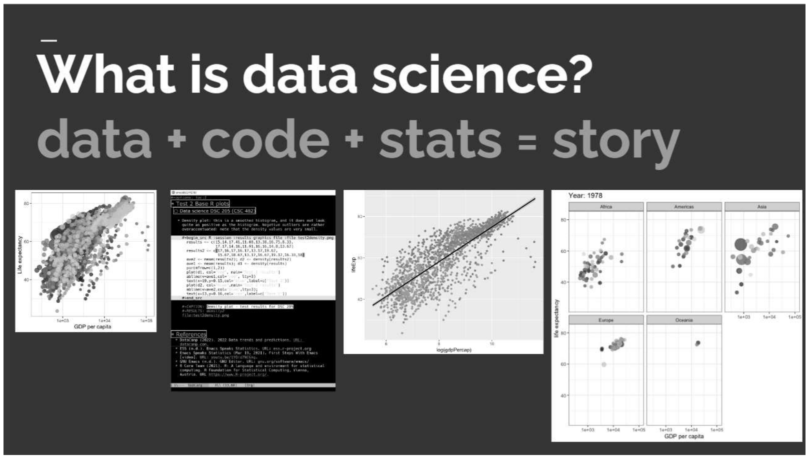
DATA, Exploratory Data Analysis, and R

Introdution to Data Visualization

Table of Contents



What is data science?

data + code + stats = story

Figure 1: data science pipeline

- Why do we analyze data?
- Data as a concept and as a practice

- The importance of metadata
- Exploratory Data Analysis
- Numbers vs. graphs
- Data analysis workflow
- Why R?
- R package management
- Download R practice file

1 WHY DO WE ANALYZE DATA?

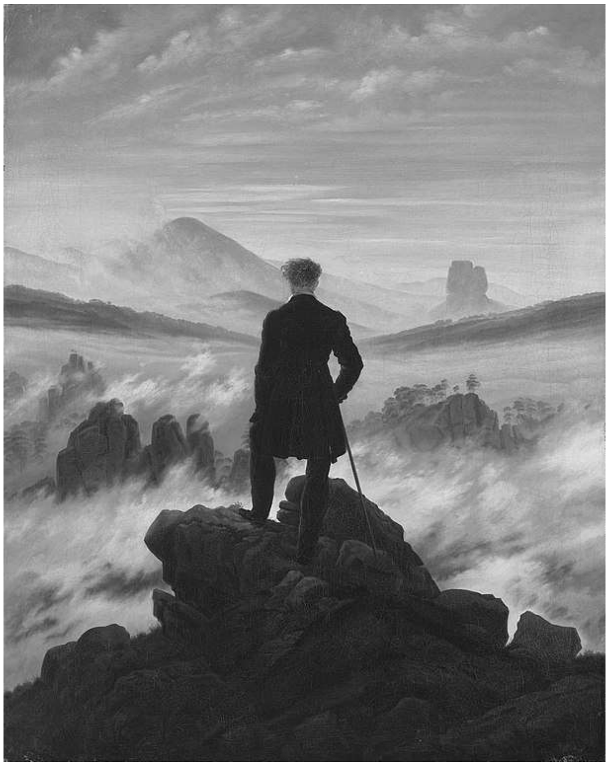


Figure 2: data science pipeline

- Well, what do you think?

  Tip: reframe WHY questions as WHAT questions -

  - What data?
  - What analysis?
  - What benefits?

  Pearson:

  1. to **understand** what has happened or what is happening;
  2. to **predict** what is likely to happen, either in the future or in other circumstances we haven't seen;
  3. to **guide** us in making decisions.

2 DATA - CONCEPT

- An **entity**, e.g.
  - family history of patient in medical study
  - competing company characteristics in marketing analysis
- An **event**, e.g.
  - demographic characteristics of voters
  - locations visited during a shopping visit
- A **process**, e.g.
  - manufacturing data from a production line
  - application information from a hiring process

3 DATA - PRACTICE

- Data structures = rectangular array of observed values
- Rows = observation of entity, event, or process

- Columns = recorded characteristic or attribute

```
## extract first six records from the mtcars data frame
head(mtcars)
```

```
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

- Data frame rows and columns have *names*
- Complete description with help(mtcars)
- Meta data supplement data frame content

4 META DATA

Figure 3: Greek goddess of peace and spring (Ειρήνη)

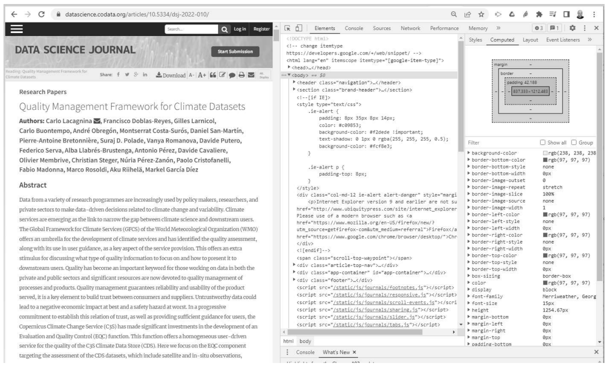
- "Data about data" (Greek μετά = 'after', 'beyond')
- Meta data in `mtcars`:
  - Original source of the data
  - Scientific paper analyzing the data
  - Description of the variables (columns)

  What could be issues with metadata?

  - **Completeness -** origin
  - **Consistency -** logic, values, (time) dependency
  - **Accuracy -** origin and validity

    > "As potentially valuable as metadata is, we cannot afford to accept it uncritically: we should always cross-check the metadata with the actual data values, with our intuition and prior understanding of the subject matter, and with other sources of information that may be available." (Pearson, 2018)

5 TODO PRACTICE: META OR NOT META?



datascience.codata.org/articles/10.5334/dsj-2022-010/

DATA SCIENCE JOURNAL

Research Papers

Quality Management Framework for Climate Datasets

Authors: Carlo Lacagnina, Francisco Doblas-Reyes, Gilles Larnicol, Carlo Buontempo, André Obregón, Montserrat Costa-Surós, Daniel San-Martín, Pierre-Antoine Bretonnière, Suraj D. Polade, Vanya Romanova, Davide Putero, Federico Serva, Alba Llabrés-Brustenga, Antonio Pérez, Davide Cavaliere, Olivier Membrive, Christian Steger, Núria Pérez-Zanón, Paolo Cristofanelli, Fabio Madonna, Marco Rosoldi, Aku Riihelä, Markel García Díez

Abstract

Data from a variety of research programmes are increasingly used by policy makers, researchers, and private sectors to make data-driven decisions related to climate change and variability. Climate services are emerging as the link to narrow the gap between climate science and downstream users. The Global Framework for Climate Services (GFCS) of the World Meteorological Organization (WMO) offers an umbrella for the development of climate services and has identified the quality assessment, along with its use in user guidance, as a key aspect of the service provision. This offers an extra stimulus for discussing what type of quality information to focus on and how to present it to downstream users. Quality has become an important keyword for those working on data in both the private and public sectors and significant resources are now devoted to quality management of processes and products. Quality management guarantees reliability and usability of the product served, it is a key element to build trust between consumers and suppliers. Untrustworthy data could lead to a negative economic impact at best and a safety hazard at worst. In a progressive commitment to establish this relation of trust, as well as providing sufficient guidance for users, the Copernicus Climate Change Service (C3S) has made significant investments in the development of an Evaluation and Quality Control (EQC) function. This function offers a homogeneous user-driven service for the quality of the C3S Climate Data Store (CDS). Here we focus on the EQC component targeting the assessment of the CDS datasets, which include satellite and in-situ observations,

```
<!DOCTYPE html>
<!-- change itemtype
https://developers.google.com/+/web/snippet/ -->
<html lang="en" itemscope itemtype="[google-item-type]">
 <head>…</head>
…<body> == $0
   <header class="navigation">…</header>
   <section class="brand-header">…</section>
   <!--[if IE]>
   <style type="text/css">
       .ie-alert {
           padding: 8px 35px 8px 14px;
           color: #c09853;
           background-color: #f2dede !important;
           text-shadow: 0 1px 0 rgba(255, 255, 255, 0.5);
           background-color: #fcf8e3;
       }

       .ie-alert p {
           padding-top: 8px;
       }
   </style>
   <div class="col-md-12 ie-alert alert-danger" style="margi
       <p>Internet Explorer version 9 and earlier are not su
   href="http://www.ubiquitypress.com/site/internet_explorer
   Please use of a modern browser such as <a
   href="https://www.mozilla.org/en-US/firefox/new/?
   utm_source=getfirefox-com&utm_medium=referral">Firefox</a
   href="https://www.google.com/chrome/browser/desktop/">Chr
   </div>
   <![endif]-->
   <span class="scroll-top-waypoint"></span>
   <div class="article-top-nav">…</div>
   <div class="app-container" id="app-container">…</div>
   <div class="footer">…</div>
   <script src="/static/js/journals/footnotes.js"></script>
   <script src="/static/js/journals/responsive.js"></script>
   <script src="/static/js/journals/scroll-events.js"></scri
   <script src="/static/js/journals/sharing.js"></script>
   <script src="/static/js/journals/slider.js"></script>
   <script src="/static/js/journals/tabs.js"></script>
```

```
background-color        rgb(238, 238, 238
border-bottom-color     rgb(97, 97, 97)
border-bottom-style     none
border-bottom-width     0px
border-image-outset     0
border-image-repeat     stretch
border-image-slice      100%
border-image-source     none
border-image-width      1
border-left-color       rgb(97, 97, 97)
border-left-style       none
border-left-width       0px
border-right-color      rgb(97, 97, 97)
border-right-style      none
border-right-width      0px
border-top-color        rgb(97, 97, 97)
border-top-style        none
border-top-width        0px
box-sizing              border-box
color                   rgb(97, 97, 97)
display                 block
font-family             Merriweather, Georg
font-size               15px
height                  1254.67px
margin-bottom           0px
margin-left             0px
margin-right            0px
margin-top              0px
padding-bottom          0px
```

Figure 4: datascience.codata.org/articles/10.5334/dsj-2022-010/

Pair exercise: Identify the different types of data and metadata in the screenshot of an online journal article.

1. Article meta data: Journal title, "Research paper", title, authors
2. Layout meta data: HTML/CSS elements
3. Browser meta data: browser data (buttons for: download, font size, print, login, register, menu options; browser console; URL)

4. Article content data: abstract + paper text, tables and figures

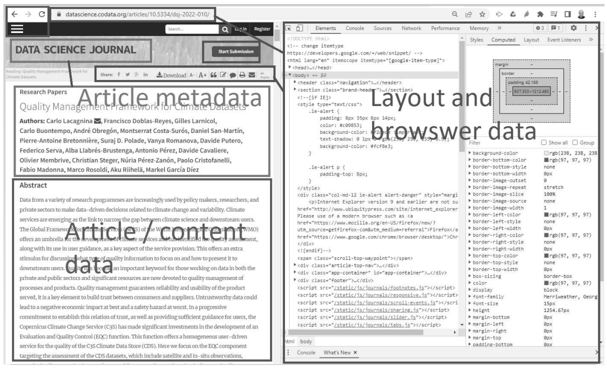
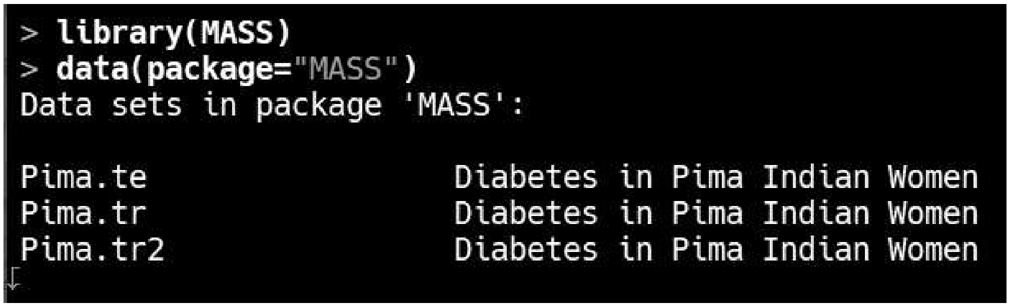


Figure 5: Solution

6 PROBLEM: MISSING VALUES



```
> library(MASS)
> data(package="MASS")
Data sets in package 'MASS':

Pima.te                 Diabetes in Pima Indian Women
Pima.tr                 Diabetes in Pima Indian Women
Pima.tr2                Diabetes in Pima Indian Women
```

Figure 6: Pima Indians data sets in the MASS package

Check out structure of Pima datasets:

```
str(Pima.te)
str(Pima.tr)
str(Pima.tr2)
```

- The MASS package contains three different versions of the Pima indians data set (diabetes in women of the Pima tribe)
- MASS metadata comments:

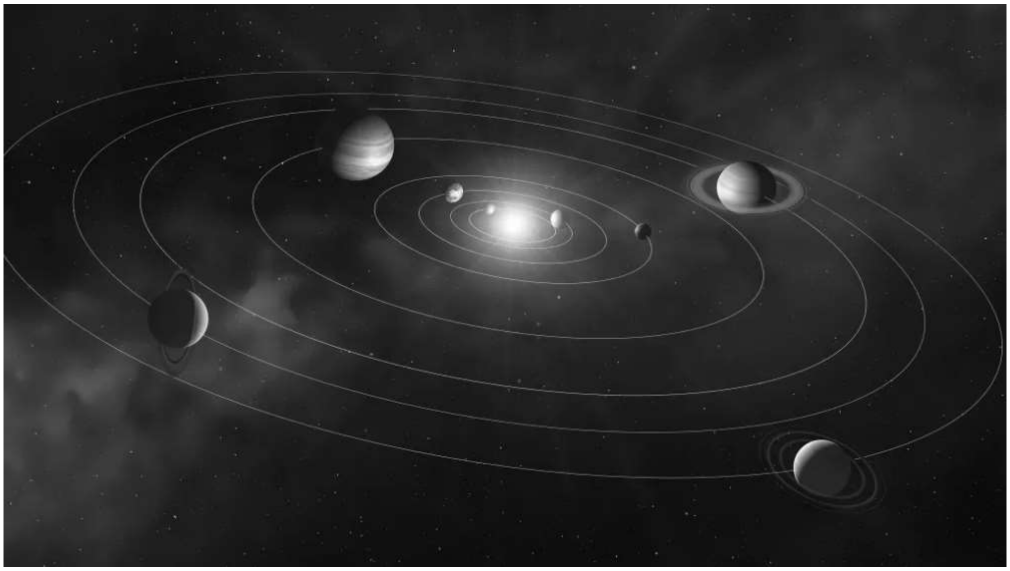
  > "The training set `Pima.tr` contains a randomly selected set of 200 subjects, and `Pima.te` contains the remaining 332 subjects. `Pima.tr2` contains `Pima.tr` plus 100 subjects with missing values in the explanatory variables."

- The kaggle.com database is yet another version: more records, one more variable - the "Metadata" information is missing
- Missing data are often coded as `0` instead of `NA` leading to errors:

  > "A number of studies characterizing *binary classifiers* have been published using [the Pima] dataset as a benchmark where the authors were not aware that data values were missing." (Pearson, 2018)

7 PROBLEM: VARIABLE DEFINITIONS

- How many planets are there orbiting the sun?



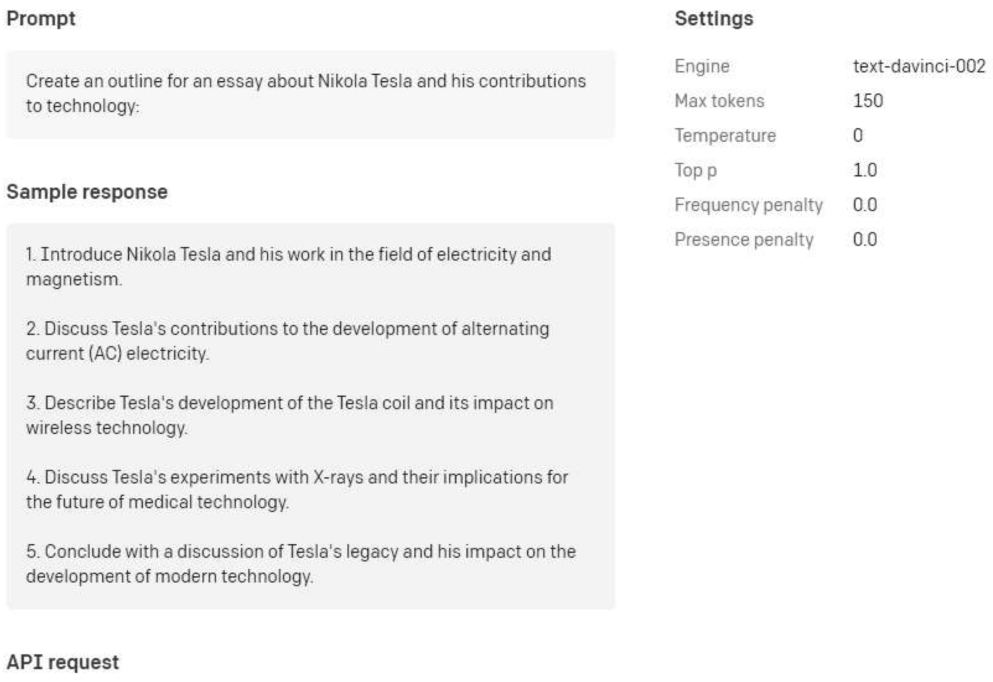
- Definitions count: e.g. *planethood* (Weintraub, 2007)
  1. the object is too small to generate nuclear fusion energy
  2. the object is big enough to be spherical
  3. the object must have a primary orbit around a star
- Unrecognized disagreements in the definition of a variable are possible between those who *measure and record* it, and those who use data in *analysis*.
- Prominent examples: when does a patient die of COVID-19? What is the cause of death? When do two patients have the same disease?

8 EXPLORATORY DATA ANALYSIS (EDA)

> "We look at *numbers* or *graphs* and try to find *patterns*. We pursue leads suggested by background information, imagination, patterns perceived, and experience with other data analyses." (Diaconis, 1985)

- Analysis is always based on exploring numbers
- Non-numerical data are converted to numbers: e.g. *categorical* variables are converted from discrete named values ("political party", "city") into counts or relative frequencies
- Each discrete value or category is also called a "level"

9 TYPES OF CATEGORICAL VARIABLES



Prompt

Create an outline for an essay about Nikola Tesla and his contributions to technology:

Sample response

1. Introduce Nikola Tesla and his work in the field of electricity and magnetism.

2. Discuss Tesla's contributions to the development of alternating current (AC) electricity.

3. Describe Tesla's development of the Tesla coil and its impact on wireless technology.

4. Discuss Tesla's experiments with X-rays and their implications for the future of medical technology.

5. Conclude with a discussion of Tesla's legacy and his impact on the development of modern technology.

API request

Settings

Engine	text-davinci-002
Max tokens	150
Temperature	0
Top p	1.0
Frequency penalty	0.0
Presence penalty	0.0

Figure 9: AI-generated outline for research topic (Source: OpenAI)

- Few levels (e.g. "Firm", "Party", "City")
- Many levels (e.g. US ZIP code with 40,000 levels)
- Exploitable sub-structure (e.g. text data[1])

10 SOME ISSUES WITH GRAPHS

- Humans are better at seeing patterns in graphs than numbers[2]

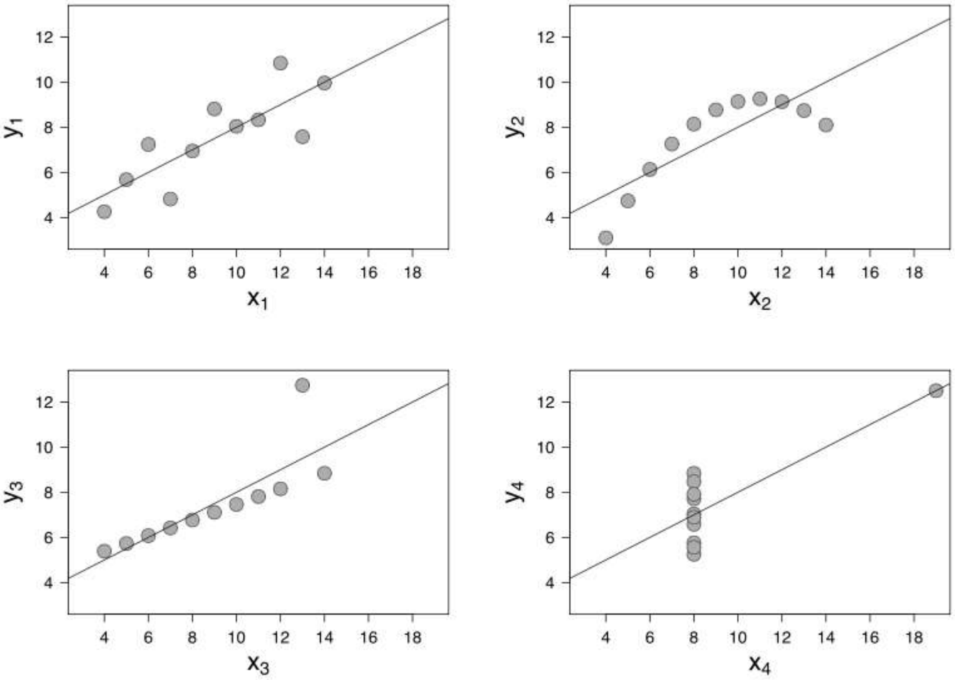


Figure 10: Anscombe dataset

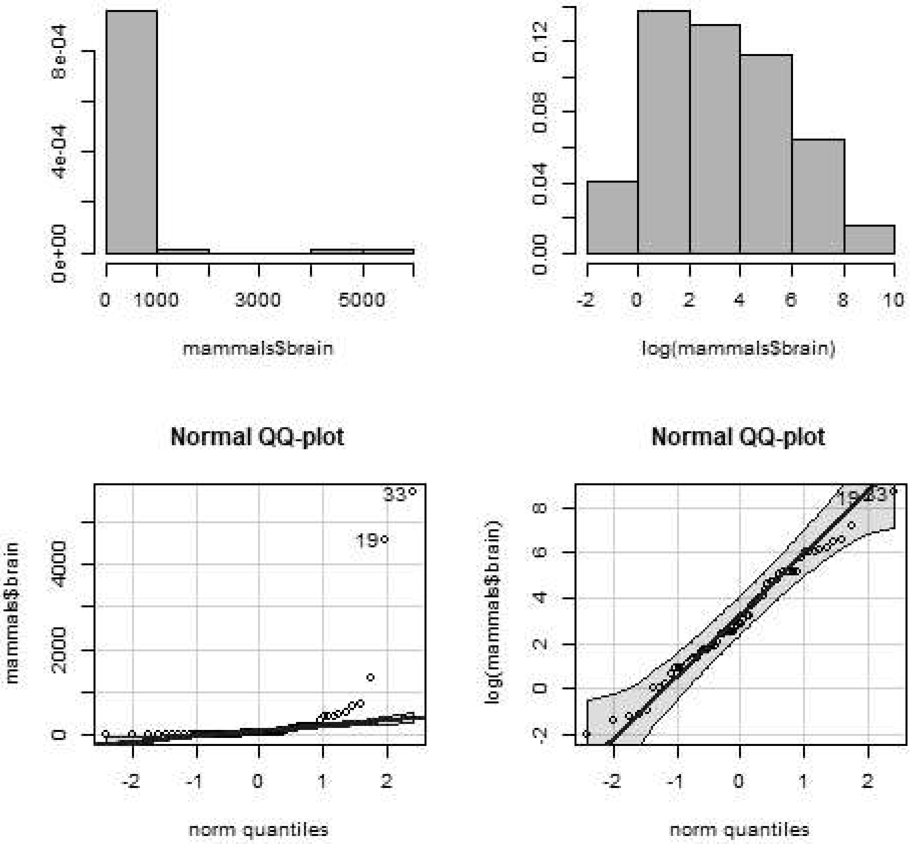
  - Use different graphs to explore and to explain - data mining is *exploratory*, data story telling is *explanatory*[3]
  - Usefulness of a graph depends on **how data** are displayed, and strongly on **which data** are chosen to be displayed

11 TODO PRACTICE: RAW VS. TRANSFORMED GRAPH DATA

- The following two sets of plots are constructed from the `brain` element of the `mammals` dataset from the `MASS` package that lists body and brain weights for 62 different animals.

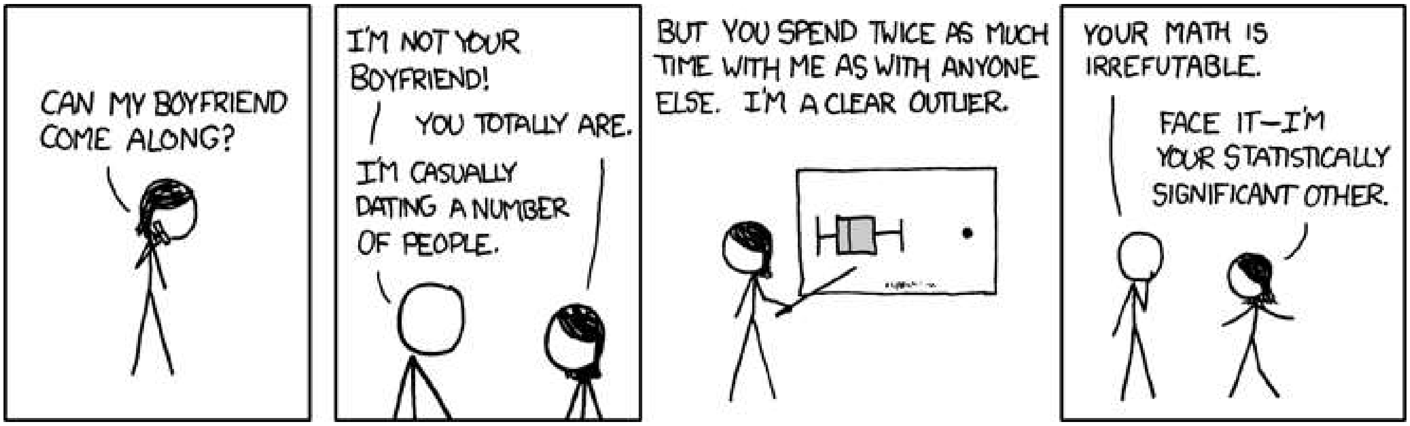
- **What do you think which graphs are more meaningful and why?**

```
library(MASS)
library(car)
par(mfrow=c(2,2))
truehist(mammals$brain)
truehist(log(mammals$brain))
qqPlot(mammals$brain)
title("Normal QQ-plot")
qqPlot(log(mammals$brain))
title("Normal QQ-plot")
```



- The plots are telling us something about the distribution of data values.
- The left-hand pair were generated from raw data values, the right-hand pair were generated from log-transformed data
- The right-hand pair suggests that the data exhibit a Gaussian (normal) distribution

12 R FOR EXPLORATORY ANALYSIS



CAN MY BOYFRIEND COME ALONG?
I'M NOT YOUR BOYFRIEND!
YOU TOTALLY ARE.
I'M CASUALLY DATING A NUMBER OF PEOPLE.
BUT YOU SPEND TWICE AS MUCH TIME WITH ME AS WITH ANYONE ELSE. I'M A CLEAR OUTLIER.
YOUR MATH IS IRREFUTABLE.
FACE IT—I'M YOUR STATISTICALLY SIGNIFICANT OTHER.

- Exploratory analysis has more use for graphical tools
- R supports many different graphical displays and plot types
- Important focus: searching for anomalies and outliers in the data

13 DATA ANALYSIS WORKFLOW

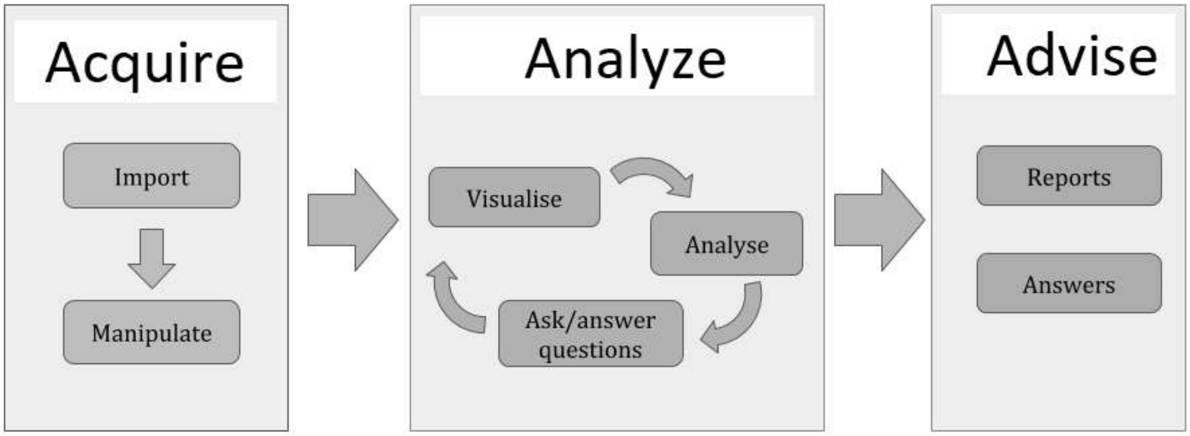


Figure 13: Data analysis workflow (emanuelaf.github.io - modified)

1. **Acquire**: make data available to the software
2. **Analyse**: perform the analysis

3. **Advise**: make analysis results available to those who need them

   - In training, the emphasis is often on (2) analysis, and pre-loaded, small, clean datasets and well-tested packages are used.
   - On the job, the emphasis is on (1) acquisition, and much time is spent importing and readying the data for analysis
   - In business, the main interest is (3) advice, hence the shift to storytelling and interpretation

14 COMPUTERS

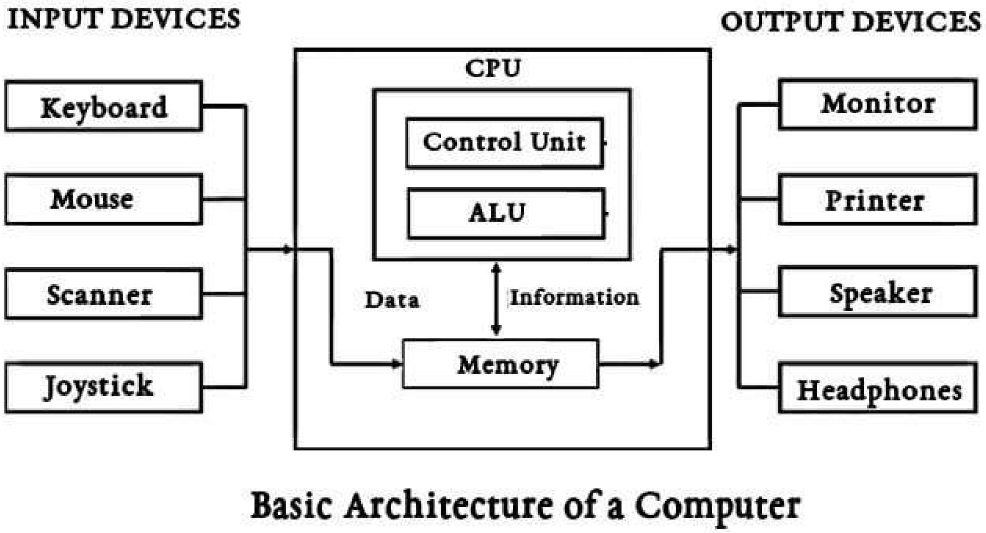


Figure 14: Von Neumann computer architecture (PSC Arivukal, 2020)

- RAM is several orders of magnitude faster than NVM
- Most R functions require raw data and results to fit in RAM
- OS and Internet impose infrastructure constraints[4]

15 WHY R?

- R is FOSS (Free Open Source Software) available for all OS
- Supported range of analysis methods ready for use
- Unix-style package and version control system
- Diverse, active community of users and developers

16 THE STRUCTURE OF R

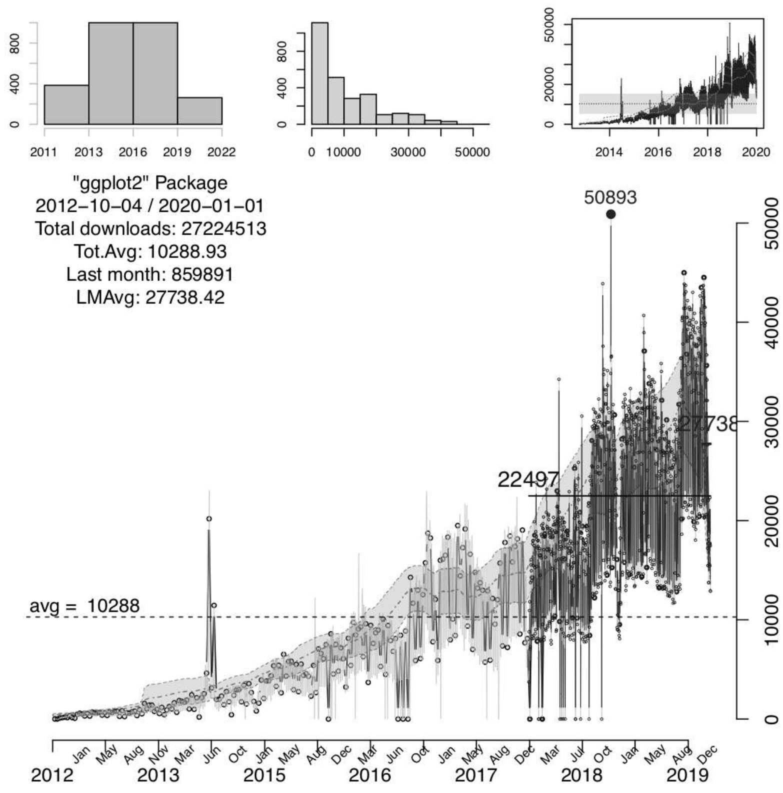


Figure 16: ggplot2 downloads from CRAN 2012-2020

1. Set of *base R packages* for basic statistics, data analysis, graphics
2. Set of *recommended packages* included in installations (like `MASS`)
3. Set of *optional add-on packages* for special purposes

Example: The optional, popular `ggplot2` graphics package was downloaded more than 272 mio. times between 2012 and 2020, with a monthly average of > 800k downloads (Source: CRAN, 2021).

17 INSTALLATION AND LOADING R PACKAGES

- We'll do this directly on the command line (see e.g. here):
- Installation = download and unpacking of binary or compilation (on Windows, when you're asked, do not compile from source):

```
install.packages("MASS")
install.packages("car")
```

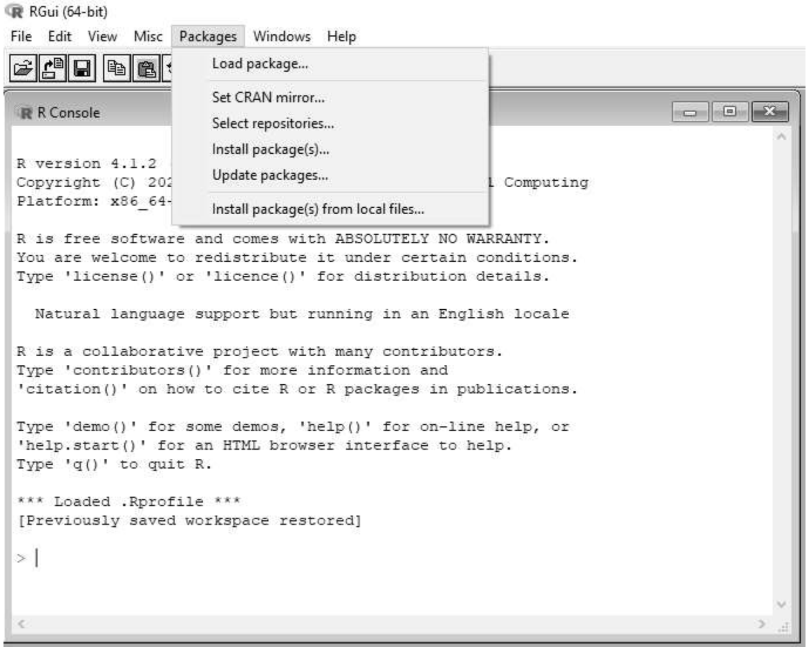
- Loading = load package (functions + datasets) into current R session:

```
library(MASS)
library(car)
```

- Alternatively, you can use the Rgui program, or the RStudio IDE

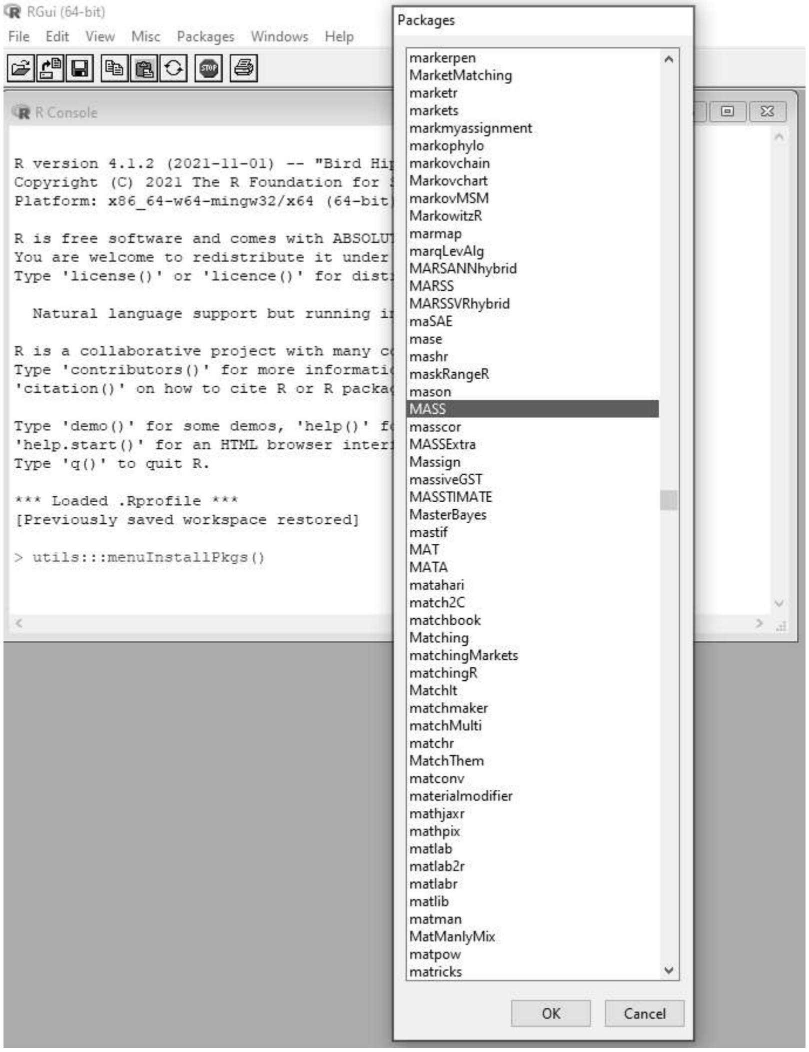
18 OPTIONAL INSTALLATION IN THE RGUI

- Start the Rgui from the CMD line terminal
- The Rgui includes a command line and graphics
- The RTerm or R program is a console only
- In the R GUI, find the tab "Packages"
- Set CRAN mirror site (closest to you)
- Install or update package from list



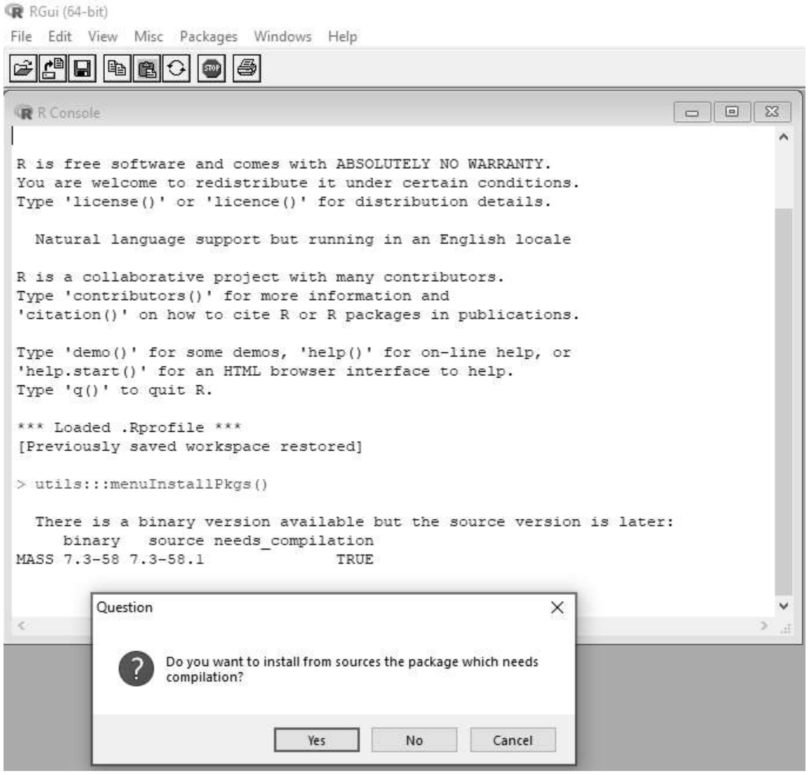
```
R version 4.1.2
Copyright (C) 202                                 l Computing
Platform: x86_64-

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

*** Loaded .Rprofile ***
[Previously saved workspace restored]

> |
```

Figure 17: Package management in the Rgui program



```
R version 4.1.2 (2021-11-01) -- "Bird Hi
Copyright (C) 2021 The R Foundation for
Platform: x86_64-w64-mingw32/x64 (64-bit

R is free software and comes with ABSOLU
You are welcome to redistribute it under
Type 'license()' or 'licence()' for dist

  Natural language support but running i

R is a collaborative project with many c
Type 'contributors()' for more informati
'citation()' on how to cite R or R packa

Type 'demo()' for some demos, 'help()' f
'help.start()' for an HTML browser inter
Type 'q()' to quit R.

*** Loaded .Rprofile ***
[Previously saved workspace restored]

> utils:::menuInstallPkgs()
```

Figure 18: Package management in the Rgui program



```
R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

*** Loaded .Rprofile ***
[Previously saved workspace restored]

> utils:::menuInstallPkgs()

  There is a binary version available but the source version is later:
     binary   source needs_compilation
MASS 7.3-58 7.3-58.1              TRUE
```

Question

Do you want to install from sources the package which needs compilation?

Yes No Cancel

Figure 19: Package management in the Rgui program

19 QUESTIONS TO ASK FROM DATA

1. Where does the dataset come from, and how is it documented?
2. How many records (rows) does this dataset contain?
3. How many fields (variables, columns) are included in each record?
4. What kinds of variables are these (e.g. numerical, categorical)
5. Are there missing values?
6. If there are missing values: are these variables always observed?
7. If there are missing values: how are they represented?
8. Are the variables included in the dataset the ones we expect?
9. Are the variable values consistent with what we expect?
10. Do the variables exhibit the relationships we expect?

20 TODO PRACTICE: A REPRESENTATIVE R SESSION

1. Open the course directory in GitHub, https://github.com/birkenkrahe/dviz
2. Open `/org/2_practice.org`
3. Open the `raw` version of the file
4. Save file as `2_practice.org`
5. Right click on the file in Explorer
6. Change `Opens with:` property to Emacs
7. Open file with Emacs from the Explorer

Summary:

21 CONCEPT SUMMARY

- Data are analysed to understand, predict, or guide decisions
- Data are entities, events or processes
- Meta data contain critical information for validation
- The data analysis workflow: acquire, analyze, advise
- R is FOSS, specialized on stats, and popular
- CRAN is the central hub for R package management

22 GLOSSARY

TERM	MEANING
Data frame	Rectangular array
Observation	Recorded event
Attribute	Characteristic
Meta data	Data about data
Data	Entity, event, process
Binary classifier	Attribute with 2 values
Missing value (`NA`)	Values that were not recorded
Categorical variable	Non-numerical, discrete
Level	Category, discrete value
Anomaly, outlier	Unusual data
CRAN	Comprehensive R Archive Network
Rgui	R console pgm with graphics
Rterm	R console (terminal) pgm only

23 REFERENCES

- CRAN (27 April 2021). Visualize downloads from CRAN Packages. Online: cran.r-project.org.
- OpenAI (2022). Example: Generate an outline for a research topic. Online: beta.openai.com.
- Pearson, R.K. (2018). Exploratory Data Analysis Using R. CRC Press.
- PSC Arivukal (July 26, 2020). Basic Computer Architecture. Online: pscarivukal.com.

- Revolutionanalytics (May 2, 2017). The Datasaurus Dozen. Online: blog.revolutionanalytics.com.

Footnotes:

[1] Text data can be normalized (reduced - e.g. parsed into words, eliminating common words like "and", "of" and punctuation marks), and converted to numbers. The numbers are analyzed mathematically, and the result is transformed back to allow interpretation of the original text data. This technique leads to impressive NLP feats (so-called transformer ML models based on massive mined data sets, like GPT-3.)

[2] The plots show Anscombe's quartet - four scatterplots which despite having different numerical values all have identical mean, variance, and standard correlation (Source: revolutionanalytics.com).

[3] This difference goes deeper than data science: explanatory research is usually confirmatory (of some theory), while exploratory research is used to construct, or build, theory. Personal note: All of my own research has been exploratory.

[4] Though they can also be enablers of education: e.g. Linux and the command line shell as a data science tool, and online REPL installations (usually Docker containers) as training grounds.

Author: Marcus Birkenkrahe

Created: 2022-08-07 Sun 16:47