Uber fares prediction

Objective

To develop a predictive model that accurately estimates Uber fares based on historical trip data, enabling more precise fare calculations and improved pricing strategies.

Background

Uber's dynamic pricing model is influenced by various factors, including trip distance, time of day, passenger demand, and local market conditions. Accurately predicting the fare amount is crucial for the company to maintain competitiveness and ensure customer satisfaction. The model aims to leverage data-driven insights to forecast fare amounts, helping to optimize pricing and enhance service efficiency.

Problem Statement

The challenge is to analyze the historical trip data and extract meaningful patterns and relationships that contribute to the fare amount. The project involves preprocessing the data, performing exploratory data analysis (EDA), engineering relevant features, and applying machine learning models to predict the fare amount. The goal is to identify the model that best captures the complexities of fare determination and provides the most accurate predictions.

Dataset Description

The dataset comprises historical trip data from Uber, which includes details about each trip, such as pickup and dropoff locations, timestamps, fare amounts, and the number of passengers.

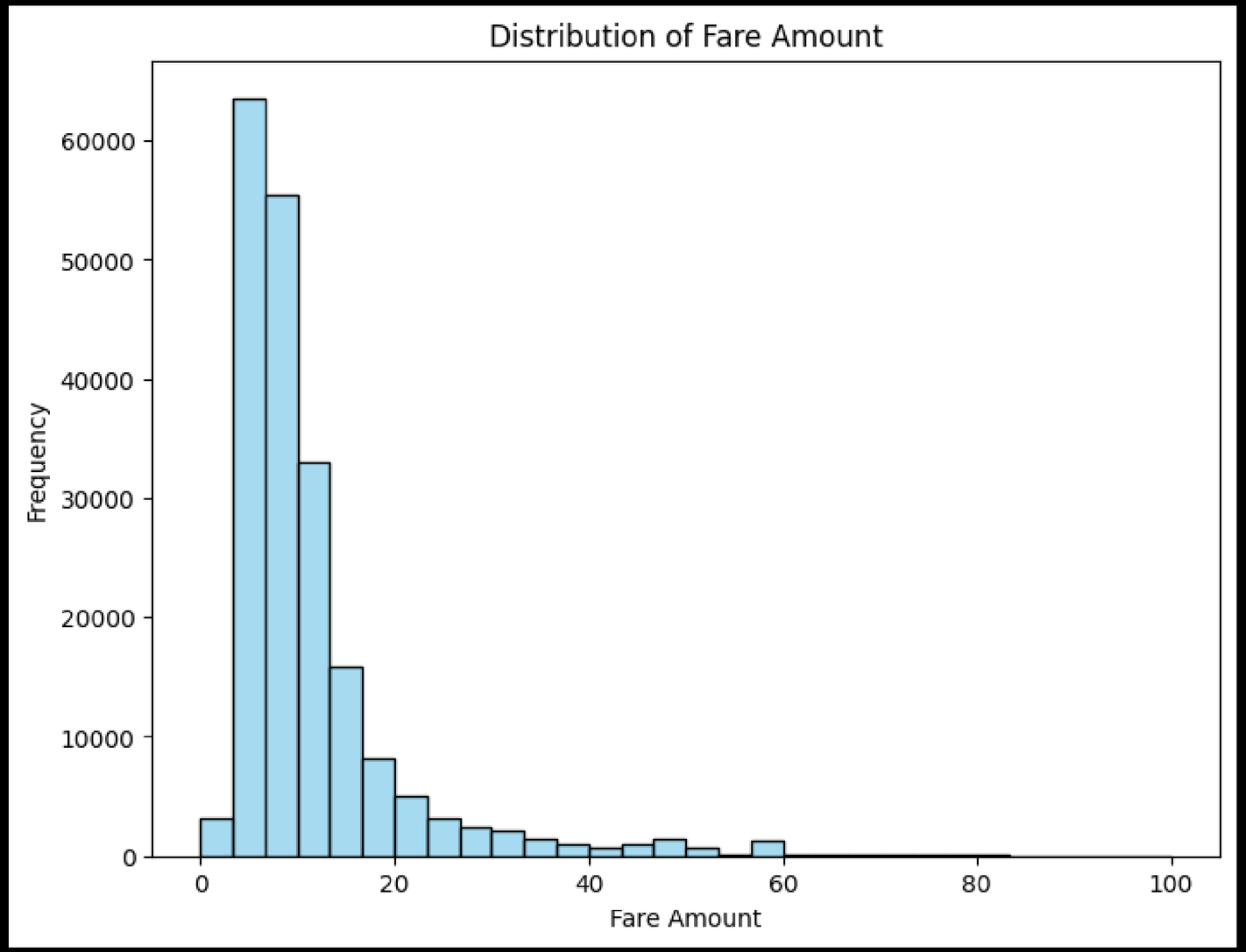
Features

- fare_amount: The fare of the trip in USD.
- pickup_datetime: The date and time when the trip started.
- pickup_longitude and pickup_latitude: Geographical coordinates of the trip's starting point.
- dropoff_longitude and dropoff_latitude: Geographical coordinates of the trip's ending point.
- passenger_count: The number of passengers on the trip.

Preprocessing Steps

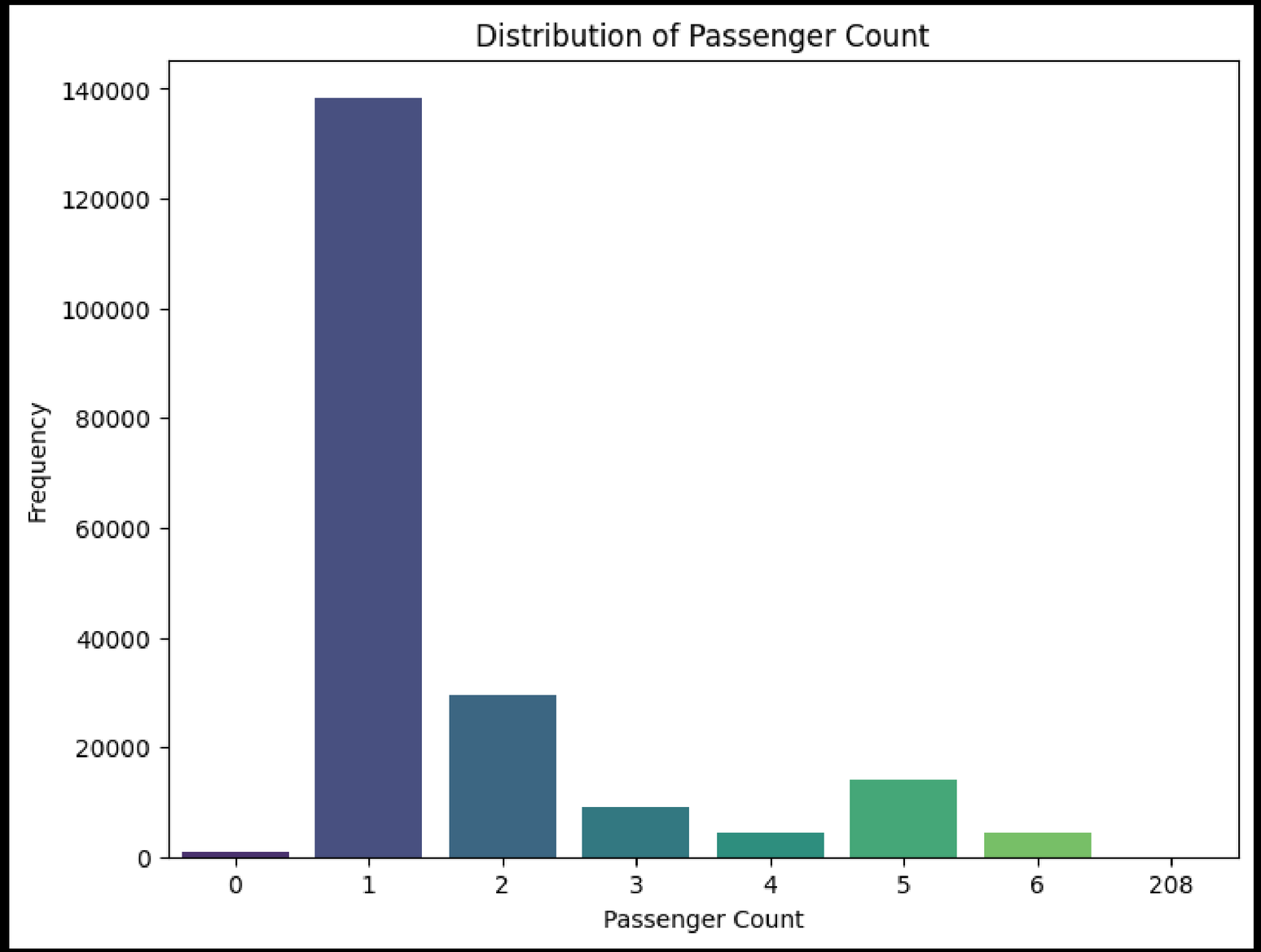
- Cleaning: Removed records with missing values and corrected anomalies in location data (e.g., coordinates outside valid ranges).
- Outlier Handling: Identified and removed outliers in fare amounts, focusing on realistic ranges to ensure accurate analysis and modeling.
- Feature Engineering: Calculated trip distance using pickup and dropoff coordinates, and extracted time-based features like the hour of the day and day of the week from the pickup datetime.

Distribution of Fare Amount



Our exploratory data analysis provided valuable insights into Uber's fare structure. The distribution of fare amounts indicates a concentration of trips within a lower fare range, highlighting a high frequency of short to medium-distance trips. This right-skewed distribution suggests that while there are outliers with high fares, possibly due to longer trips or premium services, the majority of users are utilizing Uber for more economical rides.

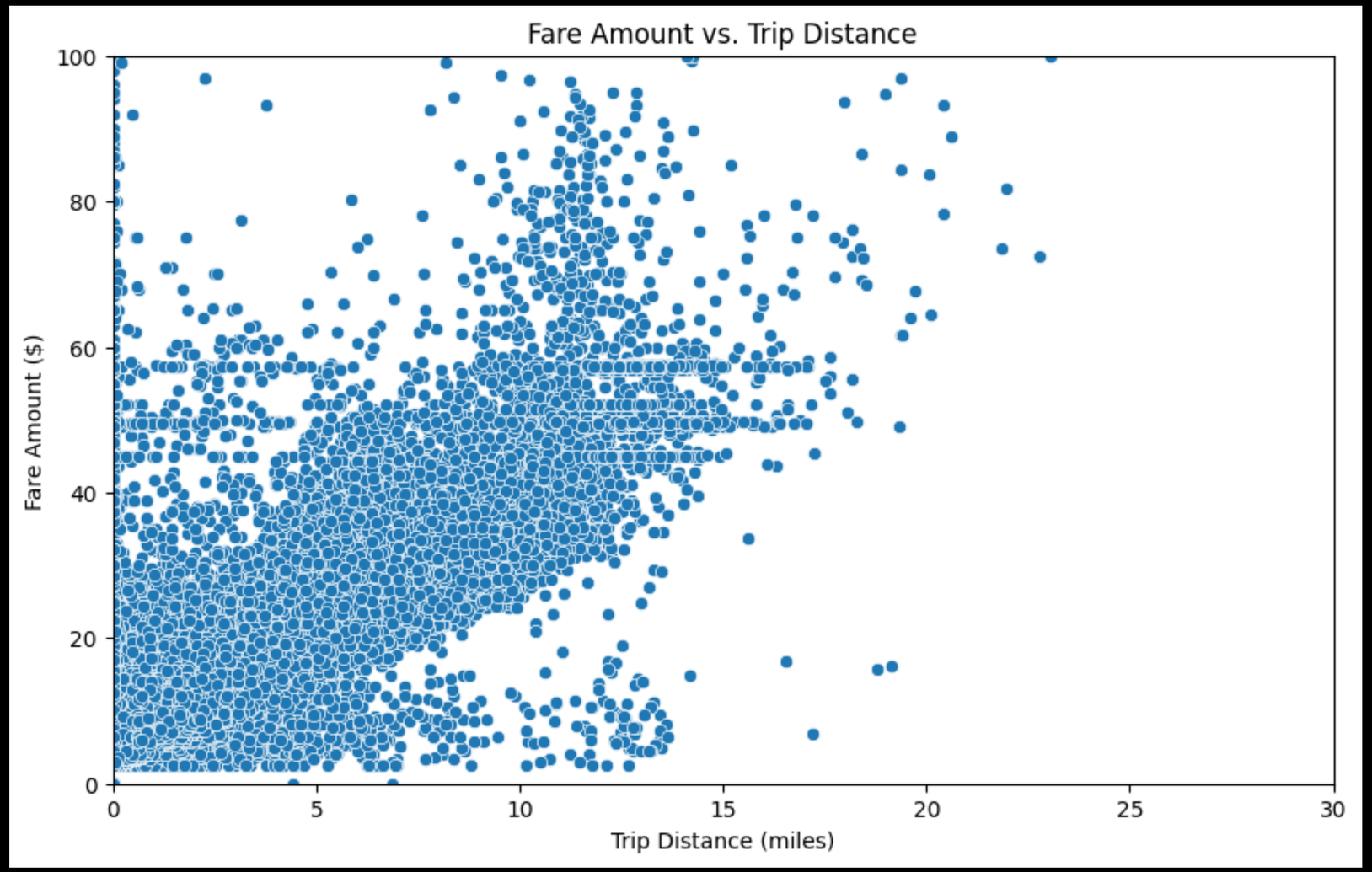
Distribution of Passenger Count



- The analysis of passenger count reveals a dominant trend of solo riders, with a significantly higher number of trips made by individuals compared to groups. This trend may reflect a user preference for personal travel or indicate that Uber’s service is commonly used for commuting purposes, where individual travel is more frequent.

- The infrequent higher passenger counts could correspond to group travel, which appears to be a less common use case. This insight could be pivotal for Uber to tailor their marketing and operational strategies to enhance the experience for solo riders while also exploring opportunities to better accommodate groups, potentially increasing market share in this segment.

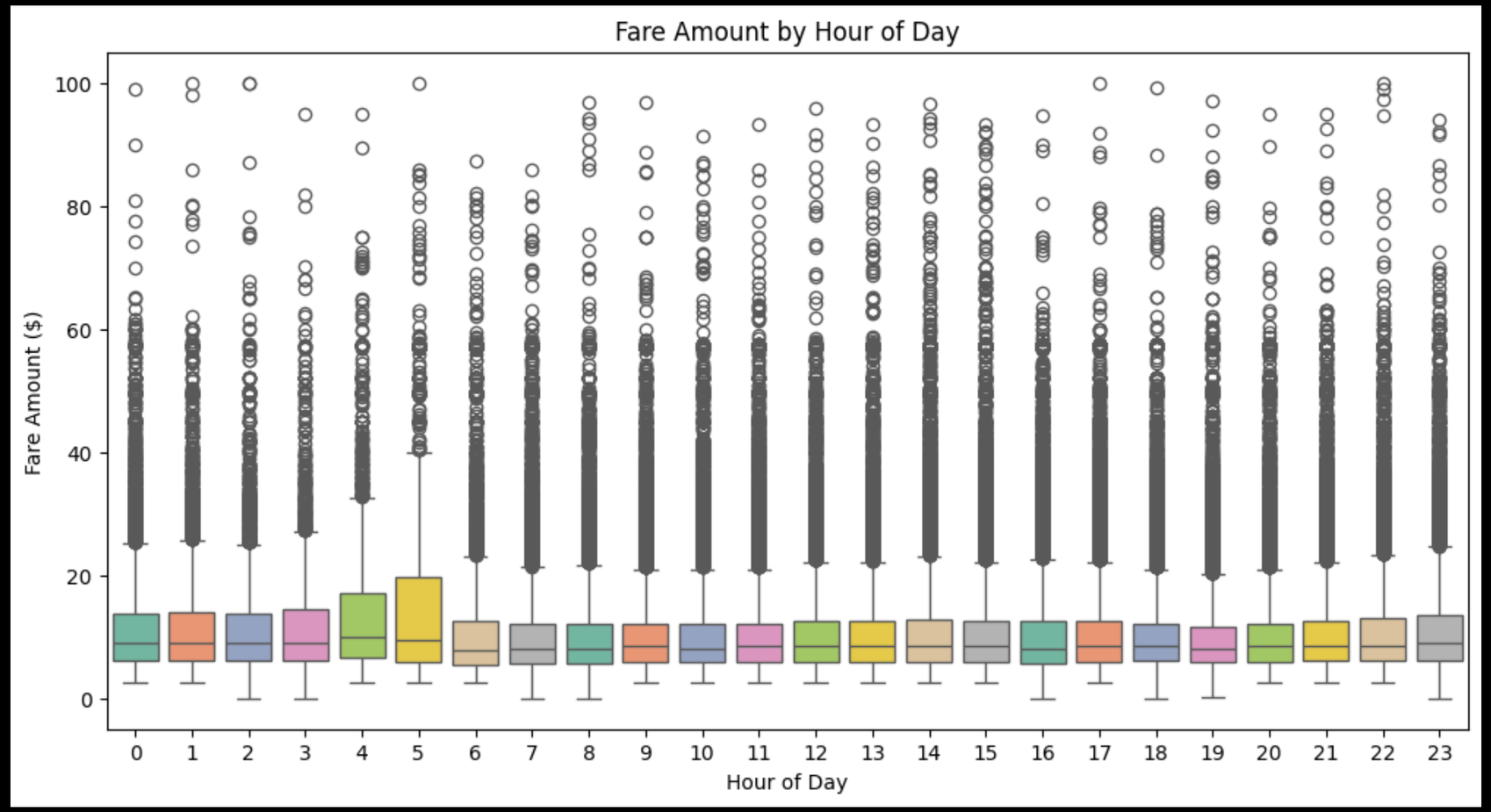
Fare Amount vs. Trip Distance



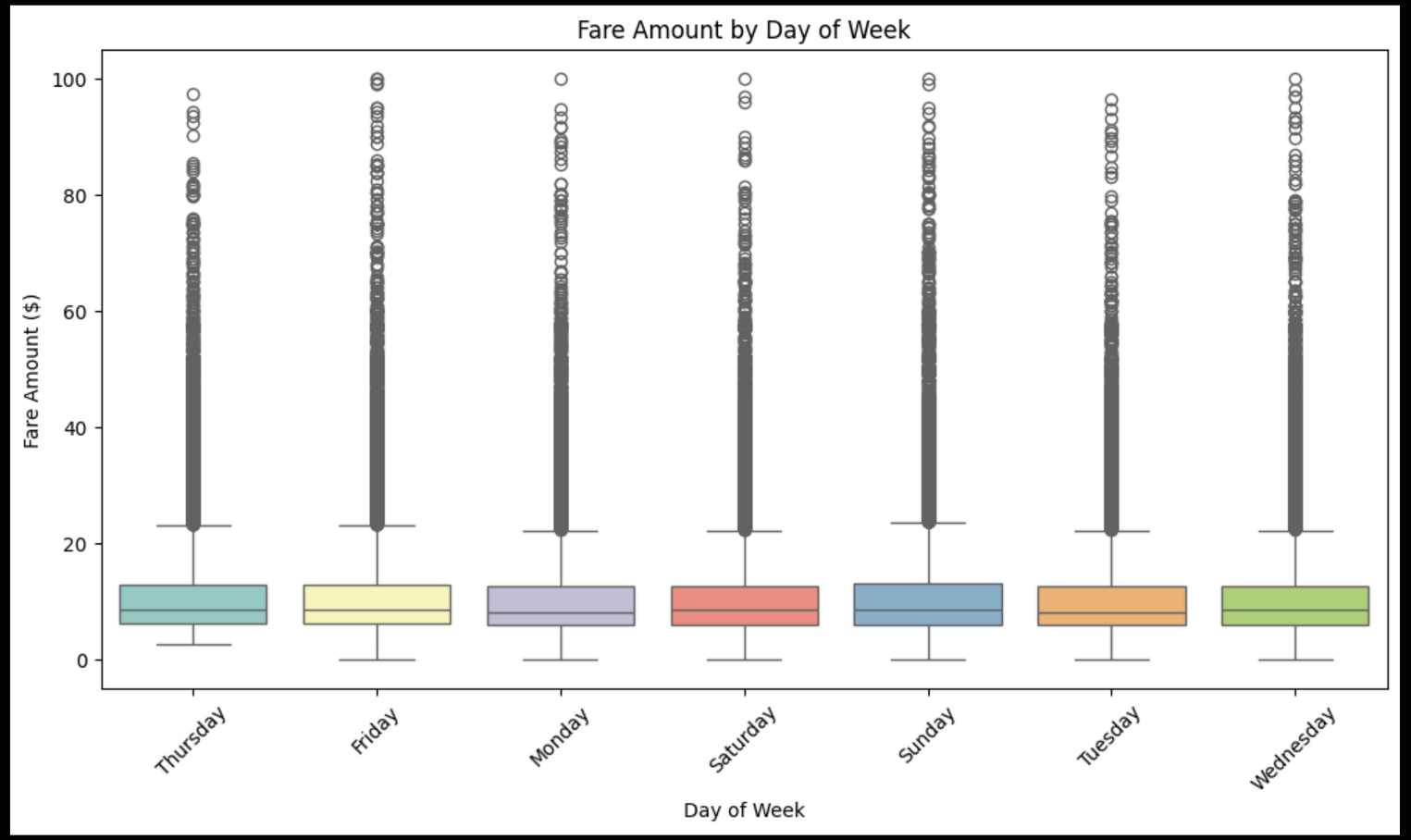
- The scatter plot of fare amounts versus trip distance clearly indicates a positive correlation between the two variables, confirming the intuitive understanding that longer trips generally result in higher fares. The spread of data points suggests a non-linear relationship, with fare increases diminishing in rate beyond a certain trip length.

- This pattern is important for Uber's pricing algorithm, indicating that while distance is a primary driver of fare cost, other factors may influence the final price, such as traffic conditions, surge pricing, and route taken. For more accurate fare predictions, a model that accounts for these complexities would be beneficial, supporting Uber in maintaining fair pricing and transparency with customers.

Fare Amount by Hour of Day Conclusion



The box plot variation across different hours of the day illustrates the dynamic nature of fare pricing. Notably, fares tend to be higher during late-night and early-morning hours, which may reflect a premium for off-peak travel times



- Workweek Trends: Median fare amounts from Monday to Friday remain relatively consistent, mirroring the routine nature of workweek commutes.
- Weekend Fluctuations: Increased fare variability on Saturdays and Sundays suggests a diversity in travel purposes, potentially longer leisure trips or special events.
- Daily Outliers: High fare outliers are present every day, indicating sporadic long-distance travel or surge pricing events, independent of the weekday.
- Strategic Implications: These observations can inform demand forecasting, driver allocation, and promotional offers to optimize service throughout the week.

Feature Engineering

Trip Distance:

- Calculated from the geographical coordinates using the Haversine formula.
- Rationale: Distance is a fundamental factor affecting fare price, as longer trips typically cost more.

Time of Day:

- Categorized into ‘Morning’, ‘Afternoon’, ‘Evening’, and ‘Night’ based on pickup_hour.
- Rationale: Fare prices may vary throughout the day with demand, capturing peak and off-peak periods.

Day of the Week:

- Extracted from pickup_datetime to identify the weekday of each trip.
- Rationale: Travel patterns can differ by day, affecting demand and fare prices.

Passenger Count Categories :

- Transformed into categorical variables to reflect individual, small group, or large group travel.
- Rationale: The number of passengers can influence fare if pricing differs by car size or service type.

Purpose of Feature Engineering:

- To improve the model's ability to predict fares by incorporating factors known to influence pricing.
- To capture both linear and non-linear relationships within the data for enhanced accuracy.

Model Building and Comparison

Evaluated Models:

Random Forest Regressor:

- A robust ensemble model that uses multiple decision trees to produce a more accurate and stable prediction.
- Pros: Good for handling non-linear data, resistant to overfitting, and provides feature importance.
- Cons: Can be computationally intensive, less interpretable due to ensemble nature.

Gradient Boosting Regressor:

- An advanced ensemble technique that builds trees sequentially, with each tree learning to correct the errors of the previous one.
- Pros: Often provides superior predictive accuracy, handles various types of data well.
- Cons: More prone to overfitting and can be sensitive to noisy data and outliers.

Performance Metrics:

Random Forest:
RMSE: 4.875
R^2: 0.729

Gradient Boosting:
RMSE: 4.528
R^2: 0.766

Model Selection:

- The Gradient Boosting Regressor was chosen as the final model based on its lower RMSE and higher R^2 value, indicating better overall predictive accuracy and fit to the data compared to the Random Forest model.
- The decision was also influenced by the model's ability to capture complex patterns in the data, which can be particularly useful in dynamic pricing environments like Uber's.

Key Findings:

- Predictive Modeling: Gradient Boosting Regressor outperformed Random Forest in predicting Uber fares with higher accuracy.
- Data Insights: EDA revealed critical factors influencing fare prices, including trip distance and time-based patterns.

Business Implications:

- Pricing Strategy: The models suggest opportunities for dynamic pricing adjustments based on time of day and trip distances to maximize profitability.
- Customer Segmentation: Identified trends can guide tailored services for solo riders versus groups, optimizing resource allocation and enhancing user experience.

Recommendations for Future Work:

- Data Enrichment: Incorporate additional contextual data like weather, special events, and traffic patterns to refine predictions.
- Model Exploration: Experiment with advanced machine learning techniques, such as deep learning, to capture more complex relationships.
- Continuous Improvement: Implement an iterative process for model updates with new data, ensuring the model remains relevant over time.