Data Mining, Analysis, & Exploration of NYC Car Accident Data, 2019-2022

Jen Arriaza

NYU School of Professional Studies

1 Introduction

This paper will examine car crash data for the past 3 years, directly sourced from NYC Open Data, and discuss the steps taken in the data mining process. The goal of this analysis is to explore the data for any interesting findings through an exploratory analysis. Any potential findings from the data may be important insight for drivers, pedestrians, and bicyclists in NYC. The analysis examined the data for patterns and any points of interest pertaining to car accidents in the five boroughs. This data is also useful for infrastructure development, and to inform policies or technologies that can make travel on roads safer. Informing technologies with real-world accident data can help to optimize developing automotive tech.

2 Main Questions

This paper will attempt to answer the following questions about the NYC car accident data: are there any points of interest where accidents occur? Which neighborhoods in NYC have more fatal accidents than others? What are the observable patterns regarding time? The initial body of research will discuss these findings in-depth, and then explore subsequent questions that emerged as data exploration progressed.

The subsequent questions were as follows: What are the seasonal trends in the car accident data? Is there any rate or proportion difference between the types of vehicles involved in accidents? Are any vehicle types involved in more fatal accidents than others? It is theorized that the rise in overall accident fatalities may be related to the increased number of SUVs on the road. Exploration of the dataset revealed the ability to answer several of these questions, which will be presented in the form of analysis results and visualizations.

3 Literature Review

The motivation for exploring the NYC car accident data is multi-faceted. With dramatic changes in the way people live and move around in the NYC through the COVID-19 pandemic, it is interesting to know how this may've affected accidents. More importantly, it is important to policymakers and safety engineers to know where more serious are occurring. Furthermore, it may be interesting to know if any trends are consistent in NYC accidents with other accident data.

A relevant topic to NYC car accidents is concerns surrounding the growing percentage of pedestrian and cyclist fatalities (Government Accountability Office, 2015). The study found results of particular interest to urban areas such as New York City and the surrounding boroughs: the National High Traffic Safety Administration reports that 73% of pedestrian fatalities occur in urban areas, while 23% occur in rural areas. This is consistent with the fact that vehicles present the greatest risk of injury and death to pedestrians and more vehicles will populate urban areas.

Also relevant to the topic of motor vehicle accidents in NYC is the larger observable trend in the usage of Sports Utility Vehicles (SUVs). A study published in March 2022 from the Insurance Institute for Highway Safety found that SUVs are more likely to kill pedestrians than standard cars. The study found that although advances in safety and driver assistance systems has reduced the number of fatalities on the road by approximately 30% since 1980, the number of pedestrians killed on U.S. roads has steadily ticked upward. Furthermore, the study reports that SUVs are substantially more lethal to pedestrians than cars (IIHS, 2022). This report is relevant more than ever as the proportion of SUVs on U.S. roads has risen around 8% since 2009, and these numbers are expected to continue rising. This paper will explore whether these trends are observable in NYC, and if so to what degrees.

4 Dataset Description

This analysis was conducted in Jupyter Notebooks using Python and Tableau. The main goal was to investigate for any patterns through an exploratory data analysis and implementation of the data mining process. The key variables are vehicle type, number of accidents, and number of people

injured—broken down by driver, pedestrian, and cyclists. This paper will explore findings from data broken down by month, year, and location data to examine any potential trends.

The data is collected by NYC Open Data and sourced from NYPD police reports. [1] The data is observational, as it is not processed or intervened in any way when collected. The sample in this analysis contains all accidents in the database starting January 2019 through the start of March 2022. The total amount of accidents in the sample is 451,310 events. Since our focus is to examine potentially certain points of interest of accidents in NYC, such as level of severities, density in geographic areas, etc.—the data was sliced into subsets for certain components of an analysis.

.

5 Data Pre-Processing, Selection, Mining, and Transformations

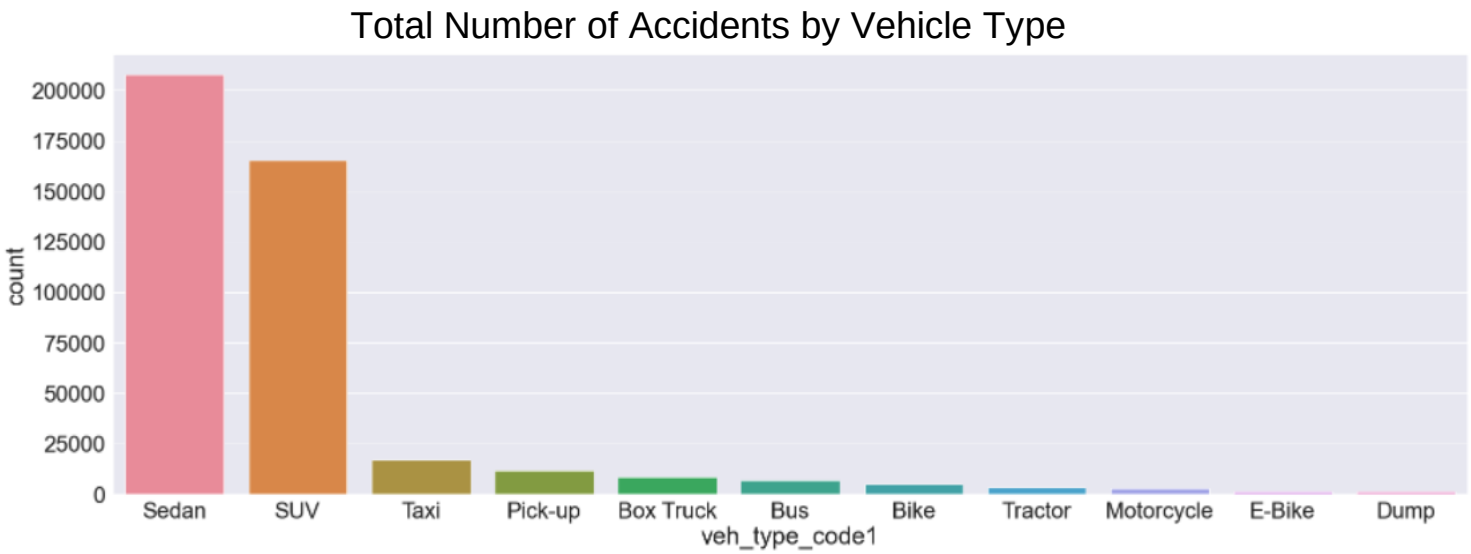
The dataset for the selected time period of January 2019 to March 2022 had a pre-cleaned and pre-selected total of over 430k records and 26 columns. The dataset was inspected for missing values, duplicates, and outliers. Most of the columns had at least one missing value, which were likely due to the information not applying for that accident—i.e., man accidents in the dataset did not involve more than two vehicles, so the columns for ‘vehicle3_type’ will be NULL. For conciseness of the analysis and ease of implementation with the very large dataset size, these columns were excluded from analysis and rows with missing values for critical features such as borough and geocoordinates were excluded.

There were 185 duplicates found in the dataset. The duplicated events did not present any observable patterns and were attributed to human error. The duplicate items were removed from the main dataset before analysis. Outliers were searched by examining percentiles, where the 1% and 99% were used to remove outliers.

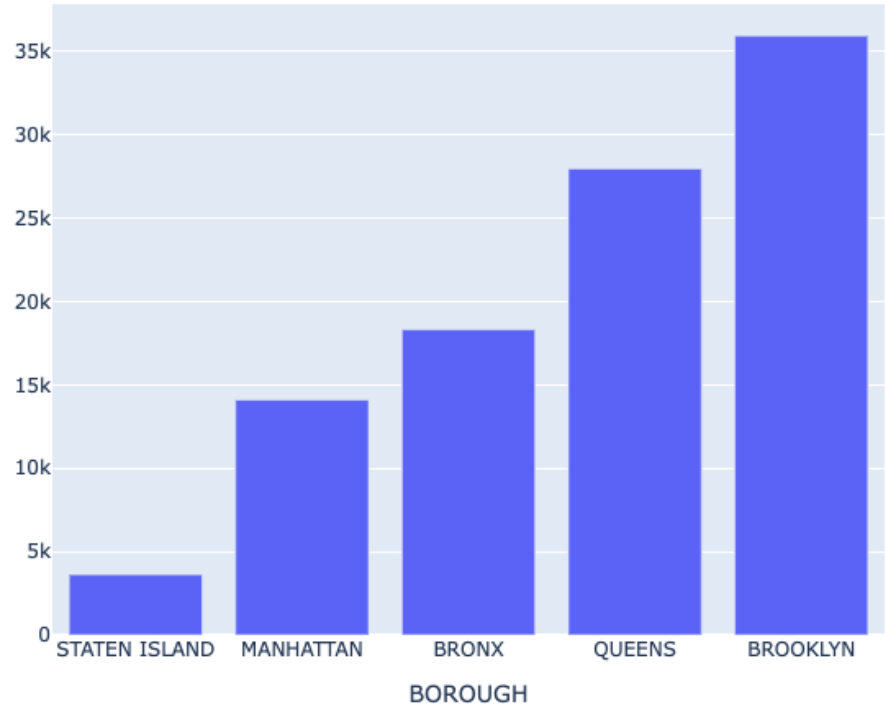
As the data was examined further, it was clear that some of the calculations were being affected by the fact that the data for 2022 was incomplete, as the data was retrieved March 2022 and the other years include all 12 months of accident data. For this reason, the accident data for 2019-2021 was selected for the predictive analytics component of this project.

The initial exploratory analysis includes all vehicle types within the accident data. To further drill-down the analysis to the most relevant data, the accident

data comparing sedans and SUVs was selected for some of the analytics components of this project. The following graph shows the distribution of accident totals aggregated by vehicle type, showing that sedans and SUVs make-up most of the data.



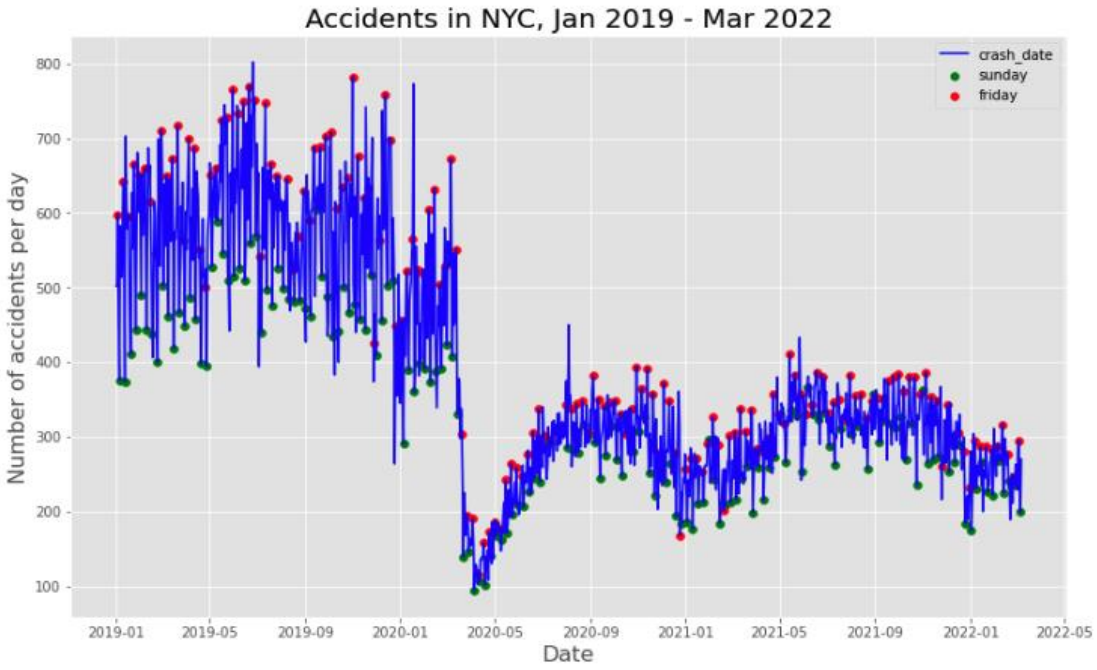
Similarly, the initial exploratory analysis includes all NYC boroughs within the accident data. It was found that Brooklyn has the highest occurrence of accidents. The machine learning component of this project sub-selected to accidents in the Brooklyn area. The below graph shows the distribution of accident totals aggregated by borough, showing that Brooklyn and Queens had the highest occurrence of accidents (Petrou, 2017).



In further exploration of the dataset, it was determined that classification of accident severity would be most useful for the predictive analytics component of the project. Accidents where no injuries and no fatalities occurred were classified as Severity level “Slight”. Accidents where at least one injury occurred were classified as “Serious”, and accidents where at least one fatality occurred were classified as “Fatal”. This would be useful for later implementation of one-hot encoding for to perform predictive classification with machine learning algorithms.

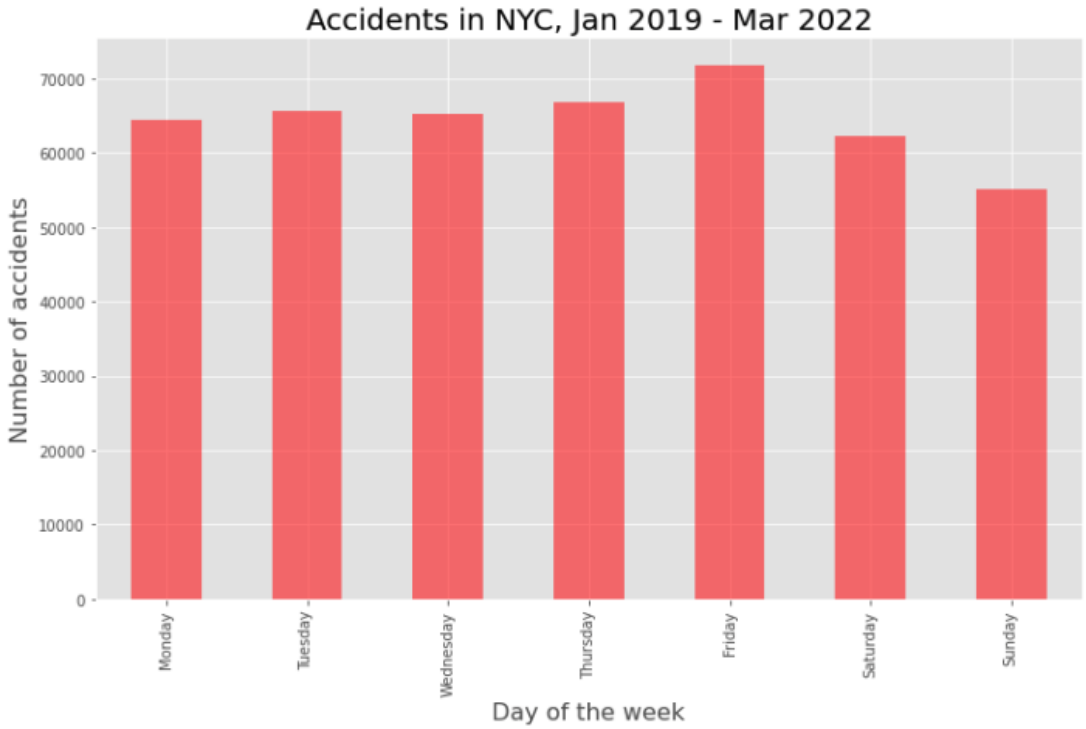
5b Data Analysis

First, a time series analysis was performed on the entire dataset to examine for any seasonal trends or patterns related to time.

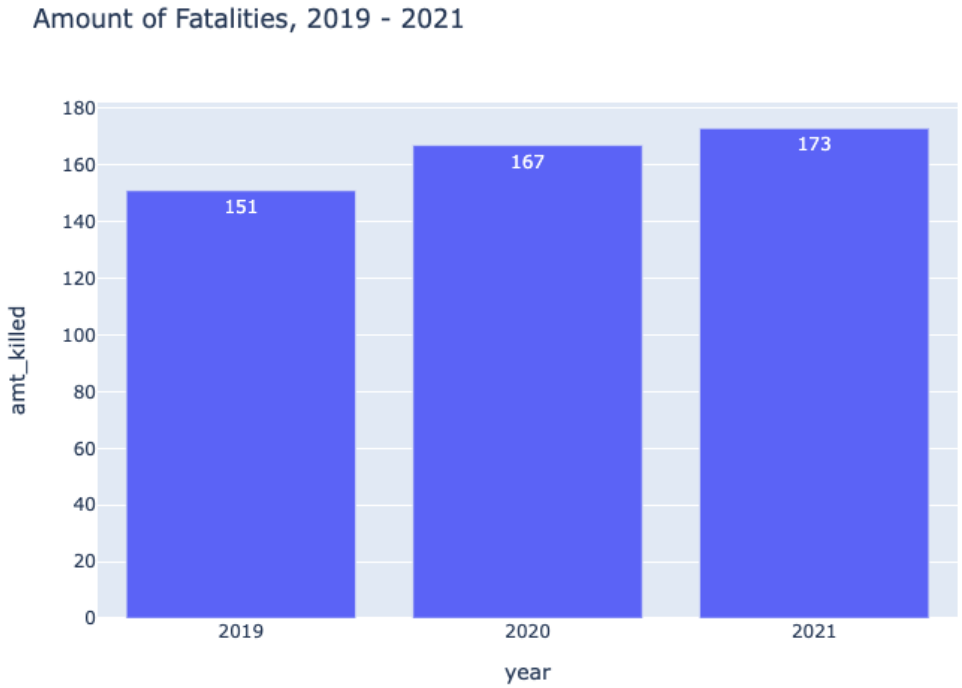


The above time series scatter plot shows accident number trends for the past 3 years. There’s a drop in March 2020 when COVID-19 quarantines began, then numbers steadily rose as restrictions eased. However, accident numbers do not look like they will reach pre-pandemic levels anytime soon—potentially reflecting the adoption of permanent remote work policies, as well as large numbers of people of moving out of NYC.

The graph also shows that the highest number of accidents occur on Friday, and the lowest number of accidents occur on Sundays. This pattern may indicate that there is correlation between the availability of business, commerce, or other entities and the overall number of accidents. The graph below indicates that there are steady totals throughout most days of the week.



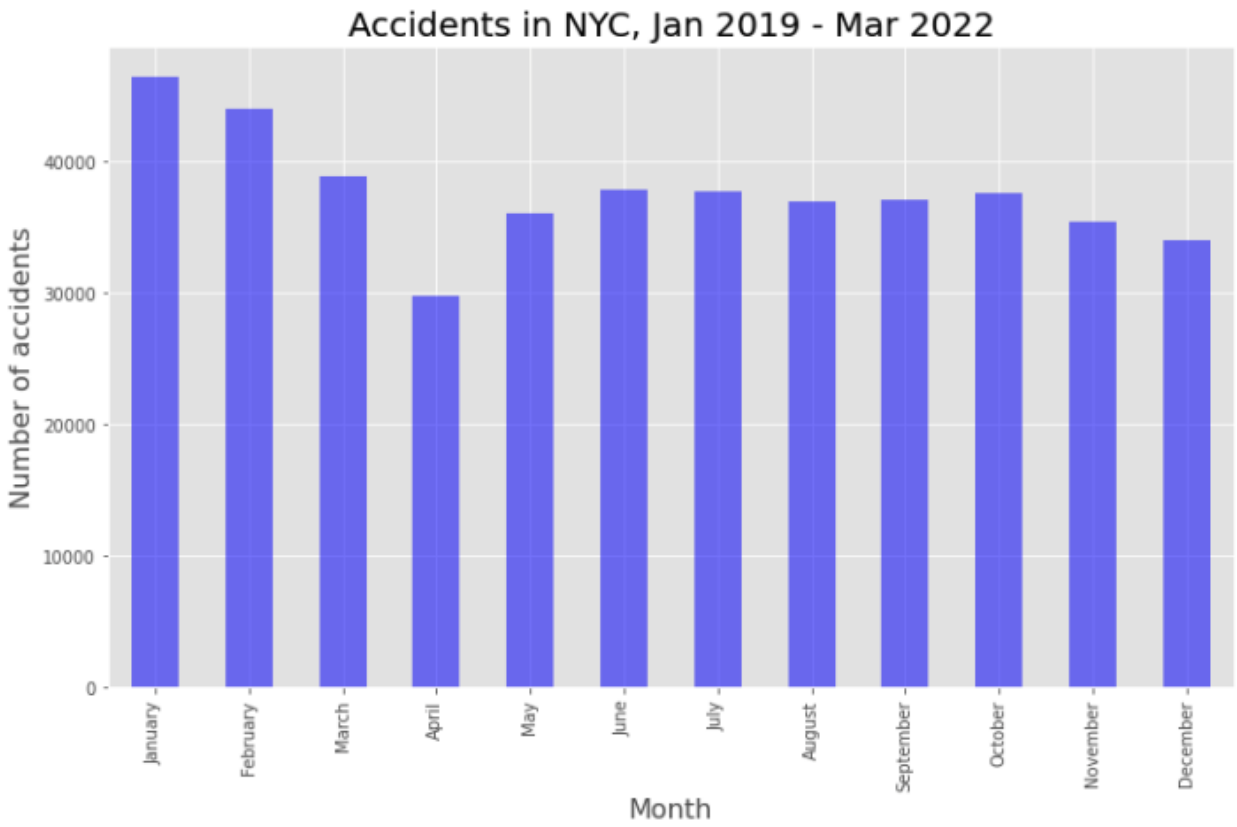
One very interesting find in the data is that although there are significantly less vehicles on the road during and after the pandemic in NYC—and subsequently less accidents—the number of accident-related fatalities is showing an upward trend in the graph below.



This contrasts with the downward trend in overall accidents. The distribution of accidents is smoothly going upwards in the bar graph. It will be interesting to

see what the totals for 2022 are and whether it is consistent with the previous years.

Further analysis also shows that there may be patterns in the number of accidents that happen according to the time of year. The below graph shows the number of accidents that happened each month since 2019.



There is an uptick of accidents in the month of January, perhaps due to weather conditions or another unknown factor, and accidents were lowest in April before steadily rising back up in the summer months.

Statistics

The data was sliced to compare the accident data between Sedans and SUVs and perform an exploratory data analysis of the subset. This will answer one of the main questions of this paper: is there a difference in observable patterns between SUVs and Sedans?

To drill-down our data to the main points of interest, we focused our analysis on Sedans and SUVs since they make up > 90% of the accident data.

The total count and mean number of accidents each year is represented in the below table.

vehicle_type	sum	mean	std
SUV	165139	41284.75	30114.900037
Sedan	207559	51889.75	35266.054655

Figure 6: The above table displays sum & mean number of accidents each year by vehicle type.

The intervals for SUVs and Sedans at 95% confidence level are 41,140 < μ < 41,430 and 51,738 < μ < 52,041 respectively.

vehicle_type	sum	mean	std
SUV	223	55.75	30.739497
Sedan	269	67.25	35.873621

Figure 7: The above table displays the sum & mean number of crash fatalities each year by vehicle type. The intervals for SUVs and Sedans at 95% confidence level are 52 < μ < 60 and 63 < μ < 72 respectively.

vehicle_type	sum	mean	std
SUV	138	34.5	19.226718
Sedan	104	26.0	14.809907

Figure 8: The above table displays sum & mean number of pedestrians killed each year by vehicle type. The intervals for SUVs and Sedans at 95% confidence level are 31< μ < 38 and23 < μ < 29 respectively.

The statistics indicate that Sedans have a greater overall number of accidents, injuries, and fatalities. This is consistent with the fact that Sedans are the most popular type of vehicle on the road in general. The high standard deviations indicate there may be high variance in the accident data, or that some outliers may exist.

To further understand the patterns between accidents involving Sedans vs. SUVs, these figures were examined in the below data visualizations.

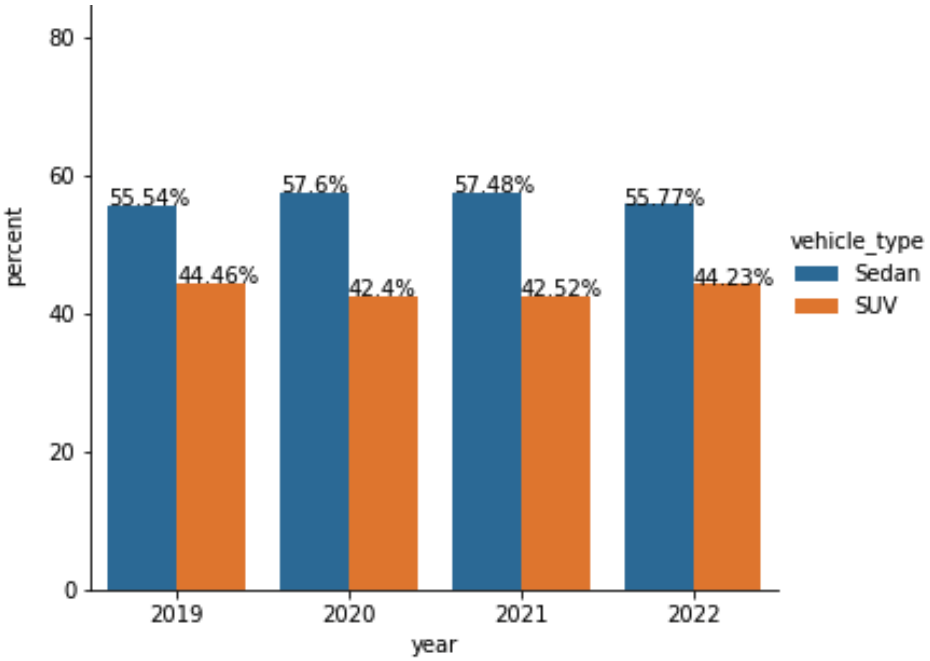


Figure 9: Sedans have a higher proportion of accidents than SUVs.

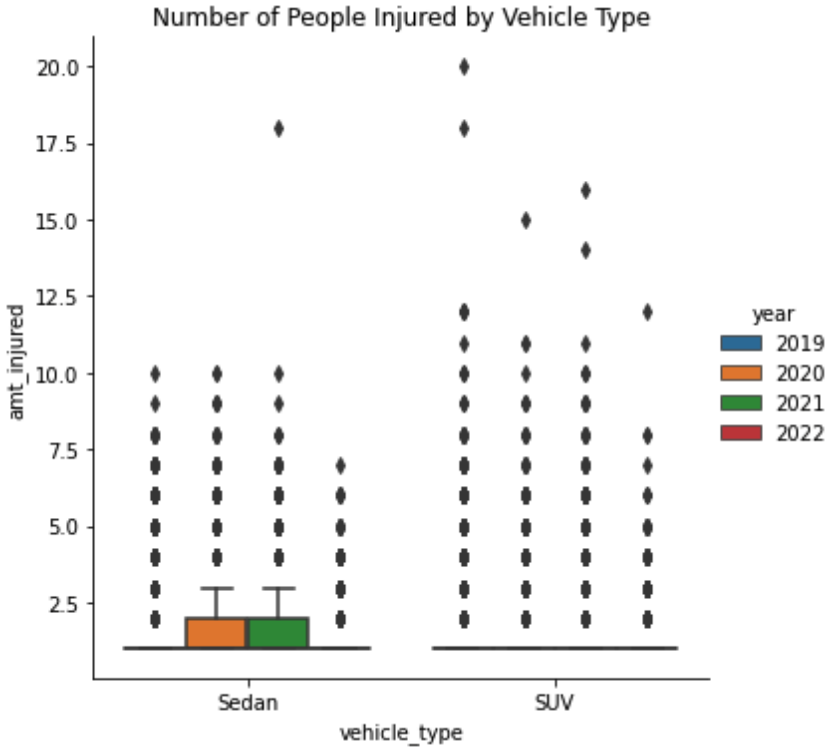


Figure 10: The above box plot indicates that Sedans have a greater amount of accident injuries, and SUVs have had unusually high data points (potential outliers) for amount of people injured.

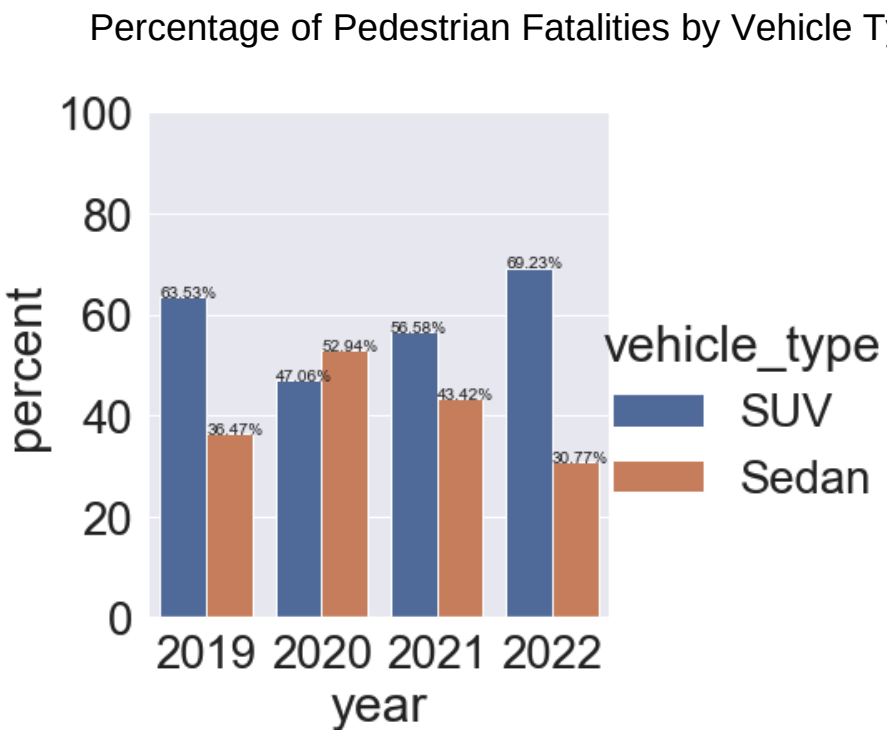


Figure 10: The graph shows that SUVs had a significantly higher rate of pedestrian fatalities for all years with exception of the year 2020.

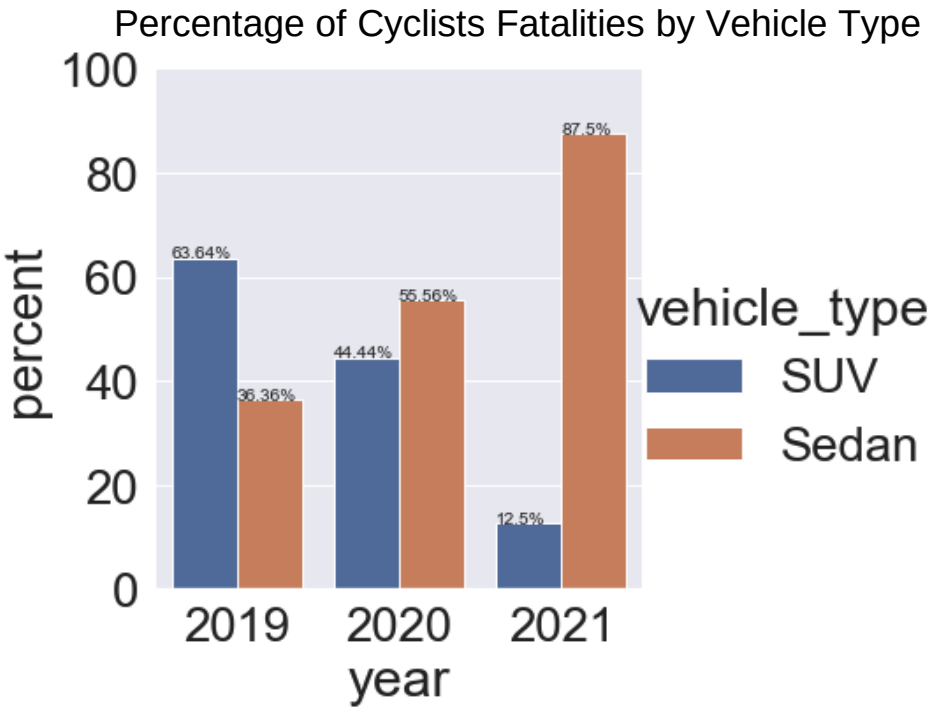


Figure 11: The graph shows that Sedans have a higher rate of cyclists fatalities, with exception of the year 2019 where SUVs accounted for significantly more.

Another intersting find in comparing the accident data between SUVs and Sedans were the observable trends for overall fatalities since 2019.

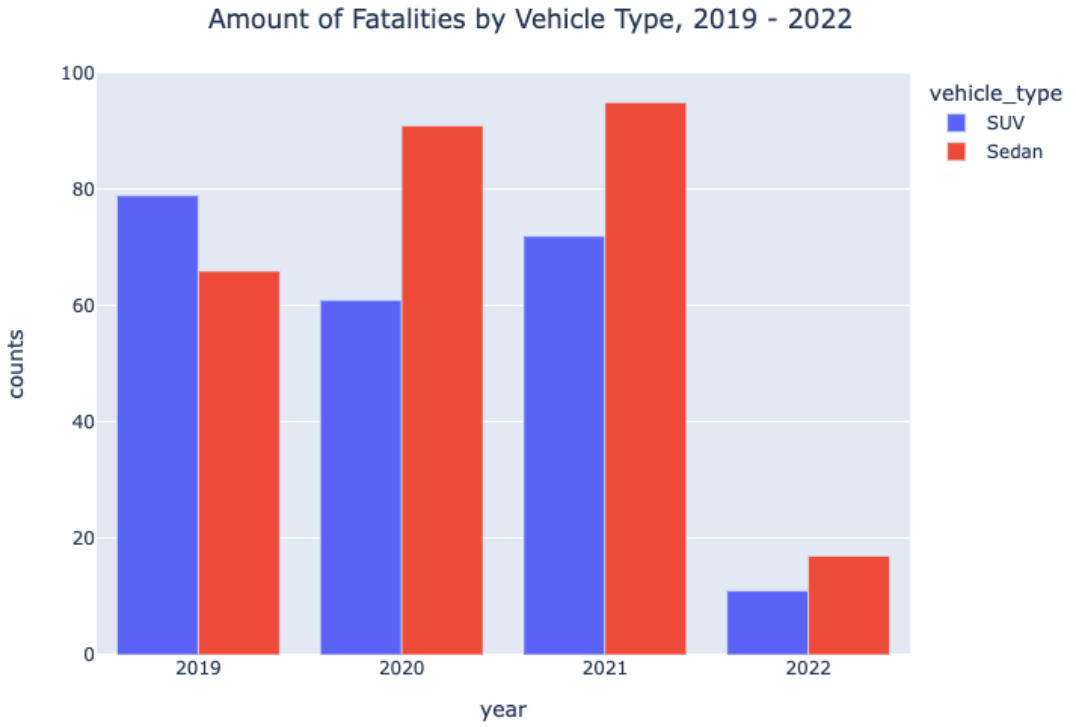
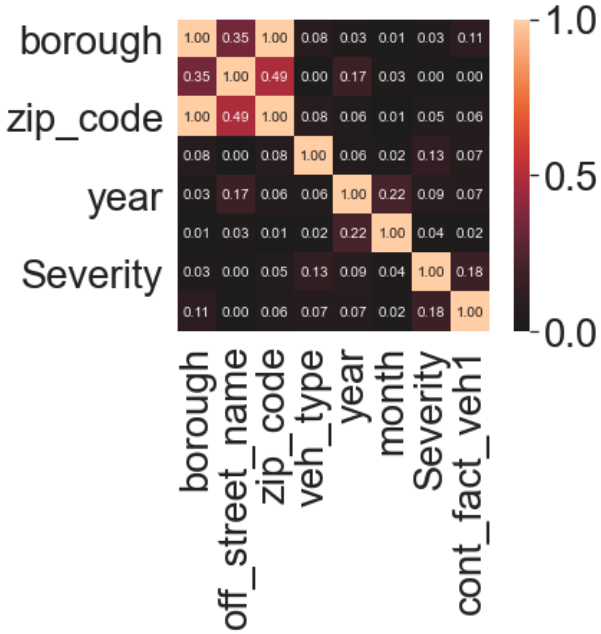
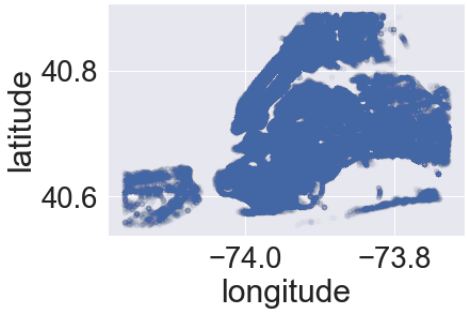


Figure 12: The above graph shows that Sedans have a higher overall fatality for all years, except 2019. This is consistent with the fact that the overall number of accidents, fatal or non-fatal, involve Sedans. However, SUVs having a higher overall rate of fatalities for an entire year was an unexpected find.

Next, the categorical variables were examined for any potential correlations. The results are depicted in the below correlation matrix. No significant correlations were found.



Next, we examined the geo-coordinates to see if there are any patterns in the locations data. The below graph represents the plotted coordinates of the accident data (Geron, 2019).



The below geospatial visualization represents the results of a density-based analysis for neighborhoods in Manhattan and Brooklyn where fatal accident data points were densely populated. The map shows that the Downtown, Chinatown, and Hudson Square are potential areas of interest for fatal accidents.

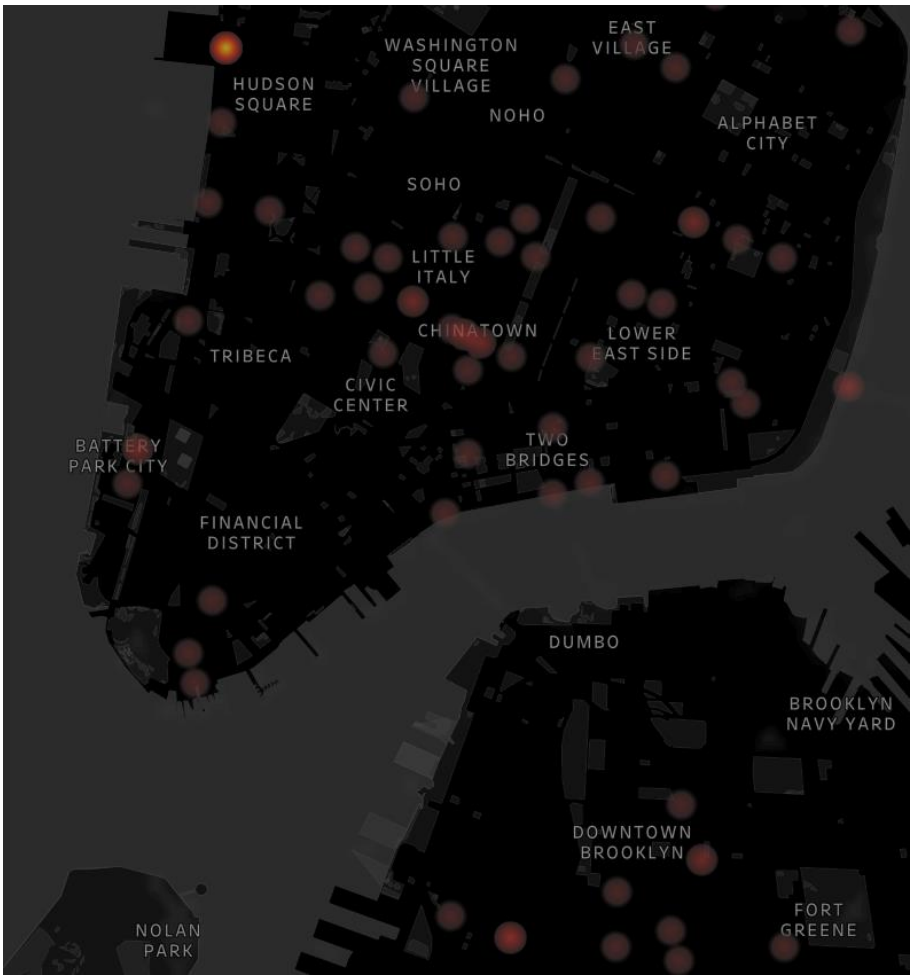


Figure 13: The above geo-spatial visualization is from a density-based analysis in Tableau, where areas of a dense number of fatal accidents occurred.

Machine Learning

Next, we implemented machine learning methods to attempt to predict the classification of an accident in terms of severity. The labelling for classification was constructed as follows:

Accident Severity

No Injuries = “Slight”

One (1) or more Injuries = “Serious”

One (1) or more Fatalities = “Fatal”

The dataset was split between a training and testing dataset in preparation for running a logistic regression analysis. The goal of the analysis is to predict a severity classification based on the dataset dimensions and their corresponding coefficients. The below screenshot depicts the algorithm implementation and results. The accuracy score for this classification algorithm was 0.829, or 82%.

Logistic Regression

```
1
2  # Logistic regression
3  lr = LogisticRegression(random_state=0)
4  lr.fit(X_train,y_train)
5  y_pred=lr.predict(X_test)
6
7  # Get the accuracy score
8  acc=accuracy_score(y_test, y_pred)
9
10 # Append to the accuracy list
11 accuracy_lst.append(acc)
12
13 print("[Logistic regression algorithm] accuracy_score: {:.3f}.".format(acc))
```

```
[Logistic regression algorithm] accuracy_score: 0.829.
```

Next, the training and test data was implemented in a K-Nearest Neighbors Classification algorithm. The accuracy score for this algorithm was 0.822, or 82%.

The scores for each algorithm were very similar, but the logistic regression results were slightly more accurate. However, neither of the results have a particular high accuracy score for classification of severity, and adjustments of parameters may be necessary for better accuracy.

6 Final Findings

In conclusion of this analysis, there are some insights we can take from the accident data. Initial findings highlighted that Brooklyn and Queens have the highest occurrences of accidents in NYC. There are observed trends of higher amounts of accidents in January between 2019 and 2022. Additionally, Fridays are when the highest number of accidents occurred. The accident data shows a drop in overall accidents since COVID-19, but general accident patterns appear to follow pre-pandemic levels. Sedans have a higher proportion of accidents than SUVs.

The most interesting and unexpected the result of the data were the observable trends in accident fatalities. Although the overall number of accidents are a at a fraction of what they were at pre-pandemic levels, there is a clear upward trend in the number of accident-related fatalities in NYC. There may be some latent factors that could be further explored to understand why this phenomenon may be occurring.

A significant finding from the analysis was the difference in means between SUVs and Sedans for number of pedestrian fatalities. While Sedans have a higher overall mean of accidents, injuries, and fatalities—the difference in mean number of pedestrian fatalities involving SUVs was found to be statistically significant. These findings indicate that cities may see more pedestrian deaths as more SUVs are on the road, and that more safety features and ordinances for pedestrians may be appropriate for SUVs in urban areas.

7 Final Notes

This project had several interesting findings, as well as a lot of takeaways in learning how to perform a more complete data analysis project. This dataset was particularly large and posed its own unique set of challenges. Thorough implementation of the data mining process will reveal patterns that are easily missed when you're only focused on summarizing the data. To extract the most value out of a dataset, the analyst must be able to have a balance between a micro-level and macro-level of thinking.

Furthermore, it was very interesting to see real-world phenomena and how patterns may or may not follow the expected turn of events. With more time to develop the analysis, it would be useful to further explore the categorical features in the dataset, such as the contributing factors to an accident. It would be most interesting to see how the observed patterns develop in the following years and whether they are consistent with this paper's findings.

8 Bibliography

1. NYC OpenData. Motor Vehicle Collisions - Crashes. Retrieved March 8, 2022. https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

2. Government Accountability Office. Report to Congressional Requesters. GAO-16-66. (November 2015). Pedestrians and Cyclists.

3. SUVs, other large vehicles often hit pedestrians while turning. Insurance Institute for Highway Safety. 2022. Retrieved March 22, 2022, from https://www.iihs.org/news/detail/suvs-other-large-vehicles-often-hit-pedestrians-while-turning

4. Géron Aurélien. (2019). *Hands-on machine learning with scikit-learn and tensorflow: Concepts, tools, and techniques to build Intelligent Systems*. O'Reilly.

5. Petrou, T. (2017). *Panda's cookbook: Recipes for scientific computing, time series analysis and data visualization using Python*. Packt Publishing.