Uber Fare Price Prediction

By~ Ashutosh Rai

Project Roadmap

Exploratory Data Analysis

Loading Data

Initially there are total 200000 rows and 9 columns in Dataset

Data Cleaning and Preparation

Handling Missing Values: Identifying and imputing or removing missing values.
Data Types and Conversion: Ensuring all data types are appropriate for analysis.

Descriptive Statistics

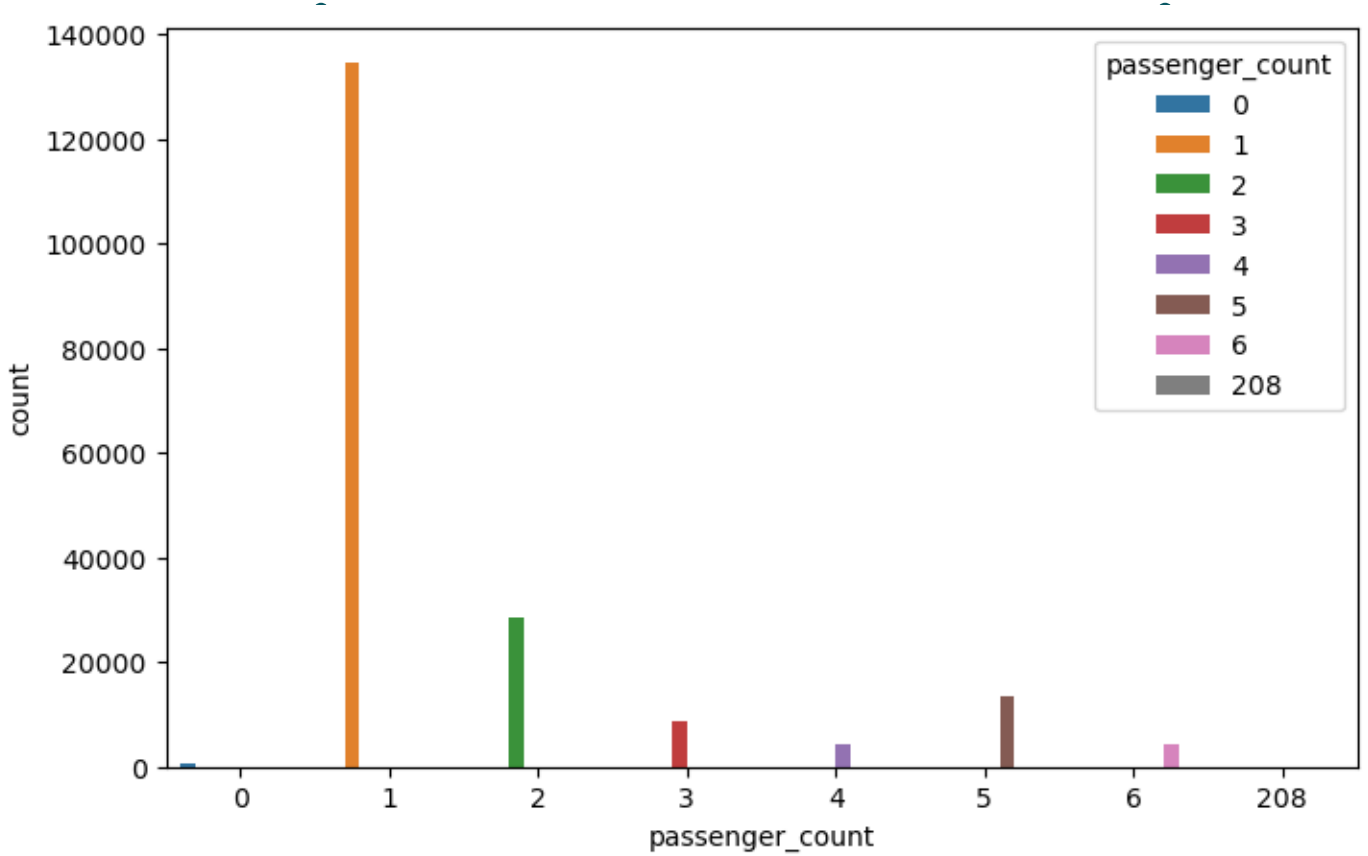
Summary Statistics: Calculating mean, median, mode, standard deviation, etc., .Distribution Analysis: Plotting histograms or density plots to understand distribution.

Identifying Patterns and Relationships

Correlation Analysis: Calculating correlation coefficients to identify linear relationships between features. violin plots..

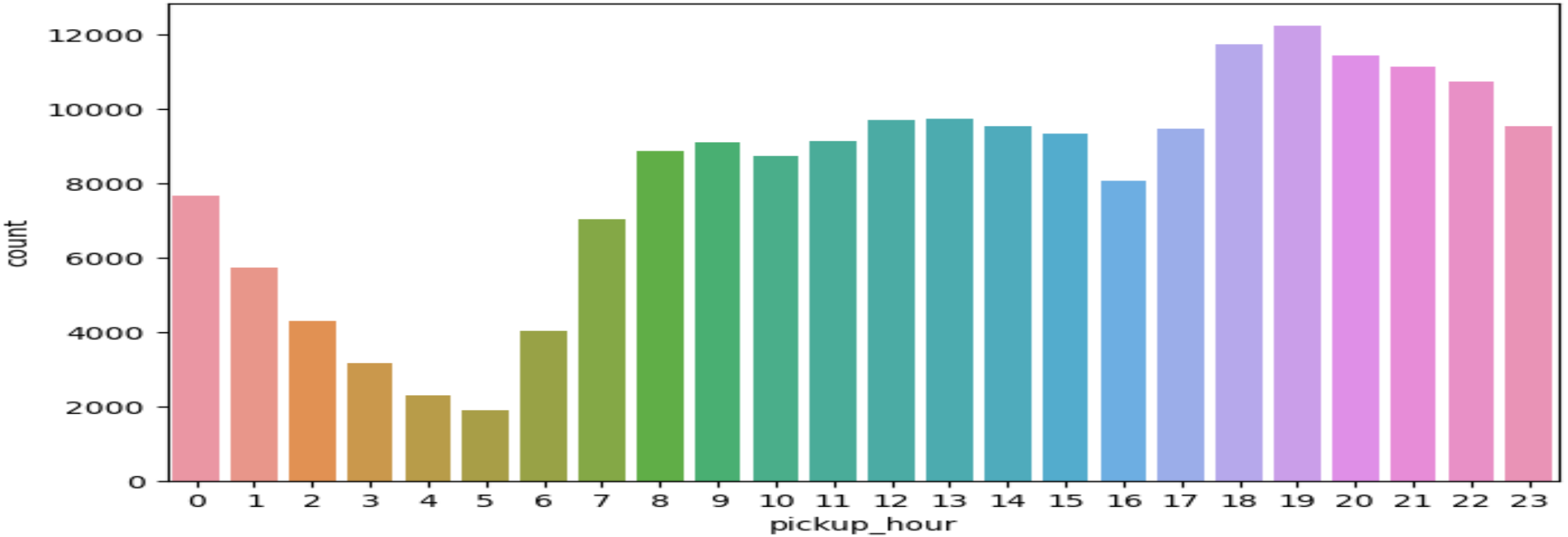
Feature Engineering and Selection

Creating New Features: Generating new features based on existing ones to improve model performance..



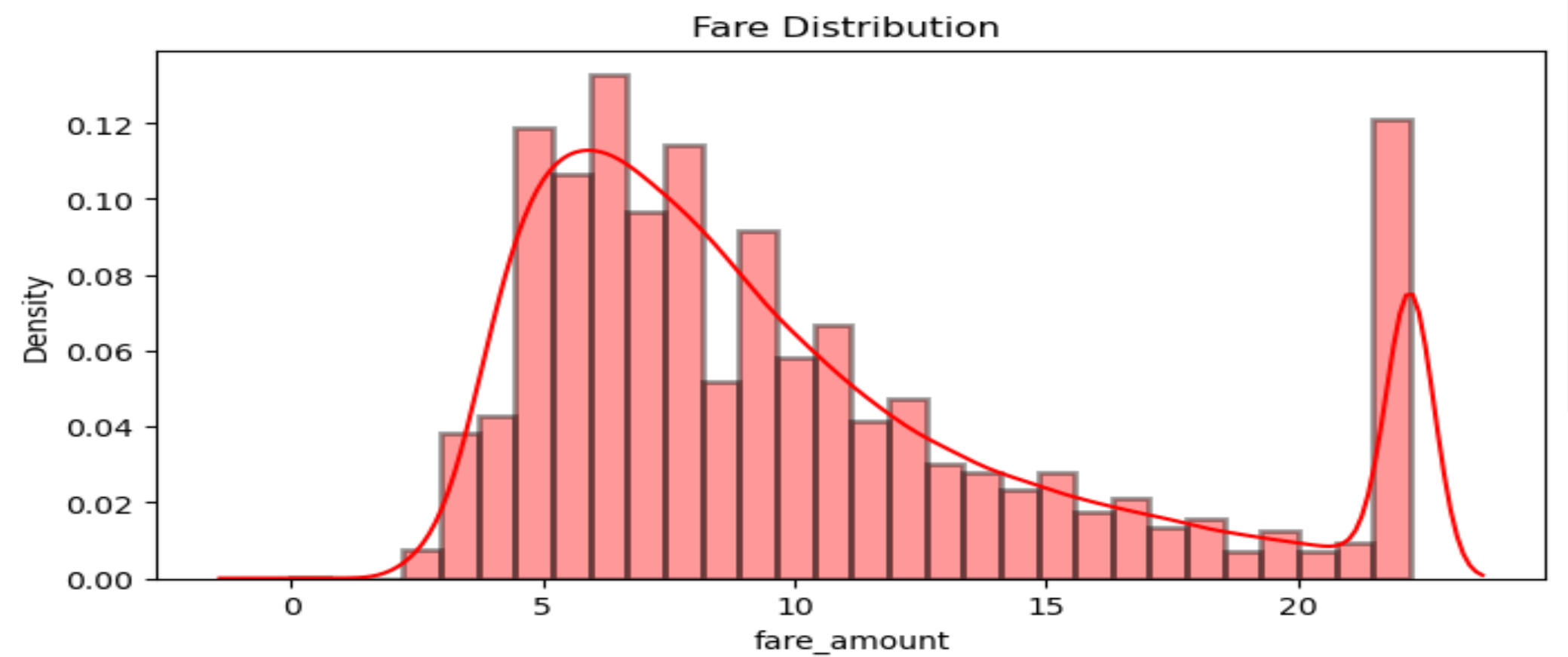
Most of the passenger are traveling alone.

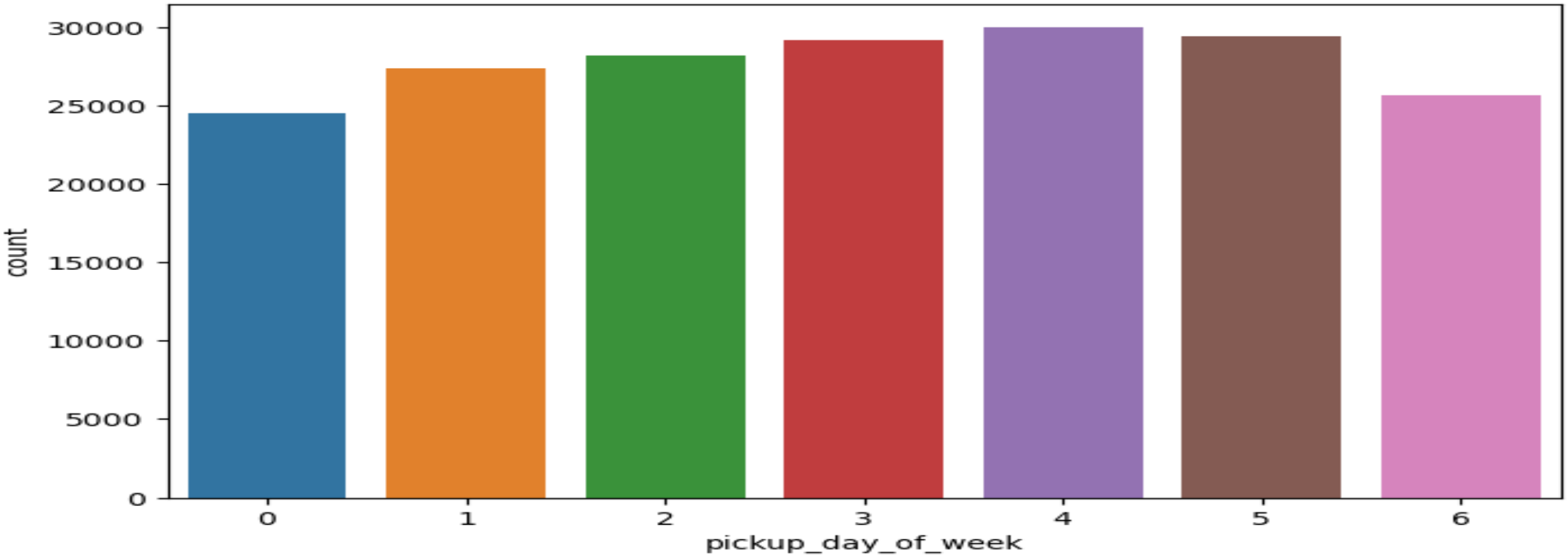
We have one value as 208 in 'passenger_count'



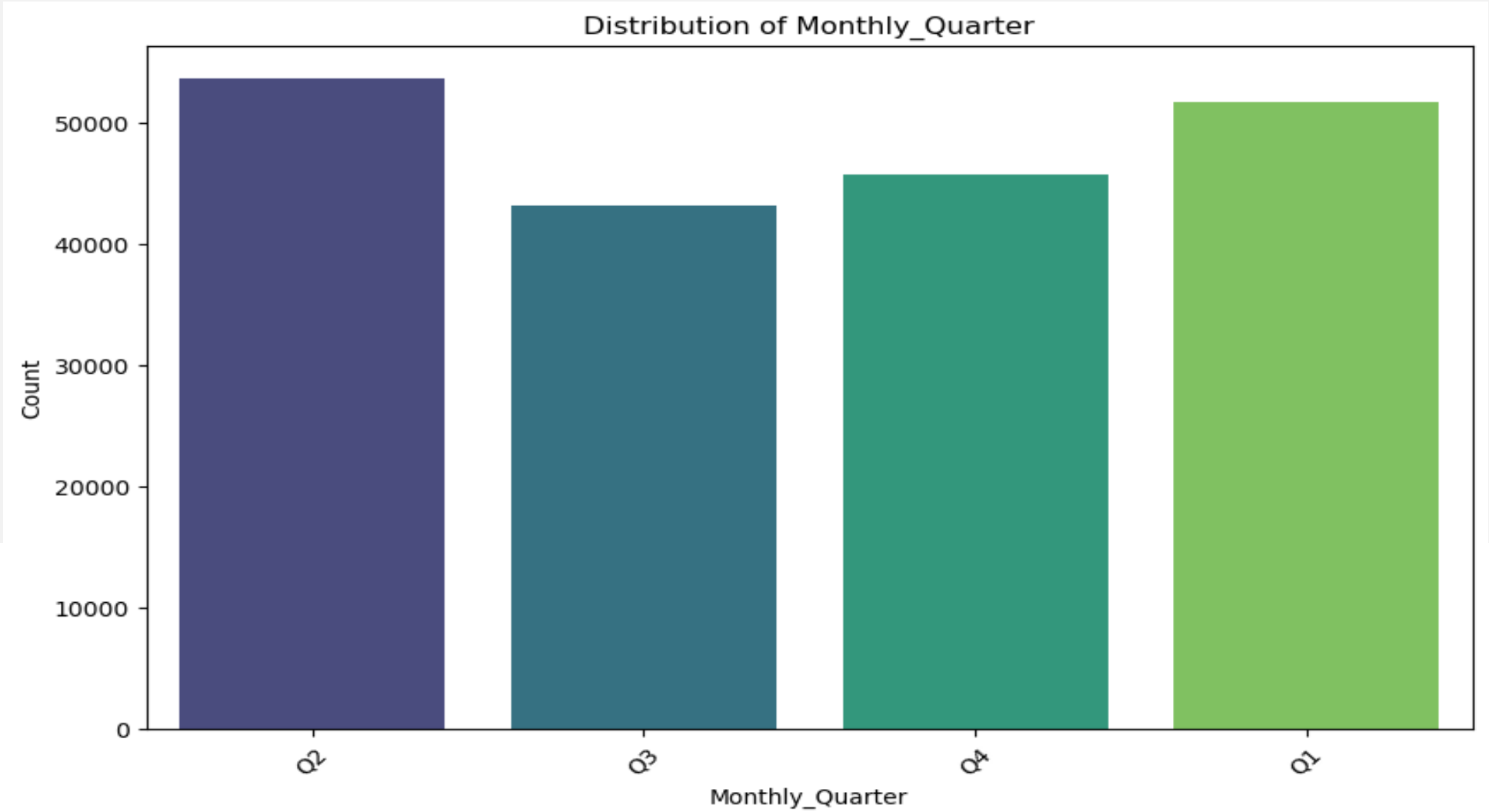
- Peak Hour at 19:00: This is our busiest time.
- Low Bookings at 00:00: Midnight sees fewer bookings. We can attract more customers with targeted coupons.
- 1:00-7:00 Low Activity: Late-night bookings are low, likely due to the timing. We should offer incentives to boost usage during these hours.
- 8:00-17:00 Improvement Potential: Bookings are low during these hours, but there's significant potential for growth. Consider offering new coupons for office and school commuters, and promote carpooling services.
- Decline from 20:00-23:00: Customer numbers drop steadily, likely due to longer wait times. Focusing on this period could improve overall performance.

Fare Distribution

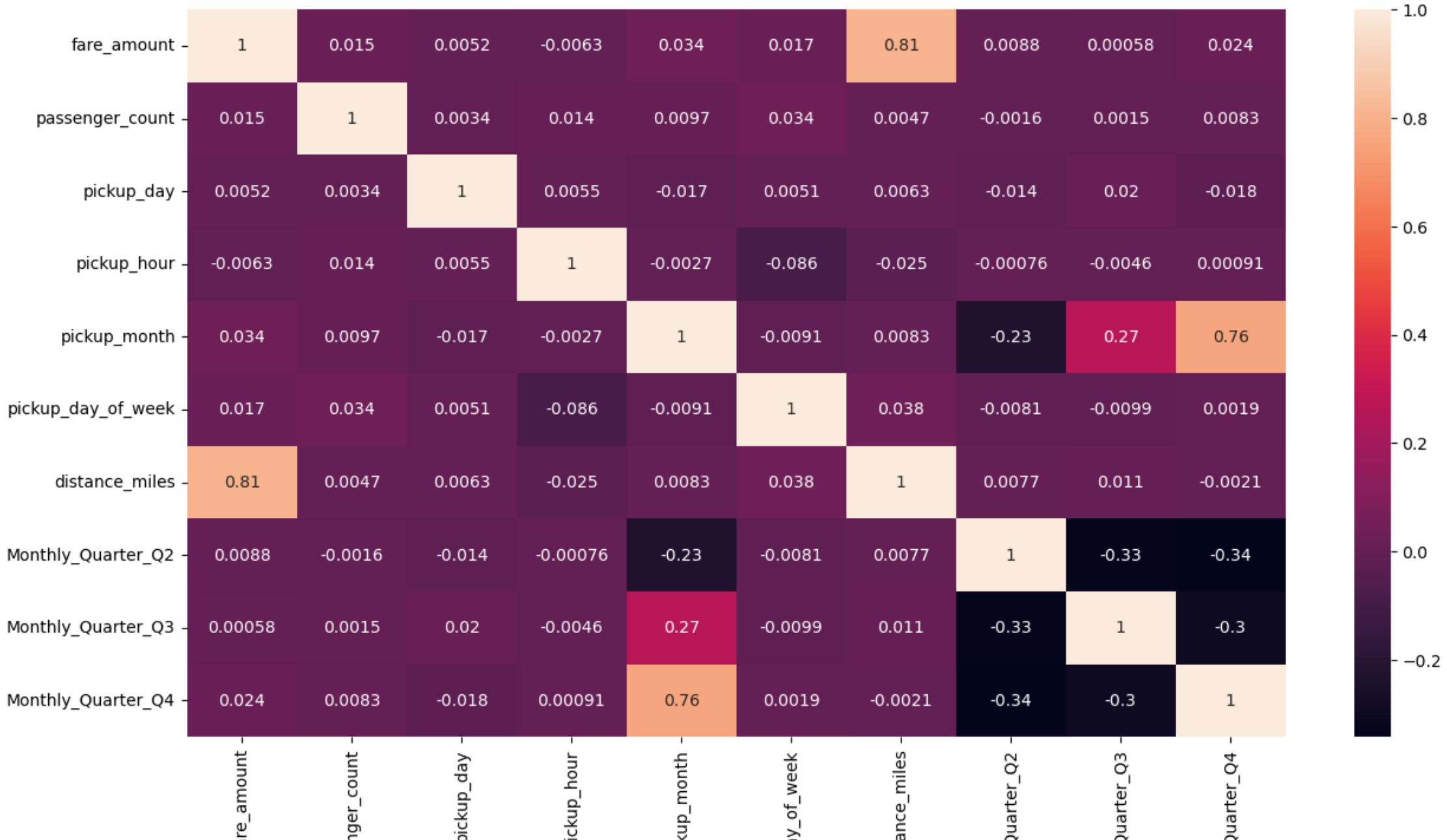


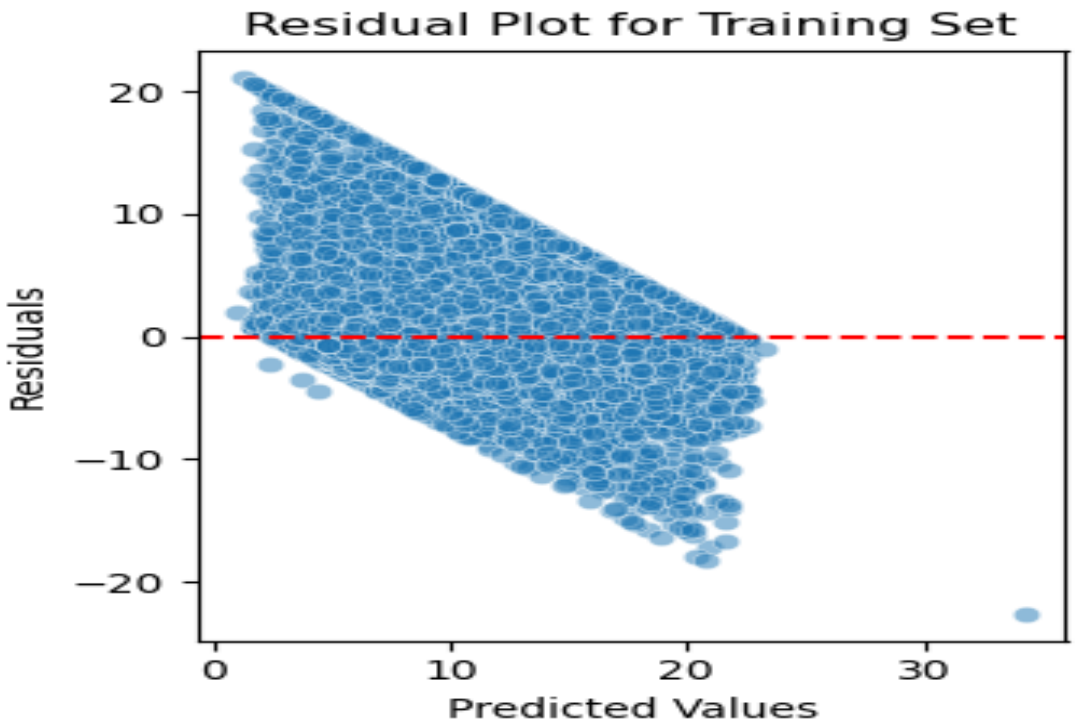


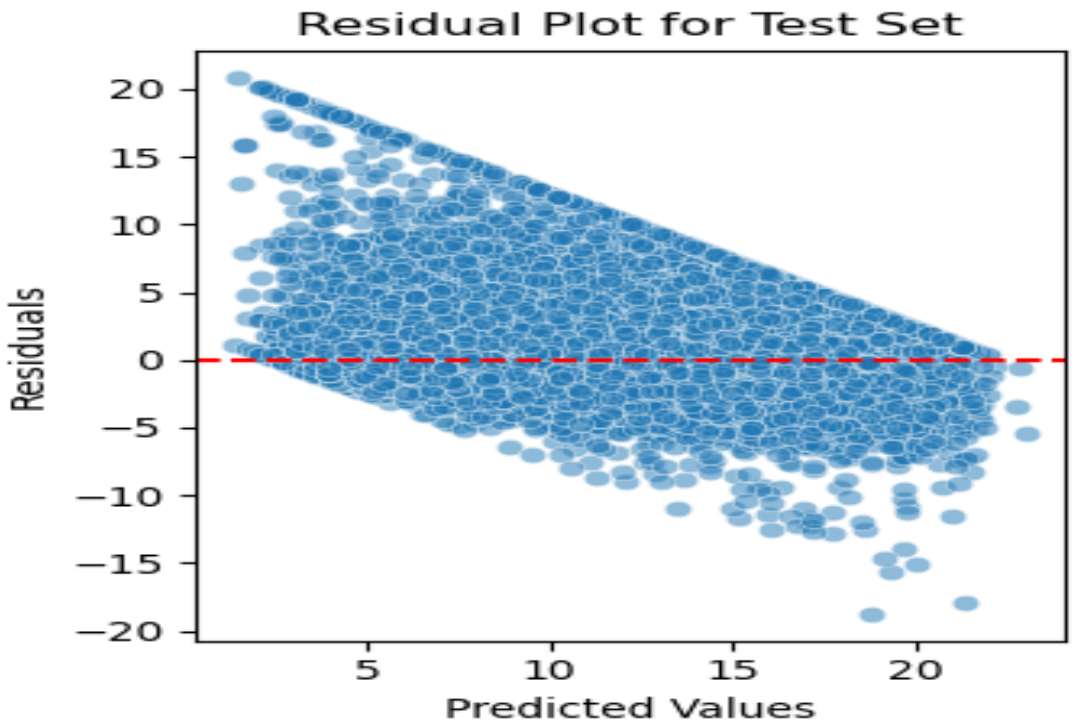
- In weekend the number of customers are compratively low
- We should put on some sttractive offers for consumers to plan their weekends.

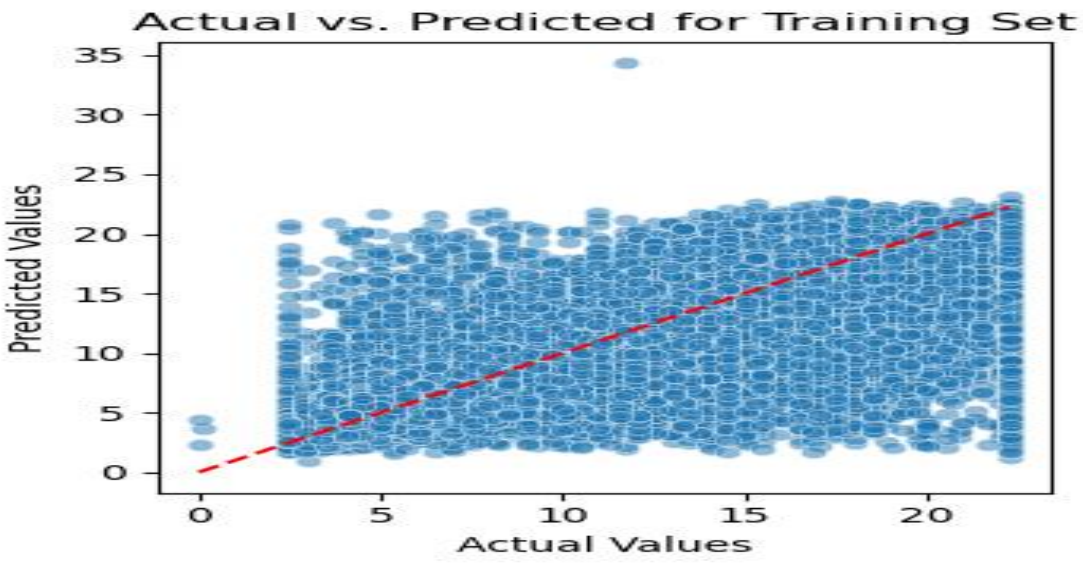

Distribution of Monthly_Quarter
Count
50000
40000
30000
20000
10000
0
Q2
Q3
Q4
Q1
Monthly_Quarter

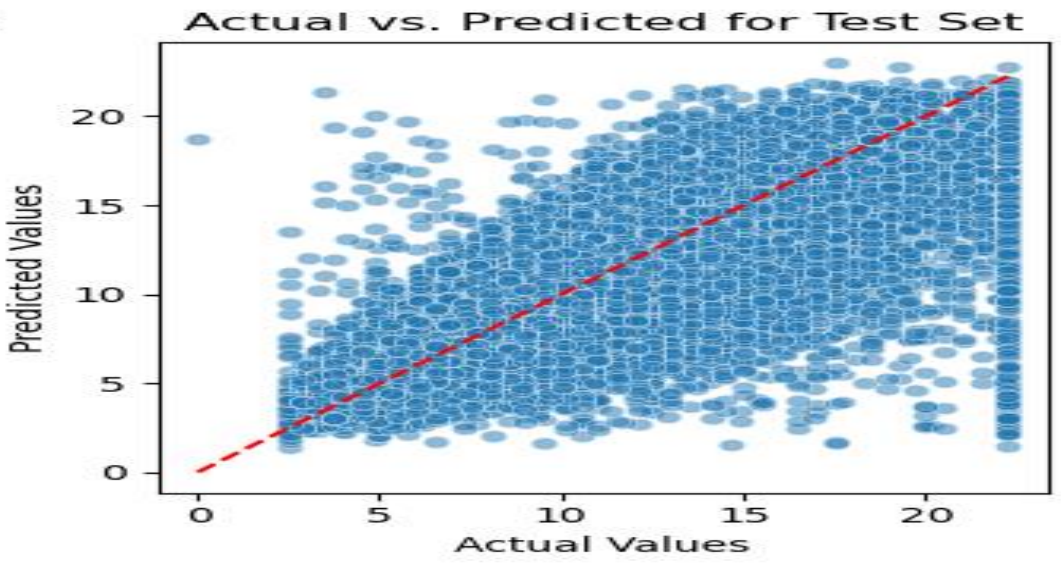
Heatmap

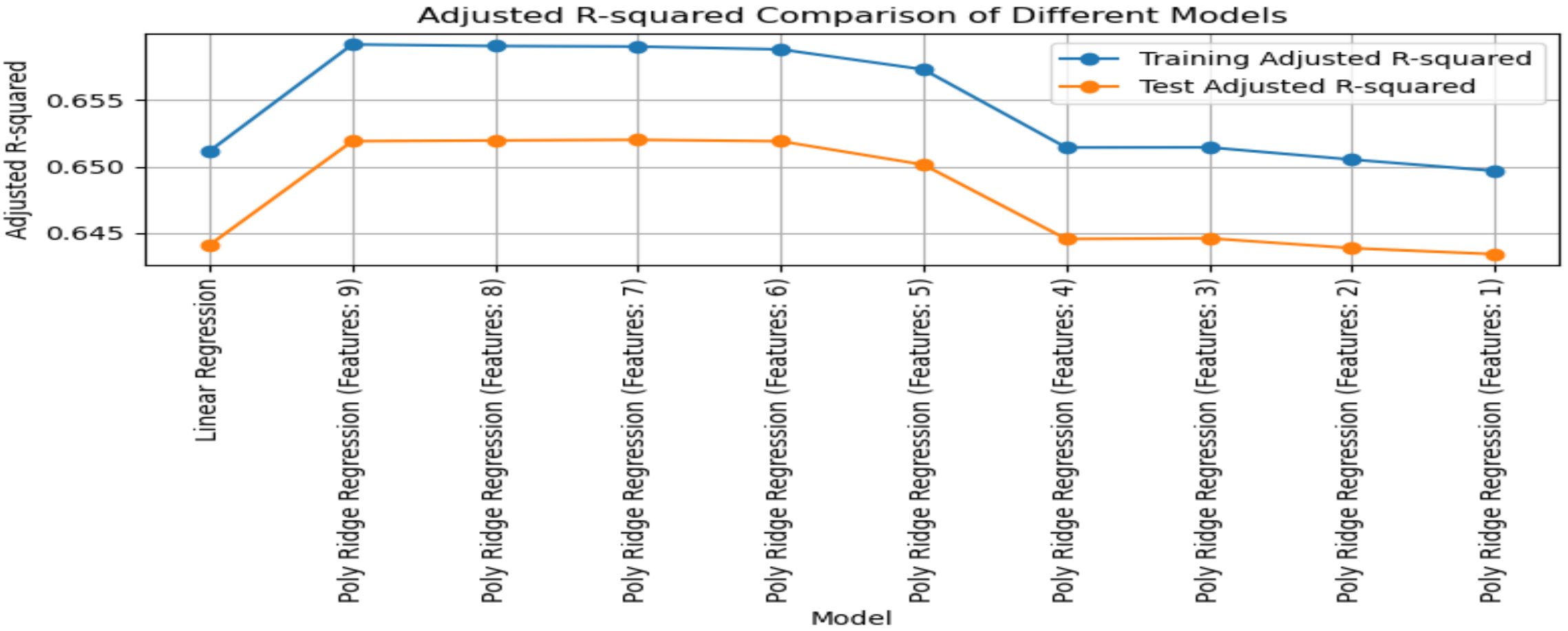

fare_amount 1 0.015 0.0052 -0.0063 0.034 0.017 0.81 0.0088 0.00058 0.024
passenger_count 0.015 1 0.0034 0.014 0.0097 0.034 0.0047 -0.0016 0.0015 0.0083
pickup_day 0.0052 0.0034 1 0.0055 -0.017 0.0051 0.0063 -0.014 0.02 -0.018
pickup_hour -0.0063 0.014 0.0055 1 -0.0027 -0.086 -0.025 -0.00076 -0.0046 0.00091
pickup_month 0.034 0.0097 -0.017 -0.0027 1 -0.0091 0.0083 -0.23 0.27 0.76
pickup_day_of_week 0.017 0.034 0.0051 -0.086 -0.0091 1 0.038 -0.0081 -0.0099 0.0019
distance_miles 0.81 0.0047 0.0063 -0.025 0.0083 0.038 1 0.0077 0.011 -0.0021
Monthly_Quarter_Q2 0.0088 -0.0016 -0.014 -0.00076 -0.23 -0.0081 0.0077 1 -0.33 -0.34
Monthly_Quarter_Q3 0.00058 0.0015 0.02 -0.0046 0.27 -0.0099 0.011 -0.33 1 -0.3
Monthly_Quarter_Q4 0.024 0.0083 -0.018 0.00091 0.76 0.0019 -0.0021 -0.34 -0.3 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Residual Plot for Training Set
Residuals
Predicted Values
20
10
0
−10
−20
0
10
20
30


Residual Plot for Test Set
Residuals
Predicted Values
20
15
10
5
0
−5
−10
−15
−20
5
10
15
20


Actual vs. Predicted for Training Set
Predicted Values
Actual Values
35
30
25
20
15
10
5
0
0
5
10
15
20


Actual vs. Predicted for Test Set
Predicted Values
Actual Values
20
15
10
5
0
0
5
10
15
20



Adjusted R-squared Comparison:

Training Adjusted R-squared:

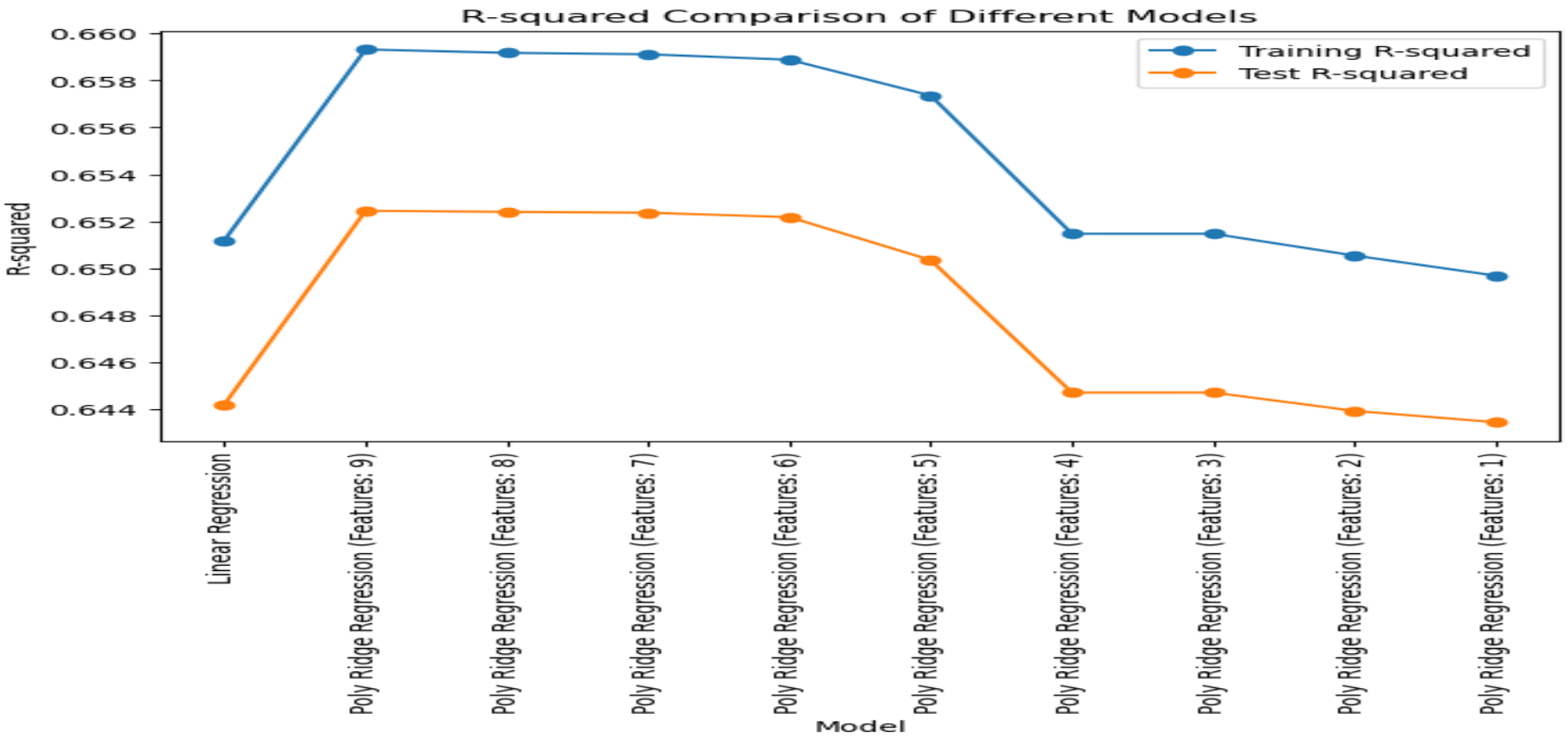
Linear Regression and Polynomial Ridge Regression with 9, 8, and 7 features exhibit the highest training adjusted R-squared values.

As the number of features decreases below 6, the training adjusted R-squared values also drop, indicating less fit to the training data.

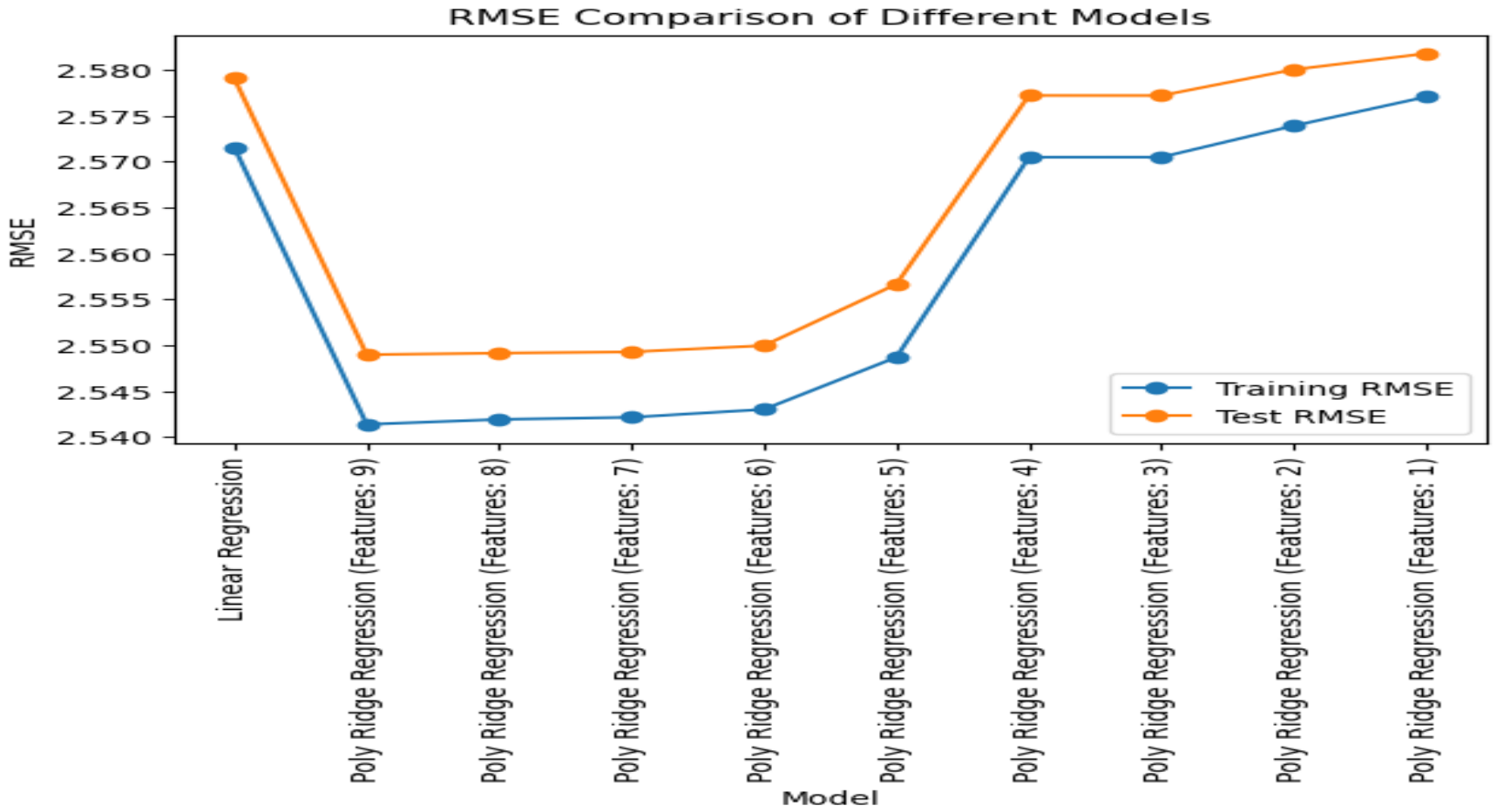
Test Adjusted R-squared:

Polynomial Ridge Regression with 8 and 7 features shows the highest test adjusted R-squared values.

The test adjusted R-squared drops significantly with fewer features, highlighting the models' reduced predictive power on unseen data with fewer features.



R-squared Comparison:
Training R-squared:
Similar to the adjusted R-squared, the training R-squared is highest for models with more features, peaking at 9, 8, and 7 features.
Test R-squared:
Polynomial Ridge Regression models with 8 and 7 features have the highest test R-squared values.
Test R-squared values decrease notably for models with fewer than 6 features.



RMSE Comparison:

Training RMSE:

The training RMSE is lowest for Polynomial Ridge Regression with 9, 8, and 7 features, indicating better fit on the training data.

RMSE values increase as the number of features decreases, indicating poorer model performance.

Test RMSE:

The lowest test RMSE values are observed for Polynomial Ridge Regression with 8 and 7 features.

Test RMSE increases with fewer features, showing the models' reduced accuracy on the test set with fewer features.

Conclusion

- Peak Hour at 19:00: This is our busiest time.
- Low Activity at Midnight (00:00): Consider using coupons to attract more customers.
- Low Bookings from 1:00-7:00: Late-night timing likely affects demand; incentives could help increase usage.
- Low Bookings from 8:00-17:00: High potential for growth by targeting commuters with coupons and promoting carpooling.
- Decline from 20:00-23:00: Focus on reducing waiting times to improve customer retention during these hours.

Recommendations:

- Targeted Promotions: Introduce coupons and discounts during low-demand hours (00:00 and 1:00-7:00) to attract
- more customers.
- Commuter Incentives: Launch special offers for office and school commuters during 8:00-17:00, and promote carpooling services to boost bookings.
- Operational Efficiency: Analyze and address factors contributing to increased waiting times between 20:00-23:00 to maintain customer satisfaction and prevent declines in bookings.