Lecture 2: Introduction to Julia

Dataset

We will use Traffic Stops from Cincinnati City.

Data description: This dataset captures all subjects of traffic stops involving motor vehicles. Time of incident, officer assignment, race/sex of stop subject, and outcome of the stop ("Action taken") are also included in this data. Individual traffic stops may populate multiple data rows to account for multiple subjects and multiple outcomes: "incident number" is the unique identifier for every one (1) traffic stop.

Filename: "Traffic_Crash_Reports__CPD__Aug2018.csv" *Make sure this file in the same directory as the ipynb file*

Setup: Use Julia 0.6.4 kernel. Install the packages CSV, Gadfly, Cairo and Fontconfig.

In [ ]:
```
Pkg.add("CSV",VersionNumber("0.2.5"));
Pkg.add("Gadfly",VersionNumber("0.8.0"));
Pkg.add("Cairo",VersionNumber("0.5.6"));
Pkg.add("Fontconfig",VersionNumber("0.1.1"))
Pkg.add("RDatasets",VersionNumber("0.4.0"))
```

Use the packages...

In [ ]:
```
using CSV, DataFrames, Gadfly, Cairo, Fontconfig;
```

Questions

Q 1: Write Julia code to load this data into memory.

Q 2: What is the size of the data?

Q 3: Create a new Dataframe by selecting the columns AGE, CRASHSEVERITY, DAYOFWEEK, GENDER, INJURIES, LIGHTCONDITIONSPRIMARY, LOCALREPORTNO, MANNEROFCRASH, ROADSURFACE, WEATHER, ZIP

From here on wards work with the new Dataframe.

Q 4: Using showcols, list the different element types in the new data frame. Also list the columns in which there are missing values.

Q 5: Remove the rows in the missing values from the Dataframe. Comment on the number of rows that have been removed in this process.

Q 6: List the unique entries in the CRASHSEVERITY column

Q 7: Find out the different types of crashes in this data.

Q 8: Find out the different types of WEATHER conditions in this data.

Q 9: Determine the number of crashes happened in each of these weather conditions using by() function.

Q 10: Find out the different light conditions in this data.

Q 11: Determine the number of crashes happened in each combination of weather and light conditions using by() function. State your observations.

Q 12: How many ZIP codes are covered in this data.

Q 13: Plot a bar graph showing the number of accidents in each of the ZIP codes

Step 14: Draw a scatter plot between weather and light conditions. State your observations. Please use set_default_plot_size(12inch, 8inch) function to adjust the figure size as needed for visibility.

Step 15: Make a plot to view the number of crashes on different days of the week. On which day of the week fewer crashes happen? On which day of the week more crashes happen?

Step 16: Make a plot to view the number of crashes reported per age-group. State your observations. State your observations.

Step 17: Use the following two lines of code to load the "iris" dataset:

```
using RDatasets

iris = dataset("datasets", "iris");
```

This dataset has information about flowers from three plant species.

Do:

1. List attributes in this data
2. Generate a scatter plot between "PetalLength" and "PetalWidth" where each point is colored based on "Sepecies". What observations can you make about the flowers from the three plant species based on this plot.

Step 18: Using IRIS dataset draw a box plot to compare the SepalWidth for the three plant species. What observations can you make based on this plot?

Step 19: Draw a violin plot for SepalWidth (similar to the box plot above) and state any new observations you may have.