New York city taxi trip analysis

Neco Darian

ABSTRACT

This report presents an analysis of the NYC taxi trip dataset, with the aim of uncovering key insights into the daily trends and peak periods for taxi operation, as well as the factors affecting the total amount paid for a trip. The report includes an overview of the dataset, data preprocessing techniques, and data analysis methods, including exploratory data analysis and linear regression modeling. Key findings from the analysis include variations in taxi demand across different days of the week, with Friday having the highest demand and Monday the lowest demand, as well as the observation that the peak period for taxi operation is during the evening hours, particularly during rush hour times. Additionally, the report includes details on the development of a linear regression model to predict the total amount paid for a taxi trip, achieving a high level of accuracy in predictions.

1 INTRODUCTION

The taxi industry plays a significant role in New York City's transportation system, providing a convenient and flexible option for millions of residents and visitors each year. In 2019 alone, there were over 240 million taxi and for-hire vehicle (FHV) trips in the city, generating over $2.9 billion in fares.

The availability of large-scale, publicly accessible datasets from sources like the NYC Taxi & Limousine Commission (TLC) has opened up new opportunities for analyzing and understanding taxi travel patterns and trends in the city. In this report, we present an analysis of a subsample of the TLC's taxi trip dataset, which includes information on over 10.9 million completed trips.

Our analysis focuses on several key aspects of taxi travel in New York City, including trip patterns, fare structures, and passenger behavior. By examining these factors, we aim to provide insights into how the taxi industry is working on a hourly, daily, and weekly bases and build a regression model for prediction of total amount of a trip.

Overall, our analysis highlights the value of data-driven approaches to understanding complex transportation systems like the New York City taxi industry. By leveraging the power of large-scale datasets, we can gain new insights into the factors driving transportation patterns and help inform policy decisions that can improve the efficiency and sustainability of urban transportation systems.

2 DATASET DESCRIPTION

The dataset used in this project was sourced from the official website of the NYC Taxi & Limousine Commission (TLC)[?]. It contains a variety of explanatory variables related to completed taxi trips in New York City, including pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

There are two datasets for this project: the Main Sample and the New Sample. The Main Sample is the original data with over 10.9 million rows, while the New Sample is a separate subsample of the data that is used for specific analyses.

The dataset contains every trip done for the month of January back in 2016.

Overall, the dataset provides a comprehensive view of taxi trips in New York City and can be used to explore various aspects of taxi travel, such as trip patterns, fare structures, and passenger behavior.

2.1 Initial analysis, cleaning and preprocessing

In this project, we start with EDA to understand the structure of the dataset and identify any potential errors or issues in the data. Firstly, we display the first 5 rows of the dataset to have a quick overview of the data. Then, we show the shape of the dataframe, which gives us an idea of the number of rows and columns in the dataset. Next, we show the data types of values that exist in the columns, which help us understand the nature of the variables we are dealing with.

To ensure the data is reliable and accurate, we search for any missing values in the dataset. Furthermore, we show some descriptive statistics of the numerical columns to identify any outliers or anomalies that may exist in the dataset. In this step, we notice that some numerical values are less than 0, which does not make sense. Additionally, we check if each of the numerical variables is within the dataset description limits. During this sequence, we find errors that do not make sense and are a result of some error.

After performing the EDA, we proceed to data preprocessing, which involves cleaning and preparing the data for further analysis. We identify some incorrect values in the dataset and decide to drop trips with wrong numeric values.

We want to handle the data in a way that we can describe a real-life taxi trip. Therefore, we set up some limitation to the dataset. Generally, most trips last more than 3 minutes, so we choose the data with duration more than 3 minutes. The provided trip data is for New York city, which makes it normal to choose the coordinates for NYC only. According to github account of "Jakebathman" **(-71.7517 - 79.7624) (40.4772 - 45.0153)** are the coordinates that we will be using to exclude trips outside this range.

We take the absolute values of all money-related negative values to make them non-negative. We also remove trips where the improvement_surcharge is not equal to 0.3 or 0, trips where mta_tax is not equal to 0 or .5, and trips where extra is not equal to .5, 1 or 0.

We do replace '99' in RateCodeID with the most frequently observed 1. Additionally, we remove rows with extremely high values of Tip_amount, Tolls_amount, fare_amount and Total_amount. According to their website the standard fare amount is 3 dollars and that will be the minimum amount for our dataset. After all these changes we will end up with 8087058 rows that we will be working on for this analysis.

3 DATA ANALYSIS

After cleaning, preprocessing, and converting the data, we are ready to answer the questions asked in this project. For Task A, we need to work with datetime and total amount data from the preprocessed dataset. We create a target datetime column, tpep_pickup_datetime, and calculate daily and hourly trends based on the time of the taxi trip start. We also create temporary additional columns to answer questions related to time.

3.1 Task A.i

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?

To answer the question, we group the trips by the day of the week using the ’tpep_pickup_datetime’ variable, which represents the date and time when the trip started. The ’count’ function was applied to this variable to count the number of trips for each weekday. The resulting dataframe was then renamed and the index was also renamed to ’weekday’.
To calculate the average demand for taxis, we divided the total number of trips by the number of days in the dataset. This was achieved by dividing the ’avg_demand’ column of the dataframe by the total number of records in the dataset. Finally, we print the results, which shows that the highest demand day is Friday, and the lowest demand is Monday(Table.1).

Weekday	Average demand
Monday	0.115688
Tuesday	0.127905
Wednesday	0.134036
Thursday	0.140346
Friday	0.0180809
Saturday	0.153640
Sunday	0.147575

Table 1: Average demand for taxis on the days of the week.

3.2 Task A.ii

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

To do this task, we first extract the hour of the day from the pickup timestamps in the dataset and categorize them into four time periods: morning (6am to 12pm), afternoon (12pm to 6pm), evening (6pm to 12am), and night (12am to 6am). We then group the data by these time periods and count the number of pickups within each period to obtain the hourly demand for taxi services.
We get the results by grouping the data by the ’day_time’ column (which contains the time periods we created) and counting the number of pickups in each group using the ’tpep_pickup_datetime’ column. The resulting dataframe, ’hourly_demand’, shows the demand for taxi services in each of the four time periods.

We see in (table.2), that the highest demand for taxi services occurs in the evening period, with over 2.3 million pickups recorded during this time. The second highest demand occurs during the afternoon period, followed by the morning and night periods. This information can be useful for taxi companies and drivers in planning their schedules and allocating resources to meet the needs of customers during peak demand periods.

Day-time	Demand
Evening	2804546
Afternoon	2358825
Morning	1867904
Night	1055783

Table 2: Peak period of the day for taxi operation.

3.3 Task A.iii

On average, how much revenue was generated in the weekdays and weekends for the business for the period covered in the dataset?
For this task, we first create a new future called "working_weekend". Using the NumPy library, we assigned the value "weekday" to each row in the dataset where the pickup date was a weekday (Monday through Friday), and "weekend" to each row where the pickup date was on a weekend day (Saturday or Sunday). This allowed us to distinguish between trips made during weekdays versus weekends. Next, we use the "value_counts()" function to count the number of trips that fell into each of these categories, resulting in the numbers shown in (table.3). From these counts, we can infer that more trips were made during weekdays, which can indicate that more revenue was generated during the weekdays compared to weekends.Lastly, we display the average revenue for each day of the week(table.4)&(figure.1).

Working_days	Average_revenue
Weekday	5651111
Weekend	2435947

Table 3: Average revenue generated in the weekdays and weekends.

Weekday	Average demand
Monday	18.870004
Tuesday	18.770447
Wednesday	18.432362
Thursday	18.621527
Friday	18.196327
Saturday	16.903955
Sunday	17.668185

Table 4: Average revenue based on days.

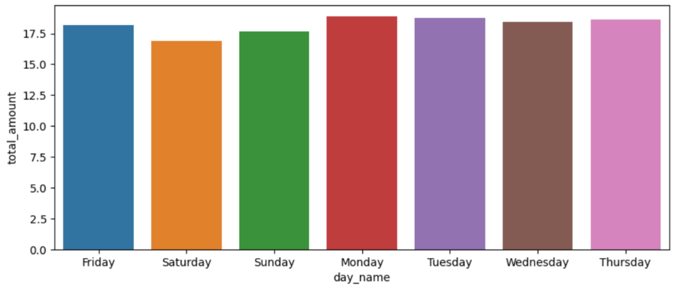

Figure 1: Average revenue based on days.

4 REGRESSION ANALYSIS

4.1 Task B

In order to achieve the most accurate prediction model, several steps were taken during the regression analysis process. Firstly, outliers were removed from the data to prevent them from negatively influencing the performance of the prediction model. These values are not necessarily mistakes, but they can significantly impact the accuracy of the model.

Next, specific variables were examined and adjusted to ensure they were appropriate for use in the regression analysis. The 'trip_distance', 'tip_amount', and 'tolls_amount' variables were adjusted by keeping only the first 97.59% of values. This approach was also applied to the 'fare_amount' variable, which was deemed to be the most influential and significant contributor to the target 'total_amount' variable.

As regression models only work with numerical variables, adjustments were made to the 'tpep_pickup_datetime' and 'tpep_dropoff_datetime' timestamp columns by extracting the day, weekday, and hour values to keep important information for the regression model. The 'store_and_fwd_flag' column was also converted from Y and N values to 1 and 0 respectively.

To identify the variables with the strongest and weakest correlation with the 'fare_amount', a correlation analysis was conducted. It was observed that the 'total_amount', 'trip_distance', 'tip_amount', and 'tolls_amount' variables had a strong positive correlation with 'fare_amount', which was logical.

As such, a linear regression model was used to identify the most noteworthy features that explained the target variable ('total_amount') and their contribution to its value. The numerical preprocessed variables 'trip_distance', 'fare_amount', and 'tolls_amount' were selected as the features (X), while the 'total_amount' column was the target variable (y).

The dataset was then split into train and test sets for further training and testing of the regression model. Finally, the Linear Regression model was trained with default properties and used to predict 'total_amount' based on the features from the test part of the dataset.

After training the linear regression model and predicting the target variable ('total_amount') based on the test dataset, the next stage is to evaluate the model's accuracy. To assess the accuracy, several metrics were used, namely MSE (mean_squared_error), RMSE (Root-mean-square deviation), R2 (coefficient of determination), and MAE (mean_absolute_error).

Using the test dataset, which consists of 20% of the initial values of the 'total_amount' variable, we will compare the predicted values ('y_pred') with the actual values ('y_test') using these statistical metrics. By comparing the predicted and actual values, we can determine the degree of accuracy of the model.

MSE measures the average squared difference between the predicted and actual values. RMSE is the square root of MSE and provides a measure of the absolute fit of the model. R2 measures the proportion of variance in the target variable predictable from the independent variables. Finally, MAE measures the average absolute difference between the predicted and actual values(table.5).

By assessing the accuracy of the model using these metrics, we can determine whether the linear regression model provides a reliable prediction of the 'total_amount' variable based on the selected features.

Metrics	Accuracy
MSE	3.9597
RMSE	1.9899
MAE	1.4763
R2	0.9760

Table 5: Accuracy score.

The R2 value of the model is 0.9759900299233435, indicating that 97.59% of the variation in the 'total_amount' variable can be explained by the variables in the model. With this information, we will proceed to find the regression coefficients(table.6) and create an equation for the model, which is:

total_amount = 1.46 + 0.03 * trip_distance + 1.1 * fare_amount + 1.24 * tolls_amount.

Features	Coefficient
trip_distance	0.032133
fare_amount	1.102305
tolls_amount	1.239618

Table 6: Features and their coefficients.

The accuracy of the trained model seems to be good enough, so we can now use it to predict the total amount paid for the trip records shown in the New Sample file. However, before doing so, we need to preprocess the New Sample data in the same way as we did with the trained dataset.

Lastly, we are tabulating the predicted values in the same order the records were arranged in the dataset and display the outcome as shown in (table.7).

Index	predicted_total_amount
0	6.993039
1	23.682490
2	11.994862
3	8.672202
4	23.685060
5	19.246601
6	8.670596
7	8.659349
8	9.789288
9	19.240495
10	10.904125
11	19.244673
12	8.664491
13	18.116983
14	11.444352
15	6.445421
16	18.683880
17	20.344728
18	41.187354
19	9.219820
20	67.597308
21	61.266551
22	10.882917
23	8.118158
24	26.470705
25	14.245421
26	17.552656
27	10.328230
28	18.678097
29	12.565936
30	9.232995
31	15.891166
32	10.892236
33	7.554474
34	23.684739
35	8.674130
36	5.326408
37	26.492556
38	6.995931
39	9.761975

Table 7: Features and their coefficients.

5 DISCUSSION

We have trained a Linear Regression model with default properties and achieved outstanding results in predicting our target variable using the listed features. Despite the availability of many other regression models, we did not need to explore them as our simple model produced a remarkable accuracy of 97.6% on the test dataset. This model serves as a valuable tool for predicting the total amount of a taxi drive and determining which features contribute most significantly to the amount value.

Prior to modeling, we preprocessed the data, making some assumptions and allowances. We observed numerous incorrect values and mistakes in the data. For example, some trip distances were either null or unrealistically large. To rectify this, we could have calculated the distance between the pick-up and drop-off geo locations, but this is also not reliable due to numerous records containing zeros or identical values.

Ideally, we could have verified and corrected values in the datasets using some calculations to obtain more precise data. However, our model produced excellent results even without such refinements.

6 CONCLUSION

Based on the analysis of the taxi trip dataset, several key findings were uncovered. Firstly, the average demand for taxis varied across different days of the week. Specifically, Friday had the highest demand while Monday had the lowest demand. Secondly, the peak period for taxi operation was observed to be in the evening hours, particularly during rush hour times.

Additionally, it was found that the average revenue generated by the taxi business was slightly higher during weekdays compared to weekends.

After preprocessing and cleaning the data, a linear regression model was created to predict the total amount paid for a taxi trip, given trip information such as time, distance, fees, and fares. The model was trained on 80% of the data and tested on the remaining 20% to ensure its generalization abilities. The final model showed a good performance in predicting the total amount paid for a taxi trip, with a MAE (1.4663) and high R2 score (0.976).

In conclusion, the analysis of the taxi trip dataset provided insights into the daily trends and peak periods for taxi operation, as well as the factors affecting the total amount paid for a trip. The developed regression model could be used to predict the total amount paid for a taxi trip, helping the taxi business to optimize their pricing strategies and increase revenue.

REFERENCES

[1] NYC Taxi & Limousine Commission (TLC). (2021). Taxi and For-Hire Vehicle Trip Records. Retrieved from https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[2] JakeB. (n.d.). State Boundaries. Retrieved from https://gist.github.com/jakebathman/719e8416191ba14bb6e700fc2d5fccc5