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.

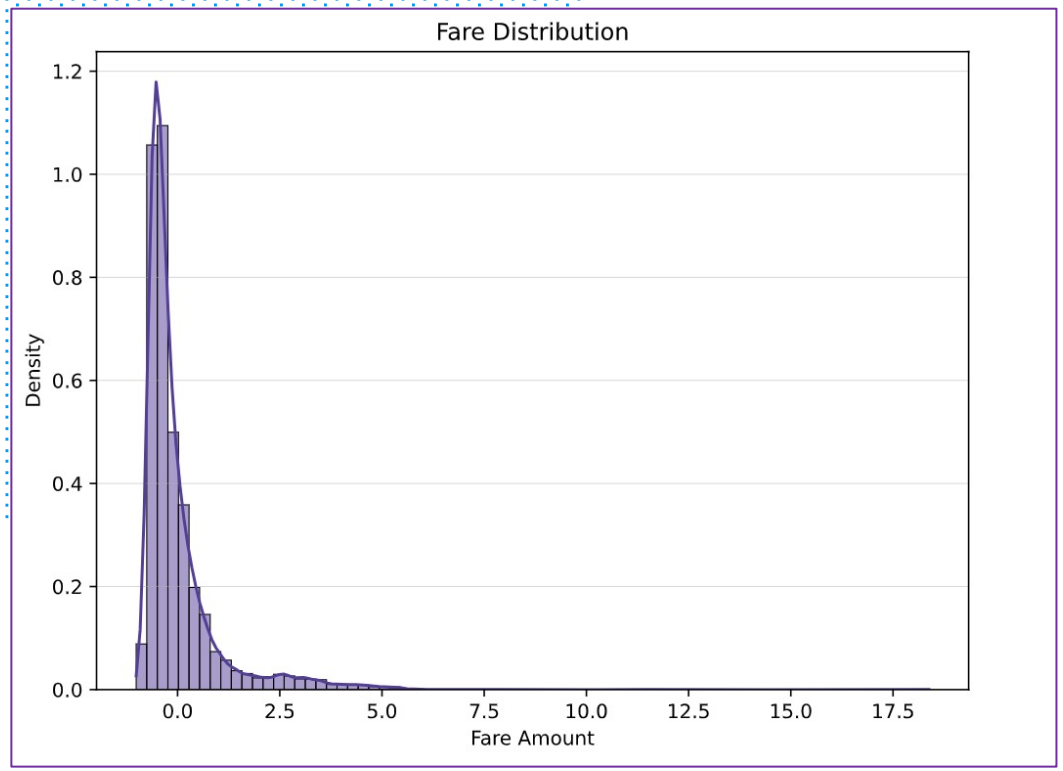
About Data:
The data of Green Taxi trip records contain 2 months which is **January** and **February** in **2021.**

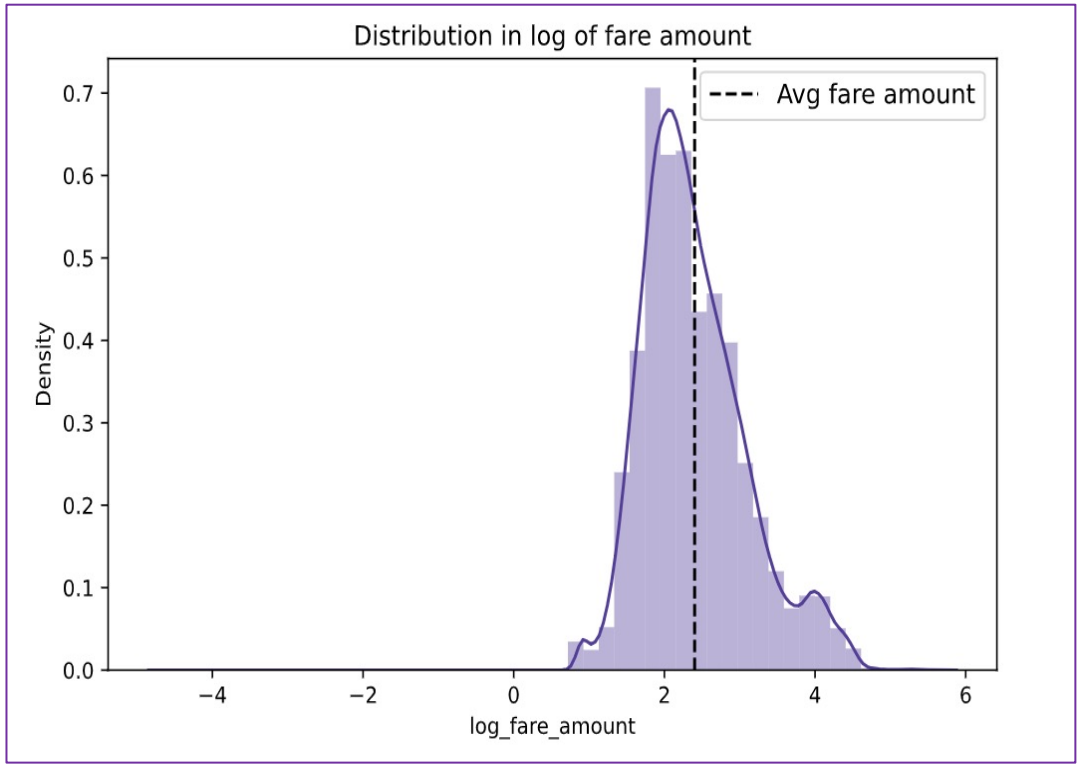
It's contained:
- 20 Features .
- 76487 Observations for **January** and 64541 for **February**.

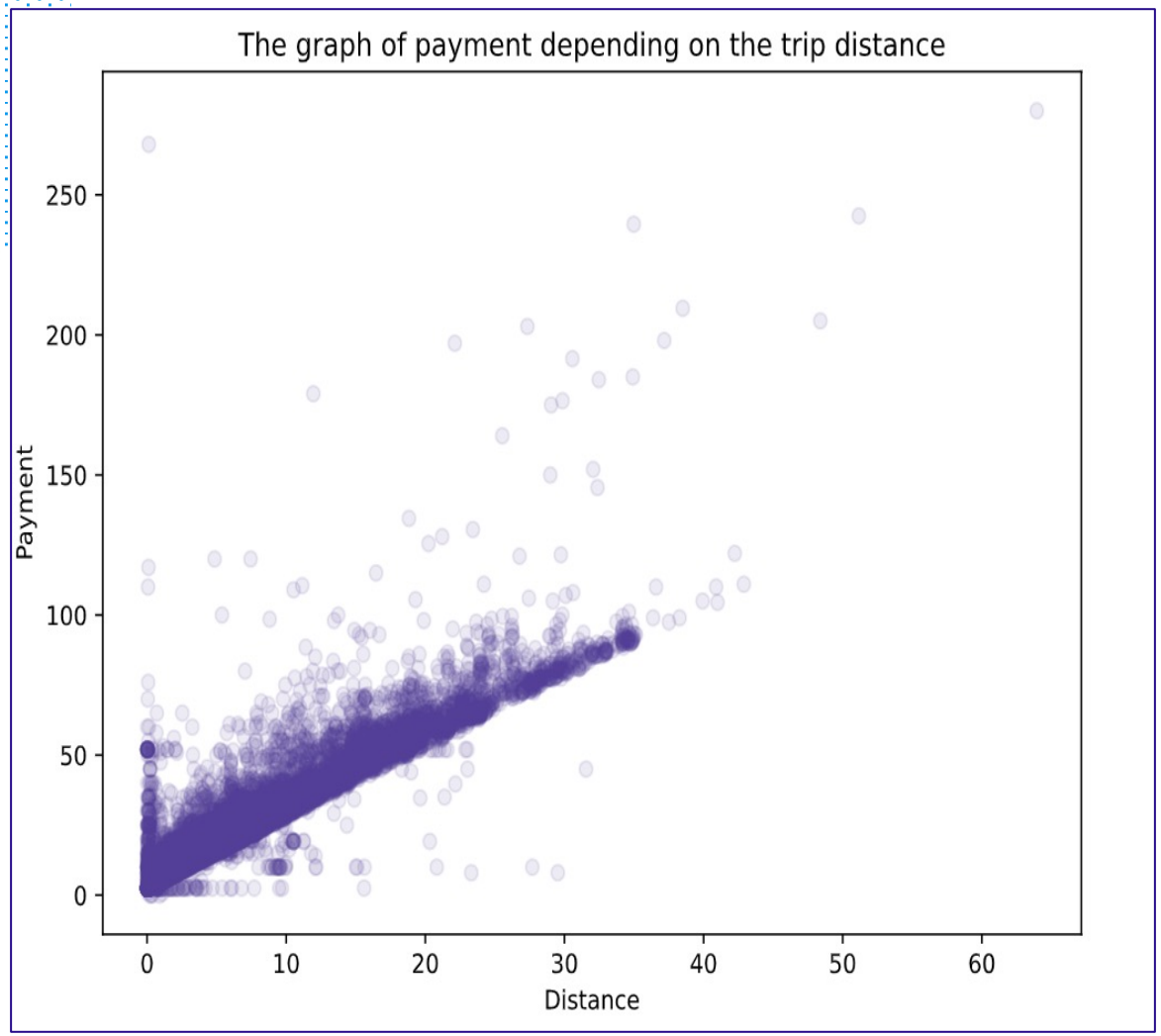
Preprocessing

1. Check null values, Duplicates, Outliers
2. Get Dummies
3. visualization

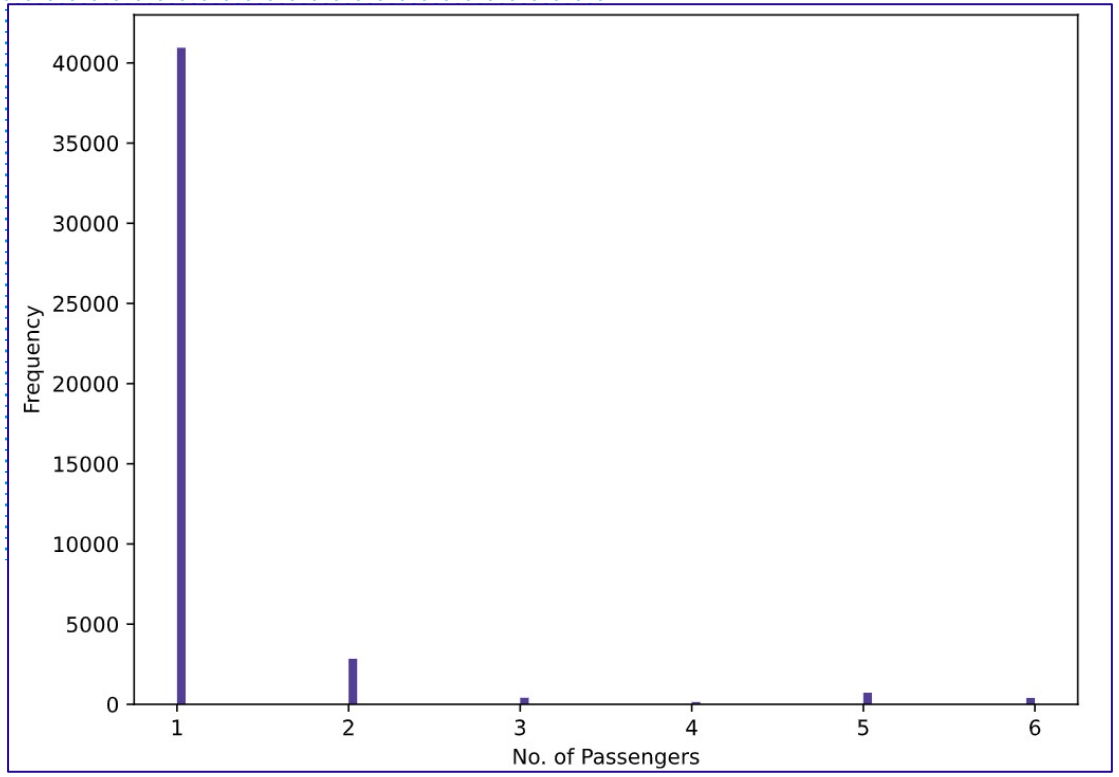
Visualize data

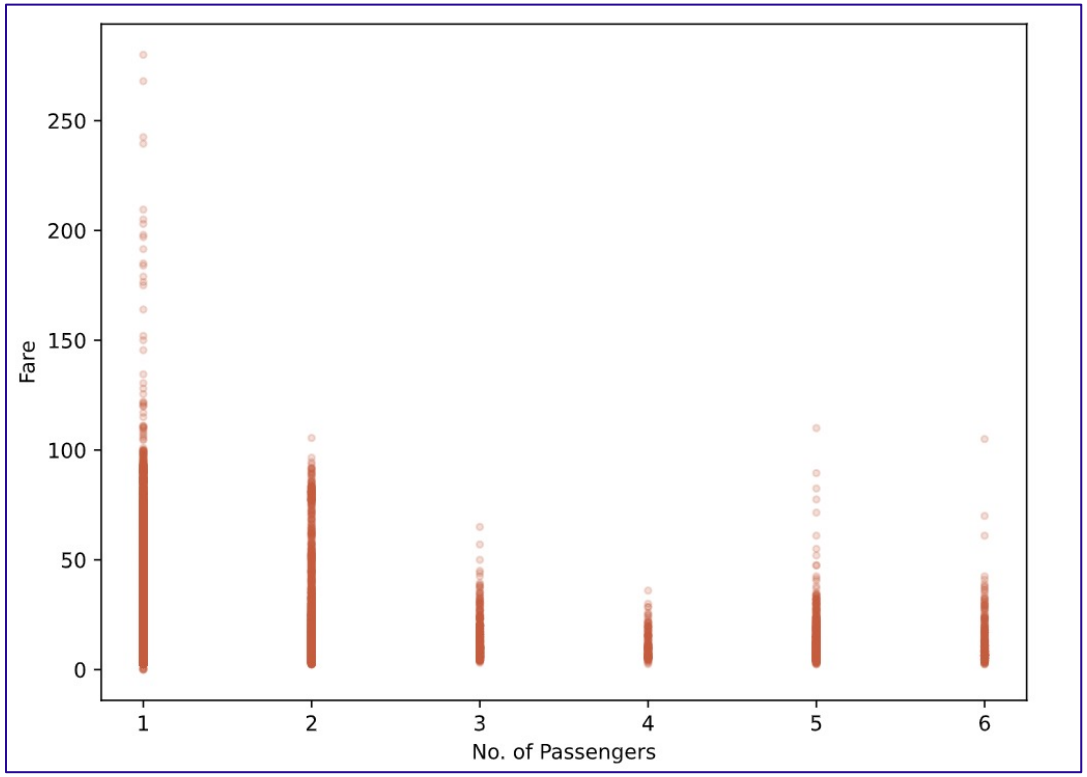




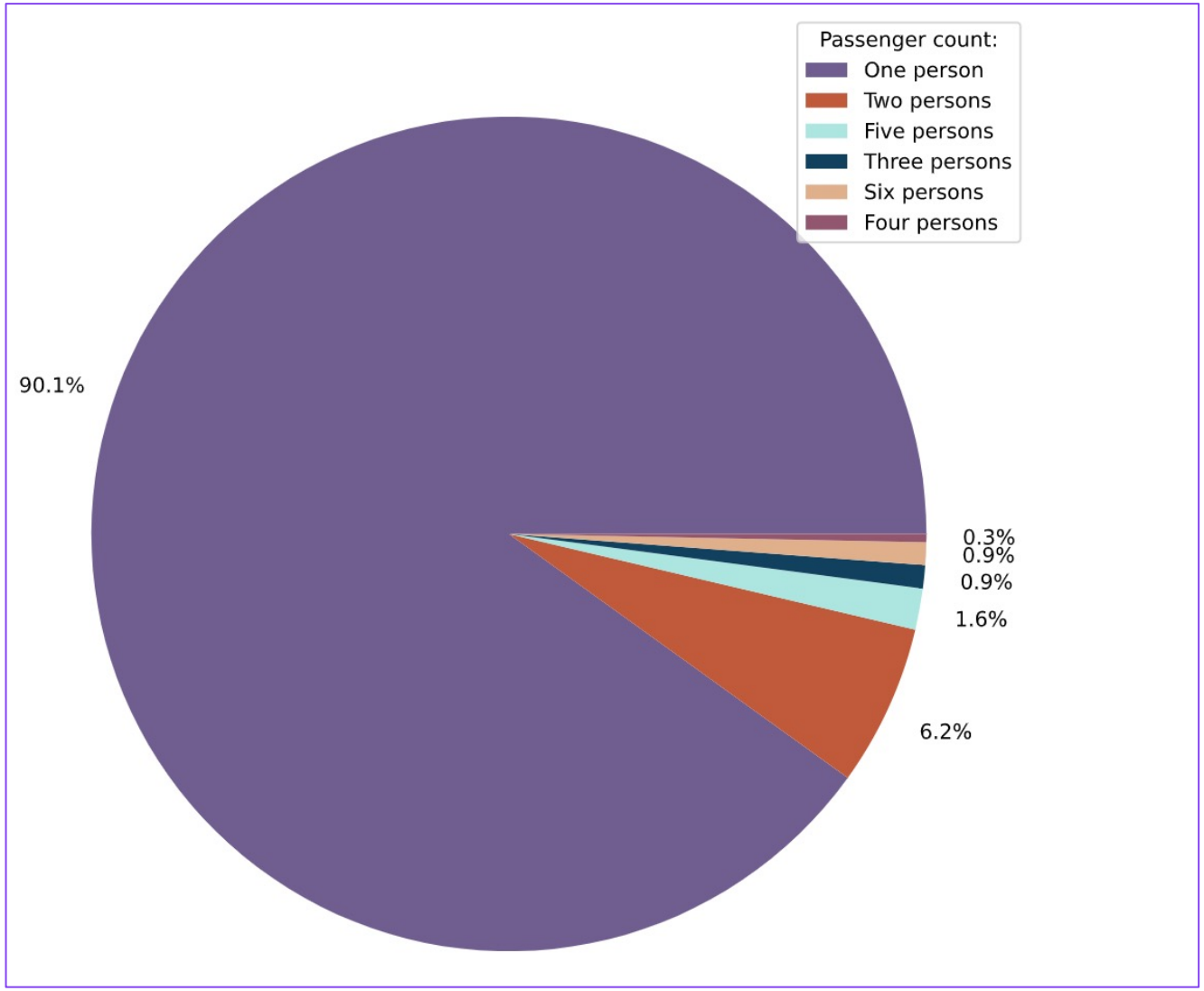

The graph of payment depending on the trip distance
Payment
250
200
150
100
50
0
0
10
20
30
40
50
60
Distance

1- Does the number of passengers affect the fare?

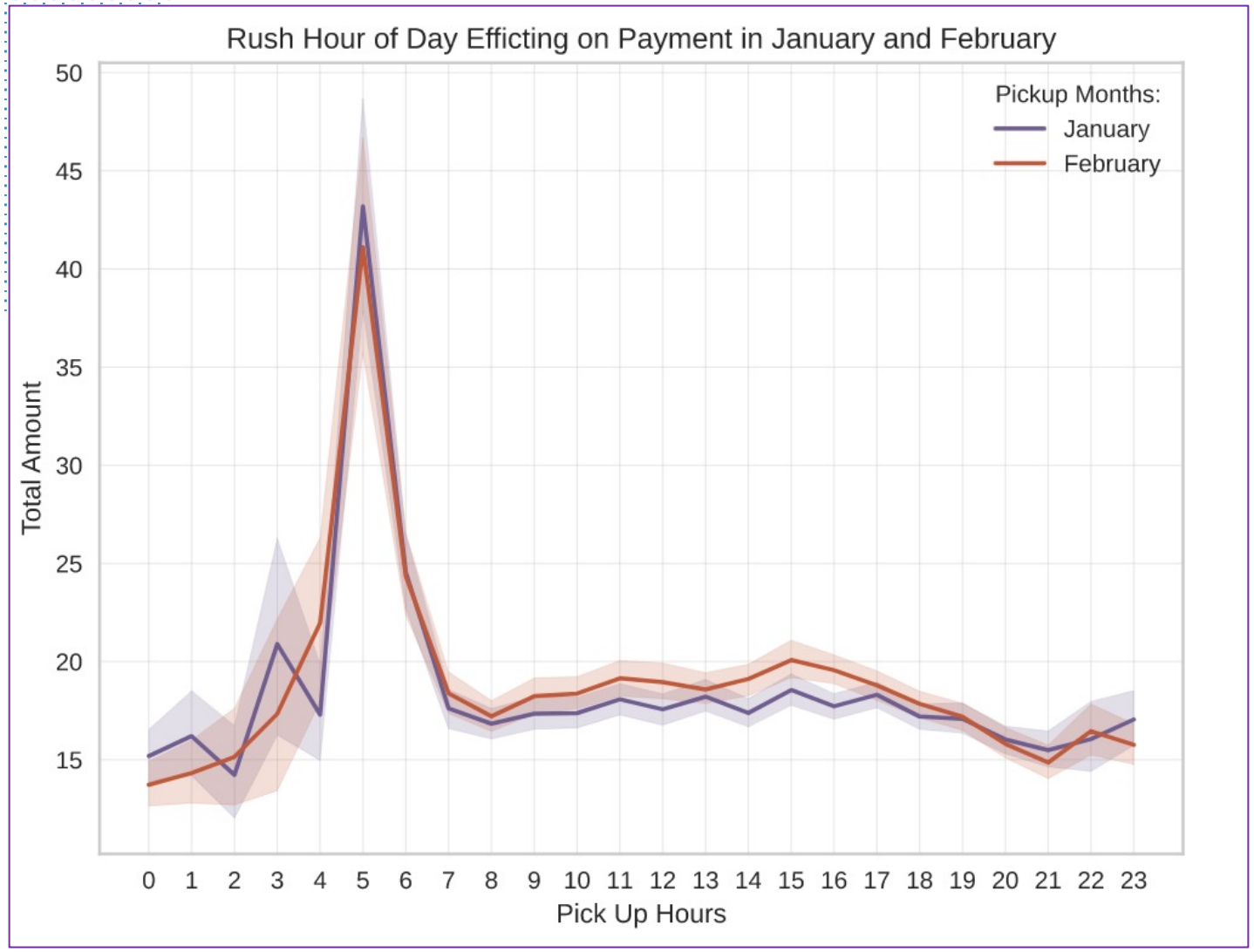




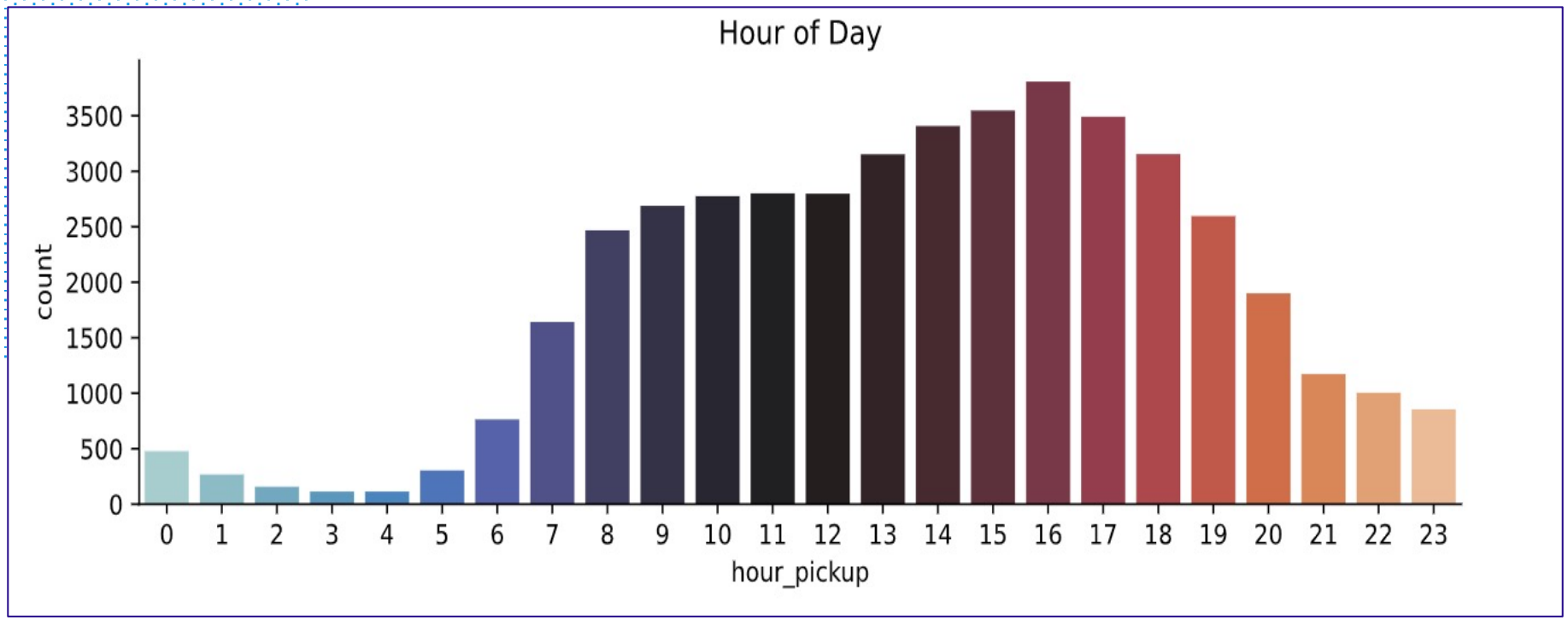
Passenger Count in Trips Distribution

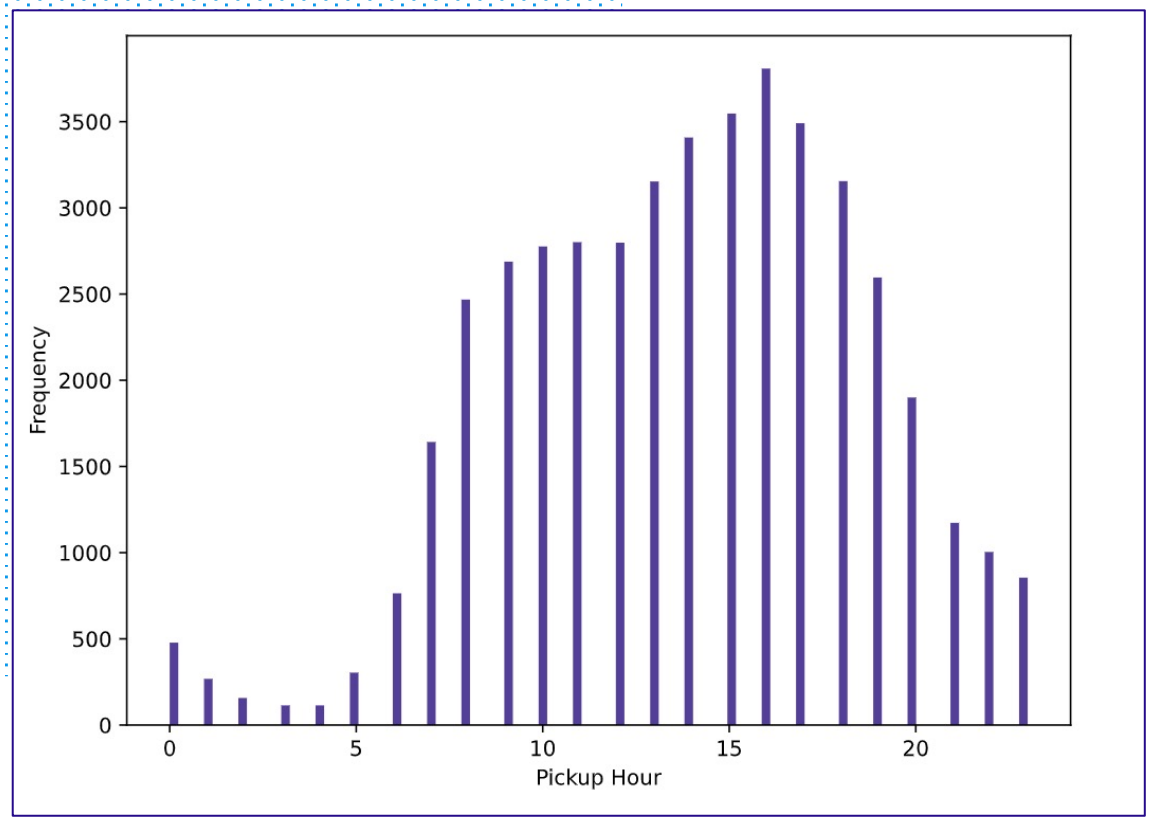


Total Amount depending on hours of the day for taxi trip.

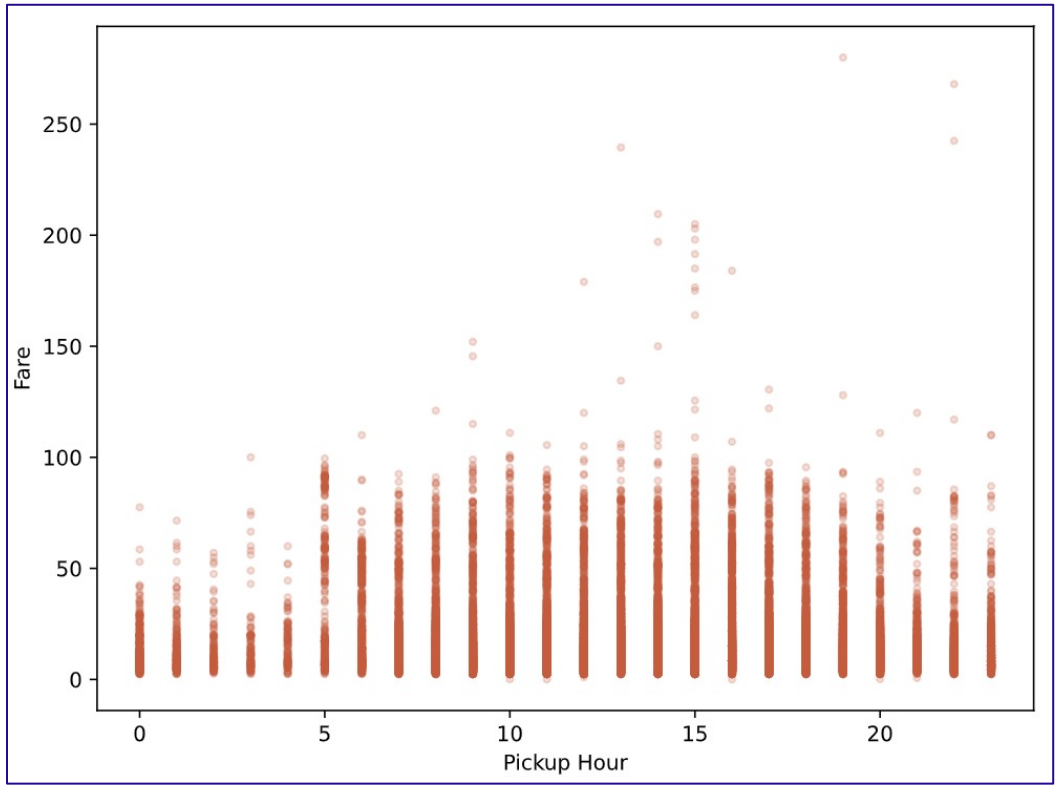


Most requested hours of the day for taxi trip.

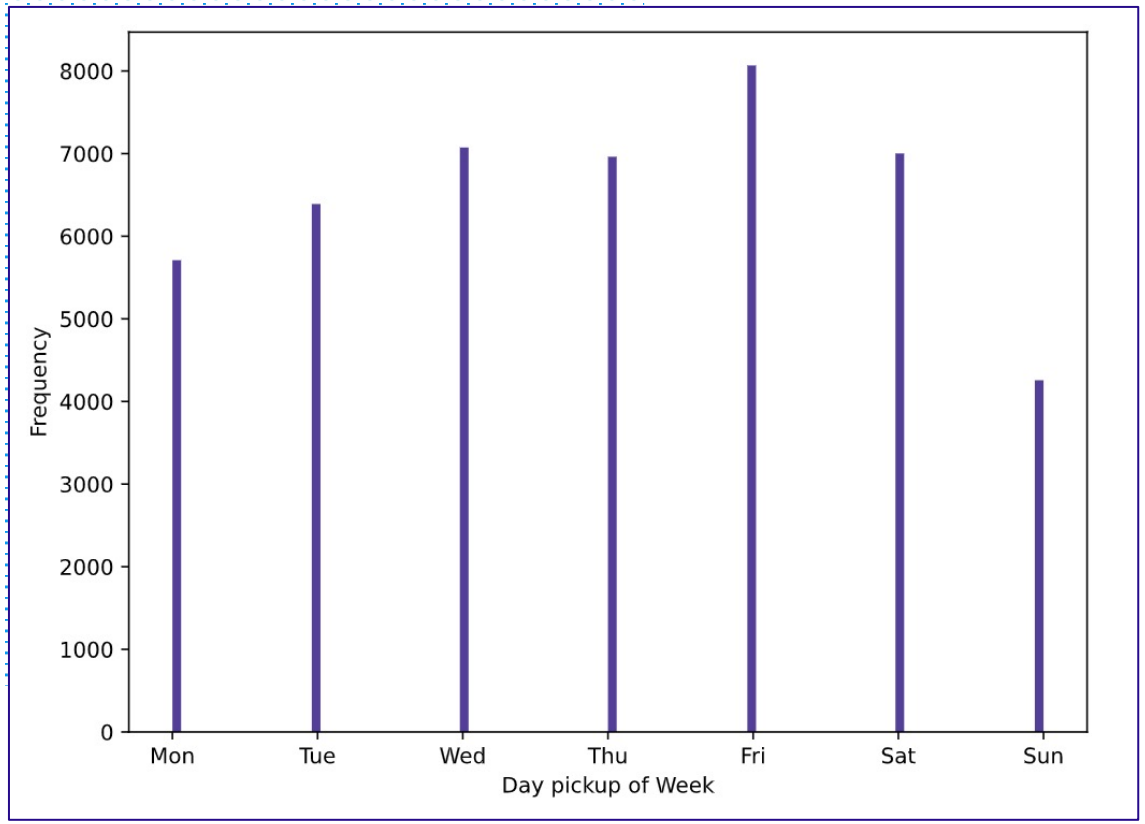


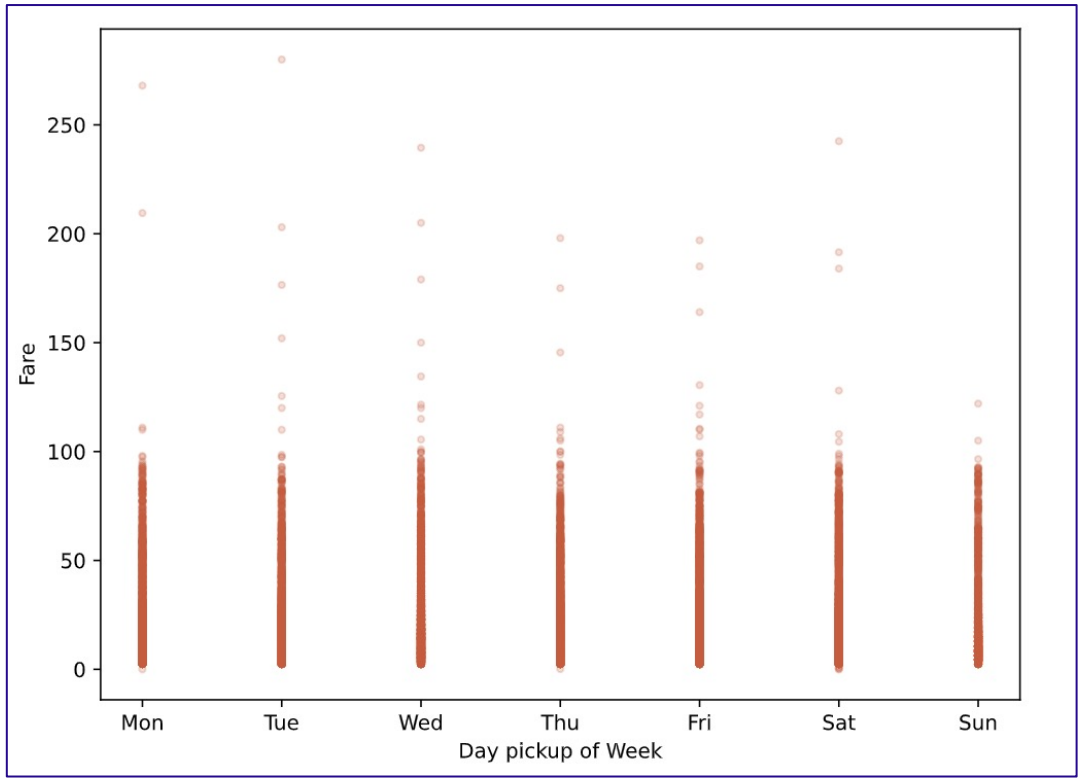


2- Does the time of pickup effect the fare?



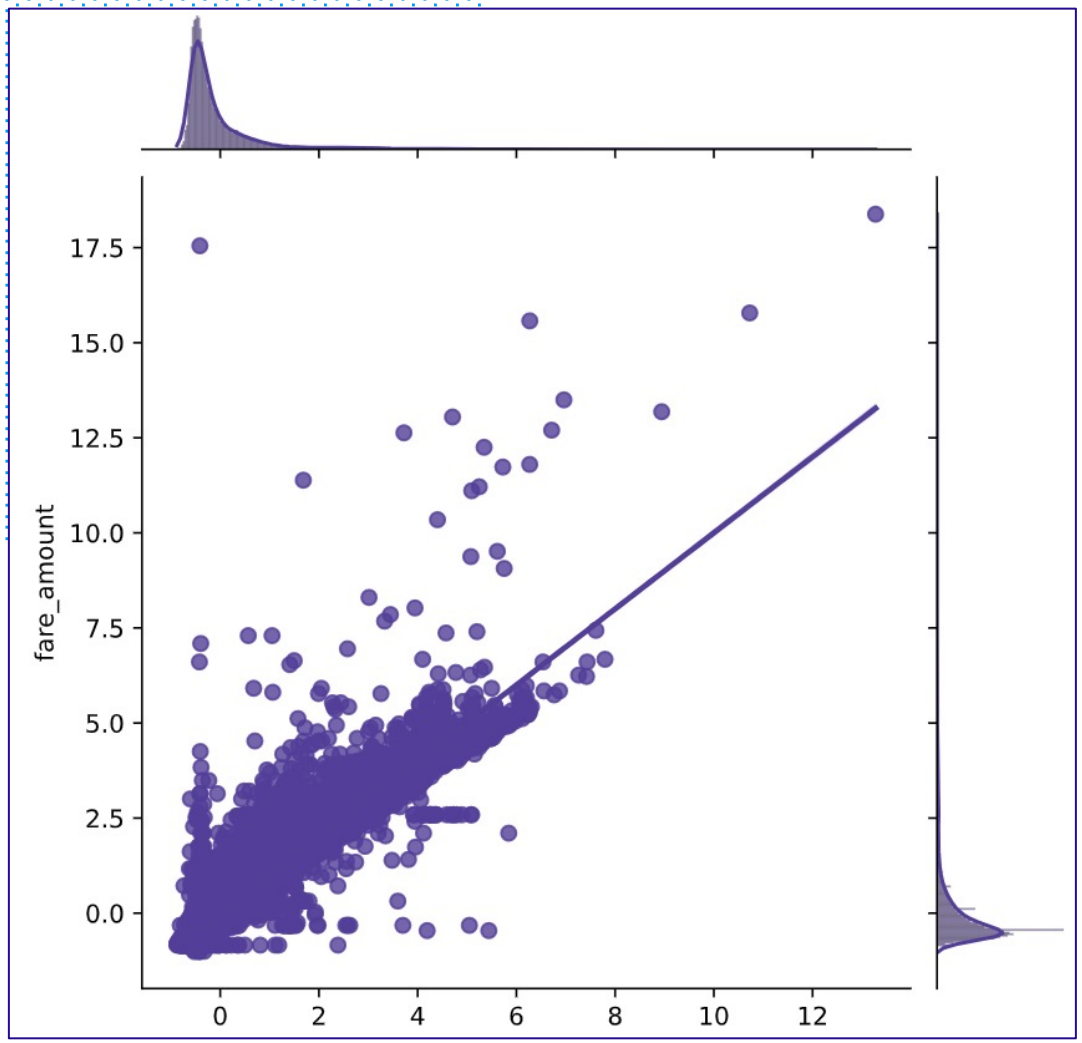
3- Does the day of the week effect the fare?





Model Building

R-sq

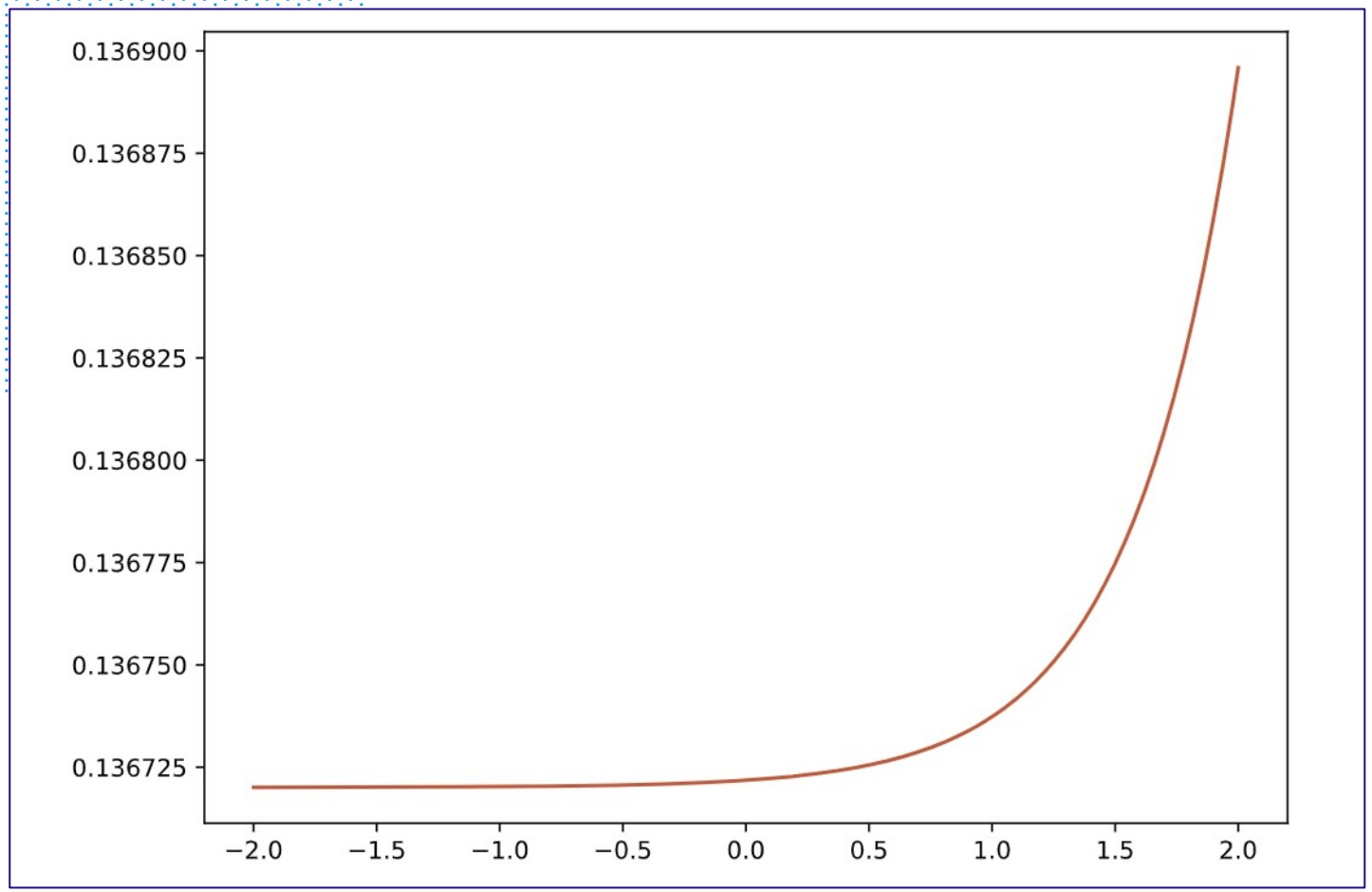


Training =0.8958

Validation=0.8873

Testing=0.8199

Ridge Regularization

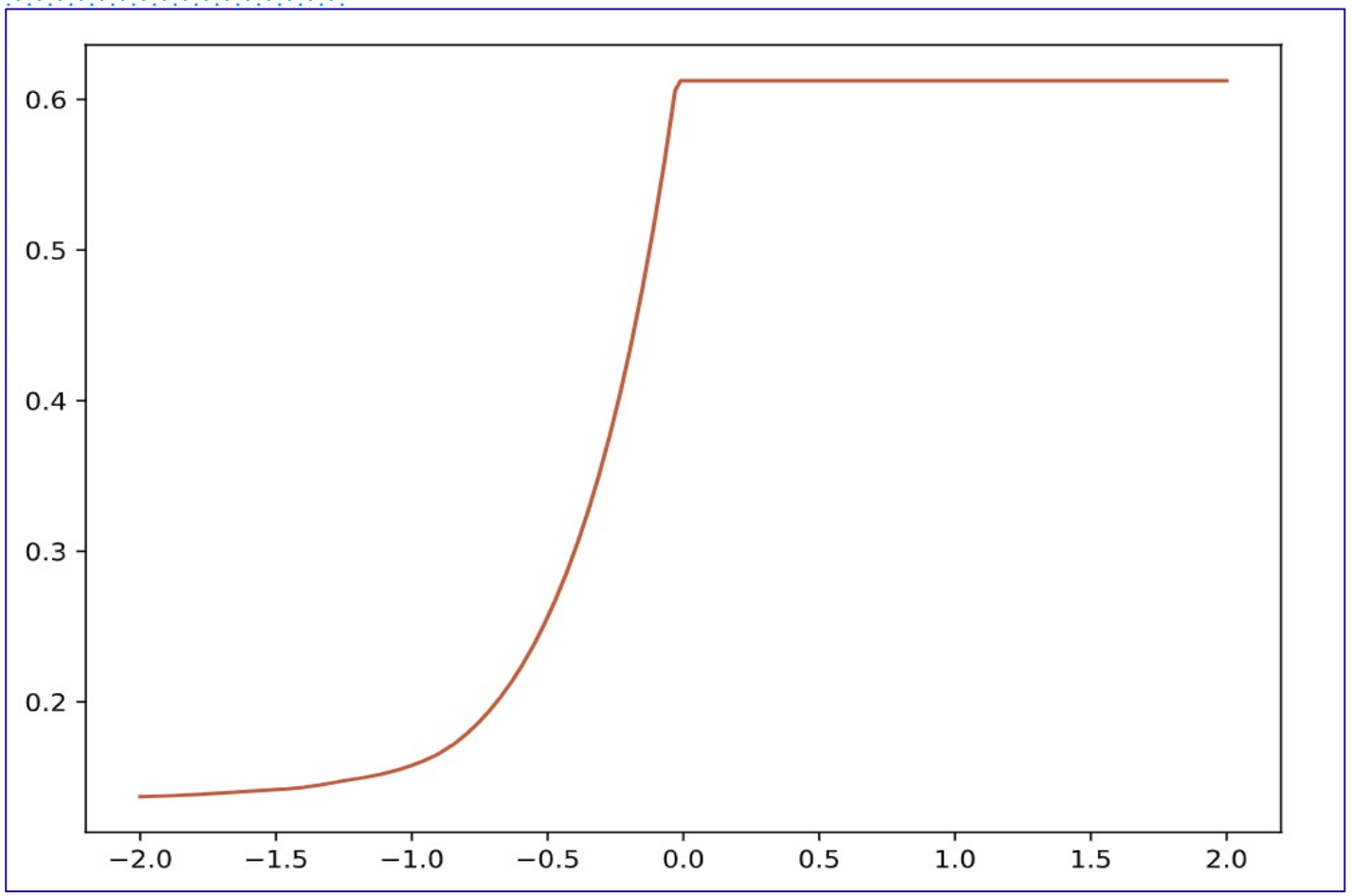


Training =0.8951

Validation=0.8867

Testing=0.8199

Lasso Regularization



Training =0.8958

Validation=0.8873

Testing=0.81995

Conclusion

In the attempts to predict the best model for Fare Amount, we made several 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