New York's Streetside Casualties

An explanatory analysis of NYC car accidents

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 (Department of Health, 2023). 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 a random sampling of 10,000 motor vehicle collisions (MVCs) out of the 2 million MVCs publicly released by the New York City Police Department (NYPD), spanning 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 any MVC in which someone is injured or killed, or which result in at least $1,000 of total property damage.

We created a binary response variable corresponding to whether or not an MVC resulted in a *casualty* (defined as a fatality *or* an injury), and additional variables corresponding to type of vehicles involved (commercial, motorcycle, etc.) and factors that contributed to the MVC (aggressive driving, impairment, etc.). A full data dictionary is available in the appendix.

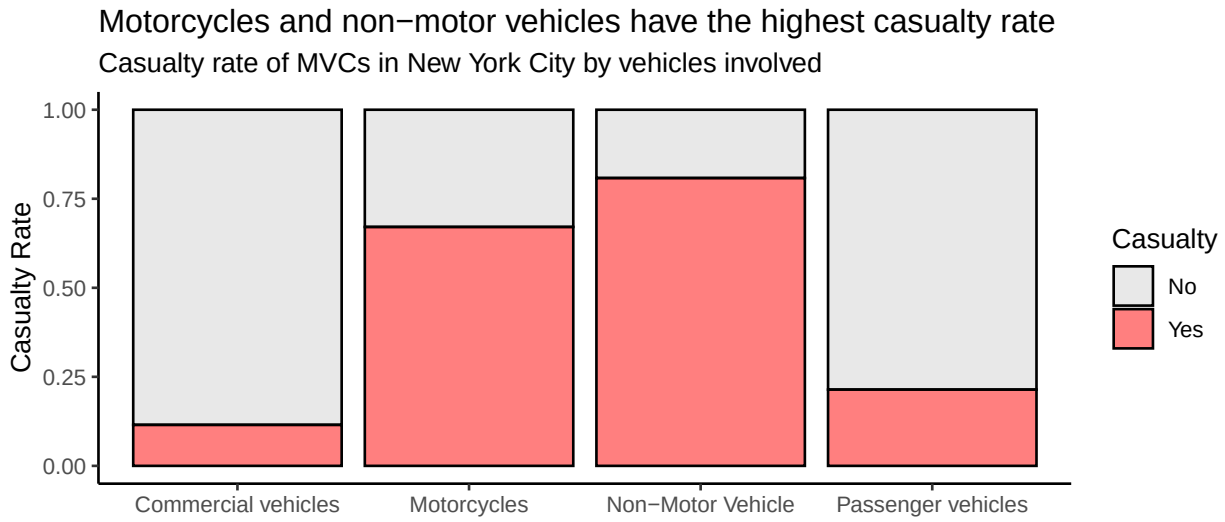
Our primary research concern is determining which characteristics of an MVC's timing and participants involved make casualties more likely.

Data Cleaning

For our analysis, we selected only MVCs between two or fewer motorists due to high levels of missingness in crashes with greater than two motorists. We then re-categorized the hundreds of vehicle types and contributing factors into fewer, larger bins, and created categorical variables for time of day and whether or not the crash took place during the weekend.

Exploratory Analysis

The following figure demonstrates the large difference in casualty rate between MVCs that did and did not involve motorcycles or non-motor vehicles. Similar bivariate analyses showed much smaller differences in casualty rates among different causal factors (see appendix).



Methodology

We fit a binary logistic regression model examining where there are associations between odds of casualty occurrence and peri-accident characteristics. Our predictors were chosen a priori using our original analyses (shown above and in the appendix) and using other research in the field (Mohamed et al., 2013; Zou et al., 2017).

Independence is reasonable because exactly one observation is made per accident, regardless of number of vehicles involved. Thus, accidents that impacted others would be considered in the same observation. Linearity, as assessed via empirical logit plots, was also reasonable (see appendix). Finally, we used likelihood ratio tests that suggested good model fit compared to other combinations of response variables and predictors ($p = 2.3 \times 10^{-16}$).

Results

Our model found four statistically significant ($p < 0.05$) predictors of casualties in MVCs. As expected from our exploratory analysis, MVCs have a much higher chance of casualty if motorcycles ($p = 1.03 \times 10^{-34}$) or non-motor vehicles ($p = 3.98 \times 10^{-94}$) were involved. These are the most significant predictors by far, with the others listed in the following table. For the complete results of the model, see the appendix.

Table 1: Significant predictors ($p < 0.05$)

Predictor	Odds ratio	P-value	95% CI
Involved motorcycle	10.25	1.2e-34	[7.11, 14.98]
Involved non-motor vehicle	18.89	2.0e-93	[14.36, 25.21]
MVC during evening	1.45	1.7e-06	[1.25, 1.7]
MVC at night	1.73	3.9e-11	[1.47, 2.04]
Failed to obey traffic rules	2.04	1.1e-17	[1.73, 2.39]
Impaired driver	1.36	2.9e-06	[1.19, 1.55]
Miscellaneous/unknown cause	1.46	1.4e-04	[1.2, 1.78]

Discussion

Whether or not an MVC involved a cyclist or motorcycle rider is by far the best predictor of the odds of a casualty occurring. However, should policy makers only be willing to make a single change to try to stem the flow of MVC casualties, improving conditions for bikers may not be the most efficient choice. While their casualty rate is the highest, in absolute number, non-motor vehicles and motorcycles combine to make up less than 200 of the 10,000 MVCs in our sampling.

To save the highest absolute number of casualties, improving evening- and night-time conditions is a better option, as the total number of casualties across times of day is much more uniform than across vehicle type.

While we were unable to investigate the specific causes of the increase in casualty rate during the later hours of the day, future research that includes light conditions from sunlight or streetlights, weather data, impairment rates, and other possible causes could further narrow in on a policy recommendation for city leaders. Additional research is also needed to investigate whether or not working to increases rates of cycling (thus having fewer cars to cause accidents) would be a net increase or decrease in absolute casualty count, especially given such high casualty rates in the accidents that *do* occur involving cyclists and motorcyclists.

While further research and additional data are needed, we hope that the insights that our model is able to predict can at least spark a conversation about potential policy decisions for urban planners in NYC and similar American cities.

Appendix

Data Dictionary

`has_casualty` (response variable): A boolean variable that is `True` when there was at least one injury or death in an MVC.

`involved_motorcycle`: A boolean variable that is `True` when one or more of the vehicles involved was a motorcycle, motor scooter, or moped (generalized to "motorcycle" in this report).

`involved_non_motor`: A boolean variable that is `True` when one of the vehicles involved was a bicycle or scooter (electric or analog).

`time_day`: A categorical variable that is `morning` if the MVC occurred between 5:00 am and 11:59 am, `afternoon` if between 12:00 pm and 4:59 pm, `evening` if between 5:00 pm and 8:59 pm, and `night` if between 9:00 pm and 4:49 am.

`weekend_weekday`: A categorical variable that is `Weekend` if an MVC occurred on Saturday or Sunday, and `Weekday` otherwise.

`yday`: A numeric variable representing the cumulative day of the year (1-365) on which an MVC occurred.

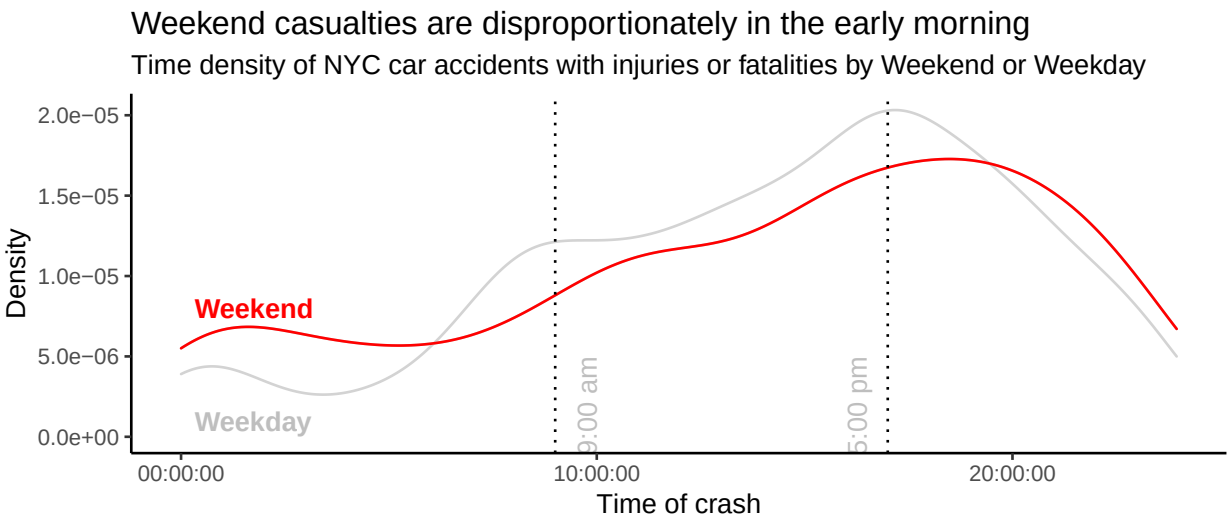
`failed_to_obey`: A boolean variable that is `True` if one of the listed causes was failure to obey traffic signs/signals/rules.

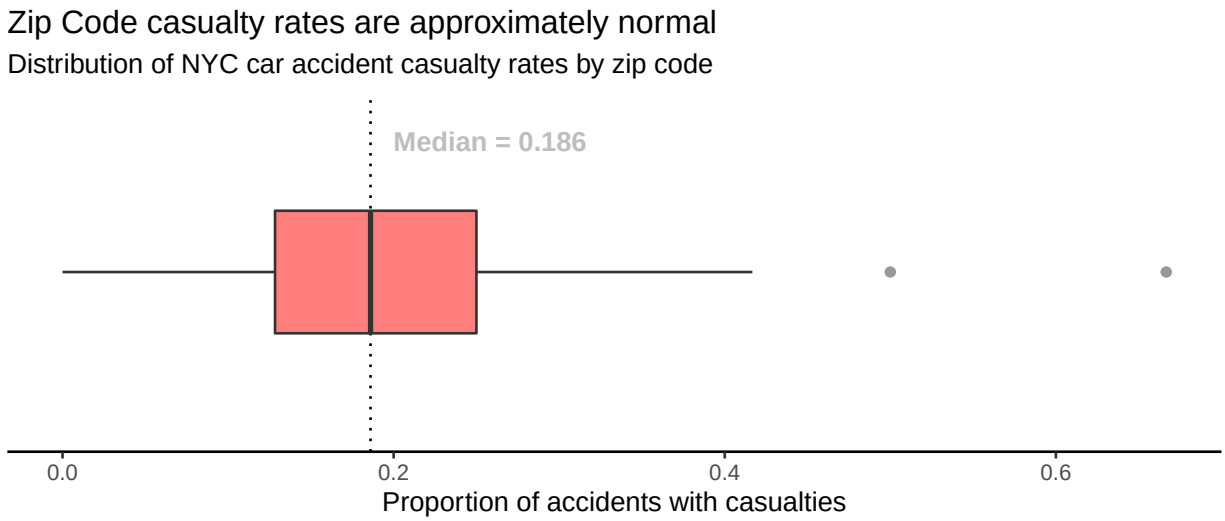
`was_impaired`: A boolean variable that is `True` if one of the listed causes was impairment, fatigue, or fatigue.

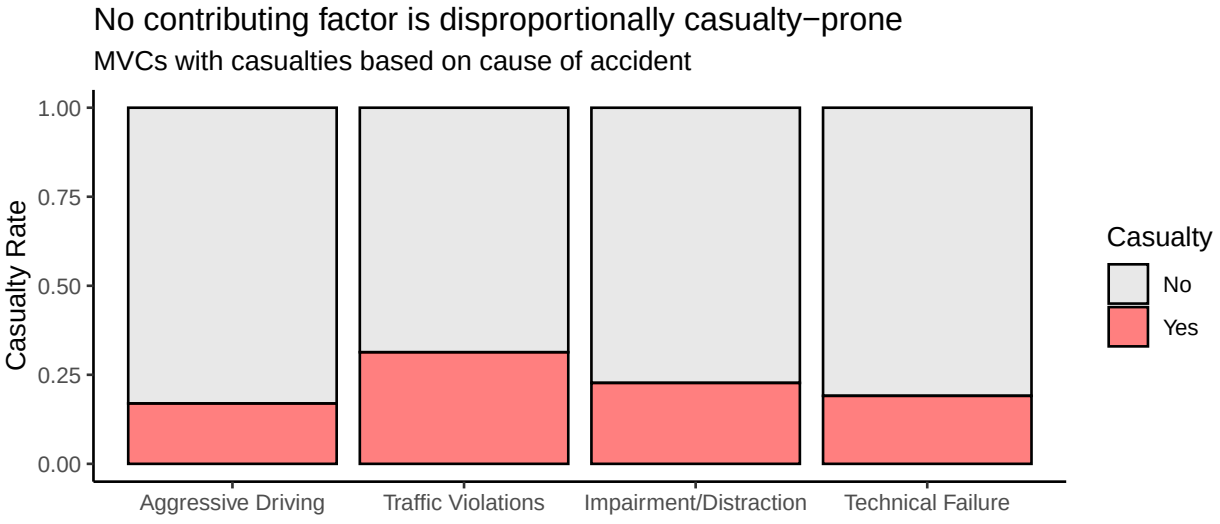
`mech_failures`: A boolean variable that is `True` if one of the listed causes was performance-unrelated mechanical failures of one or more vehicles.

`misc_cause`: A boolean that is `True` if the cause of an MVC is unknown or does not fall into any other category.

Exploratory Data Analysis







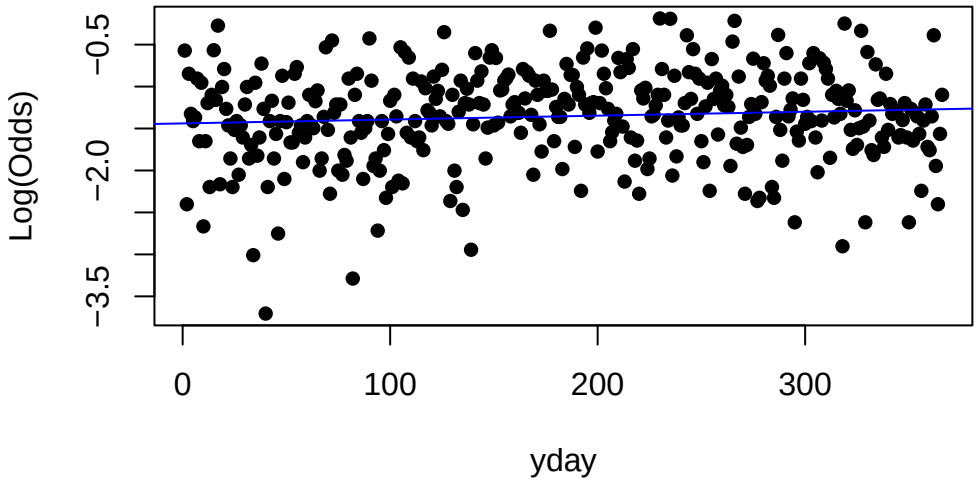
3. Methodology

a) Likelihood-ratio test between models considered

The following output shows that `Model 2` is superior ($p = 2.3 \times 10^{-16}$), and was therefore used in our final analysis.

```
  Resid. Dev Df Deviance  Pr(>Chi)
1     8524.2
2     8444.7  4   79.427 2.303e-16 ***
3     8441.7  3    3.037    0.3859
---
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. Final model

The following table is the complete model used in our analysis. Variable names are raw, as seen in our code.

Table 2: Full model output

term	estimate	p.value	odds
Baseline	-2.34e+00	7.06e-73	9.66e-02
Involved Motorcycle	2.33e+00	1.22e-34	1.02e+01
Involved non-motor vehicle	2.94e+00	2.00e-93	1.89e+01
Afternoon	1.04e-01	1.61e-01	1.11e+00
Evening	3.75e-01	1.73e-06	1.45e+00
Night	5.49e-01	3.94e-11	1.73e+00
Weekend	6.83e-02	2.75e-01	1.07e+00
Day of year	1.18e-04	6.55e-01	1.00e+00
Failed to obey	7.11e-01	1.10e-17	2.04e+00
Impaired	3.07e-01	2.94e-06	1.36e+00
Mech. Failures	1.68e-01	1.27e-01	1.18e+00
Misc./Unknown cause	3.78e-01	1.42e-04	1.46e+00

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

Mohamed, M. G., Saunier, N., Miranda-Moreno, L. F., & Ukkusuri, S. V. (2013). A clustering regression approach: A comprehensive injury severity analysis of pedestrian–vehicle crashes in New York, US and Montreal, Canada. Safety Science, 54, 27–37. https://doi.org/10.1016/j.ssci.2012.11.001

Zou, W., Wang, X., & Zhang, D. (2017). Truck crash severity in New York City: An investigation of the spatial and the time of day effects. Accident Analysis & Prevention, 99, 249–261. https://doi.org/10.1016/j.aap.2016.11.024