Understanding and Enhancing Traffic Safety in New York City: A Comprehensive Analysis

Anova Youngers

Table of Contents

Introduction

New York City (NYC) is well known as the city that never sleeps. As the most densely populated city in the United States (Consensus.gov, 2022), it is no wonder that the city is notorious for its traffic congestion. When analyzing traffic collision data in NYC, sobering trends begin to emerge. On average, 149 people are injured or killed every day, and the rate of injuries and fatalities per collision has been rising year over year. These accidents leave an indelible mark on victims and their families. Deciphering the interplay of factors influencing them has become an imperative concern for researchers, policymakers, and advocates of public safety.

Lord and Mannering (2010) aptly noted the urgency of the matter when they asserted that the aspiration to reduce incidences of traffic accidents necessitates a profound understanding of the variables that shape their occurrence. It is this understanding that not only empowers us to predict and potentially prevent such incidents but also guides the development of effective policies and countermeasures to mitigate their impact.

In 1998, the New York City Police Department (NYPD) initiated a new program, TrafficStat. Inspired by similar programs aimed at reducing combat homicides, TrafficStat provides a transparent data feed of accident data gathered by police officers to the public with the aim of reducing fatal traffic accidents. Later in 2014, NYC expanded this data-driven traffic safety initiative through Vision Zero with the goal to reduce fatalities from traffic related accidents. The data is freely available to incentivize analysis of traffic trends in the city and subsequent identification of opportunities to improve traffic safety.

The aim of this study is to use statistical methods to systematically analyze traffic collision data and discern trends in NYC, identify principal contributors to injuries or fatalities, and provide recommendations that can attenuate the toll of traffic related injuries or loss of life. The multifaceted nature of these accidents shaped by location, road conditions, vehicle type, state of the driver, and diverse road users such as pedestrians, cyclists, and motorists, underscores the complicated composition of each individual accident. Our goal is grounded in a comprehensive analysis of the past five years of NYC traffic collision data, with the overarching objective of fostering a safer, more secure urban environment for all its inhabitants.

Dataset Overview

The dataset selected for the study was the July 22, 2023, snapshot of the NYC Motor Vehicle Collision data uploaded to Kaggle. The dataset was obtained by Wierzbicki from TrafficStat data made available from NYC OpenData, which is frequently updated with new traffic statistics over time. Each row represents a traffic collision resulting in an injury, fatality, or significant property damage (greater than $1,000). This is a rich dataset with over 2 million entries of collisions that took place across the 5 boroughs of NYC from January 2014 to July 2023. Across 29 columns, descriptive information about the collision includes the following areas:

- Location (Borough, Zip Code, etc.)
- Timing (Date, Time, etc.)
- Number of injuries, fatalities and whether pedestrian, cyclist, or motor
- Number of vehicles involved and their types/classes.
- Cause of collision associated with each involved vehicle.
- Vehicle type of each involved vehicle

Data Preparation

Data preparation for analysis first involved dropping several columns not used in this study as well as reducing the scope of analysis to only collisions that occurred between 2018 and 2023. Columns dropped mostly included specifics around location of collision such as latitude, longitude, zip code, etc. Collision IDs were also discarded.

Second, any collision entries with unrecorded values for columns of interest were filled with zeroes. Following the fillings of zeroes, four additional columns were added to improve conditional and proportional evaluation of injuries and fatalities:

- *BOOLEAN* 'HAS INJURY' and 'HAS FATALITY' to indicate whether a collision resulted in any of the respective outcomes.

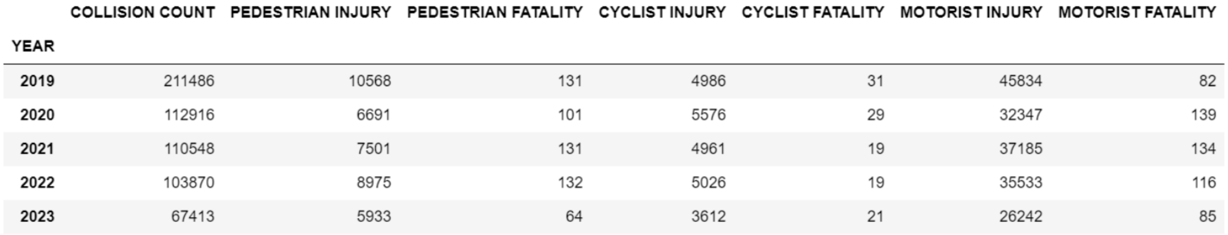
- *BOOLEAN* 'INJURY and FATALITY' if a collision resulted in both outcomes.
- *INT64* 'CARS INVOLVED' for the total number of cars involved in crash based on data present in Vehicle 1-5 related columns.

Finally, to optimize explanatory vehicles for modeling of collision as a response variable, some of the categorical columns with too many subtypes were condensed into overarching categories. Causation factor for each vehicle involved in a collision were condensed into the following subtypes: *Traffic Rules, Nature, Unknown, External, Malfunction, Attention, Substance, Medical, Electronic Device.* Similarly, vehicle types for all vehicles were classified into the following groups: *Car, Truck, Bike, Service Vehicle, Big Rig, Motorbike, Van, Trailer.*

Data Analysis and Modeling Methodology

Injuries and Fatalities

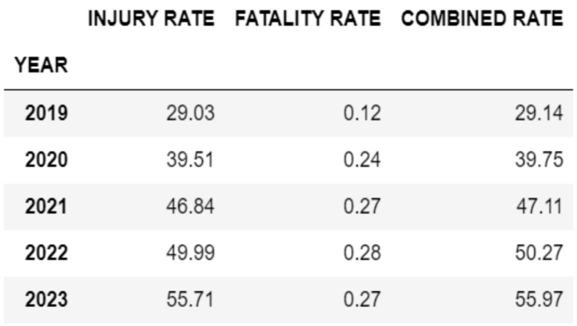
Investigating injuries and fatalities over the past 5 years, a few trends begin to take shape. The collision count has exhibited a decline since 2019, a trend that may be attributed to the extraordinary circumstances brought about by the COVID-19 pandemic. Paradoxically, the statistics reveal that the incidence of injuries and fatalities has not undergone a corresponding reduction, as is graphically illustrated in Table 1.



YEAR	COLLISION COUNT	PEDESTRIAN INJURY	PEDESTRIAN FATALITY	CYCLIST INJURY	CYCLIST FATALITY	MOTORIST INJURY	MOTORIST FATALITY
2019	211486	10568	131	4986	31	45834	82
2020	112916	6691	101	5576	29	32347	139
2021	110548	7501	131	4961	19	37185	134
2022	103870	8975	132	5026	19	35533	116
2023	67413	5933	64	3612	21	26242	85

Table 1: Collision, Injury, and Fatality counts over the last 5 years.

In actuality, the counts associated with injuries and fatalities are experiencing an upward trajectory after an initial decline, with the data from 2023 indicating a trajectory on pace to surpass that of 2022. This phenomenon becomes even more pronounced when calculating the injury and fatality rates based on the number of collisions, unveiling a consistent year-over-year increase, as shown in Table 2. It is evident that although the global pandemic moderated overall traffic volumes, it did not engender a commensurate reduction in the likelihood of injuries or fatalities.

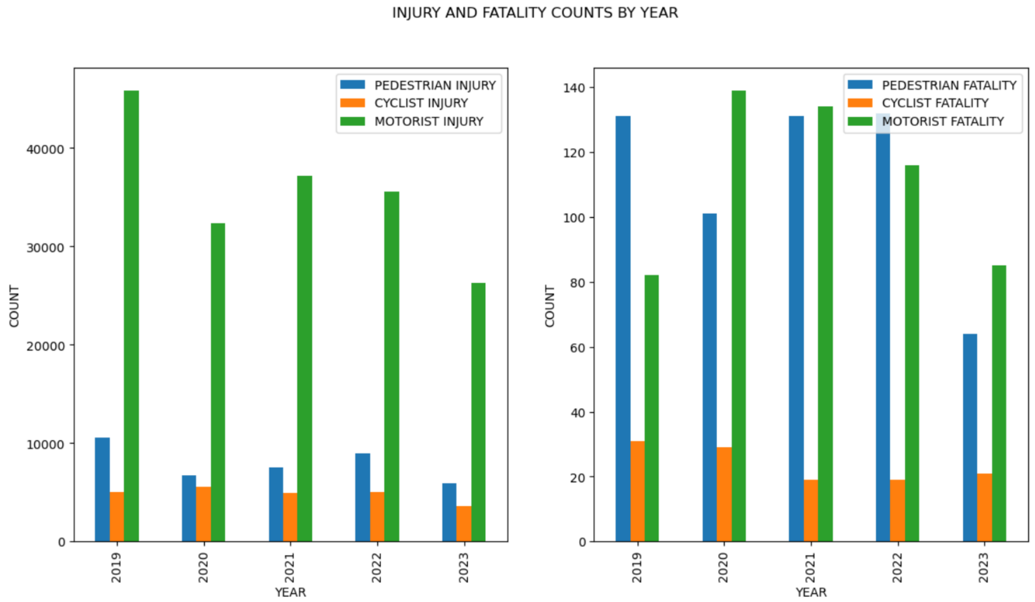


	INJURY RATE	FATALITY RATE	COMBINED RATE
YEAR			
2019	29.03	0.12	29.14
2020	39.51	0.24	39.75
2021	46.84	0.27	47.11
2022	49.99	0.28	50.27
2023	55.71	0.27	55.97

Table 2: Injury and Fatality Rates over the last 5 years

From these findings it can be deduced that the reduction of traffic volume, though it may seem an intuitive solution, may not provide a definitive remedy for augmenting traffic safety. It becomes evident that there exist other underlying factors which exert a more substantial influence on the rates of injuries and fatalities.

Continuing the analysis, an evaluation of injury or fatality rate by the affected road user: pedestrians, cyclists, or motorists, was done. An inspection of the data clearly reveals that motorists constitute the predominant demographic within the injury cases, followed by pedestrians and cyclists. However, when examining the category of fatalities, a starkly different picture emerges, with pedestrians claiming the highest count, closely trailed by motorists. Cyclist fatalities, in contrast, exhibit the lowest numbers, consistently remaining below 40 throughout the past five years (Figure 1). This delineation affords a substantive basis for drawing the conclusion that pedestrian safety merits rigorous scrutiny and assessment, particularly within the broader context of enhancing overall traffic safety.



Collisions by Time

An examination of the collision data across the calendar months reveals distinct patterns, with heightened counts observed during the months of January through March and May through August. It is noteworthy that the counts pertaining to injuries and fatalities remain relatively lower until the onset of the summer months, encompassing May through August (Figure 2). This phenomenon can likely be attributed to a reduced presence of pedestrians and cyclists on the roadways during the winter months. This temporal analysis showcases a salient seasonality in the occurrence of injuries and fatalities, characterized by elevated rates in the summer and diminished rates in the winter.

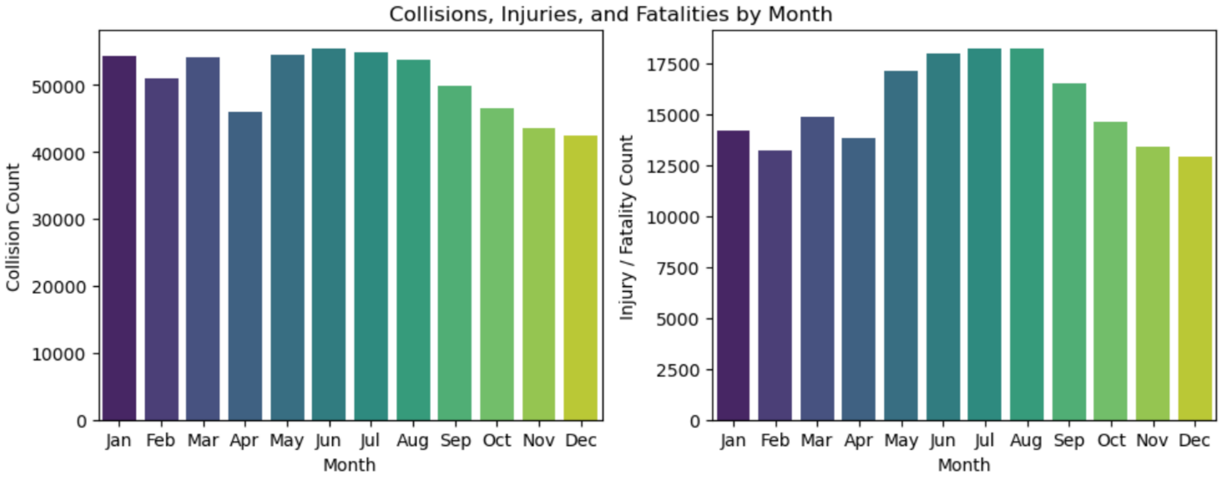


Figure 2: Collision, Injury, and Fatality counts by calendar month.

Conducting a similar investigation concerning the days of the week, a marginal upswing in collision, injury, and fatality counts on Fridays can be observed, while Saturdays and Sundays exhibit comparatively lower counts, as depicted in Figure 3. Given the alignment of the injury and fatality counts with the collision data, it is plausible to infer that the day of the week may exert a limited influence on the injury or fatality rate.

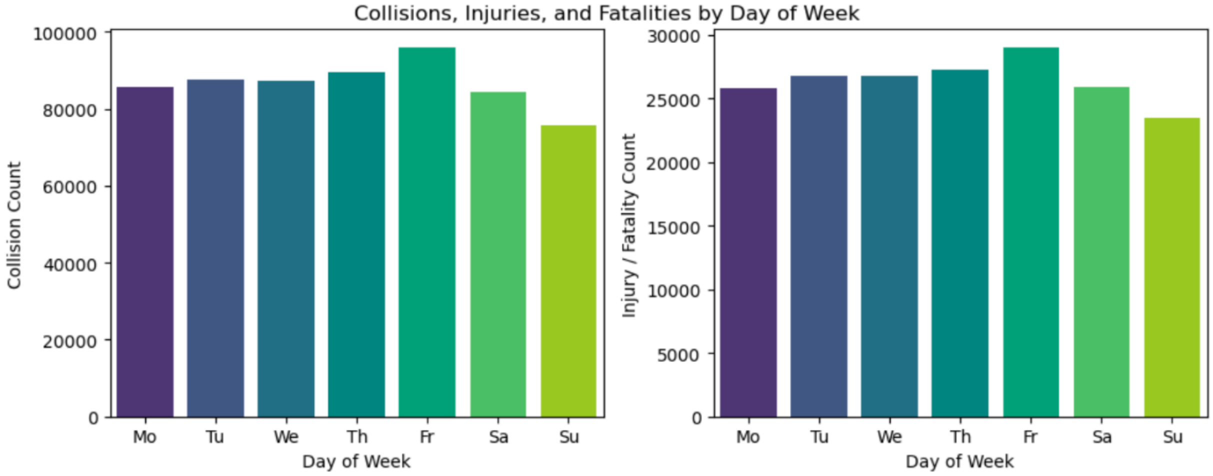


Figure 3: Collision, Injury, and Fatality counts by Day of Week

Lastly, the daytime was partitioned into four distinct time periods: Morning, Afternoon, Evening, and Night. An examination of the collision, injuries, and fatalities data reveals consistent patterns across these categories, except for the Night period, where incidents of injuries and fatalities register notably higher (Figure 4).

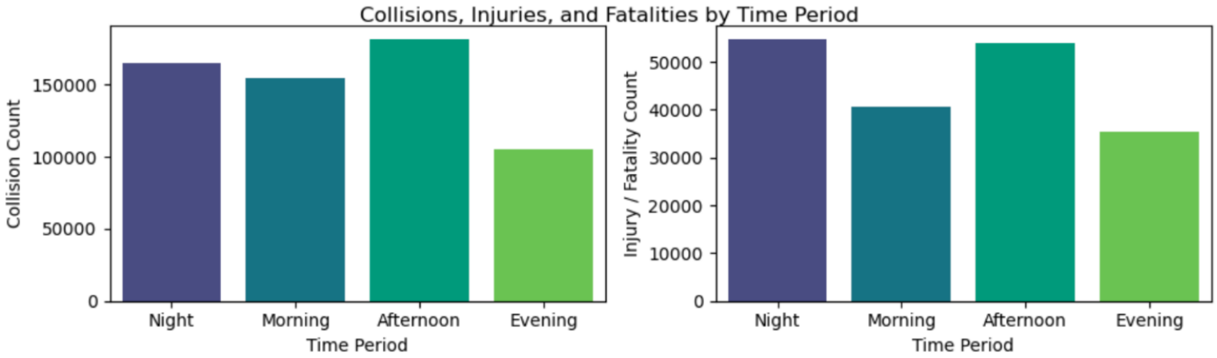


Figure 4: Collision, Injury, and Fatality Throughout 4 Segments of a Day

The crash times throughout the day resembled a gamma distribution pattern. Therefore, a gamma PDF on the crash time results was performed (Figure 5). This shows a left skewed gamma distribution with a mean collision time of 14:21 PM. Evaluating descriptive statistics on this distribution it was concluded that 68% (1-sigma distance from the mean) of collisions occurred 8:30am and 10:30pm, as indicated by the shaded region in the figure below.

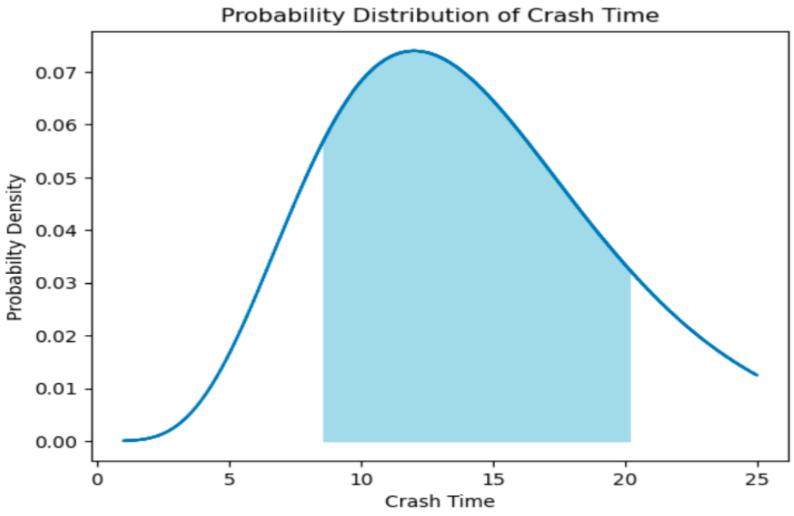


Figure 5: Gamma distribution of crash time throughout the day

Collision Behaviors Across NYC's 5 Boroughs

Figure 6 outlines that Brooklyn accounts for the highest number of collisions (~33% of the total), and Staten Island is the safest borough to drive in, accounting for only ~5% of total collisions in New York City. The evaluation of proportions for deaths and injuries in collisions between different boroughs follows similar patterns. Significantly fewer accidents in Staten Island led to injury and death versus boroughs such as Brooklyn and Queens.

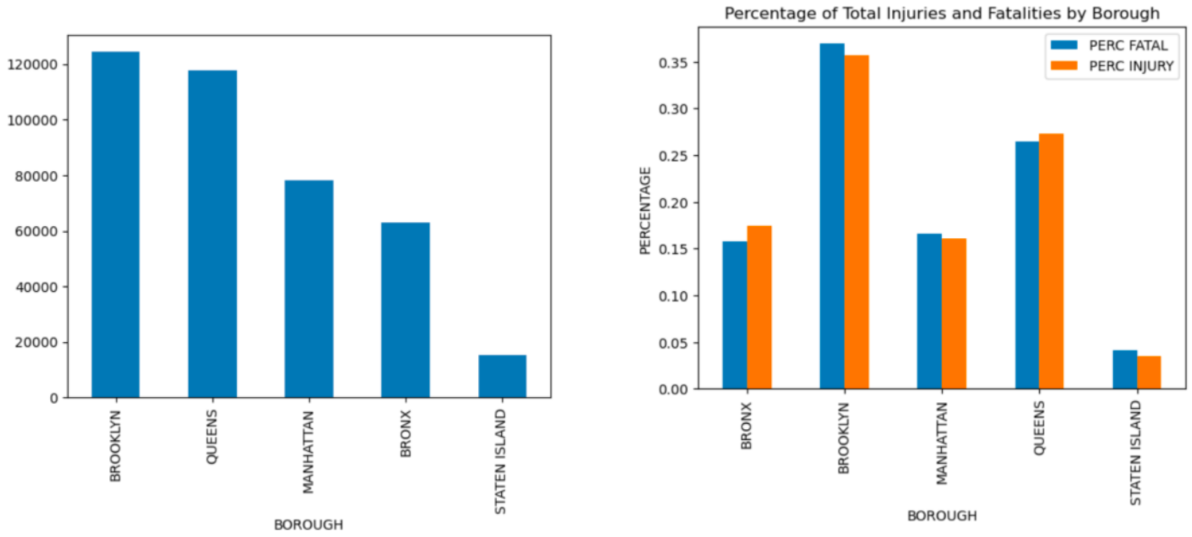


Figure 6: Counts and proportions of injuries and fatalities across different boroughs of NYC

Evaluating statistics of crash times and their distributions between boroughs yielded no significant pattern of differentiation between boroughs. The distribution of crash times across all 5 boroughs of New York City was quite similar. Figure 7 describes how the collision patterns across different boroughs follow the same gamma distribution pattern discussed earlier. One-way f-tests were used to evaluate the significance of differences in collision times between boroughs (Figures 8, 9). With a p-value of 0.00 and an F-test of 22.03, the null hypothesis was rejected, and differences were concluded to be significant. However, when applying a more practical lens to the findings, there is little difference in mean times to conclude with any notable inferences, regardless of statistical significance.

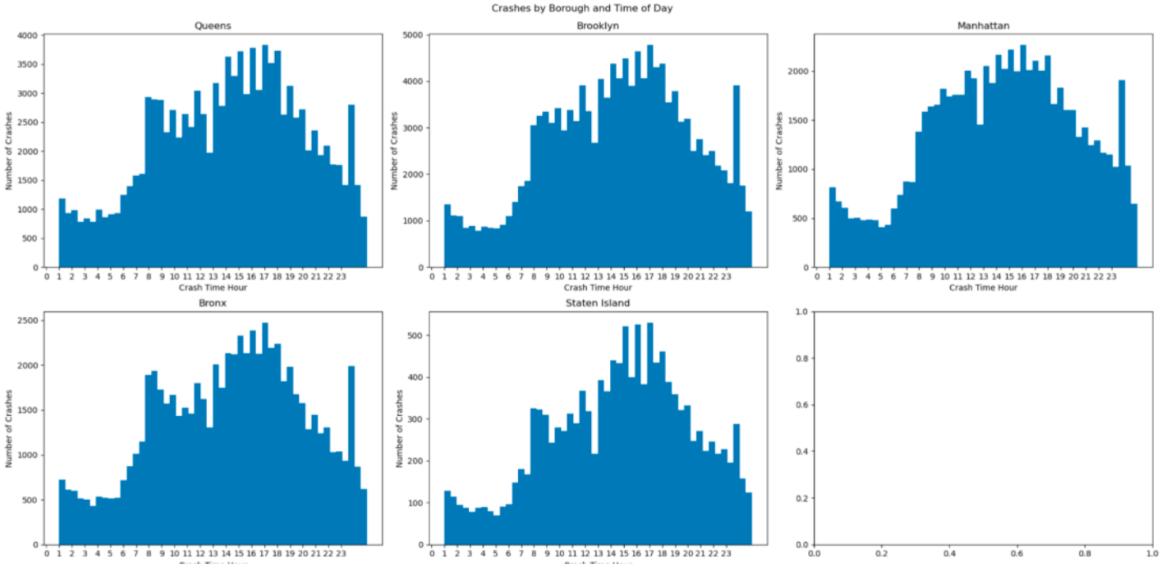
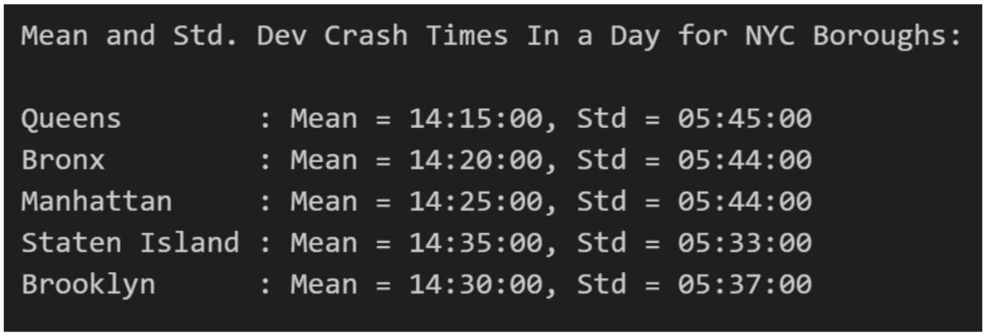


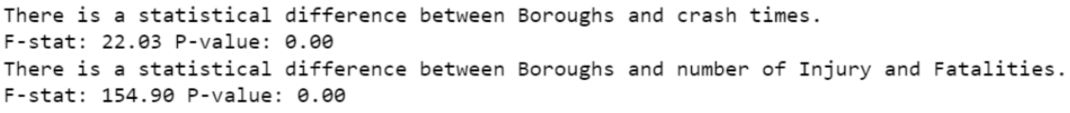
Figure 7: Gamma distribution of collisions throughout the day in different boroughs



```
Mean and Std. Dev Crash Times In a Day for NYC Boroughs:

Queens        : Mean = 14:15:00, Std = 05:45:00
Bronx         : Mean = 14:20:00, Std = 05:44:00
Manhattan     : Mean = 14:25:00, Std = 05:44:00
Staten Island : Mean = 14:35:00, Std = 05:33:00
Brooklyn      : Mean = 14:30:00, Std = 05:37:00
```

Figure 8: Mean and Standard Deviations of Collision Times Across Boroughs



```
There is a statistical difference between Boroughs and crash times.
F-stat: 22.03 P-value: 0.00
There is a statistical difference between Boroughs and number of Injury and Fatalities.
F-stat: 154.90 P-value: 0.00
```

Figure 9: One way f-test results of comparing mean collision time between boroughs

Comparably, the distribution of collision counts across the years of interest also provided similar patterns for each of the boroughs (Figure 10).

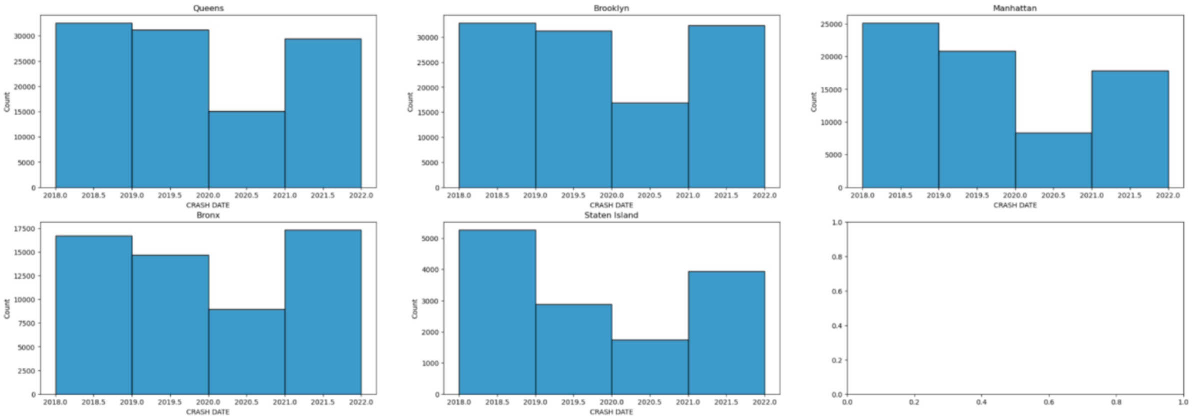


Figure 10: Distribution of Collision Counts Across Years of Interest in NYC's Boroughs

Vehicles Involved

Analyzing the relationship between the number of vehicles involved and the incidence of collisions, injuries, and fatalities, unveils noteworthy patterns, as exemplified in Figure 11.

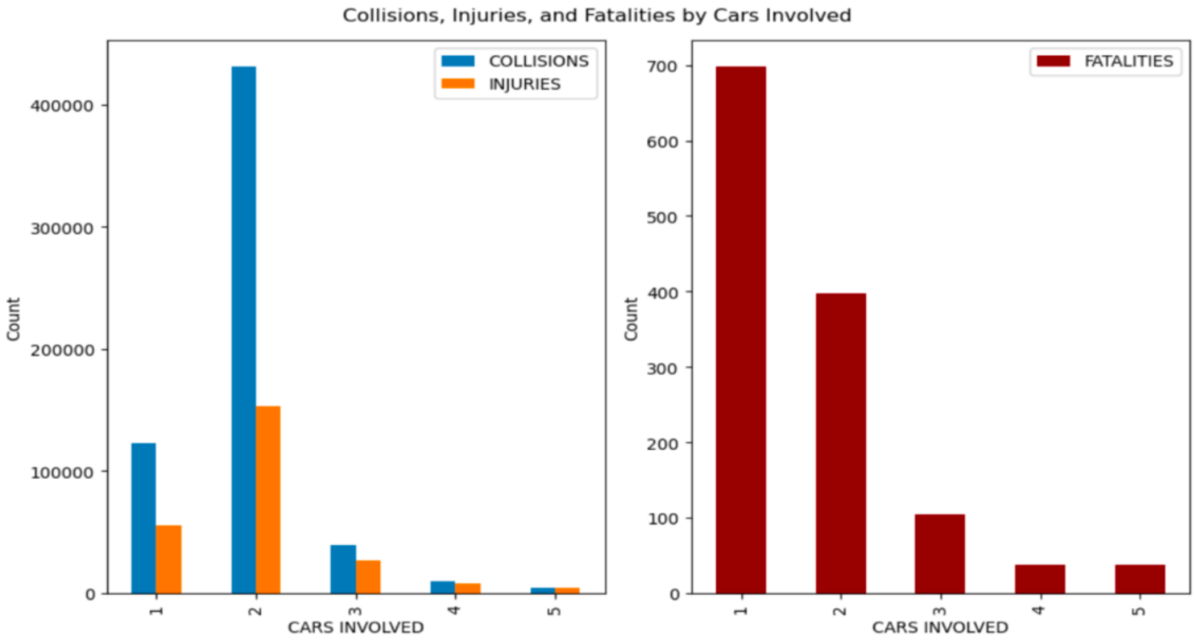


Figure 11: Bar Graph detailing Collision, Injury, and Fatality counts by Number of Cars Involved

The frequency distribution of collision and injury counts, stratified by the number of vehicles involved, reveals a right-skewed distribution. Notably, the median value of vehicles involved stands at two. Additionally, the frequency distribution of fatalities, contingent on the number of vehicles involved, conforms to an exponential distribution, with a median count of one vehicle. Employing a t-test to scrutinize the relationship between the number of vehicles involved and the figures for injuries and fatalities, it can be ascertained a high level of statistical significance, underscored by a p-value of .00 (below .01), validating the robustness of this variable's impact on collision outcomes.

Vehicle Contributing Factors

The collision dataset presents a valuable repository of information concerning the causal factors attributed to each vehicle involved in collisions. To facilitate a comprehensive analysis of these factors, they were aggregated into a unified column named "Contributing Factors". While this dataset encompasses a plethora of variables, distinct patterns emerge, as showcased in Figure 12. Notably, a select few factors emerge as prominent leaders in their influence on collision outcomes. The top three uncategorized contributing factors associated with both collisions and injuries are: "Driver Inattention/Distraction," "Following Too Closely," and "Failure to Yield Right-of-Way." The factors contributing to high fatality counts exhibit a slight variation, with "Unsafe Speed" ranking as the most influential. This points to a key insight into speeding and its correlation to fatal outcomes. One way to make streets safer may be to reduce speeding infractions.

When charting the incidence of collisions, injuries, and fatalities across aggregated or simplified categories, consistent trends come to the forefront, as illustrated in Figure 13. It becomes evident that infractions related to "Traffic Rules" overwhelmingly dominate all three metrics, exhibiting a substantial lead.

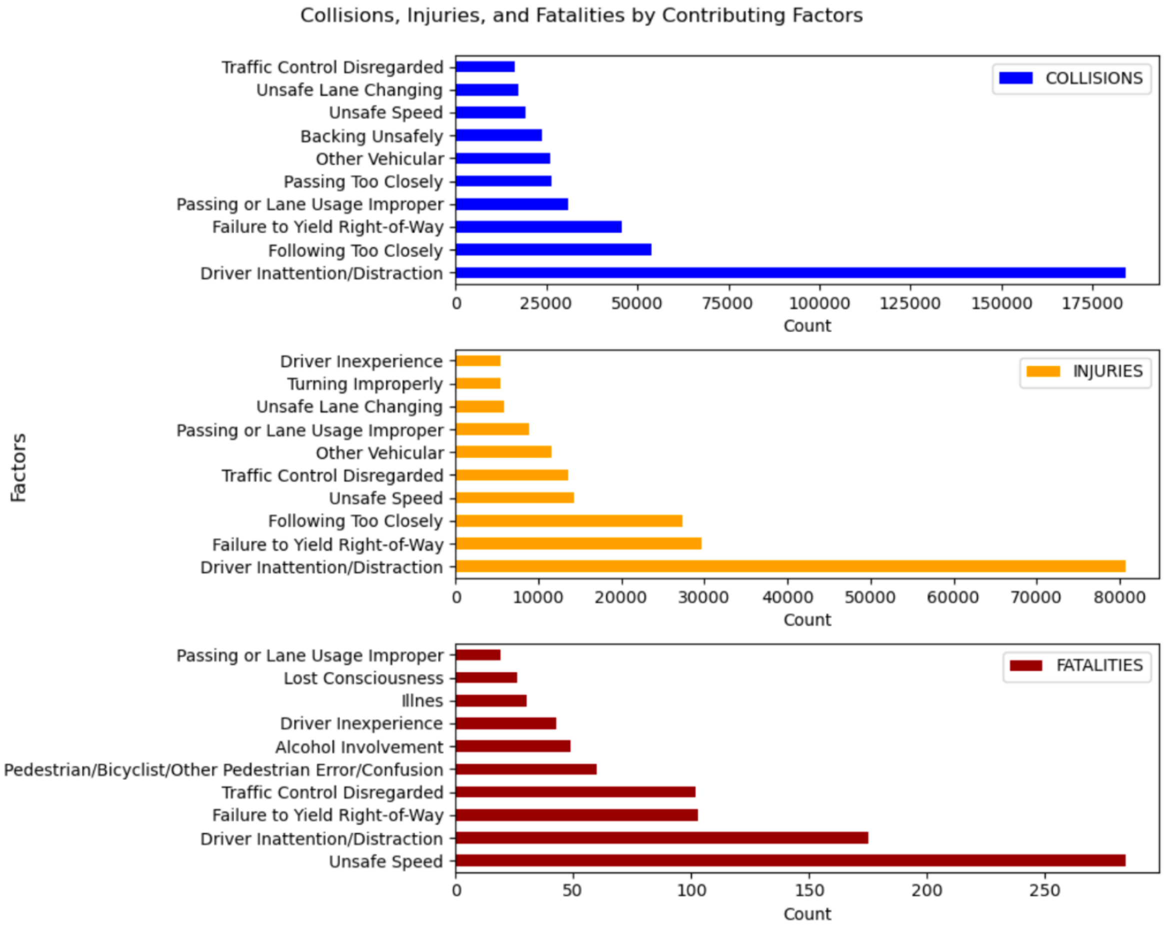
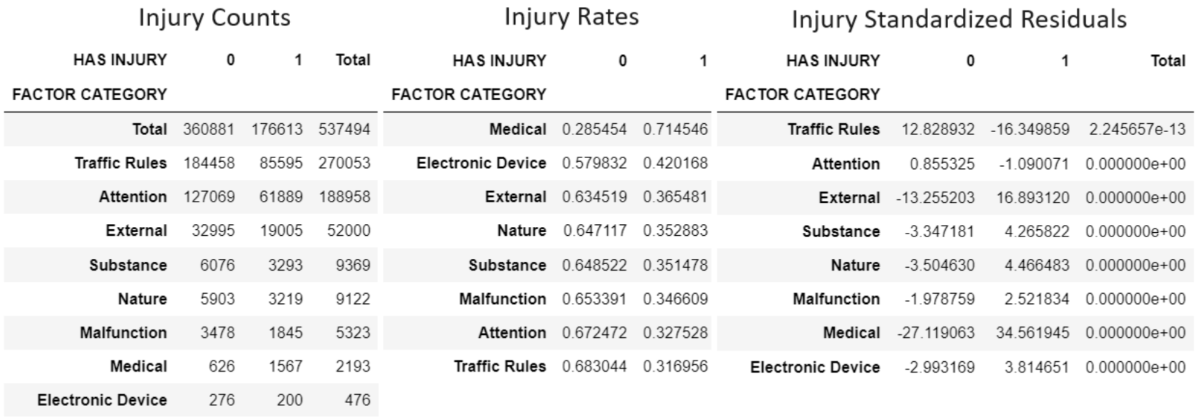


Figure 12: Bar Graph detailing Collision, Injury, and Fatality counts by True (Unaggregated) Causation Factors

To further expound upon the relevance of these categories in relation to injury and fatality rates, cross-tabulation tables were crafted to delineate counts, categories, and standardized residuals, contingent on causation factor categories. As evidenced in Figure 13, the categories of "Medical" and "Electronic Device" notably elevate the injury rate, despite contributing minimally to the overall collision count. By computing the Standardized Residuals for injuries, it is apparent that "Medical" and "External Factors" exhibit injury rates higher than would be anticipated under the assumption of equal injury rates across categories.



Injury Counts

HAS INJURY	0	1	Total
FACTOR CATEGORY			
Total	360881	176613	537494
Traffic Rules	184458	85595	270053
Attention	127069	61889	188958
External	32995	19005	52000
Substance	6076	3293	9369
Nature	5903	3219	9122
Malfunction	3478	1845	5323
Medical	626	1567	2193
Electronic Device	276	200	476

Injury Rates

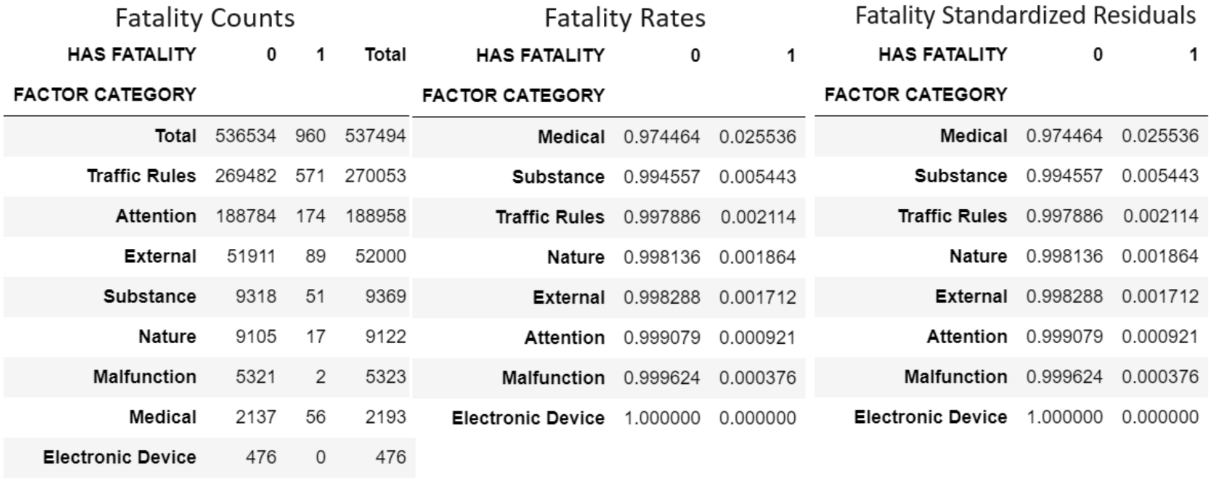
HAS INJURY	0	1
FACTOR CATEGORY		
Medical	0.285454	0.714546
Electronic Device	0.579832	0.420168
External	0.634519	0.365481
Nature	0.647117	0.352883
Substance	0.648522	0.351478
Malfunction	0.653391	0.346609
Attention	0.672472	0.327528
Traffic Rules	0.683044	0.316956

Injury Standardized Residuals

HAS INJURY	0	1	Total
FACTOR CATEGORY			
Traffic Rules	12.828932	-16.349859	2.245657e-13
Attention	0.855325	-1.090071	0.000000e+00
External	-13.255203	16.893120	0.000000e+00
Substance	-3.347181	4.265822	0.000000e+00
Nature	-3.504630	4.466483	0.000000e+00
Malfunction	-1.978759	2.521834	0.000000e+00
Medical	-27.119063	34.561945	0.000000e+00
Electronic Device	-2.993169	3.814651	0.000000e+00

Figure 13: Injury Counts, Injury Rates, and Injury Standardized Residuals by Cause Factor Category

Collisions with fatality reflect similar tendencies, with "Medical" and "Substance" factors registering as the highest contributors to fatality rates(Figure 14). Concurrently, the standardized residuals for "Medical" and "Substance" underscore elevated fatality rates, surpassing the expected values.



Fatality Counts

HAS FATALITY	0	1	Total
FACTOR CATEGORY			
Total	536534	960	537494
Traffic Rules	269482	571	270053
Attention	188784	174	188958
External	51911	89	52000
Substance	9318	51	9369
Nature	9105	17	9122
Malfunction	5321	2	5323
Medical	2137	56	2193
Electronic Device	476	0	476

Fatality Rates

HAS FATALITY	0	1
FACTOR CATEGORY		
Medical	0.974464	0.025536
Substance	0.994557	0.005443
Traffic Rules	0.997886	0.002114
Nature	0.998136	0.001864
External	0.998288	0.001712
Attention	0.999079	0.000921
Malfunction	0.999624	0.000376
Electronic Device	1.000000	0.000000

Fatality Standardized Residuals

HAS FATALITY	0	1
FACTOR CATEGORY		
Medical	0.974464	0.025536
Substance	0.994557	0.005443
Traffic Rules	0.997886	0.002114
Nature	0.998136	0.001864
External	0.998288	0.001712
Attention	0.999079	0.000921
Malfunction	0.999624	0.000376
Electronic Device	1.000000	0.000000

Figure 14: Injury Counts, Injury Rates, and Injury Standardized Residuals by Cause Factor Category

To assess the statistical significance of these categorizations on injury and fatality outcomes, a chi-squared test was executed, with the Null hypothesis of causation category factors playing no role in injury or fatality right. This Null hypothesis was rejected with a p-value of 0.000, affirming a significant statistical relationship between causation categories and collision outcomes.

Vehicle Types

A thorough examination of collision counts, injuries, and fatalities with respect to unaggregated vehicle types reveals conspicuous spikes in counts, particularly evident among the top two or three values (Figure 15). Sedans emerge as the prevailing vehicle type across all three metrics, closely followed by station wagons and S.U.Vs. This pattern can be attributed, in part, to the prevalence of these vehicle types on the road. A notable anomaly within the data is the ascendance of "Motorcycle" as the third-largest category of vehicles involved in fatal collisions, signifying an area where targeted measures may be necessary to enhance road safety and mitigate fatalities.

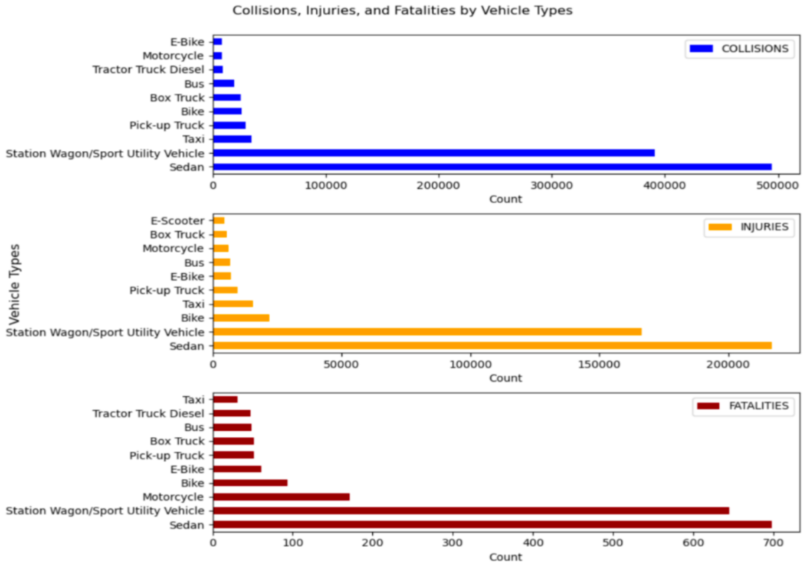
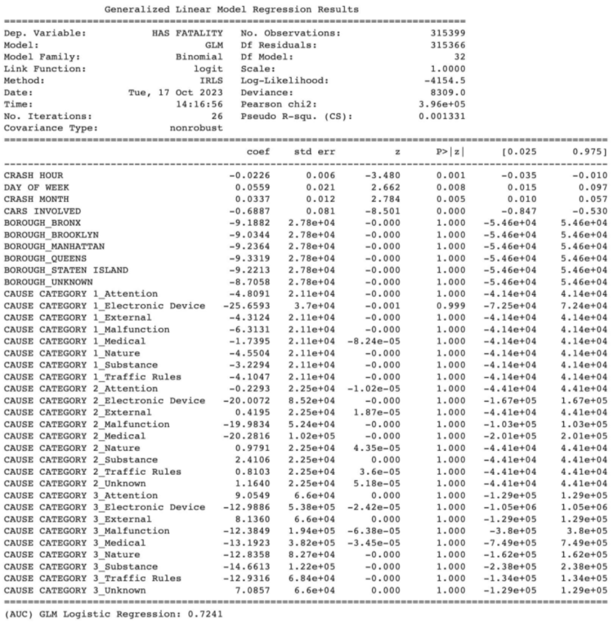


Figure 15: Collision, Injury, and Fatality counts by Vehicle Types

Regrettably, no discernible new insights materialized from charting aggregated/condensed vehicle type categories against collision counts, injuries, and fatalities. To gauge the statistical significance of the vehicle type category in relation to injury and fatality rates, a chi-squared test was executed. Not surprisingly, the outcome yielded a p-value of 0.00, affirming the unequivocal statistical significance of the vehicle type in relation to injury and fatality rates.

Predicting Fatalities Using Statistical Modeling

An attempt was made to use statistical modeling in determining the likelihood of an injury or fatality given the independent variables of time, borough, vehicle causation factor, vehicle type, and number of vehicles involved (Figure 16). Separate modeling was performed for injuries and fatalities, using the logit link function along with the Logistic Regression family for a GLM model.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:           HAS FATALITY   No. Observations:               315399
Model:                            GLM   Df Residuals:                   315366
Model Family:                Binomial   Df Model:                           32
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -4154.5
Date:                Tue, 17 Oct 2023   Deviance:                       8309.0
Time:                        14:16:56   Pearson chi2:                 3.96e+05
No. Iterations:                    26   Pseudo R-squ. (CS):           0.001331
Covariance Type:            nonrobust
====================================================================================================
                                         coef    std err          z      P>|z|     [0.025     0.975]
----------------------------------------------------------------------------------------------------
CRASH HOUR                            -0.0226      0.006     -3.480      0.001     -0.035     -0.010
DAY OF WEEK                            0.0559      0.021      2.662      0.008      0.015      0.097
CRASH MONTH                            0.0337      0.012      2.784      0.005      0.010      0.057
CARS INVOLVED                         -0.6887      0.081     -8.501      0.000     -0.847     -0.530
BOROUGH_BRONX                         -9.1882   2.78e+04     -0.000      1.000  -5.46e+04   5.46e+04
BOROUGH_BROOKLYN                      -9.0344   2.78e+04     -0.000      1.000  -5.46e+04   5.46e+04
BOROUGH_MANHATTAN                     -9.2364   2.78e+04     -0.000      1.000  -5.46e+04   5.46e+04
BOROUGH_QUEENS                        -9.3319   2.78e+04     -0.000      1.000  -5.46e+04   5.46e+04
BOROUGH_STATEN ISLAND                 -9.2213   2.78e+04     -0.000      1.000  -5.46e+04   5.46e+04
BOROUGH_UNKNOWN                       -8.7058   2.78e+04     -0.000      1.000  -5.46e+04   5.46e+04
CAUSE CATEGORY 1_Attention            -4.8091   2.11e+04     -0.000      1.000  -4.14e+04   4.14e+04
CAUSE CATEGORY 1_Electronic Device   -25.6593    3.7e+04     -0.001      0.999  -7.25e+04   7.24e+04
CAUSE CATEGORY 1_External             -4.3124   2.11e+04     -0.000      1.000  -4.14e+04   4.14e+04
CAUSE CATEGORY 1_Malfunction          -6.3131   2.11e+04     -0.000      1.000  -4.14e+04   4.14e+04
CAUSE CATEGORY 1_Medical              -1.7395   2.11e+04  -8.24e-05      1.000  -4.14e+04   4.14e+04
CAUSE CATEGORY 1_Nature               -4.5504   2.11e+04     -0.000      1.000  -4.14e+04   4.14e+04
CAUSE CATEGORY 1_Substance            -3.2294   2.11e+04     -0.000      1.000  -4.14e+04   4.14e+04
CAUSE CATEGORY 1_Traffic Rules        -4.1047   2.11e+04     -0.000      1.000  -4.14e+04   4.14e+04
CAUSE CATEGORY 2_Attention            -0.2293   2.25e+04  -1.02e-05      1.000  -4.41e+04   4.41e+04
CAUSE CATEGORY 2_Electronic Device   -20.0072   8.52e+04     -0.000      1.000  -1.67e+05   1.67e+05
CAUSE CATEGORY 2_External              0.4195   2.25e+04   1.87e-05      1.000  -4.41e+04   4.41e+04
CAUSE CATEGORY 2_Malfunction         -19.9834   5.24e+04     -0.000      1.000  -1.03e+05   1.03e+05
CAUSE CATEGORY 2_Medical             -20.2816   1.02e+05     -0.000      1.000  -2.01e+05   2.01e+05
CAUSE CATEGORY 2_Nature                0.9791   2.25e+04   4.35e-05      1.000  -4.41e+04   4.41e+04
CAUSE CATEGORY 2_Substance             2.4106   2.25e+04      0.000      1.000  -4.41e+04   4.41e+04
CAUSE CATEGORY 2_Traffic Rules         0.8103   2.25e+04    3.6e-05      1.000  -4.41e+04   4.41e+04
CAUSE CATEGORY 2_Unknown               1.1640   2.25e+04   5.18e-05      1.000  -4.41e+04   4.41e+04
CAUSE CATEGORY 3_Attention             9.0549    6.6e+04      0.000      1.000  -1.29e+05   1.29e+05
CAUSE CATEGORY 3_Electronic Device   -12.9886   5.38e+05  -2.42e-05      1.000  -1.05e+06   1.05e+06
CAUSE CATEGORY 3_External              8.1360    6.6e+04      0.000      1.000  -1.29e+05   1.29e+05
CAUSE CATEGORY 3_Malfunction         -12.3849   1.94e+05  -6.38e-05      1.000   -3.8e+05    3.8e+05
CAUSE CATEGORY 3_Medical             -13.1923   3.82e+05  -3.45e-05      1.000  -7.49e+05   7.49e+05
CAUSE CATEGORY 3_Nature              -12.8358   8.27e+04     -0.000      1.000  -1.62e+05   1.62e+05
CAUSE CATEGORY 3_Substance           -14.6613   1.22e+05     -0.000      1.000  -2.38e+05   2.38e+05
CAUSE CATEGORY 3_Traffic Rules       -12.9316   6.84e+04     -0.000      1.000  -1.34e+05   1.34e+05
CAUSE CATEGORY 3_Unknown               7.0857    6.6e+04      0.000      1.000  -1.29e+05   1.29e+05
====================================================================================================
(AUC) GLM Logistic Regression: 0.7241
```

Figure 16: Binomial Generalized Linear Model (GLM) for 'HAS FATALITY' as a Response Variable

The Area Under the Curve (AUC) for the Injury prediction model was 0.7475, and 0.7241 for fatalities. This indicates a relatively good ability of the model to predict likelihood of an injury or fatality because of a collision compared to .5 representing random guessing.

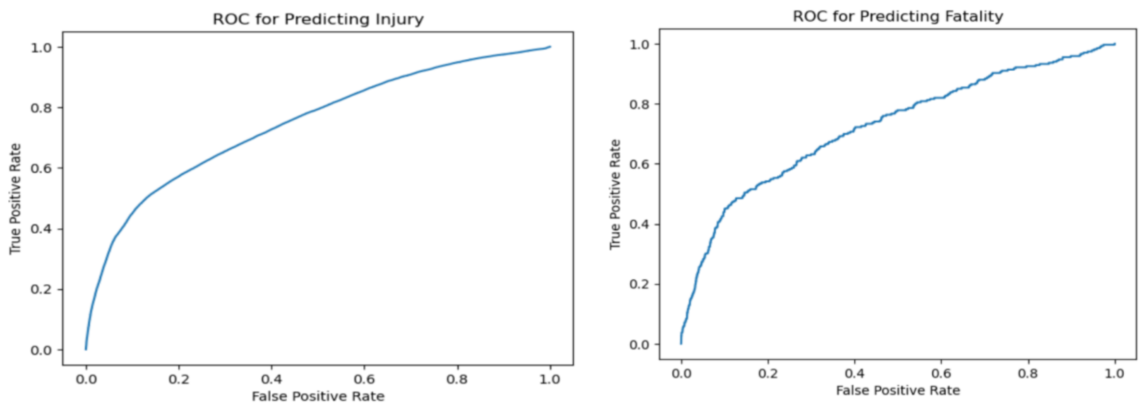


Figure 17: Area Under Curve (AUC) for Prediction of Injury and Fatality via Binomial GLM

Both models produced large coefficients and standard errors, along with p-values of close to 1 for most categorical explanatory variables used in the model. Excluding crash day, crash month and cause category of crash car 2, most other coefficients had a negative influence on possibility of injury and death. While most categorical coefficients were negative, a few stood out as having a large positive or negative impact. Electronic Devices had a large dampening effect on determining fatality pointing to the unlikely event that collisions with these cause factors result in a death. Attention, External, and Substance coefficients with the second vehicle echo similar findings in the statistical analysis, which points them being higher contributors to fatality rates.

The extreme coefficient values obtained suggest there may be multicollinearity or overfitting. The resultant AUC was found to be acceptable, and a variation of Simpson's paradox was observed where the data coefficients exhibited large deviation from the model described above.

Additional modeling experiment and analysis is required to determine the validity of the devised model and whether it can be used to accurately predict the future possibility of an injury or death in a collision in NYC. Furthermore, precision, recall and F-1 scores could also be leveraged to assess the overall fit of the model.

Conclusion

The NYC TrafficStat program, which offers open access to comprehensive traffic statistics, has paved the way for researchers to glean valuable insights and put forth strategies aimed at alleviating traffic-related issues within the city. Moreover, it presents an avenue to curtail the incidence of injuries and fatalities in traffic-related incidents.

Analysis of collisions involving injuries and fatalities in NYC between 2019 and 2023 suggest that reduction in collisions counts over the pandemic did not result in a corresponding reduction in injuries and fatalities over the same time span. On the contrary, the incidence of injuries and fatalities exhibited an upward trajectory, and the correlation of these incidents does not solely rely on traffic volume.

Examination of descriptive statics of the data shed light on the trends of collisions across different spheres of study. It was determined that seasonality (higher rates of accidents in summer months), time of day, and crash contributing factors played a role in shaping the trends of collisions. While collision incidents were not evenly distributed across New York City's five boroughs, the patterns and causality behaviors of collisions within each borough were quite similar.

Analyzing the contributing factors to collisions further enriched accident understandings. Key contributors such as unsafe speeds, driver inattention/distraction, following too closely, and failure to yield right-of-way were identified as frequent contributing factors These factors also point to potential aggressive driving behavior from NYC's residents. This study also delved into the types of vehicles involved in collisions, injuries, and fatalities. Car collisions were by far the most prevalent, but motorcycle fatalities were a statistically significant contributor. Insights like this merit deeper investigation and suggest the necessity for tailored strategies, targeting specific vehicle categories, to enhance road safety effectively.

The modeling used in this study provided essential insights into traffic collision fatalities and injuries in New York City, although they exhibited areas requiring improvement and refinement. The Logistic Regression model demonstrated a reasonable discriminative ability with an AUC of 0.7241 but showed signs of potential overfitting and complexity. For further model enhancement, a focused approach involving feature engineering, handling of imbalanced classes, and feature simplification is recommended. These refinements are crucial for improving the predictive capabilities of the models, facilitating a more effective analysis and mitigation of traffic collisions.

Based on the comprehensive analysis conducted across various dimensions of traffic collisions in New York City, and the insights gleaned from data modeling, several pragmatic strategies can be recommended to mitigate the frequency and severity of traffic incidents. Firstly, deployment targeted driver safety and defensive driving campaigns to address predominant contributing factors such as driver attention and failure to obey traffic rules. Stricter enforcement of traffic rules, coupled with technological solutions like speed cameras, could also be pivotal in curbing unsafe driving practices. Given the identified vulnerabilities of specific road users, particularly pedestrians and motorcyclists, specialized safety measures such as improved crosswalk designs, pedestrian barriers, and dedicated motorcycle lanes could help reduce fatalities in these populations.

Seasonality and timing-related insights also advocate for adaptive traffic management strategies, such as variable message signs and dynamic speed limits, to tailor traffic flow regulations in alignment with changing risk profiles throughout the day and year. Furthermore, geographical analysis underscores the necessity for area-specific safety enhancements, such as bolstered traffic calming measures in high-incidence boroughs like Brooklyn.

In leveraging the predictive capabilities of statistical models and insights, a data-driven approach should be adopted for continuous refinement and optimization of traffic safety strategies based on emerging trends and patterns. This will facilitate a responsive and adaptive traffic safety management paradigm to enhance the well-being of all road users in New York City.