New York City Taxi Cab Industry

Abstract

The taxi cab industry of New York City has seen a significant negative financial impact since 2014. The project seeks to explore the “new normal” for New York City taxi drivers and opportunities to increase revenue and strategies for maintaining current revenue streams. The project seeks to answer these questions through Exploratory Data Analysis and through Predictive Modeling.

Intro/Background of the Problem

The taxi cab industry of New York City has seen a significant negative financial impact since 2014. An important part of understanding the problem is that drivers purchase a medallion to be an authorized New York City yellow cab driver. At its peak, the medallion was valued at $1 million. However, there are now medallions sitting in inventory and the value has dropped significantly. Within the industry, there are two clear reasons why this shift has taken place and observers appear to take sides on which is to blame. The first reason, and most obvious, is the increase of ride-sharing apps (such as Uber and Lyft) have changed the way riders use the service. The second reason discussed is that the financiers let the economic bubble blow up while there were signs that the loans provided to drivers were not going to be sustainable. As a result, in January 2020, the NYC City Council has created a task force to investigate if there were any bad practices with the lending and possible solutions for the impacted drivers.

Data Source

The New York City Taxi and Limousine Commission provides monthly reports on Taxi Cab rides. The monthly reports range from January 2009 to December 2019. The variables include pick-up and drop-off date and time, number of passengers, trip distance, location IDs for pick-up and drop-off, and total fare (which also includes a breakdown of tip, surcharges, taxes, and payment method).

The data and data dictionary can be accessed at the following link:
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

Data Preparation

In order to create a master dataset, each month's CSV file needed to be downloaded. They were then imported and concatenated with Pandas in Python. Due to the size of the files, the first 10,000 rows of each month were imported. The restriction with this is being able to see standard revenue and trip numbers over time. However, the sources show that this has already been researched, tracked, and validated so the investment in data storage did not seem optimal considering the scope of this project.

There were variables that changed in measurement and existence over the timespan of 2014-2019 and these were removed for the master dataset. These variables included Vendor ID, Rate Code ID, Store and Forward Flag, Pickup Location ID, Dropoff Location ID, Pickup and Dropoff Longitude and Latitude, Payment Type, and Extra (as in extra fees). The Pickup Location ID and Dropoff Location ID were previously listed as longitude and latitude which is why they were removed. However, when researching trip information for the latest year, these variables are reintroduced. The remaining columns were then renamed to keep consistency throughout the datasets before they were concatenated. These columns are listed below:

1) Pickup_datetime
2) Dropoff_datetime
3) Passenger_count
4) Trip_Distance
5) Fare_Amount
6) Surcharge
7) MTA_tax
8) Tip_amount

9) Tolls_amount
10) Total_amount

The pickup_datetime and dropoff_datetime were formatted such as the following: 2014-01-09 20:45:25. For a more detailed analysis of time, the datetime library in Python was used to create new columns for time, month, date, and year. These new variables were then used to calculate the trip duration in seconds and these results were stored in a new column.

The next stage was to find outliers that would indicate errors or not be usable in predictive modeling or analytics. First, any rows with NA values were removed. Second, there were trips listed as taking place in the years 1970 and 2008 and these rows were removed. There were also trips with extremely high distances (including one for 284,000 miles) which were removed. There were also values of zero where this would not make sense for the dataset. Any total amount or trip distance had to be greater than zero. The total amount also had to be under $400. The total amount may seem high but there were a few rows with high dollar amounts for travel that may be important for the context of this analysis. Figure 1 shows the final result after this filter and while there are still some oddities in the data at this stage, it generally shows a trend of an increase in total amount as the distance increases.

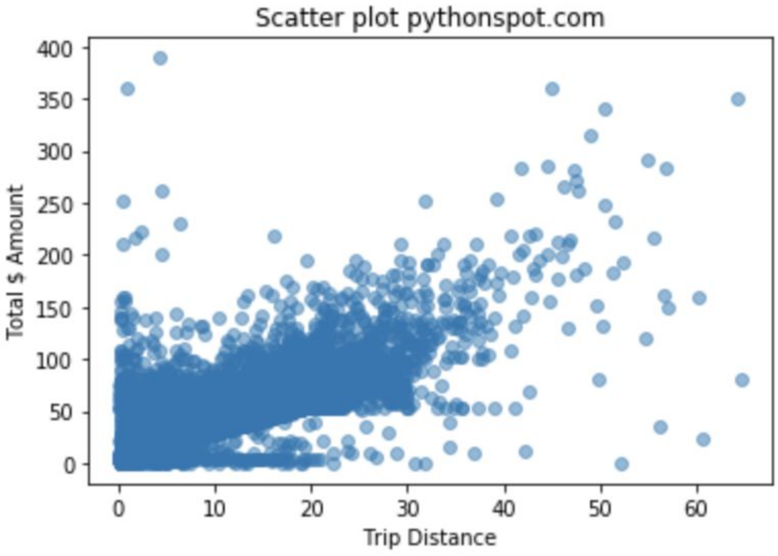


Figure 1

Another oddity in the data would be trips with high tip values but low fare amounts. In some cases, the trip record had $0 for the fare but a tip greater than $50. These rows were deemed to have errors so a new column was created called Tip Percentage which calculated the tip amount over the fare amount. Any row with a tip percentage higher than 75% was removed. This was also reflected in when comparing tip amount to trip distance as seen in figure 2. Figure 3 shows the correlation between fare amount and tip amount after these filters and the representation appears much cleaner.

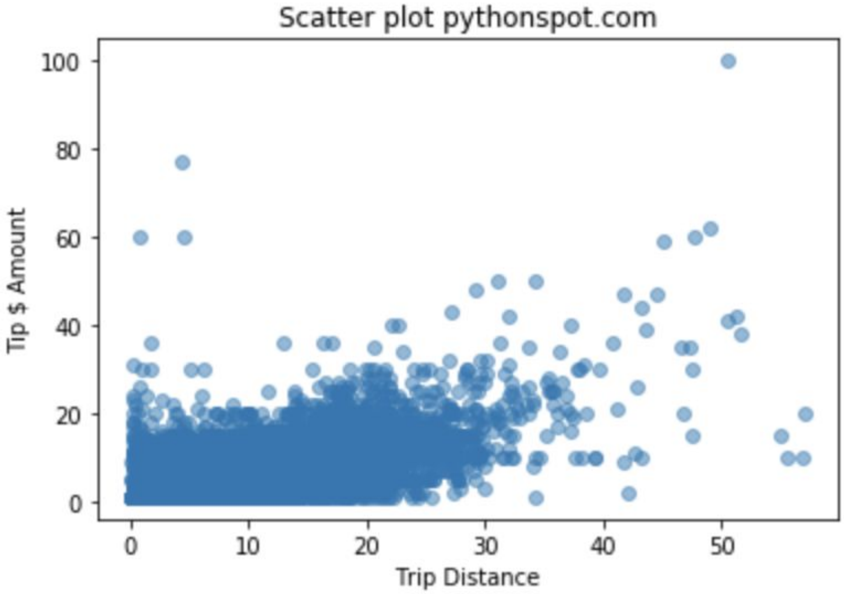


Figure 2

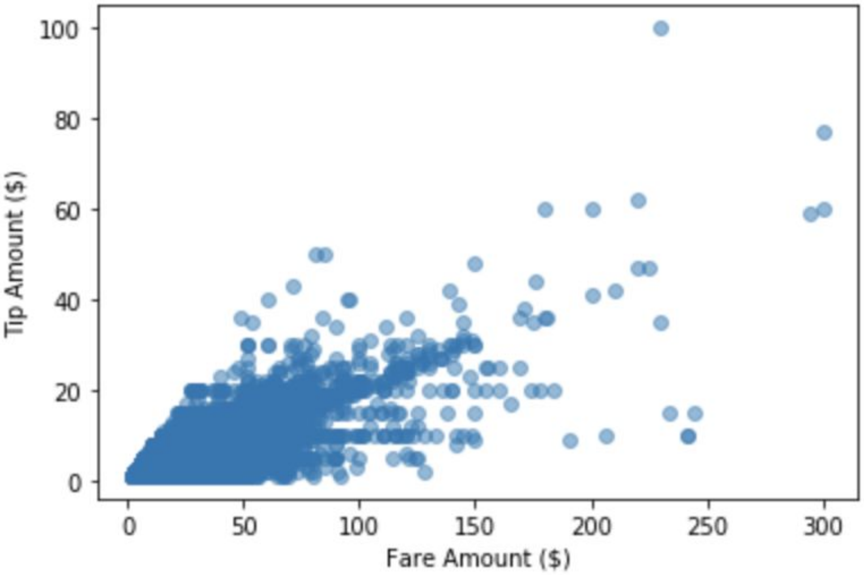


Figure 3

Data Exploration

I started by creating Linear Regression Models in R to see if there were any correlations that could be looked into further. The Linear Regression Models took all variables that were prepared against three different variables. The defining variables for the Linear Regression Model are listed below.

1) Total Amount
2) Pickup Month
3) Pickup Year

The first model based on Total Amount saw the most positive p-values compared to the other two models. However, any positive values were very small values and there did not appear to be any strong correlations that could be explored. However, this still provides some possible insight. It likely means that there have been few circumstances that have changed fares, rates, taxes throughout the years that are being explored. If there were significant increases or decreases, we would likely see some correlations here.

Since there were no significant patterns to the variables on passengers, fares, and time, the focus shifted to pickup location. For solving the business problem, if a taxicab company knew of ways to drive revenue by pickup location, the company could potentially maximize revenue by having the drivers in the best places for the most profitable trips. Because of changes in the original datasets, only the years 2017 through 2019 could be used for exploring this part of the data.

The first item to note about this perspective on the data is the difference between the frequency of trips from a particular pickup location. In Figure 4, we can see that the most frequent pickup locations are mostly in the Manhattan area (with exceptions for JFK airport and LaGuardia airport) from 2017 to 2019. When the focus is on 2019, the same detail is present as demonstrated in Figure 5.

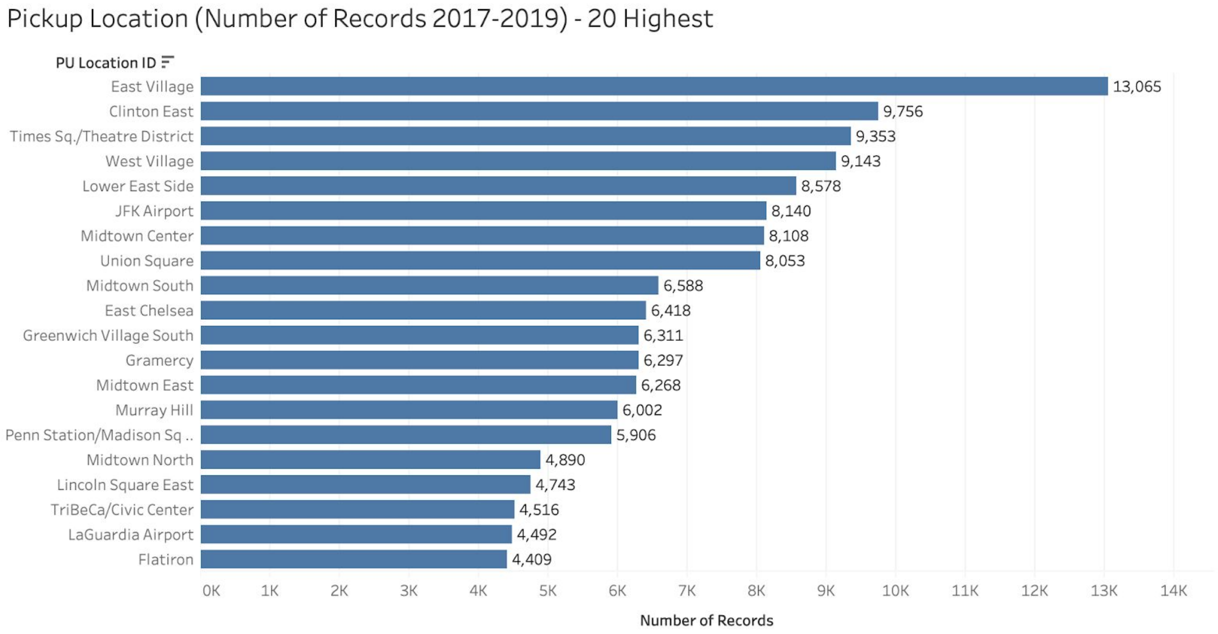


Figure 4

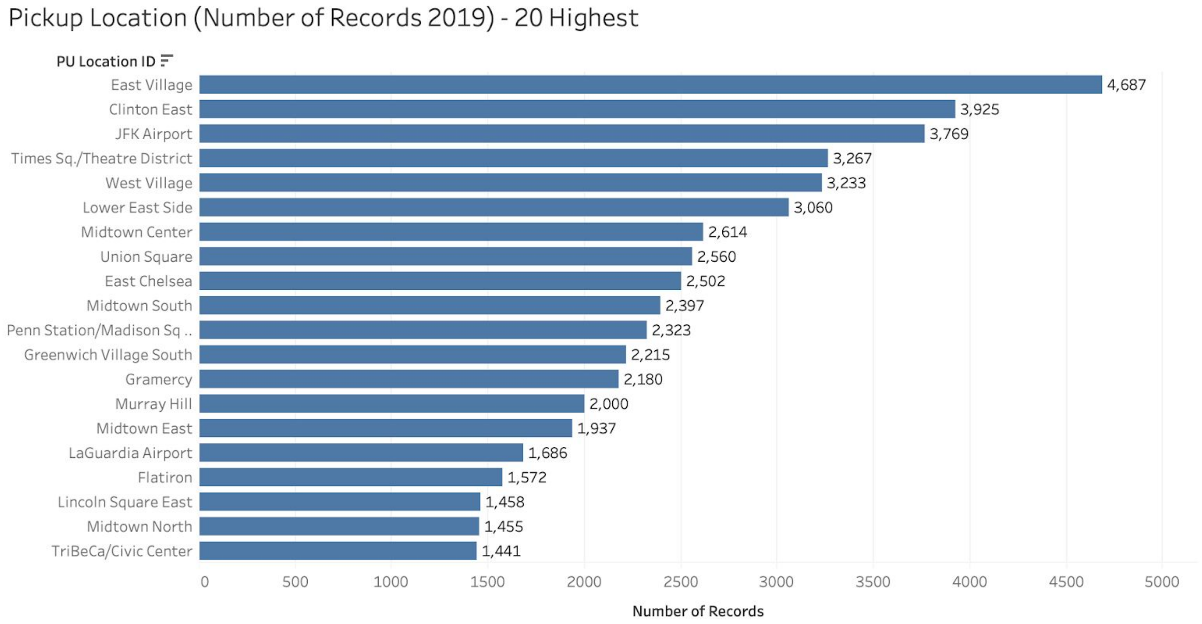


Figure 5

However, when the pickup locations with the highest average fare amounts for 2019 are collected, most of the pickup locations are in the other New York City boroughs. The significant exception seen in Figure 6 is JFK Airport.

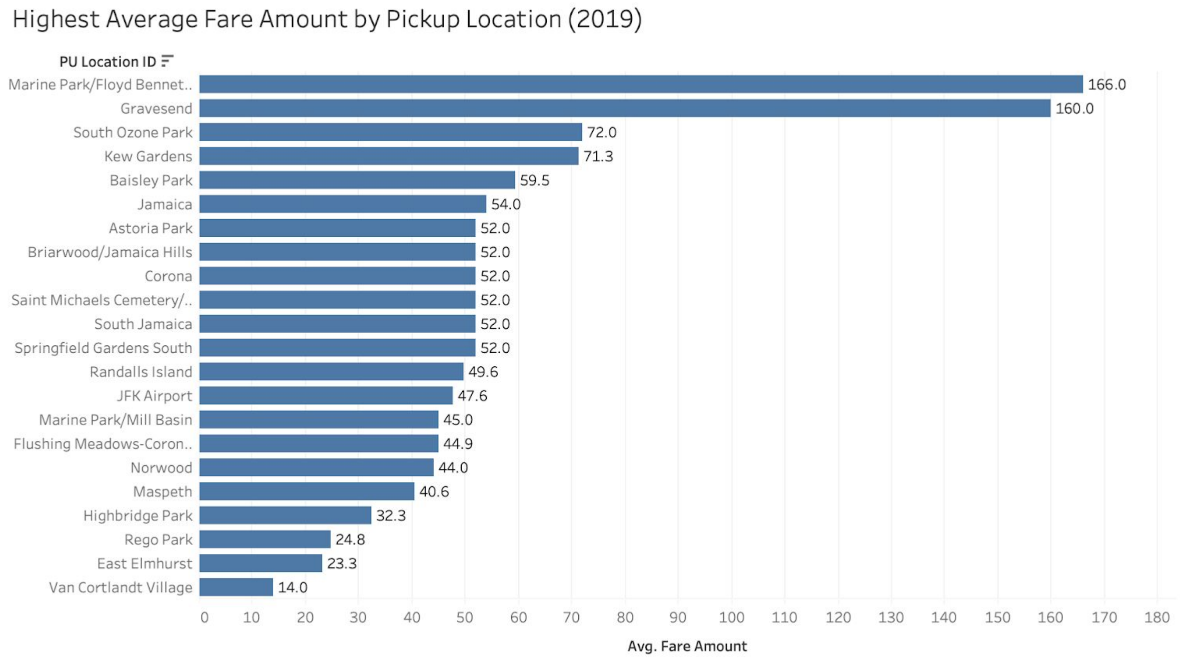


Figure 6

Another perspective was to look at tip amounts (Figure 7) and tip percentage (Figure 8). While tips are not part of a company revenue, they are a significant variable as it is a factor in driver satisfaction. Drivers will partly be motivated to go to the pickup locations where they can get the most tips. The results are a mix of locations.

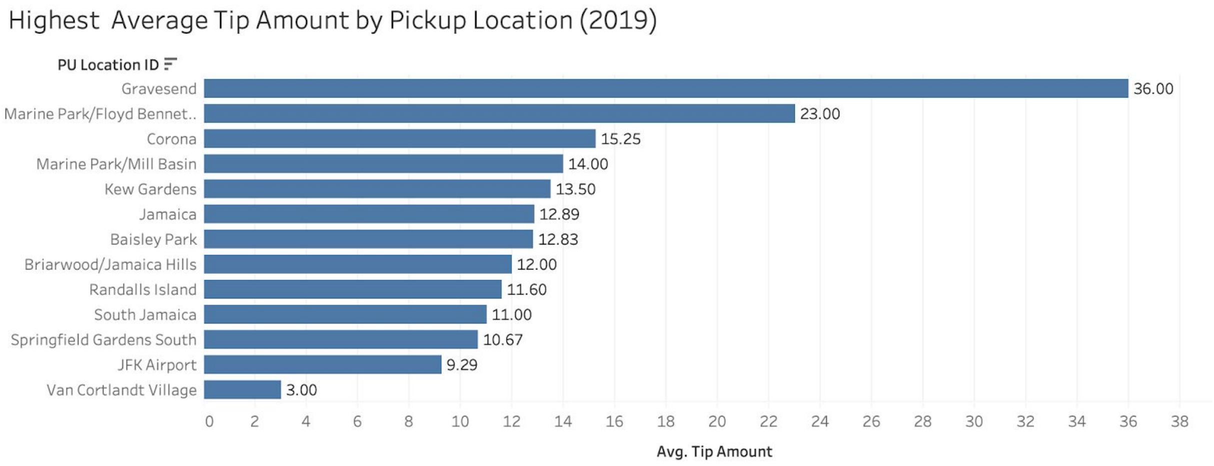


Figure 7

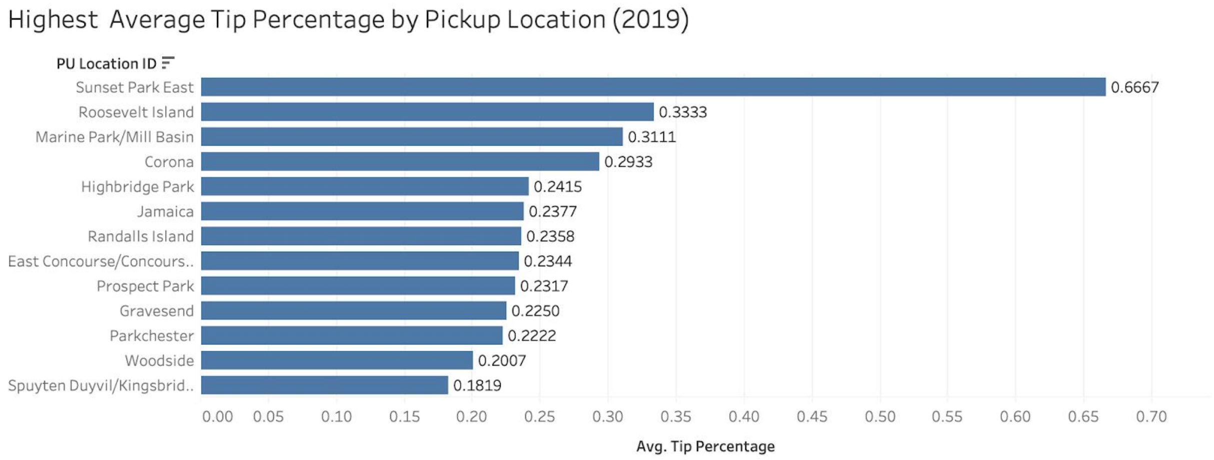


Figure 8

These results indicate that the pickup locations that cause a high number of trips are have some of the lowest average fares and the pickup locations with the highest average fares are some of the least common pickup locations. Figure 9 illustrates that this does indeed appear to be true.

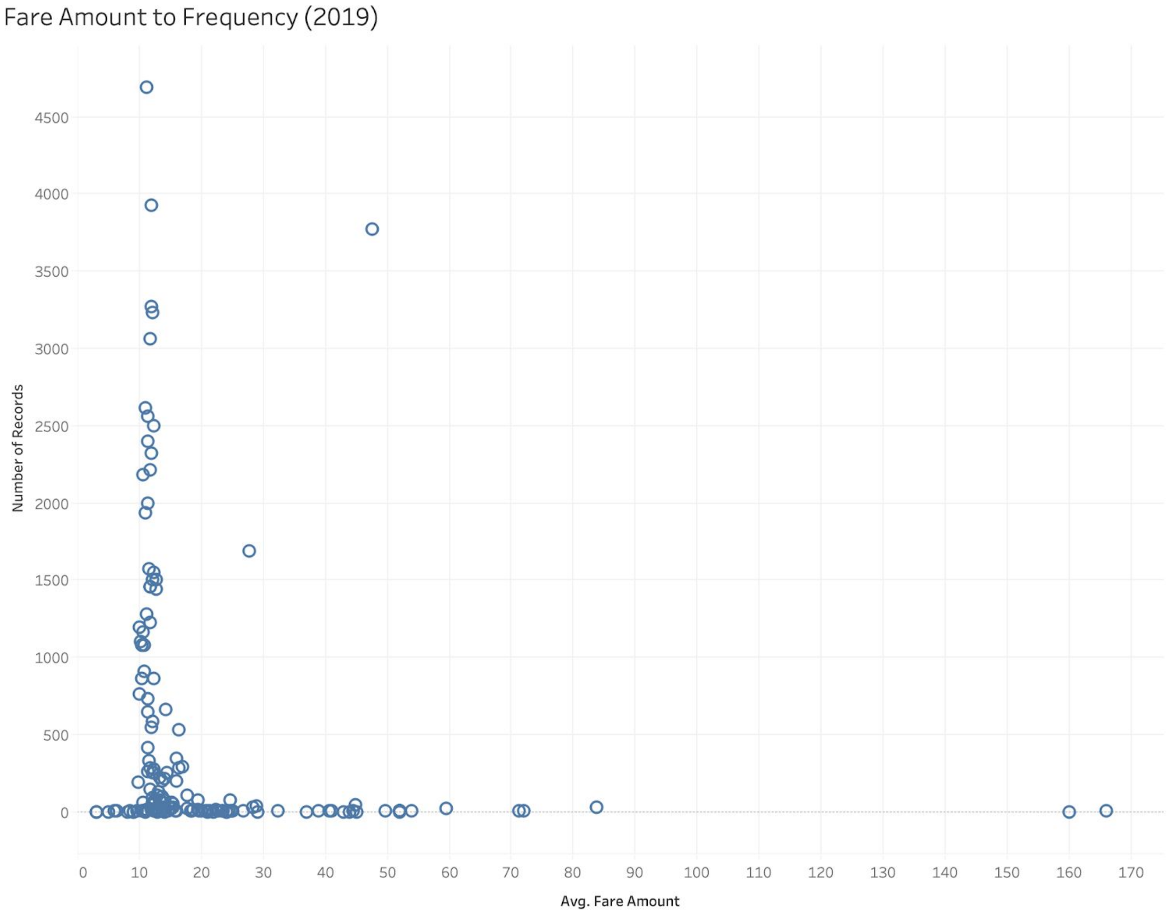


Figure 9

However, there is a chance that the locations that are frequent for pickups but low for average fare may drive more revenue for the company. While the fares may be low, their frequency may lead to a higher revenue and Figure 10 demonstrates this.

The other significant takeaway from Figure 10 is that the airports are outliers. JFK and LaGuardia have a high number of trips as a pickup location and high fares. This demonstrates that airports must be part of a taxi cab company's strategy to increase revenue.

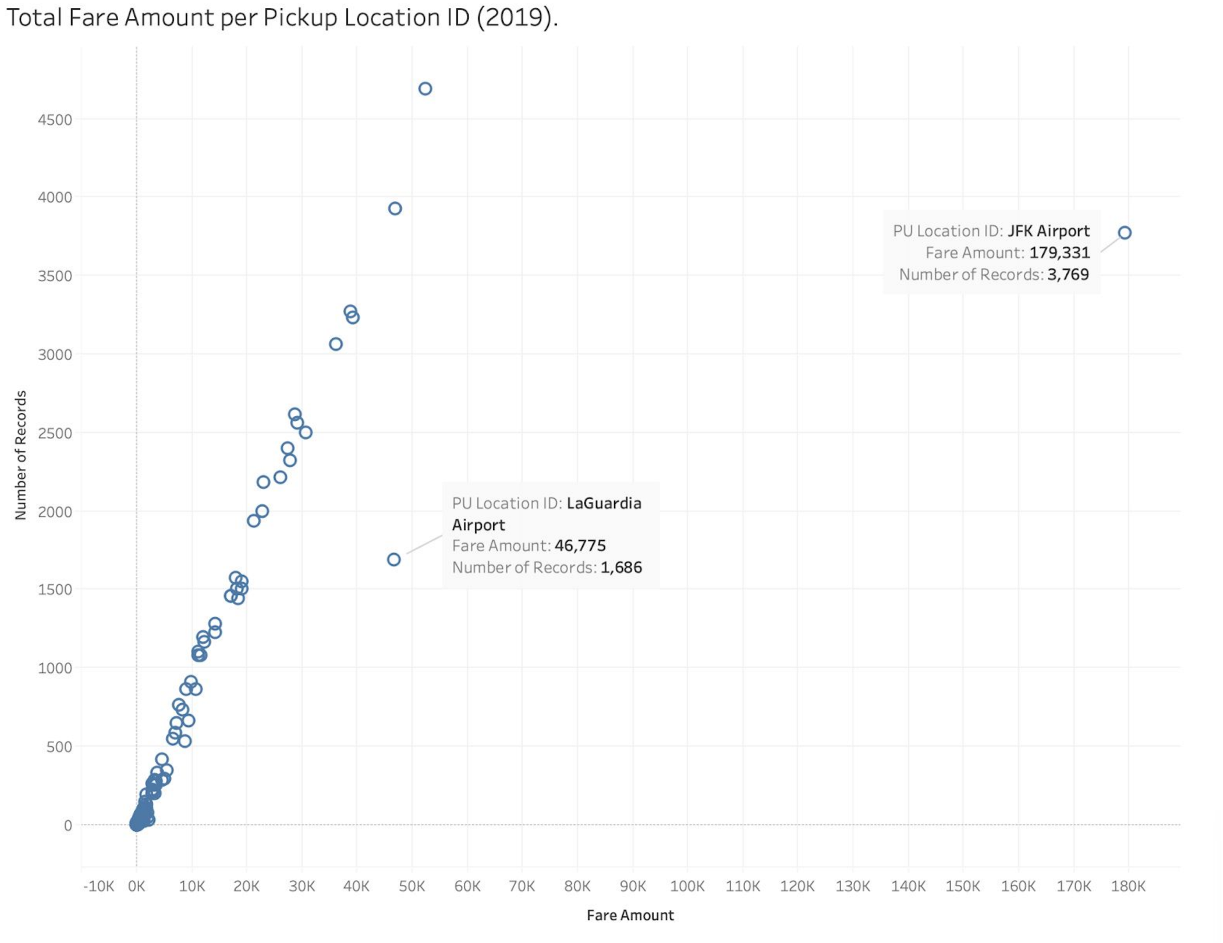


Figure 10

Machine Learning Models

The initial Machine Learning models were built in Python and consisted of a Neural Network, three Random Forest Models (10 Decision Trees, 50 Decision Trees, 100 Decision Trees), a Boosting ensemble model and a Bagging ensemble model. The Machine Learning Models are trained and tested on the following target variables:

1) Passenger Count

2) Pickup Year

3) Pickup Month

The models had a very low accuracy. The Pickup Year variable had a higher accuracy score than the Pickup Month but was still low. However, this may continue to indicate that rates and fees had small changes year over year and that there are no trends going from month to month.

Through Data Exploration, more insights were obtained when reviewing the data through the lens of the Pickup Location. For the first round of models, twenty-one locations were selected for target variables. The twenty-one target variables included the two airports, nine locations with a high number of occurrences but low average fare, and ten locations with a low number of occurrences but a high average fare. The results of these models also had low accuracy and ranged from 32.4% to 43.79%.

There was one significant factor to these models. For nearly all of the models, the accuracy score on the training data was close to 100%. The lowest was on the Bagging model which led to a higher score on the testing data. These results indicate that there was a possibility of overfitting with the training data. When trying to resolve the overfitting, two strategies come into play: adjust the size of the test and training datasets and reduce the number of feature variables.

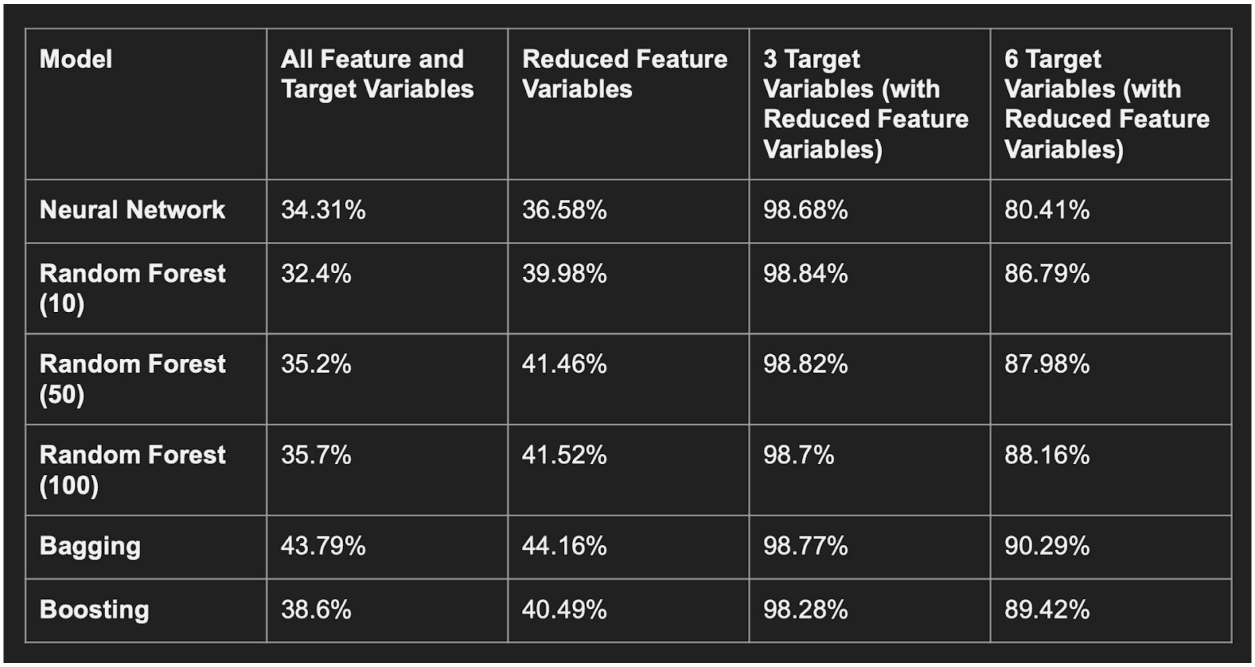
The models were run with a training to test ratio of 90/10, 80/20 and 60/40. The differences in the results on the test data were insignificant and the training sets still had close to 100% accuracy.

To determine which feature variables should remain or be removed, they were tested for correlation and covariance. Using both Pearson and Spearman correlation, no variable had a significant correlation with the Pickup Location ID. However, there were five variables that indicated a high covariance score. These variables were Trip Distance, Dropoff Location ID, Fare Amount, Tip Amount, and Total Amount. A new dataset was created with these five variables as the feature variables.

The models were tested again with fewer feature variables and the accuracy scores were still low. They were all under 50% and the Bagging Ensemble Model continued to have the highest score at 44.16%. The training scores continued to be close to 100% which indicated that if there was an overfitting problem, it had not been resolved.

Recalling one of the significant finds in the data exploration, there appeared to be three groups into which most of the pickup locations could be separated: airports, locations with high volume but low average fare, and locations with low volume but high average fare. For the next round of models, one location that fit each of these criteria were selected: LaGuardia Airport, East Village, and Astoria Park, respectively. All models achieved almost 100% accuracy on the testing data. These results signify a possibility that looking at the pickup locations through the three buckets was more important than distinguishing each pickup location individually as the differences may be insignificant. It is also important to note that some of the ensemble models used decision trees which will almost always perform better with a lower number of target variables.

For the final rounds of models, two locations for each category were selected for the target variables. JFK Airport and LaGuardia for the airports, East Village and Clinton East for high occurrence/low average fare, and Astoria Park and Jamaica for low occurrence/high average fare. The results ranged from 80.41% (Neural Network) to 90.29% (Bagging). Figure 11 outlines all results for the models that were used.



Model	All Feature and Target Variables	Reduced Feature Variables	3 Target Variables (with Reduced Feature Variables)	6 Target Variables (with Reduced Feature Variables)
Neural Network	34.31%	36.58%	98.68%	80.41%
Random Forest (10)	32.4%	39.98%	98.84%	86.79%
Random Forest (50)	35.2%	41.46%	98.82%	87.98%
Random Forest (100)	35.7%	41.52%	98.7%	88.16%
Bagging	43.79%	44.16%	98.77%	90.29%
Boosting	38.6%	40.49%	98.28%	89.42%

Figure 11

Results and Conclusion

The original goal of this project was to find a way for a taxi cab company in New York City to maximize revenue. Based on the data analysis performed, it is clear that the best approach is to focus on pickup locations. First, airports are consistent with passengers wanting rides with yellow cabs and having a high average fare. Second, drivers need to be available in Manhattan as there are many locations that have a high number of rides even though they have a low average fare. Figure 10 demonstrates that although rides from these pickup locations have a low average fare, their frequency still drives a higher revenue than locations that have a higher average fare but less demand for yellow taxi cab drivers.

References

[1]Guse, C. (2019 May 19). "Quick growth of Uber and Lyft has hammered NYC's yellow cab industry, data shows." Retrieved from https://www.nydailynews.com/new-york/ny-medallions-taxi-data-20190519-jt7y7s2ym5ad7n3hyjftq2ojdm-story.html

This article outlines some high-level data points comparing the decline of NYC taxi cab rides to the increase of Uber and Lyft rides. The article also discusses the decline in value and the increase in inventory of the NYC taxi cab medallions. The article also claims that there is no major price difference between Uber, Lyft, and NYC taxi cabs.

[2]Guse, C. (2020 Jan 30). "Driving NYC taxis out of business: How Uber and Lyft doomed the once-solid yellow cab industry." Retrieved from https://www.nydailynews.com/new-york/ny-medallion-foreclosures-taxi-bailout-plan-uber-lyft-20200130-s2mjkhjubzgptdxasoxddwdote-story.html

This article reviews the 2019 financial impact of the decline in the NYC taxi cab industry. The article also discusses talks in New York City Council to bailout drivers impacted by the decrease in value of taxi medallions.

[3]Skandul, E. (2019 Oct 16). "How New York City taxis can get ahead of Uber and Lyft." Retrieved from https://www.cityandstateny.com/articles/opinion/commentary/how-new-york-city-taxis-can-get-ahead-of-uber-and-lyft.html

The article recognizes how companies such as Uber and Lyft have popped the bubble of the New York City taxi cab Medallion and discusses potential business solutions for the taxi cab industry to be saved.

[4]Griswold, A. (2019 May 22). "No one is innocent in New York City's taxi market." Retrieved from https://qz.com/1624986/whos-to-blame-for-the-plight-of-the-taxi-driver/

The article states that Uber is not the only target for blame for the downfall of the NYC taxi cab industry and that financiers and marketers of the medallion were misleading in the sale of the commodity.

[5]Grant, C. (2019 May 21). "Taxi industry insiders — not Uber — created New York City's cab-tastrophe." Retrieved from https://thehustle.co/new-york-city-taxi-industry-lending/

The article discusses that lenders to cab drivers to finance the medallions are to blame for the devastating personal and financial impacts on those vulnerable borrowers. The article also makes a point that Uber should not be blamed for the downfall in the industry.

[6]"Responding to the Collapse of the New York City Taxi Medallion Market." Retrieved on Mar 12 2020 from https://www.ncua.gov/news/responding-collapse-new-york-city-taxi-medallion-market

The National Credit Union Association's (NCUA) response to the decline in value for taxi medallions and places blame on ride-sharing apps and provides resources to drivers who are impacted.

[7]Flamm, M. (2020 Jan 27). "Taxi-medallion task force to recommend rule changes, rescue plans." Retrieved from https://www.crainsnewyork.com/transportation/taxi-medallion-task-force-recommend-rule-changes-rescue-plans

The article discusses the city council task force that is looking into how to assist with the taxi-medallion loan crisis.

[8] Hickman, J. (2015 Nov 16). “New York City Taxi Revenue Declines Accelerate in Last Four Months.” Retrieved from https://www.thestreet.com/opinion/new-york-city-taxi-revenue-accelerates-decline-in-uber-era-13367110

The article from 2015 explores the sudden revenue declines in the New York City Taxi industry.

[9] Pesce, N.L. (2019 Aug 9). “This chart shows how Uber rides sped past NYC yellow cabs in just six years.” Retrieved from https://www.marketwatch.com/story/this-chart-shows-how-uber-rides-sped-past-nyc-yellow-cabs-in-just-six-years-2019-08-09

The article explores the data behind Uber and Lyft business growth in New York City and how it correlates with the business decline of the New York City taxi industry.

[10] Watt, C.S. (2017 Oct 20). “'There's no future for taxis': New York yellow cab drivers drowning in debt.” Retrieved from https://www.theguardian.com/us-news/2017/oct/20/new-york-yellow-cab-taxi-medallion-value-cost

The article talks to various New York City taxi drivers about the impact of the ride-sharing apps on their jobs and welfare.