, en conseqüència, la correcció a l'hora de dur a terme l'extracció de les dades de memòries anteriorment mencionada, ja no es podria fer. Finalment, també milloraria l'abast del projecte creuar les dades de les subvencions de traducció de literatura amb les del TRAC, que és una base de dades que conté totes les traduccions de la literatura catalana a altres llengües. Aquest encreuament hauria permès analitzar el nombre de traduccions dutes a terme a l'IRL dins del total d'obres traduïdes a Catalunya, entre d'altres.

7 Referències

1. *Universitat Politècnica de Catalunya – BarcelonaTech.* Universitat Politècnica de Catalunya. https://www.upc.edu/ca
2. *Institut Ramon Llull*. Institut Ramon Llull. https://www.llull.cat/catala/home/index.cfm
3. *Grau en Ciència i Enginyeria de Dades*. Universitat Politècnica de Catalunya. https://dse.upc.edu/ca
4. (2022). *Python*. Python. https://www.python.org/downloads/
5. (2020). *Altair: Declarative Visualization in Python.* Altair Developers. https://altair-viz.github.io/
6. (2022). *Fundamentos de la metodología Agile*. Wrike. https://www.wrike.com/es/project-management-guide/fundamentos-de-la-metodologia-agile/
7. (20 de setembre de 2021). *Qué es SCRUM*. Proyectos Ágiles. https://proyectosagiles.org/que-es-scrum/
8. (2022). *Access SQL: conceptos básicos, vocabulario y sintaxis*. Microsoft. https://support.microsoft.com/es-es/office/access-sql-conceptos-b%C3%A1sicos-vocabulario-y-sintaxis-444d0303-cde1-424e-9a74-e8dc3e460671
9. (2022). *SQL Tutorial*. Refsnes Data - W3Schools (W3.CSS). https://www.w3schools.com/sql/
10. (2022). *MobaXterm*. MobaXterm. https://mobaxterm.mobatek.net/
11. Ansgar, B. *HeidiSQL*. HeidiSQL. https://www.heidisql.com/
12. Vázquez, P.P. (setembre de 2020). *Visualitzation. Perception.* Pàg 16-33. Diapositives de la classe "Visualització de la Informació" del GCED
13. (10 d'abril de 2019). *7 Gestalt principles of visual perception: cognitive psychology for UX. UserTesting*. https://www.usertesting.com/blog/gestalt-principles
14. *Gestalt principles.* Interaction Design Foundation. https://www.interaction-design.org/literature/topics/gestalt-principles
15. *Te damos la bienvenida a Colab*. Google. https://colab.research.google.com/notebooks/intro.ipynb
16. (2022). *Numpy.* Numpy. https://numpy.org/
17. (2021). *Matplotlib: visualitzation with Python.* The Matplotlib Development Team. https://matplotlib.org/
18. Rodrigues, I. (17 de febrer de 2020). *CRISP-DM methodology leader in data mining and big data*. Towards Data Science. https://towardsdatascience.com/crisp-dm-methodology-leader-in-data-mining-and-big-data-467efd3d3781
19. Wirth, R; Hipp, J. *CRISP-DM: Towards a Standard Process Model for Data Mining*. http://www.cs.unibo.it/~danilo.montesi/CBD/Beatriz/10.1.1.198.5133.pdf
20. (2022). Stack Overflow. https://stackoverflow.com/questions
21. (2 d'abril de 2022). *Pandas*. Pandas_dev (NumFOCUS). https://pandas.pydata.org/

8 Apèndixs

8.1 Document “Fita inicial”

Anàlisi estadístic de les bases de dades de l’Institut Ramon Llull

Proposta de projecte i pla de treball

ESCRIT PER:

Nom: Anna Patrícia Orteu

Data: 16/02/2022

REVISAT I APROVAT PER:

Nom: Francesc Rey Micolau

Data:

Visió general del projecte i objectius

El projecte és desenvolupat a l'Institut Ramon Llull (IRL). L'objectiu principal d'aquest és analitzar una gran part de la base de dades de l'IRL, de la qual mai n'han tret profit, per tal de poder millorar les seves ofertes de cara a l'usuari final i augmentar d'aquesta manera la seva popularitat.

Dins de l'IRL existeixen varies àrees que ofereixen tot tipus de subvencions, exàmens oficials de català, campus o estades a diverses universitats, etc. Tanmateix, aquest projecte només es centrarà en 3 d'aquestes àrees, l'àrea de CREACIÓ, la de LLENGUA I UNIVERSITATS i l'àrea de LITERATURA.

Dins de l'àrea de CREACIÓ, només s'analitzaran les SUBVENCIONS. Aquesta anàlisi té com a principal objectiu donar respostes a preguntes del tipus: "Quina és la subvenció que té més revocacions?" o "Quina és la subvenció més popular i quins elements podrien donar lloc a aquesta popularitat?".

D'altra banda, dins de l'àrea de LITERATURA, també s'analitzaran només les SUBVENCIONS. Tanmateix, aquesta anàlisi serà més complexa degut a que estarà lligada a les traduccions dels llibres i les subvencions que es demanen per a fer-les. Així doncs, els objectius es trobaran alineats amb els anteriorment mencionats afegint preguntes com "Quins són els llibres que es tradueixen més i a quins idiomes?".

Finalment, dins de l'àrea de CREACIÓ s'analitzaran quatre camps:

1. Les SUBVENCIONS ofertes per aquesta àrea en concret, on les preguntes a respondre es troben altre cop alineades amb les mencionades anteriorment.
2. Els exàmens de català oferts per l'Institut, el qual rep el nom de CERTIFICACIÓ dins de l'IRL. Dins d'aquest subtema, es voldrà analitzar, per exemple, si les notes obtingudes en aquests exàmens es poden relacionar amb les hores empleades en estudiar el català per part dels participants, la llengua materna o el lloc de realització de l'examen, entre d'altres.
3. Les MEMÒRIES realitzades pels usuaris a qui se'ls ha concedit una subvenció. Aquestes són justificants de subvenció, les quals tenen com a objectiu principal demostrar que s'ha ensenyat català durant el període de la subvenció així com avaluar l'experiència. En aquest cas, l'anàlisi estarà més enfocada a saber si després de realitzar algun dels cursos els usuaris s'acaben presentant a alguna prova de certificació i a saber quina és la universitat amb la qual el Llull té més relació.
4. Les INSCRIPCIONS de l'IRL. Aquestes, fan referència a totes les estades, campus, cursos o seminaris que ofereix l'Institut. Així doncs, en aquest cas els interessarà saber elements com en quina universitat es realitzen més inscripcions, si hi ha alguna raó per a que això sigui així o quina és l'activitat més ben puntuada i com es podria promocionar, si els interessa.

Antecedents del projecte

D'una banda, mencionar que el projecte parteix des de zero, sent jo mateixa, Anna Patrícia Orteu, qui vaig proposar l'oportunitat de realitzar-lo a l'organització gràcies a la meva relació prèvia amb l'Institut Ramon Llull degut a unes pràctiques extra curriculars realitzades durant els darrers 2 anys (entre 2020 i 2022). Així doncs, aquest és totalment independent respecte qualsevol altra anàlisi que pugui estar sent realitzada a l'empresa en aquest moment. No obstant això cal mencionar que no em consta que cap altre equip estigui fent alguna anàlisi similar, ja que els usuaris de l'Institut sempre havien volgut realitzar algun projecte d'aquest tipus però ningú ho havia fet per falta de temps i coneixements.

En segon lloc, comentar que la idea inicial de realitzar el projecte com mencionat breument va ser de l'autora del projecte, és a dir, jo mateixa. Tanmateix, tot i proposar alguna idea inicial per a que l'organització sàpigues a quin tipus de preguntes o problemes podria donar resposta, la majoria de elements desitjats d'analitzar han estat mencionats per usuaris interns de l'Institut, ja que són ells els que coneixen el sistema i qui saben quins són els punts on l'organització desitjaria millorar.

Pla del treball

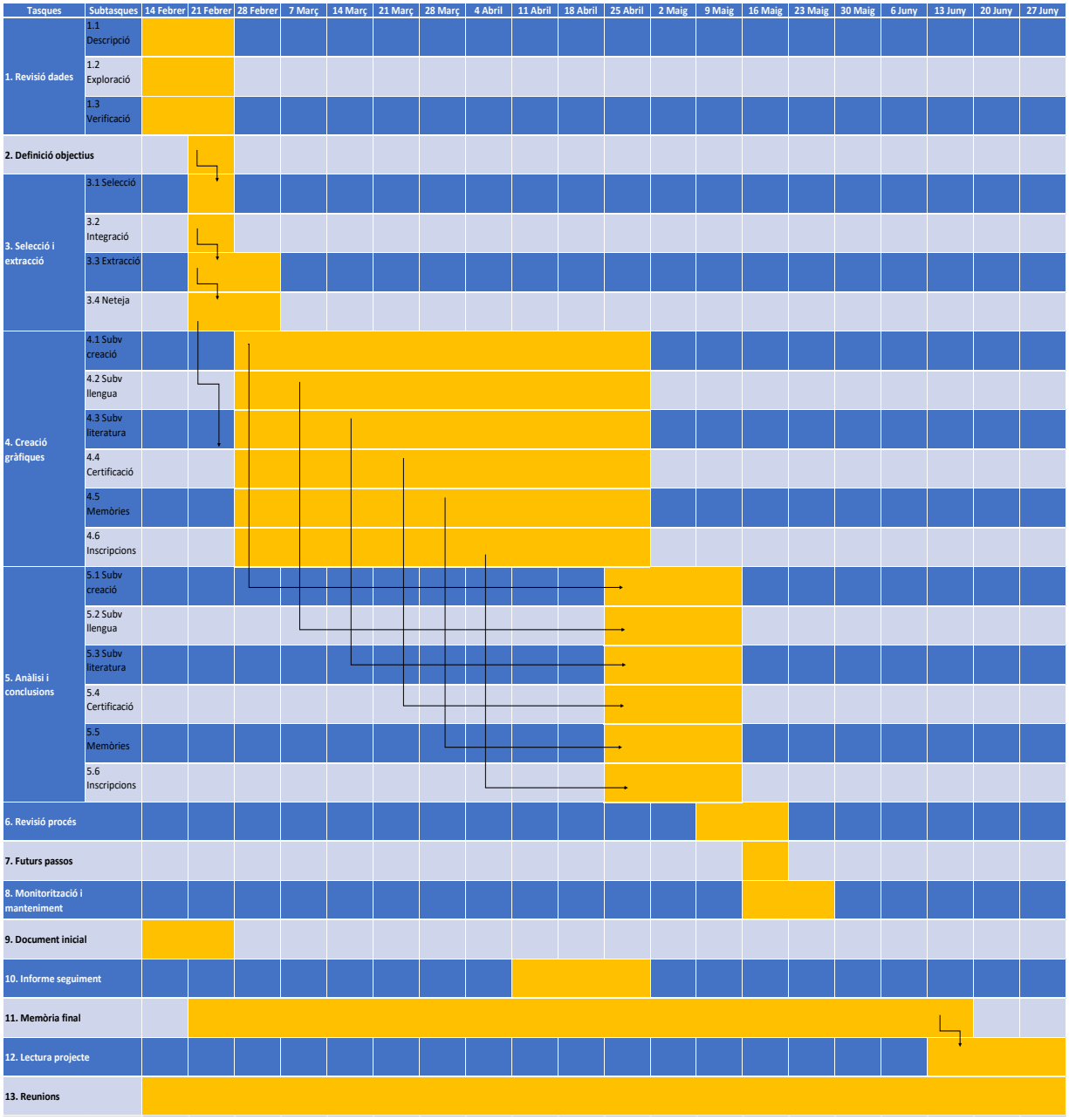
Tasques i fites – Diagrama de Gantt

A continuació, es mostra un conjunt de tasques amb els corresponents objectius o fites del projecte. La gràfica també mostra els períodes de temps de cada secció així com les subtasques incloses en cadascuna d'elles:

Tasca 1: **Revisió de les dades de l'Institut Ramon Llull – avaluació de la situació**	Subtasca 1.1: Descripció de les dades	14/02/2022 – 21/02/2022	Word per a cada àrea mencionant les dades amb les que es pot tractar i exemples d'elements que es podrien analitzar
	Subtasca 1.2: Exploració de les dades		
	Subtasca 1.3: Verificació de la qualitat		
Tasca 2: **Definició dels objectius finals a través de les reunions**		21/02/2022 - 25/02/2022	Document que especifiqui tots els objectius que pretenen assolir amb el projecte els usuaris de l'IRL
Tasca 3: **Selecció i extracció de les dades a analitzar**	Subtasca 3.1: Selecció de les dades	26/02/2022	Selects que s'ha de realitzar contra les diferents bd del Llull
	Subtasca 3.2: Integració de les dades		
	Subtasca 3.3: Extracció de les dades	27/02/2022 – 28/02/2022	Excels completament netejats preparats per a fer l'anàlisi
	Subtasca 3.4: Neteja de les dades		
Tasca 4: **Creació de les gràfiques/elements per analitzar**	Subtasca 4.1: Subvencions de creació	01/03/2022 – 30/04/2022	Obtenir totes les gràfiques o elements necessaris que permetin extreure les conclusions
	Subtasca 4.2: Subvencions de llengua		
	Subtasca 4.3: Subvencions de literatura		
	Subtasca 4.4: Certificació		
	Subtasca 4.5: Memòries		
	Subtasca 4.6: Inscripcions		
Tasca 5: **Anàlisi i extracció de conclusions**	Subtasca 5.1: Subvencions de creació	25/04/2022 – 10/05/2022	Anàlisi de les gràfiques obtingues anteriorment i extracció de conclusions. Plantejar noves gràfiques si fos necessari per a resoldre noves qüestions
	Subtasca 5.2: Subvencions de llengua		
	Subtasca 5.3: Subvencions de literatura		

	<u>Subtasca 5.4:</u> Certificació		
	<u>Subtasca 5.5:</u> Memòries		
	<u>Subtasca 5.6:</u> Inscripcions		
<u>Tasca 6:</u> **Revisió del procés**		10/05/2022 – 20/05/2022	Definició de quins processos han sigut correctes i quins es poden millorar (i fer-ho si es pot)
<u>Tasca 7:</u> **Determinar possibles futurs passos**		18/05/2022 – 20/05/2022	Definir quins elements es podrien afegir en un futur per a millorar l'abast del projecte
<u>Tasca 8:</u> **Crear una metodologia per a que es pugui monitoritzar i mantenir l'anàlisi de cara a anys posteriors**		20/05/2022 – 30/05/2022	Definir un conjunt de passos per a poder tornar a realitzar l'anàlisi de forma semi automàtica
<u>Tasca 9:</u> **Realització del document inicial**		15/02/2022 – 25/02/2022	Document que contingui els objectius del projecte i el pla de treball
<u>Tasca 10:</u> **Realització de l'informe de seguiment**		15/04/2022 – **30/04/2022**	Document parlant de l'evolució del projecte i si cal redefinir el pla de treball
<u>Tasca 11:</u> **Realització de la memòria final**		Fita: **18/06/2022**	Document descrivint la totalitat del projecte
<u>Tasca 12:</u> **Preparació i realització de la lectura del projecte**		19/06/2022 – **28/06/2022**	Presentació final (oral i suport visual)
<u>Tasca 13:</u> **Reunions**		Al llarg de tot el projecte	Decidir l'abast del projecte i modificar tot allò necessari a temps

A continuació es mostra un diagrama de Gantt que mostra les dependències entre totes les tasques anteriorment mencionades:



Pla de trobades i comunicació

Durant el primer mes del projecte es realitzarà 1 reunió setmanal amb la directora del projecte per a saber com enfocar-lo i deixar clars els objectius per part de l'Institut. Durant aquest primer mes, també es realitzaran com a mínim 3 entrevistes especials, una amb cada àrea mencionada anteriorment: creació, llengua i literatura; per a presentar-los el projecte i demanar-los que proposin idees.

Durant la resta del projecte es realitzarà 1 reunió amb la directora del projecte, Esther Coll Caldas, cada dues setmanes.

Competències genèriques

Durant el desenvolupament del projecte es potenciaran i avaluaran les següents competències genèriques:

#	Competències genèriques	Valorat
GS1	Innovació i emprenedoria	X
GS2	Context social i ambiental	X
GS3	Comunicació oral i escrita	X
GS4	Treball en equip	X
GS5	Enquesta de recursos d'informació	X
GS6	Aprenentatge autònom	
GS7	Comunicació en alguna llengua estrangera	
GS8	Perspectiva de gènere	

8.2 Document “Informe de seguiment”

Anàlisi estadístic de les bases de dades de l’Institut Ramon Llull

Informe de seguiment

ESCRIT PER:	**REVISAT I APROVAT PER:**
Nom: Anna Patrícia Orteu	Nom: Francesc Rey Micolau
Data: 12/04/2022	Data:

CONTINGUTS

1. COMENTARIS GENERALS SOBRE EL PROGRÉS DEL PROJECTE

1.1 Incidències

Durant aquesta primera fase del projecte s'ha topat principalment amb 3 incidències que han fet posposar la generació dels gràfics (la fase central del projecte). Aquesta, s'ha hagut de moure dues setmanes des de el 28 de febrer al 14 de març, respecte la planificació originalment indicada.

La primera incidència i la que ha fet posposar durant més temps l'anàlisi de les dades ha estat la no uniformització d'aquestes en una gran quantitat de cassos. Ja s'esperava trobar dades en certs cassos incompletes i no 100% netejades i que per tant s'havien de processar. Tanmateix, s'ha trobat varies variables composades per cadenes de text que mostraven múltiples valors per a referir-se a 1 mateix element, les quals s'ha hagut d'uniformitzar a 1 únic valor.

Un exemple d'aquest fet podria ser la variable "universitat" que s'utilitza en més d'un dels conjunts de dades analitzat. Aquesta podia mostrar "Universitat de California", "California University", "UCLA", "Universitat de California, Los Angeles", etc; per a referir-se a 1 única universitat. I tenint en compte que existeixen aproximadament 80 universitats, el procés es va allargar de forma considerable.

En segon lloc, la primera reunió amb els usuaris interns de l'Institut Ramon Llull per a saber què volien analitzar es va realitzar més tard del previst. Finalment, durant la proposta de projecte i pla de treball inicials, es va comentar que es treballaria amb 4 tipus de dades diferents: subvencions, memòries, inscripcions i certificacions. Tanmateix, més tard s'ha sabut que també s'havia de tractar amb dades de selecció, el qual afegeix tota una nova secció al treball.

1.2 Modificacions realitzades al Pla de Treball

Com comentat a la secció anterior, s'ha afegit una nova secció "Selecció de professorat" al treball, el qual dona lloc a la modificació del Pla de Treball. L'Institut organitza

anualment convocatòries de selecció de professorat d'estudis catalans a les universitats amb les quals col·labora per cobrir les vacants i garantir d'aquest manera la continuïtat de la docència. Totes les dades recollides d'aquestes convocatòries són les que formaran part d'aquesta secció "Selecció", que es trobarà dins de l'àrea de Creació.

2. PLA DE TREBALL ACTUALITZAT

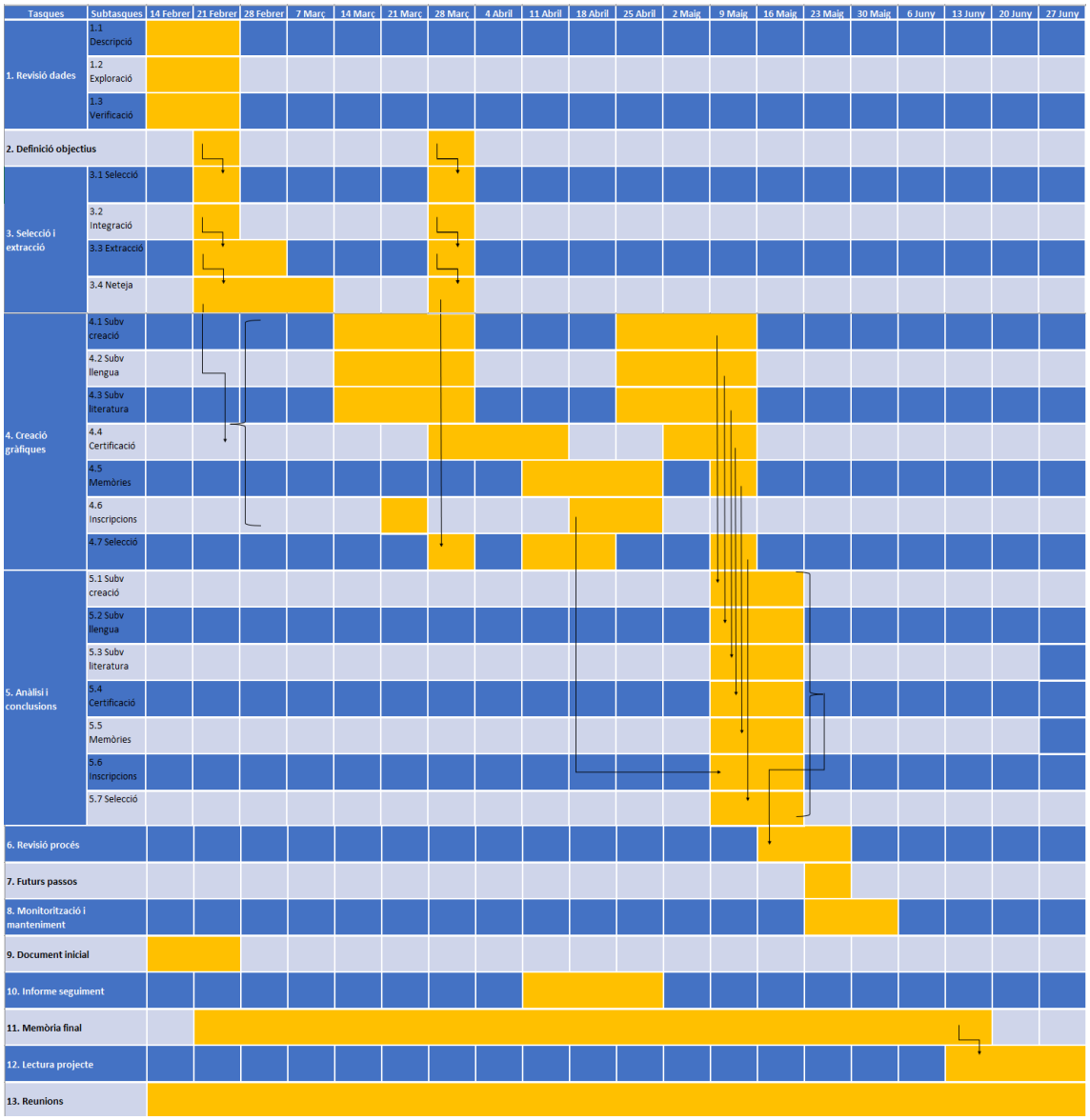
2.1 Tasques i fites actualitzades

A continuació, es mostra un conjunt de tasques amb els corresponents objectius o fites del projecte actualitzats. La gràfica també mostra els períodes de temps de cada secció així com les subtasques incloses en cadascuna d'elles:

Tasca 1: **Revisió de les dades de l'Institut Ramon Llull – avaluació de la situació**	Subtasca 1.1: Descripció de les dades	14/02/2022 – 21/02/2022	Word per a cada àrea mencionant les dades amb les que es pot tractar i exemples d'elements que es podrien analitzar
	Subtasca 1.2: Exploració de les dades		
	Subtasca 1.3: Verificació de la qualitat		
Tasca 2: **Definició dels objectius finals a través de les reunions**		21/02/2022 - 22/02/2022 + 30/03/2022 (selecció)	Saber objectius que pretenen assolir amb el projecte els usuaris de l'IRL
Tasca 3: **Selecció i extracció de les dades a analitzar**	Subtasca 3.1: Selecció de les dades	26/02/2022 + 30/03/2022 (selecció)	Selects que s'ha de realitzar contra les diferents bd del Llull
	Subtasca 3.2: Integració de les dades		
	Subtasca 3.3: Extracció de les dades	27/02/2022 – 13/03/2022 + 31/03/2022 – 02/04/2022 (selecció)	Excels completament netejats preparats per a fer l'anàlisi
	Subtasca 3.4: Neteja de les dades		
Tasca 4: **Creació de les gràfiques/elements per analitzar**	Subtasca 4.1: Subvencions de creació	14/03/2022 – 10/05/2022	Obtenir totes les gràfiques o elements necessaris que permetin extreure les conclusions
	Subtasca 4.2: Subvencions de llengua		
	Subtasca 4.3: Subvencions de literatura		
	Subtasca 4.4: Certificació		
	Subtasca 4.5: Memòries		

	<u>Subtasca 4.6:</u> Inscripcions		
	<u>Subtasca 4.7:</u> Selecció		
<u>Tasca 5:</u> **Anàlisi i extracció de conclusions**	<u>Subtasca 5.1:</u> Subvencions de creació	10/05/2022 – 20/05/2022	Anàlisi de les gràfiques obtingues anteriorment i extracció de conclusions. Plantejar noves gràfiques si fos necessari per a resoldre noves qüestions
	<u>Subtasca 5.2:</u> Subvencions de llengua		
	<u>Subtasca 5.3:</u> Subvencions de literatura		
	<u>Subtasca 5.4:</u> Certificació		
	<u>Subtasca 5.5:</u> Memòries		
	<u>Subtasca 5.6:</u> Inscripcions		
	<u>Subtasca 5.7:</u> Selecció		
<u>Tasca 6:</u> **Revisió del procés**		20/05/2022 – 25/05/2022	Definició de quins processos han sigut correctes i quins es poden millorar (i fer-ho si es pot)
<u>Tasca 7:</u> **Determinar possibles futurs passos**		23/05/2022 – 25/05/2022	Definir quins elements es podrien afegir en un futur per a millorar l'abast del projecte
<u>Tasca 8:</u> **Crear una metodologia per a que es pugui monitoritzar i mantenir l'anàlisi de cara a anys posteriors**		25/05/2022 – 30/05/2022	Definir un conjunt de passos per a poder tornar a realitzar l'anàlisi de forma semi automàtica
<u>Tasca 9:</u> **Realització del document inicial**		15/02/2022 – 25/02/2022	Document que contingui els objectius del projecte i el pla de treball
<u>Tasca 10:</u> **Realització de l'informe de seguiment**		10/04/2022 – **30/04/2022**	Document parlant de l'evolució del projecte i si cal redefinir el pla de treball
<u>Tasca 11:</u> **Realització de la memòria final**		Fita: **18/06/2022**	Document descrivint la totalitat del projecte
<u>Tasca 12:</u> **Preparació i realització de la lectura del projecte**		19/06/2022 – **28/06/2022**	Presentació final (oral i suport visual)
<u>Tasca 13:</u> **Reunions**		Al llarg de tot el projecte	Decidir l'abast del projecte i modificar tot allò necessari a temps

2.2 Diagrama de Gantt actualitzat



8.3 Taules usades

En aquest annex es pot visualitzar una llista de totes les taules d'on s'ha extret informació per a crear els conjunts de dades desitjats. També es mostra una breu descripció d'aquestes.

8.3.1 Subvencions

Per a totes les àrees en un inici es va obtenir informació de les següents taules:

Nom de la taula	Descripció de la taula
Tsubvencions	Conté informació sobre totes les subvencions de les àrees de llengua, creació i literatura, independentment de si han estat o no atorgades i de si són directes o de concurrència.
Testatsubvencio	Conté informació sobre l'estat de les subvencions, per exemple, si ha estat atorgada o revocada.
Tprefixsubvencio	Conté informació sobre el prefix de les subvencions. El prefix identifica de forma més precisa el tipus de subvenció indicant, per exemple, si es tracta d'una mobilitat o una traducció.
Ttipusresolucio	Conté informació sobre el desenllaç final d'una subvenció.
Tdisciplines	Conté informació sobre la disciplina de la subvenció, entre d'altres, si fa referència al món del circ o de la música.
Tliniessubvencio	Conté informació sobre a quina subcategoria (anomenades línies per l'IRL) es troba classificada una subvenció.
Tconvocatories	Conté informació sobre a quina convocatòria pertany la subvenció. Existeix 1 convocatòria per a cada línia i any.
Tterminis	Conté informació sobre a quin termini pertany la subvenció. Poden existir múltiples terminis dins d'una convocatòria.
Tpersones	Conté informació sobre la persona o entitat analitzada. Ha estat necessari crear múltiples lligams amb aquesta.
Tprovincies	Conté informació sobre les províncies. En trobar-se directament relacionada amb tpersones també ha estat necessària la creació de múltiples lligams amb aquesta.
Tpaisos	Conté informació sobre els països. En trobar-se directament relacionada amb tpersones també ha estat necessària la creació de múltiples lligams amb aquesta..
Tzones	Conté informació sobre la regió geogràfica analitzada (per continents). En trobar-se directament relacionada amb tpersones també ha estat necessària la creació de múltiples lligams amb aquesta.
Tdossiers	Conté informació sobre les subcategories de les subvencions "genèriques".

A més a més, per a les subvencions de literatura també s'ha establert relacions amb les següents entitats:

Nom de la taula	Descripció de la taula
Ttraduccions	Conté informació sobre les traduccions indicant, entre d'altres, l'obra traduïda i a quin idioma.
Tgeneres	Conté informació sobre el gènere de l'obra o llibre en qüestió.

Tidiomes	Conté informació sobre els diferents idiomes amb què tracta l'Institut.
Tobres	Conté informació sobre les obres que han estat associades a alguna subvenció.
Trel_autor_obra	Lliga la informació de la taula tobres amb tautor.
Tautor	Conté informació sobre l'autor original de les obres.
Tformularis_promocio	Conté informació sobre totes les subvencions de promoció que s'ha sol·licitat.

També ha estat necessari introduir relacions amb 1 vista més per al cas de les subvencions de creació i amb 3 taules per a les subvencions de llengua. En el segon cas, la nova informació no ha estat directament lligada a la consulta de les subvencions sinó a la consulta que indica les relacions de les entitats amb aquesta àrea.

Nom de la taula	Descripció de la taula
V_nom_festival	Conté informació sobre els noms dels festivals pels quals es demanen les subvencions de creació així com els països on es van celebrar aquests.

Nom de la taula	Descripció de la taula
Tany_activitat	Conté informació sobre totes les entitats relacionades amb l'àrea de llengua de l'IRL a l'any, indicant el tipus de relació.
Tsubtipuscentre	Conté informació sobre els subtipus dels centres associats al Llull.
Ttipuscentre	Conté informació sobre els tipus de centres associats al Llull.

8.3.2 Inscripcions

Per a les inscripcions ha estat necessari obtenir dades de 8 taules diferents:

Nom de la taula	Descripció de la taula
Tinscripcions	Conté informació sobre totes les inscripcions sol·licitades indicant entre d'altres el tipus d'aquestes.
Testades	Conté informació sobre les estades i els campus planificats de forma anual pel Llull.
Tliniessubvencio	Conté informació sobre a quina subcategoria (anomenades línies per l'IRL) es troba classificada una subvenció.
Tidiomes	Conté informació sobre els diferents idiomes amb què tracta l'Institut.
Tpersones	Conté informació sobre la persona o entitat analitzada. Ha estat necessari crear múltiples lligams amb aquesta.
Tprovincies	Conté informació sobre les províncies. En trobar-se directament relacionada amb tpersones també ha estat necessària la creació de múltiples lligams amb aquesta.
Tpaisos	Conté informació sobre els països. En trobar-se directament relacionada amb tpersones també ha estat necessària la creació de múltiples lligams amb aquesta.
Tzones	Conté informació sobre la regió geogràfica analitzada (per continents). En trobar-se directament relacionada amb tpersones també ha estat necessària la creació de múltiples lligams amb aquesta.

8.3.3 Certificació

Per a les certificacions ha estat necessari obtenir dades altre cop de 8 taules diferents per a formar el conjunt de dades desitjat:

Nom de la taula	Descripció de la taula
Clc_examens	Conté informació sobre tots els exàmens als quals els examinands s'han inscrit dins del Llull independentment del lloc o el resultat d'aquest.
Clc_seus	Conté informació sobre la seu (lloc) on s'ha dut a terme l'examen.
Clc_nivells	Conté informació sobre el nivell de l'examen portat a cap.
Clc_paisos	Conté informació sobre els països donada una localització. Ha estat necessari crear múltiples lligams amb aquesta.
Clc_alumnes	Conté informació sobre els examinands que s'han presentat a l'examen.
Clc_convocatories	Conté informació sobre les diferents convocatòries que existeixen.
Clc_conv_seus	Conté informació sobre les diferents convocatòries que existeixen donat un lloc.
Clc_conv_seus_nivells	Conté informació sobre les diferents convocatòries que existeixen donat un lloc i un nivell.

8.3.4 Selecció de professorat

Tot seguit es descriuran les 11 taules/vistes d'on s'ha extret informació per al cas de selecció:

Nom de la taula	Descripció de la taula
Sp_formulari	Conté totes les candidatures presentades pels candidats a l'hora de demanar una plaça.
Sp_enunciat_form	Conté informació sobre puntuacions que es tenen en compte a l'hora de seleccionar als millors candidats.
Sp_enunciat_conv	Conté informació sobre puntuacions que es tenen en compte a l'hora de seleccionar als millors candidats.
Sp_valor_form	Conté informació sobre puntuacions que es tenen en compte a l'hora de seleccionar als millors candidats.
Sp_valor_conv	Conté informació sobre puntuacions que es tenen en compte a l'hora de seleccionar als millors candidats.
Sp_valor_combo	Conté els títols de les característiques analitzades i contingudes a la taula sp_valor_form.
Sp_convocatoria	Conté informació sobre les diferents convocatòries a les quals els professors han pogut aplicar al llarg dels anys.
Sp_universitat_form	Conté informació sobre les universitats sol·licitades pels professors en les convocatòries.
Sp_universitat	Conté informació sobre les universitats que han de cobrir les places vacants.
Sp_pais	Conté informació sobre els països de les universitats.
Sp_v_idioma	Conté informació sobre els idiomes coneguts pels candidats (així com el nivell amb què els coneixen).

8.3.5 Memòries

Per als justificants de les subvencions atorgades a les universitats només ha estat necessari obtenir dades de 5 taules diferents:

Nom de la taula	Descripció de la taula
Xov_memories	Conté informació sobre tots els justificants de les subvencions atorgades a universitats.
Xov_produccio	Conté informació sobre les produccions (ex: articles) que s'ha escrit dins del marc de la memòria realitzada.
Xov_activitat	Conté informació sobre les activitats que s'han dut a terme dins del marc de la memòria realitzada.
Xov_assignatura	Conté informació sobre les assignatures que s'han cursat dins del marc de la memòria realitzada.
Xov_puntuacio	Conté informació sobre les puntuacions que se'ls ha donat a diversos aspectes de la subvenció justificada. Ha estat necessari crear múltiples lligams amb aquesta.

8.4 Descripció dels conjunts de dades

Al llarg d'aquest annex es podrà trobar una descripció variable per variable de tots els conjunts de dades utilitzats per a la creació de les gràfiques. Aquests conjunts es poden observar al GitHub del projecte dins les carpetes anomenades dades.zip.

8.4.1 Subvencions de llengua

Pel cas del conjunt de dades que conté les sol·licituds de les subvencions i el seu desenllaç final (subv_llengua_final.xlsx) es tenen les següents variables:

Id de la columna	Nom de la variable	Nombre de files no nul·les	Tipus	Descripció
0	**idsubvencio**	3392	int64	ID de la subvenció
1	**idpersona**	3392	int64	ID de la persona que sol·licita la subvenció
2	**pers_objecte**	3392	int64	ID de la persona/entitat que rep la subvenció
3	**importsolicitat**	3392	float64	Import sol·licitat
4	**importsubvencionable**	3288	float64	Import subvencionable
5	**importatorgat**	3266	float64	Import atorgat
6	**datasolicitud**	3392	datetime64[ns]	Data de sol·licitud de la subvenció
7	**datamaxjustificacio**	3392	datetime64[ns]	Data de màxima justificació de la subvenció
8	**idestatsubvencio**	3392	int64	ID de l'estat de la subvenció
9	**idtipusresolucio**	1671	float64	ID del tipus de resolució
10	**idprefix**	695	float64	ID del prefix
11	**concurrent**	3392	object	Indica si la subvenció és concurrent o directe
12	**cobrament**	3392	int64	Indica si s'ha fet el pagament
13	**ajutviatge**	3392	int64	Indica si inclou ajudes pel desplaçament
14	**importjustificat**	1499	float64	Import justificat
15	**any_directe**	3392	int64	Any de petició i atorgament de la subvenció
16	**idtermini**	695	float64	ID del termini
17	**ordre_pagament**	3392	int64	Indica si s'ha generat alguna ordre de pagament
18	**paper_online**	3392	object	Indica si la subvenció ha estat sol·licitada online o en paper
19	**nom_p**	3392	object	Nom de la persona que sol·licita la subvenció
20	**ambit_p**	531	object	Àmbit de la persona que sol·licita la subvenció

21	**tipuspersona_p**	3392	object	Tipus de persona que sol·licita la subvenció
22	**datacreacio_p**	3392	datetime64[ns]	Data d'inserció a la BD de la persona que sol·licita la subvenció
23	**sexe_p**	500	object	Sexe de la persona que sol·licita la subvenció
24	**robinson_p**	3392	int64	Indica si la persona que ha sol·licitat la subvenció vol rebre més informació de l'IRL
25	**tipus_identificacio_p**	2926	float64	Tipus d'identificació de la persona que sol·licita la subvenció
26	**nom_pobjecte**	3392	object	Nom de la persona/entitat que rep la subvenció
27	**ambit_pobjecte**	514	object	Àmbit de la persona/entitat que rep la subvenció
28	**tipuspersona_pobjecte**	3392	object	Tipus de persona/entitat que rep la subvenció
29	**idprovincia_pobjecte**	1095	float64	Data d'inserció a la BD de la persona/entitat que rep la subvenció
30	**sexe_pobjecte**	63	object	Sexe de la persona/entitat que rep la subvenció
31	**robinson_pobjecte**	3392	object	Indica si la persona/entitat que ha rebut la subvenció vol rebre més informació de l'IRL
32	**tipus_identificacio_pobjecte**	2573	object	Tipus d'identificació de la persona/entitat que rep la subvenció
33	**estat**	3392	object	Estat de la subvenció
34	**prefix**	695	object	Prefix de la subvenció
35	**tipusresolucio**	1125	object	Tipus de resolució de la subvenció
36	**datainici_termini**	695	datetime64[ns]	Data d'obertura del termini
37	**datafi_termini**	695	datetime64[ns]	Data de tancament del termini
38	**dataresolucio_termini**	616	datetime64[ns]	Data de resolució del termini
39	**datajustificaciomax_termini**	695	datetime64[ns]	Data de màxima justificació del termini
40	**actiu_termini**	695	float64	Indica si el termini es troba actiu o no
41	**idconvocatoria**	695	float64	ID de la convocatòria
42	**pais_p**	3371	object	País de la persona que sol·licita la subvenció
43	**idzona_p**	3356	float64	ID de la zona geogràfica de la persona que sol·licita la subvenció
44	**provincia_pobjecte**	1095	object	Província de la persona/entitat que rep la subvenció

45	**pais_pobjecte**	3357	object	País de la persona/entitat que rep la subvenció
46	**idzona_pobjecte**	3328	float64	ID de la zona geogràfica de la persona/entitat que rep la subvenció
47	**any_conv**	695	float64	Any de la convocatòria
48	**idlinia_conv**	3392	int64	ID de la línia
49	**descripcio_linia_subv**	3392	object	Descripció de la línia a la qual pertany la subvenció
50	**zona_p**	3356	object	Zona de la persona que sol·licita la subvenció
51	**zona_pobjecte**	3328	object	Zona de la persona/entitat que rep la subvenció
52	**dies_termini**	695	float64	Dies d'obertura del termini sol·licitat

Mentre per al conjunt de dades que conté les relacions de les diferents entitats amb l'àrea de llengua (subv_evolucioUnis.xlsx) s'observen els següents camps:

Id de la columna	Nom de la variable	Nombre de files no nul·les	Tipus	Descripció
0	**idany_activitat**	6731	int64	ID que identifica de manera única el tipus d'una entitat en un any determinat
1	**idpersona**	6731	int64	ID de l'entitat
2	**anyacademic**	6731	int64	Any acadèmic
3	**subtipuscentre**	6731	object	Subtipus de l'entitat
4	**tipuscentre**	6731	object	Tipus de l'entitat
5	**nom**	6731	object	Nom de l'entitat
6	**pais**	6455	object	País al qual pertany l'entitat
7	**zona**	6438	object	Zona geogràfica a la qual pertany l'entitat

8.4.2 Subvencions de creació

Per al cas del conjunt de dades que conté les sol·licituds de les subvencions i el seu desenllaç final (subv_creacio_final.xlsx) es tenen les següents variables:

Id de la columna	Nom de la variable	Nombre de files no nul·les	Tipus	Descripció
0	**idsubvencio**	10811	int64	ID de la subvenció
1	**idpersona**	10811	int64	ID de la persona que sol·licita la subvenció
2	**pers_objecte**	10811	int64	ID de la persona/entitat que rep la subvenció

3	**importsolicitat**	10811	float64	Import sol·licitat
4	**importsubvencionable**	9540	float64	Import subvencionable
5	**importatorgat**	8539	float64	Import atorgat
6	**datasolicitud**	10811	datetime64[ns]	Data de sol·licitud de la subvenció
7	**datamaxjustificacio**	10811	datetime64[ns]	Data de màxima justificació de la subvenció
8	**idestatsubvencio**	10811	int64	ID estat de la subvenció
9	**idprefix**	10327	float64	ID del prefix
10	**concurrent**	10811	object	Indica si la subvenció és concurrent o directe
11	**ajutviatge**	10811	int64	Indica si inclou ajudes pel desplaçament
12	**importjustificat**	8539	float64	Import justificat
13	**any_directe**	10811	float64	Any de petició i atorgament de la subvenció
14	**pax**	8172	float64	Nombre de persones que viatgen en l'activitat
15	**idtermini**	10327	float64	ID del termini
16	**datainici**	8180	datetime64[ns]	Data inici de l'activitat objecte de pagament
17	**ordre_pagament**	10811	int64	Indica si s'ha generat alguna ordre de pagament
18	**carrer**	10811	int64	Subtipus 1 de les subvencions de teatre
19	**titelles**	10811	int64	Subtipus 2 de les subvencions de teatre
20	**infantil**	10811	int64	Subtipus 3 de les subvencions de teatre
21	**en_requeriment**	4347	float64	Indica si la subvenció es troba en requeriment
22	**en_recurs**	4347	float64	Indica si la subvenció es troba en recurs
23	**datafi**	3459	datetime64[ns]	Data fi de l'activitat objecte de pagament
24	**paper_online**	10811	object	Indica si la subvenció ha estat sol·licitada online o en paper
25	**nom_p**	10811	object	Nom de la persona que sol·licita la subvenció
26	**ambit_p**	935	object	Àmbit de la persona que sol·licita la subvenció
27	**tipuspersona_p**	10811	object	Tipus de persona que sol·licita la subvenció
28	**idprovincia_p**	10012	float64	ID de la província de la persona que sol·licita la subvenció

29	**datacreacio_p**	10811	datetime64[ns]	Data d'inserció a la BD de la persona que sol·licita la subvenció
30	**sexe_p**	201	object	Sexe de la persona que sol·licita la subvenció
31	**robinson_p**	10811	int64	Indica si la persona que ha sol·licitat la subvenció vol rebre més informació de l'IRL
32	**tipus_identificacio_p**	10709	float64	Tipus d'identificació de la persona que sol·licita la subvenció
33	**nom_pobjecte**	10811	object	Nom de la persona/entitat que rep la subvenció
34	**ambit_pobjecte**	536	object	Àmbit de la persona/entitat que rep la subvenció
35	**tipuspersona_pobjecte**	10811	object	Tipus de persona/entitat que rep la subvenció
36	**idprovincia_pobjecte**	4562	float64	ID de la província de la persona/entitat que rep la subvenció
37	**sexe_pobjecte**	69	object	Sexe de la persona/entitat que rep la subvenció
38	**robinson_pobjecte**	10811	object	Indica si la persona/entitat que ha rebut la subvenció vol rebre més informació de l'IRL
39	**tipus_identificacio_pobjecte**	3265	object	Tipus d'identificació de la persona/entitat que rep la subvenció
40	**estat**	10811	object	Estat de la subvenció
41	**prefix**	10327	object	Prefix de la subvenció
42	**iddisciplina**	10327	float64	ID de la disciplina
43	**datainici_termini**	10327	datetime64[ns]	Data d'obertura del termini
44	**datafi_termini**	10327	datetime64[ns]	Data de tancament del termini
45	**dataresolucio_termini**	9357	datetime64[ns]	Data de resolució del termini
46	**datajustificaciomax_termini**	10327	datetime64[ns]	Data de màxima justificació del termini
47	**num_termini**	10327	float64	Número de termini
48	**actiu_termini**	10327	float64	Indica si el termini es troba actiu o no
49	**idconvocatoria**	10327	float64	ID de la convocatòria
50	**provincia_p**	10012	object	Província de la persona que sol·licita la subvenció
51	**pais_p**	10798	object	País de la persona que sol·licita la subvenció

52	**idzona_p**	10796	float64	ID de la zona geogràfica de la persona que sol·licita la subvenció
53	**provincia_pobjecte**	4562	object	Província de la persona/entitat que rep la subvenció
54	**pais_pobjecte**	7016	object	País de la persona/entitat que rep la subvenció
55	**idzona_pobjecte**	7015	float64	ID de la zona geogràfica de la persona/entitat que rep la subvenció
56	**disciplina**	10327	object	Disciplina de la subvenció
57	**any_conv**	10327	float64	Any de la convocatòria
58	**idlinia_conv**	10811	int64	ID de la línia
59	**descripcio_linia_subv**	10811	object	Descripció de la línia a la qual pertany la subvenció
60	**any_inici_linia_subv**	2957	float64	Any que la línia a la qual pertany la subvenció es va obrir per primer cop
61	**zona_p**	10796	object	Zona geogràfica de la persona que sol·licita la subvenció
62	**zona_pobjecte**	7015	object	Zona de la persona/entitat que rep la subvenció
63	**objecte**	10811	object	Objecte (raó) de la subvenció
64	**paisos_festivals**	5396	object	Països on es duen a terme els festivals
65	**noms_festivals**	5396	object	Noms dels festivals pels quals s'atorguen les subvencions
66	**poblacions_Espanya_festivals**	2127	object	Poblacions d'Espanya on es duen a terme els festivals
67	**dies_termini**	10327	float64	Dies d'obertura del termini sol·licitat

Mentre per al que conté la informació de les poblacions s'observen els següents camps (subv_creacio_poblacions_final.xlsx):

Id de la columna	Nom de la variable	Nombre de files no nul·les	Tipus	Descripció
0	**any_directe**	2242	Float64	Any acadèmic
1	**descripcio_linia_subv**	2242	object	Descripció de la línia a la qual pertany la subvenció
2	**poblacions_Espanya_festivals**	2242	object	Poblacions on s'ha realitzat els festivals (1 població per tupla)

3	**counts**	2242	int64	Nombre de cops que s'ha demanat alguna subvenció per a dur a terme un festival en un any i població determinats dins la mateixa línia

8.4.3 Subvencions de literatura

subv_lite_final.xlsx:

Id de la columna	Nom de la variable	Nombre de files no nul·les	Tipus	Descripció
0	**idsubvencio**	3925	int64	ID de la subvenció
1	**idpersona**	3925	int64	ID de la persona que sol·licita la subvenció
2	**pers_objecte**	3925	int64	ID de la persona/entitat que rep la subvenció
3	**importsolicitat**	3925	float64	Import sol·licitat
4	**importsubvencionable**	3659	float64	Import subvencionable
5	**importatorgat**	3256	float64	Import atorgat
6	**datasolicitud**	3921	datetime64[ns]	Data de sol·licitud de la subvenció
7	**datamaxjustificacio**	3898	datetime64[ns]	Data de màxima justificació de la subvenció
8	**idcontacte**	2592	float64	ID de la persona que l'entitat ha marcat com a contacte
9	**idestatsubvencio**	3925	int64	ID estat de la subvenció
10	**idprefix**	3873	float64	ID del prefix
11	**idtraduccions**	2046	float64	ID de la traducció
12	**concurrent**	3925	object	Indica si la subvenció és concurrent o directe
13	**ajutviatge**	3925	int64	Indica si inclou ajudes pel desplaçament
14	**importjustificat**	2885	float64	Import justificat
15	**any_directe**	3925	float64	Any de petició i atorgament de la subvenció
16	**ididioma**	1246	float64	ID de l'idioma utilitzat per a la traducció o activitat literària
17	**idtermini**	3873	float64	ID del termini
18	**tipus_promocio**	462	object	Modalitat dins la línia de "Promoció d'obres"
19	**datainici**	1691	datetime64[ns]	Data inici de l'activitat objecte de pagament
20	**ordre_pagament**	3925	int64	Indica si s'ha generat alguna ordre de pagament

21	**prorroga**	3925	int64	Indica si se li ha concedit una pròrroga a la subvenció
22	**en_requeriment**	2254	float64	Indica si la subvenció es troba en requeriment
23	**en_recurs**	2252	float64	Indica si la subvenció es troba en recurs
24	**datafi**	1625	datetime64[ns]	Data fi de l'activitat objecte de pagament
25	**datafullpagament**	1058	datetime64[ns]	Indica la data en què s'ha extret l'últim full de pagament de la subvenció
26	**datajustificacio**	1031	datetime64[ns]	Indica la data en què s'ha justificat la subvenció
27	**importapagar**	1078	float64	Import que encara queda per pagar
28	**estat**	3925	object	Estat de la subvenció
29	**paper_online**	3925	object	Indica si la subvenció ha estat sol·licitada online o en paper
30	**prefix**	3873	object	Prefix de la subvenció
31	**idioma**	1246	object	Idioma utilitzat per a la traducció o activitat literària
32	**ANY_conv**	3873	float64	Any de la convocatòria
33	**idlinia**	3925	float64	ID de la línia
34	**descripcio_linia**	3925	object	Descripció de la línia a la qual pertany la subvenció
35	**datainici_termini**	3847	datetime64[ns]	Data d'obertura del termini
36	**datafi_termini**	3847	datetime64[ns]	Data de tancament del termini
37	**dataresolucio_termini**	3318	datetime64[ns]	Data de resolució del termini
38	**datajustificaciomax_termini**	3873	datetime64[ns]	Data de màxima justificació del termini
39	**num_termini**	3873	float64	Número de termini
40	**actiu_termini**	3873	float64	Indica si el termini es troba actiu
41	**idconvocatoria**	3873	float64	ID de la convocatòria
42	**ideditorial_trad**	2046	float64	ID de l'editorial que ha realitzat la traducció
43	**idgenere_trad**	1106	float64	ID del gènere de la traducció
44	**idobra_trad**	2046	float64	ID de l'obra de la traducció
45	**idpersona_trad**	2046	float64	ID del traductor
46	**ididioma_trad**	2046	float64	ID de l'idioma de la traducció
47	**titoltraduccio_trad**	2046	object	Títol de la traducció
48	**anytraduccio_trad**	1798	float64	Any de la traducció
49	**stock_trad**	2046	float64	Nombre de còpies en paper de la traducció
50	**update_stock_trad**	2046	float64	Indica si s'ha actualitzat l'stock
51	**import_trad**	767	float64	Import de la traducció

52	**data_solicitud_trad**	767	datetime64[ns]	Data de sol·licitud de la traducció
53	**autor_obra_original**	3925	object	Autor de l'obra original de la traducció
54	**genere_trad**	1106	object	Gènere de la traducció
55	**idioma_trad**	2046	object	Idioma de la traducció
56	**idpersona_o_trad**	1003	float64	ID de la persona objecte de la traducció
57	**idgenere_o_trad**	1384	float64	ID del gènere de l'obra traduïda
58	**titolobra**	2046	object	Títol de l'obra
59	**anypublicacio_trad**	909	float64	Any de publicació de la traducció
60	**genere_o_trad**	1384	object	Gènere de la traducció
61	**nom_p**	3925	object	Nom de la persona que sol·licita la subvenció
62	**ambit_p**	605	object	Àmbit de la persona que sol·licita la subvenció
63	**tipuspersona_p**	3925	object	Tipus de persona que sol·licita la subvenció
64	**idprovincia_p**	1072	float64	ID de la província de la persona que sol·licita la subvenció
65	**idpais_p**	3761	float64	ID del país de la persona que sol·licita la subvenció
66	**datacreacio_p**	3925	datetime64[ns]	Data d'inserció a la BD de la persona que sol·licita la subvenció
67	**sexe_p**	10	object	Sexe de la persona que sol·licita la subvenció
68	**robinson_p**	3925	object	Indica si la persona que ha sol·licitat la subvenció vol rebre més informació de l'IRL
69	**tipus_identificacio_p**	3138	object	Tipus d'identificació de la persona que sol·licita la subvenció
70	**provincia_p**	1072	object	Província de la persona que sol·licita la subvenció
71	**pais_p**	3761	object	País de la persona que sol·licita la subvenció
72	**idzona_p**	3727	float64	ID de la zona geogràfica de la persona que sol·licita la subvenció
73	**zona_p**	3727	object	Zona geogràfica de la persona que sol·licita la subvenció
74	**nom_pobjecte**	3925	object	Nom de la persona/entitat que rep la subvenció

75	**ambit_pobjecte**	644	object	Àmbit de la persona/entitat que rep la subvenció
76	**tipuspersona_pobjecte**	3925	object	Tipus de persona/entitat que rep la subvenció
77	**idprovincia_pobjecte**	1284	float64	ID de la província de la persona/entitat que rep la subvenció
78	**idpais_pobjecte**	3599	float64	ID del país de la persona/entitat que rep la subvenció
79	**datacreacio_pobjecte**	3925	datetime64[ns]	Data d'inserció a la BD de la persona/entitat que rep la subvenció
80	**sexe_pobjecte**	47	object	Sexe de la persona/entitat que rep la subvenció
81	**robinson_pobjecte**	3925	object	Indica si la persona/entitat que ha rebut la subvenció vol rebre més informació de l'IRL
82	**tipus_identificacio_pobjecte**	2486	object	Tipus d'identificació de la persona/entitat que rep la subvenció
83	**provincia_pobjecte**	1284	object	Província de la persona/entitat que rep la subvenció
84	**pais_pobjecte**	3599	object	País de la persona/entitat que rep la subvenció
85	**idzona_pobjecte**	3576	float64	ID de la zona geogràfica de la persona/entitat que rep la subvenció
86	**zona_pobjecte**	3576	object	Zona de la persona/entitat que rep la subvenció
87	**objecte**	3925	object	Objecte (raó) de la subvenció
88	**dies_termini**	3847	float64	Dies d'obertura del termini sol·licitat

8.4.4 Inscripcions

inscripcio_final.xlsx:

Id de la columna	Nom de la variable	Nombre de files no nul·les	Tipus	Descripció
0	**idestada**	1162	int64	ID de l'estada
1	**idpersona**	1162	int64	ID de la persona que sol·licita realitzar una estada o campus
2	**hores_catala**	1161	object	Hores de català que la persona sol·licitant ha estudiat

3	**universitat**	1157	object	Universitat on la persona sol·licitant ha estudiat
4	**data_inscripcio**	1162	datetime64[ns]	Data d'inscripció al campus o estada
5	**beques_1**	1160	float64	Indica si fa 1 any l'usuari va rebre alguna beca per part de l'IRL
6	**beques_2**	1160	float64	Indica si fa 2 anys l'usuari va rebre alguna beca per part de l'IRL
7	**campus_1**	1160	float64	Indica si fa 1 any l'usuari va participar al campus
8	**campus_2**	1160	float64	Indica si fa 2 anys l'usuari va participar al campus
9	**irl_basic**	1160	float64	Indica si l'usuari té el nivell bàsic acreditat per l'IRL
10	**irl_elemental**	1160	float64	Indica si l'usuari té el nivell elemental acreditat per l'IRL
11	**irl_intermedi**	1160	float64	Indica si l'usuari té el nivell intermedi acreditat per l'IRL
12	**irl_suficiencia**	1160	float64	Indica si l'usuari té el nivell suficiència acreditat per l'IRL
13	**irl_superior**	1160	float64	Indica si l'usuari té el nivell superior acreditat per l'IRL
14	**andorra_a**	1160	float64	Indica el nivell d'Andorra que té l'usuari
15	**andorra_b**	1160	float64	Indica el nivell d'Andorra que té l'usuari
16	**universitat_p**	912	object	Universitat a la qual pertany el professor que acredita la sol·licitud
17	**contacte**	1161	float64	Indica si l'usuari vol ser contactat
18	**puntuacio_anterior**	1162	int64	Puntuació basada en si l'usuari ha realitzat algun campus amb anterioritat
19	**puntuacio_certificat**	1162	int64	Puntuació que té en consideració els certificats de l'alumne
20	**puntuacio_motiu**	1162	int64	Puntuació que té en consideració els motius de l'alumne per a realitzar el campus o estada
21	**puntuacio_professor**	1162	int64	Puntuació concedida al professor que du a terme l'acreditació en funció del seu palmarès

22	**puntuacio_total**	1162	int64	Puntuació total atorgada a l'alumne
23	**estada_1a**	496	float64	Indica si fa 1 any l'usuari va participar en l'estada de València
24	**estada_1b**	496	float64	Indica si fa 1 any l'usuari va participar en l'estada de les Illes Balears
25	**estada_2a**	496	float64	Indica si fa 2 anys l'usuari va participar en l'estada de València
26	**estada_2b**	496	float64	Indica si fa 2 anys l'usuari va participar en l'estada de les Illes Balears
27	**a2**	496	float64	Indica si l'usuari té el nivell bàsic (a2) acreditat pel Govern d'Andorra
28	**b1**	496	float64	Indica si l'usuari té el nivell elemental (b1) acreditat pel Govern d'Andorra
29	**b2**	496	float64	Indica si l'usuari té el nivell intermedi (b2) acreditat pel Govern d'Andorra
30	**c1**	496	float64	Indica si l'usuari té el nivell suficiència (c1) acreditat pel Govern d'Andorra
31	**idlinia**	1162	int64	ID de la línia de l'estada o campus sol·licitat
32	**any_estada**	1162	int64	Any de sol·licitud
33	**descripcio**	1162	object	Descripció de l'estada o campus
34	**codiexpedient**	1162	object	Codi d'expedient de l'estada o campus
35	**descripcio_ECCS**	1162	object	Indica si és un campus o una estada
36	**codiconveni**	1162	object	Codi del conveni de l'estada o campus
37	**ambit**	1162	object	Àmbit de la persona sol·licitant
38	**idpais**	1162	int64	ID del país de la persona sol·licitant
39	**pais**	1162	object	País de la persona sol·licitant
40	**idzona**	1083	float64	ID de la zona de la persona sol·licitant
41	**zona**	1083	object	Zona de la persona sol·licitant
42	**ididioma**	1162	int64	ID de l'idioma matern de la persona sol·licitant
43	**idioma**	1162	object	Idioma matern de la persona sol·licitant

44	**sexe**	1161	object	Sexe de la persona sol·licitant
45	**robinson**	1162	object	Indica si la persona sol·licitant vol rebre més informació de l'IRL
46	**irl_nivell**	1162	object	Nivell màxim acreditat per l'IRL obtingut per l'alumne
47	**andorra_nivell**	1162	object	Nivell màxim acreditat pel Govern d'Andorra obtingut per l'alumne

8.4.5 Certificació

certifiicacio_final.xlsx:

Id de la columna	Nom de la variable	Nombre de files no nul·les	Tipus	Descripció
0	**id_examen**	19749	int64	ID de l'examen
1	**informe_avaluacio**	19749	int64	Indica si s'ha extret l'informe de l'avaluació
2	**certificat_mobilitat_creat**	19749	int64	Indica si s'ha creat el certificat de mobilitat
3	**preu**	19749	int64	Preu que l'examinand ha de pagar
4	**import_pagat**	19749	int64	Preu pagat per l'examinand
5	**data_pagament**	8815	datetime64[ns]	Data de pagament
6	**presentat**	19749	int64	Indica si l'examinand s'ha presentat a l'examen
7	**resultat**	19749	object	Resultat de l'examen
8	**id_conv_seus_nivells**	19749	int64	ID de la convocatòria donat una seu i un nivell
9	**actiu_conv_seus_nivells**	19749	int64	Indica si la convocatòria en un nivell i seu determinats es troba oberta
10	**preu_conv_seus_nivells**	19749	int64	Indica el preu donat una seu i nivell determinats
11	**sequencial_conv_seus_nivells**	19749	int64	Indica si l'examen serà seqüencial donat una seu i nivell determinats
12	**id_nivell**	19749	int64	ID del nivell
13	**nivell**	19749	object	Nivell de l'examen
14	**codi_nivell**	19749	object	Codi del nivell
15	**preu_nivell**	19749	int64	Preu del nivell
16	**num_clausus_nivell**	19749	int64	Nombre màxim d'estudiants acceptats dins d'un nivell, seu i convocatòria

17	**id_conv_seus**	19749	int64	ID de la convocatòria donada una seu
18	**institucio_examen_conv_seus**	19749	object	Institució on s'ha realitzat l'examen donada una convocatòria i una seu
19	**poblacio_examen_conv_seus**	19749	object	Població on s'ha realitzat l'examen donada una convocatòria i una seu
20	**format_electronic_conv_seus**	19749	int64	Indica com volen extreure els examinands el resultat de la prova
21	**enregistradora_conv_seus**	19749	int64	Enregistradora de la convocatòria donada una seu
22	**id_convocatoria**	19749	int64	ID de la convocatòria
23	**nom_convocatoria**	19749	object	Nom de la convocatòria
24	**any_convocatoria**	19749	int64	Any de la convocatòria
25	**codi_convocatoria**	19749	int64	Codi de la convocatòria
26	**pagament_transferencia_actiu_convocatoria**	19749	int64	Indica si el pagament per transferència es permet en la convocatòria donada
27	**descarrega_proves_activa_convocatoria**	19749	int64	Indica si la descàrrega de les proves es troba activa donada una convocatòria
28	**certificat_mobilitat_actiu_convocatoria**	19749	int64	Indica si la descàrrega del certificat de mobilitat es troba actiu donada una convocatòria
29	**data_obertura_liquidacio_convocatoria**	19749	datetime64[ns]	Indica la data d'obertura de la liquidació donada una convocatòria
30	**data_tancament_liquidacio_convocatoria**	19749	datetime64[ns]	Data de tancament de la liquidació donada una convocatòria
31	**data_examen**	19749	datetime64[ns]	Data de l'examen
32	**data_consulta_resultats_examens_convocatoria**	19749	datetime64[ns]	Data de consulta dels resultats de l'examen donada una convocatòria
33	**data_obertura_inscripcio_convocatoria**	19749	datetime64[ns]	Data d'obertura de la inscripció donada una convocatòria
34	**data_tancament_inscripcio_convocatoria**	19749	datetime64[ns]	Data de tancament de les inscripcions donada una convocatòria
35	**id_seus**	19749	int64	ID de la seu
36	**nom_seus**	19749	object	Nom de la seu
37	**sigles_seus**	19749	object	Sigles de la seu
38	**codi_seus**	19749	int64	Codi de la seu

39	**gratuita_seus**	19749	int64	Indica si la seu on es realitza l'examen és gratuïta
40	**id_pais**	19749	int64	ID del país de realització de l'examen
41	**Pais**	19749	object	País de realització de l'examen
42	**num_pais**	19740	float64	Número del país de realització de l'examen
43	**irpf1_pais**	19749	int64	Irpf1 del país de realització de l'examen
44	**irpf2_pais**	19749	int64	Irpf2 del país de realització de l'examen
45	**id_alumnes**	19749	int64	ID de l'examinand
46	**data_naixement_alumne**	19365	datetime64[ns]	Data de naixement de l'examinand
47	**sexe_alumne**	19719	object	Sexe de l'examinand
48	**inscripcio_previa_cert_alumne**	19749	object	Indica la certificació prèvia de l'examinand si aquest ha realitzat algun examen amb anterioritat
49	**nivell_estudis_alumne**	16866	object	Nivells d'estudis de l'examinand
50	**ocupacio_alumne**	11212	object	Ocupació de l'examinand
51	**hores_catala_alumne**	11238	object	Hores de català estudiades per l'examinand abans de presentar-se a l'examen
52	**motiu_prova_alumne**	14153	object	Motiu de l'examinand per a dur a terme l'examen
53	**centre_catala_ultim_any**	9564	object	Centre d'estudi del català de l'examinand l'últim any abans de presentar-se a l'examen
54	**uni_catala_ultim_any**	19698	object	Universitat on ha estudiat l'examinand l'últim any abans de presentar-se a l'examen
55	**id_pais_alumnes**	19749	int64	ID del país de l'examinand
56	**pais_alumne**	19607	object	País de l'examinand
57	**num_pais_alumne**	19523	float64	Número del país de l'examinand
58	**irpf1_pais_alumne**	19749	int64	Irpf1 del país de l'examinand
59	**irpf2_pais_alumne**	19749	int64	Irpf2 del país de l'examinand
60	**ambit_alumne**	19749	object	Àmbit de l'examinand
61	**tipus_institucio**	19744	object	Tipus d'institució on es realitza l'examen
62	**llengua_materna_alumnes**	16822	object	Llengua materna de l'examinand
63	**Anglès**	19749	int64	Indica si l'examinand sap anglès

64	**Francès**	19749	int64	Indica si l'examinand sap francès
65	**Castellà**	19749	int64	Indica si l'examinand sap castellà
66	**Alemany**	19749	int64	Indica si l'examinand sap alemany
67	**Italià**	19749	int64	Indica si l'examinand sap italià
68	**dies_inscripcio_convocatoria**	19749	int64	Dies que una convocatòria es troba oberta
69	**edat_alumnes**	19365	float64	Edat de l'examinand
70	**nivell_pais_hcat**	11175	object	Nivell a què s'apunten més examinands donat un país i unes hores de català

8.4.6 Selecció de professorat

Pel cas del conjunt de dades que conté les sol·licituds presentades pels candidats en les diferents convocatòries i el seu desenllaç final (seleccio_final.xlsx) es tenen les següents variables:

Id de la columna	Nom de la variable	Nombre de files no nul·les	Tipus	Descripció
0	**id_formulari**	2039	int64	ID del formulari
1	**nom**	2039	object	Nom del candidat
2	**sexe**	1699	object	Sexe del candidat
3	**poblacio**	1229	object	Població del candidat
4	**any_naixement**	2020	float64	Any de naixement del candidat
5	**pais**	2039	object	País de naixement del candidat
6	**any_estudis**	1412	float64	Any en què el candidat va obtenir la llicenciatura
7	**any_master**	610	float64	Any en què el candidat va obtenir el màster
8	**universitat_master**	828	object	Universitat on el candidat va obtenir el màster
9	**admes**	2039	object	Indica si l'usuari ha estat admès
10	**exclos**	2039	object	Indica si l'usuari ha estat exclòs
11	**reserva**	2039	object	Indica si l'usuari ha estat seleccionat com a reserva
12	**seleccionat**	2039	object	Indica si l'usuari ha estat seleccionat
13	**unitat_puntuacio**	2039	float64	Indica la puntuació de la unitat didàctica (subjectiva)
14	**motiu_exclusio**	592	object	Motiu per excloure el candidat de la convocatòria
15	**professor_nomcomplert**	814	object	Nom complert del professor que acredita la sol·licitud

16	**professor_si**	2039	int64	Indica si l'usuari ha omplert el formulari
17	**acabat**	2039	int64	Indica si el formulari ha estat acabat
18	**paper**	2039	int64	Indica si el formulari s'ha realitzat en paper o telemàticament
19	**universitat1**	496	object	Universitat on el candidat va obtenir la llicenciatura
20	**idLlicenciatura**	2039	int64	Indica el tipus de llicenciatura
21	**idAltres**	2039	int64	Indica el tipus de llicenciatura si aquest no es troba dins dels típics
22	**idPostGrau**	2039	int64	Indica el tipus de postgrau
23	**idMaster**	2039	int64	Indica el tipus de màster
24	**idDoctorat**	2039	int64	Indica el tipus de doctorat
25	**data_creacio**	2002	datetime64[ns]	Data d'entrada de la sol·licitud al sistema
26	**id_convocatoria**	2039	int64	ID de la convocatòria
27	**any_academic**	2039	int64	Any acadèmic
28	**data_inici_conv**	2039	datetime64[ns]	Data inici d'inscripció a la convocatòria
29	**data_fi_conv**	2039	datetime64[ns]	Data fi d'inscripció a la convocatòria
30	**descripcio_conv**	2039	object	Descripció de la convocatòria
31	**universitats**	2038	object	Universitats de la convocatòria on els candidats volen optar (ordenades per ordre de preferència)
32	**idiomes**	1665	object	Idiomes coneguts pels candidats (ordenats de major a menor coneixença i, en segon lloc, de major a menor puntuació atorgada)
33	**nivells_idiomes**	2039	object	Nivells dels idiomes coneguts pels candidats (mateix ordre que "idiomes")
34	**punt_cert_aptitud_pedagogica**	2039	object	Puntuació atorgada per tenir el certificat d'aptitud pedagògica
35	**punt_cert_correccio_textos**	2039	object	Puntuació atorgada per tenir el certificat de capacitació per a la correcció de textos orals i escrits (K)
36	**punt_postgraus**	2039	object	Puntuació obtinguda per haver realitzat algun postgrau
37	**punt_masters**	2039	object	Puntuació obtinguda per haver realitzat algun màster

38	**punt_doctor**	2039	object	Puntuació obtinguda per haver realitzat algun doctorat
39	**punt_experiencia_cat_esobatx**	2039	object	Puntuació obtinguda per tenir experiència docent en llengua i literatura catalanes en ESO o batxillerat
40	**punt_admin_proves_oficials_cat**	2039	object	Puntuació obtinguda per haver elaborat i/o administrat proves oficials de llengua catalana
41	**punt_mat_docent_publicacio**	2039	object	Puntuació obtinguda per haver elaborat material docent per a la seva publicació
42	**punt_experiencia_trad_correccio**	2039	object	Puntuació obtinguda per tenir experiència professional relacionada amb la traducció i/o la correcció
43	**punt_cursos_altres_institucions**	2039	object	Puntuació obtinguda per haver realitzat cursos de llengua a altres institucions (mín. 60 hores)
44	**punt_curs_extensio_uni**	2039	object	Puntuació obtinguda per haver realitzat cursos d'extensió universitària (de 150 hores)
45	**punt_cert_apt_pedag_sinop postgrau**	2039	object	Puntuació obtinguda per tenir el certificat d'aptitud pedagògica (en el cas de no disposar de la titulació de postgrau equivalent)
46	**punt_cert_correc_text_sino postgrau**	2039	object	Puntuació obtinguda per tenir el certificat de capacitació per a la correcció de textos orals o escrits (K) (en cas de no disposar de la titulació de postgrau equivalent)
47	**punt_cert_altres_institucions**	2039	object	Puntuació obtinguda per haver obtingut el certificat de llengua catalana a altres institucions (mín. 60 hores)
48	**punt_cert_extensio_uni**	2039	object	Puntuació obtinguda per obtenir el certificat d'extensió universitària (de 150 hores)

49	**punt_cert_assis_introdidactica**	2039	float64	Puntuació obtinguda per tenir el certificat d'assistència i d'aprofitament del curs d'introducció a la didàctica de la llengua com a idioma estranger, organitzat per l'AVL, la XVU i l'IRL
50	**primera_titulacio**	2039	object	Primera titulació
51	**punt_primera_titulacio**	2039	float64	Puntuació obtinguda per la primera titulació
52	**segona_titulacio**	2039	object	Segona titulació
53	**punt_segona_titulacio**	2039	float64	Puntuació obtinguda per la segona titulació
54	**exper_uni_exterior**	2039	object	Indica si es té experiència docent en llengua i literatura catalanes en l'àmbit universitari a l'exterior
55	**punt_exper_uni_exterior**	2039	float64	Puntuació obtinguda per tenir experiència docent en llengua i literatura catalanes en l'àmbit universitari a l'exterior
56	**exper_uni_dins_domini**	2039	object	Indica si es té experiència docent en llengua i literatura catalanes en l'àmbit universitari dins del domini lingüístic
57	**punt_exper_uni_dins_domini**	2039	float64	Puntuació obtinguda per tenir experiència docent en llengua i literatura catalanes en l'àmbit universitari dins del domini lingüístic
58	**exper_adults_nouni**	2039	object	Indica si es té experiència docent en llengua i literatura catalanes amb adults en l'àmbit no universitari
59	**punt_exper_adults_nouni**	2039	float64	Puntuació obtinguda per tenir experiència docent en llengua i literatura catalanes amb adults en l'àmbit no universitari
60	**exper_estrangeres**	2039	object	Indica si es té experiència docent en ensenyament de llengües estrangeres
61	**punt_exper_estrangeres**	2039	float64	Puntuació obtinguda per tenir experiència docent en ensenyament de llengües estrangeres
62	**titulacio_requerida**	2039	object	Indica si es té la titulació requerida

63	**punt_titulacio_requerida**	2039	float64	Puntuació obtinguda per a tenir la titulació requerida
64	**Llicenciatura**	2039	object	Tipus de llicenciatura realitzada pel candidat
65	**punt_llicenciatura**	2039	float64	Puntuació obtinguda en funció de la llicenciatura indicada
66	**Postgrau**	2039	object	Tipus de postgrau realitzat pel candidat
67	**punt_postgrau**	2039	float64	Puntuació obtinguda en funció del postgrau indicat
68	**Master**	2039	object	Tipus de màster realitzat pel candidat
69	**punt_master**	2039	float64	Puntuació obtinguda en funció del màster indicat
70	**Doctorat**	2039	object	Tipus de doctorat realitzat pel candidat
71	**punt_doctorat**	2039	float64	Puntuació obtinguda en funció del doctorat indicat
72	**llicen_o_grau**	2039	object	Llicenciatura o grau realitzada pel candidat
73	**punt_llicen_o_grau**	2039	float64	Puntuació obtinguda en funció de la llicenciatura o grau indicats
74	**altres_llicen_o_grau**	2039	object	Altres llicenciatures o graus realitzats pel candidat
75	**punt_altres_llicen_o_grau**	1231	float64	Puntuació obtinguda en funció d'altres llicenciatures o graus indicats
76	**idioma1**	1665	object	Idioma més conegut pel candidat
77	**idioma2**	950	object	Segon idioma més conegut pel candidat
78	**idioma3**	306	object	Tercer idioma més conegut pel candidat
79	**nivell_idioma1**	1665	object	Nivell del primer idioma més conegut pel candidat
80	**nivell_idioma2**	950	object	Nivell del segon idioma més conegut pel candidat
81	**nivell_idioma3**	306	object	Nivell del tercer idioma més conegut pel candidat
82	**punt_idioma1**	1665	float64	Puntuació obtinguda per l'idioma més conegut
83	**punt_idioma2**	950	float64	Puntuació obtinguda pel segon idioma més conegut
84	**punt_idioma3**	306	float64	Puntuació obtinguda pel tercer idioma més conegut

85	**uni_opcio1**	2038	object	Primera opció d'universitat on el candidat vol treballar
86	**uni_opcio2**	1555	object	Segona opció d'universitat on el candidat vol treballar
87	**uni_opcio3**	243	object	Tercera opció d'universitat on el candidat vol treballar
88	**pais_uni_opcio1**	2038	object	País de la primera universitat escollida
89	**pais_uni_opcio2**	1555	object	País de la segona universitat escollida
90	**pais_uni_opcio3**	243	object	País de la tercera universitat escollida
91	**dies_convocatoria**	2039	int64	Dies d'obertura de la convocatòria
92	**edat_candidats**	2020	float64	Edat dels candidats
93	**fa_anys_estudis**	1412	float64	Anys que han passat entre que el candidat ha obtingut la llicenciatura i ha aplicat
94	**fa_anys_master**	610	float64	Anys que han passat entre que el candidat ha obtingut el màster i ha aplicat

D'altra banda, per al conjunt que indica el nombre d'universitats finançades i deixades de finançar de forma anual per continents (seleccio_docencia_universitats.xlsx) es descriuen les variables a continuació:

Id de la columna	Nom de la variable	Nombre de files no nul·les	Tipus	Descripció
0	**continent**	780	object	Continent
1	**pais**	780	object	País
2	**any_academic**	780	int64	Any acadèmic
3	**financament**	780	int64	Indica si les universitats van ser finançades. Existeixen dues tuples per a cada país i any acadèmic a partir de l'any 2007.
4	**nombre_unis**	780	int64	Nombre d'universitats que van ser, o no, finançades dins d'un país donat un any acadèmic

8.4.7 Memòries

memories_final.xlsx:

Id de la columna	Nom de la variable	Nombre de files no nul·les	Tipus	Descripció
0	**id**	62361	int64	ID de la memòria
1	**id_persona**	62361	int64	ID de la universitat
2	**any_academic**	62361	int64	Any acadèmic
3	**universitat_confirmada**	62361	int64	Indica si la universitat ha estat confirmada
4	**universitat**	62361	object	Nom de la universitat
5	**poblacio**	61779	object	Població on es troba la universitat
6	**codi_pais**	62356	object	Codi del país on es troba la universitat
7	**professor**	62361	object	Professor de la universitat
8	**responsable_academic**	62340	object	Responsable acadèmic de la universitat
9	**situacio_estudis_catalans_estudis_de_grau**	62033	float64	Situació dels estudis catalans a la universitat en estudis de grau
10	**situacio_estudis_catalans_estudis_de_postgrau**	41641	float64	Situació dels estudis catalans a la universitat en estudis de postgrau
11	**situacio_estudis_catalans_estudis_de_master**	57052	float64	Situació dels estudis catalans a la universitat en estudis de màster
12	**situacio_estudis_catalans_centre_de_llengues**	52071	float64	Situació dels estudis catalans a la universitat pel que fa als centres de llengües
13	**situacio_estudis_catalans_especialitzacio_catala**	62151	float64	Situació dels estudis catalans a la universitat en especialitzacions
14	**situacio_estudis_catalans_especialitzacio_catala_quina**	27133	object	Indica l'especialització de català oferta per la universitat
15	**situacio_estudis_catalans_especialitzacio_catala_credits**	27104	float64	Nombre de crèdits que atorga l'especialització feta
16	**situacio_estudis_catalans_especialitzacio_catala_graduats**	31573	float64	Nombre de graduats en l'especialització de català
17	**dades_docencia_data_inici_curs**	62250	datetime64[ns]	Data en què comença el curs
18	**dades_docencia_data_final_curs**	62250	datetime64[ns]	Data en què acaba el curs

19	**dades_docencia_nombre_estudiants_amb_estudis_catalans**	62241	float64	Nombre d'estudiants amb estudis catalans
20	**dades_docencia_nombre_estudiants_nivell_a2_endavant**	62121	float64	Nombre d'estudiants amb el nivell A2 o major
21	**dades_docencia_filologies**	41214	float64	Indica si hi ha algun estudiant realitzant algun estudi de filologia en una universitat i any donats
22	**dades_docencia_arts_humanitats**	29118	float64	Indica si hi ha algun estudiant realitzant algun estudi d'arts humanitats en una universitat i any donats
23	**dades_docencia_altres_disciplines**	26079	float64	Indica si hi ha algun estudiant realitzant algun estudi diferent d'humanitats o filologies en una universitat i any donats
24	**metodologia_canvis**	25305	float64	Indica si hi ha hagut canvis en la metodologia donat un any i una universitat
25	**formacio**	62361	int64	Indica si s'ha format a algú dins del marc de la subvenció
26	**recerca**	62361	int64	Indica si s'ha fet recerca dins del marc de la subvenció
27	**produccio_academica**	62361	int64	Indica si s'ha creat producció acadèmica dins del marc de la subvenció
28	**proves_certificacio_estudiant_presentat_examen**	62227	float64	Nombre d'estudiants presentats a exàmens de certificació
29	**proves_certificacio_universitat_seu_examen**	62072	float64	Indica si la universitat acull exàmens de certificació
30	**proves_certificacio_nombre_inscrits_basic**	36065	float64	Nombre d'estudiants inscrits a exàmens de certificació del nivell bàsic
31	**proves_certificacio_nombre_presentats_basic**	35736	float64	Nombre d'estudiants presentats a exàmens de certificació del nivell bàsic
32	**proves_certificacio_nombre_inscrits_elemental**	30838	float64	Nombre d'estudiants inscrits a exàmens de certificació del nivell elemental
33	**proves_certificacio_nombre_presentats_elemental**	30036	float64	Nombre d'estudiants presentats a exàmens de certificació del nivell elemental

34	**proves_certificacio_nombre_inscrits_intermedi**	33992	float64	Nombre d'estudiants inscrits a exàmens de certificació del nivell intermedi
35	**proves_certificacio_nombre_presentats_intermedi**	33164	float64	Nombre d'estudiants presentats a exàmens de certificació del nivell intermedi
36	**proves_certificacio_nombre_inscrits_suficiencia**	37548	float64	Nombre d'estudiants inscrits a exàmens de certificació del nivell suficiència
37	**proves_certificacio_nombre_presentats_suficiencia**	36866	float64	Nombre d'estudiants presentats a exàmens de certificació del nivell suficiència
38	**proves_certificacio_nombre_inscrits_superior**	24911	float64	Nombre d'estudiants inscrits a exàmens de certificació del nivell superior
39	**proves_certificacio_nombre_presentats_superior**	24629	float64	Nombre d'estudiants presentats a exàmens de certificació del nivell superior
40	**programes_estudiants_nombre_solicituts**	54502	float64	Nombre de sol·licituds rebudes d'estudiants per a realitzar programes
41	**programes_estudiants_nombre_persones_beneficiaries**	54062	float64	Nombre d'estudiants beneficiaris per a realitzar programes
42	**programes_estudiants_conveni**	41351	float64	Indica si s'accepten convenis
43	**programes_estudiants_conveni_quins_estudiants**	34734	float64	Nombre d'estudiants que han dut a terme algun conveni
44	**programes_estudiants_tercer_cicle**	41351	float64	Nombre d'estudiants de tercer cicle
45	**activitat**	62361	int64	Indica si s'ha realitzat alguna activitat dins del marc de la subvenció
46	**valoracio_suggeriments**	62361	int64	Indica si es valoren els suggeriments d'anys anteriors fets pels lectors
47	**id_activitat**	61737	float64	ID de l'activitat
48	**titol_activitat**	61737	object	Títol de l'activitat
49	**tipus_activitat**	61737	object	Tipus de l'activitat
50	**participants_activitat**	61583	float64	Nombre de participants en les activitats
51	**assistents_activitat**	61737	object	Nombre d'assistents a les activitats
52	**id_assignatura**	62290	float64	ID de l'assignatura
53	**lector_assignatura**	62290	float64	Lector de l'assignatura

54	**estudis_catalans_assignatura**	62290	float64	Indica si l'assignatura conté estudis catalans
55	**assignatura**	62290	object	Nom de l'assignatura
56	**compartida_assignatura**	38526	float64	Indica si l'assignatura és compartida
57	**tipus_assignatura**	60884	object	Tipus de l'assignatura
58	**llengua_assignatura**	60884	object	Llengua amb què l'assignatura s'imparteix
59	**periode_assignatura**	58918	object	Període en què l'assignatura s'imparteix
60	**curricular_assignatura**	41550	float64	Indica si l'assignatura és curricular
61	**credits_assignatura**	60884	float64	Nombre de crèdits de l'assignatura
62	**hores_assignatura**	60884	float64	Hores de l'assignatura
63	**estudiants_assignatura**	58919	float64	Estudiants apuntats a l'assignatura
64	**impartida_assignatura**	62290	float64	Indica si l'assignatura va ser impartida
65	**id_produccio**	50727	float64	ID de la producció
66	**tipus_titol**	50727	object	Tipus de la producció
67	**autor_titol**	50727	object	Autor de la producció
68	**tipus_autor_titol**	50727	object	Tipus d'autor de la producció
69	**puntuacio_curr**	62361	int64	Puntuació atorgada pel currículum d'una universitat (assignatures impartides en català)
70	**valor_puntuacio_e_assignatures**	62243	float64	Valor de les e_assignatures
71	**puntuacio_e_assignatures**	62361	int64	Puntuació atorgada per l'e_assignatures
72	**valor_puntuacio_ranking**	52222	object	Posició de la universitat en el rànquing global (extern a l'IRL)
73	**puntuacio_ranking**	62361	int64	Puntuació atorgada en funció de la posició en el rànquing global
74	**valor_puntuacio_geo**	62361	object	Localització geogràfica de la universitat
75	**puntuacio_geo**	62361	int64	Puntuació atorgada en funció de la localització geogràfica
76	**valor_puntuacio_produccio**	62243	object	Tipus i quantitat de producció realitzada
77	**puntuacio_produccio**	62361	int64	Puntuació atorgada en funció de la producció realitzada
78	**puntuacio_activitats**	62361	int64	Puntuació atorgada en funció de les activitats

79	**valor_puntuacio_e_examens**	61440	float64	Valor dels e_examens
80	**puntuacio_e_examens**	62361	int64	Puntuació atorgada pels e_examens
81	**valor_puntuacio_formacio**	62243	object	Indica si hi ha hagut formació en el marc de la subvenció
82	**puntuacio_formacio**	62361	int64	Puntuació atorgada en funció de la formació
83	**valor_puntuacio_jornades**	62361	object	Indica si hi ha hagut jornades en el marc de la subvenció
84	**puntuacio_jornades**	62361	int64	Puntuació atorgada en funció de les jornades
85	**puntuacio_finan**	62361	int64	Puntuació atorgada en funció del finançament de les universitats
86	**valor_puntuacio_aportacio_irl_prof**	61521	float64	Diners aportats per l'IRL als professors
87	**puntuacio_aportacio_univ_prof**	62361	int64	Puntuació atorgada en funció dels diners aportats per la universitat al professor
88	**valor_puntuacio_aportacio_irl_inc**	27462	float64	Diners aportats per l'IRL a la incorporació
89	**valor_puntuacio_conveni**	62129	object	Indica si hi ha hagut convenis en el marc de la subvenció
90	**valor_puntuacio_incidencies**	62361	object	Indica si hi ha hagut incidències en el marc de la subvenció
91	**valor_puntuacio_futur**	62361	object	Indica si hi haurà canvis en el futur
92	**puntuacio_total**	62361	float64	Puntuació total atorgada a una universitat en un any concret
93	**difusio_activitat_web**	62361	int64	Indica si la web s'ha utilitzat per a difondre una activitat
94	**difusio_activitat_xarxes_socials**	62361	int64	Indica si les xarxes socials s'han utilitzat per a difondre una activitat
95	**difusio_activitat_estudiants**	62361	int64	Indica si els estudiants s'han utilitzat per a difondre una activitat
96	**difusio_activitat_correu**	62361	int64	Indica si el correu s'ha utilitzat per a difondre una activitat
97	**difusio_activitat_cartells**	62361	int64	Indica si els cartells s'han utilitzat per a difondre una activitat
98	**difusio_activitat_universitat**	62361	int64	Indica si la universitat s'ha utilitzat per a difondre una activitat

99	**difusio_activitat_intern**	62361	int64	Indica si s'ha utilitzat la difusió interna per a donar a conèixer una activitat
100	**curr_espec**	62087	float64	Nombre d'assignatures impartides amb alguna especialització
101	**curr_a1_r**	62087	float64	Nombre d'assignatures impartides amb nivell A1
102	**curr_a2_r**	62087	float64	Nombre d'assignatures impartides amb nivell A2
103	**curr_b1_r**	62087	float64	Nombre d'assignatures impartides amb nivell B1
104	**curr_b2_r**	62087	float64	Nombre d'assignatures impartides amb nivell B2
105	**curr_c1_r**	62087	float64	Nombre d'assignatures impartides amb nivell C1
106	**curr_c2_r**	62087	float64	Nombre d'assignatures impartides amb nivell C2
107	**curr_cont_r**	62087	float64	Nombre d'assignatures impartides en total
108	**vp_activitats_dina_r**	53645	object	Indica el tipus de dinamització de l'activitat
109	**vp_activitats_correcta_r**	30419	object	Valoració objectiva d'una activitat
110	**vp_activitats_bona_r**	671	object	Indica com és de bona una activitat