Analysis of New York's medallion (yellow) taxis

Kristofer DeYoung

1 INTRODUCTION

This exploratory analysis aims to identify relevant factors that affect customer billing and utilization that are essential for descriptive and predictive analytics. Though rates are subject to change, the underlying payment schemes have likely remained consistent, providing valuable insights for analyzing average demand, revenue, and pricing predictions.

New York legally mandates that only 13,587 yellow taxis operate in the city, each with an individual medallion attached. These vehicles can be either owned or leased by licensed garages. The iconic yellow taxis can pick up street hails anywhere within New York City's limits, encompassing all five boroughs (Staten Island, Brooklyn, Queens, Manhattan, and the Bronx). On the other hand, green taxis must abide by specific exclusion zones, such as above W110St/E 96th St in Manhattan1. Traditional street hailing involves signaling an on-duty driver from the side of the road, but nowadays, e-hailing through apps like Curb or Arro is also common.

The Taxi & Limousine Commission (TLC) has been collecting For-Hire Vehicle (FHV) data since 2015, although the data points have varied. For instance, drop-off time, date, and location were mandated to be included in the data in 2017, in addition to the already gathered pick-up information[5]. Taxi drivers, both yellow and green, are regulated to work a maximum of 10 hours within 24 hours and 60 hours per week. They may only pick up new customers if these time limits are within the limit, or they risk a $200 fine. However, if a driver has already picked up a customer before exceeding the time limit and the ride duration spans over their limit, they will not be fined[1]. The driver's pay is based on time, distance, and utilization. These rules do not necessarily affect customer fares but ensure that drivers of larger FHV companies are fairly compensated per trip and motivated to maximize utility. Although these regulations were put in place as of January 2019, a similar scheme may have existed previously[1].

The payment system can be complex, with multiple variables impacting the total fare. Identifying the most relevant underlying factors contributing to customer billing is crucial for predictive analysis. The following factors, taken from the government website[4], are considered to potentially impact this analysis:

- $3.00 initial charge.
- Plus 70 cents per 1/5 mile when traveling above 12mph or per 60 seconds in slow traffic or when the vehicle is stopped.
- Plus 50 cents MTA State Surcharge for all trips that end in New York City or Nassau, Suffolk, Westchester, Rockland, Dutchess, Orange, or Putnam Counties.
- Plus $1.00 Improvement Surcharge.
- Plus $1.00 overnight surcharge 8 pm to 6 am.
- Plus $2.50 rush hour surcharge from 4 pm to 8 pm on weekdays, excluding holidays.
- Plus New York State Congestion Surcharge of $2.50 (Yellow Taxi) that begins, ends, or passes through Manhattan south of 96th Street
- Plus tips and any tolls, including returns that driver must pay.
- There is no charge for extra passengers, luggage, bags, or paying by credit card.
- There is no extra charge for digital hailing.
- Rate #01 - Standard City Rate, within the City limit to Nassau or Westchester.
- Rate #02 - Standard City Rate, JFK Airport.
- Rate #03 - Standard City Rate, Plus $20.00 Newark Surcharge.
- Rate #04 - Out of City Rate, The metered fare is double the amount from the City limits to your destination, Nassau or Westchester.
- Rate #05 - Out of City Negotiated Flat Rate between the Driver and Passenger.

It is also important to note that the rates mentioned above are current as of 19 December 2022, though they still provide value to an analysis; solely depending on the rate amount would be insufficient in any predictive analysis when using older data. However, rates have likely compensated for inflation; it is probable that the underlying payment schemes for pick-up scenarios still have remained relatively consistent, given the past ten years. This is also confirmed in "The New York Times" that rates increased by an average of 23% in 2022 from the last rate hike, which was in 2012.[6] Considering also that the medallion system that limits the number of yellow taxis operating in the five boroughs of New York was first implemented in 1937.[4] Further research on fare amount for 2016 could be done by sending a request to the NYC government.

2 EDA AND TRANSFORMATION

The research showed several factors relevant to the analysis, but also the potential for mistakes given the nature of how the data was collected; data such as geographical and meter is susceptible to human error and mechanical failure. Thus preceded on-side caution was prudent given the current circumstances. The data set itself is openly provided by government collecting agency[3] and spans the period of January 2016. The data set has 10 906 858 records with 19 attributes related to each fare, totaling 207 million entries. The following components have been identified:

- Time: pickup and dropoff
- Location: Rate code, trip distance, pickup and dropoff (latitude and longitude)
- Pricing: Rate code, Total amount is the aggregate of (fare amount, extra, mta tax, tip amount, tolls amount, improvement surcharge)
- Organizational: VendorID, passenger count, payment type, store and forward flag

Initial exploration showed most of the data to be either uniformly distributed or significantly skewed, with a considerable amount of data points missing. For example, all four attributes with geographical data were missing 1.5% of all records, while other locations were pointing to the middle of the Atlantic, most likely as machine faults when recording data. Other examples of extremes were trip distances of 8 million miles or fare amounts of 111 271 dollars when the average is 12.49 dollars. The precedence of such extremes only further confirmed initial predictions. In an attempt to normalize data, two different " clean " methods were conceptualized; more details are in the next section.

2.1 IQR

The initial analysis showed many outliers skewing the min and maximum data of a lot of the data. Utilizing methods such as the z-score for outlier detection was impossible since most of the data needed to be more balanced and uniformly distributed. However, mapping the amount times each record of the dataset fell outside of the Inter Quartile range(IQR) provided an overview of the number of outliers by calculating the first quartile (Q1) and third quartile (Q3) and then computing the IQR as the difference between Q3 and Q1. Next, filtering the DataFrame so that the name and count of each record outside of the range (Q1 - 1.5 * IQR) and (Q3 + 1.5 * IQR) was appended to a new column at the end of the row, effectively noting the frequency of outliers in a given row and listing which parts of the row fell outside the range. The frequency of outliers and the amount of rows influenced is displayed in Figure 1.

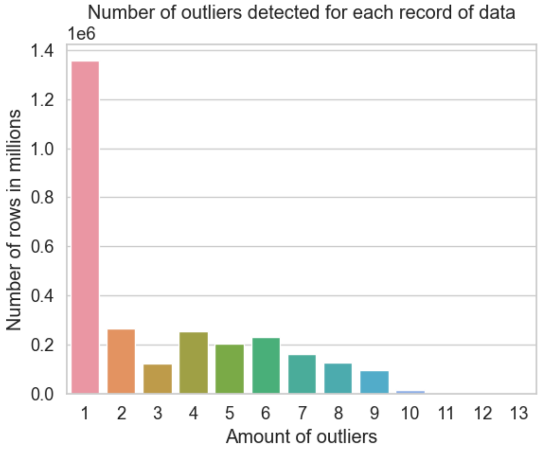


Figure 1: Number of outliers with IQR

As seen in figure 1, Using the IQR method was adequate; however, it also removed 26% of data entries resulting in 8 058 491 records left. A few general tests were conducted, varying by only removing rows with two or more outliers. However, more thorough testing could be performed, and there needed to be more significant results for the predictive analysis. IQR is more robust to handle datasets with non-normal distributions and skewed data. Thus allowing for a more accurate representation of the underlying patterns; However, the research showed that specific data points could do better with a manual approach of removal or imputation.

2.2 Timeseries

After initial exploration, adding a trip duration to the data set, a trend emerged of what can only be explained as driver error in forgetting to end the trip meter. Most interesting was that 14 508 of those trips ended within 23-24 hours of running, leading to deductions that the 60-hour work limit was reached for the respective week. The driver then proceeded with roughly 24 hours of rest depending on when the last customer was dropped off, only realizing that the meter had still been running first, returning to the cab after a day of rest. After the 24-hour mark, however, the number of mistakes had a drop off to only 18 such errors in the data set. To account for the discrepancy caused by human error, removing all trips spanning over 11 hours would account for any extreme difference. The total removed accumulated to 15 857 trips where the driver presumable forgot to stop the meter. Further exploration of time series into the positively skewed distribution showed data points spanning into March, as the data set only represented the month of January; it further reinforced the theory that these outliers are most likely due to human error, as previously described forgetting to turn off the taxi meter. The data set was limited to 01 January, including 01 February, considering any rides ending after midnight. Further research that could be done to predict whether the meter is being accidentally left running would require more factors and predictive analysis, such as; the amount of activation not resulting in cab fare. However, this is outside the scope of this analysis.

2.3 Geo-spacial data

The geographical data (ratecodeID, drop-off/pick-up longitude/latitude, presented further challenges. It remained clear that the extreme deviations in location data were errors, as pointed to the mid-Atlantic ocean. Furthermore, the initial analysis showed an average of roughly 1.5% zero value, most likely due to GPS losing connection, as these data points are automatically collected when the taxi meter is started. As detailed in the introduction, research into payment schemes showed that many factors aggregated into the total amount charged to customers were influenced by location data. Given the nature of the predictive analysis, an effort was put into preserving and recreating as much of the value as possible. Since the only indication of missing geographic location was the categorical data "ratecodeID", each airport had its own category while the standard rate was within the five boroughs and negotiated fare on other stretches; thus, it could help the analysis in recreating geodata that was null. However, the attribute "ratecodeID" also had multiple values of 99, which was not explained through preliminary research. To solve data errors in both directions, the categorical data were grouped, and average MODE was used to imputing missing values in geodata with the most commonly occurring value in each rate code group to replace missing values. Limitations were set to only remove values with 0 geodata and 99 in the rate code. Though it is not entirely accurate, it is still believed to contribute to providing value in predictive analysis. The remaining data set was limited to the longitude and latitude of the larger New York area, thus taking care of any extreme outliers. Further data analysis could be made by grouping the geodata into the boroughs or other areas of significance, such as airports and city limits, allowing cross-validation for most attributes related to charges or pricing.

2.4 Pricing

The following attributes are considered situation-dependent extras; therefore, null values are very likely; 'extra', 'mta_tax', 'tip_amount', 'tolls_amount', 'improvement_surcharge', so little was done in cleaning. It was observed that "fare_amount" and "total_amount" had a high positive skew, where extreme values mostly had no observable correlation with other attributes such as trip duration or trip distance. Even when considering that more expensive services such as limousine data would possibly skew the data set because of the typically higher expense to operate, however through initial research, we found no pricing structure for more expensive alternatives such as limousines; furthermore, the data set was supposed to be restricted yellow medallion taxis. In conclusion, the explanation is that these resulted from an error. The sample of extreme values was insignificant, so the decision was to remove all data entries outside the IQR. While also eliminating any values from the total amount that are 0 as these don't provide any data.

2.5 Other discrepancy's

The number of passengers "passenger_count" has a total of 500 null values as this a driver-entered value, and data entries otherwise looked valid; the assumption was that this is a mistake and the null value was replaced with the most frequent value or the mode of the given attribute.

2.6 Weather Anomaly

The drop in the frequency on 23-24 January was likely a result of the largest blizzard on record since 1869, where total snowfall in central park was measured up to 69.85 centimeters [7]. The consequence of the anomaly, unofficially nicknamed "snowzilla". Initially, the major declared a state winter emergency; this resulted in a travel ban on 23-24 because of the number of accidents in New York, New Jersey, and Newark, ultimately grounding air traffic.[2] Both leading up to and following this anomaly was a drop in taxi utilization, potentially compromising the data. It was also noteworthy that there was no excess use in the days following the anomaly from built-up demand; thus, we conclude that the event only negatively impacted the total usage numbers and revenue. When considering the proposed questions, it was found that the affected days would not accurately represent normal usage. Therefore, a separate instance of removing data between 22 January to 29 January was created to answer specific questions.

In summary, using multiple methods of outlier detection proved that the data was heavily cluttered with errors, empty data points, and extreme values. The result was three copies of the data set (no anomalies, Inter Quartile Range, Manual) using different data cleaning methods to perform more accurate descriptive and predictive analysis.

3 DATA ANALYSIS

3.1 Question 1

What is the average demand for taxis on the days of the week (i.e., daily trend). Which of the days has the highest and which lowest demand?

The prepared data set without weather anomalies were used to calculate the daily average, eliminating any skewed data resulting from weather anomalies. The data was first aggregated by date while counting the records for each date, adding this to a separate column called "demand". Then, the data was grouped by weekday and further aggregated by summing the demand for each record and dividing it by the frequency of each weekday in the data set. Ensuring that the data was accurately represented, considering that the frequency of weekdays is not uniform in January, as shown in the table:

	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
weather anomaly	3	3	3	3	4	4	4
no anomaly	4	4	4	4	5	5	5

The consequence of not considering such occurrences would be that merely summing and averaging and then aggregating the sums of averages together would result in Fridays, Saturdays, and Sundays having an extra day of data; Not accurately answering the question; instead, considering this resulted in the following averages seen in Figure 2.

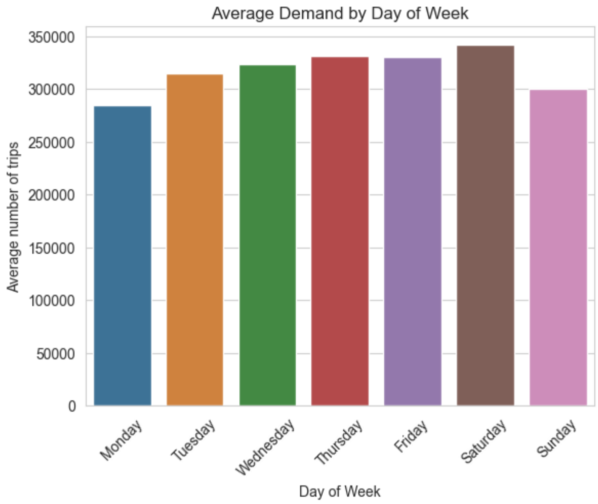


Figure 2: Average demand per day

As shown in Figure 2, the daily average is 314,860 trips. The day with the highest demand was Saturday, with 331,127 trips, representing a 10.5% increase in the number of trips. Conversely, Monday was the day with the lowest demand, with 299,593 trips.

3.2 Question 2

Which time of the day (morning, afternoon, evening, and night) is likely to be a peak period for the taxi's operation from the data?

A similar process for answering question 1 was used for this one also to prevent a record-breaking weather anomaly from further skewing the results; the impacted days were removed before averaging the demand per time of day. A new column was added, and data was aggregated while summarizing the count of "trips" of records in the data set. While again, determining that the occurrence of certain days of the week could impact the result; for example workday will have another pattern of demand vs. weekends. The day week was grouped and averaged with the frequency of occurrence in January, as represented in the table on question 1. The time was combined into a column of categorical data with the criteria as shown in the table.

	From	To
Morning	06:00	12:00
Afternoon	12:00	18:00
Evening	18:00	00:00
Night	00:00	06:00

The averages for the different periods of the day are then displayed in figure 3.

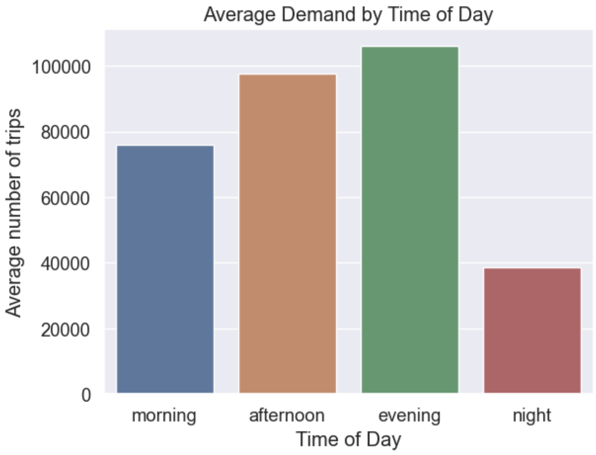


Figure 3: Average demand by time of day

In figure 3, we see that demand peaks at 106 061 trips in the evening, dropping down to 38 683 trips at night.

3.3 Question 3

On average, how much revenue was generated on the weekdays and weekends for the business for the period covered in the dataset? The previous questions referred to the usage amount, thus removing data impacted by the anomaly. However, this question referred to revenue generated on weekdays and weekends. The data set used to answer this was the one that was manually cleaned, i.e., only removed certain extremes. Similar methods were used in aggregating the data and grouping weekdays and weekends. A weekday is defined as Mon-Fri, and a weekend is Sat-Sun; the results in the table below is the weekly average for the month of January.

	Revenue ($)	percent of total
Weekends	6 852 764	36,26%
Weekdays	12 045 617	63,74%
Total	18 898 381	

The average revenue per week generated over the period is 18 898 381 $ with 36.26 % 6 852 764$ on weekends and 63.74% 12 045 617$ on weekdays generating on average 175.78% more revenue than weekends.

4 REGRESSION ANALYSIS

Since the records of that data were sequential, i.e., the first event transpiring that month is the first index, the test data was split so that the first 80% of the indexes data was used to train the data and the remaining 20% of the last record of the was testing.

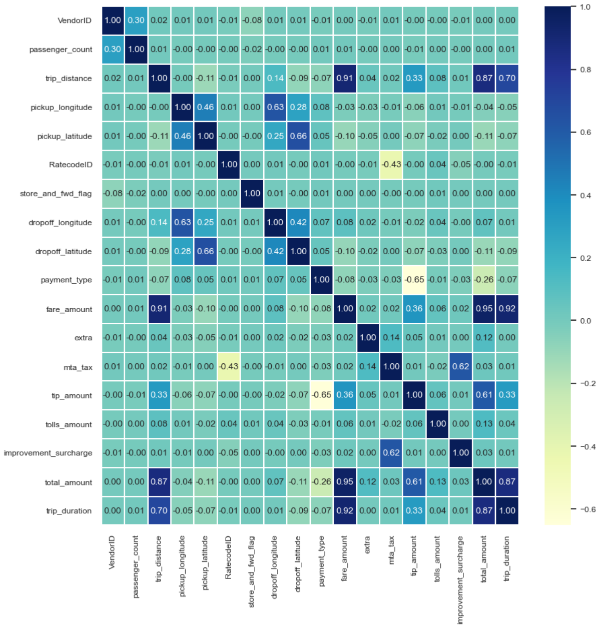
Through observation made when researching medallion taxis, the total amount is dependent on the sum of multiple attributes; thus, the most straightforward approach to the problem would be to summarize several or all of the columns into a combined column "sum of columns" or "total amount 2" on any test before making a prediction. Thus the training on such a model would yield the most accurate predictions; though the guidance is given not specify that this was not allowed, we continue the analysis under the assumption that it isn't. The multiple other variables that could impact the total amount the initial theory was, therefore that geographic location could give some could be valuable attributes for the potential predictive model; therefore, the large number of resources was utilized not only to understand the attributes but also to prepare and clean the data. Consequently, cleaning methods previously outlined yielded good results when predicting the test set that had been split after cleaning data; however, they resulted in poor results when predicting an uncleaned data set. As further elaborated in the next section, one interesting aspect was that only the simplest model outperformed in both aspects.

4.1 Correlation

A Pearson correlation coefficient was used in a correlation matrix to validate the hypothesis and potentially detect any other attributes significant to the prediction model.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

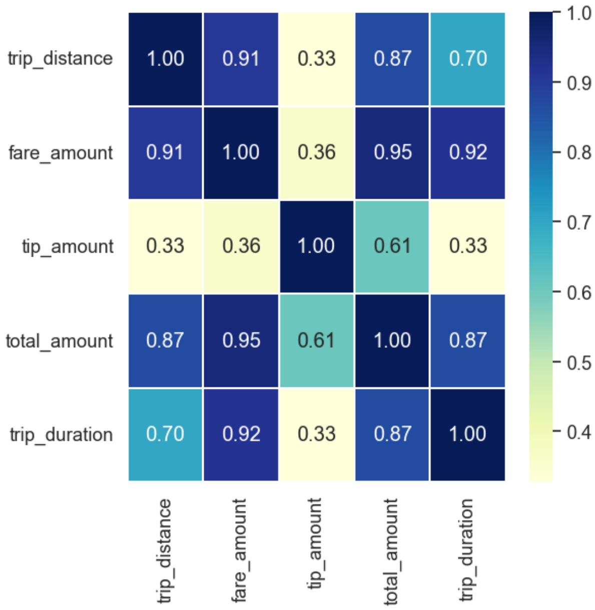
The correlation matrix shown in figure 4 is of all the attributes available.



	VendorID	passenger_count	trip_distance	pickup_longitude	pickup_latitude	RatecodeID	store_and_fwd_flag	dropoff_longitude	dropoff_latitude	payment_type	fare_amount	extra	mta_tax	tip_amount	tolls_amount	improvement_surcharge	total_amount	trip_duration
VendorID	1.00	0.30	0.02	0.01	0.01	-0.01	-0.08	0.01	0.01	-0.01	0.00	-0.01	0.01	-0.00	-0.00	-0.01	0.00	0.00
passenger_count	0.30	1.00	0.01	-0.00	-0.00	-0.00	-0.02	-0.00	-0.00	0.01	0.01	-0.00	0.00	-0.01	0.00	-0.00	0.00	0.01
trip_distance	0.02	0.01	1.00	-0.00	-0.11	-0.01	0.00	0.14	-0.09	-0.07	0.91	0.04	0.02	0.33	0.08	0.01	0.87	0.70
pickup_longitude	0.01	-0.00	-0.00	1.00	0.46	0.01	0.00	0.63	0.28	0.08	-0.03	-0.03	-0.01	-0.06	0.01	-0.01	-0.04	-0.05
pickup_latitude	0.01	-0.00	-0.11	0.46	1.00	-0.00	-0.00	0.25	0.66	0.05	-0.10	-0.05	0.00	-0.07	-0.02	0.00	-0.11	-0.07
RatecodeID	-0.01	-0.00	-0.01	0.01	-0.00	1.00	0.00	0.01	-0.00	0.01	-0.00	-0.01	-0.43	-0.00	0.04	-0.05	-0.00	-0.01
store_and_fwd_flag	-0.08	-0.02	0.00	0.00	-0.00	0.00	1.00	0.01	-0.00	0.01	0.00	0.00	-0.00	-0.00	0.01	0.00	0.00	0.00
dropoff_longitude	0.01	-0.00	0.14	0.63	0.25	0.01	0.01	1.00	0.42	0.07	0.08	0.02	-0.01	-0.02	0.04	-0.00	0.07	0.01
dropoff_latitude	0.01	-0.00	-0.09	0.28	0.66	-0.00	-0.00	0.42	1.00	0.05	-0.10	-0.02	0.00	-0.07	-0.03	0.00	-0.11	-0.09
payment_type	-0.01	0.01	-0.07	0.08	0.05	0.01	0.01	0.07	0.05	1.00	-0.08	-0.03	-0.03	-0.65	-0.01	-0.03	-0.26	-0.07
fare_amount	0.00	0.01	0.91	-0.03	-0.10	-0.00	0.00	0.08	-0.10	-0.08	1.00	0.02	0.02	0.36	0.06	0.02	0.95	0.92
extra	-0.01	-0.00	0.04	-0.03	-0.05	-0.01	0.00	0.02	-0.02	-0.03	0.02	1.00	0.14	0.05	0.01	0.00	0.12	0.00
mta_tax	0.01	0.00	0.02	-0.01	0.00	-0.43	-0.00	-0.01	0.00	-0.03	0.02	0.14	1.00	0.01	-0.02	0.62	0.03	0.01
tip_amount	-0.00	-0.01	0.33	-0.06	-0.07	-0.00	-0.00	-0.02	-0.07	-0.65	0.36	0.05	0.01	1.00	0.06	0.01	0.61	0.33
tolls_amount	-0.00	0.00	0.08	0.01	-0.02	0.04	0.01	0.04	-0.03	-0.01	0.06	0.01	-0.02	0.06	1.00	0.00	0.13	0.04
improvement_surcharge	-0.01	-0.00	0.01	-0.01	0.00	-0.05	0.00	-0.00	0.00	-0.03	0.02	0.00	0.62	0.01	0.00	1.00	0.03	0.01
total_amount	0.00	0.00	0.87	-0.04	-0.11	-0.00	0.00	0.07	-0.11	-0.26	0.95	0.12	0.03	0.61	0.13	0.03	1.00	0.87
trip_duration	0.00	0.01	0.70	-0.05	-0.07	-0.01	0.00	0.01	-0.09	-0.07	0.92	0.00	0.01	0.33	0.04	0.01	0.87	1.00

Figure 4: Correlation analysis

As expected, figure 4 shows a stronger correlation between the attributes contributing to the total amount. Although the data exploration and research found that geographical data showed potential for prediction in total amount, they showed a weak correlation with the total amount and a stronger correlation with each other. Reflecting on the limitation of the Pearson correlation coefficient is that it can not determine nonlinear relationships between variables. Thus further research could be conducted using the Kendall correlation, Spearman correlation, or Point-Biserial. However, it was worth exploring the attributes larger than 0.5 in a correlation score, as shown in figure 5.



	trip_distance	fare_amount	tip_amount	total_amount	trip_duration
trip_distance	1.00	0.91	0.33	0.87	0.70
fare_amount	0.91	1.00	0.36	0.95	0.92
tip_amount	0.33	0.36	1.00	0.61	0.33
total_amount	0.87	0.95	0.61	1.00	0.87
trip_duration	0.70	0.92	0.33	0.87	1.00

Figure 5: Correlation analysis above 0.5 correlation

When inspecting the heat matrix figure 3, the value with the strongest correlation with the total amount was "fare amount" with a score of 0.95; thus, it was chosen for the predictive model; the second highest was the combined column trip duration and "trip distance", and the last was the tip amount. As stated, the objective of this analysis was to use any attributes for prediction as "trip duration" and "trip distance" scored very similarly to each other with both "total amount" and "fare amount". The trip duration was chosen; although both these attributes scored well with the total amount, they scored 0.70 with each other and might have been ideal composition for a prediction. However that the problem did not limit the usage of the attribute that was a primary aggregate "total amount"; thus, "fare amount" and "trip distance" were chosen for the predictive model.

4.2 Training & Testing

Fare amount and trip duration were determined to have linear relationships. In contrast, the geographical longitude and latitude remained unknown. It was decided to utilize three different models, Linear, Polynomial, and Random forest, together with the four cleaning methods. The details of the clean techniques used for data preparation are outlined in the data description & preparation section. In summary, the preparation methods used are Inter quartile range(IQR), Manuel filtering(Man), the combination of aforementioned (man + IQR), and finally, "No weather anomaly" however, the removal of a whole week of data; turned out inferior results for the prediction model, so it's this dataset. It was not taken into the final results. Each model varied in the combination of data cleanpreparation, and the attributes used are specified in the following sections.

4.2.1 Linear Regression. Linear regression was trained on fare amount and trip duration, and the training and testing were done on each of three clean&preparation methods. The equation used for Linear regression was:

$$y = \beta_0 + \beta_1 x + \epsilon$$

4.2.2 Polynomial Regression. Polynomial regression was trained on drop-off longitude, drop-off latitude, pick-up longitude, pick-up latitude, fare amount, and trip duration. The training and testing were done on each of three different clean&preparation methods. The degree of the polynomial was set to 2, and the following equation was used:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \epsilon$$

4.2.3 Random forest Regression. The regression was trained on drop-off longitude, drop-off latitude, pick-up longitude, pick-up latitude, fare amount, and trip duration, and the training and testing were done on each of three different clean&preparation methods. This model was chosen for its flexibility, handling both contiguous and categorical attributes. In Random forest regression, there isn't a simple equation such as Linear and polynomial regression; instead, it uses a collection of decision trees and averages to obtain a final result.

4.2.4 Testing. Results provided in the table are R squared(RS), the coefficient of determination, and Mean Square Error(MSE).

	Linear		**Polynomial**		**Random Forest**	
	RS	**MSE**	**RS**	**MSE**	**RS**	**MSE**
Manual	0,9134	1,9611	0,9143	1,9418	0,7689	5,2355
IQR	0,9371	1,3901	0,9377	1,3773	0,7908	4,6216
Man+IQR	0,9191	1,2391	0,9199	1,2271	0,7549	3,7556

As seen in the table, the testing results concluded that the preparationscleaning method yielded the best overall scores on all models. Furthermore, it is observed that though random forest still performs admirably on all datasets, it significantly underperforms the other two models. In contrast, the Polynomial model uses the same attributes for the prediction; instead, the score is marginally higher than the simplest of the linear, which only uses two attributes. Another note is also that the previous conclusion in the correlation analysis was that "trip duration" and "trip distance" could be viable options for the prediction model, yielding an R2 score:0.9285 and an MSE: 1.58 on the test set with linear regression and on IQR dataset an impressive result considering that both these attributes are not an aggregate of the total amount.

4.3 Predictions

When validating the predictions made with the "new Sample" file, all columns with values in them were summed into a new "total amount" column while also adding a predicted value column to allow for calculating the R-squared score and Mean Squared Error. However, when predicting the total amount on the "new sample" data provided, the only model that performed within reason was the linear regression, as Index 20 on the "new sample" data provided was missing a drop-off time. Thus trip duration was zero, resulting in all other models predicting a much higher total amount for that row, consequently leading to inferior R-squared scores and large mean square errors. The top model in predicting an unknown data set ended up being both the simplest model combined with the simplest method of cleaning and preparing data result for the "New Sample" data set is an R-squared score: of 0.9824 and Mean Squared Error: of 2.78.

The prediction of the new data sample below for each record in the dataset is in the table below.

Index	**total_amount**	**predicted_total_amount**
1	5,8	6,9
2	21,3	23,8
3	11,5	11,9
4	7,8	8,6
5	25,3	23,8
6	17,3	19,3
7	9,4	8,6
8	7,8	8,6
9	9,8	9,7
10	17,3	19,3
11	11,8	10,8
12	17,3	19,3
13	9,0	8,6
14	18,0	18,2
15	12,4	11,4
16	7,0	6,3
17	20,2	18,7
18	22,0	20,4
19	36,3	34,7
20	10,8	9,1
21	68,8	69,9
22	53,3	46,7
23	12,7	10,8
24	8,8	8,0
25	28,6	26,7
26	15,4	14,3
27	19,0	17,6
28	10,3	10,3
29	20,2	18,8
30	12,8	12,5
31	8,3	9,1
32	17,2	15,9
33	9,8	10,8
34	6,8	7,4
35	25,6	23,8
36	9,3	8,6
37	4,8	5,2
38	29,8	26,7
39	7,2	6,9
40	8,8	9,7

5 DISCUSSION

The analysis conducted in this study aimed to understand the trends in taxi demand and revenue and develop a prediction model for the total amount generated from taxi trips. The study successfully identified patterns in taxi demand, with Saturdays having the highest demand, Mondays the lowest, and evenings being the peak period for taxi operations. Furthermore, it was found that weekdays generated significantly more revenue than weekends. Several prediction models were tested to find the most suitable one for forecasting the total amount generated from taxi trips. Linear, polynomial, and random forest regression models were evaluated using different attributes and data-cleaning methods. Among these, the linear regression model with fare amount and trip duration as predictors and the interquartile range (IQR) method for data cleaning emerged as the best-performing model, with high accuracy and low error rates. Although Geolocation was not effective predictive attribute in this analysis, the author believes that further research could yield better results.

6 CONCLUSION

This analysis has provided insights into taxi demand patterns, revealing that Saturdays and evening hours experience the highest demand for taxis. Furthermore, it was found that weekdays generate considerably more revenue than weekends. The linear regression model using fare amount and trip duration as predictors and the IQR method for data cleaning was identified as the most accurate model for predicting the total amount generated from taxi trips. These findings can be used to understand better and manage taxi operations, enabling more efficient allocation of resources and improved service quality. The predictive model can help stakeholders make data-driven decisions to optimize their operations and maximize revenue. Future research could explore additional factors influencing taxi demand and revenue, such as weather conditions or special events, and investigate the effectiveness of other prediction models and data-cleaning techniques.

REFERENCES

[1] [n. d.]. Fatigued Driving Prevention - Frequently Asked Questions - TLC. https://www.nyc.gov/site/tlc/about/fatigued-driving-prevention-frequently-asked-questions.page

[2] [n. d.]. January 2016 United States blizzard. https://en.wikipedia.org/w/index.php?title=January_2016_United_States_blizzard&oldid=1142226310 Page Version ID: 1142226310.

[3] [n. d.]. Taxi Fare - TLC. https://www.nyc.gov/site/tlc/passengers/taxi-fare.page

[4] [n. d.]. Taxis of New York City. https://en.wikipedia.org/w/index.php?title=Taxis_of_New_York_City&oldid=1144387741 Page Version ID: 1144387741.

[5] [n. d.]. TLC Trip Record Data - TLC. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[6] Ana Ley. [n. d.]. New York City Approves Taxi Fare Hike, Raising Average Fare 23%. ([n. d.]). https://www.nytimes.com/2022/11/17/nyregion/taxi-fare-hike-nyc.html

[7] â¢. [n. d.]. 2016 Blizzard Was NYC's Biggest Snowstorm on Record, NOAA Report Finds. https://www.nbcnewyork.com/news/local/nyc-new-york-city-blizzard-biggest-ever-january-23-2016/831660/