Project: Overall Safety of Vechicles

CUNY SPS: DATA606

Richie Rivera

May 13, 2024

Abstract

Since 2012, cars have been growing larger and heavier and this phenomenon has has been reported upon more often. For example, The Economist among other publications have published articles that are showing that cars across North American and Europe are growing larger.

This is a worrying race to the bottom as larger vehicles are worse for the environment and are seemingly much less safe for society. Across the country, there are a growing number of local governments implementing vision 0 initiatives to reduce the number of vehicular related injuries and fatalities.

In this report, we will use statistical methods to verify if larger vehicles are indeed statistically significantly more likely to injure or kill someone in the event of a crash.

Data

We will be working with the "Motor Vehicle Collisions - Crashes" dataset provided by the City of New York. This dataset contains vehicular crashes in NYC which were reported by NYPD from July 2012 through the end of April 2024. Each observation in the dataset represents one crash.

This dataset contains 2,084,770 observations with 1,651 unique entries for vehicle class. For this analysis, we will be grouping each vehicle class into either "Passenger Vehicle", "SUV/Pickup Truck", or "Other". We will also observe the fields corresponding to the number of people killed and injured per crash.

Research Question

Is there a statistically significant difference in incidents between vehicles of different body types when a crash results in an incident?

For the purposes of this analysis, an incident is a crash where at least one person is injured or killed as a result of a crash.

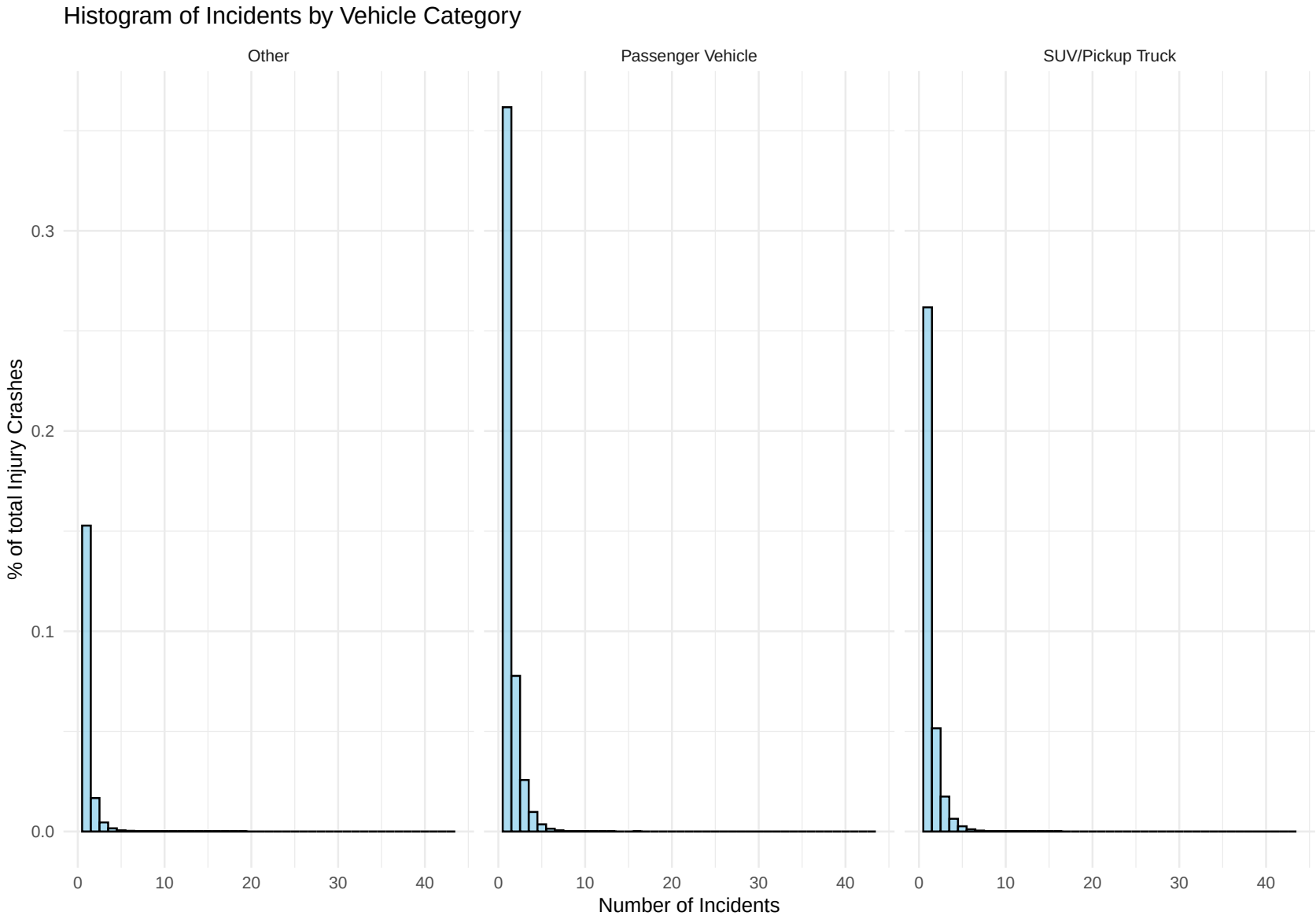
We will investigate this by looking at the sum of incidents (dependent variable) by vehicle class (independent variable).

Our null hypothesis is that there is no difference in incident rates across different vehicle classes and our alternative hypothesis is that there is a difference in incident rates across different vehicle classes.

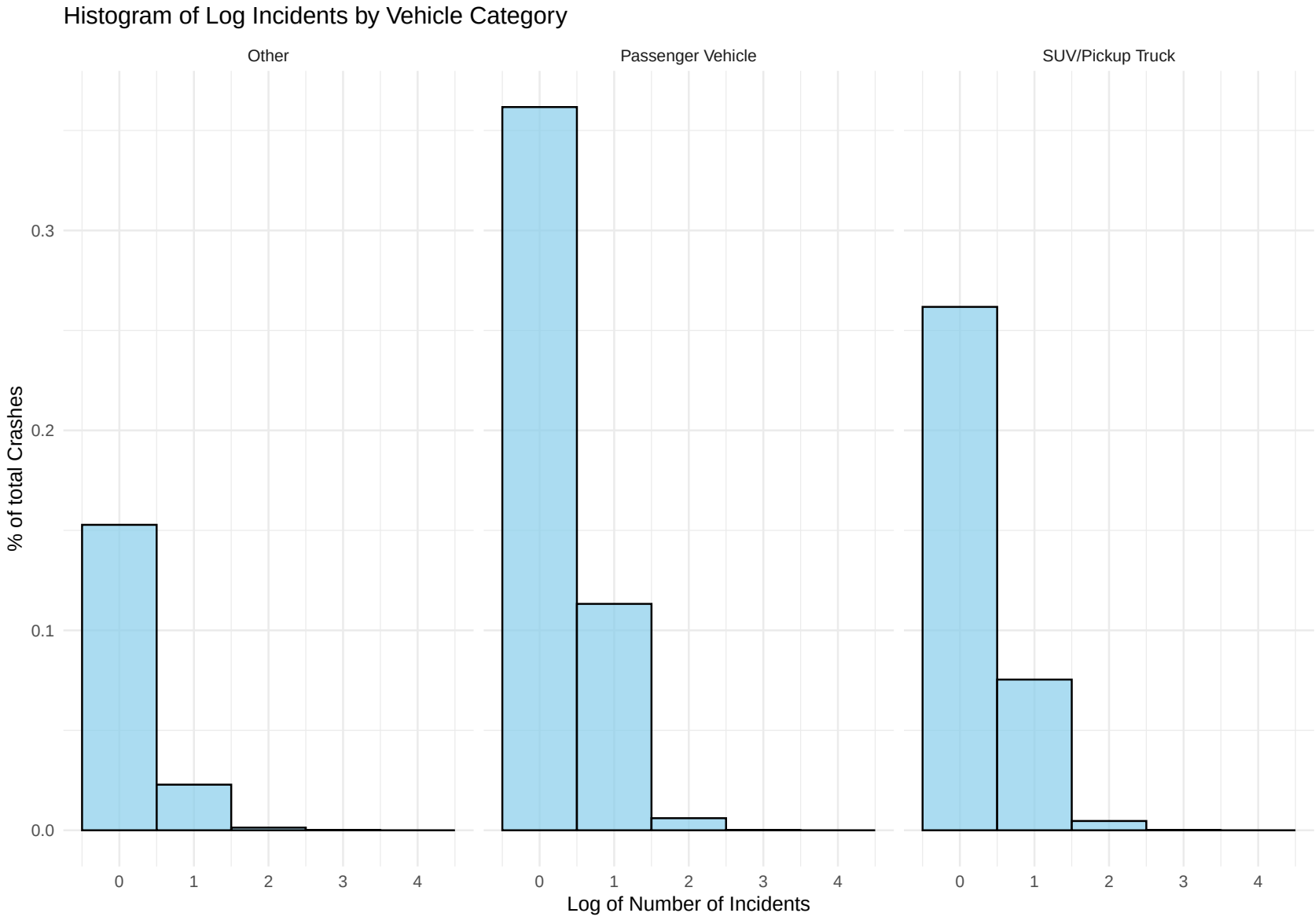
Summary Statistics

```
##
##  Descriptive statistics by group
## vehicle_category: Other
##                                vars     n mean   sd min max range se
## TOTAL.PERSONS.INJURED.KILLED      1 85048 1.22 0.80   1  40    39  0
## NUMBER.OF.PERSONS.KILLED          2 85048 0.01 0.11   0   8     8  0
## NUMBER.OF.PERSONS.INJURED         3 85048 1.21 0.80   0  40    40  0
## ------------------------------------------------------------
## vehicle_category: Passenger Vehicle
##                                vars      n mean   sd min max range se
## TOTAL.PERSONS.INJURED.KILLED      1 231084 1.39 0.85   1  43    42  0
## NUMBER.OF.PERSONS.KILLED          2 231084 0.00 0.08   0   5     5  0
## NUMBER.OF.PERSONS.INJURED         3 231084 1.39 0.85   0  43    43  0
## ------------------------------------------------------------
## vehicle_category: SUV/Pickup Truck
##                                vars      n mean   sd min max range se
## TOTAL.PERSONS.INJURED.KILLED      1 164222 1.38 0.87   1  25    24  0
## NUMBER.OF.PERSONS.KILLED          2 164222 0.01 0.08   0   5     5  0
## NUMBER.OF.PERSONS.INJURED         3 164222 1.37 0.88   0  25    25  0
```

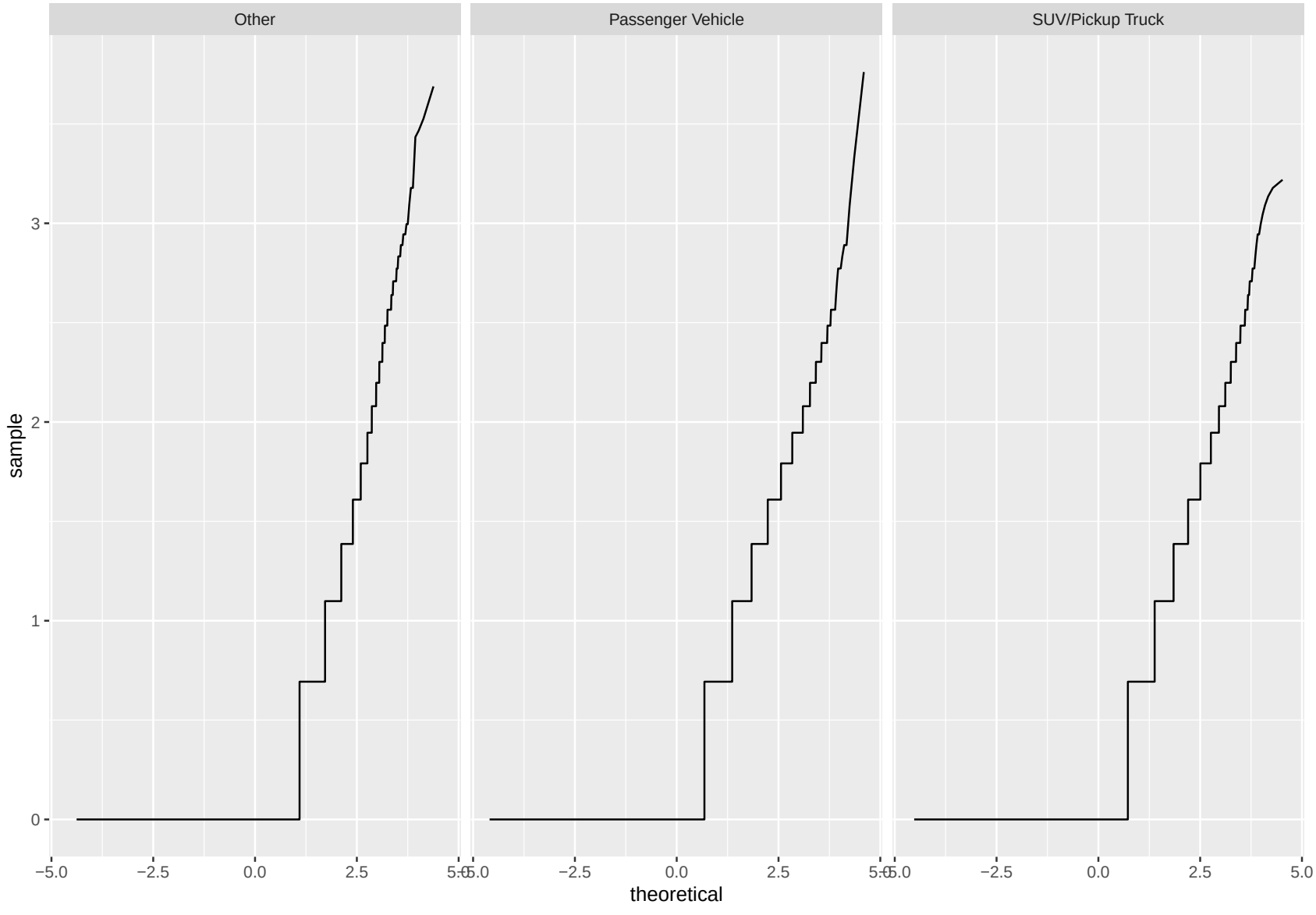
Histogram of Incidents



Histogram of Log Incidents



QQ Plot of Log Incidents



From here we can see that the distribution is somewhat normal.

Requirements for ANOVA

1. The Observations are independent within and across groups
2. The data within each group are nearly normal
3. The variability across the groups is about equal

We meet all of these requirements and can proceed with ANOVA.

ANOVA Results for log of Incidents

```
## # A tibble: 2 x 6
##   term                df  sumsq  meansq statistic p.value
##   <chr>            <dbl>  <dbl>   <dbl>     <dbl>   <dbl>
## 1 vehicle_category     2   646. 323.        2031.       0
## 2 Residuals       480351 76452.   0.159       NA       NA
```

From the summary above, we can see that the F-Value of 2031 indicates that there is a difference between the number of injuries/fatalities and the p.value of essentially 0 signifies that the result is statistically significant.

Regression of the Results

```
##
## Call:
## lm(formula = log.TOTAL.PERSONS.INJURED.KILLED ~ vehicle_category,
##     data = log_hurt_crashes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.2209 -0.2209 -0.2106 -0.1213  3.5676
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      0.121288   0.001368   88.66   <2e-16 ***
## vehicle_categoryPassenger Vehicle 0.099616   0.001600   62.26   <2e-16 ***
## vehicle_categorySUV/Pickup Truck  0.089343   0.001685   53.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3989 on 480351 degrees of freedom
## Multiple R-squared:  0.008384,   Adjusted R-squared:  0.00838
## F-statistic:  2031 on 2 and 480351 DF,  p-value: < 2.2e-16
```

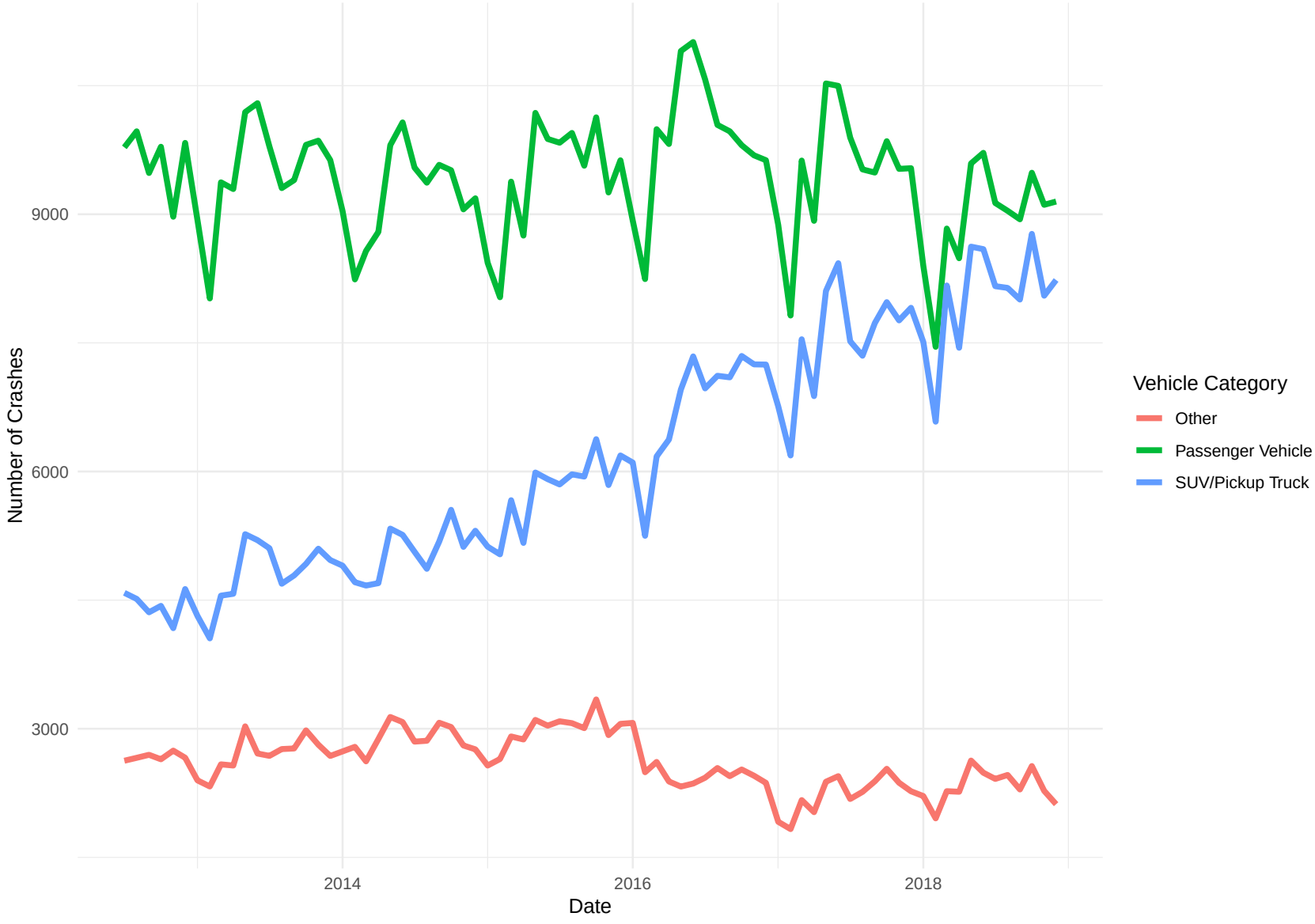
From here we can see that the coefficient for the Other category is the highest, followed by Passenger Vehicles, and SUV/Pickup Truck. Although we observe this result, our Adjusted R-squared of 0.008 suggests that only 0.8% of the observed variance is due to the vehicle category. Interpreting our p value of essentially 0, we can also know that this result is statistically significant.

Conclusion

Recalling the means from before, we found that passenger vehicles had the greatest mean of number of incidents. So these results would imply that passenger vehicles are more likely to result in an incidents (1.390 vs 1.377).

As reported by NHTSA, there were 42,939 people killed and 2,497,657 people injured in car crashes across the country in 2021. If we assume that lives are precious then understanding the human cost associated with what vehicles we choose to drive is crucial.

Total Monthly Crashes by Vehicle Category



Limitations & Next Steps

This analysis was conducted only looking at New York City. These results would likely vary across different levels of government and locations.

As a next step, an interesting observation is that although the mean injury per crash for SUV/Pickup Trucks are lower than those for Passenger Cars, it's noted that the fatality rate is higher (0.006235 vs 0.004825) which could provide evidence that they are more fatal. This would require further investigation to show evidence for or against that idea.

References

- https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes
  - Accessed May 3, 2024
- https://www.economist.com/the-economist-explains/2024/03/11/why-american-cars-are-so-big
  - Accessed May 3, 2024