Vehicle Collision Analysis in New York City

Bijay Koirala

Data Science I - CS 287

Nov 2, 2015

ABSTRACT:

Vehicle collision is a major stumbling block in an extensively populated city like New York. The systematic road and technology implemented in controlling vehicle collision does not seem to be working effectively. Therefore, the purpose of this analysis is to carefully weigh different factors that contribute and result in vehicle collision. Data from New York City Open Data was analyzed for relationships between vehicle collision and different factors like hour of the day, vehicle type etc.[1] The findings of this paper provide clear understanding of the relation between vehicle collision and contributing factors. Taking a step ahead, this paper also looked into the relationship between 311 calls and vehicle collision and found that they are highly correlated. The finding of this analysis can be utilized on preparing future vehicle collision interventions.

INTRODUCTION

Traffic accidents at New York City result in huge cost to society in terms of death, injury, lost productivity and property damage. The factors that cause accidents are not well known.[2] More than 200,000 motor vehicle collisions happen every year - an average of 3 accidents per minute.

In this paper, a thorough analysis of vehicle collisions based on time of the day, days of the week and month were made to investigate the relationship between collisions and these time factors. The collisions were also analyzed based on their locations to figure out the cause factors of collisions. Dataset containing 311 requests and vehicle collision were also analyzed to investigate correlation between them[1].

The purpose of this paper is to profoundly lay out the causes of vehicle collision in NYC. Clear understanding of the relationship between the causes of accidents will possibly provide a controlling intervention to implement and reduce accidents in the future. The primary goal of this research is to analyze the present data to find out the precise locations in New York that were safe and also the areas that encountered most of the collisions. Also to look at which reason plagues what part of New York City, so that the result could be handy to come up with recommendations.

DATA / MATERIALS

Source of Data

New York is one of the few states which provides datasets of public information easily available to interested people through its data portal. The first dataset used was related to the vehicle collision information in New York City obtained from NYC Open Data.[1] The dataset had 670,104 data of collision information from collisions between 2012 and 2015. To be specific about the data, it was a CSV file which had information regarding dates, times, locations of collisions along with number of persons, cyclists, motorists, pedestrians etc. and the reasons of collision with respect of each vehicle.

Second dataset was a dataset of over 9 million 311 requests made between 2010 and 2015, also obtained from NYC open data.[1]

Data Cleaning

The datasets obtained were already to the point and required very less cleaning. In most of the cases, empty results or values like "Unspecified" had to be filtered. Empty results for locations were filtered as well. For the dataset with 311 results, only relevant information for the purpose of the study such as dates of complaints, location information and zip codes was extracted.

Design

The main motivation behind this analysis was to address the following questions: -

1. Do vehicle collisions occur uniformly throughout the city or are they more concentrated on certain areas?
2. Are collisions affecting cyclists, motorists or pedestrians uniformly?
3. Are there specific reasons causing more collisions in certain parts of the city?
4. Are certain types of vehicles more involved in collisions in certain parts of the city?
5. Is there a correlation between numbers of collisions with the number of 311 requestss?

RESULTS

During the analysis, it was found that, 18.65 % of the accidents caused injuries and 0.13% of them were fatal. More collisions were observed during A.M. rush hours of 8 to 9. Which decreased briefly for a while then peaked again peak at 4pm and then falls gradually. It was also observed that the number of collisions were highest in Fridays, fell through Saturdays and Sundays and stayed constant throughout other weekdays.

To find out if the collisions occurred uniformly at all the locations, locations of collisions was collected, and after analysis it was found that the distribution is not uniform at all. Figure 1. below shows a plot of the areas that observed the most number of accidents and those that observed the least number of accidents. The bigger the radius of the dots, the bigger was the impact of collisions and the more compact they are, the more frequent are the accidents. The regions plotted in green are the regions that observed lowest number of collisions. Therefore those areas with green circles can be considered as safe areas. It is also observed that on certain areas, the rate of collisions in exceptionally high that might require immediate attention.

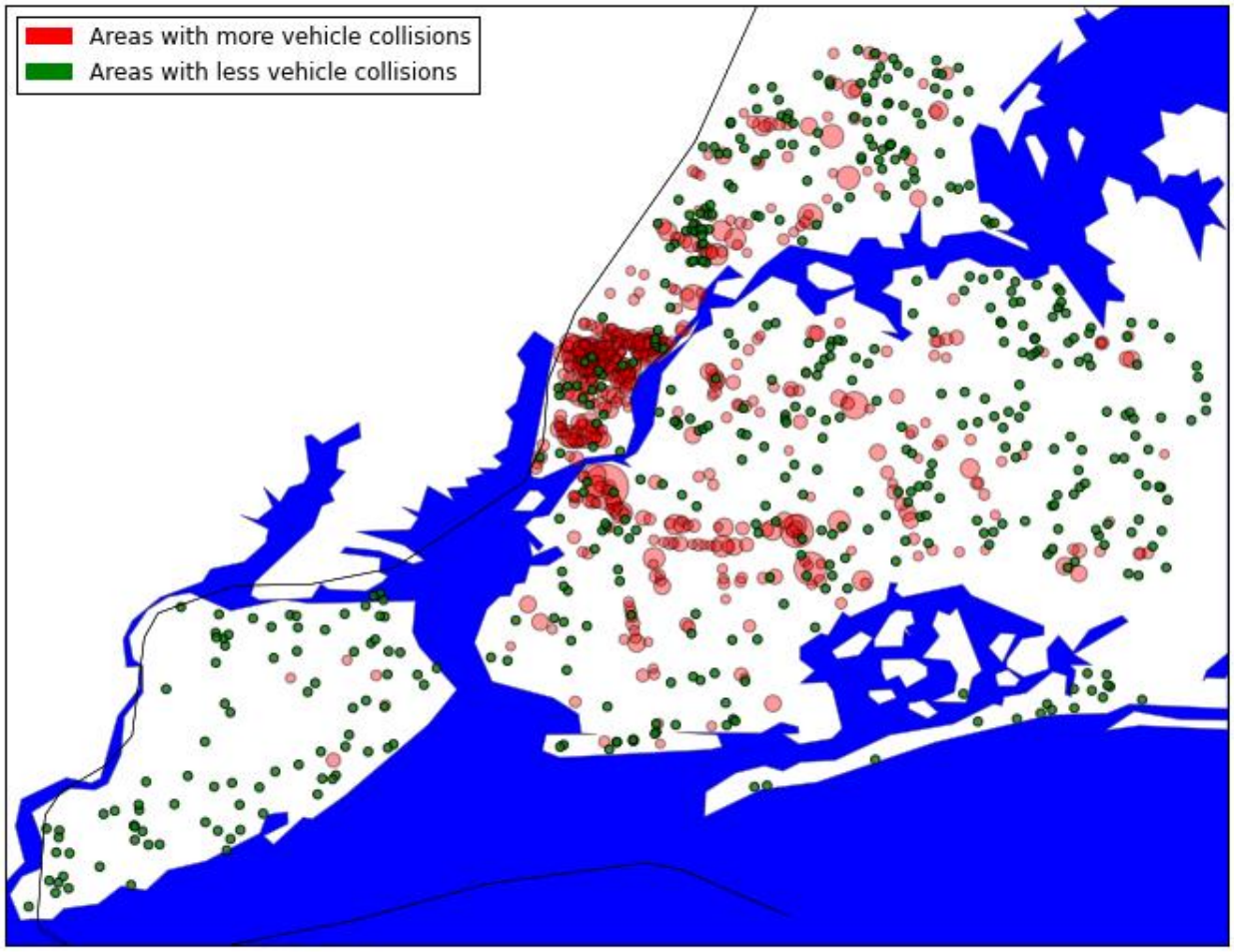

Figure 1. Plot of areas with higher collision and lower collisions in New York City.

Figure 2. Shows the impact of collisions on cyclists, pedestrians and motorists. Results indicated that even though the injuries (represented as circles) are somewhat concentrated on certain areas, the deaths (represented with squares) are distributed uniformly.

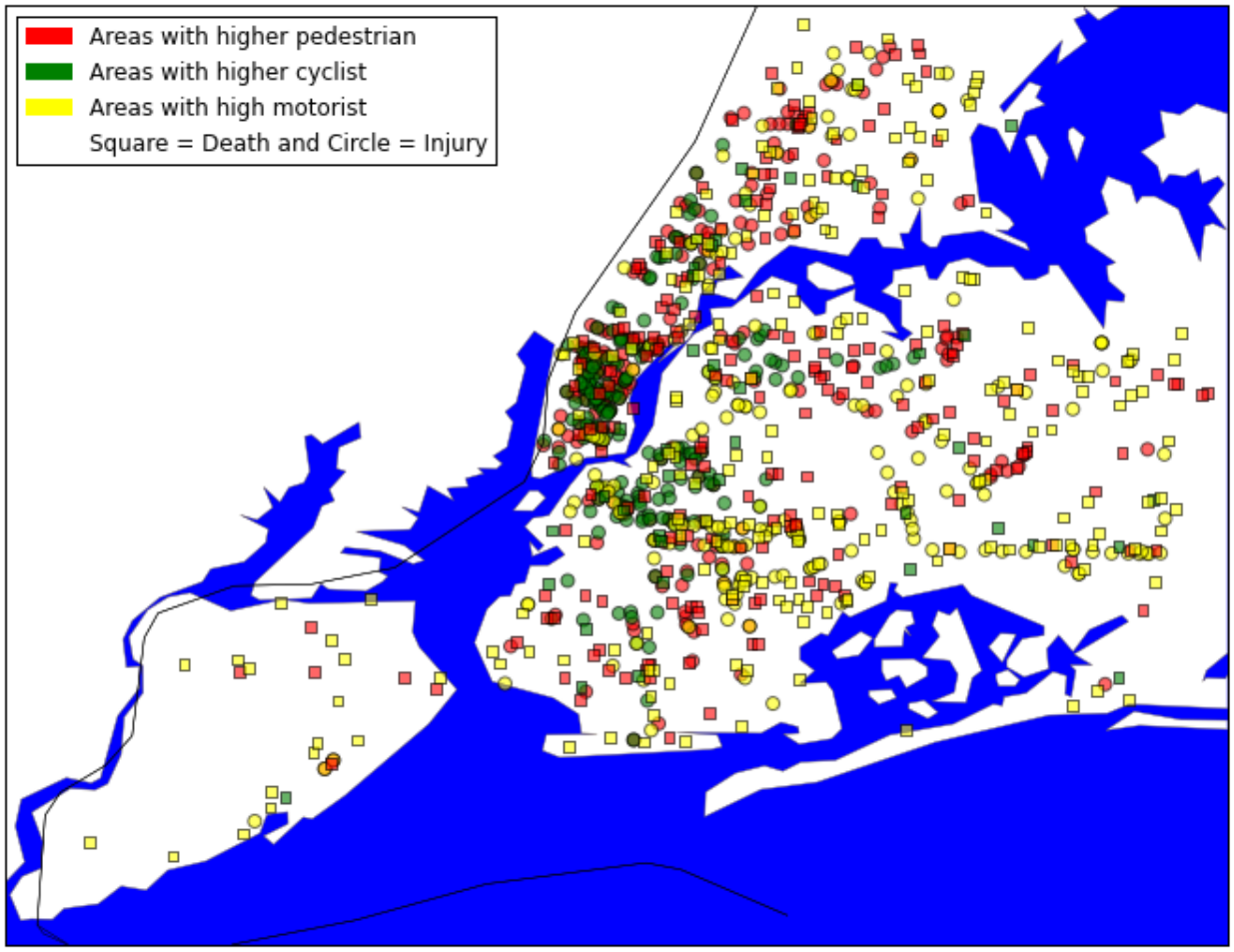

Figure 2. Plot of injury and death caused to pedestrians, cyclists and motorists.

Next analysis was to investigate if there are regions where more collisions are happening due to certain factors. The reason for the highest number of collisions was Driver Inattention / Distraction, followed by Driver being fatigued or drowsy, backing unsafely, turning improperly and losing consciousness. Alcohol Involvement, surprisingly was the 14th reason in our analysis just after "Pavement Slippery". Fatal accidents were concentrated more on regions outside of busy streets, non-fatal accidents were more concentrated in busy areas of city.[3] For this, the top 6 reasons for collisions were considered. On those considered the 100 most affecting places of all the factors was plotted to observe a pattern if present. Observations showed that downtown Manhattan area was the most heavily influenced for all the above factors. It can also be observed that accidents due to fatigue are more frequent on streets which are not very busy. People not following rules properly is causing more collisions in the lower Manhattan region.

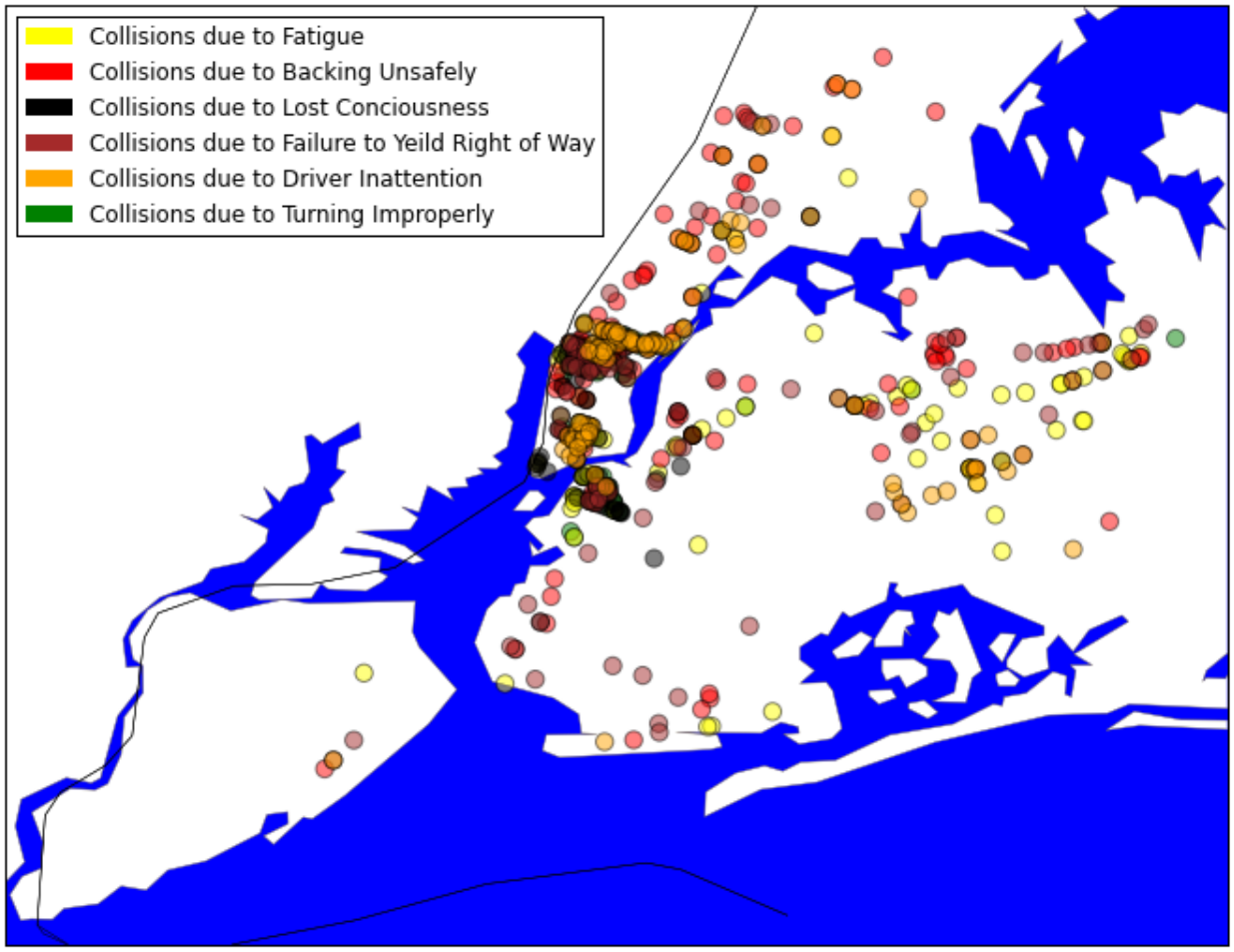

Figure 3. Major reasons of vehicle collisions and the area they affect.

Figure 4. shows the kinds of vehicle that are involved in majority of collisions. Four types of vehicles were found to cause majority of the accidents. The highest number of collisions was caused due to passenger vehicles followed by Sports Utility and those were scattered in Manhattan, Bronx and Queens. The next two vehicles in the lists were taxis and vans which were somewhat concentrated on the downtown Manhattan region.

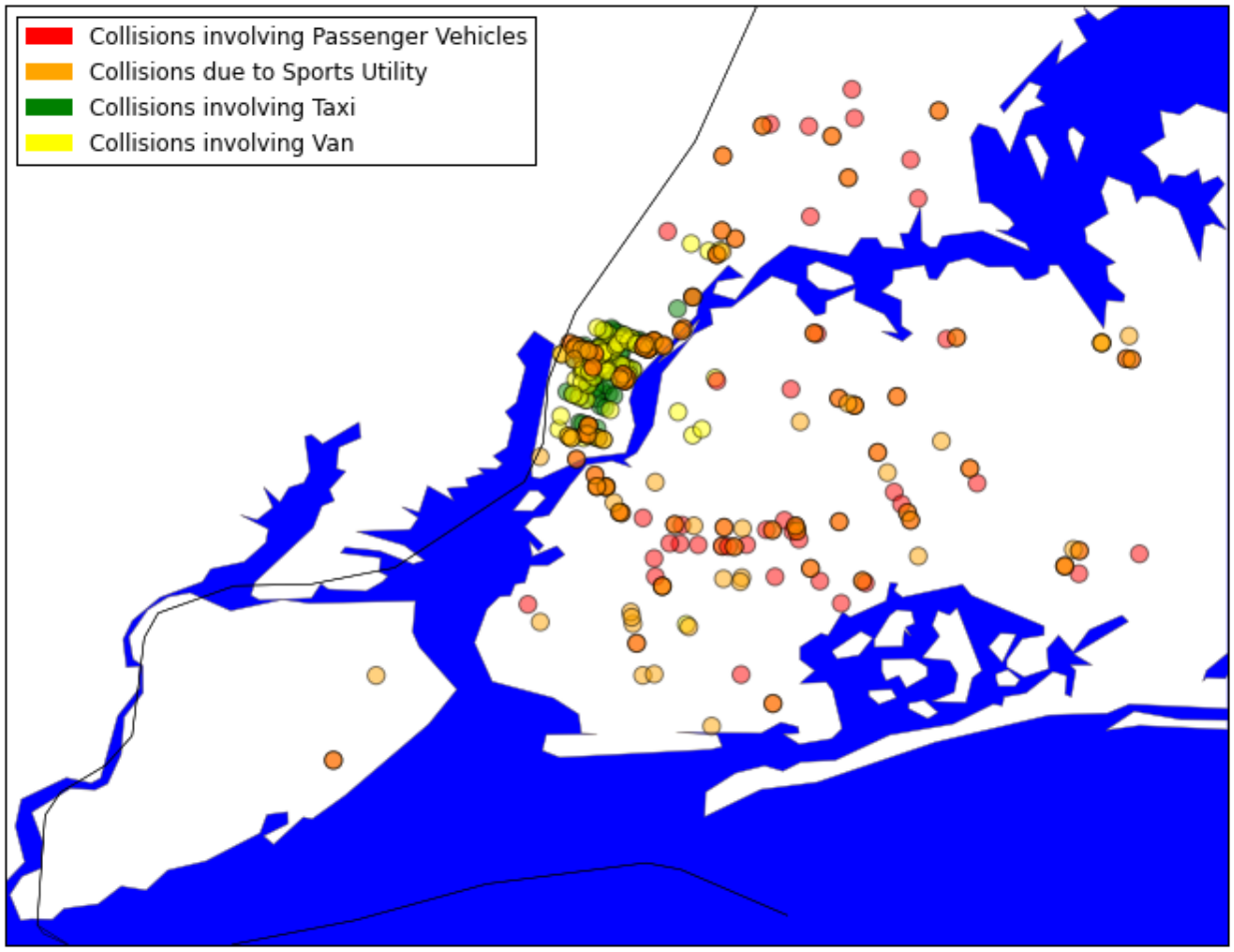

Figure 4.Type of vehicles involved in collisions and areas impacted.

Figure 5 shows the relation between the number of vehicle collisions and non-emergency calls. For this the numbers of 311 requests and vehicle collisions across each zip code of the city were analyzed. The scatterplot includes the 10 percentage sample of whole data considered for the results. The plot reveals a relationship between 311 class and number of vehicle collision. Regression analysis on the data resulted in R squared = 0.465006674248, p-value = 2.47006667825e-27. This r squared value represents that there is a relationship between the two variables.[4] To get more accurate results, same procedure for latitude longitude was carried by reducing the precision level of latitude/longitude pairs in order to divide the city a grid of over 900 parts. The analysis also formulated similar results with R squared = 0.4920006132 and p-value = 3.02073461995e-135. The decent value of R-squared and the extremely little p-values indicate that it is highly unlikely that the two do not have correlation.

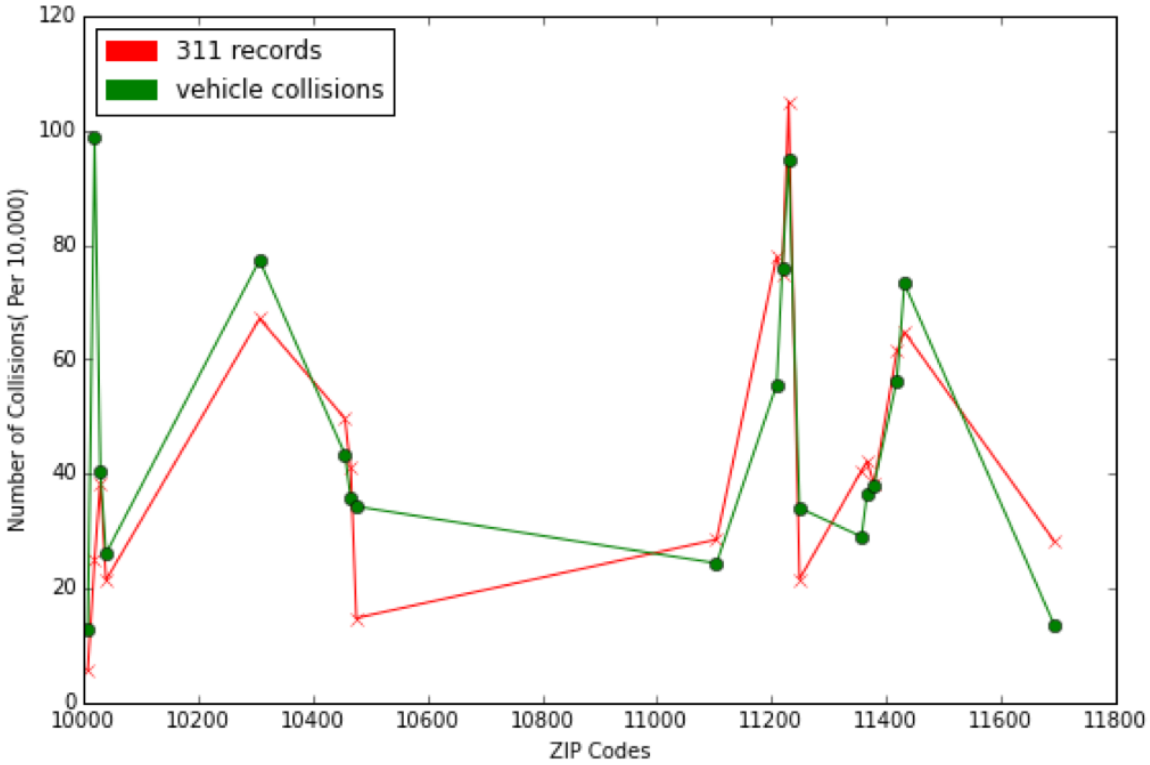

Fig 5. Sample of 311 request and vehicle collision based on ZIP codes.

DISCUSSION

Vehicle collision is not evenly distributed throughout the city as indicated from the results. Some areas in the city are more affected compared to others. Moreover, certain areas are plagued with certain causes of vehicle collisions. The results of this study might be useful for authorities to implement different techniques and bring newer interventions for positive change.

The results above provides deeper insight into vehicle collision in New York City. Even though there are many regions which need to be improved, there are certain areas that need immediate attention. The results show that there are many accidents that are being caused due to slippery pavements, many accidents being caused because of improper turning and so on as figure 3. Figure 5, which shows the correlation between 311 requests and vehicle collisions, indicates that the areas with lower 311 requests also see lower number of vehicle collisions. Very small p-value obtained strongly suggests that 311 call have an effect on number of vehicle collision. Further research into 311 call details might bring about new method to control accidents. All these factors can be addressed by working on them immediately. Speed humps might be incorporated on areas where more accidents are being occurred because of the driver being drowsy. Since fatal accidents are occurring in regions outside of busy streets, traffic regulations should be tightened on those regions and, speed should be controlled and improvements must be made traffic signs[3]. Especially in lower Manhattan area, and a few regions of Bronx seem to appear on the statistics all the times and in my opinion are the highest priorities to be dealt by the DMV.

BIBLIOGRAPHY

1 "NYC Open Data." *NYC Open Data*. Web. 20 Oct. 2015.

2 Poch, Mark, and Fred Mannering. "Negative Binomial Analysis of Intersection-Accident Frequencies." *Journal of Transportation Engineering J. Transp. Eng.*: 105-13. Print.

3 Mueller, Beth A., Frederick P. Rivara, and Abraham B. Bergman. "Urban-rural Location and the Risk of Dying in a Pedestrian-vehicle Collision." *The Journal of Trauma: Injury, Infection, and Critical Care*: 91-94. Print

4 Hoel, Paul G. *Introduction to Mathematical Statistics*. 4th ed. New York: Wiley, 1971. Print.