Datasets and metadata: Find, visualize and explore

Sónia Teixeira and Pedro Campos
University of Porto, Portugal

Abstract:

The present document describes the main achievements of Output 4 – Datasets and Metadata. It contains a Preamble, where we introduce the main goal of Output 4. Then, we make an overview of existing data and describe the datasets used for the experiments with several data visualization tools. For each data set, we define the Name, Source, Year(s)/time scope, Unity of analysis, Number of registers, and Number of variables. In addition, information concerning the main variables and the corresponding types used for analysis is also provided, as well as possible cross tabulations and further comments.

Suggested Citation: Teixeira, S., and Campos, P. (2018). Data Sources. A product of the ProCivicStat Project. Retrieved (Date) from: http://IASE-web.org/ISLP/PCS (or www.procivicstat.org until Oct 2018

For more information, extensive teaching resources, supporting papers, datasets, contacts, and our Call for Action and Recommendations: See the ProCivicStat website under the International Association for Statistics Education (IASE) website here: http://iaseweb.org/islp/pcs. You can also visit our original website at www.procivicstat.org, though it will not be updated after Nov 2018.

Acknowledgments: The authors thank the ERASMUS+ program of the European Commission for its support of the *ProCivicStat* project. However, the opinions expressed in this document reflect the authors' own views and not necessarily those of the sponsoring agency or the EC.

[1] This project has been funded with support from the European Commission. This publication reflects the views only of the authors,

Preamble

Informed citizens need to access data in order be able to explore, understand, and reason about information of a multivariate nature. However, information is not always easily accessible. Educators, for example, often do not have time to locate information sources, and many of them are not in open data.

Therefore, one of the goals of PCS was to identify relevant multivariate data sets about social, economic, environmental and health issues. Datasets have been be prepared in formats that are accessible to customised data visualizations.

In this report, we describe the following deliverables expected in Output 4. We start with an overview of existing data (04-A.1) and then we present the datasets used for the experiments with several data visualization tools (04-A2). For each data set, we define the Name, Source, Year(s)/time scope, Unity of analysis, Number of registers, and Number of variables. In addition, information concerning the main variables and the corresponding types used for analysis is also provided, as well as possible cross tabulations and further comments.

Overview of existing data

The list of existing datasets that been identified to support teaching activities is in Appendix 1. Most part of the sources are open data. To supplement this list, we recommend searching via https://toolbox.google.com/datasetsearch

Datasets Description

Sónia Teixeira, Paula Lopes, Pedro Campos

University of Porto, Portugal

(August 2017)

The present document contains a description of the datasets used for the experiments with several data visualization tools. For each data set, we define the Name, Source, Year(s)/time scope, Unity of analysis, Number of registers, and Number of variables. In addition, information concerning the main variables and the corresponding types used for analysis, is also provided, as well as possible cross tabulations and further comments.

- **PISA 2012 dataset**

Name	PISA (Programme for International Student Assessment).
Source	OECD Available in: https://www.oecd.org/pisa/pisaproducts/pisa2012database-downloadabledata.htm
Year	2012
Unity of analysis	Country
Number of registers	65
Number of variables	2

Variables:

Sex	Sex	This is a macrodata dataset. The value represent the percentage of students at each proficiency level in mathematics, by gender.
Level	Level result on PISA test	

Main variables for analysis:

All variables: Sex, Level

Possible analysis (crosstabs/graphs, etc):

All variables (Ex. bar charts in *Smart Center* and Pie chart in *infogr.am*)

Comments:

This data sets is one of the several available from PISA. In this particular data set, results show the profiles of students' performance in Mathematics. The PISA data set has been used in *Smart Center* tool and in *infogr.am* tool (data should be prepared/preprocessed in advance).

- **PIAAC 2015 dataset**

Name	PIAAC (Programme for the International Assessment of Adult Competencies)
Source	OECD Available in: http://www.oecd.org/skills/piaac/publicdataandanalysis/#d.en.408927
Year	2015
Unity of analysis	Individuals (in Austria)
Number of registers	5130
Number of variables	1328

Variables (20 variables were selected from the 1328 variables available in the data set):

CNTRYID	Country Identifier	Qualitative	Nominal
GENDER_R	Gender	Qualitative	Nominal
AGEG5LFS	Age groups in 5-year intervals	Qualitative	Ordinal
B_Q01a	Education - Highest qualification - Level	Qualitative	Nominal
G_Q04	Skill use work - ICT - Experience with computer in job	Qualitative	Nominal
G_Q05a	Skill use work - ICT - Internet - How often - For mail	Qualitative	Ordinal
G_Q05c	Skill use work - ICT - Internet - How often - Work related info	Qualitative	Ordinal
G_Q05d	Skill use work - ICT - Internet - How often - Conduct transactions	Qualitative	Ordinal
G_Q05e	Skill use work - ICT - Computer - How often - Spreadsheets	Qualitative	Ordinal
G_Q05f	Skill use work - ICT - Computer - How often - Word	Qualitative	Ordinal
G_Q05g	Skill use work - ICT - Computer - How often - Programming language	Qualitative	Ordinal
G_Q05h	Skill use work - ICT - Computer - How often - Real-time discussions	Qualitative	Ordinal
H_Q04b	Skill use everyday life - ICT - Experience with computer everyday life	Qualitative	Nominal
H_Q05a	Skill use everyday life - ICT - Internet - How often - For mail	Qualitative	Ordinal
H_Q05c	Skill use everyday life - ICT - Internet - How often - In order to better understand various issues	Qualitative	Ordinal
H_Q05d	Skill use everyday life - ICT - Internet - How often - Conduct transactions	Qualitative	Ordinal
H_Q05e	Skill use everyday life - ICT - Computer - How often - Spreadsheets	Qualitative	Ordinal

H_Q05f	Skill use everyday life - ICT - Computer - How often - Word	Qualitative	Ordinal
H_Q05g	Skill use everyday life - ICT - Computer - How often - Programming language	Qualitative	Ordinal
H_Q05h	Skill use everyday life - ICT - Computer - How often - Real-time discussions	Qualitative	Ordinal

Main variables for analysis:

CNTRYID, GENDER_R, AGEG5LFS, B_Q01a and the variables that allow to make a comparative study to evaluate the computational capacity in the work and in the everyday life that are: G_Q04, G_Q05a, G_Q05c, G_Q05d, G_Q05e, G_Q05f, G_Q05g, G_Q05h, H_Q04b, H_Q05a, H_Q05c, H_Q05d, H_Q05e, H_Q05f, H_Q05g, H_Q05h).

Possible analysis (crosstabs/graphs, etc):

B_Q01a x GENDER_R (Ex. Make Barplot in iNZight or make contingency table in JMP Student Edition)
B_Q01a x GENDER_R x AGEG5LFS (Ex. Make table in JMP Student Edition or make barplot in INZight)
G_Q04 x H_Q04b (Ex. Use the Distribution function in Analyze in JMP Student or do barplot in INZight))
G_Q04 x H_Q04b x GENDER_R (Ex. Compare the G_Q04 x H_Q04b distribution by gender using the JMP Student Edition Analyze distribution function)
G_Q04 x H_Q04b x AGEG5LFS (Ex. Make bar graphs of G_Q04 by AGEG5LFS and H_Q04b by AGEG5LFS in JMP Student Edition)
BQ01a x G_Q04 x H_Q04b (Ex. Ex. Perform Chi-square association test that relates BQ01a x G_Q04 and BQ01a x H_Q04b in JMP Student Edition)
G_Q05d x H_Q05d x AGEG5LFS x GENDER_R (Ex. Make a comparative barplot between G_Q05d x AGEG5LFS x GENDER_R and H_Q05d x AGEG5LFS x GENDER_R)

Comments: Tools used to explore this data set: *JMP Student Edition and iNZight*.

- **Migration dataset - Nigeria**

Name	Migration Household Survey 2009 (Nigeria)
Source	World Data Bank Available in: http://microdata.worldbank.org/index.php/catalog/402
Year	2009
Unity of analysis	Household – Individuals
Number of registers	3344
Number of variables	82

Variables (10 variables were selected from the 82 variables available in the data set):

HHType	Type of migrant	Qualitative	Nominal
Sex	Sex	Qualitative	Nominal
Age	Age	Quantitative	Discrete
Reasonforleaving	Reason for leaving	Qualitative	Nominal
Howlong	How long has the individual lived in his/her current location?	Qualitative and Quantitative	
Sendmoney	Does the individual send any money to your household?	Qualitative	Nominal
Migrantgroup	Type of migrant group	Qualitative	Nominal
State	State of residence	Qualitative	Nominal
Educationgroup	Education Group	Qualitative	Ordinal
MaritalState	Marital State	Qualitative	Nominal

Main variables for analysis:

Age, Sex, Reasonforleaving, Migrantgroup, Educationgroup, MaritalState

Possible analysis (crosstabs/graphs, etc):

Age x Sex x Migrantgroup (Ex. Boxplot in *iNZight*)
Migrantgroup x Reasonforleaving x Educationgroup (Ex. Circle Packing in *Raw*)
Migrantgroup x Age x Reasonforleaving (Ex. “Points”, using “Graph Builder” [1] by *JMP*)
Migrantgroup x Sex x Reasonforleaving x Howlong (Ex: Scatter Plot in *Plotly*)
Age x Sex x Educationgroup x Reasonforleaving x SendMoney (Ex. Treemap in *Tableau*)

Comments:

Tools used to explore this data set: *Plotly, Raw, Tableau, JMP and iNZight*.

[1] For more information about graph Builder see: http://www.jmp.com/support/help/Overview_of_Graph_Builder.shtml#531604

- **Gender equity dataset:**

Name	GenderME
Source	IPUMS Available in: [https://international.ipums.org/international-action/ variables/group](https://international.ipums.org/international-action/variables/group)
Year	2000 and 2010
Unity of analysis	Household – Individuals
Number of registers	340189
Number of variables	15

Variables:

Country	Country (Brazil and USA)	Qualitative	Nominal
Year	Year	Qualitative	Ordinal
Sample	IPUMS sample identifier (id)	Qualitative	Nominal
Serial	Household serial number (id)	Qualitative	Nominal
PerNum	Person number	Qualitative	Nominal
PerWt	Person weight	Qualitative	Nominal
Age2	Age, grouped into intervals	Qualitative	Ordinal
Sex	Sex	Qualitative	Nominal
EmpStat	Activity status (employment status) [general version]	Qualitative	Nominal
EmpStatD	Activity status (employment status) [detailed version]	Qualitative	Nominal
OccISCO	Occupation, ISCO general	Qualitative	Nominal
IndGen	Industry, general recode	Qualitative	Nominal
ClassWk	Status in employment (class of worker) [general version]	Qualitative	Nominal
ClassWkD	Status in employment (class of worker) [detailed version]	Qualitative	Nominal
IncTot	Total income	Quantitative	Discrete

Main variables for analysis:
Age, Sex, IncTot, OccISCO, IndGen, EmpStat
Possible analysis (crosstabs/graphs, etc):
Sex x OccISCO x IncTot (Ex. Histograms in iNZight)
Sex x Age2 x IncTot (Ex. Histograms in iNZight)
IndGen x Sex x IncTot (Ex. Histograms in iNZight)

Comments:
GenderME data set has been tested with iNZight.

- **Natural disasters dataset**

Name	NaturalDisaster
Source	Center for Research on the Epidemiology of Disasters – Emergency Events Database (EM-DAT) Available in: http://www.emdat.be/advanced_search/index.html
Year	1900-2015
Unity of analysis	Region
Number of register	4254
Number of variable	10

Variables:

Year	Year	Qualitative	Ordinal
Region	Region	Qualitative	Nominal
Disaster Type	Disaster Type	Qualitative	Nominal
Occurrence	Occurrence	Quantitative	Discrete
Total of deaths	Total of deaths	Quantitative	Discrete
Affected	Affected	Quantitative	Discrete
Injured	Injured	Quantitative	Discrete
Homeless	Homeless	Quantitative	Discrete
Total affected	Total affected	Quantitative	Discrete
Total damage	Total damage (‘000$)	Quantitative	Discrete

Main variables for analysis:

Year, Region, Disaster Type, Occurrence, Total of deaths and Homeless

Possible analysis (crosstabs/graphs, etc):

Year x Total of deaths x Disaster type (Ex. Circle Packing in *Raw*)
Total of deaths x Disaster type (Boxplots ex. in *RAW, iNZight* or *SPSS*)
Year x Region x Disaster type (Histograms/ Bar charts ex. *iNZight / Tableau*)
Homeless x Disaster type x Region (Bar charts ex. *iNZight* or *Tableau*)
Region x Occurrence x Total of deaths (Ex. Side by side Bar Charts in *Tableau*)
Disaster Type x Region x Occurrence x Total of deaths(Ex. Circle Packing in *Raw*)

Comments:

Tools used with Natural Disasters data set: *Raw, iNZight* and *Tableau*.

- **Refugees dataset 1**

Name	UNdata_Export_20160810_152035594_Refugees
Source	United Nations - Data Retrieval System Available in: http://data.un.org/Data.aspx?d=UNHCR&f=indID%3AType-Ref
Year	2013
Unity of analysis	Country
Number of registers	80428
Number of variables	6

Variables:

CountryAsylum	Country or territory of asylum or residence	Qualitative	Nominal
CountryOrigin	Country or territory of origin	Qualitative	Nominal
Refugees	Refugees	Quantitative	Discrete
RefugeesA	Refugees assisted by UNHCR	Quantitative	Discrete
TotalRefugees	Total refugees and people in refugee	Quantitative	Discrete
TRefugeesUNHCR	al refugees and people in refugee, like situations assisted l UNHCR	Quantitative	Discrete

Main variables for analysis:
All

Possible analysis (crosstabs/graphs, etc):
CountryOrigin x Refugees (Ex. Scatter Plot in *Tableau*)
CountryAsylum x CountryOrigin x Refugees (Ex. Maps in *Power BI*)
Refugees x TotalRefugees x TRefugeesUNHCR x CountryAsylum (Ex. Maps in *Power BI*)

Comments:
The tools used to test the Refugees dataset was *Tableau* and *Power BI*.

- **Refugees dataset 2**

Name	
Source	UNHCR – The UN Refugee Agency Available in: http://popstats.unhcr.org/en/persons_of_concern
Year	2013, 2014 and 2015
Unity of analysis	Country
Number of registers	30349
Number of variables	14

Variables:

Year	Year	Qualitative	Ordinal
Country	Territory of asylum/residence	Qualitative	Nominal
Origin	Country of Origin	Qualitative	Nominal
TotalPersons	Total persons pending start-year	Quantitative	Discrete
Assisted	Of which UNHCR assisted start-year	Quantitative	Discrete
Rejected	Rejected	Quantitative	Discrete
ToralPerPendYear	Total persons pending end year	Quantitative	Discrete
AssistedEnd	Of which UNHCR assisted end year	Quantitative	Discrete

Main variables for analysis:
Country, Origin, Rejected, TotalPersons, ToralPerPendYear, Assisted, AssistedEnd and Year.

Possible analysis (crosstabs/graphs, etc):
Origin x ToralPerPendYear (Ex. Barplot in Tableau)
Origin x Rejected (EX. Circle chart in Tableau)
Origin x Rejected x Year (Ex. Circle chart in Tableau)
TotalPersons x ToralPerPendYear x Year (Ex. Superimposed Line chart in Tableau)
Origin x Assisted x Assisted (Ex. Barplot in Tableau)
ToralPerPendYear x Origin x Country (Ex. Circle chart in Tableau)

Comments: The tool used to test the Refugees data set was *Tableau*

- **Refugee dataset 3**

Name	Number of refugees by country of origin and destination
Source	UNHCR The UN Refugee Agency Available in: http://popstats.unhcr.org/en/time_series
Year	2015
Unity of analysis	Country
Number of registers	3600
Number of variables	5

Variables:

ID	Node ID / Identification number of country	Qualitative	Nominal
Label	Node label / Country name	Qualitative	Nominal
Source	Source node / Country of origin	Qualitative	Nominal
Target	Target node / Country of destination	Qualitative	Nominal
Weight	Edge weight / Number of people applied for refugee status	Quantitative	Discrete

Main variables for analysis:

All variables

Possible analysis (crosstabs/graphs, etc):

All variables

Comments:

This data set is edited to make it suitable for graph visualization and network analysis exercises. The dataset is divided into two .csv files that can be used as input in graph visualization softwares. The full dataset in its original form is available on the UNHCR website.

- **Malnutrition dataset 1**

Name	JME_December_2016
Source	UNICEF Available in: http://data.unicef.org/topic/nutrition/malnutrition/
Year	1983 – 2015
Unity of analysis	Country
Number of registers	789
Number of variables	18

Variables (9 variables were selected from the 18 variables available in the data set):

Country	Country and areas	Qualitative	Nominal
Year	Year	Qualitative	Ordinal
MillDevGoalsR	Region of Millenium Development Goals	Qualitative	Nominal
SevereWasting	Severe wasting	Quantitative	Continuous
Wasting	Wasting	Quantitative	Continuous
Overweight	Overweight	Quantitative	Continuous
Stunting	Stunting	Quantitative	Continuous
Underweight	Underweight	Quantitative	Continuous
UnderPop	Under 5 population (000s)	Quantitative	Continuous
IncGroup	Income group	Qualitative	Ordinal

Main variables for analysis:
Under 5 population (000s), MillDevGoalsR, Underweight, Year, Stunting.

Possible analysis (crosstabs/graphs, etc):
UnderPop x year (Ex. Line chart in Tableau)
Underweight x year x MillDevGoalsR (Ex. Barplot in Tablue)
Stunting x year x MillDevGoalsR (Ex. Line chart in Tablue)

Comments: Tools used to explore this data set: Tableau

- **Malnutrition dataset 2**

Name	Malnutrition_hanci
Source	HANCI Available in: http://www.hancindex.org/explore-the-data/
Year	2014
Unity of analysis	Country
Number of registers	45
Number of variables	46

Variables (8 variables were selected from the 46 variables available in the data set):

Africa	Africa	Qualitative	Nominal
agexpend	Government spending on agriculture	Quantitative	Continuous
healthexpend	Government spending on health	Quantitative	Continuous
birthreg	Civil registration of live births	Quantitative	Continuous
vitamina	Vitamin A coverage	Quantitative	Continuous
wateraccess	Access to drinking water	Quantitative	Continuous
sanitaccess	Access to sanitation	Quantitative	Continuous
constRSocialSec	Constitutional right to social security		

Main variables for analysis:

Africa, agexpend, healthexpend, wateraccess, sanitaccess, birthreg, constRSocialSec and vitamin.

Possible analysis (crosstabs/graphs, etc):

Africa (Ex. Barplot in iNZight)
Country x agexpend (Ex. Barplot Tableau)
Birthreg x Africa (Ex. Boxplot in iNZight)
Country x agexpend x healthexpend (Ex. Barplot Tableau)

Comments: Tools used to explore this data set: *Tableau and iNZight*.

- **Risk of Poverty or Social Exclusion dataset**

Name	Ilc_peps01
Source	eurostat Available in: http://appsso.eurostat.ec.europa.eu/nui/show.do
Year	2006 – 2015
Unity of analysis	Country
Number of registers	40
Number of variables	30

Variables:

Sex	Gender	This is a macrodata dataset. The value represent the rate of total risk and social exclusion by year, the rate of risk and social exclusion by sex per year, and the rate of risk and social exclusion by group of age per year
Age	Age by intervals	
Year	Year (2010 and 2014)	
TotalRisk	Risk of Poverty or Social Exclusion	

Main variables for analysis:

All

Possible analysis (crosstabs/graphs, etc):

Country x Total Risk
Sex x Year x Country
Country x Year x Age

Comments: Tools used to explore this data set: *Tableau.*

- **Sex and Race discrimination dataset**

Name	cddata_IWPR
Source	IWPR – Institute for Women's Policy Research Available in: http://www.iwpr.org/publications/resources/consent-decree
Year	2000 - 2008
Unity of analysis	Company (USA)
Number of registers	510
Number of variables	90

Variables:

State	State Filed	Qualitative	Nominal
Year	Year	Qualitative	Ordinal
SexDisc	Filed as Sex Discrimination	Qualitative	Binary
RaceDisc	Filed as Race Discrimination	Qualitative	Binary
NOriginDisc	Filed National Origin Discrimination	Qualitative	Binary
AgeDisc	Additionally filed as Age Discrimination	Qualitative	Binary
ReligiousDisc	Additionally filed as Religious Discrimination	Qualitative	Binary
DisabDisc	Additionally filed as Disability Discrimination	Qualitative	Binary
RaceSexDisc	Filed as Race and Sex	Qualitative	Nominal
SexHar	Sexual Harassment	Qualitative	Binary
Pay	Pay	Qualitative	Binary
Promotion	Promotion	Qualitative	Binary
Hiring	Hiring	Qualitative	Binary
Retaliation	Retaliation	Qualitative	Binary
Pregnancy	Pregnancy	Qualitative	Binary

Main variables for analysis:
All
Possible analysis (crosstabs/graphs, etc):
Case x RaceSexDisc
Case x SexDisc
Year x ReligiousDisc (Ex: lines)
Year x NOriginDisc (Ex: lines)
State x Retaliation x Hiring x Promotion (Ex: Comparative chart)

Comments: Tools used to explore this data set: *Tableau*

- **Child Health dataset – Respiratory Infection**

Name	CoD_ARI_Dec-2015_WHO_MCEE_236
Source	WHO and Maternal and Child Epidemiology Estimation Group (MCEE) Available in: http://apps.who.int/gho/data/node.main.ChildMort?lang=en
Year	2015
Unity of analysis	Country
Number of registers	194
Number of variables	14

Variables:

nnd	Total Neonatal deaths (estimated)	Quantitative	Continuous
pnd	Total Post-Neonatal deaths (estimated)	Quantitative	Continuous
Rneo9	Neonatal death rate from Acute Respiratory Infection (per 1000 livebirths)	Quantitative	Continuous
Rpost9	Postneonatal death rate from Acute Respiratory Infection (per 1000 livebirths)	Quantitative	Continuous
Rufive9	Underfive death rate from Acute Respiratory Infection (per 1000 livebirths)	Quantitative	Continuous

Main variables for analysis:
All

Possible analysis (crosstabs/graphs, etc):
Country x nnd
Country x rufive9 x rpost9 (Ex: comparative chart)

Comments: Tools used to explore this data set: *Tableau*

- **Crime dataset - Sweden**

Name	Swedish crime
Source	Kaggle Available in: https://www.kaggle.com/mguzmann/swedishcrime
Year	1950 - 2015
Unity of analysis	Year
Number of registers	66
Number of variables	21

Variables:

Year	Year	Qualitative	Ordina
crimes.total	total number of reported crimes	Quantitative	Discrete
crimes.penal.code	total number of reported crimes against the criminal code	Quantitative	Discrete
crimes.person	total number of reported crimes against a person	Quantitative	Discrete
murder	total number of reported murder	Quantitative	Discrete
sexual.offences	total number of reported sexual offences	Quantitative	Discrete
rape	total number of reported rapes	Quantitative	Discrete
assault	total number of reported aggravated assaults	Quantitative	Discrete
stealing.general	total number of reported crimes involving stealing or robbery	Quantitative	Discrete
fraud	total number of reported frauds	Quantitative	Discrete
narcotics	total number of reported narcotics abuses	Quantitative	Discrete
drunk.driving	total number of reported drunk driving incidents	Quantitative	Discrete
population	the total estimated population of Sweden at the time	Quantitative	Discrete

Main variables for analysis:
All

Possible analysis (crosstabs/graphs, etc):
crimes.person x murder
crimes.total x fraud x Year
crimes.penal.code x drunk.driving x Year

Comments: Tools used to explore this data set: *iNZight*

- **Conflict dataset**

Name	ACLED-All-Africa
Source	*ACLED – Armed Conflict Location and Event Data Project* Available in: *https://www.prio.org/Data/Armed-Conflict/UCDP-PRIO/*
Year	2016-2017
Unity of analysis	Conflict
Number of registers	1023
Number of variables	25

Variables: (7 variables were selected from the 25 variables)

Year	The year in which an event took place	Qualitative	Ordinal
Event_Type	The type of conflict event	Qualitative	Nominal
Actor1	The named actor involved in the event	Qualitative	Nominal
Actor2	The named actor involved in the event	Qualitative	Nominal
Country	The country in which the event took place	Qualitative	Nominal
Fatalities	The number of reported fatalities which occurred during the event	Quantitative	Discrete
Interactions	Number of interactions	Quantitative	Discrete

Main variables for analysis:
All

Possible analysis (crosstabs/graphs, etc):
Year x Event_Type
Country x Fatalities
Interactions x Event_Type
Actor1 x Event_Type
Actor2 x Event_Type
Event_Type x Fatalitiesx Country

Comments: Tools used to explore this data set: *iNZight* and *Tableau*

- **Student Alcohol Consumption - Portugal**

Name	Student (Dataset for the secondary school students of the Portuguese course)
Source	UCI – Machine Learning Repository Available in: https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION
Year	2008
Unity of analysis	School
Number of register	649
Number of variable	33

Variables:

school	student's school	Qualitative	Nominal
sex	student's sex	Qualitative	Nominal
age	student's age (from 15 to 22)	Quantitative	Discrete
address	student's home address type Rural or Urban	Qualitative	Nominal
Pstatus	parent's cohabitation status	Qualitative	Nominal
higher	wants to take higher education	Qualitative	Nominal
Dalc	workday alcohol consumption	Qualitative	Ordinal
Walc	weekend alcohol consumption	Qualitative	Ordinal
Medu	mother's education	Qualitative	Nominal
famsup	family educational support	Quantitative	Discrete
G1	first period grade	Quantitative	Discrete
G2	second period grade	Quantitative	Discrete
G3	final grade	Quantitative	Discrete

Main variables for analysis:
All

Possible analysis (crosstabs/graphs, etc):
Dalc x sex (Ex. Make a contingency table and Chi-square association test)
Dalc x address (Ex. Make a bar graph and contingency table)
G3 x school (Ex. Make a plot of mean of G3 in function of school)
G3 x Medu (Ex. Make a plot of mean of G3 in function of Medu)
G1 x G3 (Ex. Calculate the mean and do the average value equality hypothesis test)
G1 x G2 (Ex. Calculate correlation and do correlation hypothesis test)
Pstatus x Dalc (Ex. Make contingency table and Chi-square association test)
School x Medu x higher (Ex. Make a Multi-way table)

Comments: Tools used to explore this data set: R Commander

Observation The dataset has two sub datasets, for Portuguese course and for maths course.

- **Air Quality dataset – Italian city**

Name	AirQualityUCI
Source	UCI – Machine Learning Repository Available in: https://archive.ics.uci.edu/ml/datasets/Air+Quality
Year	March 2004 – April 2005
Unity of analysis	Register of air quality
Number of registers	9357
Number of variables	15

Variables:

Date	Date	Qualitative	Ordinal
Time	Time (in hours)	Quantitative	Discrete
CO (GT)	True hourly averaged concentration CO in mg/m^3	Quantitative	Continuous
PT08.S1(CO)	(tin oxide) hourly averaged sensor response (nominally CO targeted)	Quantitative	Discrete
NMHC(GT)	True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3	Quantitative	Discrete
C6H6(GT)	True hourly averaged Benzene concentration in microg/m^3	Quantitative	Continuous
PT08.S2(NMHC)	(titania) hourly averaged sensor response (nominally NMHC targeted)	Quantitative	Discrete
NOx (GT)	True hourly averaged NOx concentration in ppb	Quantitative	Discrete
PT08.S3(NOx)	(tungsten oxide) hourly averaged sensor response (nominally NOx targeted)	Quantitative	Discrete
NO2(GT)	True hourly averaged NO2 concentration in microg/m^3	Quantitative	Discrete
PT08.S4 (NO2)	(tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)	Quantitative	Discrete
PT08.S5 (O3)	(indium oxide) hourly averaged sensor response (nominally O3 targeted)	Quantitative	Discrete
T	Temperature	Quantitative	Continuous
RH	Relative Humidity	Quantitative	Continuous
AH	Absolute Humidity	Quantitative	Continuous

Main variables for analysis:
Time, CO, NMHC, C6H6, NOx, NO2, T, RH

Possible analysis (crosstabs/graphs, etc):
Time x C6H6 (Ex: barplot)
Time x NO2 x CO (Ex: barplot)
Time x T x RH (Ex: barplot)

Comments: Tools used to explore this data set: Tableau

- **Health dataset - Zika**

Name	cdc_zika
Source	Kaggle Available in: https://www.kaggle.com/cdc/zika-virus-epidemic
Year	January 2016 to June 2016
Unity of analysis	Suspicious Zika case
Number of registers	107619
Number of variables	9

Variables:

location	Names specified in the country place name database	Qualitative	Nominal
location_type	ocation code is included indicating: city, district, municipality, county, state province, or country.	Qualitative	Nominal
data_field	Short description of what data is represented in the row	Qualitative	Nominal
data_field_code	This code is defined in the country data guide.	Qualitative	Nominal
value	The observation indicated for the specific report	Quantitative	Discrete

Main variables for analysis:
location, location_type, data_field, data_field_code, value

Possible analysis (crosstabs/graphs, etc):
location_type x value
Data_field x value
Location x Data_field_code

Comments: Tools used to explore this data set: Tableau

- Mental Health in Tech Survey

Name	MentalHealth
Source	Kaggle Available in: https://www.kaggle.com/osmi/mental-health-in-tech-survey
Year	2014
Unity of analysis	
Number of registers	1259
Number of variables	26

Variables:

Timestamp	Date and hour		
Age	Age	Quantitative	Discrete
Gender	Gender	Qualitative	Nominal
Country	Country	Qualitative	Nominal
SelfEmp	Are you self-employed?	Qualitative	Nominal
FamilyHistory	Do you have a family history of mental illness?	Qualitative	Nominal
Treatment	Have you sought treatment for a mental health condition?	Qualitative	Nominal
WorkInterfere	If you have a mental health condition, do you feel that it interferes with your work?	Qualitative	Ordinal
remoteWork	Do you work remotely (outside of an office) at least 50% of the time?	Qualitative	Nominal
TechCompany	Is your employer primarily a tech company/organization?	Qualitative	Nominal
leave	How easy is it for you to take medical leave for a mental health condition?	Qualitative	Ordinal
anonymity	Is your anonymity protected if you choose to take advantage of mental health or substance abuse treatment resources?	Qualitative	Nominal
coworkers	Would you be willing to discuss a mental health issue with your coworkers?	Qualitative	Nominal
supervisor	Would you be willing to discuss a mental health issue with your direct supervisor?	Qualitative	Nominal
MentPhysic	Do you feel that your employer takes mental health as seriously as physical health?	Qualitative	Nominal

Main variables for analysis:

Possible analysis (crosstabs/graphs, etc):

Comments: Tools used to explore this data set: Tableau

- **Hapiness dataset**

Name	Hapiness
Source	kaggle Available in: https://www.kaggle.com/unsdsn/world-happiness
Year	2015
Unity of analysis	Country
Number of registers	158
Number of variables	11

Variables:

Region	Region	Qualitative	Nominal
HappScore	Happiness Score	Quantitative	Continuous
Economy	GDP per Capita	Quantitative	Continuous
Family	Family	Quantitative	Continuous
Health	Life Expectancy	Quantitative	Continuous
Freedom	Freedom	Quantitative	Continuous
Trust	Government Corruption	Quantitative	Continuous
Generosity	Generosity	Quantitative	Continuous
DystRes	Dystopia Residual (Dystopia is an imaginary country that has the world's least-happy people)	Quantitative	Continuous
Hapiness	Ranking Position (highest in overall happiness)	Qualitative	Ordinal

Main variables for analysis:
All

Possible analysis (crosstabs/graphs, etc):
Region x HappScore
Country x Hapiness
Region x Trust
Region x Generosity x Family x Freedom
Region x HappScore x Health

Comments: Tools used to explore this data set: *iNZight* and *Tableau.*

- **Ecological footprint and biocapacity dataset**

Name	Ecological footprint
Source	kaggle Available in: https://www.kaggle.com/footprintnetwork/ecological-footprint
Year	2015
Unity of analysis	Country
Number of registers	188
Number of variables	21

Variables:

Region	Region	Qualitative	Nominal
HDI	Human development index	Quantitative	Continuous
GrazingF	Grazing footprint	Quantitative	Continuous
ForestF	Forest footprint	Quantitative	Continuous
CarbonF	Carbon footprint	Quantitative	Continuous
FishF	Fish footprint	Quantitative	Continuous
TotalFootP	Total Ecological footprint	Quantitative	Continuous
GrazingL	Grazing land	Quantitative	Continuous
ForestL	Forest land	Quantitative	Continuous
UrbanL	Urban land	Quantitative	Continuous
FishW	Fishing water	Quantitative	Continuous
TotalBio	Total biocapacity	Quantitative	Continuous
BioDeficit	Biocapacity deficit	Quantitative	Continuous

Main variables for analysis:

All

Possible analysis (crosstabs/graphs, etc):

Region x TotalFootP x TotalBio
Country x HDI x BioDeficit
Region x GrazingF x ForestF x CarbonF x FishF

Comments: Tools used to explore this data set: *Tableau.*

- ### Cardiovascular disease dataset - USA

Name	CVDRisk
Source	CDC – Center for Disease Control and Prevention Available in: https://chronicdata.cdc.gov/videos
Year	1998
Unity of analysis	State (USA)
Number of registers	51
Number of variables	4

Variables:

CurSmok	Current Smoking	Quantitative	Continuous
Over	Overweight	Quantitative	Continuous
FruitVegs	<5 Fruit / vegetables	Quantitative	Continuous
PhisInactive	Physically Inactive	Quantitative	Continuous

Main variables for analysis:
All

Possible analysis (crosstabs/graphs, etc):
Overweight x Physically Inactive
FruitVegs x Over
State x CurSmok

Comments: Tools used to explore this data set: *Infogr.am* and *Tableau.*

- **German Gender Pay Gap data set**

Name	Verdienststrukturerhebung (fdz_vse_cf_2006_csv.zip)
Source	Destatis Available in: https://campus-file-fdz.nrw.de/index.php?strAction=user.datasets
Year	2006
Unity of analysis	Individuals
Number of registers	60 000
Number of variables	32

Variables:

Region	Region (West/East)	Qualitative	Nominal
Wzgruppe	Distinguishes between different kinds of industries	Qualitative	Nominal
EF9	Information about the pay scale grouping	Qualitative	Ordinal
EF10	Gender	Qualitative	Nominal
EF11	Year of birth	Quantitativ	Discrete
EF12U2	Year of joining the company	Quantitativ	Discrete
Beruf	Classification of job	Qualitative	Nominal
EF16U1	Position in job	Qualitative	Nominal
EF16U2	Education	Qualitative	Ordinal
EF17	Type of contract	Qualitative	Nominal
EF18	Weekly working time	Quantitativ	Discrete
EF19	Paid normal working hours per month	Quantitativ	Discrete
EF20	Paid additional working hours per month	Quantitativ	Discrete
EF21	Monthly wage in Euro (pre-tax)	Quantitativ	Discrete

Main variables for analysis:
Region, Gender, EF9, EF16U1, EF18, EF21
Possible analysis (crosstabs/graphs, etc):
Gender x EF21
Gender x EF16U1
Region x EF21
Comments: Tools used to explore this data set: Fathom, TinkerPlots, Codap (Sample)

- **Other datasets**

- **PISA 2015 dataset - Percentage of students at each proficiency level in mathematics, by gender**

Name	PISA (Programme for International Student Assessment).
Source	OECD Available in: https://www.oecd.org/pisa/data/2015database/
Year	2015
Unity of analysis	Individuals
Number of registers	519334
Number of variables	2048

Note: New variables (demographic) related with this dataset will be available in October of 2017. On that time will be suggested the variables and possible analysis.

- **Poverty and Equity dataset**

Name	PovStats
Source	The World Bank Open Data Available in: http://data.worldbank.org/data-catalog/poverty-and-equity-database
Year	1974 – 2015
Unity of analysis	Country
Number of registers	5741
Number of variables	3

- **Climate Change dataset**

Name	Climate_Change_download0
Source	The World Bank Open Data Available in: http://data.worldbank.org/data-catalog/climate-change
Year	Between 1990 and 2011
Unity of analysis	Country
Number of registers	13512
Number of variables	3

- **Violence dataset**

Name	ucdp-onesided-14-2016_violence attacks
Source	Uppsala Conflict Data Program (violence attacks on civilians by governments and formally organize

	armed groups) Available in: http://ucdp.uu.se/downloads/
Year	1989-2015
Unity of analysis	Attack
Number of registers	870
Number of variables	11

- **Smoking dataset - USA**

Name	CurrentCigaretteSmoking
Source	CDC – Center for Disease Control and Prevention Available in: https://chronicdata.cdc.gov/videos
Year	1999
Unity of analysis	State
Number of registers	51
Number of variables	3

- **Pneumonia dataset**

Name	Pneumonia
Source	Knoema Available in: https://knoema.com/atlas/topics/Health/Deaths-among-childre under-5-by-cause/
Year	2010
Unity of analysis	Country
Number of registers	190
Number of variables	2

- **Elections dataset**

Name	Qualified_voter_listing_2015_pr
Source	kaggle Available in: www.kaggle.com
Year	2015
Unity of analysis	case
Number of registers	1686
Number of variables	14

Appendix 1

List of data sources

- **Find data:**

ABS.Stat - Australian Bureau of Statistics

ACLED - Armed Conflict Location and Event Data Project

AidData - Open data for International Development

Bureau of Labor Statistics

Cambridge Open Data

China Data Center

CIRI Human Rights Data Project

Data and Story Library

Data First

Data Revolution

Data.gov

Data.gov.in - India

Dataverse - Harvard Library

Devecondata

Developments in a globalized world

DHS Program - Demographic and health Surveys

DISC - Data and Information Services Center

EM-DAT

Europe Union Open Data Portal

European Union Agency for Fundamental Rights

Eurostat

German National Statistics Office

gesis - German official microdata

GHDx - Global Health Data Exchange

GOV.UK - Statistics

govdata - Germany

GSS The General Social Survey (NORC - U. Chicago)

GTD - global terrorism database

Human development Report_UN

Human Rights Data Analysis Group

Hunger and Nutrition Commitment Index

ICPSR - University of Michigan

IFPRI - International Food Policy Research Institute

Global Health Data Exchange - Institute for Health Metrics and Evaluation

IHSN - International Household Survey Network

ILRI datasets

Inequality project - University of Texas

IPUMS

IWPR - Institute for Women's Policy Research

Kaggle
Knoema - World data
LIS - Cross National Data Center in Luxembourg
Macroeconomics and Growth - World Bank
NBER - National Bureau of Economic research
NISRA - Northern Ireland Statistics and Research Agency – microdata teaching file 2011 Census
OECD
Office for National Statistics - Neighbourhood Statistics - UK
Office for National Statistics – Crime and Justice - UK
Open Data Philly - Philadelphia
openAFRICA
OpenDataCity (German only)
OpenEI
OpenMicrodata
OPR - Office of Population Research
Politicians
PRIO - Peace Research Institute Oslo
RatSWD - German Data Forum
SDG Indicators - Global Database
SEDLAC - Social Economic Database for Latin and the Caribbean
SHARE - Survey of Health Ageing and retirement in Europe
START - Study Terrorism and responses to terrorism - Univ. Maryland
The World Wealth and Income Database
TransMonEE - Transformative Monitoring for Enhanced Equity (UNICEF)
UCDP - Uppsala conflict data program
UCI - Machine learning repository
UK Data Service
UNdata
UNESCO
UNHCR - The UN Refugee Agency
UNICEF
UNU - WIDER (World Income Inequality Database)
US Census
World Bank (Microdata catalog)
FAO (UN Food and Agriculture)

- **Visualize and Explore data:**

Africa Development Bank
AIDS - Information from UN
Better Life Index
British Social Attitudes
Cardiovascular Disease
Census Data - German Statistics Office

Census Explorer - U.S.
Chartbook of Economic Inequality
Children – UNICEF Report
Climate Change
Climate Spirals
Compare your Income – from OECD
Constituency Explorer
Data 360
DataTopics_World Bank - Financial Inclusion
Demographic and Health indicators - US AID
EIGE - European Institute for Gender Equality
Ending Rural Hunger
Environmental accounting
Explore Climate - World Bank
Feeding America
Flowing data
Food Security
Gender pay gap
GenderStatistics
Global Forest Watch
HNP - Health, Nutrition and population Statistics (World Bank)
How different groups spend their day - NY Times
Hunger and Nutrition Commitment Index
IHME - Institute for Health Metrics and Evaluation
International Food Policy Research Institute _IFPRI
Malnutrition - UNICEF
NCD RisC - Diabetes
OECD Data
Our world in data
PISA 2015
PovcalNET – poverty calculator
Poverty and Equity
PRIO - GRID
PRIO - Sexual Violence during Armed Conflict
Radicalization in US
Refugee Project
SMART Census - data visualisations of 2011 UK census dada
SMART Center - High school curriculum materials on UK education, crime and health
Knoema - 'Smarter research with the worlds statistics in your hands'
Social Determinants of health Visualization
Statistics Explorer - OECD
Status of Women in the States
The Rhythm of Food
UCDP - Uppsala Conflict Data Program

Understanding Uncertainty - Survival
UNESCO - data for the Sustainable Development Goals
UNHCR - Operational Portal Refugee Situation
Violence against women survey - FRA
Water
When will you die?
Women in Science
Worldmapper

- **Other**

AAUW - Empowering Women
British Social Attitudes
Census - US
Country Profile - IHME
Country profiles 1 - UNICEF
Country profiles 2 - UNICEF
Department of Peace and conflict Research
Gallup_Well-Being Index_ObesityEqualPayPortal
Gender Equality DevelopmentGallup_Well-Being Index_Obesity
Gender Gap EducationGender Equality Development
Gender Pay Gap in EuropeGender Gap Education
GenderATlas - AustriaGender Pay Gap in Europe
GenderStatsGenderATlas - Austria
Global Nutrition - AfghanistanGenderStats
Global Partnership for Sustainable Development DataGlobal Nutrition - Afghanistan
Knoema - World data - EbolaGlobal Partnership for Sustainable Development Data
PayScale Human CapitalKnoema - World data - Ebola
Refugees Emergency Response Mediterranean - UNHCRPayScale Human Capital
Spurious CorrelationRefugees Emergency Response Mediterranean - UNHCR
Statistics Explained _ EuroStatSpurious Correlation
Undernutrition - DeathsStatistics Explained _ EuroStat
Visual.ONSUndernutrition - Deaths
WorldHungerVisual.ONS
WorldHunger