Predicting Taxi Fare Prices in NYC: An Analysis of Temporal and Spatial Influences

Applied Data Science Assignment 1

Jeremy Lau
Student ID: 1356056
https://github.com/jeremlll/ADS-Assignment-1

August 31, 2024

1 Introduction

This study leverages datasets from the New York City Taxi and Limousine Commission and the NYC Open Parking and Camera Violations database to forecast taxi fare prices. By evaluating various regression models, we aim to uncover how factors such as time, location, and trip distance influence fare predictions.

Datasets

The main dataset used for the investigation was obtained from the New York City (NYC) Taxi and Limousine Commission (TLC) website. The website displays the Trip Record Data of 4 different vehicle categories, arranged into separate datasets. For the purpose of a more directed investigation, we chose to focus on only the Yellow Taxi data during the analysis, opting to omit the other 3 vehicle categories. The Yellow Taxi dataset contains trip data on NYC's yellow taxi's can be hailed on the street or via certain apps in all boroughs of NYC. The other dataset used for the analysis was the NYC Open Parking and Camera Violations dataset obtained from the NYC Open Data website. The dataset contains various different parking and camera violations and arranged by the different New York Police Department (NYPD) precincts across NYC.

For a more focused investigation, we decided on restricting our analysis to a 6 month period of time, settling on the period of time between December 2023, and May 2024 as this is the most up to date period of time available for both datasets.

2 Pre-processing

To ensure that the data used in our analysis was fit for model building and visualisation, we applied various pre-processing techniques in order to improve data quality. While there were some common issues between both datasets, we ultimately employed different filtering and processing strategies for each. Table 1 provides a short summary of the total number of rows omitted in the date cleaning and filtering process for both datasets.

Dataset	Initial Size	Reduced Size
Taxi Dataset	20,169,467	16,708,696
Violations Dataset	18,910,967	3,187,236

Table 1: Summary of Dataset Size Before and After Pre-processing

2.1 Taxi Zone and NYPD Precinct Mapping

One major challenge in pre-processing was the mismatch in location types between the NYC Yellow Taxi and Open Parking and Camera Violations datasets. The Yellow Taxi dataset used 265 taxi zone IDs from a shape-file, while the Parking and Violations dataset used 77 NYPD precincts from another shape-file.

To resolve this, we created a mapping key that linked each taxi zone to a corresponding precinct. We did this by overlaying the taxi zone centroids on the NYPD precinct geo-dataframe and assigning precincts accordingly. The mapping dictionary was created with almost every taxi zone matched to a precinct. For the 3 taxi zones (1, 46, and 103) not matched, associated with airports, parks, and islands, we assigned them a unique precinct code of 0 to capture their trends while minimizing potential outlier effects.

2.2 NYC Yellow Taxi Data

On the TLC website, the Yellow Taxi datasets were arranged into separate files depending on month, which contained 19 different columns, with the most relevant ones for our research question being the Fare Amount and related costs and the date, time and location of pickup and drop-offs. For the purpose of our analysis remaining as relevant as possible, we decided to use the most recently available data over the span of a 6 month period, using data on taxi trips starting from December 2023 to May 2024. In total, the amount of rows in all of the datasets combined totaled to **20,169,467** rows.

Data Pre-processing

In the preparation for the dataset analysis, the following pre-processing steps were undertaken to ensure data quality and integrity:

- **Removal of Missing Values:** We decided to omit any rows containing completely missing values. While performing this step did not remove any rows from our dataset. It is an easy to perform standard practice in data pre-processing, eliminating any redundant rows and improving dataset quality.
- **Elimination of Negative Value Rows:** Through inspection of some rows in the dataset, we observed that there was a systematically random pattern of numerical data relating to the cost of the trip was input as negative values. While transformation steps could have been performed to potentially correct certain row values from being negative, we ultimately decided to omit these rows instead due to only a small amount of data being lost. In total, 1,993,708 rows were omitted because of this which was only 9.9% of the data.
- **Combination of Six Months of Data:** As a necessary part of the pre-processing, we combined the 6 separate datasets of the NYC yellow taxi data into a single .parquet file for easier use later on in our analysis. During this stage, we also checked that the schema for each dataset remained consistent with each other
- **Enforcement of Data Dictionary Rules, Value Checking and Logic:** As the next step

of the data cleaning and pre-processing, we enforced data dictionary rules upon the dataset, catching any unexpected values outside the expected range.:

  - The `RatecodeID` and `payment_type` values were restricted to the range of 1–6, as outlined in the dataset's documentation.
  - The `mta_tax` column was checked to ensure it was set to a constant value of 0.5, as per the dataset description.
  - The `fare_amount` column was filtered to be above $3, as per the enforced minimum taxi fare price (New York City Taxi & Limousine Commission, 2022).
  - Trip Distance: Additionally, trips with a distance below 0.25 miles were removed. Given that such short trips are unlikely to occur in reality due to the structure of the NYC taxi system, they were assumed to be input errors or trivially small outliers that may significantly affect the training of our model later on.

  This ended up reducing the size of the dataset to 17,776,484 rows.

- **Date and Time Conversion to ISO 8601 Format:** In order to prepare the date and time for later temporal analysis in conjunction with the Open Parking and Camera Violations dataset, we decided to select the widely recognised ISO 8601 format to ensure data type consistency. As a result of this, `pickup_datetime` and `dropoff_datetime` columns were converted to the ISO 8601 format.

Outlier Removal

After observing the mean, standard deviation, maximum and minimum values for each of the numerical columns, we decided that further outlier pruning would be necessary in order to prepare the dataset for model training. The 4 main features that were flagged were the `trip_distance, fare_amount, tip_amount` and the `cost_per_mile` having abnormally high values.

- **Trip Distance:** Starting with the trip distance, we observed a maximum value of 161,726.1 miles in our set of summary statistics. After plotting a binned histogram of all the data, with a bin size of 0.1, we observed that most of the data lay between the distances of 0.25 and 29 miles. Cutting out the significant outliers, reduced the dataset size by 5415 rows.
- **Fare Amount:** When plotting the binned distribution of fare amounts, we found an anomaly where over 600,000 rows were concentrated in the $70 range. Despite adjusting the bin width to 0.1, the issue persisted, with adjacent bins containing only 5,155 and 5,073 rows. To address this skew, we decided to sample 5,114 rows from the $70 bin and omit the rest, resulting in the removal of 615,068 rows.
- **Tip Amount:** Similarly to both trip distance and fare amount, the tip amount had unrealistically large tip amounts in some rows of the dataset, with a maximum of $999.99. After observing the distribution of the tip amounts, we that the majority tip amount was below $20. Hence, we decided to impose a $20 limit on the tip amount, removing a total of 26,180 rows.
- **Cost Per Mile:** Most importantly, for the outlier analysis of the cost per mile, we observed an initial mean of 8.11, with a standard deviation of 3.33 and maximum and minimum values of 0.1 and 464.286. Though the data had a right skew, the log transformation that we tried ended up giving the data a left skew instead. Ultimately, we decided upon simply cropping out values below $3 per mile and above $40 per mile as any values above and below seemed unrealistically large. This resulted in a reduction of 8871 rows.

Feature Engineering and Data Processing

Further steps were undertaken to enhance the quality of the dataset and refine the features used for modeling:

- **Removal of Pick-ups or Drop-offs outside NYC:** After observing the `LocationID` key of the taxi zones and the corresponding locations they represent, we decided to omit any data points containing the `LocationID` of 264 or 265 as they corressponded to "Unknown" locations as well as "Outside of NYC" trips respectively. This was due to the focus of our analysis remaining within NYC.
- **Omission of Irrelevant Columns:** During the feature selection stage of our dataset processing, we determined that some columns deemed irrelevant to the prediction task and were dropped to reduce dimensionality and enhance model efficiency. In total, features were dropped at this stage being, the `VendorID, RatecodeID, Airport_fee, mta_tax, improvement_surcharge`, and the `store_and_fwd_flag`, reducing the total number of columns to 13.
- **Feature Engineering:** As we were primarily interested in the cost per mile of each trip, we decided to create a new column, combining the information of both the trip distance and the fare amount in a more meaningful manner. It is also easier to observe how the variability in other areas may affect this, rather than having two separate features.

These pre-processing steps were crucial to ensure the data's integrity, providing a solid foundation for the subsequent predictive modeling phase.

2.3 NYC Open Parking and Camera Violations

The NYC Open Parking and Camera Violations, obtained from the NYC Open Data website, contains 19 different columns containing data such as the location of the violation, the fine amount, the violation type as well as the time of the violation. In order to match the 6 month period of data selected for our taxi analysis, we opted to match the same period of time, between December 2023 and May 2024, in order to maintain a relevant analysis. The dataset obtained from the NYC Open data website could also be queried within the website and provided an easy way of omitting columns that were not necessary for our analysis. In total, 7 of the 19 columns were omitted before downloading the dataset through an API link with a final total of 18,910,967 rows in the landing stream.

Data Pre-processing

In a similar way to the NYC Taxi data, the NYC Open Parking and Camera Violations dataset was cleaned and preprocessed in order to prepare the dataset for joining.

- **Removal of Missing Values:** In a similar way to the taxi dataset, we had to remove missing values from crucial rows such as `Issue Date` and `Violation Time` to ensure non-redundancy of data. In total 210 rows were removed.
- **Conversion of Date and Time into the ISO 8601 Format** Again, similarly to the taxi dataset, we had to convert the date and time of the parking or camera violation into a standardised format. The task for this dataset however, involved the combination of both the `Issue Date` and `Violation Time` columns into one, as well as the conversion of the 12 hour time format adopted by the `Violation Time` column into 24 hour time.
- **Removal of Data Outside our Date Range:** Due to difficulty downloading specific date and time ranges from the NYC Open Data Website, we had to download the data for both the years

of 2023 and 2024 and after formatting the date and time into a uniform datatype, we were able to reduce the data down to the range of December 2023 to May 2024 leaving us with 5,624,335 rows of data.

- **Removal of Invalid Precinct Numbers:** When inspecting the dataset, we found that there were a surprisingly large number of non-existent precinct numbers present. This may have been due to inputation error and were subsequently cross referenced with the NYPD precinct lookup table (NYC Open Data, 2024). This led to a reduction of 2,437,099 data rows.

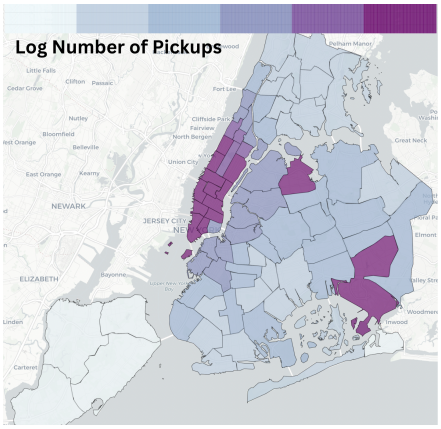
Overall, the total number of rows left for analysis was 3,187,236

3 Analysis and Visualisation

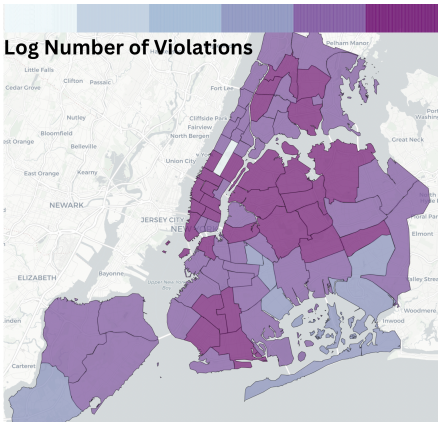
In this section, we analyze the datasets to identify trends and relationships that can provide insights into the factors influencing Cost Per Mile.

3.1 Geospatial Distribution

In analyzing the two heatmaps (Figure 1), we observed that taxi pickups were concentrated in precincts 18, 19, 20, 14, and 6, while violations were more evenly spread across the city. We interpreted the concentrated pickups in these precincts as likely high-demand areas, which may result in higher fare amounts due to traffic, frequent trips, and increased demand. In contrast, the spread of violations across various precincts introduced potential variability in predicting the cost per mile, as violations could lead to delays or fines, indirectly affecting pricing. This helped inform us by highlighting areas with consistent fares and others where violations could complicate predictions.



(a) Heatmap of Log Number of Taxi Pickups



(b) Heatmap of Log Number of Parking and Camera Violations

Figure 1: Comparison of Taxi Pickups and Violations

3.2 Temporal Analysis of Taxi Data

As we can see in Figure 2, the average cost per mile for a taxi trip usually falls between $5.8 and $10 and can vary a great deal depending on which day of the week it is. While the cost per mile for most days typically peaks at around 12-2pm for most days, it is interesting to note that on Saturdays, the most expensive times usually fall between 4-10pm. This may be attributed to people being more active at night on the weekends.

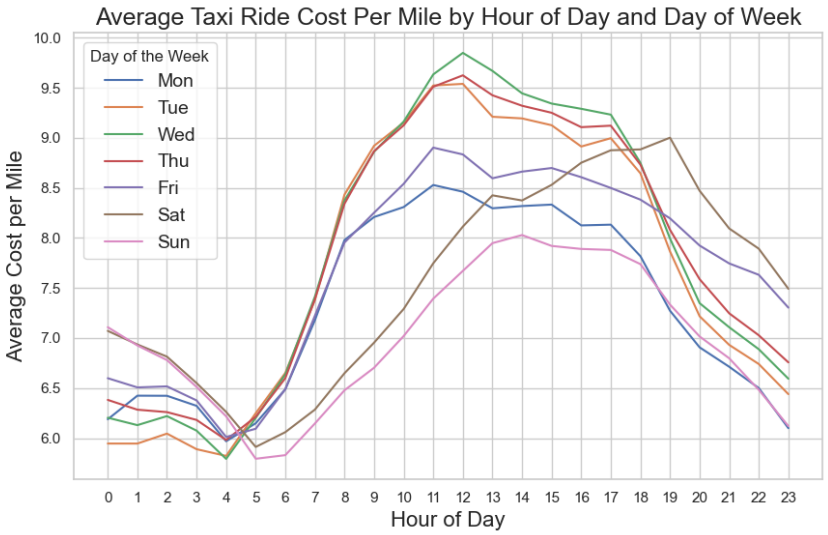


Figure 2: Distribution of the Average Cost Per Mile Throughout the Day

While the graph does capture the expected trends for the average cost per mile of taxi ride, by observing Figure 3, we are able to see a distribution of the frequency of taxi rides at different times throughout the day and different days of the week. Again as expected, we see a comparatively large proportion of taxi rides in the early hours of the morning, between 12-4am, falling on Saturday or Sunday mornings compared to a significantly lesser amount on the weekday mornings. Interestingly, we see a significant drop in both taxi ride frequency and average cost per mile at around the 4-5am time, suggesting that there may be an excess in taxi drivers or deficiency in passengers, assuming that a higher demand for taxis is correlated with a higher potential cost per mile. When comparing both figure 2 and 3, it also interesting to see that while the highest frequency of taxi rides occurs during 5-7pm, the most expensive time to ride is in the middle of the day, between 12-2pm.

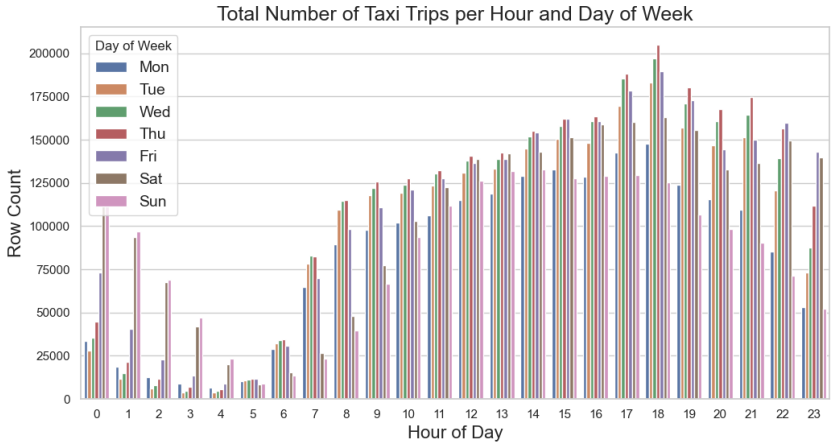


Figure 3: Distribution of Taxi Trips by Hour and Day of Week

3.3 Temporal Analysis of Parking and Camera Violations

In Figure 3, we analyzed the distribution of violations over a 24-hour period and observed an exceedingly low number of violations between 9 PM and 5 AM. This drop in violations during late-night hours introduces a potential bias in our analysis, as fewer data points may affect the reliability of predictions for trips occurring during these hours. The limited violation data could lead to underrepresentation of key factors, such as delays or penalties, that influence fare amounts or cost per mile. This trend suggests that trips during this period may behave differently from daytime trips, and further analysis would be required to address this imbalance.

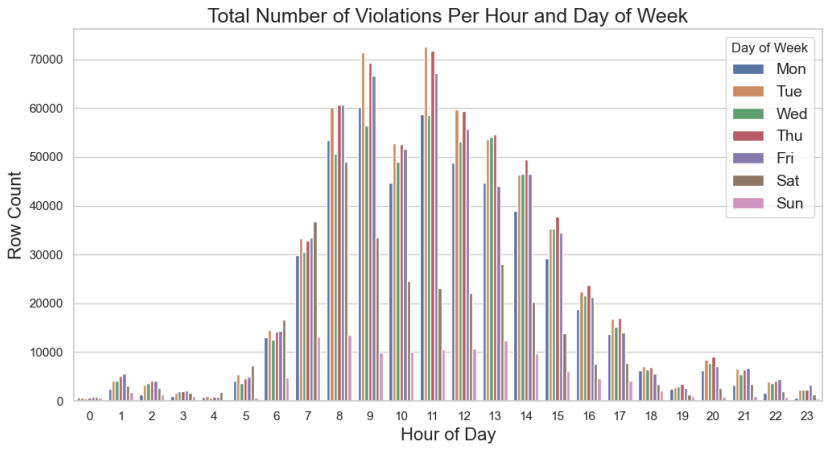


Figure 4: Distribution of Violations Issued by Hour and Day of Week

4 Modelling

To predict the cost per mile, we focused on predicting fare amount due to linear model constraints. We evaluated Linear Regression, Lasso Regression, Ridge Regression, and ElasticNet Regression, using features like trip distance, precinct, day of week, and payment type.

After splitting the data and performing grid search for hyperparameter tuning, ElasticNet Regression emerged as the best model, with an MSE of 10.7613 and an R2 of 0.9228. It slightly outperformed Lasso and Ridge regressions, which had similar R2 scores but higher MSE values. Linear Regression, while simple, had a comparable R2 (0.9226) but a higher MSE (10.7808).

The final evaluation revealed an RMSE of 3.2804 and a cross-validated RMSE of 3.3388 ($\pm$0.7191), demonstrating ElasticNet's robustness and effective regularization in minimizing overfitting and enhancing generalization.

Model	MSE	R2	Best Parameters	RMSE	Cross-validated RMSE
Linear Regression	10.7808	0.9226	None	3.2804	-
Lasso Regression	10.7630	0.9228	`alpha: 0.001`	3.2792	-
Ridge Regression	10.7624	0.9228	`alpha: 10`	3.2792	-
ElasticNet	10.7613	0.9228	`alpha: 0.001, l1_ratio: 0.5`	3.2804	3.3388 (+/- 0.7191)

Table 2: Comparison of Model Performance

5 Discussion and Recommendations

Figure 5 shows the learning curves for our model, depicting the Root Mean Squared Error (RMSE) for both the training and validation sets as a function of the training set size. The training RMSE starts relatively high and decreases as the training set size increases, indicating that the model benefits from more data. The validation RMSE also decreases but at a slower rate, which shows improved generalization with a larger dataset.

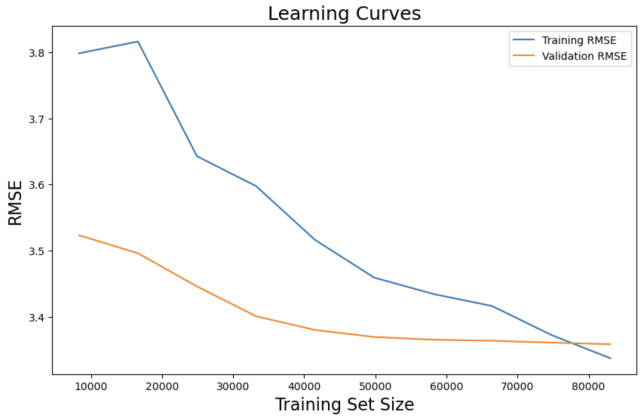


Figure 5: Learning Curve of ElasticNet Model

The gap between the training and validation RMSEs narrows as the dataset grows, suggesting that adding more data could further enhance model performance. However, the consistent reduction in both curves points to a model that is neither overfitting nor underfitting, affirming that our model training process is on track. Further data may yield marginal improvements, but the decreasing trend confirms that our model is learning effectively.

6 Conclusion

This analysis utilized taxi fare and parking violations datasets to predict the cost per mile of taxi rides. We compared Linear Regression, Lasso Regression, Ridge Regression, and ElasticNet models. Lasso and Ridge Regression performed best, both achieving an R^2 of 0.9228 and an MSE of around 10.76. Lasso Regression had the lowest RMSE at 3.2792. ElasticNet also showed good performance, with a cross-validated RMSE of 3.3388 ($\pm$0.7191).

The models highlighted that time of day, location, and trip distance are key factors in predicting taxi fares. While the analysis was successful, further exploration into the impact of parking violations could provide additional insights. Overall, the models effectively demonstrated the importance of temporal and spatial variables in fare prediction.

References

[1] Police Precincts. (2024, August 19). Retrieved from NYC Open Data website: https://data.cityofnewyork.us/Public-Safety/Police-Precincts/78dh-3ptz

[2] Taxi Fare - TLC. (2022). Retrieved August 31, 2024, from Nyc.gov website: https://home.nyc.gov/site/tlc/passengers/taxi-fare.page#:~:text=%243.00%20initial%20charge.

[3] Elastic net regularization. (2021, March 19). Retrieved from Wikipedia website: https://en.wikipedia.org/wiki/Elastic_net_regularization

[4] Wikipedia Contributors. (2018, November 21). Linear regression. Retrieved from Wikipedia website: https://en.wikipedia.org/wiki/Linear_regression