**IST687 HW 5 - Accident Analysis using JSON**

**Reminders of things to practice from previous weeks:**

Descriptive statistics mean( ) max( ) min( )

Sequence operator : (For example, 1:4 is shorthand for 1, 2, 3, 4)

?command Ask R for help with a command

Summarize variable summary( )  
Coerce to numeric as.numeric( )

**This week:** Up until now we have been working with quite simple data structures. For example, a vector is just a simple list of values; a data frame is a rectangular data structure with well-defined columns. This week we have the dual challenge of pulling a complex data structure from the web and restructuring it into a data frame. This data structure contains vehicle accident data from the State of Maryland. For the first time, we will be using *nested lists* – basically a list that contains other lists. We will use the [[ ]] double-bracket notation to get access to the lists that are nested within other lists.

**Step A: Load the data**

1. Read in the following JSON dataset

[http://data.maryland.gov/api/views/pdvh-tf2u/rows.json?accessType=DOWNLOAD](https://data.maryland.gov/api/views/pdvh-tf2u/rows.json?accessType=DOWNLOAD)

**Step B: Clean the data**

1. Remove the first 8 columns,
2. Then, to make it easier to work with, name the rest of the columns as follows:

Note, not surprisingly, it is in JSON format. You should be able to see that the first result is the metadata (information about the data) and the second is the actual data.

namesOfColumns <- c("CASE\_NUMBER","BARRACK","ACC\_DATE","ACC\_TIME","ACC\_TIME\_CODE","DAY\_OF\_WEEK","ROAD","INTERSECT\_ROAD","DIST\_FROM\_INTERSECT","DIST\_DIRECTION","CITY\_NAME","COUNTY\_CODE","COUNTY\_NAME","VEHICLE\_COUNT","PROP\_DEST","INJURY","COLLISION\_WITH\_1","COLLISION\_WITH\_2")

**Step D: Explore the data – using the dataframe you created**

1. What was the total number of accidents with injuries?
2. How many accidents happened on Sunday?
3. How many injuries occurred each day of the week?

**Step D: Explore the data – using dplyr**

*Hint: Use and explain the following lines of code (document each line)*

*df.GroupBydays <- group\_by(df, DAY\_OF\_WEEK)*

*accidents <- summarize(df.GroupBydays, count = n(), injury = sum(INJURY))*

Then answer the following questions:

1. What was the total number of accidents with injuries?
2. How many accidents happened on Sunday?
3. How many injuries occurred each day of the week?
4. In a block comment, explain if you find doing the analysis with the dataframe directly, or using dplyr easier

**Step D: Explore the distribution of the number of vehicles in accidents**

1. What is the distribution of the number of vehicles in accidents on Friday?  
   (use a histogram and quantile)
2. How does this distribution compare with the distribution of the number of vehicles in accidents on Sunday?