NYPD Motor Vehicle Collisions - Data Analysis
Data Science - Mini Project 2

Arpita Sheth
(SBU ID: 110422828)

Introduction

Analysis of data is a process of inspecting, cleaning, transforming, and modelling data with the goal of discovering useful information. In this assignment, we use Data Analysis to analyse some issues in NYC. New York City's open data legislation provides a centralised location for the City's Open Data – the **Open Data Portal**. NYC Open Data Portal consists of almost 1300+ datasets. Available data spans multifarious city operations, including cultural affairs, education, health, housing, property, public safety, social services, transportation, and more. These data power other initiatives like the NYC BigApps competition and the work of the Mayor's Office of Data Analytics, and pave the way for new initiatives to use technology and data to engage the public, guide decision-making and make government more effective.

Problem Statement

The dataset containing data about **Motor Vehicle Collisions in NYC for the year 2012-2015** is chosen for analysis. The data is provided by NYPD and consists of granular details like street name, zip-codes, borough information about accidents, number of persons killed or injured, types of vehicles involved in accidents and various factors contributing to these collisions such as driver inattention etc. We use this dataset to answer some intriguing questions like which zones are most ancient prone, how many people are killed every year, what are the most common causes of accidents and so on.

Data Acquisition

NYPD Motor Vehicle Collisions data which is a part of NYC BigApps is used. The dataset is available on NYC Open Data Portal.

Data Pre-Processing

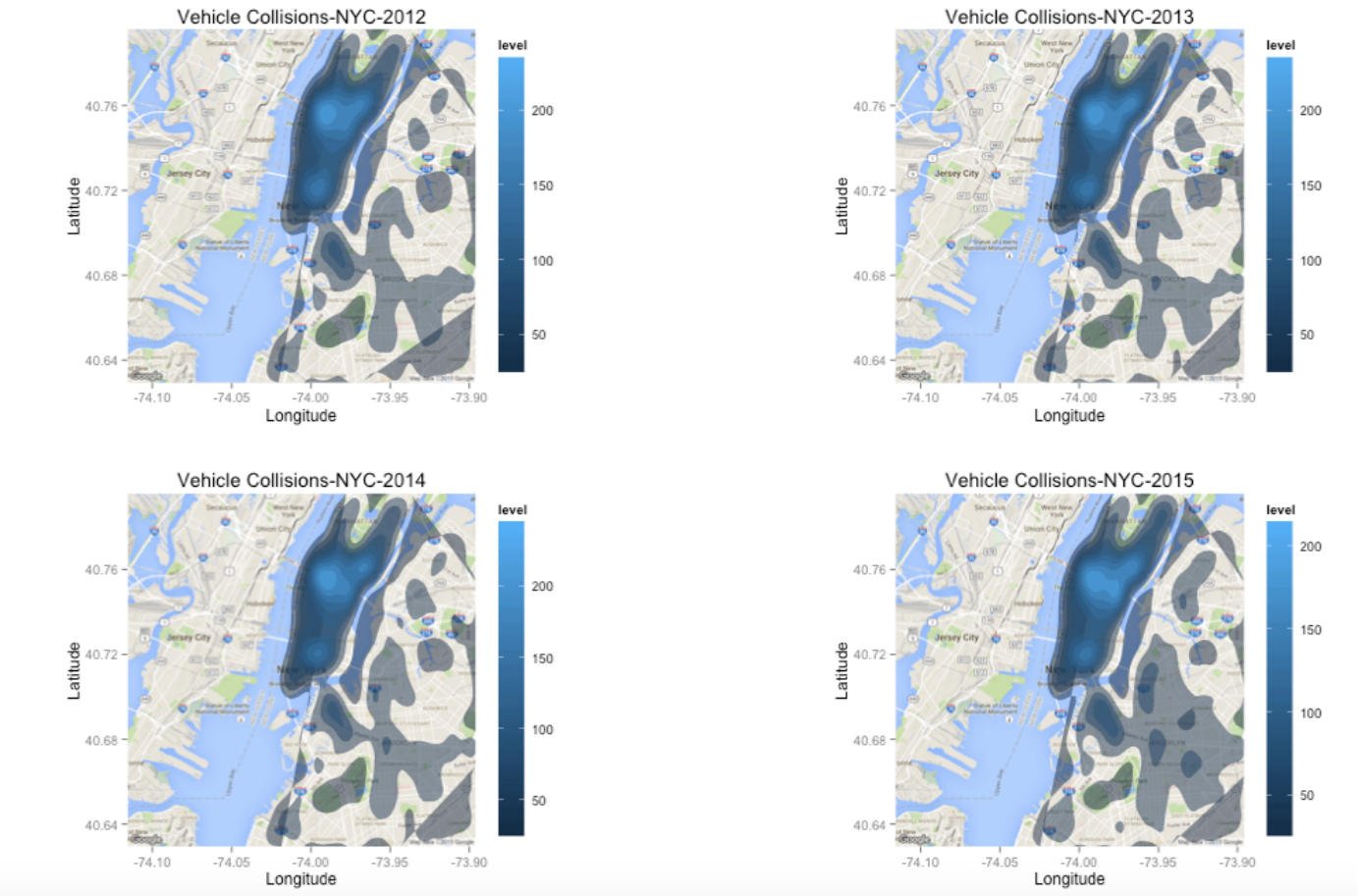
The data is a public dataset and contains lot of missing values in the records. As a prerequisite the data is cleaned to filter out records that do not contain any Location information (Lat,Long). Year information is extracted from the dates provided.

Observation, Techniques Used and Results

1. **What sort of data is provided?**

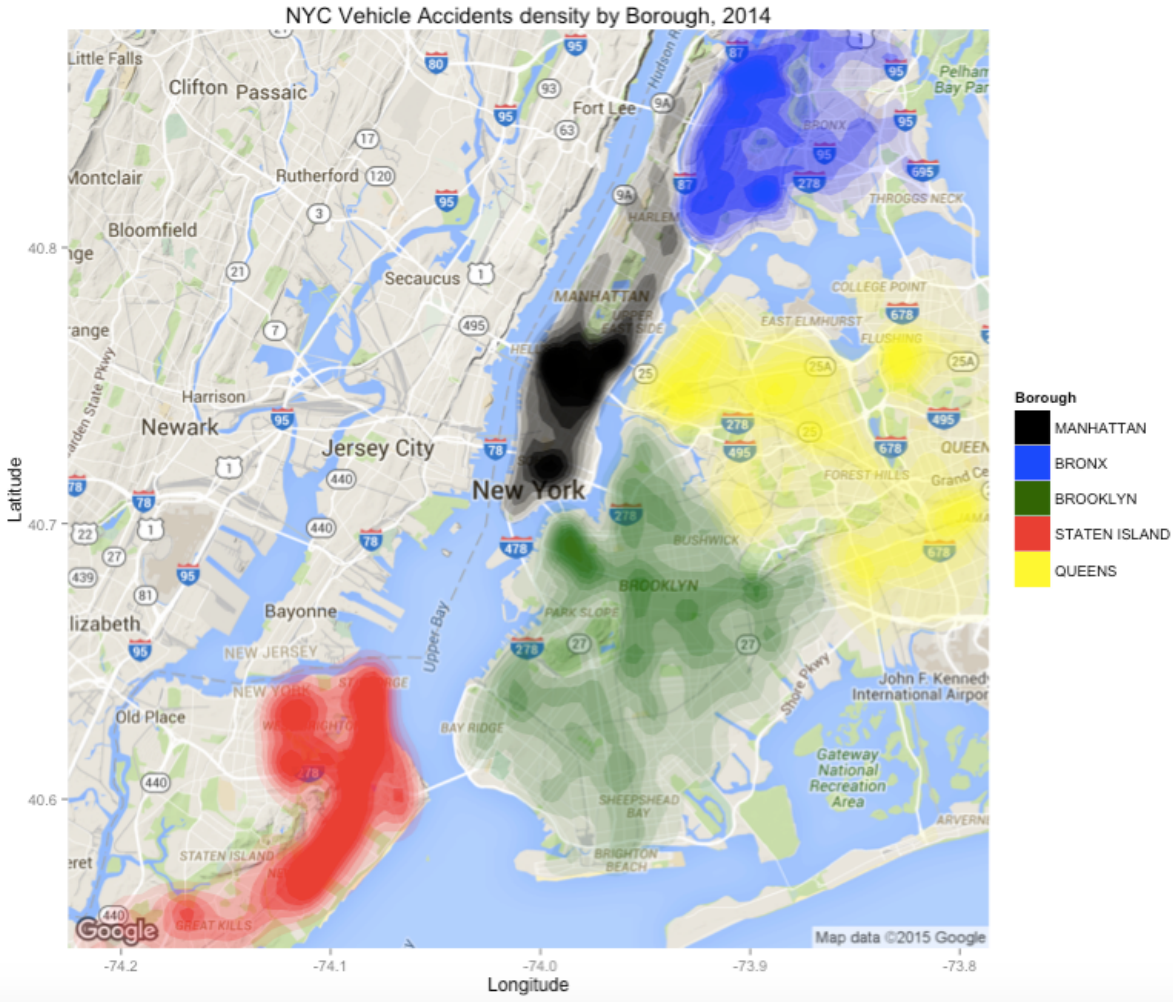
   The given dataset is categorical spatial data. Spatial Data Analysis like Bubble chart, Heat Maps, Geo-charts, 2D density plots can help in visualising various features of the dataset.

2. The dataset ranges from 2012 to 2015. We separate the dates provided into relevant year groups. We convert the point data into a **2d continuous field of point density** for each year. It shows the density of accidents in various regions for each year on the map of New York City. The density plot illustrates that there are some areas of relatively high density close to the centre of Manhattan. (Plot shown below). ggmap package in R is used to get the terrain map of NYC.



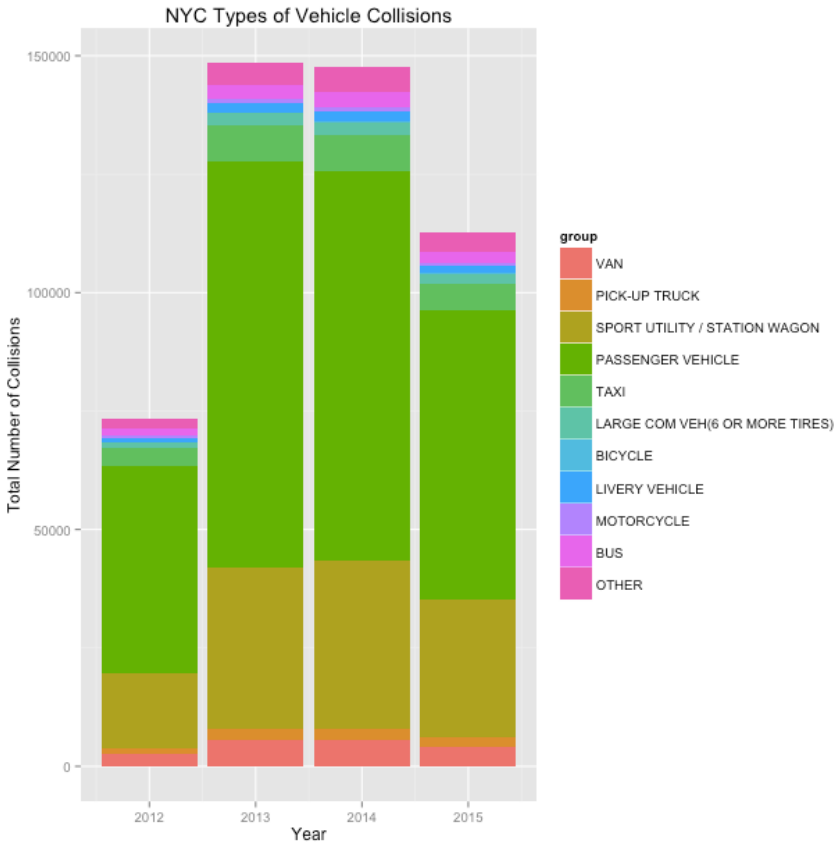
3. **What does the accident density distribution over various Boroughs look like?**

   The Data is classified in **5 boroughs** namely Bronx, Brooklyn, Manhattan, Queens and Staten Island. We **visualise zones of interest** from the point data. We use density plot to map the collisions borough wise on the map of New York City. Various colours indicate various boroughs and the colour density level indicates the concentration of accidents in each area.

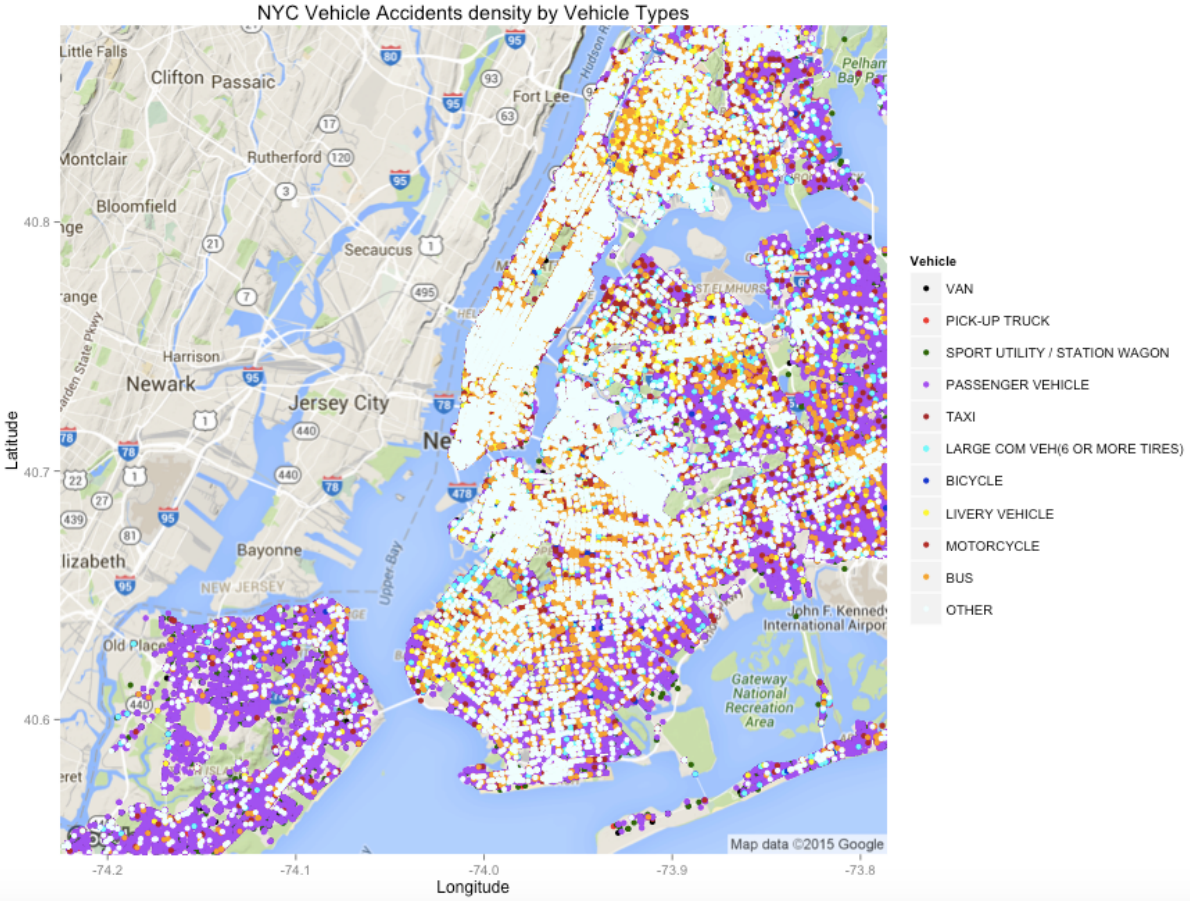


4. **Which types of vehicles are more involved in the collisions? Does the trend vary over years?**

   The dataset contains data about the vehicles involved in collisions. Plotting this data for all years as a stacked bar chart shows that **Passenger Vehicles** form a major portion of vehicles involved in accidents followed by Pick-Up Trucks. We can also see that passenger vehicles are mostly involved in accidents over all the years and the trend remains same.



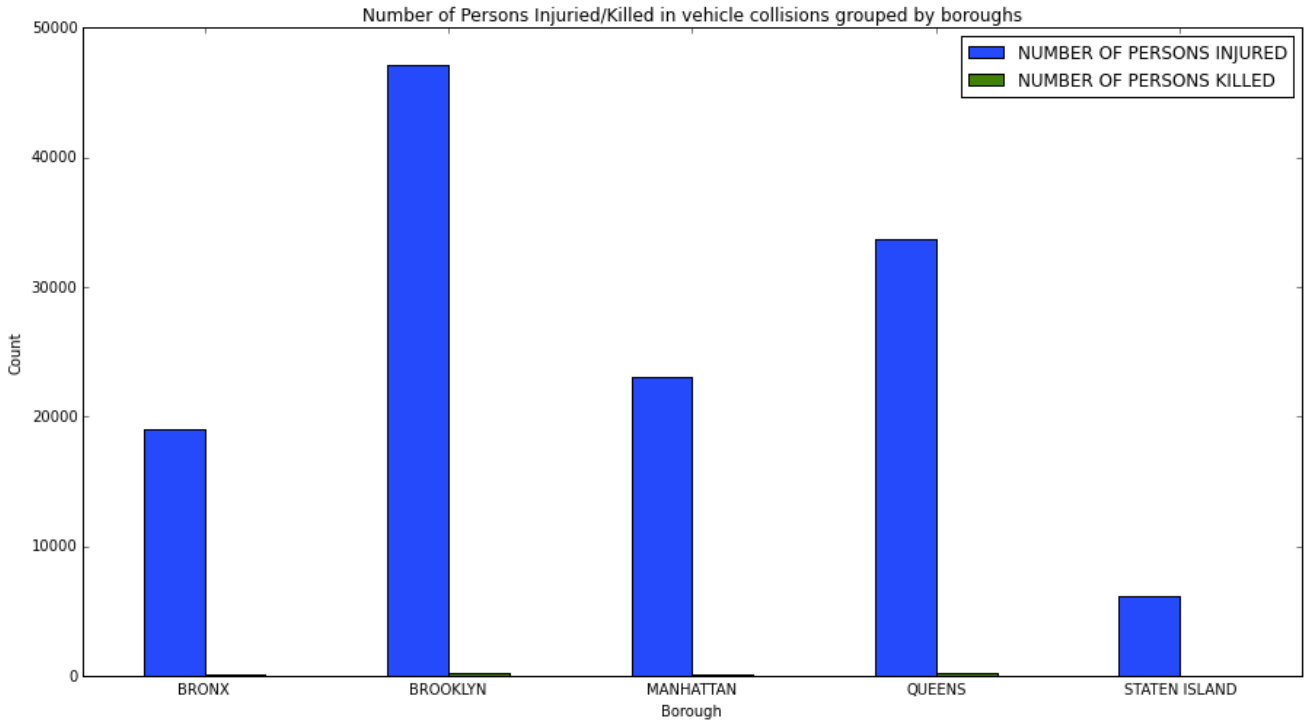
5. **What do the above mentioned vehicular statistics look like when mapped over regions of the city?**



The accident density can be visualised using vehicle distributions over different areas of New York City using bubble chart. From the figure in #5, it can be seen that Passenger Vehicle collisions are very prevalent in all boroughs but they are mostly dominant in **Staten Island** region shown by purple dots.

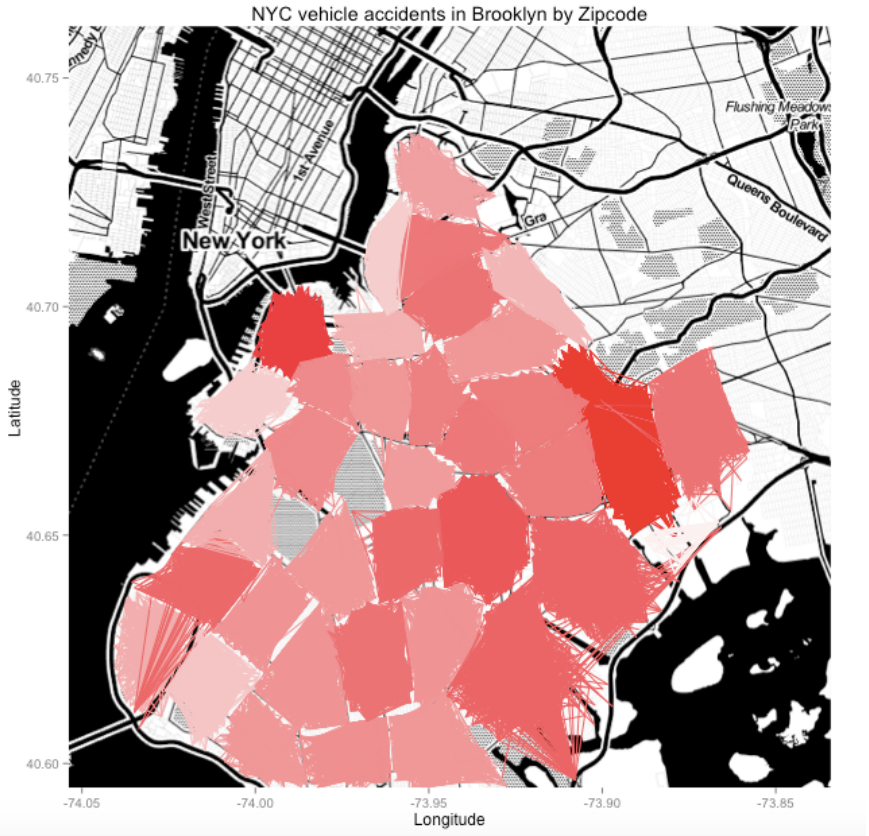
6. **How many fatal accidents or injuries have occurred in each borough since 2012?**

   A histogram showing the number of persons injured and killed in each borough indicates that Brooklyn has a much higher number of accidents with around 47k people injured.



7. **Knowing that Brooklyn has highest number of accidents, can we get an area wise distribution?**

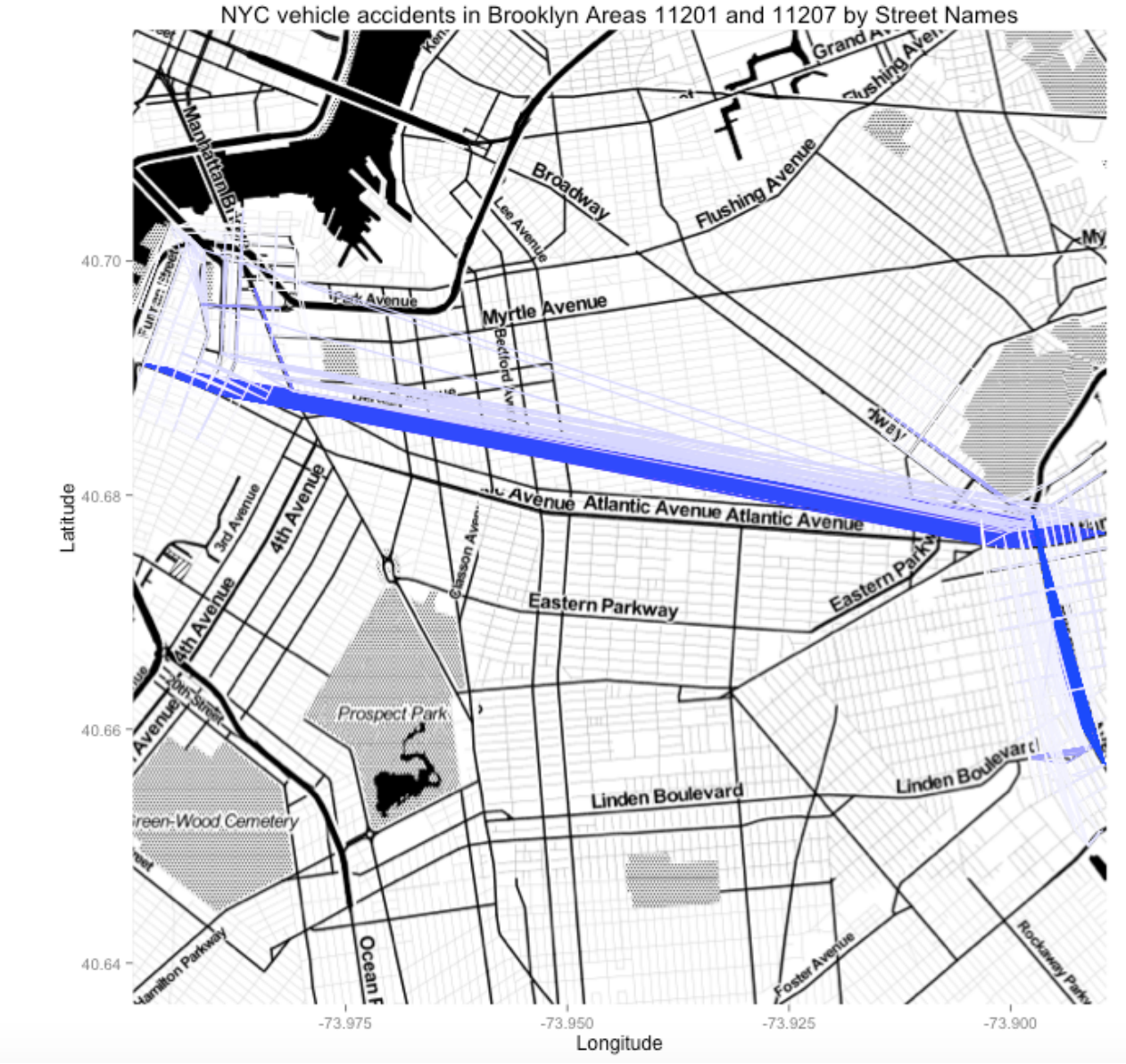
   Mapping the categorical data according to zip codes using **heat map** in the Brooklyn borough shows high density accident prone areas in **bright red**.



8. Figure in #7 shows that zip-codes 11201 and 11207 have the highest accident density. 11201 is **Downtown Brooklyn** where Brooklyn Bridge is situated. 11207 is **East NY** consisting on Brooklyn Pennsylvania Avenue. This is logical because Brooklyn Bridge area experiences a lot of traffic congestion and construction work, so collisions are quite common in this area.

9. **What are the main streets or intersection points which are highly dangerous and accident prone in the above mentioned areas?**

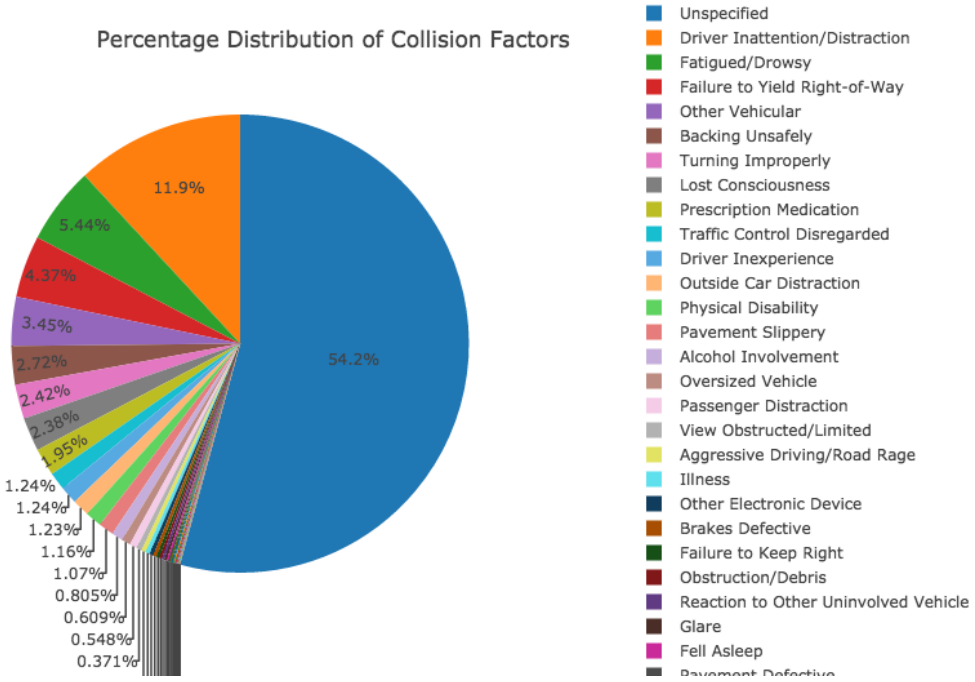
   Analysing further with the data provided for these zip-codes, a heat map based on street data and terrain map of Brooklyn shows which streets experience most number of collisions in this area and which intersections are dangerous.



10. The above figure shows that most accidents happen on the **Pennsylvania Avenue** joining Downtown Brooklyn with East NY. Pennsylvania Avenue is followed by **Atlantic Avenue.**

11. What are the most common factors contributing to accidents?

Another important information provided in the dataset is contributing factors in vehicle collisions. Various factors contributing to accidents include driver inattention, fatigue,drowsiness, etc. It can be represented as follows:



As the statistics indicate, **Driver inattention or distraction** is the most commonly specified factor in vehicle collisions.

12. What supervised learning models can be used on the given categorical data?

1. We can use the naive Bayes classifier to classify the collision incidents according to boroughs, types of vehicles involved, time of collision etc.
2. For doing Naive Bayes classification on the data, the data is split into train set and test set. In the given data set the data from 2012 to 2014 is treated as train set and we treat the observations in 2015 as the test set. Naive Bayes is a simple probabilistic classifier based on Bayes theorem. We develop a prediction model based on various factors like time and contributing vehicle from the train set and use it for classification on the test set.

13. **Hypothesis Testing**

1. The dataset contains another important information i.e Time of the day when collision occurred. Interesting traffic and accident trends can be found by looking at the data. For e.g one would expect more accidents during the on-peak hours or evening time i.e office hours. We can use hypothesis testing to consider our hypothesis. **t-test** is very commonly used for hypothesis testing.
2. For our testing we consider null hypothesis and an alternative hypothesis and perform a t-test.
   H_0 = The accident distribution is uniform across all the times of day. Accidents are equally likely at all times.
   H_1 = Accident incidents have different distribution depending on time of the day. e.g morning and evening,etc.

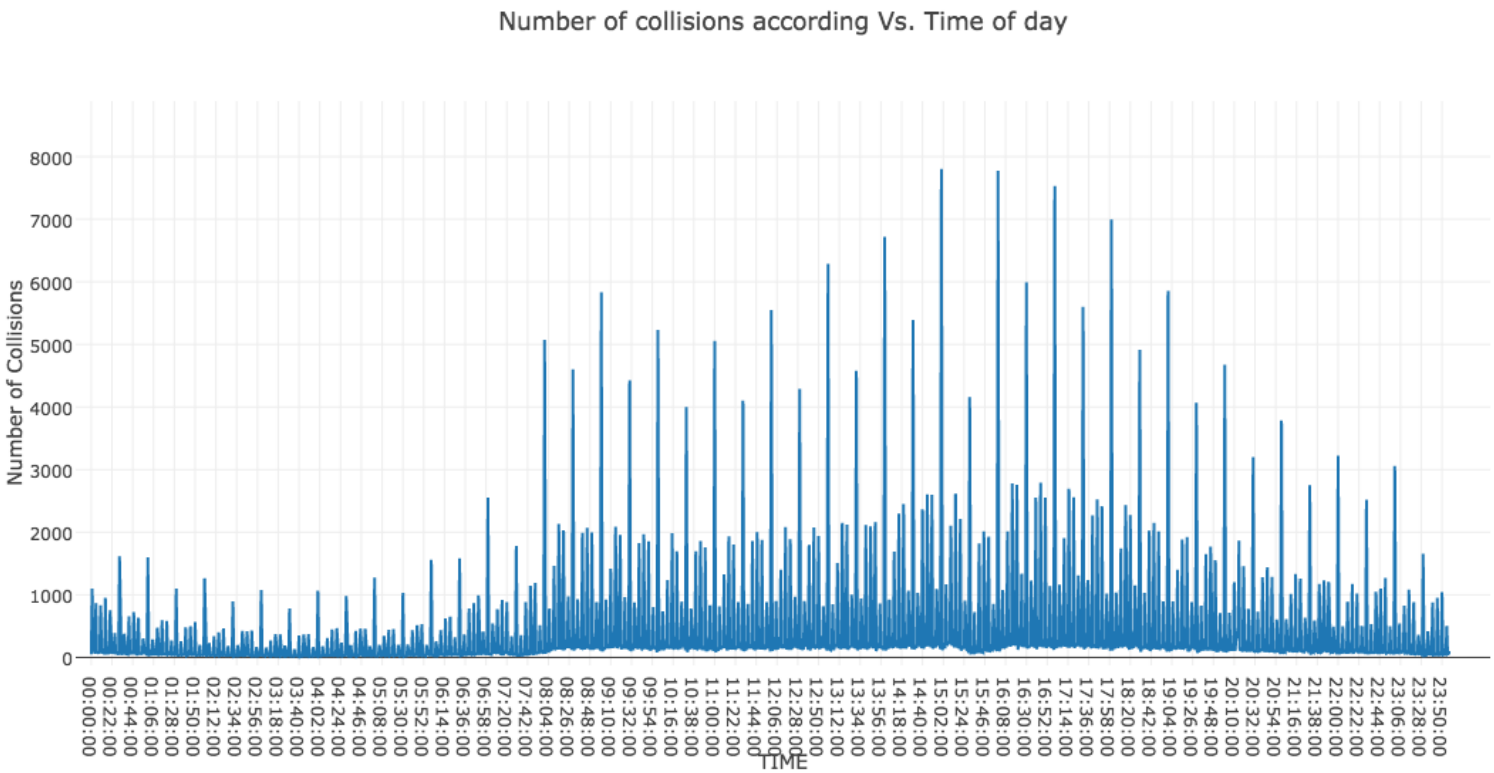
Results of t-test:

data: morningData$TIME and eveningData$TIME
t = -406.84, df = 397400, p-value < 0.0002
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-5.081581 -5.032854
sample estimates:
mean of x mean of y
11.11683 16.17405

Because the p-value is very low, we reject the null hypothesis and conclude that there's a statistically significant difference. The greater the magnitude of T (it can be either positive or negative), the greater the evidence against the null hypothesis that there is no significant difference. It proves our hypothesis that accidents are more likely during evening time which makes sense because, the light is hazy and traffic congestion is higher.

14. **What time has the highest accident frequency if the distribution is not uniform as proved by hypothesis testing?**

It can be seen that accidents are most frequent in the evenings ,closer to 16:30.



Conclusions

From the aforementioned analysis, some facts and patterns can be stated about NYC vehicle collisions. It can be inferred that Brooklyn has maximum number of collisions and they are concentrated mostly near the Brooklyn Bridge and Pennsylvania Avenue. Manhattan and Queens also have huge accident frequency. Passenger Vehicles are involved in almost 50% ob the accidents. Driver Inattention or distraction and Drowsiness are major contributing factors. Also, the accident frequency is higher in the evening since these are the busy hours for the city.

References

Leman, A. Statistics and Visualization [PDF document]. Retrieved from Lecture Notes Online Web site: https://blackboard.stonybrook.edu/bbcswebdav/pid-3495709-dt-content-rid-21975973_1/courses/1158-CSE-590-SEC01-81617/05-Statistics-II%20%281%29.pdf

Tufte's Design Principles. Retrieved from http://www.sealthreinhold.com/school/tuftes-rules/rule_one.php

R ggmap documentation. Retrieved from http://stat405.had.co.nz/ggmap.pdf

R spatial Data Visualization. Retrieved from: https://stablemarkets.wordpress.com/2014/09/01/spatial-data-visualization-with-r/

R spatial maps guide. Retrieved from: https://www.nceas.ucsb.edu/~frazier/RSpatialGuides/ggmap/ggmapCheatsheet.pdf

R GIS tutorial. Retrieved from https://pakillo.github.io/R-GIS-tutorial/

Plotly. Retrieved from https://plot.ly/

Statistical Hypothesis Testing(n.d.). Wiki: https://en.wikipedia.org/wiki/Statistical_hypothesis_testing

RStudio Installation - R Documentation Installation link: https://support.rstudio.com/hc/en-us/articles/200552306-Getting-Started

NYC Open Data Portal : https://data.cityofnewyork.us/view/m666-sf2m