Data Analysis Project

(Due 12 noon Friday March 8, 2019)

For the Data Analysis Project you are tasked with sourcing and analyzing a data set and writing a concise report describing your work. The report is worth 25% of your final grade for the course.

There are three stages to the project, described here with their relevant weighting (marks out of 25):

Stage 1. Source an interesting data set (8 Marks)

In this digital information age there is no shortage of publicly available data available for analysis. You are required to find an interesting data set that you will then analyse and report on as a group. You can decide what 'interesting' means, but here are a few requirements:

- The data must be uploaded into R (usually via a .csv file or similar). You will need to submit the data file along with your final report (see "Submission Instructions" below).
- The data shouldn't have more than 500 observations. You may choose to subset observations from a larger data set, either randomly or by certain criteria.
- You must describe the context and origin of the data and describe the key variables of interest (this may be a subset of all the variables in the original data set). **Do not** include or describe variables that you do not use in the analysis.
- You must state one or two research questions of interest, which the data can be used to begin to explore using the methods covered in Data Analysis (see Stage 2).
- These sites list a number of data sets that you may find interesting (although you are not limited to these sites):
    - https://data.world/datasets/regression
    - https://www.kaggle.com/datasets?sortBy=relevance&group=public&search=regression&page=1&pageSize=20&size=all&filetype=all&license=all
    - http://archive.ics.uci.edu/ml/datasets.html?format=&task=reg&att=&area=&numAtt=&numIns=&type=&sort=nameUp&view=table

Stage 2. Analyse an interesting data set (12 Marks)

Having posed your research question(s) in Stage 1, proceed to analyse your data with a view to answering the research questions. Your analysis will be limited to the methods we have covered in Data Analysis, namely:

- Illuminating visualizations of the data;
- Suitable numerical summaries displayed in an informative way;
- Appropriate linear models fitted to the data (and assumptions checked);
- Relevant confidence intervals for estimates and model parameters and model selection.

All of this analysis must be conducted using **tidyverse** functions in R and contained in an **R Markdown** file (.Rmd). No graphs or summaries or any other output can be 'copied and pasted' from another source, but all of your analysis must be *reproducible* from the R Markdown file.

Stage 3. Write a report on an interesting data set (5 Marks)

You will then describe your work in Stages 1 and 2 in a group report written using R Markdown such that:

- all R output, including figures and tables, are appropriately labeled;
- R code should **not** be included in the body of the report; and
- the report should be no longer than **7 pages**.

Your report should include:

- An appropriate Title;
- An Introduction detailing the data set and question(s) of interest (see points in Stage 1);
- An Exploratory Analysis of the data (Stage 2);
- A Formal Analysis of the data using linear models and confidence intervals (Stage 2); and
- Conclusions summarizing the findings and limitations of your analysis.

Submission Instructions

You must submit **your report** as a **.pdf** file **and** its corresponding **.Rmd** file **and** its corresponding data file (usually **.csv**) via the *DA Project Submission* links in the "DA Project" section on the Data Analysis Moodle page. Please include the "*matriculation number*" in the .pdf/.Rmd/.csv filenames.

Note on Declaration of Originality

Every student must make a *Declaration of Originality*. These forms will be included together in a Moodle form alongside the .pdf / .Rmd submission links and must be completed before 12 noon Friday March 8, 2019.