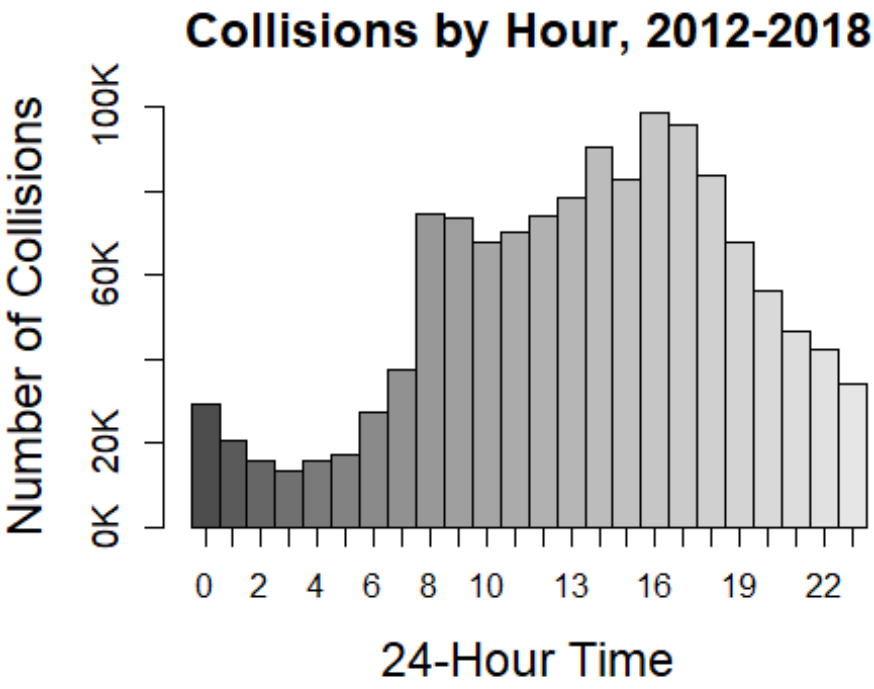
NYC Council Collision Report

Sanata Sy-Sahande

August 16, 2018

Best and Worst Driving Times

To identify driving times, I plot below the frequency of crashes by hour across all observations in the dataset, from 2012-2018. The best driving times (times when crashes are at their lowest levels) are in the early morning. The worst driving times are during morning and evening rush hour (max at 16h).



Factors Impacting Injury Rates

The dataset identifies 50+ contributing factors to collisions, ranging from road conditions and vehicle-specific factors to driver behavior. To identify which factors are most associated with injury and fatal crashes, I compared the injury and mortality rates rates across all the factors. The greater the rates, the more likely this factor is in predicting high injury and mortality rates.

First, we look at the top contributing factors overall in the table below. These make up 90 percent of all factors cited. We see that improper lane passing, driver distraction, and vehicle problems make up half of all contributing factors.

	Pct
Passing or Lane Usage Improper	31
Driver Inattention/Distraction	10
Other Vehicular	8
Backing Unsafely	7
Pedestrian/Bicyclist/Other Pedestrian Error/Confusion	7
Brakes Defective	7
Reaction to Uninvolved Vehicle	5
Turning Improperly	3
Passing Too Closely	3
Fell Asleep	3
Unsafe Lane Changing	3
Failure to Yield Right-of-Way	3

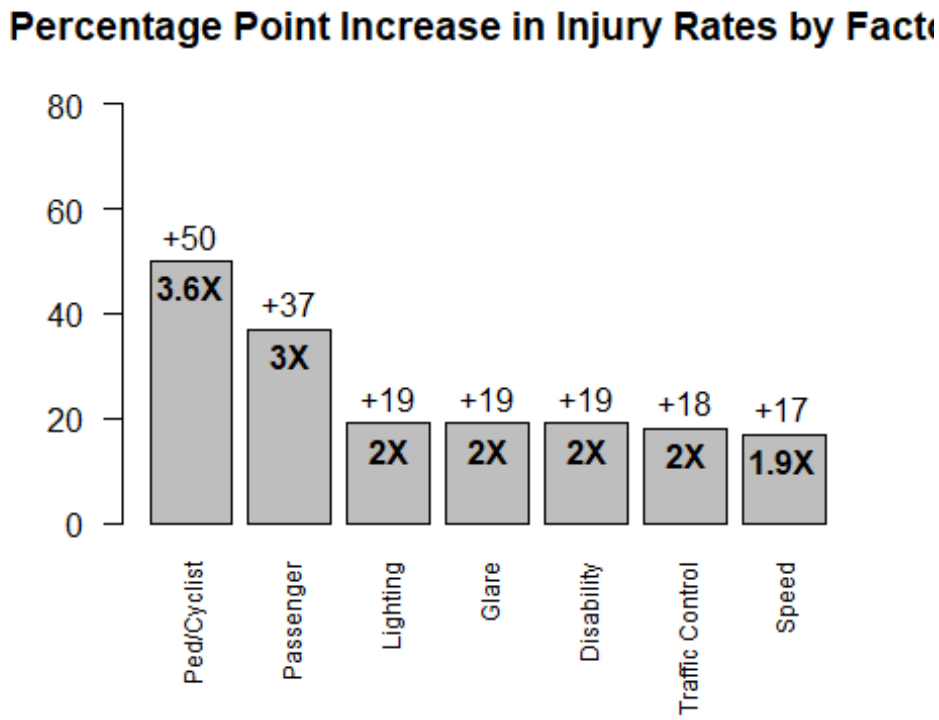
Next, we look at injury and mortality rates, as well as crashes that involved both. This is important to establish a baseline so that we can determine what counts as a "high" mortality or injury rate. Fatal crashes are the rarest, making up .1% of all cases.

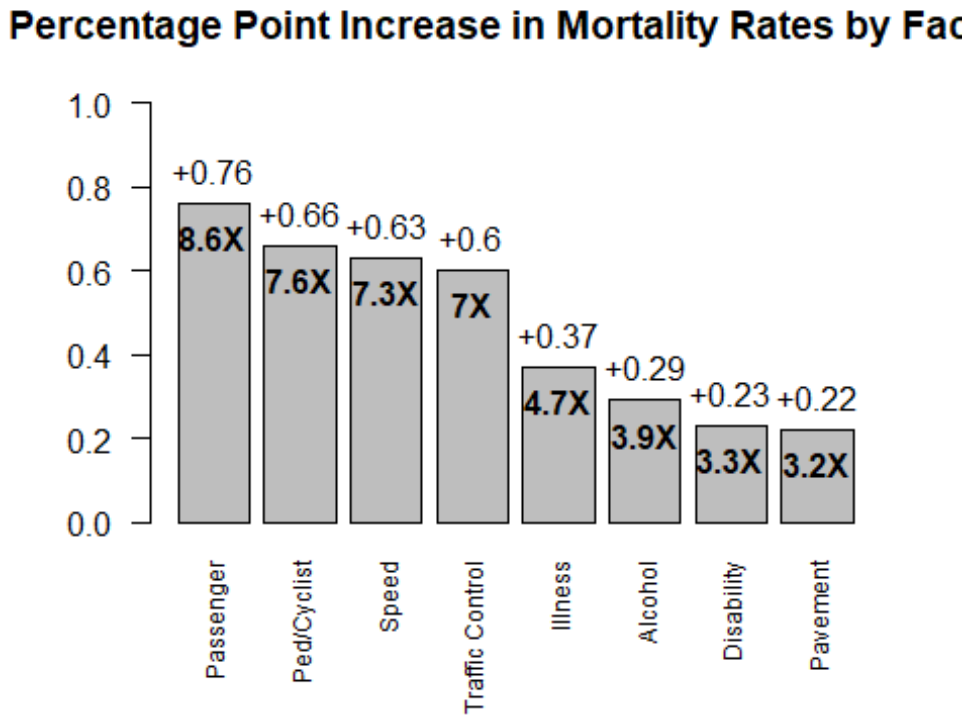
	victim_mean	injury_mean	fatal_mean
Percentage	18.9	18.8	0.1

I then identified the factors that had injury and mortality rates that were at least twice as high as the baseline rates shown above. The barplots below show the factors associated with the greatest increases in injury and mortality rates. The bars indicate how much higher the rate was for each factor relative to the baselines presented above, denoted by the + sign. For a more intuitive interpretation, I also indicate how many times higher the injury and mortality rates are for each factor relative to the baseline, denoted by the X sign.

For injury rates, the most significant factors were collisions involving pedestrian or cyclists, passenger distractions, lighting defects and glare, physical disability, ignored traffic controls, and unsafe speeds. Notably, pedestrian confusion and passenger distractions more than tripled injury rates.

For mortality rates, the most significant factors were similar to those for injury rates, with the addition of alcohol, illness, and pavement defects. Speed and ignored traffic controls mattered much more, increasing mortality rates by a factor of 7.

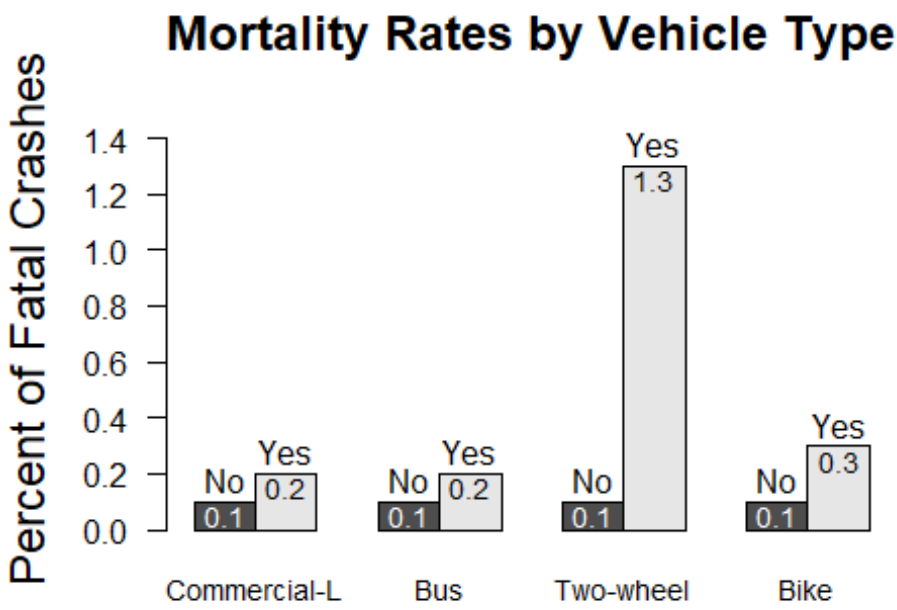




Effect of Vehicle Type on mortality rates

Out of 13 vehicle types, I identified the four best predictors of fatal crashes. Below, I compare the mortality rates in crashes with and without the vehicle type. Two-wheel

motor vehicles (scooters and motorcycles) had the greatest positive difference, resulting in mortality rates that are 10 times higher than the average.



Exploratory Analysis

I tried a different method by turning the prompt into a classification problem: using vehicle type or contributing factor, which items best predict whether a crash will result in injuries or fatalities? Once I identify these factors, the next steps would be to include them in a logistic regression with injury or fatality as the dependent variable, fit the model to a random sample of the dataset as my training data, and get out of sample predictions using the remaining observations to validate.

I only show the first step here for vehicle types and injuries: I calculate the rates at which we observe each vehicle type in injury crashes vs. no-injury crashes and determine the extent to which they are over- or under-represented in crashes with injuries using two-sample proportion t-tests.

I then plot these results and estimate the distance between each point and a y=x line, which indicates equal probability of being observed in either type of crash. The results of this analysis largely replicate the findings of the first method, adding greater confidence to my estimates.

```
d <- sqrt(((log(ptest.inj$prop1/ptest.inj$prop2))^2)/2)
ptest.inj$dist <- d

plot(log(ptest.inj$prop1), log(ptest.inj$prop2), pch= ifelse(ptest.inj$dist >
```

```
=0.64, 16, 1 ),
     main = "Rates of Contributing Factors (Logged)", xlab = "Injury", ylab =
"No Injury")
text(log(ptest.inj$prop1[ptest.inj$dist >=0.64]), log(ptest.inj$prop2[ptest.i
nj$dist >=0.64]) - 0.3, substr(rownames(ptest.inj)[ptest.inj$dist >=0.64], 1,
20), cex = 0.75)
lines(c(-10, 10), c(-10, 10), lty = 2)
```