NYC Taxi Research Project

Navigating NYC: Taxi Fare Insights & Travel Tips for Tourists

Jenny Mai
Github Repository

September 29, 2024

1 Introduction

In the bustling streets of New York City, where every minute counts and every block is a new adventure, the humble yellow taxi is more than just a mode of transportation—it's an essential part of the city's pulse. For tourists navigating this urban jungle, understanding the patterns behind taxi fares can be the difference between a seamless journey and an unexpected detour in both time and cost. This report dives deep into the intricate world of NYC taxi data, offering comprehensive analysis and actionable insights that will empower visitors to make informed decisions, ensuring their New York experience is as smooth as the ride itself.

2 Preprocessing

2.1 Descriptive Statistics and Outlier Analysis

Trip Distance: The initial dataset showed a mean trip distance of approximately 4.47 miles, with a high standard deviation indicating the presence of extreme outliers. The majority of trips (median: 1.8 miles) are short, fitting within typical NYC taxi operations. For relevant analysis, trips below 0.2 miles and above 75 miles were excluded as outliers, ensuring that the data reflects realistic taxi rides. This threshold was chosen based on typical NYC taxi trip distances, where extremely short trips are likely misreported, and very long trips are rare and often associated with exceptional circumstances.

Fare Amount: The mean fare was $8.62, with erroneous and extreme values such as a negative fare and a maximum of $5901.74. A more representative median of $12.8 was used to guide outlier analysis. After outlier removal, only fare amounts between $3 and $200 are retained. This range was selected to exclude erroneous data points, such as negative fares or excessively high amounts, which are likely due to data entry errors or non-standard trips.

Passenger Count: The average passenger count was 1.38, with outliers ranging from 0 to 9. To maintain data integrity, trips with zero passengers or counts exceeding the legal taxi capacity (6 passengers) were excluded from the analysis. This decision is based on the assumption that trips recorded with zero passengers are erroneous, and that taxis exceeding their legal capacity do not represent typical rides. Excluding these helps maintain the integrity and applicability of the analysis.

2.2 Missing Values and Data Cleaning

Handling Missing Data: With a robust dataset of approximately 19.3 million records, missing data was handled by removing rows with missing values, resulting in a loss of about 500,000 entries, primarily in non-critical columns.

Outlier Detection & Data Cleaning: Outliers and invalid values were removed to ensure the analysis reflects realistic and reliable data.

2.3 Data Transformation

Feature Scaling: Numerical features were standardized using the `StandardScaler` to ensure consistency in model inputs, particularly for algorithms sensitive to feature magnitude. This step is critical to achieving accurate and reliable predictions. Standardizing features is crucial for models like linear regression, which assume that features are normally distributed and have similar scales. This step helps prevent features with larger magnitudes from dominating the model training process.

Log Transformation: Right-skewed numeric features, such as trip duration and distance, were log-transformed to reduce skewness in the data, making it more suitable for linear models. This transformation helps in stabilizing variance and improving the model's ability to learn from the data.

2.4 Feature Engineering

Temporal Features: Key time-based features, including trip duration, pickup/dropoff hours, and day of the week, were extracted to identify patterns relevant to tourists, such as optimal travel times.

Geospatial Features: Features indicating whether a trip started or ended at key locations like airports or tourist attractions were created. This allows for specific insights into how trips associated with these locations differ, providing valuable information for tourists planning their travel.

Environmental Features: Weather data, including daily average temperature and precipitation, was integrated to analyze the influence of environmental conditions on taxi trip outcomes. This helps tourists understand how weather might affect their travel plans and costs. The weather data can be accessed [here].

3 Analysis and Geospatial Visualisation

3.1 Geospatial Analysis

Average Fare Amount by Drop-off Location: Based on Figure 1, higher fares are notably concentrated in specific areas, particularly in the southern parts of Staten Island, where darker shades indicate more expensive drop-offs. This suggests that trips ending in these locations tend to be longer, potentially from central Manhattan or involving tolls, leading to increased travel costs.

Trip Density by Pickup Location: Figure 2 shows that taxi pickups are most concentrated in central Manhattan, particularly around business districts and major transit hubs, reflecting high demand in these areas. Smaller but significant pickup densities are also seen at key locations like JFK Airport in Queens, highlighting these areas as major transit points for both locals and tourists.

Peak Pickup Hours by Location: Figure 3 indicates that peak taxi activity varies by location. For example, Manhattan's business districts experience peak pickups from 3 PM to midnight, coinciding with work commutes and evening events. In contrast, Brooklyn's nightlife areas, such as Williamsburg,

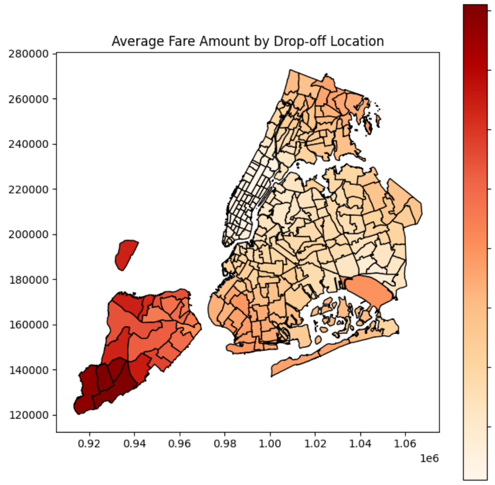

Figure 1: Mean Fare Amount by Drop-off Location

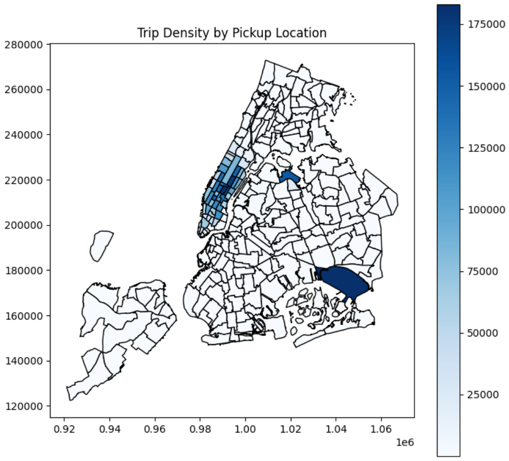

Figure 2: Trip Density by Pickup Location

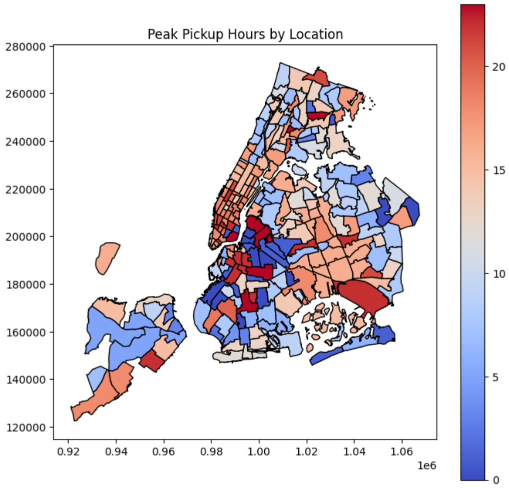

Figure 3: Peak Pickup Hours by Location

peak between 10 PM and 3 AM. Tourists should consider these peak times when planning their outings to avoid higher fares and longer wait times.

3.2 Temporal Analysis

Average Trip Duration by Hour of Day: Trip durations tend to peak between 4 PM and 6 PM, coinciding with rush hours, where increased traffic leads to longer travel times. Conversely, shorter trip durations are observed between midnight and 6 AM, reflecting lighter traffic. Tourists should plan their travel during these off-peak hours to minimize travel time.

Trip Density by Time of Day and Day of the Week: The heatmap indicates that taxi trip density is highest from Wednesday to Saturday between 5 PM and 8 PM, particularly in tourist-heavy areas. This suggests a strong demand for taxis during these times, and tourists should plan accordingly to avoid peak-hour congestion.

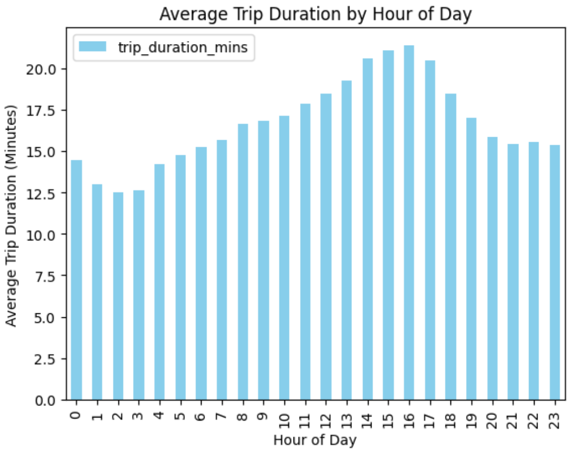

Figure 4: Average Trip Duration by Hour

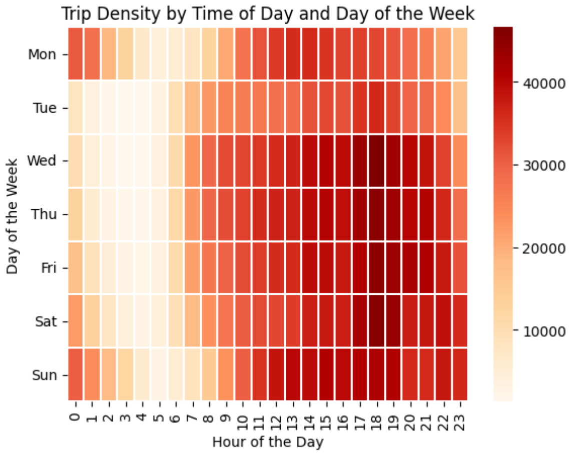

Figure 5: Daily & Weekly Trip Density

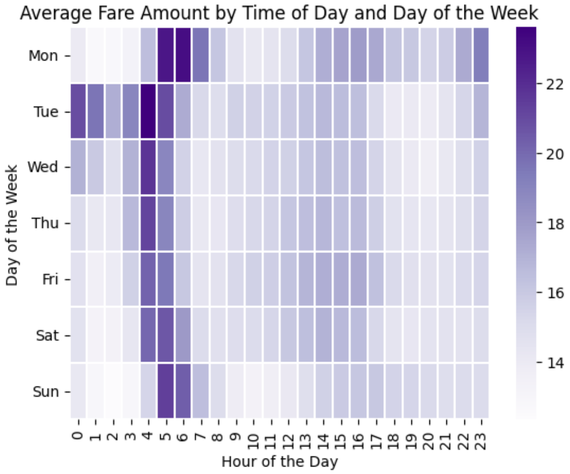

Figure 6: Daily & Weekly Fare Amount

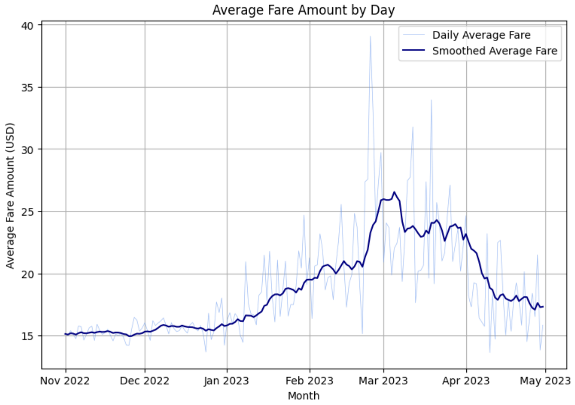

Figure 7: Mean Fare Amount over Time

Fare Amount by Time of Day and Day of the Week: Higher fares are observed during the early morning hours (4-7 AM) on Mondays and Tuesdays, likely due to airport transfers and lower taxi availability. Tourists catching early flights during these times should anticipate higher travel costs and consider booking their taxis in advance.

Seasonal Fare Increase: There is a notable increase in average taxi fares from late January 2023, peaking in mid-March 2023. This trend likely corresponds to seasonal demand, such as winter tourism and events, affecting overall taxi fares. Tourists visiting during these months should budget for higher taxi expenses, particularly around major events in March.

3.3 Distribution Analysis

Fare Amount Distribution: The fare distribution is highly skewed, with most fares between $5 and $20, but with peaks for airport flat rates and a few long-distance trips. This suggests that while most trips are affordable, certain types of trips, such as airport transfers, can significantly increase costs.

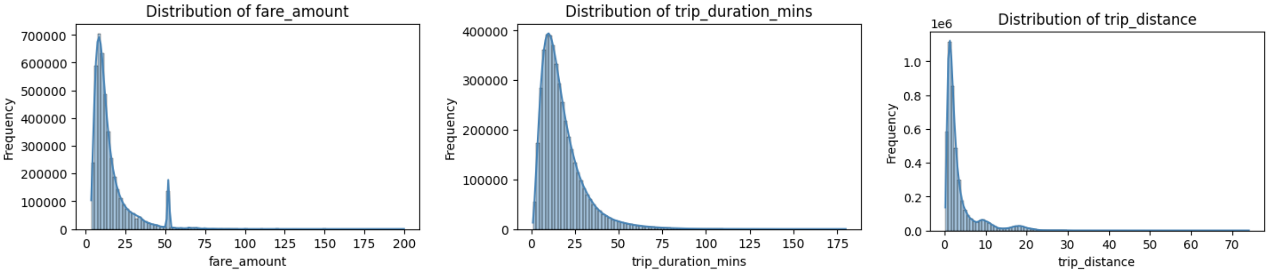

Figure 8: Distribution of `fare_amount`, `trip_duration_mins`, and `trip_distance`

Trip Duration Distribution: Most trips last between 5 and 30 minutes, with fewer longer trips. This suggests that tourists can expect relatively short trips to most destinations within the city.

Trip Distance Distribution: The majority of trips cover 1-7 miles, with longer trips being less common but present. Tourists should expect that most of their taxi rides will be short and relatively inexpensive unless traveling to distant locations like airports.

Weather Variables: The average temperature and precipitation data indicate typical winter and spring weather patterns, with colder temperatures and minimal precipitation. This suggests that weather conditions during these seasons have a limited impact on taxi usage and fares.

3.4 Correlation Analysis

Fare Amount & Trip Distance: A strong positive correlation (0.93) between fare_amount and trip_distance confirms that longer trips result in higher fares, a key consideration for tourists planning their travel budgets.

Total Amount & Fare Amount: The near-perfect correlation (0.98) between total_amount and fare_amount highlights that additional fees and surcharges are consistent across trips, ensuring that fare predictions are reliable.

Trip Duration & Trip Distance: The positive correlation (0.76) between trip duration and distance. This correlation is expected, as longer distances generally result in longer trip durations. Understanding this relationship is crucial for predicting not just costs but also the time commitment involved in a trip, which is particularly important for tourists planning their schedules.

Is Airport Trip & Distance-Time Interaction: A strong correlation (0.88) between airport trips and the distance-time interaction underscores the longer, more expensive nature of these trips, which is critical information for tourists planning airport transfers.

Congestion Surcharge & Total Amount: The moderate correlation (0.74) indicates that congestion surcharges contribute significantly to total fare amounts, particularly during peak traffic times. Tourists should be aware of these additional costs when traveling during busy periods.

4 Modelling

4.1 Prediction Variables Assumption

To improve the user experience, particularly for tourists unfamiliar with fare prediction models, it is recommended to develop a user-friendly application. This app would allow users to input trip details and automatically estimate necessary variables for fare prediction. The app should:

Perform Routing Calculations: Calculate trip distance and duration based on start and end locations (PULocationID, DOLocationID), considering traffic and time of day.

Generate Fare Components: Identify the appropriate RatecodeID (e.g., standard, JFK flat rate), and include surcharges like mta_tax, tolls, improvement_surcharge, congestion_surcharge, and airport_fee. Flag trips involving airports or major attractions.

Incorporate Calendar-Based Variables: Adjust for holidays and significant events.

Leverage Historic Weather Data: Incorporate average temperature for the time of year using historical data.

4.2 Performance Metrics

The models were evaluated using Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) to assess their predictive accuracy because RMSE emphasizes larger errors, MAE provides a straightforward average of all errors, and R^2 indicates the proportion of variance explained by the model:

Model	**RMSE**	**MAE**	**R^2**
Linear Regression	7.46	3.82	0.78
Lasso Regression	7.58	3.94	0.77
Random Forest Regressor	7.97	4.65	0.78
Gradient Boosted Trees (GBT)	3.96	1.88	0.94

Table 1: Performance metrics for different regression models.

Analysis: Gradient Boosted Trees (GBT) demonstrated the best performance, achieving the lowest RMSE and MAE, with a high R^2 value of 0.94. GBT was chosen for its ability to handle complex, non-linear interactions between features, which is crucial in capturing the variability in taxi fare data. This model effectively captures complex interactions and patterns in the data, making it the most reliable for predicting taxi fares, especially for tourists who need precise estimates. However, the model's predictions are based on historical data, which may not fully account for future changes in predictors. Additionally, these models may suffer from overfitting, specifically when the training data includes noise or outliers that do not represent typical taxi trips. This could lead to predictions that are less accurate for new or unseen data.

4.3 Feature Analysis and Model Interpretation

Linear & Lasso Regression: Both models identified trip_distance and RateCodeID as the most significant predictors of fare amount, which aligns with intuitive expectations that longer trips cost more. However, Lasso's feature reduction process highlighted the trade-off between simplicity and predictive power, particularly with variables like PULocationID, improvement_surcharge, is_holiday_season, and avg_temp.

Random Forest Regressor: This model emphasized the importance of trip_distance, trip_duration_mins, and the distance_time_interaction feature. While it captured the interplay between distance and time well, the inclusion of numerous features with minimal impact slightly inflated error metrics. These features were expected to be influential based on domain knowledge—trip distance and duration naturally drive fare amounts. The inclusion of interaction terms captures the combined effect of these variables.

Gradient Boosted Trees (GBT): GBT's iterative learning approach allowed it to balance the influence of various features, achieving superior performance by refining predictions at each stage. This model's ability to integrate and adjust for errors incrementally explains its dominance in predictive accuracy.

4.4 Diagnostic Evaluation

Residuals vs. Predictions: GBT showed the least heteroscedasticity, indicating that it captures variance uniformly across predictions. Linear and Lasso Regression models displayed patterns of heteroscedasticity, reflecting their limitations in handling data variability.

Residual Distributions: GBT's residuals were tightly centered, indicating its superior fit to the data. Other models exhibited right-skewed residuals with longer tails, reflecting larger prediction errors, particularly in outlier cases.

Q-Q Plots: GBT's residuals aligned closely with the normal distribution, particularly in the middle quantiles, further confirming its ability to handle the data's non-linear aspects more effectively than the other models.

4.5 Comparative Insights

Gradient Boosted Trees emerged as the most effective model for predicting taxi fares in NYC, particularly for tourists. Its ability to capture non-linear relationships and adjust predictions iteratively makes it the most reliable tool for generating accurate fare estimates, crucial for tourists planning their budgets. It is also important to note that the predictions generated by the GBT model can exhibit variability due to fluctuations in underlying factors such as weather, traffic, and taxi availability, which are not perfectly predictable. This inherent variability means that while the model is highly accurate on average, individual predictions may occasionally deviate from actual fares, particularly in cases where the input features deviate from the typical patterns seen in the training data.

Linear and Lasso Regression models, while interpretable and simpler, lack the flexibility to capture the complexities of NYC taxi fare data, leading to higher errors and less accurate predictions.

Random Forest, despite its ability to model non-linear relationships, struggled with variance and complexity, resulting in less precise predictions compared to GBT.

5 Recommendations

Based on the analysis conducted, here are tailored recommendations for tourists using NYC yellow taxis:

1. **Plan for Higher Fares During Early Morning Airport Transfers (4 AM to 7 AM on Weekdays):** If you're planning to catch an early flight on a Monday or Tuesday, be prepared for higher taxi fares. This time frame sees the highest average fare amounts, likely due to longer trips to airports and increased demand. Consider booking your taxi in advance or exploring other transportation options during these hours to avoid peak pricing.
2. **Leverage Mid-Morning (8 AM to 11 AM) for Cheaper Rides:** To save on taxi fares, plan your trips during the mid-morning hours between 8 AM and 11 AM. Fares are generally lower during this time, and traffic is less intense compared to rush hours, providing a balance between cost and travel time.

3. **Avoid High-Density Times Like Friday and Saturday Evenings (5 PM to 8 PM):** If possible, avoid taking taxis on Fridays and Saturdays between 5 PM and 8 PM, as this period experiences the highest trip density. This could lead to longer wait times, higher fares, and potentially slower trips due to traffic congestion. Instead, consider using alternative transport options like subways or buses during these peak hours.

4. **Consider Off-Peak Times for Leisure Activities:** For visits to tourist attractions, plan your trips outside the peak hours of 5 PM to 8 PM to avoid higher fares and crowded conditions. Early afternoons (1 PM to 3 PM) on weekdays might be a better time for sightseeing trips, offering a more relaxed experience with potentially lower fares.

5. **Strategic Planning for Weekend Outings (Saturday and Sunday Afternoons):** If you're planning to explore NYC on weekends, aim to travel during the early afternoon on Saturdays and Sundays. Taxi fares are generally stable during this time.

6. **Use Taxis During Late Night Hours (Midnight to 3 AM) for Clubbing or Late Dinners:** For those enjoying NYC's nightlife, consider using taxis between midnight and 3 AM. Fares are relatively low during these hours, and traffic is minimal, ensuring a quicker and more affordable ride back to your hotel or accommodation.

7. **Prepare for Slightly Higher Fares Around Major Holidays and Events (e.g., Presidents' Day):** If your visit coincides with major holidays like Presidents' Day or large-scale city events or peak periods like March, be aware that taxi fares might be higher than usual. Planning your transportation needs during these times can help you budget more effectively and avoid unexpected costs.

8. **Utilize Gradient Boosted Trees for Fare Predictions with Caution:** While GBT provides accurate fare estimates, the inherent variability in taxi fares due to factors like traffic and demand means that predictions should be treated as approximate. The recommendation to use GBT predictions as a guide rather than an exact figure is based on the model's high accuracy but also on the understanding that real-world conditions introduce uncertainty. Additionally, the model predicts the fare amount only, so other fees should be added to estimate the total travel cost. Given the inherent variability in taxi fares due to traffic, demand, and other factors, it is advisable to use it as a starting point for planning your travel budget, with an understanding that actual fares may vary.

6 Conclusion

This report offers key insights into NYC taxi data, aimed at helping tourists make informed travel decisions. Through detailed analysis of trip distances, fare amounts, and temporal as well as geospatial patterns, we identified the primary factors affecting taxi fares and trip durations.

Preprocessing steps ensured the data was accurate and relevant, while visualizations highlighted critical patterns, such as peak times and fare variations by location. The Gradient Boosted Trees (GBT) model proved the most effective for fare prediction, though its estimates should be used as guidelines due to variability in real-world conditions.

In summary, tourists can use the findings in this report to optimize their taxi experiences in New York City, avoiding unnecessary expenses and ensuring smoother, more cost-effective journeys.

References

[1] National Oceanic and Atmospheric Administration (NOAA), *Climate Data Online*, Available at: `https://www.weather.gov/wrh/Climate?wfo=okx`.

[2] Scikit-learn, *Scikit-learn: Machine Learning in Python*, Available at: `https://scikit-learn.org/stable/documentation.html`.

[3] New York City Taxi and Limousine Commission, *NYC Taxi and Limousine Service Trip Record Data*, Available at: `https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page`.

[4] New York City Taxi and Limousine Commission, *NYC Yellow Taxi Trip Record Data Dictionary*, Available at: `https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf` (Accessed: 24 August 2024).

[5] The University of Melbourne, *COMP30027: Machine Learning*, School of Computing and Information Systems, University of Melbourne, 2024.