

New York's Streetside Casualties

An explanatory analysis of NYC car accidents

Drew LoPolito and Mitch Harrison

Introduction

Instructions:

This section includes an introduction to the project motivation, data, and research question. Describe the data and definitions of key variables. It should also include some exploratory data analysis. All of the EDA won't fit in the paper, so focus on the EDA for the response variable and a few other interesting variables and relationships.

Grading Criteria

The research question and motivation are clearly stated in the introduction, including citations for the data source and any external research. The data are clearly described, including a description about how the data were originally collected and a concise definition of the variables relevant to understanding the report. The data cleaning process is clearly described, including any decisions made in the process (e.g., creating new variables, removing observations, etc.) The explanatory data analysis helps the reader better understand the observations in the data along with interesting and relevant relationships between the variables. It incorporates appropriate visualizations and summary statistics.

What we have:

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:

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

weekend_weekday: a binary variable corresponding to whether a MVC occurred during the weekend or the w

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

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 concern is determining which characteristics of the timing and participants involved make a motor vehicle collision (MVC) more likely to involve casualties.

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.

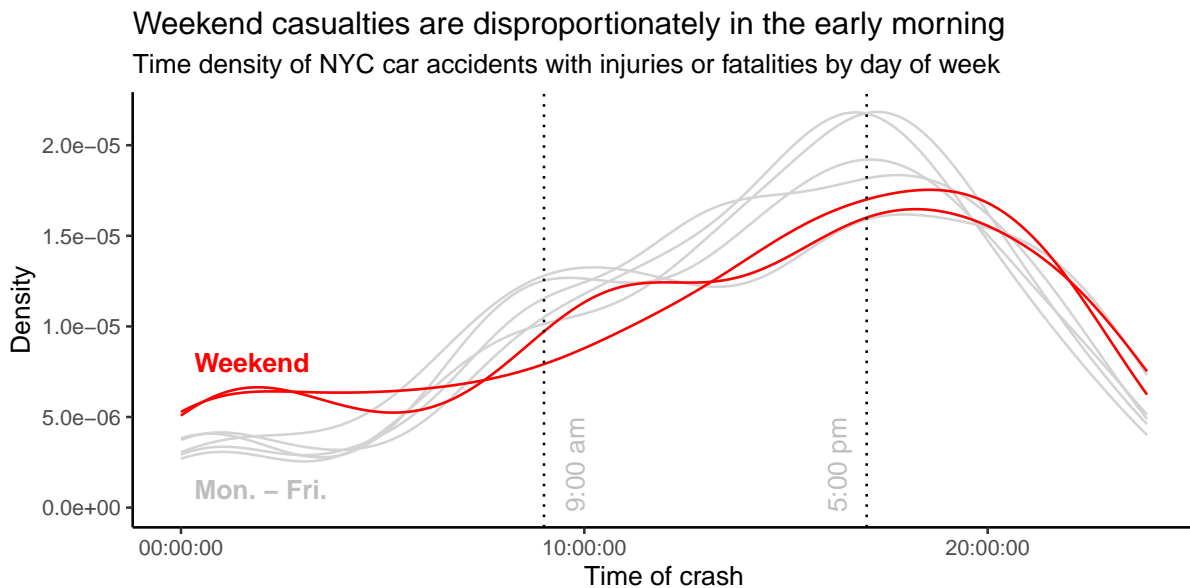
The day of the year (as measured by the number of days after January 1st, i.e., “Julian Day”) is also included. During exploratory analysis, we found that accidents had disproportionately few casualties in the winter months and more around the middle of the year. Because of this difference, we include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

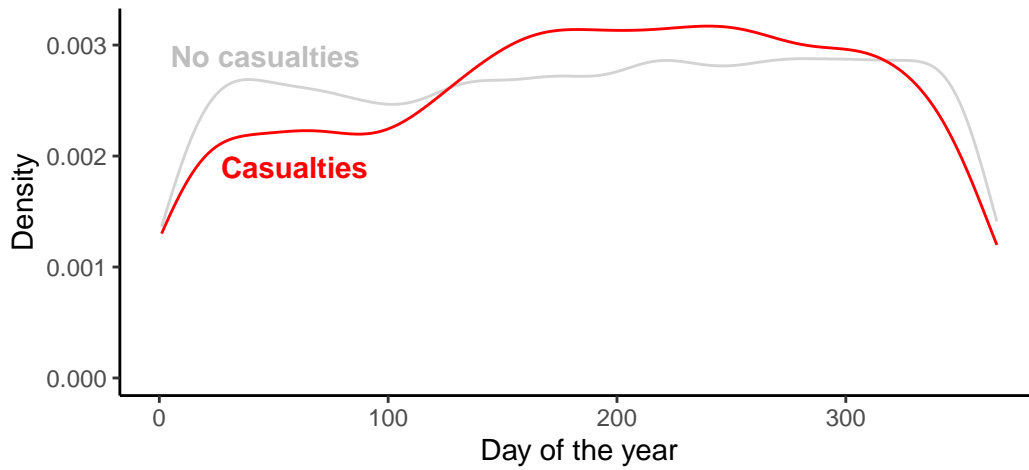
We also thought that some streets could have a disproportionately high casualty rate and thereby have a statistically significant impact on our model. However, we ran into many of the same problems that we did with zip codes. In the 10,000 observations we selected from the larger dataset, there were 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



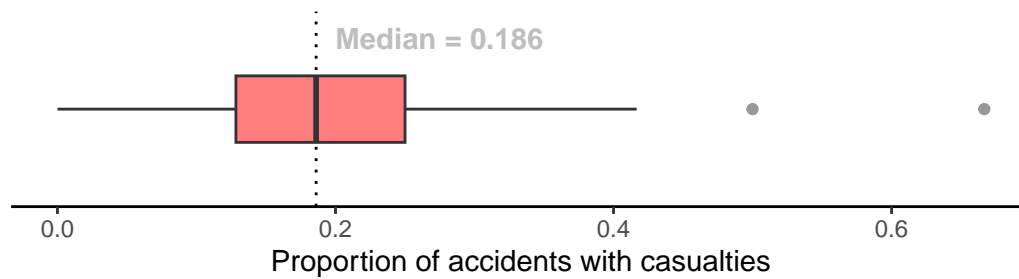
Colder months have lower vehicular casualty proportion

NYC car accident density over all years by level of severity



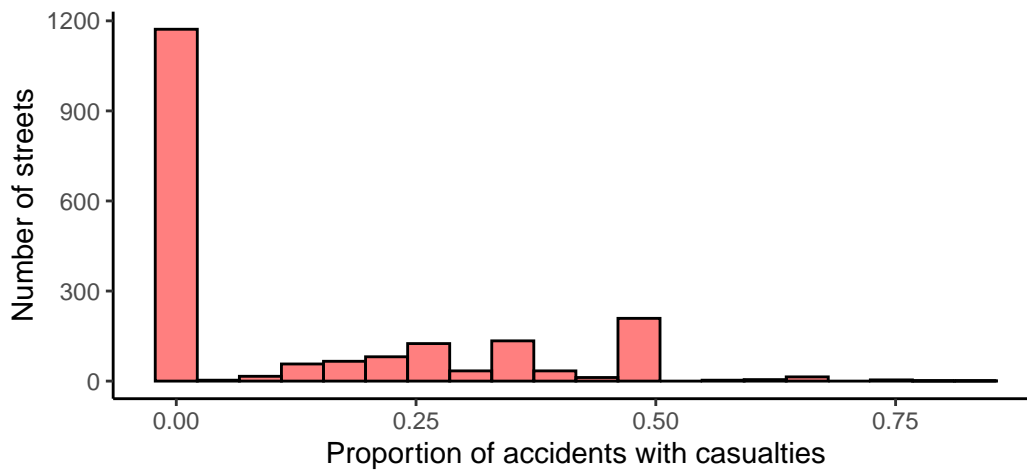
Zip Code casualty rates are approximately normal

Distribution of NYC car accident casualty rates by zip code



Most streets have no casualties

Distribution of NYC car accident casualty rates by street



Instructions:

This section includes a brief description of your modeling process. Explain the reasoning for the type of model you're fitting, predictor variables considered for the model including any interactions. Additionally, show how you arrived at the final model by describing the model selection process, interactions considered, variable transformations (if needed), assessment of conditions and diagnostics, and any other relevant considerations that were part of the model fitting process.

Grading criteria

The analysis steps are appropriate for the data and research question. The group used a thorough and careful approach to select the final model; the approach is clearly described in the report. The model selection process was reasonable, and addressed any violations in model conditions were discussed and/or fixed. The model conditions and diagnostics are thoroughly and accurately assessed for their model. If violations of model conditions are still present, there was a reasonable attempt to address the violations based on the course content.

What we have:

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.

Results

Instructions:

This is where you will output the final model with any relevant model fit statistics. Describe the key results from the model. The goal is not to interpret every single variable in the model but rather to show that you are proficient in using the model output to address the research questions, using the interpretations to support your conclusions. Focus on the variables that help you answer the research question and that provide relevant context for the reader.

Grading Criteria

The model fit is clearly assessed, and interesting findings from the model are clearly described. Interpretations of model coefficients are used to support the key findings and conclusions, rather than merely listing the interpretation of every model coefficient. If the primary modeling objective is prediction, the model's predictive power is thoroughly assessed.

What we have

Discussion

Instructions:

In this section you'll include a summary of what you have learned about your research question along with statistical arguments supporting your conclusions. In addition, discuss the limitations of your analysis and provide suggestions on ways the analysis could be improved. Any potential issues pertaining to the reliability and validity of your data and appropriateness of the statistical analysis should also be discussed here. Lastly, this section will include ideas for future work.

Grading criteria

Overall conclusions from analysis are clearly described, and the model results are put into the larger context of the subject matter and original research question. There is thoughtful consideration of potential limitations of the data and/or analysis, and ideas for future work are clearly described.

What we have

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