

# DATA, Exploratory Data Analysis, and R

## Introduction to Data Visualization

### Table of Contents

- [1. WHY DO WE ANALYZE DATA?](#)
- [2. DATA - CONCEPT](#)
- [3. DATA - PRACTICE](#)
- [4. META DATA](#)
- [5. PRACTICE: META OR NOT META?](#)
- [6. PROBLEM: MISSING VALUES](#)
- [7. PROBLEM: VARIABLE DEFINITIONS](#)
- [8. EXPLORATORY DATA ANALYSIS \(EDA\)](#)
- [9. TYPES OF CATEGORICAL VARIABLES](#)
- [10. SOME ISSUES WITH GRAPHS](#)
- [11. PRACTICE: RAW VS. TRANSFORMED GRAPH DATA](#)
- [12. R FOR EXPLORATORY ANALYSIS](#)
- [13. DATA ANALYSIS WORKFLOW](#)
- [14. COMPUTERS](#)
- [15. WHY R?](#)
- [16. THE STRUCTURE OF R](#)
- [17. INSTALLATION AND LOADING R PACKAGES](#)
- [18. OPTIONAL INSTALLATION IN THE RGUI](#)
- [19. QUESTIONS TO ASK FROM DATA](#)
- [20. PRACTICE: A REPRESENTATIVE R SESSION](#)
- [21. CONCEPT SUMMARY](#)
- [22. GLOSSARY](#)
- [23. References](#)



- 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?

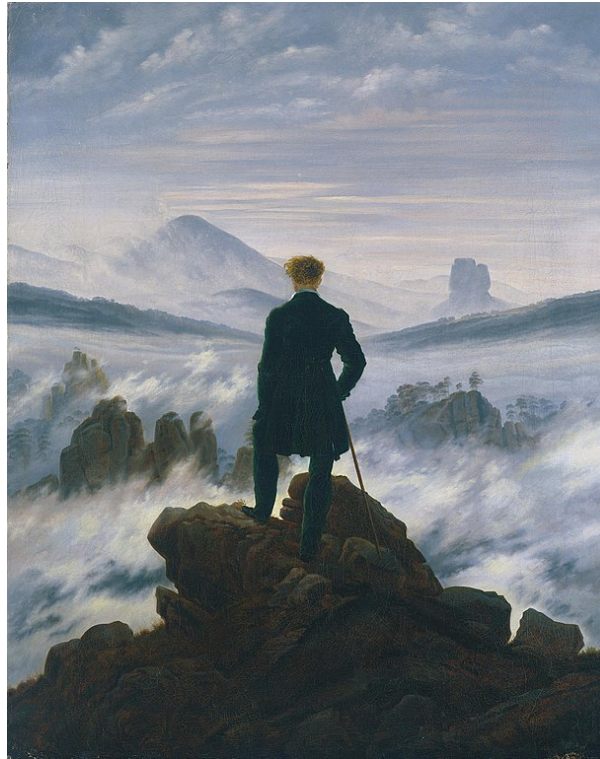


Figure 2: Caspar David Friedrich, Wanderer above a sea of fog (1818)

- Well, **what do you think?**

Reframe the question:

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

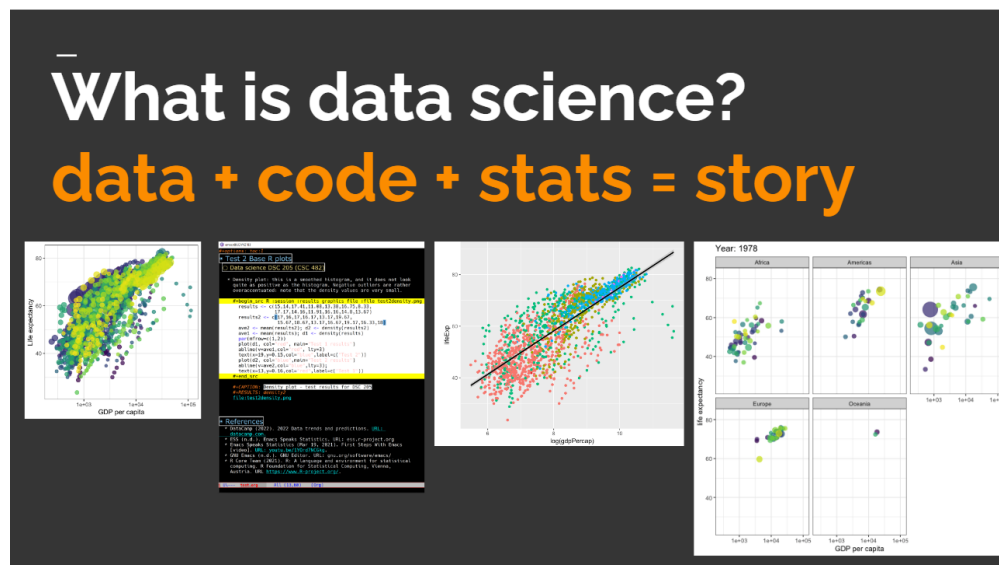


Figure 3: collapsed data science pipeline

- 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 4: Greek goddess of peace and spring "Eirene"

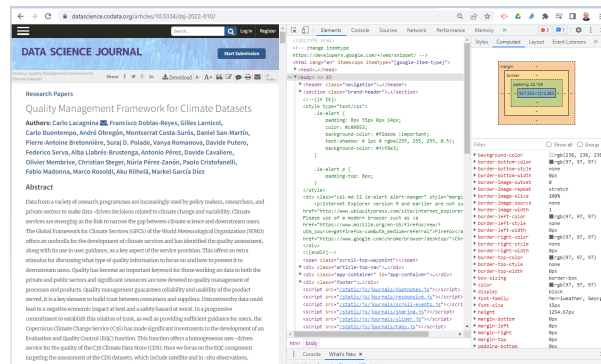
- "Data about data" (Greek 'meta'= 'after', 'beyond')
- Meta data for mtcars:
  - Original source of the data
  - Scientific paper analyzing the data
  - Description of the variables (columns)

### 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 PRACTICE: META OR NOT META?

Figure 5: [datascience.codata.org/articles/10.5334/dsj-2022-010/](https://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

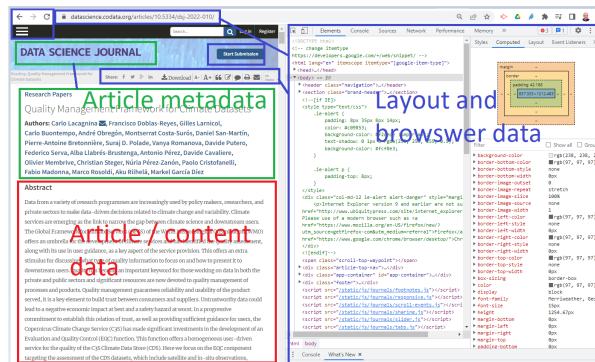


Figure 6: Solution

## 6 PROBLEM: MISSING VALUES

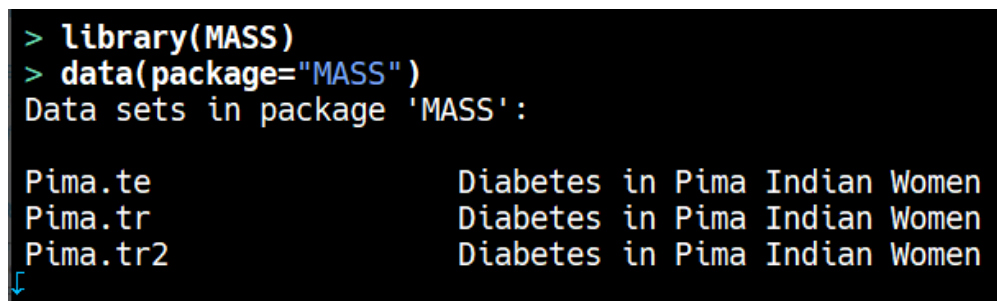


Figure 7: Pima Indians data sets in the MASS package

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:

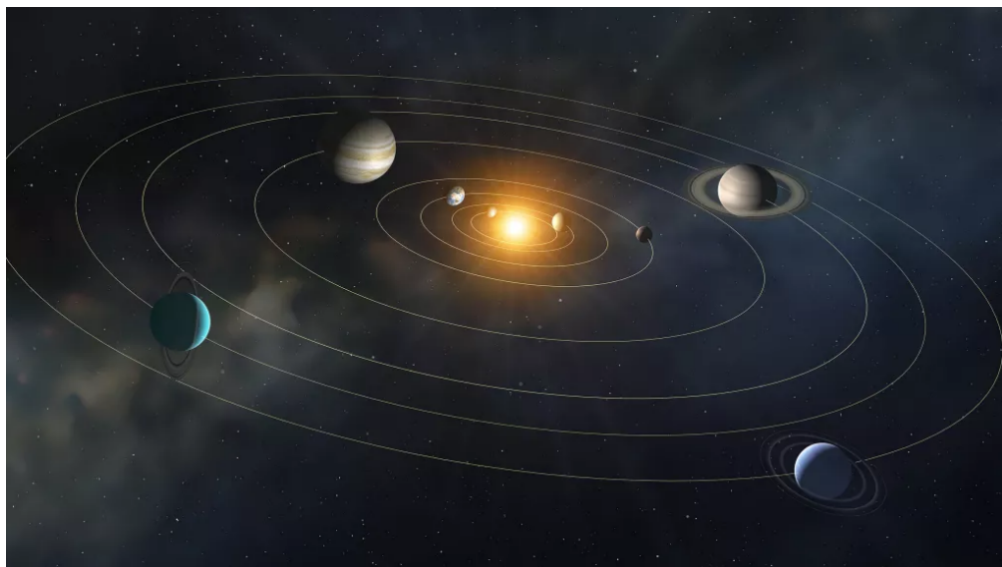
"The training set `Pima.tr` contains a randomly selected set of 200 subjects, and `Pima.te` contains the remaining 332 subjects. `Pima.tr2` contains `Pima.tr` plus 100 subjects with missing values in the explanatory variables."

- The [kaggle.com database](#) is yet another version: more records, one more variable - the "Metadata" information is missing
- Missing data are often coded as  $\emptyset$  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* (quantification)
- *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*
- In R, each discrete value or category is also called a *level*.

```
fv <- factor(c("male", "female"))  
fv
```

```
[1] male   female  
Levels: female male
```

## 9 TYPES OF CATEGORICAL VARIABLES

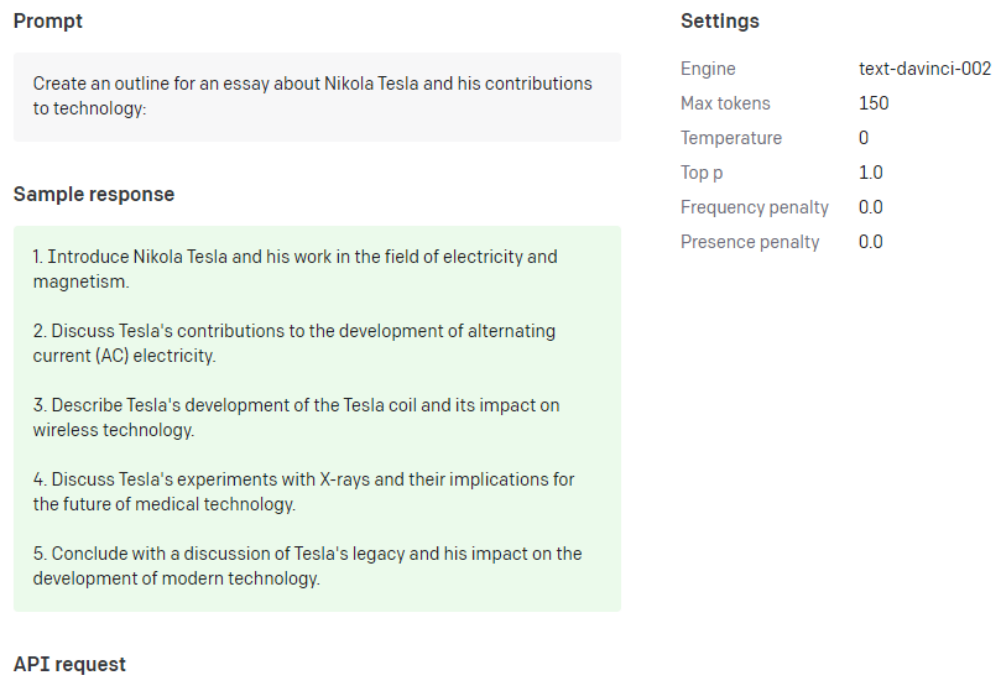


Figure 10: AI-generated outline for research topic (Source: OpenAI)

- 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<sup>1</sup>)

## 10 SOME ISSUES WITH GRAPHS

- Humans are better at seeing patterns in graphs than numbers<sup>2</sup>



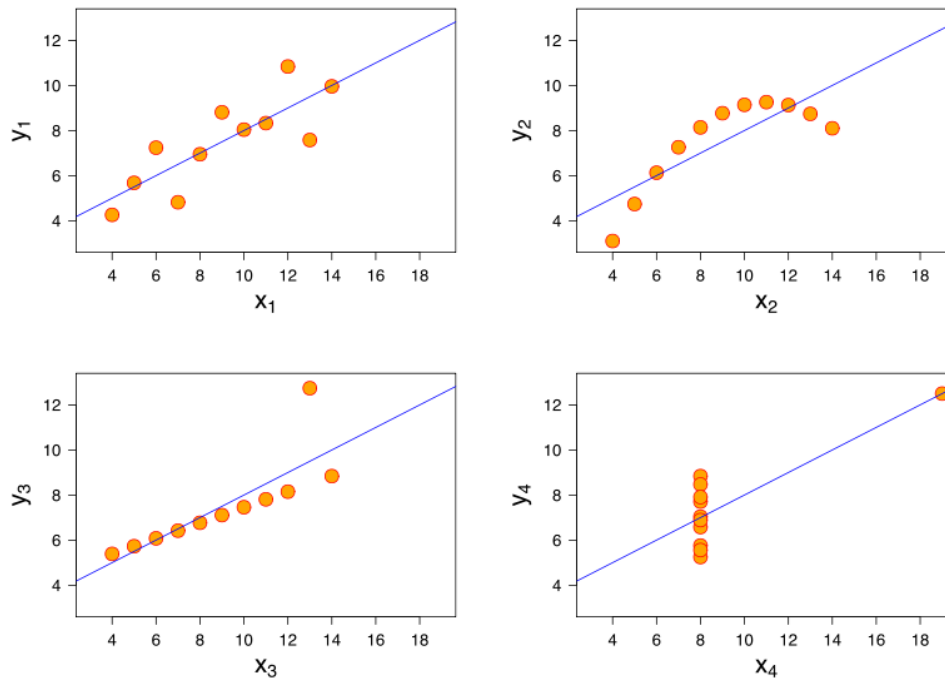


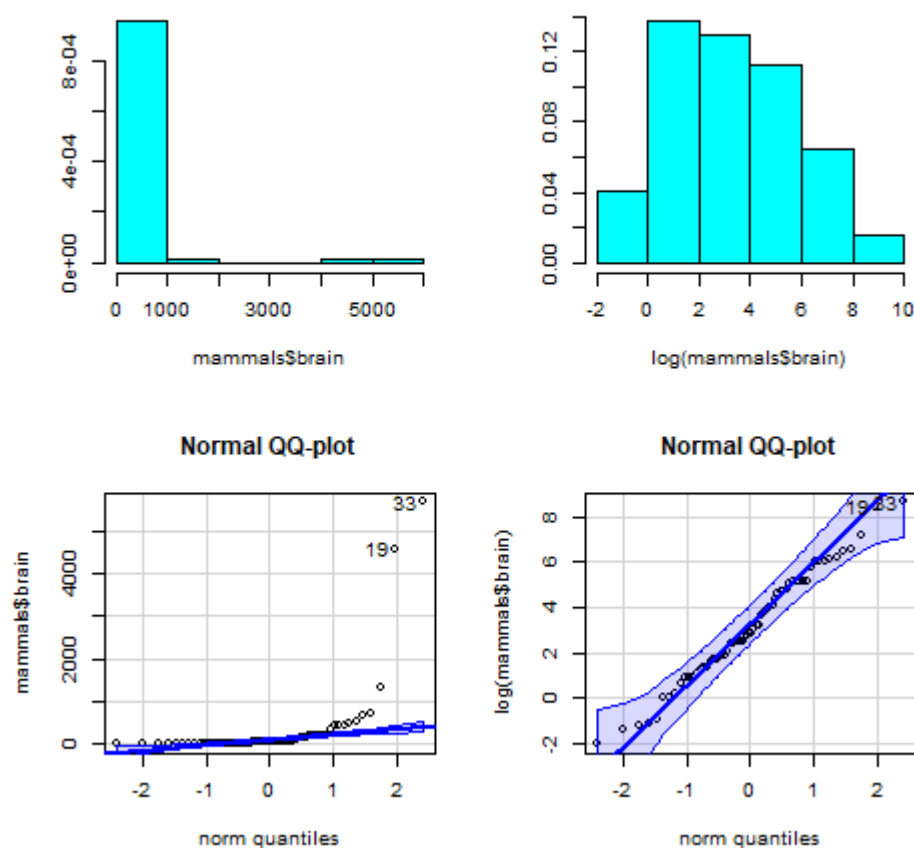
Figure 11: Anscombe dataset

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

## 11 PRACTICE: RAW VS. TRANSFORMED GRAPH DATA

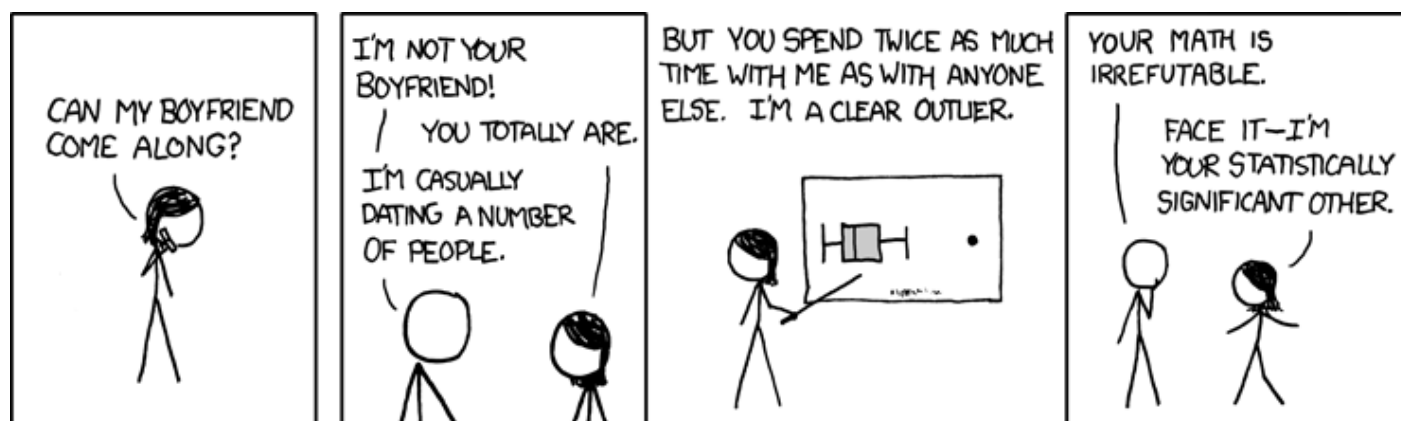
- The following two sets of plots are constructed from the brain element of the `mammals` dataset from the `MASS` package that lists body and brain weights for 62 different animals.
- **What do you think which graphs are more meaningful and why?**

```
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 tell 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



- 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

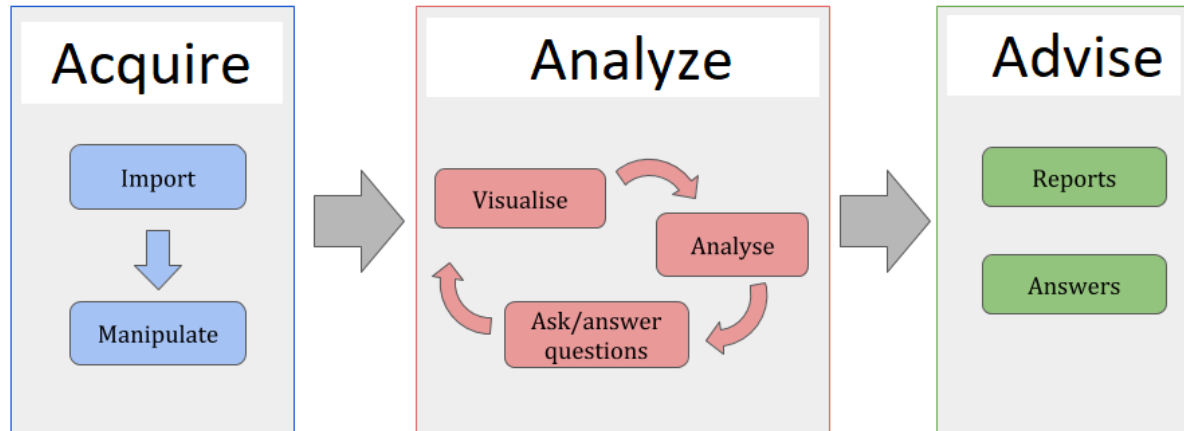


Figure 14: Data analysis workflow (emanuelaf.github.io - modified)

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 for decision-making support, hence the shift to storytelling and interpretation

## 14 COMPUTERS

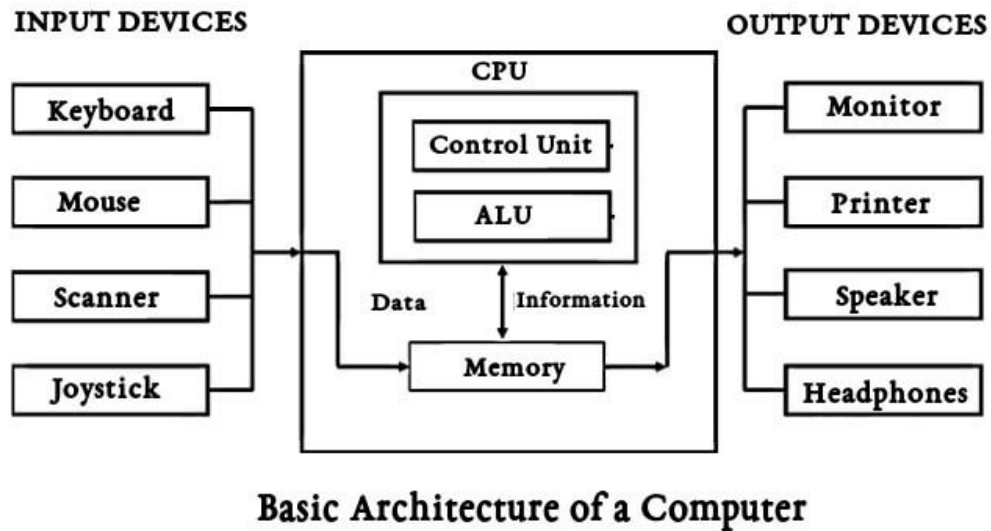


Figure 15: Von Neumann computer architecture (PSC Arivukal, 2020)

- 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<sup>4</sup>

## 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

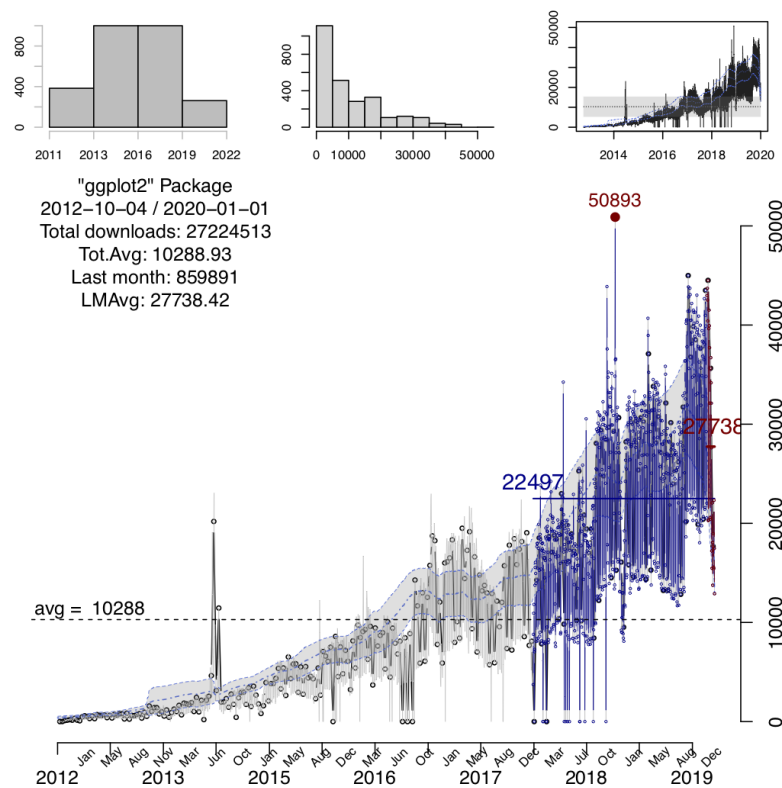


Figure 17: ggplot2 downloads from CRAN 2012-2020

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

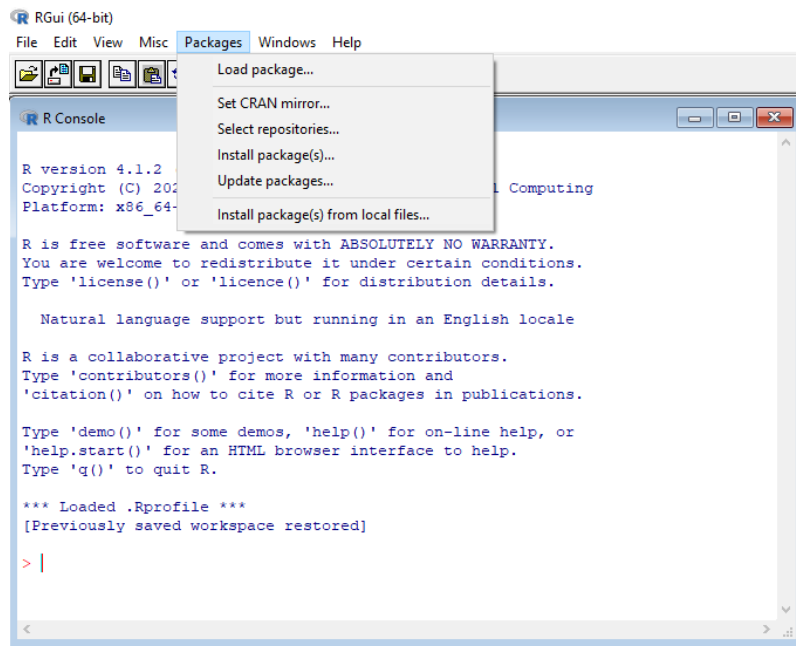


Figure 18: Package management in the Rgui program



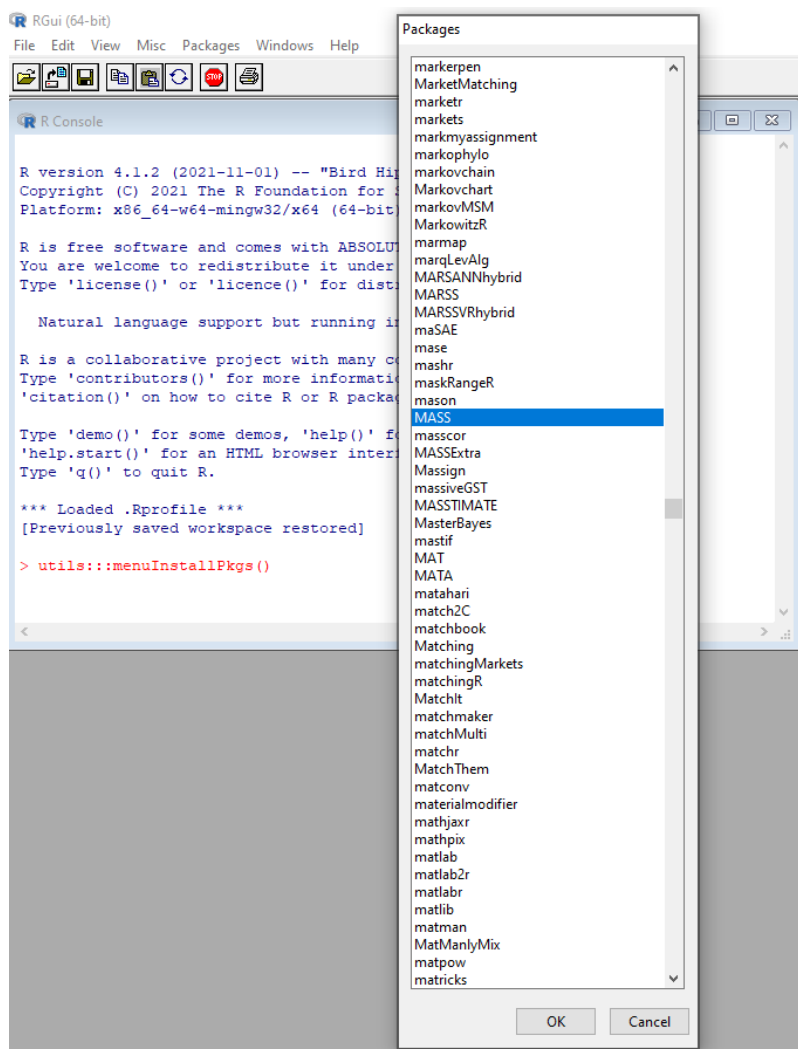


Figure 19: Package management in the Rgui program

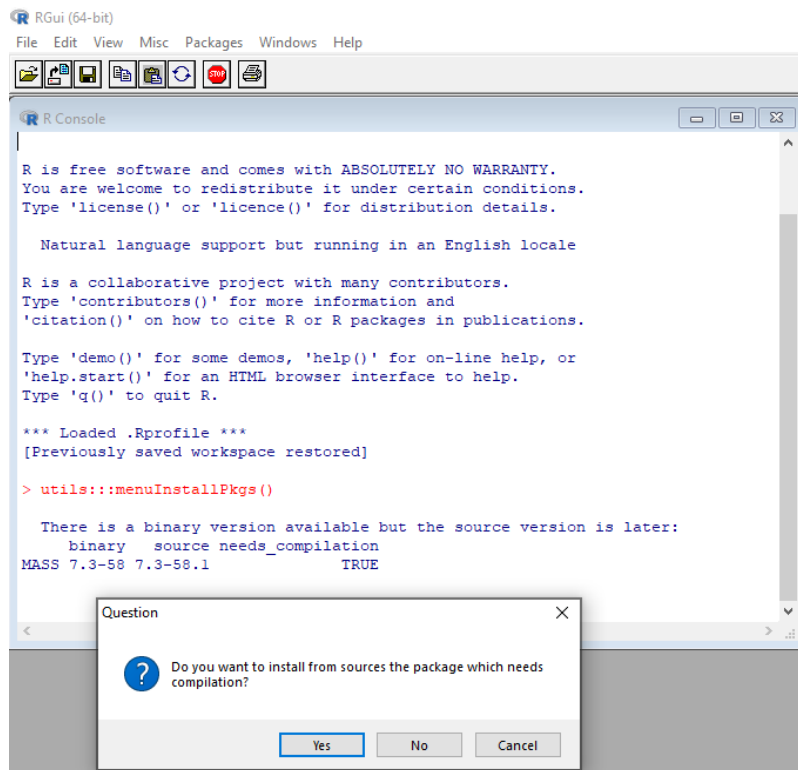


Figure 20: Package management in the Rgui program

## 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? (NA)
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 PRACTICE: A REPRESENTATIVE R SESSION



1. Open the course directory in GitHub, <https://github.com/birkenkrahe/dviz>
2. Open /org/2\_data\_eda\_R\_practice.org
3. Open the raw version of the file
4. Save file as 2\_data\_eda\_Rpractice.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](https://cran.r-project.org).
- OpenAI (2022). Example: Generate an outline for a research topic. [Online: beta.openai.com](https://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](https://pscarivukal.com).

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

## Footnotes:

<sup>1</sup> 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](#).)

<sup>2</sup> 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).

<sup>3</sup> 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.

<sup>4</sup> 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: 2022-09-22 Thu 18:03