***Statistics for Humanities Researchers – day 1***

**Read the following instructions with the Day1Exercise1.Rmd file open.**

**The exercises should have 2 outputs:**

* **This document with answers to the questions. Cut and paste any plot you’re asked to produce.**
* **The Rmd file with your code written in and “knitted”**

Lab 1: Approaching a dataset

Exercise 1) – Setting up a package in R and using simple functions

1. Install the “fueleconomy” package. This will give you access to the data which we will use in this lab. The data are from the US Environmental Protection Agency database, which presents miles per gallon fuel economy for most makes and models sold in the US

Exercise 2) – Classifying variables

1. Which variables (if any) in the data set are nominal variables?
2. Which variables (if any) in the data set are ordinal variables?
3. Which variables (if any) in the data set are interval variables?

Exercise 3) – Subsetting

1. Make a new data frame which only includes data on the following makes
   1. Chevrolet
   2. Dodge
   3. Ford
   4. Mitsubishi
   5. Nissan
   6. Toyota
   7. Volvo
   8. Volkswagen
   9. BMW
2. What is the time range of the data set? What are the earliest and latest years in the data?

Exercise 4) – Basic scatterplots

1. Using ggplot, make a basic scatterplot which shows fuel economy for city driving, in each make in you new data frame, across each year represented in the data
2. Comment on the visual features of the plot. Is there anything about the plot which could be improved so that you could better interpret the data?

Exercise 5) – Excluding data

1. Modify the data frame so that you exclude electric and hybrid cars from the data set

Exercise 6) – Recoding variables

1. Make a new variable which codes cars by region of production. So for instance, Chevrolet, Dodge, and Ford are produced in the US; Mitsubishi, Nissan, and Toyota are produced in Asia; and Volkswagen, Volvo, and BMW are produced in the EU. Make a variables that codes this information.

Exercise 5) – fitting curves

1. Using ggplot, make a new scatterplot that captures the relation between year and fuel economy for city driving for each region of production
2. Now make a scatterplot that captures the same relation, assuming that the relationship follows a straight line
3. Now make a scatterplot that captures the same relation, assuming the relationship is quadratic (ask for help if you need it).

Exercise 6) – Aggregating data

1. Using the aggregate function, make a table that represents the mean fuel economy of cars produced by each region. Here we just mean the average, we’ll talk about this more tomorrow. Turn your table into a new data frame by giving it a sensible name.
2. Using the aggregate function, make a table that represents the standard deviation of fuel economy of cars produced by each region. Standard deviation is a measure of variability, we’ll talk more about this tomorrow. Turn your table into a new data frame by giving it a sensible name.
3. Using the aggregate function, make a table that represents the mean fuel economy of cars produced by each region, grouped by the number of cylinders in each car. Turn your table into a new data frame by giving it a sensible name.
4. Using the aggregate function, make a table that represents the mean fuel economy of cars produced by each region, grouped by the number of cylinders in each car. Turn your table into a new data frame by giving it a sensible name.

Exercise 7) – Barplots

1. Using the aggregated data produced in exercises 6a, use ggplot to produce a barplot that represents average fuel economy in each region. Using the data frame produced in 6b, include error bars in your plot, which represent 1 standard deviation.
2. Using the aggregated data produced in exercises 6c, use ggplot to produce a barplot that represents average fuel economy in each region, grouped by the number of cylinders in the cars. Using the data frame produced in 6d, include error bars in your plot, which represent 1 standard deviation.
3. Compare the error bars in these two plots. What is the main difference between the two? How might you explain this difference?

Exercise 8) Compare barplots and boxplots. What are the advantages of choosing the one or the other? What are the drawbacks?