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)

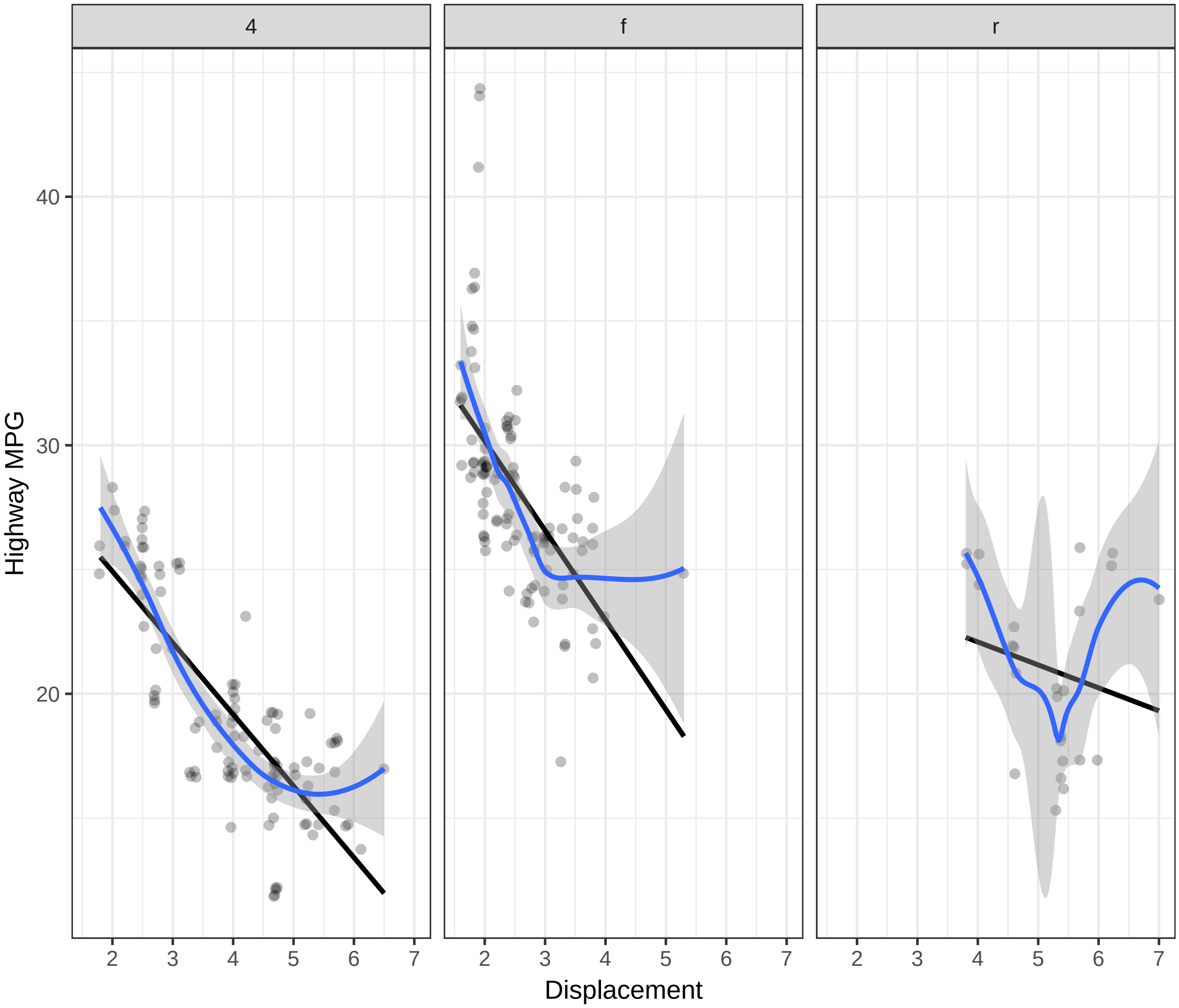
1 Learning `ggplot2` (45 points)

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) (15 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 Generating data and advanced density plots (30 points)

(a) (10 points) Create a data frame (or a tibble) named `df` with 500 rows and 4 variables: `a, b, c` and, `d`. Each variable should contain data generated randomly from a *different type* of distribution (e.g. `rnorm` generates normally distributed data randomly; there are several other similar commands available that you need to lookup).

The data frame will look something like the following example:

a	b	c	d
1.2	0.4	-0.1	4.9
0.9	1.3	0.9	-0.7
…			

Create data frame (or a tibble) named `df2` from the data frame `df` data by "reshaping" the data into two columns: `groupVar` and `value`. The variable `value` will contain all of the random values from `df`. The variable `groupVar` will contain the original associated variable name. The new data frame will have 2,000 rows. (Hint: if you are cool, checkout `dplyr` and the command `gather` from the `tidyverse`. If you are old school, but not quite obsolete, checkout the `reshape2` package and the `melt` command). Example data frame would look something like the following:

groupVar	value
a	1.2
a	0.9
b	0.4
b	1.3
…	

Note: please do not "print out the data" as a part of your homework submission, you may use the "head" function to show a small excerpt if desired; otherwise, code is sufficient.

(b) (20 points) Plot the densities of each distribution overlaid on each other on one plot. Each density should have some level of transparency and be colored differently. (Hint: the reshaping of the data you completed in (a) will work very well with `ggplot2`)

3 House prices data (15 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.

In a later assignment you will see this data set again. In preparation of that, perform exploratory data analysis and visualization of the data however you choose. Using `ggplot2`, create at least 3 different visualizations of the data that you think are informative. You do not have to analyze every variable, but I encourage you to play around with different possibilities and present the *best* ones. Also, please attempt to try different types of visualizations (e.g., scatter plots, distributions, bar charts, parallel plots, etc.) You might want to check out `https://www.r-graph-gallery.com/ggplot2-package.html` for some ideas.

4 Missing Data (10 points)

The `freetrade` data frame from the `Amelia` package has economic and political data on nine developing countries in Asia from 1980 to 1999. The 9 variables include year, country, average tariff rates, Polity IV score, total population, gross domestic product per capita, gross international reserves, a “dummy” variable for if the country had signed an IMF agreement in that year, a measure of financial openness, and a measure of US hegemony. Unfortunately, this data has missing values.

Explore the “missingness” in the `freetrade` using your choice of methods, e.g. from packages `VIM`, `mice`, `Amelia`, and/or others.

5 Extra credit (10 points)

Using the data from Question 4, implement your own statistical test (e.g. ANOVA, *t*-test, chi-square test, etc.) to determine if the missingness in the `tariff` variable is independent by `country`. Does your answer change if you remove Nepal or if you remove the Philippines? Discuss why. (Note: a short description of using the chi-square goodness of fit test is available in the course website.)