STOR 565 Project 1 Instructions

Machine Learning, Spring 2021

Introduction

This document outlines the guidelines and expectations for Project 1 of STOR 565 (Spring 2021). The purpose of this project is to give you experience performing exploratory data analysis on a real dataset and communicating about data. These are both very important skills for being successful as a data scientist and practitioner of machine learning.

Format of the Project

You will be randomly assigned to groups of 3 students. The group will then select a dataset to analyze either from the list below or another source (see below for guidelines). You will then be asked to answer a series of questions about the data using R and compile your analyses into a single report. A template and list of questions to answer in your report is included in the accompanying R markdown file *Project1.Rmd*. A single final report for each group should be turned in on March 18 before the start of class. Please turn in the R markdown file, knitted PDF file, dataset, and any other files necessary to reproduce your results.

In addition to the 25 questions listed in the template, we ask that you include the following in your report:

- Abstract (no more than 1 paragraph)
  - Give a high level overview of what your report contains.
- Introduction (1-2 paragraphs)
  - Discuss the dataset you will be exploring. Where is it from? What kind of information does it contain? Why is it interesting? What is its potential importance?
  - Discuss a few questions about the data that you have before starting the analyses. Talk about how you will address these questions in your analyses.
- Conclusion (1-2 paragraphs)
  - Give an overview of the findings of your analysis. Were your questions answered? Did any new questions arise?

Grading

The project will be worth 100 points total with the following breakdown:

- Abstract: 5 points
- Introduction: 10 points
- Questions: 75 points (3 points each)
- Conclusion: 10 points

You will find that most of the questions are somewhat open-ended and involve both writing some code as well as providing a written response. In order to receive full points on a given question, you must fully address the question, justify your answers, and present your response in a clear manner. For details on what the latter means, please refer to the next section.

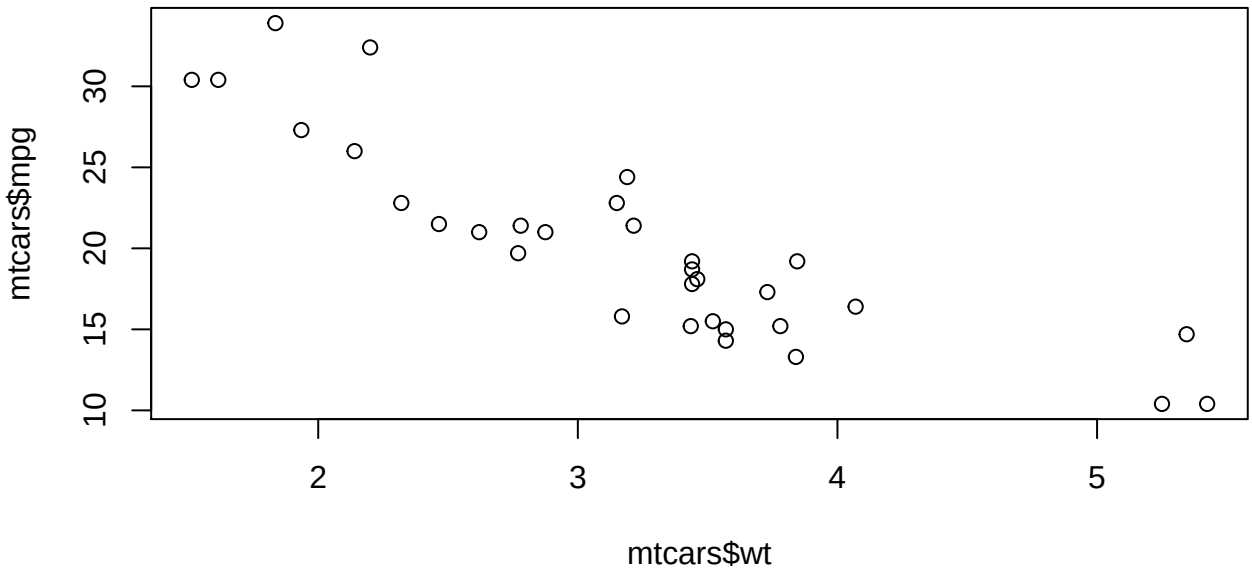
Expectations

As discussed above, one of the points of this project is to give you practice at communicating about technical material in a coherent and easily understandable way. As such, we expect you to answer all questions in complete, grammatical sentences. Where appropriate, you should restate the question in your answer so that it is clear what questions you are answering based on your response.

Aside from clear writing, another important aspect of communicating about data is visualizing data. The point of visualizing data is to make it easier for someone to understand some aspect of the data without

examining the data in detail. In order to achieve this, it is important to ensure that your graphics are clean and properly labeled. Here is an example of a poor visualization.

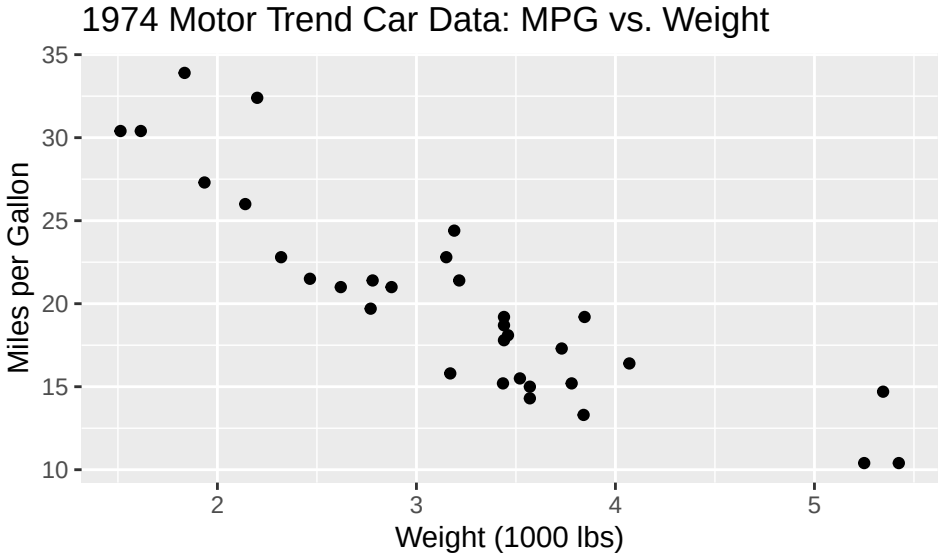
```
plot(mtcars$wt, mtcars$mpg)
```



Why is this a poor visualization? For someone who has worked with the *mtcars* dataset before, it may be clear what is being plotted. However, someone who has not seen the data before cannot quickly identify what information is being displayed. Moreover, someone who does not know R may be confused about what the dollar sign notation in the axis labels means.

Here is one way to display this information in a clearer way.

```
ggplot(mtcars) +
  geom_point(aes(mtcars$wt, mtcars$mpg)) +
  xlab("Weight (1000 lbs)") +
  ylab("Miles per Gallon") +
  ggtitle("1974 Motor Trend Car Data: MPG vs. Weight")
```



Just adding labels and cleaning up your plots can significantly improve the experience of the reader and

overall interpretability of your results. All plots should be properly labeled and sized to fit nicely on the knitted PDF. Note that you can change the size and alignment of a figure in R markdown by setting options such as *fig.width=5*, *fig.height=3*, and *fig.align="center"* in the header of the code chunk in which the figure is produced. The use of ggplot for generating graphics is not necessary but is strongly encouraged.

Finally, we ask that you include comments in your code. This is common practice in professional data science and can make it much easier for someone else to work with your code.

Datasets

Your group should pick one dataset to analyze throughout the project. You may either pick one of the datasets in the list below or submit one of your own. Datasets for this project should contain at least 50 samples and 10 numerical variables. Moreover, there should be at least one categorical variable that could reasonably correspond to the class of each sample (for example digit in the MNIST data). If you would like to use your own dataset, please submit it to the IA (Kevin O'Connor) for consideration by Thursday 2/25. In your email, please give a general description of the data including whether you have worked with it before and in what capacity. If you will be using one of the suggested datasets below, it will not be necessary to submit anything for consideration.

The following are the suggested datasets for the project:

- MNIST handwritten digits
  - This dataset is a collection of grey-scale images of handwritten digits. You can read more about the dataset here: http://yann.lecun.com/exdb/mnist/.
  - You may easily load the data into R using the *dslabs* package (https://rdrr.io/cran/dslabs/src/R/read_mnist.R).
- TCGA gene expression data
  - An anonymized dataset of gene expression values from 2000 genes measured for 217 breast cancer tumors. Each tumor has been labeled according to its subtype: *Normal* or *Basal*.
  - You can read more about the Cancer Genome Atlas here: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga.
  - This dataset can be found in the file *TCGA_sample.txt*.
- The Federalist papers
  - A dataset containing the frequency of occurence of 70 different words and authorship for the 85 Federalist papers, written in 1787.
  - Historical evidence shows that Jay authored 5 papers, Madison 14, and Hamilton 51. A further 3 were joint efforts between Madison and Hamilton (labeled *COL*). The remaining 12 are disputed (labeled *DIS*).
  - This dataset can be found in the file *Federalist.txt*.