Project Instructions

Intro to Statistics and Data Analysis with R (0560.1823)

Adi Sarid

May 2020

Background

The following document contains instructions to the project in the Introduction to Statistics and Data Analysis with R course.

The project has a weight of 30% of you final grade.

Goal

The goal of the project is to demonstrate and practice the different elements we have been talking about, which are a part of most data analysis/data science projects.

Methods

In this project you will handle the different phases of a data analysis project:

1. Data **Import** (reading the data into R).
2. Data **Tidying** (arranging the data into something you can work with)
3. Understanding the data:
    a. **Transforming** variables.
    b. **Visualizing** (using ggplot2 to show distribution of variables, relationships between variables, and to hypothesize).
    c. **Modelling**: using a few of the tools we have learned during the course (like hypothesis testing, regression, analysis of variance, etc.) to examine your hypothesis.
4. **Communicate** your findings via an oral and a written report

Instructions and Schedule

The project should be performed **in pairs**.

Choosing a dataset

First, you should select a dataset on which you will perform the project. I recommend using a data set from either Kaggle or from tidytuesday. You can select something else.

In any case, please do not choose something "too popular" (e.g., no built-in `R` datasets). **Check with me before you start**, so that I can confirm your dataset: email adi@sarid-ins.co.il with:

- The dataset name
- Source (a url with the data and documentation of the dataset)
- A direct link to download the raw data you will be using

Please choose a dataset and email me the details by **June 7th 2020** (you can do it earlier and start the work on the project once I email you back with the green light).

Consultation

We will have a lab session in which the pairs will be able to present their work to me (not to the entire class), and ask questions about specific models, difficulties they had with coding, or any other related questions.

Submission

Final submissions should be made by **July 10th 2020.**

Please submit your file to moodle as `statintro_final_studentname_studentID.zip` which bundles an Rmd version, data files, and a knitted html version of your report. The Rmd should compile standalone in every computer.

Grading

You will be graded along the following lines:

- Data import and tidying (10%): Your ability to use the proper methods to import the data, and tidy it towards the next stages.
- Visualizations (20%): Your ability to utilize visualizations to articulate your hypothesis and to illustrate different patterns and relationships in the data. You should be able to match the proper types of charts to what ever it is you are trying to show.
- Modelling (20%): Your ability to match the appropriete statistical tests/models to the problem, verifying (or highlighting) certain assumptions which are valid or invalid in this case. Please provide at least two relevant models/hypothesis tests that we learned.
- Communication, documentation, explanations (20%): You should be able to explain the different steps you are doing, lead the reader in a logical and appealing manner, explain your results, and highlight the research or business implications of your findings.
- Code (15%): Readability, proper use and proper documentation of code.
- Oral exam (15%): A one-on-one session with each student (5 minutes), just to make sure you are familiar with the work (that you actually worked as a pair).

Good luck!

Appendix: Questions and answers

Some more questions and answers.

How should you report the results?

In tests such as t-test or goodness of fit, you should explain in plain text what you are doing, what assumptions the test entails and if they indeed hold in this case or not. The add the code chunk and include the output.

For example, in linear regression, you should also report a qqplot of the residuals and check homoscedasticity.