New York City Taxi Accidents in 2020
(January - August)

Rachel Molent
Student ID: 1354088
Github repo with commit

August 25, 2024

1 Introduction

New York City reached an average of 819 taxi related injuries per month in 2019 [**one**], demonstrating the need for data driven insights to increase the safety of taxi trips in NYC. One proposed method of doing so, is by reducing congestion. In 2017 The Metropolitan Transportation Authority (MTA) introduced congestion tolls, as well as in December of 2020[**two**]. Further congestion tolls were introduced in mid-2024 with the goal of reducing congestion by a further 17%[**three**].

The following investigation compares two different machine learning models to explore the factors contributing to these taxi accidents and why congestion tolls may be effective in improving traffic safety. These insights could be used by the NYC government and the MTA city planning, to reduce the number of taxi accidents.

1.1 Data

The two main datasets used in this report were from the TLC Taxi Trip Record Data. These datasets were released by the NYC Taxi & Limousine Commission[**four**]. The full year of 2020 was downloaded for both the green and the yellow taxis, which were later merged in the pre-processing stage. Many of the variables were unrelated to taxi accidents, so from the combined dataset, only 6 out of 19 features were kept out:

- Pick Off Location Id
- Pick Up Date and Time
- Trip Distance (in miles)
- Drop Off Location Id
- Drop Off Date and Time
- Congestion (True or False)

The combined datasets contained 26383268 instances, however, after initial stages of pre-processing, around 2.6% of the data, leaving 25705901 instances of valid data entries to work with.

An external data was obtained from Kaggle[**five**], containing eight months of vehicle accident records from January of 2020. This contained 4279 instances involving taxis, which where used for further analysis. Most variables were relevant to the investigation topic, however, individual accidents recorded by the type of vehicle and extent of injury were replaced by the combined sums, as well as the vehicle types being combined into the total number of vehicles involved in each crash. This reduced this initial 33 features into 14 features as follows:

- Crash date
- Crash time

- Borough
- Zip code
- Latitude
- Longitude
- Location
- Fatalities
- Injuries
- Contributing factor (vehicle 1)
- Contributing factor (vehicle 2)
- Contributing factor (vehicle 3)
- Contributing factor (vehicle 4)
- Contributing factor (vehicle 5)

Dataframe Shapes		
Dataset	Total Features	Total Instances
NYC Taxi Records (Yellow and Green)	6	25705901
NYC Taxi Accidents	14	4279

2 Preprocessing

2.1 Raw Layer

- Filtered the accidents dataset to ensure the accidents being analysed involved taxis
- Created consistent schema by case folding and converting data to suitable datatypes
- Removed Null feature 'ehail fee' from the green taxi accidents dataset
- Removed redundant features
- Renamed columns
- Merged taxi trip datasets
- Validity checking. Filtered the datasets so that distance > 0. Removed unknown taxi zones
- Summed the number of accidents, injuries, fatalities and number of vehicles involved in each accident.

2.2 Curated Layer

- Exploratory Data Analysis including scatterplots
- Used Boxplots to determine which transforms to apply to the data
- Removed outliers
- Standardised Data
- Descriptive Statistics
- Creating new variables and aggregated dataframes
- Grouped data into periods of weeks
- Imputed missing boroughs due to small sample size of taxi accidents (include stats length of df)

Curated Layer Data Summary: Calculations performed on the original taxi accidents dataset was used to create new features such as speed and duration. For analytic and graphical purposes, the data was

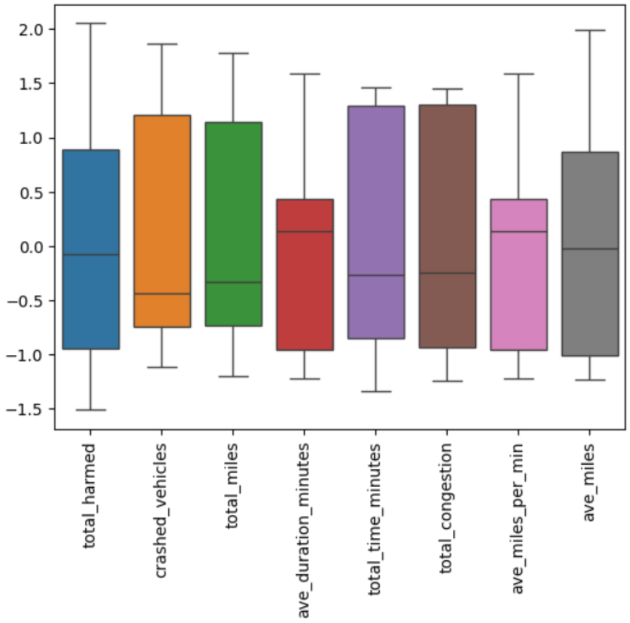


Figure 1: Distribution of Standardised Predictors

grouped into blocks of weeks of which there were 34 weeks in total, with 9 features in the taxi accidents dataframe.

Dataframe Features post-processing	
NYC Taxi Records (Yellow and Green)	NYC Taxi Accidents
• week • congestion surcharge • drop off location ID • pick up location ID • taxi colour • Drop off date and time • Pick up date and time • trip distance (in miles) • duration (in min) • speed (in miles per min) • pick up date • pick up hour	• week • total harmed • crashed vehicles • total miles • average duration (in minutes) • total duration (in minutes) • total congestion • average miles (per min) • total average miles

Dataframe Shapes		
Dataset	Total Features	Total Instances
NYC Taxi Records (Yellow and Green)	12	25705901
NYC Taxi Accidents	9	4279

3 Preliminary Analysis

3.1 Exploratory Data Analysis

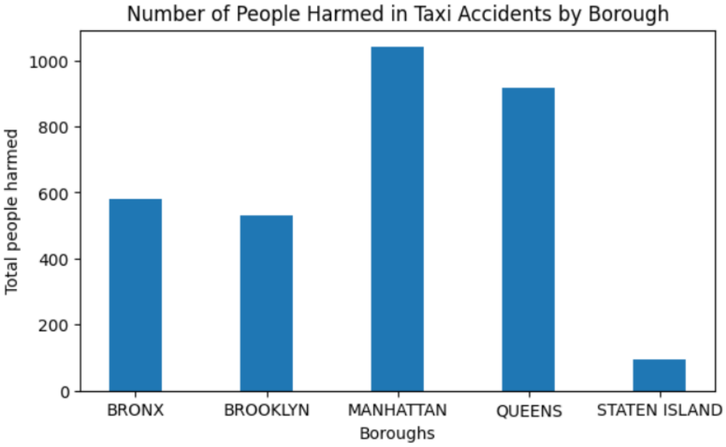


Figure 2: Significantly more accidents occurring in Manhattan and Queens compared to other boroughs. Low accident rate in Staten Island.

From the exploratory data analysis shown in Figure 2, it was evident that there is a disparity between the number of accidents that occur in each borough. This lead to further geospatial analysis through the use of choropleth maps.

3.2 Analysis and Geospatial Visualisation

Due to the excessive number of instances in the taxi dataset, a sample of 5% of the data was used for plotting, corresponding to roughly 1285296 instances.

In Figure 3, Manhattan differed greatly from the remaining suburbs, which remained entirely within the 0-1,015 $USD bracket. Manhattan was the only borough with areas reaching higher total congestion fees. Several areas within Manhattan reached medium-high total fee brackets, whilst 2 areas fell within the highest bracket of 5,077-6,092 $USD in total fees. Since Manhattan had the highest number of people harmed in taxi accidents in January to August of 2020, as seen in Figure 2, this highlights a likely relationship between taxi accidents and congestion.

4 Modelling

From Table 1, Elastic Net regression, Linear regression and Support Vector Regression under-performed. This is unsurprising since these models assume linearity, which was not true of all the predictors.

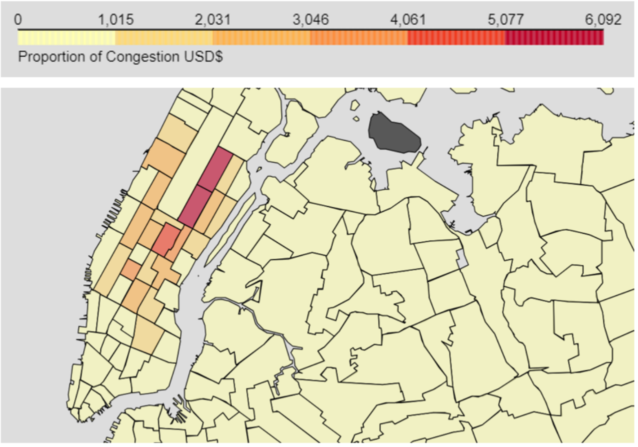


Figure 3: Total Congestion Fees Per Borough (Jan-Aug, 2020)

Table 1: Evaluation Metrics

Model	MSE	$R\hat{2}$
Elastic Net Regression	0.132	0.882
Linear Regression	0.149	0.881
Random Forest Regression	0.014	0.977
Support Vector Regression	0.068	0.885

However, most predictors roughly had a linear relationship with the response after transformation, so the model is still able to perform adequately, but even after transformation and standardisation, some predictors had a skewed distribution, such as 'average duration in minutes' in figure 1. Additionally, these models are not were not complex enough to capture all the relationships within the data and were likely under fitting, as seen in Figure 4.

The more complex model, Random Forest Regression, performed very well with an $R\hat{2}$ score of 0.977. This model is well suited to the data given that there is a very large number of instances to train on, however, achieving an $R\hat{2}$ score, so close to 1 is likely a sign of overfitting.

Linear Regression model was chosen to test the strength of correlation between predictors and the response variable and to see whether this relationship could be well approximated by a linear model. However, the model was inadequate, so Random Forest Regression was used as a more complex alternative that does not require the linearity assumption.

5 Discussion

The goal of this paper was to use machine learning models to determine what factors were contributing to taxi accidents, with a particular emphasis on whether congestion tolls are able to improve traffic safety. Since the introduction of congestion tolls from 2017 onwards, the number of injuries caused by taxi accidents has dropped from 819 a month in 2019[**six**] to 435 a month according to my analysis. Although it is important to consider the role that COVID 19 and lockdowns played in this reduction, as traffic may have reduced over this time. Further to this, total congestion showed strong correlation

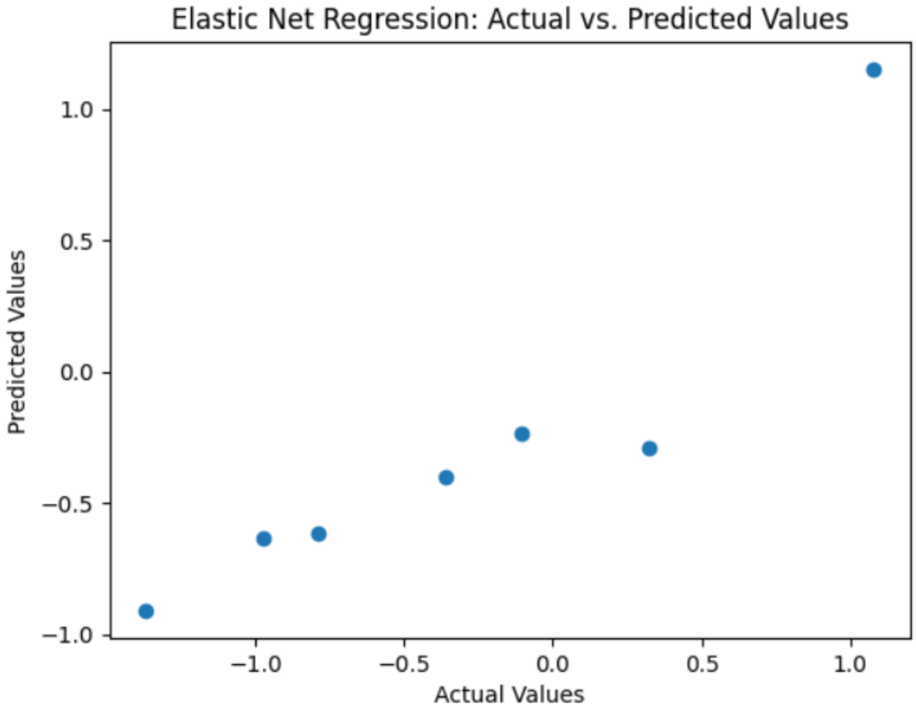


Figure 4: Evaluation of the Elastic Net Performance

to taxi accidents through 5s, Linear Regression and via geospatial analysis, particularly for Manhattan. However, Figure 2 demonstrated Queens also has a very high rate of harm caused by taxi accidents. Further analysis is needed into the casual effects for this borough, however, since Manhattan's injury rates caused by taxis per month has decreased since the introduction of congestion tolls, it is suggested that Queens also consider adopting a similar method as a short term strategy to improve public safety.

6 Recommendations

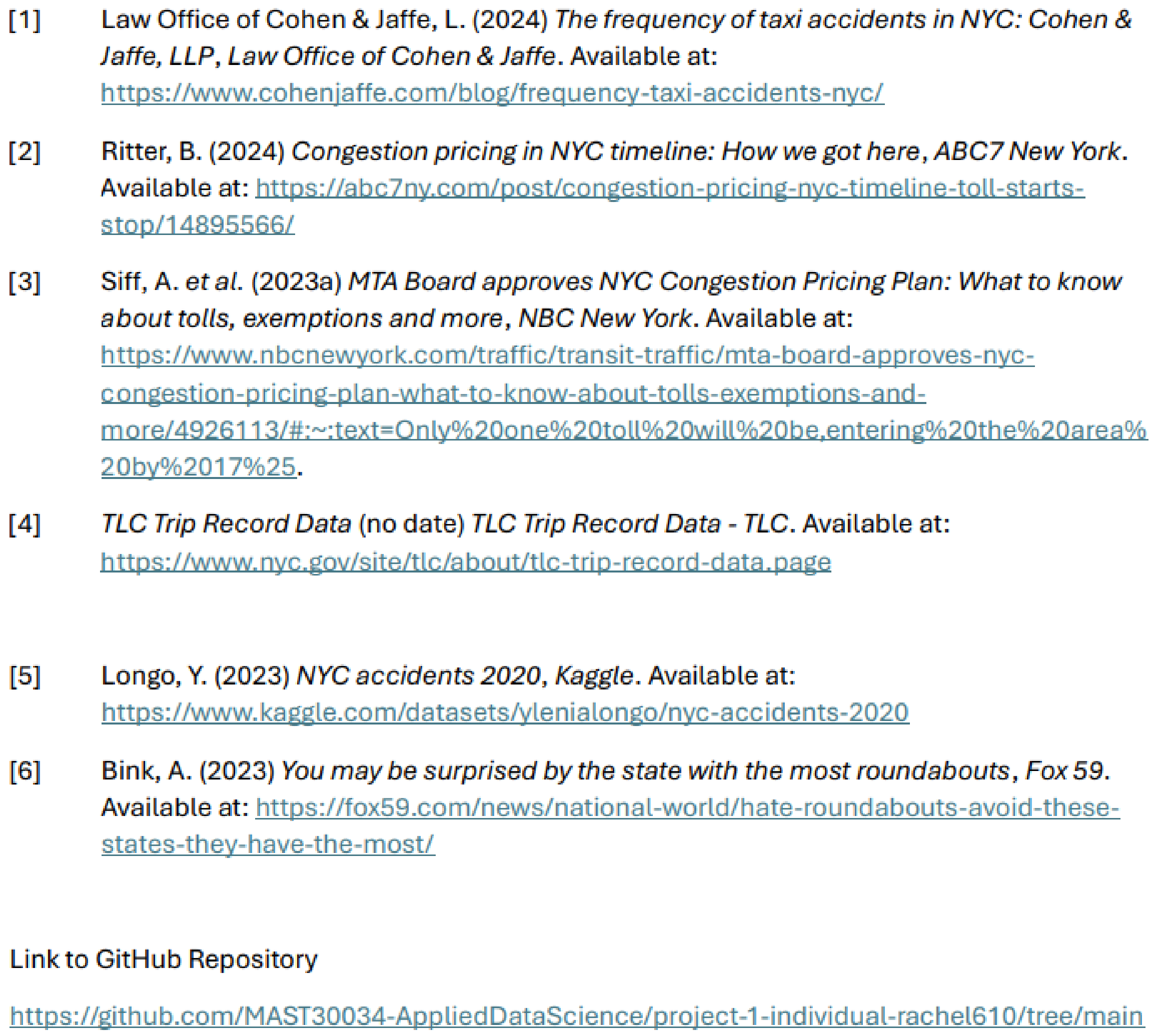
Improvements into the safety of taxi users in NYC would most effectively be focused on Manhattan, due to their inflated rate of taxi accident related harm and it's disproportionate rate of congestion. However, Queens is also a high risk area for taxi accidents that needs further research and could perhaps benefit from the congestion toll strategy implemented in Manhattan. Both boroughs should particularly focus on improving traffic flow which may be regulated through timed traffic lights or the implementation of a larger number of roundabouts, since NYC has a low rate of roundabouts per population and area[**six**]. Additionally, encouraging more availability and use of public transport and bike lanes to reduce the number of vehicles on the road could have a great impact.

While this paper supports the conclusion that congestion tolls are effective in reducing congestion and thereby improve public safety, it is suggested that they only be used as a temporary solution. It is recommended that the funds gained from the tolls be used to create services that will decrease the city's dependence on cars. This may include, improving the accessibility of public transport, introducing more roundabouts to improve the flow of traffic and installing better cycling facilities. These implementations are likely to reduce the overall number of vehicles on the road, and thereby, alleviate congestion burdens and consequently, vehicular accidents.

7 Conclusion

This report delved into the contributing factors pertaining to taxi accidents in NYC over the months January to August of 2020. The major finding of this report is the strong correlation between congestion and taxi accidents, highlighting a crucial factor to address in improving traffic safety. While congestion tolls are likely to bring down the number of people harmed in taxi accidents, it is not recommended as a long term solution due to reasons of impracticality and public nuisance.

Given the role of COVID 19 and the various fluctuations in congestion tolls over the years, it would prudent to explore a wider time frame of data, in order to minimise the effect of extraneous factors influencing the data. It is recommended that further evaluation metrics be studied such as F1-scores and confusion matrices, to better understand why some of the machine learning models do not perform well in certain ranges, and better improve their performance.



[1] Law Office of Cohen & Jaffe, L. (2024) *The frequency of taxi accidents in NYC: Cohen & Jaffe, LLP*, *Law Office of Cohen & Jaffe*. Available at: https://www.cohenjaffe.com/blog/frequency-taxi-accidents-nyc/

[2] Ritter, B. (2024) *Congestion pricing in NYC timeline: How we got here*, *ABC7 New York*. Available at: https://abc7ny.com/post/congestion-pricing-nyc-timeline-toll-starts-stop/14895566/

[3] Siff, A. *et al.* (2023a) *MTA Board approves NYC Congestion Pricing Plan: What to know about tolls, exemptions and more*, *NBC New York*. Available at: https://www.nbcnewyork.com/traffic/transit-traffic/mta-board-approves-nyc-congestion-pricing-plan-what-to-know-about-tolls-exemptions-and-more/4926113/#:~:text=Only%20one%20toll%20will%20be,entering%20the%20area%20by%2017%25.

[4] *TLC Trip Record Data* (no date) *TLC Trip Record Data - TLC*. Available at: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[5] Longo, Y. (2023) *NYC accidents 2020*, *Kaggle*. Available at: https://www.kaggle.com/datasets/ylenialongo/nyc-accidents-2020

[6] Bink, A. (2023) *You may be surprised by the state with the most roundabouts*, *Fox 59*. Available at: https://fox59.com/news/national-world/hate-roundabouts-avoid-these-states-they-have-the-most/

Link to GitHub Repository

https://github.com/MAST30034-AppliedDataScience/project-1-individual-rachel610/tree/main

Github Repo: https://github.com/MAST30034-AppliedDataScience/project-1-individual-rachel610