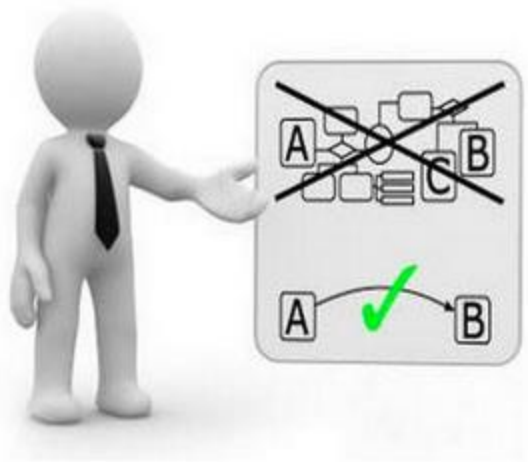
Introducción al control y gestión de la calidad de los datos de investigación.

Yusnelkis Milanes Guisado. PhD

Biblioteca / CRAI

Agenda

- Importancia de la calidad de los datos
- ¿Qué entendemos por calidad de los datos?
- Dimensiones de la calidad de los datos
- Consecuencias de los errores en los datos
- Cómo elaborar la documentación relativa a los datos
- Check list_Gestión de calidad de los datos
  - Calidad de los ficheros de datos
  - Calidad de la documentación sobre los datos
  - Calidad de valores de los parámetros en tu dataset
  - Información temporal
- Métodos y herramientas útiles



2 niveles de magnitud

Grandes datasets – estudios multicéntricos, Investigación multiescala, Big data

Datasets pequeños.
Productores de datos

Desafíos_ Datos Big Data

Los usuarios de datos no son necesariamente productores de datos.

Es muy difícil medir la calidad de los datos

Desafíos_ Datos Big Data

AIDS

Articles & Issues | For Authors | Journal Info

Outline

Download

Cite

Share

Favorites

Permissions

SUPPLEMENT ARTICLES

Power of Big Data in ending HIV

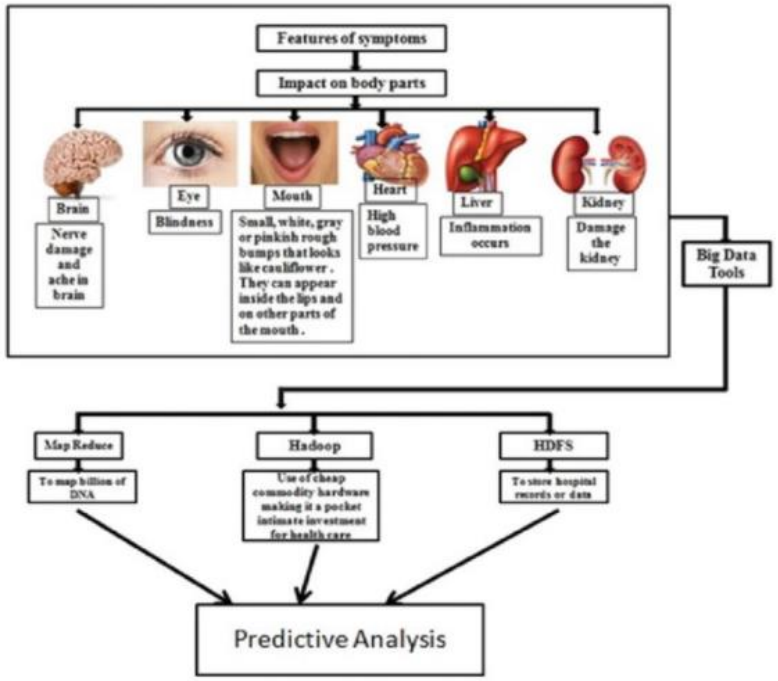
Olatosi, Bankole[a,b]; Vermund, Sten H.[c]; Li, Xiaoming[a,d] Author Information

AIDS: May 1, 2021 - Volume 35 - Issue - p S1-S5
doi: 10.1097/QAD.0000000000002888

FREE

Abstract

The articles in this special issue of *AIDS* focus on the applic
so-called Big Data science (BDS) as applied to a variety of H
research questions in the sphere of health services and epi
Recent advances in technology means that a critical mass o
health data with actionable intelligence is available for opt
outcomes, improving and informing surveillance. Data science will play a key but complementary role in supporting current efforts in prevention, diagnosis, treatment, and response needed to end the HIV epidemic. This collection provides a glimpse of the promise inherent in leveraging the digital age and improved methods in Big Data science to reimagine HIV treatment and prevention in a digital age.



Desafíos_ Datos Big Data

Using Big Data to Improve HIV Treatment Program Outcomes in South Africa

Big data approaches to answering the Government's 3 questions

	Data Science Analytical approach	Databases used
VLD: Do people who are on HIV treatment, get their HIV viral load detected as per South Africa's HIV treatment guidelines?	• Create a temporal patient database with consecutive lab results, per facility • Compare VL tests performed at specific time intervals against the number of HIV treatment clients at facility	TIER.net dhis2 NATIONAL HEALTH LABORATORY SERVICE Harmonised master list of health facilities
VLS: Are people on HIV treatment in SA virally suppressed?	• Use temporal patient database with consecutive VL lab results, per facility • Check VLS status disaggregated by sub population	Temporal set of patient data Harmonised master list of health facilities
CD4 recovery: Does this viral suppression lead to improved health for HIV patients?	• Use temporal patient database with consecutive CD4 lab results, per facility • Check CD4 status disaggregated by sub population • Determine temporal change	Temporal set of patient data Harmonised master list of health facilities
Spatial distribution: Are there spatial patterns?	2 types of spatial correlation analyses: • Moran's I • Geary's c	VLD and VLS results from above Harmonised master list of health facilities

TIER.net HIV Electronic Register

- Three Interlinked Electronic Registers (TIERs)
- Since 2011
- 3-tiered electronic patient management system
- Captures **patient-level data** on HIV counselling and testing, pre-HIV-treatment and HIV-treatment services

WORLD BANK GROUP

NATIONAL HEALTH LABORATORY SERVICE

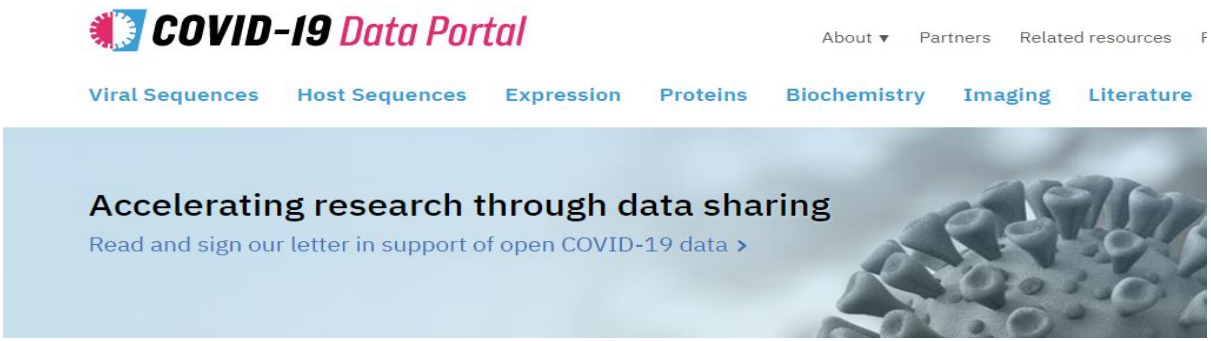
- NHLS is the largest diagnostic pathology service in South Africa
- Supports national and provincial health departments
- Provides laboratory and related public health services to over 80% of the population through a national network of laboratories
- Samples to NHLS laboratory, test performed and results via SMS printer to facility
- Manual transcription to patient file
- Houses a Corporate Data Warehouse (CDW) on all **laboratory tests and their results**
  - For HIV: viral load and CD4 test results
  - NO unique client identifiers

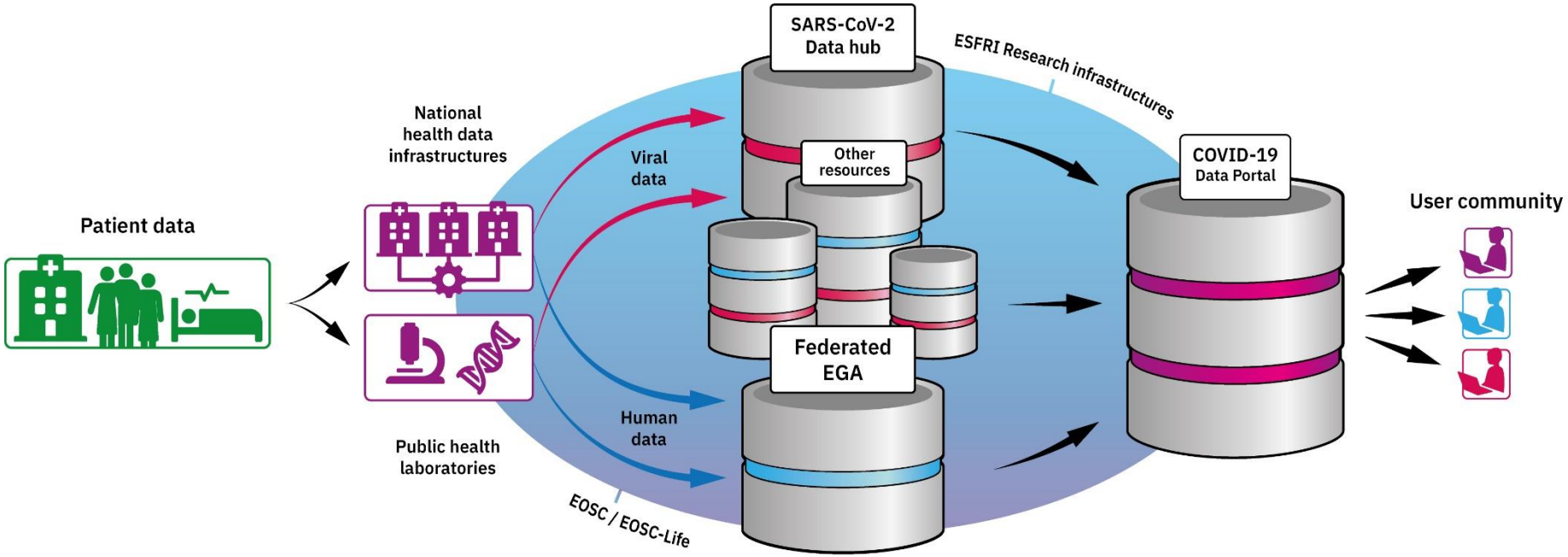
dhis2

- District Health Information System
- South Africa's health management information system
- Summarises data from 'tick registers' and patient that are completed daily
- Data in DHIS based on national indicator set for health service monitoring
- NOT patient level monitoring
- Includes aggregate HIV data (number of patients and types of services, in aggregated form) on HIV testing, HIV treatment and other HIV services

https://thedocs.worldbank.org/en/doc/317541541431615271-0090022018/original/3.SouthAfricaBigDataAnalyticsPPT.pdf

Desafíos_ Datos Big Data





https://www.covid19dataportal.org/the-european-covid-19-data-platform

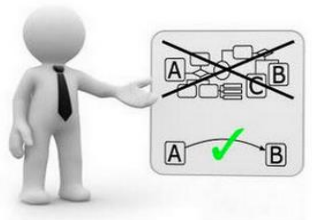
Desafíos_ Datos COHORTES

Control de Calidad _ Cohorte nacional VIH/SIDA

A1 | ERRORCODE

	A	B	C	D	E
1	**ERRORCODE**	**TABLE**	**DESCRIPTION**	**ACLARACION**	**RECORDSIDENT**
2	AC001	TRATAMIENTOS	Patient has no record in table BAS	Revisar paciente: tiene información en la tabla tratamiento pero no	1
3	AW009	TRATAMIENTOS	Missing art_id	Revisar registro sin código de tratamiento	422
4	BC002	IDENTIFICACION	AIDS_Y=0, but AIDS-defining records in tblDIS	Revisar primera enfermedad de Sida	1
5	B_COHERE007	IDENTIFICACION	MODE is homo/bisexual or heamophilac and GENDER=2 (Fe	Revisar categoría de trasmisión y sexo	1
6	CATC002	ANALÍtICAS	CD4_D after DROP_D in tblLTFU	Revisar fecha de CD4 y fecha de muerte o última visita	30
7	CC_TRAT001	TRATAMIENTOS	Paciente con informacion en tblCC_tratamientos y no en Id	Paciente sin información en tabla identificación	1
8	CEPATC002	SAT-ENoS o FIBROSCÁN	CEP_D after DROP_D in tblLTFU	Revisar fecha de ENoS y fecha de última visita	2
9	CEPC003	SAT-ENoS o FIBROSCÁN	CEP_D after L_ALIVE in tblLTFU	Revisar fecha de ENoS posterior a fecha de muerte o última visita	10
10	CEPW001	SAT-ENoS o FIBROSCÁN	Duplicate Records, same CEP_ID, CEP_SPEC, and CEP_D	Revisar registro duplicado mismo ENoS o Fibroscan y misma fecha	5
11	CEP_ATC006a	SAT-ENoS o FIBROSCÁN	CEP_ID not coded as coding lists on table definition	CoRIS: revisar codificación	7
12	CW002	ANALÍtICAS	Duplicate Records for same date	Revisar mediciones de CD4 duplicadas	8
13	C_COHERE004	ANALÍtICAS	All patients listed in BAS should have at least one entry in tbL	Paciente sin ninguna medición de CD4	10
14	C_COHERE007	ANALÍtICAS	CD4_D > CENS_D	Revisar fecha de CD4 y fecha de muerte o última visita	41
15	C_COHERE008	ANALÍtICAS	CD4_D > L_ALIVE	Revisar fecha de CD4 y fecha de muerte o última visita	39
16	CoRIS	IDENTIFICACION	Incluido en la actualizacion anterior	Revisar si el paciente ha sido correctamente excluido	530
17	DATC006DIS_ID	ENFERMEDADES SIDA	DIS_ID not coded as coding lists on table definition	Revisar Enfermedades de Sida	1
18	DC002	ENFERMEDADES SIDA	AIDS-defining records, yet AIDS=0 in tblBAS	Tienen registros en la tablas DIS	1
19	DW001	ENFERMEDADES SIDA	Duplicate Records for same DIS_ID and same DIS_D	Revisar enfermedades de sida duplicadas	1
20	DW003	ENFERMEDADES SIDA	Miscoded DIS_ID - as in code list attached to table definitio	Revisar Enfermedades de Sida	1
21	EA	tblART_EA	Efecto Adverso no especificado	Especificar el Efecto Adverso correspondiente	39
22	EC_VR	tblART_EA	El campo de Ensayo o Vida Real estÃ¡ vacÃ-o o es deconoci	El campo ART_CT debe de estar relleno a partir del 2014	3172
23	EVENTO_MUERTE	tblENOS	Muere por neoplasia, causa hepatica o infarto y no esta en	Revisar y rellenar ENO correspondiente	1
24	HCVA	SEROLO &SAT-HEP	No tiene test de anticuerpos VHC en los 12 meses tras la inc	En los 6 primeros meses después de entrar un paciente en la cohort	333
25	HCVR	SEROLO &SAT-HEP	No tiene Carga Viral VHC y tiene un test de anticuerpos VHC	Pacientes con un test de anticuerpos de VHC positivo y no tienen ni	4
26	LATC002	PERFIL BASICO/CD8	LAB_D after DROP_D in tblLTFU	Revisar fecha de analítica (Perfil básico, CD8) o fecha de última visi	112
27	LF_COHERE002	FIN SEGUIMIENTO	L_ALIVE > CENS_D	Revisar causa de muerte y última visita	9
28	LVW005	SEROLO &SAT-HEP	Missing VS_R	Revisar resultado de test o serología	3
29	LVW006	SEROLO &SAT-HEP	Missing VS_V	Revisar resultado de test o serología	11

Summary | AC001 | AW009 | BC002 | B_COHERE007 | CATC002 | CC_TRAT001 | CEPATC002 | CEPC003 | CEPW001 | CEP_ATC006a | CW002 | C_COHERE004 | C_C



Desafíos_ Datos Biga Data

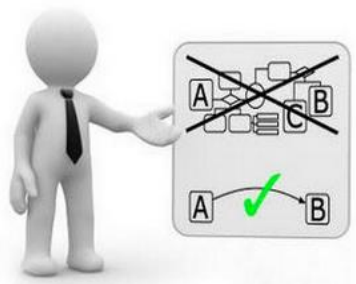
En la actualidad, la calidad de datos de big data se enfrenta a los siguientes desafíos:

- **La diversidad de fuentes de datos** aporta abundantes tipos de datos y estructuras de datos complejas y aumenta la dificultad de la integración de datos.

- **El volumen de datos es tremendo**, y es difícil juzgar la calidad de los datos dentro de un tiempo razonable.

- **Los datos cambian muy rápido** y la "puntualidad" de los datos es muy corta, lo que requiere mayores requisitos para la tecnología de procesamiento.

- **No hay demasiados estándares de calidad de datos unificados** y aprobados y la investigación sobre la calidad de datos de big data.



¿Datos de mala calidad ?

- **Incompletos**
- **Fluctuantes**
- **No comparables**
- **Ruido**

Gestión de datos de investigación

¿Por qué le sería útil tener habilidades gestionando los datos de su investigación…?

- **Para guardar sus archivos** en formatos a los que todos puedan acceder, sin importar si tienen acceso a cierto software o no

- **Para mantener su computadora organizada**

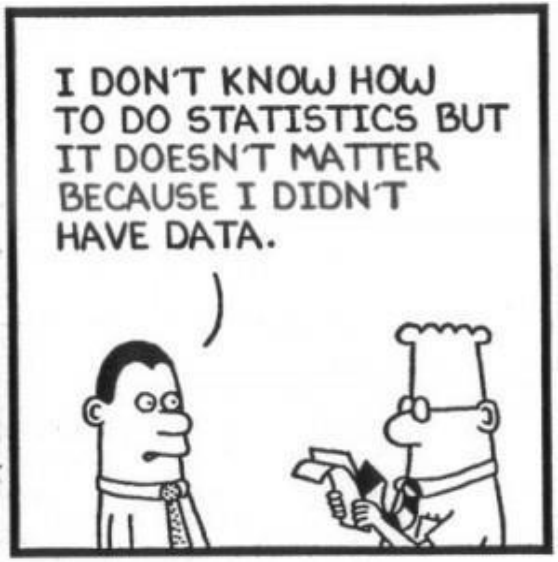
- **Para documentar su proceso de investigación, decisiones y cambios**.

  Si no lo anota, ¡lo olvidará!

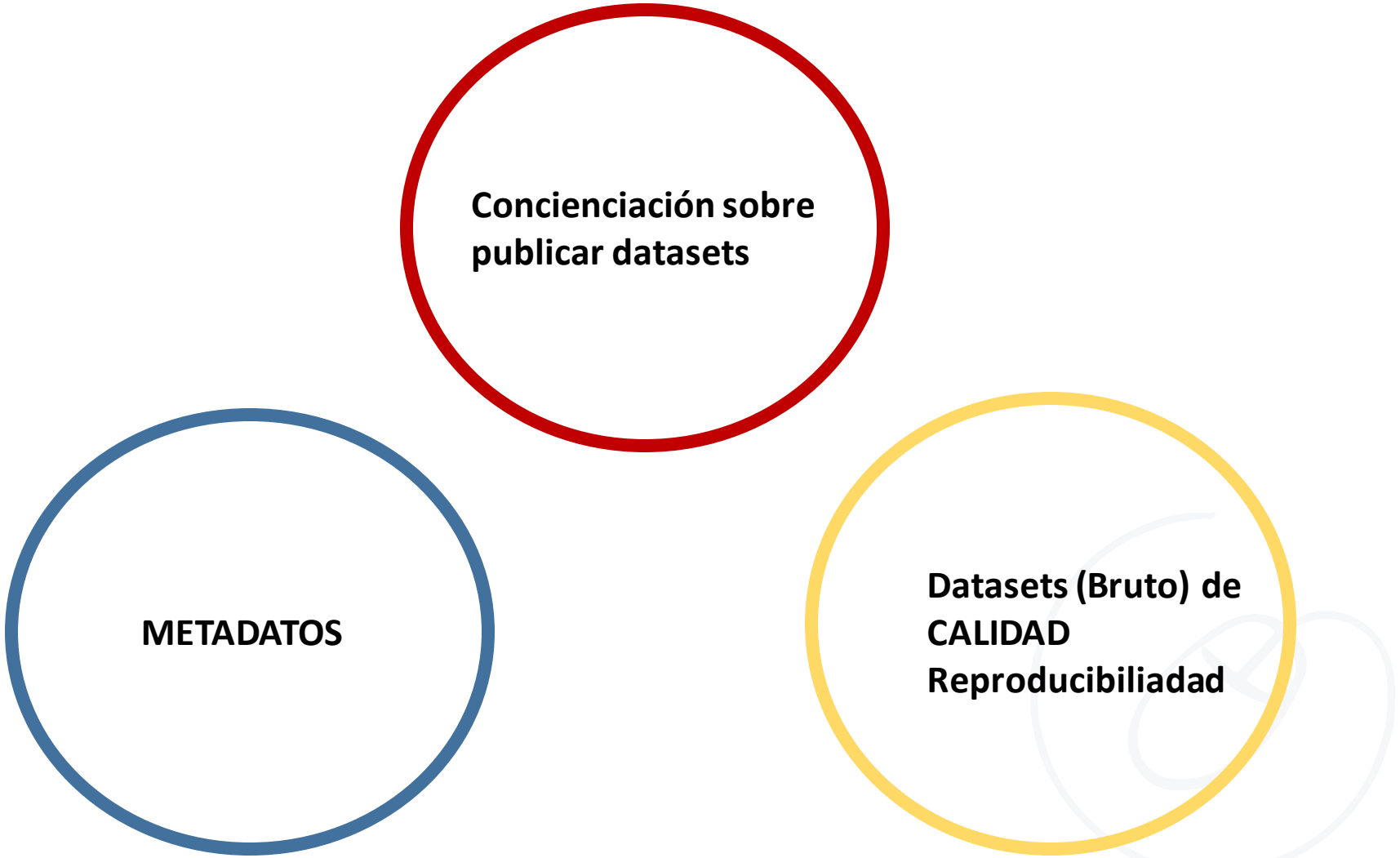
- **Para realizar copias de seguridad de sus datos con regularidad**, en varios lugares y en más de un tipo de medios, teniendo en cuenta una estrategia de seguridad

- **Para reutilizar sus datos** cuando lo necesite, usted y otros investigadores.

- **Para garantizar reproducibilidad de investigación de calidad**



Ciencia reproducible



Datos de investigación

“datos que son recolectados, observados o creados para ser analizados y producir resultados de investigación originales”

•Numéricos, descriptivos o visuales.

•Encontrarse en estado bruto o analizado, pueden ser experimentales u observacionales.

Plan de Gestión de Datos de Investigación

Planes de Gestión de Datos de Investigación

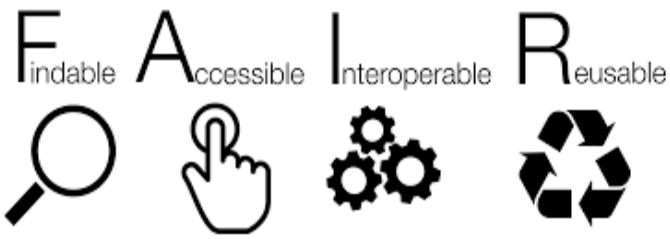
- *"Documentos que describen que harás con tus datos durante tu investigación y una vez que termines con tu proyecto"*

EL DMP debe contener solo la información más importante y preferiblemente **no debería exceder los 10.000 caracteres (incluidos los espacios).**

Plan de Gestión de Datos de Investigación

- ¿En qué **tipo de datos** se basa la investigación?
- ¿Los datos van a ser derivados, se van a crear de cero, recopilar o reutilizar?
- ¿Qué **formatos** se manejarán?
- ¿Cuál es el **tamaño total esperado de los datos** recopilados?
- ¿Cómo se va a establecer la **estructura de las carpetas**?
- ¿Cómo se va a definir la **nomenclatura** de archivos?
- ¿Se implementarán **estándares** específicos, como convenciones de nomenclatura o estructuras de codificación estandarizadas?

Datos FAIR



- se pueden **encontrar** en Internet,
- son **accesibles** (derechos y licencias claros),
- están en un **formato utilizable**,
- **se identifican de una manera única y persistente** para que se pueda hacer referencia a ellos.

Herramientas para valorar si tus datos son FAIR:

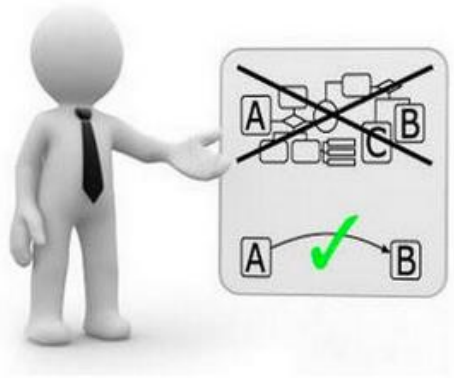
FAIR data self-assessment tool (https://ardc.edu.au/resources/working-with-data/fair-data/fair-self-assessment-tool/)

FAIR-Aware (https://fairaware.dans.knaw.nl/)

Calidad de los datos

Los datos tienen calidad cuando:

- Se usan según el contexto
- Son útiles
- Fáciles de entender y representar
- Sobre todo, CORRECTOS.
- **Deben permitir la REPRODUCIBILIDAD de la investigación**



Calidad de los datos

Los datos tienen calidad cuando son confiables:

- **Exactos:**
  - Son precisos
  - El valor o su representación refleja la info. de origen.
  - No hay ambigüedad.
- **Consistentes:**
  - Después de procesar los datos, sus conceptos, dominios de valor y formatos, todavía coinciden como antes de procesar
  - Durante todo el tiempo permanecen verificables
  - Evita info. contradictoria al hacer cruce

Calidad de los datos

Los datos tienen calidad cuando son confiables:

- **No Duplicados:**
- **Integridad:**
  - El formato de los datos es claro y cumple los criterios definidos
  - Los datos son consistentes con la estructura e integridad del contenido
- **Complitud:**
  - Si una deficiencia de un componente afectará la precisión y la integridad de los datos

Consecuencias de los errores en los datos

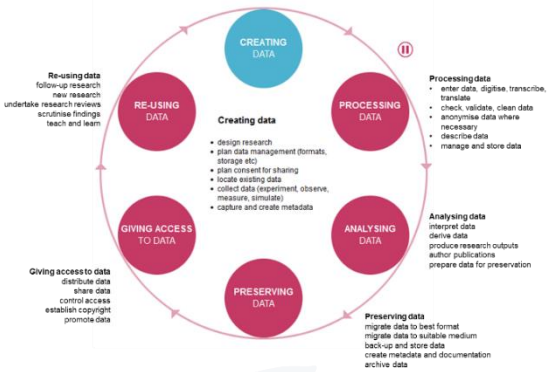
- Proyectos Fallidos
- Valores anómalos, ausentes, RESULTADOS SESGADOS
- Pérdida de tiempo y efectividad en la investigación
- Decisiones erróneas
- Dificulta el Compartir los datos (Data-Sharing)
- No – Estándares
- RUIDO
- Costes

¿Motivos de errores en los datos?

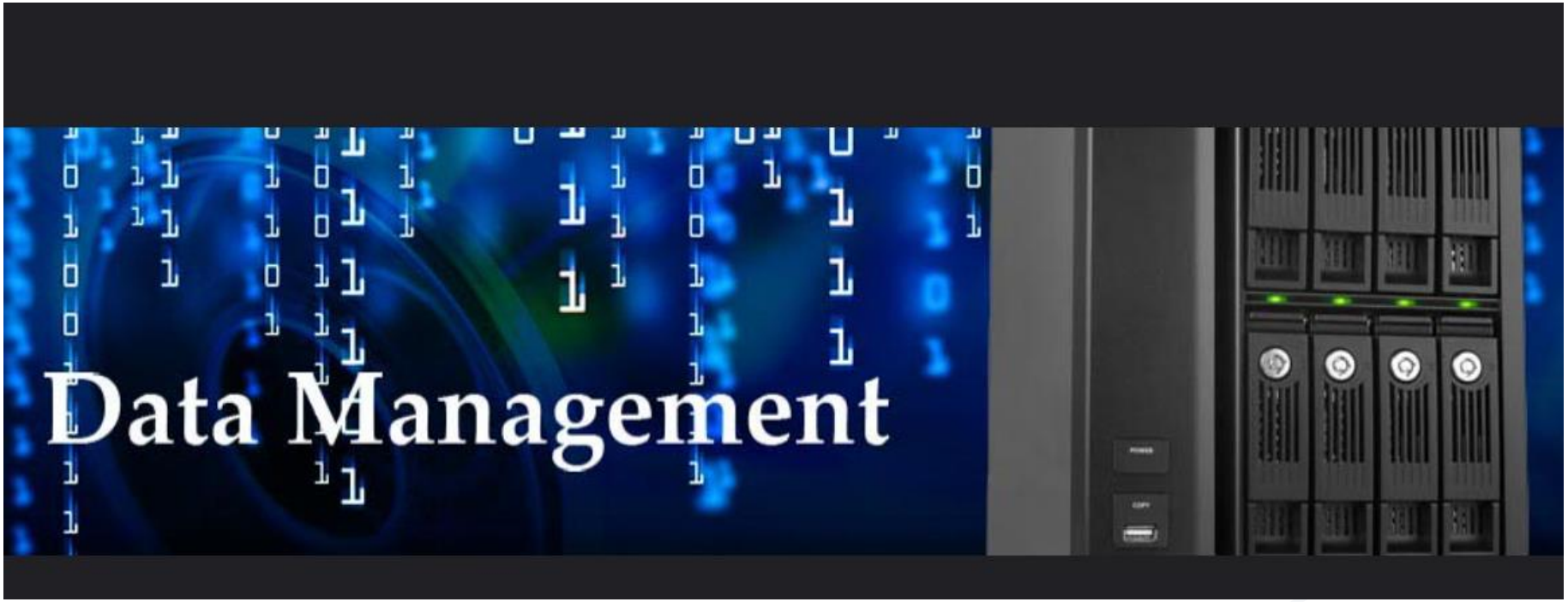
- Datos de entrada (Humanos)
- Datos externos
- Errores arrastrados o de carga de otros sistemas (migraciones)
- Errores en la codificación del sistema de variables
- Sesgos en el diseño de encuestas
- Errores en la integridad de los datos al transcribir entrevistas o cuestionarios
- Falta de normalización de los datos

¿Cuándo realizar el control de la calidad ?

- Antes de la recogida de datos: durante el diseño de los instrumentos de recolección de datos y documentación relativa a los datos
- Durante la recolección de los datos
- Preparación de los datos
- Durante el Análisis (Análisis Exploratorio de Datos (AED) y Análisis final)
- Antes de depositar los datos, resultados y compartirlos.



Ciclo de vida de los datos



"Because Good research needs Good data"

DCC. Digital Curation Center.

¿Cómo identificar los errores en nuestros datos?

- 1) Verificación manual
- 2) Análisis estadístico

  2.1) Análisis exploratorio de datos
  2.2.) Visualización exploratoria de datos

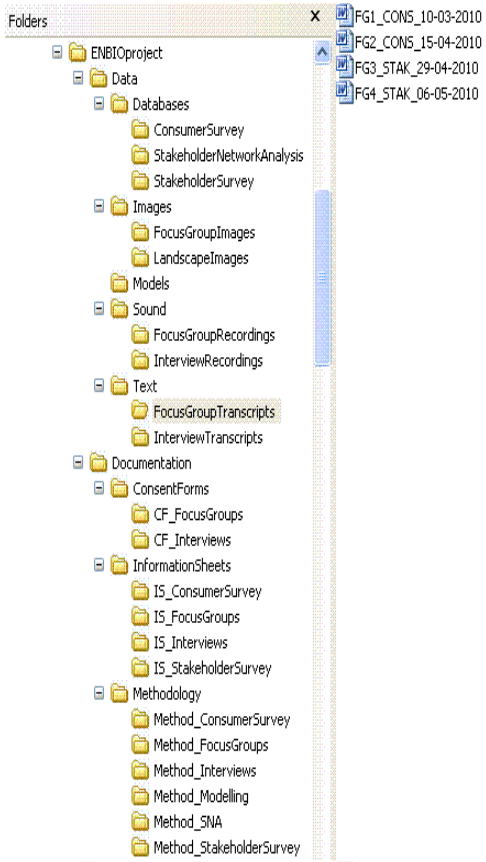
3) Análisis de correspondencia con la documentación de los datos (Diccionario de datos, Cuestionario; Guía del usuario, etc.)

Dimensiones a evaluar

- **Archivos y carpetas**
- **Documentación de los datos**
- **Temporalidad de los datos**
- **Valores de los datos, transformación y representación**

Checklist _ CALIDAD_ARCHIVOS Y CARPETAS

Archivos y carpetas



Checklist _ CALIDAD_ARCHIVOS Y CARPETAS

1. Comprueba la integridad de los ficheros (tamaño, número de ficheros)

2. Los nombres de los ficheros son descriptivos y consistentes

3. Comprueba que el formato de tus datos es apropiado (lo más abierto)

4. La organización de las carpetas es consistente y apropiada

5. El encabezado de las tablas es completa y coherente con la documentación

6. El versionado es apropiado.

Checklist _ CALIDAD_ARCHIVOS Y CARPETAS

Ejemplo:

01_Modelosdata2021_V3_20201121_YMG.

- ✓01 – Paquete de trabajo
- ✓**Modelosdata2021** – Datos del experimento, actividad, etc.
- ✓V3 – Versión
- ✓20201121 – Fecha
- ✓YMG – Autor(a)

Title:		Vision screening tests in Essex nurseries	
File Name:		VisionScreenResults_00_05	
Description:		Results data of 120 Vision Screen Tests carried out in 5 nurseries in Essex during June 2007	
Created By:		Chris Wilkinson	
Maintained By:		Sally Watsley	
Created:		04/07/2007	
Last Modified:		25/11/2007	
Based on:		VisionScreenDatabaseDesign_02_00	
Version	Responsible	Notes	Last amended
00_05	Sally Watsley	Version 00_03 and 00_04 compared and merged by SW	25/11/2007
00_04	Vani Yussu	Entries checked by VY, independent from SK	17/10/2007
00_03	Steve Knight	Entries checked by SK	29/07/2007
00_02	Karin Mills	Test results 81-120 entered	05/07/2007
00_01	Karin Mills	Test results 1-80 entered	04/07/2007

Fuente: UK Data Archive. (2017). Version control & authenticity. Retrieved June 22, 2017, from http://www.data-archive.ac.uk/create-manage/format/versions

Documentación relativa a los datos

https://bitbucket.org/ukda/ukds.tools.textanonhelper/wiki/Home

Documentación

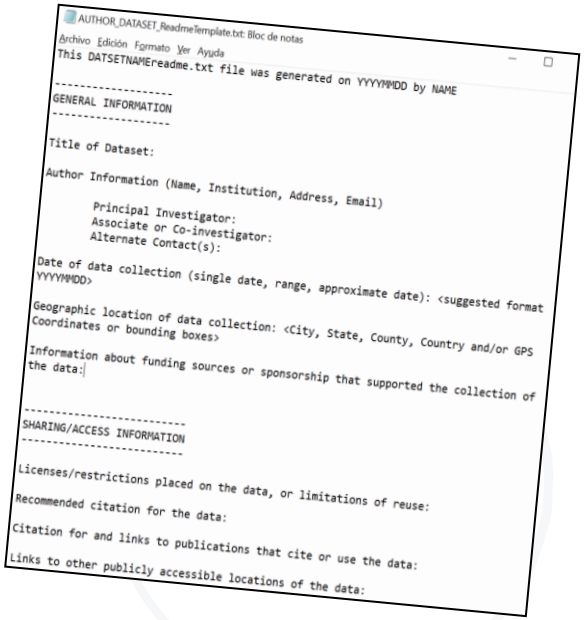
Debe acompañar a los datos para que estos se puedan comprender y reutilizar.

- El **contexto** de la recopilación de datos: historia del proyecto, objetivos e hipótesis
- **Métodos de recopilación de datos**: muestreo, proceso de recopilación de datos, instrumentos utilizados, hardware y software utilizado, escala y resolución, cobertura temporal y geográfica y fuentes de datos secundarias utilizadas
- **Estructura de los archivos de datos y relaciones entre archivos**
- **Validación de datos**, verificación, limpieza y procedimientos llevados a cabo para asegurar su calidad
- **Cambios** realizados en los datos a lo largo del tiempo desde su creación original e identificación de las diferentes versiones
- **Información sobre el acceso**, condiciones de uso o confidencialidad

Documentación

Niveles fundamentales para la documentación

- 1. **A nivel de proyecto:** Se documenta objetivos del estudio, preguntas de investigación, metodologías, instrumentos de medida, etc.
- **2. A nivel de Base de datos y Fichero:** Se documenta cómo todos los ficheros que conforman el data-set se relacionan. Se incluye un fichero **"readme.txt"** con la información relevante.
- **3. A nivel de Variables e ítems:** Se incluye un fichero tipo diccionario, no sólo con los nombres de las variables, sino con sus respectivas etiquetas explicando su significado en el contexto del estudio.


AUTHOR_DATASET_ReadmeTemplate.txt: Bloc de notas

Archivo Edición Formato Ver Ayuda

This DATSETNAMEreadme.txt file was generated on YYYYMMDD by NAME

GENERAL INFORMATION

Title of Dataset:

Author Information (Name, Institution, Address, Email)

Principal Investigator:
Associate or Co-investigator:
Alternate Contact(s):

Date of data collection (single date, range, approximate date): <suggested format YYYYMMDD>

Geographic location of data collection: <City, State, County, Country and/or GPS Coordinates or bounding boxes>

Information about funding sources or sponsorship that supported the collection of the data:

SHARING/ACCESS INFORMATION

Licenses/restrictions placed on the data, or limitations of reuse:

Recommended citation for the data:

Citation for and links to publications that cite or use the data:

Links to other publicly accessible locations of the data:

Documentación

Readme.txt

Fichero en que **se describe la información necesaria para que los conjuntos de datos sean comprensibles y reutilizables**: autoría, título, descripción, metodología, proyectos financiadores, cobertura temporal y geográfica, derechos de uso y privacidad, etc.

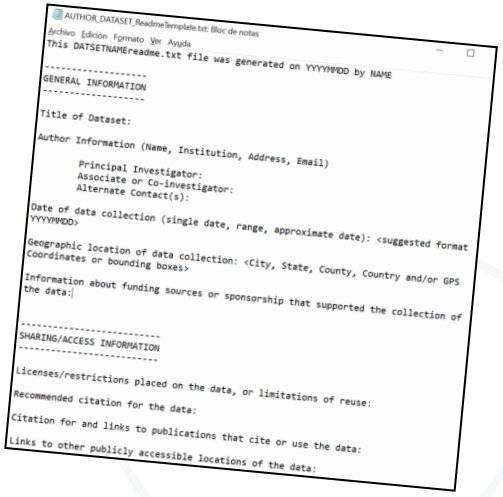
https://edatos.consorciomadrono.es/readme.xhtml

Guide to writing "readme" style metadata

A readme file provides information about a data file and is intended to help ensure that the data can be correctly interpreted, by yourself at a later date or by others when sharing or publishing data. Standards-based metadata is generally preferable, but where no appropriate standard exists, for internal use, writing "readme" style metadata is an appropriate strategy.

Want a template? **Download one** and adapt it for your own data!

- Best practices
- Recommended content
  - General information
  - Data and file overview
  - Sharing and access information
  - Methodological information
  - Data-specific information
- References
- Related information



https://data.research.cornell.edu/content/readme#fileoverview

Documentación

Diccionario de datos

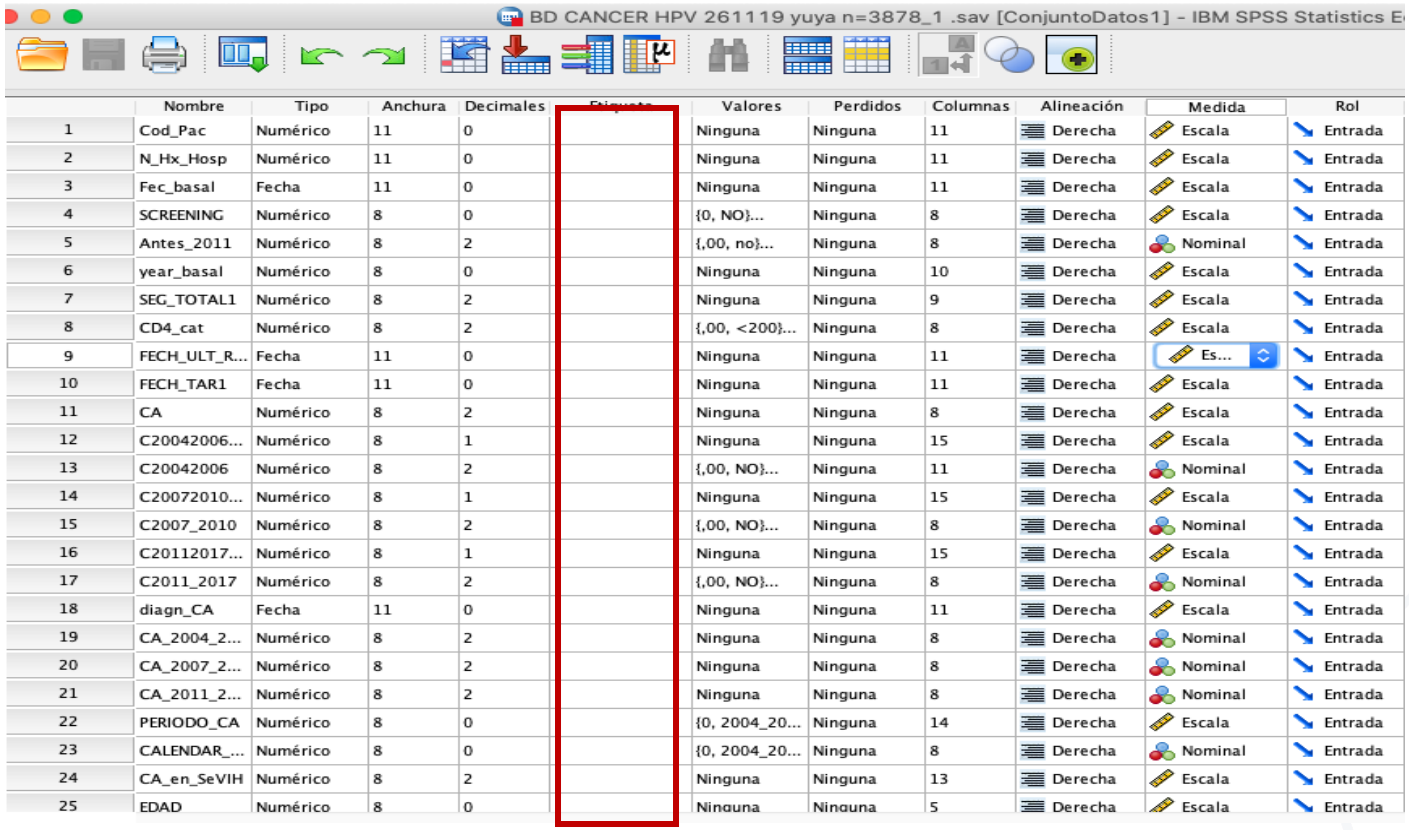
A1 | fx | Variables

	A	B	C	D	E	F	G
1	**Variables**	**Código de la variable**	**Grupo de la variable**	**Definición conceptual**	**Tipo de variable**	**Naturaleza de la variable**	**Definición operativa**
2	Género	Género	Demográficos	Identidad de género	Independiente	Cualitativa politómica	Mujer= 1; Hombre= 2, Otros= 3, Prefiero no decirlo= 4
3	Edad	Edad	Demográficos	Edad del paciente en el momento de realización del estudio.	Independiente	Cuantitativa continua	Edad (años)
4	Nivel de estudio	Estudios	Demográficos	Nivel máximo de estudios alcanzados	Independiente	Cualitativa politómica	Sin estudios= 0; Primaria= 1; Secundaria= 2; Bachillerato o Grado profesional= 3; Universitario= 4
5	Nivel socio-económico	SocioEcon	Demográficos	Nivel socio-económico (Renta Familiar)	Independiente	Cualitativa politómica	<12.000€=0; 12001-24000€=1; 24.001-36.000=2; 36.001-50.000=3; >50.000€=4
6	Pregunta MHA 1	MHA_1	Modificaciones en los hábitos alimentarios	1.- ¿Ha hecho usted un cambio restrictivo en su dieta debido a su intolerancia alimentaria?	Independiente	Cualitativa politómica	No he cambiado mi dieta nunca= 0; Lo hice después de hacerme la prueba=1; Lo hice incluso antes de hacerme la prueba= 2
7	Pregunta MHA 2	MHA_2	Modificaciones en los hábitos alimentarios	2.- ¿Ha hecho usted los cambios en la dieta siguiendo las pautas de algún médico o nutricionista?	Independiente	Cualitativa politómica	No he cambiado mi dieta nunca= 0; He hecho los cambios que me ha indicado mi médico= 1; He hecho los cambios que me ha indicado mi especialista= 2; He hecho lo cambios que me ha indicado mi nutricionista= 3
8	Pregunta MHA 3	MHA_3	Modificaciones en los hábitos alimentarios	3.- Sí ha hecho usted algún cambio en la dieta, ¿lo mantiene en el tiempo?	Independiente	Cualitativa politómica	No he cambiado mi dieta nunca= 0; Con frecuencia consumo alimentos que me sientan mal= 1; Puntualmente consumo alimentos que me sientan mal= 2; Sigo mi dieta a diario y evito los alimentos que me sientan mal= 3
9	Pregunta MHA 4	MHA_4	Modificaciones en los hábitos alimentarios	4.- ¿Qué motivos principales cree que influyen en no poder cambiar su dieta para evitar alimentos no recomendados? (puede marcar varias respuestas)	Independiente	Cualitativa politómica	Los alimento recomemdados son más caros= 1; El sabor de los alimentos adaptados es diferente al que estoy acostumbrado= 2; Si como fuera de casa tengo menos opciones para elegir un plato que me guste= 3; Es difícil de encontrar una oferta variada en mis tiendas habituales= 4; No considero que sea necesario cambiar mi alimentación= 5
10	Cumplimiento de MHA	Cumplimiento	Modificaciones en los hábitos alimentarios	Adherencia al cambio dietético recomendado. Se calcula mediante la suma de las preguntas MHA_1 a MHA_3, se considera incumplidor si alguna de las	Independiente	Cualitativa dicotómica	Cumplidor= 0; Incumplidor= 1

RecogidaDatos | Diccionario de variables | Explorar

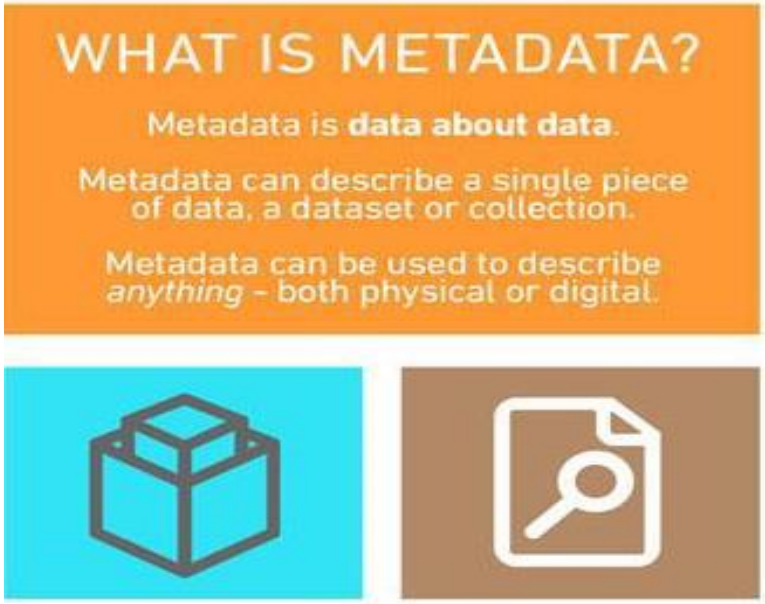
Documentación

Base de datos_Codificación



	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	Cod_Pac	Numérico	11	0		Ninguna	Ninguna	11	Derecha	Escala	Entrada
2	N_Hx_Hosp	Numérico	11	0		Ninguna	Ninguna	11	Derecha	Escala	Entrada
3	Fec_basal	Fecha	11	0		Ninguna	Ninguna	11	Derecha	Escala	Entrada
4	SCREENING	Numérico	8	0		{0, NO}...	Ninguna	8	Derecha	Escala	Entrada
5	Antes_2011	Numérico	8	2		{,00, no}...	Ninguna	8	Derecha	Nominal	Entrada
6	year_basal	Numérico	8	0		Ninguna	Ninguna	10	Derecha	Escala	Entrada
7	SEG_TOTAL1	Numérico	8	2		Ninguna	Ninguna	9	Derecha	Escala	Entrada
8	CD4_cat	Numérico	8	2		{,00, <200}...	Ninguna	8	Derecha	Escala	Entrada
9	FECH_ULT_R...	Fecha	11	0		Ninguna	Ninguna	11	Derecha	Es...	Entrada
10	FECH_TAR1	Fecha	11	0		Ninguna	Ninguna	11	Derecha	Escala	Entrada
11	CA	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
12	C20042006...	Numérico	8	1		Ninguna	Ninguna	15	Derecha	Escala	Entrada
13	C20042006	Numérico	8	2		{,00, NO}...	Ninguna	11	Derecha	Nominal	Entrada
14	C20072010...	Numérico	8	1		Ninguna	Ninguna	15	Derecha	Escala	Entrada
15	C2007_2010	Numérico	8	2		{,00, NO}...	Ninguna	8	Derecha	Nominal	Entrada
16	C20112017...	Numérico	8	1		Ninguna	Ninguna	15	Derecha	Escala	Entrada
17	C2011_2017	Numérico	8	2		{,00, NO}...	Ninguna	8	Derecha	Nominal	Entrada
18	diagn_CA	Fecha	11	0		Ninguna	Ninguna	11	Derecha	Escala	Entrada
19	CA_2004_2...	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Nominal	Entrada
20	CA_2007_2...	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Nominal	Entrada
21	CA_2011_2...	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Nominal	Entrada
22	PERIODO_CA	Numérico	8	0		{0, 2004_20...	Ninguna	14	Derecha	Escala	Entrada
23	CALENDAR_...	Numérico	8	0		{0, 2004_20...	Ninguna	8	Derecha	Nominal	Entrada
24	CA_en_SeVIH	Numérico	8	2		Ninguna	Ninguna	13	Derecha	Escala	Entrada
25	EDAD	Numérico	8	0		Ninguna	Ninguna	5	Derecha	Escala	Entrada

Metadatos



Estándares
http://www.dcc.ac.uk/resources/metadata-standards

Proporcionan información sobre el origen de un conjunto de datos completo.

- **Título:**
- **Autor / investigador principal**:
- **Colaboradores (contributor):**
- **Identificador:**
- **Tipo de datos:**
- **Derechos:**
- **Fechas:**
- **Idioma:**
- **Lugar:**
- **Resumen de contenido y palabras**
- **Investigación**.
- **Relaciones:**

Checklist _ CALIDAD_DOCUMENTACION

- Ejemplos: The *National Survey of Sexual Attitudes and Lifestyles, 2010-2012*

National Survey of Sexual Attitudes and Lifestyles, 2010-2012: Open Access Teaching Dataset

Details | Documentation | Resources | Access data

Details

Title:	National Survey of Sexual Attitudes and Lifestyles, 2010-2012: Open Access Teaching Dataset
Alternative title:	Natsal-3
Study number (SN):	8786
Access:	These data are open
Persistent identifier (DOI):	10.5255/UKDA-SN-8786-1
Principal investigator(s):	University of Manchester, Cathie Marsh Institute for Social Research (CMIST), UK Data Service

Documentation

Title	File name	Size (MB)
Open Teaching Dataset - Codebook	8786_natsal_open_codebook_feb21.pdf	0.35
Open Teaching Dataset - User Guide	8786_natsal_open_user_guide_feb21.pdf	0.41
Questionnaire from NATSAL 2010-2012 Study (SN7799)	8786_natsal_3_questionnaire.pdf	0.56
UK Data Archive Citation File for Study 8786	UKDA_Study_8786_Information.htm	0
UK Data Archive Data Dictionaries	ukda_data_dictionaries.zip	0.01
UK Data Archive ReadMe File for Study 8786	read8786.htm	0

https://ukdataservice.ac.uk/news-and-events/newsitem/?id=5782

Checklist _ CALIDAD_DOCUMENTACION

- La documentación coincide con los archivos
- El conjunto de datos y su contenido se describen claramente
- La información geoespacial y temporal está completa y se describe
- Las variables y unidades siguen estándares o están bien definidas
- Se proporciona publicación o manuscrito que describe los datos
- Metodología, calibraciones y algoritmos proporcionados.
- Problemas / limitaciones conocidos claramente descritos
- Las citas están debidamente referenciadas

Temporalidad _ Datos

Fecha y hora (Calendario, unidades de tiempo y extensión temporal, resolución y límite acorde a los estándares)

Recomendado ISO standard date formats: yyyy-mm-dd or yyyymmdd

En caso de usar otro, debe estar codificado.

	A	B	C	D	E	F	G	H	I
1	ccaa_iso	fecha	num_casos	num_casos_prueba	num_casos_pru	num_casos_prueb	num_casos_prueba_elisa	num_casos_prueba_descon	
2	AN	01/01/2020	0	0	0	0	0	0	
3	AR	01/01/2020	0	0	0	0	0	0	
4	AS	01/01/2020	0	0	0	0	0	0	
5	CB	01/01/2020	0	0	0	0	0	0	
6	CE	01/01/2020	1	0	0	1	0	0	
7	CL	01/01/2020	0	0	0	0	0	0	
8	CM	01/01/2020	0	0	0	0	0	0	
9	CN	01/01/2020	0	0	0	0	0	0	
10	CT	01/01/2020	3	3	0	0	0	0	
11	EX	01/01/2020	0	0	0	0	0	0	
12	GA	01/01/2020	0	0	0	0	0	0	
13	IB	01/01/2020	0	0	0	0	0	0	

https://datos.gob.es/es/catalogo/e05070101-evolucion-de-enfermedad-por-el-coronavirus-covid-19

Valores de las variables

Checklist _ CALIDAD_Valores de las variables

1. Comprueba el rango de los valores de cada parámetro/variable.

2. Visualiza (plot, map, or both)

3.Estadística descriptiva

4. Utiliza códigos para los valores ausentes (Missing Values _ N/A)

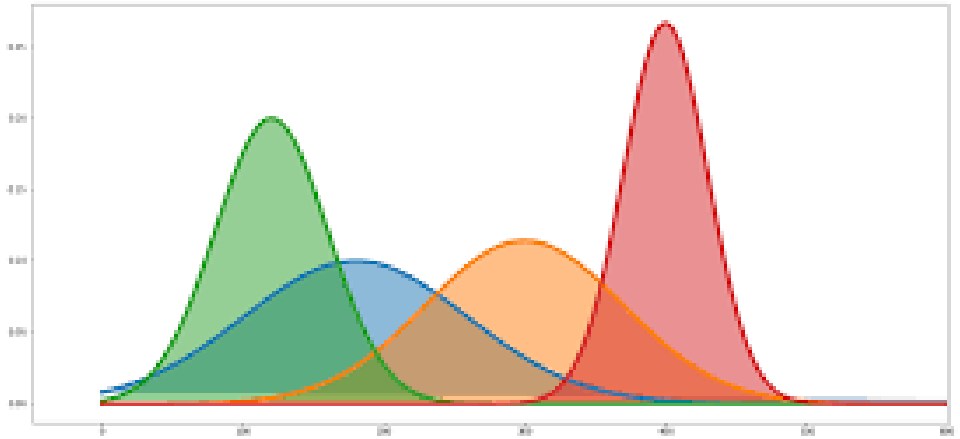
5. Valores para los campos codificados definidos.

6. ¿Necesitan transformación tus datos ?

CALIDAD_Valores de las variables

Comprueba el rango de los valores de cada parámetro/variable.

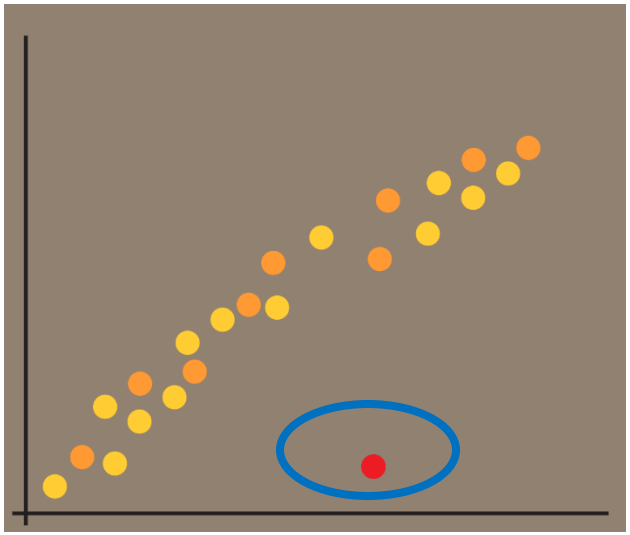
- ¿Valores anómalos?
- Comprueba Máx. y Min.
- ¿Cómo se distribuyen ?



CALIDAD_Valores de las variables

Visualiza (plot, map, or both)

- Relevancia de la visualización exploratoria



Explorar

Plots, trends, timelines, etc.

Analizar

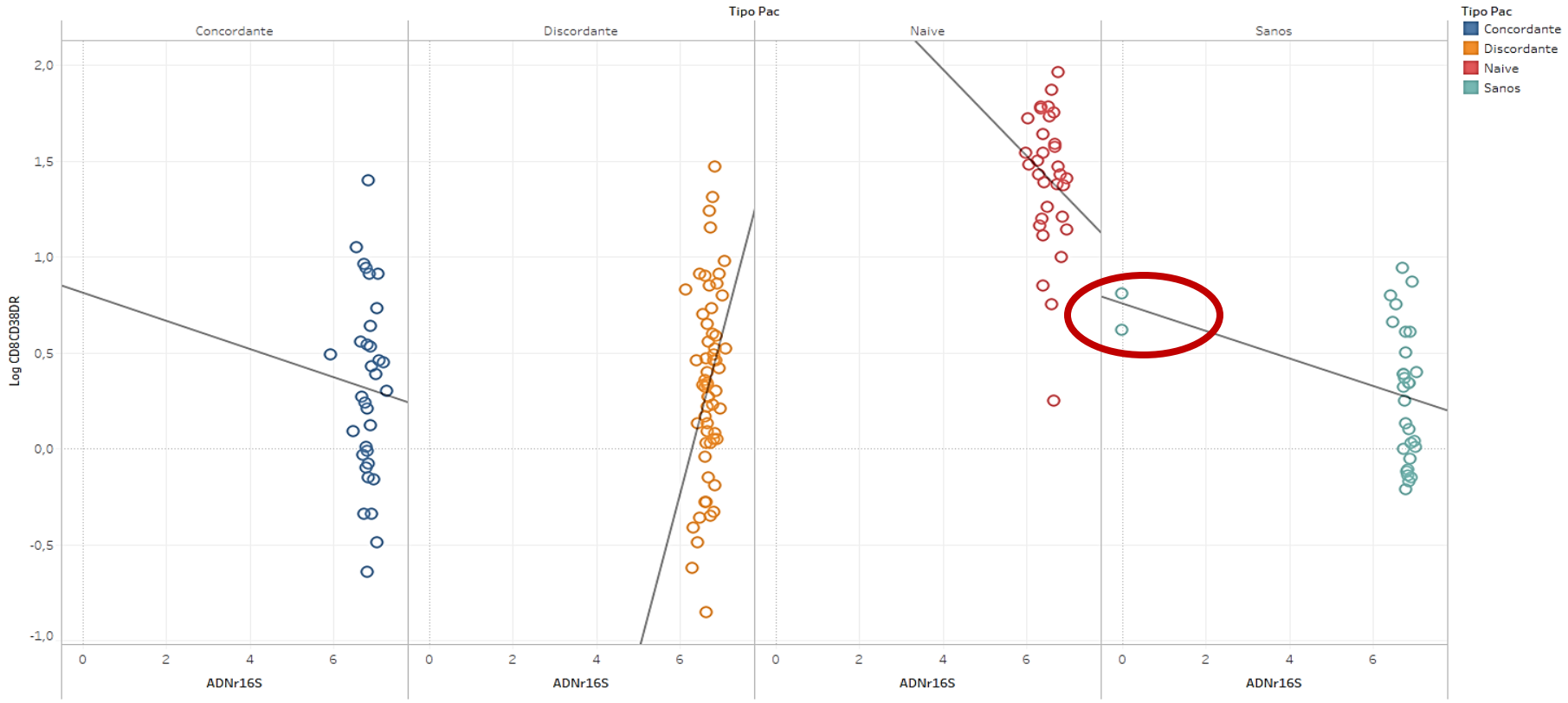
Desarrollar y testar hipótesis

Descubrir errores en los datos, anomalías.

Encontrar patrones

CALIDAD_Valores de las variables

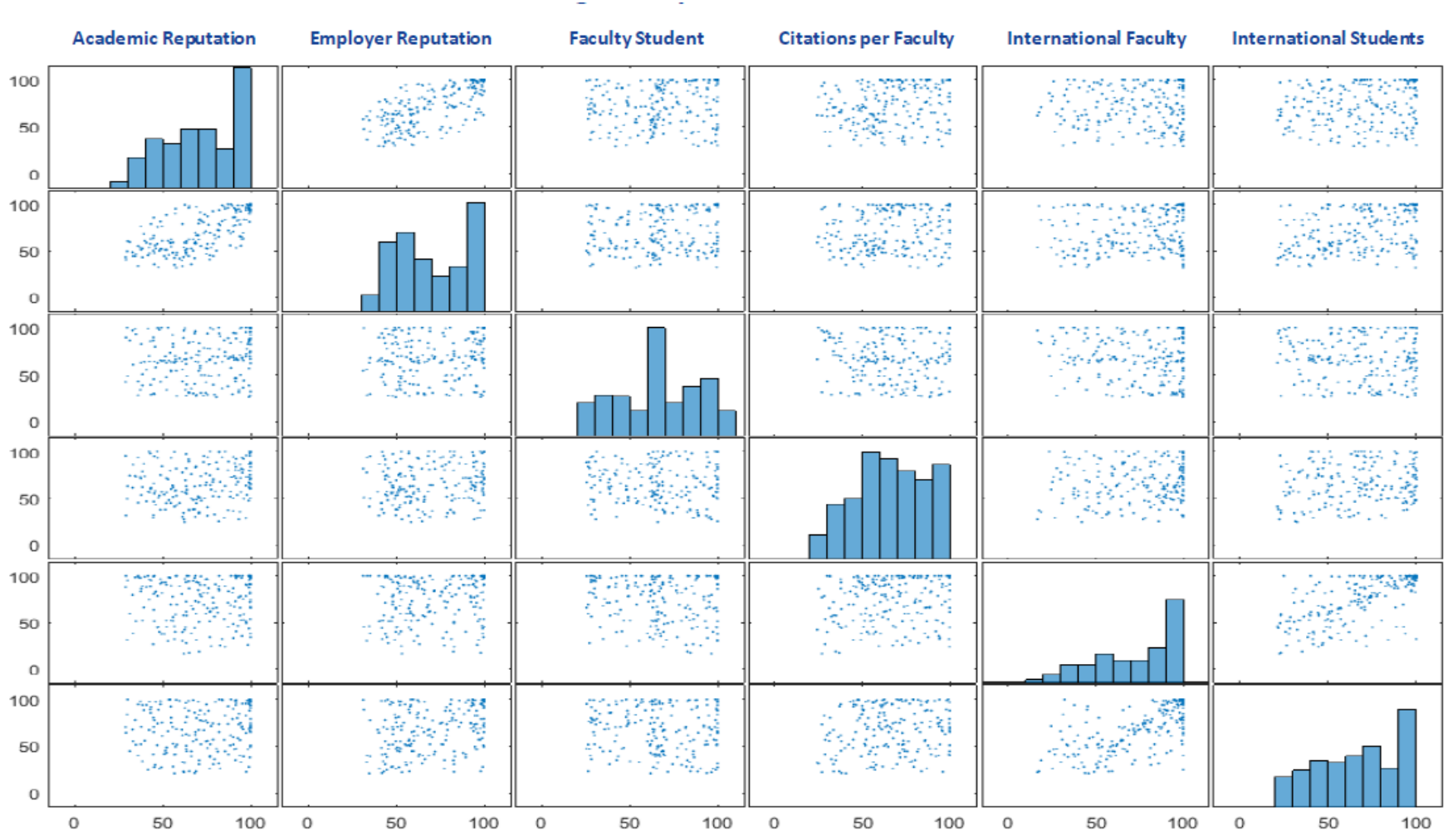
CD8CD38DR_ADN16S



ADNr16S vs. Log CD8CD38DR desglosado por Tipo Pac. El color muestra detalles acerca de Tipo Pac.

Paneles		Línea		Coeficientes				
Fila	Columna	valor p	DF	Término	Valor	StdErr	valor t	valor p
Log CD8CD38DR	Concordante	0,845093	31	ADNr16S	-0,0734909	0,373002	-0,197026	0,845093
Log CD8CD38DR	Discordante	0,015575	57	ADNr16S	0,829366	0,332609	2,49352	**0,015575**
Log CD8CD38DR	Naive	0,367706	30	ADNr16S	-0,225215	0,24625	-0,914578	0,367706
Log CD8CD38DR	Sanos	0,0541853	29	ADNr16S	-0,0717642	0,0357623	-2,0067	**0,0541853**

CALIDAD_Valores de las variables



CALIDAD_Valores de las variables

Herramienta para la visualización exploratoria interactiva

https://public.tableau.com/s/.

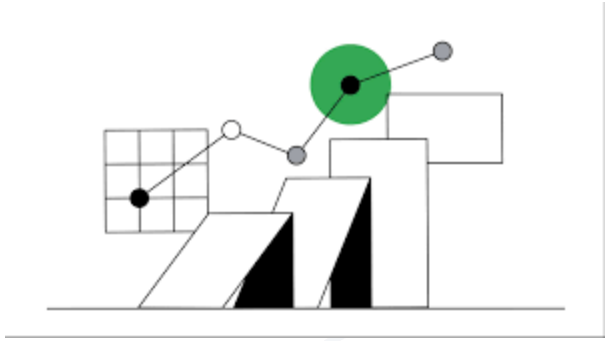
CALIDAD_Valores de las variables

La **visualización efectiva de los datos** puede dar lugar a la necesidad de transformar los datos.

Al preparar los datos para la visualización, surgen preguntas relacionadas a la escala y granularidad.

Por ejemplo:

¿Debería un gráfico de líneas tener ocurrencias diarias a lo largo del eje Y, o ser suavizado (promediado) para mostrar puntos por semana o mes?



La respuesta depende de lo que vale la pena mostrar en los datos.

Transformar y normalizar datos

- **Necesidad de transformar datos**
- **Normalización.**

http://dam.ukdataservice.ac.uk/dataskills/longitudinaldata/4/story_html5.html

Transformar y normalizar datos

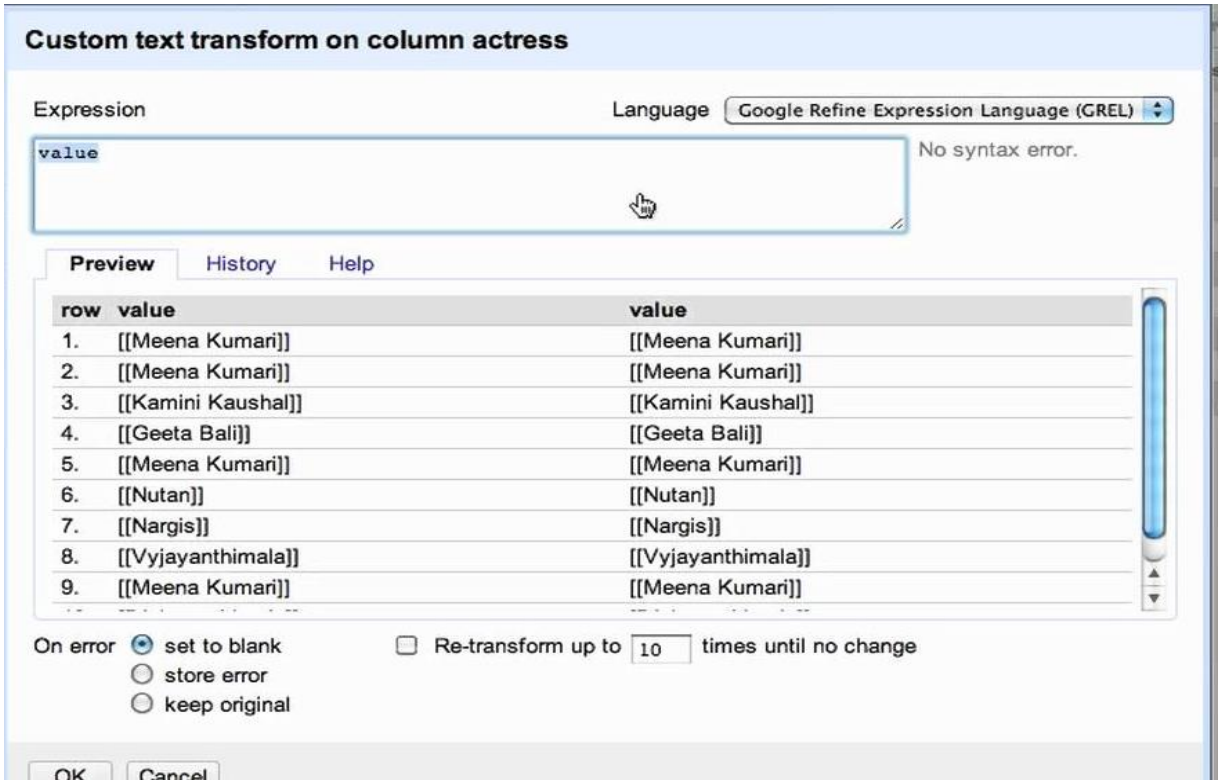
•**Normalización estadística:** usar una fórmula o un algoritmo para transformar las variables medidas en diferentes escalas en una escala común para que puedan ser comparables (manzanas con manzanas) o analizadas en un modelo estadístico elegido;

•**Normalización de bases de datos:** eliminar la duplicación e inconsistencia:

- •Por ejemplo, **dividir las tablas grandes en grupos más pequeños** y vincular los campos entre tablas a través de una clave o ID común.

Transformar y normalizar datos

- **Limpiar, Transformar datos (OpenRefine)**



https://openrefine.org/

Transformar y normalizar datos

- **Limpiar, Transformar datos (OpenRefine)**

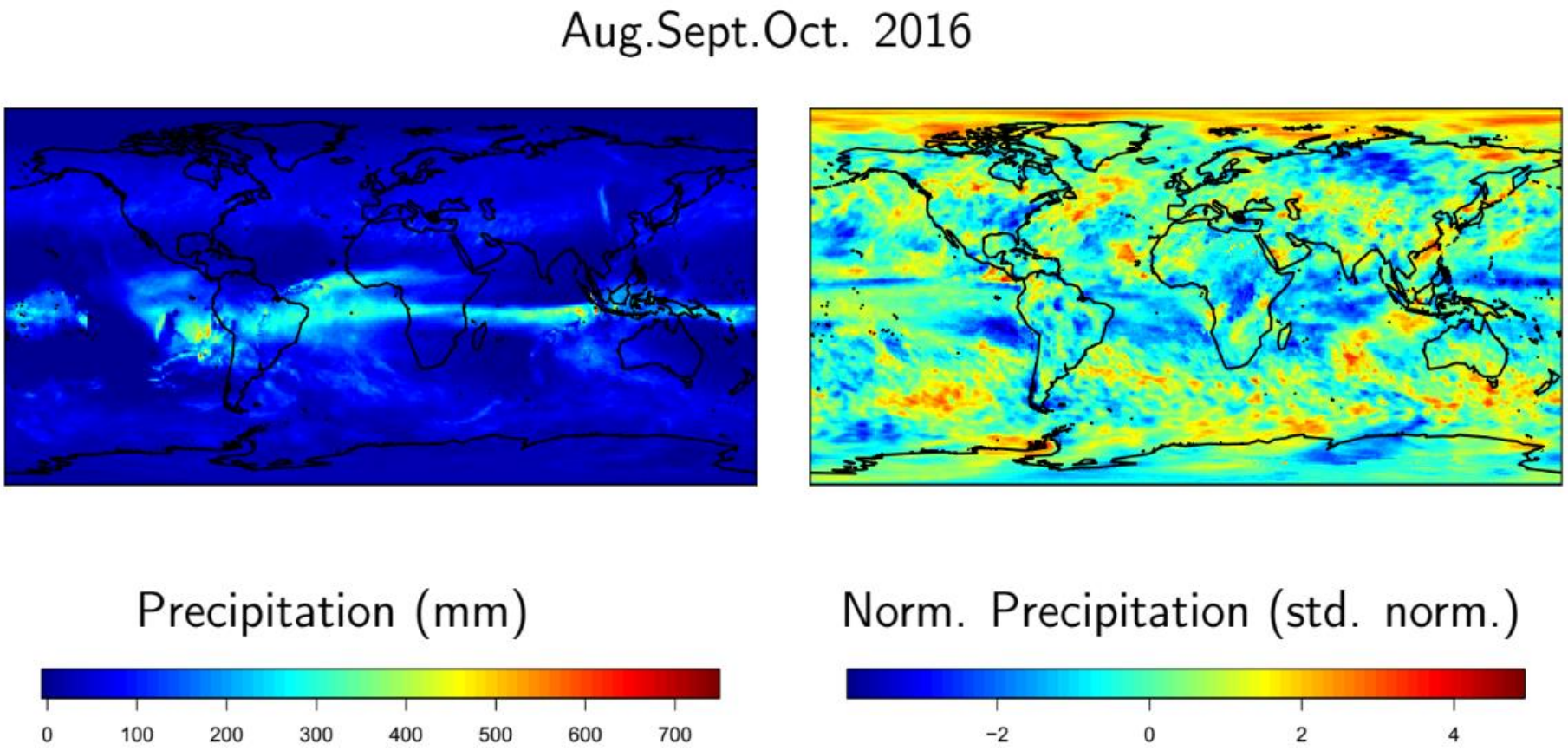
11285 rows Extensions: Zemanta ▾ Freebase ▾ RDF ▾ CK

Show as: **rows** records Show: 5 10 25 **50** rows « first ‹ previous **1 - 50** next ›

All		Capital or Reven	Directorate	Transaction Nur	Date	Service Area	Expenses Type	Amount	Supp
1.	Revenue	Community Wellbeing & Social Care	5105695746	05.04.2013	Youth & Community	Operational Equipment	120	REDACTE PERSONA
2.	Revenue	Community Wellbeing & Social Care	5105695746	05.04.2013	Youth & Community	Operational Equipment	80	REDACTE PERSONA
3.	Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments edit	695.89	REDACTE PERSONA
4.	Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments	695.89	REDACTE PERSONA
5.	Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments	695.89	REDACTE PERSONA
6.	Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments	695.89	REDACTE PERSONA
7.	Revenue	Community Wellbeing & Social Care	5105698650	24.04.2013	Leaseholds by LA	Accommodation Costs - Leaseholder Payments	695.89	REDACTE PERSONA
8.	Revenue	Chief Executive, Schools & Learning	5105698316	19.04.2013	L&A Commissioned Activity	Bought in Prof Services - Curriculum (Schools)	250	REDACTE PERSONA
9.	Revenue	Chief Executive, Schools & Learning	5105698318	19.04.2013	L&A Commissioned Activity	Bought in Prof Services - Curriculum (Schools)	710	REDACTE PERSONA
10.	Revenue	Economy & Environment	5105695879	05.04.2013	IW Biological Record	General Materials	220.2	REDACTE PERSONA
11.	Revenue	Chief Executive, Schools & Learning	5105696514	12.04.2013	Adult Services Training	Training and Conferences	150	REDACTE PERSONA
12.	Revenue	Community Wellbeing & Social Care	5105695832	10.04.2013	Short Breaks	Payments to Voluntary and Other Associations	1,260.00	REDACTE PERSONA
13.	Capital	Resources	5105696504	12.04.2013	Capital Receipts	External Design and Supervision Fees	400	REDACTE PERSONA
14.	Capital	Resources	5105696505	12.04.2013	Capital Receipts	External Design and Supervision Fees	1,350.00	REDACTE PERSONA
15.	Revenue	Economy & Environment	5105696707	12.04.2013	Schools Reorganisation	Security of Buildings	300	REDACTE PERSONA
16.	Revenue	Economy & Environment	5105696707	12.04.2013	Schools Reorganisation	Security of Buildings	300	REDACTE

https://openrefine.org/

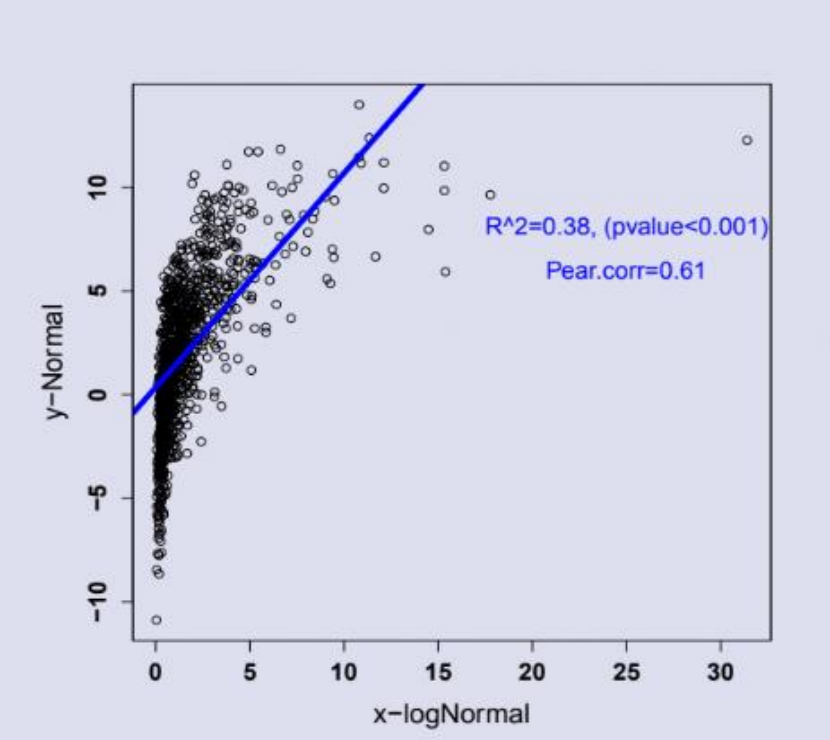
Transformar y normalizar datos

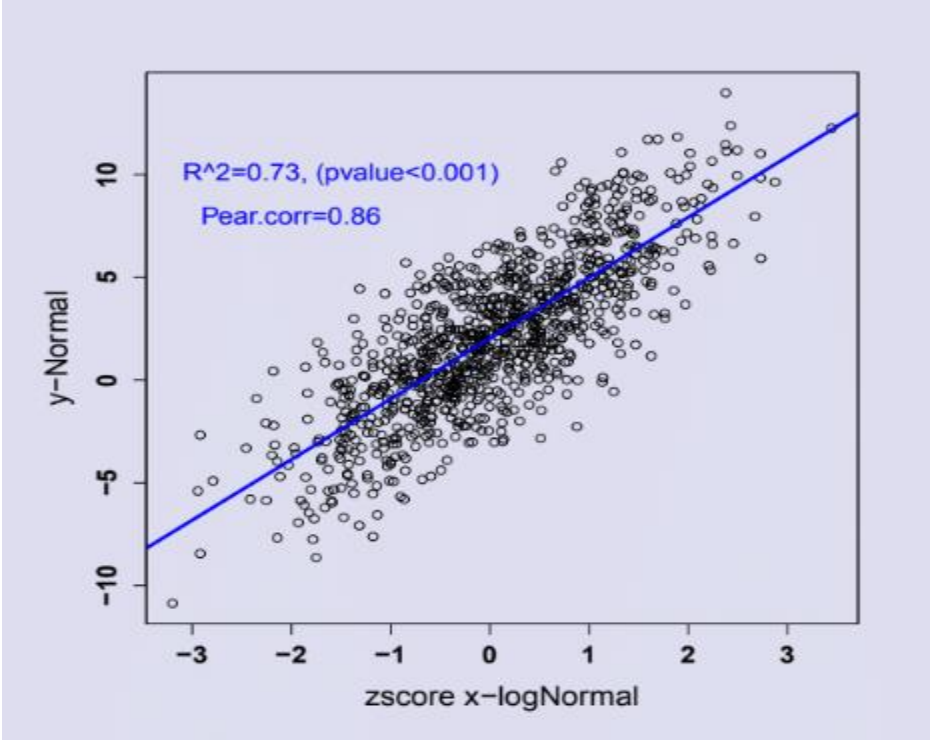


Métodos de normalización de los datos

- **Variables normales: transformaciones lineales:** (s-score, Min, max)
- **Variables no_normales**: Non_linear transformation (Quantile normalization, log transformation, box-cox, Johnson method)

Métodos de normalización de los datos





Missing values

Missing Values

- **Usa una notación consistente para los campos sin valores**
- **Utiliza un campo separado (para revisión de calidad)**
- **Tratamiento adecuado de los N/A**
- **Para campos numéricos, usar valor extremo como ej: -9999**
- **Para campos carácter, "NA"**

Missing Values

Don't know and refusal

When asked a question, survey participants may respond 'do not know' or refuse to answer. Such responses are recorded using specific codes, often something distinctive compared to other values such 99 or 999.

To help with analysis of this dataset, most 'not answered' and 'don't know' values are pre-set as missing values in the SPSS and Stata versions. The main exception is the sexual attitudes variables where the category labelled 'Depends/Don't Know' is kept as a valid value.

Ejemplo: No_Responses en diseño de encuestas

Tratamiento de los Missing Values

¿Por qué es importante ?

- **Reduce el poder estadístico de tus datos**
- **Genera estimación sesgada**
- **Reduce la representatividad de la muestra**

Ejemplo: No_Responses en diseño de encuestas

Tratamiento de los Missing Values

Pasos a considerar

1.¿Con qué tipo de datos estoy tratando? (variable, p. ej: categorical vs. continua; fuente, p. ej: datos estadísticos oficiales, encuestas, etc.)

2.Identificar patrones y recodificarlo correctamente

3.Cuál es la distribución?

4.Escoger el método adecuado para análisis y/o imputación

Value	Description
-9	Missing by error
-8	Not applicable to the respondent
-7	Proxy respondent not eligible for question
-2	Refused
-1	Don't know

Ejemplo: Missing values

STAGE 1: Skip pattern in survey

P. ej.: Community Innovation Survey 2014 –firms data

2.1 During the three years 2012 to 2014, did your enterprise introduce:

	Yes *1*	No *0*	
Goods innovations: New or significantly improved goods (exclude the simple resale of new goods and changes of a solely aesthetic nature)	☐	☐	*INPDGD*
Service innovations: New or significantly improved services	☐	☐	*INPDSV*

If no to all options, go to section 3
Otherwise, go to question 2.2

2.2 Who developed these product innovations?

Tick all that apply

	Goods innovations		Service innovations	
Your enterprise by itself	☐	*INITGD*	☐	*INITSV*
Your enterprise together with other enterprises or organisations*	☐	*INTCGD*	☐	*INTOSV*
Your enterprise by adapting or modifying goods or services originally developed by other enterprises or organisations*	☐	*INADGD*	☐	*INADSV*
Other enterprises or organisations	☐	*INOTHGD*	☐	*INOTHSV*

**: Include independent enterprises plus other parts of your enterprise group (subsidiaries, sister enterprises, head office, etc.). Organisations include universities, research institutes, non-profits, etc.*

Ejemplo: Missing values

Técnicas más utilizadas para tratar los Missing values

-Borrar: listwise, pairwise

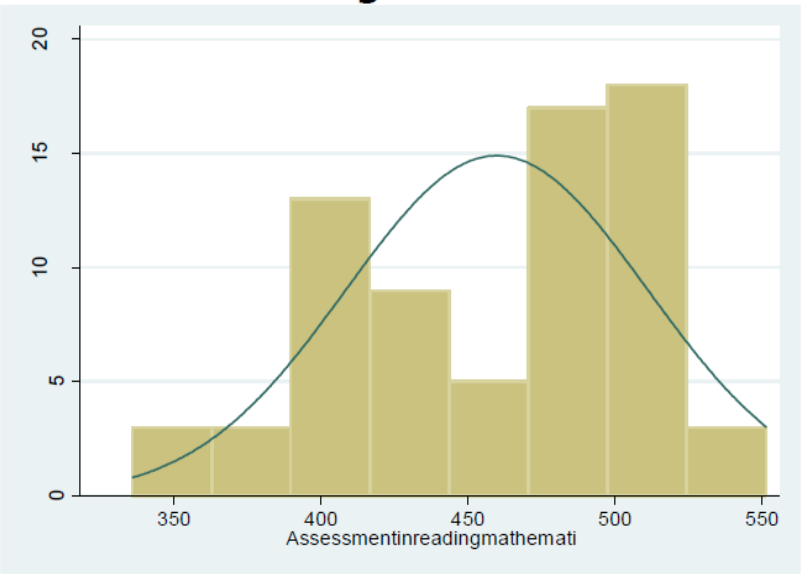
-Ignorar los Missing values

-Imputación simple: mean/median/mode substitution, hotdeck, single regression, etc.

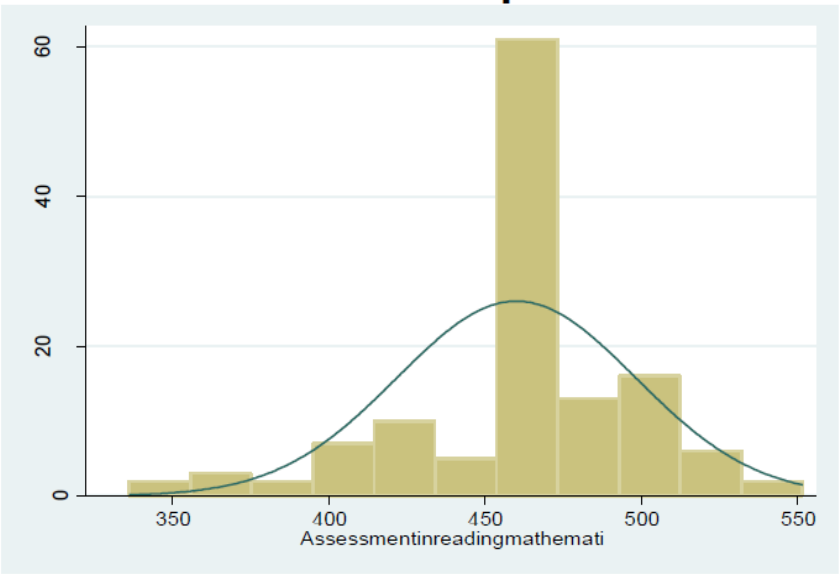
-Basado en modelos (Expectation Maximization Maximum Likelihood, Multiple Imputation)

Ejemplo: Missing values

With missing values



After mean imputation



Source: author elaboration, 2016 GII data

Advantage: simple and straightforward
Disadvantage: distorts distribution, attenuates variance=>biased estimates, it modifies relationships between variables

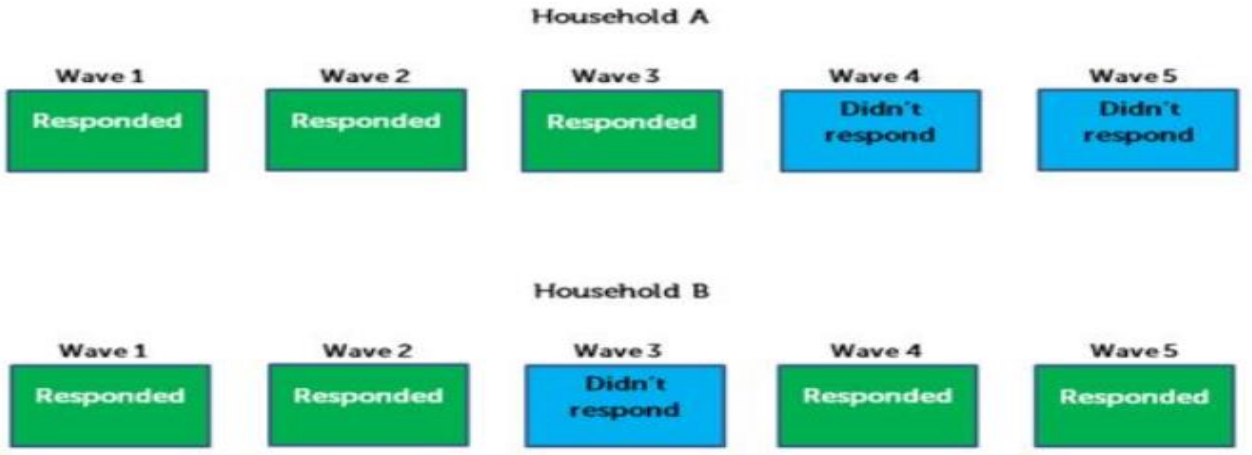
Ejemplo: No_Responses en diseño de encuestas

1. Dos tipos de **No-respuesta** en encuestas:

1.1 unit non-response *
1.2 item non_response **

*1.1 Ocurre cuando el possible encuestado no respondió la encuesta, ya sea por negarse o porque no pudo ser contactado.

* Por tanto, el encuestador, durante el tiempo de seguimiento, no podrá seguir el mismo número de sujetos



http://dam.ukdataservice.ac.uk/dataskills/longitudinaldata/4/story_html5.html

Ejemplo: No_Responses en diseño de encuestas

- **Incluir la información del número de encuestados y motivo como variable en el dataset**

- **Razones de No – respuesta**
- **Controlar los sujetos que abadonaron y los que se volvieron a incluir durante el seguimiento de la encuesta**

Frequencies

Value label	Value	Absolute frequency	Relative frequency
full interview	1	47732	61.74%
proxy interview	2	3262	4.22%
refusal	10	4155	5.37%
other non-intvw	11	3225	4.17%
ill/away during survey period	14	579	0.75%
too infirm/elderly	15	253	0.33%
language difficulties	16	230	0.30%
youth interview	21	4899	6.34%
youth: oth non-int	23	1728	2.24%
child under 10	24	11246	14.55%
	Total	77309	100.00%

http://dam.ukdataservice.ac.uk/dataskills/longitudinaldata/4/story_html5.html

Ejemplo: No_Responses en diseño de encuestas

- **Considerar la transformation de datos en función de un sistema de ponderación (weigthing)**

- **Grupos infra o sobre representados**
- **Necesidad de ajustarlo para evitar sesgos y falta de precisión en los resultados**

http://dam.ukdataservice.ac.uk/dataskills/longitudinaldata/4/story_html5.html

Ejemplo: No_Responses en diseño de encuestas

- **Incorporarlo en "The User Guide", apartado metodológico.**

Survey methodology variables

Survey weight

Many survey datasets contain variables called survey weights. These variables are made by the data collectors for you to apply when analysing data. We use weights to make sample data better represent the population it's designed to reflect by adjusting for over and under-represented cases. Under and over-representation can result from the complex sampling methods used in large scale surveys and other issues including non-response.

In this teaching dataset, the survey weight variable **Total_wt** is for use when analysing the total sample (including the boosts of younger people). The weight accounts for unequal selection probabilities from the sample design (selection weight) and for non-response by adjusting the distribution of age, sex and regional profiles to match the general population (poststratification weight).

For more information about the weight variable, see the Natsal-3 documentation.

http://dam.ukdataservice.ac.uk/dataskills/longitudinaldata/4/story_html5.html

Ejemplo: No_Responses en diseño de encuestas

1.2 item non_response **

- **Cuando el encuestado realize la encuesta pero no responde todas las preguntas (Missing or No-válidos)**
- **Codificar los valores:**

Value	Description
-9	Missing by error
-8	Not applicable to the respondent
-7	Proxy respondent not eligible for question
-2	Refused
-1	Don't know

http://dam.ukdataservice.ac.uk/dataskills/longitudinaldata/4/story_html5.html

Ejemplo: No_Responses en diseño de encuestas

1.2 item non_response **

- **Evitar confusions y asumir que ej: Código -8 (No applicable) es debido a errores en los datos**

- **La razón más común es que no todas las preguntas son respondidas por todos los encuestados.**

- **Solución a mano: Ir al cuestionario para esa pregunta**
- **en específico.**

Value	Description
-9	Missing by error
-8	Not applicable to the respondent
-7	Proxy respondent not eligible for question
-2	Refused
-1	Don't know

http://dam.ukdataservice.ac.uk/dataskills/longitudinaldata/4/story_html5.html

Ejemplo: No_Responses en diseño de encuestas

Transformaciones, otros ejemplos

- **Los datos cualitativos**, como transcripciones de entrevistas, **pueden ser transformados en datos cuantitativos** aplicando técnicas de codificación textual y de categorización
- **Variables cuantitativas** pueden ser convertidas en categóricas o nominales como la edad.
- Variables de muchos niveles de una encuesta, pueden ser transformadas a menos niveles, según necesidad del propio análisis.

http://dam.ukdataservice.ac.uk/dataskills/longitudinaldata/4/story_html5.html

Tratamiento de datos sensibles

- Los **datos personales** son aquellos relacionados con una persona viva, lo que permite que ésta sea identificada.

- **Los datos confidenciales o sensibles** son datos personales sobre: origen racial o étnico, opiniones políticas, creencias religiosas, membresía sindical, salud física y mental, vida sexual, delitos y procedimientos judiciales.

Anonimización y seudonimización (ver Coursera curso)

Será necesario **anonimizar o seudonimizar** datos para compartir dichos datos con investigadores y terceros sin comprometer la privacidad del usuario cuando:

- se quiera compartir o ceder datos a destinatarios con los que no se ha firmado un acuerdo de confidencialidad.
- se quiera publicar datos abiertamente.
- la reducción de la calidad de la información sea aceptable y no afecte al uso de los datos.

Una vez que los datos se anonimizan, no sería necesario el consentimiento. Si **las personas pudieran ser identificadas de alguna forma es contar con un formulario de consentimiento informado**, firmado por los participantes.

UK Data Service. *Format your data. "Create well organised and sustainable data"*
https://www.ukdataservice.ac.uk/manage-data/format/file-formats.aspx

Anonimización y seudonimización

El proceso de seudonimización y anonimización implica que los datos sobre los individuos se vean alterados a través de distintos procesos: **pueden ser suprimidos, sustituidos, distorsionados, generalizados o agregados**. Es importante que esto no afecte al posterior uso de la información.

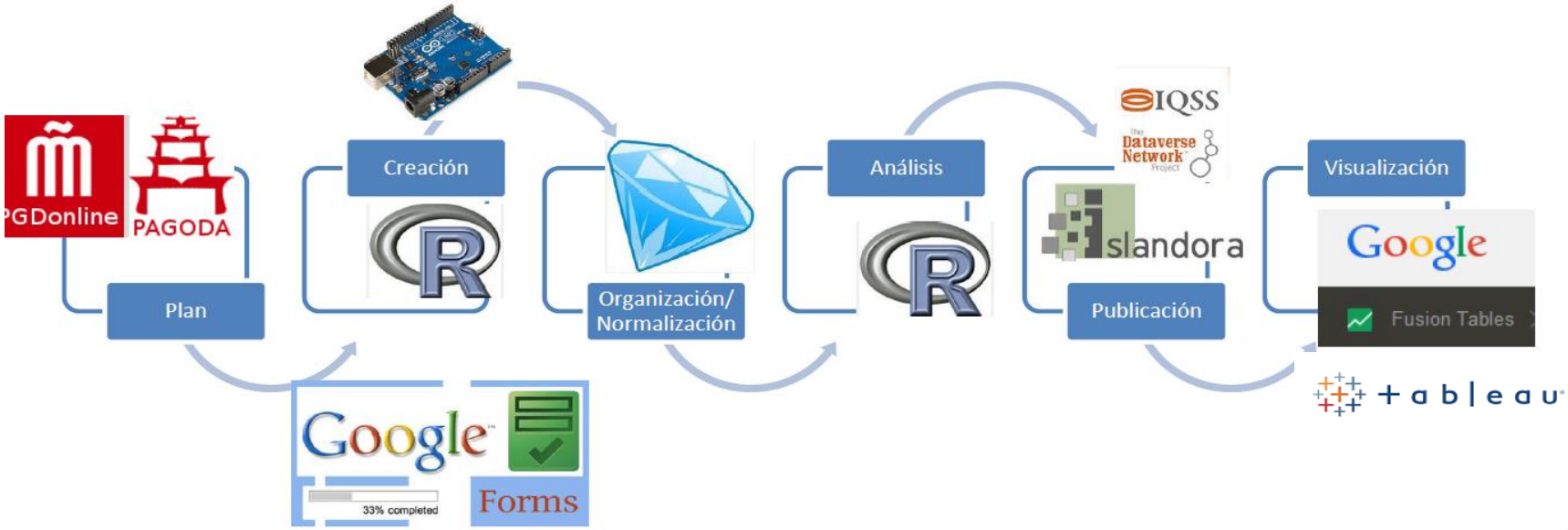
Ejemplos de eliminación y generalización de datos:

id	Zipcode	Age	National.	Disease
1	13053	28	Russian	Heart Disease
2	13068	29	American	Heart Disease
3	13068	21	Japanese	Viral Infection
4	13053	23	American	Viral Infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart Disease
7	14850	47	American	Viral Infection
8	14850	49	American	Viral Infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	Cancer

~~id~~	Zipcode	Age	National.	Disease
~~1~~	130**	<30	*	Heart Disease
~~2~~	130**	<30	*	Heart Disease
~~3~~	130**	<30	*	Viral Infection
~~4~~	130**	<30	*	Viral Infection
~~5~~	1485*	≥40	*	Cancer
~~6~~	1485*	≥40	*	Heart Disease
~~7~~	1485*	≥40	*	Viral Infection
~~8~~	1485*	≥40	*	Viral Infection
~~9~~	130**	3*	*	Cancer
~~10~~	130**	3*	*	Cancer
~~11~~	130**	3*	*	Cancer
~~12~~	130**	3*	*	Cancer

Alicia Fatima Gomez Sanchez, & Elli Papadopoulou. (2020, November). *OpenAIRE tools and resources: supporting research data management services for libraries and researchers*. http://doi.org/10.5281/zenodo.4317181

Herramientas en la Gestión de datos



Adaptado de: Luis Urbine . Taller de datos. [*14º Workshop de REBIUN de Proyectos Digitales: los horizontes de los repositorios (Universidad de Córdoba, 2015)*]

Check-list control calidad _ Resumen

- **Revisa la organización del contenido y los descriptores de tus archivos** para asegurarte que no faltan elementos clave.
- **Ordena los registros por parámetros clave** para resaltar posibles discrepancias .
- **Verifica la validez de los valores medidos.** Busca valores imposibles o outliers (Ejemplo: Un pH de 74; Una altura de 2.50, etc.)
- **Verifica el marco temporal de tus datos**. Genera gráficos exploratorios de series temporales para detectar valores anómalos o posibles lagunas en tus datos.

Check-list Control de calidad

- **Valores ausentes o Missing values en tu data.** No olvides codificarlos adecuadamente, e incluir los códigos en la documentación que acompaña a tus datos en el proyecto.
- **Comprueba el tipo de datos, escala, tamaño de las imágenes**, etc. Tipologías errores dan lugar a análisis fallidos.
- Como parte del propio Análisis Exploratorio de Datos (AED)**, revisa las estadísticas descriptiva de tu dataset** (media, mediana, cuartiles, valores mínimos y valores máximos)
- **Elimina todo parámetro variable que no aporte información relevante**

¿Preguntas?

Gracias

Yusnelkis Milanes Guisado

ymilgui@upo.es

UNIVERSIDAD

PABLO DE OLAVIDE

SEVILLA