Project Instructions

Intro to Statistics and Data Analysis with R (0560.1823)

Adi Sarid

2022-05-09

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 40% 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** (use `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. **Communicating** your findings via a written report

Instructions and Schedule

The project should be performed **in pairs** (same groups of homework submissions).

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 (kaggle.com) or from tidytuesday (https://github.com/rfordatascience/tidytuesday), or government data (https://data.gov.il/dataset/). You can select something else.

In any case, please do not choose something “too popular” (e.g., no built-in `R` datasets, and no data sets that we’ve worked on in the lectures).

In your work you must document:

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

Consultation

I’m dedicating a weekly reception hour, Thursdays 09:00, in zoom. You can bring questions regarding the project, coding, `R` , etc. Please coordinate in advanced (send me an email if you want to join the reception hour).

Submission

Final submissions should be made by **June 10th 2022.**

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, tidying, and transformations (20%): Your ability to use the proper methods to import the data, tidy it and apply required transformations 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.

- Modeling (20%): Your ability to match the appropriate 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. For example, make sure you start with data description, research questions, hypothesis, etc.

- Code (20%): Readability, proper use and proper documentation of code. You may use tidyverse code or base R.

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. Then 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.

Where can I see examples for projects?

You can see examples for projects from previous semesters here (https://github.com/adisarid/intro_statistics_R/tree/master/project/examples).