Statistical Analysis on New York Motor Vehicle Collisions

Abstract

Motor vehicle collisions cause an average of 38,000 deaths each year and an additional 2.35 million are injured or disabled. There were 33,654 fatal motor vehicle crashes in the United States in 2018 in which 36,560 deaths occurred, resulting in 11.2 deaths per 100,000 people. With such a high toll on human life, we wanted to understand what are the different factors that cause injuries during motor vehicle collisions. We wanted to analyze and present insights that would identify high-probability factors causing injuries during collisions and raise awareness about the next steps that can be taken to reduce the damage from them. For our study, we chose the NYC Motor Vehicle collision dataset covering a rich dataset of all the collisions that occurred in New York City in 2019. We probed deeper into the data and set up statistical tests to understand how differences in crash times, weather conditions, seasonality and driver's alcohol and drug intake influence an injury developing from vehicular collisions. Further, we have represented a comprehensive summary describing the associations of various factors in causing an injury from a collision. Using a logistic regression model at its core, we have substantiated our results. Finally, we have presented our learnings from the analysis and highlighted the scope and future potentials of our work.

Introduction

In order to comprehend the underlying factors in motor vehicle collisions, we chose New York City's Open Data collisions dataset. Our initial plan was to conduct the study for Seattle city so that our learnings and insights are more relatable but owing to less rich data availability, we opted for NYC data. Courtesy to the initiative Vision Zero, there was a greater emphasis on the collection of more traffic data and work towards eliminating traffic fatalities. Consequently, our data spanned for a longer time frame and also covered a lot many variables in comparison to the Seattle collisions dataset.

What are the factors that affect if a collision results in an injury in New York City? Do different boroughs affect whether an injury occurs out of a collision? Are collisions more injurious during the rainy season than the summer season? Does the NYC traffic at different times of the day also impact if collisions result in injuries? We started with much curiosity and wanted to get deeper into the problem of understanding how motor vehicle collisions are occurring for New York City in 2019. Our motive behind this study was to prepare a firm base of learnings and insights. We aimed at understanding the association of different factors such as crash time, season and weather conditions, contributing factors such as driving under the influence and other factors leading to injury from a collision.

Following thorough discussions, we implemented four distinct tests on our dataset that revolved around the following goals:

- What is the impact of time of the day on collisions resulting in an injury and how does it vary across the five different boroughs of New York City?
- What is the impact of seasons and weather components namely snowfall, precipitation and visibility on a collision causing an injury across the five boroughs?
- What is the impact of DUI (*Driving under the influence*) on a collision leading to an injury?
- How significant different factors such as location (borough), time of the day, seasonality, weather, contributing causes, vehicles types, and road types become in causing an injury when a collision has happened

With the above motive, the response that we have studied throughout is whether a collision had an injury or not. Being a binary variable, we opted for logistic regression for verifying our learnings. We performed missing value treatment and feature engineering on the data to preserve the maximum of our information. We then did an exploratory data analysis to highlight the factual statistics of data and examine the distribution of our key variables. The remainder of the paper discusses our data, hypotheses, statistical methods used, results and conclusions.

Dataset Description

Primary dataset - Motor Vehicle Collision Data – Crashes (2019)

The Motor Vehicle Collision data on crashes is an observational data obtained from NYC Open Data. It contains records of crash events that have occurred in NYC. We obtained a subset of data for the year 2019. The following are the key attributes from the dataset

Attribute	Description
Collision Id (Primary Key)	Unique id of a crash event
Crash Date	Date of the crash
Crash Time	Time of occurrence of a crash
Borough	NYC Borough name in which crash occurred: Manhattan, Bronx, Queens, Brooklyn, Staten Island
Zip Code	The zip code corresponding to the crash location
On-Street Name	Street name where the crash occurred
Off-Street Name	Street name where the crash occurred off-street
Longitude, Latitude	Geolocation corresponding to a crash address
Number of People Injured	Total number of people injured in the crash if any
Number of People Killed	Total number of people killed in the crash if any

Contributing Factor	Factors that contributed to the cause of the crash such as Driver Inattention, Turning Improperly, etc.
Vehicle Type	Type of vehicle involved in a crash such as Sedan, Station Wagon, etc.

Secondary dataset Motor Vehicle Collision Data – Persons (2019)

The motor vehicle collision data on persons is obtained from NYC open data and contains information on people involved in crashes. The following are the key attributes from the dataset

Attribute	Description
Victim Id(Primary Key)	Id of the person involved in the crash
Victim Type	Occupant/Bicyclist/Pedestrian involved in a crash
Collision Id	Collision Id corresponding to the accident

Secondary dataset NOAA Daily Weather Data (2019)

Weather data is obtained from the National Oceanic and Atmospheric Administration (NOAA), which contains information on daily weather summaries. The following are the key attributes from the dataset

Attribute	Description
Date (Primary Key)	Date of the observed weather information
Precipitation	Precipitation observed (in inches)
Snowfall	Amount of snowfall that occurred (in inches)
Fog	Indicates the presence of fog/heavy fog
Smoke or haze	Indicates the presence of smoke or haze
Blowing snow	Indicates the presence of blowing snow

The following are some of the key steps that we have taken to address inconsistencies and prepare the dataset for statistical analysis

- **Missing value treatment** - For 2019, 35% of collisions had missing borough information. To explore the variation of collisions across different boroughs, we required borough information and dropping one-third of the collisions didn't seem feasible as it would mean a significant information loss.

Solution - For collisions where we had the latitude-longitude information but missing boroughs, we performed reverse geocoding to obtain the corresponding borough information. This resulted in a drop in missing boroughs from 35% to 6%. Considering the number of collisions, we proceeded by dropping the ~6% of observations for our analyses.

- **Data collation -** The analysis requires us to use attributes from different datasets. For eg, to analyze the impact of environmental factors on collision, the corresponding weather attributes need to be denoted for each collision.
  Solution - The three datasets were combined into a single dataset using collision ID as a primary key. The Weather and Persons data of crashes are combined using the Collision Id. The weather data is combined using the date of the crash. The dataset aggregated contained 197,855 records of which 47% of the collisions resulted in injury.

- **Increasing the Scope of the Categorical Data -** The categorical features such as contributing vehicular factors, vehicle types and street names had a lot of levels and as such cannot be used in the models as it would lead to a large number of factors
  Solution - On-street and off-street name attributes were combined into a single attribute called Road Type. Based on the number of injuries, top 15 vehicular factors, top 10 vehicle types and top 5 road types were retained and the others were grouped together

- **Increasing the Scope of the Cyclical Data -** The 24 hours time period of the day and the months of the year as such are very granular and may not present tangible inferences.
  Solution - In order to improve the scope of cyclical data, the time of the day is grouped into 4 time segments and months were tagged to one of the four seasons.

Exploratory Data Analysis

Below are some plots used in our exploratory data analysis that informed our analyses.

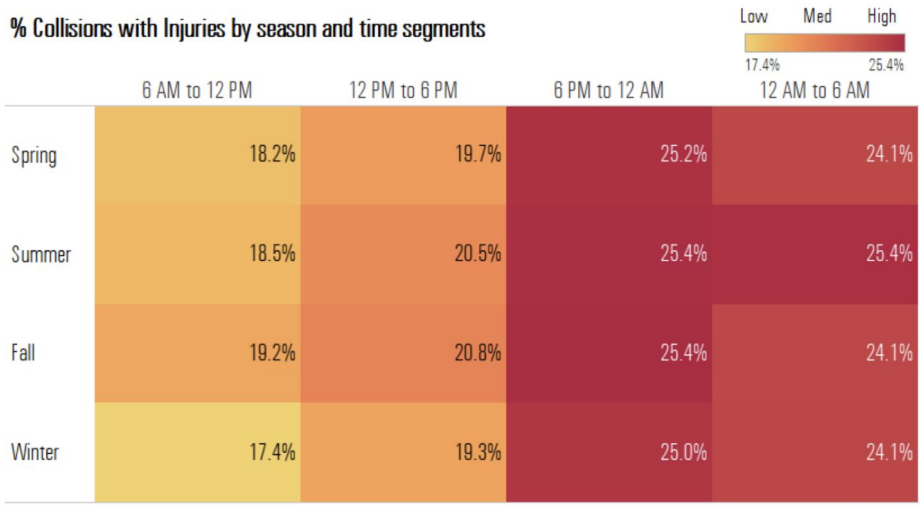

Figure 1: Percentage of collisions resulting in injuries across season and time segments in 2019

The heatmap in Figure 1 summarizes the % of collisions resulting in injuries by season and time segments. There is a clear pattern observed among the time segments as the % of collisions resulting in injuries is much higher during dark (6PM to 12 AM and 12 AM to 6 AM) compared to the morning and afternoon hours.

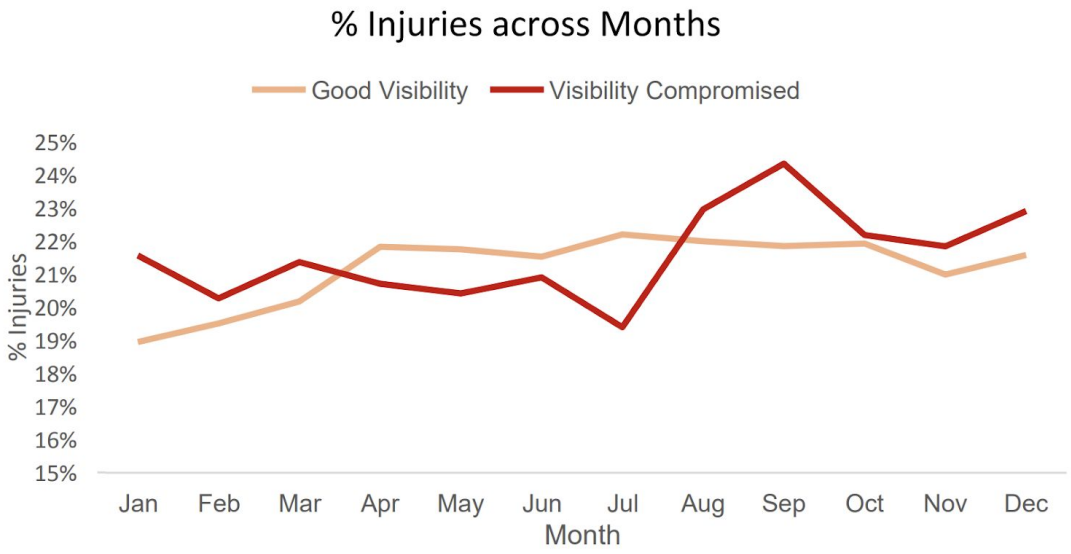

Figure 2: Impact of low visibility on percentage of collisions resulting in injuries across months in 2019

Low visibility index does not seem to have any impact on the proportion of injuries during summer months. However, Figure 2 shows increased trend during Autumn and Winter seasons.

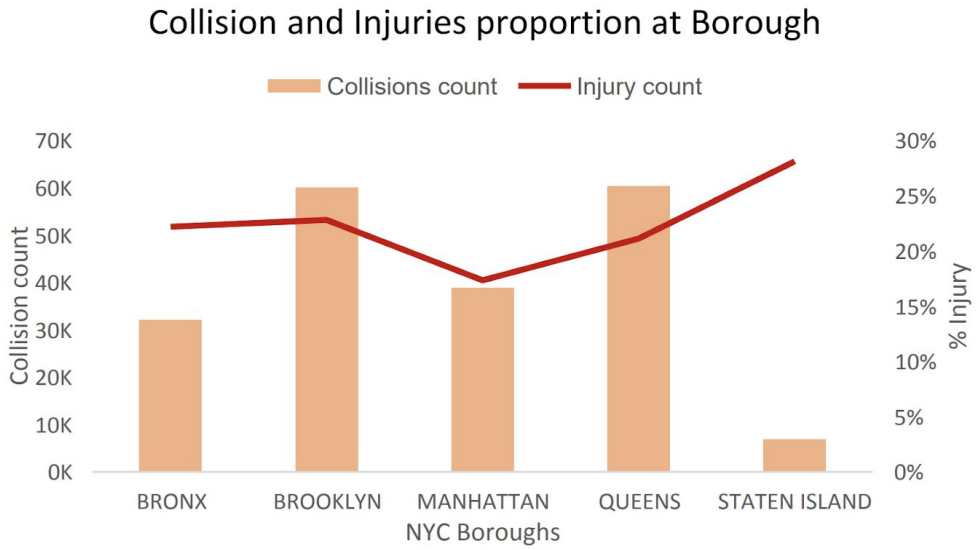

Figure 3: Total number of collisions and percentage of collisions resulting in injuries across Boroughs

Staten Island has the highest proportion of injuries(~28%) out of total collisions, whereas Manhattan observes the lowest (17%) proportion of injuries in 2019

Statistical Methods

Given we had the data at a collision level and the corresponding outcome for each collision, we analyzed the following questions in order to understand the impact of various vehicular as well as non-vehicular factors on a collision resulting in an injury.

1. Is there an association between the time segment of a day and a collision resulting in injuries across the boroughs of New York City?

In order to answer the above question, we used a GLM logistic regression model with:

- Response Variable: Collision resulted in an injury or not (binary)
- Predictors: Time segment of the day (factor) and Borough (factor)

We decided to go ahead with these variables as it was important to assess the effect of time of a day on the outcome of a collision, as the severity of a collision happening during day time might be different compared to a collision during midnight. For the analysis, we grouped time of the day into four segments (6 AM to 12 PM, 12 PM to 6 PM, 6 PM to 12 AM and 12 AM to 6 PM) to reflect the varying trends in the flow of traffic across NYC.

In order to assess the association of time segment with collisions resulting in injuries, we performed a deviance test to test the composite hypothesis, treating time segment as a factor variable:

- Full model: Injury ~ Borough(factor) + time segment(factor)
- Reduced model: Injury ~ Borough(factor)

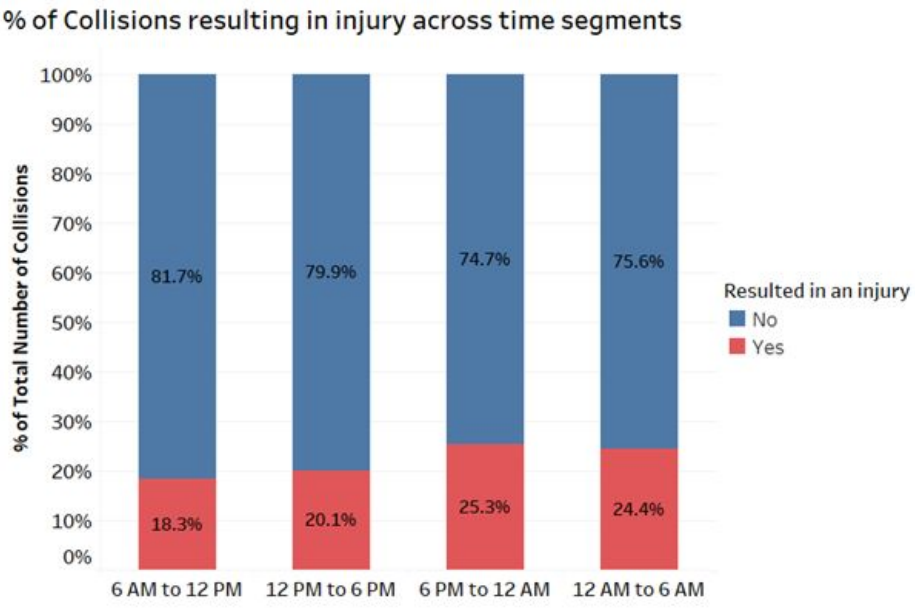

Figure 4: Percentage of collisions resulting in injuries across time segments

2. Do seasons and weather conditions have an association with collisions resulting in injuries across New York City?

In order to answer the above question, we used a GLM logistic regression model with:

- Response Variable: Collision resulted in an injury or not (binary)
- Predictors: Seasons (factor), Precipitation (in inches), Snowfall (in inches), Low visibility (binary)

We decided to go ahead with these variables as it was important to assess the effect of seasons on the outcome of a collision, as the severity of a collision happening during summer might be different compared to a collision during winter season under snowy conditions. For the analysis, we grouped all months of 2019 into four seasons (Spring, Summer, Fall, and Winter) to reflect the varying weather conditions. If the weather condition for a day observed fog, smoke haze or blowing snow, the day is given a low visibility flag of 1.

In order to assess the association of season with collisions resulting in injuries, we performed a deviance test to test the composite hypothesis, treating season as a factor variable:

- Full model: Injury ~ season(factor) + precipitation + snowfall + Low visibility
- Reduced model: Injury ~ precipitation + snowfall + Low visibility

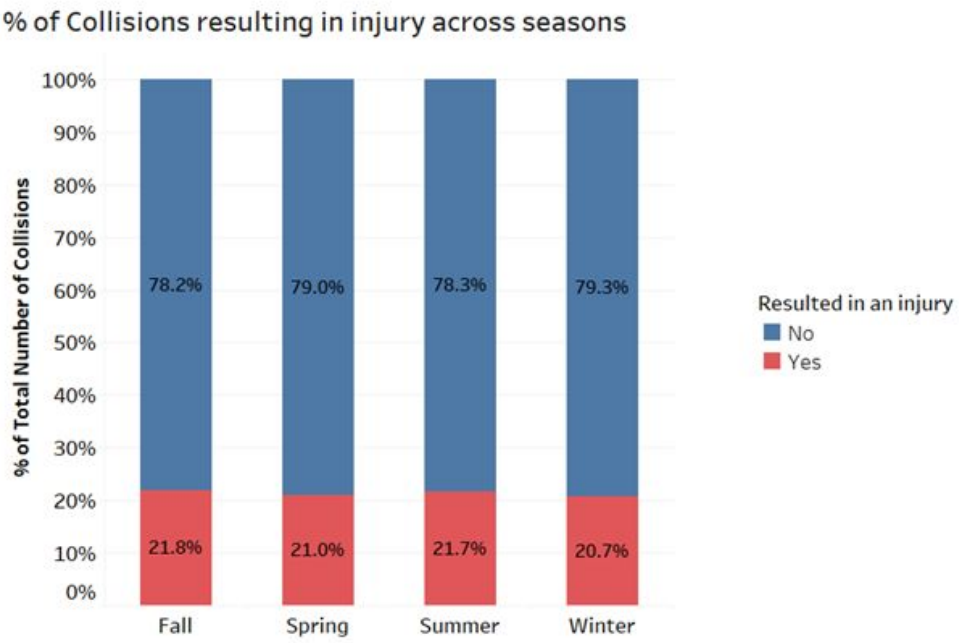

Figure 5: Percentage of collisions resulting in injuries across seasons of the year

3. Is there an association between Driving Under Influence (DUI) and a collision resulting in injuries across the boroughs of New York City?

In order to answer the above question, we used a GLM logistic regression model with:

- Response Variable: Collision resulted in an injury or not (binary)
- Predictors: DUI indicator (binary) and Borough (factor)

We decided to test the association of the DUI indicator as we believed that the severity of a collision while driving under the influence of alcohol/ drugs would be higher than the severity for collisions where other factors were involved. For the DUI indicator, we created a binary indicator flag that is set to 1 when the contributing vehicular factor is either 'Alcohol Involvement' or 'Drugs (Illegal)' and 0 otherwise.

In order to assess the association of DUI with collisions resulting in injuries, we performed a deviance test to test the composite hypothesis with:

- Full model: Injury ~ Borough(factor) + DUI Indicator (Binary)
- Reduced model: Injury ~ Borough(factor)

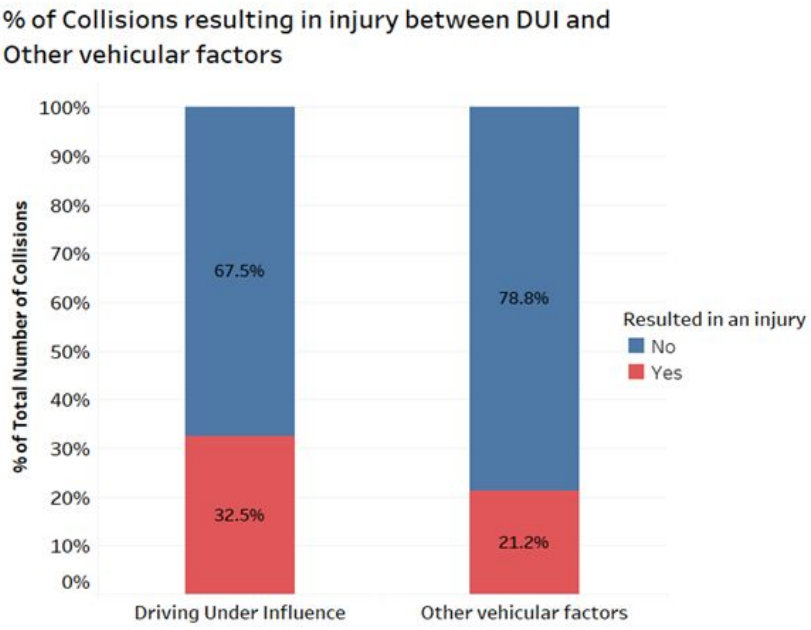

Figure 6: Percentage of collisions resulting in injuries for Driving Under Influence vs. Other factors

4. Do Boroughs, seasons, time segments, weather conditions, road type, victim type, vehicular factors, and vehicle types have an association with collisions resulting in injuries across New York City?

In order to answer the above question, we used a GLM logistic regression model with:

- Response Variable: Collision resulted in an injury or not (binary)
- Predictors:
  - Borough (factor)
  - Time segment (factor)
  - Season (factor)
  - Pedestrian (indicator)
  - Bicyclist (indicator)
  - Motor vehicle occupant (indicator)
  - Precipitation (inches)
  - Snowfall (inches)

- Low visibility (indicator)
- Vehicular factors (factor)
- Vehicle type (factor)
- Road type (factor)

The motivation behind this regression model is to understand the association and also do a relative comparison of the magnitude of the effect of these variables to answer questions such as i)Do highways have higher odds of collisions resulting in injuries compared to other road types? ii) what are the odds of collisions resulting in injuries for sedans vs trucks iii) what are the odds of collisions resulting in injuries for a bicyclist vs. motor vehicle occupant, and so on.

In order to test the presence of multicollinearity, we calculated the variance inflation factor(**VIF**) for all the features used in the model.

In order to assess the overall significance of all the model terms, we performed a deviance test as follows:

- Full model: Injury ~ Borough + time segment + season + pedestrian + bicyclist + occupant + precipitation + snowfall + low visibility + vehicular factors + road type + vehicle type
- Null model: Injury ~ 1

For all the four questions, we used Logistic regression models as the assumptions of the models were satisfied and the response variable was a binary value.

Under null hypothesis, the deviance test for each question has an approximate chi-squared distribution with p degrees of freedom(χ^2_p), where p is the difference between the number of parameters in the two model

Model assumptions:

The following are the assessment of the assumptions of the logistic regression model that we have considered for answering all our questions. These assumptions are taken from the standard set of assumptions for any GLM.

- **Independence:** We cannot assume independence of observations in our collisions data, since this is observational data and not a randomized experiment. There could be several confounding factors that might have led to a collision as well as the severity (injury) of the collision. Extreme weather conditions on a particular day or heavy traffic during a particular time segment could have had an impact on subsequent collisions. However, since the collisions data captures the data for the entire 2019 across the boroughs of NYC, we do not expect high correlations among variables that would cause a multi-collinearity problem and have a significant impact on the model outcome.

- **Large Sample size**: For 2019, the collisions dataset has 197,855 observations (collisions) and 42,156 collisions resulting in injuries. As both these numbers can be considered as being relatively large, the large sample size assumption is satisfied.

- **Mean-variance relationship:** For logistic regression, we do not need to assess the mean-variance relationship because it must be true due to the outcome variable being binary. The residual plot for the logistic regression model is shown below in Figure 4.

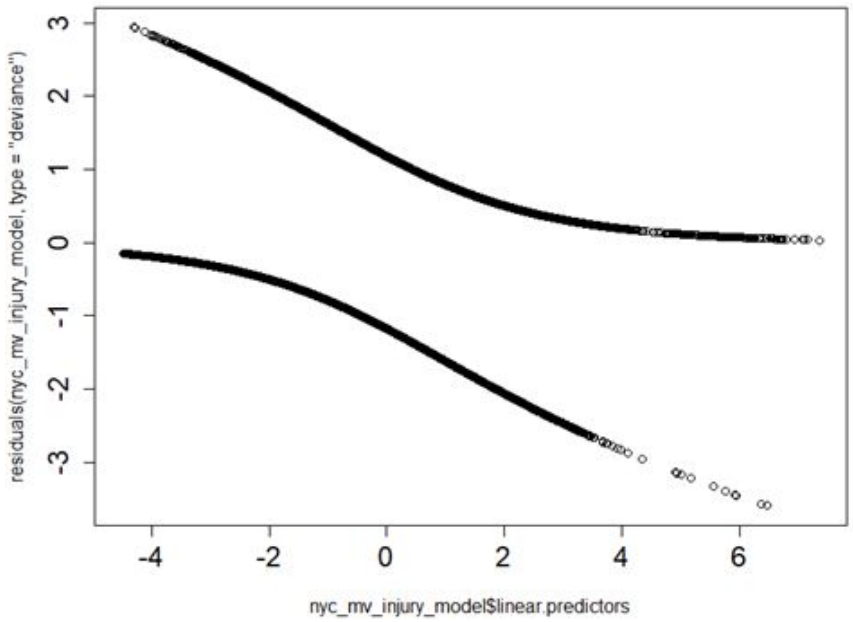

Figure 7: Residual plot for the logistic regression model

Results

The results of our analyses for the four questions posed are discussed below.

1. Is there an association between the time segment of a day and a collision resulting in injuries across the boroughs of New York City?

Logistic Regression model:

Predictors	Coefficient Estimate	Std. Error	Z Statistic	p-value (>|z|)	Odds Ratio
(Intercept)	-1.343	0.016	-83.645	0.000000e+00	-
12AM to 6AM	-0.043	0.020	-2.179	2.930180e-02	0.96
12PM to 6PM	-0.301	0.014	-21.738	9.063571e-105	0.74
6AM to 12PM	-0.412	0.015	-26.781	5.390582e-158	0.66
BRONX	0.316	0.019	16.588	8.477598e-62	1.37

BROOKLYN	0.349	0.017	21.020	4.317238e-98	1.42
QUEENS	0.253	0.017	15.077	2.311439e-51	1.29
STATEN ISLAND	0.633	0.030	20.985	9.062231e-98	1.88

Deviance Test

Analysis of Deviance Table

Model 1: Is_injury ~ (Borough)
Model 2: Is_injury ~ (time_segment) + (Borough)
 Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1 197850 204320
2 197847 203403 3 917.94 < 2.2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Conclusion

- The odds for a collision resulting in an injury are lower by 4.19% during 12 AM to 6 AM and by 34% during 6 AM to 12 PM with reference to time segment 6 PM to 12 AM
- With reference to Manhattan borough, the odds of for a collision resulting in an injury is 88% higher for Staten Island and 41% higher for Brooklyn
- The p-value across the four time segments and 5 boroughs are < 0.001 and hence we can reject the null hypothesis. We conclude that there is an impact of time of the day on a collision resulting in injury across all the 5 boroughs
- With the deviance value of 917, we can say that time segment has a significant effect on a collision resulting in an injury

2. Do seasons and weather conditions have an association with collisions resulting in injuries across New York City?

Logistic Regression model:

Predictors	Coefficient Estimate	Std. Error	Z Statistic	p-value (>|z|)	Odds Ratio
(Intercept)	-1.286	0.012	-110.828	2E-16	-
Winter	-0.069	0.016	-4.306	1.67E-05	0.93
Spring	-0.046	0.015	-2.988	0.002808	0.95
Summer	0.005	0.015	0.318	0.750298	1.00
Precipitation in inches	0.047	0.020	2.385	0.017059	1.05

Snowfall in inches	-0.075	0.020	-3.799	0.000145	0.93
Low Visibility	0.006	0.014	0.432	0.665797	1.01

Deviance Test:

Analysis of Deviance Table

Model 1: Is_injury ~ Precipitation_in + Snowfall_in + is_LowVisibility
Model 2: Is_injury ~ winter + spring + Summer + Precipitation_in + Snowfall_in +
is_LowVisibility
Resid. Df Resid. Dev Df Deviance
1 197851 204943
2 197848 204912 3 30.473

p-value = 1.097597e-06

Conclusion:

- Reference to Fall seasons, the odds of a collision resulting in an injury is lower by 6.7% in Winter and lower by 4.52% in Spring
- The odds of a collision resulting in injury are higher by 4.8% for each additional 1 inch of precipitation
- The odds of a collision resulting in injury are lower by 7.26% for each additional 1 inch of snowfall
- Summer with reference to the fall season and 'Low Visibility' does not have a significant effect on collisions resulting in injury
- The p-value across the seasons is < 0.001 and hence we can reject the null hypothesis. We conclude that there is an impact of season on a collision resulting in injury
- We also observe significant deviance of 30.473 indicating that the seasons have a significant effect on collisions resulting in injury

3. Is there an association between Driving Under Influence (DUI) and a collision resulting in injuries across the boroughs of New York City?

Logistic Regression model:

Predictors	Coefficient Estimate	Std. Error	Z Statistic	p-value (>|z|)	Odds Ratio
(Intercept)	-1.560	0.013	-116.490	<2e-16	1.78
DUI indicator	0.543	0.047	11.640	<2e-16	1.72
BRONX	0.302	0.019	15.890	<2e-16	1.35
BROOKLYN	0.340	0.017	20.560	<2e-16	1.40

QUEENS	0.242	0.017	14.500	<2e-16	1.27
STATEN ISLAND	0.610	0.030	20.290	<2e-16	1.84

Deviance Test:

Analysis of Deviance Table

Model 1: factor(d_i_ind) ~ DUI_ind + Borough
Model 2: factor(d_i_ind) ~ Borough
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1 197849 204655
2 197850 204782 -1 -126.4 < 2.2e-16 ***

Conclusion:

- Driving Under the Influence increases odds of injuries or death by 72% during motor vehicle collisions
- Staten Island collisions has increased odds ratio 84% compared to Manhattan under DUI presence; whereas Brooklyn has 40% increased odds of injuries compared to Manhattan
- Under the presence of Boroughs and keeping the boroughs constant, DUI indicator impacts still stay high with increased odd of 72% injuries
- We also observe a high deviance indicating that the DUI have a significant effect on collisions resulting in injury

4. Do Boroughs, seasons, time segments, weather conditions, road type, victim type, vehicular factors, and vehicle types have an association with collisions resulting in injuries across New York City?

Logistic Regression model:

Predictors	Coefficient Estimate	Std. Error	Z statistic	Pr (>|z|)	Odds Ratio
(Intercept)	-2.503	0.100	-24.942	< 2e-16	0.08
Bronx	0.543	0.024	22.183	< 2e-16	1.72
Brooklyn	0.534	0.022	24.681	< 2e-16	1.71
Queens	0.472	0.022	21.257	< 2e-16	1.6
Staten Island	0.883	0.036	24.286	< 2e-16	2.42

Fall	0.035	0.018	1.886	0.059262	1.04
Summer	0.070	0.018	3.872	0.000108	1.07
Winter	-0.047	0.019	-2.536	0.011206	0.95
12 AM to 6 AM	0.353	0.023	15.179	< 2e-16	1.42
12 PM to 6 PM	0.065	0.017	3.897	9.72E-05	1.07
6 PM to 12 AM	0.251	0.018	13.764	< 2e-16	1.29
Victim is a Bicyclist	3.320	0.043	78.046	< 2e-16	27.65
Victim is a pedestrian	4.290	0.038	112.687	< 2e-16	72.95
Victim is a motor vehicle occupant	-0.084	0.085	-0.995	0.319686	0.92
Precipitation in inches	0.027	0.023	1.137	0.255603	1.03
Snowfall in inches	-0.061	0.023	-2.630	0.008533	0.94
Low Visibility	-0.034	0.016	-2.045	0.040825	0.97
Weekend	0.030	0.015	2.044	0.041	1.03
Alcohol Involvement	0.674	0.055	12.258	< 2e-16	1.96
Backing Unsafely	-1.115	0.049	-22.896	< 2e-16	0.33
Driver Inattention/Distraction	0.123	0.023	5.328	9.94E-08	1.13
Failure to Yield Right-of-Way	0.455	0.030	15.242	< 2e-16	1.58
Following Too Closely	0.400	0.028	14.378	< 2e-16	1.49
Other Vehicular	0.113	0.041	2.730	0.006329	1.12
Passing or Lane Usage Improper	-0.462	0.042	-11.054	< 2e-16	0.63
Pedestrian/Bicyclist/Other Pedest	0.227	0.094	2.410	0.015953	1.25
Reaction to Uninvolved Vehicle	0.241	0.052	4.613	0.000003	1.27
Traffic Control Disregarded	1.320	0.044	30.345	< 2e-16	3.74
Turning Improperly	0.018	0.048	0.369	0.712475	1.02
Unsafe Lane Changing	-0.334	0.046	-7.215	5.38E-13	0.72
Unsafe Speed	1.009	0.047	21.662	< 2e-16	2.74

Unspecified vehicular factor	-0.286	0.024	-11.682	< 2e-16	0.75
Bike	0.226	0.097	2.320	0.02032	1.25
Box Truck	-0.538	0.070	-7.710	1.26E-14	0.58
Bus	-0.019	0.064	-0.295	0.768178	0.98
Motorcycle	1.713	0.076	22.413	< 2e-16	5.55
Pickup Truck	-0.176	0.053	-3.305	0.000951	0.84
Sedan	0.178	0.035	5.091	0.000000356	1.2
Station Wagon/Sport Utility Vehicle	0.057	0.035	1.617	0.105939	1.06
Taxi	0.306	0.045	6.768	1.31E-11	1.36
Unknown vehicle	-0.060	0.113	-0.530	0.59645	0.94
Avenue	0.124	0.035	3.554	0.000379	1.13
Boulevard	0.256	0.040	6.322	2.59E-10	1.29
Highway	0.517	0.037	13.984	< 2e-16	1.68
Road	0.255	0.046	5.601	2.13E-08	1.29
Street	-0.171	0.036	-4.777	0.00000178	0.84

Deviance Test:

Analysis of Deviance Table
Model 1: Is_injury ~ 1
Model 2: Is_injury ~ Borough + season + time_segment + Bicyclist + Pedestrian +
Occupant + Precipitation_in + Snowfall_in + Is_LowVisibility +
Is_weekend + Contributing_factor_f + Vehicle_type_f + Road_type_f

	Resid. Df	Resid. Dev	Df	Deviance	
1	197854	204974			
2	197809	158737	45	46236	p-value: 0

Test for Multicollinearity:

Predictors	GVIF	Df	GVIF^(1/(2*Df))
Borough	1.389	4	1.042

Season	1.043	3	1.007
Time segment	1.070	3	1.011
Bicyclist	1.359	1	1.166
Pedestrian	1.188	1	1.090
Occupant	1.623	1	1.274
Precipitation	1.425	1	1.194
Snowfall	1.064	1	1.031
Low visibility	1.477	1	1.215
Weekend	1.022	1	1.011
Contributing_factor	1.295	14	1.009
Vehicle type	2.097	9	1.042
Road type	1.357	5	1.031

Variable Importance:

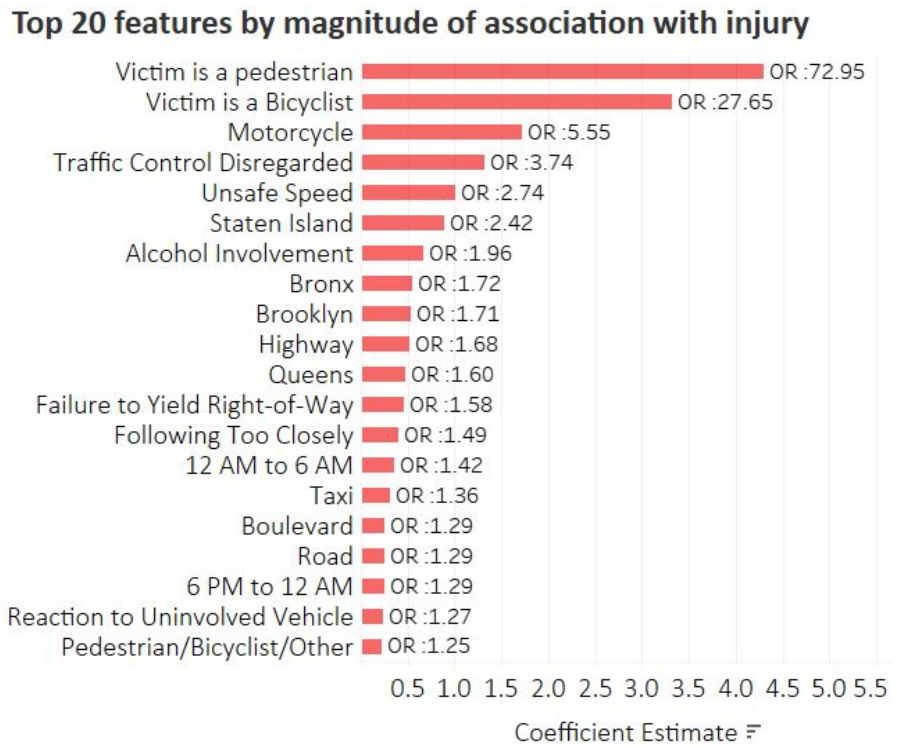

Figure 8: Top 20 features by magnitude of association with injury

Conclusion:

- The odds of a collision resulting in injury is higher by ~140% in Staten Island compared to Manhattan
- Given the victim in a collision is a pedestrian, the odds of getting injured is 72 times higher than the other victim types
- Factors like traffic control disregarded and unsafe speeds have the highest odds of injuries among vehicular factors. Similarly, among the vehicles, motorcycles have an odds of 5.5 times compared to other vehicles in a collision resulting in an injury
- Among road types, the odds ratio of a collision resulting in an injury is the highest for Highways
- Since none of the variables in the model have a high VIF (>5), the test for multicollinearity is satisfied
- The large deviance value and the p-value being equal to 0 suggests that the overall model is highly significant

Conclusions and Discussion:

The results from the hypothesis tests allow us to conclude that 'Time of the Day', 'Seasons' and 'Driving Under Influence' are significant factors contributing to collisions resulting in injuries in NYC. From the logistic regression model we can infer that factors such as 'Traffic Disregard' and 'Unsafe Speed' each increase the odds of collisions resulting in injuries in NYC by 200%. Highways have the highest odds of 68% collisions resulting in injuries among other road types. Among all the boroughs of NYC, Staten Island has the highest odds of injuries during a motor vehicle collision (~ 142%) and the highest number of DUI incidents. Other significant factors like the time period 12AM-6AM and the Summer season have higher odds of collisions resulting in injuries. Results also show that pedestrians are most prone to being injured in a motor vehicle collision.

We faced some limitations with regards to the interpretation of the results of the model and hypothesis tests. Because of the nature of our data and hypothesis, we had a lot of categorical predictor variables. For these categorical variables the estimated coefficients are always calculated against a reference category and hence interpretability of variable importance becomes difficult when modelling with multiple such predictors. With regards to the interpretation of odds ratio, magnitude of association between the variables may be misunderstood because of the presence of a reference level, leading to unrealistic estimates of intervention benefit.

Future Work

For our future work, we would like to improve our feature importance methodology by testing out different techniques such as pseudo partial correlation metric, adequacy metric, c-statistics and information values. We would also like to use Generalized Linear Mixed models to better capture the cyclic nature of seasons and time of the day features for enhanced inference and prediction on a wider time frame.

References

[1] Motor Vehicle Collisions - Crashes: Nyc Open Data Police (NYPD) - https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95
[2] Motor Vehicle Collisions - Person: Nyc Open Data Police (NYPD) - https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Person/f55k-p6yu
[3] Daily Summaries Station Details: National Centers for Environmental Information- Ncei - https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USW00014732/detail
[4] Ranking predictors in logistic regression – Paper http://www.mwsug.org/proceedings/2009/stats/MWSUG-2009-D10.pdf