Traffic Predictions

Data 621 Final Project

Tommy Jenkins, Violeta Stoyanova, Todd Weigel, Peter Kowalchuk, Eleanor R-Secoquian, Anthony Pagan

20-Dec-2019

Abstract

The purpose of this project is to build various models in an attempt to predict the trip duration of yellow taxis in New York City.

Using the techniques we've learned in the class, like classification, model diagnostics and transformation, we will explore data to find new patterns. And just like what is required in the Kaggle contest, we will try to predict the duration of each trip in the test set. We will build multiple linear regression modeling and then summary to interpret the results. We’ll further analyze the results by adding discrimination in the model and then assess the discrimination with ROC curve.

Introduction

“The Partnership for New York City's one-page study asserted that “excess congestion” deprives the five boroughs and the suburbs of Long Island, Westchester and Rockland counties and northern New Jersey $20 billion annually,” according to Crain’s. Moreover, emergency services reports life or death consequences based on NYC traffic conditions, according to multiple sources. We seek to analyze publicly accessible datasets in the service of better understanding the factors that influence traffic as encapsulated in the Kaggle competition: New York City Taxi Trip Duration . In a recent article, Emil Skandul, writing for City and State NY, cited the importance of Mayor De Blasio hiring a technologically savvy commissioner to lead the Taxi and Limousine Commission to rival Lyft and Uber. In all of these regards, we believe being able to predict traffic through the proxy of taxi trip duration can be highly relevant to the welfare of the NYC population.

Method

Using the techniques we've learned in the class, like classification, model diagnostics and transformation, we will explore data to find new patterns. And just like what is required in the Kaggle contest, we will try to predict the duration of each trip in the test set. We will build multiple linear regression modeling and then summary to interpret the results. We’ll further analyze the results by adding discrimination in the model and then assess the discrimination with ROC curve.

We will consider following the competition evaluation metrics and results format, only in cases where it fully agrees with the recommendations of our DATA 621 coursework.

Data Exploration

The features we will initially consider can be found under the data tab in the competition page and we may choose to augment the data with information such as weather.

File descriptions

- train.csv - the training set (contains 1458644 trip records)

- test.csv - the testing set (contains 625134 trip records)

Data fields

- id - a unique identifier for each trip

- vendor_id - a code indicating the provider associated with the trip record

- pickup_datetime - date and time when the meter was engaged

- dropoff_datetime - date and time when the meter was disengaged

- passenger_count - the number of passengers in the vehicle (driver entered value)

- pickup_longitude - the longitude where the meter was engaged

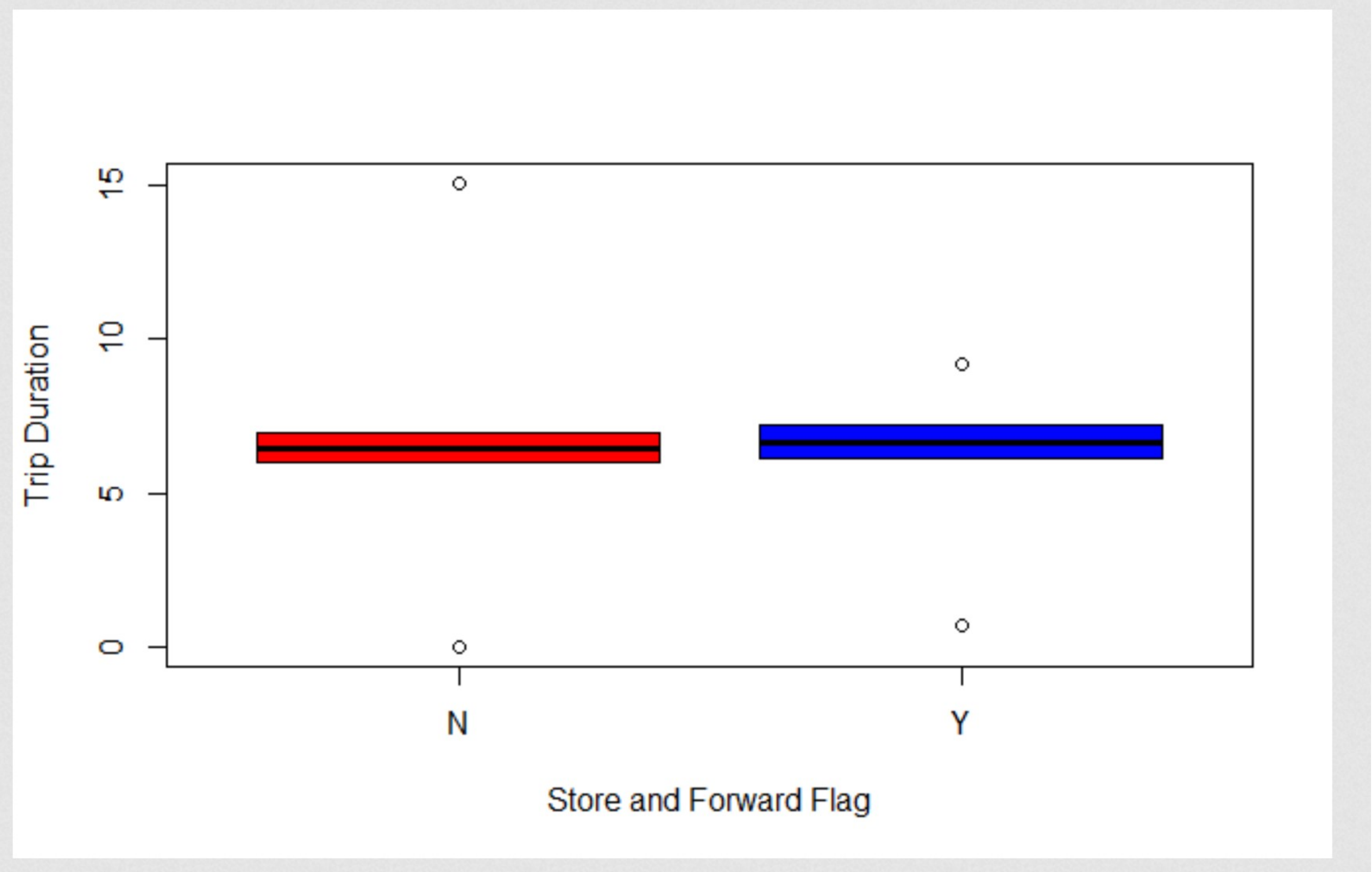
- pickup_latitude - the latitude where the meter was engaged

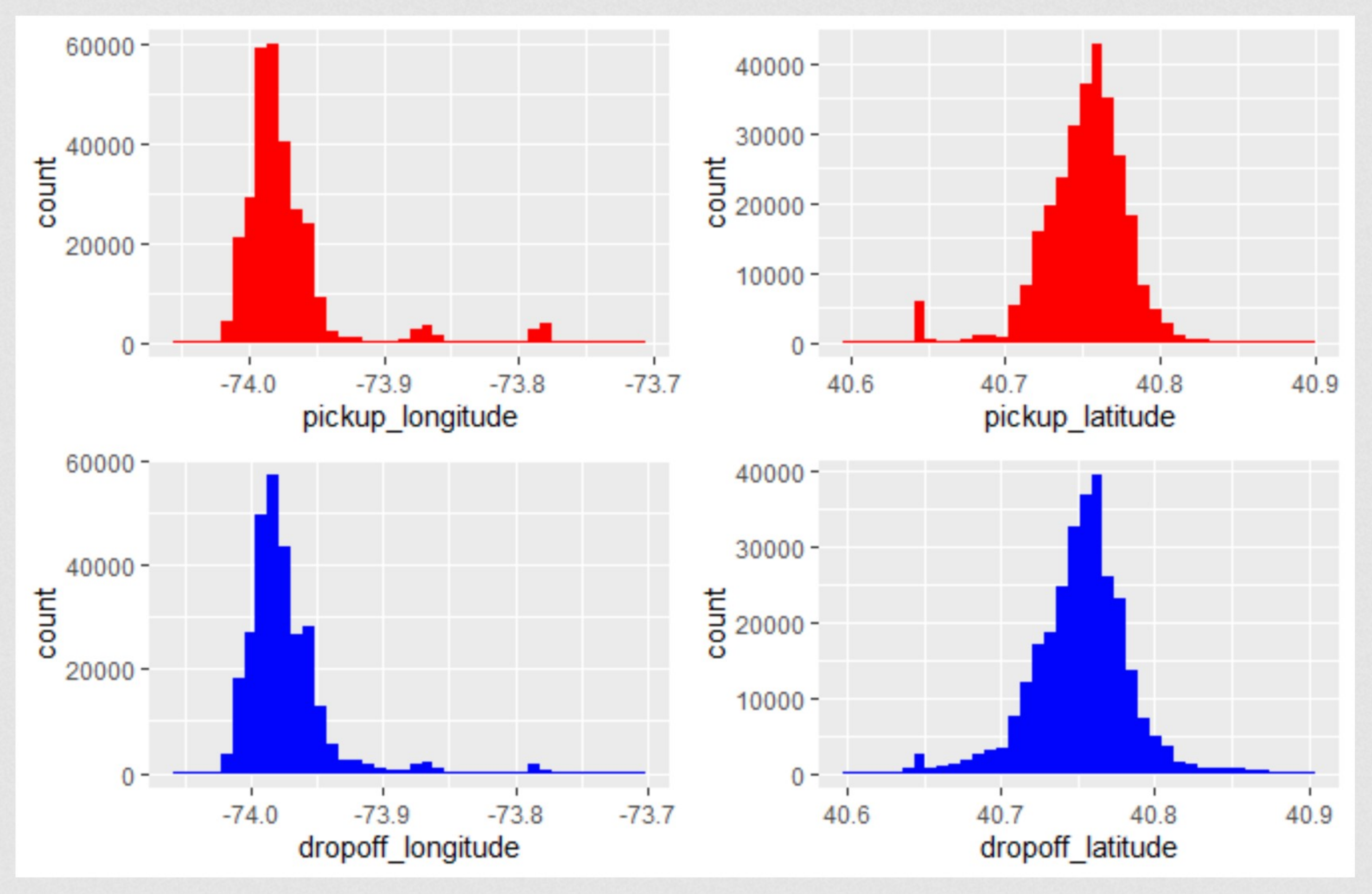
- dropoff_longitude - the longitude where the meter was disengaged

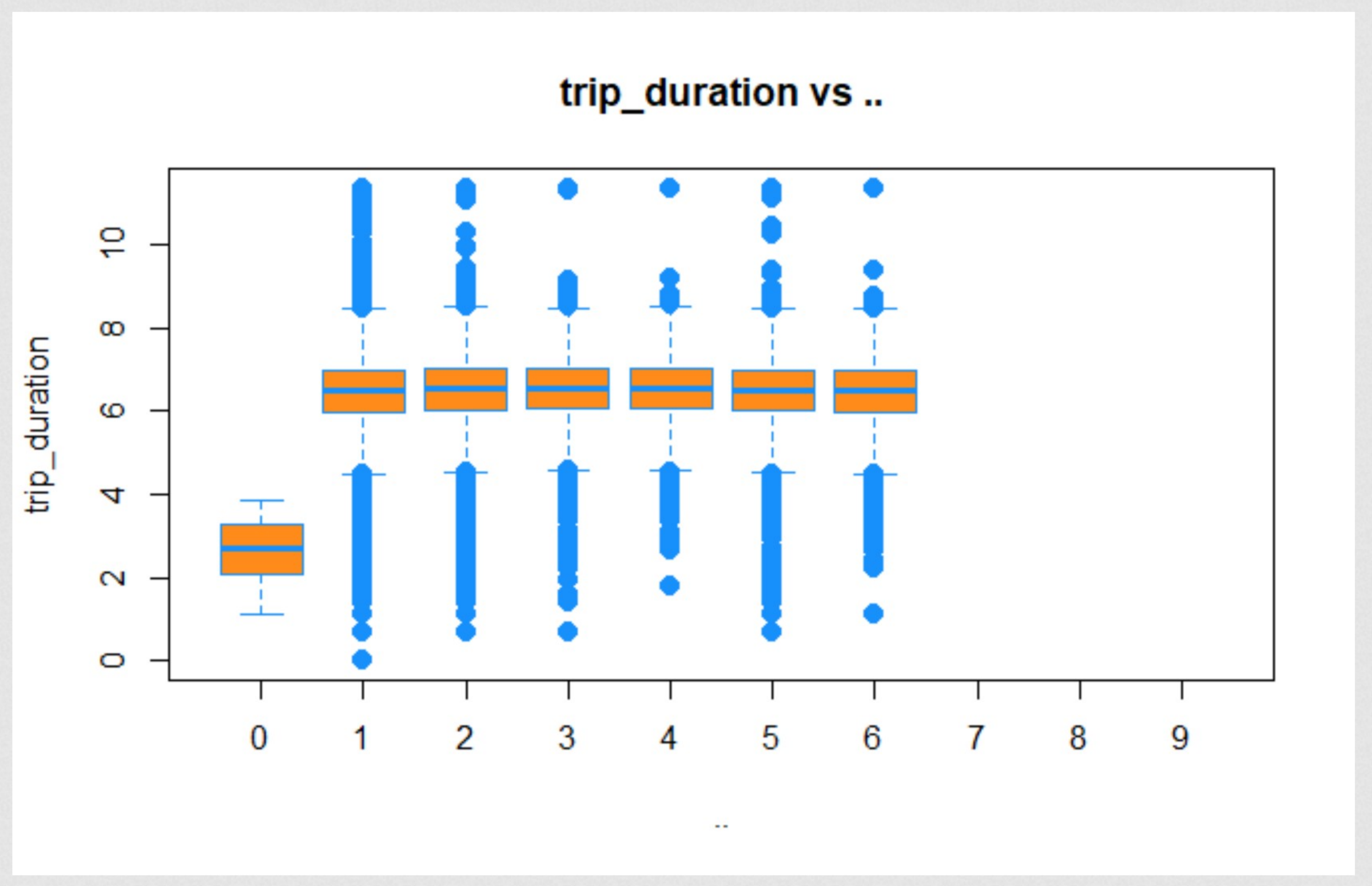
- dropoff_latitude - the latitude where the meter was disengaged

- store_and_fwd_flag - This flag indicates whether the trip record was held in vehicle memory before sending to the vendor because the vehicle did not have a connection to the server - Y=store and forward; N=not a store and forward trip

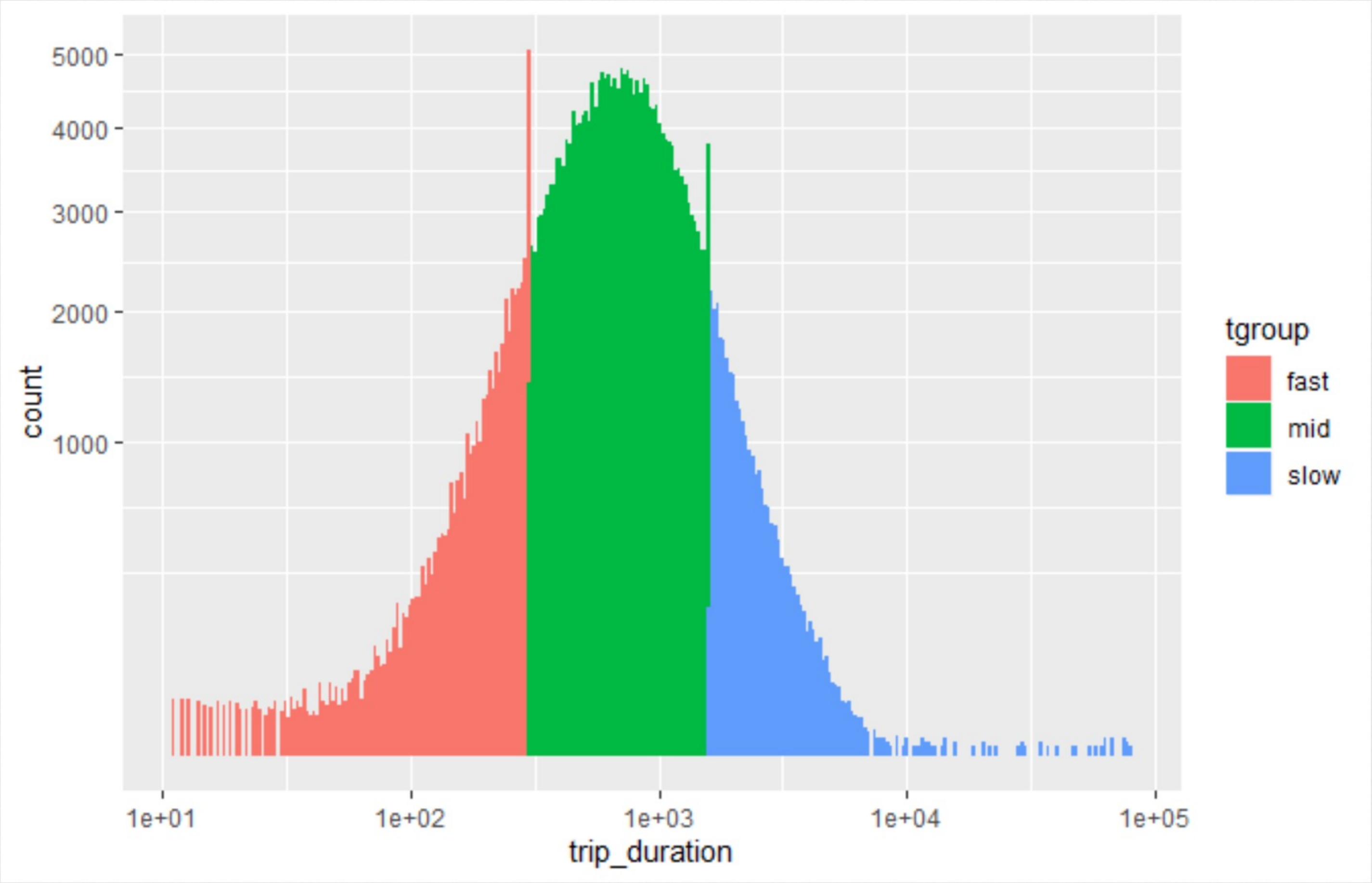
- trip_duration - duration of the trip in seconds

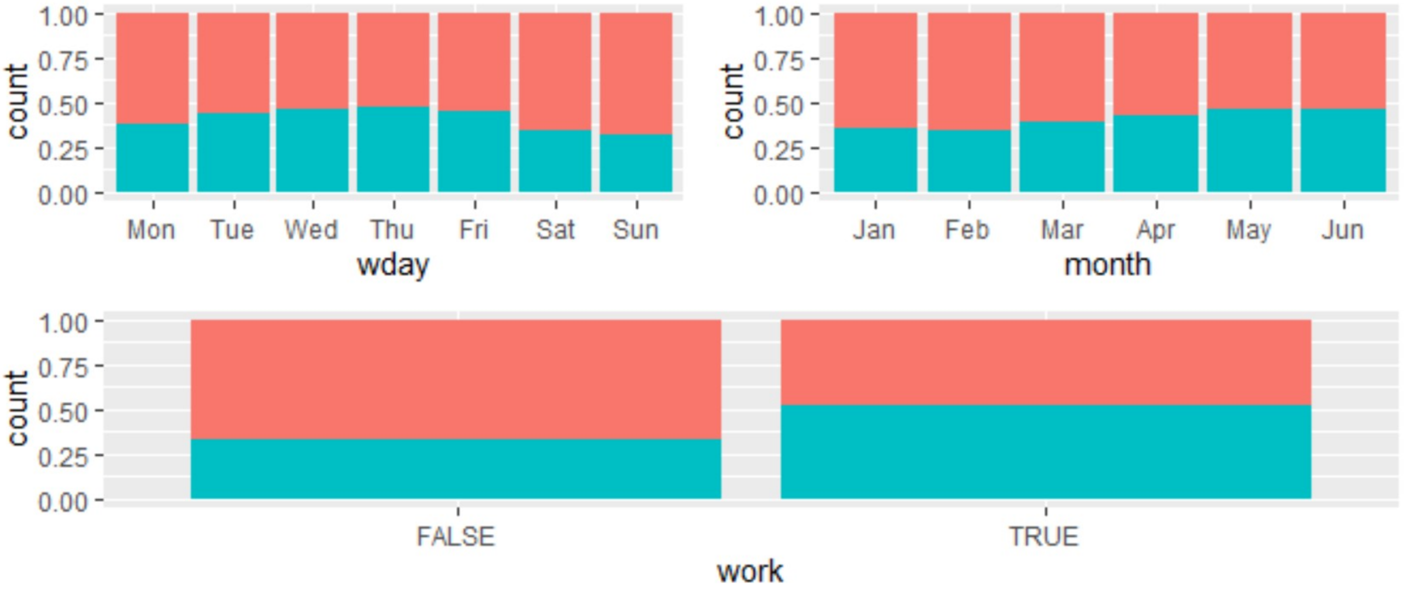

Trip Duration
0
5
10
15
N
Y
Store and Forward Flag

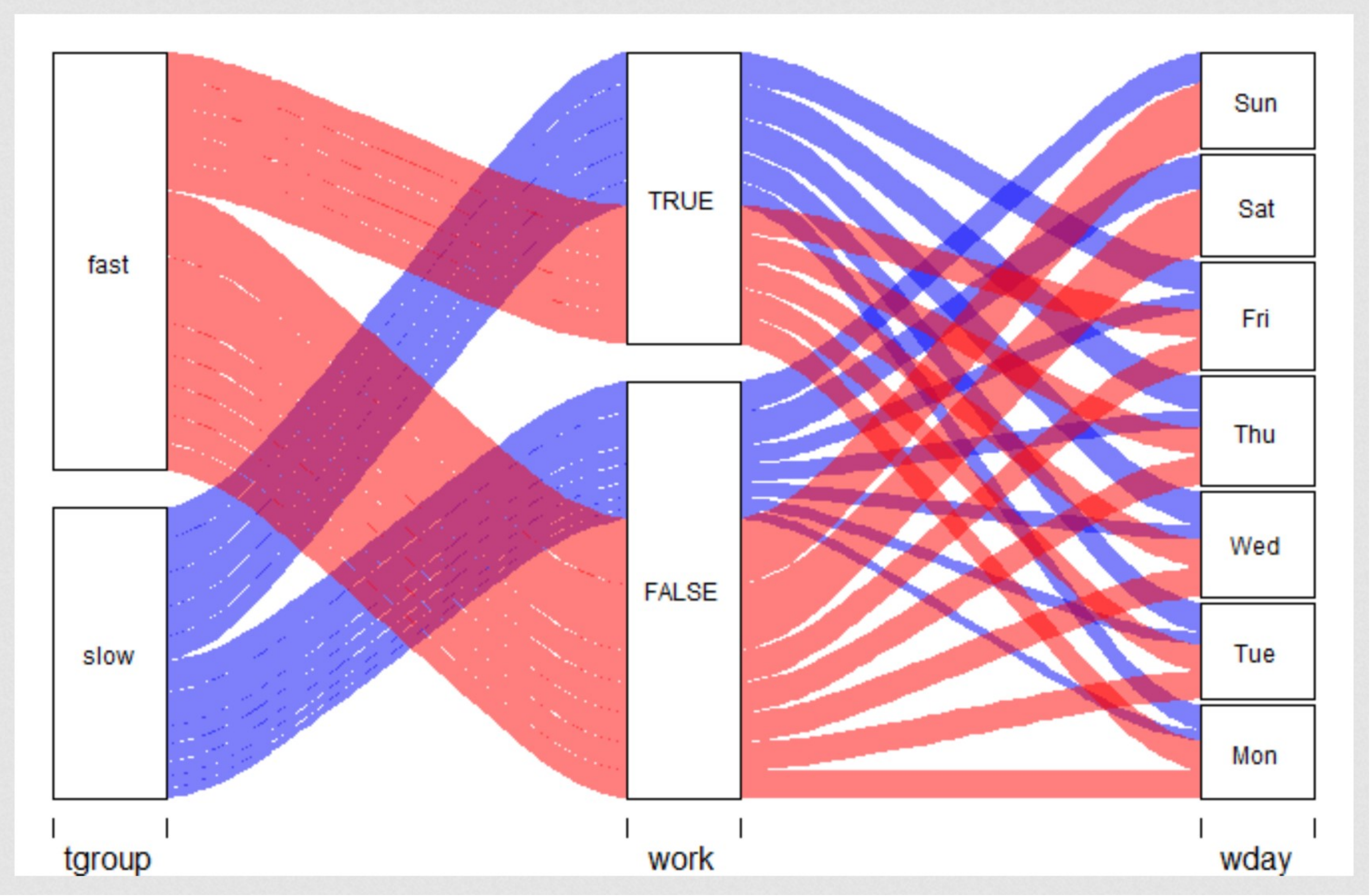

60000
40000
20000
0
count
-74.0
-73.9
-73.8
-73.7
pickup_longitude
40000
30000
20000
10000
0
count
40.6
40.7
40.8
40.9
pickup_latitude
60000
40000
20000
0
count
-74.0
-73.9
-73.8
-73.7
dropoff_longitude
40000
30000
20000
10000
0
count
40.6
40.7
40.8
40.9
dropoff_latitude


trip_duration vs ..
trip_duration
0
2
4
6
8
10
0
1
2
3
4
5
6
7
8
9
..

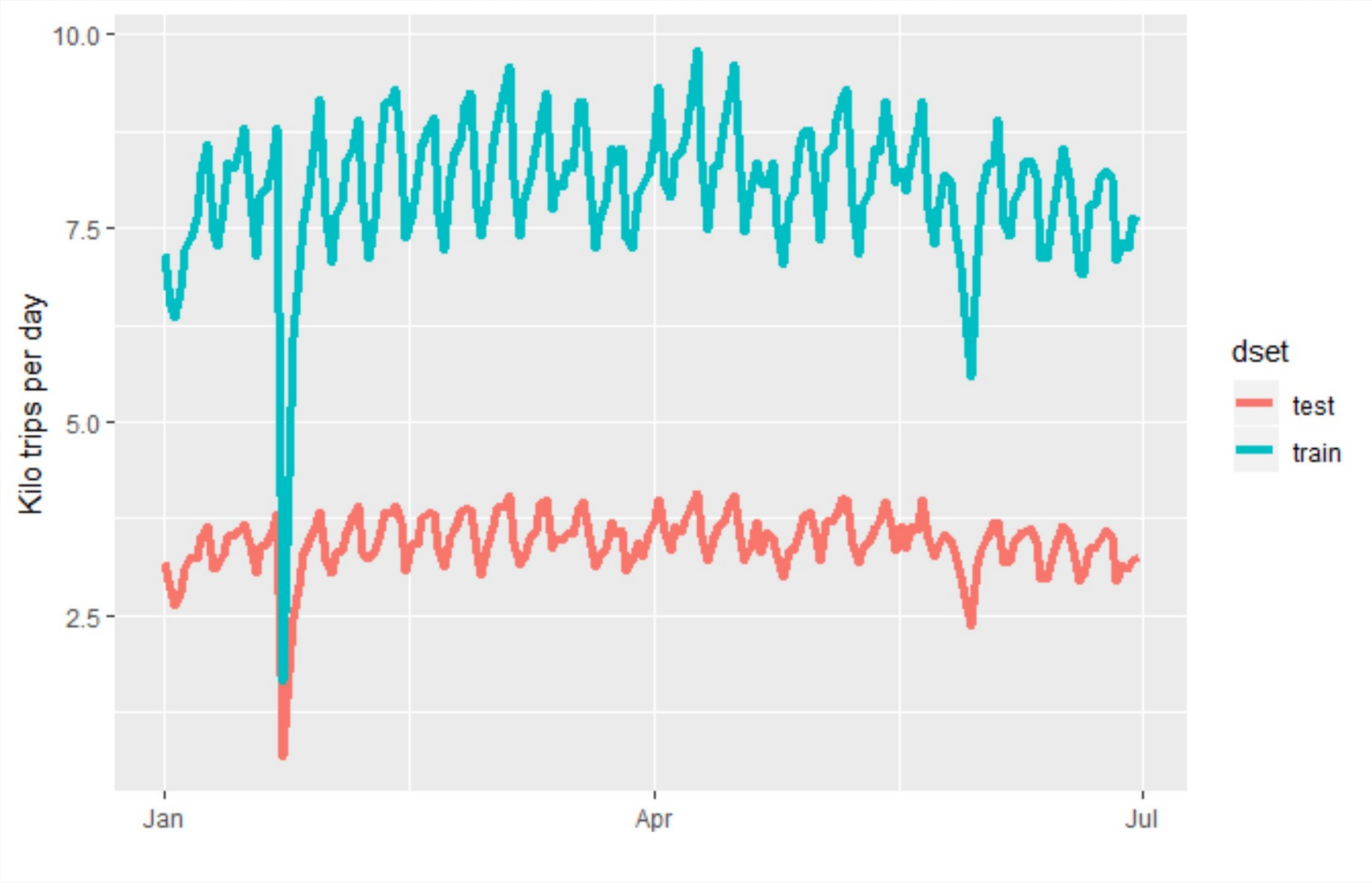
An excursion into classification

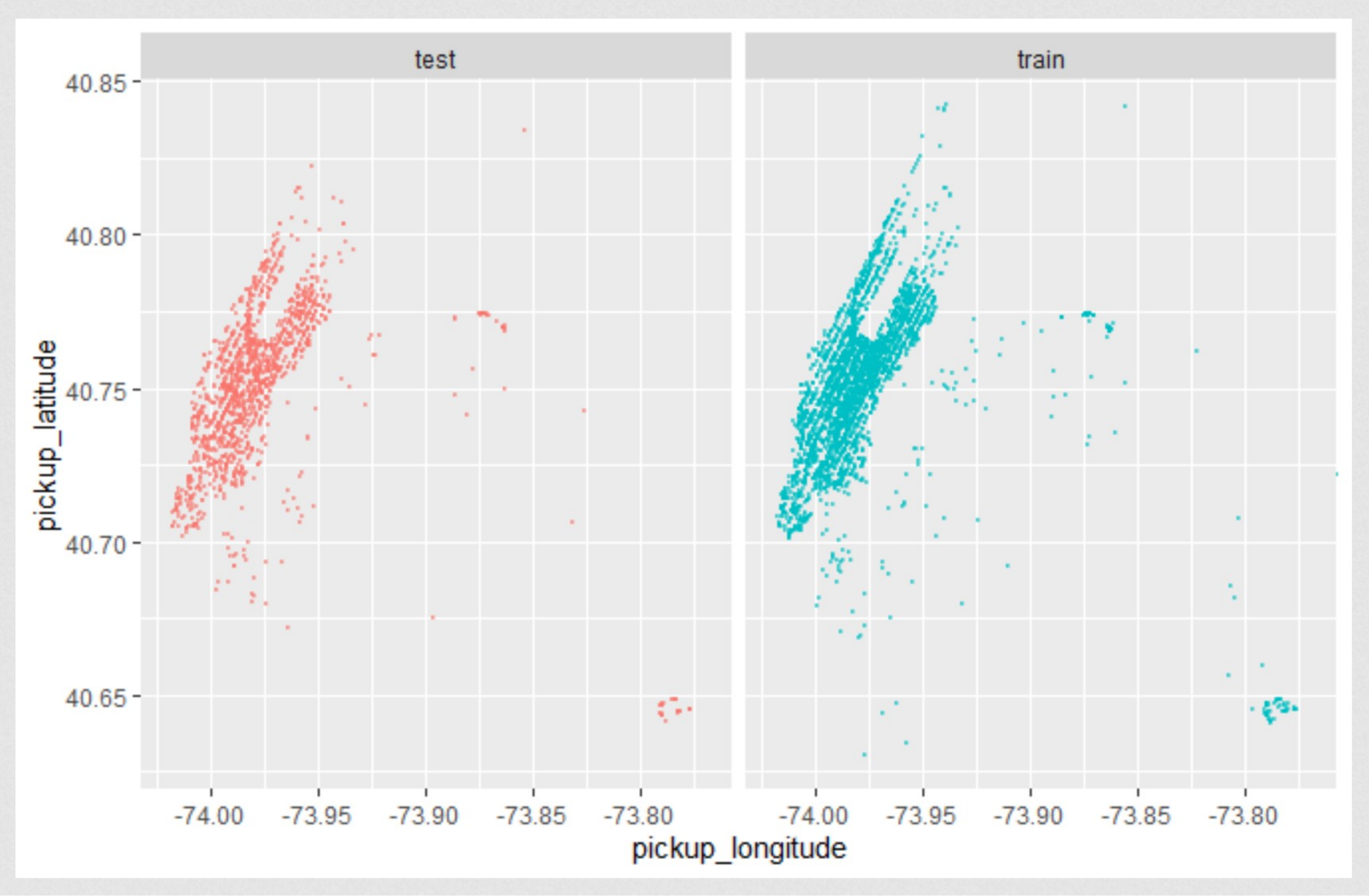


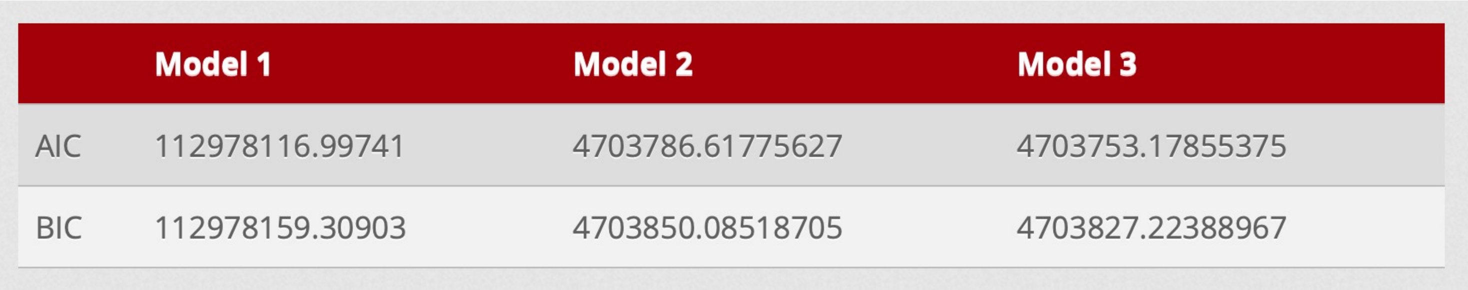

count
1.00
0.75
0.50
0.25
0.00
Mon
Tue
Wed
Thu
Fri
Sat
Sun
wday
count
1.00
0.75
0.50
0.25
0.00
Jan
Feb
Mar
Apr
May
Jun
month
count
1.00
0.75
0.50
0.25
0.00
FALSE
TRUE
work


fast
slow
TRUE
FALSE
Sun
Sat
Fri
Thu
Wed
Tue
Mon
tgroup
work
wday

Model, correlation





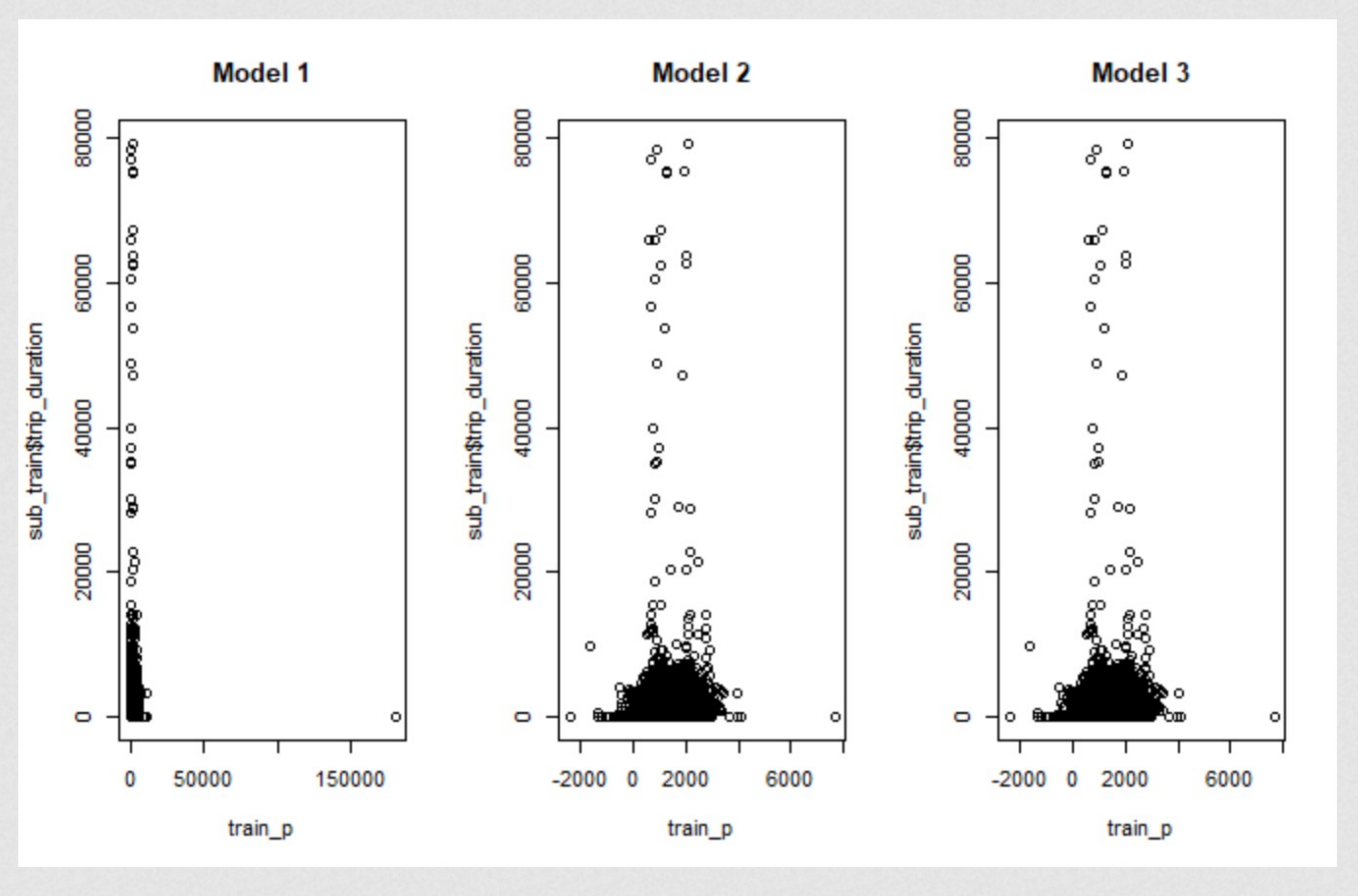


	Model 1	Model 2	Model 3
AIC	112978116.99741	4703786.61775627	4703753.17855375
BIC	112978159.30903	4703850.08518705	4703827.22388967

Selection Model Result

Conclusion

With 3 models computed, we select the model with the lowest combination of AIC and BIC. From the table, we can see the model to pick is model 1.



Conclusion

Model 1 showed the best result. We can observe its performance by plotting the datasets Vendor_ID values against the predicted values.