MAST30034 2021 Sem 2 Project 1

Predicting the Amount of Taxi Pickups in NYC Zones (2019)

Student Name: Seth Ng Jun-Jie
Student ID: 1067992

August 15, 2021

1 Introduction

This report aims to depict the relationship between the number of taxi pickups and statistics of both crime and taxi trips happening within zones across the boroughs of New York City; the Bronx, Brooklyn, Manhattan, Queens, and Staten Island, and a "sixth borough"; the EWR Airport. This report attempts to predict the amount of taxi pickups of each zone from October to December in 2019 given the corresponding zone's crime and taxi statistics from 2019's first nine months, specifically studying yellow and green taxi trips. The crime statistics considered regard arrests, shootings, and complaints in each zone. This report is directed mainly towards taxi businesses, taxi drivers, police, and even data analysts interested in New York taxi data.

Datasets for both yellow and green taxis are used as using both kinds would lessen the imbalance for if just yellow taxi data were used, particularly since yellow taxis mainly drive around Manhattan whereas green taxis populate areas that lack yellow taxis *(Barron)*. **2019 data** was used since New York's crime index has immensely decreased since 1990 and steadily decreasing for more than 20 years *(Disaster Center)* so it would be best to use the latest data, but not 2020 as it may result in interaction interference from COVID-19. The **other taxi statistics, and the number of arrests, shootings, and complaints** are plausible attributes to affect the amount of pickups in a zone. The **gender, race, and age attributes selected** are due to gender, racial, and age stereotypes that affect people's actions through fear and judgement. **Taxi businesses and taxi drivers** are the target audience as they would be able to predict which zones would have more demand for them than others. The amount of crime in a zone in New York could strike fear in people and prevent them from going out, and therefore, decrease the amount of taxi pickups within that zone in the coming month.

Assumptions: This report assumes that there are no unknown confounding variables (interaction terms) outside the attributes used, that the NYPD datasets include all crime incidents that have been publicly announced on the news, and that there are no sudden incidents that decrease demand for taxis, such as COVID-19.

Dataset Shape: 27 datasets (12 months of yellow taxis, green taxis, and 2019 data for arrests, shootings, and complaints) were used in this report, roughly 7.97GB in size. None of these datasets have a Gaussian distribution in terms of the number of pickups, arrests, shootings, and complaints as they cannot be negative and all have a mean that is equal or close to 0. *(Figure 1)* shows the distributions for a few of the features from the first 9 months of 2019, with the last 3 months of 2019 following similar shapes.

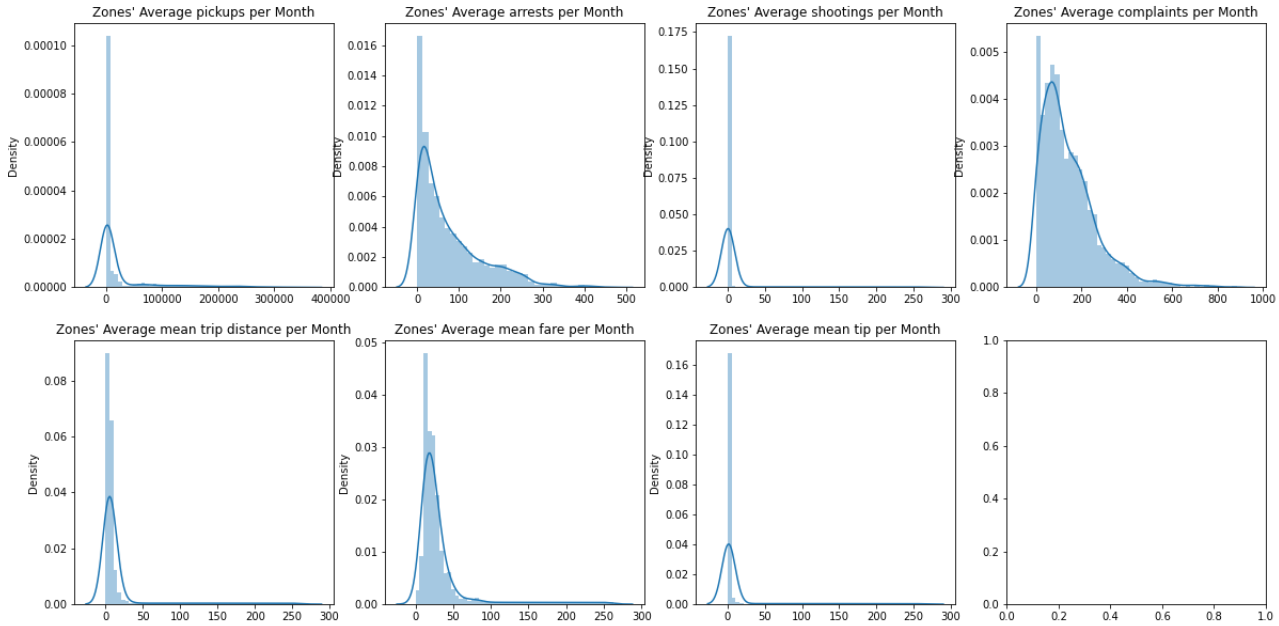

Figure 1: Distributions of zones' average pickups, arrests, shootings, complaints, trip distance, fare, and tip per month (Jan-Sep 2019).

2 Preprocessing

2.1 NYC TLC Yellow and Green Datasets (2019)

Each instance in the yellow and green taxi datasets has 18 and 20 features respectively, of which only 5 were retained from both datasets to use for visualisation and modeling; trip distance, fare amount, tip amount, payment type, and pickup location. After filtering the unwanted attributes, each month's respective yellow and green CSV files were combined. Trips with distances of zero were left out from the mean trip distance, however, not from the other continuous attributes as trips with no payments may prove to show a trend. Pickup locations and payment types were the only two features with potential 'unknown' values. Data with unknown payment types were included as part of the features as there could be some correlation to it, however, data with unknown pickup locations were excluded as it cannot be useful without a dedicated zone.

2.2 Arrests, Shootings, and Complaints (2019)

The arrests and shootings datasets each contain 19 features whereas the complaints datasets have 35. Out of the 73 features 20 were retained, which collectively contained the number of arrests, shootings, and complaints, the number of perpetrators and victims (if applicable) in each age, gender, and ethnicity group, if a murder occurred, and each incident's location. Unknown age, gender, and ethnicity groups were included as part of the features as an unidentifiable perpetrator may have an effect. Crime data with unknown locations were left out for the same reasons as the taxi ones. Affirmation if a murder occurred and the location of the incidents were present for all instances.

2.3 Outlier Analysis

If a zone does not have information on the number of taxi pickups nor on any crime throughout the required months then that zone is considered redundant. Throughout the months, zones 103 and 104 had zero pickups and zone 1 had zero crime.

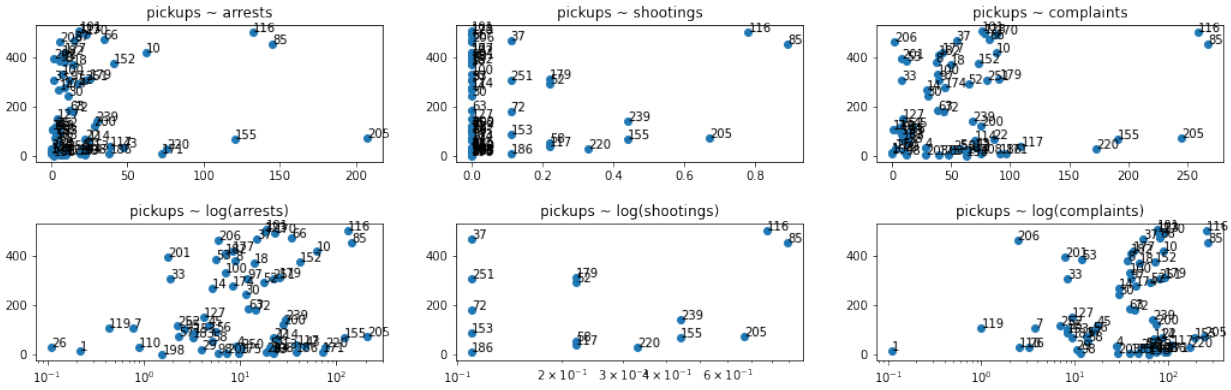

Figure 2: The average number of pickups against the average number of each crime and their log transformations in the lower quartile range (Jan-Sep 2019).

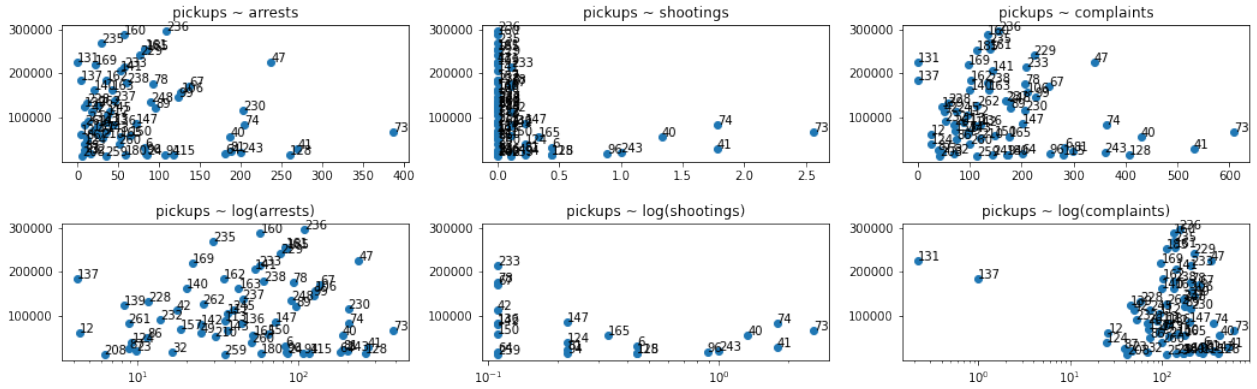

Figure 3: The average number of pickups against the average number of each crime and their log transformations in the upper quartile range (Jan-Sep 2019).

Further investigation was carried out through scatter plots for the lower and upper quartile ranges for the first 9 months as shown above. It can be seen that the log scaled plots (lower three plots in both figures *[2][3]*), particularly with arrests and complaints, seem to have more of a trend than the original plots with the average number of pickups. From *(Figure 2)*, zone 109 can be considered an outlier, as well as zones 131 and 137 from *(Figure 3)*. Therefore, zones 1, 103, 104, 109, 131, and 137 will be excluded from the training and test sets.

2.4 Training Sets and Test Sets

Of the 20 crime features and 5 taxi features mentioned above, those that are discrete and have n¿=2 values are stored as n features, each storing its category's count, e.g. features "credit", "cash", "no charge", etc. were made from the trip payment type feature. This process transformed 25 features into 95 features. These 95 features of each zone, 257 excluding the outlier zones, for each month were then compiled and made into their own CSV file. These would be used as the training (2313 instances) and test sets (771 instances) for the statistical model later on.

The number of pickups of each non-outlier zone for each month were sorted into 8 bins (of intervals [0, 100, 450, 850, 1550, 3250, 14000, 90000, 350000]) of relatively equal amounts. The end of each CSV file then included the respective pickup bin number according to each month and zone, which the 95 features would need to predict in the modeling process.

3 Analysis

3.1 Preliminary Analysis

Geospatial visualisations were made for the first 9 months of 2019 for the amount of pickups, arrests, shootings, and complaints. From *(Figure 4)*, it is observed that even though green taxi datasets were included to reduce the imbalance of pickups between Manhattan and the other boroughs it still is not enough to achieve an equilibrium, with Brooklyn having the most range on pickups throughout its zones. Staten Island appears to have little presence in complaints *[7]* and almost none in taxi pickups *[4]* and shootings *[6]* compared to the other boroughs. Majority of the zones were found to have little to no shooting incidents *(Figure 6)*. *(Figures 8 and 9)* has Manhattan as their centre, as if the further the zone from Manhattan the higher the average fare and trip distance respectively.

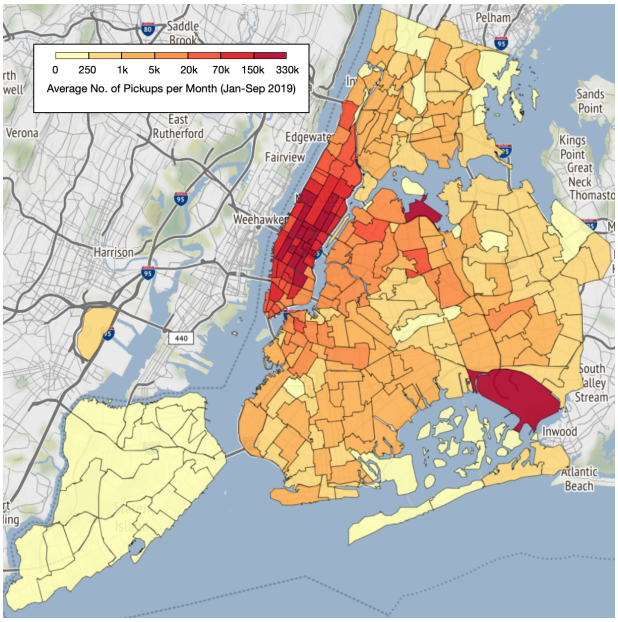

Figure 4: Average number of pickups a month (Jan-Sep 2019).

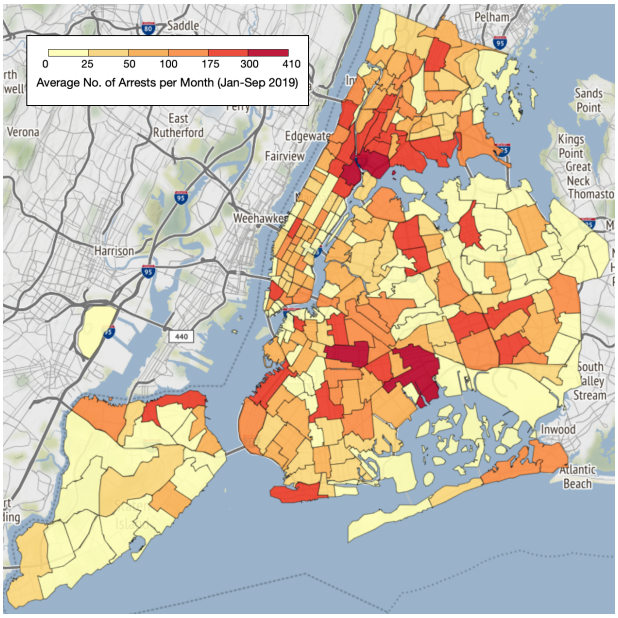

Figure 5: Average number of arrests a month (Jan-Sep 2019).

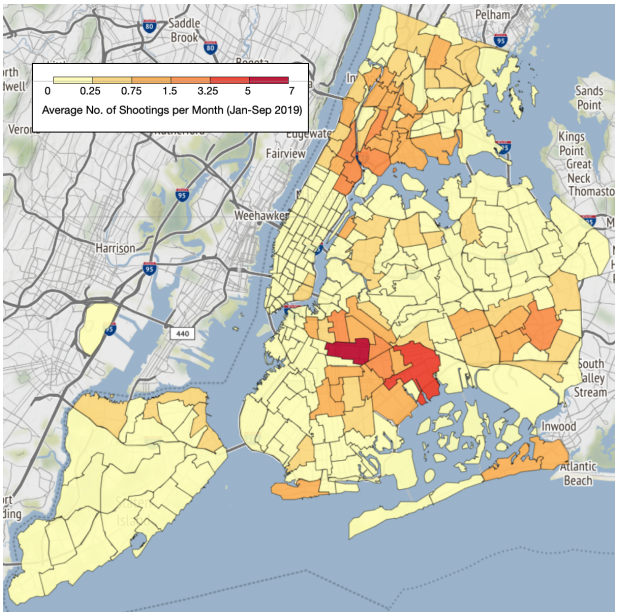

Figure 6: Average number of shootings a month (Jan-Sep 2019).

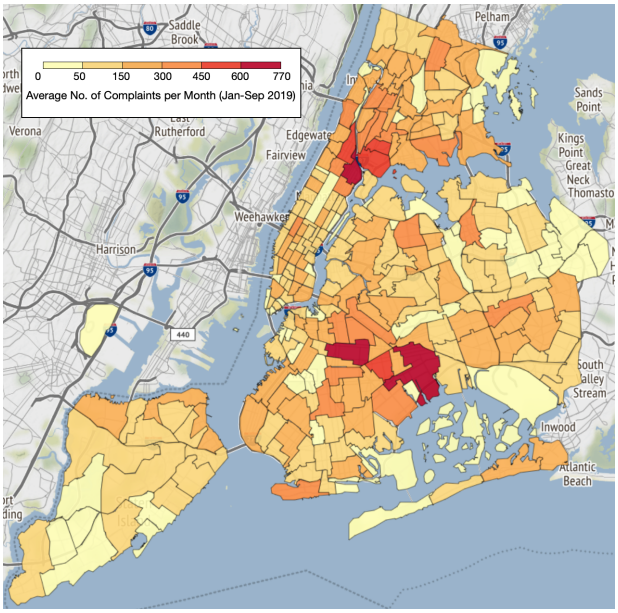

Figure 7: Average number of complaints a month (Jan-Sep 2019).

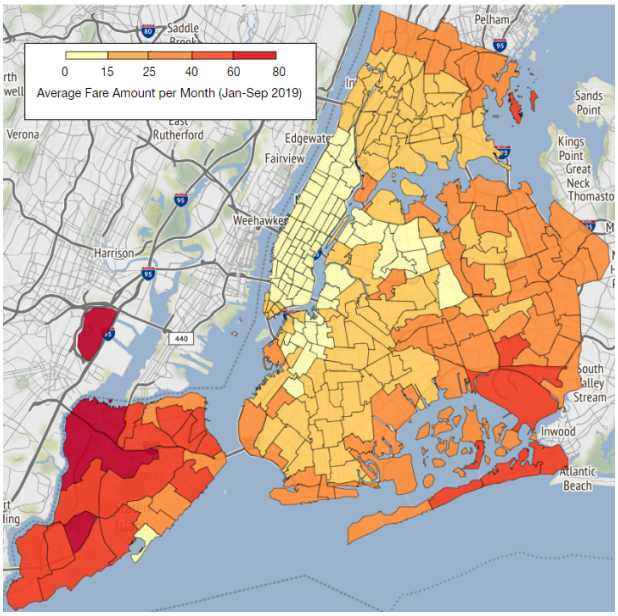

Figure 8: Average fare amount a month (Jan-Sep 2019).

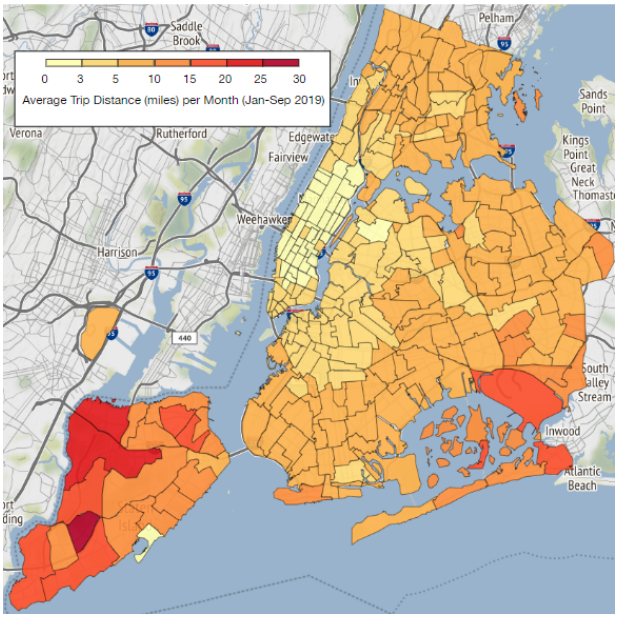

Figure 9: Average trip distance (miles) a month (Jan-Sep 2019).

3.2 Attribute Analysis

It is seen that a high number of shootings *[6]* are strongly correlated with both a high number of arrests *[5]* and complaints *[7]*, judging by the zones with 3.25 shootings per month. It is also shown that a high volume of pickups *[4]* tends to be correlated with medium to low levels of the three crimes.*(Figures 8 and 9)* as discussed above show that fare amount and trip distance correlate with one another. Their densities for Staten Island and Manhattan zones are inverses to that of *(Figure 4)*.

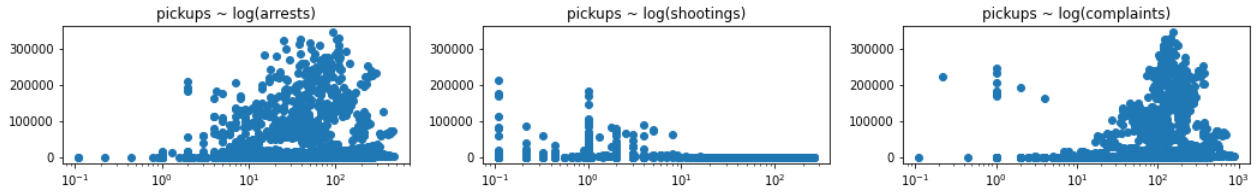

Figure 10: Average number of pickups against the log transformations of the average of each crime per month (Jan-Sep 2019).

(Figure 10) shows that for arrests and complaints the plots follow a non-linear trend, almost like an exponential curve. With both forming similar curves, the number of arrests and complaints may be more correlated than *(Figures 5 and 7)* portray them to be. For shootings on the other hand, it seems that both its zone map *[6]* and its log plot *[10]* do not seem to correlate much with any of the other non-crime data and since it already correlates well with arrests and complaints it may not be entirely useful due to its lack of density.

4 Statistical Modelling

4.1 Model

The model chosen for this report was a logistic regression model *(Towards Data Science)*. This model was chosen due to the curve-like relationship in the plots with log scales on the x-axis in *(Figures 2 and 3)*. Sklearn's "LogisticRegression" function *(Scikit Learn A)* was used to implement the model and it was trained using cross-validation with k = 10 folds on the the 95 features for the 257 non-outlier zones in the training set. In order to establish optimal parameter values, the function "GridSerach" was used *(Scikit Learn B)*. It takes in a classifier and a dictionary of parameters and their possible values. The logistic regression model was passed into the function along with a parameter dictionary for penalty terms, inverse regularisation strengths, and solver methods. After excessive computations, the optimal model achieved an accuracy of 65.76%.

4.2 Refinement

Feature selection was conducted using the "SelectFromModel" function *(Scikit Learn C)*, which filters features of an already fitted model that it deems important according to their correlation coefficients within the training set. The function resulted in only 20 of the 95 features being relevant to the modeling process. These were the number of arrests and complaints, the gender, age group, and race of arrests perpetrators, the number of people arrested for felonies and misdemeanors, average taxi trip distance and fare amount, and taxi payment type. Using just the 20 features, the same logistic regression model from before achieved a similar accuracy of 65.62%. While it did not improve the overall accuracy, it managed to maintain the accuracy while harshly reducing the redundancy of the training set.

Furthermore, through inspection on the confusion matrix of the latest model and feature set, it was found that the majority of incorrect classifications occurred in bins 3, 4, and 5. Since zones from the other bins were predicted quite accurately, the bin intervals were rearranged ([0, 50, 300, 775, 1750, 5000, 17000, 100000, 350000]) so that bins 3, 4, and 5 were allocated more zones/data from the other bins. By doing so, the refined model resulted in an accuracy of 70.17%, with classification improvements in bins 3 and 5.

4.3 Discussion

The feature selection process proved the theory that shooting statistics are irrelevant in the model and can be sufficiently covered by arrests and complaints. It was not predicted that arrests statistics would be far more superior than complaints statistics, however, logically speaking it does sound realistic. Average trip distances and fare amounts are indeed beneficial in the model due to their negative correlation with the number of pickups. While the intervals of the original bins were of almost equal frequency they certainly were not of equal width. The lower and higher bins probably had more obvious patterns, but zones around the median likely had too similar feature values, making them harder to classify.

5 Recommendations

With the drastic decrease in relevant features, taxi companies and police officers now have a more specified list of information to obtain to use the model. Police making arrests may not want to put down any unknown values for the perpetrator and could also note down any other extra details that may provide additional value to the model, seeing as that statistics on arrests are correlated to number of pickups. Taxi businesses as well should either ensure their drivers always record distance, fare amount, and payment type for their trips or invest in an automatic electronic records system to do so.

6 Conclusion

This report implemented nothing but crime and supplementary taxi data to predict the average number of pickups in zones across New York City. Having a prototype model at 70% accuracy, this concept has the potential to grow considerably and be used as a powerful tool for taxi businesses and drivers on a monthly basis, possibly even on a daily basis. More variable datasets, such as weather, events, and population, can be included to make a more complex and efficient model with high levels of capabilities that could adapt towards not just other cities but other countries as well.

References

[1] Logistic Regression - Detailed Overview - Towards Data Science. Accessed August 14, 2021
https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

[2] Disaster Center. Accessed August 13, 2021
https://www.disastercenter.com/crime/nycrime.htm

[3] J. Barron - The New York Times. Accessed August 13, 2021
https://www.nytimes.com/2018/09/03/nyregion/green-cabs-yellow-uber.html

[4] NYPD Arrests Data (Historic) - NYC Open Data. Accessed August 13, 2021
https://data.cityofnewyork.us/Public-Safety/NYPD-Arrests-Data-Historic-/8h9b-rp9u/data

[5] NYPD Shooting Data (Historic) - NYC Open Data. Accessed August 13, 2021
https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8/data

[6] NYPD Complaints Data (Historic) - NYC Open Data. Accessed August 13, 2021
https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Historic/qgea-i56i/data

[7] sklearn.linear_model.LogisticRegression - Scikit Learn A. Accessed August 13 2021
https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[8] sklearn.model_selection.GridSearchCV - Scikit Learn B. Accessed August 13 2021
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[9] sklearn.feature_selection.SelectFromModel - Scikit Learn C. Accessed August 13 2021
https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFromModel.html

[10] "Taxi Fare." Taxi Fare - TLC. Accessed August 5, 2021.
https://www1.nyc.gov/site/tlc/passengers/taxi-fare.page.