STA 404/504 Homework 1

Due Monday Feb 10th, by 5:00pm, submit through canvas

Learning Objectives:

* Reading data into R
* Identifying basic R object characteristics
* Building basic visualizations
* Writing short descriptions of plot stories

The dataset “student-mat.csv” on canvas is about student’s achievement in the Math subject in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. This is one of the available datasets. (Cortez and Silva, 2008)

You may refer to this website for more information about the attributes and the data: <https://archive.ics.uci.edu/ml/datasets/student+performance>

Please follow the following steps to load the dataset into R, and complete the questions. After you finish, please submit:

1. **An R code file** showing your work and clearly mark the question number using comments. The instructor will be able to replicate your results by running the code you provided.
2. **A pdf file** that is converted from word containing the code, the outputs (including the plots) and your answers to the questions.

Note: Alternatively, you may use Rmarkdown to generate a pdf, or html file, including all your answers and discussions. In this case, please upload the .Rmd file, as well as the knitted html/pdf file.

First, you may run the following code to load the tidyverse package and read the data into R. After running it, R will show the column specifications.

library(tidyverse)

math <- read\_csv("student-mat.csv") #if you set the working directory to where you store the data

or

math <- read\_csv("the file path on your computer/student-mat.csv")

check whether “math” is a data frame, by

is.data.frame(math)

Then, answer the following questions (show the answers with R code and output, when creating the graphs, make sure appropriate title, labels or notations are appropriately displayed):

1. [1pt] How many students are recorded in the dataset for school Gabriel Pereira? How many are there for the school Mousinho da Silveira?
2. [1pt] What percent of the students choose school based on school reputation?
3. [1pt] For all the students in the data with a first period grade greater than 10, what percent of them are female?
4. [1pt] What is the average and standard deviation of second period grade for the students who study 5 to 10 hours per week and has extra-curricular activities?
5. [2pt] Students spend various time travel to school and their weekly study various. Based on your analysis, what kind of travel time and study time has the highest average final grade?

For the following questions, focus on the students in the school Gabriel Pereira.

1. [3pt] Create a graph that shows the relationship between the students first period grade and their final grade. Write a few sentences discussion about your opinion about such relationship.
2. [3pt] Create and describe a plot that tells us about the distribution of the number of school absences.
3. [3pt] How does the final grade differ by students’ travel time? Use a side-by-side style of plot to help tell a story of how they differ. Please make sure the label you use for travel time is reader friendly, do not use 1,2,3,4, etc to represent the students’ travel time type. Describe in a few sentences what this display tells us about the relationship between the variables. Why did you choose this type of display?
4. [2pt] When you do the plot in part 8, any problems do you find? How you will edit the graph? Provide a new graph that satisfies your need.
5. [3pt] Create an additional plot of your choosing that tells an interesting story about this data. Describe the plot in a few sentences. The chosen plot or associated-story should be different than those already mentioned above.

The items below will be considered for grading

* Correct values are used in the calculation.
* Axis labels and titles are correct and complete.
* Labeled units.
* The plots are correct, with professionalism.
* Proper grammar in the write-up.
* The story told is correct.

Reference

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.