Case Study Instructions

Data Analysis and Visualization in R (IN2339)

Daniela Klaproth-Andrade, Felix Brechtmann, Julien Gagneur

Why do we do a case study in this course?

- Opportunity to practice you R knowledge and to participate in a data analysis project.
- Practice you soft skills by working in groups and presenting results.
- Chance to improve by one grade step (0.3) in the final and repeat exam.

This year's case study: The city of Barcelona

- You are going to investigate data related to the administration, urban environment, population, territory, economy and business in the city of Barcelona
- Information of the Open Data BCN portal, the Ajuntament de Barcelona's open data service
- From Kaggle https://www.kaggle.com/xvivancos/barcelona-data-sets

Dataset

The dataset contains 17 .csv-files and is grouped into four categories:

1. Demography
   - births.csv: Births by nationalities and by neighborhoods (2013-2017)
   - deaths.csv: Deaths by age groups and by neighborhoods (2015-2017)
   - population.csv: Population by neighborhoods, by age groups and by genre (2013-2017)
   - unemployment.csv: Registered unemployment by neighborhoods and genre (2013-2017)
   - immigrants_by_nationality.csv: Immigrants by nationality and by neighborhoods (2015-2017)
   - immigrants_emigrants_by_age.csv: Immigrants and emigrants by age groups and by neighborhoods (2015-2017)
   - immigrants_emigrants_by_destination.csv: Immigrants and emigrants by place of origin and destination (2017)
   - immigrants_emigrants_by_destination2.csv: Immigrants and emigrants by place of origin and destination, respectively, and by neighborhoods (2017)
   - immigrants_emigrants_by_sex.csv: Immigrants and emigrants by sex by neighborhoods (2013-2017)
   - most_frequent_baby_names.csv: Most common baby names in the city by sex (1996-2016)
   - most_frequent_names.csv: Most common names of the inhabitants of Barcelona by decade of birth and sex.
   - life_expectancy.csv: Life expectancy by gender (2006-2013)

Dataset

2. Accidents
    - `accidents_2017.csv`: Accidents handled by the local police in 2017
3. Environment
    - `air_quality_Nov2017.csv`: Air quality information including measures of Tropospheric Ozone (O3), Nitrogen dioxide (NO2) and suspended particles ((PM10)
    - `air_stations_Nov2017.csv`: Air quality measure stations in the city
4. Transport
    - `bus_stops.csv`: Bus stops by day and night including airport bus stops
    - `transports.csv`: Public transports including underground, train, cable car, tramcar, etc.

Data access and further information on the dataset

- The data can be downloaded from Moodle or from Kaggle: https://www.kaggle.com/xvivancos/barcelona-data-sets
- See Kaggle for a more detailed explanation of the data and the features in each file

Formalities: registration to the case study

- The case studies are done in groups of **exactly 4** students:
  - Find your groups on your own
  - No more and no less students than 4
  - Use the Slack workspace for communication
- To register for participation upload exactly **one .csv-file** "registration.csv" containing your matriculation number, first and las name(s)
  - One file for the whole group including the information of every group member
  - Use the script registration_to_case_study.R from Moodle to easily create that file and upload the file to Moodle
- The deadline to submit the registration file is on **Jan 9, 2022 at 23:59**
- We will not accept
  - registrations after the deadline or
  - registration files which are not in the required format
  - registration files which do not contain the complete correct required information

Formalities: Submission via Moodle

- For passing the case study you will have to submit the **final** version of your case study via Moodle
  - One submission per group
  - Submission deadline on **Jan 23, 2022 at 23:59**
- Submit one `.zip`-file group named as `submission.zip`. The `.zip`-file should contain:
  - R markdown file containing all analysis
  - `.pdf`-file generated from the R markdown file containing **at most 5 pages**
- All submissions will be checked by the tutor team. No bonus applied for `pdf`-files with more than 5 pages.

Formalities: Optional presentation of your case study

- A few volunteers will virtually present their final version of the case study during the last lecture on Feb 1st 2022
  - This is a chance to show what you have accomplished and listen to interesting presentations from other groups

Your task in the case study

The submission should contain the following for passing the case study:

1. Motivation and goals

- What is the concrete goal and focus of your case study? To which questions/problem do you aim to find an answer with the performed analysis?

2. Data Preparation
    - The needed data preparation steps (e.g. merging, filtering, subsetting) should be contained in the `.Rmd`-file but not neccesarily in the `.pdf`-file
3. Data Analysis
    - Include at least 1 **descriptive plot**
    - Come up with at least 1 hypotheses/claim and support it with a **demonstrative plot**
    - Make at least 1 **statistically supported claim** and visualize it
    - Show an example where controlling for **cofounding factors** was necessary to support a claim or invalidate the hypothesis **or** implement one prediction task and show its performance
4. Conclusion
    - Finish with a conclusion recapping the main findings

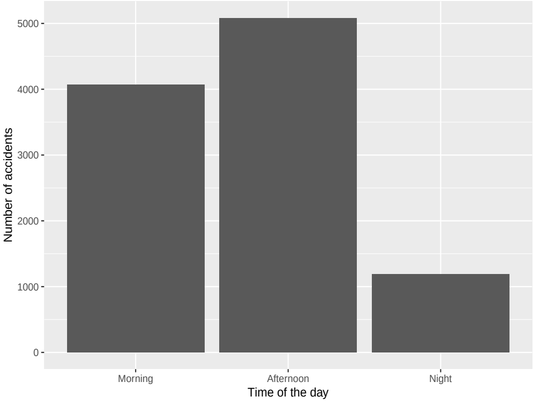
A few recommendations for the sumbmission

- Think about story telling: build up an exciting and clear story where you dig step by step into details. The ultimate goal is to explore the data, make claims and support these claims.
- The `.Rmd`-file has to be reproducible and contain your complete code for the analysis. However, in the compiled `.pdf`-file you can ommit chunks of code by setting `echo=FALSE` in the chunk options. In this way, you can show only relevant code: visualizations and plots are usually more interesting than long chunks of code.
- Read and follow other advices described in the script
  - e.g. slide titles, legible labels, color guidelines

Template

- We created a template `.Rmd`-file to help you conduct your analysis with the needed requirements
  - Download it from Moodle!
- Here is a first example of a plot based on the accidents dataset:

In 2017 more vehicle accidents occurred during the afternoon in Barcelona



FAQs

Q: Are we are allowed to use any additinal data?

A: Yes, any publicly available dataset. Include the source.

Q: Is the bonus also applicable for the repeat exam?

A: Yes, applicable for final and repeat exam.

Q: Are we allowed to form groups of more or less than 4 people?

A: No, groups with more or less than 4 people will be splitted or merged together.

Q: Is the bonus from last year also applicable for this year?

A: The bonus from last year (WS2021) is applicable. The bonus from WS1920 or earlier is not applicable anymore.