Roadway Crashes Prediction Model using: Different Analytical and Statistical Modeling Methods.

Alia Kasem [1] and Lina Kasem[1]

[1]Affiliation not available

February 28, 2022

Abstract:

This paper's main objective is to develop and validate a framework prediction based on the use of various classifiers that operate on merged attributes to obtain more effective crash predictions from classifiers' outcomes. In order to systematically classify the crash through feature selection techniques, a data-driven approach was adopted to investigate the best potential classifier to predict the contributing factors and the number of injuries/fatalities per collision. To produce diversity in learning techniques, by using the following methods Bayes Network, Naïve Bayes, J48, and K-Nearest Neighbor (KNN) for the year of 2019 and the 100 intersections with the high-risk for collisions generated from the NYPD Collision data using the 2014-2019 dataset. The findings indicate the classifier's accuracy was highest for the Naïve Bayes with 81.59%, followed by the Bayes Network with an accuracy of 81.57%, and J48 resulted in an accuracy of 80.81%, and KNN performed the lowest accuracy of 80.20%.

Overall, the methodology and outcomes offer new insights into crash detection and be a useful method to improve intervention efforts relevant to safer transportation planning. The geospatial analysis was performed, providing insight into areas with a frequency of different types of collisions that could establish a pattern based on reoccurring locations within the same area/zip code. The geospatial analysis indicates that boroughs of Brooklyn and Queens have the highest numbers of collisions. The research also analyzed the relationship between the number of collisions and the conditions of the pavement. The findings suggest that the condition of the pavement has an impact on the number of crashes. The total number of accidents was 55,827 and 3,622 intersecting with a poorly graded condition of pavement; 227 total fatalities and 127 intersecting with a poorly rated condition of the pavement.

Keywords:

Bayes Network, Naïve Bayes, J48, KNN and Geospatial Analysis, High-Risk Intersections, and Crash prediction.

Introduction

At over 27,000 persons per square mile, New York has the highest population density of any major city in the United States [27]. In developing safer roadways, planners and engineers must understand that the city is very diverse in transportation modes. The risk defines a probability of multiple levels of damages from a transportation planning procedure, such as injury, fatalities, liability, and property damages; all these elements

create roadway vulnerabilities. As part of the roadway safety plan, the probability of an outcome, and any potential severity of the outcome if it occurs. Collision prediction is a critical component of transportation planning; it plays a role in the decision-making process and funding allocation to certain areas. (FHWA Manual, 2010)

About 1.4 million households own a car compared to 3.1 total households based on the census tract. According to New York City Economic Development Corporation's (NYCEDC) study from 2018, the information is derived from Metro System Ridership and 2011-2015 ACS Microdata; the following are the average percentage of the vehicle ownership borough base: Manhattan, 22 percent, Brooklyn 44 percent, Queens 62 percent, Bronx 40 percent, and Staten Island 83 percent. With diverse vehicle types and roadway characteristics, 27 percent of commuters' travel via truck, car, or van. (NYCEDC, 2018)

As a safety plan, NYC has adopted Vision Zero to achieve zero fatality. Based on the 2014 Vision Zero Year 4 Report, over 250 people were killed, and 4,000 were seriously injured in traffic crashes. Compared to the June 2020 Vision Zero report, there was a decrease in the number of fatalities to a total of 220 [4].

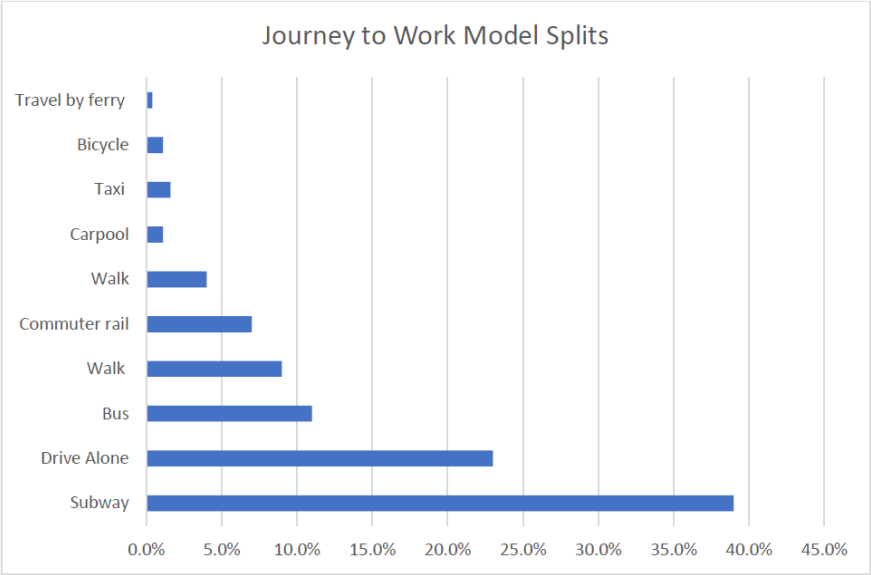


Figure1. Journey to Work Model Splits

According to America Census Tract data [5], the motorist is the second-highest mode used to commute to work, as shown in the figure above.

since the adaption of Vision zero, NYC has witnessed several implementation measures

to develop safer intersections such as:

* Increasing the number of bike mileage.

* Increasing the Leading Pedestrian Intervals (LPI).

* Impalement a total of 158 Street Improvement Projects (SIP).

* Installation of the speed camera within the school zone.

* An increase in the number of violations, with a total of 81,609 summonses were issued to drivers for failure to yield.[4]

This study reviewed the highest number of contributing factors to crashes based on location. The purpose of the research study is to identify high-risk locations identified by the contributing factors and the number of collisions resulting in a higher risk of potential collisions. In order to classify high-risk crash locations, this analysis analyzed the highest number of recurring crashes, identified by the contributing risk factors and the number of collisions that result in a location having a higher risk of potential collisions. The analysis uses various statistical models to generate risk predictions by integrating multiple types of data. The data used historical collisions to determine high-risk locations; NYC-Open-Data crash data was used as the primary source for collision data.

Systemic Safety Approach

As outlined in the Federal Highway Administration's (FHWA) Systemic Safety Project Selection Tool are the following:

- Identifies safety concerns based on an evaluation of data at the system level.
- Establishes common characteristics (risk factors) of locations where severe crashes frequently occur.
- Given high aggregate numbers of target crashes but low average density per site tends to deploy lower-cost countermeasures to many sites to affect a large number of locations. (Preston et al., 2013)

The above principles can be described as the methodology to identify high-risk locations for crashes based on a data-driven approach (Preston et al., 2013). These steps assist in finding the type of crashes and the risk factor associated with the collision. The functionality of this model might have the capability to provide some mitigation measures.

Literature Review

Collision Predictors have been tested in several analyses using similar methods and data frames. The subsequent research used the analogous regression approach to obtain a model and or used the spatial method.

A. Iranitalab, A. Khattak Prediction models for crash severity, allow various agencies to predict the severity of a reported crash with uncertain severity or the severity of crashes that could occur sometime in the future. This paper focuses on the following: comparison of the performance in forecasting traffic crash occurrence of four statistical and machine learning approaches, including Multinomial Logit (MNL), Nearest Neighbor

Classification (NNC), Support Vector Machines (SVM), and Random Forests (RF); developing a crash cost-based approach to the comparison of crash severity prediction methods; and exploring the impact on the performance of crash severity prediction models of data clustering methods containing K-means Clustering (KC) and Latent Class Clustering (LCC). Crash data from Nebraska, United States was collected from the 2012-2015 records, and two-vehicle crashes were extracted as the study data.

The dataset was divided into subsets for training/estimation (2012–2014) and validation (2015). The accurate predictor and proposed approach indicated that NNC had the highest predictive results in both the total and the more severe crashes. RF and SVM had the next two sufficient results, and the worst approach was MNL. Data clustering did not impact SVM's prediction output, but KC improved MNL, NNC, and RF's prediction performance, while LCC improved MNL and RF but decreased NNC's performance. Overall, the accurate prediction rate had almost the exact opposite effects relative to the proposed solution, indicating that neglecting the crash costs could lead to a misjudgment in the choice of the correct prediction process.

C. Zhang et al. (2014). Focuses on zonal crash prediction and safety assessment are essential in transportation, safety planning, and safety diagnostics. Geographic Information System (GIS) based framework for data integration and safety assessment was introduced. The research established a crash prediction model using a Negative binomial method. The proposed data showed: the crash frequency is correlated with road and traffic characteristics, such as the average free flow rate, the average daily traffic within the zone and the total length of the road within the zone, and social-economic and demographic characteristics, such as the overall population, the percentage of households with high incomes. Three datasets were used for this zonal prediction model: The Traffic Analysis Zones (TAZ) dataset, the traffic and roadway dataset, and the accident datasets. According to the negative binomial model estimation,

It is possible to estimate the total number of crashes in a TAZ by its Mean expression.

J. Lee et al. Crash simulations have played a key role in detecting crash hotspots and evaluating safety response initiatives. A variety of macro-level crash prediction models have recently been developed to integrate highway safety considerations into the long-term transport planning process. In this study, the authors defined a series of intersection crash models with macro-level data for seven spatial units for total, severe, pedestrian, and bicycle crashes. The study found that the bicycle crash models performed better in zip-code tabulation area data, and the census-based data-based pedestrian crash models outperform opposing models. It was also identified that intersection crash models could be significantly enhanced for macro-level entities by including random effects. For each crash type, three types of models were developed as follows: Model type (1): crash prediction models with only micro-level variables; Model type (2): micro-level variables and macro-level random-effects crash prediction models; and Model Type (3): crash prediction models of micro-level variables and random-effect macro-level variables. The results showed that Macro-level random-effects only intersection crash forecast models and those with macro-level random effects and variables outperform those with only intersection-level variables. By adding macro-level random effects, the intersection crash prediction models can be improved. With ZIP Code Tabulation Areas (ZCTA) based data, the intersection crash prediction models for total, severe, and bicycle crash models have the highest performance. With Census Tract based data, the intersection crash prediction model for pedestrian crashes performs the best. Finally, pedestrian and bicycle crash simulation findings suggest that multiple macro-level variables may be a robust surrogate exposure for such crashes.

H. Huang et al. (2016) Generally, the level of zonal protection is measured by comparing separate macroscopic variables to aggregated crash statistics on a specific spatial scale. The level of zonal protection was measured by comparing separate macroscopic variables to aggregated crash statistics on a particular spatial scale. The micro-level perspective, in which zonal crashes are determined by summing up all road entities' predicted crashes within the zones of interest. The purpose of this study was to compare these two forms of models for zonal crash prediction. The Bayesian macro-level spatial model with a conditional auto-regressive prior and the Bayesian micro-level spatial joint model was developed and empirically tested. The Optimized Hot Zone Detection method suggested using the benefits of separate macro and micro screening results. The study focused on a three-year data collection for the urban road network. Results have demonstrated that

the micro-level model has better overall performance and predictive efficiency the shows an insight into the micro-factors that directly correlate to the severity of an injury and contributes to more direct counter steps. Whereas the macro-level crash review has the benefit of requiring fewer comprehensive details, offering additional input on non-traffic infrastructure problems, and being an essential method for integrating safety concerns into long-term transportation planning.

M. Abdel-Aty et al. (2013). research with the Geographic Information System (GIS) can analyze crashes with different geographical units. Macro-level safety research is at the center of Transport Safety Planning (TSP) and is crucial in many aspects of safety investment strategy and decision-making. The choice of the spatial unit may vary depending on the variable of the model. In this study, three different crash models for Traffic Analysis Zones (TAZs), Block Groups (BGs), and Census Tracts (CTs) were investigated. Models were developed for total crashes, severe crashes, and pedestrian crashes in this region. The research's primary objective was to analyze and examine the impact of zone heterogeneity (scale and zoning) on these particular types of crash models. These models have been developed based on different road characteristics and census variables. The importance of the explanatory variables was found to be inconsistent between models focused on different zoning systems. Although the variation in variable significance across geographic units has been established, the findings also indicate that the coefficients are rational and explainable across all models.

This analysis's significant results are the coefficients are compatible if these variables are essential in models with the same response variables, even if the geographic units are different. The number of significant variables is influenced by response variables and also by geographic units.

J. Lee et al. (2014) Many crash modeling experiments have concentrated only on the areas where the crash occurred. This study focused on the residence features correlated with the origin of the drivers triggering a traffic accident, the so-called at-fault drivers. Intuitively, it is rational to conclude that the number of at-fault drivers is related to the at-fault driver's residence location's socio-demographic characteristics. Therefore, the key purpose of this analysis is to establish the relationship between the number of at-fault drivers and the zone characteristics of the residence from which the at-fault drivers originated. The Bayesian Poisson-lognormal model was introduced to assess the residential zones' contributing factors to the number of accidents based on faulty drivers.

This research indicates that the crash's occurrence is impacted by road/traffic causes and by many demographic and socio-economic factors of the residential zones. In order to evaluate the relationship between the zonal characteristics and the number of at-fault drivers, the Bayesian Poisson lognormal model was used in this analysis, and the result showed that the exposure measures as 'Log of population,' and 'Proportion of commuters using non-motorized modes' of residence zones were positively associated with the number of at-fault drivers.

On the other hand, 'Proportion of elderly people,' 'Proportion of people working at home,' 'Proportion of commuters whose travel time is less than 15 min' and 'Median Family income' in residential areas have had a negative association with the number of at-fault drivers.

S. Alkheder et al. In this research Analysis, WEKA (Waikato Climate for Knowledge) data-mining software was used to build the artificial neural network (ANN) classifier. The traffic accident data was used in two separate ways to create two classifiers. For testing and validating the first classifier (training set), 90% of the data was used to train the second classifier, and the remaining 10% was used to assess it (testing set). The experimental results showed that the established ANN classifiers could predict accident severity with moderate accuracy.

The average performance of the training and testing results for model estimation was 81.6 percent and 74.6 %, respectively. Traffic accident data were grouped into three clusters using the k-means algorithm to increase the ANN classifier's estimation accuracy. The results after clustering demonstrated a substantial change in the ANN classifier's prediction accuracy, especially for the training data collection.

In this work, to verify the efficiency of the ANN model, the ordered probit model was also used as a

benchmark. The dependent variable (i.e., degree of injury) has been modified from ordinal to numerical (1, 2, 3, 4) for (minor, moderate, severe, death). The R tool has been used to execute the ordered probit. The ordered probit model revealed how likely this accident would occur in each type of accident (minor, moderate, severe, death). The precision of 59.5 percent was achieved from the probit model, less than the ANN's precision of 74.6 percent.

Z. Elassad et al. The objective of this research is to develop and validate an ensemble fusion structure based on various base classifiers running on fused features and a Meta classifier that learns from the base classifier results to achieve more effective crash predictions. A data-driven methodology was introduced to examine the ability to the convergence of four real-time and continuous categories of features, namely physiological signals, driver maneuvering inputs, vehicle kinematics, and weather covariates, to systematically classify the most robust precursors of the accident using feature selection techniques. Also, a resampling-based scheme, including Bagging and Boosting, is being applied to produce variety in learner combinations comprising Bayesian Learners (BL), k-Nearest Neighbors (kNN), Vector Machine Support (SVM), and Multilayer Perceptron (MLP). In order to ensure that the proposed system provides efficient and stable choices, an imbalance-learning approach was introduced using the Synthetic Minority Oversampling Technique (SMOTE) to solve the issue of class imbalance as a crash case. The results reveal that Boosting has shown the best efficiency within the merger scheme and can reach a maximum of 93.66 percent F1 score and 94.81 percent G-mean with Naïve Bayes, Bayesian Networks, K-NN, and SVM with MLP as a Meta-Classifier. Overall, the methods and results offer new insights into accident detection and can be harnessed as a promising technique to improve intervention efforts relevant to intelligent transport traffic systems.

C. Chen et al. This Research examines driver injury severities in rear-end crashes in New Mexico. Rear-end collisions are a major type of traffic collision in the U.S. A thorough analysis of the function that results in casualties and fatalities is of practical necessity. Decision Table (DT) and Naïve Bayes (NB) approaches have also been commonly used but independently to address classification problems in several fields, except for road safety studies. The study used a two-year rear-end crash dataset; this paper uses a decision table/Naïve Bayes (DTNB) hybrid classifier to determine the probabilistic attributes and predict driver accident results for rear-end crashes. Test results show that the hybrid classifier performs well, as shown by various efficiency metrics, such as precision, F-measurement, receiver operating characteristic (ROC), and area under the ROC curve (AUC). Fifteen significant attributes are significant in predicting driver injury severity, including weather, lighting conditions, road geometry, and driver behavior. The extracted decision rules indicate that heavy vehicle presence, a comfortable traffic environment, low lighting conditions, two-lane rural roads, damaged vehicle injury, and two-vehicle collisions would increase the likelihood of fatal injuries sustained by drivers.

D. Khera et al. Road traffic is an essential part of life, but repetitive accidents cause serious physical harm and property damage. Road

Traffic Accidents (RTAs) were identified as a major public health issue. The analysis was intended to examine various taxonomies methodologies using the Road Accident Tool WEKA and TANAGRA. The performance was determined by the Naive Bayes, ID3, and Random Tree algorithms. Comparison of data mining algorithm results based on error rate, computational time, precision value, and accuracy. The concept comparison using the WEKA experimenter showed that Naive Bayes surpasses Random Tree and ID3 algorithm with a 50.7 percent accuracy compared to 45.07 and 25.35 percent accuracy using the TANAGRA model. using TANAGRA Random tree outperforms Naive Bayes and ID3 algorithms with an accuracy of 92.95%, 67.6%, and 57.74% respectively.

ACI, and ÖZDEN et al. The analysis of this paper focused on predicting the severity of the motor vehicle accident injuries in Adana, Turkey, using different statistical models to predict the severity of the accidents. The methods were used in this study are K-Nearest Neighbor (KNN), Naive Bayes, Multilayer Perceptron, Decision Tree (DT), Support Vector Machine (SVM). The analysis uses the Regional Traffic Division's traffic accident reports and the Regional Directorate of Meteorology's weather data during 2005-2015; the collision data was recorded in severity as fatal or not fatal. This study's main goal was to determine if the weather

condition impacts the collision's severity and any factors for traffic accidents.

The results indicate that DTC and KNN algorithms yielded slightly more accurate results in classifying fatal instances in both datasets. The accuracy for the KNN was 90.3%, followed by the DTC, which was 90.2%, SVM was recorded at 88.4, MLP was recorded at 87.2%, and the LR was the lowest 82.4%. The analysis of the prediction model's importance measured the Area Under Curve based on the input ranking. The temperature variables were higher based on the methods, identifying that the Maximum Temperature and weather parameters negatively affected all models' classification performance.

Theofilatos, Chen, and Antoniou et al. The main purpose of this study was to investigate the influence of real-time traffic and weather parameters on the frequency of crashes on freeways; no studies are comparing the prediction efficiency of machine learning (ML) and deep learning (DL) models to the best of the results knowledge. The present study contributes to existing information by comparing and validating ML and DL approaches to forecast real-time crash events. Real-time traffic and weather data from Greece's Attica Tollway were connected to historical crash data. The comprehensive data set was divided into subsets of training/estimation (75%) and validation (25%), then structured. Initially, using the training data collection, the ML and DL prediction models were trained/estimated. The models were subsequently compared on the test set based on their performance parameters (accuracy, sensitivity, precision, area under the curve, or AUC). K-nearest neighbor, Naive Bayes, decision tree, random forest, support vector machine, external neural network, and, ultimately, deep neural network were the models considered. Overall, though it outperformed all other candidate variants, the DL model appears to be more fitting. More precisely, relative to other models, the DL model managed to achieve a balanced efficiency among all parameters (total precision = 68.95 percent, sensitivity = 0.521, accuracy = 0.77, AUC = 0.641). However, it is unexpected that, although being much less complex than other models, the Naive Bayes model achieved good efficiency. The analysis results are beneficial since they offer an initial insight into ML and DL models' success.

Hossain and Muromachi et al. Due to recent advances in information systems and traffic sensor technology, the idea of measuring the crash risk for a concise time window in the near future is gaining more practicality. Although some real-time crash prediction models have been suggested, they are still primitive and require substantial real-life improvements. This manuscript investigates the current frameworks' main limitations and proposes alternatives with an updated structure and simulation system to solve them. It uses a random multinomial log model to define the most relevant predictors and the most appropriate detector positions to acquire data to construct such a model. The model was developed using high-resolution detector data obtained from Shibuya 3 and Shin-juku 4 expressways under the authority of Tokyo Metropolitan Expressway Corporation Limited, Japan. The model was then added to the Bayesian belief net (BBN) to create the real-time crash prediction model. It was explicitly designed for the freeway's simple stretches and estimates the probability of establishing a dangerous traffic situation for a particular 250-meter-long road portion within the next 4-9 minutes. The performance appraisal findings reveal that the model can accurately classify 66 percent of possible accidents with a false alarm rate of less than 20 percent at an average threshold value. The findings reveal that the model will correctly identify 66 percent of the accidents with a false alarm rate of less than 20 percent at a threshold value of 4.56 percent. If the threshold value is increased to 7%, the model will still forecast 58% collisions and 87% normal traffic conditions with 82% average classification precision. Moreover, in the event of a threshold value as high as 14%, with just less than 3% false warning, the model classifies 30 percent of the crash cases.

Alajali, Zhou, and Wang The study aims to introduce an accurate intersection traffic prediction by including additional data sources other than road traffic volume data into the prediction model using various decision trees. The analysis also benefits two sets of data collected from the reports of road accidents and roadworks happening near the intersections. All of the data are driven from the Victorian Government Data Directory maintained by the State of Victoria in Australia. The first dataset is the intersection traffic volume. Additionally, this study uses the sensors installed at the traffic signal, reflecting the intersections' real-time traffic volume. They are focusing on the Central Business District (CBD). The second data source was the accident data set, which consists of several attributes, such as accident ID, location coordinates,

road number, date, time, and an accident type. The data set also includes the condition of the road surface. As noted earlier, this study aims at the prediction of traffic, particularly at intersections. Addressing several traffic- issues associated with traffic predictions. Pointing out traffic forecasting is a complex and nonlinear problem. One of the significant problems this study will address is the Recurrent patterns vs. Non-recurring events in traffic predictions. Therefore, the implementation of -Spatiotemporal real-time information related to non-recurring events can help predict accurate traffic. A distinctive characteristic of traffic flow is the existence over time of repeated trends, which are useful for traffic prediction models. These events require a nonlinear model, along with the patterns of time and space variations. Another element that this study points out is scalability, which is a crucial traffic prediction requirement since it relies on a large amount of historical data and real-time data sources. The study provides a variety of factors in terms of related work. The traffic modeling for this analysis only evaluates the short-term traffic prediction can be divided into two major segments, and they are Parametric methods and Nonparametric methods. The parametric methods: it is a flexible family of models and estimates the model parameters based on the training data; the paper's abstract gives the topic a solid, concise sense by explicitly outlining the approaches used to solve the proposed problems. However, to make the introduction more concrete, the authors may wish to make multiple references to support the argument made in identifying the methods of the decision trees and how they are going to be deployed to address the presented problems in regards to intersection traffic predictions. The authors might want to include another statement with descriptions of some of this technology's implementations, along with relevant sources. In the related work, the study identifies the methodologies in which will be used to address the intersection traffic predictions. As was pointed out earlier, the study depends on two modeling methods. They are parametric methods and Nonparametric methods. Both of these models provide numerous benefits, and they are as follows. The parametric methods' benefits are as follows: It is a flexible family of models and estimates the model parameters based on the training data. Auto-Regressive Integrated Moving Average (ARIMA), introducing a space-time ARIMA model for urban traffic forecasting. This introduces the proposal of Support Vector Regression (SVR), used for traffic speed prediction. SVR is one of the methods that achieve good accurate results. On the other hand, the study points out the benefits of Nonparametric methods, and they are the following: It is widely adopted for predicting traffic flow, as they are considered to be more suitable than other methods for capturing the nonlinear and stochastic nature of traffic flow. Also, this method uses the k-Nearest Neighbors (KNN) for one stop-ahead traffic prediction. With supporting, methods of Radial Basis Function (RBF) and Artificial Bee Colony (ABC), considered the nonlinear correlation between spatial and temporal features. Also, with a proposed Bayesian Multivariate adaptive Regression (MAR) method for accurate and interpretable traffic prediction. Fast Incremental Model Trees with Drift Detection (FIMT-DD) algorithm. They highlight the use of the FIMT-DD traffic analysis and visualization process. This work expands the FIMT-DD approach by incorporating incidents and roadwork data with regular traffic volume data to forecast the correct traffic at intersections. This paper's objective gives a clear understanding of an approach. However, related work does not identify the purpose of using both methods. Which is assumes that the use of the parametric methods to calculate the errors in the dataset. Also, it is assuming that the nonparametric method was used to calculate the confidence mean. The research equipment is quite standard and appropriate for the research, particularly because the main focus of the paper intersection traffic predictions, possibly, in the related work should have related the use of the decision trees. One important aspect of the research paper underlines some limitations of previous studies related to traffic prediction that considers an online learning approach that focuses on typical and atypical traffic conditions.

The outlined benefits of using Decision Trees are as follows: the decision Tree will allow decision-makers to manage priority improvements regarding transportation safety measures. Ensemble decision trees were developed to increase decision tree models' performance by combining multiple weak predictors to obtain more accurate predictions.

This research focused on if accidents and roadworks can dramatically influence traffic patterns at intersections; a novel approach to intersection traffic prediction has been suggested, which involves integrating several data sources for model training. Three ensemble decision tree algorithms (GBRT, RF, and XGBoost) have

been adopted to train the prediction and model in a batch learning form. It has also introduced an interactive learning system in which the FIMT-DD algorithm was introduced to update the real-time model. The authors may also want to provide a more comprehensive discussion about each decision tree's result and how each would benefit through traffic prediction at the intersections.

Data Used

The data frames used for the study were as follow the five years (2014-2019) of Collision data derived from the New York City Police Department. Extracting High-Risk Locations for Collision Citywide (100 Intersections) and extracting 2019 to generate a prediction model. As a supporting dataset, the population dataset derived from the American Community Survey, 5-year estimate. Next, using the PUMA tabular data extracting the total Population per zip code and using the Public Use Microdata Area (PUMA) shapefile (Cartographic Boundary Shapefiles) New York to ArcMap. Using the Neighborhood Tabulation Areas (NTAs) using the Borough boundary shapefile from NYC OpenData to clip to shoreline ensures that the analysis focused on the City Right of Way. Also, this analysis uses MapPLUTO data from BYTES of the Big Apple for the land use identifications.

Safety Performance Functions (SPFs) explain the statistical associations between crash predictions and the importance of identifying the crash factors. It is important to consider from the literature the causes that increase the level of risk for these types of collisions in order to build a risk-based predicting model. This section began by examining the modeling strategies that have been used to explain roadway collisions.[6] The study then examined the literature on risk factors for roadway collisions, including accident characteristics of roadway design (Pavement condition) associated with increased risk.

Methodology

Python

The datasets were arranged by merging three different Land use tables, Population per zip code, Vision Zero Priority Intersections, and Collision data. For land-use, the main focus was to review the demographic data based on the crashes' location. The analytical tool to organize and simplify the data will be Python (package: Pandas), to understand the need for vision zero to include street geometry/ location as a collision factor. Using Python Data cleaning with Pandas and NumPy using the following functions:

- Dropping columns
- Changing the index of the data frame
- Arranging and organizing fields in the data
- Combining strings with NumPy clean columns
- Renaming columns and skipping rows with empty cells
- Calculating the number of collisions per zip code.
- Graphing and plotting

Python-Pandas was used to identify the locations with the highest number of collisions.

ArcGIS

Using ArcGIS, joined the two spatial data, they are the Map PLUTO data from BYTES of the Big Apple for land use identifications and Borough boundary shapefile from NYC Open Data. These two datasets are mainly used to visualize data with the NYC Motor Vehicle Collisions as the primary data analysis to the spatial features. This study used GIS to identify collision locations based on the intersection of the highest number of collisions, in which it was considered high-risk locations. The data sets were NYPD collision, as a CSV file. The ArcGIS Process began by uploading the data files. For the collision data, the display X, Y data option was used to form the longitude and latitude and extract the CVS file into a shapefile. During this step, a changed to the Projected Coordinate System from NAD_1983_StatePlane_New_York_Long_Island_-FIPS_3104_Feet to WGS84 (DD). Followed by using the Arc-Toolbox, using the Overlay tool, spatial join to the Collision Data, and Borough boundary shapefile. The Targeted Feature was the collision data and the join. Features were the Borough boundary shapefile, joining option- used JOIN_ONE_TO_ONE; Match option used to the 'With-In,' Following the proximity tool using the Intersect the collision with a Zip Code to visualize the relationship between the collision location and the total number of collisions per zip code.

WEKA

Using 2019 Crash Data, with the Binary number of Person Collisions, the following method was used to forecast the best algorithm to predict the number of injuries and fatalities in collisions:

For the Bayes Network: choose the Bayes Classifier, Bayes-Net, with the default K2 search algorithm, set init-As-Naïve-Bayes to TRUE, random-Order to TRUE max-Nr-Of-Parents to 2. run with 10-fold Cross-Validation.

For the Naïve Bayes: choose the Bayes Classifier, Naïve Bayes, with the default algorithm, set use Supervised-Discretization to TRUE, run with 10-fold Cross-Validation.

For the J48 decision tree: classify tab, followed by trees, choose J48, with binary splits set to TRUE, reduced error pruning set to TRUE, min-Num-Obj set to 60, with the other options set their default WEKA setting using 10-fold Cross-Validation.

For K-Nearest Neighbor:

- Choose the Lazy Classifier, IBK, with the default setting, except for the binary-Splits set to TRUE.
- Set reduced-Error-Pruning to TRUE.
- Run with 10-fold Cross-Validation.

The same WEKA method mentioned above was used for predicting the contributing factor for collisions, with the top 100 collisions over five years (2012-2019), using the training set.

Results and Discussion

A fundamental measure of traffic safety is crash prediction models. This research paper's objective was to provide multiple statistical models to predicate and compare crashes at high-risk locations and the latest year of collision data. The selection of intersections was identified based on the number of injured and killed and compared to the Vision Zero Priority Intersections and the population records.

General background derived from the 5-years data collection:

- The total killed in motor vehicle collisions from 2014- 2019 is 1,481
- The total Injured in motor vehicle collisions from 2014 to 2019 is 34,6881
- The total Killed in Pedestrians collisions from 2014 to 2019 is 795

- The total injuries in Pedestrian collisions from 2014 to 2019 is 65,053
- The total Killed in cyclist collisions from 2014-01-01 to 2019 is 121
- The total injuries in cyclist collisions from 2014 to 2019 27,85

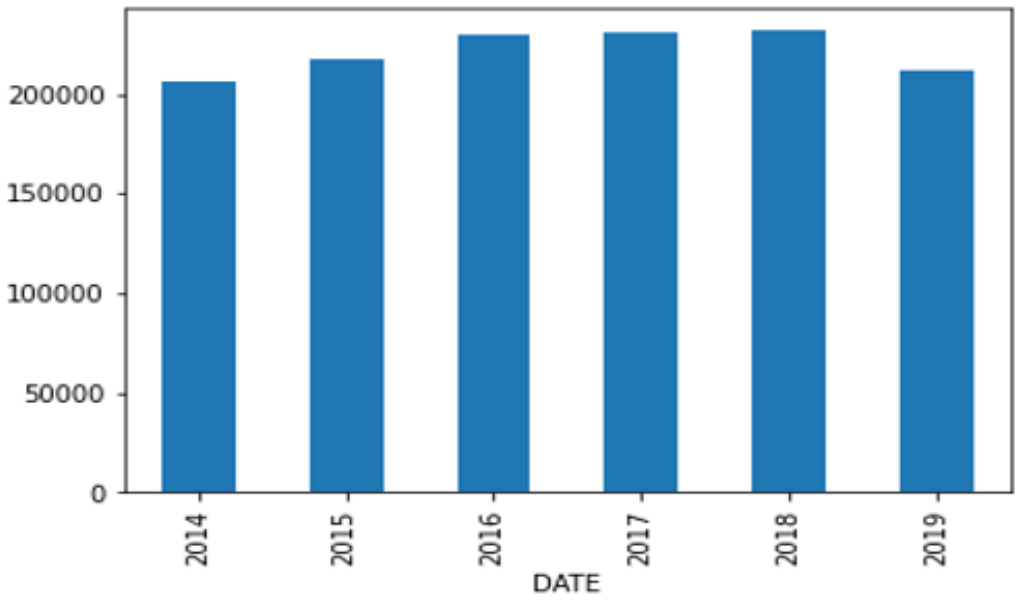


Figure 2. Total Collisions 2014-2019

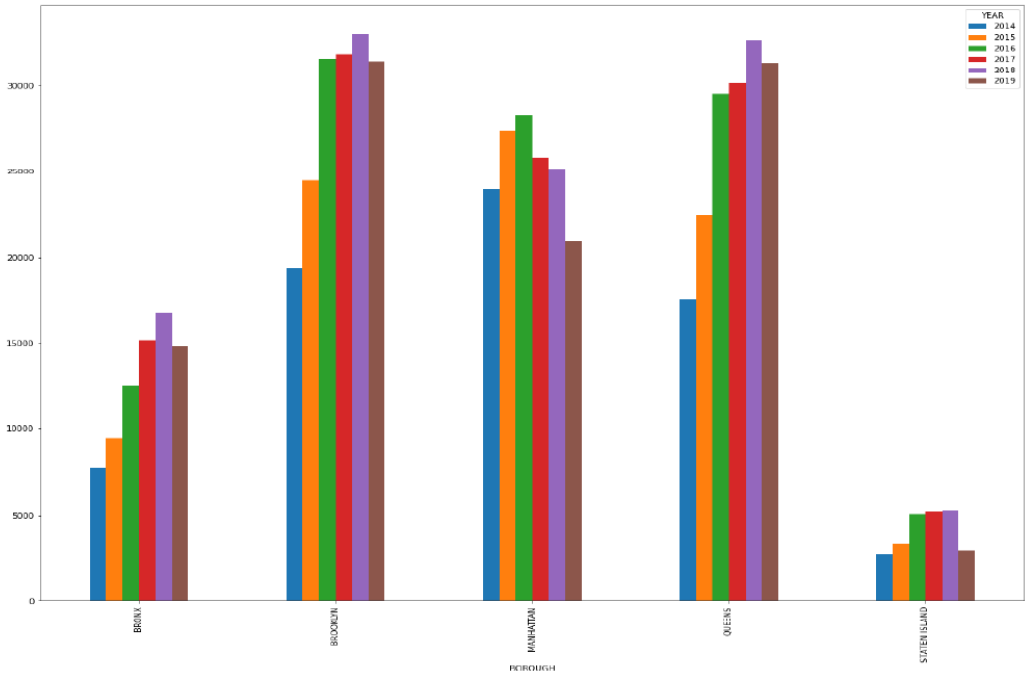


Figure 3. Number of Collision by Borough

Figure 3. reflects the collision rates by borough and by year, indicating that Brooklyn and Queens have the highest number of collisions, followed by Manhattan and the Bronx, with the lowest number of collisions in Staten Island.

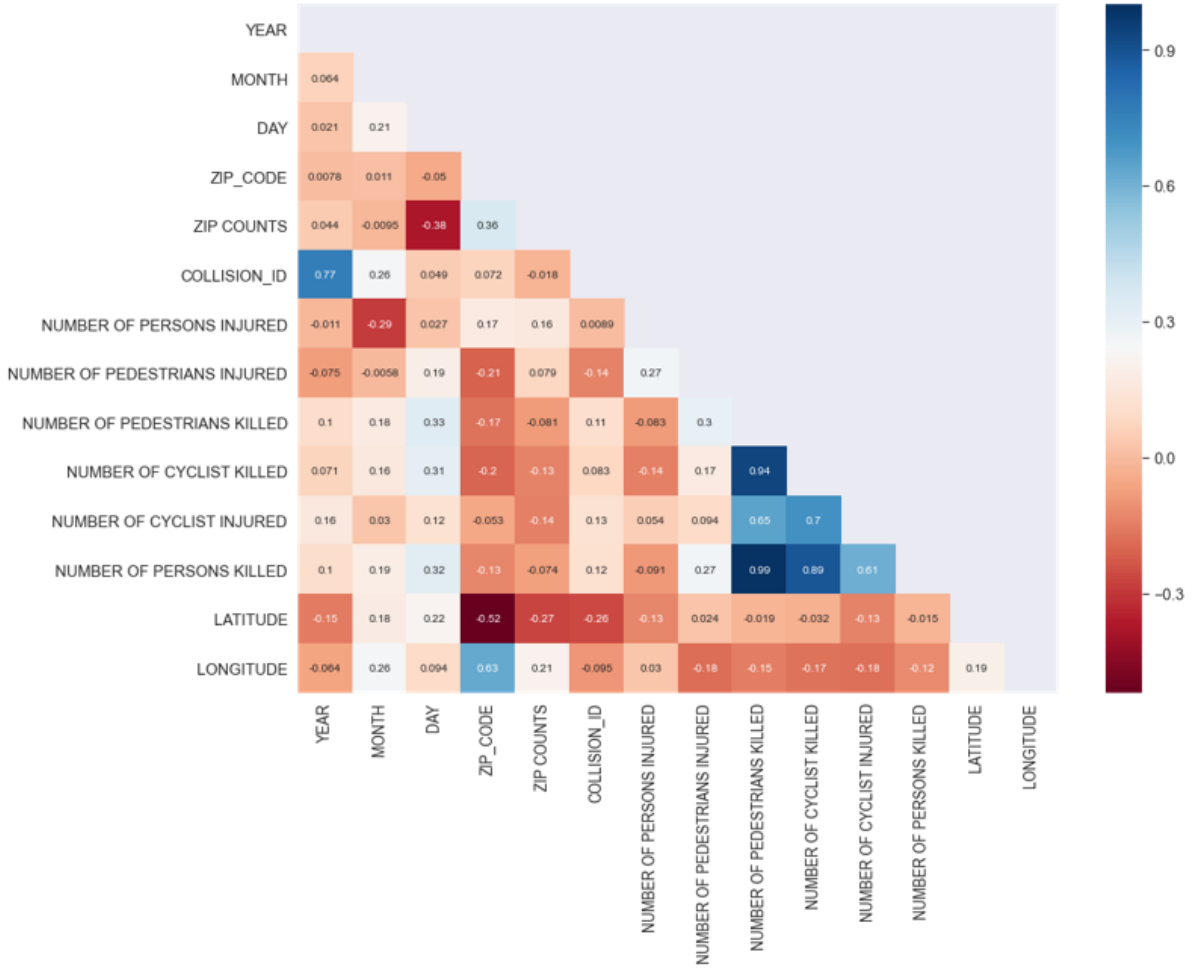


Figure 4. Correlation of data variables

Figure 1: By using these variables, it will allow for a better production outcome. The correlation between the Collision data reflected in the heatmap above shows each square indicates the correlation between variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables; such as the number

By using these variables, it will allow for a better production outcome. The correlation between the Collision data reflected in the heatmap above shows each square indicates the correlation between variables on each axis. Correlation ranges from -1 to +1. Values closer to zero means there is no linear trend between the two variables; such as the number of pedestrians killed, and the number of cyclists killed. Also, it shows a close relationship between the person killed and the pedestrian killed. However, the heatmap reflects a negative relationship between the zip count and the number of injured. As noted earlier, the number that is close to 1 indicates a more positively correlated correlation, which indicates a strong relationship. The numbers that are closer to -1 has more of a diverse correlation, but instead of both increasing, one variable will decrease

as the other increases. The larger the number and darker the color, the higher the correlation between the two variables.

The variables were used for both spatial analysis, and predictions models were the followings:

- Vehicles type code 1
- Vehicles type code 2
- Borough
- Year
- Zip Code
- Zip Code Counts
- Number of persons killed
- Number of persons injured
- Number of Pedestrians killed
- Number of Pedestrian injured
- Number of cyclists killed
- Number of cyclists injured
- Latitude, Longitude
- contributing_factor_vehicle_1

An overview of the collisions per zip code was carried out as part of the research, as reflected in the map below the zip code within Brooklyn: East New York, Flatbush, and East Williamsburg. For Queens are South Ozone Park and South Jamaica (near John F. Kennedy International Airport), and for Manhattan, the areas are East Village, West Village, and Flatiron District. These are the largest number of crashes, noted that they are highly dense areas with a wide range of income, with a variety of land use that will be discussed later in the analysis.

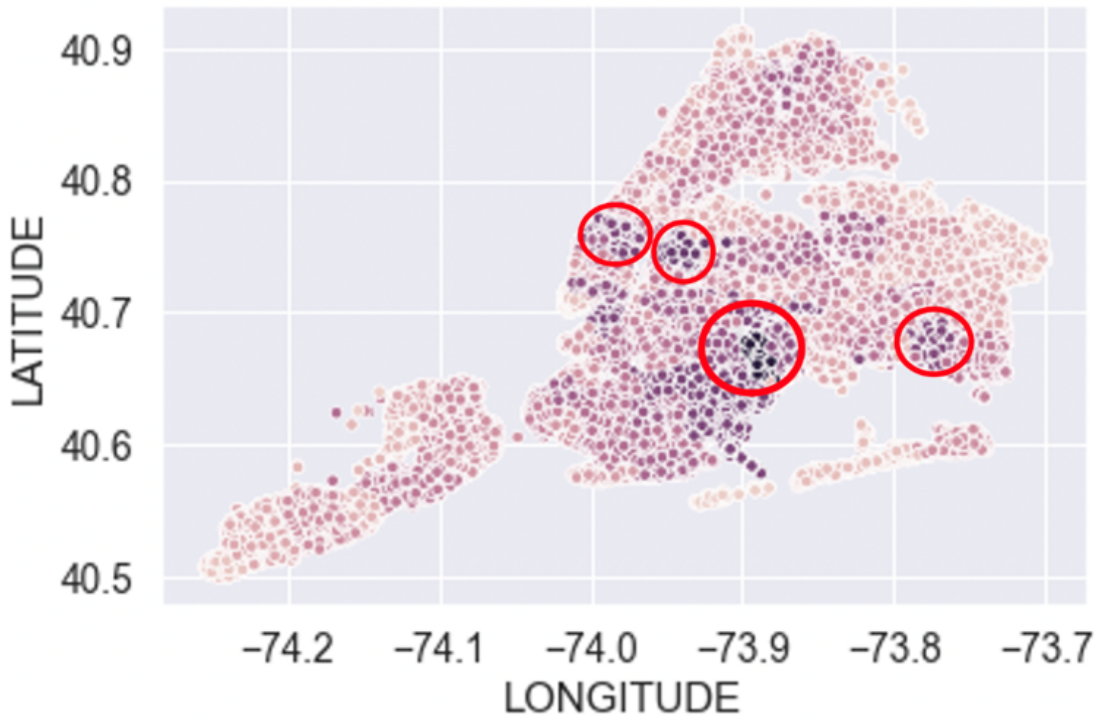


Figure 5. Collision Cluster based on neighborhood

Spatial Features and Analysis

The collision analysis started by examining the total number of crashes per zip code to classify intersections considered high-risk collision locations around the area. The study focused on 100 intersections with the highest number of collisions over five years. Additionally, the analysis examined 2019 crash results, creating a statistical model for persons injuries and fatalities. The number of injured and killed is reported as the cumulative number of accidents and fatalities for travel modes. The following maps indicated the different types of injuries and killed based on location, area, and transportation mode. The number of crashes was assigned per zip code and normalized by the total population; to determine the locations' high-risk collisions. (See appendix B)

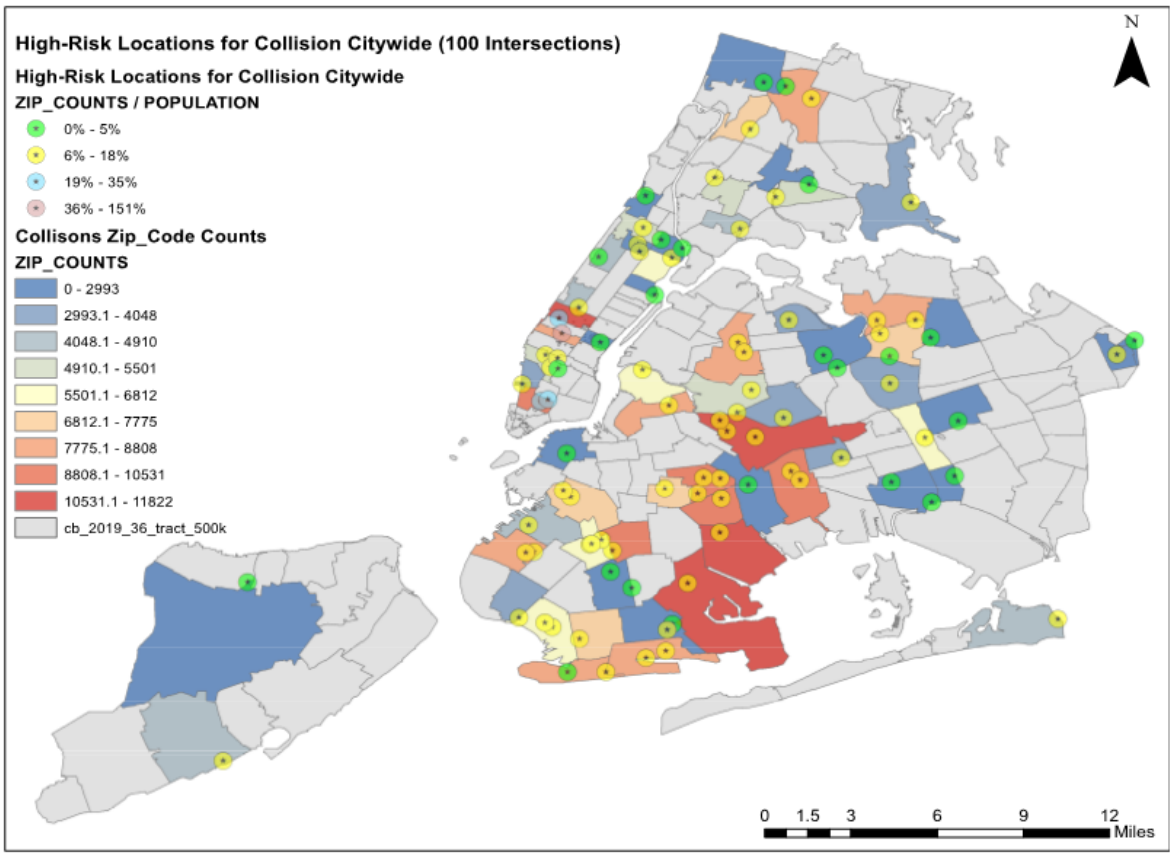


Figure 6. Zip Code Count/ High-Risk Locations for Collision Citywide (100 Intersections)

As shown in figure 6, the largest number of crashes were reported in Brooklyn, followed by Queens and Manhattan, with crashes ranging from 8,808 to 10,531 per zip code. The findings showed that most accidents occurred in residential areas of Brooklyn, Staten Island, and Queens; however, Manhattan does not have high-risk collisions within the residential area. The zip code count normalized by population indicated Brooklyn and Queens have a range of 6%-18% per zip code, which is considered low. On the other hand, Manhattan reflects a 19%-35% per zip code of collision normalized by population, considered average. Furthermore, the light pink marks indicate a high range of 36% -151% per zip code (High-rise buildings), indicating the area to be highly dense.

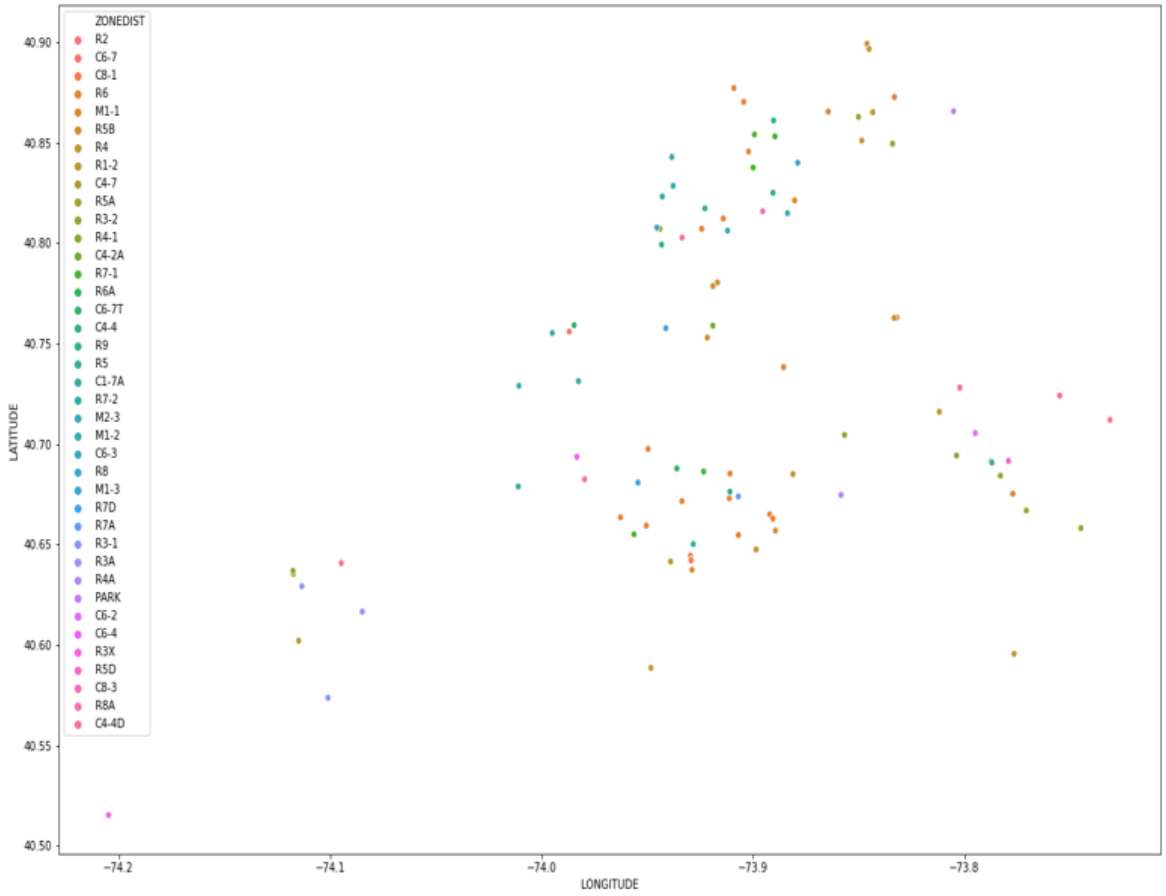


Figure 7. Land Use for the High-Risk Locations for Collision Citywide (100 Intersections)

Figure 7 shows the land-use for the top 100 High-Risk locations citywide. Brooklyn has the highest number of collisions; most of the crashes occurred within a residential area, followed by commercial and manufacturing. Queens had the second high-risk location, mainly residential, followed by manufacturing and commercial. In the Bronx, the land-use was residential, commercial, manufacturing, and park. In Manhattan, most collisions occurred within a commercial and manufacturing area, followed by a residential area. The borough of Staten Island was only residential. (See Table 1-5).

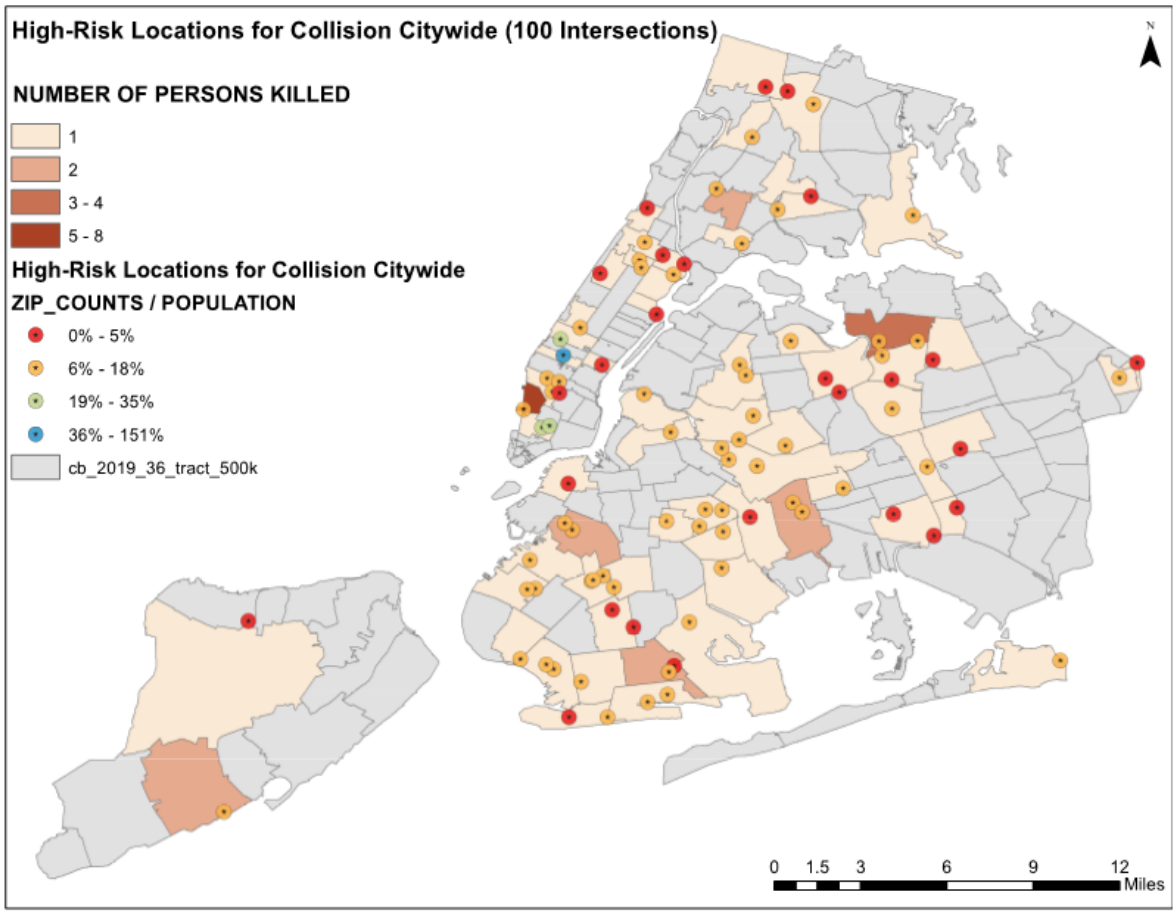


Figure 8. Number of Persons Killed High-Risk Locations for Collision Citywide

After evaluating both land use and zip code crash counts, an estimated number of fatalities ranged from 1-8. Suggesting that only Manhattan had a total of 5-8 fatalities with a median range of 6%-18% of the population normalized by zip code counts. Followed by Queens, with a fatality scale of 3-4, the zip code is similar to Manhattan. Staten Island shows that two people were killed, and the zip code count is 6%-18%. With a one-person range, the population normalized by zip code count was reported as 0%-5 %, a very low range. The Bronx had a low number of two killed, with a zip code count of 6 %-18%.

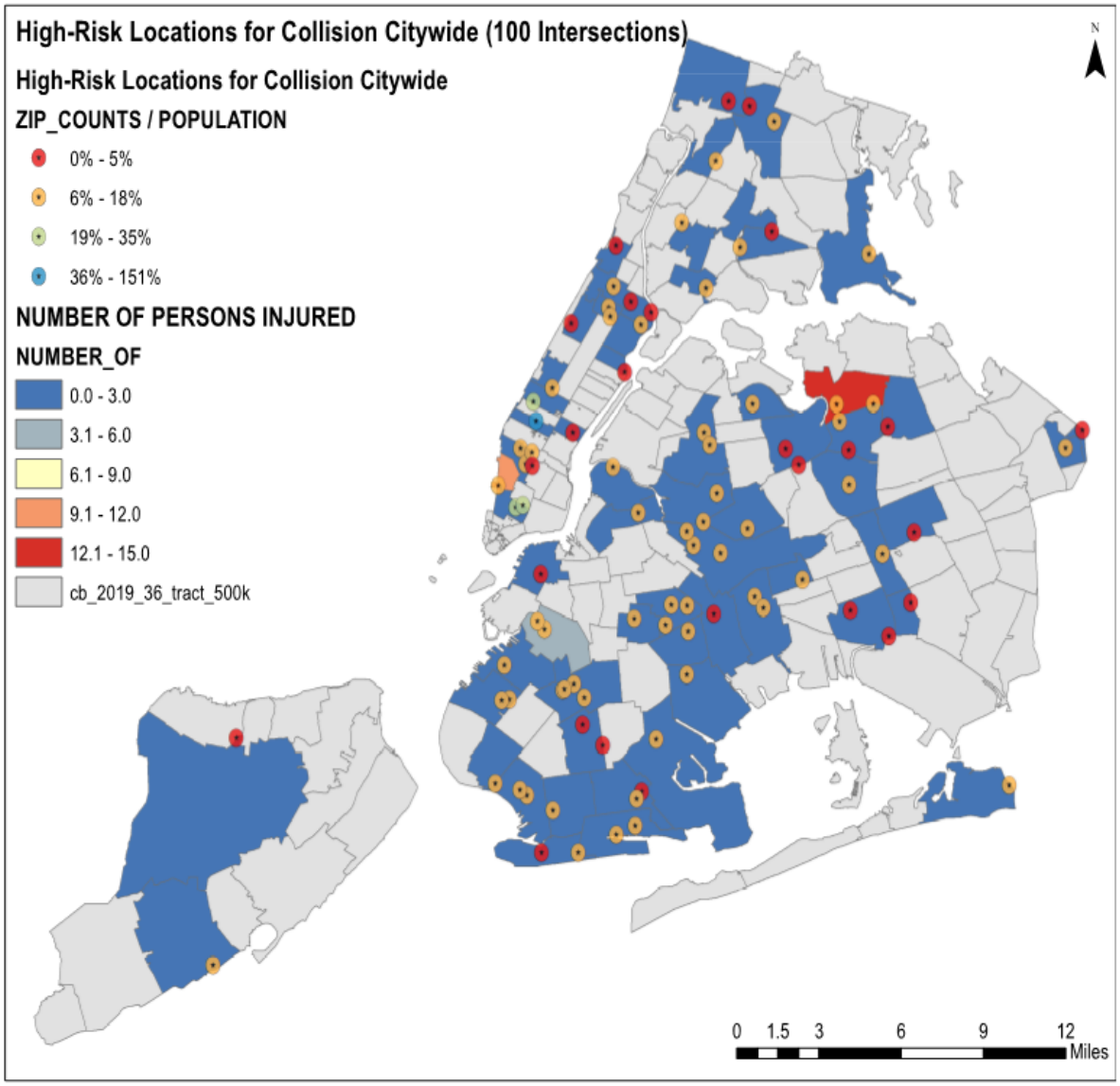


Figure 9. Number of Persons Injured High-Risk Locations for Collision Citywide

As seen in figure 9, Manhattan had a high range of 9-12 injures with a medium range of 6%-18% of zip code crash counts normalized by population. Queens had a location ranging from 12-15 injuries; the zip code count was similar to Manhattan. Staten Island had 0-3 persons injures. The zip code count was 6%-18% (this study assumes that 0 counts indicate an occurrence of collision). Also, Brooklyn had a range of 3-6 injures; the zip code counts were similar to Manhattan, Queens, Staten Island, and the Bronx.

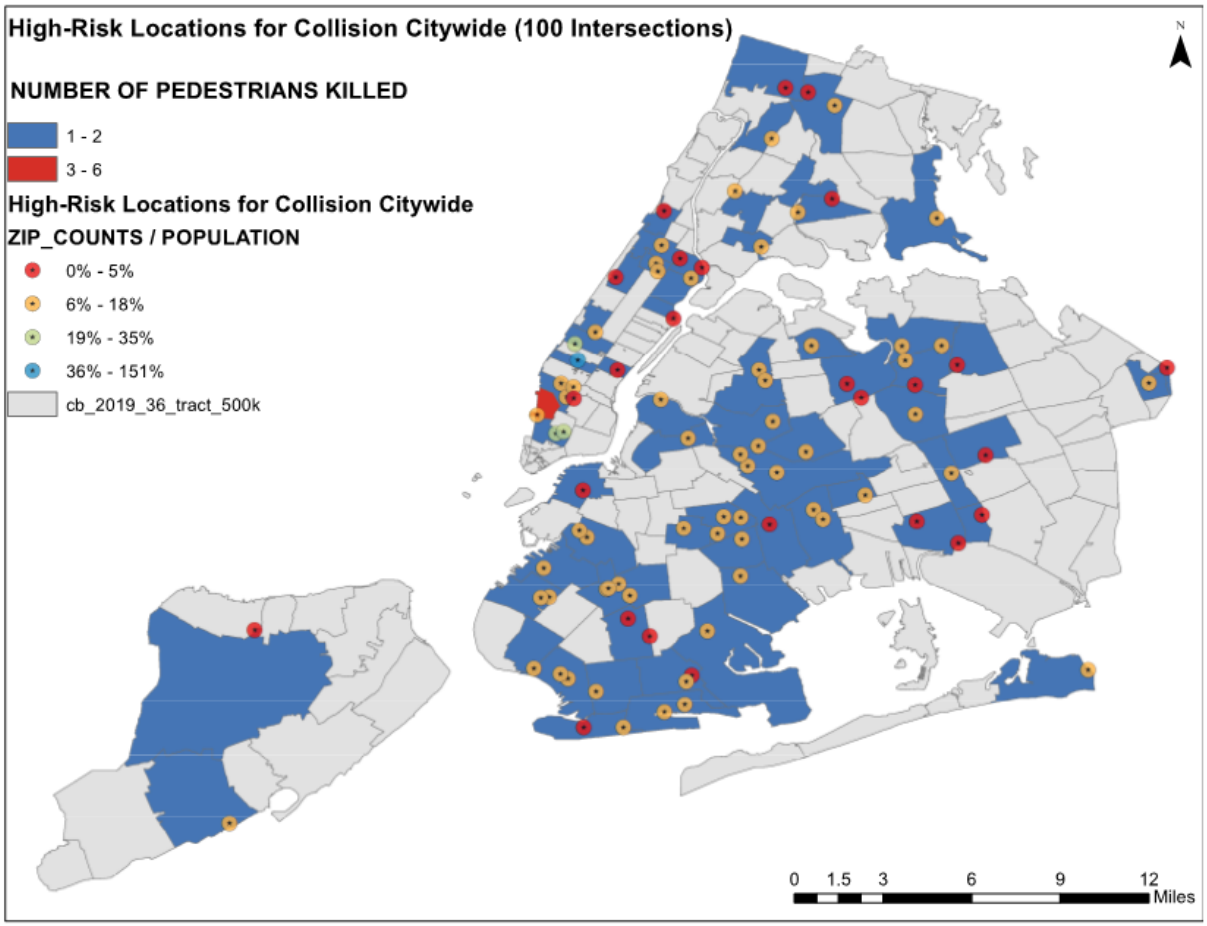


Figure 10. Number of Pedestrian Killed High-Risk Locations for Collision Citywide

Figure 10 shows the majority killed range between 1-2 within the four boroughs with the zip code count, ranging between 0% -5% and 6%- 18%. Except for Manhattan, the range killed was 3-6, and the overall rate of zip-code crash counts ranged from 6%- 18%. Within proximity, the zip code count, normalized by populations, range with a high percentage of 19% to 35% zip code count, range of 36% -151% within a close distance from the location where the 3-6 pedestrian killed.

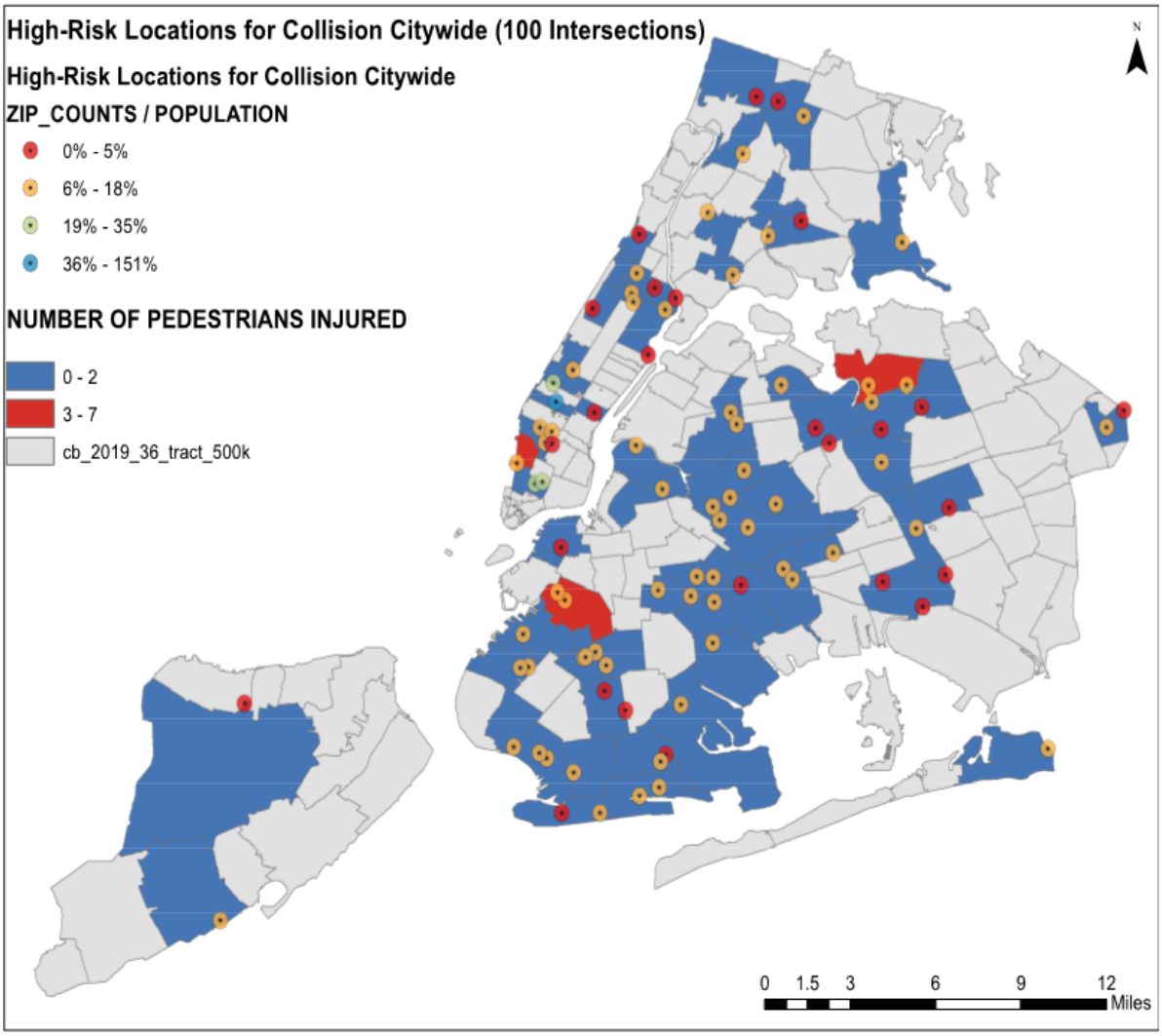


Figure 11. Number of Pedestrian Injured High-Risk Locations for Collision Citywide

As shown in figure 11, the number of pedestrians injured in Brooklyn, Queens, and Manhattan had crashes ranging from 2-7. The zip code count is 6%-18%. As shown in figure 6, Manhattan had a location where the zip code crash counts ranged from 8,808 to 10,531, a high number of collisions within one zip code. The area square footage is comparably minor to the zip code location in Brooklyn and Queens. With a range of 0 to 2 in the Bronx and Staten Island.

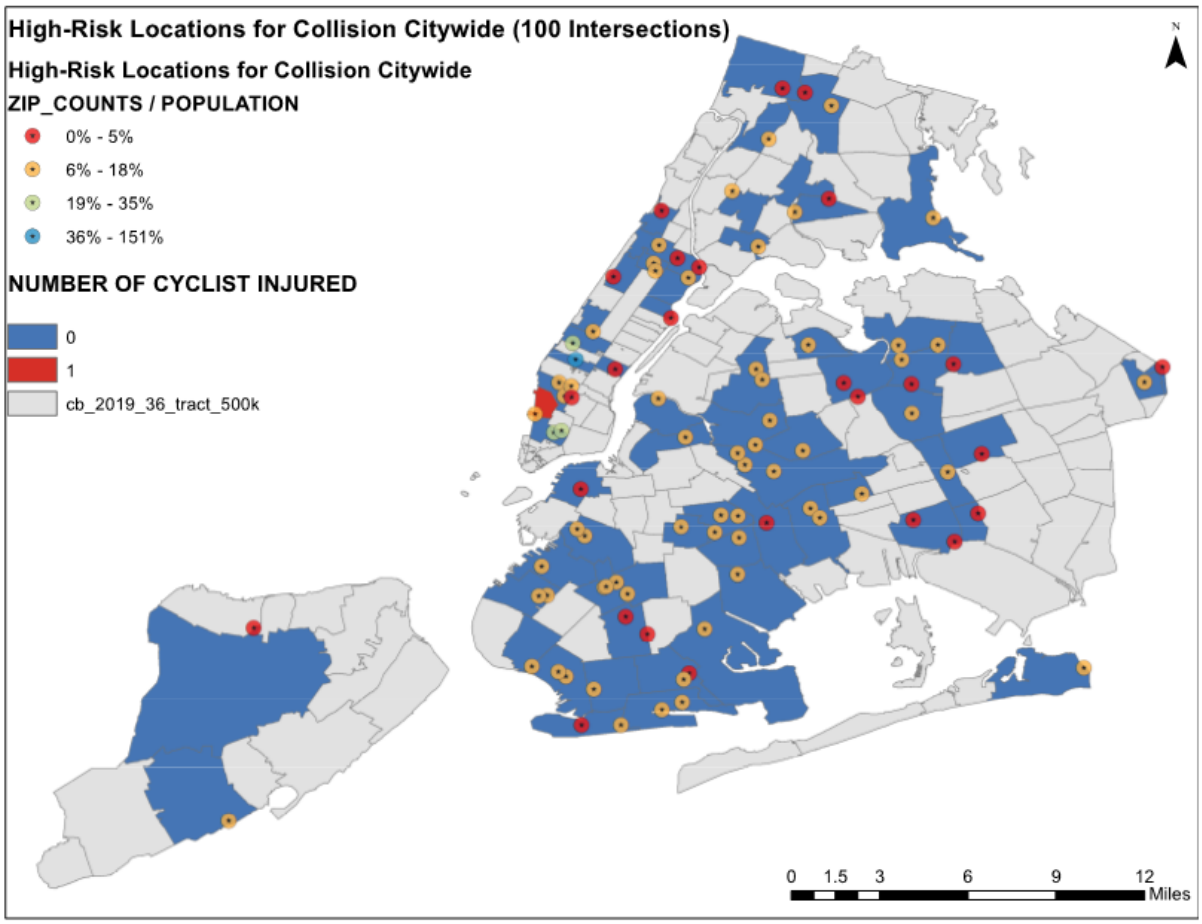


Figure 12. Number of Cyclist Injured High-Risk Locations for Collision Citywide

Figure 12 shows a pattern of reoccurrence of collision within the same zip code in Manhattan. The zip code count is 6%-18% within proximity of the zip code count, normalized by populations, ranges from a high percentage of 19% to 35% to the location of high-risk intersection. The percentage of cyclists injured was very low compared to other transportation modes.

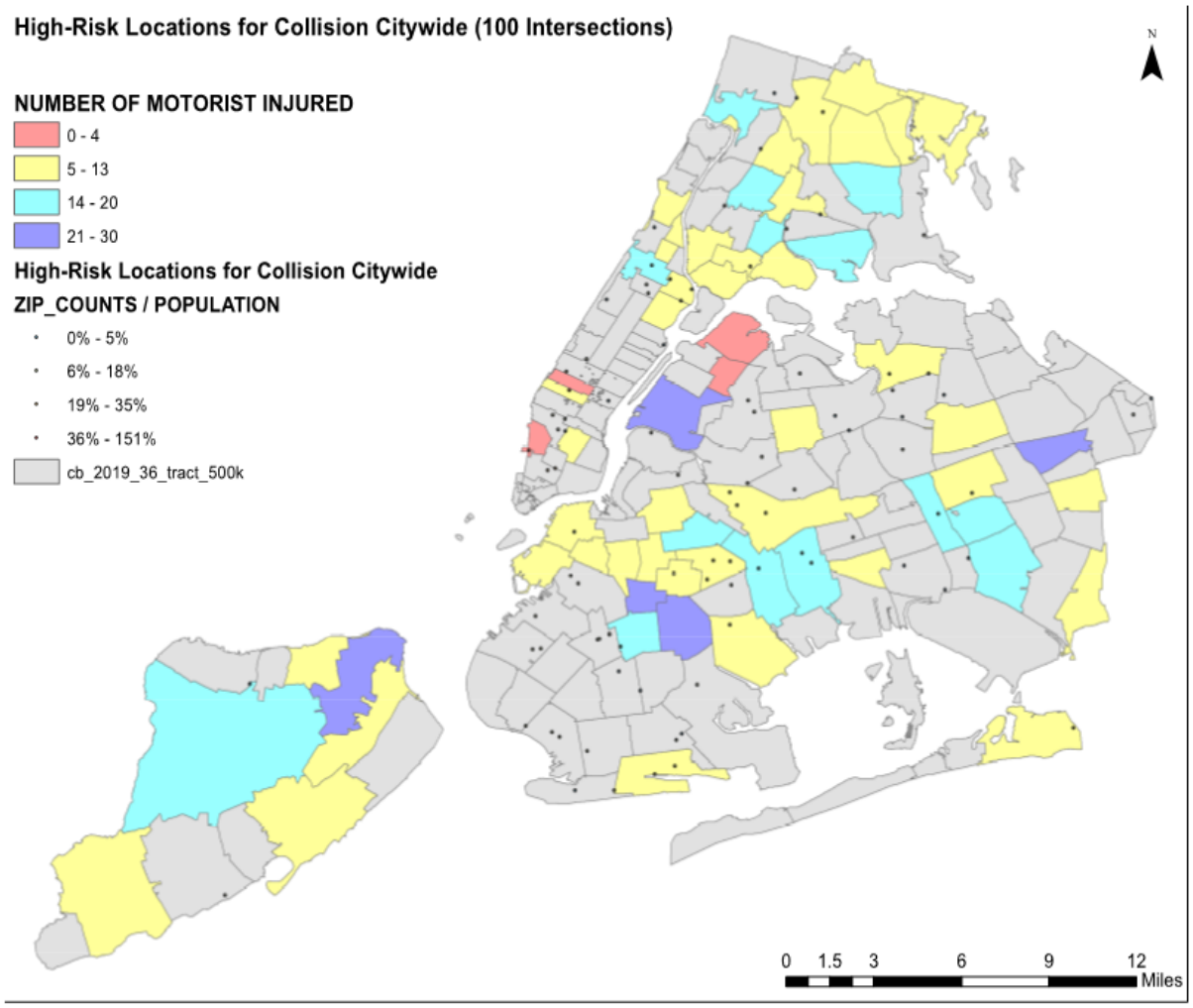


Figure 13. Number of Motorist Injured High-Risk Locations for Collision Citywide

Figure 13 shows the number of motorist injurers relatively high compared to other transportation modes with a total range of 1 to 30 injuries per crash. Staten Island has a range of 5-13, 14-20, and 21-30 injuries. The Bronx has an additional zip code with 14-20 injuries. With zip code crash counts normalized by a population that ranges between 0%-5% and 6%- 18% within the crashes area ranging between 5-13 and 14-20 injuries. Queens, Brooklyn, and Staten Island had a range of 21-30 crashes. A consistent range of zip code crash counts normalized by population ranges between 6%-18%, similar to other transportation modes.

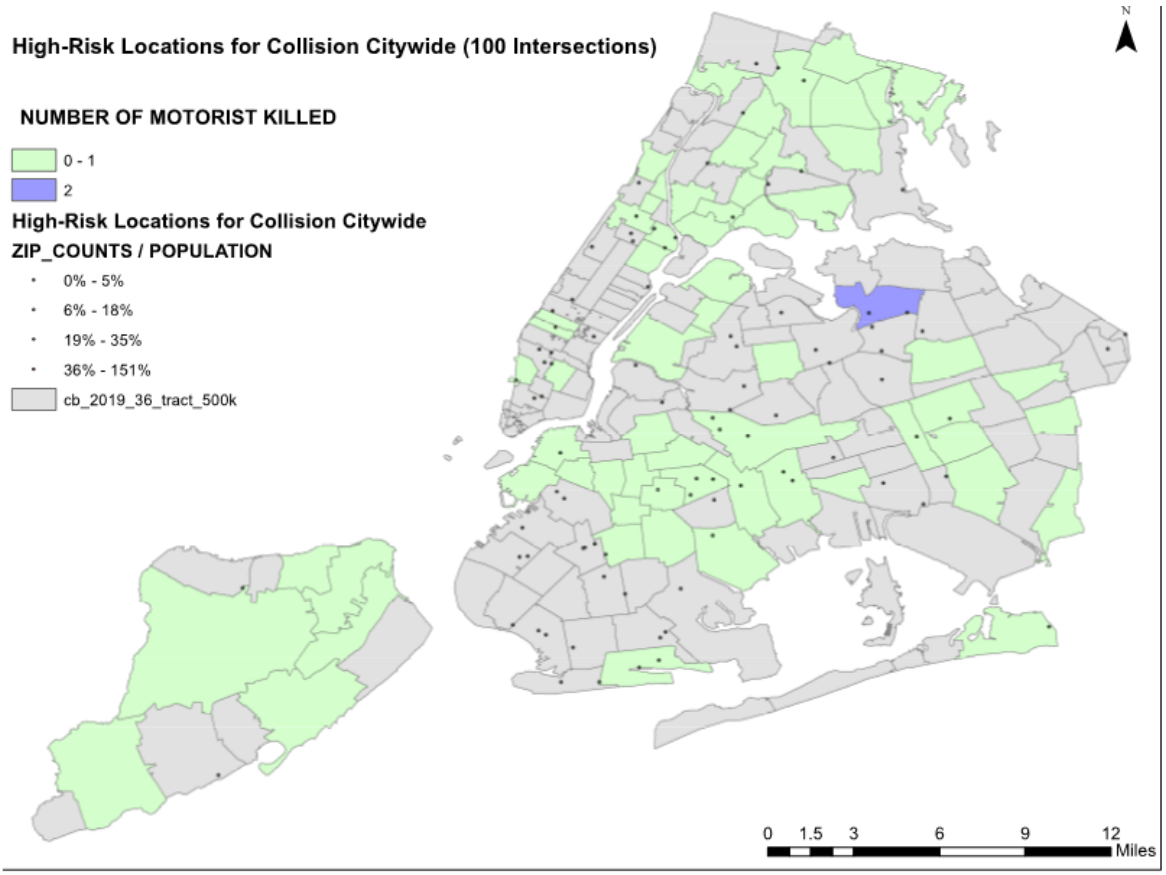


Figure 14. Number of Motorist Killed High-Risk Locations for Collision Citywide

As shown in Figure 14, Motorist fatalities ranged from 0 to 1 throughout the five boroughs, with increased motorist fatalities in Staten Island and the Bronx. Except in Queens with a zip code area range of 2 fatalities, a zip code crash counts normalized by population ranges between 6%-18%. Manhattan had a crash pattern within the same area; however, the motorist fatalities were comparable to the other transportation modes; the fatalities were low. Manhattan was very similar to Brooklyn by collision pattern. The numbers of motorist fatalities collisions were also low, and the zip code area was limited compared to the other modes.

Overall Spatial Analysis

The study identified collision patterns across the five boroughs, with different transportation modes, by recognizing crash frequency within a certain location. The borough of Brooklyn reflected a pattern of collisions within the following: zip codes: 11234, 11236, 11235,11231, 11217, and 11203 had a These zip codes had a different type of land use; however, the majority occurred within a residential zone. As shown in Table 1 below:

BOROUGH	ZONEDIST	ZIP_CODE
BROOKLYN	R4	11208
BROOKLYN	M1-1	11207
BROOKLYN	C8-1	11203
BROOKLYN	R6	11238
BROOKLYN	R6	11213
BROOKLYN	R4	11235
BROOKLYN	R6A	11221
BROOKLYN	C6-4	11201
BROOKLYN	R4	11236
BROOKLYN	R7D	11216
BROOKLYN	R6	11212
BROOKLYN	M1-1	11203
BROOKLYN	R7A	11212
BROOKLYN	R6	11233
BROOKLYN	R5	11233
BROOKLYN	R6	11206
BROOKLYN	R5	11203
BROOKLYN	R5	11231
BROOKLYN	R6	11207
BROOKLYN	R6	11225
BROOKLYN	C8-1	11203
BROOKLYN	R7-1	11226
BROOKLYN	R6	11233
BROOKLYN	R6A	11221
BROOKLYN	R6	11207
BROOKLYN	R5	11233
BROOKLYN	R8A	11217
BROOKLYN	R4	11203
BROOKLYN	R6	11207

Table 1. Brooklyn Land-Use

Queens had a high number of collisions with an established pattern of collisions in a particular area of various modes of transport; the following zip code had high collisions: 11354, 11432, 11365, 11366, and 11412. A more significant number of collisions occurred in both the commercial and manufacturing areas. This recognition allows for better planning, engineering, and decision-making, as seen in Table 2.

BOROUGH	ZONEDIST	ZIP_CODE
QUEENS	M1-1	11101
QUEENS	R5B	11105
QUEENS	M1-1	11354
QUEENS	R5D	11434
QUEENS	M1-1	11373
QUEENS	R6	11354
QUEENS	R5B	11105
QUEENS	C4-2A	11103
QUEENS	R4-1	11435
QUEENS	R5	11433
QUEENS	R3-2	11434
QUEENS	R4	11435
QUEENS	R3-2	11434
QUEENS	R3-2	11434
QUEENS	M1-3	11101
QUEENS	C6-2	11432
QUEENS	R6	11434
QUEENS	R2	11427
QUEENS	R4-1	11385
QUEENS	R4	11691
QUEENS	R3-2	11433
QUEENS	R3-2	11422
QUEENS	R2	11429
QUEENS	R4A	11417
QUEENS	R2	11366

Table 2. Queens Land-Use

Manhattan indicated a pattern of crashes with a single zip code for all transportation modes, 10014, zoned as a manufacturing area. However, other zip codes have been classified as areas a high risk of collision: They are 10029, 10018, 10036, and 10115—the zoning for these zip codes was a different commercial zone type; a transition of land uses from commercial to residential zoning, as shown in Table 3.

BOROUGH	ZONEDIST	ZIP_CODE
MANHATTAN	R7-2	10032
MANHATTAN	C1-7A	10018
MANHATTAN	C6-7T	10036
MANHATTAN	R9	10029
MANHATTAN	C1-7A	10009
MANHATTAN	C6-3	10027
MANHATTAN	R7-2	10039
MANHATTAN	C4-7	10027
MANHATTAN	C4-4D	10035
MANHATTAN	M2-3	10014
MANHATTAN	C6-7	10036
MANHATTAN	R7-2	10030

Table 3. Manhattan Land-Use

The Bronx had a low number of collisions compared to Brooklyn, Queens, and Manhattan. The zip codes with high-risk for collisions were 10466, 10458, 10463, 10470, and 10473. Most crashes occurred within a commercial and manufacturing area. However, the highest number of crashes occurred within a different residential zone type. As shown in Table 4.

BOROUGH	ZONEDIST	ZIP_CODE
BRONX	R8	10460
BRONX	R6	10467
BRONX	R7-1	10457
BRONX	R4	10466
BRONX	R7-1	10457
BRONX	M1-2	10454
BRONX	C4-4	10451
BRONX	R6	10475
BRONX	M1-2	10474
BRONX	R4	10466
BRONX	PARK	10464
BRONX	R6	10463
BRONX	R6	10455
BRONX	R4-1	10469
BRONX	R6	10454
BRONX	R4	10469
BRONX	R6	10454
BRONX	C8-3	10455
BRONX	R7-1	10457
BRONX	C4-4	10459
BRONX	C4-4	10458
BRONX	M1-1	10473
BRONX	M1-1	10457
BRONX	R6	10463
BRONX	R5A	10461
BRONX	R4	10461

Table 4. Bronx Land-Use

Staten Island had the lowest number of collisions compared to the other boroughs. However, Staten Island had the highest number of motorist collisions, both injured and killed. The following zip code developed a crashes pattern that indicated high-risk collision areas; 10304, 10314,10311, and 10309; the zoning criteria were all in different residential zonings, as seen in Table 5.

BOROUGH	ZONEDIST	ZIP_CODE
STATEN ISLAND	R3-1	10306
STATEN ISLAND	R3-2	10310
STATEN ISLAND	R3-1	10310
STATEN ISLAND	R3X	10309
STATEN ISLAND	R1-2	10301
STATEN ISLAND	R2	10301
STATEN ISLAND	R3A	10304
STATEN ISLAND	R3-2	10310

Table 5. Staten Island Land-Use

Land Use and Zoning Criteria

As noted earlier, the highest number of collisions occurred in locations with various zoning areas (R-Residential, C-Commercial, M-Manufacturing), such as R3, R4, R5, R6, R7A, R6A, R8, and R9. R3 zoning is for one to two-story, semi-detached, and detached that accommodate low-density districts. R4 and R5 have attached houses and apartment houses, of two stories homes within a transit zone. These are mainly mapped in Brooklyn, Queens, Bronx, and Staten Island. R5 allows for a third story, semi-detached or detached, with off-street parking and widely zoned for Brooklyn and Queens. R5 is used as a transiting zone between low- and high-density neighborhoods. R6 and R7 are built for medium-density areas, and they are located mostly in Brooklyn, Queens, and the Bronx; they are buildings for multiple family buildings. R6 allows for 13 stories, and R7 allows 14 stories with open space as a requirement. R7, mostly located in Manhattan and the Bronx, requires a lower parking area. R8 are apartment buildings, ranging from mid-rise, 8 to 10 story buildings; considered to be zoned for a high-density area. R8 are mainly found in the Bronx, Grand Concourse, Brooklyn Heights, and widely mapped in Manhattan's borough. R9, located in a high-density area, mainly in the borough of Manhattan R9, requires off-street parking. Some of the significant collisions occurred within commercial areas throughout the five boroughs.

The following commercial-zoned are C1, C4, C6, and C8. C1 are mapped within a high dens area with a residential component, a commercial overlay area; the commercial use is limited to one to two stories. To be noted, C1 also can be mixed with R8, R9, and R10. C4 is mapped in regional commercial centers, mainly located in Queens, Bronx, and Staten Island, outside of the business district. Some of the C4 mapped building includes a residential component. C6 is considered as a wide range of high bulk commercial uses that require a central location. C6 is mostly located in Manhattan, Downtown Jamaica, Queens, and Downtown Brooklyn. Most of C6 are built as corporate headquarter, large hotels, and department stores as part of the zoning criteria; noting, it must include a public plaza with the built right of way. C8 is a land-use transitioning from commercial to manufacturing uses. C8 is widely used for commercial services, such as car wash, automobile showroom, and repair shop. These zones are found in Bay Ridge, Brooklyn, and Castleton Corners, Staten Island; the C8 zone does not allow for any housing or residential use.

As part of the analysis, a reoccurrence collision within the manufacturing zone areas. The manufacturing zones: M1-1, M1-2, M1-3, and M2-3. M1-1, M1-2, and M1-3 are manufacturing areas from the Garment District in Manhattan and Port Morris in the Bronx and Red Hook, Brooklyn, and College Points, Queens:

these are built as a multi-story loft or one-two story warehouse. M1 must include a buffer between M2 or M3 districts and adjacent residential or commercial districts. These districts are subject to parking requirements based on the square footage of the building. M2-3, which is only located in Manhattan. [26]

Contributing Factors

One of the variables included in this study was collision contributing factors, using data from the New York City Police Department from Five Years (2014-2019). The following information was extracted from the NYPD dataset for the top 25 contributing factors for collisions. As indicated, human error is the top cause for collisions; however, part of the most recurrent causes are geometric design obstacles such as oversized vehicles, slippery pavement and view obstruction, and limited visibility. However, the NYPD reports do not indicate the pavement condition/roadway conditions as contributing factors.

- Driver Inattention/Distraction
- Failure to Yield Right-of-Way
- Following Too Closely
- Backing Unsafely
- Other Vehicular
- Passing or Lane Usage Improper
- Passing Too Closely
- Turning Improperly
- Unsafe Lane Changing
- Fatigued/Drowsy
- Traffic Control Disregarded
- Driver Inexperience
- Reaction to Uninvolved Vehicle
- Unsafe Speed
- Alcohol Involvement
- Pavement Slippery
- Lost Consciousness
- Prescription Medication
- View Obstructed/Limited
- Oversized Vehicle
- Outside Car Distraction
- Pedestrian/Bicyclist/Other Pedestrian Error/Confusion
- Passenger Distraction
- Physical Disability
- Aggressive Driving/Road Rage

This crash prediction analysis reviewed the intersection contributing factors to investigate these collision causes— see appendix for contributing factors visualization. It indicates the most recurrence based on the borough and the year. Most importantly, the two factors that could be mitigated are the Slippery Pavement and view obstruction, limited visibility, and oversized vehicle.

The oversize vehicle collisions are the highest cause of crashes. As part of the analysis, the truck route was analyzed to measure the number of crashes that intersect with a truck route. This process would allow for a cross-examination of the mitigation measures and course of actions that NYC-DOT took over the course of the 5-years, based on the data collection—using the High-Risk Locations for Collision Citywide (100 Intersections)

As reflected in figure 2. 2015 and 2016, the highest number of oversized vehicle collisions is in Brooklyn's

borough. (See appendix A. for visualization)

Traffic collisions are mostly caused by motorists' poor decision-making, including speeding, inability to comply, and distracted driving. However, some traffic accidents are caused by undefined attribution factors, such as poor road conditions. Poor road conditions could lead to more than an unpleasant or uncomfortable ride; roads can deteriorate to the point that they become unsafe. That includes potholes, poor road surfaces, cracked asphalt, exposed rebars, boreholes, and road fractures. The Agency performs continuous evaluation; the ratings are based on a scale from 1 to 10, and results are grouped in the following categories: Good, Fair, and Poor. According to TRIP (National Transportation Research Nonprofit) [28], every year, below-standard roads cost drivers in the New York metro area an average of almost $800 daily in wear and tear on their cars. As per the police reporting/records, the street condition is not a contributing factor causing the collision. However, as part of this analysis, it is essential to identify whether the roadway condition could be a contributing factor for collisions. To address the question of how many injuries and fatalities occur within a close distance to a poorly graded roadway, the data sources used were New York Police Department (NYPD) Collision Data Collection and Street Pavement Rating from the New York City Department of Transportation. The pavement Ratings are scaled from 1 to 10. The results are divided into the following categories:

- Good- ratings of 8 to 10
- Fair - ratings of 4 to 7
- Poor - ratings of 1 to 3

As part of the spatial analysis, a 1000-ft buffer (which is the length of the roadway segment based on the pavement ratings) was developed to obtain the number of crashes that occur within the distance of a poorly rated roadway segment. The Pavement Rating Data is recoding as roadway segments, ranging from 100ft to 1000ft, and the Collision Data is recoding as points at the intersection. The analysis studied a collision that occurred in the year 2019.

As part of the analysis, a calculation was performed to obtain the total number of motorist injuries and fatalities. This analysis also reviewed the total pedestrian fatalities and injuries with the 1,000-ft distance within proximity of a poorly rated roadway segment.

In 2019, the total number of injuries was 55,827, and 3,622 intersect with a poorly rated pavement condition; 227 total fatalities and 127 intersected with a poorly rated pavement condition. A total of 10,182 pedestrians, a 6,267 intersected with the poor rated pavement, and 123 total pedestrian fatalities and 83 intersects with a poorly rated roadway.

WEKA Results

The results compare five supervised data mining algorithms using WEKA (Waikato Environment for Knowledge Analysis) 3.8.4 data mining software. Performance is evaluated by algorithms such as Naïve Bayes, Bayes Network, K-Nearest Neighbor (KNN), J48, and Random Tree. Comparison of data mining algorithm results based on error rate, processing time, precision value, and accuracy. This analysis applies the following predictors to evaluate the accuracy metrics; to determine which is the most accurate predictor. [7]

A Bayesian classifier is founded on the notion that predicting the values of features for elements of that class is the function of a (natural) class, and it is a probabilistic graphical model that represents information about a set of random variables [8]. For this study, the Bayes Network and Naïve Bayes algorithm are used for the predictive modeling.

Decision trees work to evaluate an instance of data by constructing a tree, beginning at the root of the tree and progressing to the leaves (roots) before a prediction can be made. The method of constructing a decision

tree works by greedily choosing the best split point to make predictions and repeating the process until a fixed depth is reached by the tree. It is pruned after the tree is developed in order to boost the capacity of the model to generalize to new data [9]. The J48 algorithm is used to classify various applications and to obtain correct classification results. Random Forest is an Ensemble Learning Algorithm that works by constructing a multitude of decision-making trees at training time and producing the predicted class. [10]

The K-Nearest Neighbors algorithm operates by storing and querying the full training dataset to find the most related training patterns when making a prediction [11]. The k-NN algorithm is used for the calculation of constant variables in k-NN regression. Also, this algorithm uses a weighted average of the closest k neighbors, weighted by their distance inversely. This algorithm works as follows: Root Mean Squared Error dependent. This is achieved with the aid of cross-validation. The key downside to KNN being increasingly slower as data volume expands makes it an inefficient alternative in conditions where predictions need to be done quickly [12].

Classifier Name	**Accuracy**	**Time in Seconds**
Bayes Network	81.57%	0.22
Naïve Bayes	81.59%	0.03
J48	80.81%	6.94
KNN	80.20%	0.03

Table 6. Performance Classifier results, using WEKA

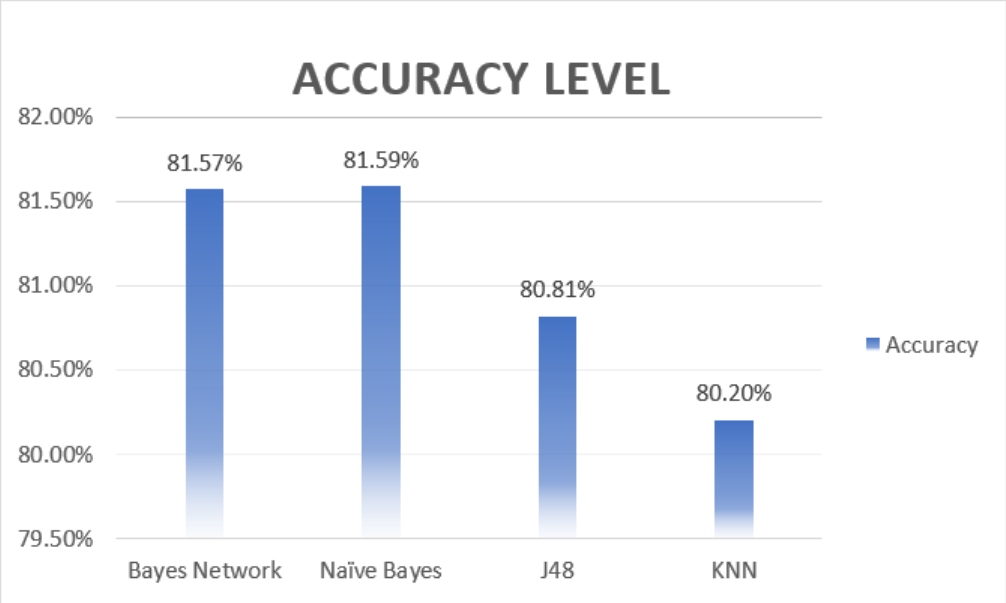


Figure 15. Comparison of Accuracy Percentage for: Bayes Network, Naïve Bayes, J48, and KNN

The Accuracy rate (AC) is the percentage of accurate forecasts, depending on the confusion matrix, and it can be determined by the following equation:

$$AC = \frac{TN + TP}{TP + FP + FN + TN'}$$

Where:

- TN: True Negative,
- TP: True Positive,
- FP: False Positive
- FN: false Negative.

As shown in Figure 15. the Naive Bayes classifier is more effective when compared to other Basic classifiers in predicting number of injuries. As illustrated in Table 1. The error rates and accuracy of each classifier are shown, showing the accuracy for the Naïve Bayes is about 81.59%. For the Naïve Bayes it took 0.03 seconds to run compared to the other classifiers with KNN having the same time and Bayes Network being 0.22 seconds, and the most time was 6.84 seconds for the J48 Decision Tree.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

As shown above, the Root Mean Squared Error (RMSE) is the root mean squared error deviation that is determined by the square root of the Mean Squared Error (MSE). Generally, the RMSE is used to calculate the difference between the actual values and the predicted value of the model. the RMSE represents the standard deviation of the difference between the predicted values and the actual values. It is preferable that the value of the RMSE be low (cigar).

Classifier Name	Mean Absolute Error (MAE)	Root mean squared error (RMSE)
Bayes Network	0.2808	0.3761
Naïve Bayes	0.2798	0.3763
J48	0.3061	0.389
KNN	0.2676	0.4031

Table 7. Accuracy results of batch learning methods

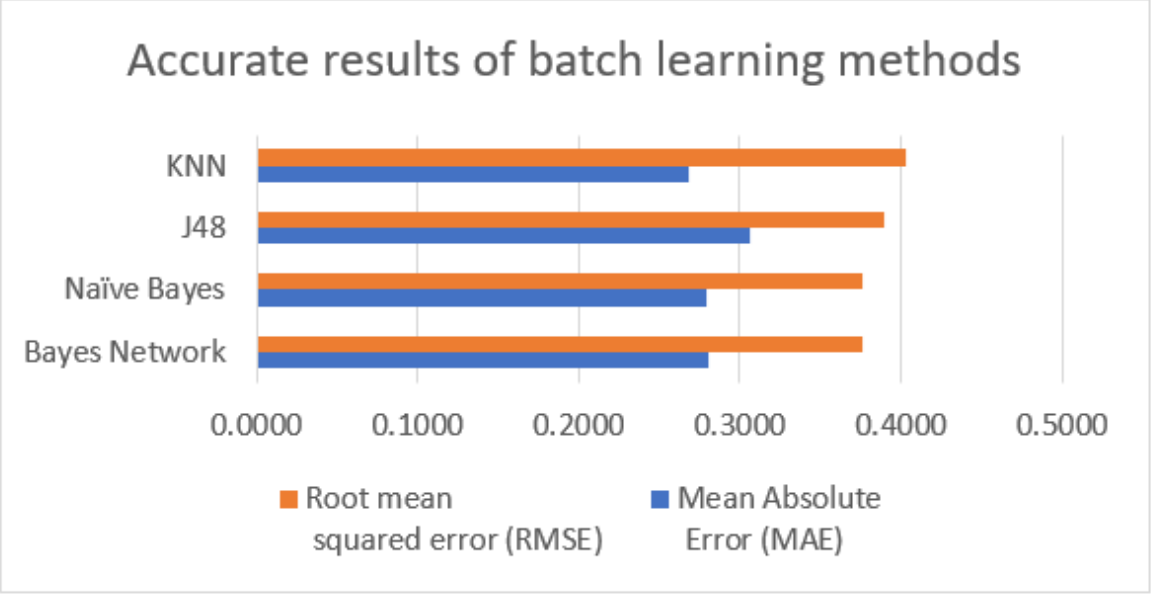


Figure 16. Accurate results of batch learning methods

As shown in Figure 16, the result of the batch learning method, as shown in Table 7, comparing the accuracy of the Bayes Network, Naïve Bayes, J48 Decision Tree, and KNN methods using the RMSE and MAE. The best RMSE was reported for Bayes Network, followed by the Naïve Bayes. The MAE follows a different trend, KNN shows the best results followed by Naïve Bayes, and the Bayes Network.

Classifier Name	ROC area	Precision	Recall	F-measure
Bayes Network	0.715	0.727	0.178	0.286
Naïve Bayes	0.716	0.741	0.172	0.279
J48	0.663	0.757	0.110	0.192
KNN	0.669	0.579	0.165	0.256

Table 8. Comparison of classification algorithm, detailed accuracy by class.

Precision (P) is the fraction of the positive observations correctly predicted from the overall positive observations predicted, and it can be determined by the following equation:

$$P = \frac{TP}{TP + FP'}$$

Recall (R) is the fraction of the positive observations correctly predicted from the overall observations in the class, and it can be determined by the following equation:

$$R = \frac{TP}{TP + FN'}$$

F-measure: The Precision and Recall material that can be interpreted together instead of separately. The F-Measure values produced by the harmonic mean of the columns of Precision and Recall accomplishing this since the harmonic mean provides the average of two different factors generated per unit. Therefore, F offers both the degree of classification precision and how stable it is (less data loss) [13]: where the precision is and where the recall is, and it can be determined by the following equation:

$$\mathrm{F-measure} = \frac{2 \times P \times R}{P + R}$$

ROC area is the predictive efficiency of the various classification algorithms is determined by the ROC field curve. One of the critical assessment metrics used to choose the right classification algorithm is the area under the ROC curve. If the area below the curve approaches 1, it means that the classification has been correctly carried out.

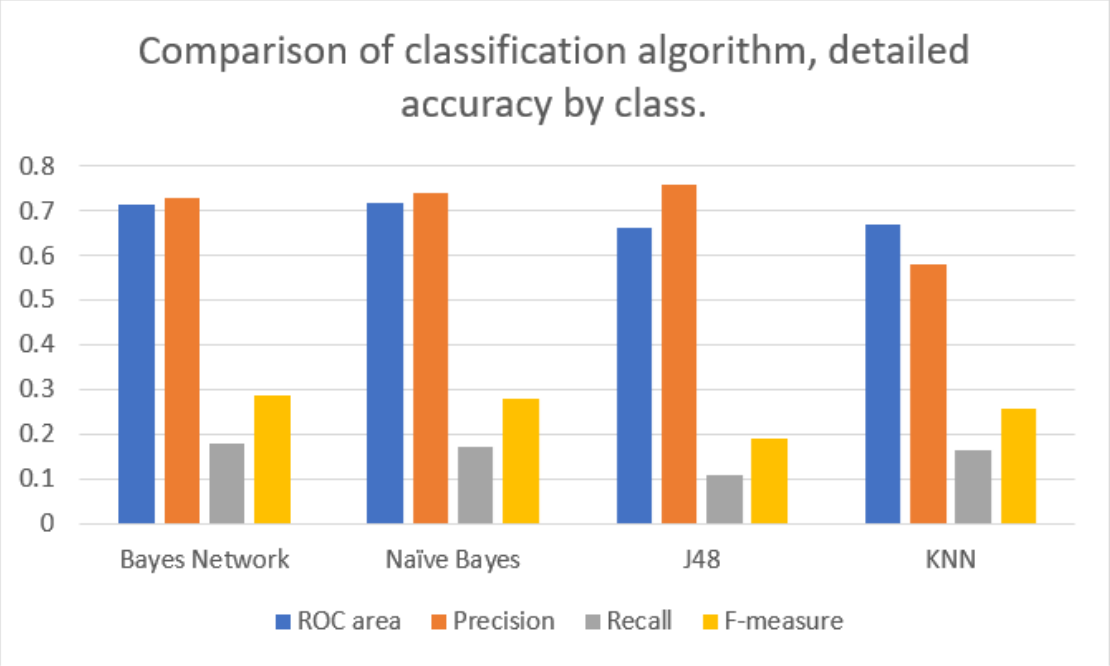


Figure 17: Comparison of classification algorithm, detailed accuracy by class.

As shown in table 8, and figure 17, it shows that the ROC area is highest at the Naïve Bayes followed by Bayes Network. The algorithms that produced the best results for the precision criterion are the J48 followed by Naive Bayes and Bayes Network. In conclusion, Naïve Bayes and Bayes Network are the best algorithm to predict the number of injuries, in reference to RMSE, ROC area, accuracy, F-measure, and recall statistical measures. The only exclusion is the precision value, where KNN had the best precision. The KNN showed the best accuracy to predict the number of killed in a collision.

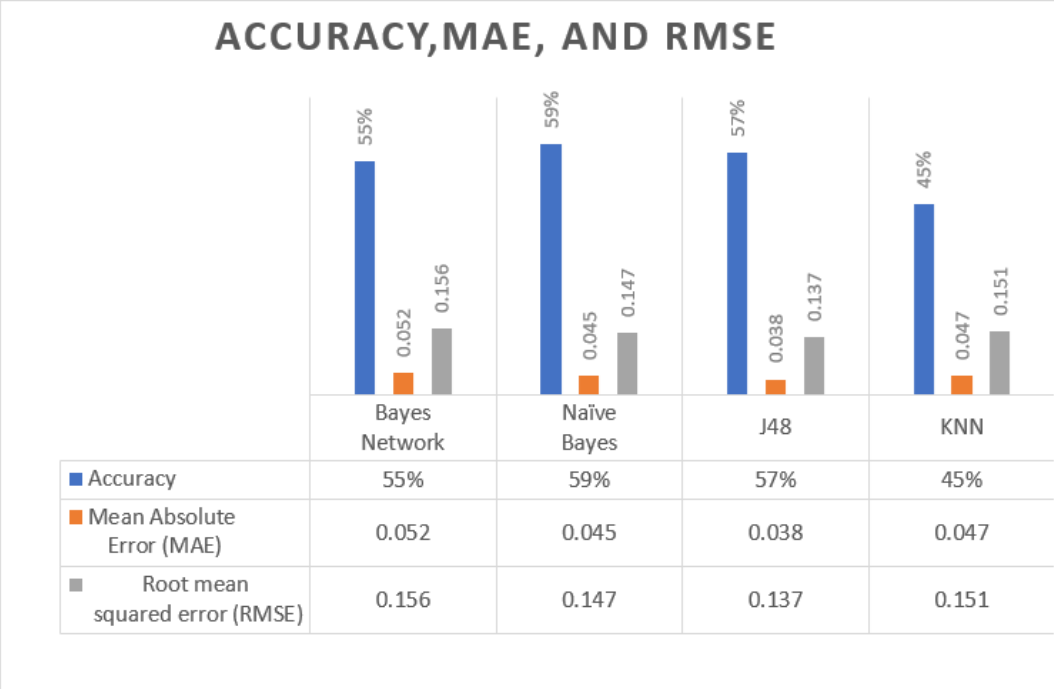


Figure 18: Accuracy Results, and results of batch learning methods

To predict the Contributing factor for the top one hundred collisions. BayesNet, Naïve Bayes, J48, and KNN classification algorithms were used. Figure 18 shows the accuracy, MAE, and RMSE, for each algorithm. Naïve Bayes had the best accuracy level of 59%, with the lowest MAE. Concluding Naïve Bayes is the most fitting algorithm to predict a collision's contributing factor.

Assumptions

The analysis does not consider all the dataset's present assumptions, creating a bias in the results. For example, the roadway ratings are subjective, and there are no universal requirements for the measurement. The number of crashes that occurred within the intersected poor-rated roadway segments appears to have a larger number compared to the total number of fatalities and injuries; however, the NYPD raw data might include some errors, which the analysis did not account for. For example, collision data might have been recorded twice with a different point location because it is recorded at borough boundaries, resulting in counting the crash twice. There are other contributing factors that both data sets do not account for, such as the roadway/intersection geometry, traffic, and pedestrian volume. The dataset lacked some of the driver's information and the vehicle characteristic, and street conditions. With additional information, a more precise, detailed analysis could be carried out in the future.

Discussion and Conclusion

Traffic crash prediction and accuracy modeling are essential and useful tools for planning, engineering, decision-making, and developing roadway safety programs. The analysis was able to identify a pattern of high-risk locations within New York City. By studying historical collision data and deploying many statistical

models to provide accuracy for the number of injured, killed, and contributing factors; by using the following methods Bayes Network, Naïve Bayes, J48, and KNN for the year of 2019 and the 100 intersections with high-risk for collisions derived from the NYPD Collision data using the 2014-2019 dataset. The result of these statistical model Performance classifier showed Naïve Bayes with the highest of 81.59%. Followed by the Bayes Network with a total accuracy of 81.59% and with J48 resulted in an accuracy of 80.81%. KNN performed the lowest with an accuracy of 80.20%.

Additionally, the 100 intersections with the high-risk for collisions were analyzed based on the contributing factors; using the same statistical classifier, Naïve Bayes had the best accuracy level of 59%. The Bayes Network performed at 55% accuracy, followed by J48 of 57%, and KNN with the lowest accuracy level was 45% accuracy. The 100 intersections did not perform as well as the 2019 dataset using the collision's severity.

Furthermore, the study performed a geospatial analysis identifying high-risk locations by using the 100 intersections. The geospatial analysis result was able to identify the following zip code as a problematic area for a diverse type of collisions, as shown in Table 9 below:

BOROUGH	ZIP_CODE
Brooklyn	11234
Brooklyn	11236
Brooklyn	11235
Queens	11354
Queens	11432
Queens	11365
Manhattan	10014
Manhattan	10029
Manhattan	10018
Bronx	10466
Bronx	10458
Bronx	10463
Staten Island	10314
Staten Island	10311
Staten Island	10304

Table 9. High-Risk Area by zip-code

The geospatial analysis results indicated that most collisions occurred within a variety of residential zoning areas. Manhattan is the only borough with most collisions occurring within a manufacturing zone. The roadway pavement condition was analyzed based on the collision that occurred throughout 2019. This research was conducted mainly to identify if the roadway condition is a contributing factor to the collision. The following were the results: the total number of injuries was 55,827, and 3,622 intersecting with a poorly rated pavement condition; 227 total fatalities and 127 intersected with a poorly rated pavement condition.

References

1. Highway Safety Improvement Program. (n.d.). Retrieved from https://safety.fhwa.dot.gov/hsip/resources/fhwasa09029/fhwasa09029.pdf

2. New Yorkers and Their Cars. (n.d.). Retrieved December 01, 2020, from https://edc.nyc/article/new-yorkers-and-their-cars

3. VISION ZERO YEAR FOUR REPORT. (n.d.). Retrieved from https://www1.nyc.gov/assets/visionzero/downloads/pdf/vision-zero-year-4-report.pdf

4. VISION ZERO YEAR SIX REPORT. (n.d.). Retrieved from www1.nyc.gov/assets/visionzero/downloads/pdf/vision-zero-year-6-report.pdf

5. "Table B08201. Household Size by Vehicles Available – Universe: Households". 2009 American Community Survey. United States Census Bureau. Archived from the original on February 11, 2020.

6. Safety Performance Function Development Guide: Developing JurisdictionSpecific SPFs. (n.d.). Retrieved from https://safety.fhwa.dot.gov/rsdp/downloads/spf_development_guide_final.pdf

7. Frank, E. (n.d.). The WEKA Workbench. Retrieved from https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf

8. Vala, K. (2019, March 03). Probabilistic Graphical Models: Bayesian Networks. Retrieved December 01, 2020, from https://towardsdatascience.com/probabilistic-graphical-models-bayesian-networks-d8f0d51b14bf

9. (n.d.). Retrieved December 01, 2020, from https://www.saedsayad.com/decision_tree.htm

10. Abhishek SharmaHe is a data science aficionado. (2020, May 12). Decision Tree vs. Random Forest - Which Algorithm Should You Use? Retrieved December 01, 2020, from https://www.analyticsvidhya.com/blog/2020/05/decision-tree-vs-random-forest-algorithm/

11. D, R. (2019, April 02). K-Nearest Neighbors. Retrieved December 01, 2020, from https://medium.com/@rndayala/k-nearest-neighbors-a76d0831bab0

12. F-Score. (2019, May 17). Retrieved December 01, 2020, from https://deepai.org/machine-learning-glossary-and-terms/f-score

13. Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. (n.d.). Retrieved December 01, 2020, from https://www.tandfonline.com/doi/abs/10.1080/19439962.2016.1152338

14. C. ACI and C. OZDEN, "Predicting the Severity of Motor Vehicle Accident Injuries in Adana-Turkey Using Machine Learning Methods and Detailed Meteorological Data", IJISAE, vol. 6, no. 1, pp. 72-79, Mar. 2018.

15. Theofilatos, A., Chen, C., & Antoniou, C. (2019). Comparing Machine Learning and Deep Learning Methods for Real-Time Crash Prediction. Transportation Research Record, 2673(8), 169–178. https://doi.org/10.1177/0361198119841571

16. Roy, A., Hossain, M., & Muromachi, Y. (2018). Enhancing the Prediction Performance of Real-Time Crash Prediction Models: A Cell Transmission-Dynamic Bayesian Network Approach. Transportation Research Record, 2672(38), 58–68. https://doi.org/10.1177/0361198118797802

17. Alajali, Walaa; Zhou, Wei; Wen, Sheng; Wang, Yu. 2018. "Intersection Traffic Prediction Using Decision Tree Models." Symmetry 10, no. 9: 386.

18. Cuiping Zhang, Xuedong Yan, Lu Ma, Meiwu An, "Crash Prediction and Risk Evaluation Based on Traffic Analysis Zones", Mathematical Problems in Engineering, vol. 2014, Article ID 987978, 9 pages, 2014. https://doi.org/10.1155/2014/987978

19. Iranitalab, Amirfarrokh, and Aemal Khattak. "Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction." Accident Analysis & Prevention, Pergamon, 6 Sept. 2017, www.sciencedirect.com/science/article/pii/S0001457517302865?via=ihub.

20. Lee, J., Abdel-Aty, M., & Cai, Q. (2017, March 21). Intersection crash prediction modeling with macro-level data from various geographic units. Retrieved December 01, 2020, from https://www.sciencedirect.com/science/article/pii/S0001457517301070

21. Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., & Abdel-Aty, M. (2016, June 22). Macro and micro models for zonal crash prediction with application in hot zones identification. Retrieved December 01, 2020, from https://www.sciencedirect.com/science/article/pii/S0966692316303222?casa_token=Pab7jEcIz7MAAAAA%3AYFUfkOEa9tN3NVOoljxtiuHqaQx7Lp3APCveD6hFU3VcRLIPshovL0wmznxNAiW25mIFvw

22. Chen, C., Zhang, G., Huang, H., Wang, J., & Tarefder, R. (2016, August 06). Examining driver injury severity outcomes in rural non-interstate roadway crashes using a hierarchical ordered logit model. Retrieved December 01, 2020, from https://www.sciencedirect.com/science/article/pii/S0001457516302111?casa_token=meCuCJ7kRfEAAAAA%3Ah1AtR5Z0peArwGepOShPHTIxyBecRwiE-GjDyb2cMNg4nwhgXp7VXcMg5PFs–qWDIUHng

23. Elassad, Z., Mousannif, H., & Moatassime, H. (2020, July 10). A real-time crash prediction fusion framework: An imbalance-aware strategy for collision avoidance systems. Retrieved December 01, 2020, from https://www.sciencedirect.com/science/article/pii/S0968090X20306239?via=ihub

24. BYTES of the BIG APPLE. (n.d.). Retrieved December 15, 2020, from https://www1.nyc.gov/site/planning/data-maps/open-data.page

25. (n.d.). Retrieved December 15, 2020, from https://www1.nyc.gov/site/planning/planning-level/topics.page

26. About Zoning. (n.d.). Retrieved December 15, 2020, from https://www1.nyc.gov/site/planning/zoning/about-zoning.page

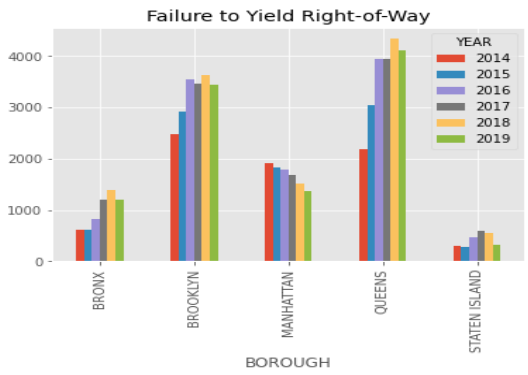
27. Zoning and Land Use Application (ZoLa). (n.d.). Retrieved December 15, 2020, from https://www1.nyc.gov/site/planning/data-maps/zola.page

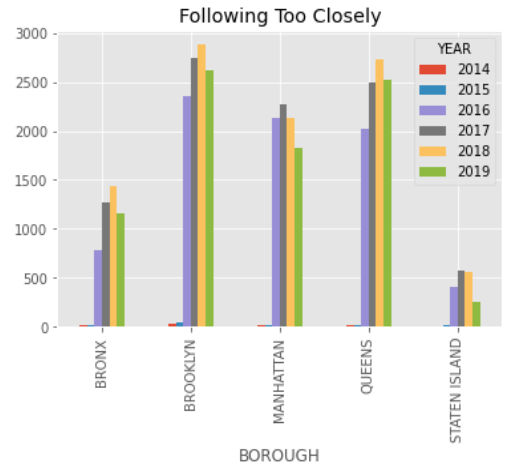
28. Calgary, O. (n.d.). Street Pavement Rating. Retrieved December 15, 2020, from https://data.cityofnewyork.us/Transportation/Street-Pavement-Rating/2cav-chmn

29. Public-Safety. (n.d.). Retrieved from https://data.cityofnewyork.us/Public-Safety/Vision-Zero-View-Data/v7f4-yzyg

30. Motor-Vehicle-Collisions-Crashes. (n.d.). Retrieved from https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

Appendix A





Appendix B.