TLC Trip Record Data Prediction

Presented by:

Rehab AL Zaidi & Amal Al Thaqafi

Contents

Introduction

The purpose of this project is to predict the fare amount of the trip using linear regression algorithm. We worked with data provided by TLC Trip Record Data.

About Data:

The data of Green Taxi trip records contain 2 months which is **January** and **February** at **2021.**

It's contain:

- 20 Features .
- 76487 Observations for **Jan** and 64541 for Feb.

Preprocessing

1

Check null values,
Duplicates,
Outliers

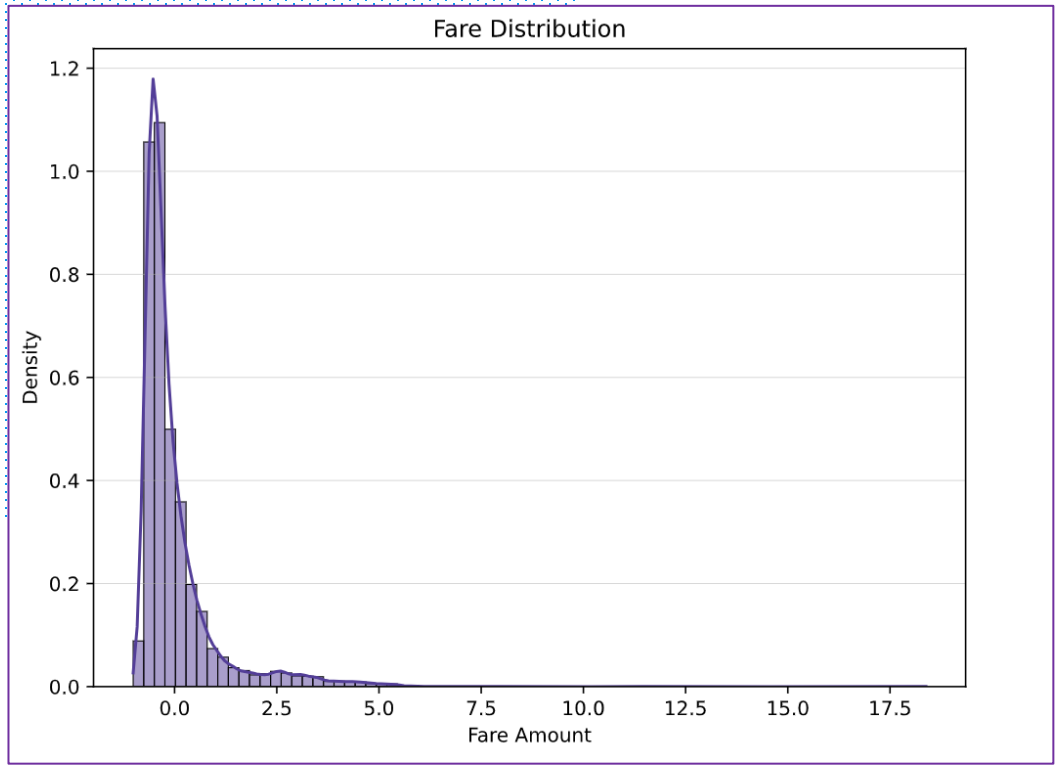
2

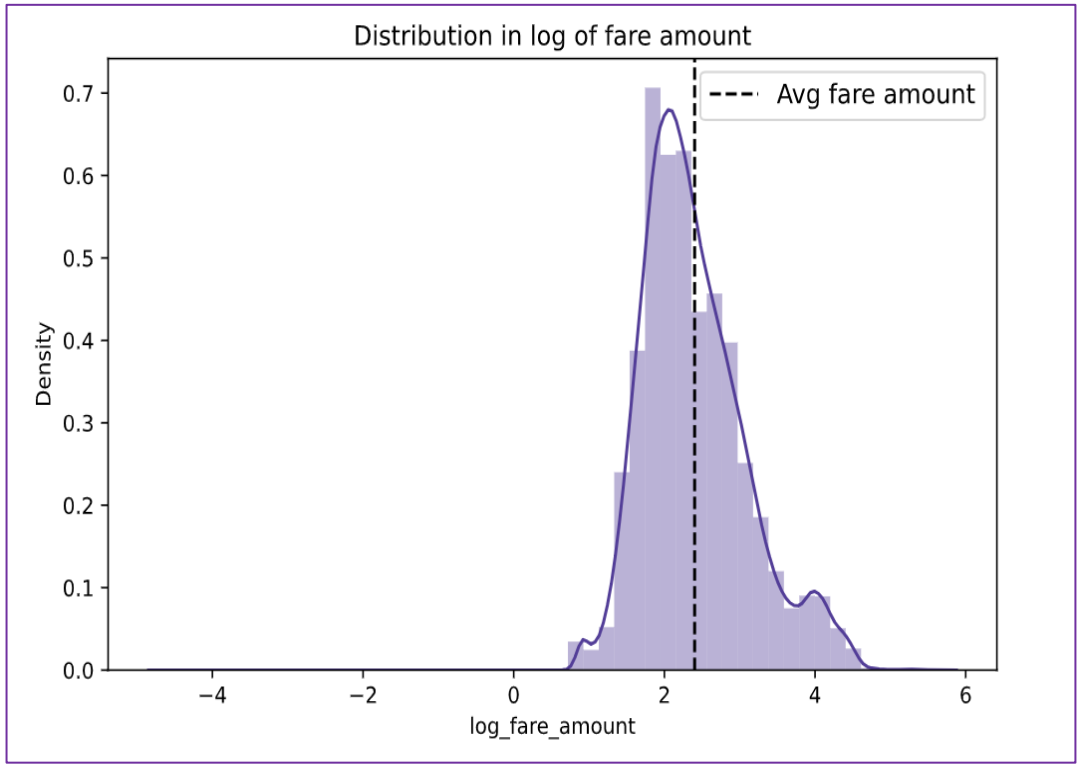
Get
Dummies

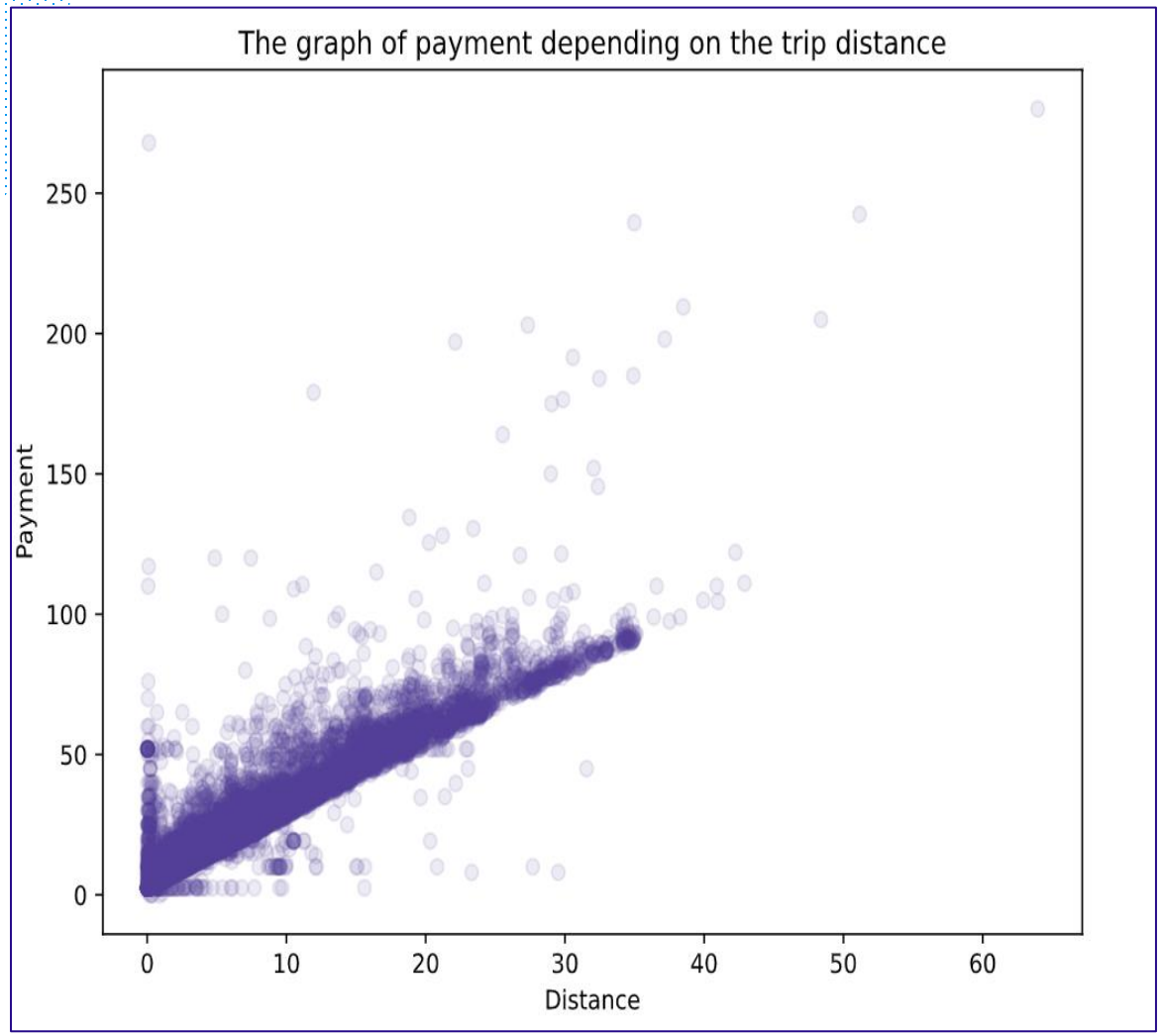
3

visualization

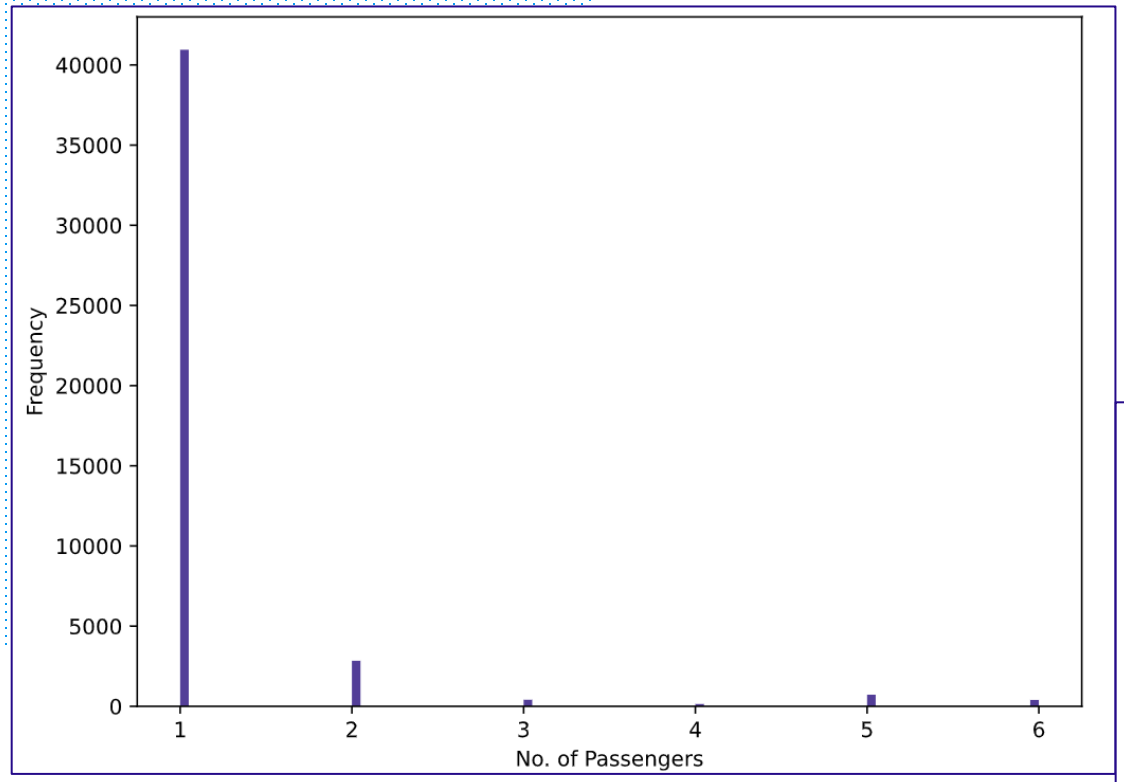
Visualize data

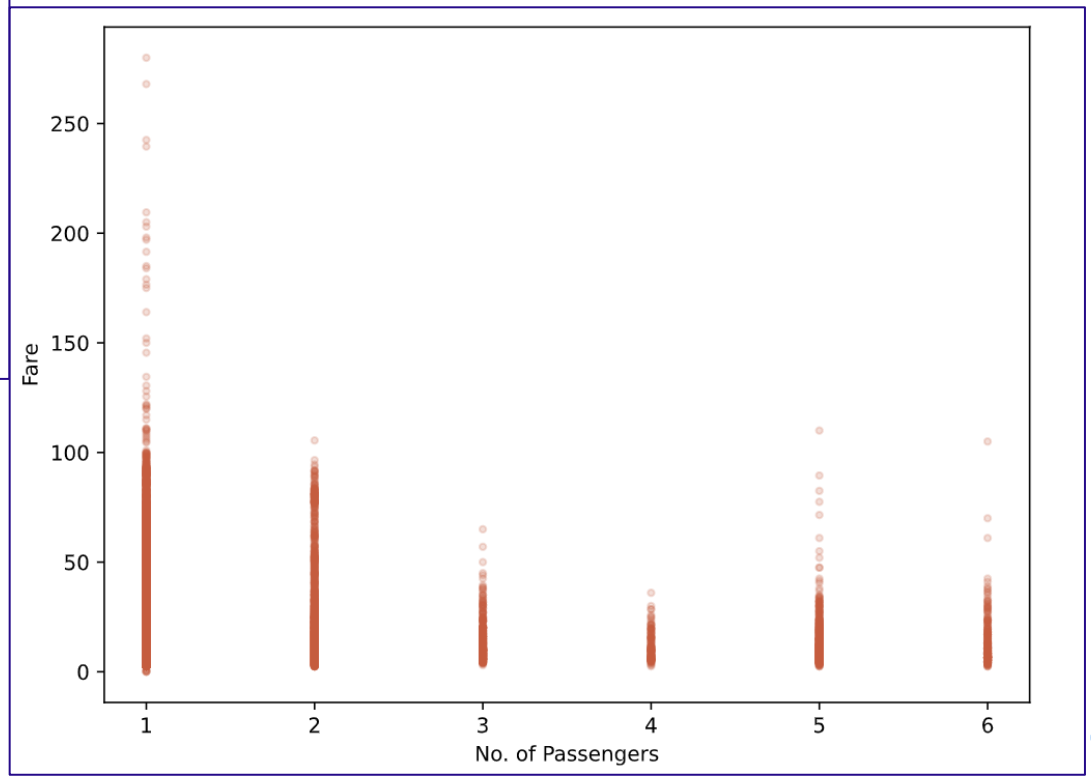




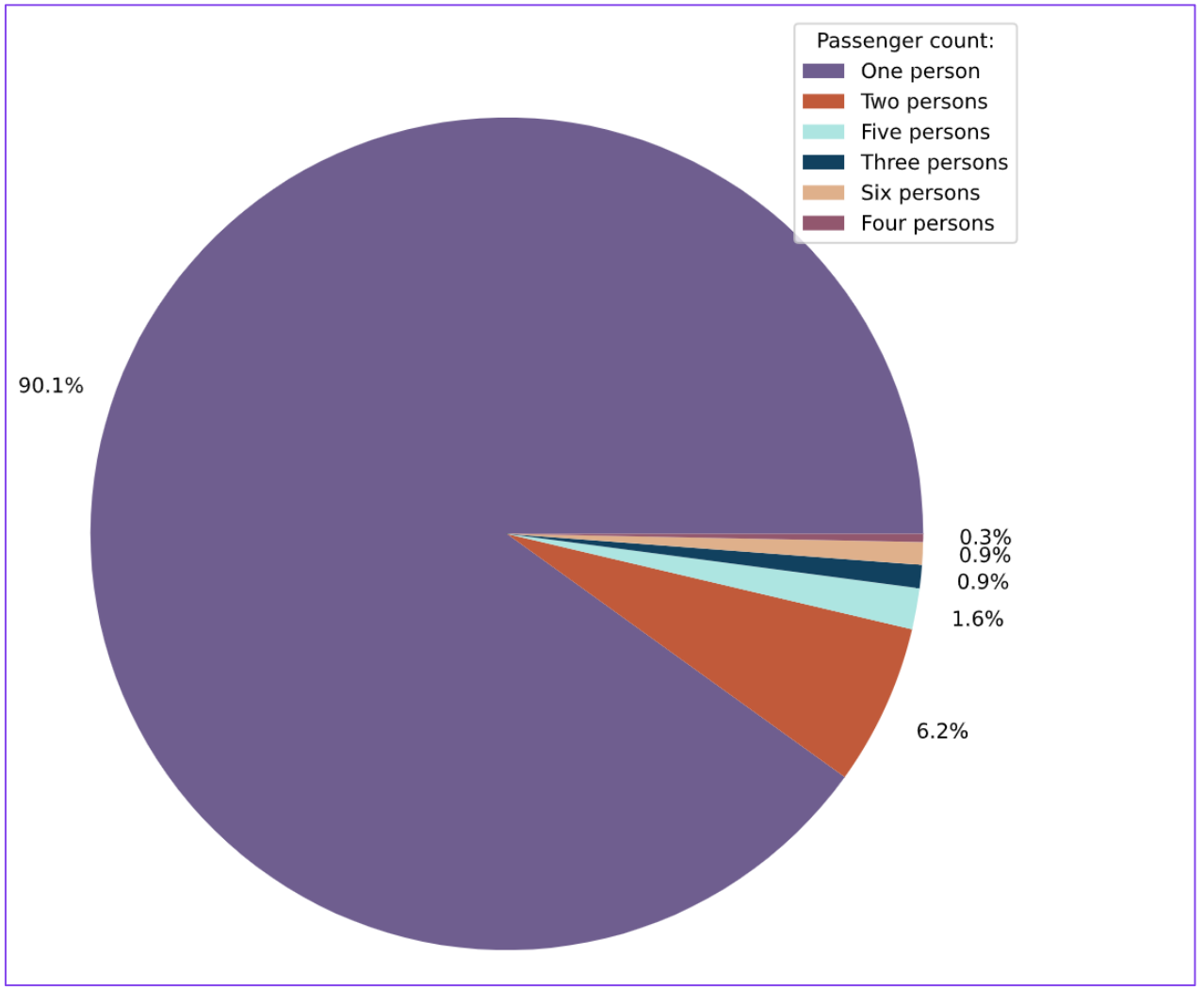

The graph of payment depending on the trip distance
Payment
250
200
150
100
50
0
0
10
20
30
40
50
60
Distance

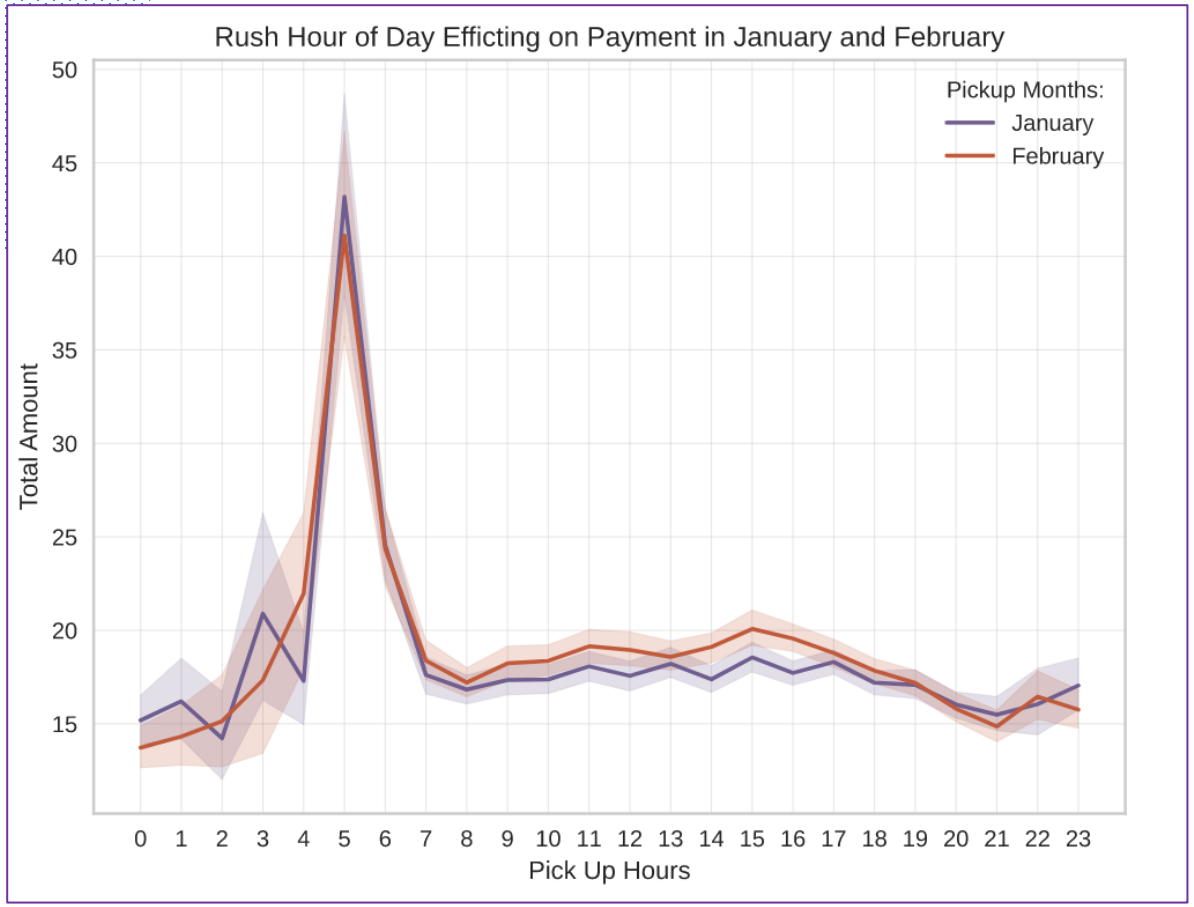
1- Does the number of passengers affect the fare?



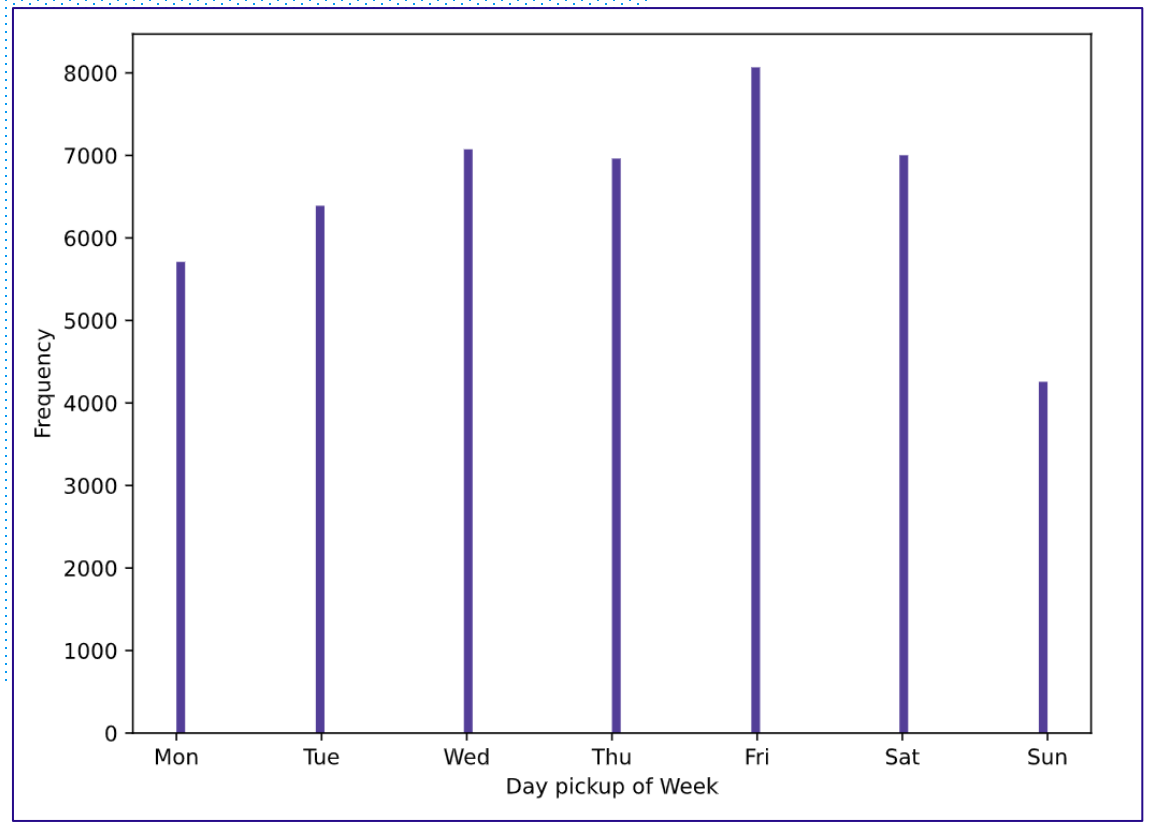


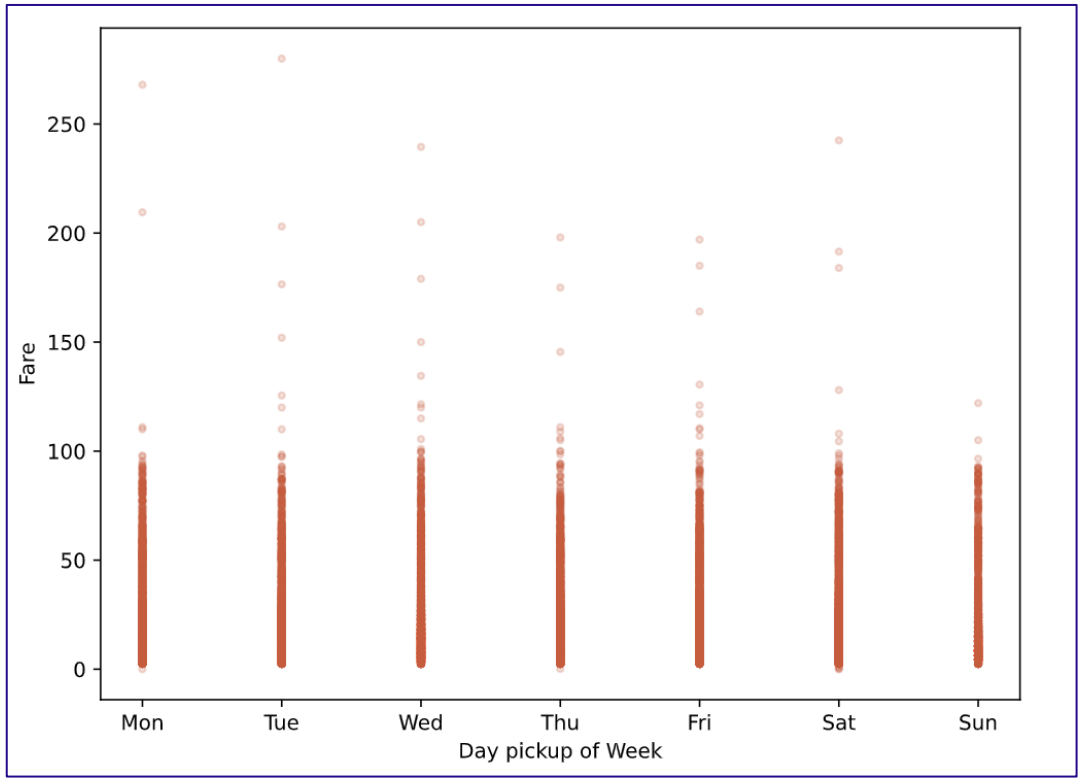
passenger count in trips distribution



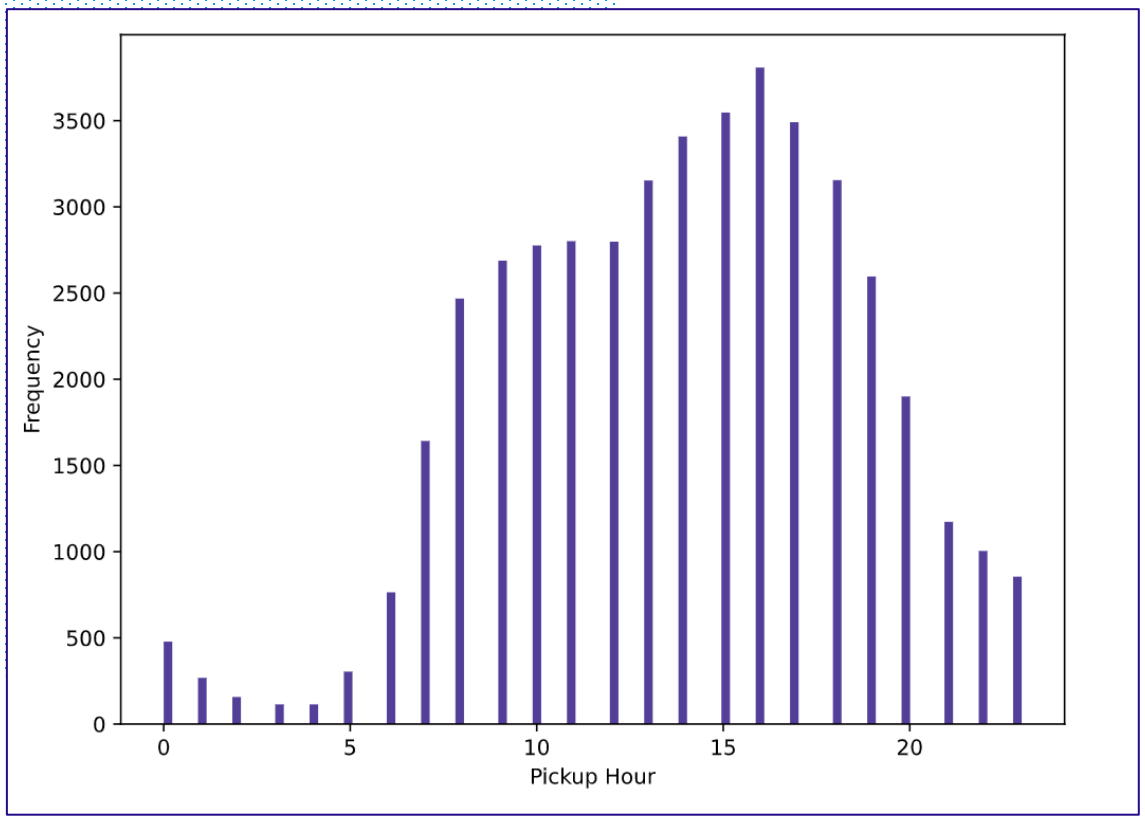

Rush Hour of Day Efficting on Payment in January and February
Pickup Months:
January
February
Total Amount
50
45
40
35
30
25
20
15
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Pick Up Hours

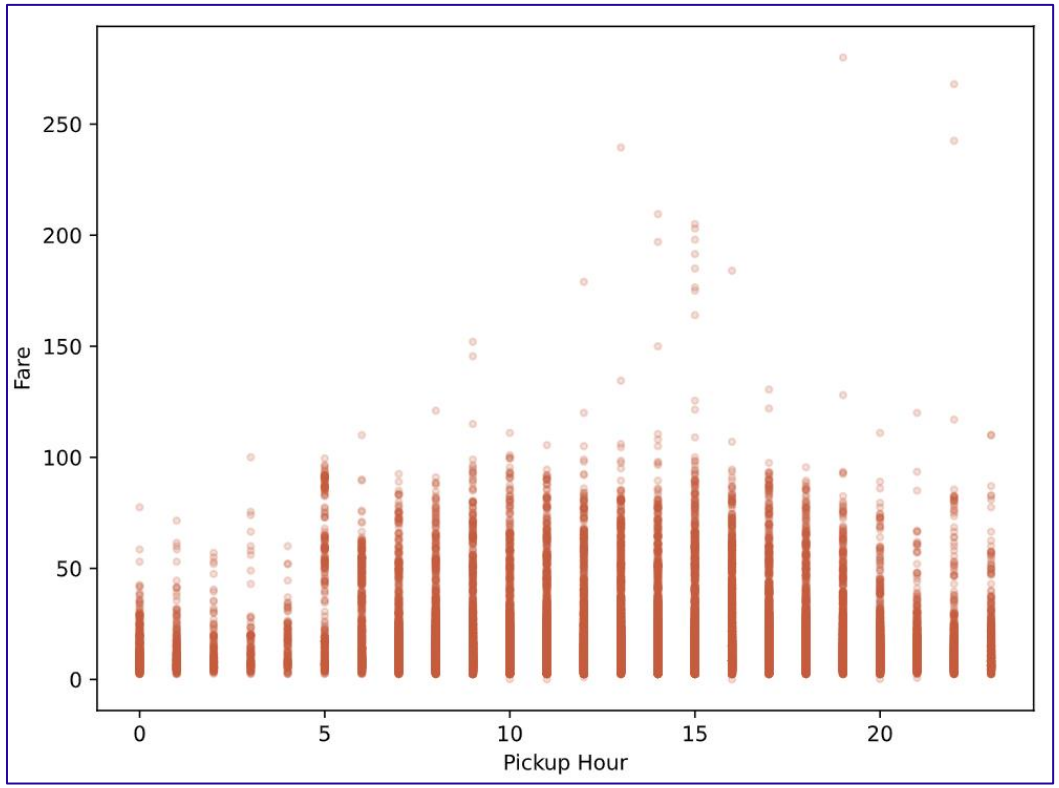
3- Does the day of the week affect the fare?



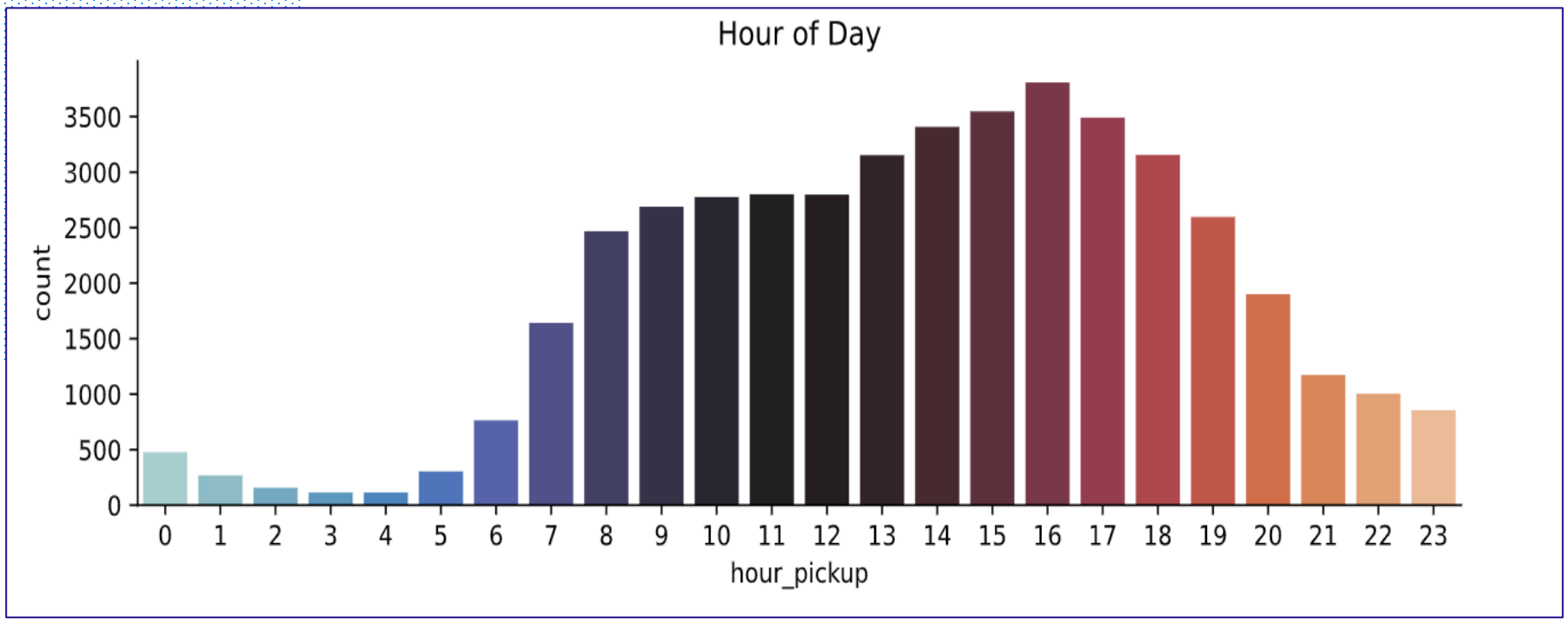


2- Does the time of pickup affect the fare?

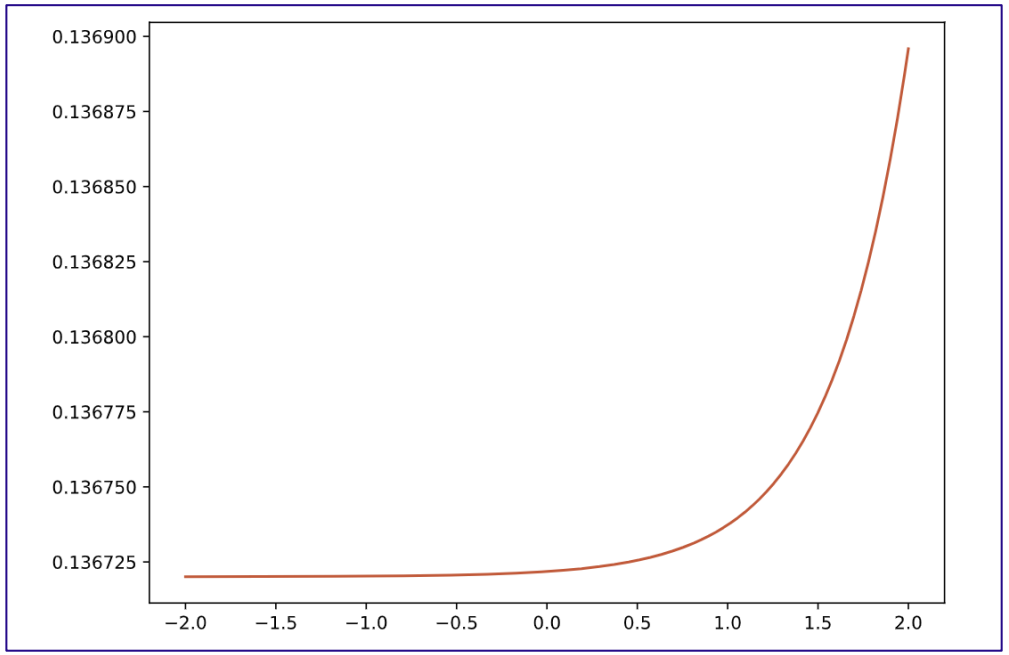




Taxi trip by Rush hour of the day



Ridge Regularization

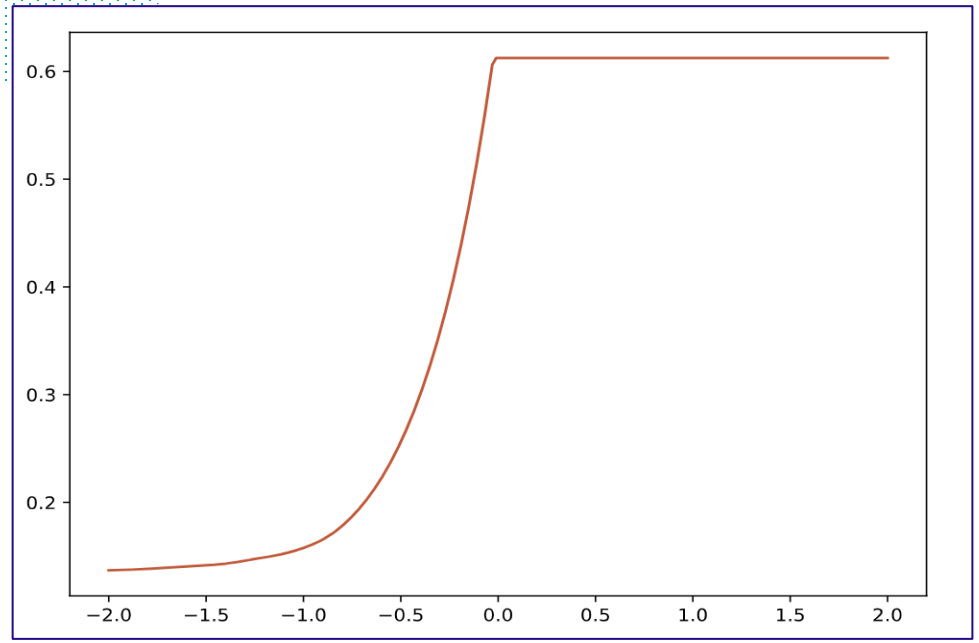


Training =0.8951

Validation=0.8867

Testing=0.8199

Lasso Regularization

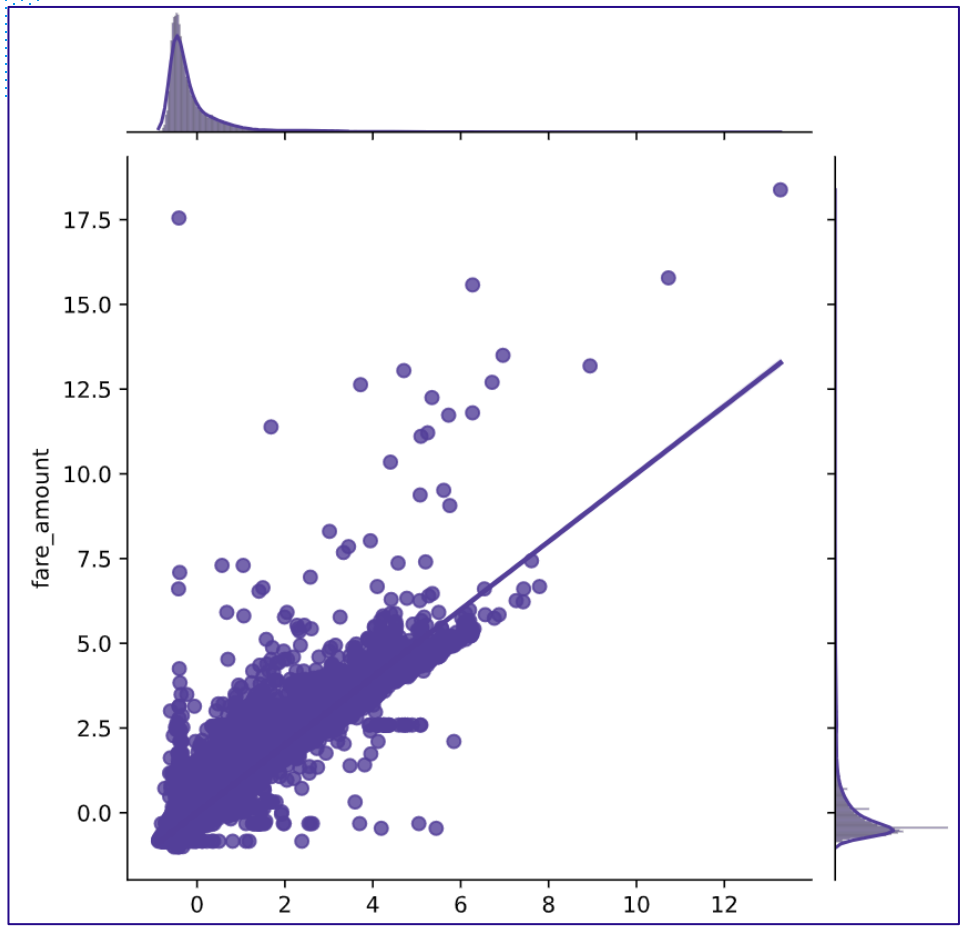


Training =0.8958

Validation=0.8873

Testing=0.81995

Model Building



Conclusion

In the attempts to predict the best model for Fare Amount, we made a number of models such as: The Ridge, Lasso, RMSE, MAE,MSE, Feature Engineering.

The best score is:

R-sq of training set = 0.8958
R-sq of validation set = 0.8873
R-sq of Test set = 0.8199

THANK YOU