Lecture 2: Introduction to Julia

Dataset

We will use Traffic Crash Reports data from Cincinnati City.

Data description: Traffic Crash Reports are records in the event of a CPD response to a traffic crash. The source of this data is the City of Cincinnati Police Department. The column names for this data are self explantory.

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, RDatasets;
```

Questions

Write Julia code to answer the following questions:

Q 1: Load this data (Traffic_Crash_Reports**CPD**Aug2018.csv) into memory.

In [ ]:
```
data =
```

Q 2: What is the size of the dataset? How many data points and how many attributes?

In [ ]:
```

```

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

*Use the new_data Dataframe for **Q4** and **Q5**.*

In [ ]:
```
new_data =
```

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

In [ ]:

Q 5: Create a new dataframe 'newdata_nomissing' by removing the rows in the missing values from the new_data Dataframe. How many rows have been removed in this process?

In [ ]:
```
new_data_nomissing =
```

*For the following questions until **Q15** use new_data_nomissing dataframe.*

Q 6: Generate a list of the different types of crashes in this data.

In [ ]:

Q 7: Generate a list of the different types of WEATHER conditions in this data.

In [ ]:

Q 8: Determine the number of crashes happened in each of these weather conditions using by() function.

In [ ]:

Q 9: Generate a list of the different light conditions in this data.

In [ ]:

Q 10: Determine the number of crashes happened in each combination of weather and light conditions using by() function. State which combination of weather and ligt conditions result in most number of crashes.

In [ ]:

Q 11: How many ZIP codes are covered in this data.

In [ ]:

For the following questions that involve generating plots, you may use the white_panel theme.

In [ ]:
```
white_panel = Theme(
    panel_fill=colorant"white",
    default_color=colorant"blue",
    major_label_font_size=26pt,
    minor_label_font_size=22pt,
    major_label_color=colorant"black",
    minor_label_color=colorant"black"
);
```

Q 12: Plot a bar graph showing the number of accidents in each of the ZIP codes

In [ ]:

Q 13: Generate a scatter plot between weather and light conditions. State which combinations of weather and light conditions appear to have significantly higher number of crashes. Please use set_default_plot_size(12inch, 8inch) function to adjust the figure size as needed for visibility.

In [ ]:

Q 14: Generate a plot to view the number of crashes on different days of the week. On which day of the week do fewer crashes happen? On which day of the week do the highest number of crashes happen?

In [ ]:

Q 15: Generate a plot to view the number of crashes reported per age-group. Which age group is involved in the most number of crashes?

In [ ]:

Q 16: Load the "iris" dataset using the following command.

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.

In [ ]:

Q 17: Using Iris dataset, generate a box plot to compare the SepalWidth for the three plant species. Flowers from which species has generally longer sepalwidths?

In [ ]:

Q 18: Using Iris dataset, generate a violin plot for SepalWidth (similar to the box plot above). What new observations can you make from this plot, compared to the box plots you generated in response to **Q 17**.

In [ ]: