**Using Big Data Deliverables**

**Michelle Zokou**

**CTEC 298**

**Dr. Bemley**

**10/16/2019**

1. **Introduction**

The New York Times just broke a new story about vehicular incident and collisions in New York, and the state’s lack of policies and planning to address the issues. Most people commute using some means of transportation; however, they do not realize there are chances of not coming back home safely. The are many people getting injured as well as losing their lives in New York city. In New York city, there are different types of accidents caused by vehicles, motorcycles, and bicycles. Even though there are few accidents happening off the road, almost all accidents happen on the road. Therefore, our team focuses on tackling vehicular accidents occurring on the street, not off the street. There are various types of vehicles involved into the accidents, to mention some: passenger vehicle, taxi, sport utility, and van. The accidents have occurred on small streets, main highways, and big intersection. Some drivers have caused the accidents as a result of health problems.

1. **Summary of CTEC 128**

My team and I created a concept map for the project and according to the concept map, the type of road where the accidents occur are streets, avenues, parkways, boulevard, and highways. Those help to determine whether the accidents happen in small residential areas or on main streets. The people involved in accident include persons, pedestrians, cyclists, and motorists. To avoid confusion between persons and pedestrians, persons represent people who were driving vehicles, whereas pedestrians represent people who were crossing the street or those who were outside vehicles. People in this research paper represents pedestrians, persons, cyclists, and motorists. The top three most causes are driver distraction, Failure to yield right of way, and fatigue(drowsy). Passenger vehicle, taxi, sport utility was involved in the accidents. The concept map has used as a bridge to determine research questions. Based on the concept map, we were able to determine few research questions. To mention some: which borough has experienced the most incident? What was the most frequent factor? What was the health problems related to the drivers? Which type of vehicle has caused the most accident? The final data is focused on factors causing the accident and regions that hosts many accidents in New York city. Since there are only five regions (Bronx, Brooklyn, Manhattan, Queens, and Staten Island), we organized the data set according Boroughs. This helps us to compare different features of the regions. To reduce the accidents happening on the road, it is essential to know the core causes which can lead us to a conclusion. Knowing the cause will enable us to address the most occurring problem. The pivot table that helped us decide which region has maximum injured people.

1. **Description of CTEC 128 Material Submitted**

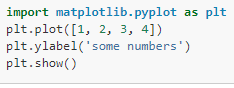
The materials there were submitted for CTEC 128 were a PowerPoint, an excel spreadsheet with the data used for research for vehicular incidents and collisions in New York, and a final data science paper. The original data set contains 477, 733 rows and twenty-nine columns. Variables, which are organized in columns includes lists of ID number, date, time, borough, zip code, accidents on the street, accidents off the street, persons injured, persons killed, motorist killed, motorists injured, cyclists killed, cyclists injured, vehicle type, and factors causing the accident. I made many changes to the original data as some are less reliable and do not help to predict possible causes. I have removed the variables lacking location names because the analysis relies on location, specifically boroughs. I removed the data without time specification as I would like to know whether it occurs day time or night time. Since there are few involvements of fourth and fifth vehicle, I deleted types and factors of vehicle four and five. The very first PowerPoint I submitted via share drive gives a visual of the graphs and excel spreadsheet. The final data science report submitted contains the overall research for the topic of for vehicular incidents and collisions in New York.

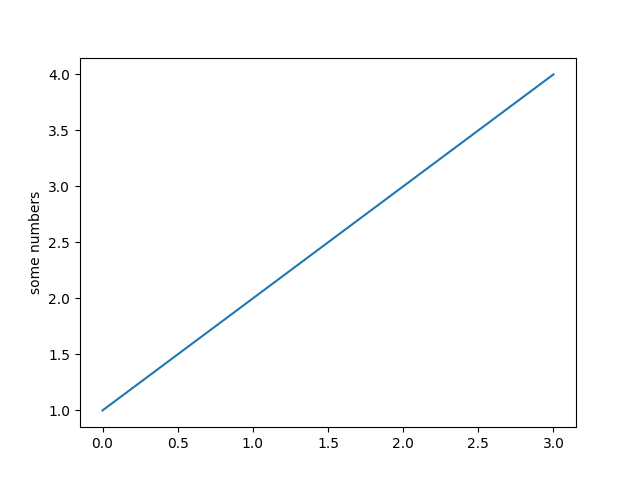
**IV. Description of the plot deliverables**

The utilization of data visualization enables you to show information in a pictorial or graphical organization. This enables you to examine and change your information. The 6 distinct plots utilized for the plot expectations comprised of, dissipate plots, reference diagrams, histograms, pie plot, stack plot, and multiplot chart. Your structured presentations are utilized to analyze information among gatherings and measure changes after some time. A histogram is utilized to speak to information given as certain gatherings. A dissipate plot is utilized to analyze at least two factors or think about the progressions after some time from two gatherings. A pie diagram is utilized to envision unmitigated information. A stack plot is utilized to track changes after some time. A multiplot chart indicates subplots for at least 2 arrangements of information. In the documentation gave, there is a model and how to make each plot.

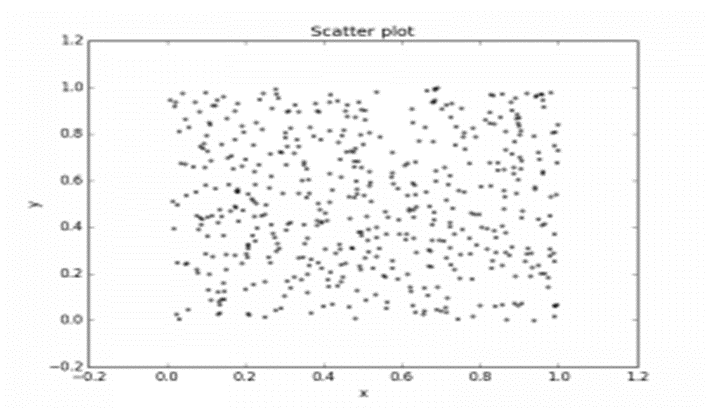
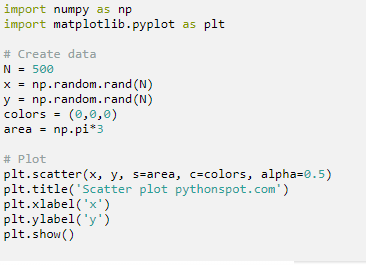
**Matplotlib Tutorial: Python plotting**

You import the pyplot module of the matplotlib in library under the alias plt. Pyplot is a module in the matplotlib package, that is why sometimes, you see matplotlib plot in code. The module provides an interface that allows you to implicitly and automatically create figures and axes to achieve the desired plot.



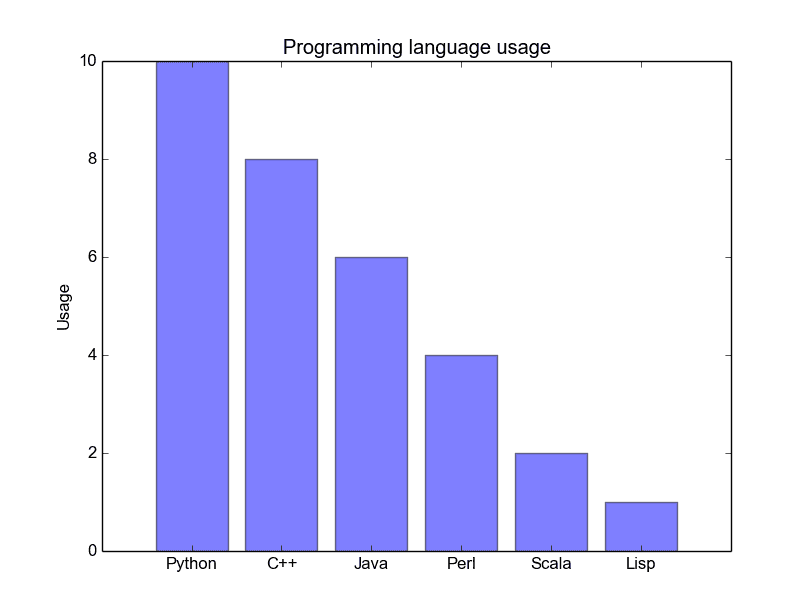


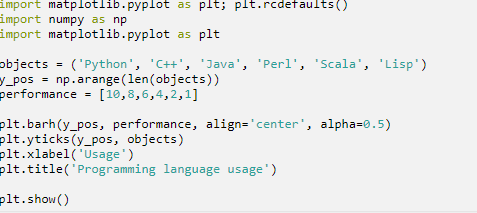
If you provide a single list or array to the plot () command, matplotlib assumes it is a sequence of y values, and automatically generates the x values for you. Since python ranges start with 0, the default x vector has the same length as y but starts with 0. Hence the x data are [0,1,2,3].



I used Matplot using Jupyter to a built-in function to create scatterplots called scatter (). A scatter plot is a type of plot that shows the data as a collection of points. The position of a point depends on its two-dimensional value, where each value is a position on either the horizontal or vertical dimension.

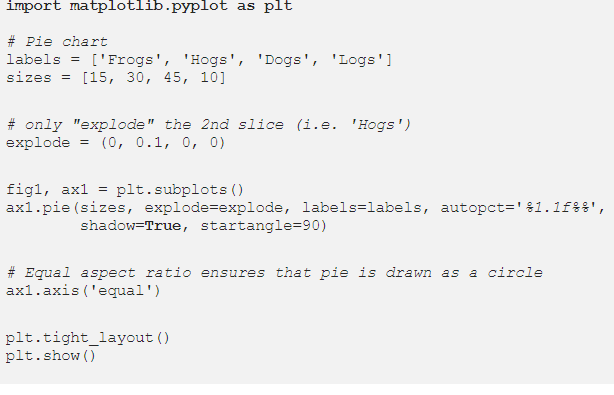
**Bar Graph**

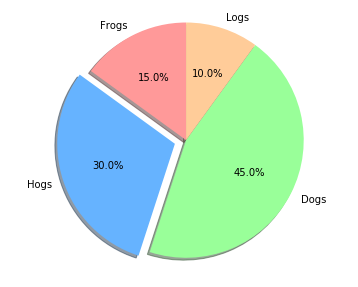




Bar plot is a plot that represents categorical data with rectangular bars with lengths proportional to the values that they represent. A bar plot shows comparisons among discrete categories. One axis of the plot shows the specific categories being compared, and the other axis represents a measured value.

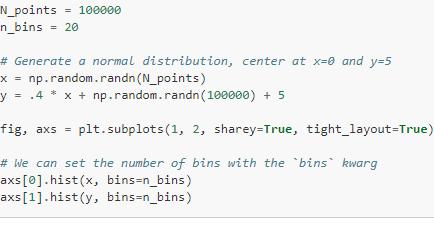
**Pie plot/chart**

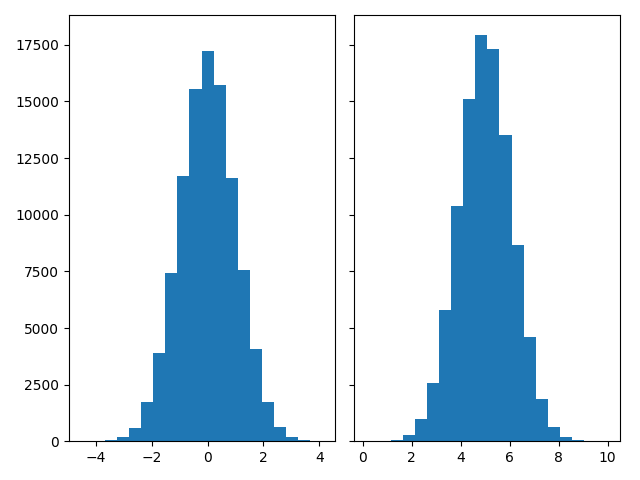




Pie object is a circular statistical chart, which is divided into sectors to illustrate numerical proportion. Data visualized by the sectors of the pie is set in values. The sector labels are set in labels. The sector colors are set in marker color.

**Histogram Graph**





Histogram graph is an accurate graphical representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable. To contrast a histogram, the first step is to bin the range of values that is, divide the entire range of values into a series of intervals and then count how many values fall into each interval.

**Conclusion**

Big data analytics is the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. I think ctec 298 was a good class because, I learn so much in so little. I learn to used matplotlib, I learn to use python which I had no idea how to use it in the pass, but after taking ctec 298 I enjoy python a lot better.

**References**

* ” Matplotlib, <https://matplotlib.org/>.
* “Pandas Foundations.” DataCamp, <https://www.datacamp.com/courses/pandas-foundations>?
* New York Vehicular Collision data set from 2015 to 2017