New York's Streetside Casualties

An explanatory analysis of NYC car accidents

Drew LoPolito and Mitch Harrison

Introduction

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. 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 (code shown in Methods). 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, which brought our total number of observations down to approximately 9,300.

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:

`crash_date`: the date on which the MVC occurred.

`crash_day`: a categorical variable corresponding to the day of the week on which the MVC occurred, generated from the original `crash_date` variable.

`yday`: a numeric variable ranging from 1 to 365 corresponding to the numerical day of the year on which the MVC occurred.

`crash_time` the time (in 24-hr time) at which the MVC occurred.

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

`num_casualties`: a numeric variable corresponding to the number of casualties, i.e., injuries or fatalities, resulting from an MVC. This was generated from the original `number_of_persons_injured` and `number_of_persons_killed` variables.

`has_casualty`: a binary variable corresponding to whether or not a MVC resulted in at least one casualty.

`vtype1` and `vtype2`: categorical variables corresponding to the types of each vehicle (if applicable) involved in the crash, with categories of "Commercial vehicles," "Passenger vehicles," "Motorcycles," "Non-Motor Vehicle," and "Other/Unknown."

`factor1` and `factor2`: categorical variables corresponding any notable factors which were listed as potentially contributing to the crash (if applicable), with categories of "Aggressive/Reckless Driving," "Distraction/Inattention/Fatigue," "Failure to Obey Traffic Signs/Signals/Rules," "Impaired," "Other Technical/Mechanical Factors," and "Other/Unknown."

Our primary research question of interest is whether an accident having casualties or not depends on the type(s) of

Read Data/Libraries

Methodology

To explore the factors that lead to casualties in car accidents, we will construct a binary logistic regression model. For the purposes of this analysis, we define "casualties" as injuries or fatalities, not just fatalities. We are only examining accidents involving two or fewer vehicles, which account for over 90% of accidents while avoiding the complexities involved with correcting for several vehicle types at once.

We anticipate that the vehicle type involved in the accident will play a significant role, but our dataset has over 100 individual vehicle types. To simplify our model and avoid overfitting, we consolidated these into five larger categories: passenger vehicles, commercial vehicles, motorcycles, non-motor vehicles, and others/unknown. We also anticipated a significant impact on accident severity by reason for the accident. Specifically, we wanted to investigate impaired driving to other causes. Like the vehicle type, however, we risked overfitting and over-complicating our model because of the sheer number of accident causes. Like our vehicle type variable, we consolidated accident causes into five categories: impaired, distraction/inattention/fatigue, aggressive/reckless driving, failure to obey traffic signs/signals/rules, technical/mechanical failures, and other/unknown.

After consolidating categories, we investigate both the time of day and the day of the year at which an accident occurred. Grouping by day of the week, we find that the time at which injuries with casualties occur varies, with fewer casualties around 5:00 pm and later at night during the weekend. Because of this difference, we include both the time of day and weekday in our model.

The day of the year (as measured by the number of days after January 1st, i.e., "Julian Day") is also included. During exploratory analysis, we found that accidents had disproportionately few casualties in the winter months and more around the middle of the year. Because of this difference, we include the day of the year as a predictor in our model.

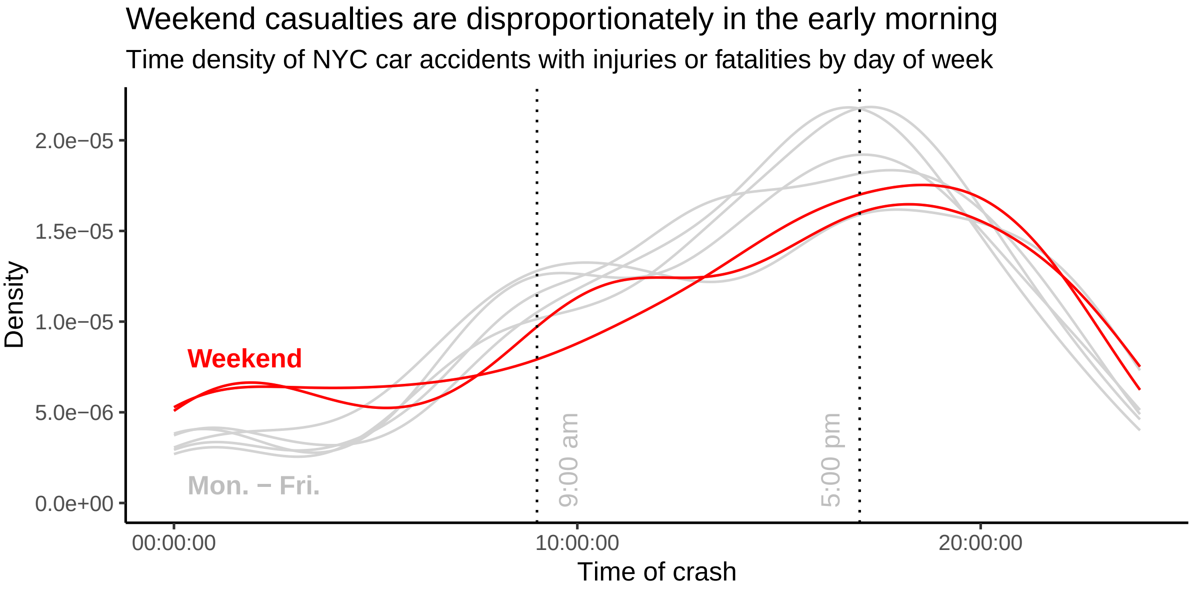
We hoped to include the borough as a predictor but could not. So many observations were missing a value for the borough that including it in our complete-case model would have dropped a substantial portion of our overall dataset. If we could confirm that the borough data was MCAR, then we wouldn't be concerned about the final accuracy of our model. However, because we could not verify the missingness mechanism (and due to the general rarity of MCAR data that isn't explicitly designed as such), we do not include the borough in which the accident occurs in our model.

While we hoped that the zip code might provide valuable explanatory insight, we found two issues with its inclusion in our model: first, there are 100 unique zip codes in the dataset. As a categorical variable, this would make our model much more complex to calculate and communicate. Second, the proportion of accidents with casualties is approximately normally distributed around a median of 1.191. Therefore, specific zip codes are both unlikely to make an appreciable difference in an explanatory model and are highly likely to increase its complexity dramatically. Thus, we do not include zip code in our final model.

We also thought that some streets could have a disproportionately high casualty rate and thereby have a statistically significant impact on our model. However, we ran into many of the same problems that we did with zip codes. In the 10,000 observations we selected from the larger dataset, there were over 1000 unique streets, many of which had minimal accidents. To combat these small number of observations, we used LaPlace succession, adding one to each of the number of accidents and making it a non-casualty accident. This technique allowed us to correct for low accident counts. After applying LaPlace succession, we found that the large majority of streets had no casualties at all, and the remainder were distributed around 0.25. Because of the massive complexity that adding a 1000-level categorical variable would add to our model, combined with an unremarkable distribution of accident proportion after applying LaPlace succession, we are not including streets as a predictor in our final model.

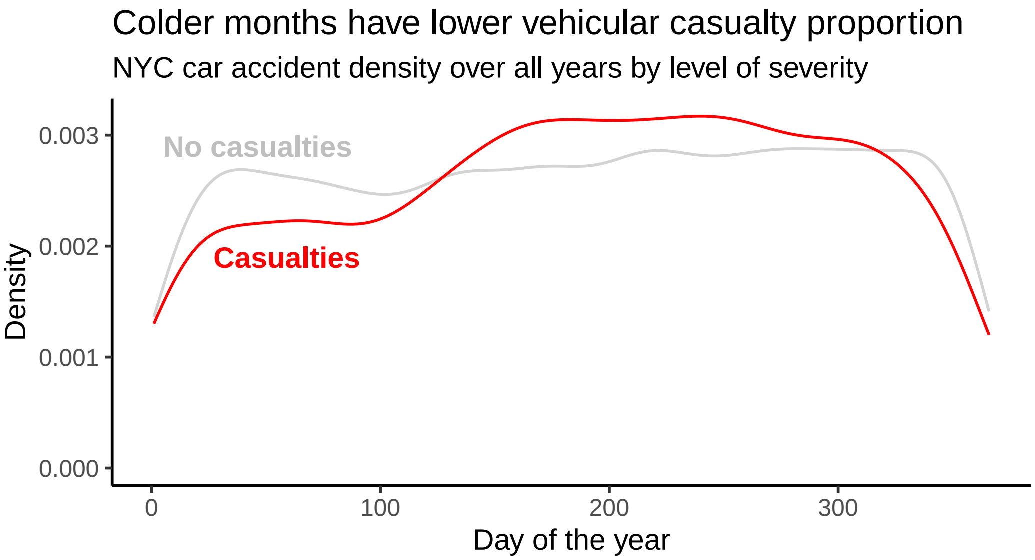
Exploratory Analysis

Time of Day



Of accidents with one or more casualties, those happening on Saturdays and Sundays appear to occur disproportionately late at night and with a notably smaller peak around 5:00 pm. Because of this difference, we constructed a new variable to denote whether or not an accident happens on a weekend or weekday. We will build two models, one with an interaction term between the time of day and weekend/weekday and one without, and compare their performance before selecting a final model.

Day of the year

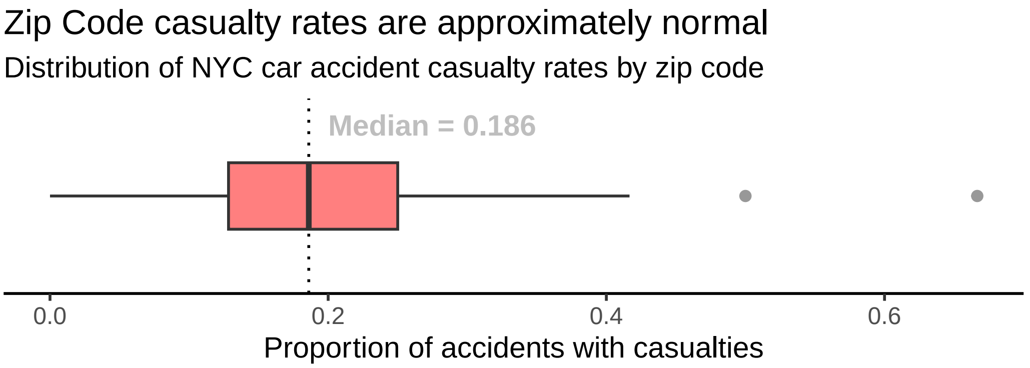


The variance in density between levels of severity of accidents through all years is relatively low. However, we see disproportionately severe accidents around the late summer/early spring months and minor accidents in the early winter months. This difference is enough to include it as a potential predictor of interest in our final model.

Borough

We have far too many missing values to include borough in our model. If we could verify that borough was Missing Completely At Random (MCAR), we could include it in our complete-case model. Still, unfortunately, we are unable to verify the missingness mechanism and are therefore uncomfortable with possible performance impacts by using borough in our final model.

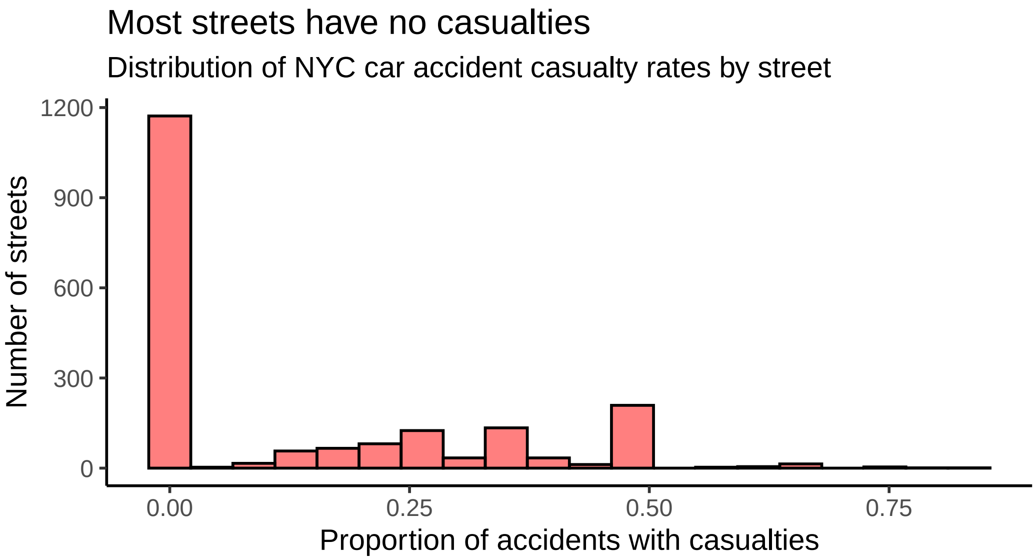
Zip codes


Zip Code casualty rates are approximately normal

Distribution of NYC car accident casualty rates by zip code

Building a box plot to observe the distribution of accident proportion by zip code, we see that casualty rates are approximately normally distributed about a median of 0.186. Because of the large absolute number of zip codes in the dataset and the unlikely impact of a normally distributed proportion having a significant impact on the model, we will not use zip code as a predictor in the final model.

Streets

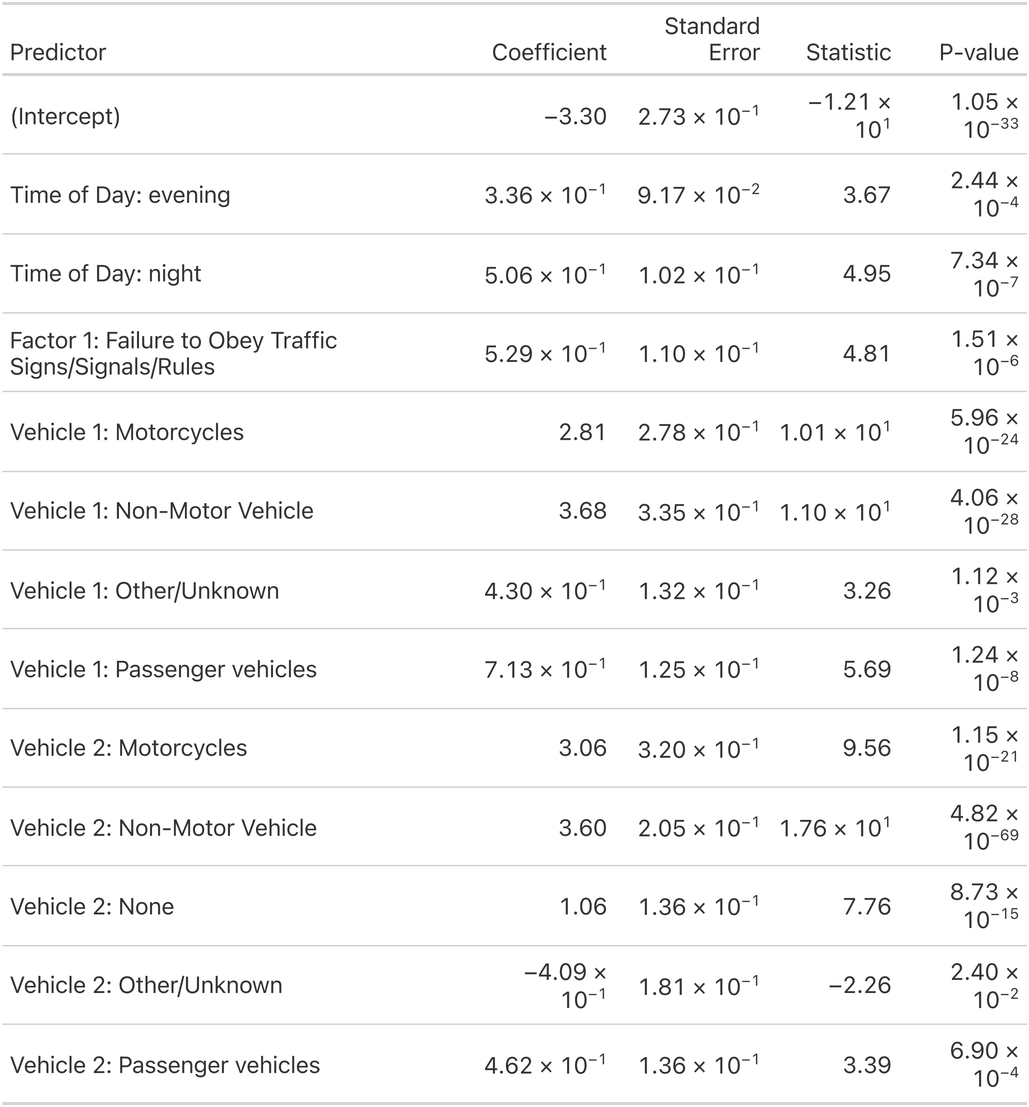

Most streets have no casualties

Distribution of NYC car accident casualty rates by street

After applying LaPlace succession, we see that the vast majority of accidents had no casualties, and most others were distributed around 0.25, with another peak around 0.5, partially due to the mathematical consequences of LaPlace succession. Because so many streets followed an unremarkable distribution and because adding street as a predictor would introduce massive model complexity, we will not use the street on which an accident occurs as a predictor in the final model.

Results

Discussion



Predictor	Coefficient	Standard Error	Statistic	P-value
(Intercept)	-3.30	2.73×10^{-1}	-1.21×10^{1}	1.05×10^{-33}
Time of Day: evening	3.36×10^{-1}	9.17×10^{-2}	3.67	2.44×10^{-4}
Time of Day: night	5.06×10^{-1}	1.02×10^{-1}	4.95	7.34×10^{-7}
Factor 1: Failure to Obey Traffic Signs/Signals/Rules	5.29×10^{-1}	1.10×10^{-1}	4.81	1.51×10^{-6}
Vehicle 1: Motorcycles	2.81	2.78×10^{-1}	1.01×10^{1}	5.96×10^{-24}
Vehicle 1: Non-Motor Vehicle	3.68	3.35×10^{-1}	1.10×10^{1}	4.06×10^{-28}
Vehicle 1: Other/Unknown	4.30×10^{-1}	1.32×10^{-1}	3.26	1.12×10^{-3}
Vehicle 1: Passenger vehicles	7.13×10^{-1}	1.25×10^{-1}	5.69	1.24×10^{-8}
Vehicle 2: Motorcycles	3.06	3.20×10^{-1}	9.56	1.15×10^{-21}
Vehicle 2: Non-Motor Vehicle	3.60	2.05×10^{-1}	1.76×10^{1}	4.82×10^{-69}
Vehicle 2: None	1.06	1.36×10^{-1}	7.76	8.73×10^{-15}
Vehicle 2: Other/Unknown	-4.09×10^{-1}	1.81×10^{-1}	-2.26	2.40×10^{-2}
Vehicle 2: Passenger vehicles	4.62×10^{-1}	1.36×10^{-1}	3.39	6.90×10^{-4}

Figure 1: Model Output