ISE 5103 Intelligent Data Analytics
Homework #2

Instructor: Charles Nicholson

See course website for due date

Learning objective: Explore and visualize data.
Submission notes:

1. You will submit a PDF file with your solutions. Additionally, you will provide the R code you created to address the problems. The PDF is primarily what will be graded. The grader *may* view your R code, but should never *have* to in order to find your solutions.

2. In the PDF, clearly identify each problem (e.g., Problem 1a, Problem 2b, etc.) Also, note that only *relevant* and informative computer output should be provided.

3. Make sure to *provide comments* on what your R code is doing. Keep it clean and clear!

4. You will submit your complete R script. Note: include `library` commands to load *all* packages that are used in the completion of the assignment. Place these statements at the top of your script.

5. Do not zip your files for submission. Submit exactly two files. Name the files "LastName-HW1" with the appropriate file extension (that is, .pdf for the write-up and .R for the script)

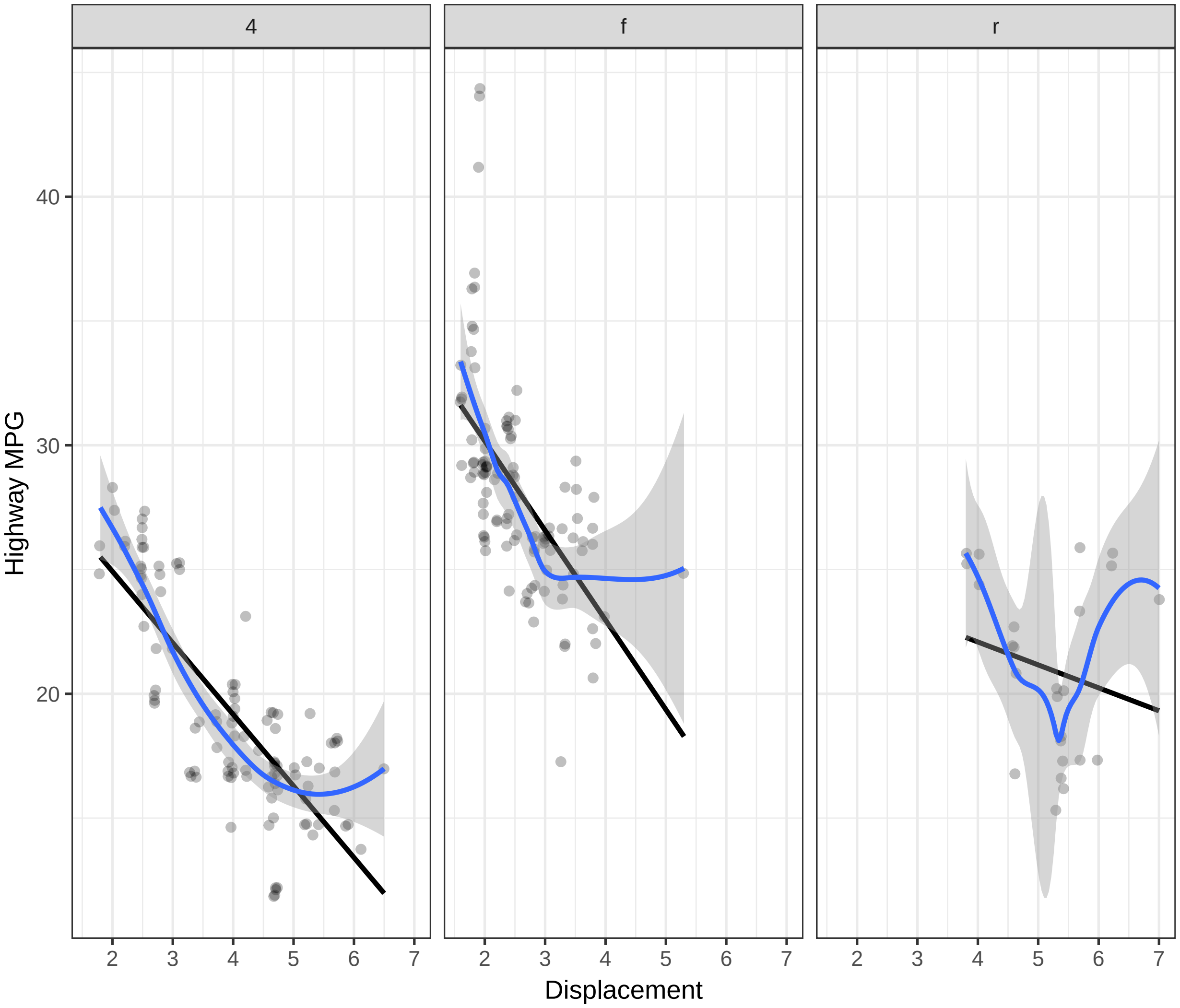
1 Learning `ggplot2` (50 points)

For this problem you will read through and work some of the exercises from Chapter 3 of the online book "R for Data Science". The book can be found here: `http://r4ds.had.co.nz/`. These questions are relatively easy, but the material in the book is great for learning `ggplot2`. Please provide any related code and graphs along with your answers for each problem.

(a) (30 points) Please address the following questions from Chapter 3 of "R for Data Science":
   - 3.2.4 Exercises #4, #5
   - 3.3.1 Exercises #3, #4, #6
   - 3.5.1 Exercises #4

(b) (20 points) After reading this chapter, you should be ready to reproduce the plot in Figure 1 using the same `mpg` data from above. Please do so. Make sure you notice the *jitter* and *alpha* levels, notice that there is both a *loess* smoothing and a linear smoothing (in black), and also, that the x and y axes are labeled.

Figure 1: Please reproduce this visualization for the `mpg` data.



2 House prices data: Exploratory Data Analysis and Visualization (50 points)

The housingData.csv file in the course website is real data associated with 1,000 residential homes sold in Ames, Iowa between 2006 and 2010. The data set includes over 70 explanatory variables – many of which are factors with several levels. The file housingVariables.pdf provides a concise explanation of the variables and the factor levels in the data.

We will use this data set again in class. In preparation of that, perform some basic exploratory data analysis and visualization of the data to get an idea of what is here.

Specifically, using ggplot2, create at least 5 different, non-trivial, visualizations of the data that you believe are informative.

You do not have to analyze every variable! However, I encourage you to play around with different possibilities and present the *best* ones. For each visualization, you must comment briefly (1-3 sentences) on what is useful/informative in the visualization.

Various possible visualizations include (but are not limited to) scatter plots with trend lines, sploms, parallel histograms, ridgeline polots, overlaid density plots, stacked bar charts, parallel plots, heatmap of correlations, missing value visualizations, tree maps, etc.

You might want to check out https://www.r-graph-gallery.com/index.html for some ideas.

Question: What does the professor mean when he says: "non-trivial" visualizations?

Answer: What I mean is *push yourself to find something interesting in the data.* Do not simply produce 5 scatterplots and call it a day. Use different visualizations. Use color, alpha, jitter, other layers, and comparisons to help find and tell a story. Grading on "non-trivial" is entirely subjective – and I do *not* like lazy when it comes to visualizations... Just do something great and you'll be fine.