# Analysis o Effects of Socio Economic factors leading to Traffic Collisions

**Naga Jyothi K, Neha Kumari, Madhumita Deepika Duvvuri**

**5/13/2016**

**Statistics, Data Analysis, and Decision Modeling**

**Professor. John Wang**

## Acknowledgements

This research was supported by Saint Peter's University, Jersey City (New Jersey). We are immensely grateful to Professor John Wang (Statistics, Data Analysis and Decision Modelling) for his assistance with data modelling and data analysis methods; and also for his help in moderating this paper and in that line helped us on improving the manuscript significantly. This research has given us a great opportunity to learn the techniques and data modelling methods used to perform analysis on the major factors causing Motor-Vehicle-Collisions. We hope this thesis will relate to various other studies and find it intelligible.

# Contents



## Table of Figures

# 1 Abstract

The goal of this paper is to perform analysis on the traffic collisions in the city of New York in the past five years. This analysis primarily focuses on the sociological, geological and economical factors and their effects leading to the traffic collisions in New York City. A main objective of this document was to determine the average number of collisions every year and the duration of highest rate of collisions on a monthly time-scale.

The average and rate of collisions were calculated based on the contributing factors in the various boroughs in New York. These collisions were identified in parts of the city where the traffic volume is available. The detailed traffic collision data can be used to identify the prime locations where the collisions occur and also the time of the year when the number of collisions are high.

Another primary objective of the analysis was to use the data to provide information related to the problem identification or causes for these collisions. We will be considering the characteristics of the driver (alcohol involvement, drowsy, inexperience, lost consciousness), characteristics of the vehicle driven (type of the vehicle, brakes defective, steering failure), characteristics of the collision (aggressive driving, reaction to other uninvolved vehicle) and characteristics of the environment in which the accident occurred (weather, pavement slippery, animal action). This information can be used later to find a solution or avoid the traffic collisions and help to decrease the rate of accidents.

The traffic collision data are available in NYC Open Data – NYPD Motor Vehicle Collisions. This database is regularly updated on daily basis so the count of the collisions in a given year will continuously change for a substantial time after the end of that year.

# 2 Introduction

With the increasing number of vehicles and the increase of population in the world today, traffic collisions have become a very common sight to us. A traffic collision or accident (commonly known as MVC motor-vehicle-collision) occurs when a vehicle collides with another vehicle or object like animal, pedestrian, stationary objects or architectural obstacles which usually result in injury, property damage, and/or death. There are many statistical numbers relating to collision which we should be aware of and might be helpful in bringing awareness among people, and thus hope to decrease the rate of collisions happening in the United States.

*"On an average, there are about 6 million motor vehicle collisions on the roads of the US, annually."* (see National Center for Statistics and Analysis 2014)

More than 3 million people get injured and more than 2 million of these injuries are permanent. There are in excess of 40,000 people who get killed due to the collisions every year(see National Center for Statistics and Analysis 2014). "Every 12 minutes, a person dies because of a traffic collision on US roads. And every 14 seconds, a MVC results in an injured victim." (reads the first line from National Center for Injury Prevention and Control)

Many of these collisions can be avoided if the drivers are more responsible. About 40% of the deaths caused by traffic collisions are due to drunken driving. Around 30% can be attributed to driving above the regulated speed limits and 33% or more can be because of reckless driving which usually results for the vehicle to go off the road or collide with another. In most of the collisions, the deceased were found not be wearing their seat belts at the time of the accident. The mandatory regulations are also not being followed, which intensifies the magnitude of the collision and people end up losing their lives or even end up severely injured.

The other causes for traffic collisions can be due to distractions like people talking to other passengers while driving, drivers playing with radio while driving, eating or drinking while driving and the highest percentage is because of talking or texting on the cell phone while at the wheel of the car. Looking at the above causes we can say that the collisions can be easily avoided if the drivers paid more attention on driving and less attention on other activities.

One of the major reasons for these collisions can be the population i.e. the crowded cities. Traffic-choked and crowded north eastern cities like Washington, D.C., New York, Baltimore, Boston, Newark, Philadelphia, Hartford and Massachusetts are most likely to have more number of collisions. According to the Fatality Analysis Reporting System (FARS) run by the National Highway Traffic Safety Admission (NHTSA), approximately 30,000 fatal motor vehicle collisions occur every year, and continue to be the leading cause of death for young adults (NHTSA, Fatality Analysis Reporting System, 2010). To understand the reasons behind the collisions and how the fatality rate can be reduced we have performed analysis on collisions in the city of New York.

New York is one of the busiest cities and there is no other place like this in the world. Millions of people come to this city every year for many reasons from business to pleasure, added to the heavy population of New York. This makes it a very exciting place to live and work making it a very crowded place. Now where there are lots of people, there will be a great deal of traffic and congestion on the roads. As the City's population and tourism have grown, the New York City Department of Transportation has worked to improve the safety, efficiency, and liveability of the streets by supporting travel options that require less space such as walking, riding the bus, and bicycling, while maintaining appropriate traffic flow for cars and trucks. These steps, including pedestrian safety projects, bus lanes, bicycle lanes, and plazas, are core to the City's safety and sustainability goals and fundamentally allow more people to access and enjoy the Central Business District without adding more cars on the road.

Permanently or temporarily reducing the capacity of roadways generally increases motor vehicle congestion, shifts vehicle travel to other routes or times of day, or causes some traffic to disappear as travellers shift to other modes or otherwise change their pattern of travel. Blocking lanes for construction of buildings, subways, underground utility infrastructure, or road repairs can also reduce road vehicular capacity depending on how many lanes are blocked. A large number of people arriving by subway or bus increases pedestrian volumes, which can reduce motor vehicle road capacity by as much as a third as turning vehicles must wait for pedestrians crossing at intersections. Motor vehicle traffic congestion is also significantly affected by how curb space is managed.

All these conditions put together give us the reason why many people find themselves involved in collisions every year in New York. As the roads of New York are congested, motor-vehicle accidents are unavoidable and many occur every year which usually result in simple shocks to serious injuries or even fatalities. The causes can be anything like driver negligence, driving conditions, alcohol or drugs, faulty vehicle parts or a variety of possibilities.

In this analysis, we will discuss the various factors like sociological, geographical and economic responsible for the motor vehicle collisions occurring in the city of New York. We will first discuss about the various variables that other studies talk about. When we look at a collision report we have several variables we need to consider. These variables can be placed in the following categories: characteristic of the person killed/injured and environment in which the accident has occurred. The statistics on the reasons of collisions can vary from study to study based on the factors considered. For instance, around 37% of all nationwide fatalities in motor vehicle related incidents in 2006 involved alcohol (AlcoholAlert, 2010). A study based on all collisions states that 57% of accidents are mainly because of driver factors, 27% to combined roadway and driver factors, 6% to vehicle and driver factors, 3% only to roadway factors, 3% to combined roadway, driver and vehicle, 2% only to vehicle factors and 1% to combined roadway and vehicle factors (Lum & Reagan, 1995).

The major contributing factor is the driver's behaviour while driving a vehicle. To avoid a collision it is required that the driver pays proper attention and also be attentive about the condition of the vehicle. There are many variables we need to take into account while analysing the collision data and these are described in the data section.

# 3 Developing the Data

We will analyse a multitude of the predictor variables. We shall discuss the data in a general perspective, the response variable, the predictor variables and the regression model.

Before we get to know about the different variables, we can discuss on how the data was retrieved. The data was collected from NYC Open Data which is a data hub for more than 1300 datasets. The details of the Motor Vehicle Collisions in New York City are provided by the Police department (NYPD). The NYPD completed a MV-104 report dating back to July 1st 2012 and the details of collisions are available from that date until the present day. The data which is available online makes it possible for us analyse and figure out an approach to prevent or avoid traffic collisions. The contributing factors available here provide us the major reason why collisions occur. To analyse this we have chosen the predictor variables considering the characteristics of the variables.

## 3.1 Driver Characteristics

There are many characteristics of a driver that could affect the odds of a collision to occur. They can be age, gender, distractions and actions. Driving is a complicated task, which involves several cognitive, physical and sensory skills working together. Distractions can be defined as a secondary activity that makes the driver to deviate his attention while driving. They have the potential to reduce the driver's performance and have serious consequences for road safety. According to the NHTSA Fatality Analysis Reporting System - 2010, most of the reported collisions are caused by driver's inattention and is estimated to 25 percent of the collisions.

Another major contributor in vehicle collisions is age. A study using multivariate logistic regression, shows that the odds ratio (OR) that a fatal injury occurs due to a collision increased with age (Bedard, Guyatt, Stones, & Hirdes, 2002). Gender can also be considered to be a characteristic that has significance.

In the same research study, the majority of crashes were among male drivers who were younger than 30 years, versus females of the same age. Actions done by the driver before the accident are also important to consider. Actions could involve many things like changing lanes, following another vehicle improperly, obscured vision, drug usage or fatigue and using of phone while driving.

## 3.2 Vehicle Characteristics

There are several characteristics of a vehicle which contribute to the analysis. The major ones for this paper are: the type of vehicle and whether it had airbag deployment. The type of vehicle is considered to be a characteristic to look at as some vehicles are rated as safer than others. Airbag deployment is also an important characteristic among researchers. Many studies reveal that airbags did not appear to be protective for drivers younger than forty years old and may have been unfavourable to drivers older than 60 years old (Bedard, Guyatt, Stones, & Hirdes, 2002).

## 3.3 Environmental Characteristics

Environment in which a collision occurred plays a very critical role. There are many environmental issues to look at when a motor-vehicle collision takes place. They could include weather, road condition, lighting, time of day, day of week, and season or quarter of the year of the crash. Environment also contributes to driver's distraction, which is considered as a major factor in collisions. Weather includes heavy rain, hail, snowstorms, high winds, blowing sand, fog, and other atmospheric effects. Weather effects always influences the driver in many ways; visibility, the ability to control the vehicle, and even the ability to hear. So, the possibility of a collision to occur during these times is higher (NHTSA, Trend and Pattern Analysis of Highway Crash Fatalities by Month and Day, 2005).

Road conditions include construction, ice, potholes and wetness which cause more crashes because it becomes hard for the driver to steer the vehicle if he is not familiar with the conditions. Time of day also can be considered when dealing with motor-vehicle collisions as during the early hours in the morning and evening the traffic in a congested city will be more which might lead to collisions.

## 3.4 Crash Characteristics

Crash characteristics mean the direction of accident and how many vehicles were involved in the crash. The various directions of a collision are head on, side impact, angle and rear-end impact. It was found that 65% of all crashes involved in front impacts. Multiple vehicle collisions are an issue heard across the world. They occur more often, but are not the most deadly type of collisions.

## 3.5 Zip Characteristics

Studies have reported that localities with higher population have a greater risk of collisions due to congestion. In this analysis, we will look at which borough has the maximum rate of collisions in New York City.

The major amount of the data that was provided from NYC Open Data forum was categorical. Categorical data is defined as a form of discrete data that describes some characteristic or attribute of that particular data. In most of the data, the variables describe several attributes ranging from what was the factor leading to crash, which intersections the accidents happened in, to whether there was a fatality or not, what type of vehicles were involved. This raw data was reconstructed into various categorical groups so that we can get a better sense of the predictor variables when we perform statistical analysis on this data. Using Microsoft Excel, we created a filtered view of the data which interpreted the word variables provided from the forum into numeric variables which could be used in the next process of analyzing the data in an easy way.

The data collected was divided into three parts for the analysis. The first part will be the contributing factors (driver's characteristics, environmental characteristics). The second part will be the vehicle characteristics and third will be the zip characteristics. This data was imported to the database software package Microsoft SQL Server and by using this software we can create and edit the large databases. Many queries are performed on these databases to get the appropriate data for our analysis. The data retrieved from NYC Open Data was huge which covered around 720,000 collisions, and using SQL seemed the best way to collect, organize and parse the data to get what was required for this study. These 720,000 data entries involved all collisions which occurred from July, 2012 in the city of New York.

# 4 Analytical Framework: Analyzing the data

After gathering and partitioning the bulk of data into three parts, the next step will be to understand which variables can be taken as the response variable and predictor variables. Before doing this we need to understand the statistics of the collision data which was retrieved based on the several contributing factors, vehicle types and zip characteristics.

The data entries of motor vehicle collisions maintained by the NYPD have several key pieces of information attached to them. These include the causes of the crash, various types of vehicles which were involved (motorcycles, buses, taxis, etc.), how many people were injured or killed (whether they are motorists, passengers, pedestrians, or cyclists), and the place of the collision (zip). We have used all the above metrics to analyze traffic accidents over the last few years, beginning with an overview of all collisions across the five boroughs (Bronx, Brooklyn, Manhattan, Queens and Staten Island).

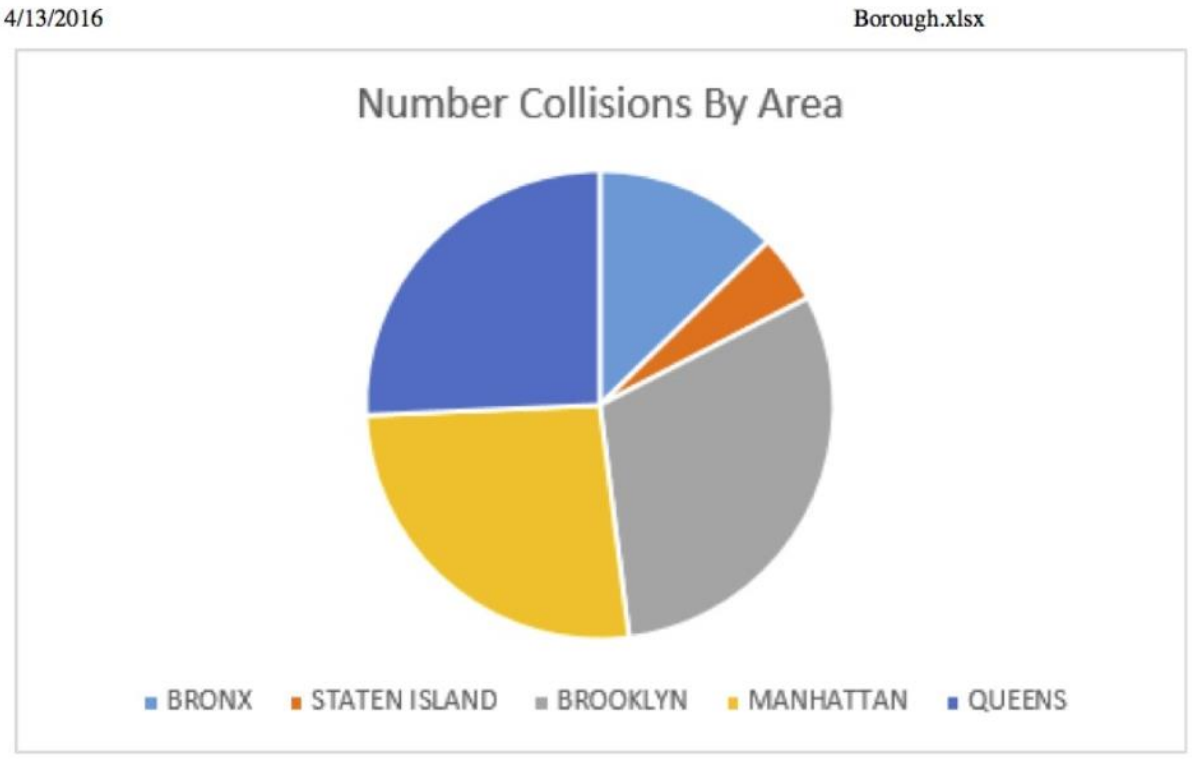


Figure 1: Percentage breakdown of collisions in five boroughs in New York

From the percentage breakdown of collisions in the five boroughs, the maximum percentage or rate of collisions was seen in Manhattan and Brooklyn than the other boroughs over the years. But if we look closely they were just more densely concentrated in this part of the city. In fact, Queens had marginally more than the other boroughs, accounting for 25.69% of the total number.

All five boroughs saw an increase in traffic collisions from March to September (summer time), with Queens, Brooklyn and Manhattan showing the most obvious upward deviations during this period on the graph Number of Accidents versus Months (for each borough). The time series below gives us the idea on how the number of collisions increase or decrease at a particular time of the day.

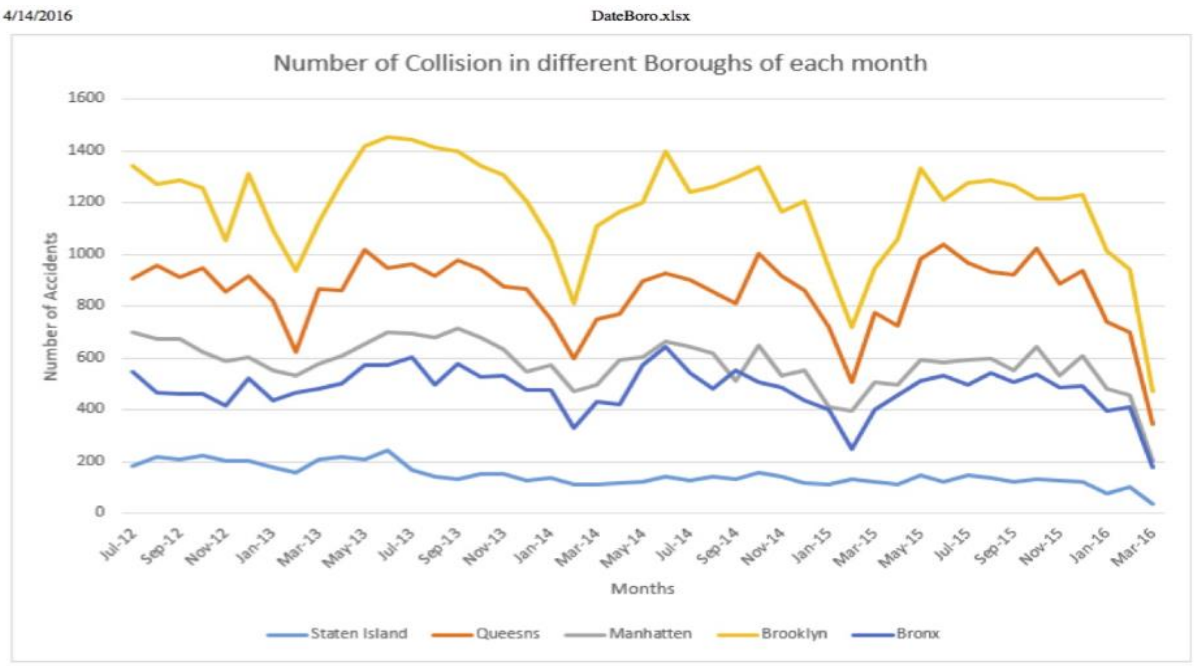


**Figure 2: Monthly line graph - Rate of Collision in each Borough**

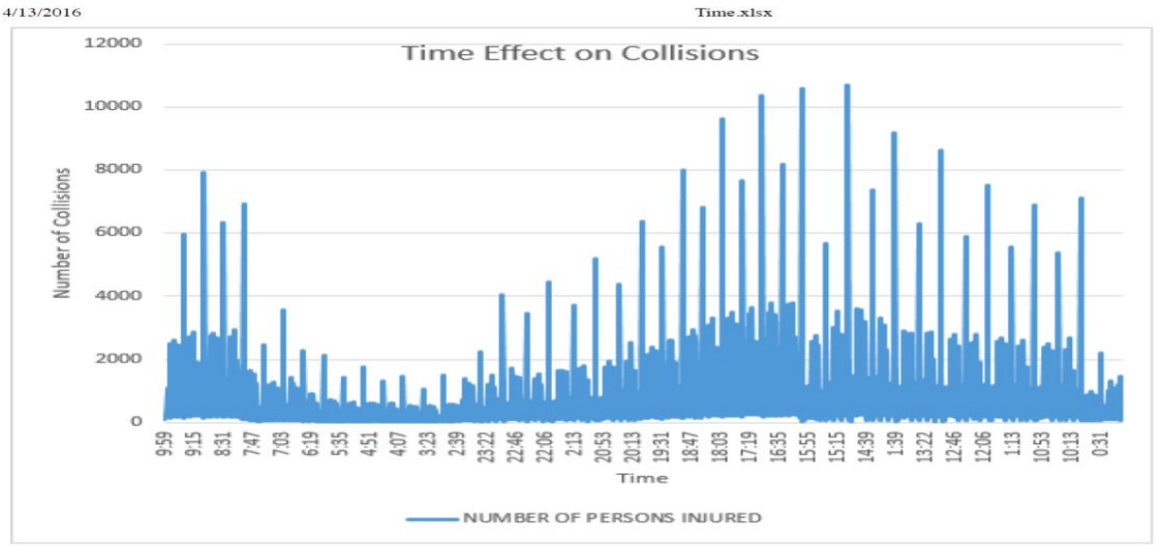


**Figure 3: Collision rate based on time of the day**

Except for Staten Island, all four boroughs the line chart changes dramatically due to the collisions that happened in December 2014, followed by January 2015 (it is the lowest number in the entire period covered). If we observe clearly collisions happening during summer are much higher than winter period. From the time series chart, we can say that the incidence of traffic collisions rises sharply from 7 to 9 a.m., when hundreds of thousands of people are commuting into and around the city to get to work and then levels off until 3 p.m. when it begins to rise to its highest level with people commuting back home from work or out of the city. Statistics from the New York State Department of Motor Vehicles echo this finding: Traffic accidents are most likely to occur between 3 and 6 p.m. This gives us better sense of how traffic accidents in New York City are distributed across time and geographical space.

According to the statistics now we know where collisions happen most in New York City, but how does the rate of collisions change by vehicle involved and the cause of crash? The most vulnerable type of vehicle in an NYC traffic accident is the bicycle – 77.5% of collisions involving bikes resulted in an injury. Scooters and motorcycles followed in second and third place.

Taxi collisions taking fifth place, are more likely to involve injuries than regular passenger vehicles. This could be because taxi drivers in New York frequently don't wear seat belts nor are they legally obliged to. Many taxi passengers forego the use of seat belts too which causes severe injuries.

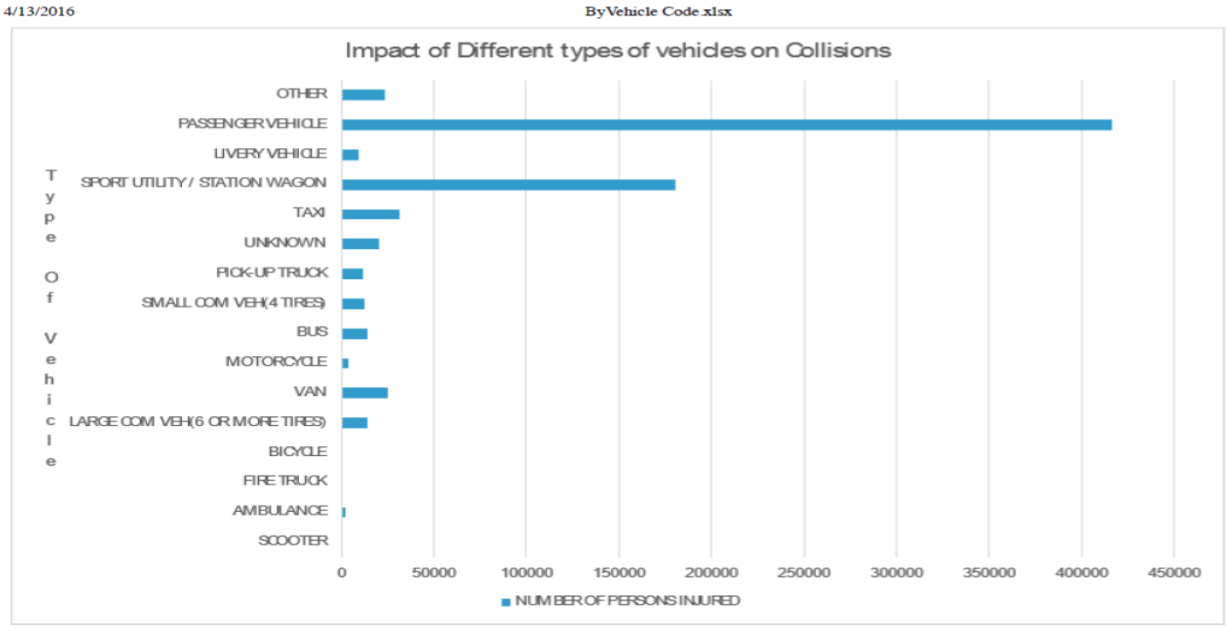


**Figure 4: Impact of different vehicle types on Collisions**

The likelihood of a traffic accident causing an injury or death depends on several contributing factors, one of which is the vehicles involved in the crash. Accidents involving taxis that resulted in injuries showed some interesting differences between boroughs as well. In the Bronx, 30.5% involved injuries, compared to only 17.1% in Manhattan. This leads us back to a question we posed at the beginning of this article:

*When we observe types of accidents by people injured in which type of accident will it be considered?*

For instance, Interstate 495 also known as the Long Island Expressway, had the most collisions in Queens in April 2015 – 54 people were injured in 230 accidents, 30 of whom were motorists, 23 passengers, and one a pedestrian. How can we classify the type of accident based on the people injured or killed? There were 223,141 traffic collisions in New York City between March 2014 and April 2015, almost a quarter of which resulted in an injury or death (NHTSA 2013). This statistic can be reduced if the rate or number of collisions can be reduced.

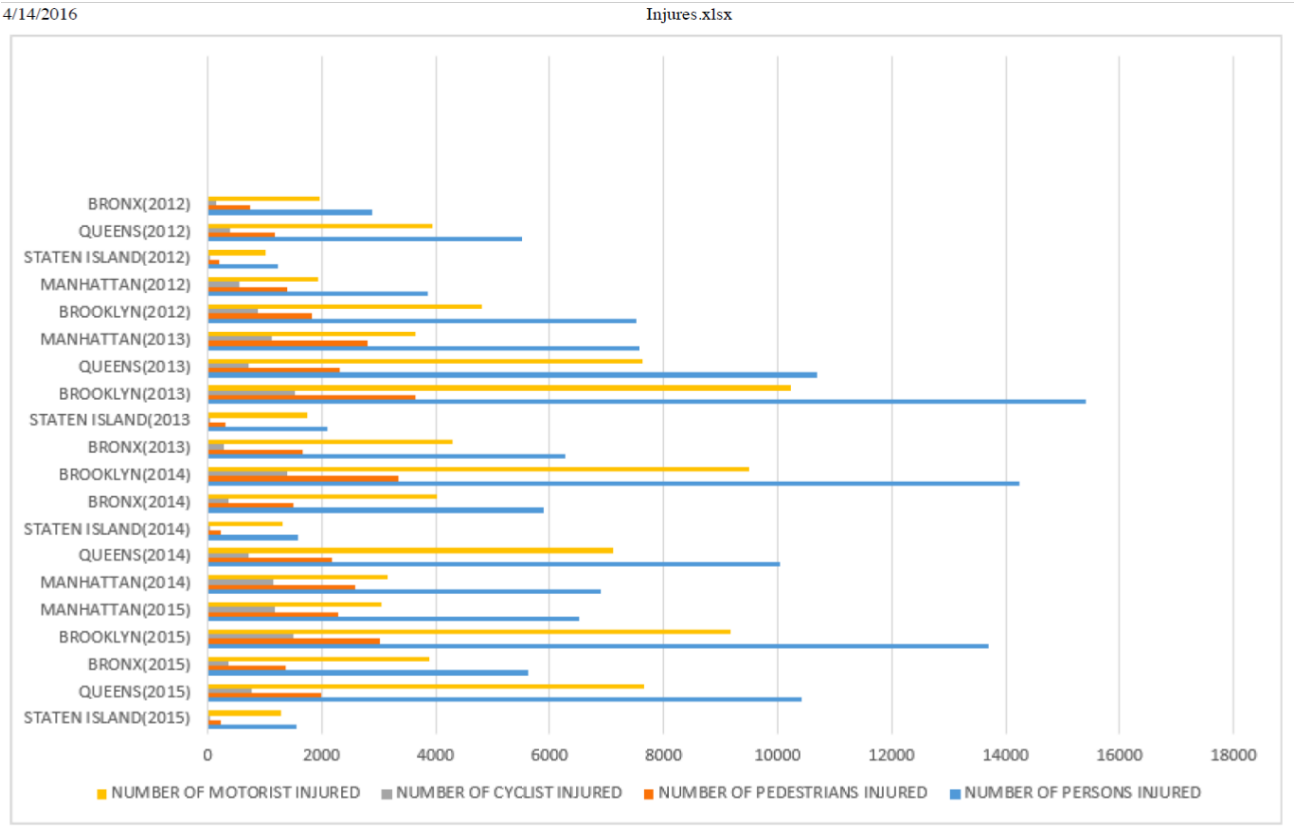


**Figure 5: Number of People Injured in each borough**

Why might the percentage of collisions resulting in at least one injury be considerably lower in Manhattan than the other four boroughs? The answer is probably speed – or rather a lack of it. According to GPS data, the average speed of a New York City taxicab in 2012 was 10.2 miles per hour. This doesn't make injuries from collisions impossible, of course (drivers can still fail to yield to a pedestrian), but it might explain why. If you're unlucky enough to get caught in a collision in Manhattan, you're less likely to get hurt than if you were in the Bronx, Brooklyn, or Queens.

The number of collisions that occur throughout the nation can determined mainly from the cause or the nature of the crash. If this cause can be avoided or at least minimized to a certain extent then the number of causalities will come down. There are many factors which lead to a collision and from the discussion in our paper the major factor comes down to distracted driving or driver's inattention while on road. According to a 2009 report by the National Highway Traffic Safety Administration (NHTSA), one out of every five crashes across the nation that resulted

in an injury was caused by distracted driving, and it was the cause of 16% of all fatal accidents. Interestingly, in New York City, driver's distraction did not cause as many injuries as drivers being distracted by their passengers. The NYPD's figures show that 18.5% of accidents caused by driver distraction resulted in at least one injury, compared to 67.3% of accidents caused by passenger distraction. This fits with research conducted by the NHTSA, which showed that passenger distraction was the most common cause in a sample of 7,000 car accidents (NHTSA 2013).

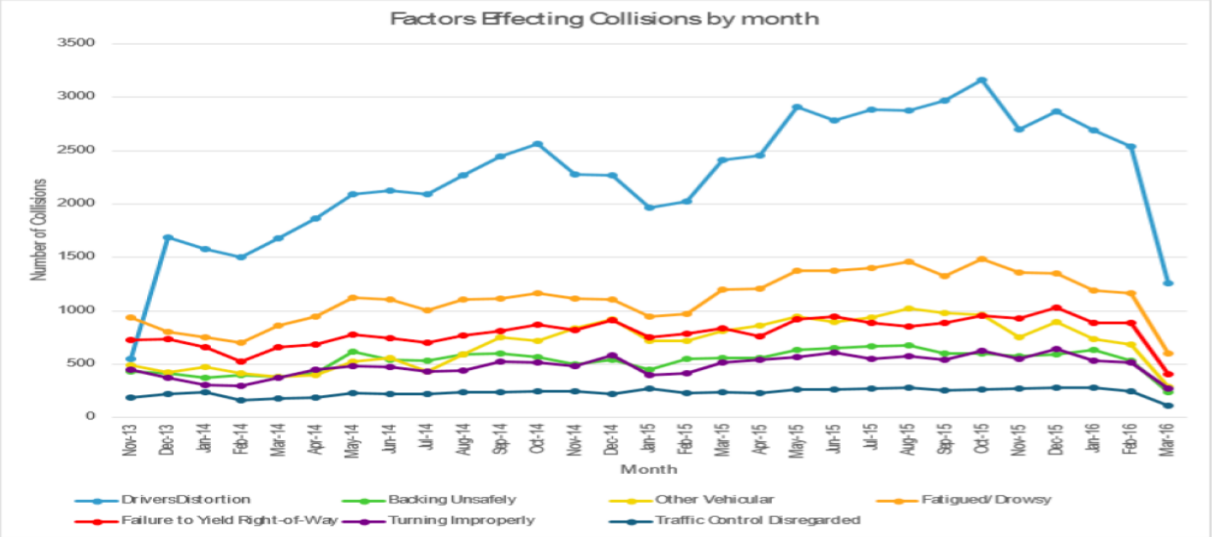


**Figure 6: Monthly Line graph based on Factors relating to Collisions**

People try to multitask while they are driving like talking on phone, texting, adjusting the radio station or even eating. This multitasking reduces their concentration on driving and the road. This is not just an opinion but it is a research finding from the National Highway Traffic Safety Administration. Now, citing data on the impact of cell phone calls on driving (even hands free phones), the National Traffic Safety Board is pushing to ban all use of all electronic devices other than GPS systems while driving, except for emergencies—a move that many drivers (and hands-free device makers) are sure to protest.

# 5 Methodology

## 5.1 Different Variables and Definitions

Methodological issues have been identified in the traffic collision frequency literature over the years. These issues shown have a potential source of error in terms of incorrectly specifying statistical models which may lead to erroneous collision-frequency predictions and incorrect inferences relating to the factors that determine the rate of collisions.

### *Over-dispersion*

One notable characteristic of this data is that the variance exceeds the mean of the collision counts. This poses some problems because the properties of the most common count-data modelling approach (the Poisson regression model which is discussed below) restrict the mean and variance to be equal for the particular data. When we have over-dispersed data present, to estimate a common Poisson model can result in an inconsistent parameter estimates. This will lead to erroneous inferences regarding the contributing factors that determine the rate of collision (Park and Lord, 2007).

### *Under-dispersion*

It is noted that sometimes collision data can be characterized by under-dispersion, where the mean of the count data entities is greater than the variance, especially when the sample mean value is very low. Previous work has shown that many traditional count-data models produce incorrect parameter estimates in the presence of under-dispersed data (see Oh et al., 2006; Lord et al., 2009).

### *Time-Varying Explanatory Variables*

Because collision data are considered over some time period, the fact that explanatory variables may change significantly over this time period is not usually considered due to the lack of detailed description of data within the time period. Ignoring the potential within-period variation in explanatory variables may result in the loss of potentially important explanatory information. This can introduce error in model estimation as a result of unobserved heterogeneity (see Washington et al., 2010)

### *Low Sample Mean and Small Sample Size*

Because of the large costs associated with the data collection process, crash data are often characterized by a small number of observations. In addition, crash data for some roadway entities may have few observed crashes which results in a preponderance of zeros. Data characterized by small sample size and low sample-mean can cause estimation problems in traditional count-frequency models. For example, with small sample sizes, the desirable large-sample properties of some parameter-estimation techniques (for example, maximum likelihood estimation) are not realized. With low sample means (and a preponderance of zeros), the distribution

of crash counts will be skewed excessively toward zero which can result in incorrectly estimated parameters and erroneous inferences.

### *Injury Severity and Crash Type Correlation*

Traffic-Collision data are often classified according to their injury severity or collision type. For example, classifying an accident by the most severely injured person could result being classified as fatal, incapacitating injury, non-incapacitating injury, possible injury and no injury. Classifying these accidents by collision type could include outcomes such as rear-end, single-vehicle run-off-the-road, right-angle, and sideswipe among others. The most common modelling approach is to consider the frequency or the count of all collisions (including all severity and collision types together), and deal with the injury severities or collision types separately once the total number of collisions is determined.

However, some researchers have sought to develop separate models for various injury severities and collision types. A potential statistical problem can be shown because there is a correlation among injury severities and collision types. For example, an increase in the number of collisions that are classified as incapacitating injury will also be associated with some change in the number of collisions that are classified by other injury types, which sets up a correlation among the various injury-outcome models. This necessitates the need for a more complex model structure to account for the cross-model correlation

## 5.2 Functional Form

The functional form of the model determines the relationship between the dependent variable and the explanatory variables and is considered as a critical part of the modelling process. Many of the count-data models assume that predictor variables influence the dependent variables in a linear fashion. However, there is a body of work that suggests that non-linear functions better characterize the relationships between the rate of collisions and the explanatory variables that are considered. These non-linear functions can often be quite complex and may require involved estimation procedures (Bonneson and Pratt, 2008).

## 5.3 Modelling Methods for Analyzing Collision-Frequency Data

The data and methodological issues related to the count data gives us a wide variety of methods have been applied over the years (many of which could compromise the statistical validity of an analysis if not properly addressed). For analyzing the collision data the following modelling methods have been considered in this study.

### *Poisson Regression Model*

As the data related to collisions has non-negative integers, the application of standard ordinary least-squares regression (which assumes a continuous dependent variable) is not appropriate. Given that the dependent variable is a non-negative integer, most of the recent thinking in the field has used the Poisson regression model

as a starting point. In a Poisson regression model, for example the probability of roadway entity (segment, intersection, etc.) i having yi number of collisions per some time period (where yi is a non-negative integer) is given by:

$$P(y_i) = \frac{EXP(-\lambda_i)\lambda_i^{y_i}}{y_i!}$$

where *P(yi )* is the probability of roadway entity *i* having *yi* number of collisions per time period and *i* is the Poisson parameter for roadway entity *i*, which is equal to roadway entity i's expected number of collisions per year, *E[yi ].* Poisson regression models are estimated by specifying the Poisson parameter $\lambda_i$ (the expected number of collisions per time period) as a function of explanatory variables. The most commonly used functional form for Poisson is:

$$\lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i),$$

where *Xi* is a vector of explanatory variables and *B* is a vector of estimable parameters. The Poisson model is the starting point for collision-frequency analysis for many years, researchers have often found that accident data exhibit characteristics that make the application of the simple Poisson regression (as well as some extensions of the Poisson model) problematic. Specifically, Poisson models cannot handle over-dispersion and under-dispersion and they can be adversely affected by low sample means and can produce biased results in small samples.

### *Negative Binomial Regression Model*

The Negative Binomial Regression model is an extension of the simple Poisson model to remove the possibility of over-dispersion in the collision data. The Negative Binomial/Poisson gamma model assumes that the Poisson parameter follows a gamma probability distribution. This model results in a closed-form equation and the mathematics to manipulate the relationship between the mean and the variance structures is relatively simple. Negative Binomial Regression model is derived by rewriting the Poisson parameter for each observation i as

$$\lambda_i = EXP(\boldsymbol{\beta}\mathbf{X}_i + \varepsilon_i)$$

where $EXP(\varepsilon_i)$ is a gamma-distributed error term with mean 1 and variance $\alpha$. The addition of this term allows the variance to differ from the mean as

$$VAR[y_i] = E[y_i][1 + \alpha E[y_i]] = E[y_i] + \alpha E[y_i]^2.$$

The Poisson regression model is a limiting model of the negative binomial regression model as $\alpha$ approaches zero, which means that the selection between these two models is dependent upon the value of $\alpha$. The parameter $\alpha$ is often referred to as the over-dispersion parameter.

The Poisson-gamma/Negative Binomial Regression model is probably the most frequently used model in crash-frequency modelling. However, the model does have its limitations, most notably its inability to handle under-dispersed data, and dispersion-parameter estimation problems when the data are characterized by the low sample mean values and small sample sizes (Lord et al., 2009).

## 5.4 Regression Analysis

As discussed, the aggregate or collective approach to collision data analysis is characterized by frequency modelling. The frequency modelling enumerates the relationship between observed count of collisions and existing geometric, roadway, and traffic conditions on a given stretch of a roadway. Studies have used varied sets of variables depending upon the scope of research. Focus of these frequency models is generally two-fold on: (a) modelling methodology and (b) the parameters used as dependent and independent variables. In other words, the past research in this area has been diverse methodologically as well as empirically.

### *Poisson versus Negative Binomial*

The traffic-collision frequency data have been analyzed using a number of statistical methodologies. Initially multiple linear regression was used for model formulation. However as per our discussion stated above, linear regression models do not describe the nature of the frequency data adequately. Poisson or Negative Binomial (NB) regression models are better suited for defining the random, discrete, and nonnegative nature of collision occurrence. The best known example of Poisson regression model is log-linear model. For Poisson-distributed data, essentially a generalized linear model (GLM) is suggested and specifies how the size of a cell frequency depends on the levels of categorical variables for that cell. The nature of this specification relates to the association and interaction structure among the categorical variables.

A Poisson model formulation requires the mean and variance of the collision data to be equal (also called 'equi-dispersion'). Therefore, the Negative Binomial model, which has all the desirable statistical properties and also relates the above stated limitation, is the most popular model formulation for frequency predictive models. The data considered for a study can show an extra variation that is greater than the mean. This condition lead to an over-dispersion in data and Negative Binomial Regression model is more flexible in this situation than Poisson regression (we can still use Poisson regression for modelling but the standard errors will be biased in this case). The Negative Binomial distribution has one parameter more than the Poisson regression that adjusts the variance independently from the mean. And we can say the Poisson distribution is a special case of Negative Binomial distribution. Both regression models have been applied on the collision data of New York city and the regression functions are given below:

1. *Poisson Model regression results:*

```
Call:
glm(formula = FullReg$collisions ~ factor(FullReg$factorcode) +
    factor(FullReg$vehiclecode) + factor(FullReg$zipcode), family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.3817  -0.9795  -0.0766   1.0322  12.5308

Coefficients:
                             Estimate Std. Error  z value Pr(>|z|)
(Intercept)                  5.483863   0.032342  169.558  < 2e-16 ***
factor(FullReg$factorcode)2  2.608664   0.033193   78.590  < 2e-16 ***
factor(FullReg$factorcode)3  3.617024   0.032467  111.406  < 2e-16 ***
factor(FullReg$factorcode)4  2.914938   0.032894   88.616  < 2e-16 ***
factor(FullReg$factorcode)5  2.648476   0.033149   79.896  < 2e-16 ***
factor(FullReg$factorcode)6  3.423141   0.032557  105.143  < 2e-16 ***
factor(FullReg$factorcode)7  1.165635   0.036679   31.779  < 2e-16 ***
factor(FullReg$factorcode)8  1.437528   0.035622   40.355  < 2e-16 ***
factor(FullReg$factorcode)9  1.684675   0.034870   48.314  < 2e-16 ***
factor(FullReg$factorcode)10 1.187146   0.036579   32.454  < 2e-16 ***
factor(FullReg$factorcode)11 0.068471   0.044569    1.536    0.124
factor(FullReg$factorcode)12 0.263847   0.042592    6.195 5.84e-10 ***
factor(FullReg$factorcode)13 -2.060999  0.098223  -20.983  < 2e-16 ***
factor(FullReg$vehiclecode)2 -1.932339  0.008595 -224.831  < 2e-16 ***
factor(FullReg$vehiclecode)3 -2.546331  0.011352 -224.303  < 2e-16 ***
factor(FullReg$zipcode)2     -3.616991  0.031270 -115.670  < 2e-16 ***
factor(FullReg$zipcode)3     -2.008931  0.014668 -136.956  < 2e-16 ***
factor(FullReg$zipcode)4     -1.052370  0.009910 -106.197  < 2e-16 ***
factor(FullReg$zipcode)10    -5.679050  0.089946  -63.139  < 2e-16 ***
factor(FullReg$zipcode)11    -2.163508  0.015711 -137.704  < 2e-16 ***
factor(FullReg$zipcode)12    -0.066180  0.007249   -9.129  < 2e-16 ***
factor(FullReg$zipcode)13    -0.852938  0.009221  -92.497  < 2e-16 ***
factor(FullReg$zipcode)14    -1.174887  0.010377 -113.225  < 2e-16 ***
factor(FullReg$zipcode)16    -3.496136  0.029569 -118.237  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 392216  on 342  degrees of freedom
Residual deviance:   1297  on 319  degrees of freedom
AIC: 3367.5

Number of Fisher Scoring iterations: 4
```

**Figure 7: Poisson Regression Model - Collision Data**

*Obsevations*:

The residual deviance in Poission model is 1297, there is an evidence of over-dispersion. As a result of overdispersion standard errors are underestimated. As a remedy negative binomial model is used

## *2. Negative Binomial Model regression results*

```
Call:
glm.nb(formula = FullReg$collisions ~ factor(FullReg$factorcode) +
    factor(FullReg$vehiclecode) + factor(FullReg$zipcode), init.theta = 58.18479487,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7704  -0.6043  -0.0224   0.6899   2.9841

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    5.55836    0.05518 100.740   <2e-16 ***
factor(FullReg$factorcode)2    2.45550    0.05931  41.400   <2e-16 ***
factor(FullReg$factorcode)3    3.48880    0.05799  60.163   <2e-16 ***
factor(FullReg$factorcode)4    2.79291    0.05881  47.492   <2e-16 ***
factor(FullReg$factorcode)5    2.66519    0.05896  45.201   <2e-16 ***
factor(FullReg$factorcode)6    3.29096    0.05817  56.570   <2e-16 ***
factor(FullReg$factorcode)7    1.02041    0.06383  15.986   <2e-16 ***
factor(FullReg$factorcode)8    1.37983    0.06219  22.186   <2e-16 ***
factor(FullReg$factorcode)9    1.63549    0.06131  26.677   <2e-16 ***
factor(FullReg$factorcode)10   1.12469    0.06326  17.779   <2e-16 ***
factor(FullReg$factorcode)11  -0.02514    0.07107  -0.354   0.7236
factor(FullReg$factorcode)12   0.13914    0.06953   2.001   0.0454 *
factor(FullReg$factorcode)13  -2.12409    0.11645 -18.241   <2e-16 ***
factor(FullReg$vehiclecode)2  -1.90630    0.02414 -78.959   <2e-16 ***
factor(FullReg$vehiclecode)3  -2.48944    0.02631 -94.634   <2e-16 ***
factor(FullReg$zipcode)2      -3.57229    0.05415 -65.966   <2e-16 ***
factor(FullReg$zipcode)3      -1.98531    0.04136 -47.997   <2e-16 ***
factor(FullReg$zipcode)4      -1.05296    0.03759 -28.010   <2e-16 ***
factor(FullReg$zipcode)10     -5.63386    0.10421 -54.062   <2e-16 ***
factor(FullReg$zipcode)11     -2.14036    0.04216 -50.766   <2e-16 ***
factor(FullReg$zipcode)12     -0.06063    0.03531  -1.717   0.0860 .
factor(FullReg$zipcode)13     -0.86041    0.03703 -23.234   <2e-16 ***
factor(FullReg$zipcode)14     -1.17723    0.03795 -31.020   <2e-16 ***
factor(FullReg$zipcode)16     -3.50917    0.05370 -65.343   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(58.1848) family taken to be 1)

    Null deviance: 50018.32  on 342  degrees of freedom
Residual deviance:   401.32  on 319  degrees of freedom
AIC: 2843.9

Number of Fisher Scoring iterations: 1


              Theta:  58.18
          Std. Err.:  8.46

 2 x log-likelihood:  -2793.942
```

**Figure 8: Negative Binomial Regression Model – Collision Data**

```
> anova(f)
Analysis of Deviance Table

Model: Negative Binomial(58.1848), link: log

Response: FullReg$collisions

Terms added sequentially (first to last)


                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                        342      50018
factor(FullReg$factorcode)  12    20163       330      29855 < 2.2e-16 ***
factor(FullReg$vehiclecode)  2    14332       328      15523 < 2.2e-16 ***
factor(FullReg$zipcode)      9    15122       319        401 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: Analysis of Variance Table

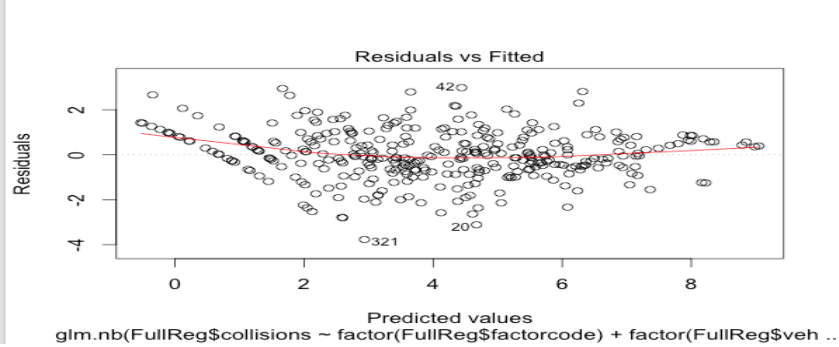


Figure 10: Residuals vs Fitted graph

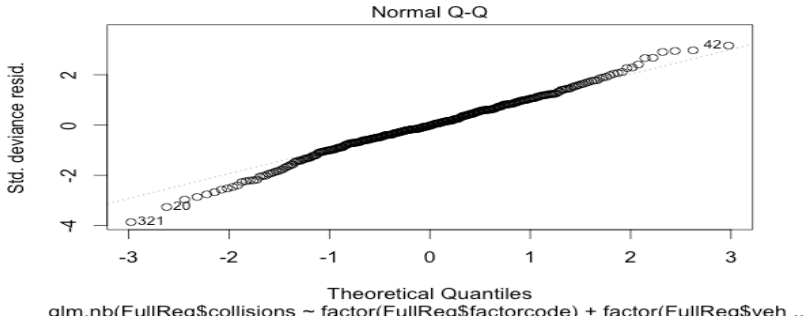


Figure 11: Normal Q-Q Plot

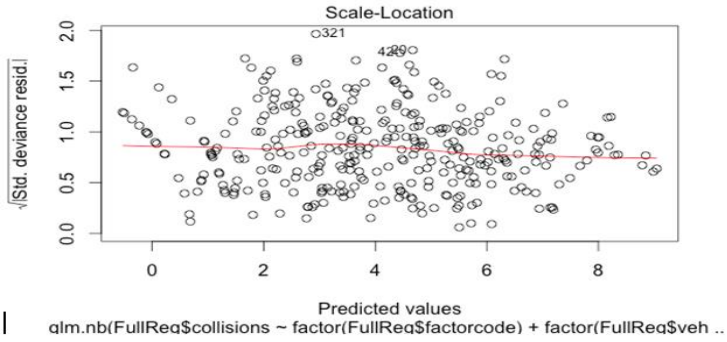


**Figure 12: Standard deviance residuals**

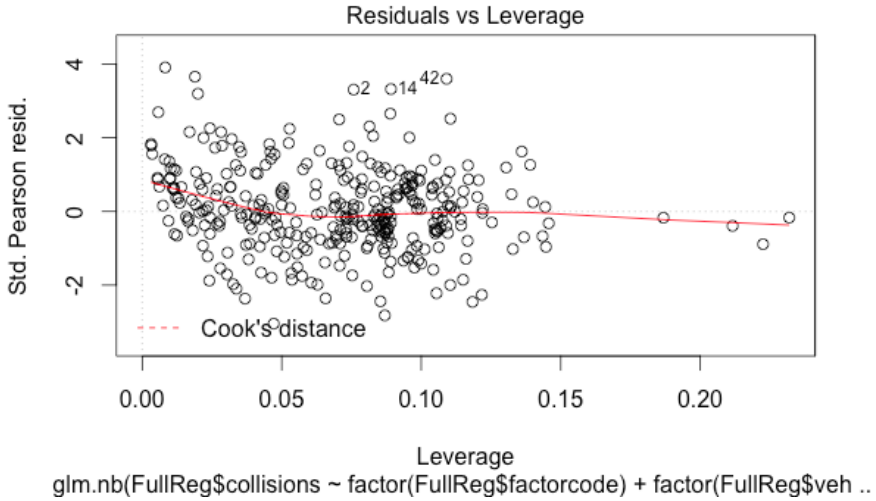


**Figure 13: Cook's Distance**

As we can see the output from Negative Binomial with Contributing Factors, Zip codes and Vehicle codes looks significant. Though P values for most of the Zip codes and Vehicle codes show that these predictor variables are significant as coefficient values are negative (maybe these can be neglected in the future models). Looking at the model we can state that only Contributing Factors have coefficients which look significant for this model to be effective.

The predictor variables are categorized into many categories, the details of the significant predictor variables are shown below:

| **Predictor Variable** | **Category** | **Description** |
|---|---|---|
| **Factor** | 1 | Failure to Yield Right-of-Way, Backing Unsafely, Traffic Control Disregarded, Turning Improperly,Failure to Keep Right,Aggressive Driving/Road Rage,Following Too Closely,Unsafe Speed,cell phone(hand held),Failure to Keep Right. |
| | 2 | Fatigued/Drowsy,Physical Disability,Prescription Medication,Lost Consciousness, Reaction to Other Uninvolved Vehicle,Illness,Fell Asleep. |
| | 3 | Alcohol Involvement, Driver Inattention/Distraction, Other Electronic Device,Passenger Distraction, Driver Inexperience, Drugs (Illegal) |
| | 4 | Other Vehicular, Outside Car Distraction, Cell Phone (hands-free), Oversized Vehicle, Animals Action, Outside Car Distraction, Pedestrian/Bicyclist/Other Pedestrian Error/Confusion |
| | 5 | Traffic Control Device Improper/Non-Working, Passing or Lane Usage Improper, Pavement Defective, View Obstructed/Limited, Lane Marking Improper/Inadequate, Obstruction/Debris, Pavement Slippery, Tow Hitch Defective, Shoulders Defective/Improper,Pavement Defective |
| | 6 | Brakes Defective, Steering Failure,Glare, Tire Failure/Inadequate, Tire Failure/Inadequate, Accelerator Defective, Windshield Inadequate, Headlights Defective |
| **Vehicle Code** | 1 | Sport Utility/station wagon, taxi, passenger vehicle,small commercial(4 tyres) |
| | 2 | Livery Vehicle, Bus,pick-up truck,fire truck,large commercial vehicle(6 or more tyres),ambulance |
| | 3 | other,unknown,motorcycle,bicycle,scooter |
| **Zipcode** | 1 | 10001,10002,10009,10010,10011,10012,10013,10016, 10017,10018,10019,10021,10022,10023,10025,10027, 10028,10029,10032,10034,10036,10038,10039,10065, 10075 |
| | 2 | 10128 |
| | 3 | 10304,10305,10306 |

| | | |
|---|---|---|
| | 5 | 10451, 10452, 10453, 10454, 10455, 10457, 10458, 10459,10462,10468 |
| | 11 | 11101,11102,11103,11105,11106 |
| | 12 | 11201,11203,11204,11206,11207,11208,11209,11210, 11211, 11212, 11213,11216,11218,11219,11220,11221, 11222, 11225, 11226, 11228, 11230, 11233,11234,11236, 11239, 11249 |
| | 13 | 11354, 11355, 11357, 11368, 11372, 11373, 11374, 11377, 11385 |
| | 14 | 11412,11413,11415,11417,11418,11419,11420,11421, 11423,11432,11435,11436 |

## *Observations:*

Lets take an example of predicting the number of collisions for Factor code2(illness), vehicke code2(commercial vehicles) and zip code1(10001-10100)

Using the negative binomial regression model we can predict the number of collisions.

**5.5 +2.455 -1.906 = 6.049**

**exp^6.049 approx. 424**

***Predicted number of accidents for this criterion is approximately 424.***

- The model is a good fit at significance level of 0.05. The model shows factor coefficients are positive indicating that increase in the factor (factor 1,2,3,4,5,6) leads to an increase in the number of road accidents.
- Vehicle code 2 and 3 are also significant statistically. Vehicle code 2 is mainly the category of heavy commercial vehicles. The negative coefficient sign indicates that commercial vehicles are less involved in the road accidents. Similar conclusions can be made for vehicle code 3.
- Zip codes are also significant statistically and have negative coefficient signs indicating as the area increases the numbers of collisions are likely to come down.
- Deviance is a measure of goodness of fit of a generalized linear model (GLM). Or rather, it's a measure of badness of fit–higher numbers indicate worse fit.
- R reports two forms of deviance – the null deviance and the residual deviance. The null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean).
- The null deviance is calculated from an intercept-only model with 342 degrees of freedom. The null deviance shows how well the response variable is predicted by a model that includes only the intercept (grand mean). The null deviance is large indicating the intercept-only model is not a good model.
- Then we see the residual deviance. The residual deviance is really small compared to null deviance, indicating the goodness of fit of this model.

- We are also shown the AIC and maximum likelihood (2*log likelihood). The likelihood can be used to compare the models.
- Log likelihood ratio test is used to compare the Poisson and Negative Binomial models. The residual deviance of Poisson model is 1297 on 319 degrees of freedom where as for Negative Binomial model residual deviance is 401.32 on 319 degrees of freedom. The difference noted is 895.68. The difference is huge; therefore using Negative Binomial is justifiable.

# 6 Conclusion

In conclusion, we can say that this model from negative binomial regression is moderately useful at this stage; it can be considered the beginning of a model that could detect the odds of a motor-vehicle collision to occur and which factors are majorly responsible for a collision. It may be beneficial to eventually look at the statistics from this model and perform deeper analysis in a later stage. The hope of this study is to become part of a larger study that may be able to be used to help further policies to create better improved regulations so that the overall collision rate is reduced considerably.

# 7 References


- National Center for Statistics and Analysis. (2014, December). 2013 motor vehicle crashes: Overview. (Traffic Safety Facts Research Note. Report No. DOT HS 812 101). Washington, DC: National Highway Traffic Safety Administration.
- NHTSA. (2010). Fatality Analysis Reporting System. Retrieved from National Highway Traffic Safety Administration: http://www.nhtsa.gov/
- AlcoholAlert. (2010). 200g Drunk Driving Statistics. Retrieved from AlcoholAlert! - Intervention at the Point of Consumption: http://www.alcoholalert.com/drunkdriving-statistics-2006.html
- Lum, H., & Reagan, J. A. (1995). Interactive Highway Design Model: Accident Predictive Module. Public Roads Magazine, 14-17.
- Bedard, M., Guyatt, G. H., Stones, M. J., & Hirdes, J. P. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities. Accident Analysis & Prevention, 717-727.
- NHTSA. (2005). Trend and Pattern Analysis of Highway Crash Fatalities by Month and Day. Washington DC: National Center for Statistics and Analysis.
- NYPD Motor Vehicle Collisions data. (2016). Retrieved from NYC Open Data: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95
- Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2013). Applied Linear Regression Models: Fifth Edition. New York: McGraw-Hill Irwin.
- NHTSA. (2013). Traffic safety facts for 2013 - A Compilation of Motor Vehicle Crash Data from the Fatality Analysis Reporting System and the General Estimates System. Retrieved from National Highway Traffic Safety Administration: http://www.nhtsa.gov/
- Park, E.-S., Lord, D., 2007. Multivariate Poisson-lognormal models for jointly modelling crash frequency by severity. Transportation Research Record 2019, 1-6.
- Lord, D., Geedipally, S.R., Guikema, S., 2009. Extension of the application of Conway Maxwell-Poisson models: analyzing traffic crash data exhibiting under-dispersion. Submitted to the 89th Annual Meeting of the Transportation Research Board, Washington, D.C.
- Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2010. Statistical and Econometric Methods for Transportation Data Analysis. Second Edition, Chapman Hall/CRC, Boca Raton, FL.
- Bonneson, J.A., Pratt, M.P., 2008. Procedure for developing accident modification factors from cross-sectional data. Transportation Research Record 2083, 40-48.