Lancaster University

Quantitative Research Methods

January 30, 2023

Professor Patrick Rebuschat

p.rebuschat@lancaster.ac.uk

Our plan for today

Lancaster University

Homework assignment discussion

Data wrangling

- Reminder from last session
- Steps 0, 1 and 2 from Handout Session 3

Exploratory data analysis

- Short introduction
- Steps 3 and 4 from Handout Session 3

Our schedule

Lancaster University

Date	Topics
Jan 16	Introduction to quantitative research methods using R
Jan 23	Data management and data wrangling
Jan 30	Exploratory data analysis
Feb 6	Data visualization
Feb 13	No class, please complete the mid-term assignment (everybody)
Feb 20	Significance testing. Hypothesis tests for continuous variables: two groups.
Feb 27	Tests for discrete variables: Analysing contingency tables
Mar 6	Correlation and linear regression
Mar 13	Analysis of Variance (ANOVA) and tests for *N* groups
Mar 20	Multiple regression

Lancaster University

Solution to homework 2

Session 2 homework

Lancaster University

FASS512

Session 2 homework task

To complete this homework task, you will need to download the `language_exams` data file from our Moodle page into your working directory.

In the file, you will find the (made-up) scores and ages of 475 students who took an intermediate Portuguese language course at university. Students were tested three times: first in September to check their Portuguese proficiency at the beginning of the course, then again in January as part of their mid-term examination, and finally in June as part of their final examination. On each occasion, students had to complete three subtests to respectively assess their Portuguese vocabulary, grammar and pronunciation. The scores for exams 1, 2 and 3 are composite scores, i.e. each combines the results of the three subtests.

Your task is to run a basic analysis of the exam data using an R script. In your script, please include all the steps, including the command that loaded the data. Please also include sections to make your script very clear, as well as comments.

1. How many observations and columns does the data file contain?
2. Run commands to display the first and the last six lines of the table.
3. What is the average age of participants?
4. What type of variable is student_id?
5. What is the rounded mean score on exam 3?
6. What is the difference between the mean scores on exams 1 and 2?

Please save the script to discuss at the next session.

Session 2 homework

Lancaster University

```
# Script for homework task 2 ------------------------

# First, I install the tidyverse package

install.packages('tidyverse')
library(tidyverse)

# Here, I load the data, which I previously
downloaded from our Moodle page.
# I then view the table, just to make sure all
imported well.

language_exams <- read.csv('language_exams.csv')
View(language_exams)

# Question 1: How many observations and columns does
the data file contain?
# To answer this, I can check the Environment tab, or
use the following command.
# Answers: 474 obs. of  5 variables

str(language_exams)

# Question 2: Run commands to display the first and
the last six lines of the table

head(language_exams)
tail(language_exams)
```

Session 2 homework

Lancaster University

```
# Question 3: What is the average age of
participants?
# Answer: 21.5865

mean(language_exams$age)

# Question 4: What type of variable is student_id?
# Answer: "integer"

class(language_exams$student_id)

# Question 5: What is the rounded mean score on exam
3?
# Answer: 61

round(mean(language_exams$exam_3))

# Question 6: What is the difference between the mean
scores on exams 1 and 2?

mean(language_exams$exam_2) -
mean(language_exams$exam_1)

# Last but not least, we need to save the script.
# Answer: I have just clicked the Save icon in the
script editor. :-)
```

Session 2: Data management and data wrangling

Basics of data management, data wrangling. Second steps in R, including tidyverse.

Slides for Session 2

Handout 2 FASS512: Second steps in R

Please download this handout to complete during class.

Data sets to download for this session

Please download these data sets for our session (Fogarty, 2019; Winter, 2019). You need them to complete the handout.

R Script: The commands we used in Session 2

FASS512 Session 2 homework task

Here you will find the homework task and the data set that goes with it.

The solution (e.g., scripts) will be posted in this folder at the beginning of the NEXT session.

RScript: Commands used to solve homework task 2

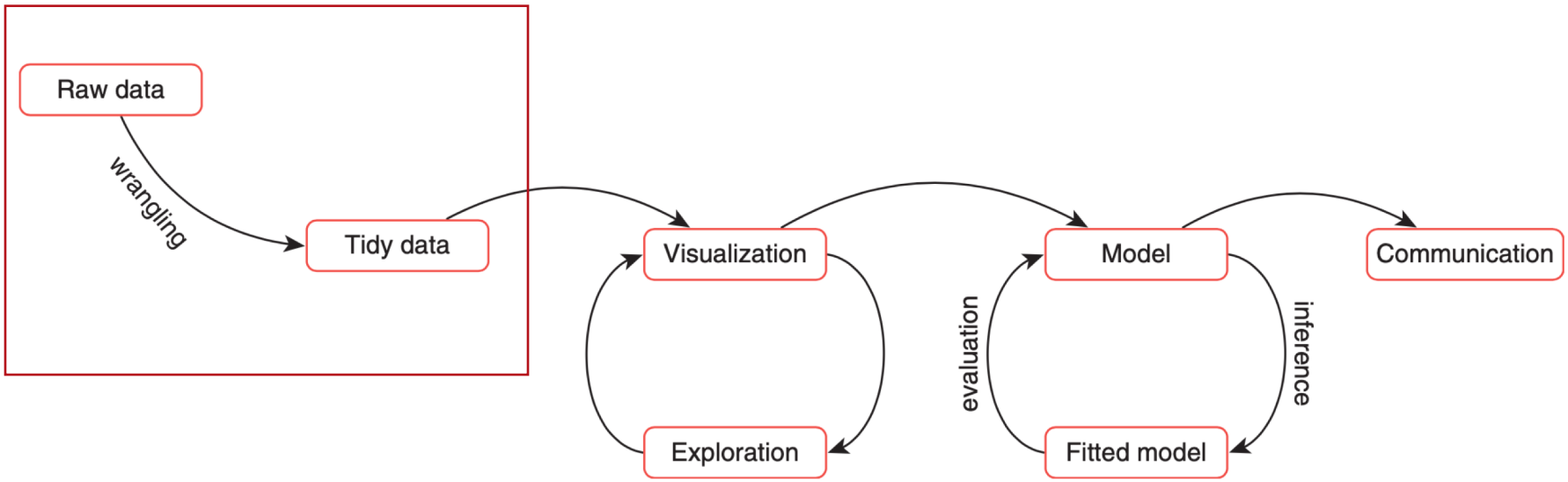
Restricted Available from **30 January 2023, 16:00**

The tidyverse style guide

Lancaster University

Data wrangling: A reminder

The Data Science Workflow



- Data science: the combined application of computational tools and statistical methods to (all aspects of) data analysis.

Reproduced from Andrews (2021)

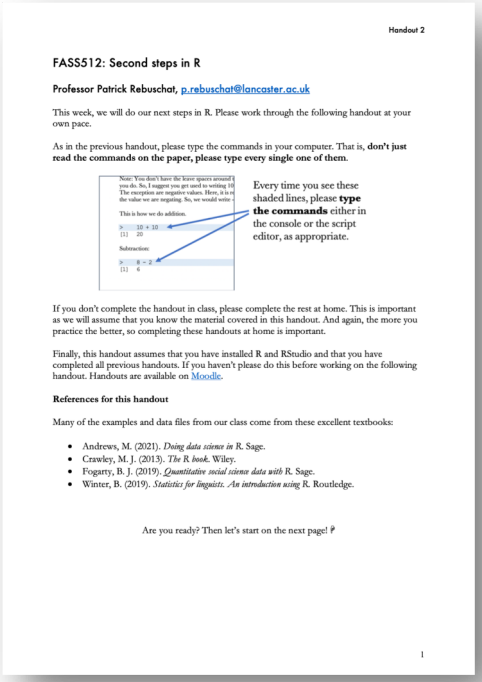
Handout 3

Let's complete the first part of the handout and do some wrangling…

Step 0: Installing and loading packages

Step 1: What is a tibble?

Step 2: Data wrangling in the `tidyverse`



Handout 2

FASS512: Second steps in R

Professor Patrick Rebuschat, p.rebuschat@lancaster.ac.uk

This week, we will do our next steps in R. Please work through the following handout at your own pace.

As in the previous handout, please type the commands in your computer. That is, **don't just read the commands on the paper, please type every single one of them**.

Every time you see these shaded lines, please **type the commands** either in the console or the script editor, as appropriate.

If you don't complete the handout in class, please complete the rest at home. This is important as we will assume that you know the material covered in this handout. And again, the more you practice the better, so completing these handouts at home is important.

Finally, this handout assumes that you have installed R and RStudio and that you have completed all previous handouts. If you haven't please do this before working on the following handout. Handouts are available on Moodle.

References for this handout

Many of the examples and data files from our class come from these excellent textbooks:

- Andrews, M. (2021). *Doing data science in* R. Sage.
- Crawley, M. J. (2013). *The R book*. Wiley.
- Fogarty, B. J. (2019). *Quantitative social science data with* R. Sage.
- Winter, B. (2019). *Statistics for linguists. An introduction using* R. Routledge.

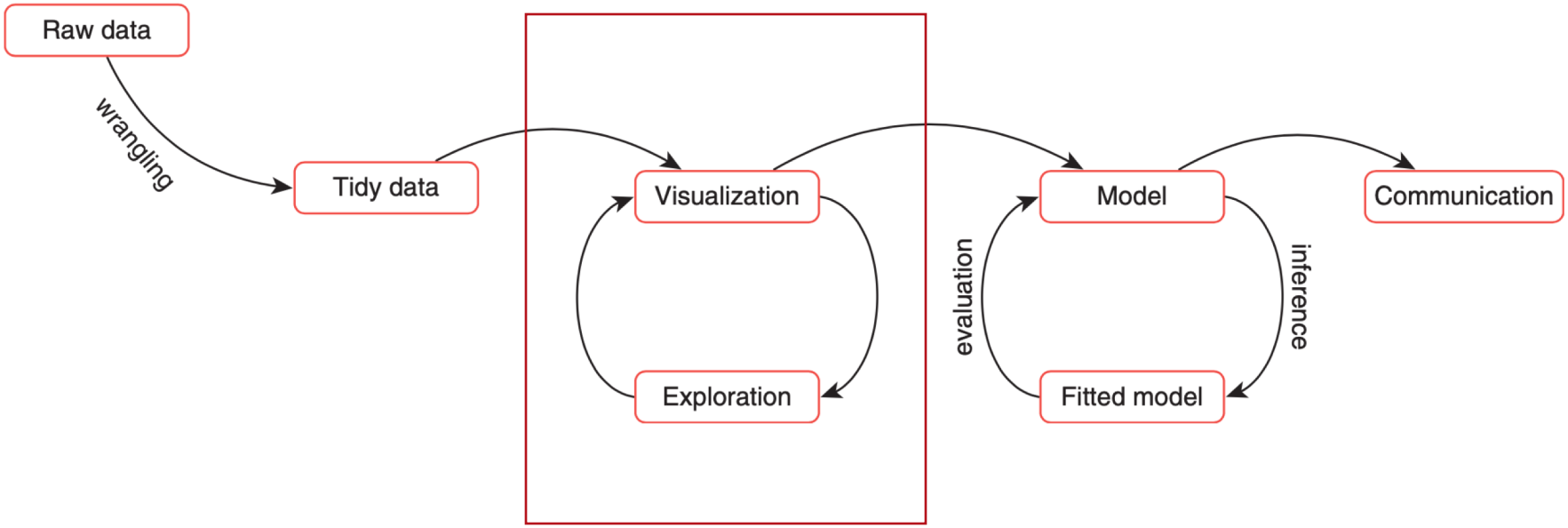
Are you ready? Then let's start on the next page! P

1

Lancaster University

Exploratory data analysis

The Data Science Workflow



- Data science: the combined application of computational tools and statistical methods to (all aspects of) data analysis.

Reproduced from Andrews (2021)

Tukey (1977)

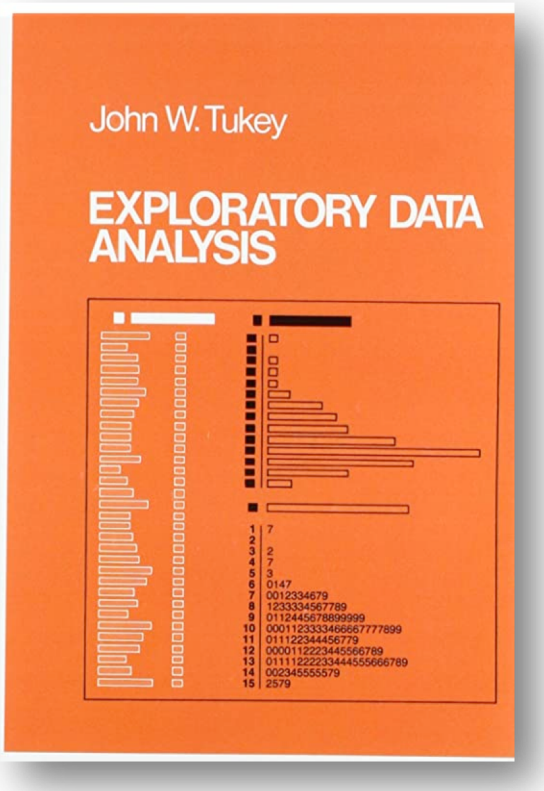
Lancaster University

Exploratory data analysis

- Aim: to discover potentially interesting patterns and behaviors in the data.

Confirmatory data analysis:

- Aim: to propose and test models of our data



Tukey (1977)

Lancaster University

Exploratory data analysis

- Detectives looking for evidence at the scene of a crime

Confirmatory data analysis:

- Courts making a prosecution case and evaluating evidence for and against

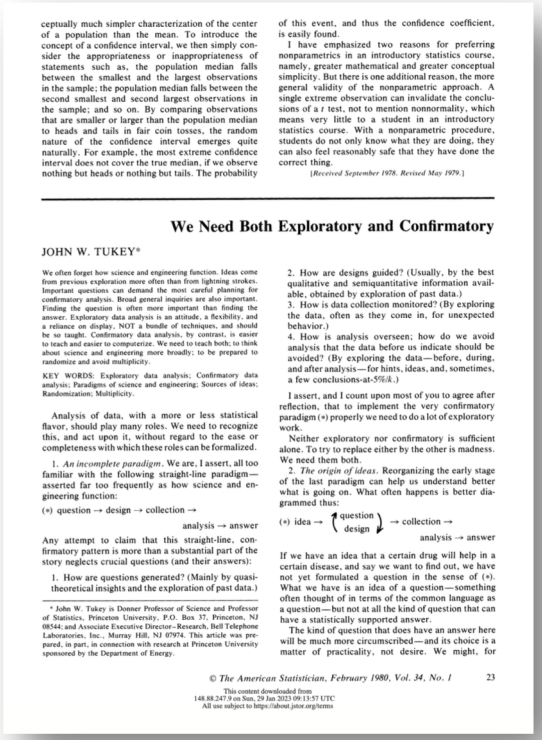
Image reproduced from here.

Tukey (1980)

We need both EDA and CDA.

EDA as an “attitude”

“Exploratory data analysis is **an attitude, a flexibility, and a reliance on display**, NOT a bundle of techniques, and should be so taught. Confirmatory data analysis, by contrast, is easier to teach and easier to computerize.” (Tukey, 1980, p. 23)



ceptually much simpler characterization of the center of a population than the mean. To introduce the concept of a confidence interval, we then simply consider the appropriateness or inappropriateness of statements such as, the population median falls between the smallest and the largest observations in the sample; the population median falls between the second smallest and second largest observations in the sample; and so on. By comparing observations that are smaller or larger than the population median to heads and tails in fair coin tosses, the random nature of the confidence interval emerges quite naturally. For example, the most extreme confidence interval does not cover the true median, if we observe nothing but heads or nothing but tails. The probability of this event, and thus the confidence coefficient, is easily found.

I have emphasized two reasons for preferring nonparametrics in an introductory statistics course, namely, greater mathematical and greater conceptual simplicity. But there is one additional reason, the more general validity of the nonparametric approach. A single extreme observation can invalidate the conclusions of a t test, not to mention nonnormality, which means very little to a student in an introductory statistics course. With a nonparametric procedure, students do not only know what they are doing, they can also feel reasonably safe that they have done the correct thing.

[Received September 1978. Revised May 1979.]

We Need Both Exploratory and Confirmatory

JOHN W. TUKEY*

We often forget how science and engineering function. Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis. Broad general inquiries are also important. Finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught. Confirmatory data analysis, by contrast, is easier to teach and easier to computerize. We need to teach both; to think about science and engineering more broadly; to be prepared to randomize and avoid multiplicity.

KEY WORDS: Exploratory data analysis; Confirmatory data analysis; Paradigms of science and engineering; Sources of ideas; Randomization; Multiplicity.

Analysis of data, with a more or less statistical flavor, should play many roles. We need to recognize this, and act upon it, without regard to the ease or completeness with which these roles can be formalized.

1. *An incomplete paradigm.* We are, I assert, all too familiar with the following straight-line paradigm—asserted far too frequently as how science and engineering function:

(*) question → design → collection → analysis → answer

Any attempt to claim that this straight-line, confirmatory pattern is more than a substantial part of the story neglects crucial questions (and their answers):

1. How are questions generated? (Mainly by quasi-theoretical insights and the exploration of past data.)
2. How are designs guided? (Usually, by the best qualitative and semiquantitative information available, obtained by exploration of past data.)
3. How is data collection monitored? (By exploring the data, often as they come in, for unexpected behavior.)
4. How is analysis overseen; how do we avoid analysis that the data before us indicate should be avoided? (By exploring the data—before, during, and after analysis—for hints, ideas, and, sometimes, a few conclusions-at-5%/k.)

I assert, and I count upon most of you to agree after reflection, that to implement the very confirmatory paradigm (*) properly we need to do a lot of exploratory work.

Neither exploratory nor confirmatory is sufficient alone. To try to replace either by the other is madness. We need them both.

2. *The origin of ideas.* Reorganizing the early stage of the last paradigm can help us understand better what is going on. What often happens is better diagrammed thus:

(*) idea → { question / design } → collection → analysis → answer

If we have an idea that a certain drug will help in a certain disease, and say we want to find out, we have not yet formulated a question in the sense of (*). What we have is an idea of a question—something often thought of in terms of the common language as a question—but not at all the kind of question that can have a statistically supported answer.

The kind of question that does have an answer here will be much more circumscribed—and its choice is a matter of practicality, not desire. We might, for

* John W. Tukey is Donner Professor of Science and Professor of Statistics, Princeton University, P.O. Box 37, Princeton, NJ 08544; and Associate Executive Director–Research, Bell Telephone Laboratories, Inc., Murray Hill, NJ 07974. This article was prepared, in part, in connection with research at Princeton University sponsored by the Department of Energy.

Tukey (1977, 1980)

Lancaster University

To explore, we can use summary statistics and data visualizations.

Example: **Stem-and-leaf display**

```
stem_example <- c(12, 24, 15, 15, 12, 24, 29, 22, 21, 25, 30,
39, 45, 50, 51)
stem(stem_example)
```

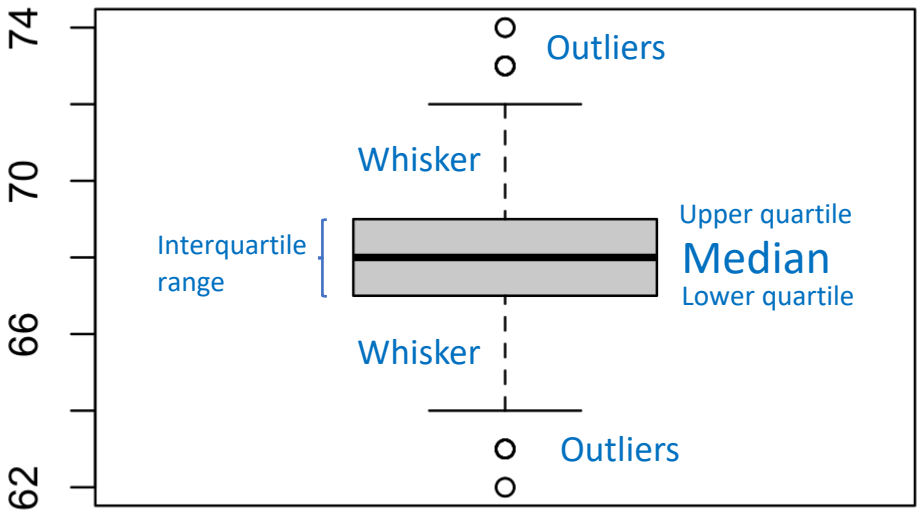
```
  The decimal point is 1 digit(s) to the right of the |

  1 | 2255
  2 | 124459
  3 | 09
  4 | 5
  5 | 01
```

Tukey (1977, 1980)

Example: **Boxplot**

- Upper quantile: 75%
- Lower quantile: 25%
- Median: 50%
- Interquartile range: Range between lower and upper quartile.
- Whisker length: 1.5*IQR from upper and lower quantiles respectively
- Outliers: Extreme values, each bubble represents one observation



Lancaster University

Types of data

Types of data

Lancaster University

We can classify data types in different ways.

1. Continuous data
2. Ordinal data
3. Count data
4. Categorical data

Continuous data

Lancaster University

- Variables can take any value in a continuous metric space.
- Between any two values (e.g., 0 and 100) can exist an infinite number of other values.
- Continuous data can be ordered.
- Example: height, weight, age; speed, time, distance.

Ordinal data

Lancaster University

- Values of a variable that can be ordered but there is no natural sense of distance between these values.
- First, second, third… These values have a natural order, but there is no sense of distance between them.
- Example: Students scoring in first, second, third place in a test. The first might score 100, the second 45, and the third 10. Or they might score 99, 98, 97. → The data can be ordered but the distance is

Count data

Lancaster University

- Tallies of the number of times something has happened.
- Values of count variable are ordered and also have a true sense of distance.
- Example: In an exam, a score of 87 correct answers is as far as from a score of 90 as a score of 97 is from 100.

Categorical data

Lancaster University

- Each value takes one of a finite number of values that are categorically distinct.
- Values of categorical variables are usually names, labels. Hence, also known as **nominal** data.
- Values of categorical variables cannot be placed in order, nor is there a natural sense of distance between them.
- Example: nationality, country, occupation, experimental conditions (experimental vs. control condition)

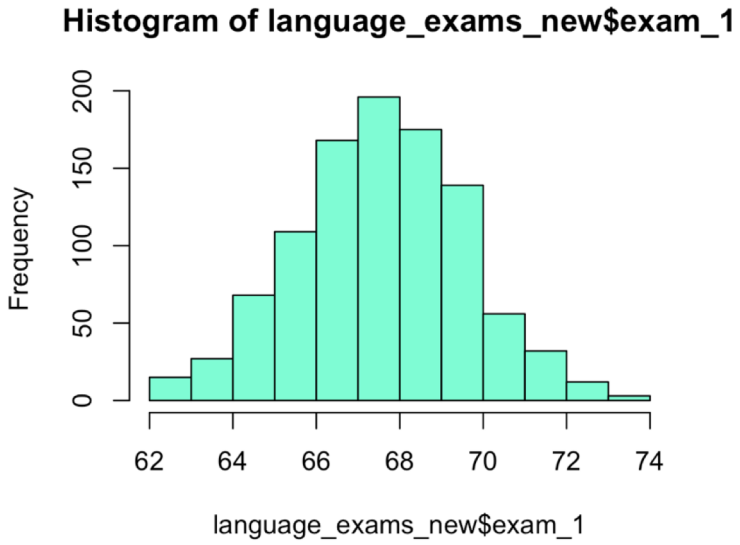
Lancaster University

Characterizing distributions

Characterizing distributions

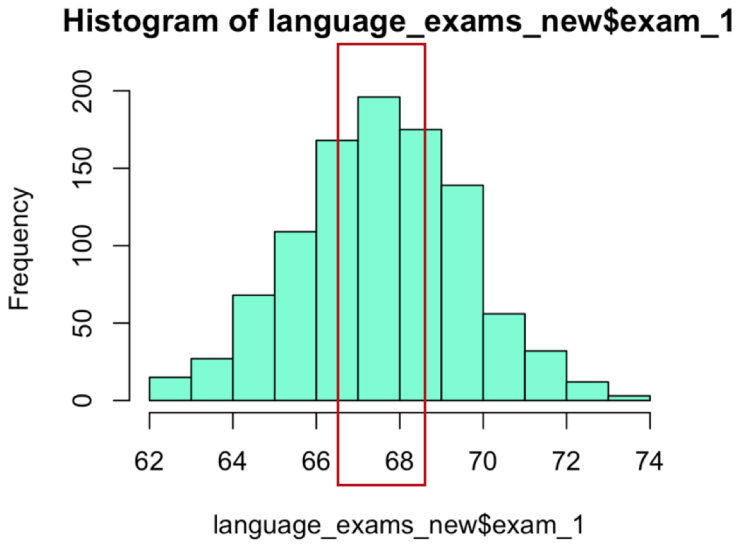
We can describe (univariate) distributions in terms of three major features:

- Location (central tendency)
- Spread (dispersion)
- Shape



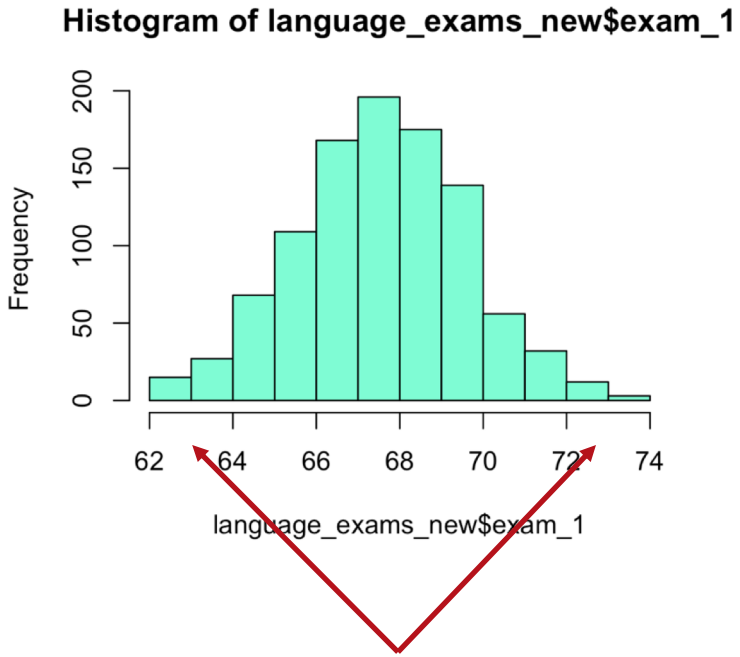
Location (central tendency)

- The location of a distribution describes where most of the value fall.
- Example: Most of the exam scores in the histogram.



Spread (dispersion)

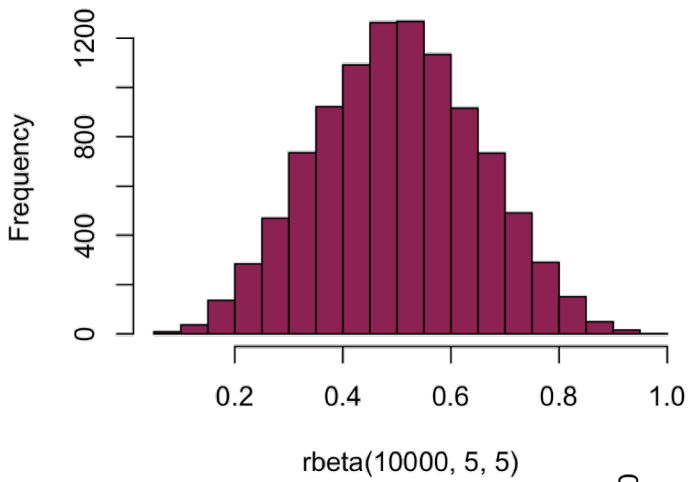
- Tells us how dispersed or spread out the distribution is.
- Are most of the scores clustered around the central value? → Short ”tails”
- Are they more spread out? → Long “tails”



”tails” of the distribution

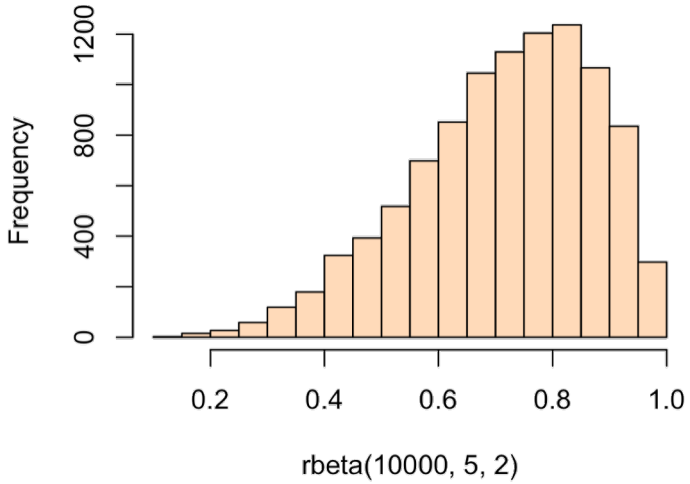
Shape: Skewness

normal

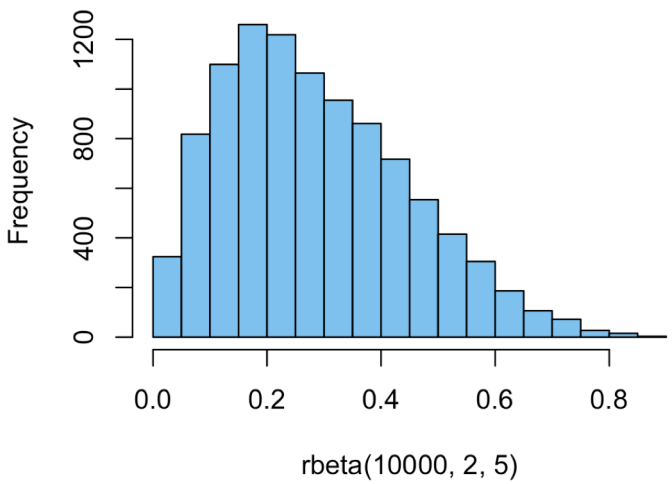


How much (a)symmetry there is in the distribution.

negative skew

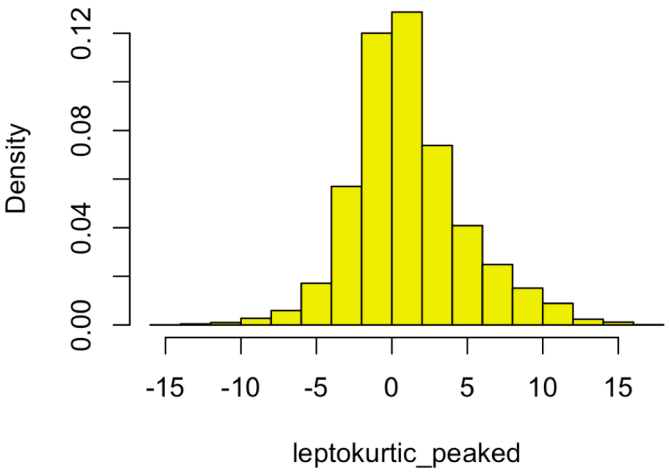


positive skew



Shape: Kurtosis

- Peaked: leptokurtic
- Flat: platykurtic



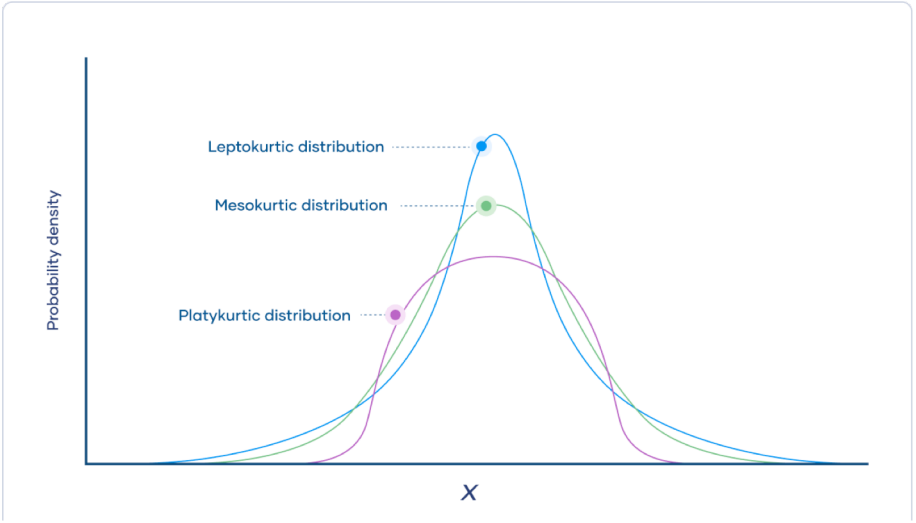


Image reference

Lancaster University

Summary statistics

Summary statistics (**models** of our data)

Lancaster University

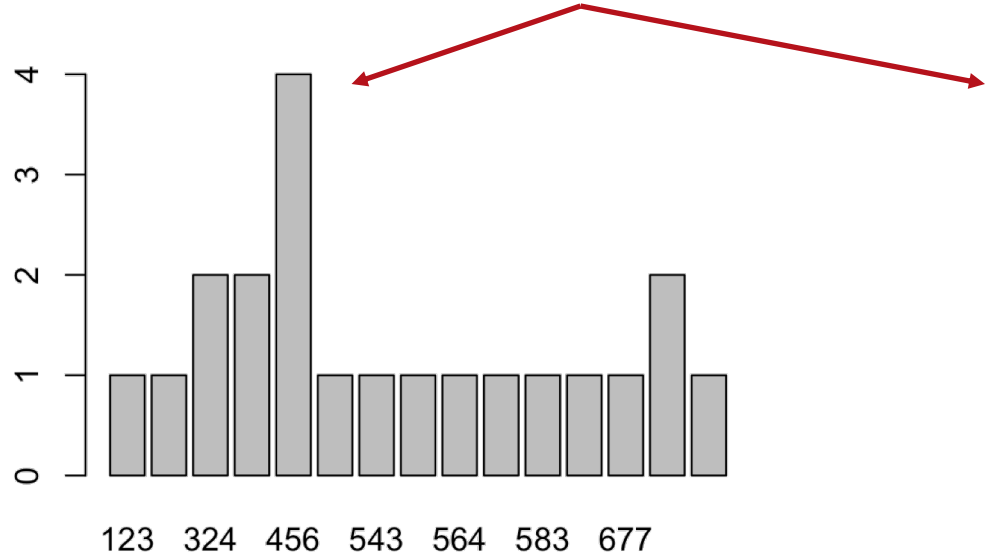
1. Measures of central tendency: mean, median, mode
2. Measures of dispersion: variance, standard deviation, range measures

Lancaster University

Measures of central tendency

Mode

The value with the highest frequency.



	Frequency	Percent
123	1	5
322	1	5
324	2	10
345	2	10
456	4	19
465	1	5
543	1	5
546	1	5
564	1	5
567	1	5
583	1	5
663	1	5
677	1	5
876	2	10
2890	1	5
Total	21	100

Can be used for categorical, continuous, count, and ordinal data.

Arithmetic mean

Lancaster University

- The mean: Sum of all observations (x_1 … x_n) divided by the number of observations (N)

$$\frac{x_1 + x_2 + x_3 + \ldots + x_n}{N}$$

- Can be calculated for continuous data and count data
- The most widely used measure of central tendency

Median

Lancaster University

- The middle point in a sorted list of values.
- Can be used for continuous, ordinal, and count data

To calculate the median:

- First sort the values, then find the middle point.

Median

Lancaster University

If there's an **odd** number of values:

- There is only one point in the middle of the sorted list of values. That's the median.

Data 2	Data 2 Ordered
123	123
543	324
456	456
546	489
876	543
324	546
489	876
Median	489

Median

Lancaster University

If there's an **even** number of values:

- Find the two points in the middle of the sorted list.
- The median is the arithmetic mean of these two points.
- Here, (465 + 564) / 2 = 514.5

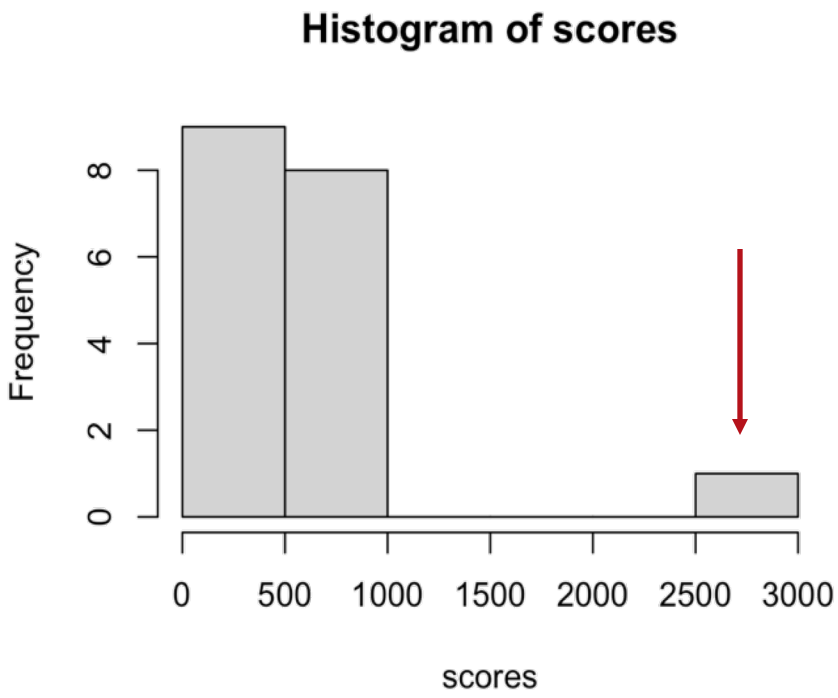
Data 1	Data 1 Ordered
564	324
677	345
345	368
465	465
368	564
583	567
324	583
567	677
Median	514.5

Median

Data 1	Data 1 Ordered		Data 2	Data 2 Ordered
564	324		123	123
677	345		543	324
345	368		456	456
465	465		546	489
368	564		876	543
583	567		324	546
324	583		489	876
567	677			
Median	514.5		Median	489

Robust measures of central tendency

- The **mean** is the most widely used summary statistic.
- But: It's not a very "robust" statistic as it's easily affect by extreme values (outliers).



Compare:

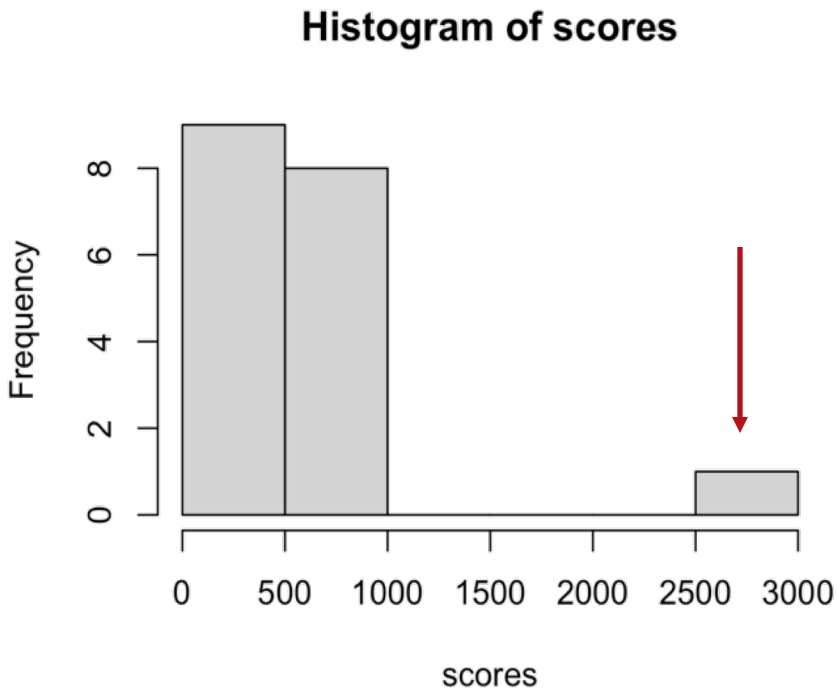
- M with outlier = **612**
- M without outlier = **499**

Robust measures of central tendency

- The **median** is a more robust statistics; it's not easily affected by extreme scores.

Compare:

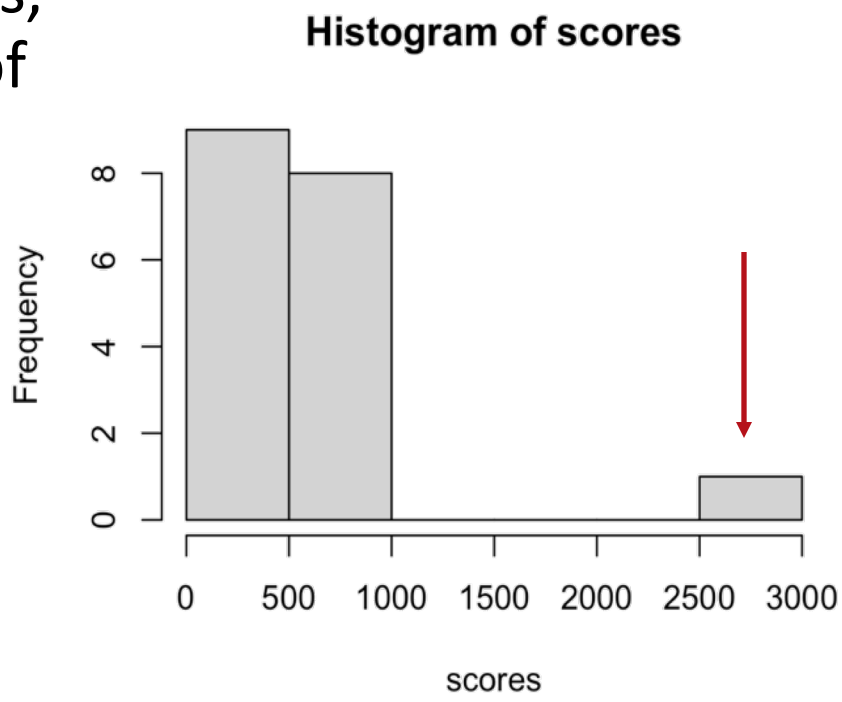
- Median with outlier: **465**
- Median without outlier: **461**



Compare mean and median

When there are extreme values, the median is a better model of our data.

- M *with* outlier = 612
- Median *with* outlier: 465
- M *without* outlier = 499
- Median *without* outlier: 461



So why not just use the median?

More robust alternatives to the standard mean

Lancaster University

- **Trimmed mean**: Extreme values are removed before calculating the mean as normal (`trim()` function)
- **"Winsorized" mean**: Extreme values are replaced with values at the thresholds of the extremes, e.g. the 90th percentile.

So, why not just always used these?

Whatever you choose, justify your decision and be transparent about it in the report.

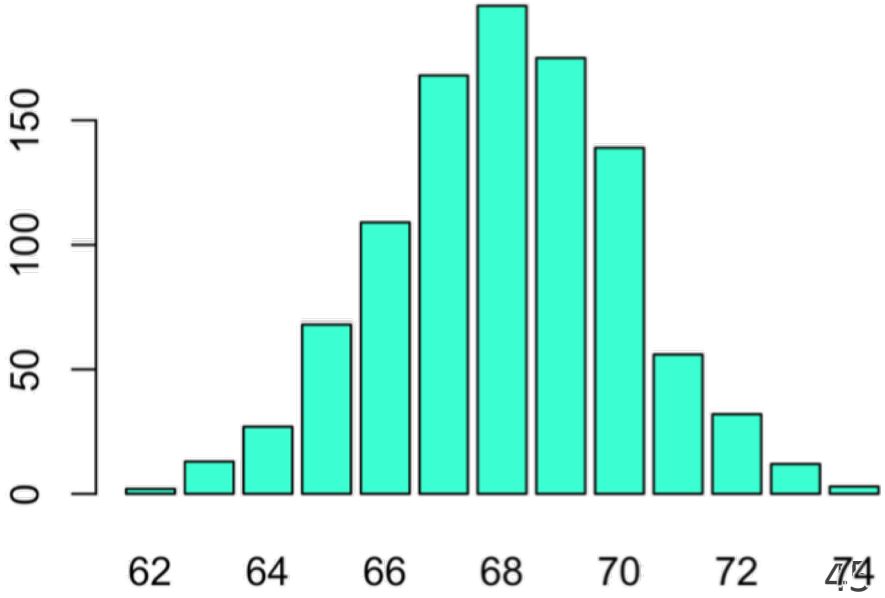
Lancaster University

Measures of dispersion

Measures of dispersion

- Tell us how the observations in our variables are spread out.
- Provide information about the variability in our data.

Best practice:
Report a measure of central tendency and a measure of dispersion (e.g., **M** and **SD**)



Standard deviation

$$SD = \sqrt{\frac{\sum (x - \bar{x})^2}{N - 1}}$$

Measure of dispersion around the **mean**

To calculate the SD:

1. Calculate the mean for the sample: $\bar{x}$
2. Calculate, for each data point (x), its difference from the mean and square this value: $(x - \bar{x})$
3. Sum up the squared differences: $(x - \bar{x})^2$
4. Divide the "sum of squares" $(x - \bar{x})^2$ by the number of observations minus 1 (N – 1).
5. Take the square the root of this number.

Range

Lancaster University

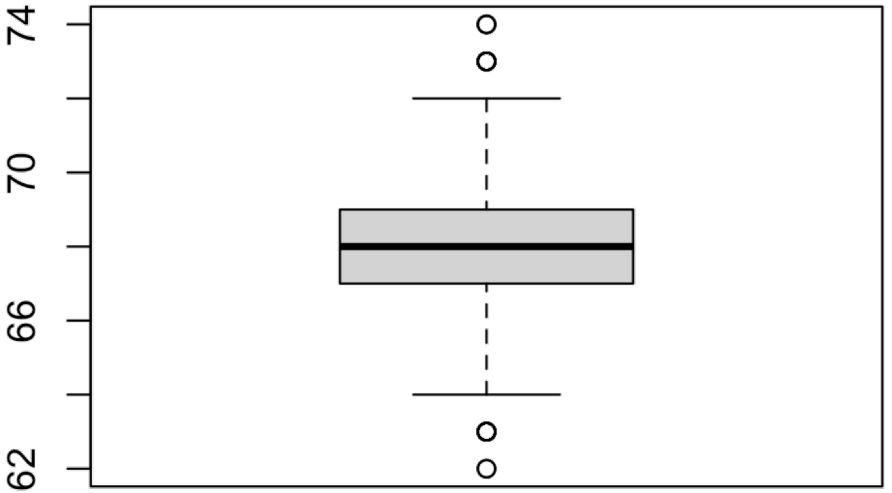
- Distance between the smallest and largest data point.
- Not very informative as it exclusively based on the most extreme values…

Example: age range in our handout

- Minimum age = 18
- Maximum age = 25
- Range = 7

Interquartile range

- Another way of measuring dispersion is by means of the interquartile range (IQR).
- This divides the sample data into quartiles (Q1, Q2, Q3 and Q4).
- Difference between Q1 and Q3 is called the **interquartile range**.



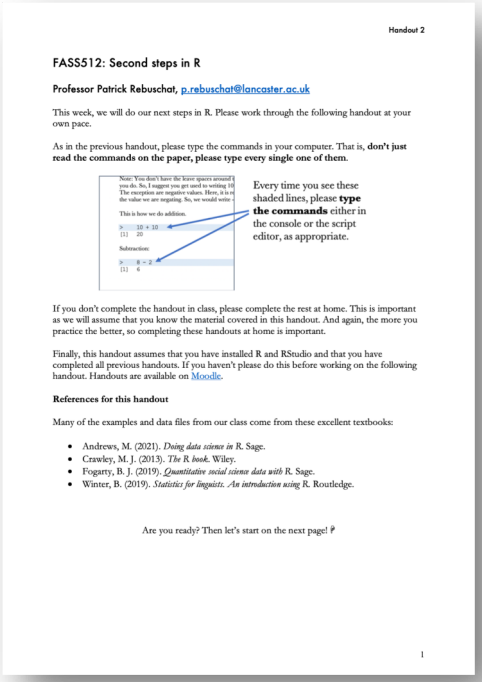
Lancaster University

Practical: Data wrangling and exploratory data analysis

Handout 3

Step 3: Exploratory data analysis

Step 4: Our first graphic explorations



Handout 2

FASS512: Second steps in R

Professor Patrick Rebuschat, p.rebuschat@lancaster.ac.uk

This week, we will do our next steps in R. Please work through the following handout at your own pace.

As in the previous handout, please type the commands in your computer. That is, **don't just read the commands on the paper, please type every single one of them.**

Every time you see these shaded lines, please **type the commands** either in the console or the script editor, as appropriate.

If you don't complete the handout in class, please complete the rest at home. This is important as we will assume that you know the material covered in this handout. And again, the more you practice the better, so completing these handouts at home is important.

Finally, this handout assumes that you have installed R and RStudio and that you have completed all previous handouts. If you haven't please do this before working on the following handout. Handouts are available on Moodle.

References for this handout

Many of the examples and data files from our class come from these excellent textbooks:

- Andrews, M. (2021). *Doing data science in* R. Sage.
- Crawley, M. J. (2013). *The R book*. Wiley.
- Fogarty, B. J. (2019). *Quantitative social science data with* R. Sage.
- Winter, B. (2019). *Statistics for linguists. An introduction using* R. Routledge.

Are you ready? Then let's start on the next page!

1

Lancaster University

Questions?

Lancaster University

Quantitative Research Methods

January 30, 2023

Professor Patrick Rebuschat

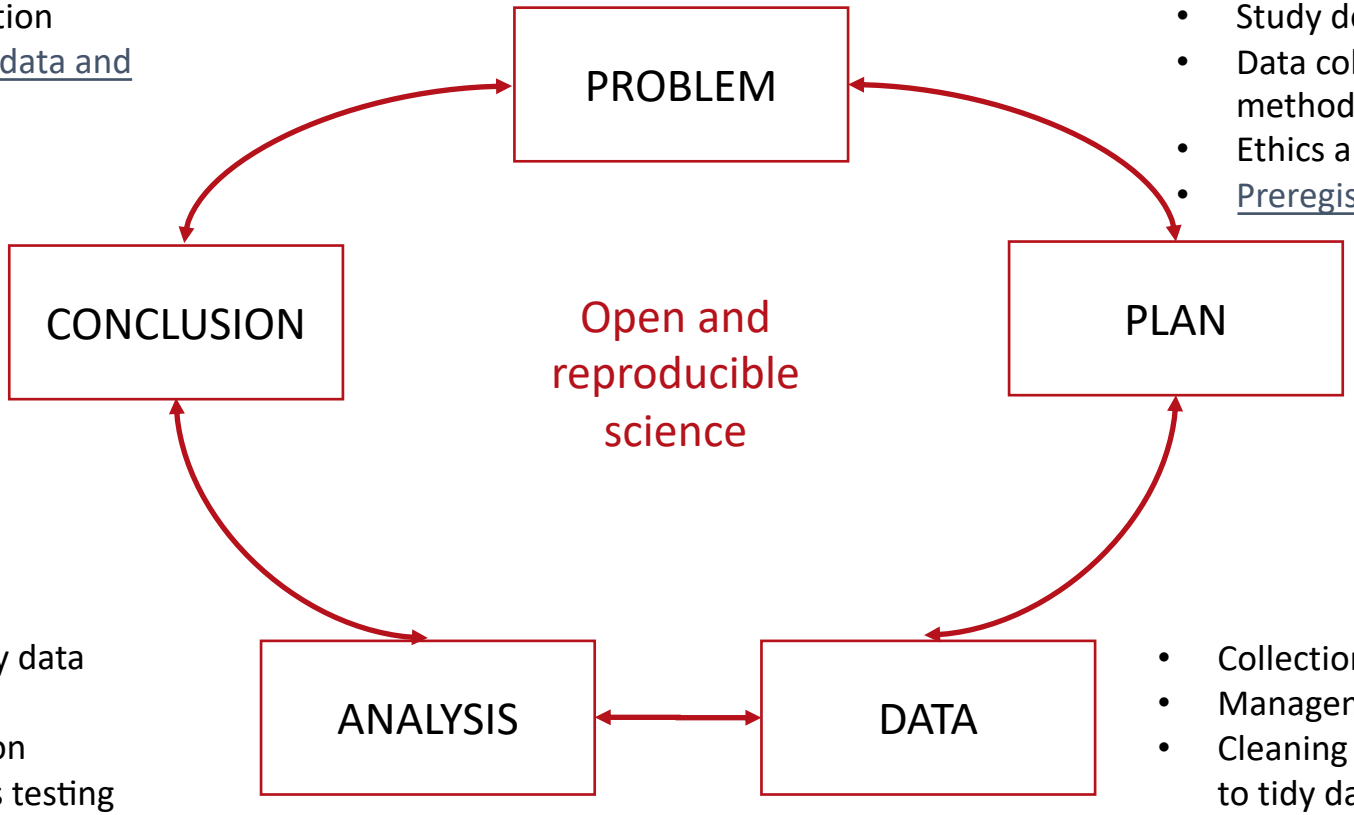
p.rebuschat@lancaster.ac.uk

The PPDAC Cycle

Adapted from Spiegehalter (2019)

- Understanding and clearly defining the problem
- Deciding on strategy to answer the question(s)?

- What to measure and how?
- Study design?
- Data collection methods?
- Ethics approval
- Preregistration?



- Interpretation
- Conclusion
- Dissemination
- Sharing of data and scripts
- New ideas

- Exploratory data analysis
- Visualization
- Hypothesis testing

- Collection
- Management
- Cleaning (from raw to tidy data)