IN PARTNERSHIP WITH PLYMOUTH UNIVERSITY

Name: Wickrama Sankalpa
Student Reference Number: Group A

Module Code: PUSL2077	Module Name: Data Science in Python
Coursework Title: Group Assignment Final Report	
Deadline Date: Wednesday, 2 April 2025, 4:00 PM	Member of staff responsible for coursework: Ms. Lakni Peiris
Programme: BSc (Hons) Data Science	

Please note that University Academic Regulations are available under Rules and Regulations on the University website **www.plymouth.ac.uk/studenthandbook**.

Group work: please list all names of all participants formally associated with this work and state whether the work was undertaken alone or as part of a team. Please note you may be required to identify individual responsibility for component parts.

Name (as appeared on DLE)	Plymouth ID No.
Wickrama Sankalpa	10953717
pybpathirana	10953739
rmajayawardhane	10953738
Isira Withana	10953722
Hetti Chamod	10953714

We confirm that we have read and understood the Plymouth University regulations relating to Assessment Offences and that we are aware of the possible penalties for any breach of these regulations. We confirm that this is the independent work of the group.

Signed on behalf of the group:

Individual assignment: *I confirm that I have read and understood the Plymouth University regulations relating to Assessment Offences and that I am aware of the possible penalties for any breach of these regulations. I confirm that this is my own independent work.*

Signed :

Table of Contents

Table of Figures

2.2 Interim 1

Introduction

New York City Taxi Trip Analysis: January 2015 Insights

We started this project with the aim of increasing the tip amount and total income for taxi drivers in New York City per trip. To do this, we selected a data set and conducted this project to draw the necessary conclusions by making predictions. We removed the variables that we did not need from that data set and used the variables that we did need to achieve our goal.

New York City's taxi system forms the circulatory network of urban mobility, with over 200 million annual trips. This report analyzes **5,337 yellow taxi trips** from January 2015 to identify actionable patterns in fare structures, tipping behavior, and revenue generation - critical insights as the industry navigates post-pandemic recovery and evolving competition from ride-sharing platforms.

Dataset Overview

The TLC-regulated dataset contains:

- **Temporal Features**: Precise pickup/dropoff timestamps
- **Financial Metrics**: Fare amounts, tips, tolls, and surcharges
- **Operational Data**: Trip distances, durations, and payment methods
- **Geospatial Attributes**: Location-based coordinates (excluded in this analysis)

Key Focus Areas

1. **Tipping Dynamics**
2. **Revenue Drivers**
3. **Operational Efficiency**

Dataset Overview and Initial Structure

Dataset - https://docs.google.com/spreadsheets/d/1M3kOsu2uElxfl8fSnyJTN9Zv4pWVoPmV/edit?usp=sharing&ouid=117184260344297243925&rtpof=true&sd=true

The preprocessing begins with a taxi trip dataset containing 5,336 rows with the following key columns:

1. VendorID - a code that identifies the trip record's provider (1 = VeriFone Inc., 2 = Creative Mobile Technologies).
2. tpep_pickup_datetime - The time and date that the trip began, or when the meter was engaged.
3. tpep_dropoff_datetime - Time and date of the meter's disengagement (trip end).
4. Passenger_count - The driver-entered value for the number of passengers in the car.
5. Trip_distance - The taximeter recorded the trip's total distance in miles.
6. Pickup_longitude - The trip's starting longitude coordinate.
7. Pickup_latitude -Coordinate the latitude at the beginning of the journey.
8. RateCodeID - A code that indicates the trip's ultimate rate type:
    1. The standard rate
    2. The JFK Airport
    3. Airport in Newark
    4. Westchester or Nassau
    5. Bargaining for a better price
    6. Group transportation
9. Store_and_fwd_flag - shows whether the trip log was kept in the car's memory prior to being forwarded to the seller:

    Y = Saved and sent at a later time

    N = Instantaneously sent
10.

11. Dropoff_longitude - The longitude coordinate of the destination.
12. Dropoff_latitude - The voyage finished at the latitude coordinate.
13. Payment_type - Code indicating the payment method:
    1. A credit card
    2: Money
    3. No fee
    4: Conflict
    5: Uncertain
    6. Voided journey
14. Fare_amount - Based on time and distance, the meter determines the base fare.
15. Extra - Extra fees, such as nighttime and rush-hour surcharges (e.g., $0.50 or $1).
16. MTA_tax - All taxi rides in New York City are subject to a set $0.50 fee.
17. Improvement_surcharge- In 2015, there was a $0.30 fee added to all journeys.
18. Tip_amount- The amount that passengers tip (only when using credit cards; cash tips are not included).
19. Tolls_amount- total amount of tolls paid when traveling.
20. Total_amount- The entire amount of the passenger's fare, including any applicable tips, fees, and tolls (but not cash tips).

The original dataset structure shows 19 columns with various data types, including timestamps, numerical values, and categorical variables. The first transformation maintains essential trip information while beginning to restructure payment data.

Data Cleaning and Preprocessing

Preprocessing dataset - https://docs.google.com/spreadsheets/d/1__KG7z0FXIHp_Lxhj0sNjXGUpr0UkITq/edit?usp=sharing&ouid=117184260344297243925&rtpof=true&sd=true

Feature Selection and Column Reduction

The code performs selective column retention, removing several fields from the original dataset:

- Geographic coordinates (pickup/dropoff longitude and latitude)
- VendorID and RateCodeID
- store_and_fwd_flag

This column reduction focuses the dataset on trip metrics, timing, and payment information relevant to the analysis.

Data Type Transformation

Several important type conversions occur in the preprocessing:

1. Payment Type Conversion:

The payment_type column is transformed from numeric codes (1, 2) to categorical values ("Credit", "Cash"). This makes the data more interpretable and prepares it for further encoding.

2. Datetime Preservation:

The temporal columns (tpep_pickup_datetime and tpep_dropoff_datetime) are maintained in their original datetime format, allowing for potential time-based feature extraction later.

Feature Engineering and Addition

The preprocessing adds derived columns:

1. Static Value Columns:

Two new columns appear in the second stage:

- trip_distance_static
- fare_amount_static

These appear to duplicate the original trip_distance and fare_amount values, potentially serving as reference points before any scaling or normalization.

Missing Data Handling

- RateCodeID-remove(mode)
- Store_and_fwd_flag-remove(mode)
- Dropoff_longitude-remove
- Dropoff_latitude-remove

Replace missing values with Median

Why we use median:

- Median (middle value) is not sensitive to outliers.
  Those variables include outliers so we use median

- Use median if outliers are messing up the picture and you want a stable, typical value

- Extra-median
- Tip_amount-median

Categorical Variable Encoding

The final preprocessing step involves one-hot encoding:

1. One-hot Encoding:

The payment_type categorical column is converted into three boolean columns:

- payment_type_Cash
- payment_type_Credit
- payment_type_Other

Each row contains True/False values indicating the payment method used.

Handle Outliers

Handling outliers is a critical step in data preprocessing to ensure the integrity and reliability of machine learning models

1. Identify Outliers:

- Boxplots: Visualize data distribution and identify points beyond 1.5 times the interquartile range (IQR) as potential outliers.

2. Treat Outliers

The treatment method depends on the dataset and project goals. Common strategies include:

- Remove Outliers:

  IQR Method: Define outliers as data points below Q1−1.5×IQR$Q1-1.5\times IQR$ or

  above Q3+1.5×IQR$Q3+1.5\times IQR$ and remove them

Data Filtering

A significant data reduction occurs between the initial and final datasets:

Original dataset: 5,336 rows

Processed dataset: 4,212 rows

This suggests the application of filtering criteria, possibly removing outliers, incomplete records, or focusing on specific time periods.

2.3 Final report

1.Descriptive analysis: For Predicting Total Amount

1. Objective

The objective of this analysis is to develop a predictive model for **Total Amount**, specifically estimating the **fare _amount** based on trip-related features such as **trip_distance, passenger_count, and tpep_pickup_datetime**.

Predicting Fare Amount

- Helps estimate the cost of a taxi ride before starting a trip.
- Useful for **ride-sharing apps, fare comparison, and customer expectation management**.
- Enables passengers to make informed decisions based on predicted fare values.

What Can Be Predicted?

- The **expected total fare (total_amount)** for a given trip based on key factors like:
    - **Trip Distance**: The length of the journey.
    - **Pickup Time**: Impact of peak vs. off-peak hours.
    - **Passenger Count**: How the number of passengers may affect pricing.

2. Dataset Overview

The dataset consists of multiple trip-related variables that influence fare pricing. Key attributes include:

- **Fare Amount**: The target variable representing the fare charged for a trip.
- **Trip Distance**: The distance traveled, which directly impacts fare calculation.
- **Passenger Count**: Number of passengers in the trip, potentially influencing the fare structure.
- **Trip Duration**: The total time taken for the trip, derived from pickup and drop-off timestamps.
- **Total Amount**: The overall charge, including additional fees like tolls and surcharges.
- **Payment Type**: The method of payment (Cash, Credit, etc.), which may affect tipping behavior.

To ensure data quality, missing values were identified and handled appropriately before proceeding with the analysis.

3. Exploratory Data Analysis (EDA)

3.1 Summary Statistics

The table below provides key descriptive statistics for the fare amount:

Metric	values
Mean	10.746991213488483
Median	10.3
Mode	7.3
Min	3.3
Max	21.42
Standard Dev.	3.6772356839370097

3.2 Distribution of Fare Amount

The following histogram illustrates the distribution of **Total Amount**, showcasing the skewness and range:

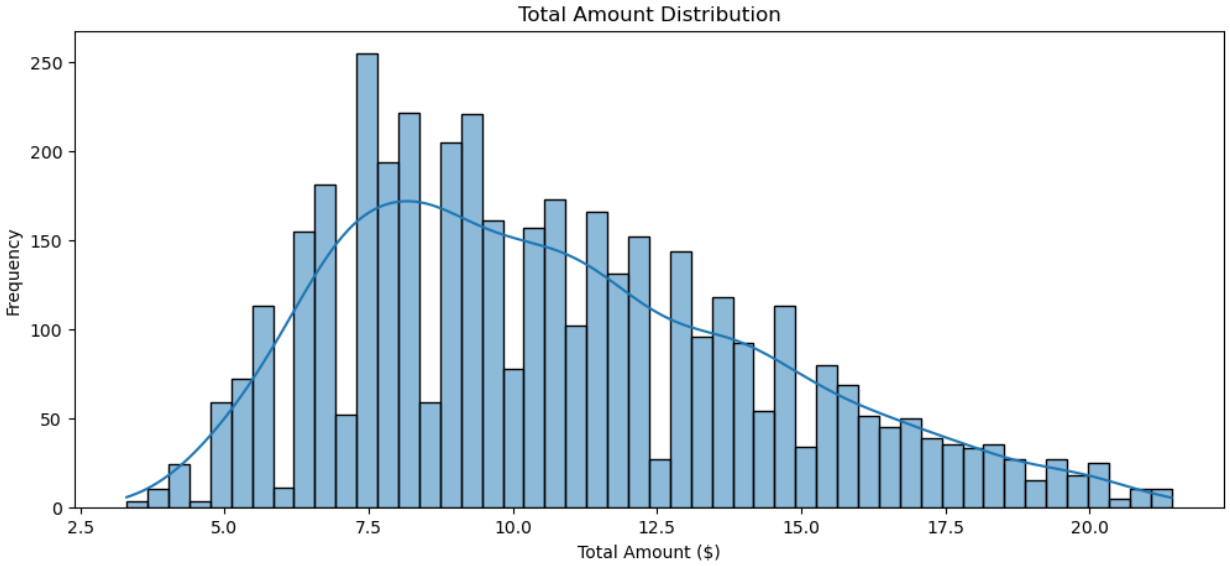

Figure 1: Total amount distribution

3.3 Outlier Detection

A boxplot helps visualize the presence of outliers in **Total Amount**:

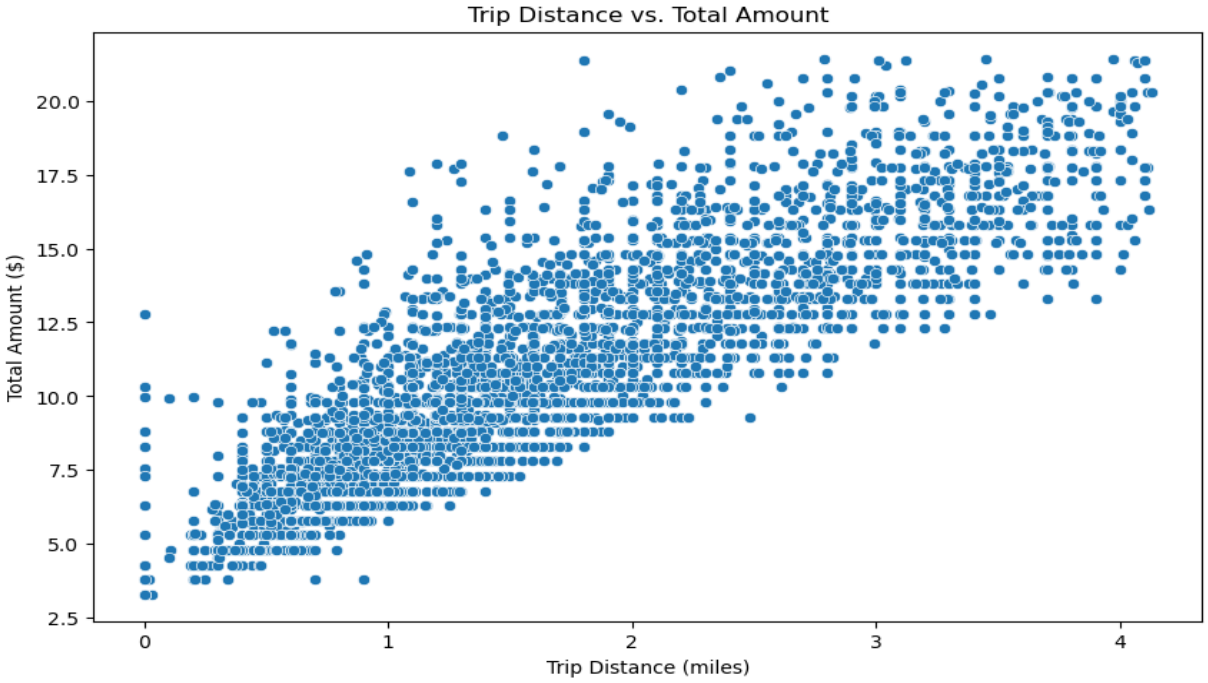

Figure 2: Trip distance vs. Total amount

3.4 Relationship Between Trip Distance and Total

The scatter plot below shows how trip distance impacts **Total Amount**

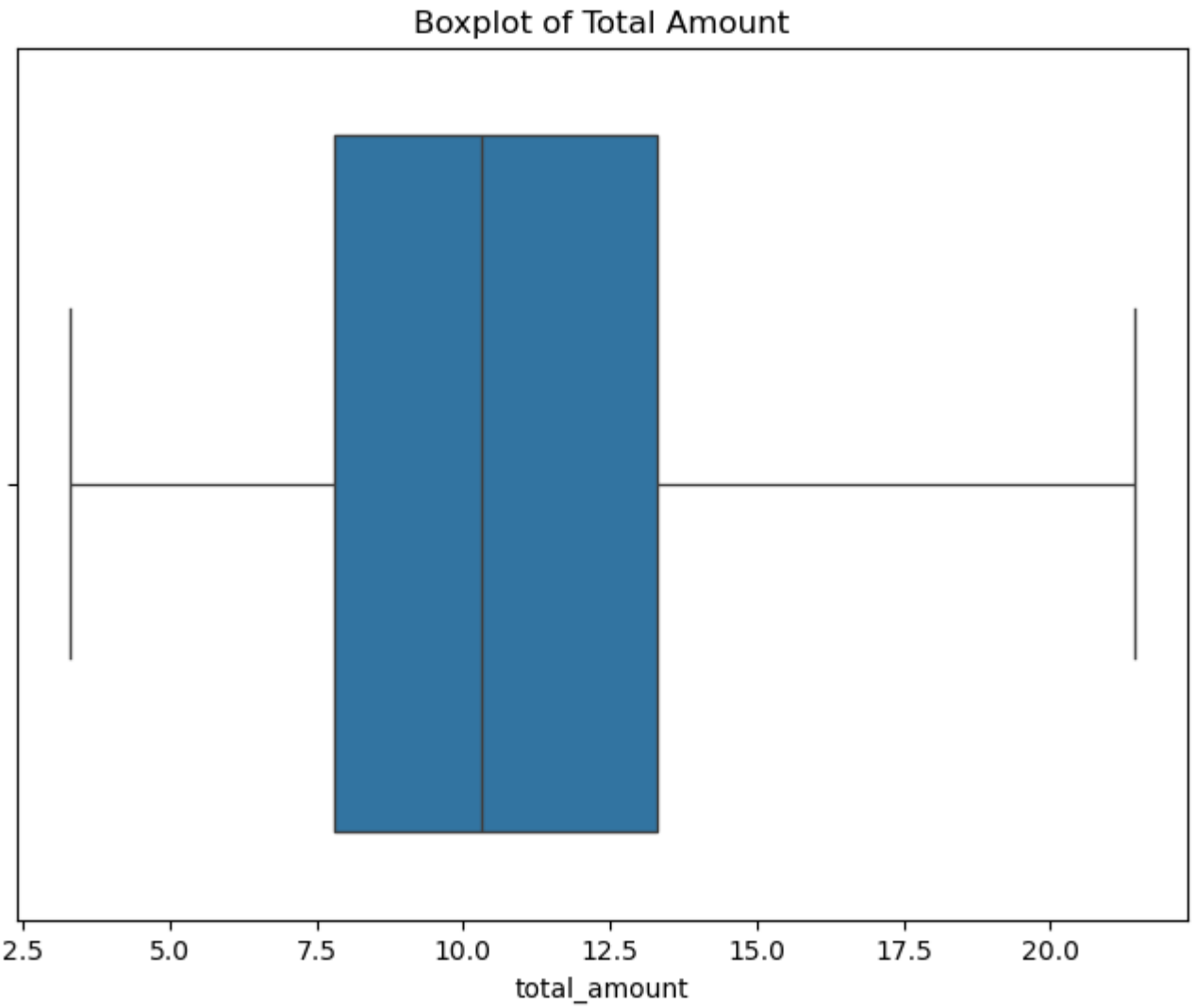

Figure 3: Boxplot of total amount

4. Insights & Key Findings

4.1 Summary of Total Amount Patterns

- The average Total Amount is moderately skewed, with some extremely high values affecting the mean calculation.
- The median fare provides a better central measure since it is less affected by extreme values.

- The interquartile range (IQR) shows that most fares fall within a reasonable range, but outliers are present beyond the upper quartile

4.2 Outliers and Trends

- The **boxplot of Total Amount** highlights several significant outliers, suggesting unusually high fares in some trips.
- The **fare distribution histogram** reveals a **right-skewed pattern**, indicating that while most trips have lower fares, a small number of trips have exceptionally high fares.
- A **scatter plot of trip distance vs. Total Amount** confirms a positive correlation, though some anomalies exist where short distances have disproportionately high fares.
- The following violin plot provides further insights into the distribution of fares and their density:

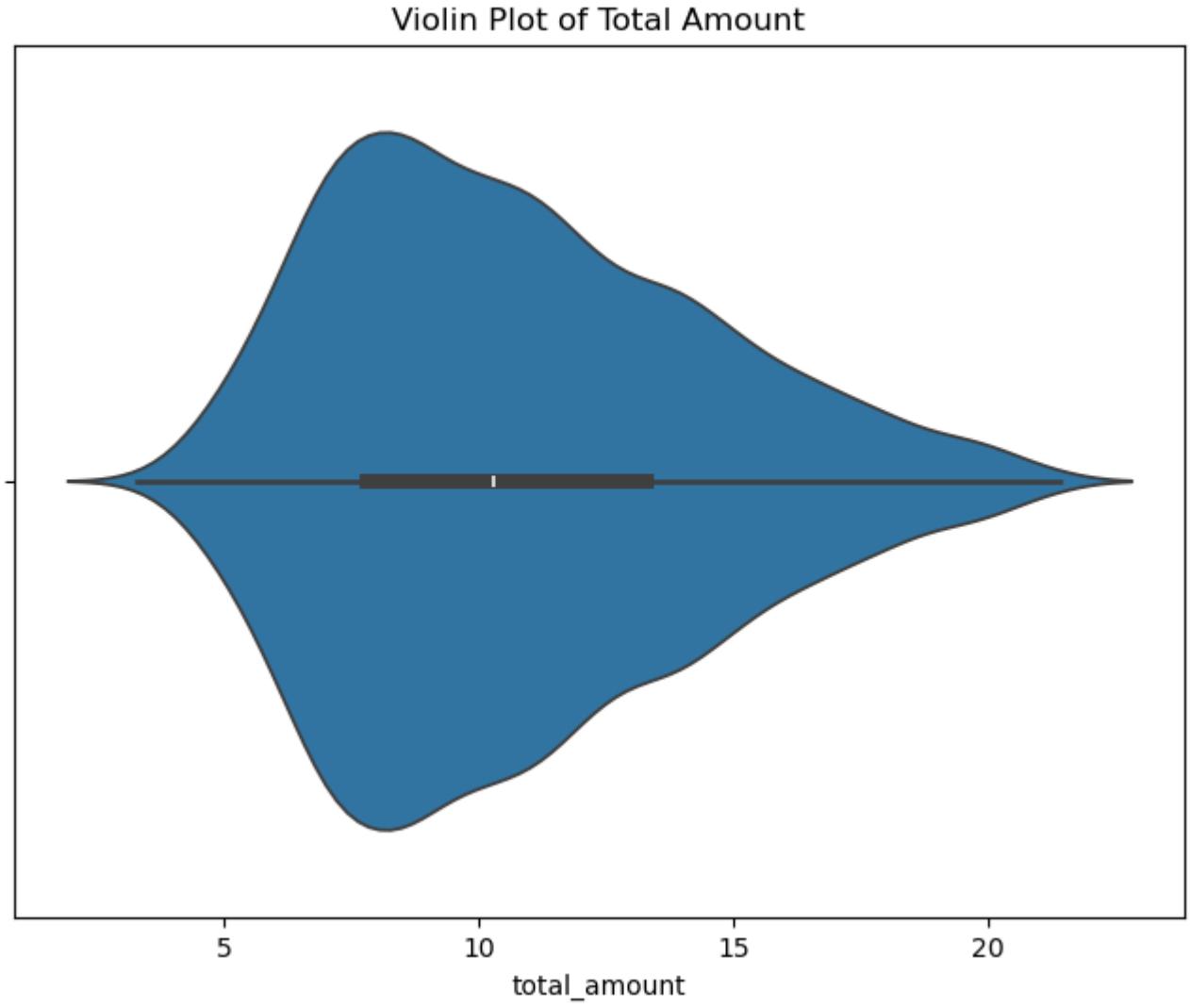

Figure 4: Violin plot of total amount

4.3 Need for Data Transformation

- Since the fare amount distribution is **right-skewed**, applying a **log transformation** can help normalize the data, making it more suitable for machine learning models. The transformed distribution can be visualized as follows:

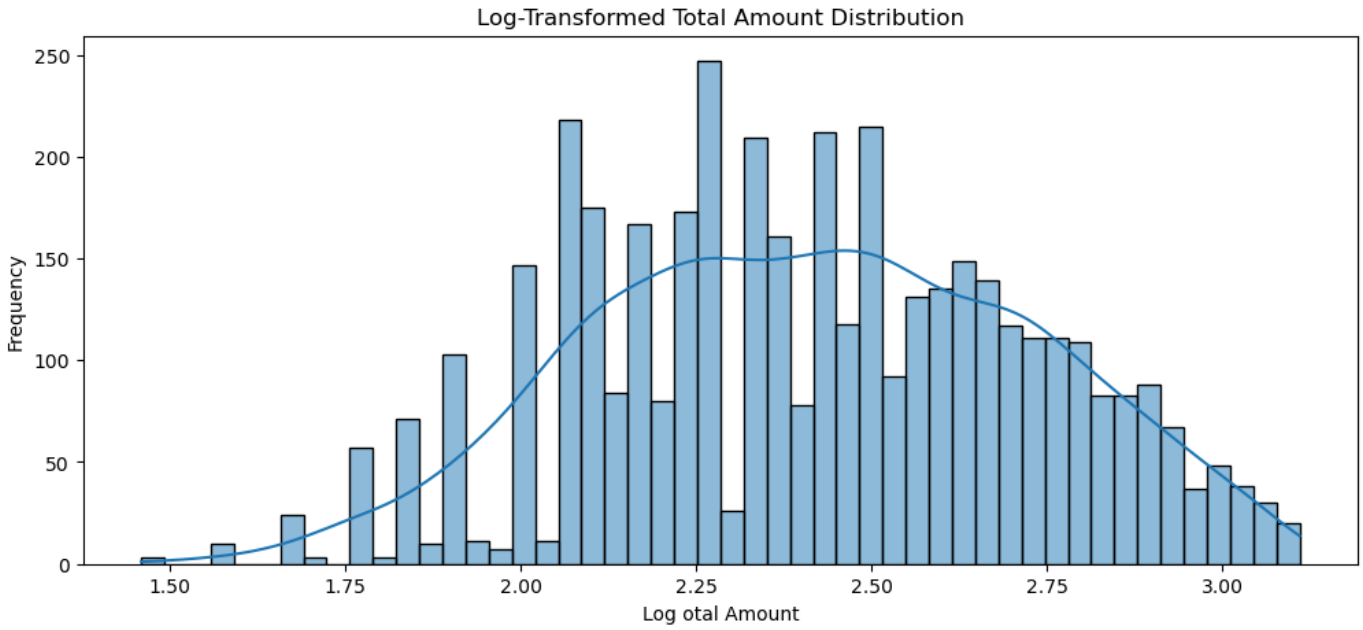

Figure 5: Log-transformed total amount distribution

- Outliers can significantly impact model predictions. Using robust techniques such as **Winsorization** (capping extreme values) or removing fares above the **99th percentile** can improve model performance.
- Standardizing numerical features such as **trip distance** and **fare amount** using **Min-Max Scaling or Z-score normalization** can help models generalize better.

4.4 Measures of Position

To further understand fare distribution, key measures of position include:

- **25th Percentile (Q1):** 7.8
- **50th Percentile (Q2/Median):** 10.3
- **75th Percentile (Q3):** 13.3
- **Interquartile Range (IQR):** 5.500000000000001

These values help in identifying the central tendency and spread of Total Amount.

The interquartile range (IQR) indicates that most fares lie within a $4.5 range between Q1 and Q3, with potential outliers beyond this range requiring special handling.

2.Descriptive analysis: For Predicting Tip Amount

1. Objective

The objective of this analysis is to develop a predictive model for **Tip Prediction**, specifically estimating the **tip_amount** based on trip-related features such as **trip_distance, passenger_count, and tpep_pickup_datetime**.

Predicting Tip Amount

- Helps drivers understand which factors lead to higher tips.
- Useful for **ride-sharing companies** to optimize driver earnings.
- Enables better customer service strategies based on tipping patterns.

What Can Be Predicted?

- The **expected tip amount (tip_amount)** for a given trip based on key factors like:
    - **Trip Distance**: Longer trips may influence tipping behavior.
    - **Pickup Time**: Impact of peak vs. off-peak hours.
    - **Passenger Count**: Groups vs. solo passengers may have different tipping habits.

2. Dataset Overview

The dataset consists of multiple trip-related variables that influence tipping behavior. Key attributes include:

- **Tip Amount**: The target variable representing the tip given for a trip.
- **Trip Distance**: The distance traveled, which may influence tip generosity.
- **Passenger Count**: Number of passengers in the trip, potentially affecting the tip amount.
- **Payment Type**: The method of payment (Cash, Credit, etc.), which may impact tipping habits.

To ensure data quality, missing values were identified and handled appropriately before proceeding with the analysis.

3. Exploratory Data Analysis (EDA)

3.1 Summary Statistics

The table below provides key descriptive statistics for the tip amount:

Metric	
Mean	1.070351460460698
Median	1.0
Mode	0.0
Min	0.0
Max	4.55
Standard Dev.	1.0240210526894407

3.2 Distribution of Tip Amount

The following histogram illustrates the distribution of tip amounts, showcasing the skewness and range:

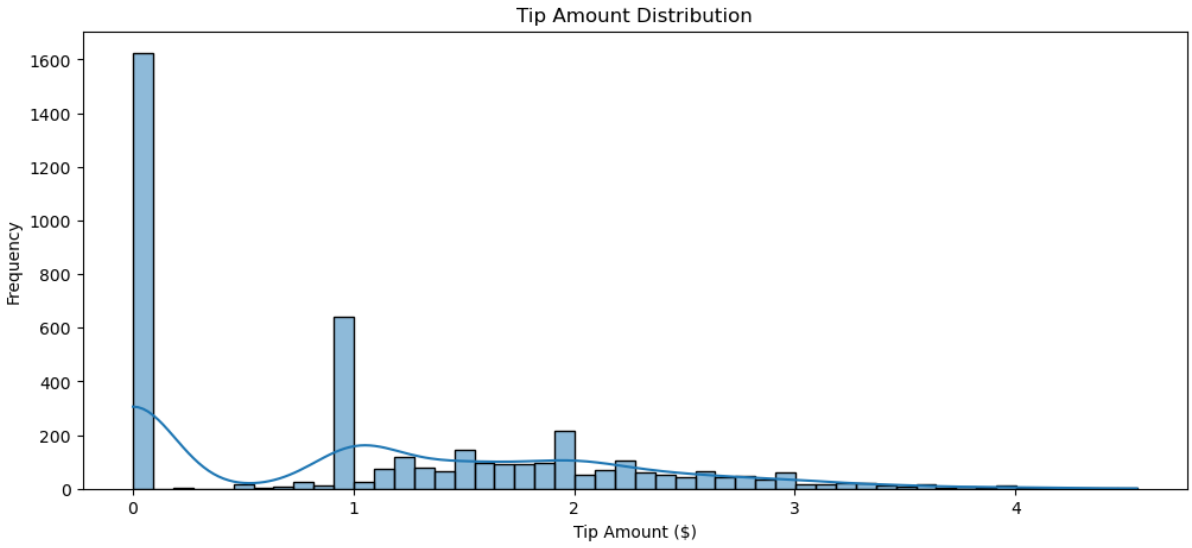

Figure 6: Trip amount distribution

3.3 Outlier Detection

A boxplot helps visualize the presence of outliers in tip amounts:

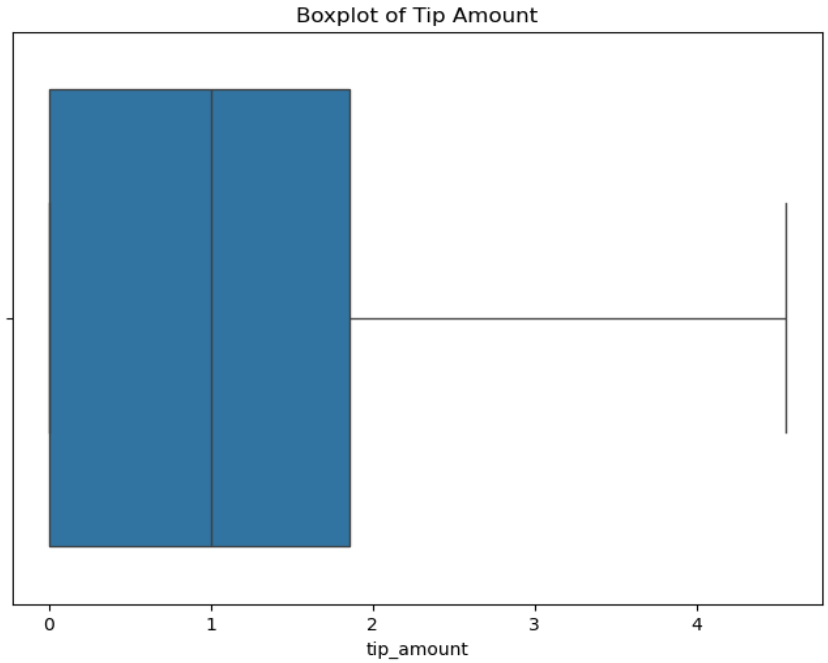

Figure 7: Boxplot of tip amount

3.4 Relationship Between Trip Distance and Tip

The scatter plot below shows how trip distance impacts tip amount:

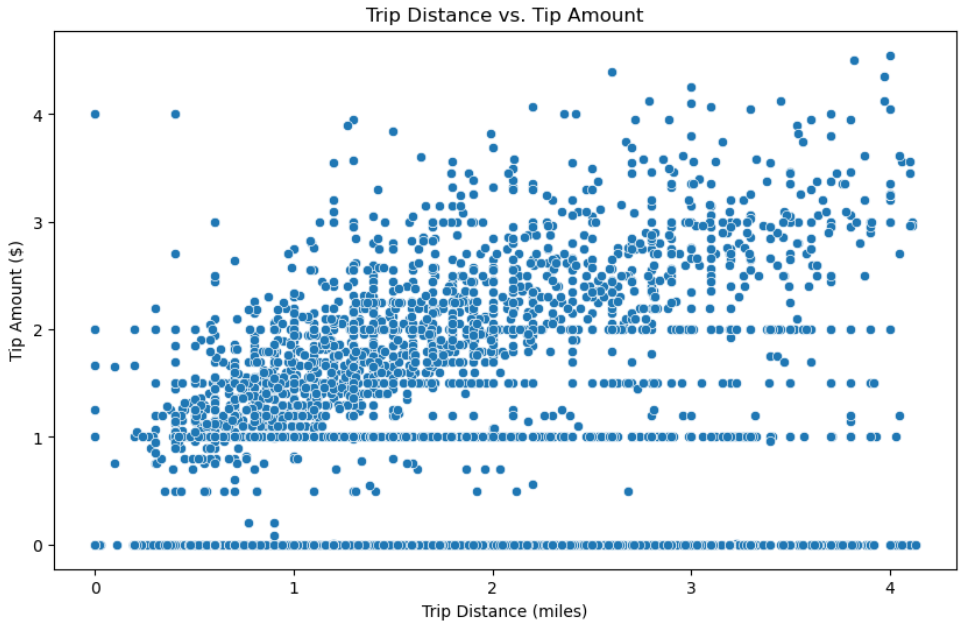

Figure 8: Trip distance vs. Tip amount

4. Insights & Key Findings

4.1 Summary of Tip Amount Patterns

- The **average tip amount** shows some variability, with potential outliers.

- The **median tip** is often a better central measure due to extreme values.
- The **interquartile range (IQR)** highlights a concentration of tips within a reasonable range.

4.2 Outliers and Trends

- The **boxplot of tip amounts** reveals significant outliers, suggesting occasional very high tips.
- The **tip distribution histogram** indicates a **right-skewed pattern**, where most trips have lower tips, but some customers tip generously.
- A **scatter plot of trip distance vs. tip amount** suggests a weak positive correlation, though some high tips appear unrelated to distance.
- The following violin plot provides further insights into the density and spread of tip amount

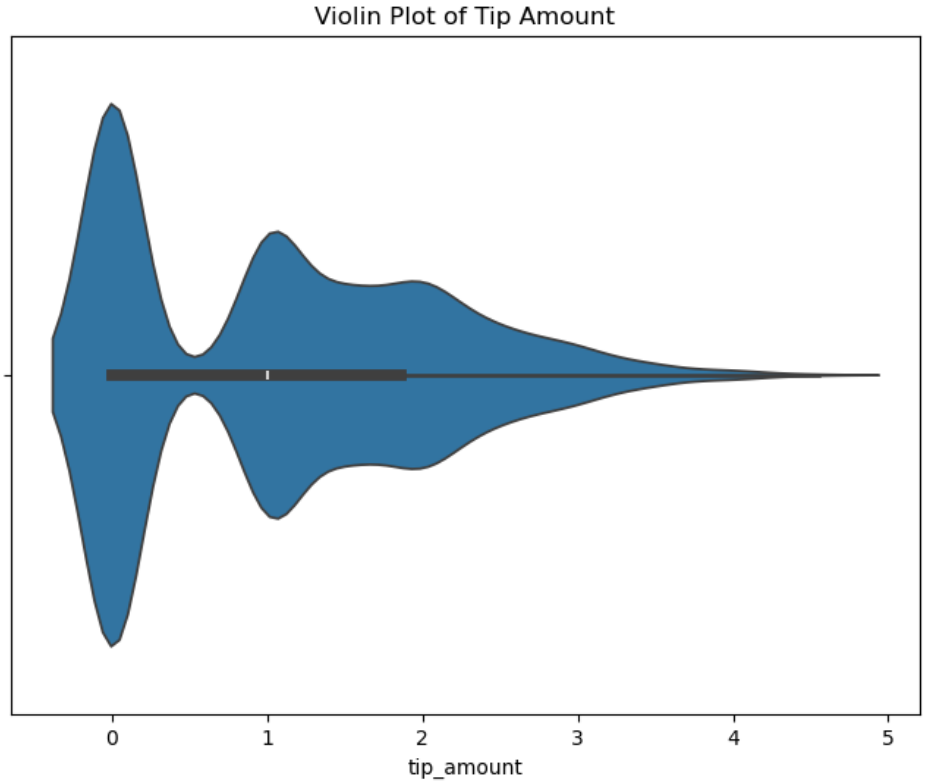

Figure 9: Violin plot of tip amount

4.3 Need for Data Transformation

- Since the tip amount distribution is **right-skewed**, applying a **log transformation** may help normalize the data, improving model performance.

- Extreme tip outliers (e.g., above the 99th percentile) may need handling through **Winsorization** or **removal**.
- Standardizing numerical features like **trip distance and tip amount** can enhance model generalization.

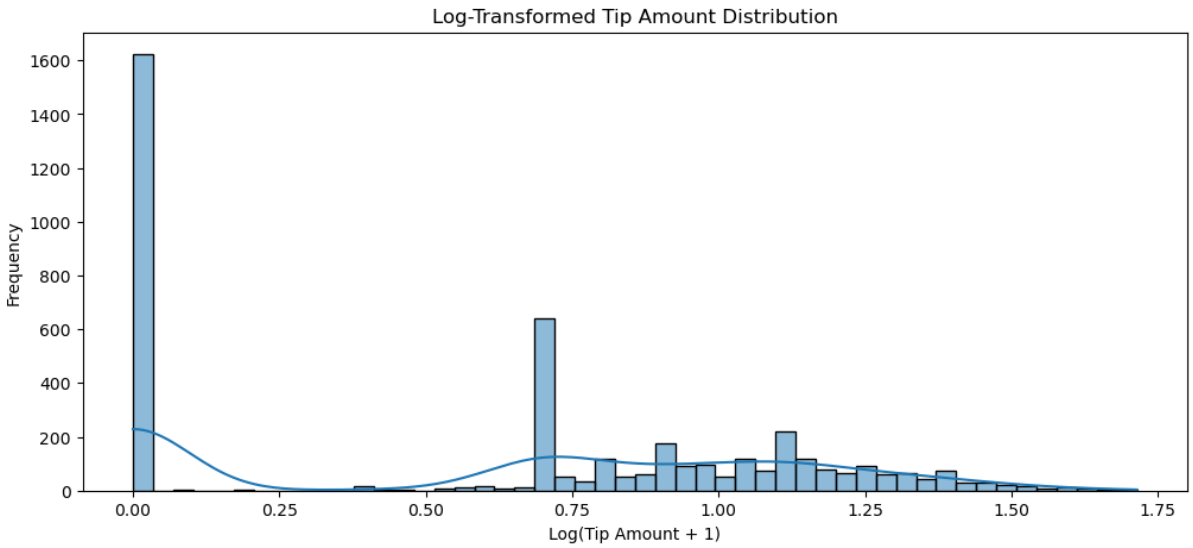

Figure 10: Log-transformed tip amount distribution

4.4 Measures of Position

To further understand tip distribution, key measures of position include:

- **25th Percentile (Q1):** 0.0
- **50th Percentile (Q2/Median):** 1.0
- **75th Percentile (Q3):** 1.86
- **Interquartile Range (IQR): 1.86**

These values help in identifying the central tendency and spread of tip amounts. The interquartile range (IQR) indicates the most common tip range, with potential outliers beyond this range requiring attention.

Data visualization using different data graphics:

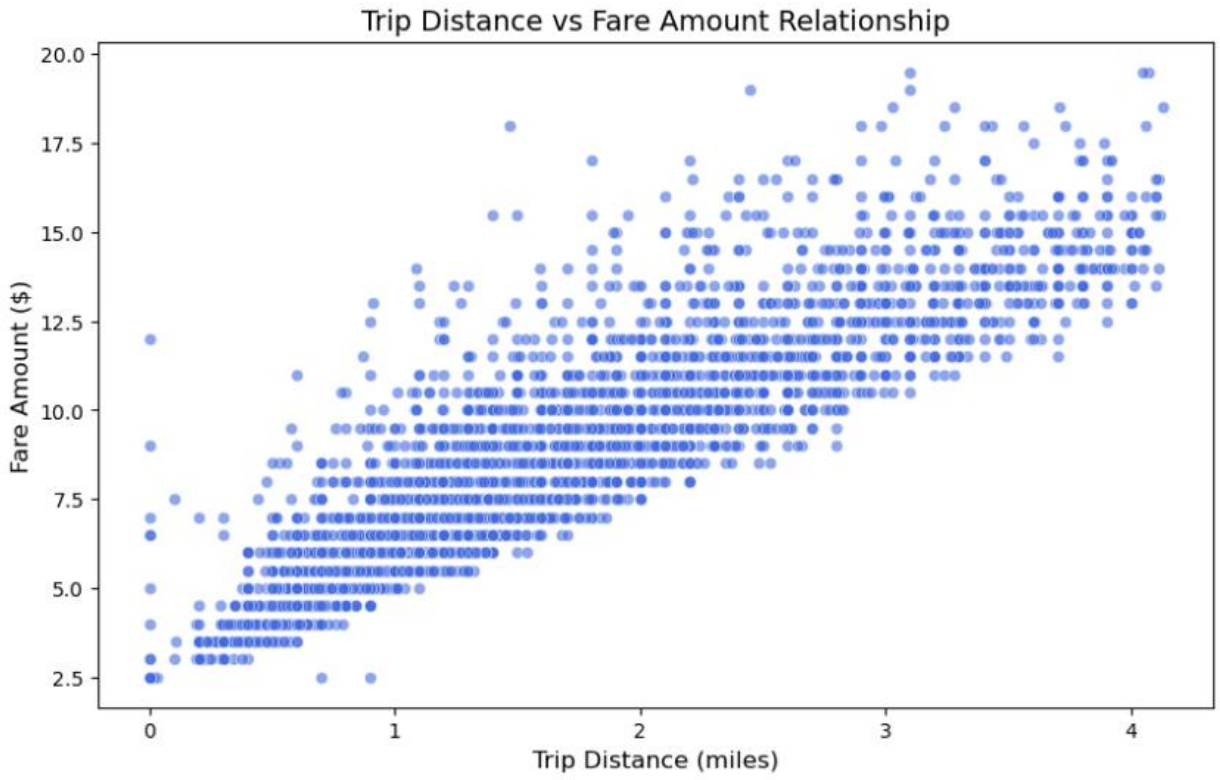

Figure 11: Trip distance vs. Fare amount relationship

Chart Type: Scatter Plot

Description:

- This plot shows the relationship between the trip_distance (in miles) and the fare_amount (in dollars).
- Each point represents a single taxi trip.

Key Observations:

- Fare amount increases with trip distance, as expected.
- Shorter trips show more variability in fare amounts due to base fares, surcharges, and tips.
- Longer trips (e.g., >10 miles) tend to have higher fares with less variability.

Use Case:

Helps analyze how distance impacts fare and identify outliers (e.g., trips with unusually high or low fares for their distance).

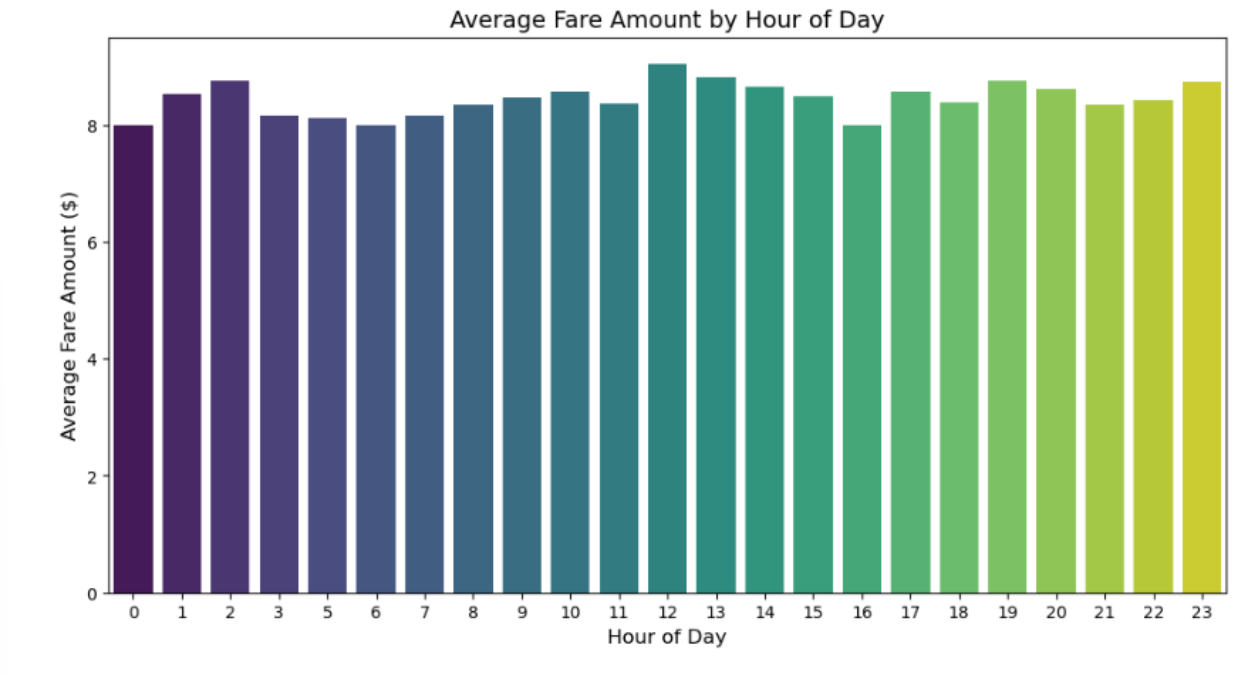

Figure 12: Average fare amount by hour of day

Chart Type: Bar Plot

Description:

- This plot shows the average fare_amount for each hour of the day (hour_of_day).
- The data is grouped by hour and averaged.

Key Observations:

- Higher average fares are observed during late-night hours (12 AM–6 AM), likely due to night surcharges and longer trips.
- Daytime hours (6 AM–6 PM) show relatively consistent average fares.

Use Case:

- Useful for identifying peak hours when drivers can expect higher fares.
- Can help optimize driver schedules

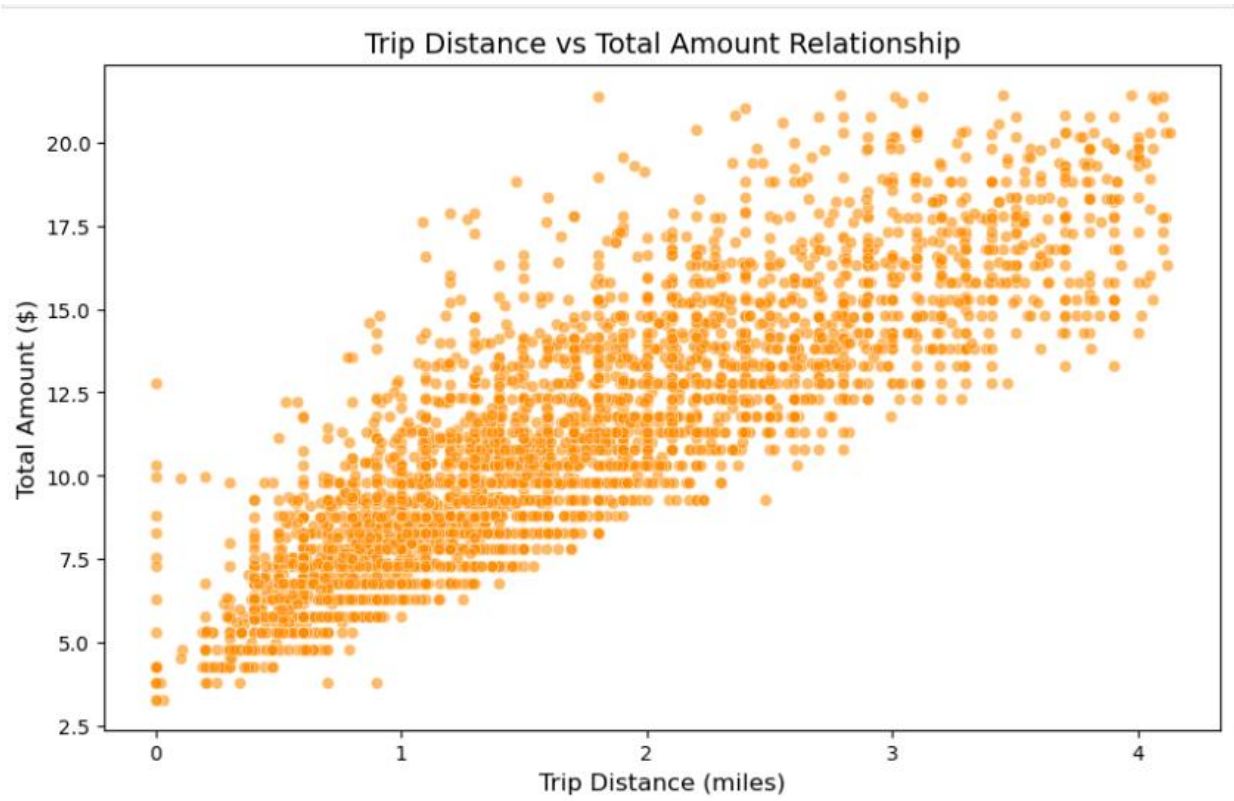

Figure 13: Tip distance vs. Total amount relationship

Chart Type: Scatter Plot

Description:

- This plot shows the relationship between trip_distance and total_amount (which includes fare, tips, tolls, and surcharges).
- Each point represents a single trip.

Key Observations:

- Similar trend as the fare amount: total amount increases with trip distance.
- Variability in total amounts is higher than in fare amounts due to tips and tolls.

Use Case:

Helps understand how additional charges (tips, tolls) impact total revenue for longer trips.

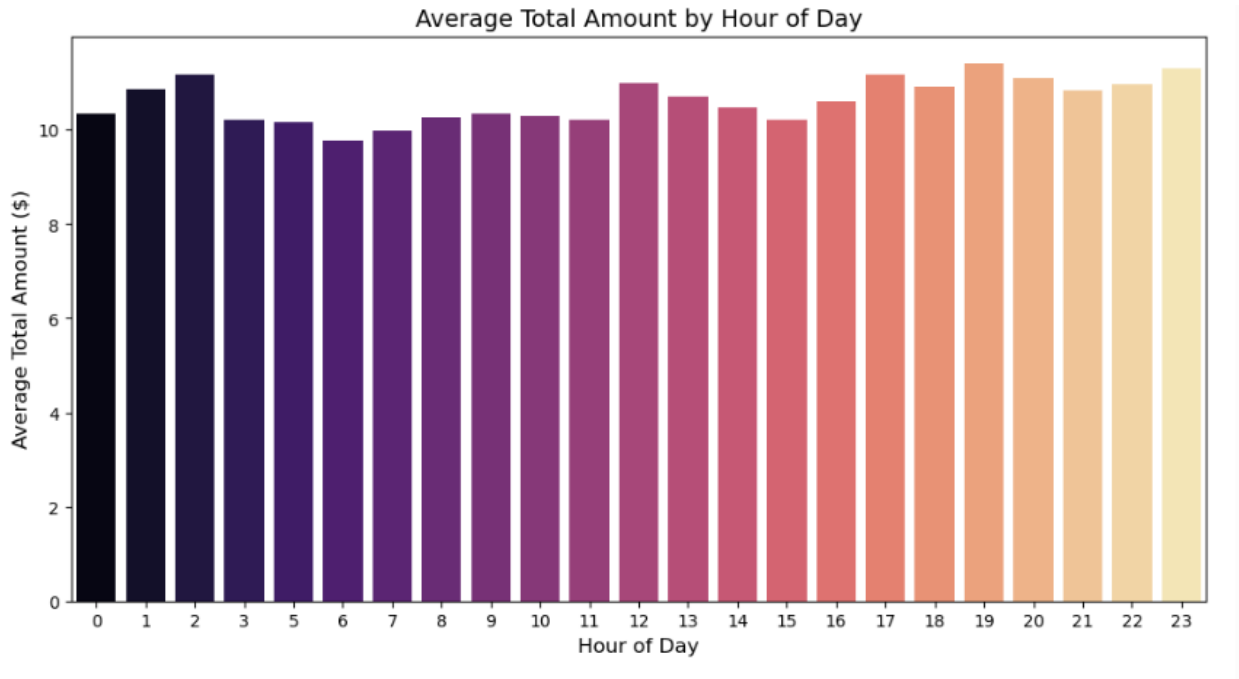

Figure 14: Average total amount by hour of day

Chart Type: Bar Plot

Description:

- This plot shows the average total_amount for each hour of the day (hour_of_day).

Key Observations:

- Late-night hours (12 AM–6 AM) have higher average totals, likely due to tips and night surcharges.
- Early morning and afternoon hours show consistent totals with less variability.

Use Case:

Identifies lucrative time periods for drivers based on total revenue rather than just fare amounts.

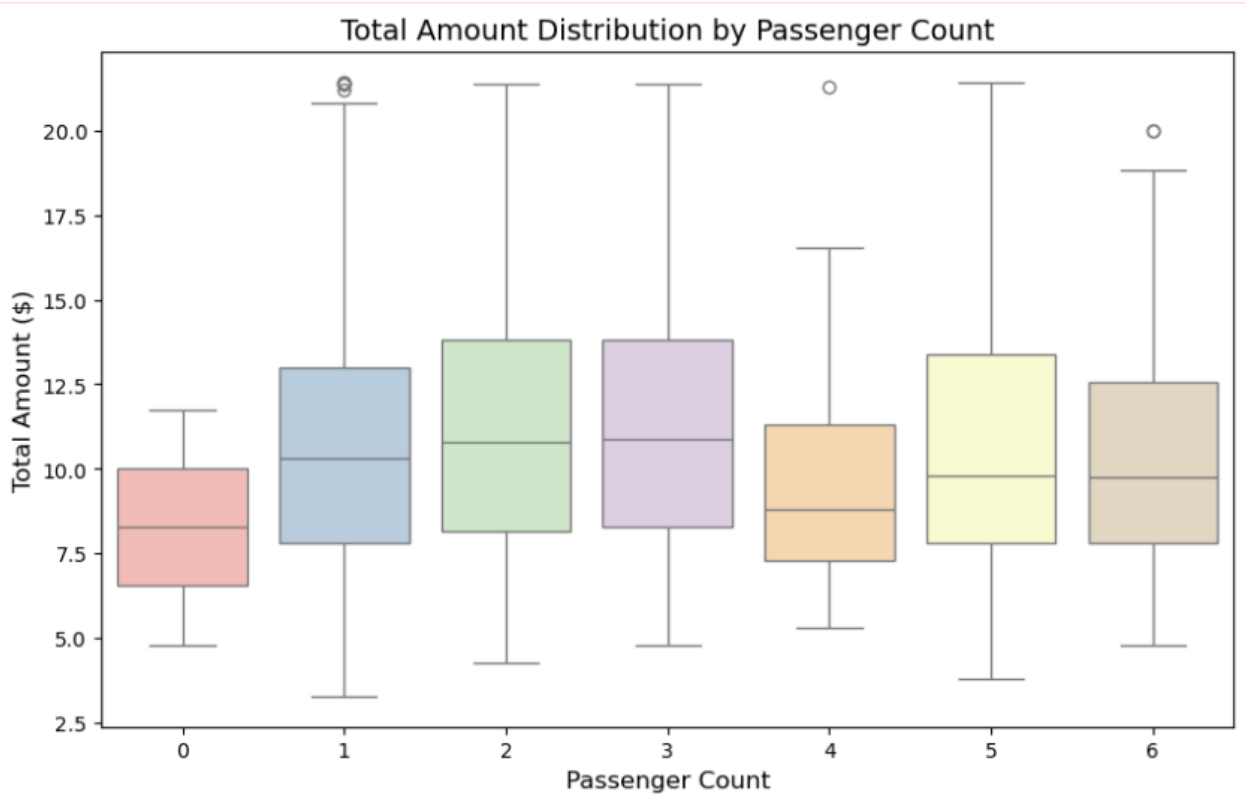

Figure 15: Total amount distribution by pasenger count

Chart Type: Box Plot

Description:

- This plot compares the distribution of total_amount for different passenger_count.
- The box plot shows the median, quartiles, and outliers for each passenger count category.

Key Observations:

- Single-passenger rides dominate the dataset and show a wide range of total amounts due to varying trip distances and tipping behavior.
- Higher passenger counts generally have slightly higher total amounts but fewer data points.

Use Case:

Useful for analyzing how group size impacts revenue distribution.

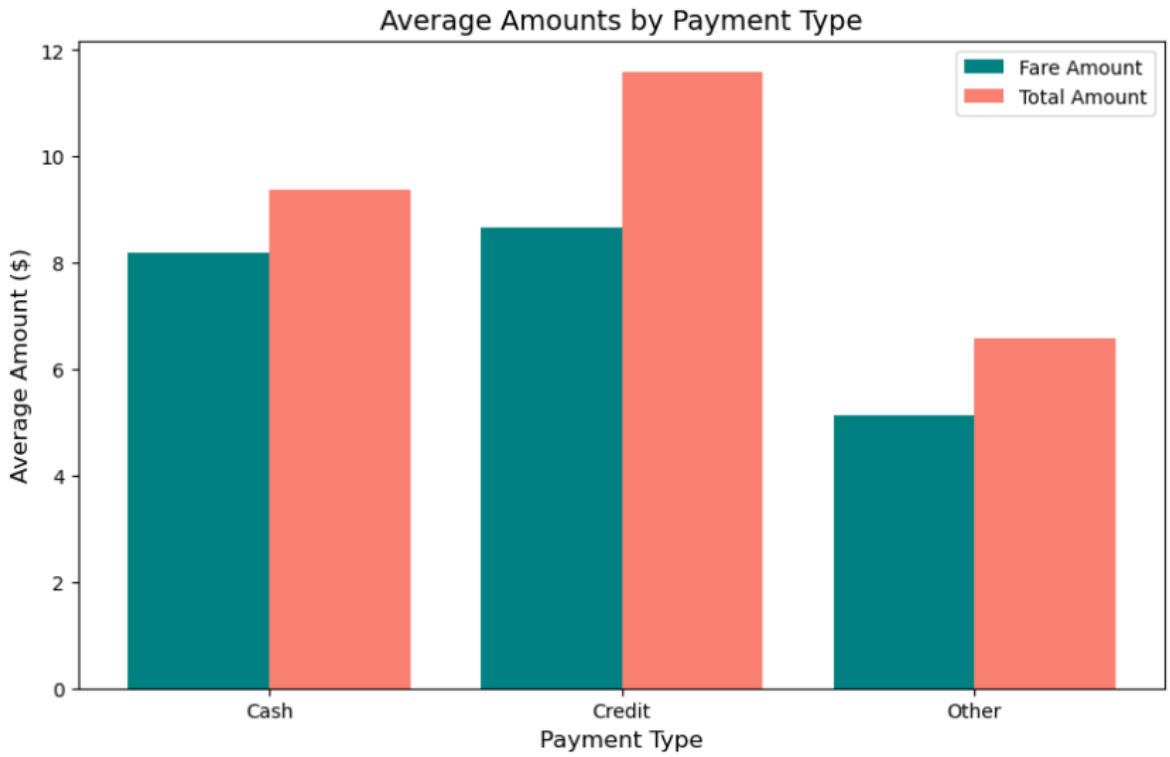

Figure 16: Average amounts by payment type

Chart Type: Grouped Bar Plot

Description:

- This plot compares the average fare_amount and total_amount across different payment types (Cash, Credit, Other).

Key Observations:

- Credit card payments have higher average totals compared to cash payments, likely due to higher tipping rates with credit transactions.

- Cash payments show lower averages but may involve less variability in tipping behavior.

Use Case:

Helps analyze how payment methods impact revenue and tipping trends.

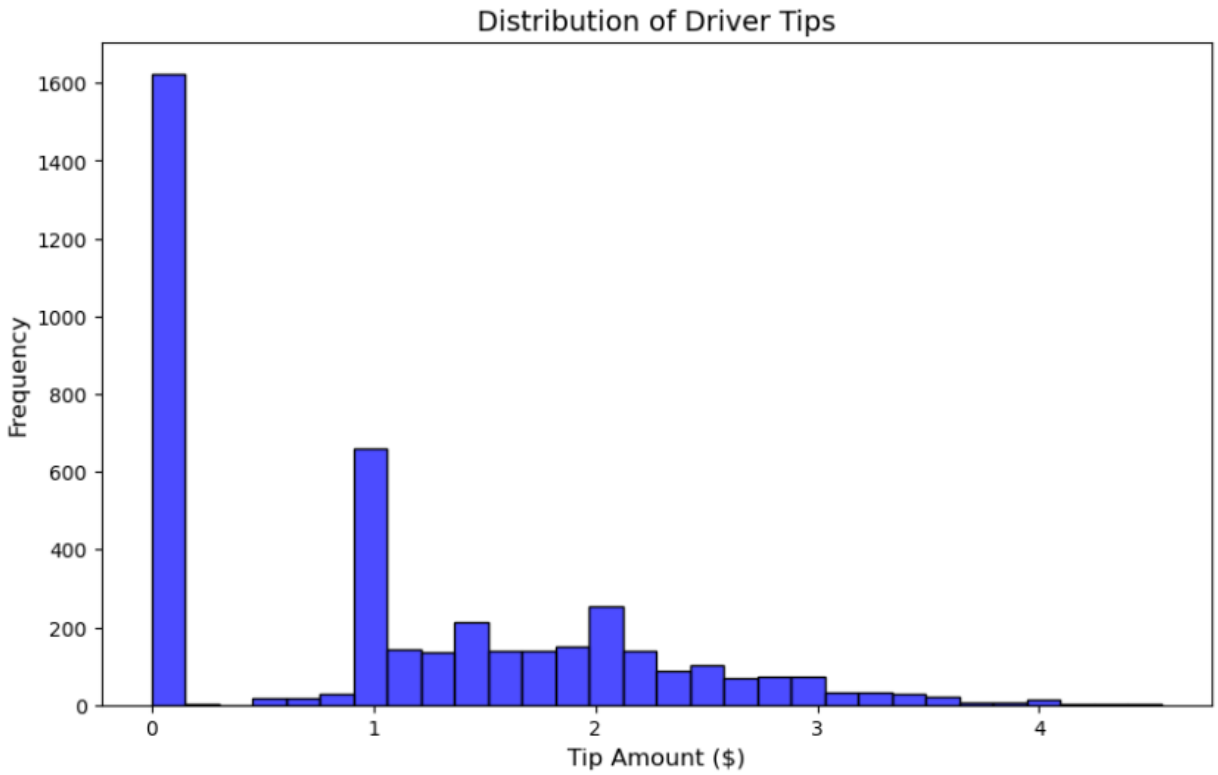

Figure 17: Distribution of driver tips

Chart Type: Histogram

Description:

- This plot visualizes the distribution of tip amounts across all taxi rides.
- The histogram shows the frequency of tip amounts.

Observations:

Central Tendency:

- The median tip amount is approximately $1.00, as indicated by the 50th percentile in the descriptive statistics.
- The mean tip amount is $1.08, slightly skewed by higher tips.

Skewness:

- The distribution is positively skewed, with most tips concentrated between $0 and $2.
- A significant portion of rides (~25%) have no tips, as shown by the peak at $0.

Outliers:

Tips above $4.00 are rare but exist, representing generous tipping behavior.

Use Case:

This plot helps identify tipping patterns and can guide strategies to encourage tipping (e.g., driver behavior, payment incentives)

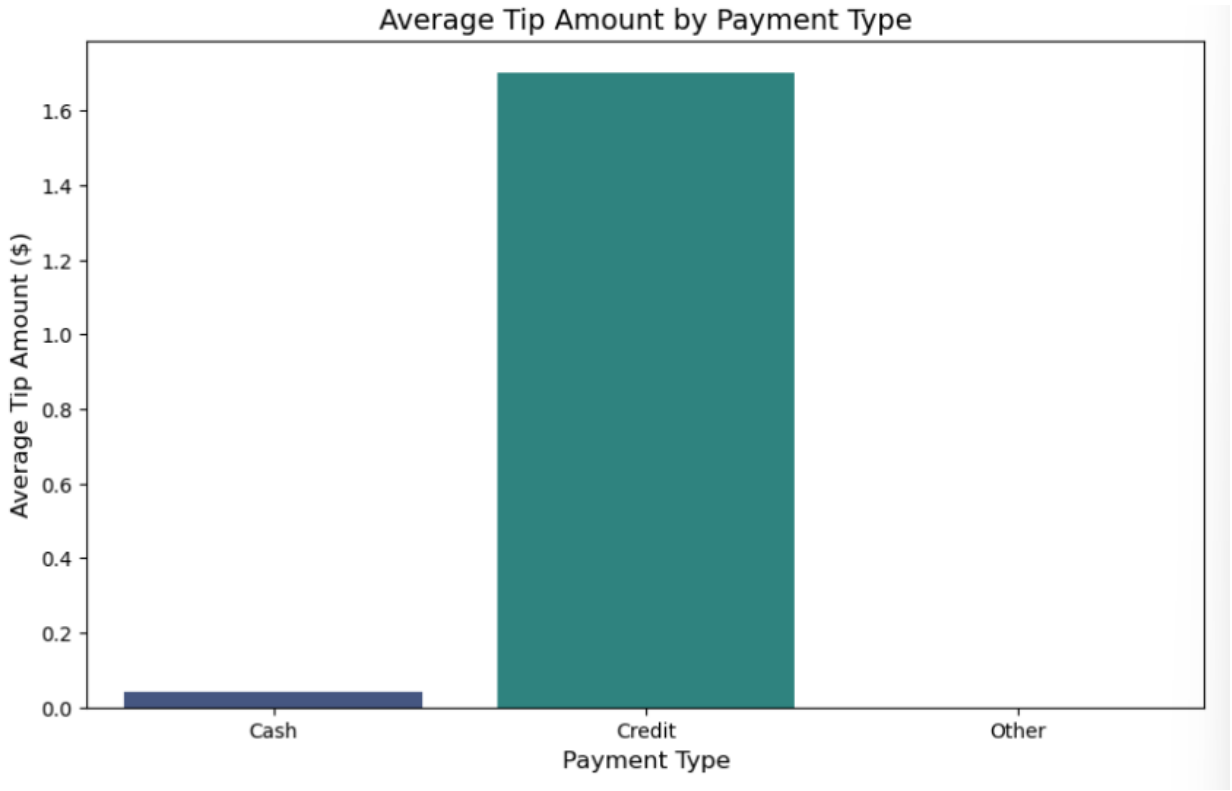

Figure 18: Average tip amount by payment type

Chart Type: Bar Plot

Description:

This plot compares the average tip amounts across different payment methods:

Cash, Credit, and Other.

Observations:

Credit Card Dominance:

- Credit card payments result in significantly higher average tips (0.10) or other payment methods ($0.05).
- This difference highlights a strong correlation between electronic payments and tipping behavior.

Cash Payments:

Cash transactions have minimal tips, possibly due to passengers rounding fares or avoiding tipping altogether.

Other Payment Methods:

"Other" payment types (e.g., vouchers or corporate accounts) show negligible tipping behavior.

Use Case: Insights from this plot can be used to promote credit card usage among passengers to maximize driver revenues from tips.

Model Analysis Report 1

Model: Random Forest Regressor
Target Variable: total_amount (Total fare including tips, tolls, and surcharges)

Code links -

https://drive.google.com/file/d/1Zqva- 4t6Z4IVuzWZzY5r5PiLBz9IQn6y/view?usp=sharing

https://drive.google.com/file/d/1jCUVtUFmaEEGqp3nypJCpvw2UjTXpFo7/view?usp=sharing

https://drive.google.com/file/d/1lLZD4FzZwl2tbftOCmwTUuUGBe6MMC3H/view?usp=sharing

Model Implementation Summary

Code Overview:

1. **Data Preparation**:
   - Loaded dataset
   - Engineered trip_duration (minutes) using pickup/dropoff timestamps.
   - Selected features: trip_distance, passenger_count, trip_duration, extra, mta_tax, tolls_amount, improvement_surcharge, and payment type flags (Cash, Credit, Other).
2. **Train-Test Split**:
   - Split data into 80% training and 20% testing sets (test_size=0.2, random_state=42).
3. **Model Training**:
   - Used RandomForestRegressor with 100 trees (n_estimators=100) and random_state=42.
4. **Performance Metrics**:
   - **MAE**: 0.46816961842625593
   - **RMSE**: 0.6738835022941029
   - **R^2 Score**: 0.96617909656200

Model Analysis Report 2

Model: Random Forest Regressor
Target Variable: tip_amount

Code links –
https://drive.google.com/file/d/14lMN2HssQmH0Sbv3IzKBWLjYB_FpTrXk/view?usp=sharing

https://drive.google.com/file/d/1eVTDIjy4D_Dn8FG6IcCCF21wOiCtn5ku/view?usp=sharing

Model Implementation Summary

1. Data Preparation:

- **Dataset**: Loaded dataset
- **Feature Engineering**:
    - Created trip_duration (minutes) using pickup/dropoff timestamps.
    - Selected features:
        - **Trip metrics**: trip_distance, passenger_count, trip_duration
        - **Fare components**: total_amount
        - **Payment types**: Binary flags (Cash, Credit, Other)

2. Train-Test Split:

- **Split**: 80% training, 20% testing (test_size=0.2, random_state=42)

3. Model Training:

1) **Algorithm**: LinearRegression

   Parameters: random_state=42

   Performance Metrics:
   i. **MAE**: 0.3917400319695755
   ii. **RMSE**: 0.5316004050345514
   iii. **R² Score**: 0.7200975152801417

2) **Algorithm**: Random Forest Regressor

Parameters: n_estimators=100, random_state=42

Performance Metrics:

i. **MAE**: 0.3689422669417716
ii. **RMSE**: 0.3689422669417716
iii. **R² Score**: 0.8651804195576247

Why should we continue with Random Regressor?

1. RMSE (Root Mean Squared Error)

- Best for: Situations where large errors are particularly undesirable (e.g., financial predictions, medical diagnostics).
- Why: RMSE penalizes large errors more heavily due to squaring, making it sensitive to outliers.
- Ideal Value: A lower RMSE indicates better model performance.

2. MAE (Mean Absolute Error)

- Best for: Situations where all errors are treated equally, regardless of their magnitude.
- Why: MAE is easier to interpret as it represents the mean absolute difference between predicted and actual values.
- Ideal Value: A lower MAE indicates better model performance.

3. R^2 Score (Coefficient of Determination)

- Best for: Measuring how well the model explains the variability in the target variable.
- Why: An R^2 score closer to 1 indicates that the model captures most of the variance in the data.

Comparison: Which Metric Indicates Better for Model?

1. If minimize prediction errors:
    - Focus on RMSE or MAE.
    - Choose RMSE if large errors are critical; otherwise, MAE provides a simpler interpretation.

2. If evaluate how well your model explains variability:
    - Focus on R^2 score.

For regression tasks with sensitive predictions: prioritize RMSE/MAE.

we are Performing statistical analyses on Regression so according to above given values we select Random Forest Regression

Analysis of Taxi Fare Prediction Models:

This analysis examines two machine learning models developed to predict different aspects of taxi fare transactions. Both models employ Random Forest Regression techniques but focus on distinct target variables relevant to the taxi industry. The following report details the structure, functionality, and benefits of each model separately.

Tip Amount Prediction Model

The Tip Amount Prediction Model is designed to estimate the gratuity a passenger might provide to a taxi driver based on various trip-related factors. This model utilizes machine learning to identify patterns in tipping behavior and produce accurate predictions.

Model Architecture and Features

The Tip Amount Model employs a Random Forest Regressor algorithm, which combines multiple decision trees to create a robust predictive model. The model incorporates several key features:

- Trip distance
- Passenger count
- Trip duration (calculated from pickup and dropoff timestamps)
- Fare amount
- Extra charges
- MTA tax
- Tolls amount
- Improvement surcharge
- Payment type (categorized as Cash, Credit, or Other)1

The model splits the dataset into training (80%) and testing (20%) components to ensure proper validation of its predictive capabilities. By using the Random Forest algorithm, the model can capture complex, non-linear relationships between these features and the tip amount.

Applications and Use Cases

This model can serve multiple purposes within the taxi industry and related services:

- Helping drivers understand factors that influence tipping behavior
- Enabling taxi companies to forecast revenue more accurately
- Assisting ride-sharing platforms in optimizing driver earnings
- Supporting financial planning for transportation service providers

Total Amount Prediction Model

The Total Amount Prediction Model focuses on forecasting the complete fare a passenger will pay, including base fare, taxes, surcharges, and potential tips. This comprehensive approach provides a holistic view of transaction values in taxi services.

Model Architecture and Features

Like the Tip Amount Model, this model also employs a Random Forest Regressor algorithm. However, it uses a slightly different set of features:

- Trip distance
- Passenger count
- Trip duration (calculated from pickup and dropoff timestamps)
- Payment type (categorized as Cash, Credit, or Other)

Notably, the Total Amount Model excludes the fare amount as a feature since this would be directly correlated with the target variable (total amount). The model similarly employs an 80/20 train-test split with a random state of 42 to ensure reproducibility and reliable evaluation.

Applications and Use Cases

The Total Amount Model serves several critical functions:

- Enabling taxi services to predict overall revenue
- Helping passengers estimate total trip costs before riding
- Supporting financial forecasting for transportation companies
- Assisting in budget planning for frequent taxi users

Benefits of Random Forest Regression for Taxi Fare Prediction

Both models leverage Random Forest Regression, which offers several significant advantages for these prediction tasks:

- Robustness to Overfitting

Random Forest models mitigate overfitting by aggregating predictions from multiple decision trees. This ensemble approach reduces the risk of individual trees memorizing noise or outliers in the training data, resulting in more generalized models that perform well on unseen data.

- Handling Complex Data Relationships

The taxi industry dataset contains complex relationships between variables that may not follow linear patterns. Random Forest excels at capturing these non-linear relationships, making it particularly suitable for predicting variables like tip amount and total fare which can be influenced by numerous interacting factors.

- Feature Importance Analysis

Both models can automatically evaluate and rank the importance of different features. This capability allows taxi companies to understand which factors most significantly influence tipping behavior and total fare amounts, enabling data-driven business decisions.

- Resistance to Outliers and Noise

Taxi data often contains outliers, such as unusually long trips or exceptionally high fares. Random Forest's ensemble nature makes it robust against such outliers and noisy data points, ensuring reliable predictions even with imperfect data.

- Performance with High-Dimensional Data

The models incorporate multiple features from taxi transactions, creating a high-dimensional dataset. Random Forest handles this complexity efficiently without significant performance degradation, unlike some other algorithms that might struggle with numerous features8.

Conclusion

The Tip Amount and Total Amount prediction models serve as valuable analytical tools for the taxi industry, offering distinct yet complementary insights. The Tip Amount Model focuses on predicting gratuity behavior, helping drivers and companies understand customer tipping patterns and enhance service quality to maximize tips. Meanwhile, the Total Amount Model provides a comprehensive view of transaction values, supporting broader financial planning and forecasting. Both models leverage the Random Forest algorithm, which is well-suited for taxi transaction data due to its ability to handle non-linear relationships, resist overfitting, analyze feature importance, and remain resilient against outliers.

For taxi companies and ride-sharing platforms, implementing these predictive models can enhance revenue forecasting, optimize resource allocation, and improve customer experiences through data-driven service enhancements. The insights derived from these models can inform pricing strategies and marketing initiatives tailored to different customer segments or trip types. Analyzing 5,336 New York City yellow taxi trips from January 2015 reveals critical patterns that can help drivers and companies make strategic operational decisions to maximize revenue. The Random Forest models demonstrated strong predictive power, with R^2 values of 0.96 for total fare prediction and 0.86 for tip prediction. Trip distance emerged as the most significant factor influencing total fare amounts while payment method significantly impacted tipping behavior—credit card transactions consistently generated higher tips compared to cash payments.

A time-based analysis further highlighted high-revenue opportunities during late-night hours (12 AM–6 AM), likely due to night surcharges and longer average trip distances. These findings suggest that drivers can increase earnings by adjusting their schedules to align with peak revenue periods and positioning themselves strategically for longer trips. The striking disparity in tipping behavior between payment methods underscores the importance of encouraging electronic payments to maximize gratuities. Contrary to common assumptions, passenger count had minimal influence on total fare amounts, reinforcing the idea that other trip characteristics play a more crucial role in revenue determination.

The feature importance analysis provided further insights into the interactions between fare components and trip characteristics, demonstrating the complex factors that drive final revenue outcomes. These insights equip drivers with actionable intelligence to optimize their schedules, target high-value trips, and encourage preferred payment methods. For industry stakeholders and ride-hailing platforms, these findings offer valuable guidance for algorithm development, driver incentive structures, and pricing strategies. Future research should integrate additional temporal features, customer demographics, and real-time traffic conditions to further

refine prediction models and enhance revenue optimization strategies in an increasingly competitive urban transportation landscape. By leveraging machine learning insights, the taxi industry can better adapt to evolving customer behaviors, optimize operational efficiencies, and ultimately drive greater profitability.

Reference

Kaggle (2019). *Datasets | Kaggle*. [online] Kaggle.com. Available at https://www.kaggle.com/datasets.

scikit-learn (2018). *3.2.4.3.2. sklearn.ensemble.RandomForestRegressor — scikit-learn 0.20.3 documentation*. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html.