STAT 405 Group 1 Final Project Report

Bill Huang, Patricia Hashimoto, Samuel Kwan, Anna Tutuianu, James Qu

Contents

Background and Introduction

Introduction

The COVID-19 pandemic has created an unusual backdrop for car crash data analysis both in New York City and nationwide. Starting in 2020, the national number of car crashes and traffic injuries declined, but the number of fatal crashes increased. From 2019 to 2020, the number of fatal crashes in the U.S. increased by 6.8% (NHTSA 2022b). The National Highway Traffic Safety Administration estimated that 42,915 people died in traffic accidents in 2021 (compared to 38,824 in 2020) (NHTSA 2022b). This analysis investigates the motor vehicle collisions that occurred in New York City from 2012 to April 2022.

New York City's efforts to reduce traffic fatalities make collision data from the city noteworthy, and the city's wide range of publicly accessible traffic data offers multiple metrics to connect. The city began a transportation policy, Vision Zero, in 2014 with the intention of eliminating all traffic deaths, and this policy has become a national model (Hu 2022). The program has involved reducing speed limits on some streets from 30 m.p.h. to 25 m.p.h., creating a network of automated speed cameras, and redesigning streets to be more pedestrian- and cyclist-friendly (Hu 2022). New York City officials have also lobbied state lawmakers for local control of city streets under the belief that city authority to set speed limits, expand red light cameras, and increase hours for school zone speed cameras could make streets safer (Hu 2022). In May 2022, responding to traffic fatality rates that have reached an eight-year high, the city debuted a $4 million billboard and media campaign to stop speeding, which has increased during the pandemic, and Mayor Eric Adams has pledged $904 million to improve the city's streets plan with measures like intersection redesigns, new protected bike lanes, and more pedestrian areas in the next five years (Hu 2022). The high toll of car crashes on human life makes the causes of, trends in, and effects of policies and current events on traffic fatalities urgent issues of investigation, and New York City's unusually accessible and complete data makes questions about these issues possible to answer.

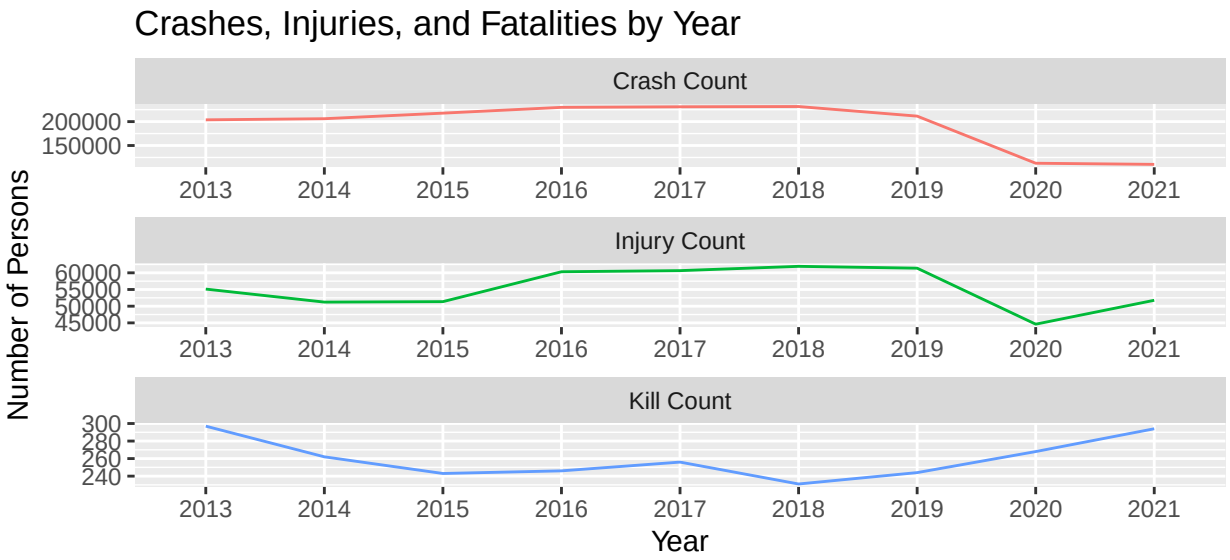
Data Sources

The primary dataset in this analysis contains approximately 1.9 million observations of motor vehicle collisions in New York City from 2012 to April 2022. The 29 columns include information like date, time, contributing factors, vehicle types, borough, zip code, and other details about each crash. This information comes from the New York Police Department: when a collision that causes an injury, a death, or more than $1000-worth of property damage, the New York Police Department files a Police Accident Report with a form MV-104AN (Police Crash Report Submission Instructions, p. 1). The information in the primary dataset comes from these forms.

The following report includes comparisons with other data sources, including comparisons with national car crash data, and secondary datasets that add another dimension to our primary dataset. Secondary data included to provide context and background include records of New York City moving violation summonses and New York City streets' speed limits, and secondary data included more comparison include New York state speeding ticket records and information on distracted driving across the United States.

First Look at Primary Dataset

The following plot shows the number of traffic collisions, injuries, and fatalities in recorded in the primary data set. Though our secondary data does not include comparably detailed numbers for national motor vehicle collision data that these trends could be tested against, the decrease in crashes and injuries in 2020 and the increase in fatalities reflect the national trends during the pandemic (NHTSA 2022b). Though New York City's traffic environment is not representative of the United States', the consistency in these trends suggests that analysis of New York City's traffic data may provide insight into pandemic era changes that have affected fatality rates and driver behavior.

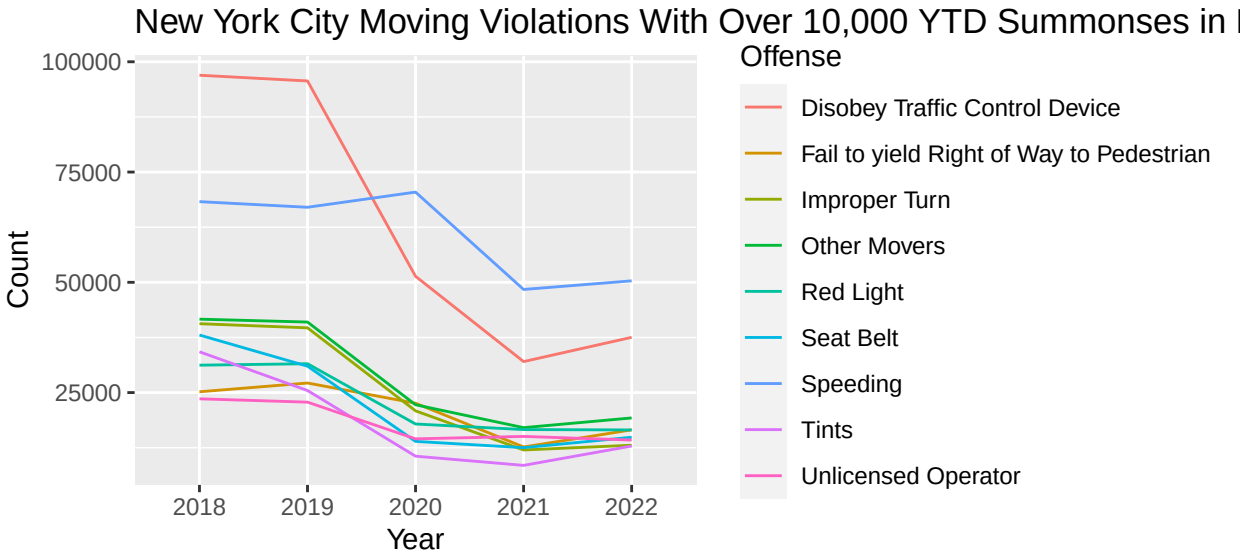


Context

The primary data set used in this report, like any large collection of non-experimental observations, does not only or wholly represent car crashes in New York City. Because information on each collision comes from a police report, the data reflects policing practices and civilian reporting practices. Though Section 605 of New York's Vehicle and Traffic Law makes failure to report an accident or failure to give correct information in connection with an accident report a misdemeanor (Police Crash Report Submission Instructions, p. 2), it is unlikely that all accidents are reported. A 2010 phone survey in Washington D.C. estimated that approximately 30 percent of motor vehicle collisions, which tend to be less severe, are not reported to the police (Davis & Co. 2015). While Washington D.C. reporting rates may differ from New York City reporting rates, New York City drivers may not always report car accidents, and their patterns of reporting are likely correlated with variables such as injury, value of damage, and certain contributing factors.

Additionally, because an officer at the scene of the collision is typically responsible for filling out the Police Accident Report form, policing patterns likely affect which crashes are recorded. For example, if the police department polices certain areas more heavily, they may file more reports for collisions in those areas while collisions in other areas may go unreported at higher rates.

To demonstrate how reporting practices and policing patterns affect outcomes, a plot of the most common New York City moving violation summonses by year is included below. Because this plot uses data through May 2022, each point uses data from January through May of its given year to maintain consistency. Summons data, which which follows from report data, demonstrates to a higher degree the strengths and weaknesses of data from police reports. The summons data, even more so than the primary data set, does not represent all collisions in New York City but instead demonstrates the laws and practices that lead to consequences for collisions. For example, despite the spiking traffic fatalities in 2021, the total number of tickets issued decreased substantially. While this decrease is partially attributable to a reduction in the number of cars on the streets during the pandemic, New Yorkers like City Councilman Ben Kallos point to reduced traffic enforcement (Hu 2021). The NYPD blame reduced enforcement on officer illness, a hiring freeze, and officer assignment to cover protests (Hu 2021). The numerous factors that contribute to traffic collisions, fatalities, and reporting, especially without data on traffic volume, require caution in interpreting the causes of pandemic-era trends.



Analysis on Reasons leading to Crashes

Reasons for crashes

The WHO reported that Every year the lives of approximately 1.3 million people are cut short as a result of a road traffic crash. Between 20 and 50 million more people suffer non-fatal injuries, with many incurring a disability as a result of their injury. ## The top risk factors behind these crashes include:

- Speeding
- DUI
- Distracted Driving
- Nonuse of helmets/seatbelts
- Unsafe road infrastructure/vehicle
- etc.

We specifically looked at **Speeding** and **Distracted Driving** in our research due to the limitation of our dataset.

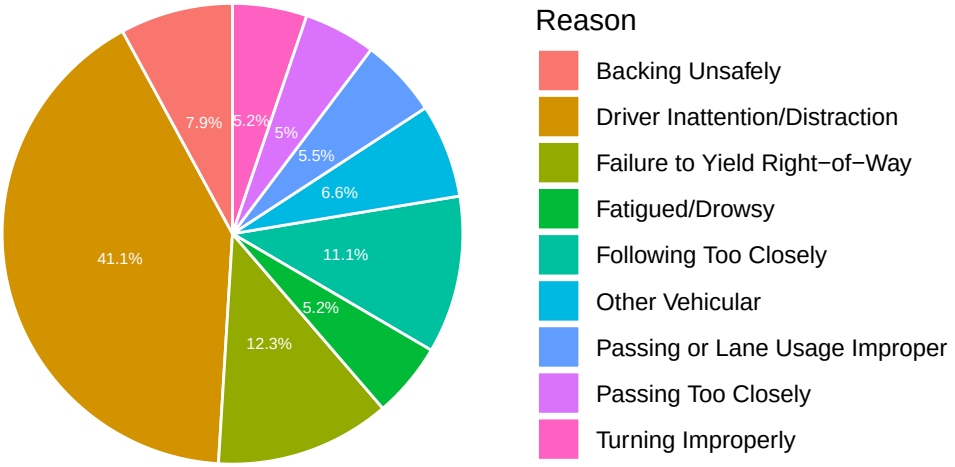
Overview of the reasons in our dataset

We took the top 9 reasons for crashes from our dataset. Some of their corresponding values are as follow:

Reason	Value	Percentage
Driver Inattention/Distraction	369831	41.1%
Failure to Yield Right-of-Way	110559	12.3%
Following Too Closely	99495	11.1%
Backing Unsafely	71268	7.9%
Other Vehicular	59175	6.6%
Passing or Lane Usage Improper	49801	5.5%
Fatigued/Drowsy	47202	5.2%
Turning Improperly	47013	5.2%
Passing Too Closely	45348	5%

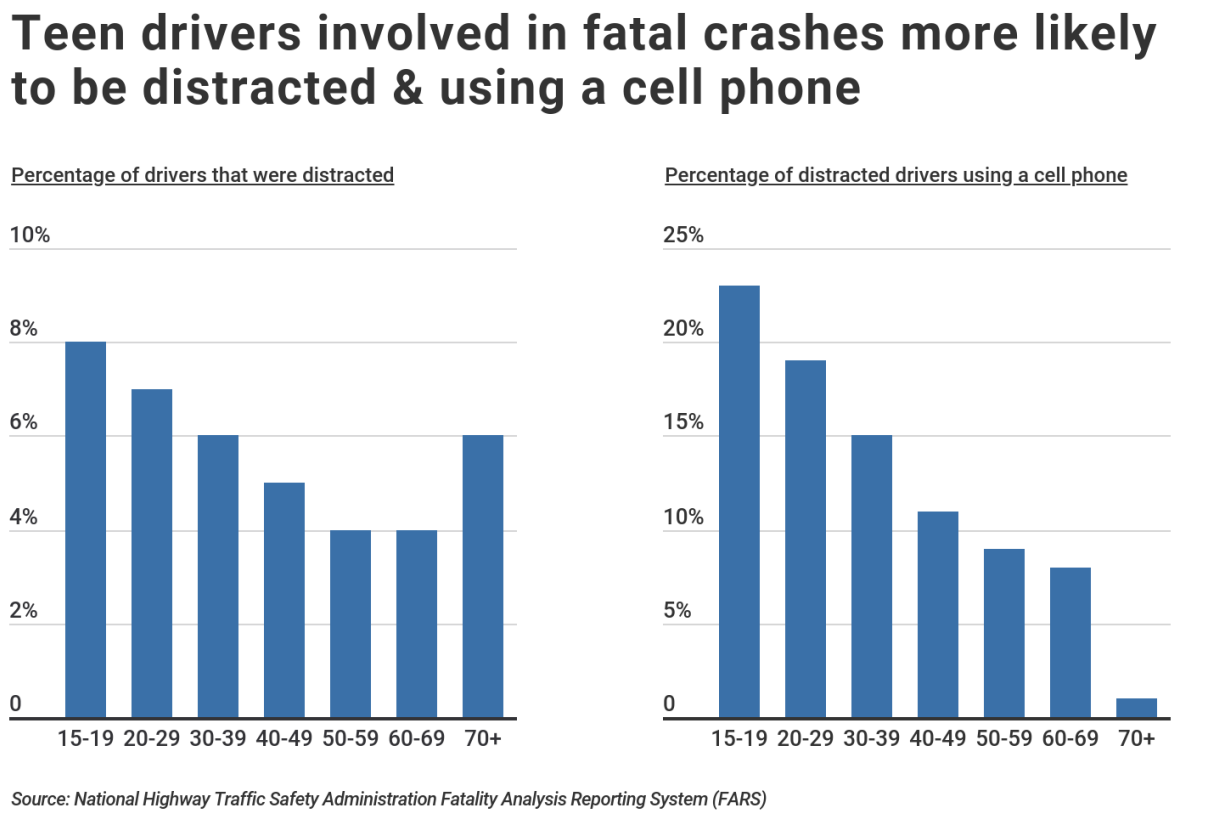
Distracted Driving | Our own graph

From our table, we decided to plot the percentages into a pie chart. As we can see, “distracted driving” has the highest occurrence with 41% of all crashes.



Distracted Driving | External data

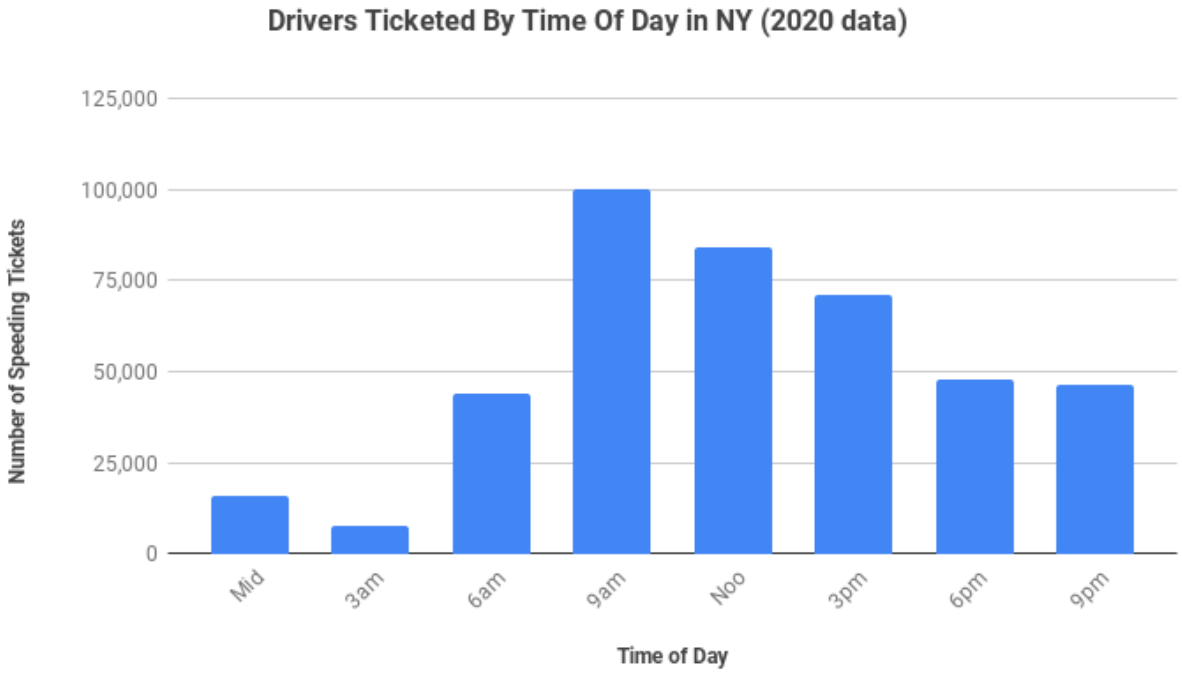
On the other hand, we found a graph that describes the situation on distracted driving nationwide in 2017.



From the graph on the left, if we add up all the percentages from all age groups, we would actually get roughly 40%. This shows that distracted driving may well become the biggest reason for road crashes, taking up almost half of all occurrences of crashes in both cases.

Speeding | External data

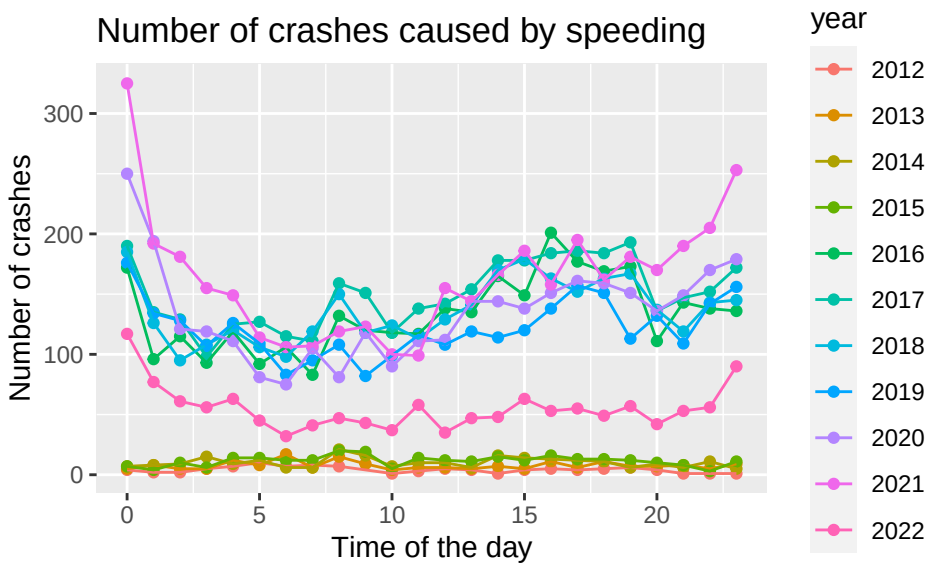
From *New York Speeding Ticket Data*, The article analyzed data from the New York State Department of Motor Vehicles to determine the most common times and locations that police issue speeding tickets. The data extracts the speeding ticket count from different times in New York from 2009 to 2020.



From the graph, we can see that there is a maximum number of tickets handed out at around 9am. This may be due to the morning rush hour, where people tend to drive faster in order to get to work on time.

Speeding | Our own graph

We extracted all the data with "speeding" as the contribution factor of the crash, and then we plotted the count of the crashes at each time of the day.



As we can see, there appears to be an increasing trend since 6am until 9am, where most of the trend hit a local max at around 8 or 9am in our graph. Recall that we see a maximum number of speeding tickets

	One Vehicle	Two Vehicles	Three Vehicles	Four Vehicles	Five Vehicles
Cumulative Amount of Vehicles	1883213.00	1558524.00	127272.00	28120.00	7535.0
Cumulative Percentage of Total Collisions Recorded	99.43	82.28	6.72	1.48	0.4
Amount of Vehicles	324689.00	1431252.00	99152.00	20585.00	7535.0
Percentage of Total Collisions Recorded	17.14	75.56	5.23	1.09	0.4

handed out at around 9am, which implies there may be more occurrences of speeding. Here, the graph seems to imply that more speeding can potentially cause more crashes.

Although we find some interesting trends for the reasons behind the crashes, more rigorous models need to be implemented to further investigate the correlation, as we did in our later research.

Analysis on Types of vehicle involved in Crashes

In this section, we will analyze the types of vehicles that are involved in the collisions.

Amount of Vehicles in collisions

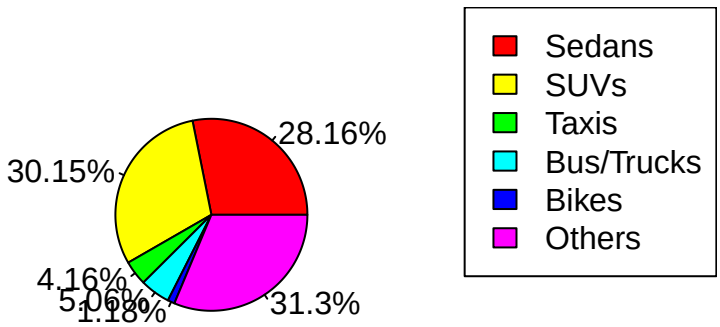
We first take a look at the amount of vehicles that are involved in each crash. The table below records the cumulative and individual percentage and amount of vehicles involved in the crashes. For example, 82.24% of the crashes recorded involve at least two vehicles while 75.51% of the crashes recorded involves exactly two vehicles. To no surprise, two vehicle crashes takes up the majority of the crashes recorded.

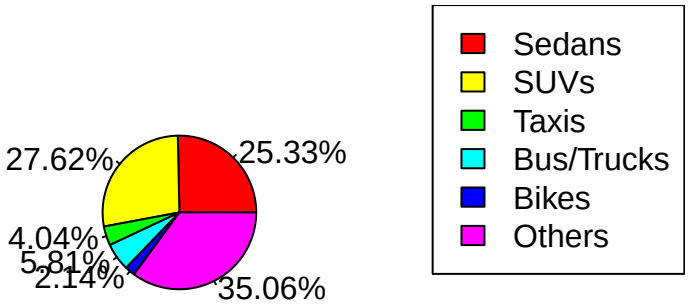
Please note that 99.42% of the crashes recorded involved at least one vehicle. This suggests that 0.58% of crashes in the data did not have the amount of vehicles record since logically, at least one vehicle should be involved in a collision.

Types of Vehicles

To categorize the type of vehicles that were involved in the crashes, we divided the types of vehicles into 5 main categories:
- Sedans
- SUVS
- Taxis
- Trucks
- Bikes

Due to the majority of the data being one or two car accidents, we present two pie charts that show the distribution of the first and second vehicle involved in each crash. The largest categories of vehicles involved in crashes were SUVs and Sedans.

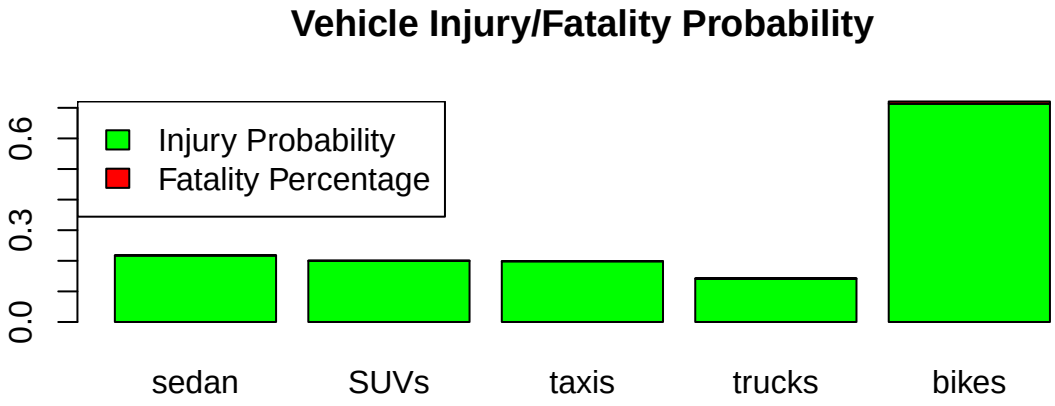




Injury and Fatality Percentage

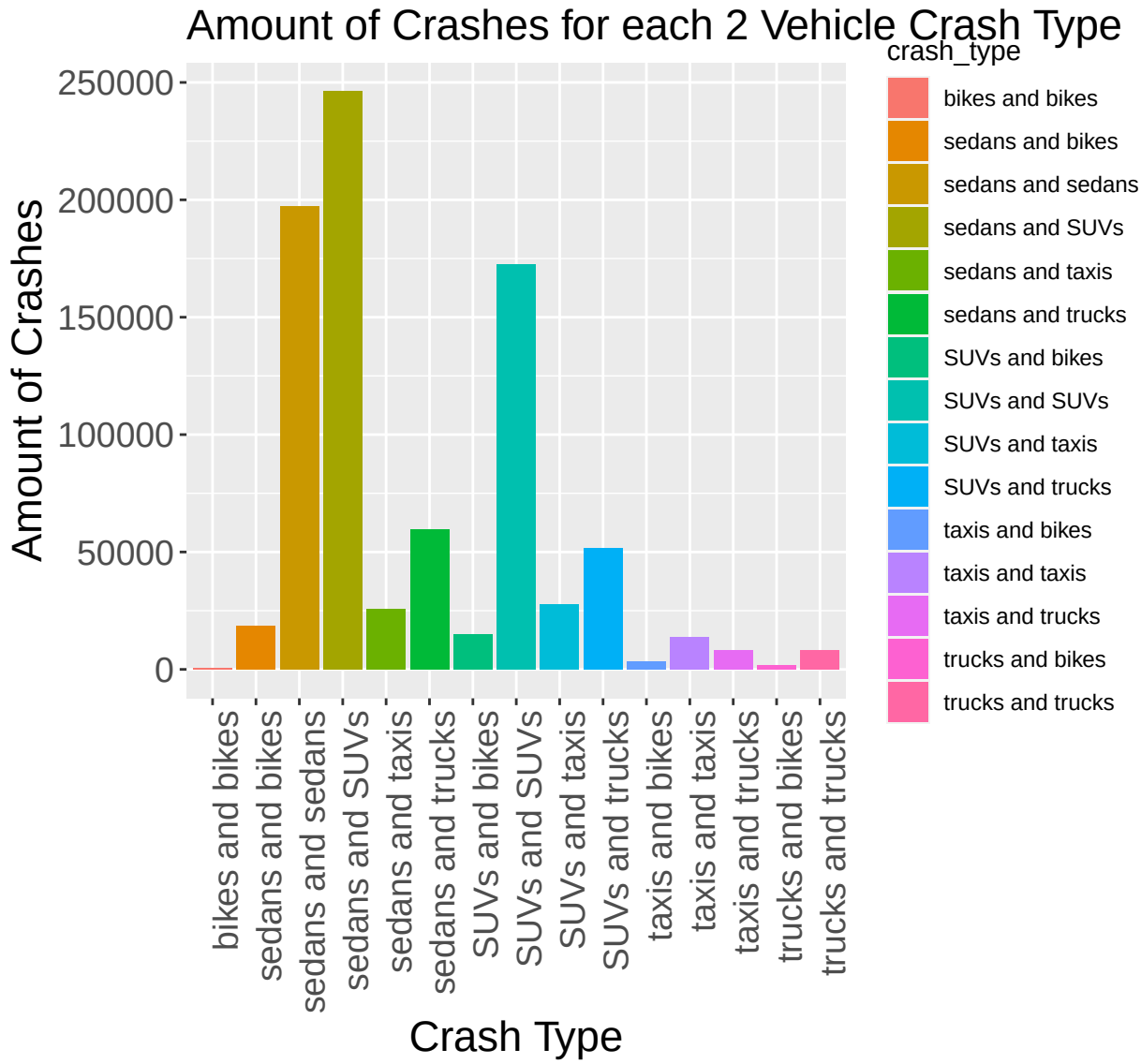
The below bar graph shows the percentage of drivers for each type of vehicle who experienced a injury or a fatality. While the fatality rate for most crashes is less than 1%. Bikes and motorcycles, who have the

highest injury rate by a great margin had significant fatality rate of over 1%.



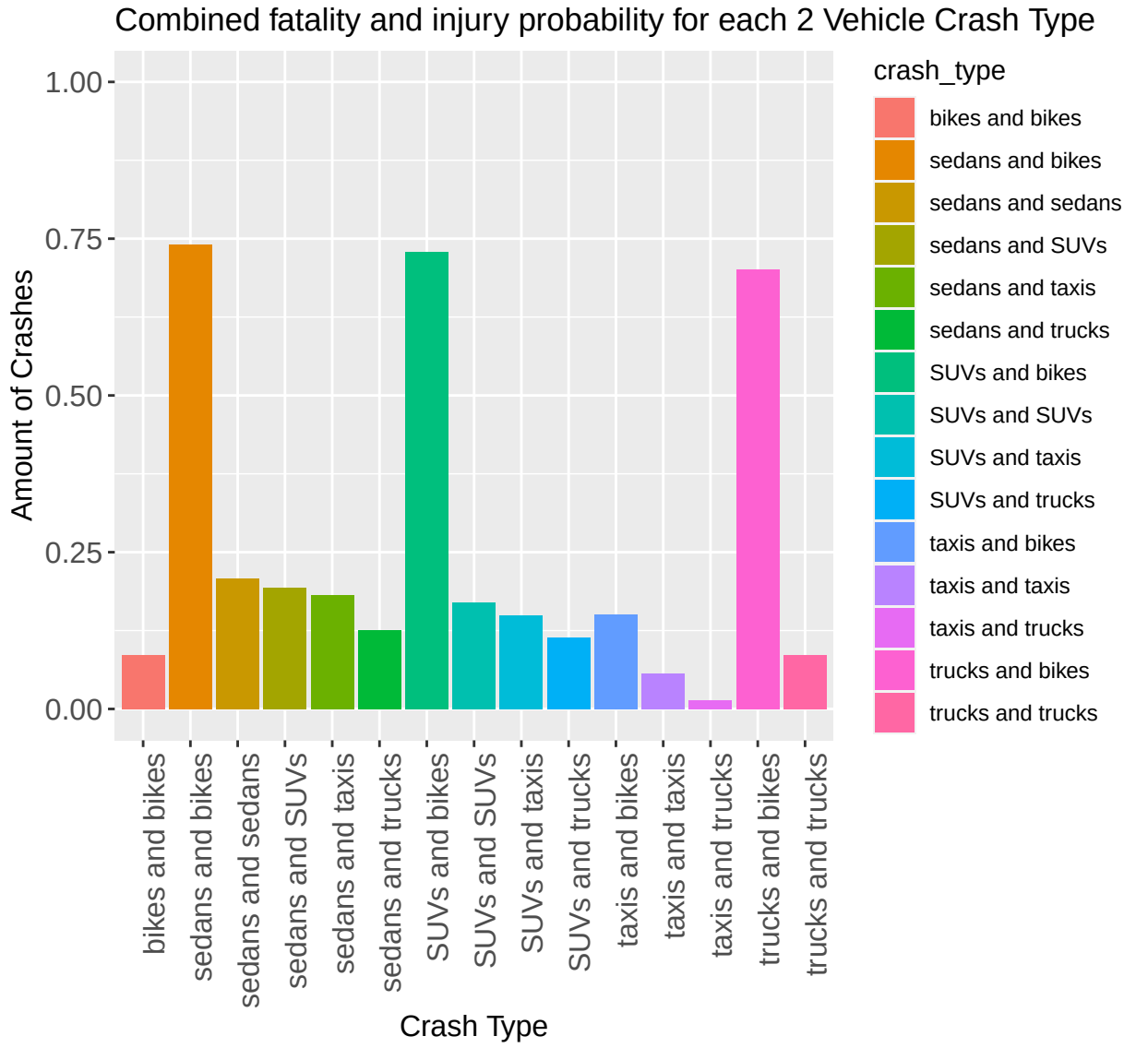
Vehicle Distribution in Two Vehcile Collisions

Since, two vehicle collisions appeared most frequently, we analyzed further on the distribution of vehicles that are involved in two vehicle crashes specifically. Similar to the overall distribution of vehicles, SEDANs and SUVS dominate the total amount.



Two Vehicle Collision Injury Rate + Mortality Rate

In order to further analyze the effects of two vehicle collisions, we analyzed the injury rate and the mortality rate by adding them together. This is because the mortality rate was very low in the previous graph, so it is practically not significant. Crashes with motorcycles still remain with the highest injury and fatality rate.



Final word on the type of vehicles.

This data set did not provide anything extraordinary or completely surprising. However, it did not further consolidate the idea that SUVs and Sedans will have the most crashes but with a medium fatality + injury rate while bikes and motorcycles would be the opposite.

Regression Analysis and Killer Plot

Motivation and Variables for Regression

As discussed in previous sections, analysis of our data corroborates past studies in that it suggests the existence of correlations between variables in our data and the probability of getting injured or killed during a car crash. We are interested in validating if these correlations are potentially casual, and if so, what insights they can give to enhance safety in the street.

We will be focusing on "Injury and Fatality" as our response variable. This variable is a binomial variable that holds the value 0 if there are no person injured or killed, or the value 1 if there's at least one person injured or killed, as a result of a car crash. For the predictor variable, we will be investigating "Vehicle Types" and "Crash Factor/Reason", which are both categorical variables provided in our main dataset.

Model in Use

For our regression analysis, we used a binomial logistic regression model in which the log odds are of a generalized linear regression form. The major reason we picked this variable is because our response variable is binomial. We can observe no multicollinearity in our predictor variables because both "Vehicle Types" and "Crash Factor/Reason" are categorical. There also aren't any autocorrelation because of the nature of car crashes. For this model, we are also assuming linearity of our predictor variables and log odds

Logistic Regression Model Explained

Before we delve into our regression process, let us first look at the structure of the logistic regression function. Using "Injury and Fatality" as Y and "Vehicle Types" or "Crash Factor/Reason" as X, we have the following function for explaining Y using X, where $x_1...x_n$ are individual cases for the categorical variable X.

$$Y = e^{a+BX} \rightarrow log(Y) = a + BX = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

Given this expression, we deduct the value of Y, or the probability of at least one injury or death occurring, to be:

$$p = \frac{1}{1 + e^{-logit(p)}}$$

where the logistic function is given by:

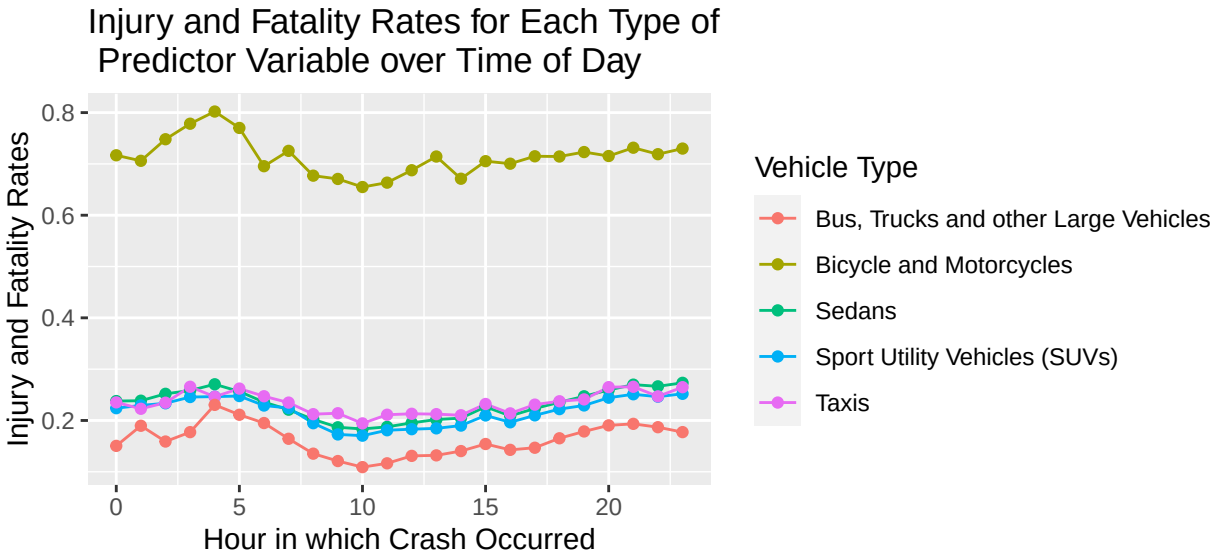
$$logit(p) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$$

Regression Analysis on Vehicle Type

The regression we did using vehicle type as the predictor variable has the following categories. Where the value "Buses, Trucks and other Large Vehicles" is explained by the constant term.

Predictor Variables	Categories
	Bus, Trucks and other Large Vehicles
x_1	Bicycle and Motorcycles
x_2	Sedans
x_3	Sport Utility Vehicles (SUVs)
x_4	Taxis

We split the data into groups based on the hour the crash occurred and calculated the rate of injuries and fatalities for each hour. This gives us enough data to perform our regression. The below is a graph that demonstrates the injury and fatality rates for each vehicle type according to time of day. We can observe that the injury and fatality probability of crashes involving a bicycle or motorcycle is highest throughout the day.



Regression Result for Vehicle Type:

The result of the regression is given in the table below. We can see that the p value for all of the coefficients, including the constant term, exhibit statistical significance, indicating casual relationships between vehicle type and injury and death.

```
##                         Estimate  Std. Error    z value Pr(>|z|)
## (Intercept)          -1.7646364 0.008527412 -206.93691        0
## VEHICLETYPEcom_bikes  2.6398506 0.017000521  155.28058        0
## VEHICLETYPEcom_sedan  0.5214949 0.009138277   57.06709        0
## VEHICLETYPEcom_suv    0.4270004 0.009129603   46.77097        0
## VEHICLETYPEcom_taxi   0.5661234 0.012003765   47.16216        0
```

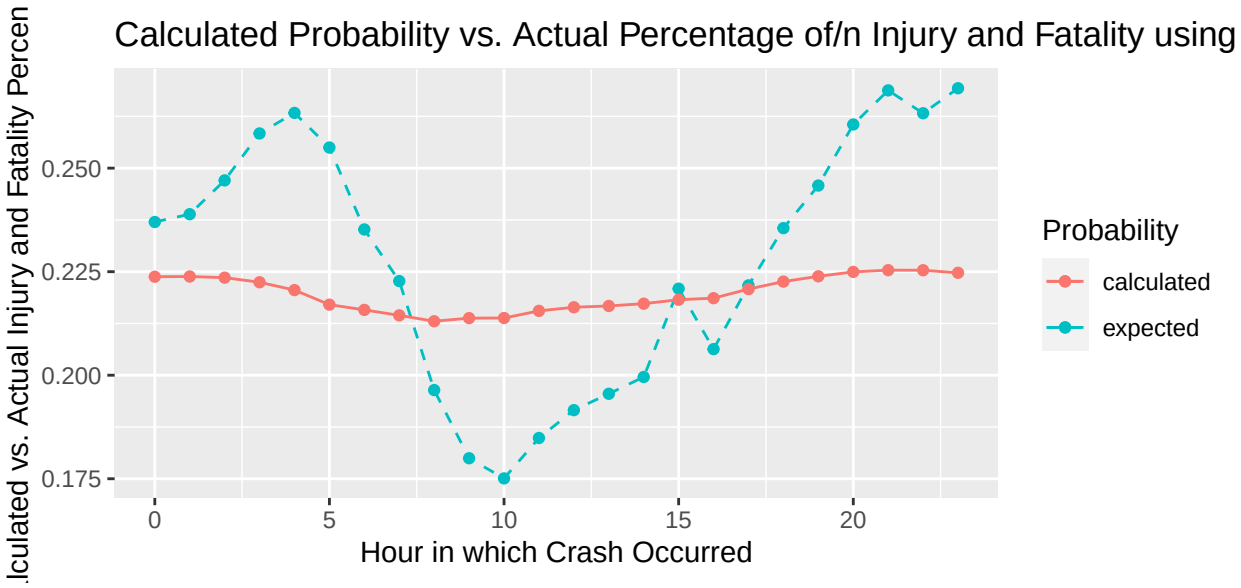
We can thus deduce the resulting predictor function to be:

$$p = \frac{1}{1 + e^{-logit(p)}}$$

where:

$$logit(p) = -1.764636 + 2.639851 * x_1 + 0.521495 * x_2 + 0.427000 * x_3 + 0.566123 * x_4$$

The graph below demonstrates a comparison between the actual and predicted overall probability of an injury or death occurring using the regression model we fitted above at each time of the day. As we can observe, though this model is far from being a perfect match due to the simplicity of its nature, it is able to capture the general trend of injury and death probabilities falling during the middle of the day and rising during night time.

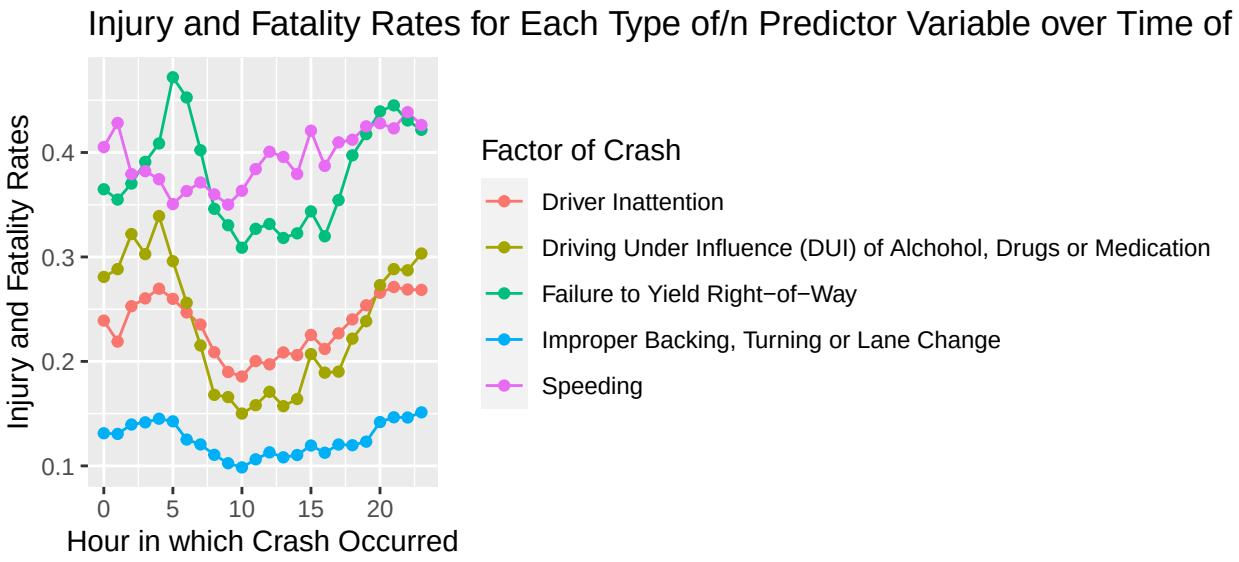


Regression Analysis on Factor of Crash

Similar to that above, the regression using vehicle type as the predictor variable has the following categories. Where the value "Driver Inattention" is explained by the constant term.

Predictor Variables	Categories
	Driver Inattention
x_1	Driving Under Influence (DUI) of Alchohol, Drugs or Medication
x_2	Failure to Yield Right-of-Way
x_3	Improper Backing, Turning or Lane Change
x_4	Speeding

We performed similar operations on the predictor variable "Factor of Crashes". The below is a graph that demonstrates the injury and fatality rates for each vehicle type according to time of day. We can observe that the injury and fatality probability of crashes involving "Failure to Yield Right of Way" and "Speeding" are relatively higher; that of "Improper Backing, Turning or Lane Change" is relatively low, partially explainable by the low speed of involved vehicles at which these crashes occur.



Regression Result for Factor of Crash:

The result of the regression is given in the table below. We can see that the p value for all of the coefficients, including the constant term, exhibit statistical significance, indicating casual relationships between factor of crash and injury and death.

```
##                                  Estimate  Std. Error     z value
## (Intercept)                   -1.23276809 0.003943362 -312.618534
## CRASHFACTORDUI                 0.08367509 0.012947477    6.462656
## CRASHFACTORFailure to Yield    0.67002133 0.007423518   90.256582
## CRASHFACTORImproper Maneuver -0.76852317 0.007885262  -97.463242
## CRASHFACTORSpeeding            0.81611278 0.014310457   57.029123
##                                         Pr(>|z|)
## (Intercept)                   0.0000000000000000
## CRASHFACTORDUI                0.0000000001028811
## CRASHFACTORFailure to Yield   0.0000000000000000
## CRASHFACTORImproper Maneuver 0.0000000000000000
## CRASHFACTORSpeeding           0.0000000000000000
```

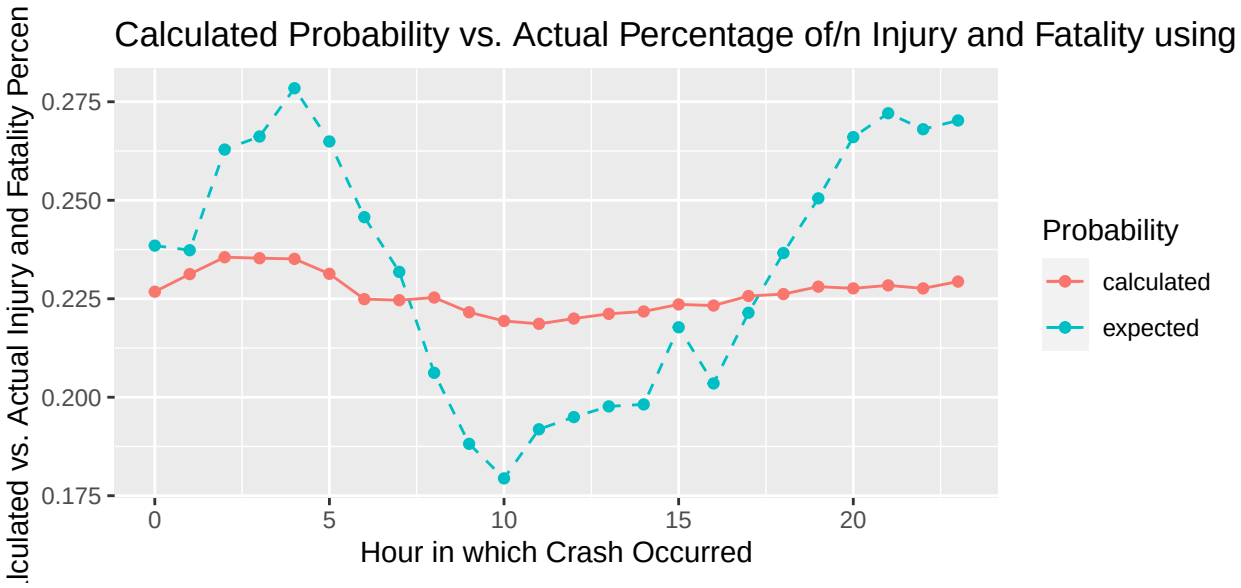
We can thus deduce the resulting predictor function to be:

$$p = \frac{1}{1 + e^{-logit(p)}}$$

where:

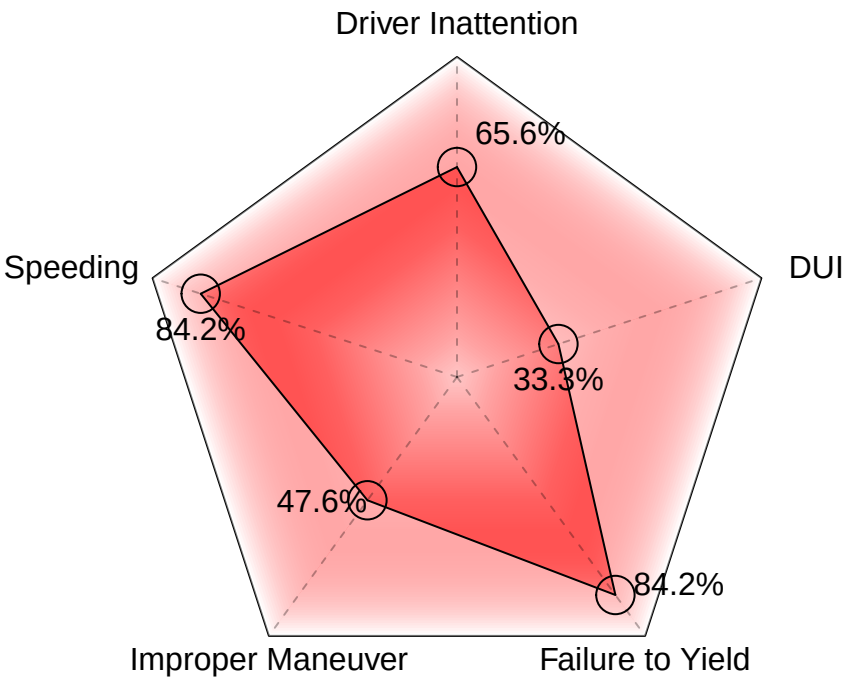
$$logit(p) = -1.2208663 + 0.0721312 * x_1 + 0.6825204 * x_2 - 0.7687210 * x_3 + 0.8121122 * x_4$$

We can see that for factor of crash, we can also identify a general trend of injury and death probabilities falling during the middle of the day and rising during night time using the fitted regression model.



Killer Plot

Resulting our analysis are realizations of the significance that external factors play in determining the outcome of a crash. We are thus prompted to design a method for drivers to anticipate the potential injury and death probability of crashes based on our data set so as to advise them to make wiser decisions. As such, we divised our killer plot displayed below.



Killer Plot For Bike and Motorcycles at 10 O'clock

Our killer plot is pentagonal shaped, with each angle representing one potential factor of crash. With in the pentagon there's another irregular pentagon. The distance between each vertex of the internal pentagon and its corresponding external pentagon represents the probability of an injury or death occurring for this particular factor of crash. For this particular graph, we are using the specific vehicle type "Bike and Motorcycles" and the time of day "10 O'clock". Thus, for example, we can infer from the graph that if you ride a bike or motorcycle at 10 O'clock in New York and get crashed due to inattention, you have a 65.6% probability of getting injured or killed. Note that we are only displaying one specific case and we can change the vehicle type and time of day to meet the needs of the reader. Our plot provides a means for drivers to know in advance what to be ware of before they hit the road. Continuing on our provided example, bike riders should certainly be cautious of speeding and failure of yielding right at 10 O'clock in the morning as these both yield high percentages of injuries and fatalities if a crash occur. The credibility of this plot is given by the statistically significant causal relationship we previously analyzes and the considerable size of our data set.

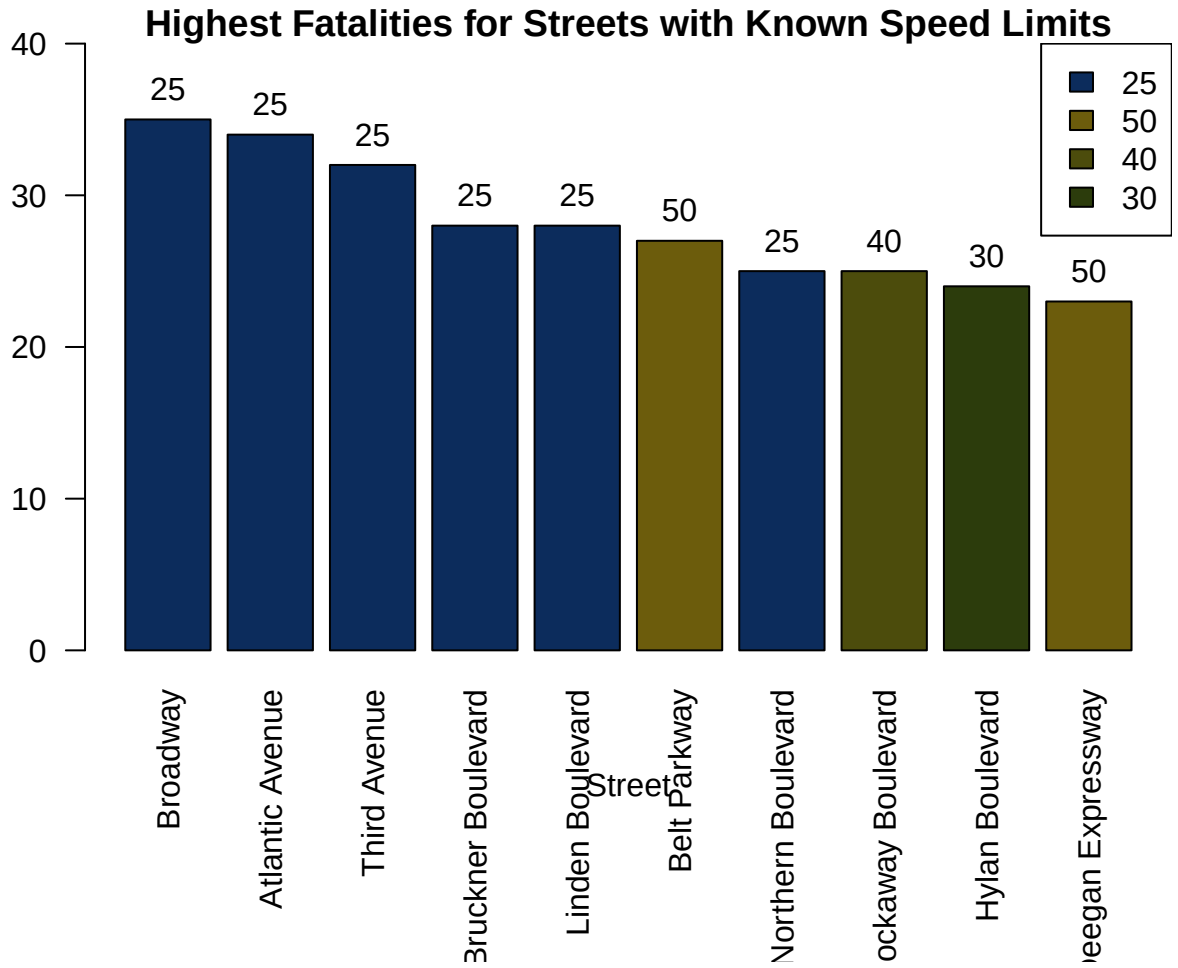
Limitations

As highlighted in the **Context** section of this report, factors that determine which crashes are reported and which crashes lead to consequences for drivers complicate interpretation of the primary data set, but these factors also capture meaningful information about policing and driver behavior. While these factors may be useful for analyses combining them with other data, the New York City motor vehicle collisions data set has limitations caused both by these factors and lack of recorded data. One unambiguous limitation of the

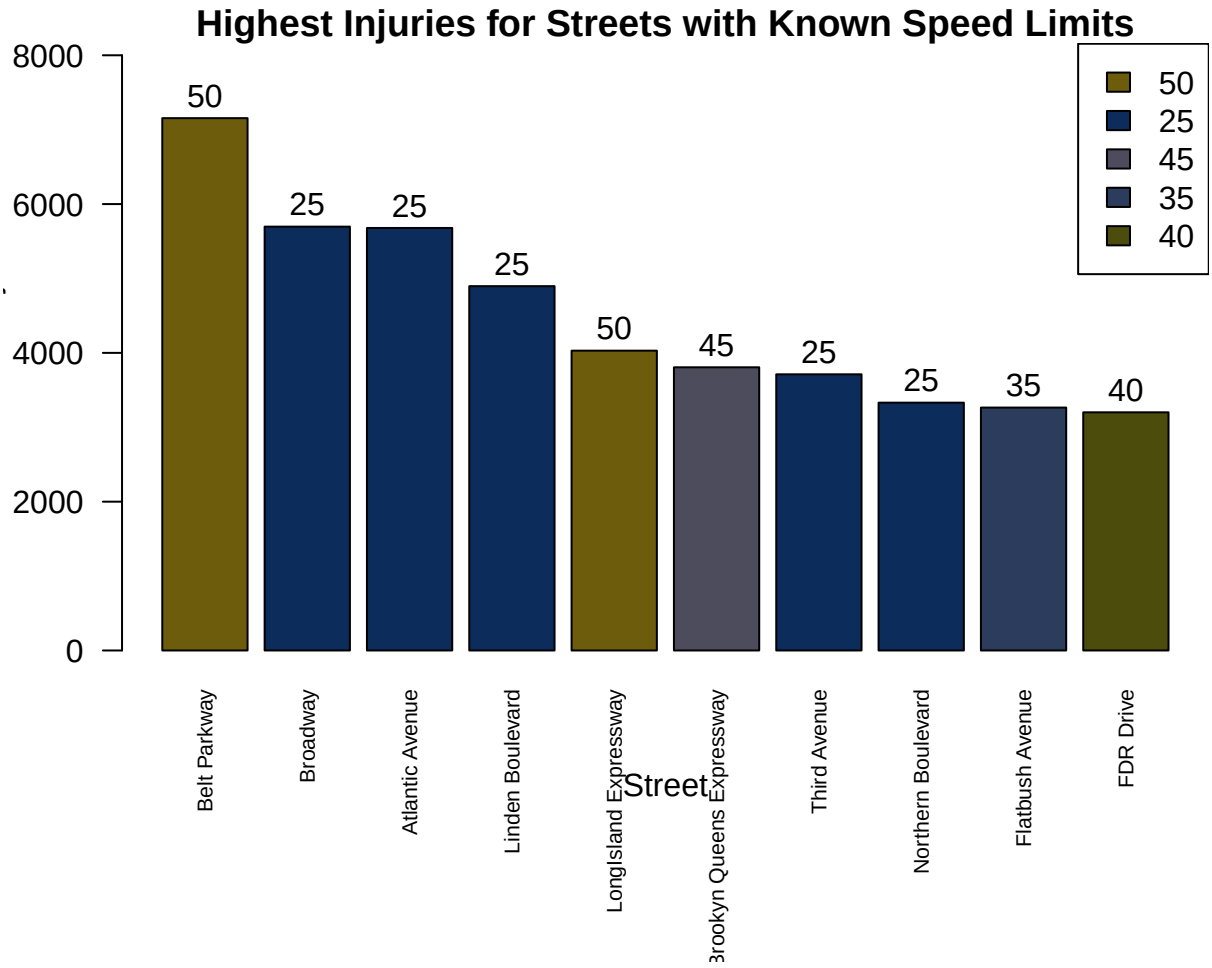
data set is the inconsistency in location variables. The columns include latitude, longitude, zip code, the street the crash in on, the street the crash is off, the cross street, coordinates, and the borough, but many observations only include data for a few of these location variables. Making the street and cross street of one collision compatible to compare with the borough or latitude and longitude of another would require several additional data sets and geographic coding.

Additionally, the primary data set includes five columns for contributing factors rather than having separate columns for each factor that could contribute. As a result, factors are only recorded if the reporting officer or a person involved in the crash identifies a factor as contributing. For example, if the identified cause of the crash was a car turning improperly, the report may not include information about the speed of the car or the weather conditions.

Included below are two plots illustrating the speed limits for the New York City streets with the highest numbers of traffic injuries and fatalities. These plots demonstrate that recorded data like speed limit does not capture the level of risk on a street: the streets with the highest number of injuries and fatalities are not those with high speed limits. In fact, the streets with the numbers of fatalities have relatively low speed limits, suggesting that further analysis may be necessary to determine if cars traveling at high speeds amongst traffic moving at low speeds may be a frequent cause of fatal accidents. Factors not included on these plots, like speeding, reckless driving, and the volume of traffic on these streets, contribute to injury and fatality numbers, and the speed limits do not represent the actual speeds of the cars involved in these collisions. These plots demonstrate the limitations of data that does not capture each crash-involved vehicle's speed or the traffic volume while also raising questions about what additional data besides Police Accident Report data is necessary to inform policy changes like reductions in speed limit.



Speed Limit and Injuries



Works Cited

CDC. "Crash Deaths in the US: Where We Stand." Centers for Disease Control and Prevention, 18 July 2016, https://www.cdc.gov/vitalsigns/motor-vehicle-safety/index.html.

Hu, Winnie. "De Blasio Vowed to Make City Streets Safer. They've Turned More Deadly." The New York Times, 30 Sept. 2021. NYTimes.com, https://www.nytimes.com/2021/09/30/nyregion/traffic-deaths-nyc.html.

Hu, Winnie. "De Blasio Vowed to Make City Streets Safer. They've Turned More Deadly." The New York Times, 30 Sept. 2021. NYTimes.com, https://www.nytimes.com/2021/09/30/nyregion/traffic-deaths-nyc.html.

Hu, Winnie. "New York Streets Are Nearly Empty, but Speeding Tickets Have Doubled." The New York Times, 16 Apr. 2020. NYTimes.com, https://www.nytimes.com/2020/04/16/nyregion/coronavirus-nyc-speeding.html.

M. Davis & Co. National Telephone Survey of Reported and Unreported Motor Vehicle Crashes. no. Findings Report. Report N. DOT HS 812 183)., July 2015, p. 244.

Newly Released Estimates Show Traffic Fatalities Reached a 16-Year High in 2021 | NHTSA. 2022a, https://www.nhtsa.gov/press-releases/early-estimate-2021-traffic-fatalities.

"New York Speeding Ticket Data." Rosenblum Law, Nov. 2021, https://traffictickets.com/new-york/traffic-tickets/speeding-tickets/data/, https://traffictickets.com/new-york/traffic-tickets/speeding-tickets/data/.

NHTSA Releases 2020 Traffic Crash Data | NHTSA. 2022b, https://www.nhtsa.gov/press-releases/2020-traffic-crash-data-fatalities.

NYPD. Motor Vehicle Collisions - Crashes | NYC Open Data. NYC OpenData, May 2022, https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95.

Routhier, Sarah. "Teen Drivers and Texting [Best and Worst States]." CarInsurance.Org, 6 June 2022, https://www.carinsurance.org/teen-driver-phone-use-study/.

Traffic Data - Archive 2018. 2018, https://www1.nyc.gov/site/nypd/stats/traffic-data/traffic-data-archive-2018.page.

Traffic Data - Archive 2019. 2019, https://www1.nyc.gov/site/nypd/stats/traffic-data/traffic-data-archive-2019.page.

Traffic Data - Archive 2020. 2020, https://www1.nyc.gov/site/nypd/stats/traffic-data/traffic-data-archive-2020.page.

Traffic Data - Archive 2021. 2021, https://www1.nyc.gov/site/nypd/stats/traffic-data/traffic-data-archive-2021.page.

Traffic Data - Summonses. 2022, https://www1.nyc.gov/site/nypd/stats/traffic-data/traffic-data-moving.page. Police Crash Report Submission Instructions. p. 102.

"VZV_Speed Limits." NYC Open Data, https://data.cityofnewyork.us/Transportation/VZV_Speed-Limits/7n5j-865y. Accessed 22 June 2022.