Info 573: Data Science I: Theoretical Foundations
Lab 1: Crash course in R
Instructor: Ben Althouse; Contact: bma85@uw.edu

In this lab you will practice basic data exploration. Here we will be loading data into R and exploring the data using tables, plots and other basic statistics. Either RStudio or the console may be used for this lab.

You may work with a partner on this lab, however, you will be asked to submit a copy of your analysis code to Canvas at the end of class - each individual must submit their own version of their code – *please be sure to put your name in the code!* Keep track of all the commands you run using a text editor or R script.

You should comment your code as you run through this exercise. You can do this in R using the # character. Please answer the questions posed in the exercise/lab by adding comments to your R script.

First we need to go through a series of preliminary tasks to prepare the R environment to do our analysis.

Creating a Working Directory / Folder for Your Work: In RStudio click on the Session option at the top and then Set Working Directory → Choose Directory.... Select a folder where you will store data and save figures produced in this analysis. I recommend creating a new folder specifically for this class or assignment. No matter how the working directory has been set, you can find out what it is (i.e., where R thinks it is working) with the function getwd(). Try this in your Console.

Starting an R Script to Document Code: You want to create a text file to document/record all code you write for this lab. You can begin with your favorite text editor, the R console, or RStudio to do this. In RStudio click on the Create File icon just below File. Choose the R Script option. You should see a blank document created. Write all code in this document! Save your progress as you go along. RStudio make it easy to execute code written in your script in the Console using the button on the upper right of your document. You can also use (Ctrl+Enter PC, Command+Return) to execute the active line.

Downloading and Importing the Dataset: Download the data required for this lab from Canvas: seatbelts.csv. Seatbelts were not mandatory in the UK before January 31, 1983. This dataset summarizes the road fatalities in the UK from January 1969 through 1984. We wish to examine trends in road fatalities both before and after the mandatory seatbelt law took effect.

Dataset description: Table below gives the variables in the dataset.

year.month	Decimal year of observation (corresponds to months)
year	Year of observation
DriversKilled	Number of drivers killed
drivers	Number of drivers killed or seriously injured
front	Number of front seat passengers killed or seriously injured (includes driver)
rear	Number of rear seat passengers killed or seriously injured
kms	Distance driven
PetrolPrice	Price of petrol
VanKilled	Number of commercial van drivers killed
law	Law is in effect (1/0, yes/no)

Load the data: you can load the data using:

```
seatbelts <- read.csv(“seatbelts.csv”)
```

Basic Exploratory Analysis: Now let’s get to work! The suggested steps will guide you through some basic exploratory analysis of the data and to produce some basic visualizations.

1. Data Cleaning: The seatbelts data is stored as a R data frame. Recall this is a specific data structure helpful for storing data about individual cases and variables measured on those cases. Verify the data is store in this format by running the following command:

```
class(seatbelts)
```

You can also explore the data frame itself more using the following:

```
# Remember this is a comment!
# Get some information about a data frame:
dim(seatbelts)
colnames(seatbelts)
# data frame attributes including column names
# how many cases are in the observed data?
# what variables are observed for each month?
```

Next, let’s look at the data types and summaries for each of the variables in our dataset.

```
# Get some information about the variables in the data:
summary(seatbelts)
```

2. Computing Averages: Suppose we want to describe the average number of deaths per year. We can find the mean age using the mean() function. What is the result?

```
mean(seatbelts[,"DriversKilled"])
```

But is this the mean by year? Or over all years? Let's subset the data and see:

```
mean(seatbelts[seatbelts[,"year"]>=1969 &
seatbelts[,"year"]<1970,"DriversKilled"])
```

Not the same. We can use the by() command here:

```
by(seatbelts[,"DriversKilled"], seatbelts[,"year"], mean)
```

What was the average number of fatalities in 1970? 1978?

What was the average number of rear seat fatalities in 1972? 1980?

3. Exploring Relationships I: Visually explore relationships between your variables. Plot the relationship between drivers killed or seriously injured and petrol price and kilometers traveled. What do you see? What hypotheses might you make after seeing these relationships?

```
plot(seatbelts[,"kms"], seatbelts[,"drivers"])
```

4. Exploring Relationships II: Consider a research question that asks about the implementation of the seatbelt law. Did it have an effect? Did it reduce fatalities?

Question: What descriptive and visual tools might we use to explore this? Examine the mean fatalities before and after the implementation of the law. Remember to subset your data (hint: there are two variables you can use to do this to do this).

Produce a figure to get some visual intuition about the response to the seatbelt law. Was it a gradual decline? Sharp? Does it appear large in magnitude or small? (Hint, `abline()` is useful for marking specific dates on time series plots)

5. Extra Credit Question: Be creative! What additional research questions might you want to explore in this data? See what you can discover about the relationships that exist between variables in this dataset. Produce a table or figure to communicate your findings and describe in words what you found and why it's interesting.