MVP

TLC TRIP RECORD DATA

Done by:
Arwa Alolyani Maitha Alqahtani

Submit to: Mr.Ali El-kassas

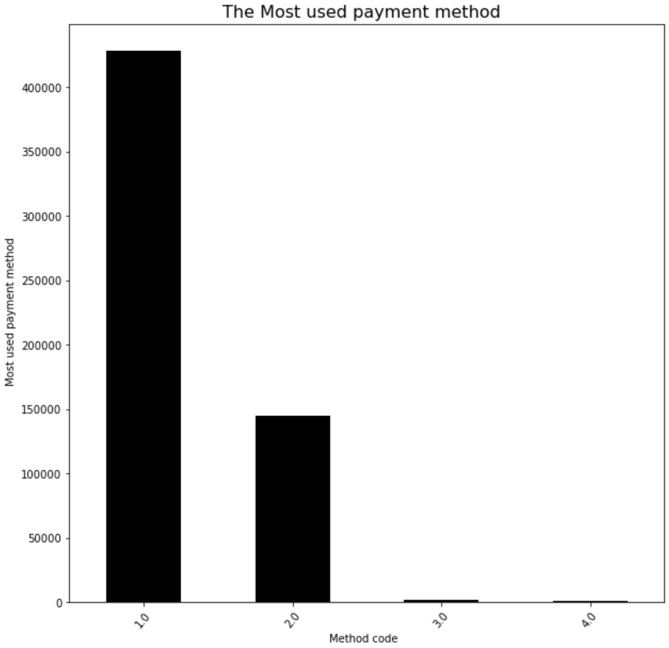
TLC TRIP RECORD DATA

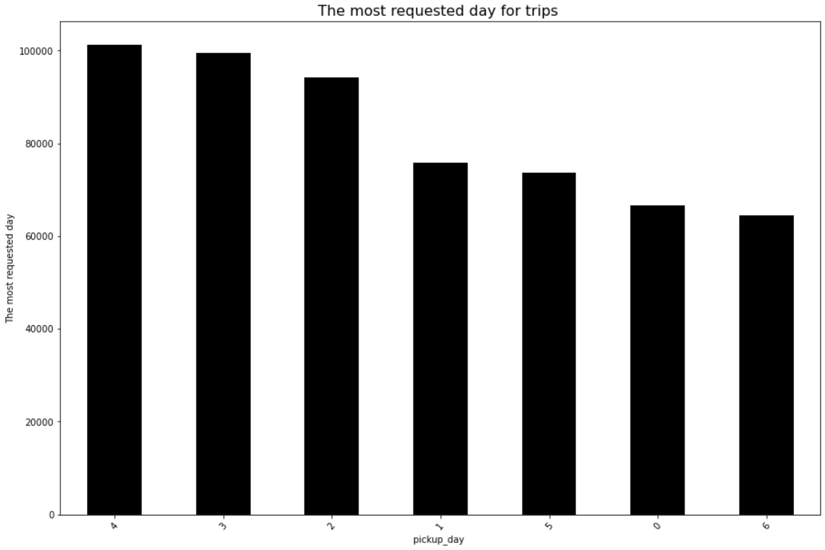
In our study, we selected a sample of basic data containing more than 6 million rows and created a million-row sample.
Our plans are to divide the data into three sections: training, verification, and testing.
Testing and Verification contains 200,000 rows and 18 columns, while training contains 600,000 rows and 18 columns.
This data is lacking coordinates for each location, so the NYC Taxi Zones (GeoJSON) file was downloaded from NYC OpenData.

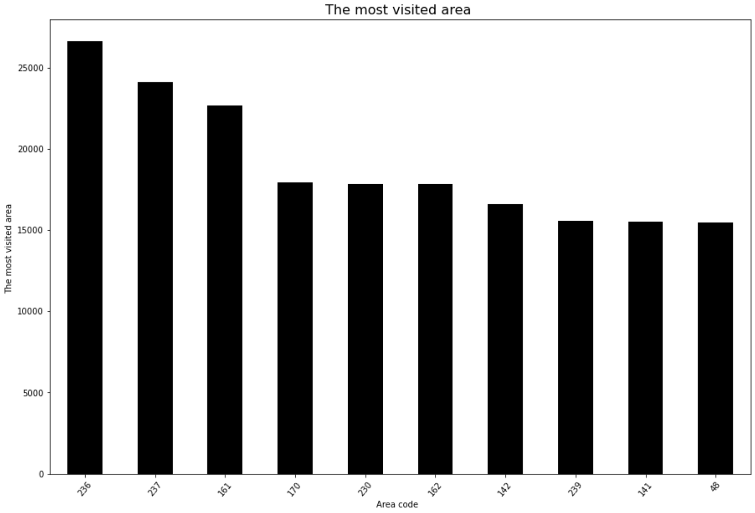
process:

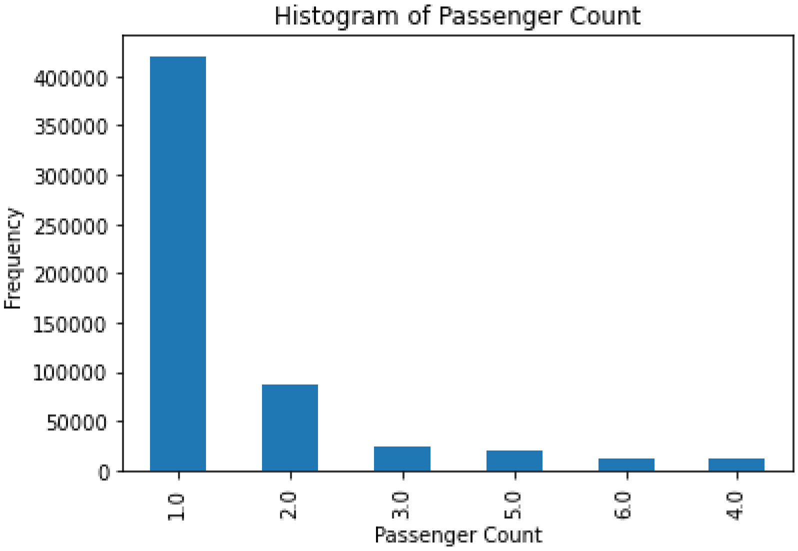
To begin with, we cleaned the data by removing outliers and nulls, converting datetime from object type to datetime type, adding new columns for day and hour, and then extracted latitude and longitude coordinates by geometric column transformation from NYC Taxi Zones database.
The second stage of data extraction analysis:

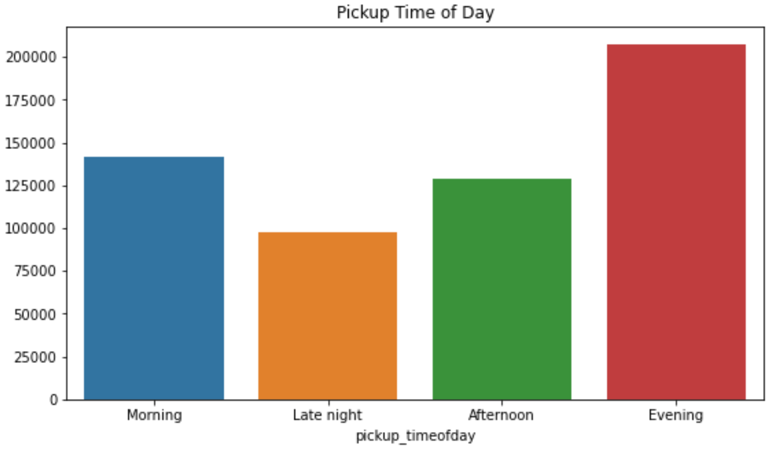
The following data was extracted
1- Preparing passengers for each flight
2- Total price for trips
3- the most visited area
4- the most requested day for trips
5- the number of trips per hour
6- the most payment method used
7- A color caller for the data, which shows how closely they are related to each other

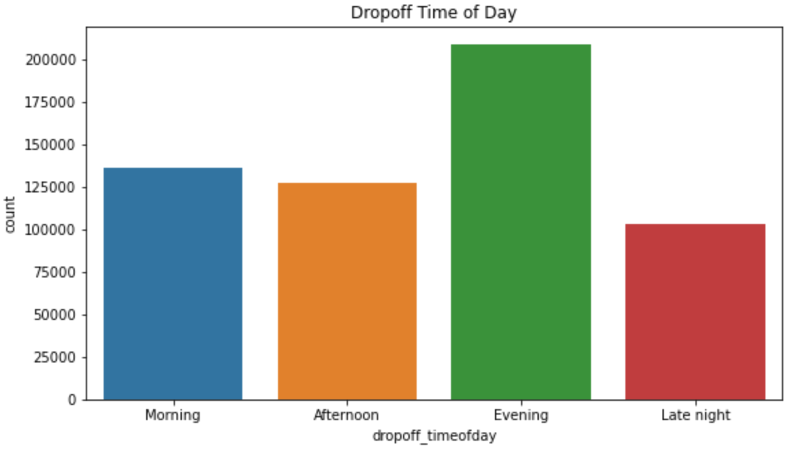


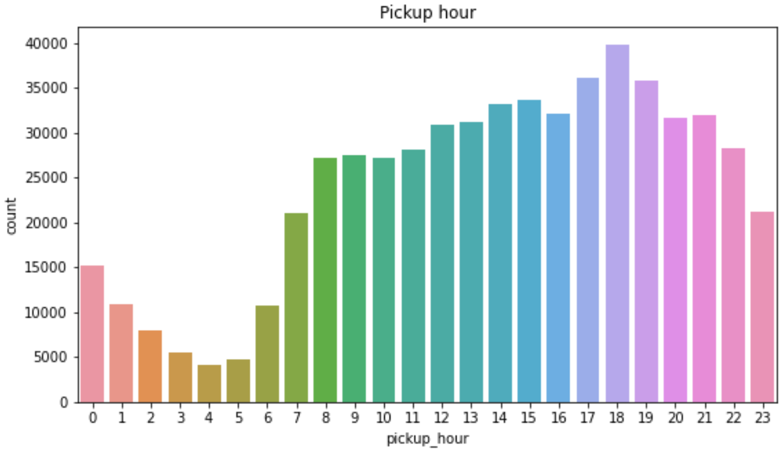


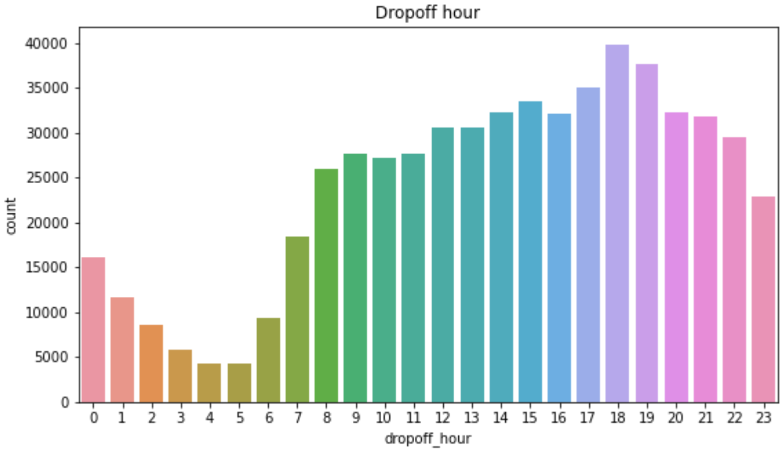

The most visited area
The most visited area
0
5000
10000
15000
20000
25000
236
237
161
170
230
162
142
239
141
48
Area code

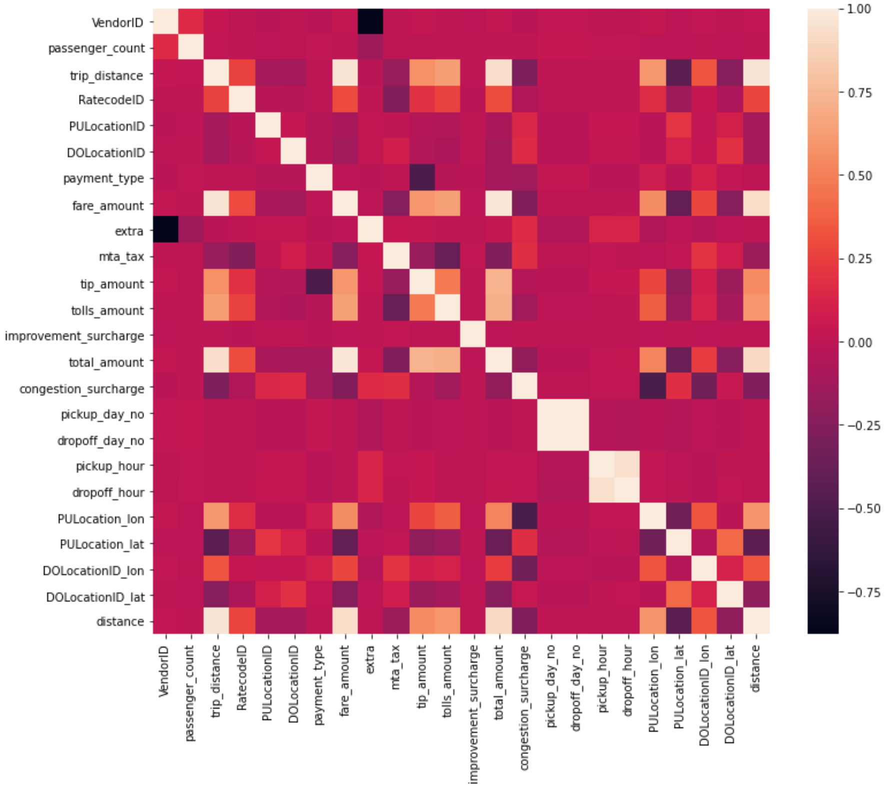

Histogram of Passenger Count
Frequency
0
50000
100000
150000
200000
250000
300000
350000
400000
1.0
2.0
3.0
5.0
6.0
4.0
Passenger Count


Pickup Time of Day
0
25000
50000
75000
100000
125000
150000
175000
200000
Morning
Late night
Afternoon
Evening
pickup_timeofday


Dropoff Time of Day
count
0
25000
50000
75000
100000
125000
150000
175000
200000
Morning
Afternoon
Evening
Late night
dropoff_timeofday

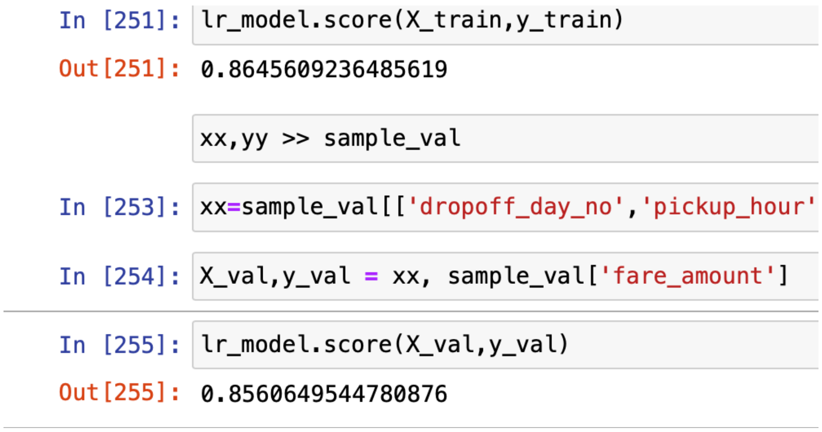

Pickup hour
count
0
5000
10000
15000
20000
25000
30000
35000
40000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
pickup_hour


Dropoff hour
count
0
5000
10000
15000
20000
25000
30000
35000
40000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
dropoff_hour



The third stage is the application of machine learning to data, and in this stage there are several interconnected sequential steps

1- Apply the model to the training data and print the ratio
2- Apply the R2 and print the ratio
3- Apply the model to the verification data and print the ratio
4- Apply the R2 and print the ratio



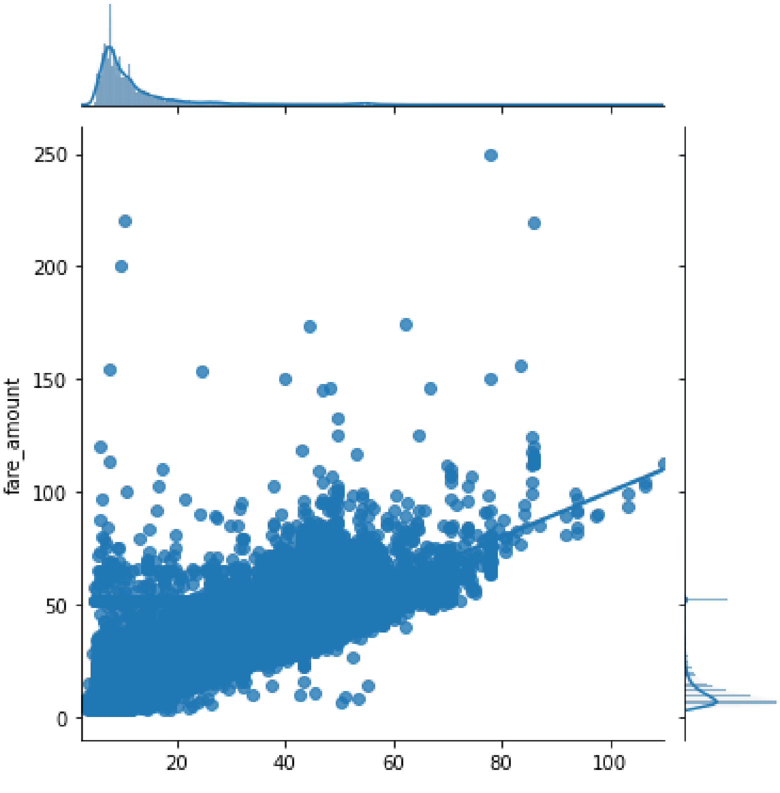
```
In [251]: lr_model.score(X_train,y_train)
Out[251]: 0.8645609236485619

          xx,yy >> sample_val

In [253]: xx=sample_val[['dropoff_day_no','pickup_hour'

In [254]: X_val,y_val = xx, sample_val['fare_amount']

In [255]: lr_model.score(X_val,y_val)
Out[255]: 0.8560649544780876
```


fare_amount