FREQUENCY AND SEVERITY OF VEHICLE CRASHES

1. Introduction

1.1. Research Questions

Vehicle crashes have been a significant global concern, resulting in tragic loss of life, severe injuries, and substantial economic damages. Given the constant presence of vehicles and inherent risks of traffic travel, understanding and mitigating the factors contributing to possibility of collisions can be cardinal to enhancing public traffic safety, developing more effective transportation policies, and implementing better safety precautions.

This project aims to investigate the contributing factors associated with vehicle crashes. We focus on two core research questions. Firstly, we aimed to **predict the frequency of crashes,** enabling proactive identification of high-risk factors. Secondly, we will **predict the severity of crashes,** which is crucial for optimizing emergency response and safety precautions.

The project integrates multiple open datasets from the New York City Police Department. The primary source is the "Motor Vehicle Collision - Crashes" dataset, which includes all police-reported collisions from July 2012 to October 2025, with details on location, injuries and fatalities, contributing factors, and vehicle types. To integrate with weather conditions, we incorporate "Historical Weather Data" dataset from regional meteorological stations as compiled by the National Climatic Data Center, providing temperature, wind, precipitation, and notable weather events over the same time span. In addition, we also integrate "Automated Traffic Volume Counts" dataset, which offer sample vehicle volumes across NYC, to represent traffic conditions. Furthermore, we also take into consideration two supplementary datasets, "Motor Vehicle Collision - People" and "Motor Vehicle Collision - Vehicles", which are described to have additional context to the primary data, linked by the ID of each collision. The final integrated dataset creates a comprehensive foundation for the project to model the contributing factors of crash frequency and severity prediction.

1.2. Literature Review

Recent research on modeling the frequency and severity of vehicle crashes highlights that the accurate prediction relies significantly on a wide range of explanatory factors (Ali et al., 2024). The literature for this question generally classifies the variables in seven primary categories: weather condition, temporal factors, crash details, human factors, vehicle characteristics, roadway and/or environment, and traffic.

Regarding **weather conditions**, specific variables including rain, snow, cloudy, clear, foggy, and stormy weather are often taken into account (Wei et al., 2017 and Effati et al., 2024). Effati et al. (2024) also points out the necessity of using quantitative meteorological data, including parameters such as precipitation level and wind speed, as a supplementary procedure to police-reported weather information, which is also applied in this project. Regarding **temporal factors,** important factors usually are time of the day, days of the week (Ali et al., 2024), and holidays or non-holiday (Effati et al., 2024). In terms of **crash details,** the literature consistently points to type of collision (rear-end, sidewipe, head-on, etc) (Zhu et al., 2021; Ali et al., 2024), and the number of vehicles involved in a crash (Wei et al., 2017; Zhu et al., 2021). **Human factors** are commonly said by the literature to play a pivotal role in crash modelling, especially for crash severity models. Zhu et al. (2021) names human factors derived from police-report data as causality factors, using variables such as dangerous speeding, following too closely, failure to yield, improper turn, etc as primary causality variables. Other human-related variables also include occupant/driver age and genders, safety equipment and ejection, substance usage such as drug and alcohol (Ma et al., 2009; Kaufman et al., 2023; Ferenchak 2023; Ali et al., 2024). Similarly, **vehicle factors** are also pointed out as influential variables in crash severity studies. The variables include vehicle model year and if motorcycle is involved (Zhu et al., 2021). Madushani et al (2021) and Ali et al. (2024) also highlight **roadway conditions** (such as wet, dry, ice, pavement markings, etc) as a favourable parameter. **Traffic** volume and speed is also suggested as important variables in studies by Wei et al. (2017), Effati et al. (2024), Lee et al. (2023).

Some studies also point out the significant difference between before and after Covid quarantine order on the vehicle crash patterns. Sedaskis (2021), Ferenchak (2023), Kaufman et al. (2023), and Lee et al. (2023) all notice a decrease in the total number of collisions and an increase in the rate of injury severity in their research. Lee et al. (2023) states that the decline in crash frequency is strongest during peak hours, and the increase in severity is attributed to risky driving behaviors, including speeding, aggressive driving, and substance abuse. Farenchak (2023) also highlights the link between worsening severity after Covid to bad driving behavior observed on less congested roads and the association with social factors caused by the Covid pandemic.

Macro-level factors such as the socio-economic characteristics of the neighborhood, driver cognitive analysis, etc are also studied by some research (Pljakić et al. 2019), but not covered within the scope of this project.

1.3. Methodology

From literature review, the variables that are applicable in this study are: Weather: Precipitation, Heavy Fog, Thunder, Sleet, Hail, Glaze, Average Temperature; Temporal factor: Is_Weekend, Is_Holiday, Month; Crash Details: Number of vehicle involved,; Human factor: Driver distraction, Driver impairment, Driver Behavior; Roadway: If roadway being a contributing factor; Vehicle factors: vehicle type; Traffic: Average Traffic Volume, Busy Hour; Covid: Before and after Covid (i.e., before and after 2020-03-21, the first day of mandatory quarantine in NYC).

Research Question 1 utilized a multiple linear regression model to answer the question, “Is there a relationship between weather, temporal factors, and/or traffic levels, and the number of vehicle crashes per day in New York City?” A log-linear model was selected for this use case, for reasons described later in section 3.1: Analysis. Several research studies have already been conducted on the effect of the COVID-19 pandemic on collision rates in various locations, such as Connecticut (Doucette et al., 2021), the central southern United States (Ferenchak 2023), and Greece (Sekadakis et al. 2021).

Our modeling process for this research question was designed to extract meaningful variables from the data available to create a predictive, but not unnecessarily large, model. First, forward stepwise selection individually added variables based on insights from exploratory data analysis, with variables being kept or removed based on significance and collinearity. Next, backward stepwise selection considered variables not immediately evident in EDA. Finally, model diagnostics were used to assess the quality and appropriateness of the resulting model.

Research question 2 utilized a logistic regression model to predict the probability of a crash having injury, as this is fundamentally a classification task. This is a binary model with two-class outcomes as is-severity (crash having injury or fatal) and not-severity (crash having no injury or fatal). Although some research concludes that statistical models such as logistic regression model may generate low accuracy, due to their limited assumptions regarding data distribution, compared to more advanced machine learning methods (Zhang et al., 2018), logistic regression model remains an essential starting point to approach the research question, given its interpretability, baseline performance, and clear insight into the key predictors before transitioning to more complex modeling techniques.

Our modeling process for severity of vehicle crash followed a structured approach to ensure both interpretability and statistical accuracy. The process starts with multicollinearity assessment using Variance Inflation Factor (VIF). This is followed by both forward and backward stepwise variable selection using AIC, BIC and Likelihood Ratio test to achieve the most suitable set of features. After that, we will explore significant interaction effects, again using AIC, BIC, and LR tests to determine the new variables' goodness of fit. Finally, the resulting model's assumptions and fit will be accessed through model diagnostics and discussion on further tuning.

2. EDA Result

2.1. Overall data exploration

Our first exploratory analysis was on our primary dataset, the "Motor Vehicle Collision - Crashes" dataset. This dataset contains over 2 million rows and 29 columns. Except for counts of people injured or killed, none of these are numerical variables for which metrics such as minimum, maximum, median, or mean would make sense.

One variable that could be correlated with probability of injury was "Contributing Factor", which could have up to five values per row (collision). This variable has 61 possible values, with the most common value being Unspecified, and the most common specified value being Driver Inattention/Distraction. Most factors have relatively few associated collisions, so we decided to group these values into a smaller number of discrete categories before including this variable in the model.

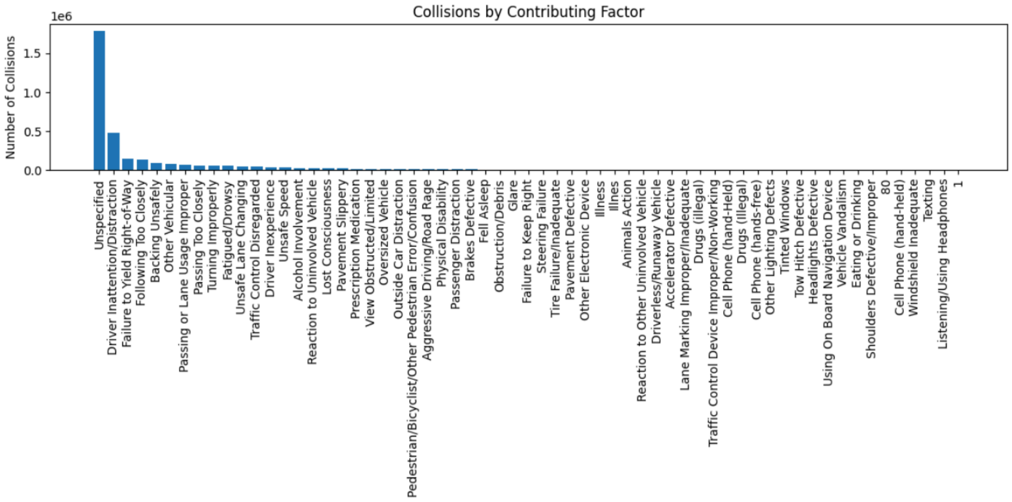
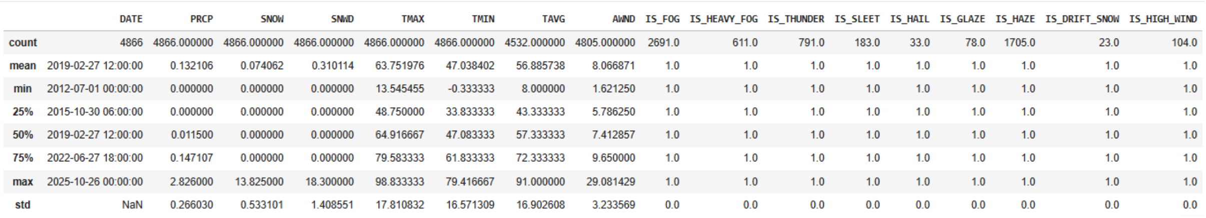


Figure 2.1.1: Collisions by Contributing Factors

Next, we examined the NOAA_Weather_2012_2025_NYC dataset. This consists of daily records from the NOAA Climate Data Online database, taken from stations included in the New York, NY US location with start and end dates encompassing the full date range of our Crashes dataset and at least 90% coverage. The portion of the dataset that we read in has over 100,000 rows and 18 columns: station name; date; seven continuous variables describing precipitation, snowfall/accumulation, temperature, and wind; and nine binary variables indicating the presence of extreme weather events such as heavy fog, sleet, or high wind. Our original plan was to join this data to the crash dataset by date and nearest latitude/longitude, but the datasets were too large to allocate enough memory in our coding environment for the join. Instead, we aggregated the data by date, calculating the average of each continuous variable and the maximum of each binary variable across all stations for that day (that is, did this event happen at any station in the area).

Below are the descriptive statistics for our weather set. The average daily temperature ranges between 8 and 91 degrees Fahrenheit, and the average wind speed ranges from 1 to 30 miles per hour. The most common severe weather event was fog, with 2,691 days, and the least common was drift snow, with only 23 days. About 50% of recorded days had at least 0.01 inch of precipitation, and only 7% had at least 0.01 inch of snow. Histograms were also created for each continuous variable, and can be viewed in the code file.



	DATE	PRCP	SNOW	SNWD	TMAX	TMIN	TAVG	AWND	IS_FOG	IS_HEAVY_FOG	IS_THUNDER	IS_SLEET	IS_HAIL	IS_GLAZE	IS_HAZE	IS_DRIFT_SNOW	IS_HIGH_WIND
count	4866	4866.000000	4866.000000	4866.000000	4866.000000	4866.000000	4532.000000	4805.000000	2691.0	611.0	791.0	183.0	33.0	78.0	1705.0	23.0	104.0
mean	2019-02-27 12:00:00	0.132106	0.074062	0.310114	63.751976	47.038402	56.885738	8.066871	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
min	2012-07-01 00:00:00	0.000000	0.000000	0.000000	13.545455	-0.333333	8.000000	1.621250	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
25%	2015-10-30 06:00:00	0.000000	0.000000	0.000000	48.750000	33.833333	43.333333	5.786250	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
50%	2019-02-27 12:00:00	0.011500	0.000000	0.000000	64.916667	47.083333	57.333333	7.412857	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
75%	2022-06-27 18:00:00	0.147107	0.000000	0.000000	79.583333	61.833333	72.333333	9.650000	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
max	2025-10-26 00:00:00	2.826000	13.825000	18.300000	98.833333	79.416667	91.000000	29.081429	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
std	NaN	0.266030	0.533101	1.408551	17.810832	16.571309	16.902608	3.233569	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 2.1.2: Descriptive statistics of weather dataset

The next dataset we explored was the Automated Traffic Volume Counts dataset. The initial dataset contained over 1.8 million rows and 14 columns: traffic volume counted over a 15-minute interval, data about the date, time, and location of each traffic sample, and some irrelevant IDs. Like the weather dataset, this data was aggregated by date, keeping the average traffic volume per date. The graph below of daily average traffic readings shows a handful of extreme outlier days, as well as some missing days during COVID that will need to be interpolated before modeling.

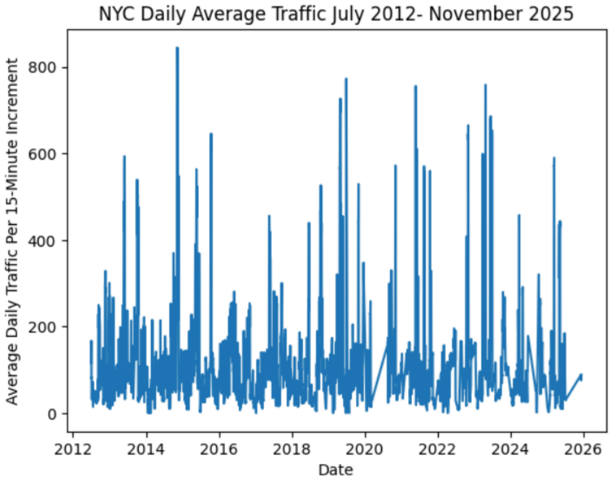


Figure 2.1.3: NYC Daily Average Traffic July 2012 - November2025

2.2. EDA specific to frequency of vehicle crashes (research question 1)

The people- and vehicle-specific datasets are not applicable to the prediction of crashes per day, so the EDA for the first research question focused on the first three datasets: crashes, weather, and traffic.

Upon graphing the number of vehicle crashes per day over time, it was evident that the overall magnitude decreased significantly at the onset of the COVID-19 pandemic, and has remained low since. After looking more closely at the volumes for March 2020, a binary ‘PRECOVID” variable was created to designate dates before March 21, 2020.

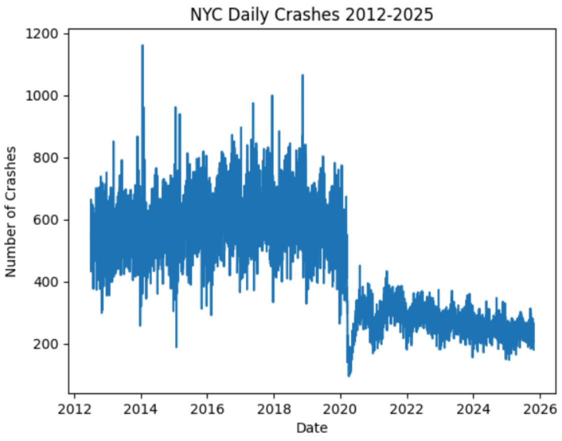


Figure 2.2.1: NYC Daily Crashes 2012-2015

Next, we calculated the correlation between each of the independent variables and the natural log of daily crashes. This revealed that most correlations with log of crashes, the bottom-most/rightmost variable in the chart, were very weak; all but the PRECOVID indicator had a magnitude of less than 0.2. Correlation values marginally increased when selecting only pre- or post-COVID data points, which suggested that interactions between PRECOVID and these variables may be significant. The numerical correlation values are recorded in the code file.

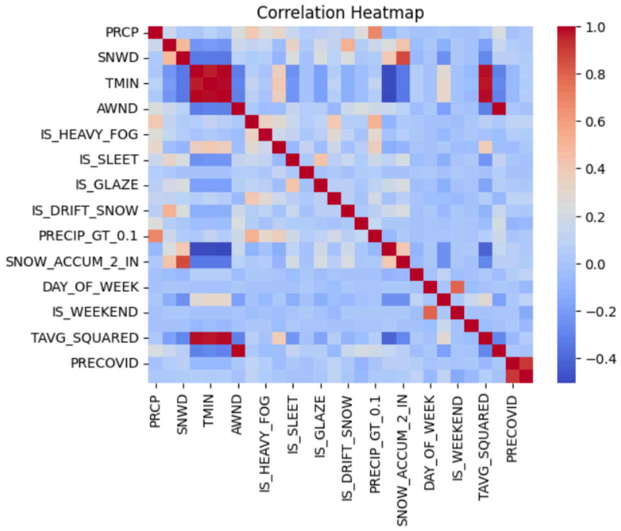


Figure 2.2.2: Correlation heatmap

The following variables were extracted from existing variables, but none were found to have correlation with our response variable above 0.2:

- Whether the snow accumulation (SNWD) was greater than 2 inches
- Whether the average temperature was less than or equal to 32 degrees (freezing)
- Whether precipitation was greater than 0.1 inch
- Month of year
- Day of week
- Whether the day of week was a weekend (Saturday or Sunday)

Additionally, continuous variables were transformed by squaring, taking the square root, or taking the natural log, and then were tested for correlation with both pre-COVID and post-COVID data points. Aside from average traffic volume post-COVID, none of these transformations increased correlation by more than 0.02. It is also worth noting that the post-COVID data has a commonly reoccurring traffic value of approximately 20, which could be the imputed value for early-COVID days.

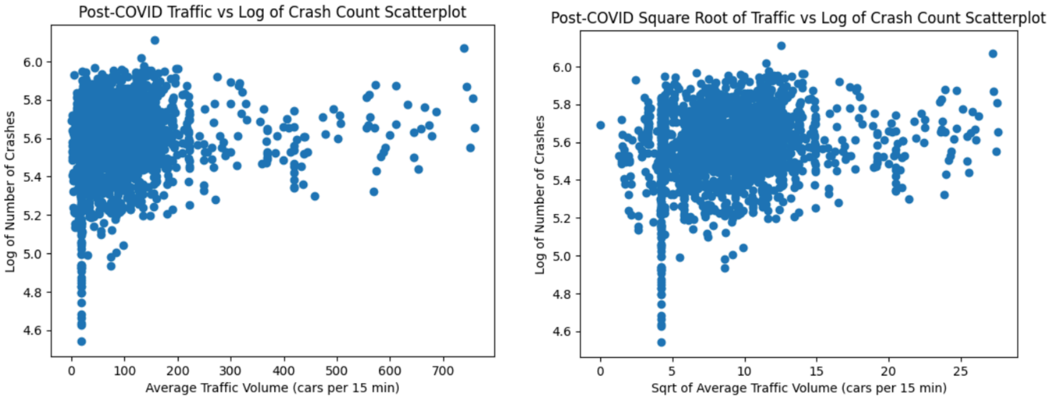


Figure 2.2.3: Post-Covid Traffic and SQRT(Pre-Covid Traffic) v.s. Log(Crash Count)

Finally, we used box plots to visualize the distribution of daily crashes across days of the week and months of the year. Especially when separating into pre-COVID and post-COVID dates, crash count is visibly lower on weekends (day values 5 and 6).

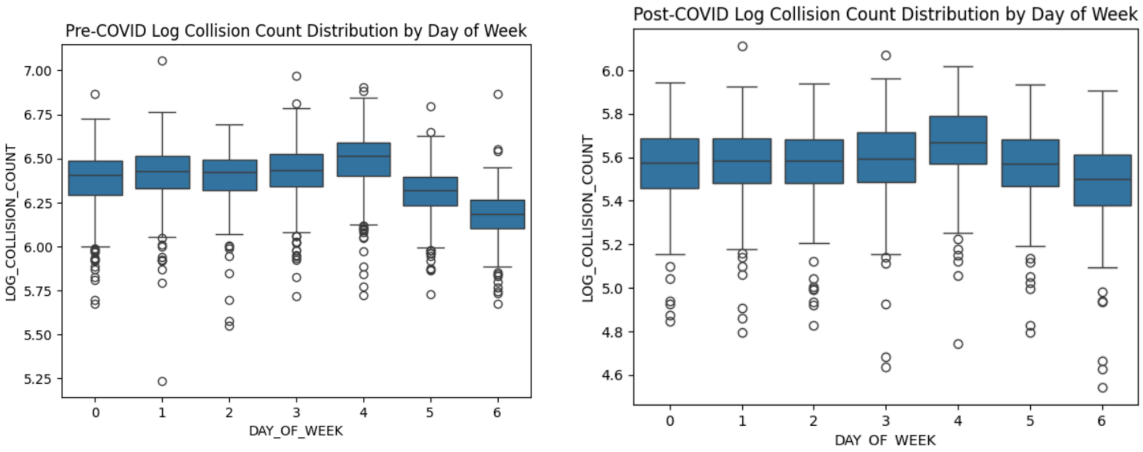


Figure 2.2.4: Pre-Covid and Post-Covid Log Collision Count Distribution by Day of Week

Looking at the distribution of daily crashes by month, there is a slight seasonal pattern with higher medians in May and June. The post-COVID data also has a lower median and many low outliers in April, but this is most likely due to very low volumes during April 2020, in the first weeks of the pandemic-related lockdowns.

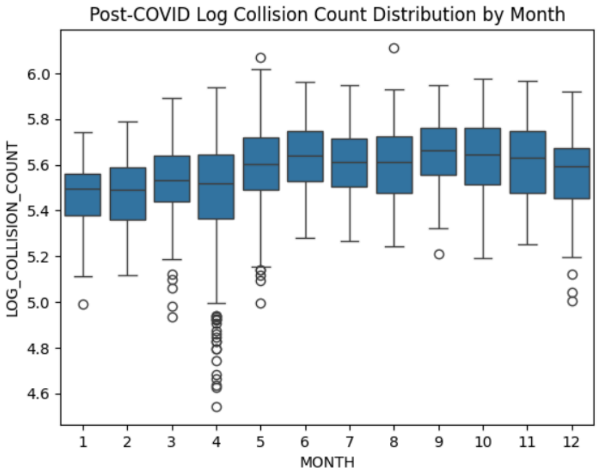


Figure 2.2.5: Post-Covid Log Collision Count Distribution by Month

2.3. EDA specific to severity of vehicle crashes (research question 2)

The goal of this research question is to find predictors for severity of the accident (i.e. if the collisions have at least one injury or one fatality), therefore, the response variable is abbreviated as IS_INJURED, which is binary by nature: 0 if the collision results in at least no injury and no fatality; 1 if there is at least one injury or one fatality. Looking at the numbers of injuries and fatalities per collision, approximately 75% of the collisions have no injury nor fatality, which can be considered an imbalance classification problem. As the number of injuries increases, the frequency of crashes with this many injuries decreases, with the highest number of injuries reported as 43 persons.

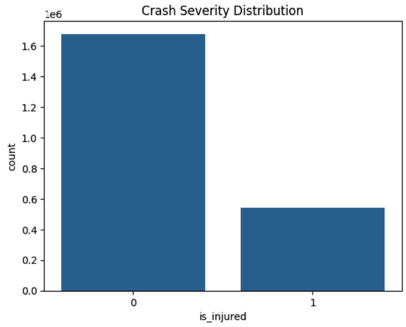


Figure 2.3.1: Distribution of Crash Severity (Injury vs No-Injury)

Following the variable sets suggested from literature review, we investigate the corresponding variables that are accessible within the datasets using data exploration techniques and association tests between the target variable and each candidate predictor, including Chi-square test for categorical variables and Independent T-test for continuous variables. The sets of variables that will further be used for predictors in model fitting is as follow:

Variable Name	**Codes/Values**	**Description**	**Asso. Test Sig**
Weather Factors			
BELOW_FREEZING	1 - Below Freezing 0 - Above Freezing	P(Injury|BELOW_FREEZING = True): 0.1937 P(Injury|BELOW_FREEZING = False): 0.2471	< 0.05
VISION_BLOCK	1 - Is Vision Block 0 - No Vision Block	P(Injury|VISION_BLOCK = True): 0.2371 P(Injury|VISION_BLOCK = False): 0.2514	< 0.05
SNOW_ACCUM_2_IN	1 - Snow over 2 inches 0 - Snow < 2 inches/no snow	P(Injury|SNOW_ACCUM_2_IN = True): 0.1721 P(Injury|SNOW_ACCUM_2_IN = False): 0.2467	< 0.05
Temporal Factors			
IS_WEEKEND	1 - Weekend 0 - Weekdays	P(Injury|IS_WEEKEND = True): 0.2508 P(Injury|IS_WEEKEND = False): 0.2405	< 0.05
CRASH MONTH	Months	12 months of the year	< 0.05
IS_HOLIDAY	1 - Holiday margin 0 - Not in holiday margin	P(Injury|IS_HOLIDAY = True): 0.2474 P(Injury|IS_HOLIDAY = False): 0.2428	< 0.05
BUSY HOUR	1 - Is busy hours 0 - Not in busy hours	P(Injury|BUSY HOUR = True): 0.2358 P(Injury|BUSY HOUR = False): 0.2472	< 0.05
Roadway factors			
Is_Roadway	1 - If roadway is a factor 0 - Roadway is not a factor	P(Injury|Environment_Roadway = True): 0.2514 P(Injury|Environment_Roadway = False): 0.2619	< 0.05
Traffic Factor			
Avg Traffic Volume	Average Traffic Volume	Min: 0; Max:763; Mean: 129 25%: 67.3; 50%: 110.9; 75%:168.1	< 0.05
Crash Detail			
Number of Vehicle Involved	1 - Single-Crash 2 - Two vehicle involved 3 - More than 3 vehicles	Proportion:1 - 20%; 2 - 70%; 3 - 8%	< 0.05
Vehicle Types			
Is_Sedan	1 - Sedan is involved 0 - No sedan is involved	P(Injury|Is_Sedan = True): 0.2426 P(Injury|Is_Sedan = False): 0.2447	< 0.05
Is_Coupe	1 - Coupe is involved 0 - No Coupe is involved	P(Injury|Is_Coupe = True): 0.2279 P(Injury|Is_Coupe = False): 0.2432	< 0.05
Is_Motorcycle_Scooter	1 - Motorcycle is involved 0 - No Motorcycle is involved	P(Injury|Is_Motorcycle_Scooter = True): 0.7579 P(Injury|Is_Motorcycle_Scooter = False): 0.2156	< 0.05
Is_Bicycle	1 - Bicycle is involved 0 - No Bicycle is involved	P(Injury|Is_Bicycle = True): 0.7869 P(Injury|Is_Bicycle = False): 0.2384	< 0.05
Is_Public	1 - Public vehicle is involved 0 - No Public vehicle is involved	P(Injury|Is_Public = True): 0.1775 P(Injury|Is_Public = False): 0.2454	< 0.05

Is_Truck_Van	1 - Truck/Van is involved 0 - No Truck/Van is involved	P(Injury|Is_Truck_Van = True): 0.1681 P(Injury|Is_Truck_Van = False): 0.2534	< 0.05
Is_Emergency	1 - Emergency vehicle is involved 0 - No Emergency is involved	P(Injury|Is_Emergency = True): 0.1291 P(Injury|Is_Emergency = False): 0.2439	< 0.05
Is_Other	1 - Other vehicle is involved 0 - No other vehicle is involved	P(Injury|Is_Other = True): 0.1857 P(Injury|Is_Other = False): 0.2743	< 0.05
Human Factor			
Driver_Distraction	1 - Driver Distraction is involved 0 - Driver Distraction not involved	P(Injury|Driver_Distraction = True): 0.2713 P(Injury|Driver_Distraction = False): 0.256	< 0.05
Driver_Behavior	1 - Driver Behavior is involved 0 - Driver Behavior not involved	P(Injury|Driver_Behavior = True): 0.2683 P(Injury|Driver_Behavior = False): 0.2549	< 0.05
Driver_Impairment	1 - Driver Impairment is involved 0 - Driver Impairment not involved	P(Injury|Driver_Impairment = True): 0.2273 P(Injury|Driver_Impairment = False): 0.2654	< 0.05
Vehicle_Defect	1 - Vehicle Defect is involved 0 - Vehicle Defect not involved	P(Injury|Vehicle_Defect = True): 0.2194 P(Injury|Vehicle_Defect = False): 0.2648	< 0.05
Covid			
PRECOVID	1 - Before '2020-03-21' 2 - After '2020-03-21'	P(Injury|PRECOVID = True): 0.1944 P(Injury|PRECOVID = False): 0.3916	< 0.05

Table 2.3.1: Summary of variables used in Research Question 2 model

The EDA and association test show little difference in precipitation level on the probability of injury for a crash, but a significant difference when comparing categories within Below_Freezing, Vision_Block, and Snow_Accum_2_In. The exploratory data surprisingly contradicted some findings from literature by highlighting how the probability of crash injury is, in fact, smaller in critical weather conditions. Similarly, the probability of injury is also displayed to be moderated during busy hours, which is defined based on traffic patterns on the data, which is 7 - 9 and 16 - 18 for weekdays and 14 - 18 for weekends, although slightly higher during holidays and the mid-year period from June to September also sees a slightly higher probability of injury.

The next factor is traffic effect - Average number of Traffic Volume. The EDA result is similar to some other research in the literature: injury often occurs in lower traffic volume sections.

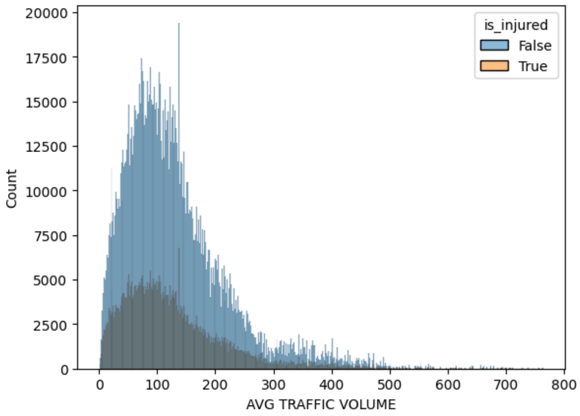


Figure 2.3.2: Distribution of crashes with injuries compared to average traffic volume

This statistic is also supplemented by the crash details, in which single-vehicle crash doubles two-vehicle crash in the severity probability.. The dataset only reports up to 5 vehicles at most, divided into 3 groups: one-vehicle, two-vehicle, and over-two-vehicle crashes, in which crashes involving two-vehicles account for 73% of the collisions but have the lowest probability of injury.

Similar to the number of vehicles involved, vehicle types (sedan, truck, motorcycles,...) can also be derived from the dataset. A crash can have utmost 5 different vehicle types, therefore, we construct an independent variable for each type instead of combining them into one categorical variable. 60% of vehicles involved in crashes is Sedan, but unprotected vehicles such as personal mobility, bicycle, and motorcycle and scooter accounts for the highest probability of injury.

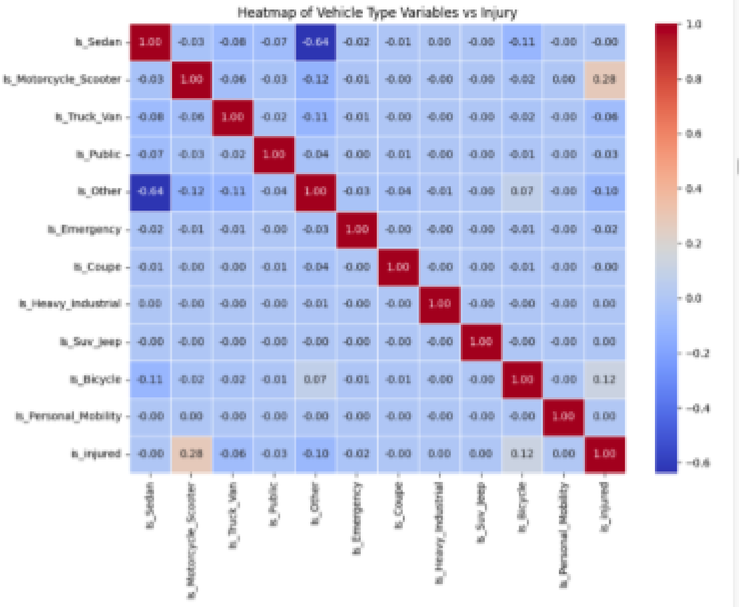


Figure 2.3.3: Correlation between Vehicle Types and Injured Probability

The EDA on the left chart of Figure 2.3.4 also points out crashes that involve both motorcycle/scooter and sedan have much higher injury rates (~75%) compared to crashes that only involved sedans (22%). Figure 2.3.5 shows that given high injury probability in both Motorcycle and Bicycle individually, the injury probability is highest at around 81% if a crash involved both vehicles.

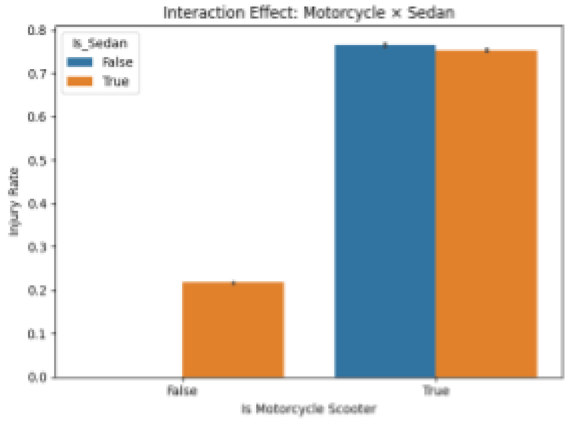


Figure 2.3.4: Injury Rate of Sedan and Sedan x Motorcycle

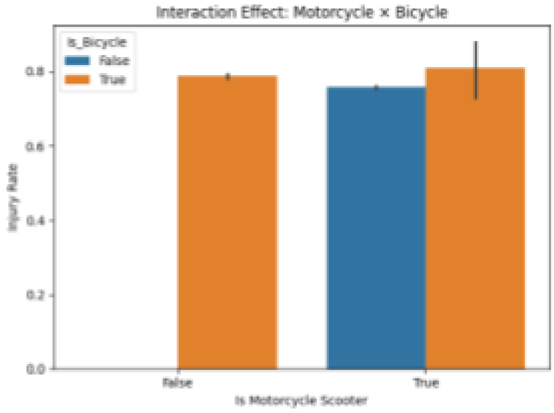


Figure 2.3.5: Injury Rate of Bicycle and Bicycle x Motorcycle

Similar to EDA in research question 1, there is also a big increase in the injury probability post-Covid. Between 2012 to early 2020, the probability of injury shows a relatively stable pattern (0.1 and 0.3), which is followed by a relatively abrupt discontinuity in early 2020 (time of the first Covid outbreak and quarantine), then significantly increased to a much higher level (0.3 and over 0.5). The rate appears to be more volatile compared to preCovid period (Figure 2.3.6). In addition, Figure

2.3.7 shows a different pattern in injury rate between weekend/weekday pre-Covid and post-Covid, highlighting that median injury rate is higher on weekends than weekdays in post-Covid, while it is on the contrary pre-Covid.

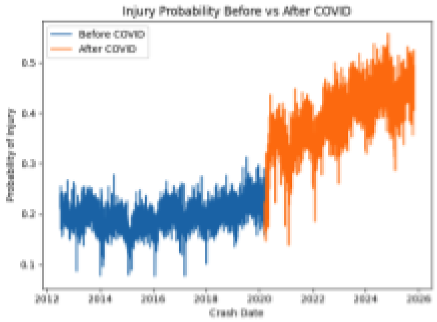


Figure 2.3.6: Average injury probability before & after Covid

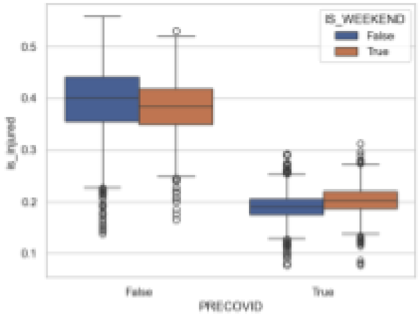
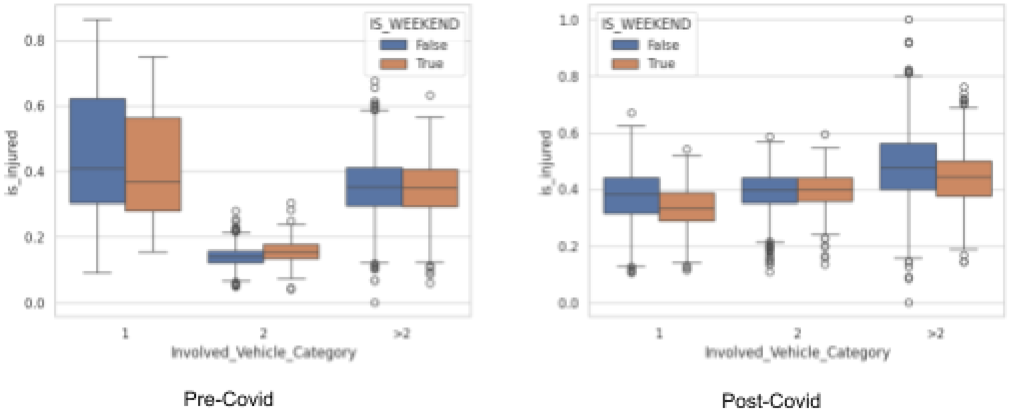


Figure 2.3.7: Average injury probability between weekend and weekdays before & after Covid

We also notice that the relative risk of injury on the number of involved vehicles changed after Covid (Figure 2.3.7). The injury rate appears to increase with crashes with two vehicles and multi vehicles (>2). In both periods, the injury rate for multi-vehicle crashes appear to be slightly lower on weekends.



Pre-Covid

Post-Covid

Figure 2.3.7: Average injury probability, categorized by number of vehicles involved and weekend

Regarding human factors in figure 2.3.8, the EDA suggests that Driver Behavior (aggressive driving, speeding, failure to yield/turn,...) and Driver Distraction are associated with slightly higher probability of injury. On the other hand, Driver Impairment (illness, drug, alcohol involved) and Vehicle Defect shows a lower probability when present.

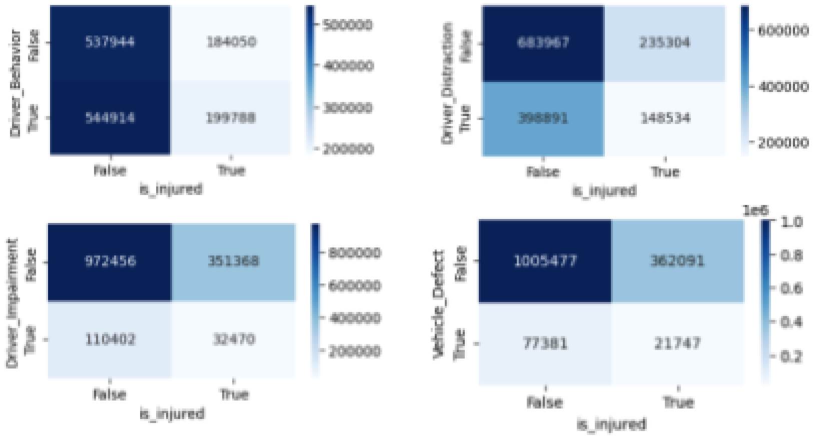


Figure 2.3.8: Human Factors contingency table

3. Analysis

3.1. Research Question 1

The number of crashes per day is an event count that cannot be negative, so a standard linear model, which can predict positive or negative values, is not appropriate. A Poisson distribution was also considered, but it assumes an equal mean and variance in the distribution, which is invalid for this dataset, as seen in the table below:

Data Selection	Mean of Collision Count	Variance of Collision Count
All Dates	455.52	32,614.33
Pre-COVID	591.87	10,432.53
Post-COVID	267.6	2,238.76

Table 3.1.1: Mean and Variance of collision count

Ultimately, the natural log was chosen as a link function, that is, the linear model will predict the natural log of the number of crashes per day. This is similar to a Poisson regression, but does not make any assumptions about mean and variance.

An initial model was created using forward stepwise variable selection, in each step selecting a variable that had a high correlation or visible relationship with the target variable in EDA. This resulted in a model with the COVID indicator, month (as a dummy-encoded categorical variable with 11 coefficients), the day-of-week-is-weekend binary variable and its interaction with the COVID indicator, average temperature, and the square-root-transformed average traffic volume as well as its interaction with the COVID indicator. However, VIF values showed a high correlation between the temperature, month, and traffic variables. This resulted in a model with only five coefficients, two of which were interaction terms.

After this, backwards stepwise regression was used to determine other variables that may be significant. Variables were eliminated, up to two at a time, with the highest p-values until no variables with p-values above 0.05 remained. VIF values were then calculated to ensure there was no excess collinearity between variables. Finally, we tried adding interactions with the pre-COVID indicator and traffic, keeping only those with statistically significant coefficients. This resulted in a model with nine variables and some interactions, as follows:

log(COLLISION COUNT) = 5.5004 + 0.9035 * PRECOVID - 0.0547 * IS_WEEKEND - 0.1243 * (PRECOVID*IS_WEEKEND) - 0.0553 * IS_HOLIDAY - 0.0907*(PRECOVID*IS_HOLIDAY) + 0.0096 * AVG_TRAFFIC_SQRT -0.0071* (PRECOVID * AVG_TRAFFIC_SQRT) + 0.048 * PRCP - 0.0209 * IS_HEAVY_FOG + 0.0346 * IS_THUNDER - 0.0557 * IS_SLEET - 0.0391 * BELOW_FREEZING

The model coefficients can be interpreted as follows:

- A hypothetical post-COVID weekday that isn't a holiday, has 0 precipitation and no weather events, has an average temperature above 32 degrees Fahrenheit, and has an average traffic volume of 0 is predicted to have $e^{5.5004}$ = 244.8 crashes. It doesn't make sense to have crashes on a day with 0 traffic, or for a day to have 0 traffic in the first place, so this has limited actual meaning.
- A day pre-COVID is expected to have ($e^{0.9035}$ -1) = 150% more crashes than an otherwise identical day post-COVID.

- A weekend day is expected to have ($e^{0.0547}$ -1) = 5.6% fewer crashes than an otherwise identical weekday post-COVID, ($e^{0.0547 + 0.1234}$ - 1) = 20% fewer crashes than an otherwise identical weekday pre-COVID.
- A holiday is expected to have ($e^{0.0553}$ -1) = 5.6% fewer crashes than an otherwise identical non-holiday post-COVID, ($e^{0.0553 + 0.0907}$ -1) = 15.7% fewer crashes than an otherwise identical non-holiday pre-COVID
- As the square root of average traffic volume per 15 minutes increases by one vehicle, the number of crashes is expected to increase by ($e^{0.096}$ -1) = 0.96% post-COVID, ($e^{0.096-0.0071}$ -1) = 0.25% pre-COVID.
- If the daily precipitation increases by 1 inch, the number of crashes is expected to increase by ($e^{0.048}$ - 1) = 4.9%.
- A day with heavy fog is expected to have ($e^{0.0209}$ - 1) = 2.1% fewer crashes than an otherwise identical day without heavy fog.
- A day with thunder is expected to have ($e^{0.0346}$ - 1) = 3.5% more crashes than an otherwise identical day without thunder.
- A day with sleet is expected to have ($e^{0.0557}$ - 1) = 5.7% fewer crashes than an otherwise identical day without sleet.
- A day with an average temperature below freezing is expected to have ($e^{0.0391}$ - 1) = 4% fewer crashes than an otherwise identical day with an average temperature above freezing.

This model has an R-squared value of 0.857, and an adjusted R-squared value of 0.856, meaning that about 86% of the variance in the data is explained by the model. However, models with all variables except the pre-COVID indicator, fit to pre- and post-COVID data points, had adjusted R-squared values of 0.278 and 0.094 respectively, indicating minimal predictive power apart from the pre-COVID indicator. Other diagnostic measures also show that this model is not strong enough to be used in a real-life scenario. For example, the below graph of actual vs predicted values demonstrates that the model outputs the same prediction value for inputs of varying actual values, so some factors that affect daily collision rates are not present in this model. Additionally, while the magnitude of actual values explains the difference in residual rates pre-COVID vs post-COVID, residuals are higher in 2020 and 2021 than in 2024 and 2025, suggesting that residuals are not independent.

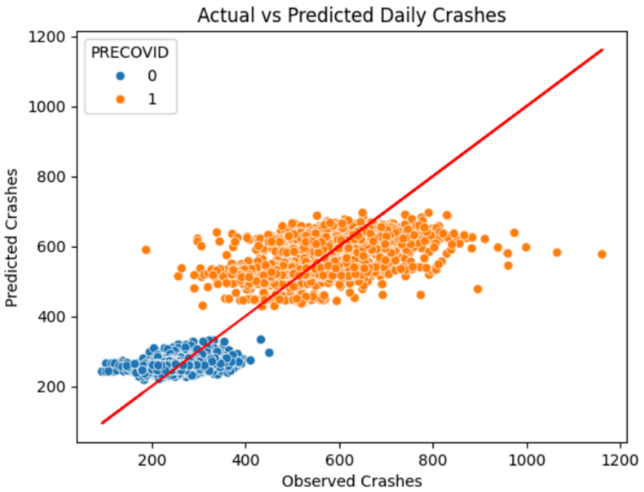


Figure 3.1.1: Plot of observed vs predicted crashes, with y=x reference line

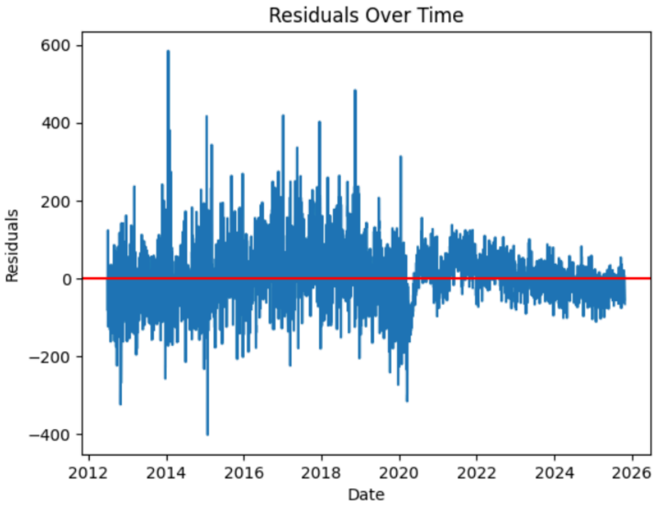


Figure 3.1.2: Plot of residual by date

3.2. Research Question 2

Variable Selections

In this research question, we also apply both forward and backward stepwise for variable selections. An initial model was created using forward stepwise variable selection, in each step, we will add the variables, which have significant association test in EDA, for each factor group. This result in a model with: COVID, Month, is weekend, is holiday

An initial model was created using forward stepwise variable selection, in each step selecting a variable that had a high correlation or visible relationship with the target variable in EDA. This was followed by backward stepwise to remove redundant variables. The method used was comparing AIC, BIC, and LR tests between models. Then it was repeated for selecting interactive terms. Some variables that are eliminated from this process is: Is_Personal_Mobility (Vehicle type), traffic volume * Covid, month * Covid, interaction of Motorcycle_Scooter/Bicycle with each of the human factors (driver distraction, driver behavior, driver impairment, vehicle defect), average temperature, Avg Traffic Volume * human factors. That resulted in the final model as followed:

$log(\frac{p(injury=1)}{1-p(injury=1)})$ = -1.0300 - 0.0647*BELOW_FREEZING - 0.1026*SNOW_ACCUM_2_IN + 0.0228*VISION_BLOCK - 0.0933*IS_WEEKEND + 0.1477*Involved_Vehicle_Category_2 + 0.9030*Involved_Vehicle_Category_gt2 + 0.5751*BEFORE_COVID - 0.0218*CRASH_MONTH_2 - 0.0536*CRASH_MONTH_3 - 0.0347*CRASH_MONTH_4 - 0.0130*CRASH_MONTH_5 - 0.0045*CRASH_MONTH_6 - 0.0192*CRASH_MONTH_7 - 0.0159*CRASH_MONTH_8 + 0.0143*CRASH_MONTH_9 + 0.0195*CRASH_MONTH_10 + 0.0018*CRASH_MONTH_11 + 0.0211*CRASH_MONTH_12 + 0.0179*IS_HOLIDAY + 0.1353*BUSY_HOUR - 0.2833*Is_Sedan + 2.0673*Is_Motorcycle_Scooter + 3.1717*Is_Bicycle - 0.2110*Is_Public - 0.4419*Is_Truck_Van + 0.0947*Environment_Roadway - 0.1448*Is_Other - 0.5942*Is_Emergency - 0.1595*Is_Coupe - 0.0634*Vehicle_Defect + 0.3102*Driver_Distraction + 0.2940*Driver_Behavior + 0.3583*Driver_Impairment + 0.1520*(IS_WEEKEND * Involved_Vehicle_Category_2) - 0.0761*(IS_WEEKEND * Involved_Vehicle_Category_gt2) - 0.0837*(IS_WEEKEND * BEFORE_COVID) -

1.5465*(Involved_Vehicle_Category_2 * BEFORE_COVID) - 0.8890*(Involved_Vehicle_Category_gt2 * BEFORE_COVID) - 0.2213*(BUSY_HOUR * Involved_Vehicle_Category_2) - 0.1031*(BUSY_HOUR * Involved_Vehicle_Category_gt2) + 0.5253*(Is_Sedan * Is_Motorcycle_Scooter) - 2.0794*(Is_Motorcycle_Scooter * Is_Bicycle) + 0.1310*(IS_WEEKEND * Involved_Vehicle_Category_2 * BEFORE_COVID) + 0.1959*(IS_WEEKEND * Involved_Vehicle_Category_gt2 * BEFORE_COVID) - 0.0001*AVG_TRAFFIC_VOLUME

Model Interpretation

As the change in odds ratio in logistics regression = ($e^{coefficient}$ -1)*100% the model can be interpreted as: holding other variables constant:

Weather Factor: Below_Freezing can decrease the injury odds by ($e^{-0.067}$- 1) =6.3%; Snow Accumulation ≥ 2 inches can decrease injury odds by ($e^{-0.1026}$- 1) =9.7%, showing that visible hazardous winter conditions may imply more cautious driving that reduces severity. However, Vision Block conditions (hail, snow drift, glaze) appear to increase injury odds by ($e^{0.028}$- 1) =2.3%, a modest but meaningful driving caution.

Temporal Factor: Most monthly effects are small for both positive and negative impacts - no single month dramatically changes the odds of injury. Weekend crashes appear to be less severe, with the odds of injury decreased by ($e^{-0.0933}$- 1) =8.9%. Meanwhile, holidays slightly increase injury odds by ($e^{0.0179}$ - 1) = 1.8%, and usual busy hours increase injury by ($e^{0.1353}$ - 1) = 14.5% for a self-crash. Compared to this baseline, crashes during busy hours that involve two vehicles are $e^{-0.2213}$-1 = 29.8% lower, and crashes during busy hours that involve more than two vehicles are $e^{-0.1031}$-1 = 9.7% lower.

Roadway factors (slippery pavement, road defective, etc.) increase the injury likelihood by ($e^{0.947}$-1) = 9.93%.

Vehicle Type has some of the strongest predictors from this model. Vehicles with the highest risk of injuries when involved in a crash are “unprotected” vehicles: motorcycles ($e^{2.0673}$ -1) = 8 times higher than otherwise, and bicycles ($e^{3.1717}$ - 1) = 23.8 times higher than otherwise. Meanwhile, “well-protected” vehicles with heavier mass appears to be safer in a crash: Sedan reduces injury odds by ($e^{-0.24833}$ - 1) = 25%, Truck/Van by ($e^{-0.4419}$ - 1) = 36%, Emergency vehicles by ($e^{-0.5942}$ - 1)=45%; Public transport by ($e^{-0.211}$ - 1) = 19%, and Coupe by ($e^{-0.1595}$-1) = 15%. A crash that involved both a sedan and a motorcycle, the odds of injury increase by $e^{0.5253}$-1 = 69% higher than that expected from each vehicle, but a crash involved two most “dangerous” vehicle - bicycle and motorcycle - actually has a $e^{-2.0794}$-1 = 87.5% lower risk of injury compared to each type individually. That implies crashes by bicycle and motorcycle can be mostly safe, if they are not interacted with heavier vehicles.

Human Factors: driver unsafe behaviors strongly increase the risk of injury, in which driver distraction increases injury odds by ($e^{0.31001}$ -1) = 36%, driver behavior (aggression, speeding, failure to yield/turn,...) by ($e^{0.2940}$- 1) = 34%, driver impairment (alcohol, drug, illness,...) by ($e^{0.3583}$ - 1) = 43%.. Meanwhile, vehicle defect (brake, steering, tire defect, etc) appears to lower the injury likelihood, following the assumption that drivers with vehicle defects may commute with safer speed and more caution.

Crash details: crashes that involve two vehicles have a higher injury likelihood to a self-crash by ($e^{0.1477}$ - 1) = 25.9%, and crashes that involve two vehicles are associated with a 147% higher injury rate compared with self-crash. However, this is moderated by busy hour factors as discussed above, and also Covid factors.

Covid: the main effect of BEFORE COVID while holding other variables constant indicates that crashes pre-Covid has injury odds of $e^{0.5752}$ - 1 = 78% higher than post-Covid, which is contrary to the pattern we found via EDA when comparing average probability of injury before and after Covid.

This difference is then explained by the existence of interaction effects of Covid with other variables: Is_Weekend, Involved Vehicle = 2, and Involved Vehicle > 2. That means, if a crash happens on a weekday, pre-covid, and is a single-crash, it will have a 78% higher risk of injury. Using it as a baseline for interaction terms with Covid, the risk of injury after Covid increases if crashes involve two-vehicle, multi-vehicle, and weekends. Specifically, before Covid, weekends have a slightly lower impact on the odds of injury compared to after Covid, with an effect of $e^{-0.0837}$-1 = 8% decrease in injury odds; 2-vehicle crashes before Covid is $e^{-1.5465}$- 1 = 79% less likely to have injury compared to 2-vehicle crashes after Covid; multi-vehicle crashes before Covid is also $e^{-0.8890}$- 1 =59% lower in injury risk, compared to otherwise. Considering the interaction of all three variables (number of vehicles involved, weekend, Covid), the model shows a higher injury likelihood for both 2-vehicle ($e^{0.1310}$-1 = 13.9%) and multi-vehicle crashes ($e^{0.1959}$-1=22%) on weekends compared to weekday, however, they are small adjustments that moderates the stronger main effects above. Given two-vehicle crashes are the most common, when interactions are applied, the combining odds for most crash types are higher after Covid. Therefore, we can conclude that Covid is an important variable if such a model is deployed in a real-world scenario studying historical data back to multiple years before 2020 in New York City.

Model Diagnostics

With the standard classification threshold of 0.5, we achieved a model that classifies a random injury crash above a random non-injury crash correctly 67.8% of the time (AUC = 0.681), which is a relatively weak model. The model's accuracy is 68.1% but in this classification problem where response = 1 is rare, accuracy is not a strong diagnostics method. From the confusion matrix, the false positive rate = 11,791/(242,989+11,791) = 4.6%, the false negative rate = 125,040/(125,040+48,603) = 72%. This imbalance indicates that the model is too cautious that it misses 72% of actual injury cases, which is also reflected in 0.80 precision at the cost of 0.28 recall. In the current context of predicting injury, for most scenarios, this is actually dangerous because the cost of missing an injury is much higher than not catching one.

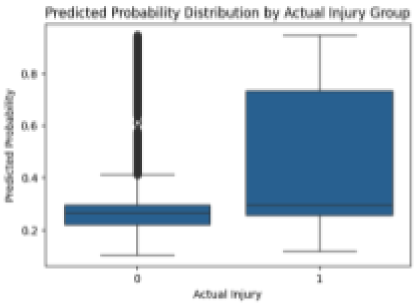
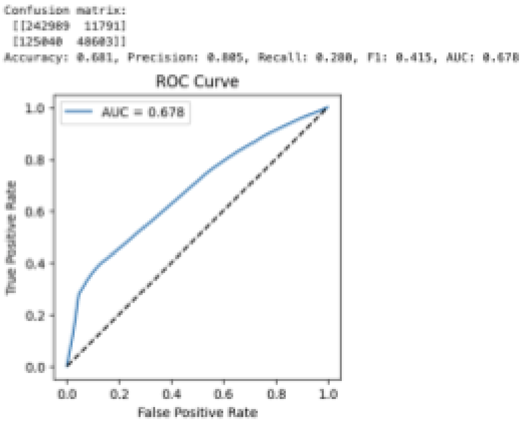


Figure 3.2.1: Predicted Probability Distribution by Actual Injury Group



```
Confusion matrix:
 [[242989  11791]
 [125040  48603]]
Accuracy: 0.681, Precision: 0.805, Recall: 0.280, F1: 0.415, AUC: 0.678
```

Figure 3.2.2: Confusion Matrix and ROC Curve

Classification Threshold Tuning

One approach to finding a better balance between false negative rates and false positive rates is lowering the classification threshold. We found that at threshold = 0.28, the false positive and false negative are at the most balanced with FPR = 0.381215 and FNR = 0.389 (figure 3.2.3). Model precision reduces to 0.522 and recall increases to 0.61 (figure 3.2.4). Hence, threshold = 0.28 is suggested to be used as a baseline model for this problem.

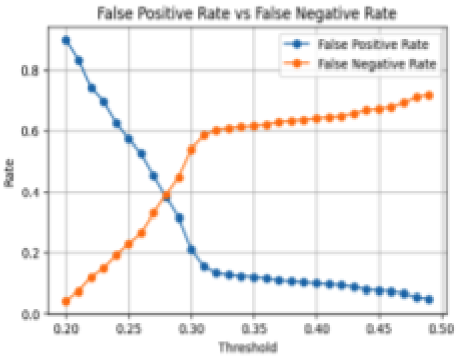
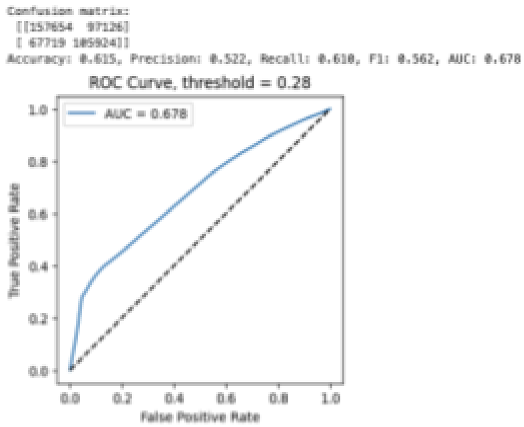


Figure 3.2.3: FPR and FNR by threshold



```
Confusion matrix:
 [[157654  97126]
 [ 67719 105924]]
Accuracy: 0.615, Precision: 0.522, Recall: 0.610, F1: 0.562, AUC: 0.678
```

Figure 3.2.4: ROC curve at threshold = 0.28

4. Discussion

4.1. Research Question 1

The resulting model demonstrates a statistically significant relationship between some weather conditions, day of week and pre-COVID vs post-COVID time as temporal factors, and traffic levels with the number of vehicle crashes per day in New York City. However, the features' correlation with the target variable, as well as the large decrease in adjusted R-squared between this original model and models trained only on pre-COVID or post-COVID data, indicate that most of the predictive power of the model is due to the COVID-period indicator variable.

The results of this analysis suggest further opportunities to improve the model fit by adding other variables and data sources. For example, there is a generally linear decrease in residuals between 2020 and 2023. This could be due to lower traffic during times of quarantine regulations, which could be modeled by a binary or categorical variable marking the days and/or levels of quarantine enforcement in New York City. Furthermore, the significance of holiday indicator and traffic variables encourages further exploration of events that could affect traffic, such as sports events or tourism seasons. This would be a more robust measure of expected traffic levels, especially because the traffic data we used doesn't sample every location at the same frequency, which affects the daily average. Finally, we did not collect any data on social measures that are taken to reduce collisions, such as adjustment of speed limits, level of police activity, and even anti-distracted-driving PSA campaigns. This would be an especially interesting research question, as unlike weather and day of week, these factors are controllable.

4.2. Research Question 2

The model demonstrates a strong predictive relationship between vehicle types (especially "unprotected" vehicles such as motorcycles and bicycles) and driver factors (distraction, behavior, impairment) to the odds of injury for a vehicle crash, highlighting that the best way to reduce risk of injury in a crash is using protected vehicle (sedan, coupe, public transport), being highly cautious if using motorcycles or bicycles, and also avoid distraction, driving with impaired well-being, and deficient driving practices. On the other hand, external factors like weather and temporal pattern is also smaller. The models also show how Covid changed the likelihood of injury by interacting with crash types and weekends, and some nuanced patterns such as pairing of high-risk vehicles.

Regarding the tuning result for this model at threshold = 0.28, in most settings, we should design the threshold to be at this threshold or even lower to avoid as many false negatives as possible. This is important for applications in emergency medical response, vehicle's safety procedure activation, and general safety. However, reducing the threshold without careful further analysis could also cause some issues. The cost of false positives may be high when considering scenarios such as: alert fatigue and loss of trust in the system, costs of controlling intervention and disruption, and insurance fraud investigations. Therefore, the threshold at 0.28 compromises an equilibrium point for false positive and false negative, suggesting that it could be considered a baseline model for future improvements. Future improvement for this research question should focus on further refinement on the variables to improve the performance (AUC score) of the model, as well as analysis for the optimal threshold applied to a specific context. More advanced models or resampling methods may also be beneficial for better prediction of injury.

5. References

City of New York. (2025). *Motor vehicle collisions - crashes* [Dataset]. data.cityofnewyork.us. https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes

National Centers for Environmental Information. (2025). *Daily summaries location details* [Dataset]. https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/locations/CITY:US360019/detail

City of New York. (2025), *Automated Traffic Volume Count* [Dataset] https://data.cityofnewyork.us/Transportation/Automated-Traffic-Volume-Counts/

City of New York. (2025). *Motor vehicle collisions - vehicles* [Dataset]. data.cityofnewyork.us. https://catalog.data.gov/dataset/motor-vehicle-collisions-vehicles

City of New York. (2025). *Motor vehicle collisions - person* [Dataset]. data.cityofnewyork.us. https://catalog.data.gov/dataset/motor-vehicle-collisions-person

Ali et al., 2024, Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review, https://doi.org/10.1016/j.aap.2023.107378

Wei et al., 2017, Truck crash severity in New York city: An investigation of the spatial and the time of day effects, http://dx.doi.org/10.1016/j.aap.2016.11.024

Effati et al., 2024: Considering the Reliability of Police‑Reported Weather Information on Freeways Traffic Crash Severity Analysis: Proposing a Mixed Statistical and Geospatial Solution, https://doi.org/10.1007/s40890-024-00218-w

Zhu et al., 2021, Crash Injury Severity Prediction Using an Ordinal Classification Machine Learning Approach, https://doi.org/10.3390/ijerph182111564

Ma et al., 2009, Analysis of the logistic model for accident severity on urban road environment, https://doi.org/10.1177/1687814018805581

Kaufman et al., 2023, Increase in Motor Vehicle Crash Severity: An Unforeseen Consequence of COVID-19, https://doi.org/10.1177/00031348211047466

Ferenchak 2023, Impacts of COVID-19 on Motor Vehicle Crash Frequency and Severity by Functional Classification and Land Use Context, DOI: 10.5507/tots.2023.011

Madushani et al., 2021, Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers, https://doi.org/10.1016/j.treng.2023.100190

Lee et al., 2023, Changes in traffic crash patterns: Before and after the outbreak of COVID-19 in Florida, https://doi.org/10.1016/j.aap.2023.107187

Sedaskis et al., 2021, Analysis of the impact of COVID-19 on collisions, fatalities and injuries using time series forecasting: The case of Greece, https://doi.org/10.1016/j.aap.2021.106391

Pljakić et al. 2019, The influence of traffic-infrastructure factors on pedestrian accidents at the macro-level: The geographically weighted regression approach, https://doi.org/10.1016/j.jsr.2022.08.021

Doucette et al., 2021, Evaluation of motor vehicle crash rates during and after the COVID-19-associated stay-at-home order in Connecticut, https://doi.org/10.1016/j.aap.2021.106399

Zhang et al., 2018, Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods, https://doi.org/10.1109/ACCESS.2018.2874979

APPENDIX

Appendix 1: Detailed description for human factor variable

Variable	If `Contributing Factor` in
Driver Distraction	'Driver Inattention/Distraction', 'Passenger Distraction', 'Outside Car Distraction', 'Cell Phone (hand-Held)', 'Cell Phone (hand-held)', 'Cell Phone (hands-free)', 'Texting', 'Using On Board Navigation Device','Listening/Using Headphones', 'Eating or Drinking','Other Electronic Device', 'Reaction to Uninvolved Vehicle', 'Reaction to Other Uninvolved Vehicle','View Obstructed/Limited', 'Glare'
Driver Impairment	'Fatigued/Drowsy', 'Fell Asleep', 'Lost Consciousness', 'Illness', 'Physical Disability', 'Alcohol Involvement', 'Drugs (illegal)', 'Drugs (Illegal)', 'Prescription Medication'
Driver Behavior	Failure to Yield Right-of-Way', 'Following Too Closely', 'Unsafe Speed', 'Passing or Lane Usage Improper', 'Unsafe Lane Changing','Turning Improperly', 'Driver Inexperience', 'Passing Too Closely', 'Aggressive Driving/Road Rage', 'Backing Unsafely','Failure to Keep Right', 'Traffic Control Disregarded','Traffic Control Device Improper/Non-Working', 'Lane Marking Improper/Inadequate','Pedestrian/Bicyclist/Other Pedestrian Error/Confusion', 'Oversized Vehicle'
Vehicle Defect	'Brakes Defective','Accelerator Defective','Steering Failure', 'Tire Failure/Inadequate','Headlights Defective','Other Lighting Defects', 'Tinted Windows','Windshield Inadequate','Tow Hitch Defective', 'Vehicle Vandalism', 'Driverless/Runaway Vehicle' ,'Other Vehicular'

Appendix 2: Detailed description for vehicle factor variable

Variable	If `Vehicle Type` in
Is_Sedan	"sedan", "4 dr", "car", "subn", "suburban", "station wa", "wagon", "smart"
Is_Coupe	"2 dr", "coupe","convertible"
Is_Suv_Jeep	"suv", "jeep", "hrv"
Is_Truck_Van	"van", "sprinter", "econoline", "cargo van", "minivan", "work van", "vanette", "van bus", "truck", "pick", "pickup", "box truck", "tractor tr", "semi", "freight", "flatbed", "dump", "tanker", "cement", "mixer", "stake", "garbage", "refuse", "rolloff", "tow", "utility tr", "u-haul", "uhaul"
Is_Public	"bus", "mta", "school bus", "shuttle", "coach", "omni", "trolley", "access-a-r", "access a r"
Is_Emergency	"ambul", "ambu", "ems", "fdny", "fire", "ladder", "rescue", "police", "nypd", "fd", "emt"

Is_Motorcycle_Scooter	"motorcycle", "motorbike", "moped", "mo ped", "scooter", "dirt bike", "motorscoot", "e-scooter", "e scooter", "kick scoot", "vespa"
Is_Bicycle	"bicycle", "pedal bike", "citibike", "bike", "e-bike"
Is_Heavy_Industrial	"fork", "bobcat", "backhoe", "loader", "bulldozer", "excav", "payloader", "tractor", "skid", "roller", "crane", "construction"
Is_Personal_Mobility	"hoverboard", "unicycle", "one wheel", "segway"

Appendix 3: Roadway

Variable	**If `Contributing Factor` in**
Is_Environment_Roadway	'Pavement Slippery', 'Pavement Defective', 'Obstruction/Debris', 'Shoulders Defective/Improper', 'Animals Action'