New York's Streetside Casualties

An explanatory analysis of NYC car accidents

Drew LoPolito and Mitch Harrison

Introduction

Project Motivation

Citizens of large cities all over America suffer injury and death in motor vehicle crashes. In New York, motor vehicle accidents are among the top five reasons for hospitalizations statewide. We hope that motivated policymakers in NYC and other metropolitan areas could use our explanatory models to craft traffic policy, shift police resources, better target traffic citations, and turn our insights into potentially lifesaving urban development and planning.

Dataset

Our dataset is composed of harvested and compiled data from New York City Police Department (NYPD) open access data on all police reported motor vehicle collisions (MVC) in all five boroughs of New York City from July 1st, 2012 through April 24th, 2023. The police report from which individual MVC observations in our dataset hail (MV104-AN) is required to be filled out for MVC where someone is injured or killed, or which result in at least $1,000 of total property damage.

We created a number of new variables by manipulating the dataset, as well as re-categorizing/cleaning some of the existing variables for practical use in modeling (for example, the original variables for factors contributing to the accident for each motorist and vehicle type of each motorist contained roughly 100 categories).

The following are the variables of interest from our dataset, with new or re-categorized variables noted:

`has_casualty` (New): a binary variable corresponding to whether a MVC resulted in at least one casualty. This variable was generated from the original `number_of_persons_injured` and `number_of_persons_killed` variables. **This is our response variable of interest.**

`weekend_weekday` (New): a binary variable corresponding to whether a MVC occurred during the week or weekend. This variable was generated from the original `crash_date` variable.

`yday` (New): a numeric variable ranging from 1 to 365 corresponding to the numerical day of the year on which the MVC occurred. This variable was generated from the original `crash_date` variable.

`time_day` (New): a categorical variable with levels of "morning" (5 AM to 12 PM), "afternoon" (12 PM to 5 PM), "evening" (5 PM to 9 PM), and "night" (9 PM to 5 AM) corresponding to the time of day at which a MVC occurred. This variable was generated from the original 24-hr `crash_time` variable.

`vtype1` and `vtype2` (Recategorized): categorical variables corresponding to the type(s) of each vehicle involved in the crash, with categories of "Passenger vehicles," "Commercial vehicles," "Motorcycles," "Non-Motor Vehicle," "Other/Unknown," and "None" (only applies to `vtype2`).

`factor1` and `factor2` (Recategorized): categorical variables corresponding to any factor(s) which potentially contributed to the crash for respective vehicles, with categories of "Aggressive/Reckless Driving," "Failure to Obey Traffic Signs/Signals/Rules," "Impairment/Distraction/Fatigue," "Performance-unrelated Technical/Mechanical Factors," and "Other/Unknown."

Our primary research concern is determining which characteristics of a MVC's timing and participants involved make casualties more likely.

Data Cleaning

Notably, only one MV104-AN form is filled out for all involved in an accident, meaning each observation in our dataset represents a unique MVC.

Our initial dataset contained approximately 1.9 million observations, found here. This was too large to push to git, so we sampled 10,000 observations from the dataset completely at random (**Appendix 1a**). The original dataset, as well as our randomly sampled dataset, contained data on crashes involving 1 to 5 motorists. However, 93% of crashes in the dataset occurred between 2 or fewer motorists (98% between 3 or fewer and 100% between 4 or fewer). Due to high levels of missingness in crashes with 3 or more motorists, as well as their low real-world frequency in New York City (where the kind of highway pile-ups which generate MVC with 3 or more motorists aren't generally observed), we decided to examine exclusively MVC between 2 or fewer motorists by removing all observations involving more than two vehicles, which brought our total number of observations down to approximately 9,300.

Next, we used the original counts of persons killed and persons injured to create a count of casualties (defined as injuries **and** casualties). We then created binary variables for MVC injury, fatality, and casualty, the last of which, is our response variable. We used these three variables to create a ordinal category of accident severity, with levels of "no casualties," "injury," and "fatal."

We also cleaned/manipulated the time data, which initially was composed of a date column and a 24-hr time column. We used this data to create a variable for MVC time of the day, day of the week, numerical day of the year (1-365), and weekday or weekend timing.

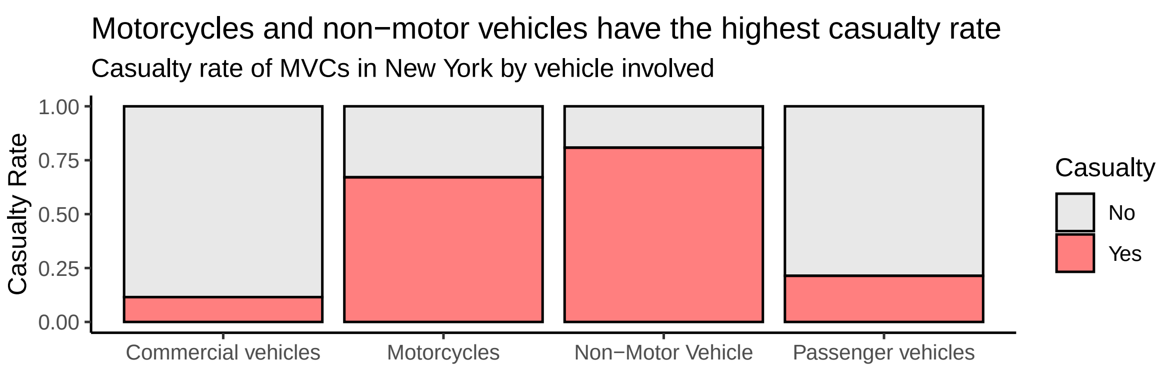
Next, we cleaned the original variables corresponding to vehicle types, involved in the accident as well as factors in each vehicle which may have contributed to the MVC. 453 MVCs contained an observation for second contributing factors, but not second vehicles, which we believe indicated not that there was a second vehicle but that the vehicle type hadn't been recorded, so we replaced those missing values with "unknown." Additionally, 59 MVCs had missing observations for all vehicles, which we removed since each MVC must involve at least one vehicle and this represented only 0.638% of the dataset.

Our initial dataset had 121 unique levels for vehicle type. In order to make these types interpretable for EDA and potentially in our model, we consolidated them into 5 larger categories: "Passenger vehicles," "Commercial vehicles," "Motorcycles," "Non-Motor Vehicle," and "Other/Unknown" (also a category of "None", which only applies to the second vehicle type).

The initial dataset also had 53 unique contributing factors. In order to make these types interpretable for EDA and potentially in our model, we consolidated them into 5 larger categories: "Aggressive/Reckless Driving," "Failure to Obey Traffic Signs/Signals/Rules," "Impairment/Distraction/Fatigue," "Performance-unrelated Technical/Mechanical Factors," and "Other/Unknown."

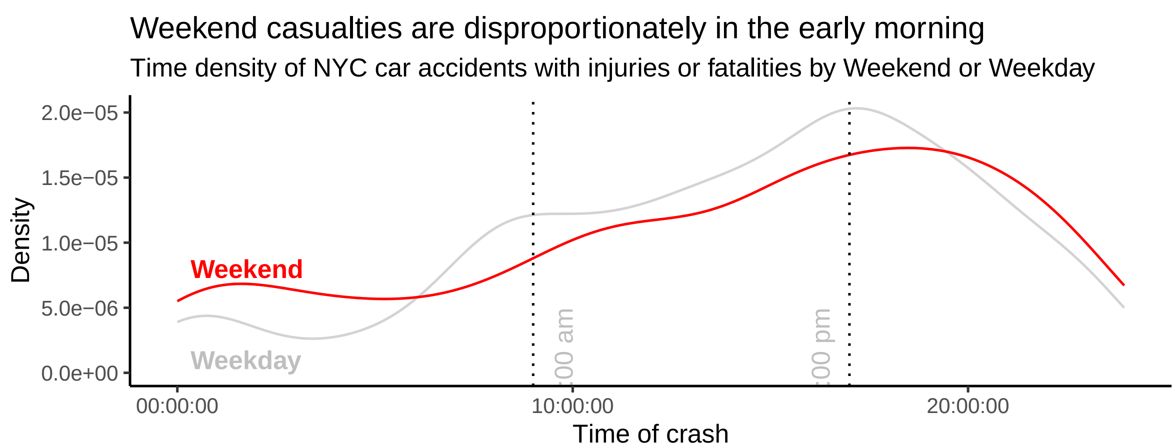
Exploratory Analysis

We can see from the visualizations below that vehicle type seems to have a significant impact on MVC casualty rate, particularly with regard to motorcycles and non-motor vehicles.

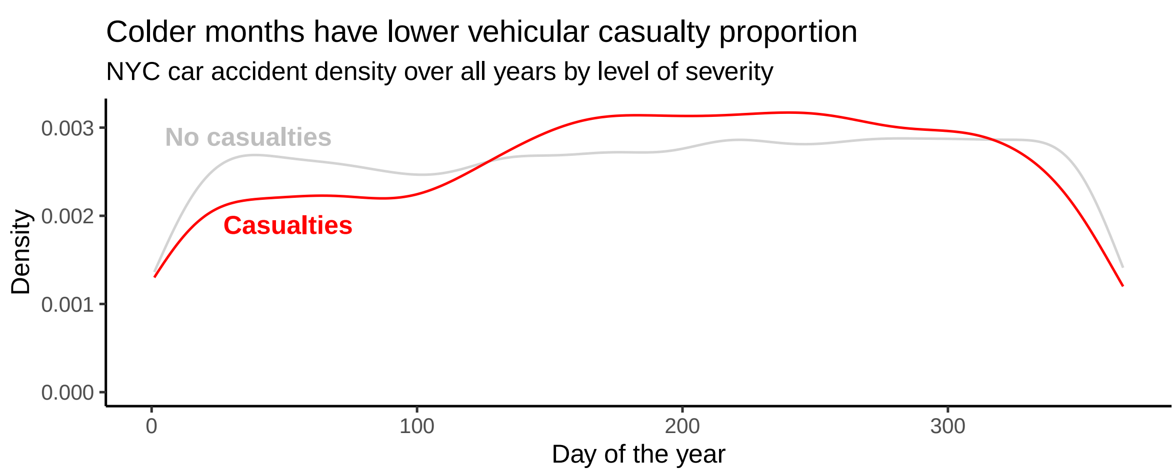


A similar analysis (**Appendix 2c**) shows no significant difference in casualty rates among categories.

To explore how timing might effect the occurrence of casualties, we explored the relationship between day of the week and casualty occurrence, finding that weekend MVC casualties were more heavily concentrated in the early morning hours than weekday MVC casualties. This was interesting as it provided some credence to the intuitive thought that early morning weekend drivers are more likely than early morning weekday drivers to be leaving parties/going out, therefore making them more likely to be impaired and get into a serious MVC. Additionally, the proportion of early morning (12 AM to 4 AM) weekend drivers with "Impairment/Distraction/Fatigue" listed as a factor was 0.2685617, fairly different from that of early morning weekday drivers, 0.1746032.



We also visualized densities of casualty occurrence and non-occurrence by numerical day of the year. We observed proportional casualty density to be lower in the colder winter months, and higher in the summer and fall.



Further analysis was done to explore whether zip code seemed to impact casualty rates (**Appendix 2a**), but this only revealed that these rates were approximately normally distributed. We also visualized MVC count by borough (**Appendix 2b**), seeing no significant trend other than high missingness for `borough` data.

Methodology

We intend to explore the factors related to timing and characteristics of drivers which might contribute to casualties in MVCs. Our outcome variable of interest, `has_casualty`, is a binary variable, thus the logical option in our case is to fit a logistic regression model to investigate our primary research concern.

We anticipate that the vehicle type involved in the MVC will play a significant role in whether or not a MVC has casualties, particularly in the context of the visualizations from our EDA. We will absolutely want to include the types of the vehicles involved in the crash, `vtype1` and `vtype2`, due to EDA as well as the intuitive fact that MVCs involving vehicles which leave a driver more exposed are far more likely to leave the driver with a casualty.

EDA showed less significant differences in casualty frequency between types of contributing factors to an accident (`factor1` and `factor2`) than it did between types of vehicles. Although we might expect that different contributing factors might contribute differently to accident severity (e.g., we would expect that driving drunk would be more likely to cause an accident with casualties than if someone fails to come to a complete stop at a stop sign), we are not completely sure that these factors will be useful in the fit of our model. Therefore, `factor1` and `factor2` will be *considered* in our model selection process, and we will treat them as a "package deal," since it stands to reason that if one of the vehicle factors is useful, it would be inappropriate to omit the other

We are very interested in the time status of MVCs and how that may affect casualty likelihood, including time of day (`time_day`), weekend or weekday (`weekend_weekday`), and numerical day of the year (`yday`). In our EDA, we saw that the density of MVCs with casualties over the course of a day varied based on `weekend_weekday`, and also observed a change in relative densities of MVCs with and without accidents according to `yday`. Therefore, we will include `time_day`, `weekend_weekday`, and `yday` in our model. As we observed that weekend MVC casualties were more heavily concentrated in the early morning hours than weekday MVC casualties, we will also *consider* an interaction term between `weekend_weekday` and `time_day`, `weekend_weekday*time_day`.

We hoped to include `borough` as a predictor but could not due to high missingness. We also considered using `zip_code`, but this variable had over 100 unique levels and the proportion of accidents with casualties by zip code is approximately normally distributed around a median of 1.191, so we would not expect this to be a useful predictor.

We constructed and compared 3 models using a likelihood ratio test at the $\alpha = 0.05$ significance level to select our final model, each with `has_casualty` as the response variable. Predictors in `model_1` (the base model) are `vtype1`, `vtype2`, `time_day`, `weekend_weekday`, and `yday`, `model_2` adds `factor1` and `factor2`, and `model_3` adds `weekend_weekday*time_day` to `model_2`.

Based on our likelihood ratio test (**Appendix 3a**), `model_2` fits our data better than model 1, while the difference between `model_2` and `model_3` is insignificant. We move forward with `model_2` (as it has fewer predictors than `model_3`) and test if it satisfies assumptions of independence and linearity.

We will consider the assumption of independence between observations to be upheld because regardless of the number of vehicles involved, each accident receives only a single observation in the dataset. That removes the possibility that multiple rows correspond to the same accident and, thereby, the same accident conditions. If numerous reports of the same accident occur, there is still only one report made, so it is reasonable to assume that each observation is independent of the others, regardless of how many people reported it or how many cars were involved. From the plot showing `yday` compared to the predicted log odds of MVC casualty (**Appendix 3b**), there appears to be a fairly strong linear relationship between the two, so the linearity assumption for `model_2` does seem to be reasonable. Thus, we will proceed with `model_2` as our final model.

Results

```
# A tibble: 23 x 6
   term                         estimate    odds std.error statistic  p.value
   <chr>                           <dbl>   <dbl>     <dbl>     <dbl>    <dbl>
 1 (Intercept)                  -3.28e+0  0.0376  0.272     -12.1    1.69e-33
 2 vtype1Motorcycles             2.80e+0 16.5     0.279      10.1    8.25e-24
 3 vtype1Non-Motor Vehicle       3.68e+0 39.5     0.335      11.0    5.11e-28
 4 vtype1Other/Unknown           4.32e-1  1.54    0.132       3.27   1.06e- 3
 5 vtype1Passenger vehicles      7.14e-1  2.04    0.125       5.70   1.21e- 8
 6 vtype2Motorcycles             3.07e+0 21.5     0.319       9.61   7.32e-22
 7 vtype2Non-Motor Vehicle       3.59e+0 36.4     0.205      17.6    4.87e-69
 8 vtype2None                    1.06e+0  2.89    0.136       7.78   7.45e-15
 9 vtype2Other/Unknown          -4.06e-1  0.667   0.181      -2.24   2.50e- 2
10 vtype2Passenger vehicles      4.61e-1  1.59    0.136       3.39   7.07e- 4
11 time_dayafternoon             9.19e-2  1.10    0.0754      1.22   2.23e- 1
12 time_dayevening               3.26e-1  1.39    0.0795      4.10   4.15e- 5
13 time_daynight                 4.70e-1  1.60    0.0845      5.56   2.75e- 8
14 weekend_weekdayWeekend        3.43e-2  1.03    0.0636      0.540  5.89e- 1
15 yday                          5.10e-5  1.00    0.000267    0.191  8.48e- 1
```

16	factor1Failure to Obey Traffic~	5.29e-1	1.70	0.110	4.81	1.48e- 6
17	factor1Impairment/Distraction/~	1.99e-1	1.22	0.0953	2.08	3.72e- 2
18	factor1Other/Unknown	-1.48e-1	0.863	0.0924	-1.60	1.10e- 1
19	factor1Performance-unrelated T~	2.04e-2	1.02	0.138	0.148	8.82e- 1
20	factor2Failure to Obey Traffic~	3.14e-1	1.37	0.289	1.09	2.77e- 1
21	factor2Impairment/Distraction/~	1.10e-2	1.01	0.237	0.0463	9.63e- 1
22	factor2Other/Unknown	2.89e-1	1.33	0.204	1.42	1.56e- 1
23	factor2Performance-unrelated T~	1.34e-1	1.14	0.286	0.471	6.38e- 1

The final model fit can be assessed by the ROC AUC estimate (**Appendix 4a**). We found an ROC AUC of 0.7073, indicating that the model is fairly well fitted to our data for predicting MVS that did and did not result in casualties. This suggests that the model may be somewhat useful for predicting the likelihood of casualties in MVCs. However, our research question was designed to asses the extent to which certain variables are associated with MVC casualty occurrence, not primarily with predicting said occurrences.

Our model output shows a number of interesting findings. For ease of interpretation in terms of odds ratios, we added a column of odds ratios for each slope, the exponentiated logit slope estimates. Statistical significance for a slope estimate was defined as having an associated p-value < 0.05.

Regarding vehicle type, all categories of the variable had statistically significant slope estimates, and the reference category for all was "Commercial vehicles." As expected based on EDA, the slopes for "Motorcycles" and "Non-Motor Vehicle" were the most significant and carried the highest odds ratios. Contextually, comparing two MVCs in NYC, one between two commercial vehicles and one where the first motorist is on a motorcycle and the second in a commercial vehicle, we predict that the MVC involving the motorcyclist has, incredibly, 16.482 times the odds of resulting in a casualty as compared to the former, all other model predictors held constant.

In terms of timing, numerical day of the year was not found to be significantly associated with casualty occurrence, nor was whether the MVC occurred on a weekend. MVC occurrence in the evening or nighttime (as compared to morning) was significantly associated with casualty occurrence. For two MVCs in NYC, one occurring in the morning and one at night, we predict that the latter has 1.6 times the odds of resulting in a casualty, all all other model predictors held constant.

None of the second driver factors were significantly related to casualty occurrence, however, the first driver failing to obey traffic signs/signals/rules or being impaired/distracted/fatigued (as compared to driving aggressively or recklessly) were both significantly associated with casualty occurrence, carrying odds ratios of 1.697 and 1.22, respectively.

Discussion

Our primary research concern is determining which characteristics of a MVC's timing and participants involved make casualties more likely. The model output shows that vehicle type significantly impacts the odds of a casualty occurring. Specifically, Motorcycles and Non-Motor Vehicles are associated with a *much* higher likelihood of a casualty than Passenger vehicles. Additionally, both time of day and weekday/weekend variance has a statistically significant impact on the probability of a casualty.

However, the day of the year is *not* a statistically significant predictor of casualties. It should be noted that even though secondary contributing factors are included in the model, none of them have a statistically significant effect on whether or not an MVC resulted in a casualty.

Our final model shows that several factors play a role in determining whether an MVC in New York City results in any casualties. The variables that significantly contribute to the likelihood of a casualty include the types of vehicles involved, the time of day, and whether the MVC occurred on a weekend.

However, there are several limitations to this analysis that must be acknowledged. Firstly, the data used in this study was harvested and compiled from NYPD open access data, which means that the reliability and validity of the data may not be ideal, with some incidents of MVCs potentially not reported. It is also worth noting that the dataset is limited to New York City and may not be generalizable to other metropolitan areas. Future studies should attempt to replicate this analysis in other urban areas to determine if the findings are consistent across different locations.

Finally, the study's limitations must be considered when interpreting the results of the analysis. While the model provides statistical evidence to support the hypothesis that certain characteristics of an MVC increase the likelihood of a casualty, it is important to note that correlation does not necessarily imply causation. Additionally, the statistical analysis used in this study assumes that the data meets certain assumptions, such as independence and linearity, which may not be valid in all cases.

As for future work, the analysis could be improved by collecting more a greater number of variables (driver demographics and weather conditions, for example). Future studies could also attempt to identify which specific combinations of variables are most strongly associated with a higher likelihood of a casualty, rather than treating each variable as an independent factor. Overall, further investigation is needed to fully understand the complex relationship between MVC characteristics and casualty likelihood.

Appendix

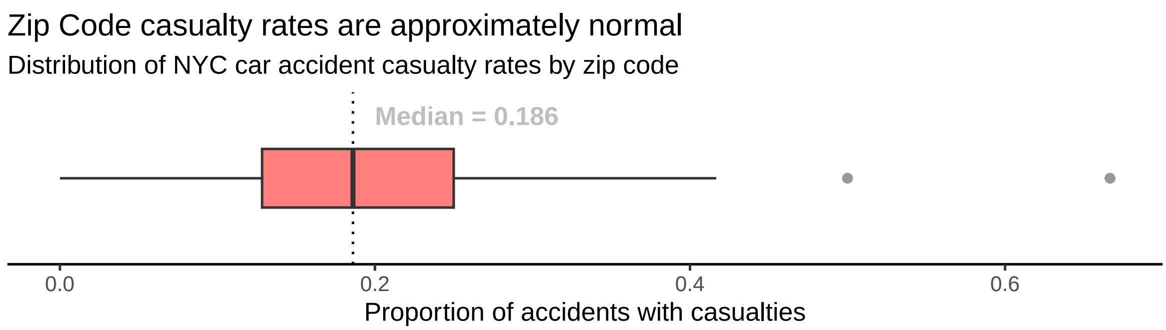
1. Data Cleaning

a) Sample code for sampling/exportation of large original dataset

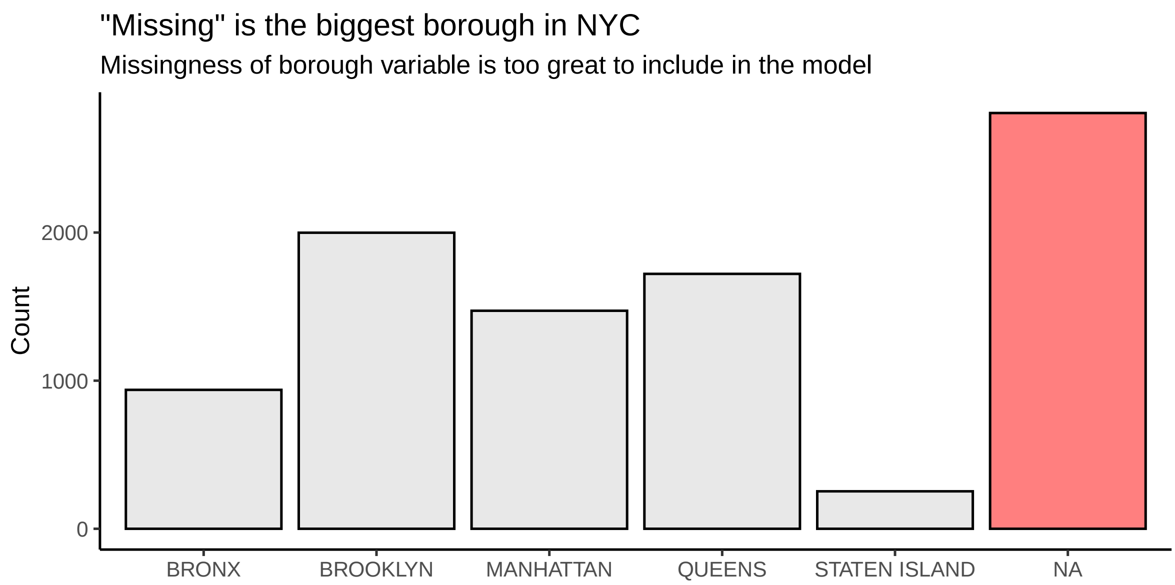
```
crashes_original <- read_csv(<filename>)
crashes <- sample_n(crashes_original, 10000)
write_csv(crashes, <"crashes">)
```

2. Exploratory Data Analysis

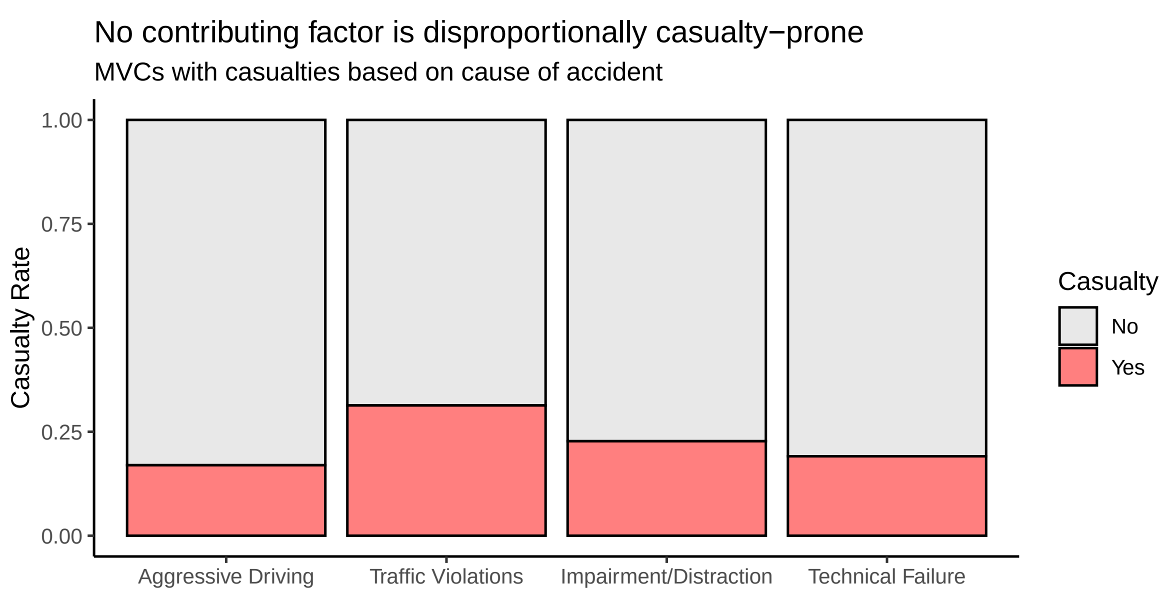
a) Visualization of Casualty Rates by Zip Code



b) Visualization of MVC by Borough



c) Proportion of casualties by contributing factor



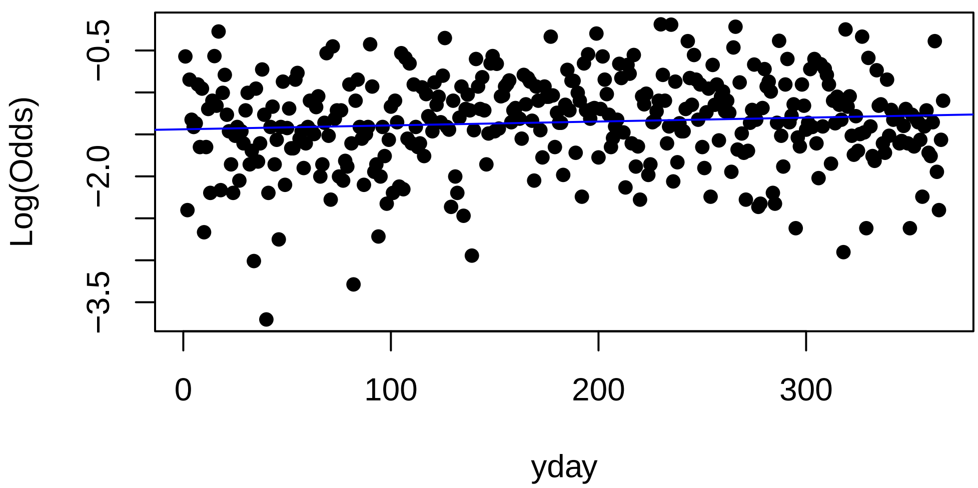
3. Methodology

a) Likelihood-ratio test between models considered

```
Analysis of Deviance Table

Model 1: has_casualty ~ vtype1 + vtype2 + time_day + weekend_weekday +
    yday
Model 2: has_casualty ~ vtype1 + vtype2 + time_day + weekend_weekday +
    yday + factor1 + factor2
Model 3: has_casualty ~ vtype1 + vtype2 + time_day + weekend_weekday +
    yday + factor1 + factor2 + weekend_weekday * time_day
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      9175     8259.6
2      9167     8194.9  8   64.716 5.493e-11 ***
3      9164     8192.4  3    2.562    0.4641
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b) Numerical Day of the Year vs. Log-odds of Casualty Plot



4. Results

a) ROC Assessment

```
# A tibble: 1 x 3
```

```
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.707
```

5. Works Cited

Department of Health. All Injuries in New York State. (n.d.). Retrieved May 4, 2023, from https://www.health.ny.gov/statistics/prevention/injury_prevention/all_injury.htm