**Homework Week 3: Data Visualization with R**

In this homework we are going to visualize a dataset (i.e., dataframe) included with the ggplot2 package, called ‘mpg’. It describes the mileage (in miles per gallon) of cars from different manufacturers (manufacturer).

The dataset describes different models (model), from different years (year), various numbers of cylinders (cyl), automatic or manual transmission (trans), 4-wheel, front or rear-wheel drive (drv), the cars’ class (class), and the cars’ mileage in a city (cty) and on the highway (hwy)!

For this homework, you only need to hand in the .r file containing the code that answers each part of the assignment. If you need to describe or discuss something, put it as a comment in the .r file. Make sure the .r file runs completely without problems.

**Assignments**

Start with loading up the ggplot2 package and looking at a summary of the mpg dataframe.

Hint: Use the commands str() and summary() for the whole dataframe; they are both useful and give somewhat different information.

**You get 1 point for completing each of the following assignments correctly.**

1. Create two boxplots: one showing the city mileage for each car manufacturer and one showing the hway milage for each car manufacturer. Describe what you see in the boxplots.

2. Create a new data frame called mpgsmall containing a subset of the mpg data, with only the information about the following manufacturers: volkswagen, ford, honda, toyota.

Hint: There are different ways how to create such a new data frame. You already encountered the use of which(); but there are other ways: Have a look at the command subset(). And there's yet another way (and probably even more...), namely using the %in% keyword in combination with a list in your selection command to select multiple discrete values of a variable.

3. Create a single graph that shows the distributions for city mileage for each of the manufacturer in mpgsmall with differently colored density plots. Then do the same, but with a histogram using faceting.

Discuss which plot is clearer and allows you to draw conclusions more easily.!

4. Create a scatter plot to show the relation between city and highway mileage in the mpgsmall dataset, using a different symbol for each class of car.

Discuss what you see.

5. Come up with a better plot to examine whether the relationship in the previous graph is different for different classes, using different colored lines. Make sure the lines are easy to read.

6. Create a bar chart showing the mean highway mileage for each class in mpgsmall using different colors for the various classes.

7. Add error bars to the previous chart.

8. Add a nice color theme from ggthemes and make the labels of the chart pretty for publication (make sure the theme can show the labels of the axes, so do not use WSJ theme).

9. Change the bar chart so that it shows a city mileage bar next to each highway mileage bar.

Hint #1: For this, you will need to reshape the mpgsmall dataset so cty and hwy are stacked below each other in long format with a variable coding for whether the mileage is cty or hwy.

Hint #2: In order to do this, you will need to have a column coding for the idvar in the reshape command. You can do that with this line:

mpgsmall$id <- row.names(mpgsmall)

Hint #3: Your long data set will have a new variable coding for whether the mileage is s cty or hwy. This new variable needs to be transformed into a factor in order for ggplot to be able to handle it as a discrete variable.

10. Go nuts in creating a really cool visualization from the mpg dataset using ggplot that differs a lot from the plots in the assignment so far. But be sure that the plot is still elegant and readable.