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:

Project Motivation

Insert a blurb using a few citations about why this project was important.

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, `num_casualties`, (defined as injuries **and** casualties). We then created binary variables for MVC injury, fatality, and casualty, the last of which, `has_casualty`, is our response variable of interest. 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 `time_day`, day of the week, `crash_day`, numerical day of the year (1-365), `yday`, and weekday or weekend timing, `weekend_weekday`.

```
Number of levels in vehicle_type_code_1: 84
```

```
Number of levels in vehicle_type_code_2: 93
```

```
Number of levels in contributing_factor_vehicle_1: 52
```

```
Number of levels in contributing_factor_vehicle_2: 38
```

```
[1] 121
```

```
[1] 53
```

```
# A tibble: 1 x 1
      n
  <int>
1   453
```

```
# A tibble: 1 x 1
      n
  <int>
1    59
```

```
            n
1 0.006379068
```

Next, we cleaned the original variables corresponding to vehicle types, `vehicle_type_code_(1 or 2)`, involved in the accident as well as factors in each vehicle which may have contributed to the MVC, `contributing_factor_vehicle_(1 or 2)`. 453 MVCs contained an observation for `contributing_factor_vehicle_2` but were NA for `vehicle_type_code_2`, which we believe indicated not that there wasn't a second vehicle but that the vehicle type hadn't been recorded, so we replaced the `vehicle_type_code_2` NAs by indicating that it was simply unknown. Additionally, 59 MVCs had missing observations for `vehicle_type_code_1`, 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_code_(1 or 2)`. In order to make these types interpretable for EDA and potentially in our model, we renamed the variables `vtype(1 or 2)` and 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 `vtype2`).

The initial dataset also had 53 unique levels for `contributing_factor_vehicle_(1 or 2)`. In order to make these types interpretable for EDA and potentially in our model, we renamed the variables `factor(1 or 2)` and 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."

```
[1] "Aggressive/Reckless Driving"
[2] "Failure to Obey Traffic Signs/Signals/Rules"
[3] "Impairment/Distraction/Fatigue"
[4] "Other/Unknown"
[5] "Performance-unrelated Technical/Mechanical Factors"
```

```
[1] "Aggressive/Reckless Driving"
[2] "Failure to Obey Traffic Signs/Signals/Rules"
[3] "Impairment/Distraction/Fatigue"
[4] "Other/Unknown"
[5] "Performance-unrelated Technical/Mechanical Factors"
```

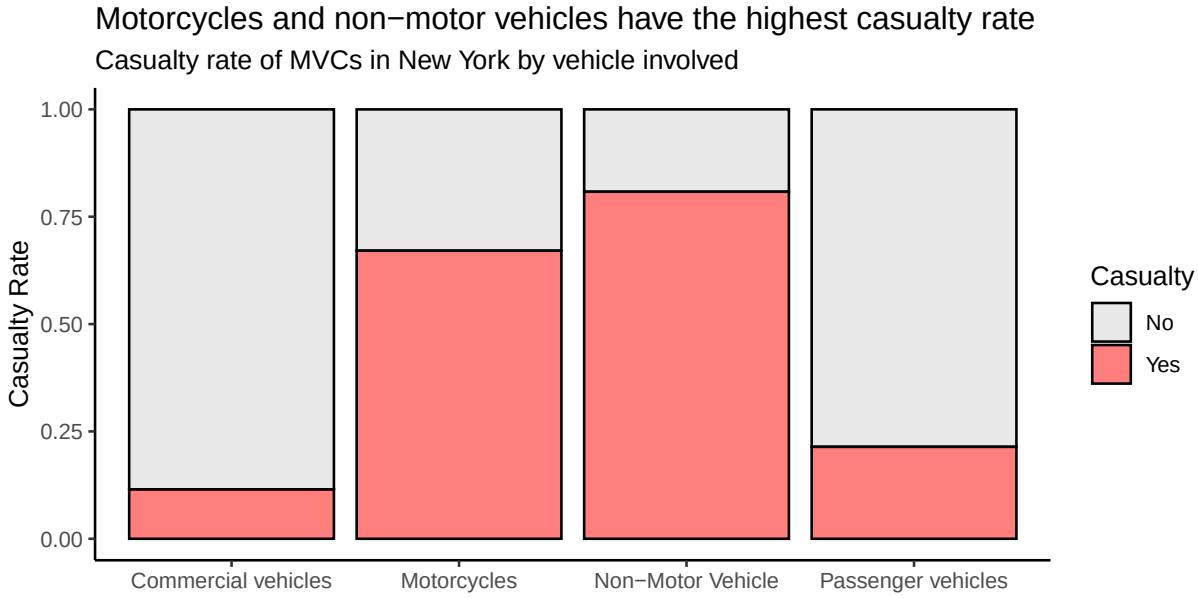
```
[1] "Commercial vehicles" "Motorcycles"         "Non-Motor Vehicle"
[4] "Other/Unknown"       "Passenger vehicles"

[1] "Commercial vehicles" "Motorcycles"         "Non-Motor Vehicle"
[4] "None"                "Other/Unknown"       "Passenger vehicles"
```

Exploratory Analysis

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.

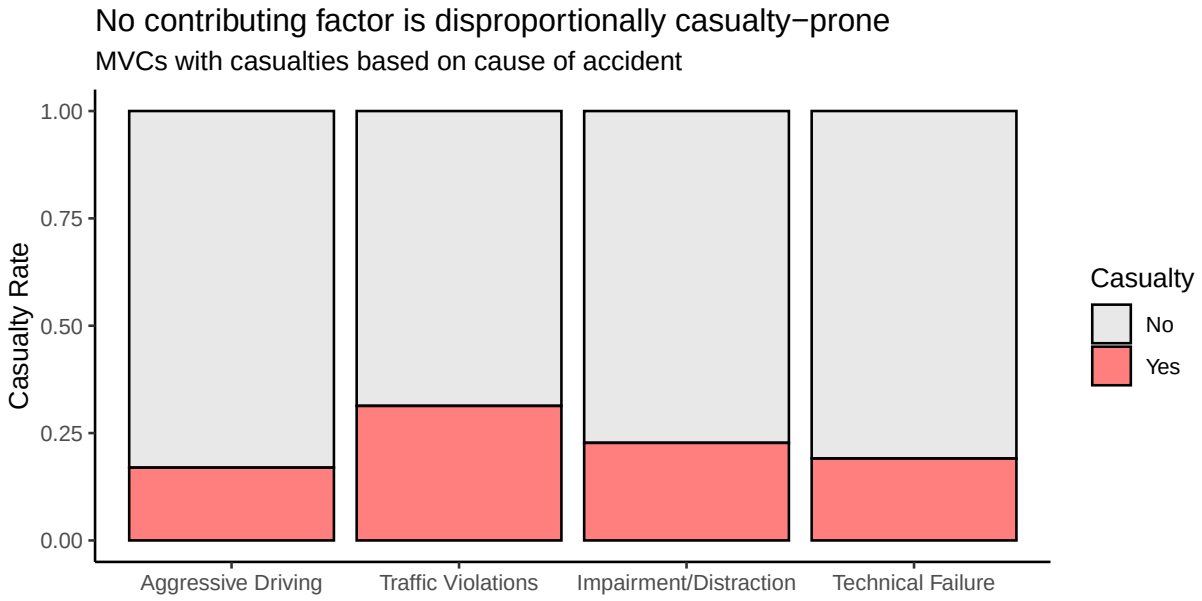
Below I was trying to make charts of the casualty proportions of accidents by vehicle types 1 and 2. The gist is clear, but if you could clean them up a bit to make them look good that would be awesome.

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.



Same here, I was trying to make charts of the casualty proportions of accidents by `factor1` and `factor2`. The gist is clear, but if you could clean them up a bit, *particularly the x-axis labels*, to make them look good that would be awesome.

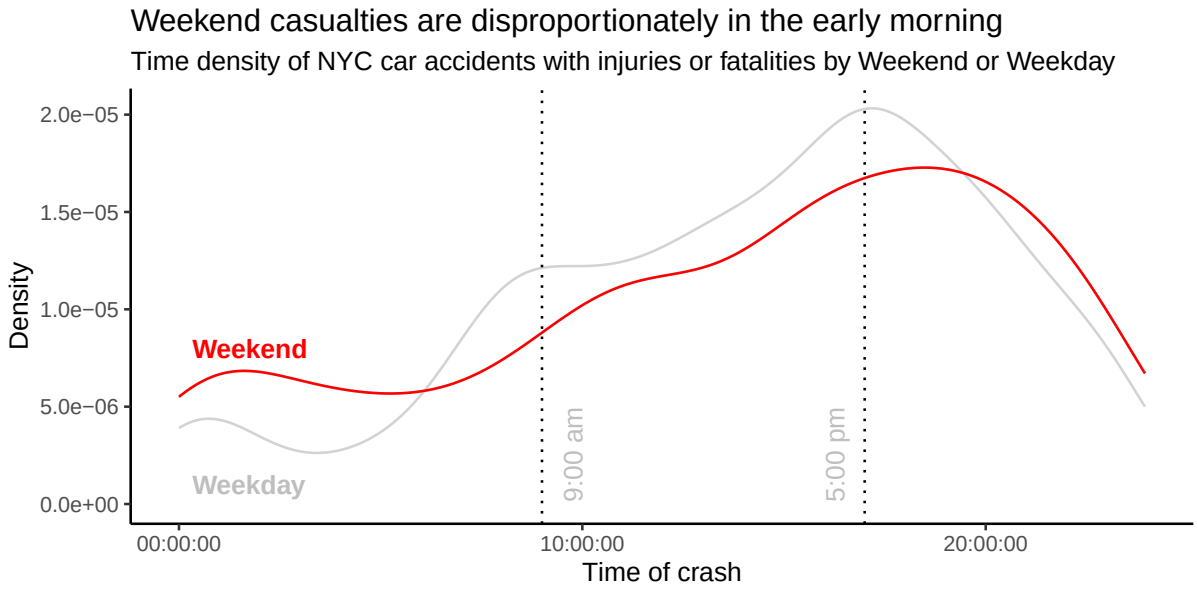
The visualizations below do not seem to show any one driver related error or issue having a remarkable impact on casualty rates as compared to any others.



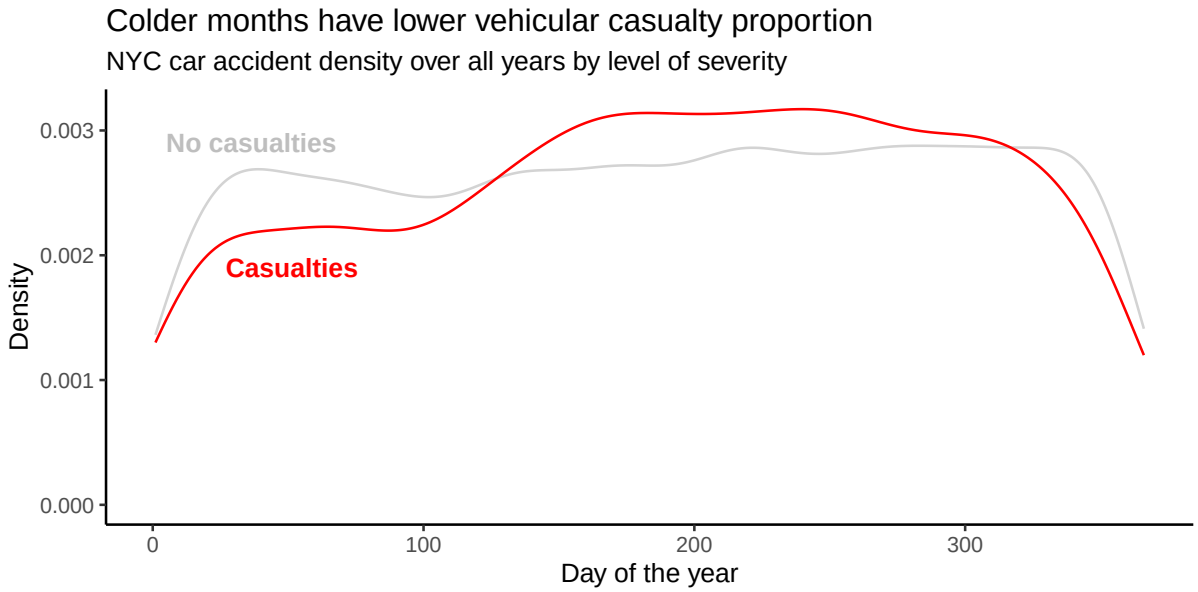
To explore how timing might effect the occurrence of casualties, we explored the relationship between day of the week and casualty occurrence, finding that weekend MVC casualties were more heavily concentrated in the early morning hours than weekday MVC casualties. This was interesting as it provided 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.

```
          n
1 0.1746032
```

I made another density plot using your code but with the `weekend_weekday` variable instead of individual days for this, we can use whichever but it seems a bit clearer with the latter in my opinion. See below:



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.

Lastly, we visualized the number of streets with

I dont think we need this plot below. To be honest, it shows that something like 300 streets have casualty rates around 0.5, which is pretty high and probably would indicate that we should include street somehow as a predictor...

Methodology

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:

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

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 we will fit a logistic regression model to investigate our primary research concern.

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

Model 1

```
# A tibble: 26 x 5
   term                                      estimate std.e~1 stati~2  p.value
   <chr>                                        <dbl>   <dbl>   <dbl>    <dbl>
 1 (Intercept)                              -3.30e+0 2.73e-1 -12.1   1.03e-33
 2 yday                                      4.21e-5 2.67e-4   0.158 8.75e- 1
 3 time_dayafternoon                         1.54e-1 8.62e-2   1.78  7.46e- 2
 4 time_dayevening                           3.37e-1 9.17e-2   3.67  2.38e- 4
 5 time_daynight                             5.10e-1 1.02e-1   5.00  5.84e- 7
 6 weekend_weekdayWeekend                    1.60e-1 1.31e-1   1.22  2.22e- 1
 7 factor1Failure to Obey Traffic Signs/Signa~ 5.29e-1 1.10e-1   4.81  1.49e- 6
 8 factor1Impairment/Distraction/Fatigue     1.99e-1 9.53e-2   2.09  3.68e- 2
 9 factor1Other/Unknown                     -1.47e-1 9.24e-2  -1.59  1.13e- 1
10 factor1Performance-unrelated Technical/Mec~ 1.92e-2 1.38e-1   0.139 8.89e- 1
# ... with 16 more rows, and abbreviated variable names 1: std.error,
#   2: statistic
```

Model 2

```
# A tibble: 23 x 5
   term                                      estimate std.e~1 stati~2  p.value
   <chr>                                        <dbl>   <dbl>   <dbl>    <dbl>
 1 (Intercept)                              -3.28e+0 2.72e-1 -12.1   1.69e-33
 2 yday                                      5.10e-5 2.67e-4   0.191 8.48e- 1
 3 time_dayafternoon                         9.19e-2 7.54e-2   1.22  2.23e- 1
 4 time_dayevening                           3.26e-1 7.95e-2   4.10  4.15e- 5
 5 time_daynight                             4.70e-1 8.45e-2   5.56  2.75e- 8
 6 weekend_weekdayWeekend                    3.43e-2 6.36e-2   0.540 5.89e- 1
 7 factor1Failure to Obey Traffic Signs/Signa~ 5.29e-1 1.10e-1   4.81  1.48e- 6
 8 factor1Impairment/Distraction/Fatigue      1.99e-1 9.53e-2   2.08  3.72e- 2
 9 factor1Other/Unknown                     -1.48e-1 9.24e-2  -1.60  1.10e- 1
10 factor1Performance-unrelated Technical/Mec~ 2.04e-2 1.38e-1   0.148 8.82e- 1
# ... with 13 more rows, and abbreviated variable names 1: std.error,
#   2: statistic
```

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.

Predictor	Coefficient	Standard Error	Statistic	P-value
(Intercept)	−3.30	2.73×10^{-1}	-1.21×10^{1}	1.03×10^{-33}
Time of Day: evening	3.37×10^{-1}	9.17×10^{-2}	3.67	2.38×10^{-4}
Time of Day: night	5.10×10^{-1}	1.02×10^{-1}	5.00	5.84×10^{-7}
Factor 1: Failure to Obey Traffic Signs/Signals/Rules	5.29×10^{-1}	1.10×10^{-1}	4.81	1.49×10^{-6}
Factor 1: Impairment/Distraction/Fatigue	1.99×10^{-1}	9.53×10^{-2}	2.09	3.68×10^{-2}
Vehicle 1: Motorcycles	2.81	2.78×10^{-1}	1.01×10^{1}	6.65×10^{-24}
Vehicle 1: Non-Motor Vehicle	3.68	3.35×10^{-1}	1.10×10^{1}	4.37×10^{-28}
Vehicle 1: Other/Unknown	4.31×10^{-1}	1.32×10^{-1}	3.27	1.08×10^{-3}
Vehicle 1: Passenger vehicles	7.12×10^{-1}	1.25×10^{-1}	5.68	1.31×10^{-8}
Vehicle 2: Motorcycles	3.06	3.19×10^{-1}	9.56	1.14×10^{-21}
Vehicle 2: Non-Motor Vehicle	3.59	2.05×10^{-1}	1.75×10^{1}	5.98×10^{-69}
Vehicle 2: None	1.06	1.36×10^{-1}	7.75	8.99×10^{-15}
Vehicle 2: Other/Unknown	-4.11×10^{-1}	1.81×10^{-1}	−2.27	2.33×10^{-2}
Vehicle 2: Passenger vehicles	4.61×10^{-1}	1.36×10^{-1}	3.38	7.16×10^{-4}

Figure 1: Model Output

What we have

Appendix

1. Data Cleaning

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

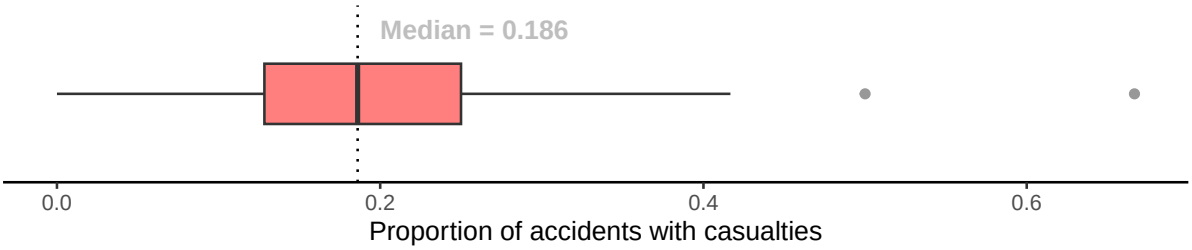
```
crashes_original <- read_csv(<filename>)
crashes <- sample_n(crashes_original, 10000)
write_csv(crashes, <"crashes">)
```

2. Exploratory Data Analysis

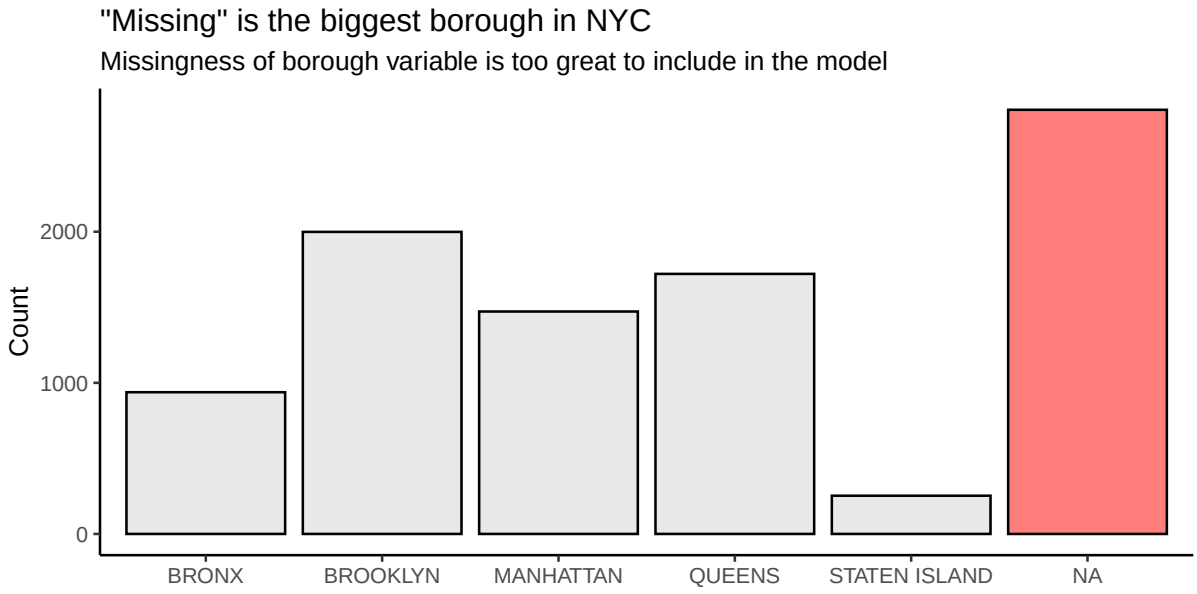
a) Visualization of Casualty Rates by Zip Code

Zip Code casualty rates are approximately normal

Distribution of NYC car accident casualty rates by zip code



b) Visualization of MVC by Borough


"Missing" is the biggest borough in NYC
Missingness of borough variable is too great to include in the model
Count
2000
1000
0
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
NA

c.