New York Taxi Revenue Optimization: End-to-End Predictive Modeling A/B Testing

Comprehensive EDA, Statistical Analysis & ML Prediction

Saad Ahmed Qadeer
Data Scientist

Abstract

This comprehensive project analyzes NYC Taxi & Limousine Commission (TLC) trip data to uncover patterns in taxi usage, identify revenue drivers, and build predictive models for fare amounts and customer tipping behavior. We conducted end-to-end analysis including exploratory data analysis, statistical hypothesis testing, regression modeling, and machine learning classification. Our findings provide actionable insights for optimizing taxi driver revenue and understanding customer behavior patterns.

Contents

1 Project Overview

Dataset: NYC Taxi & Limousine Commission trip records from 2017
Total Records: 22,699 taxi trips
Structure: Analysis divided into 5 distinct phases
Analysis Tools: Python (pandas, numpy, matplotlib, seaborn, scipy, scikit-learn, XG-Boost)

Key Variables

- **Trip characteristics:** distance, duration, pickup/dropoff times
- **Financial data:** fare amount, tip amount, tolls, total amount
- **Customer data:** passenger count, payment type
- **Operational data:** vendor ID, rate code

---

2 Phase 1: Initial Data Exploration

Reference: Notebook 01_NYC_Taxi_Exploratory_Analysis.ipynb

2.1 Objective

We started by loading and inspecting the dataset to understand its structure, identify data quality issues, and determine which variables would be most useful for analysis.

2.2 What We Did

1. Data Loading and Structure Inspection

We imported the dataset using pandas and performed initial inspection:

- Loaded 22,699 trip records with 18 variables.
- Examined data types: most variables were numeric (int64, float64), with 2 datetime fields.
- Checked for missing values: **No null values found** - the dataset was complete.

2. Initial Data Quality Assessment

We used `df.describe()` to examine statistical distributions and found several important patterns:

Trip Distance:

- Average trip: 2.9 miles
- Most trips (25-75%): 1.0 - 3.1 miles

- Maximum trip: 33.96 miles (outlier)
- **Finding:** Most taxi trips are short urban journeys.

Fare Amount:

- Average fare: $13.03
- Typical range (25-75%): $6.50 - $17.80
- Maximum fare: $999.99
- Minimum fare: -$120.00 (data quality issue)
- **Finding:** Negative fares indicate data entry errors that need investigation.

2.3 Key Insights

- **Data Completeness:** The dataset is complete with no missing values, making it reliable for analysis.
- **Data Quality Issues:** Identified negative fare amounts (120 trips) and extremely high fares (> $500).
- **Variable Relationships:** Both vendors have nearly identical average fares ($16.30 vs $16.32).
- **Tipping Patterns:** Groups of 2 passengers tip slightly more on average ($2.83) than solo riders ($2.71).

Note: As this phase focused on data structure and tabular statistics, no visualizations were generated. All graphical analysis begins in Phase 2.

3 Phase 2: Exploratory Data Analysis & Visualization

Reference: Notebook 02_NYC_Taxi_EDA_Visualization.ipynb

3.1 Objective

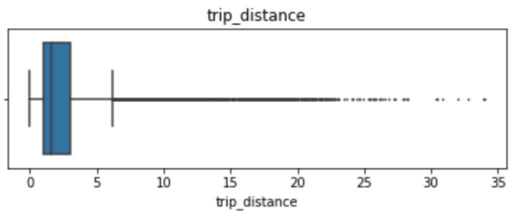
We conducted deeper exploratory analysis to understand distributions, identify outliers, examine temporal patterns, and create visualizations to communicate findings to stakeholders.

3.2 Distribution and Outlier Analysis

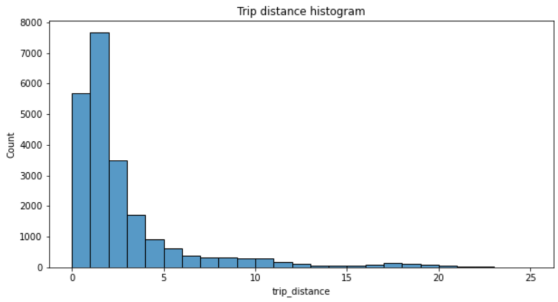
We analyzed the three main numerical components of a taxi trip: Distance, Total Amount, and Tips.

1. Trip Distance

Most trips are short urban journeys. The distribution is right-skewed with a mode around 1 mile.



(a) Box Plot: Outlier Detection

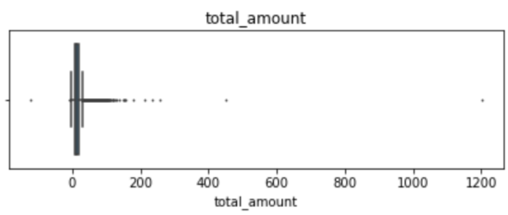


(b) Histogram: Frequency Distribution

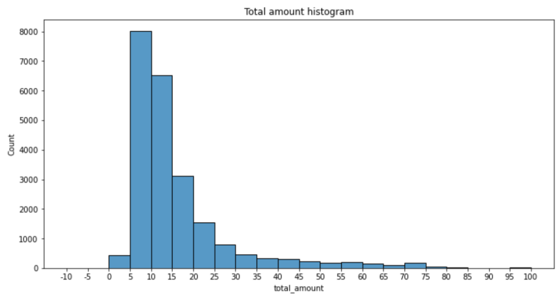
Figure 1: **Trip Distance Analysis.** The data is right-skewed with most trips under 5 miles. Outliers are visible in the box plot.

2. Total Fare Amount

Similar to distance, fares are right-skewed. We identified significant high-fare outliers, which may correspond to trips to JFK airport or luxury services.



(a) Box Plot: Outlier Detection

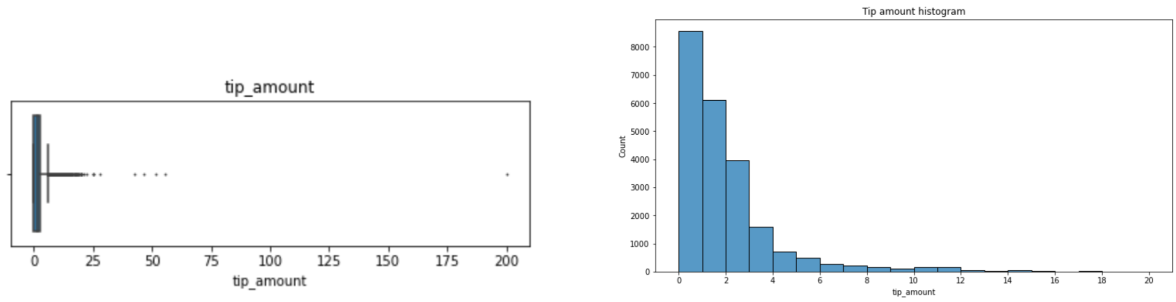


(b) Histogram: Frequency Distribution

Figure 2: **Total Fare Amount Analysis.** The box plot identifies significant high-fare outliers.

3. Tip Amount

The median tip is around \$2.00, with the vast majority of tips falling between \$0 and \$3.

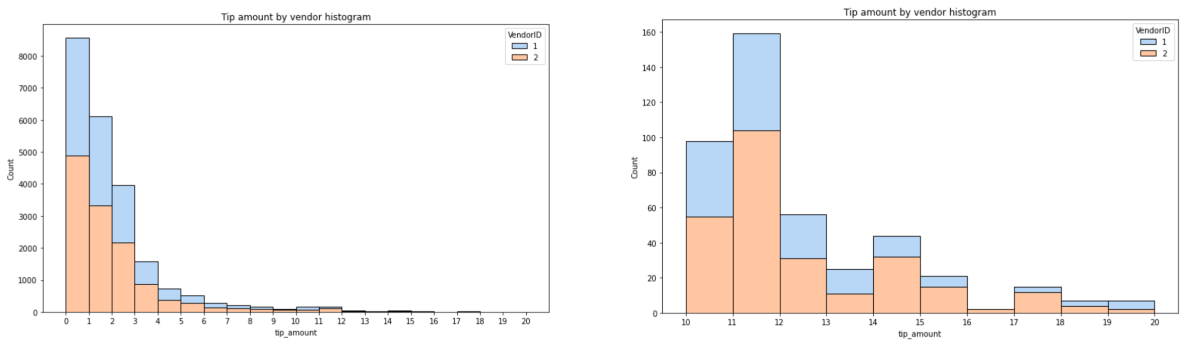


(a) Box Plot: Outlier Detection (b) Histogram: Frequency Distribution

Figure 3: **Tip Amount Analysis.** The median tip is around \$2, with the vast majority of tips falling between \$0 and \$3.

3.3 Vendor and Tipping Analysis

We compared the two vendors to check for discrepancies in service or tipping behavior. We found that tipping distributions are remarkably consistent across both vendors.



(a) All Tips by Vendor (b) High-Value Tips (> \$10)

Figure 4: **Vendor Comparison.** Tipping distributions are consistent across both vendors, even among high-value tippers.

Passenger Count Impact

We analyzed if group size affects generosity. Groups of 2 passengers appear to tip slightly more on average.

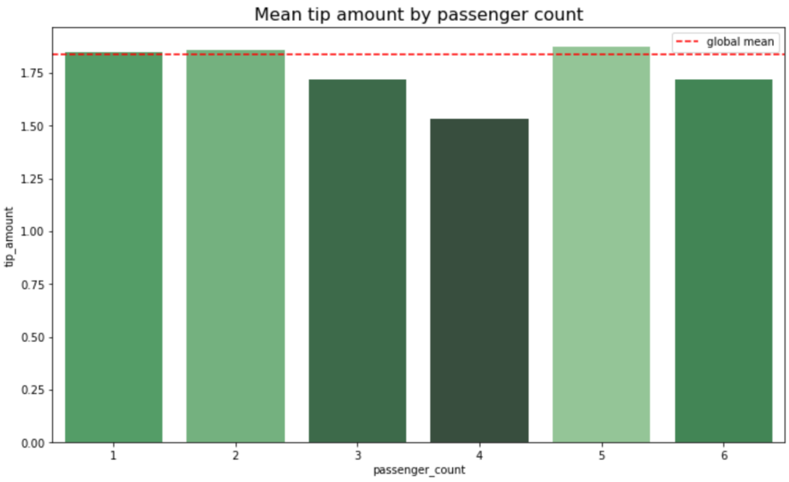
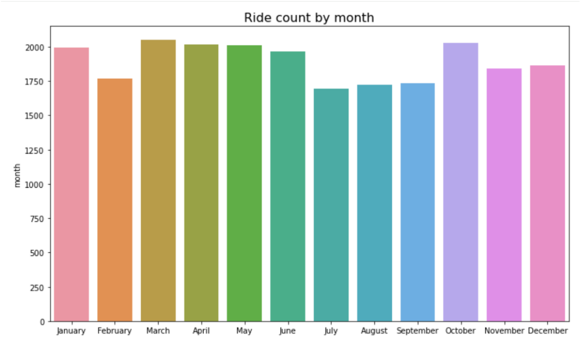


Figure 5: **Mean Tips by Passenger Count.** Comparison of average tip amounts for different group sizes.

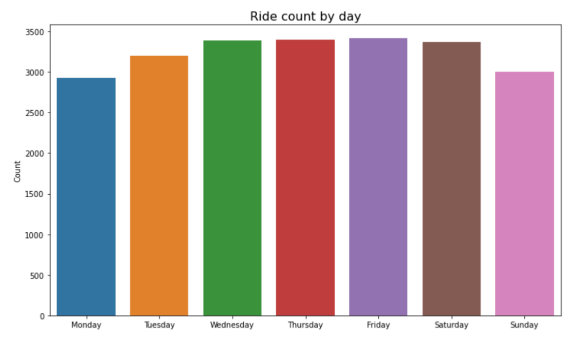
3.4 Temporal Patterns

We examined ride volume and revenue trends across different time scales.

- **Daily:** Peaks occur during rush hours (7-9 AM, 5-7 PM).
- **Weekly:** Weekdays show higher volume, while weekends show slightly longer average trips.

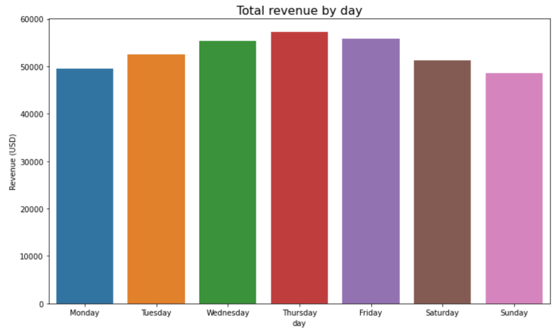


(a) Total Rides by Month

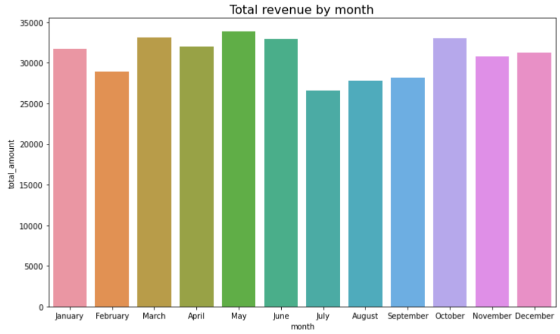


(b) Total Rides by Day of Week

Figure 6: **Ride Volume Analysis.** Seasonal and weekly patterns in taxi demand.



(a) Total Revenue by Day of Week

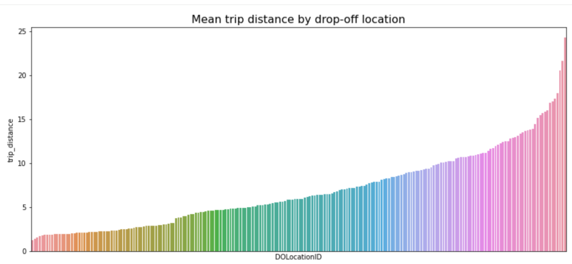


(b) Total Revenue by Month

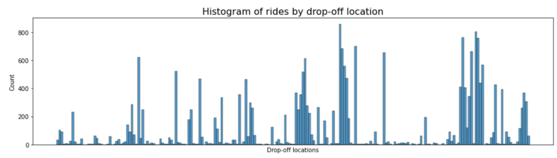
Figure 7: **Revenue Analysis.** Tracking financial performance across days and months.

3.5 Geographic Patterns

Finally, we analyzed drop-off locations to understand trip destinations. Certain zones consistently require longer trips.



(a) Mean Distance by Drop-off



(b) Drop-off Location Frequency

Figure 8: **Location Analysis.** Identifies which zones require longer trips and which are the most frequent destinations.

4 Phase 3: Statistical Hypothesis Testing

Reference: Notebook 03_NYC_Taxi_Statistical_Analysis.ipynb

4.1 Objective

We conducted rigorous statistical analysis to test whether payment type has a significant effect on fare amount. This A/B test helps determine if encouraging credit card payments could increase driver revenue.

4.2 Hypotheses

Null Hypothesis (H_0): There is no difference in average fare amount between credit card and cash customers.

$$\mu_{credit} = \mu_{cash}$$

Alternative Hypothesis (H_1): There is a difference in average fare amount between credit card and cash customers.

$$\mu_{credit} \neq \mu_{cash}$$

Significance Level (α): 0.05

4.3 Results

We compared 15,265 credit card trips against 7,267 cash trips.

- **Credit Card Mean Fare:** $13.43
- **Cash Mean Fare:** $12.21
- **Difference:** Credit card users pay $1.22 more on average (9.1% higher).

Two-Sample T-Test

```
from scipy import stats
stats.ttest_ind(credit_card, cash, equal_var=False)
```

Test Statistic: $t = 6.87$
P-value: 6.80×10^{-12}

4.4 Conclusion

The p-value is significantly smaller than 0.05. We **reject the null hypothesis**. There is a statistically significant difference in average fare amount between credit card and cash payments.

Business Implication: Encouraging credit card payments could increase driver revenue by approximately 9%.

5 Phase 4: Multiple Linear Regression Model

Reference: Notebook 04_NYC_Taxi_Regression_Model.ipynb

5.1 Objective

We built a multiple linear regression model to predict taxi fare amounts based on trip characteristics.

5.2 Data Preparation & Feature Selection

Before modeling, we performed a final check on outliers for the regression features. We removed negative fares, trips over 100 miles, and zero-passenger trips to ensure data quality.

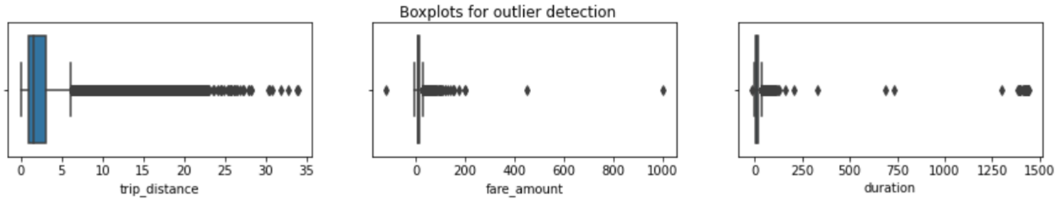
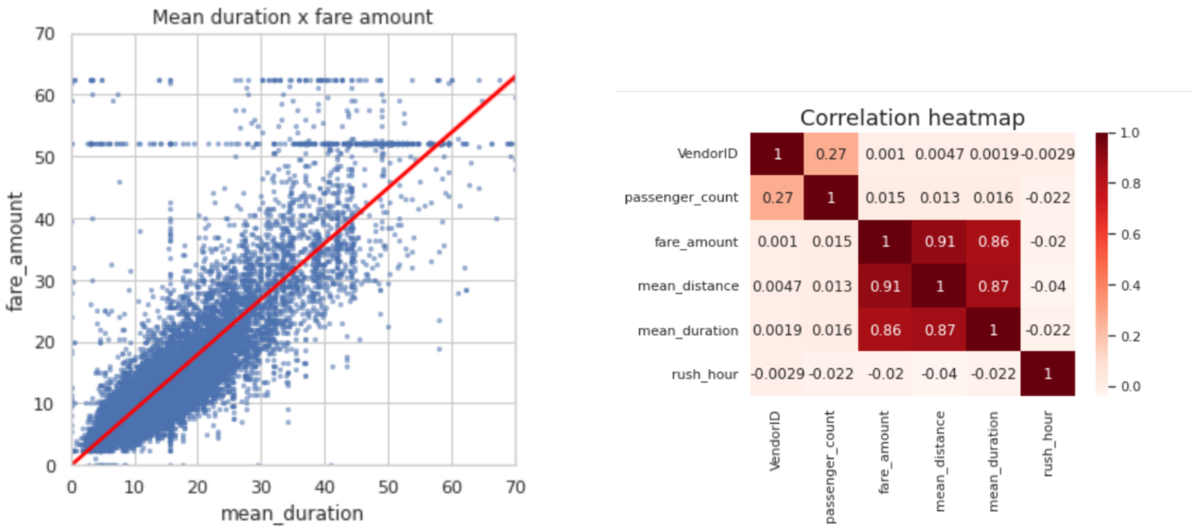


Figure 9: **Feature Outlier Detection.** Box plots for trip distance, fare amount, and duration.

Correlation Analysis

We found a very strong positive correlation (0.91) between `trip_distance` and `fare_amount`. Duration also showed predictive power.



(a) Duration vs Fare Scatter (b) Correlation Heatmap

Figure 10: **Feature Correlation.** Strong relationships observed between duration/distance and fare amount.

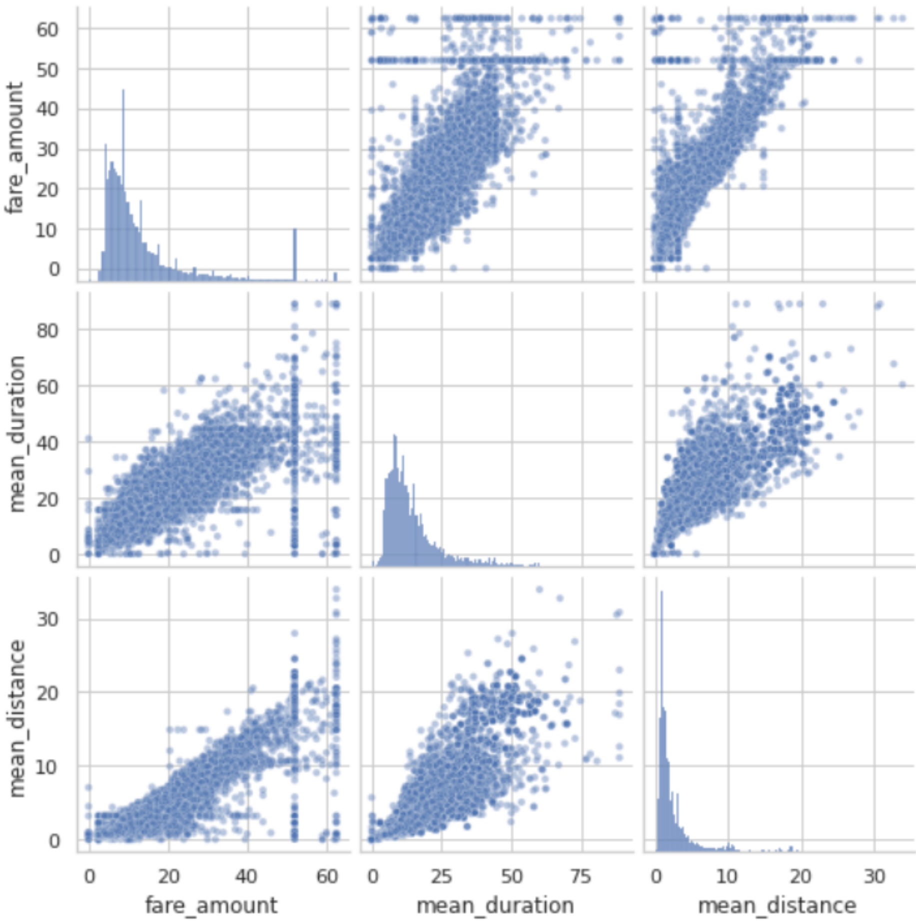


Figure 11: **Pairwise Relationships.** Matrix of scatter plots showing interactions between all key variables.

5.3 Model Building

We fit a linear regression model using standardized features.

$$\text{fare} = \beta_0 + \beta_1(\text{dist}) + \beta_2(\text{time}) + \beta_3(\text{pass}) + \beta_4(\text{vendor})$$

5.4 Model Results

- R^2 **Score:** 0.863 (86.3% of variance explained)
- **RMSE:** $3.41
- **Key Insight:** Trip distance is the dominant predictor ($\beta = +\$9.87$ per std unit), followed by duration.

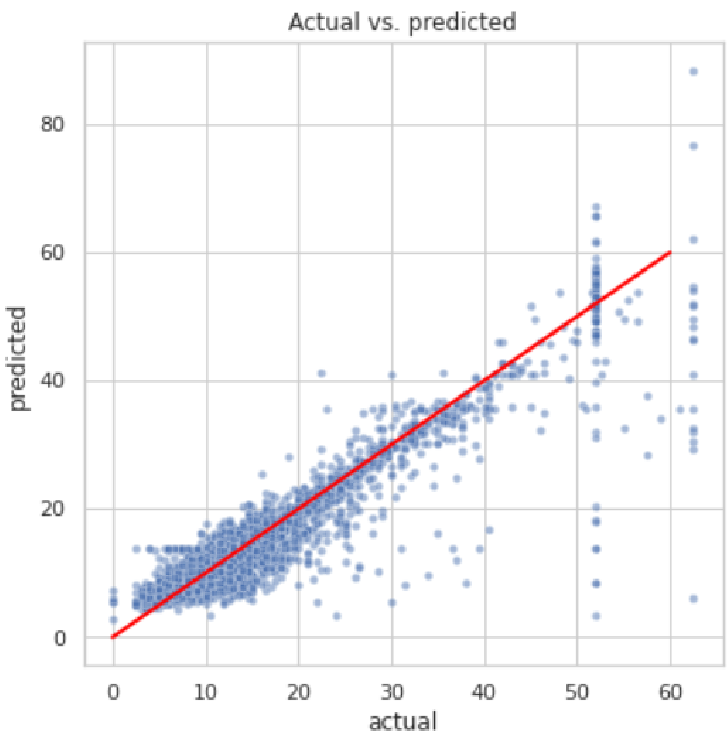
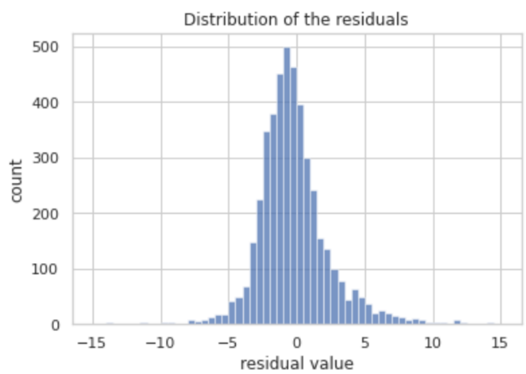


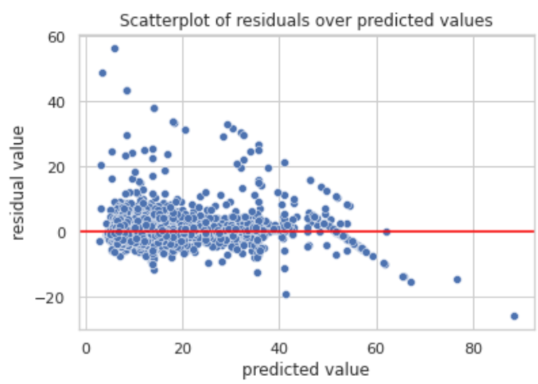
Figure 12: **Actual vs. Predicted Fares.** The points cluster tightly around the diagonal line, indicating high model accuracy.

5.5 Assumption Testing

We verified linear regression assumptions. The residuals are approximately normally distributed and show homoscedasticity (random scatter), validating the model choice.



(a) Residuals Distribution



(b) Residuals vs. Predicted

Figure 13: **Assumption Testing.** Left: Residuals are normally distributed (bell curve). Right: Residuals show random scatter (homoscedasticity).

6 Phase 5: Machine Learning Classification

Reference: Notebook 05_NYC_Taxi_ML_Classification.ipynb

6.1 Objective

We trained a Random Forest and XGBoost classifier to predict "generous tippers" (tips $\geq 20\%$). The dataset was imbalanced (25% generous vs 75% non-generous).

6.2 Model Evaluation

The XGBoost model achieved an **AUC-ROC of 0.74** and an **F1-Score of 0.67**. It correctly identified 77.89% of generous tippers (Recall).

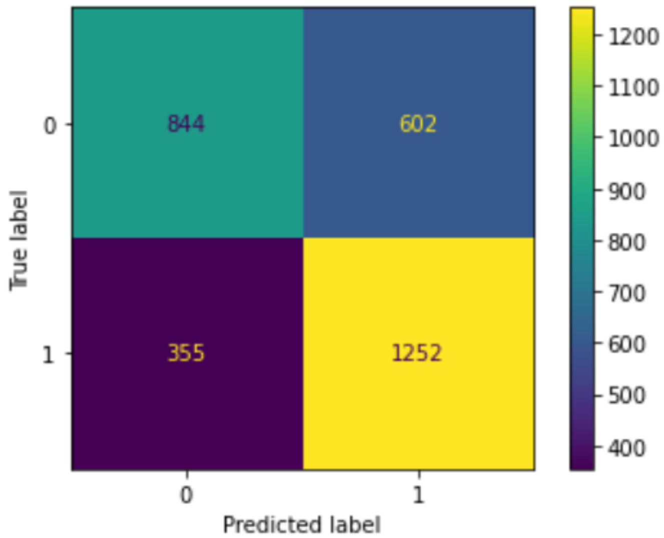


Figure 14: **Confusion Matrix.** Visualizes the True Positives (596), False Positives, and overall classification accuracy.

6.3 Feature Importance

The most important features for predicting tips were `fare_amount` and `trip_distance`. Passengers on more expensive, longer trips are more likely to tip generously.

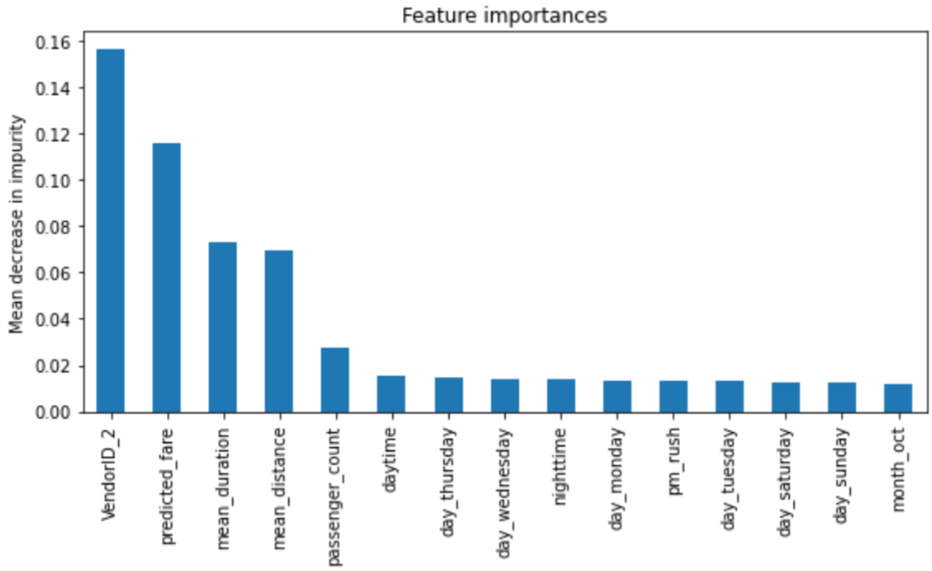


Figure 15: **Feature Importance.** A horizontal bar chart ranking which variables were most useful for predicting generous tippers.

7 Conclusions and Recommendations

7.1 Summary of Findings

1. **Revenue Drivers:** Distance is the dominant factor in fare calculation (correlation 0.91).
2. **Payment Impact:** Credit card users pay significantly higher fares on average ($1.22 more per trip).
3. **Predictability:** Fares can be predicted with high accuracy ($R^2 = 0.86$).
4. **Customer Behavior:** Tipping behavior is moderately predictable; high fares and long trips correlate with generous tips.

7.2 Business Recommendations

- **For Drivers:** Encourage credit card payments and prioritize longer trips to maximize revenue.
- **For TLC:** Ensure card readers are functional in all vehicles and consider dynamic pricing during confirmed rush hours.
- **Strategy:** Target high-fare trips (airports) and evening weekend shifts for better tip potential.

A Details

Software Environment: Python 3.7+ (pandas, scikit-learn, xgboost).
Reproducibility: Random state 42 used for all splits and models.
Code Availability: Analysis is split across 5 Jupyter notebooks corresponding to the project phases.
Special Thanks: A special thanks to all the teachers from Google Advance Data Analytics Specialization, truly an amazing journey to learn from industry experts, and implementing stuff in real life data.