NOVA IMS Information Management School

MGI

Mestrado em Gestão de Informação
Master Program in Information Management

Nortear

A platform to support higher education applicants' decisions

Rodrigo Fernando Ferreira Dias da Silva

Project Work report presented as partial requirement for obtaining the Master's degree in Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NORTEAR - A PLATFORM TO SUPPORT HIGHER EDUCATION APPLICANTS' DECISIONS

by

Rodrigo Silva

Project Work report presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Advisor: Prof. Doutor Flávio Luís Portas Pinheiro

November 2021

ACKNOWLEDGEMENTS

Firstly, I would like to thank the people of Portugal and its representatives for creating a welcoming atmosphere for students from outside Portugal and the European Union. I do not take it for granted and I thank you the opportunity for having such a transformative journey especially during this tough moment we are living in these last two years. Secondly, I would like to acknowledge the patience and encouragement given by my advisor, these were immensely helpful in the moments I needed the most. And lastly, but not least, I would like to thank NOVA University Lisbon – and I mean its whole academic body – for preparing this wonderful course. It allowed me to have a career change and today I am happy with the transformation it brought to my live.

ABSTRACT

Data can have immense public value, and although Open Data has been made available on the internet mostly by governments and public institutions having as a main justification the “public interest”, most of what is turned public, is done so in formats not suitable to the eyes of most of the public which they are aimed to. The high level of digital literacy needed to turn this data into useful information continues to keep virtually inaccessible what was supposed to be Open Data. This work aims exclusively on data over the education sector, but this situation is also true for data involving other sectors of organized society. This Open Data which is virtually “locked data” have grave relevance to students, parents, and society. This work aims to give a humble step towards exploring the lack of accessibility for information extracted from Open Data on the Portuguese educational sector to collect data that are relevant to future graduates on the choice for a university aiming to create a platform where these students can access visualizations linked to key decision factors relevant when choosing a place to study, including information about the city, the country, and the university. This way, this work intends also to create encouragement to more initiatives which can help to “translate” data for the good of the broad public.

KEYWORDS

Open Data; Higher Education; Data Visualizations; Decision Support Systems; Universities; Portugal.

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

One of the characteristics that best define humans and dissociate them from other animals is the ability of giving the available information, to reach a conclusion, and from that conclusion to decide what is thought to be the best direction from a set of possibilities. For humans, reason can supress instincts, and soon we realized that a well pondered decision can mean the difference between life and death.

The dire importance of the decision-making process gave birth to a whole group of sciences responsible for dealing with data and information. And as became evident that the quality of the decision is as good as the quality of the data acquired, early applicants of data as means of helping the decision-making process, devised methods to take care of data procurement, storage, and visualization. Books of the 18th century showing meticulously written weather tables are still preserved and the first weather charts based on that data date back from the 19th century. This is an early example of a science which depends on data and that has evolved in pace with its importance for the economy. Today, billions are spent every year with predictive weather models as today this science has become of great importance, as many more economical activities depend on weather forecast.

And what has changed in the 21st century? As the weather forecast example settles the case for a minded decision process, there has always been a hunger for new technologies that could perfect the activity. So, when computer technology became sufficiently advanced and omnipresent in businesses, this served as a pretext for the advent of what is called today Digital Transformation. Companies of many areas have been forced to embrace the idea of obtaining more and more data over their customers, employees, and even prospective customers as means of improving their business decisions.

The question stands of how Digital Transformation is helping the commoner. As businesses started build up intelligence from data, the public sector followed the lead. Governments and public bodies have their own structure to hold vast amounts of data on many aspects of the economy that affects its citizens. By the force of law, governments must make public a huge share of the data they hold as a way of being transparent with their population. Although, it is still questionable how the people can take advantage of this Open Data, and this is the main issue brought by this Work Project study.

1.1. CONTEXT

On summarizing the knowledge taught during the lectures of Information Management, especially the Specialization in Knowledge Management and Business Intelligence, one might conclude that this course is on the value of data and its ubiquitous and savage nature. Although, this course is also about how to use technology to give access to data, using it as a source of value.

The concept of generation of value has been very much connected with the success of companies or their capacity of creating solutions which cooperate with their long-term prosperity, generating wealth for its shareholders, and answering timely to the society's needs. Lately, companies have been using data as never before to deliver value.

All things considered, it is impossible not to wonder: if data is so valuable to companies, why individuals are not making the same use of it? Going straight to the point: could future graduate students who are

looking for college courses in Portugal be provided with data in form of visualizations to reach better decisions on choosing where to study?

This work is a humble step towards demonstrating ways of democratizing Open Data in Portugal and giving it meaning and purpose. The definition of purpose for the sake of this work will be helping those who are looking into continue their academic journey in a Portuguese public institution to reach a sound decision, especially when they are moving from their hometown. As for "to give it meaning", it goes as translating raw Open Data into a web visualization tool which aims to help the student on his research for a new place to study.

1.2. Motivation, Background and Problem Identification

Public colleges in Portugal, which are institutions partially financed with state funds, they provide the state with data over many aspects of its operation, being the most basic, for instance, the number of students enrolled, and their basic data, such as gender, nationality, age, and course enrolled. This and further data are published in one or more of the websites which are specialized in publishing Open Data. What happens is that, although data is made public, there is little effort on making this data palatable to the broad public. More even, when there is an effort on creating visuals, no explanation is given on the relevance of the information when it is needed to understand the data.

As it will be shown, this is a problem with Open Data initiatives worldwide. Many of the platforms in Portugal are financed with public resources, such as DADOS.GOV, which is a catalogue of data sources spread in other websites on the Internet, and there are not enough resources destined to improve the accessibility of the data, or to foster engagement among the public.

Therefore, there is a fair amount of data available which can be used to inform students so they can be better prepared to pursue their further course of studies, although, an effort needs to be made to prepare the data and contextualize it with the public it is destined to.

1.2.1. Background: Education, Economy, and Fulfilment

Education is the heart of a country in the sense that talents formed during the graduation years will be employed by the sectors of the economy to develop the solutions to the problems a nation has. They deliver oxygen to a country's economy. It has been researched that if a student is happy with its experience, the chances increase of immerse itself in the culture he is inserted (Anderson & Lawton, 2015). To live a good experience during graduation years might be a matter of fit also and this fitting can be related to personal traits and the proper research source would show the student the better choices for him.

1.2.2. The Data Available as Part of The Problem

Much of the data found during the first phase of research for this work is open and available online, although, in formats which are hard to grasp. There are data available in PDF reports which, even though aimed to present information in context, are of hard comprehension.

Data is also spread over a couple of websites belonging to the government of Portugal or to non-profit organizations. This situation also difficult the research for information.

1.2.3. The Visualizations Available as Part of The Problem

Even when the data sources found show charts, they show them individually; this means the student who is looking for information will have to navigate between a few charts before finding what he wants. As learned, this is not the best option when one needs to reach a decision. Therefore, the solutions available do not attend to the public demand.

1.2.4. Graduates Needs and Data Sources

The aim is to conciliate the needs and anxieties of future graduates with the data available and not available. There are plenty of academic work over the preferences or choices of future graduates when going through the options of universities. Researching these academic works, it may be possible to list the main concerns they have and link them with the right set of data as shown in Table 1.

Graduates Concerns	Group of data related
Can I find a job while living in this city?	Economy, Employment
How long does it take to find a job in this city?	Employment
How this university's reputation influences in my chances on the work market?	Employment
How this course's reputation influences in my chances on the work market?	Employment
Are there good options of leisure in the city?	City, Leisure
What is the number of papers published by year by the University of Porto?	Education, University

Table 1 - Example of future graduates' concerns and data available

There might be questions which have no available data in order to answer them and for this reason we are going to need to request institutions to provide us with the data. In the chapter 3 (Methodology) it is going to be discussed how it is going to be done.

1.2.5. Current Initiatives

There are ongoing initiatives abroad trying to tackle the lack of data and information over the educational system. In the USA, the website datausa.io is a platform where a variety of data over different subjects – inclusive education – Is collected and displayed in the form of charts.

Also, the website legacy.dataviva.info shows a variety of diagrams and charts over the educational sector and other info on different topics, such as occupations, work force, and economic data.

These are excellent examples of what can be done in the field, and both are an inspiration to be followed, although, more examples might be found during further research, and they can bring more ideas to enrich this work.

Although Open Data platforms are not perfect, they provide the basics so the civil society can build up from it. The data these institutions publish are the essential building blocks for a platform where we can improve this data and attack the issues presented in the problem definition.

There was never a better time to start this initiative. Today we have the data, the skills, and the technology to transform data into actionable information to the broad public.

1.3. Study Objectives

With all the tools we have at our disposal, we must as a society to use it in favour of our fellow citizens. Business Intelligence can go beyond business being a data intelligence framework at the service of society, improving our daily decisions.

In this work, the basic achievement is to use open-source tools and Data Science techniques to create a platform which will leverage the use Open Data to improve the odds of a student to find a good fit when choosing a college and city where he would continue its academic education.

1.3.1. Proposed Solution

If data is not being made available in a format that the general public, and in special the future graduates, can rapidly grasp and use it to their best interest, it becomes clear to this author, especially in face of the recent learnings throughout the last twelve months, that a Data Visualization solution can be devised to deal with this challenge.

To create a platform with excellent standards of usability and the proper visualizations based on the proper data is a huge challenge, but quality is essential in data visualization if we want our public to reach better outcomes during the decision-making process.

1.3.2. The Data Needed

Acquiring the right set of data perfectly aligned with the most important questions students have, is a must. Data that are kept private in universities might be requested in order to provide better visualizations.

To know in detail what drives the students' decisions, and their preferences will impact the choices for data to be explored in this work. Therefore, this is going to be reviewed during further research. However, there are a set of data that are known and necessary for the study as listed in Table 2.

Set of Data	Source
Economy	PORDATA
Education	DGEEC, PORDATA
Migration	INE
Health	PORDATA
Environment	dados.gov
Crime	dados.gov, INE
Housing	INE

Table 2 - Sets of Data and Sources

These sources are being reviewed in order to define which data is going to be collected and how. Some of these sources have both information about the same subject, in this case there must be a method to determine what source to use. This will be defined in the right moment soon.

1.3.3. Telling a Story

Having understood the students desires and, moreover, acquired the proper data, then a technique of visualization must be chosen in order to help the decision process. The aim of any visualization is to transform data into information (aggregated data with meaning) in order to communicate an idea (Sharda et al., 2017). This choice must follow the best practices taught during the course and further research.

Dashboarding has a few techniques, although one which might work well for this project is storytelling. For a single concern such as "Am I going to find work fast living in Porto?", a set of charts could be displayed having on their titles the story to be told. An example of this technique might be found in the sketch below in Figure 1.

The Figure 1 shows a simple sketch of how a story need to be told. Of course, this is just an illustration of the storytelling process, and the actual layout of the solution is going to be discussed in the next section. Although, this is enough to illustrate an aspect of the preoccupation of the student and help it to reach a conclusion in a logic fashion: "If the population is stable, the unemployment is decreasing, and there are more jobs available, so I might be able to find a job easily".

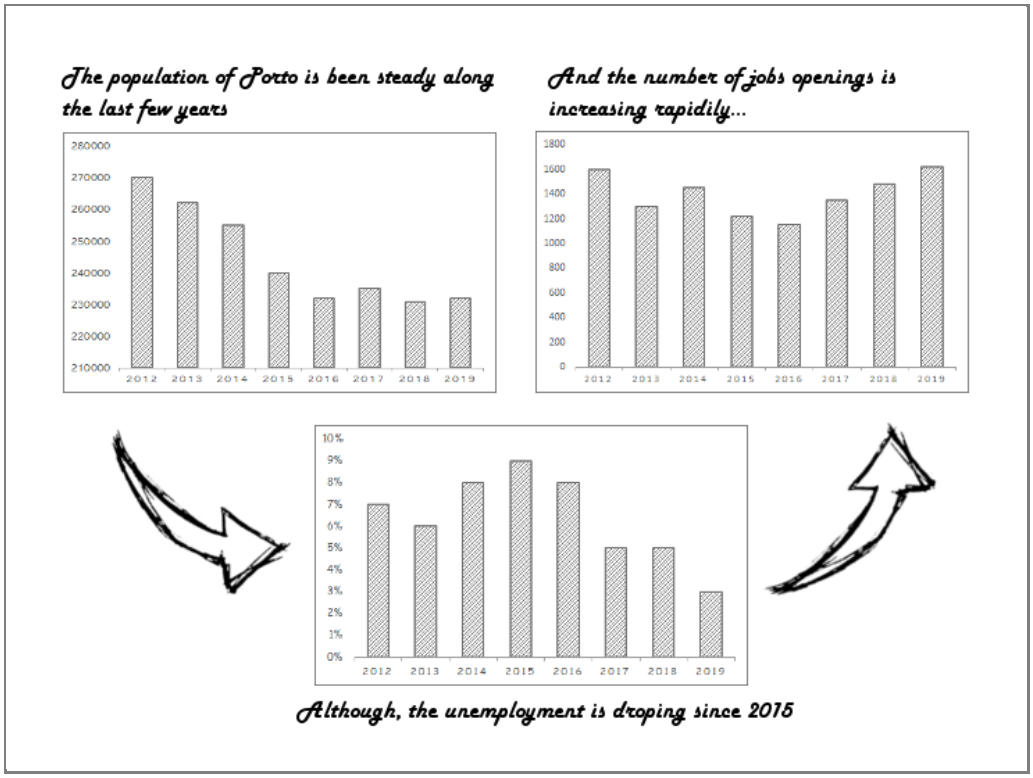


Figure 1 - Example of Storytelling

1.3.4. Chosen Universities and Cities

The public Portuguese universities are going to be chosen using as a guideline two of the most prestigious institutions which benchmark the quality of the education around the world. These are the

QS World University Ranking (QS WUR), and The Times Higher Education WUR (THE WUR). For both publications, the years of 2020 and 2021 will be considered. Given the nature of the events of the last two years, it would be fair not to consider only the year of 2021. The list of public Portuguese universities that figure in at least one of these rankings are listed below in Table 2.

University	City	QS WUR	THE WUR
University of Porto	Porto	357º	451º
University of Aveiro	Aveiro	586º	701º
University of Coimbra	Coimbra	431º	701º
University of Lisbon	Lisbon	357º	551º
Nova University of Lisbon	Lisbon	428º	376º
University of Algarve	Faro		401º
University of Beira Interior	Covilhã		701º
ISCTE University Institute of Lisbon	Lisbon		701º
University of Minho	Braga	596º	401º
Polytechnic Institute of Porto	Porto		401º [1]
University of Trás-os-Montes and Alto Douro	Vila Real		401º [1]
University of Évora	Évora		401º [1]

Table 3 – The Portuguese Universities Included in the Work Project

The Polytechnic Institute of Porto, the University of Trás-os-Montes and Alto Douro, and the University of Évora were not present in the 2021 edition of the THE WUR, although in 2020 they were ranked well having in consideration the other contenders and surely they cannot be left out of this list.

The twelve universities chosen are spread along the Portuguese territory in 9 different cities as shown in the map of Figure 2 surrounded by the red mark.

[1] 2020 publication

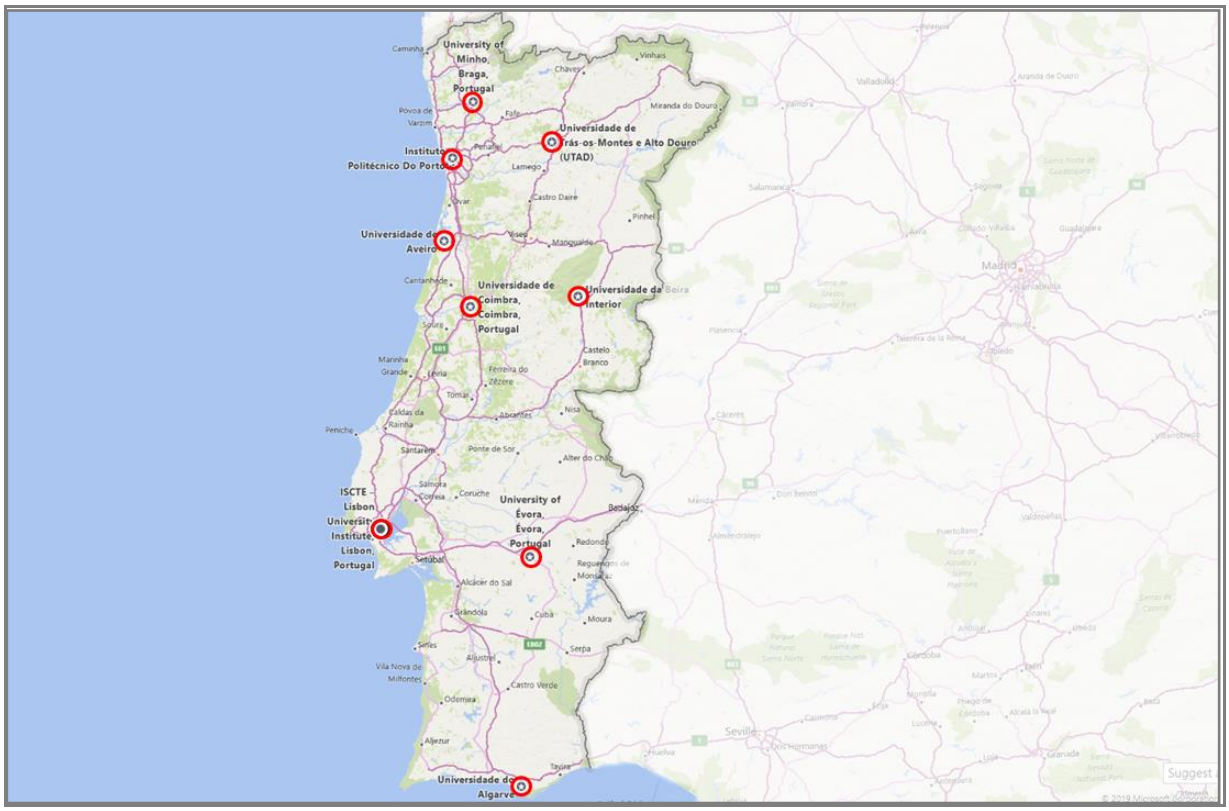


Figure 2 - The cities where the universities are located

1.3.5. Universities, cities, and data

As said, the data is being chosen according to the preferences and necessities of the future graduates. These needs might include at least, but not only, the data listed in "The Data Needed" and they may surround – or extrapolate, as it is expected – information about cities and universities listed in "Chosen Universities and Cities".

However, at this moment it can be asserted that at least the set of data listed in Table 4 is going to be needed throughout this work.

Category	Sets of Data
Universities	Unemployment of former students, Number of Students, Research Production, Yearly Investment, Nationality of Students, World Ranking, Ranking by Subject Taught,
Cities	Housing, Leisure options, Weather, Health, Cost of living

Table 4 - Category and Specific data needed

Every set of data picked will need to have a justification for its choice connected to the students' needs. In the course of this work, if it is not possible to acquire some of the sets of data listed, the

consequences for the objectives of this work need to be clear and recognized in the written part of the work.

1.3.6. Keeping the Solution Alive

After the solution have been deployed there is the concern data will become obsolete and consequently the benefits will not be maintained in long-term. A solution must be found to deal with it and some ideas are discussed in the chapter “3. Methodology”.

2. LITERATURE REVIEW

In this chapter it is going to be shown how actual research supports that the democratization of data through the use of a data-driven decision support system can lead to the improvement of choices of prospective college students when looking for a university, and a new place to live. Furthermore, it is demonstrated by current literature that this data-driven decision, when accurate, can lead to positive impacts that go beyond the sphere of the students' lives.

2.1. DEMOCRATIZATION OF DATA

2.1.1. What is Data Democratization?

There is a slight difference in terms of meaning when authors refer to data democratization. Patil & Mason (2015, p. 6) have a data focused approach, explaining it as making data widely available whether within companies among employees – having only the law as a restraint – or outside by governments, when these create platforms where data is made available to the public. Meanwhile, for other authors data democratization occurs when data serve a purpose in society (Axelrod, 2019; Sawicki & Craig, 1996; Treuhaft, 2006). According to them, truly democratic data is not seen as widespread data that requires technical expertise to be understood, yet democratic data is data that is made meaningful to the not well versed in technology. This work will follow this last interpretation of data democratization.

2.1.2. A Little History

In the 90s' the increasing availability of data gained steam in the United States of America when the president Bill Clinton signed an Executive Order that created a special infrastructure for storing and making publicly available the governments' geospatial data. This paved the way for the advent of companies in the USA specialized in collecting and interpreting raw data in order to engage communities to participate in public policy (Sawicki & Craig, 1996).

By 2013, four years after the president of the United States of America, Barak Obama, released his memo over transparency and open government, the USA agencies had already made public around 270 000 datasets (Luna-Reyes et al., 2019). In Europe, the European Commission approved in 2011 a document on the reuse of the commission's documents (European Commission, 2011) which opened the way for the release of the EU Open Data Portal in 2012, and today the European Union have a detailed plan on how to deal with open data throughout the continent (European Commission, 2021).

2.1.3. The Social Relevance

Together with the growing amount and availability of data, non-profit companies specialized in intermediating data grew both in number and in importance in the USA and they play an important role in communities' participation and social change (Treuhaft, 2006).

2.1.4. The Context Within the Current Work

This author sees this Work Project as the work of one of these mentioned data intermediaries; on making public information which today is only readable to those who have the proper technical skills, is expected that this work will also be a tool for social change in Portugal.

2.2. OPEN DATA

2.2.1. What is Open Data?

Authors agree on the underlying idea that open data from a variety of sources (mostly publicly obtained) is data which is made available so individuals can manipulate it and use it as a mean of increasing governments' transparency and public participation (Gurstein, 2011; Janssen et al., 2012).

Although an understanding of what is open data has been widely formed, the same cannot be said about a definition of standards for open data, neither of what the word "open" means; in other words, data that is said to be opened might be subject to rules that limit what can be done with it (Kitchin, 2014).

Having these disparities in mind, this author must therefore analyse the benefits and setbacks of open data.

2.2.2. Benefits and Limitations of Open Data

As seen before, open data creates a base in which public participation and transparency can be fund, although, open data can also foment entrepreneurship as it can be used not for direct profit, but instead to test new business models (Lakomaa & Kallberg, 2013).

Even though companies can easily see benefits in open data, these companies have professionals specialized in data analysis, for this reason in some cases is observed that to open data to be used for social good it needs intervention so information can be available to the broader public, otherwise open data might just empower the rich instead of helping the poor (Gurstein, 2011).

2.2.3. Open Data in Portugal

Portugal has a vast amount of open data initiatives. Whether they have been created by foundations or maintained by the government, these initiatives store data related to many aspects of Portuguese society, especially those which are going to be subject of analysis in this work and can be seen in Table 1.2.

In 2013 the World Wide Web Foundation released the Open Data Barometer. It was designed to check adoption of open data taking into consideration three dimensions as listed in Table 5 (Open Data Barometer, 2021).

Dimension	Description
Readiness	For open data initiatives
Implementation	Of open data programs
Impact	That open data is having on business, politics, and civil society

Table 5 - Dimensions of Open Data Adoption

Since the last edition in 2018 focused on the leaders, Portugal was therefore not listed. Even though, in the release of its 4th edition, in 2016, which counted with 100 countries, Portugal was listed in the 34th place as shown in Table 6 below.

Rank	Country	Score	Readiness	Implementation	Impact
1	United Kingdom	100	99	100	94
2	Canada	90	96	87	82
3	France	85	100	71	88
4	USA	82	96	71	80
5	South Korea	81	95	59	100
6	Australia	81	85	78	78
7	New Zealand	79	92	58	99
8	Japan	75	84	60	89
9	Netherlands	75	94	64	68
10	Norway	74	77	71	73
34	**Portugal**	**42**	**58**	**47**	**16**

Table 6 - Open Data Barometer's Ranking for Open Data Adoption

As seen above, Portugal scored poorly in "impact", a measure that stands for how data is being used to cause change. This is clearly evidence that this work is going in the right direction making use of the available data for the benefit of Portuguese society. Also, there are opportunity to expand the number of initiatives that make use of Open Data.

2.3. The Importance of Open Data for the Education Sector

Education is the heart of a country in the sense that talents formed during the graduation years will be employed by the sectors of the economy to develop the solutions to the problems a nation has. They deliver oxygen to a country's economy. It has been researched that if a student is happy with its experience, the chances increase of immerse itself in the culture he is inserted (P. H. Anderson & Lawton, 2015). To live a good experience during graduation years might be a matter of fit also and this fitting can be related to personal traits and the proper research source would show the student the better choices for him.

2.4. The Relevant Data to Make Available

Acquiring the right set of data perfectly aligned with the most important questions students have is a must. Data that are kept private in universities might be requested in order to provide better visualizations.

2.4.1. The Source of Data Available

To know in detail what drives the students' decisions, and their preferences will impact the choices for data to be explored in this work. Therefore, this is going to be reviewed during further research. However, there are a set of data that are known and necessary for the study as listed in Table 7.

Set of Data	Source
Economy	PORDATA
Education	DGEEC, PORDATA
Migration	INE
Health	PORDATA
Environment	dados.gov
Crime	dados.gov, INE
Housing	INE
Academic Production	dados.gov

Table 7 – Necessary Sets of Data and Sources

These sources are being reviewed in order to define which data is going to be collected and how. Some of these sources have information about the same subject, in this case there must be a method to determine what source to use. This will be defined in the right moment soon.

2.4.2. Data Over Education Available

Table 8 shows the different sub-sets of data under the group "Education". More data is being analysed and collected and will be added to the table during this process.

Sub-Sets of Data	Data	Source
Docents	Age, Gender, Area	PORDATA, dados.gov
Institutions	R&D Expenses (by economical activity), Researchers	dados.gov
Students	Gender, Formation, Scholarship Grants, Region	PORDATA, dados.gov
Portugal	Expenses with R&D, Scientific Production	dados.gov,INE

Table 8 - Subsets of Data and Sources

2.4.3. Student Concerns

It is possible to assert that one of the main drivers that lead prospective students to choose a determined institution as its reputation (Agrey & Lampadan, 2014; Briggs, 2006, 2006; Dias, 2013; Tavares & Cardoso, 2013). Also, it is clear the connection prospective students draw between what they describe as reputation of an institution and their own chances of having a successful career (Conard & Conard, 2000). More even, to have an "attractive career" looks like a common goal among Portuguese students (Dias, 2013).

The choice of a Major is linked to career opportunities (Malgwi et al., 2005). When dealing with expectations over higher education, anxious about their career prospects (Teixeira et al., 2015).

When turning again the eye to the institutions, students also seem to favour those which are the best in a certain subject, have good facilities, and its graduates get good jobs (Tavares & Cardoso, 2013). In this sense, as employment and career is an almost omnipresent factor, institutions which offer the possibility of an internship often outscore those which does not by the point of view of prospective students (Silva et al., 2016).

Studies differ when the subject the how close to home students want to be. Some studies say that at some extent the location is important, but it does not need to be close to home (P. Anderson, 1999), on the other hand other studies show some preference for being close to their hometown (Drewes & Michael, 2006).

2.4.4. Translating Concerns into Data

Whether a prospective undergraduate is asked to list the reasons behind a choice for a particular institution or a major in special, it is impossible not to draw a connection between these motives and a certain anxiety over the following themes: employability, and career. Therefore, the dataset showed in Table 9 had to be examined.

Dataset Name	Source
"Caracterização dos desempregados registados com habilitação superior - junho de 2020"	DGEEC

Table 9 – Datasets over Employment

From this dataset it was possible to extract information over unemployment of early graduates with level of detail enough to allow it to be converted in the following charts listed in Table 10.

Visual	Group	Description
Unemployment	Employment	Unemployment among early graduates
Comparison	Employment	Unemployment compared with other universities with similar numbers
Courses - Lower	Employment	Courses with a lower level of unemployment
Courses - Higher	Employment	Courses with a higher level of unemployment
Areas - Lower	Employment	Areas with a lower level of unemployment
Areas - Higher	Employment	Areas with a higher level of unemployment

Table 10 - Employment themed visuals

These concerns also open space for looking into where former graduates of a certain institution are living and working. Linked-In has an area of its website which shows precisely that, and from that set of data the charts shown in Table 11 were also built.

Visual	Group	Description
Where They Live	Former Students	Unemployment among early graduates
Where They Work	Former Students	Areas with a higher level of unemployment

Table 11 - Former students: where they live and work

Speaking about reputation, rankings are source of knowledge when people research. Although, if you grew up in a country the word of mouth will have made its way to convince people of which are the more reputable institutions. Still, companies which assemble rankings they have a methodological way of ranking institutions; therefore, it was important to research the ranking of each of the institutions studied. In fact, the universities chosen for this study are the 12 Portuguese universities which are figured in international rankings. Table 12 shows a list of charts based on 4 of the most know university rankings.

Visual	Group	Description
Times Higher Education World University Rankings	Rankings	University rankings from 2017 to 2021
QS Ranking	Rankings	University rankings from 2018 to 2021
EDUNIVERSAL	Rankings	Master's courses rankings
QS Masters Rankings	Rankings	Master's courses rankings

Table 12 - Rankings charts

Students are also concerned with the academical environment and whether their personal traits are reflected on the environment they will find when starting a new academic journey. Table 13 shows a set of visuals chosen to give the students a dimension of the environment.

Visual	Group	Description
Gender	Academic Environment	Unemployment among early graduates
Gender on bigger courses	Academic Environment	Unemployment compared with other universities with similar numbers
Nationalities	Academic Environment	Courses with a lower level of unemployment
Scholarships granted	Academic Environment	Courses with a higher level of unemployment
Offer of student dormitories	Academic Environment	Areas with a lower level of unemployment
Academic Body	Academic Environment	Areas with a higher level of unemployment

Table 13 - Academic Environment themed visuals

Finally, students who are moving to a new town are frequently concerned about many of the aspects of the new place they are going to live. Some of them plan to live in this new place during the cycle of their studies, others are open to possibilities and depending on their experience they might bring their family to live with them or even stay and make their own family on the new place they call home. Table 14 show many charts built from datasets related to the city. Some datasets might indicate how lively the city is, others how the city council is concerned with matters that are fond to students nowadays, such as how the town recycles its waste.

Visual	Group	Description
Accommodation Cost	City Data	Accommodation Costs on Uniplaces in comparison with other towns in Portugal
Student Population	City Data	Student population in town in comparison with other towns in Portugal
Population Density	City Data	Population density in comparison with other towns in Portugal
Unemployment	City Data	Unemployment in comparison with other towns in Portugal
Buying Power	City Data	Buying power in comparison with other towns in Portugal
Climate	City Data	Climate in the chosen's university city
Distances	City Data	Distances between the chosen's university city and other Portuguese cities
Gini Coefficient	City Data	Gini coefficient power in comparison with other towns in Portugal
Net Income	City Data	Net income in comparison with other towns in Portugal
Environment	City Data	Investment in environment in comparison with other towns in Portugal
Recycling Index	City Data	How much this town recycles in comparison with other towns in Portugal
Average Rent Prices	City Data	Average rent prices in comparison with other towns in Portugal
Culture and Sport	City Data	Investment in culture and sport in comparison with other towns in Portugal
Crime	City Data	Crime ratings in comparison with other towns in Portugal

Table 14 - Academic Environment themed visuals

2.4.5. The Pursue for Data

The sources of these datasets are websites of institutions such as the "Instituto Nacional de Estatística" (INE) and the "PORTDATA", as a few examples. Although these websites make a great contribution on making these datasets available and most of them are update regularly, it is not easy to search through

the vast list of datasets involving a whole number of subjects. In the case of INE's website, you can filter the data by subject, although you need still navigate among a big list of datasets for the lack of other filters. When a dataset is found, there are the possibility of filter the data, but this is not intuitive enough, what makes it hard to extract the needed data. When you finally extract the data, the file is made available in the way it is shown on the screen, with many merged headers that surely will have to be dealt with during the ETL process.

It is not easy to find a raw dataset, and by "raw" it is meant not exporting exactly what is being shown on the screen, but a table in the format of a data table. Although "PORTDATA" make available raw datasets, it does not have the same amount of information as INE has, and they also are not updated in the same frequency.

2.4.6. Discarded Data

As there is plenty of data over teachers age, and gender both for private and public higher education institution, this author must admit he was tempted to draw a connection between age and, for instance, a low adoption of ICT tools, which could impact the factors which lead a student to choose a higher education institution, such as its excellence in a determined field, reputation, or having good equipment for teaching. Nevertheless, studies show that there is lack of evidence between age of teaching staff and adoption of new and disruptive techniques (Martí-Parreño et al., 2016). Furthermore, pedagogy encourages the use of ICT tools as an inherent part of the learning process (Loveless, 2011).

2.5. VISUALIZATION OF DATA

Having understood the students desires and, moreover, acquired the proper data, then a technique of visualization must be chosen in order to help the decision process. The aim of any visualization is to transform data into information (aggregated data with meaning) in order to communicate an idea (Sharda et al., 2017). This choice must follow the best practices taught during the course and further research.

Besides the choice for a visualization technique, one might still wonder why visualization is such an important matter when dealing with democratizing data. This topic deserves further analysis.

2.5.1. The Importance of Visualizations When Democratizing Data

As said before, data is truly democratic when the society can make use of it. Governments around the world are making Open Government Data widely available for its citizens who can have instantaneous access through the internet to a wide variety of raw data on diverse aspects of their country, although, as data is made available as is – mostly given the costs involved on the tasks of processing the data and creating visualizations (Ding et al., 2011) – most of the interested parts do not have enough technical knowledge to understand the data and rely on thirty parties to create visualizations which might be the medium which can eliminate this problem translating data into useful information to the broad public (Graves & Hendler, 2014).

2.5.2. Considerations Over Ways of Presenting Data

Visualizations can be a powerful tool for persuasion (Hahn, 2018) and although in this work it is not intended to push a graduate towards any decision, it is important that the message transferred through the visualization technique chosen is clear. Also, according to Hahn (2018) the choice of visualizations should consider the public to which it is intended. For this reason, in this section some visualization techniques are going to be explored contextualized with the challenge ahead.

3. METHODOLOGY

The devise of a solution to a problem normally involves the use of a known method. If in one hand methods can fairly accelerate the development of a project, on the other, given the peculiar nature of some problems, they need a revamp of existing methods or completely new ones.

This work deals with a common problem: to provide a software solution to a known problem. And both the solution developed, and the questions it is designed to answer should be based on academic research as pictured on Figure 3.

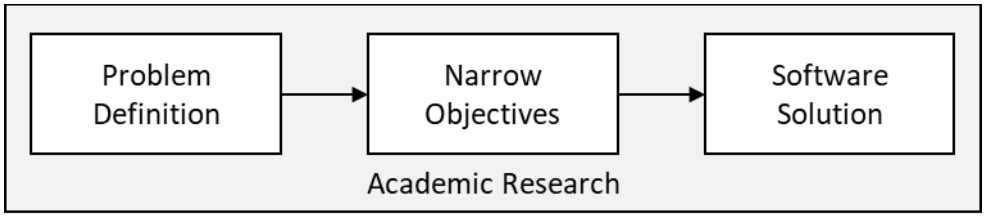


Figure 3 - Methodology Architecture

In the present work, the chapters 1 and 2 dealt with the definition of the problem and the narrowing of the objectives. In this chapter it is going to be presented how the software solution was assembled so the objectives could be reached.

3.1. Software Development Methodology

In software development, many methodologies were assembled to deal with the delivery of a software product. Not so long ago the now infamous Waterfall Methodology used to be the golden standard of software development methodologies before the arrival of the agile methods that were embraced by both companies and developers.

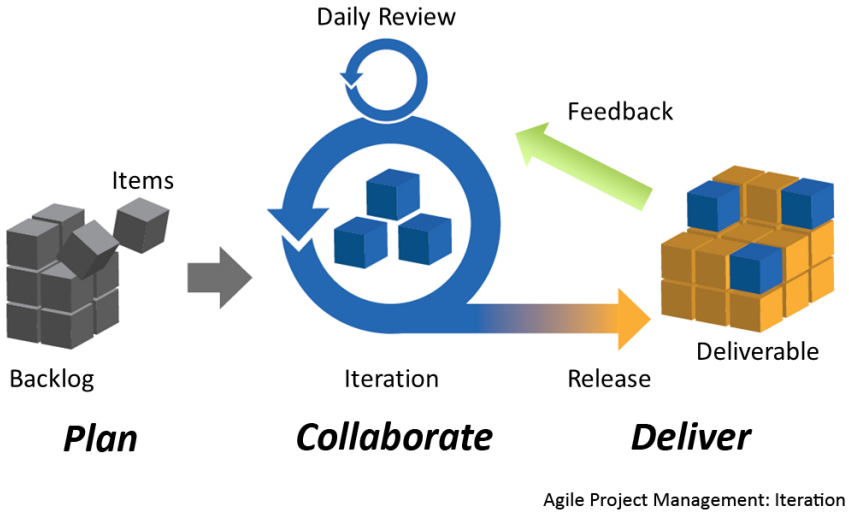


Figure 4 - Agile Methodology Diagram

As shown in Figure 4, in an agile methodology, the idea is to deliver value to the client at the end of each iteration that lasts around two weeks. By “delivery of value” what is meant is that a deliverable would be ready for production by the end of the iteration. This means that the client does not have to

wait until all the parts of the whole are ready – which would be the case in a Waterfall methodology – but it can start to profit from the delivery of part of the product right away.

In this work, what was done was to treat each part of the product as an item of the backlog and, although there was no production environment at the end of an iteration, the item was tested and was ready to be assembled with the other parts. Every item of the backlog was treated as a requirement the these are going to be presented next.

3.2. REQUIREMENTS

Having as a goal to deliver value in iterations the requirements were planned as listed in Table 1.6. Even though it was not possible to define the role of "client" to check the value of the deliveries due to some constraints, this method allows not only the delivery of value, but also to stablish a check point to verify if something must be changed to improve the quality of the deliverable.

The requirements were divided into Functional and Non-Functional as they represent two very distinct types of needs. Functional requirements are those which affect directly how the user interacts with the solution on a day-by-day basis. Non-functional requirements are not as pervasive from the point of view of the user, as they represent hidden aspects of the solution, such as the technology to be used to host the website, for instance.

Among the functional requirements, there are different kinds of categories varying in accordance with what it affects as also shown in Table 15. These categories can be related to "data", "visuals", or even "look and feel".

Requirement	Category
The system must have a configurable data injection framework	Data
The system must have a website to host the visualizations	Visuals
The website must allow the user to view all the visuals of an institution in a single page	Visuals
The website must have e "Welcome" home page	Look & Feel
The website must have a webpage dedicated to compare the data of the cities	Visuals

Table 15 - Functional Requirements

Non-functional requirements are important also for the experience of the user, as they can affect performance, for instance. Therefore, it is important to carefully ponder these hidden aspects of the solution. Normally, tests are carried out to determine if non-functional requirements are not affecting negatively the experience of the user. Here these tests were necessary in some cases, and they will be discussed in the following chapters.

The Table 16 shows a list of non-functional requirements considered during this project.

Requirement	Category
The system data must be based on a free SGBD	Non-functional
The visuals must be based on a free solution	Non-functional

Table 16 - Non-Functional Requirements

3.3. TECHNOLOGY

With the requirements defined, it is time to decide which technologies can help to near the gap between desire and reality. In companies, most of the time it is not a single project which defines the technology to be used in an initiative. Companies have constraints such as the existence of legacy environments and technologies that they cannot ditch without substantial investment. Normally, new projects make use of existing architecture, and this must be considered even during the requirements gathering phase.

Academic work is “happier” in this sense, as students are allowed to choose from a set of tools and technologies in accordance with their supervisor not having to harness their creativity and will to learn different technologies.

Tool	Reason/Characteristics
Visual Studio Code	A visual IDE with support for Python and integration with Git
Git	To keep the source code. It allows backups of the source code and versioning
Bootstrap	CSS library for webpage layout design with advanced options
Apache Server	Web server for hosting the webpage
Python 3	Programming language with high abstraction
Highcharts	JavaScript library for building charts
CSS	Programming language for webpage layout design
HTML5	Programming language for building responsive webpages
Django	Data framework and template management library
PostgreSQL	An Open-Source database
Docker	Development Environment
Trello	Project organization App (Kanban)
Ubuntu 20.04 LTS	The OS which will hold all the rest

Table 17 - Technologies used for the project

Table 17 shows a set of technologies which were chosen to help to deliver the previously listed requirements. These tools affect both the frontend, or what the user sees and interacts with, and the backend, which can be the hidden aspects of the solution – although they might the user experience.

3.4. ARCHITECTURE

A technology architecture helps to control basically where every aspect of a solution fit on the grand scheme of things. For big companies it brings control on an otherwise too complex to manage park of technologies. Companies such as banks, they have an area specialized in managing how a solution fit in its existing architecture. The architecture staff even take part in the approval of solutions to be implemented.

As an academic solution is simpler, the architecture diagram is not as meaningful, although it helps on bringing a better understanding on how each one of the technology choices are placed as shown in Figure 5.

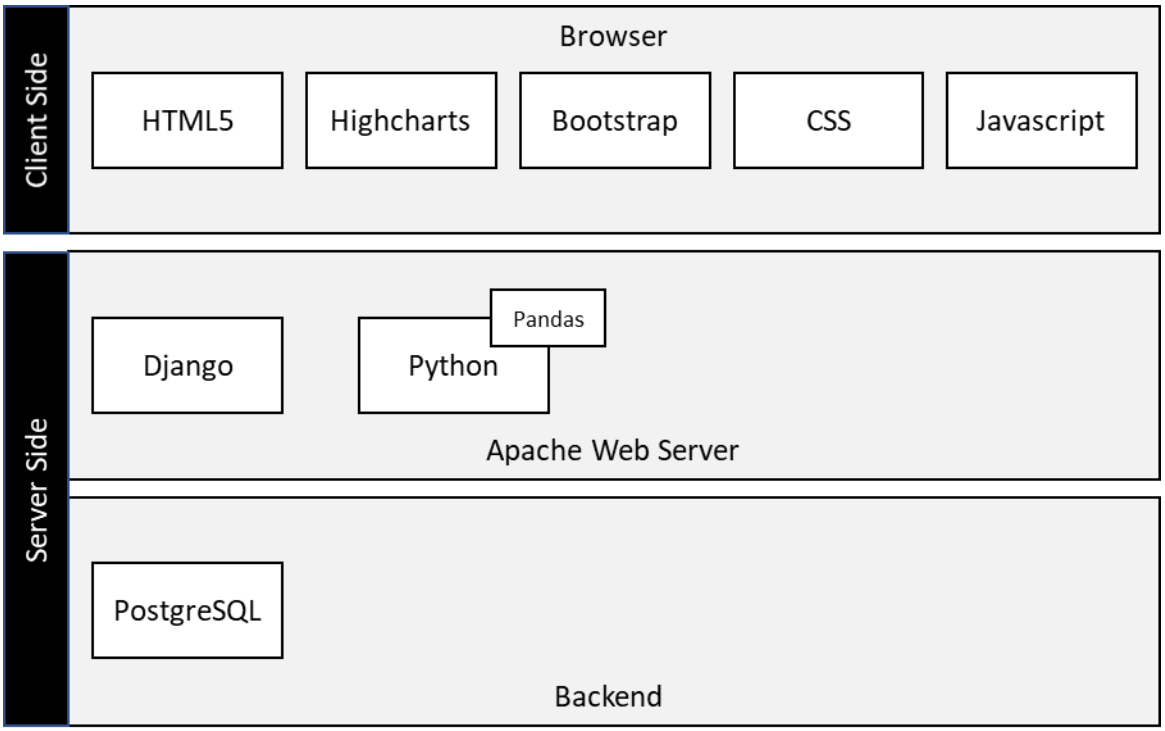


Figure 5 - System's Architecture

3.5. DATA

Before discussing the data model or how the system search for data to assemble the charts, it is important to show how data should be handled.

As is already known what data should be looked for, there will be a process for collecting this data and a decision on where to store this data. Data can be found in various formats, and this is one important aspect that will guide the development of the solution, which must be able to access data available not only in different formats, but also in different levels of quality.

3.5.1. Data Storage

Speaking about data storage, from the start it is possible to assert that data to be collected for this project appears to be structured data, which means data that is already presented in lines with the same columns, and data types are the same for every line among the columns. For cases like this, data could be handled in a relational database, which would allow the modelling of a Kimball model. Although, as this project is being also developed having in mind the possibility of future development, the choice for relational database technology could contemplate a solution which could handle semi-structured or non-structured data in the future and this makes the case for PostgreSQL as a preferable choice, as shown in Table 17.

3.5.2. Data Procurement

In this phase the data needed is gathered and more should be learned on which kind of data the system must be able to handle. This input might also affect the technologies listed in table 1.8.

As an example, in chapter 2 it was presented that one of the main concerns of those who are looking for a continuation of their academic trajectory is how the course to be chosen will improve their position in the work market. One dataset found that might help on defining this is named "Unemployment Among Early Graduates". It is a structured dataset in the format of an Excel spreadsheet with the extension "xlsx". From this observation and knowing that a data pipeline solution will have to be created to insert data into a relational database, it is possible to look for solutions which would handle many kinds of different file formats and that would allow different types of trans-formations, or even permit the use of data science tools. In this case, the use of Python and the library Pandas seem to be a sound choice.

This same process was carried out for every one of the datasets found. Although in this project we are dealing basically with "xls", "xlsx", "csv", and "json" file formats, the technologies chosen open space for different kinds of datasets.

3.5.3. Data Quality

As all the data needed was gathered and the technologies to handle data were defined, it is time to know better the datasets and learn how we can improve them.

Improvements on the data quality might involve elimination of null values, cleaning lines which are not part of the data – headers and footers –, reduction of columns, setting the right data types for columns, filtering lines which are not part of the problem analysed, unify names that might be non-uniform, such as differences of letter casing which might cause problems on grouping data (E.g. "Universidade nova de Lisboa" and "Universidade Nova de Lisboa" might be grouped separately).

3.5.4. Data Transformation

Data transformation aims to create new measures or guarantee existing measures have the right format and are grouped in the right way. More even, data can be pivoted or unpivoted, joined with different datasets, translated, and missing values can be replaces using a wide variety of data science algorithms.

In the case of this project, data was transformed and made ready for being loaded into the relational data warehouse and more over the data transformed will be discussed in the next chapter.

3.6. Software Development

All the knowledge gathered so far culminates on how to achieve the specified goal with a software solution designed to use data and Business Intelligence solutions to improve the decision-making process of future students.

Here the process which was used to create the system is described, as so is the process of developing the software. The data transformations will be described in detail as many of other aspects that impacted or guided the development process.

3.6.1. Project Management Methodology

As explained previously, this project is being developed using an Agile approach. Although it is not possible to simulate all the aspects of the methodology, an organization of the requirements and tasks is straightforward and there are plenty of tools designed to help on this task.

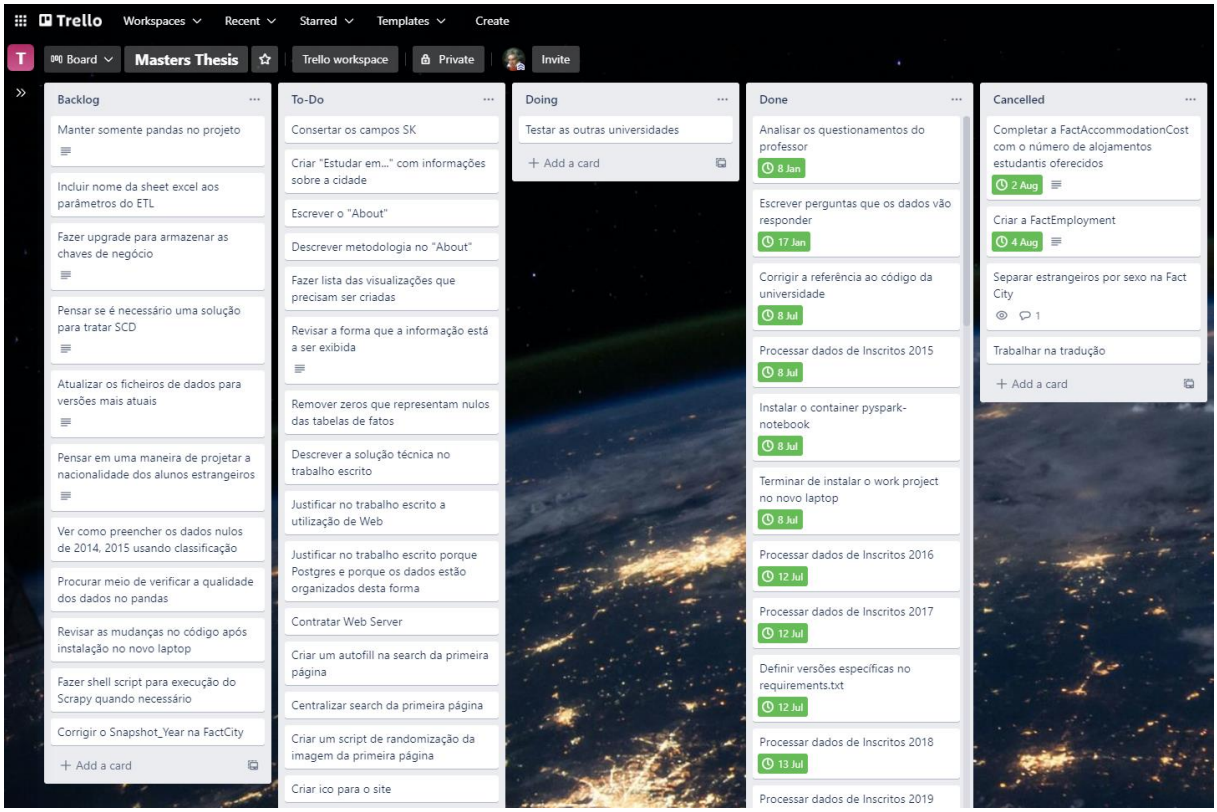


Figure 6 - View of the project organization tool Trello

To better organize the requirements and activity backlog, an account was created on Trello, where a Kanban method of organization can be assembled. This method allows to a division of tasks as “To-do”, “Doing”, and “Done” as shown in Figure 6. Trello also permits that on each activity one can attach documents, links, or notes. Also, you can specify a date in which you think you’re going to finish a certain activity. This comes handy if you want to control the delivery date of the requirements.

In this project, Trello helped on providing a better view of what still had to be done in order to deliver each part of the project. From a perspective of this author, this is a fast approach in comparison with Microsoft Project, which complexity would not add additional value to the work, besides, most of the companies are migrating to Agile approach in which Trello fits well.

3.6.2. Development Environment

A Development Environment needs to be stable, and it is preferable that it is well known by the development community given that it guarantees help resources will be plenty. As for "stability", most of the time it means that the version chosen should be kept on the whole development cycle. This is true for companies, and this should also be true for any project which has a tight schedule as one does not want to waste time fixing problems brought in new versions of the software. In the case of this project, this should apply to Visual Studio Code, and Docker, both described in Table 18. Although, it must be admitted here that this rule was not followed entirely as will be explained.

Tool	Reason/Characteristics
Visual Studio Code	A visual IDE with support for Python and integration with Git
Git	To keep the source code. It allows backups of the source code and versioning
Docker	Virtual Development Environment

Table 18 - Development Environment Software

Visual Studio Code, seen in Figure 7, became a great IDE in the last few years and as it couples well with Git and Docker, there was no reason to look any further.

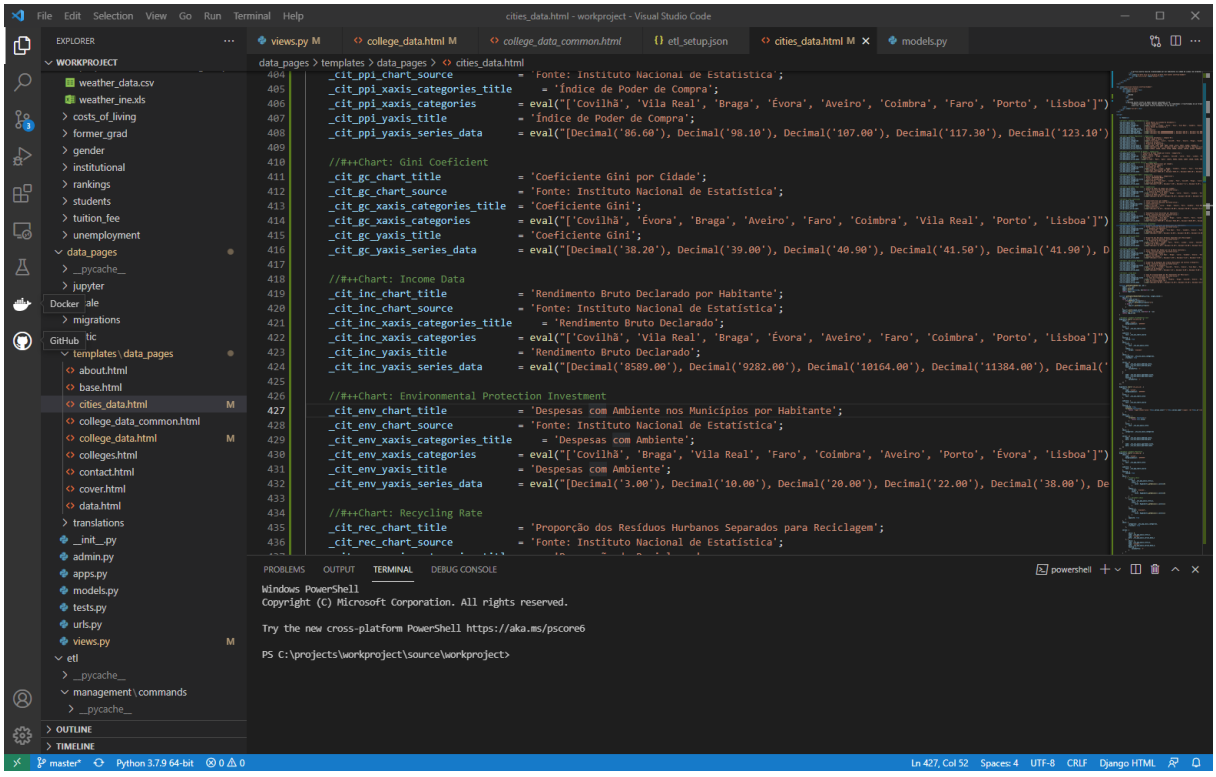


Figure 7 - Visual Studio Code Interface

All those who are fond of coding like its simplicity – even if it is full of advanced features – and research with developers has shown that it became one of the favourites IDEs in the market. One of the advanced features of Visual Studio Code, is the possibility of enhancing its capabilities by including extensions. This project made use of two extensions: the Git Hub extension, and the Docker extension.

Docker allows developers to create environments which are copies of production environments with little effort. There are thousands of "images" created by the community. One can download any of these images for free and from them create containers where you can do anything your software need to be done, without having to change the configuration of your development machine. For instance, for this project, two containers were created from two different images. You can see them on a print

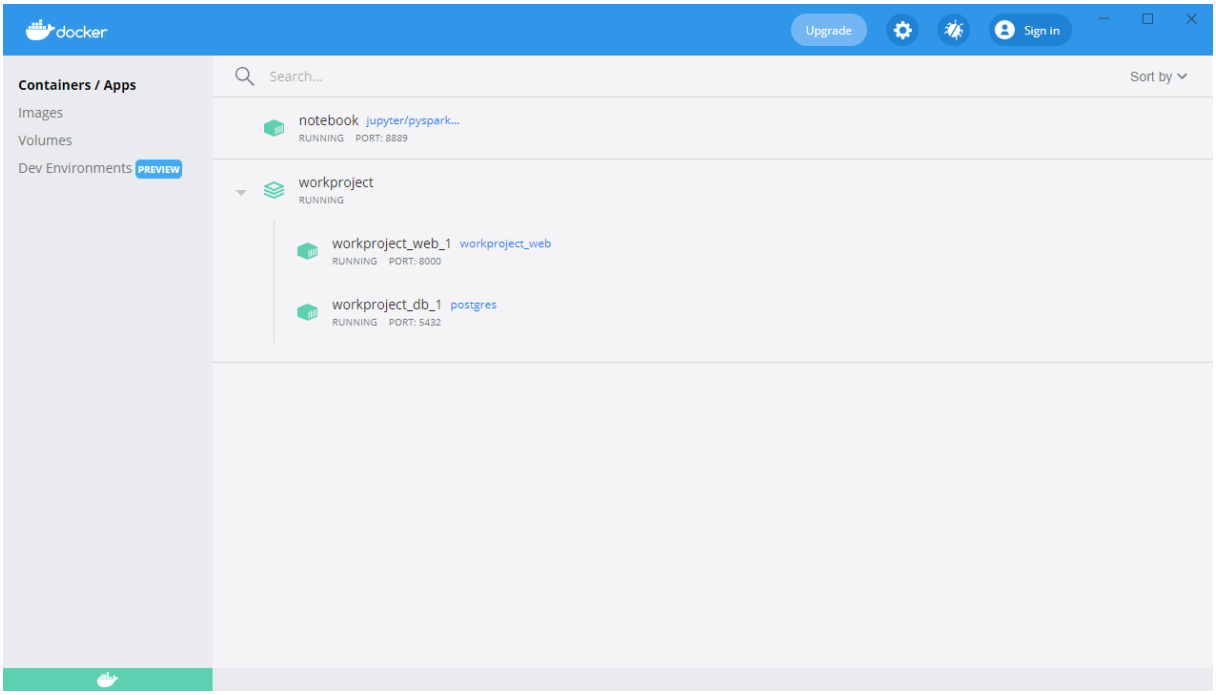


Figure 8 - Docker Interface and containers list

screen of Docker interface in Figure 8 and a description can also be found on Table 19. These two containers are part of a composition, as you can also see in Figure 8 they are in a single tree. This helps with organization and make the process of starting up them easier. Also, the two of them were created for different reasons as described in Table 19.

Container	Image	Description
workproject_web_1	python:3.9.6	Hosts the Apache webserver
workproject_db_1	postgres:latest	Hosts the PostgreSQL database
notebook	jupyter/pyspark-notebook:ubuntu-20.04	Notebooks for testing data transformations and python code

Table 19 - Docker Containers used during the project development

Figure 9 shows how the containers are organized in Docker. The two main containers, which will be deployed to production later in this work, are organized in a logical structure in Docker called "Compose". There it is possible to organize one or more containers which are part of the same solution.

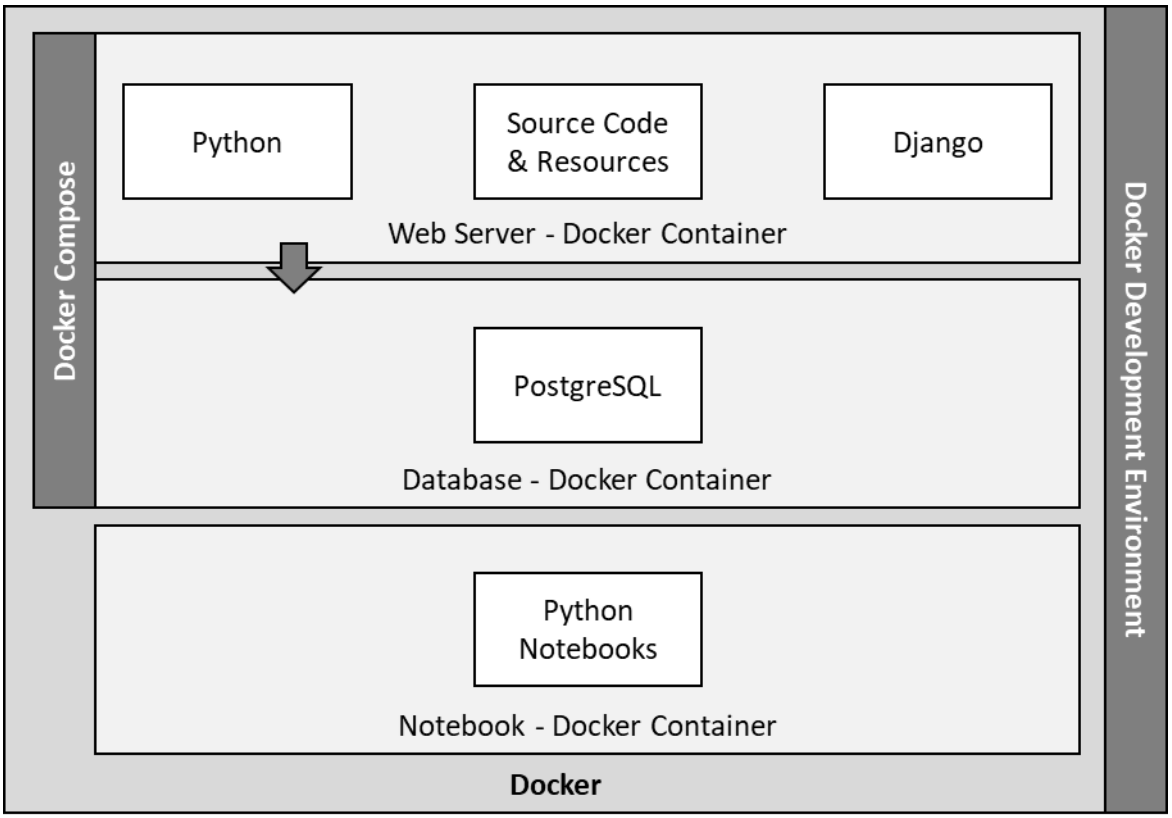


Figure 9 - Docker Development Environment used

3.6.3. Source Control

The source code being modified is shared with the container, most specifically the webserver container, using a "Volume", which is another handy feature of Docker. Volumes permit the developer to share folders and its contents seamlessly with the container. This allows that a change made on the source code on Visual Studio Code is automatically made available to the container, therefore ready to be executed by the webserver and automatically shown in the web browser after the code is saved. There is even the possibility of enable auto-save, so the webserver automatically restarts and make the change available even faster. Although, this feature is buggy as from time to time the server fails to restart. More about that on a future chapter.

As soon as a feature is ready and it is tested, we must guarantee the source code is saved and versioned. It is important that project code is not subject to failures in the machine where the work is being carried out, or failures on the network, or Docker containers. To reduce this risk, the use of Git is of extreme importance.

Git is a command line version control system and GitHub is an online service which provides a free Git Server where files can be stored and versioned. Once a set of code is ready, or even if it is not ready, but you are calling the day, for instance, you should "push" the code to the GitHub server.

Figure 10 shows the code as it is displayed in GitHub. There you can manage the repository security settings, branches, collaboration, and many other things. Also, the code can be public or private. This project now is a private project, although the desire is to make it public if it become possible to transform it into a framework for Open Data utilization.

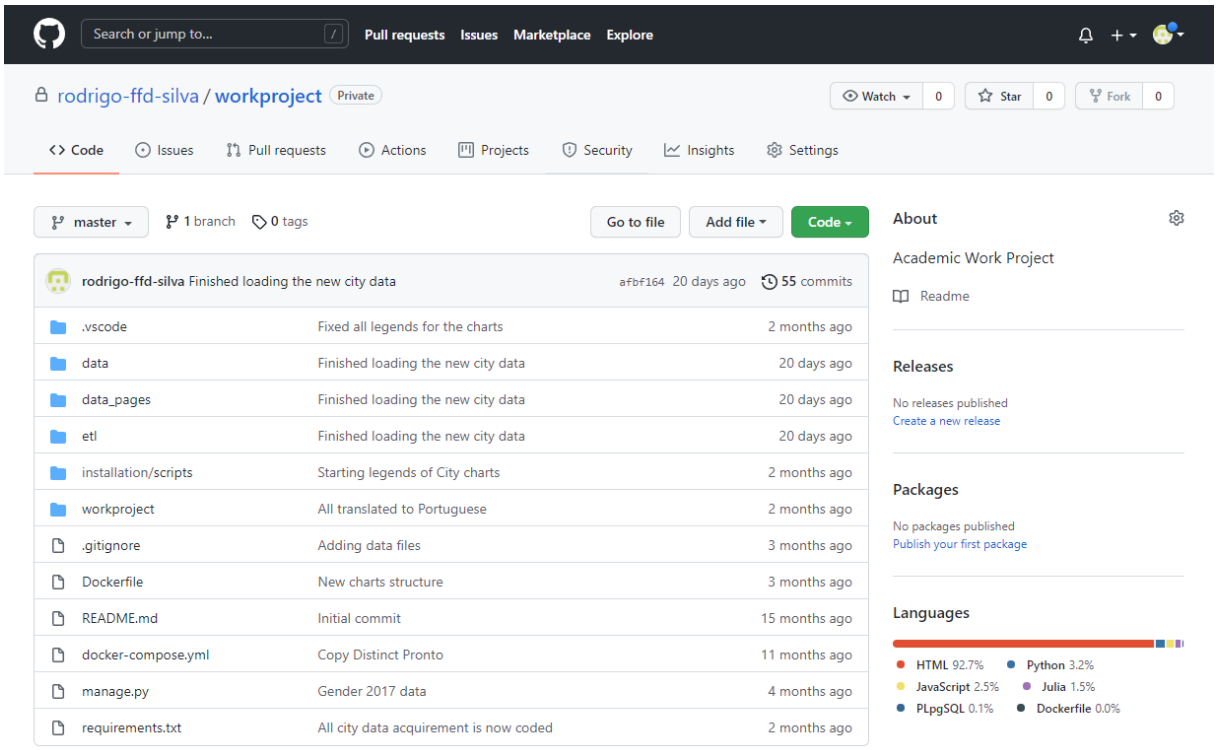


Figure 10 - GitHub Interface

Visual Studio Code makes it easy to work with Git and GitHub, as it offers a GitHub extension with all the commands needed.

This was a simple project with one developer, although Git is a complete solution also designed for development teams. It is fully prepared for parallelism which is a scary subject in companies. Although, for this project, the only concern was to publish code as soon as a functioning version was ready, and the day was finished.

3.6.4. Web Server

Once there was a working code, tests were needed to check if the whole integration was oiled. With the web server container up and running, looking at its logs it shows the address where it is running, which normally is something like “http://localhost:8000” as shown on Figure X. A local web server would, by default, be found on port “8080”, although the container choses port “8000” so it avoids conflicts with web servers outside the containers that might have been installed previously.

Of course, the home page shown in Figure 11 is from an advanced stage of the work, and the normal is to be welcomed with a default Django page when you first test the integration. Then, thinking about a scenario of production deployment, you would start the server after the code is deployed and the page below should be shown, but this will be discussed further in this work. Although, the first time the server was started a default Django welcome screen was shown, and it meant that the integration described in Figure 5 was working well and we are ready to go forward.

It is important to say that after installation of the containers, even before starting to code, when you access the home page at localhost, the server doing the work is a small webserver which is included with Django. Before the deployment to production, the Apache server should be installed, as this will

be the server used in the production environment. In fact, a pre-production environment could be created so the system could work in an environment more similar to that it will run in production.

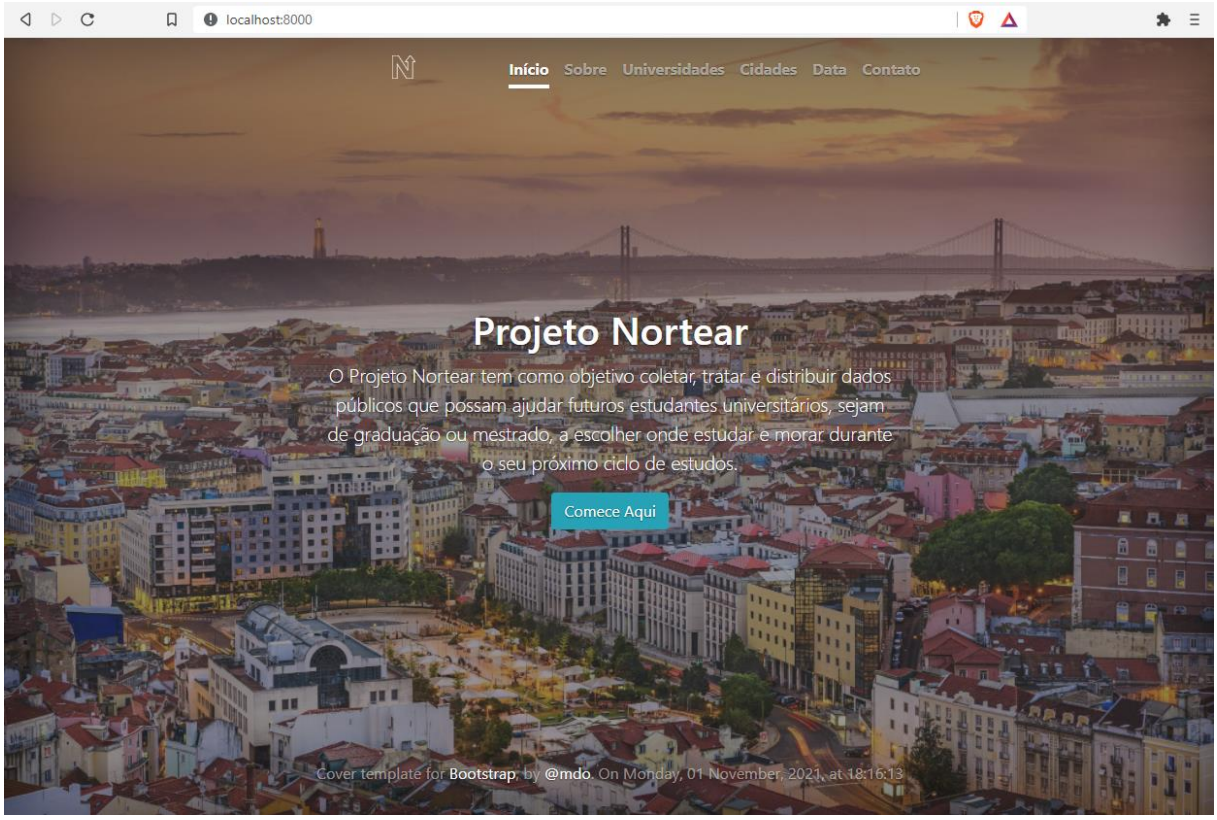


Figure 11 - The first page of the project running on Django web server

3.6.5. Ingestion Platform

All the data collected for this work comes from files. An ingestion solution fits well the scenario of this project given that when data is ever updated, it is possible to download the new file from the provider and, if no changes were made to the structure of the file (data columns, headers, and footers), we only need to upload the new file to the server and execute the ingestion pipeline to update the data used.

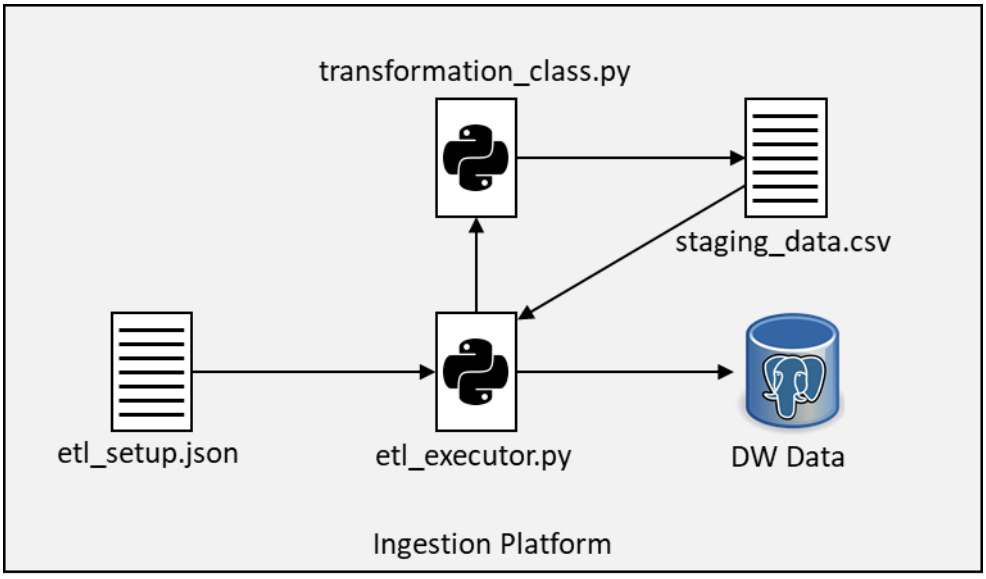


Figure 12 - The Ingestion Platform

The execution pipeline was built from scratch as a challenge for this work, although there are plenty of open-source solutions in the market which can also be used to design ingestion pipelines, such as the Pentaho Data Integrator. Although, as this project needed something simple, at first it was opted as a good solution to build something for it as a way of learning something new. Even though the task seemed simple – at first – there were moments it became so buggy that the challenge was almost left behind. But the task was manageable in the end, although a mind note stays that one should mind the weight it needs to carry.

The ingestion pipeline was written in Python, and it is composed basically of three parts as shown in Figure 12: the orchestrator, which contains the main script (etl_executor.py), and the configuration file (etl_setup.json); the staging generator, which contains the transformation class (which name is described in the configuration file and followed by the extension ".py"), and the staging data file; and, the data warehouse itself which holds the data used in the visualizations.

The project ended up with 20 pipelines, each loading one or more tables of the model. Every pipeline has its own configuration described in json within the file "etl_setup.json". As shown in an example in Figure 13, the configuration describes basically the data source, the transformation class, and the data destination. It also describes the columns to be used to lookup for the dimension keys when recording fact tables.

```
{
  "transformation_name": "Former Graduates Employers",
  "preprocessing_transformation_class": "FormerGraduatesEmployers",
  "data_source": "/code/data/former_grad/OndeMoramTrabalham.xlsx",
  "data_sink": "/code/data/former_grad/OndeTrabalham.csv",
  "extension": "*.csv",
  "location": "/code/data/former_grad/OndeMoramTrabalham.xlsx",
  "DBLookup": {
    "FactFormerGradEmployers": {
      "type": "copy_lookup",
      "origin_destination_pairs": [
        ["Universidade","fk_college_id"],
        ["Empresa","company_name"],
        ["Ano de Consulta","publication_year"],
        ["Origem dos Dados","source"],
        ["Ex-Alunos","former_students_count"]
      ],
      "destination_key": [
        "fk_college_id",
        "company_name",
        "publication_year",
        "source"
      ],
      "lookup_columns_n": {
        "Universidade": {
          "table":"DimCollege",
          "weak_match": true,
          "comparisson_column":"college_name_pt_pt",
          "column_to_obtain":"id_college"
        }
      }
    }
  }
},
```

Figure 13 – Sample of the pipeline configuration "Former Graduates Employers"

3.6.6. Data Model

Firstly, it was decided that a Kimble Star Schema was a good fit for the processed data. And Django would be able to deal well with it. The reason why this is a good fit is that we can easily separate the date into dimensions and facts and create a data warehouse that can both serve the website, which is

the main product of the present project, and provide future initiatives with a model that can be explored using other tools such as Power BI.

In order for any table to be created – and therefore for any pipeline to be run –, they need first to be specified on the model file, which is part of the configuration for a Django data model. Figure 14 shows an example of the model file and one of the fact tables. The example shows the configuration of a fact table. As shown, within the file a whole set of configurations might be set, such as foreign keys, primary keys, field sizes and precisions, and a whole set of other options are allowed by both Django and PostgreSQL.

```
193
194 class FactMastersRankings(models.Model):
195     fk_college = models.ForeignKey(DimCollege, on_delete=models.PROTECT)
196     fk_course = models.ForeignKey(DimCourse, default=-1, on_delete=models.PROTECT)
197     publication_year = models.IntegerField()
198     rank = models.IntegerField()
199     ranking_name = models.CharField(max_length=120)
200     class Meta:
201         db_table = "FactMastersRankings"
202         constraints = [
203             models.UniqueConstraint(fields=['fk_college', 'fk_course', 'publication_year', 'ranking_name'], name='unique_mastersranking')
204         ]
205
```

Figure 14 – Sample of the model FactMasterRankings

Table X shows a list of fact tables created on the model file. Each fact table has at least one dimension associated with it. Each one of these tables were defined first in the model file before they were created. Django has a set of commands for creating the tables and it stores the versions of every snapshot of the database structure, allowing us for rolling back any bad changes, which was helpful during the development of the project.

Fact Table	Description
FactAccomodationCost	Cost of student accommodation
FactCollege	Numbers over the colleges
FactFormerGradEmployers	Where former students work
FactFormerGradResidence	Where former students live
FactRankings	University Rankings
FactMastersRankings	Masters Rankings
FactTuitionFee	Cost of tuition fee
FactCity	Numbers over the cities
FactWeather	Weather data over the cities
FactStudents	Data over the students and colleges
FactCityDistances	Distances between cities
FactUnemployment	Unemployment Numbers
FactTourismAttraction	Distances between cities and TAs

Table 20 - Fact Tables Descriptions

Both fact tables and dimension tables, both listed in tables 20 and 21, respectively, have different kinds of data types. These data types were carefully thought so they would occupy less space in the storage

and would not weight in terms of performance. And in this sense one important knowledge for anyone who deals with databases is to know how much storage each data type occupies. This might not be such a big deal for this small project, but it future profs it for any further development and this is certainly a best practice to keep throughout any future projects as a developer.

Dimensions	Description
DimChosenColleges	List of Chosen Colleges
DimCollege	List of Colleges
DimCourse	List of Courses
DimCity	List of Cities
DimCourse	List of Courses
DimArea	List of areas of study
DimTourismAttraction	List of tourism attractions
DimLocation	List of country locations

Table 21 - Fact Tables Descriptions

Once the model is defined in the model file, Django has two commands to automatically upload the model to the database: "./manage.py makemigrations", which updates the DDL script files; and, "./manage.py migrate", which effectively updates the data tables in the database. This way Django takes care of any DDL database command. During the project, Django showed to be of great help making the process of updating the model easy.

3.6.6.1. Main Takeaways of the Data Model building

Table 22 lists a series of decisions to be made when creating the model. These decisions need to consider how the data is queried, besides future maintenance of the model. As it was decided to use Django, some constraints had to be imposed, for instance, the use of foreign keys, which is not something required when speaking of a Kimbal model, although it comes handy in this case.

Decision	Motivation
Use of a Kimball Star Schema	A known model. Efficient and easy to maintain.
Manage the model using Django	A framework for dealing with data can make the work easier
Create Foreign Keys	The Django queries make good use of foreign keys
Be careful with data types	Well defined data types can save storage and query time
Define primary keys	In dimensions so they can be linked to fact tables
Build tables with Django models	Django can maintain the model and make changes easy

Table 22 - Takeaways of Model Building

Every choice might bring with itself hurdles that need to be overcome. Table 23 shows some of the hurdles faced when dealing with the data model, especially when obtaining data with Django.

Challenge	How to overcome
Django do not allow composite keys	Create unique indexes to guarantee unicity
Define primary keys	Universities, for instance, had names written in English and Portuguese. Sometimes with all capital letters, other times with camel case, some transformations had to be made to guarantee that a single university was linked to a single key
Django do not support Stored Procedures	Using native queries in Django

Table 23 - Challenges faced during Model Building

3.6.7. Visualizations

Up to this point all that was created had the final aim of showing data in the form of visuals so there was a way of generating knowledge and from this knowledge action towards an objective, the student objective of choosing the best place for him or her to live and study. Although, data alone cannot reach this goal. Data needs a form of expression, and this form of expression are the visuals. These visuals are of dare importance if we want to communicate well the data collected to the public it is intended to.

3.6.7.1. Technology

Firstly, there are several ways of transforming data into visuals. Nowadays, Power BI and Tableau are well known self-service BI tools, and this author is himself a Power BI Certified Professional. Using Power BI as an example, you can build visualizations and publish them on Power BI Service so you can make them available to other people – though some resources are limited for free license users – and even can embed a visual in a webpage using an available resource. Figure 15 show an example of a Power BI visual created on Power BI Desktop, uploaded on Power BI Service, and embedded in a web page.

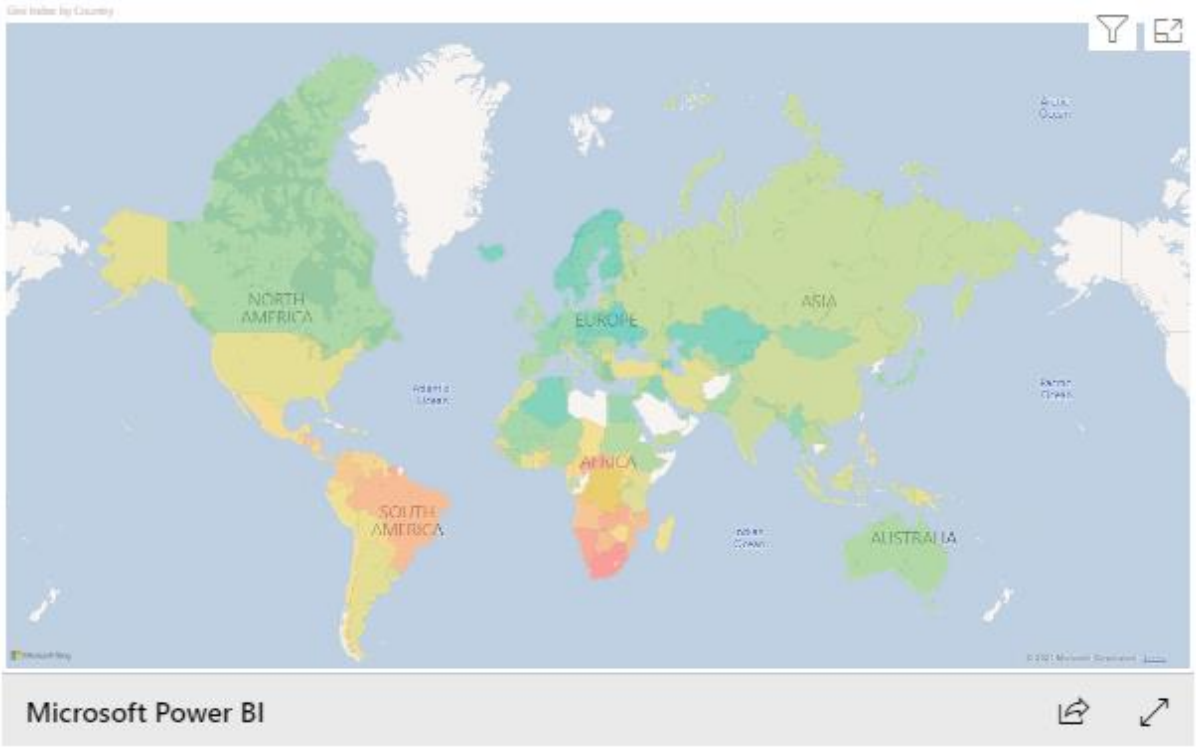


Figure 15 – Exemplo de Visualização com o Power BI embedded

Although this would work well, this author would like to explore other options that would not depend entirely on a 3rd party sets of free features available, as Microsoft can at any time limit the features that are available in the free license of Power BI Service. But, as important as this argument is, the most important drive for choosing a completely different approach is learning. There is the opportunity of learning something new besides Power BI and at the same time measure if there are any advantages on having more control over the solution. It can be verified whether this approach is simply more complex or if there would be advantages on using a visualization library specially created for web.

Following the explained rationale, it was decided to prospect some web visuals solutions. Table 24 shows a list of tools that were prospected for this project. The main requirements for choosing a solution would be: it must be free for non-commercial use; as Power BI, it must be attractive and user friendly; and, it must be of fairly easy implementation – a challenge had been agreed, although not without completely forgetting the limitations of the developer.

The attractiveness factor had a great weight here as some of the libraries are created for scientific research and most of the research do not need to concern themselves with beautiful visuals. Although, here it is important to keep in mind that as it is believed this project can help people on reaching their academic goals, it is important to eliminate bias people have on using “ugly” solutions. This project is going to be made available for a broad public, so beauty and functionality must work together.

Library	Bottom-line
ApexCharts.js	Beautiful visuals and easy to use
Chartist	Simple, not interactive
Google Charts	Crude visuals, scientific

Charts.js	Crude visuals, scientific
D3.js	Crude visuals, but complete
Highcharts	Beautiful visuals and easy to use

Table 24 - Web Visualization Libraries

For all those reasons, the choice was between ApexCharts and Highcharts. In the end HighCharts was chosen as the library to be used. It allows for a high variety of visuals, is open source, has beautiful visuals, it is free for academic purposes, and has a vast documentation, and the same cannot be said of ApexCharts.

3.6.7.2. From data to visuals

The spirit of the technical solutions of this project is to aim at building a software as close to a real scenario as possible. Having this in mind it was decided to feed the visuals with live data straight from the database using Django queries. Highchats is fed with data using arrays and json objects depending on the intended visual. So, queries had to be built in Django "views" class, shown here in Figure 16.

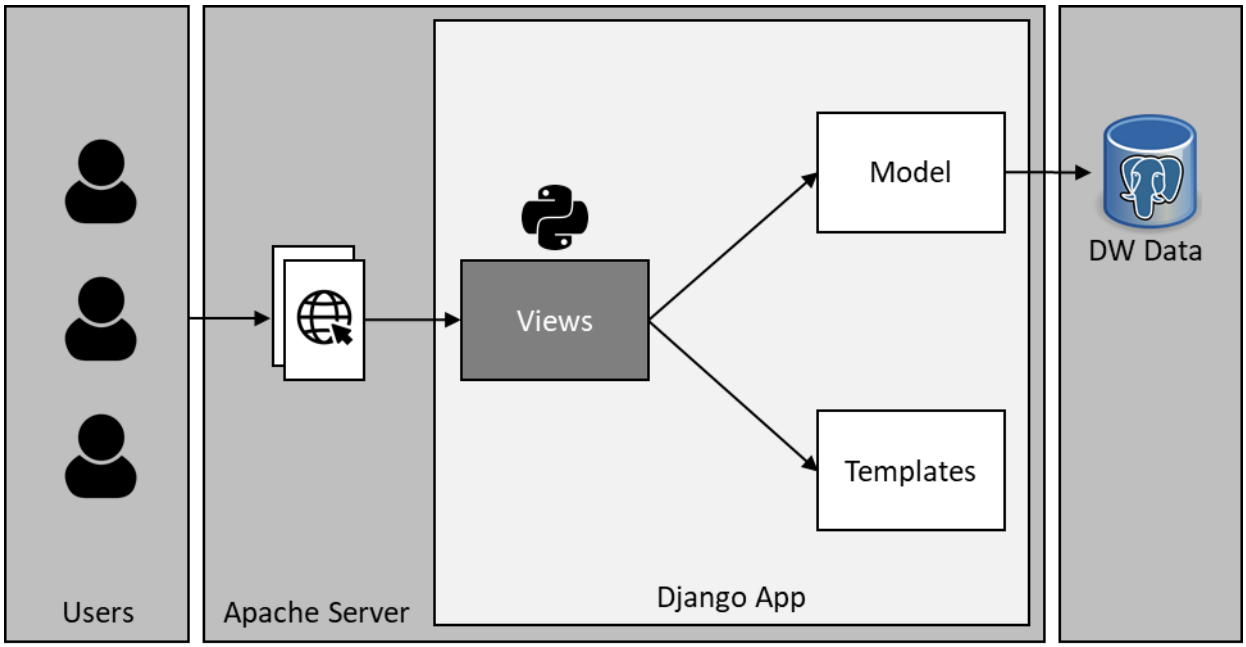


Figure 16 – How data feed the visuals

In Django architecture, the "Views" layer is responsible for obtaining data and transferring the data to templates. These templates are where Highcharts charts are coded in JavaScript and in these scripts the data will be added dynamically using Django markers brought from the View layer.

Although, before any view can feed the charts in the templates with data, queries must be coded to obtain the data that is stored in PostgreSQL. The file "views.py", which represents the View layer in Django, holds all the queries. Figure 17 has a extract of a piece of code responsible for obtaining the data that feeds the chart "Unemployment Among Early Graduates – Area Higher".

```
247
248     #++Chart: Unemployment among early graduates - Area Higher
249     unp_area_high_data = VWUnemploymentByArea.objects \
250         .values('area','unemployment_ratio','top12_area_ratio') \
251         .filter(college__startswith=college_name) \
252         .filter(unemployment_ratio__gt=0) \
253         .order_by('-unemployment_ratio')[:10]
254
255     unp_area_high_chart_title = _('Desemprego entre Recém-Graduados - Áreas')
256     unp_area_high_chart_source = _('Source: Direção-Geral de Estatísticas da Educação e Ciência')
257     unp_area_high_yaxis_title = college_name
258     unp_area_high_xaxis_categories = [data["area"] for data in unp_area_high_data]
259     unp_area_high_yaxis_data = [data["unemployment_ratio"] for data in unp_area_high_data]
260     unp_area_high_last_data = len(unp_area_high_yaxis_data) -1
261     unp_area_high_yaxis_top12_data = [data["top12_area_ratio"] for data in unp_area_high_data]
262     unp_area_high_text_summary = _('O gráfico mostra as áreas de estudo com o maior índice de desemprego na %(college_name)s comparado com o índice médio das 12 melhores universidades Portuguesas.' \
263                                    ' A área de %(first_area)s tem o maior índice, com %(first_data)s%% dos alunos inscritos nos centros de emprego, enquanto a área de %(last_area)s se encontra com o'
264                                    ' menor índice dentre as maiores, com %(last_data)s%% dos alunos à procura de emprego.') % \
265                                        {'college_name': college_name, 'first_area': title_it(unp_area_high_xaxis_categories[0]), 'first_data': unp_area_high_yaxis_data[0],
266                                         'last_area': title_it(unp_area_high_xaxis_categories[unp_area_high_last_data]), 'last_data': unp_area_high_yaxis_data[unp_area_high_last_data] }
267
```

Figure 17 – Extract of code showing how data is obtained with Django

What this code does is to query a view which data was especially prepared for this chart. It also transforms the data brought by Django into a format that can be interpreted by Highcharts.

3.6.7.3. Choices

Here it will be shown some examples of charts and how they were tuned up to better express the data they are built upon.

There are 33 visuals created for each of the 12 universities that are part of this project. Each visual had to be designed having in mind the data which is available in a way that it can be as informative as possible. Also, and more importantly, they should communicate the information related to the main concerns students have, which were described in the Literature Review.

As an example, lets analyse the chart “Areas – Lower” shown below in Figure 18. This chart shows the data over the courses with the lower level of unemployment for NOVA University of Lisbon. One important caveat was not only to show the data for the university, but to stabilish a comparison so that that number can be put on perspective, and this really inform the viewer something new. In this case it shows us that the course on “Health and Social Protection” have a higher rate of former student working on the area than on the average of the 12 best ranked Portuguese Universities.

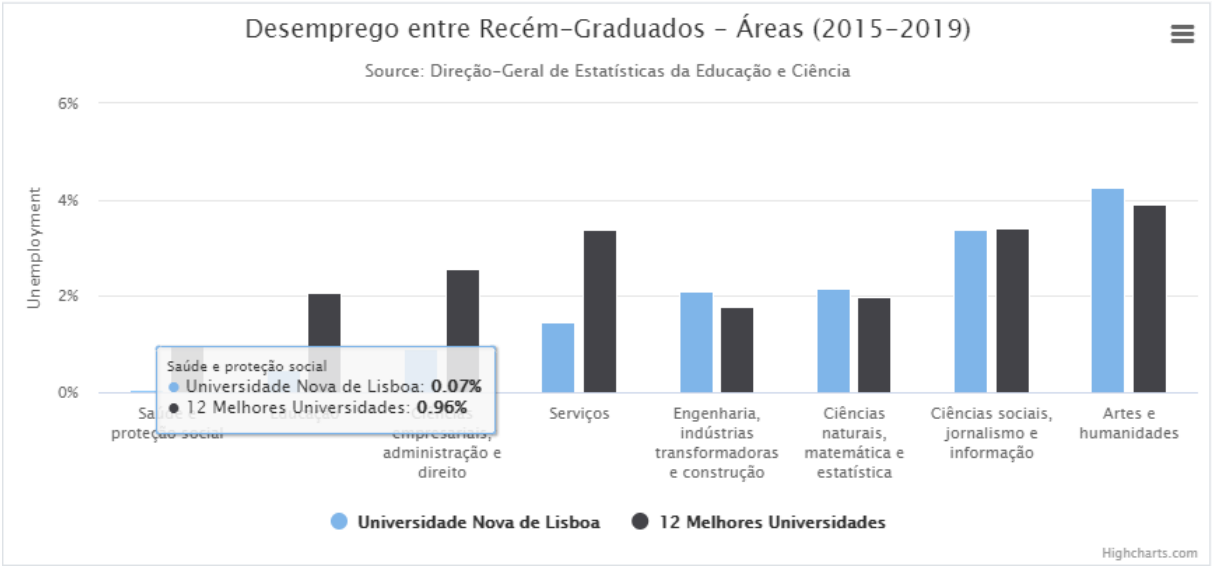


Figure 18 – Bar chart showing the general visual of the charts and tooltips

Highcharts have many layouts for charts, although the choice had to be for a layout which was simpler and yet attractive and interactive. In terms of interactivity, every chart has a tooltip that shows the

data in detail, as shown in Figure 18. Also, the data can be exported using the menu on the top right corner of the visual, a simple functionality that adds more to the user experience.

For group of charts Academic Environment, one of the hardest technical solutions was to make one of the data spots differ apart of the others using a colour, as shown in Figure 8. It was important on every chart that showed data on multiple universities, that the university whose data is being explored now was told apart of the others. The reason why this was hard it was because the data point received by Highcharts had to have a colour tag. As the data points were being brought from Django, either the data point should be stuffed with the colour in the database and converted to a string, or a python code should stuff the colour on the json format data point. Both solutions were developed for different visuals, although they both worked well as seen on Figure 19.

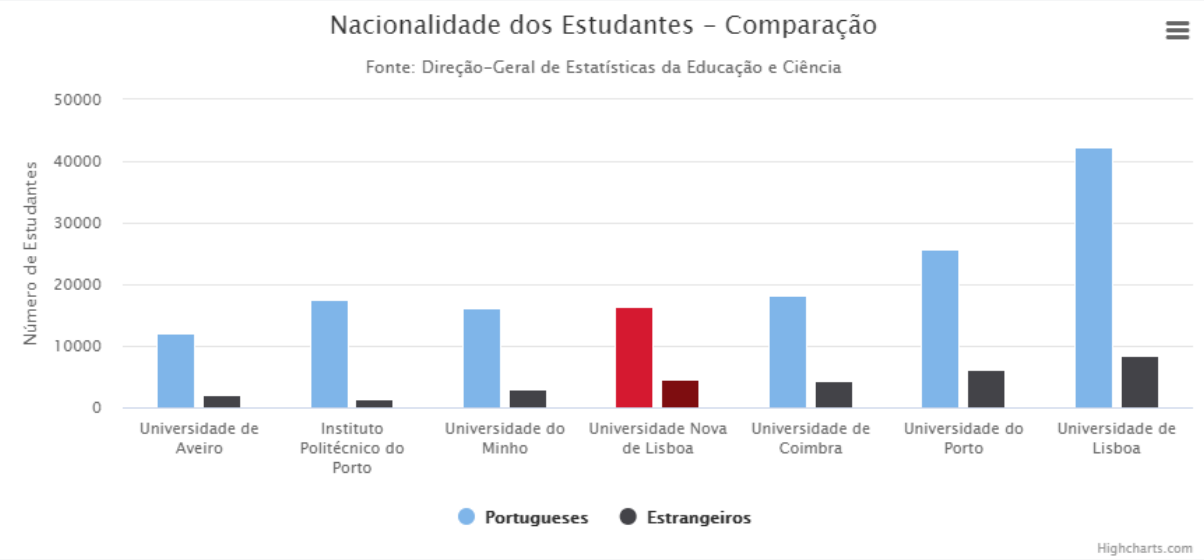


Figure 19 – Bar chart showing how a data point can be highlight

Another example come from the need of expressing ranking data. In Figure 20 the masters courses of NOVA University Lisbon ranked on the QS Business Masters Rankings are shown along the years. Here there was the problem of returning to Highcharts exactly the format it needed to display the chart as intended, which means to compare one or more master's courses throughout the period available. As Hiighcharts neede a JSON object with specific format, once more the problem had to be solved using Python. The data returned using Django had to be run in a python function which converted the data into the Highcharts format. In the end all worked well, although with help of coding.

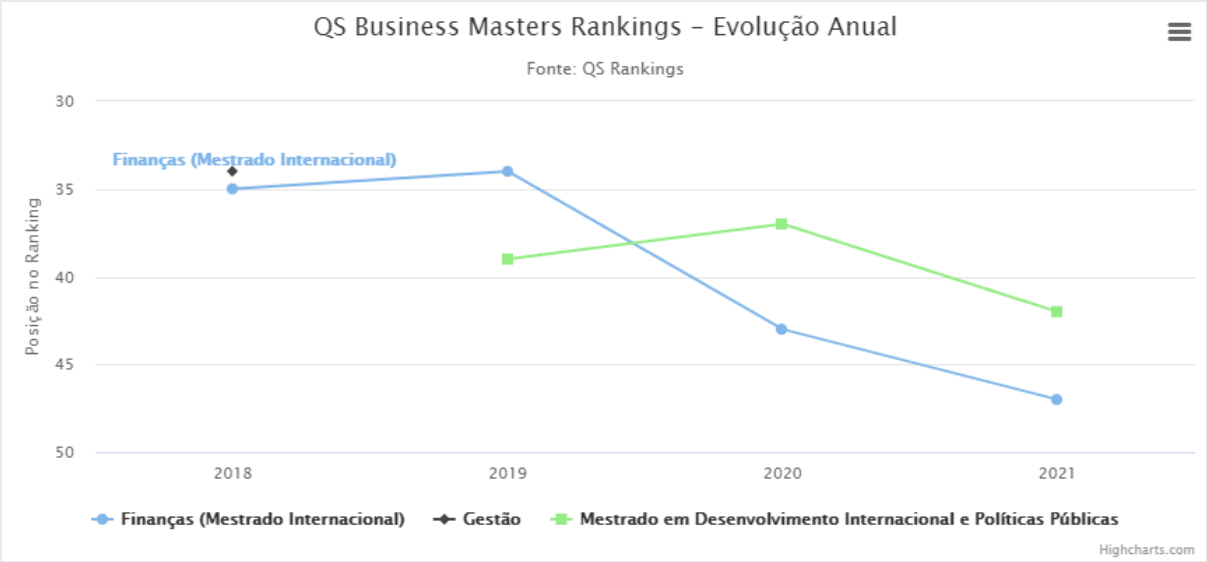


Figure 20 – Line chart showing the challenge of showing ranked data

Another decision was in terms of formatting. The better rankings were positioned in the top of the chart. A tooltip was also created so the numbers can be checked on each time point. This visual gives us an idea of how well each one of the courses developed throughout the time in the ranking. For instance, it is clear that the Master's in Finance is decreasing steadily in the ranking since 2019.

3.6.8. Deployment in Production

The production environment is where the website and visuals will be available to the broad public on the internet. As all websites, there was the need for choosing a service provider to host the website. Although, in this case the project had already been developed and had a particular architecture. Any provider would need to attend the previously defined architecture. Therefore, it was decided that a Virtual Private Server (VPS) was the best choice, as almost all aspects of the server configuration are under the responsibility of the developer.

3.6.8.1. Choosing a VPS Provider

The configuration of the VPS had to attend to minimal specs. These minimal specs are determined by the minimal specs of the technologies which are part of the architecture of the project.

Table 25 shows the components of the architecture which contain minimal specs described in the website o their publishers. As it is noticeable, Apache and Python do not present minimal specs. The reason found during research is that it depends mostly on the number of visitors to the website and how heavy the web resources are. For now, given this project is a really small one, these values wont be considered and only Ubuntu and PostgreSQL will be considered.

Component	Memory	Processor Speed	Cores	Disk Space
Apache	-	-	-	-
PostgreSQL	2GB	1GHz	1	512MB
Python	-	-	-	-
Ubuntu 20.04 LTS	1GB	1GHz	1	3GB

Table 25 - Minimal Specs for each component of the project's architecture

Having in mind the minimal specs to be aimed at, research was conducted to understand the VPS market for small projects. There are a good number of companies offering VPS service, although the different jargons and the stress on some technologies often make the choice confuse. For instance, from the chosen four companies, two do not let it clear if they offer a private IP address together with the package, what is essential for a VPS web server. As visible in Table 26, one of the providers do not clarify where the servers are located, an important characteristic which affects performance if the service is being offered in a local market. For a customer not used to buy these services, the lack of information makes the difference, and a provider called Hostinger, which clearly informed all the specs which were necessary for the project ended up being chosen.

Provider	Memory	CPU vCores	IP	Disk Space	Location	Cost
Hostinger	2GB	2	Yes	40GB SSD	Netherlands	€20/MTh
OVHCloud	4GB	2	?	80GB SSD	France	€12/MTh
Amen	4GB	2	?	40GB SSD	?	€13/MTh
WebTuga	4GB	2	Yes	100GB SSD	Portugal	€25/MTh

Table 26 - VPS Service providers

3.6.8.2. Registering a Domain

A name had to be chosen for the project, and from some choices that came up, the name “Projecto Nortear” seemed to sound well in Portuguese. As the VPS provider also offer a domain registration service and the domain “projectonortear.pt” was available for €1 a year, so it was also registered.

3.6.8.3. Configuring the Virtual Private Server

The project was developed, as described previously using a Docker Container, this Docker container had a setup, and the aim was to create a production server that closely resembled the development environment. This is possible to be reached if we copy all the versions of the tools and even the version of the SO and install the same versions on the VPS. The importance of this step is to guarantee that all the code developed is fully compatible with the production server and we are not going to confront unknown problems.

Table 27 shows a comparison between the versions of software on the development and production environments (VPS). Some discrepancies are noticeable and are marked in the table, and these differences must be addressed. Although, it is important to say that it was only considered really a change when the MAJOR version was different. A change in the MAJOR version means a change in the first number of the version and normally what it means is that there is significant difference between this and the previous MAJOR version.

Ubuntu was chosen for production without any fear because it is built under the Debian architecture. Something that works on Debian might well work on Ubuntu, although in the case of this project it needed to be proven during testing.

Resource	Dev Version	VPS Version
OS	Debian 10	Ubuntu 20.04 LTS
Apache Server	-	2.4.41
Python	3.9.6	3.8.10
PostgreSQL	13.3	12.9
Django	3.2.6	3.2.9
psycopg2-binary	2.9.1	2.9.1
Pyarrow	4.0.1	6.0.0
Pandas	1.3.0	1.3.4
Xlrd	1.2.0	2.0.1
Openpyxl	3.0.7	3.0.9
django-pivot	1.8.1	1.8.1

Table 27 - VPS Service providers

As shown in Table 27, PostgreSQL also shows a change in the MAJOR version. What happens is that the version which is supported by Canonical's Long Term Support for Ubuntu 20.04 LTS is version 12.9. The version of PostgreSQL could be upgraded to 13.3 or even 14, although this is not a good practice for production work, as we might lose the support provided by Canonical for this version of the database running along with Ubuntu 20.04 LTS. This is mostly a sensitive issue when speaking of security patches.

The Apache Server was not used during development. Although, after having configured Apache 2 for the first time after the website was deployed on the VPS server, it would be advisable to prepare a server in development environment to get used to the challenges of its configuration. The reason why Apache was not used during development is that Django offers a built-in small server, and that was enough for development, but Django's team stresses that it is neither safe nor efficient to run the built-in server in production.

3.6.8.4. Deploying the website

There were three ways to deploy the website: create an FTP server and copy the files; use git to *pull* the files to the VPS server; or the more professional, create a DEVOPS process to take care of the task. Given how hard it would be for this student to set up a DEVOPS process, it was decided to use the second idea, although not without a hint of the third idea. It means that a shell script was created to pull any code from the master brunch in GitHub and scan the files replacing the "dev paths" by the "prod paths" and replacing the Django configuration file. It worked well as an ad-hoc solution for the problem faced.

Companies have a robust process which involves approval hierarchies and automated scripts to transfer the files and change the needed configuration particularities, in the case of this project, although this would be interesting as a learning experience, this would add too much development weight to the solution.

3.6.8.5. Setting up the DNS servers

DNS servers "tell the Internet" to which IP – or server – a internet address should point to. At this point was necessary to access the DSN configuration available in Hostinger's configuration portal. After the configuration is done it can take up to 48 hours until the information gets fully propagated to the whole internet so people around the world are all able to access the website. But from this point, if all the prior configurations were done correctly, you might be able to see the Apache welcome screen when accessing projectonortear.pt.

3.6.8.6. Configuring Apache and Django

Now that the source code is in the machine, Apache server is running, and the DNS is configured and ready, it is time to set up the server so it starts to serve the webpage created. Here things get tricky, and it is not easy to describe the configuration without a fair amount of code. So, the config files will be left in the attachments, and here the configuration files will be listed in Table 28 in the order they were configured to have a magnitude of the work done.

Configuration File	Context	Reason
/etc/apache2/sites-available/workprojpp.conf	Apache Server	Create a virtual host for the website, define static files path, enable WSGI to work with Django, implement Django as a daemon (better performance), and set language and locale as pt_PT
/etc/apache2/sites-available/000-default.conf	Apache Server	Point all requisitions to port 80 to port 443
/srv/sites/projectonortear.pt /workproject/settings.py	Django	Database connection configuration, and Hosts allowed
/srv/sites/projectonortear.pt /etl/etl_setup.json	Project	Fix the paths to point to the paths of the production environment
/etc/hosts	Ubuntu	Setup the server to point to the domain name

Table 28 - Configuration files

After all configuration was done, it was needed to enable the website configured in the virtual host, enable the WSGI module for Django, enable the rewrite module, and restart Apache Server. More about these commands will be added on the attachments section.

3.6.8.7. Executing the Project installation files and commands

After the project files are deployed and Apache server is configured to access and execute them, before we can open the website on the browser, there is still a list of installation procedures to be followed which are described in Table 29.

Procedure	Context	Motivation
Create a user on PostgreSQL	PostgreSQL	A user is needed to both the SQL installation scripts, and to add this user to Django the configuration file
Create the data model	Django / PostgreSQL	A set of Django commands will create the data model on PostgreSQL from the project configuration file models.py
Create the views and procedures	PostgreSQL	Some visualizations require views and procedures which are included in the “installation/scripts” directory
Process the pipelines	Django	The pipelines are contained inside the project configuration file etl_config.json. These pipelines will fill the tables with data so the visualizations can work
Execute needed inserts and updates	PostgreSQL	Some hardcoded inserts and updates are needed and they are contained inside the file inserts.sql

Table 29 - Project Installation Procedures

After these procedures were all done, the website and visualizations were available. Of course, this was not as smooth as described here and a good amount of try and error was needed to come to a comprehensive list of the procedures which are described above.

Although the website was made available, and the visualizations were validated for all colleges. It is still needed to add an SSL certificate, and this was the next step.

3.6.8.8. Setting up the SSL Certificate

An SSL certificate can encrypt any info exchanged with the website users, besides avoiding attackers from creating a copy from the website, and stablish trust with end users when the browser does not show a message saying “this connection is not trustful” or something similar (Labs, 2021).

The SSL certificate used in this website is a free 90-day certificate that needs to be renewed every 90 days. This is a basic certificate aimed to protect the main domain (projectonortear.pt) and the www domain (www.projectonortear.pt). If it was decided to create a mail server, for instance, with a domain name such as mail.projectonortear.pt, this certificate would not encrypt the domain. A different kind of certificate would be needed in this case, and these are called wildcard certificates (Cloudflare, 2021a).

The configuration is made on Apache server after the certificate is generated on the certification authority’s website. The certification authority chosen was ZeroSSL as they can provide the free 90-day certificates. After applying for a certificate, and validate you are the owner of a domain, you need to publish the certificates to the server and point Apache configuration to the path where the certificates are located. More on the configuration file in the attachments.

3.6.8.9. Setting up the firewall

The configuration of the firewall for a VPS server is an extremely sensitive matter. Although this is an academic work, this is a subject it was decided to not do experiments, even though this served as a learning experience anyway. Table 30 show a list of ports that must be left open as they are necessary for either access the website or access the administrative interface.

Port/Config	Utility	Use
22	SSH	Access the server via terminal; Server administration.
443	HTTPS	Secured website access.
80	HTTP	Unsecured website access (redirected to HTTPS automatically).
STABLISHED, RELATED	Any TCP	Allow exchange of internet packages from connections started by the server itself. Needed for Git, for example.
Echo Reply, Echo Request	Ping	Allow ping requests.
IO	Localhost	Allow communication between resources in the server.
LOG	Dropped	Will log all the dropped packages
DROP	All	Defines that all the other packages will be dropped.

Table 30 - Ports and Resources for Firewall Configuration

The last rule is extremely important, and it must be put in the end as the last rule. It means that all rules above it are allowed and all requests to any other ports should be dropped.

The rules for ports 80 and 443 also include a configuration for prevent less sophisticated and mostly common distributed denial-of-service (DDOS) attacks. DDOS attacks happen when attackers try to flood a server with a number of requests it cannot handle forcing the service to go down (Cloudflare, 2021b).

4. RESULTS AND DISCUSSION

Here it is going to be discussed a little over what was achieved in each front of the work. Although this was planned and executed as an academic Work Project, much here resembles what is made in companies in a sense of delivering a work which is whole and functional but not complete as its completeness is guided by ever changing factors. And each one of the pieces that compose this work can be upgraded and improved and all was designed so it could be like this.

4.1. THE DATA

The beating heart of this Business Intelligence initiative is the data available. This is also one of the constraints of this work. Many datasets were obtained from different sources to compose the Data Warehouse created in the end. Although the need for Open Data can be classified as a constraint, whether speaking of the city's economic and environmental situation or the data available over the institutions and courses, there were not much that was not found when dealing with the students' needs and anxieties when looking for a course and a new place to live. In fact, the data found was even larger than this work could deal with. There are still more data to be explored and surely more visualizations that can be built, what is expected to be done in the future.

4.2. THE ARCHITECTURE

The collection of technologies assembled for this work seems to work well together. What was done here is far from being something new, of course, in terms of web server architecture, but there were a series of steps and tunings and configurations that needed to be done so the system could work well, and it should be working well for a long time as it is prepared to be.

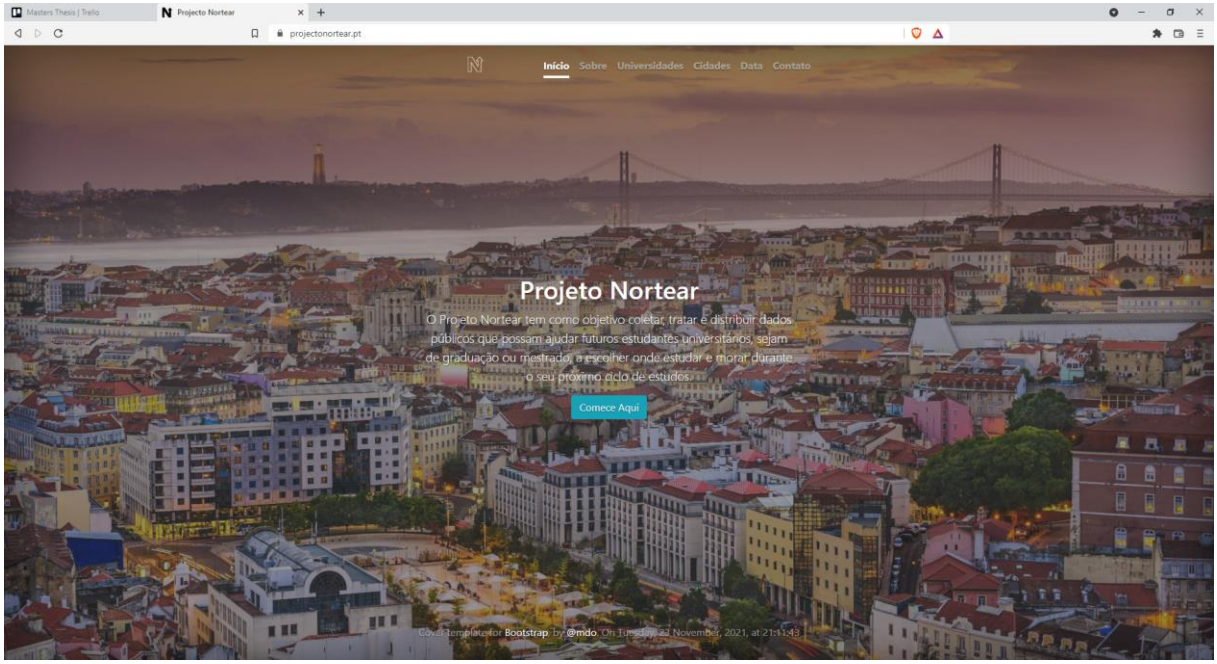


Figure 21 – The homepage of Projecto Nortear

4.3. The website

Having in mind the objective of making Open Data available in the format of useful information, the objective was accomplished. *Projecto Nortear* have a professional enough look and feel and a modern template library which allows it to be perfected and expanded gradually. The templates are set up to the simplest of choices available. This decision was made to improve speed of development and focus on what mattered the most, although they can be easily improved for the sake of user experience.

The images chosen were all available in public domain, and the proper credit is given in each image. Some colleges do not have images in public domain, so it was decided to show an image of the city which host it.

4.4. The Visuals

As said before, the datasets obtained were of great value and they allowed to building visuals that can transmit an idea clearly, helping on the decision-making process. Although the data was limited by what was open, there is the feel that with time more can be extracted from the data available and this work can become something bigger. The visuals, although addressing the main ideas that they should address, can still be better explored. There are dozens of techniques to explore the data available and more can still be done with the chosen charts library.

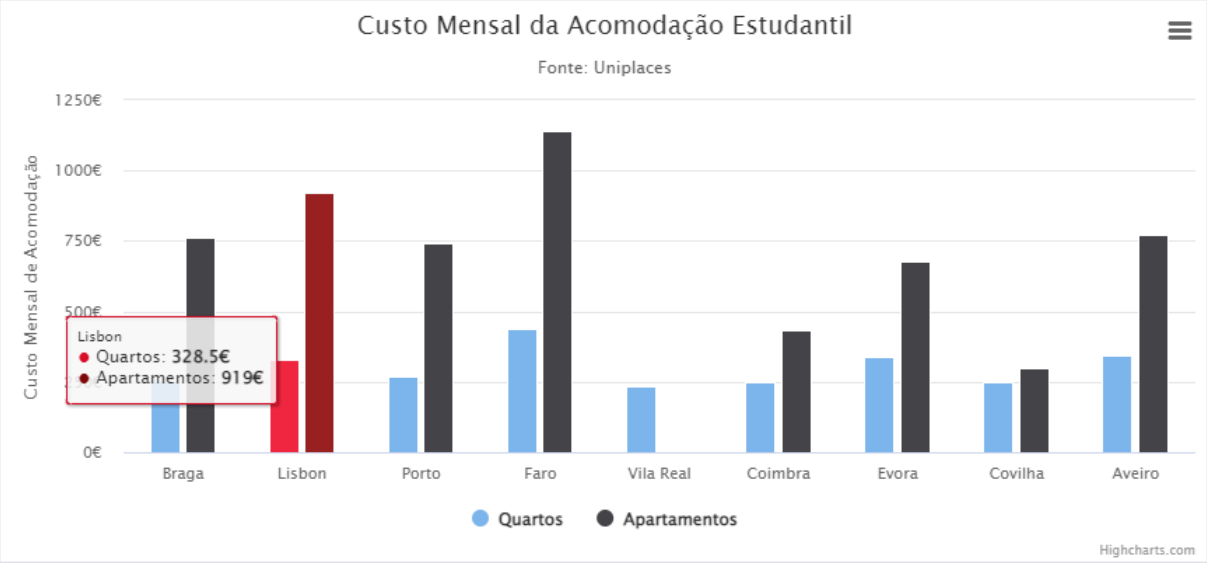


Figure 22 – Example of bar chart visual

4.5. The Back-end

The back-end is by definition something hidden, but that makes possible that the project is now able to show data in form of visuals that comes straight from the Data Warehouse. The data injection pipeline allows for the data to be updated as new datasets are made available by the responsible institutions. The performance is until now surprisingly good on the production environment. Although this should be put on proof as users start to navigate the website.

Speaking about coding, there are still changes that need to be done. There is no pretension to say the back-end is perfectly coded or the injection pipeline has a great performance. Things still need to be

done and problems still need to be addressed, but for now each piece of software is doing the work it should do with the expected output.

5. CONCLUSIONS

It is possible to universalize Open Data using open-source software solutions and, of course, knowledge. Given the problem at hand, more relevant data was collected from the "data lake" of Open Data than was previously anticipated by this author. Data over higher degree public institutions and socio-economic statistics are widely available, and they can not only address anxieties higher degree students have, but also be used to address the needs for information of many other groups of society.

Free and unlimited solutions for creating visualizations are widely available for non-profit initiatives. What is really lacking is the knowledge needed to create information from the data deluge that is observable even in the world of Open Data. More initiatives like this work need to exist and if they are well funded there might be a chance of offering the commoner, or common citizen – here referred as someone without digital literacy –, a place in the data-driven society where good decisions are pondered decisions, and there is little space for waste.

But another observation is that the commoner maybe, as indicated by some authors, should receive a minimal digital literacy education, as this would also help to close the gap of digital literacy. This education could not only teach over how to deal with raw data, but more importantly, how to read data and extract information from it not letting be fooled by some bad ways of presenting data. Informed citizens can reach better decisions, improve their lives and therefore the society.

6. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

Even though it is insisted here that Open Data was no limitation, it is reasonable to think that much data is still inside the institutions and therefore, a future upgrade to this project could try to collect data from the institutions to improve the information provided. Also, there are still hundreds of Open Data datasets which could constitute a source of value for this work, for instance, there is data available over how much money public academic institutions invest on research, although, there is literature showing that in some cases students are not fond of institutions with focus on research, and are looking for experiences which lead to work market insertion (Drewes & Michael, 2006; Silva et al., 2016). But still, it is valuable information.

Although the technology chosen for building the charts allows for beautiful and engaging charts to be built, there is complexity in the back-end code that could maybe be left behind if embedded Power BI charts were included. This is subject to further research given limitations imposed by Microsoft's platform when speaking of free Power BI Service accounts. The matter of cost is important for this project that needs to cost little enough to be self-maintainable.

Speaking about the back-end, this was a piece of the project which was hard to build and even though it brought immense knowledge in terms of Python and Django, it should be investigated if a solution such as Pentaho Data Integrator would do a better job on ingesting data. Although, the way it was built really makes sense when considering the whole architecture, as this makes all the projects' pieces a more cohesive whole. Maybe what also can be investigated is if the pipeline solution should be expanded and even became a Django pug-in for ETL or ELT functionality.

The layouts of the pages were built with little effort so much of the focus was aimed on what was harder and more important, although much still can be done with Bootstrap, the CSS library which govern how things are shown, to help build more attractive pages so the website can impress and create an engaging experience.

7. BIBLIOGRAPHY

Anderson, P. H., & Lawton, L. (2015). Student motivation to study abroad and their intercultural development. Frontiers: The Interdisciplinary Journal of Study Abroad.

Axelrod, J. (2019). Data democratization. American City & County Exclusive Insight. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,cookie,shib,uid&db=a9h&AN=135663038&lang=pt-pt&site=eds-live&scope=site

Ding, L., Lebo, T., Erickson, J. S., Difranzo, D., Williams, G. T., Li, X., Michaelis, J., Graves, A., Zheng, J. G., & Shangguan, Z. (2011). TWC LOGD: A portal for linked open government data ecosystems. Web Semantics: Science, Services and Agents on the World Wide Web, 9(3), 325–333.

EU Open Data Portal (2019). The European Union Data Portal. Retrieved from http://data.europa.eu/euodp/en/about

Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone? First Monday, 16(2).

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information Systems Management, 29(4), 258–268.

Kitchin, R. (2014). The data revolution: Big data, open data, data infrastructures and their consequences. Sage.

Lakomaa, E., & Kallberg, J. (2013). Open Data as a Foundation for Innovation: The Enabling Effect of Free Public Sector Information for Entrepreneurs. IEEE Access, 1, 558–563. https://doi.org/10.1109/ACCESS.2013.2279164

Luna-Reyes, L. F., Najafabadi, M. M., Zuiderwijk, A., & Hinnant, C. C. (2019). The US open data initiative: The road ahead. Information Polity: The International Journal of Government & Democracy in the Information Age, 24(2), 163–182. Retrieved from bth.

Open Data Barometer (2016). The Open Data Barometer: A global measure of how governments are publishing and using open data for accountability, innovation and social impact. Retrieved from https://opendatabarometer.org/4thedition/?_year=2016&indicator=ODB

Patil, D. J., & Mason, H. (2015). Data Driven. O'Reilly Media, Inc.

Sharda, R., Delen, D., & Turban, E. (2017). Business Intelligence, Analytics, and Data Science: A Managerial Perspective. Retrieved from https://books.google.pt/books?id=kIIbvgAACAAJ

Sawicki, D. S., & Craig, W. J. (1996). The democratization of data: Bridging the gap for community groups. Journal of the American Planning Association, 62(4), 512–523.

Treuhaft, S. (2006). The democratization of data: How the internet is shaping the work og data intermediaries. Retrieved from http://hdl.handle.net/10419/39261

Agrey, L., & Lampadan, N. (2014). Determinant factors contributing to student choice in selecting a university. *Journal of Education and Human Development*, *3*(2), 391–404.

Anderson, P. (1999). Factors influencing student choice in higher education. *Perspectives: Policy & Practice in Higher Education*, *3*(4), 128–131.

Anderson, P. H., & Lawton, L. (2015). Student motivation to study abroad and their intercultural development. *Frontiers: The Interdisciplinary Journal of Study Abroad*.

Axelrod, J. (2019). Data democratization. *American City & County Exclusive Insight*, N.PAG-N.PAG. a9h.

Briggs, S. (2006). An exploratory study of the factors influencing undergraduate student choice: The case of higher education in Scotland. *Studies in Higher Education*, *31*(6), 705–722. https://doi.org/10.1080/03075070601004333

Cloudflare. (2021a). *Types of SSL certificates | SSL certificate types explained*. Cloudflare. https://www.cloudflare.com/en-gb/learning/ssl/types-of-ssl-certificates/

Cloudflare. (2021b). *What is a DDoS attack?* Cloudflare. https://www.cloudflare.com/en-gb/learning/ddos/what-is-a-ddos-attack/

Conard, M. J., & Conard, M. A. (2000). An Analysis of Academic Reputation as Perceived by Consumers of Higher Education. *Journal of Marketing for Higher Education*, *9*(4), 69–80. https://doi.org/10.1300/J050v09n04_05

Dias, D. (2013). Students' Choices in Portuguese Higher Education: Influences and Motivations. *European Journal of Psychology of Education*, *28*(2), 437–451.

Ding, L., Lebo, T., Erickson, J. S., Difranzo, D., Williams, G. T., Li, X., Michaelis, J., Graves, A., Zheng, J. G., & Shangguan, Z. (2011). TWC LOGD: A portal for linked open government data ecosystems. *Web Semantics: Science, Services and Agents on the World Wide Web*, *9*(3), 325–333.

Drewes, T., & Michael, C. (2006). How do students choose a university?: An analysis of applications to universities in Ontario, Canada. *Research in Higher Education*, *47*(7), 781–800. https://doi.org/10.1007/s11162-006-9015-6

European Commission. (2011). *2011/833/EU: Commission Decision of 12 December 2011 on the reuse of Commission documents*. https://www.eumonitor.eu/9353000/1/j9vvik7m1c3gyxp/vivi69stinnd

European Commission. (2021). *A European Strategy for data*. https://digital-strategy.ec.europa.eu/en/policies/strategy-data

Graves, A., & Hendler, J. (2014). A study on the use of visualizations for Open Government Data. *Information Polity: The International Journal of Government & Democracy in the Information Age*, *19*(1/2), 73–91.

Gurstein, M. B. (2011). Open data: Empowering the empowered or effective data use for everyone? *First Monday*, *16*(2), Article 2.

Hahn, P., Allison. (2018). Data visualization. *Salem Press Encyclopedia*. http://widgets.ebscohost.com/prod/customerspecific/ns000290/authentication/index.php?url=https%3a%2f%2fsearch.ebscohost.com%2flogin.aspx%3fdirect%3dtrue%26AuthType%3dip%2ccookie%2cshib%2cuid%26db%3ders%26AN%3d129814577%26lang%3dpt-pt%26site%3deds-live%26scope%3dsite

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, *29*(4), 258–268.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage.

Labs, K. (2021). *What is an SSL certificate – Definition and Explanation*. https://www.kaspersky.com/resource-center/definitions/what-is-a-ssl-certificate

Lakomaa, E., & Kallberg, J. (2013). Open Data as a Foundation for Innovation: The Enabling Effect of Free Public Sector Information for Entrepreneurs. *IEEE Access*, *1*, 558–563. https://doi.org/10.1109/ACCESS.2013.2279164

Loveless, A. (2011). Technology, pedagogy and education: Reflections on the accomplishment of what teachers know, do and believe in a digital age. *Technology, Pedagogy and Education*, *20*(3), 301–316. https://doi.org/10.1080/1475939X.2011.610931

Luna-Reyes, L. F., Najafabadi, M. M., Zuiderwijk, A., & Hinnant, C. C. (2019). The US open data initiative: The road ahead. *Information Polity: The International Journal of Government & Democracy in the Information Age*, *24*(2), 163–182. bth.

Malgwi, C. A., Howe, M. A., & Burnaby, P. A. (2005). Influences on Students' Choice of College Major. *Journal of Education for Business*, *80*(5), 275–282. https://doi.org/10.3200/JOEB.80.5.275-282

Martí-Parreño, J., Seguí-Mas, D., & Seguí-Mas, E. (2016). Teachers' Attitude towards and Actual Use of Gamification. *Procedia - Social and Behavioral Sciences*, *228*, 682–688. https://doi.org/10.1016/j.sbspro.2016.07.104

Open Data Barometer. (2021). *About The Open Data Barometer* [The Open Data Barometer]. https://opendatabarometer.org/barometer/

Sawicki, D. S., & Craig, W. J. (1996). The democratization of data: Bridging the gap for community groups. *Journal of the American Planning Association*, *62*(4), 512–523.

Sharda, R., Delen, D., & Turban, E. (2017). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*. Pearson. https://books.google.pt/books?id=kIlbvgAACAAJ

Silva, P., Lopes, B., Costa, M., Seabra, D., Melo, A., Brito, E., & Dias, G. (2016). Stairway to employment? Internships in higher education. *Higher Education (00181560)*, *72*(6), 703–721. https://doi.org/10.1007/s10734-015-9903-9

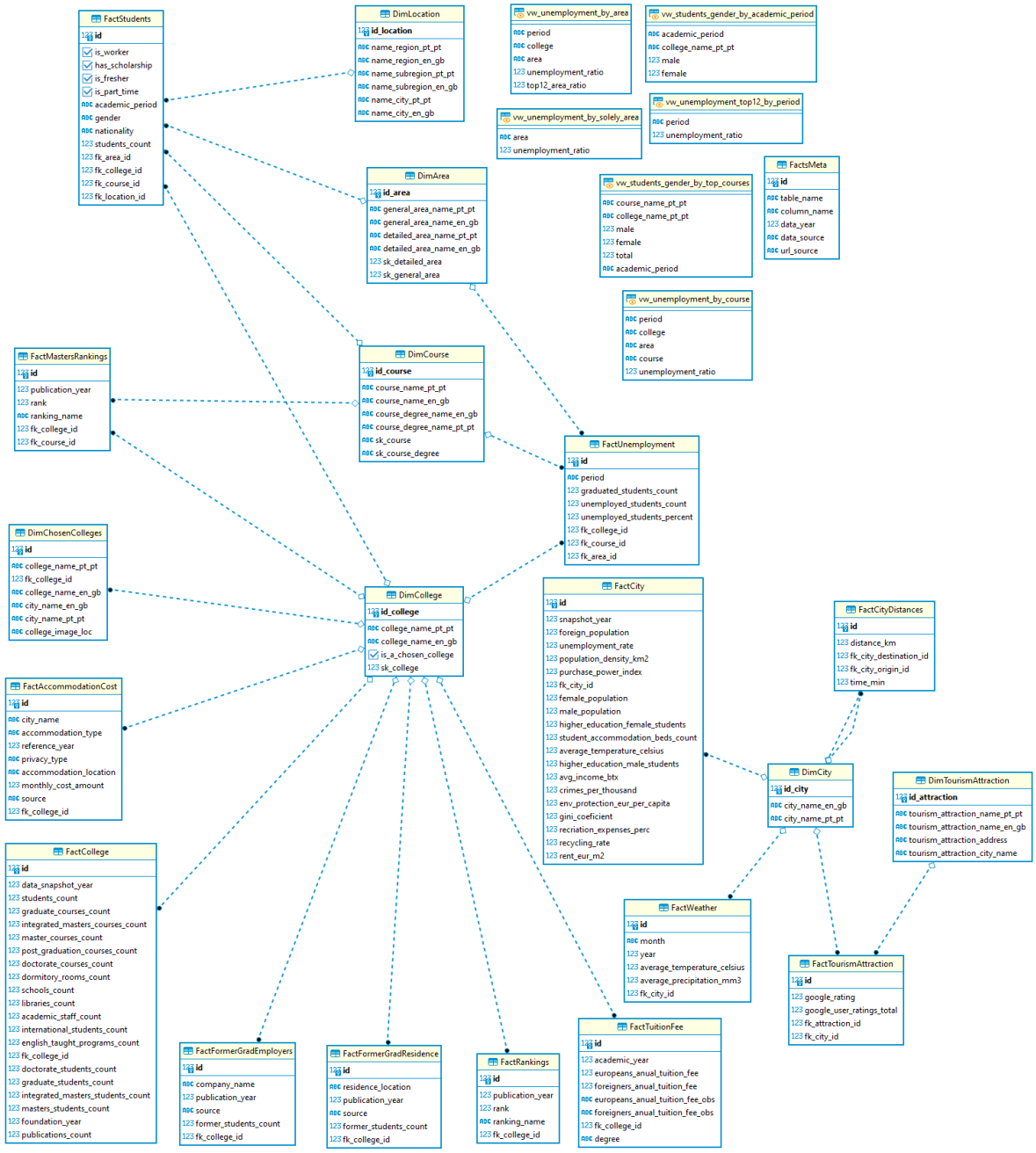
Tavares, O., & Cardoso, S. (2013). Enrolment choices in Portuguese higher education: Do students behave as rational consumers? *Higher Education (00181560)*, *66*(3), 297–309. https://doi.org/10.1007/s10734-012-9605-5

Teixeira, C., Gomes, D., & Borges, J. (2015). Introductory Accounting Students' Motives, Expectations and Preparedness for Higher Education: Some Portuguese Evidence. *Accounting Education*, *24*(2), 123–145. https://doi.org/10.1080/09639284.2015.1018284

Treuhaft, S. (2006). *The democratization of data: How the internet is shaping the work og data intermediaries*. University of California, Institute of Urban and Regional Development (IURD). http://hdl.handle.net/10419/39261

8. ANNEXES

8.1. DATA MODEL



8.2. Django Development Settings

```
"""
Django settings for workproject project.

Generated by 'django-admin startproject' using Django 2.2.14.

For more information on this file, see
https://docs.djangoproject.com/en/2.2/topics/settings/

For the full list of settings and their values, see
https://docs.djangoproject.com/en/2.2/ref/settings/
"""

import os

# Build paths inside the project like this: os.path.join(BASE_DIR, ...)
BASE_DIR = os.path.dirname(os.path.dirname(os.path.abspath(__file__)))


# Quick-start development settings - unsuitable for production
# See https://docs.djangoproject.com/en/2.2/howto/deployment/checklist/

# SECURITY WARNING: keep the secret key used in production secret!
SECRET_KEY = 'g@ulw-++(p-sndestmkjs%zccjdg$@)_$^emhv&-(@5a*z@a_c'

# SECURITY WARNING: don't run with debug turned on in production!
DEBUG = True

ALLOWED_HOSTS = ['192.168.1.101','localhost','192.168.1.13']


# Application definition

INSTALLED_APPS = [
    'django.contrib.admin',
    'django.contrib.auth',
    'django.contrib.contenttypes',
    'django.contrib.messages',
    'django.contrib.sessions',
    'django.contrib.staticfiles',
    'django.contrib.sites',
    'data_pages',
    'etl',
]

SITE_ID = 1   #define the site id

MIDDLEWARE = [
    'django.middleware.security.SecurityMiddleware',
    'django.contrib.sessions.middleware.SessionMiddleware',
    'django.middleware.locale.LocaleMiddleware',
    'django.middleware.common.CommonMiddleware',
    'django.middleware.csrf.CsrfViewMiddleware',
    'django.contrib.auth.middleware.AuthenticationMiddleware',
    'django.contrib.messages.middleware.MessageMiddleware',
    'django.middleware.clickjacking.XFrameOptionsMiddleware',
]

ROOT_URLCONF = 'workproject.urls'

```

```
60  TEMPLATES = [
61      {
62          'BACKEND':
    'django.template.backends.django.DjangoTemplates',
63          'DIRS': [
64              os.path.join(BASE_DIR,
    'data_pages/templates/data_pages'),
65              os.path.join(BASE_DIR,
    'data_pages/templates/data_pages/colleges_data'),
66              ],
67          'APP_DIRS': True,
68          'OPTIONS': {
69              'context_processors': [
70                  'django.template.context_processors.request',
71                  'django.contrib.auth.context_processors.auth',
72
    'django.contrib.messages.context_processors.messages',
73                  'django.template.context_processors.debug',
74
75
76              ],
77          },
78      },
79  ]
80
81  WSGI_APPLICATION = 'workproject.wsgi.application'
82
83
84  # Database
85  # https://docs.djangoproject.com/en/2.2/ref/settings/#databases
86
87  DATABASES = {
88      'default': {
89          'ENGINE': 'django.db.backends.postgresql',
90          'NAME': 'postgres',
91          'USER': 'postgres',
92          'PASSWORD': 'postgres',
93          'HOST': 'db',
94          'PORT': 5432,
95      }
96  }
97
98  DEFAULT_AUTO_FIELD = 'django.db.models.AutoField'
99
100 # Password validation
101 # https://docs.djangoproject.com/en/2.2/ref/settings/#auth-password
    -validators
102
103 AUTH_PASSWORD_VALIDATORS = [
104     {
105         'NAME':
    'django.contrib.auth.password_validation.UserAttributeSimilarityVal
    idator'
    ,
106     },
107     {
108         'NAME':
    'django.contrib.auth.password_validation.MinimumLengthValidator',
109     },
110     {
111         'NAME':
    'django.contrib.auth.password_validation.CommonPasswordValidator',
112     },
113     {
114         'NAME':
    'django.contrib.auth.password_validation.NumericPasswordValidator',
115     },
116 ]
117
118
```

```
119  # Internationalization
120  # https://docs.djangoproject.com/en/2.2/topics/i18n/
121
122  LANGUAGE_CODE = 'pt'
123
124  LANGUAGES = [
125    ('pt', 'Portuguese'),
126    ('en', 'English'),
127  ]
128
129  TIME_ZONE = 'UTC'
130
131  USE_I18N = False
132
133  USE_L10N = True
134
135  USE_THOUSAND_SEPARATOR = True
136
137  USE_TZ = True
138
139
140  # Static files (CSS, JavaScript, Images)
141  # https://docs.djangoproject.com/en/2.2/howto/static-files/
142
143  STATIC_URL = '/static/'
144
145
```

8.3. Firewall Configuration In Production

```
Chain INPUT (policy ACCEPT)
num  target     prot opt source               destination
1    ACCEPT     all  --  anywhere             anywhere
2    ACCEPT     all  --  anywhere             anywhere             ctstate RELATED,ESTABLISHED
3    ACCEPT     tcp  --  anywhere             anywhere             tcp dpt:ssh
4    ACCEPT     tcp  --  anywhere             anywhere             tcp dpt:http limit: avg 25/min burst 100
5    ACCEPT     tcp  --  anywhere             anywhere             tcp dpt:https limit: avg 25/min burst 100
6    ACCEPT     icmp --  anywhere             anywhere             icmp echo-request
7    ACCEPT     icmp --  anywhere             anywhere             icmp echo-reply
8    ACCEPT     tcp  --  anywhere             anywhere             tcp dpt:postgresql
9    LOG        all  --  anywhere             anywhere             limit: avg 5/min burst 5 LOG level debug prefix "iptables denied: "
10   DROP       all  --  anywhere             anywhere

Chain FORWARD (policy ACCEPT)
num  target     prot opt source               destination

Chain OUTPUT (policy ACCEPT)
num  target     prot opt source               destination
root@projectonortear:~#
```

8.4. APACHE CONFIGURATION IN PRODUCTION

```
<VirtualHost *:443>

        ServerName projectonortear.pt
        ServerAlias www.projectonortear.pt
        ServerAdmin rodrigo.ffd.silva@gmail.com

        DocumentRoot /srv/sites/projectonortear.pt/workproject

        #<If "%{HTTP_HOST} == 'www.projectonortear.pt'">
        #       Redirect permanent / https://projectonortear.pt/
        #</If>
        #Alias /.well-known/pki-validation/1D0114B7793AE3C249C1F19B0C1E08C8.txt /srv/sites/projectonortear.pt/workproject/
        .well-known/pki-validation/1D0114B7793AE3C249C1F19B0C1E08C8.txt

        #<Directory /srv/sites/projectonortear.pt/workproject/.well-known/pki-validation>
        #         Require all granted
        #</Directory>

        Alias /awstatsclasses "/usr/share/awstats/lib/"
        Alias /awstats-icon "/usr/share/awstats/icon/"
        Alias /awstatscss "/usr/share/doc/awstats/examples/css"
        ScriptAlias /awstats/ /usr/lib/cgi-bin/
        Options +ExecCGI -MultiViews +SymLinksIfOwnerMatch

        Alias /static/ /srv/sites/projectonortear.pt/data_pages/static/

        <Directory /srv/sites/projectonortear.pt/data_pages/static>
                Require all granted
        </Directory>

        <Directory /srv/sites/projectonortear.pt/data_pages/static/images>
                Require all granted
        </Directory>

        Alias /favicon.ico /srv/sites/projectonortear.pt/data_pages/static/images/favicon.ico

        <Directory /srv/sites/projectonortear.pt/workproject>
                <Files wsgi.py>
                        Require all granted
                </Files>
        </Directory>

        WSGIScriptAlias / /srv/sites/projectonortear.pt/workproject/wsgi.py
        WSGIDaemonProcess projectonortear.pt lang='pt_PT.UTF-8' locale='pt_PT.UTF-8' python-path=/srv/sites/projectonortear.pt
        WSGIProcessGroup projectonortear.pt

        SSLEngine on
        SSLCertificateFile /srv/sites/projectonortear.pt/installation/certs/certificate.crt
        SSLCertificateKeyFile /srv/sites/projectonortear.pt/installation/certs/private.key
        SSLCACertificateFile /srv/sites/projectonortear.pt/installation/certs/ca_bundle.crt

        ErrorLog ${APACHE_LOG_DIR}/error.log
        CustomLog ${APACHE_LOG_DIR}/access.log combined

</VirtualHost>
```