New York City Taxi Trip Duration

Siyi Yu

Jilin University
Computer Science And Technology

April 26, 2021

Overview

Project Introduction

- Project Introduction

Data Analysis

- Info
- Spearman Correlation
- Simple Attributes
- Complex Attributes
- Labels

Model selection

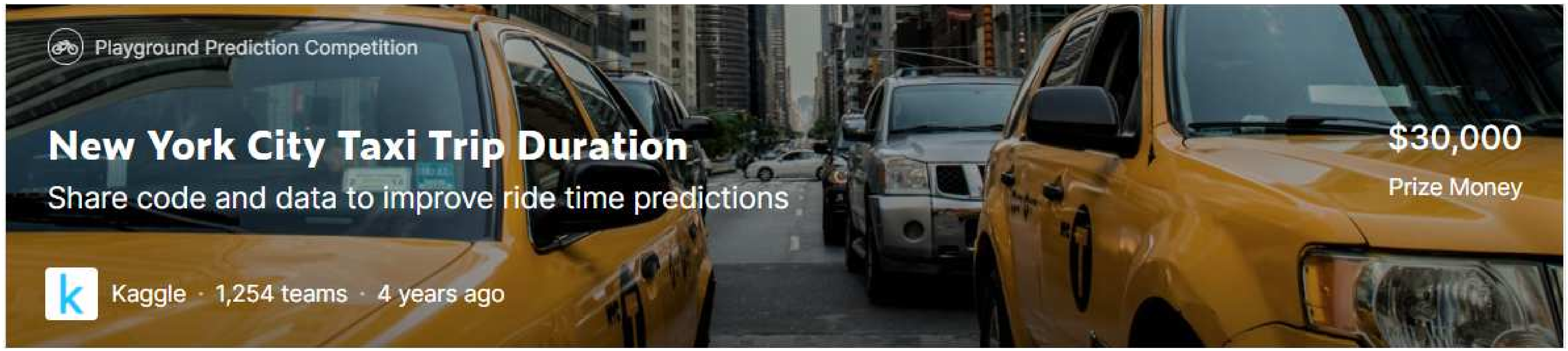
- Models

Results

- Results

Project Introduction

Project Introduction


Playground Prediction Competition

New York City Taxi Trip Duration

Share code and data to improve ride time predictions

$30,000

Prize Money

Kaggle · 1,254 teams · 4 years ago

Intro

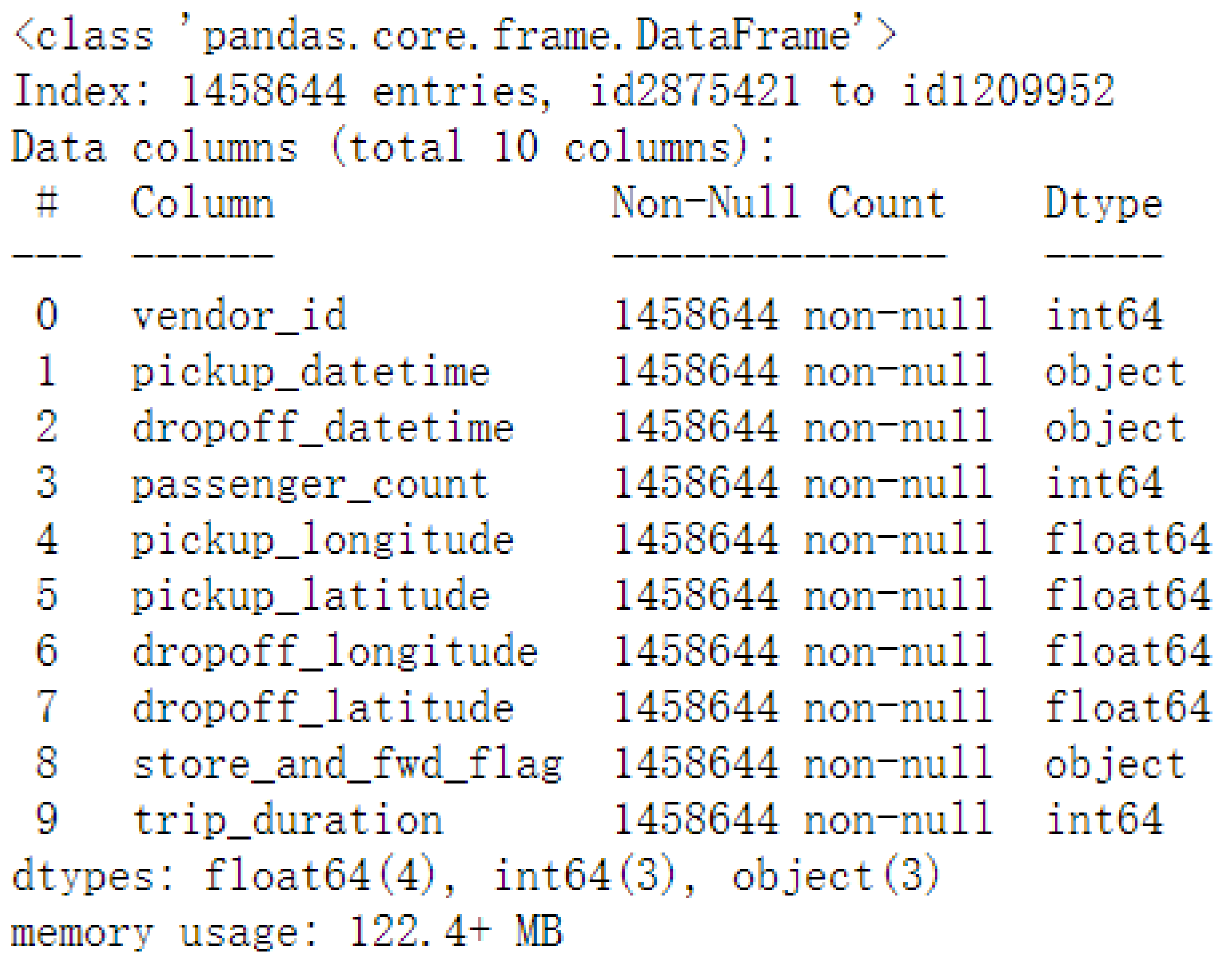
Data Fields for This Project

- id, vendor_id, store_and_fwd_flag, passenger_count
- pickup_datetime,dropoff_datetime
- pickup_longitude,pickup_latitude,dropoff_longitude,dropoff_latitude
- trip_duration

Data Analysis

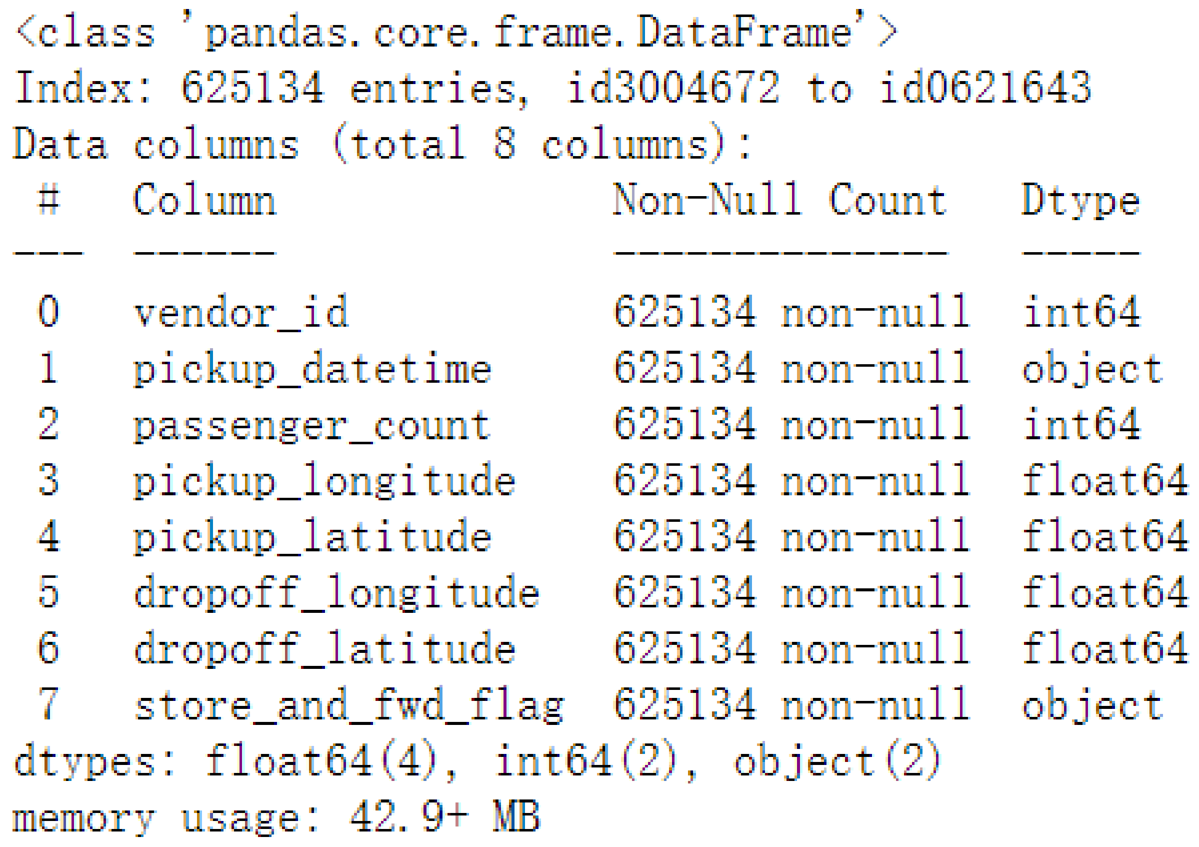
Info

train:



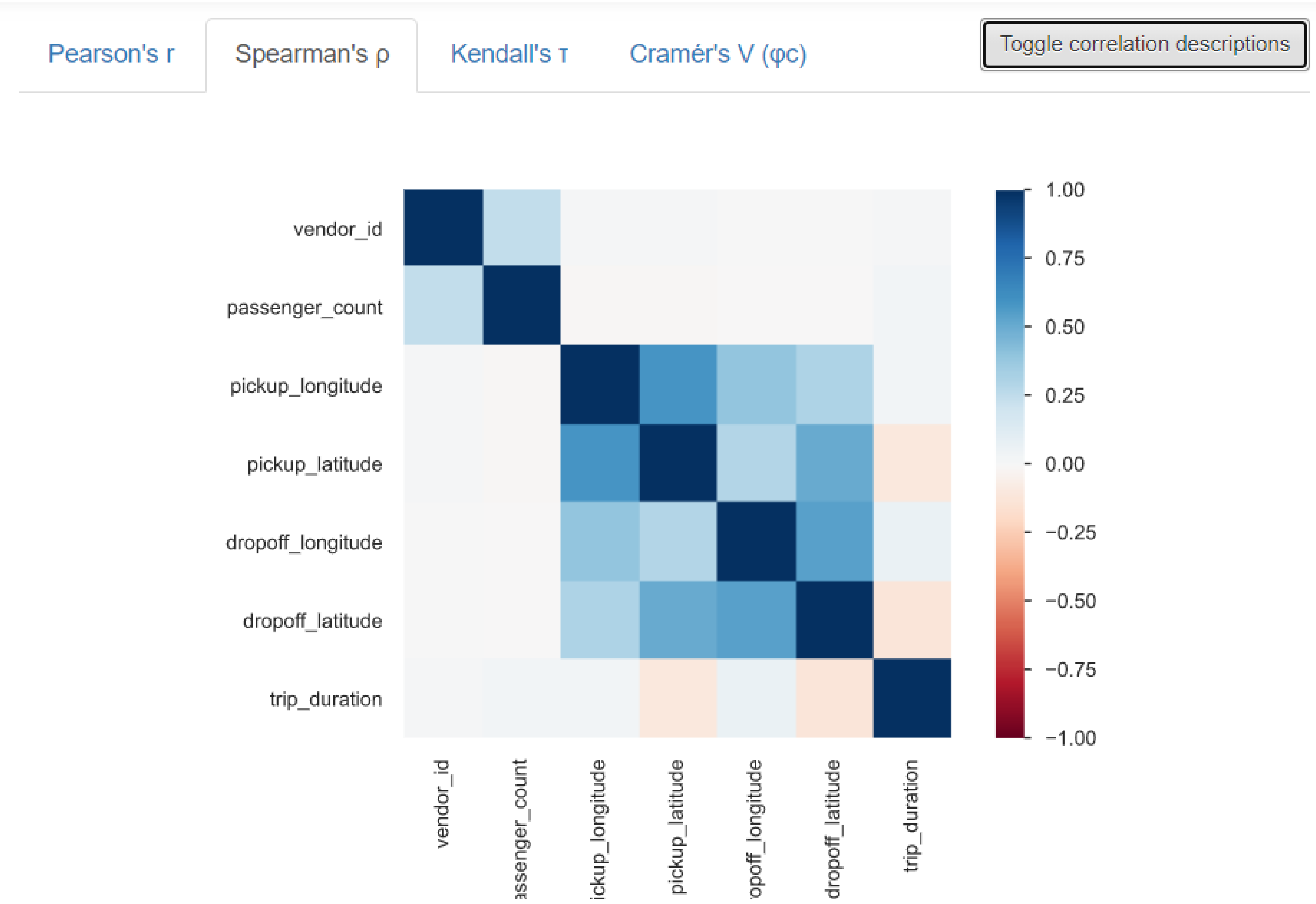
```
<class 'pandas.core.frame.DataFrame'>
Index: 1458644 entries, id2875421 to id1209952
Data columns (total 10 columns):
 #   Column              Non-Null Count    Dtype
---  ------              --------------    -----
 0   vendor_id           1458644 non-null  int64
 1   pickup_datetime     1458644 non-null  object
 2   dropoff_datetime    1458644 non-null  object
 3   passenger_count     1458644 non-null  int64
 4   pickup_longitude    1458644 non-null  float64
 5   pickup_latitude     1458644 non-null  float64
 6   dropoff_longitude   1458644 non-null  float64
 7   dropoff_latitude    1458644 non-null  float64
 8   store_and_fwd_flag  1458644 non-null  object
 9   trip_duration       1458644 non-null  int64
dtypes: float64(4), int64(3), object(3)
memory usage: 122.4+ MB
```

test:

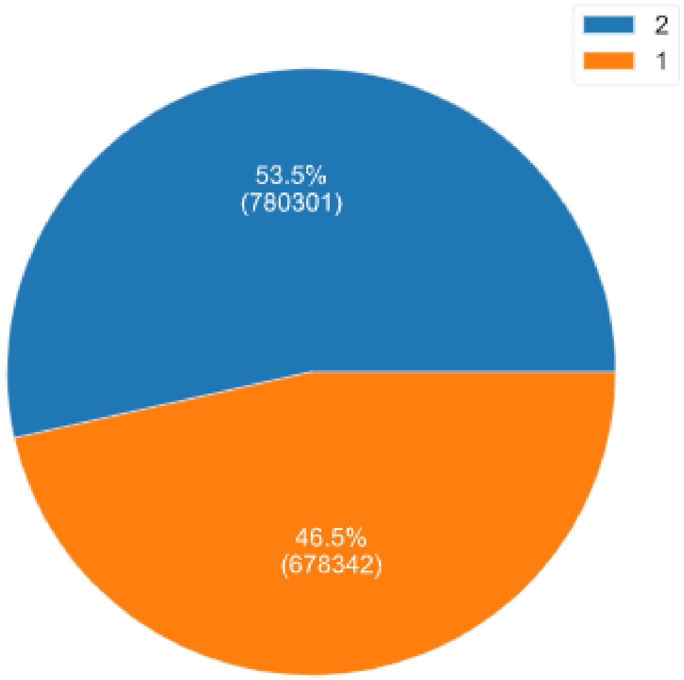


```
<class 'pandas.core.frame.DataFrame'>
Index: 625134 entries, id3004672 to id0621643
Data columns (total 8 columns):
 #   Column              Non-Null Count   Dtype
---  ------              --------------   -----
 0   vendor_id           625134 non-null  int64
 1   pickup_datetime     625134 non-null  object
 2   passenger_count     625134 non-null  int64
 3   pickup_longitude    625134 non-null  float64
 4   pickup_latitude     625134 non-null  float64
 5   dropoff_longitude   625134 non-null  float64
 6   dropoff_latitude    625134 non-null  float64
 7   store_and_fwd_flag  625134 non-null  object
dtypes: float64(4), int64(2), object(2)
memory usage: 42.9+ MB
```

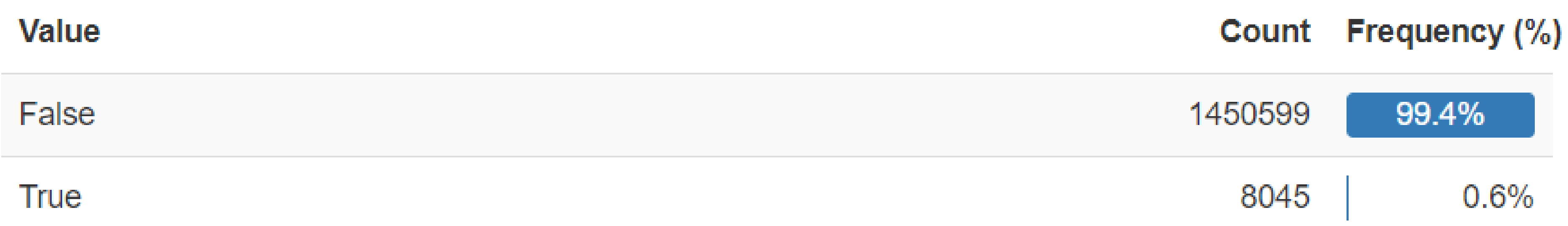
Spearman Correlation



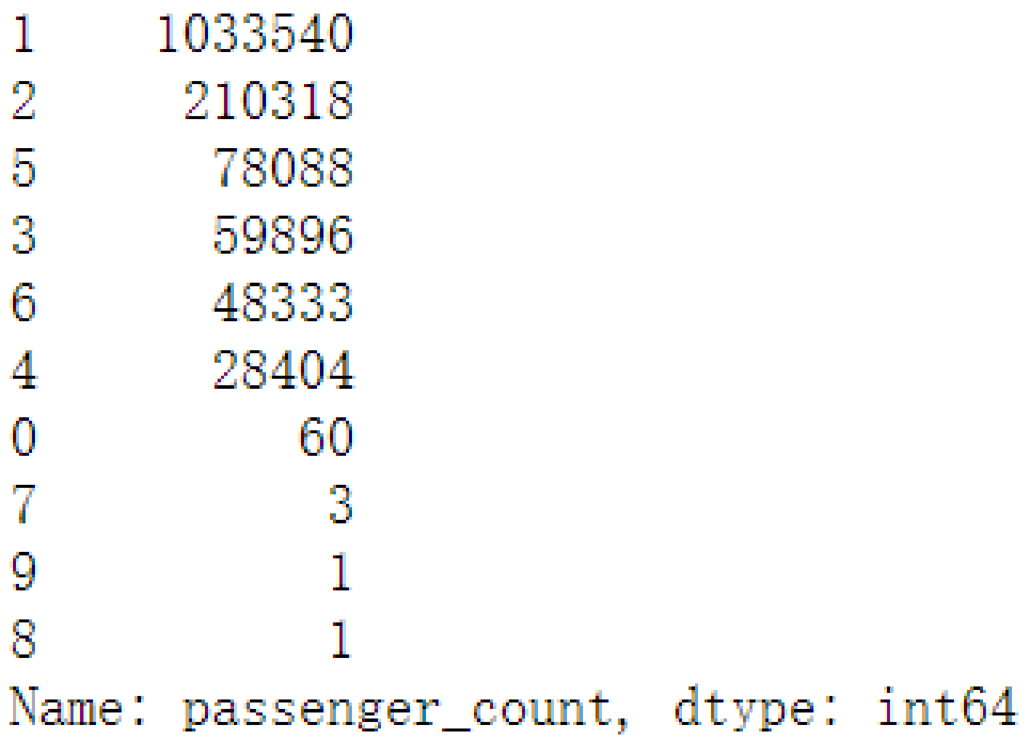
Simple Attributes

- id:Determine the uniqueness of id
- vendor_id:
    - 

    - Convert int64 into string

- store_and_fwd_flag:
- 

Value	Count	Frequency (%)
False	1450599	99.4%
True	8045	0.6%

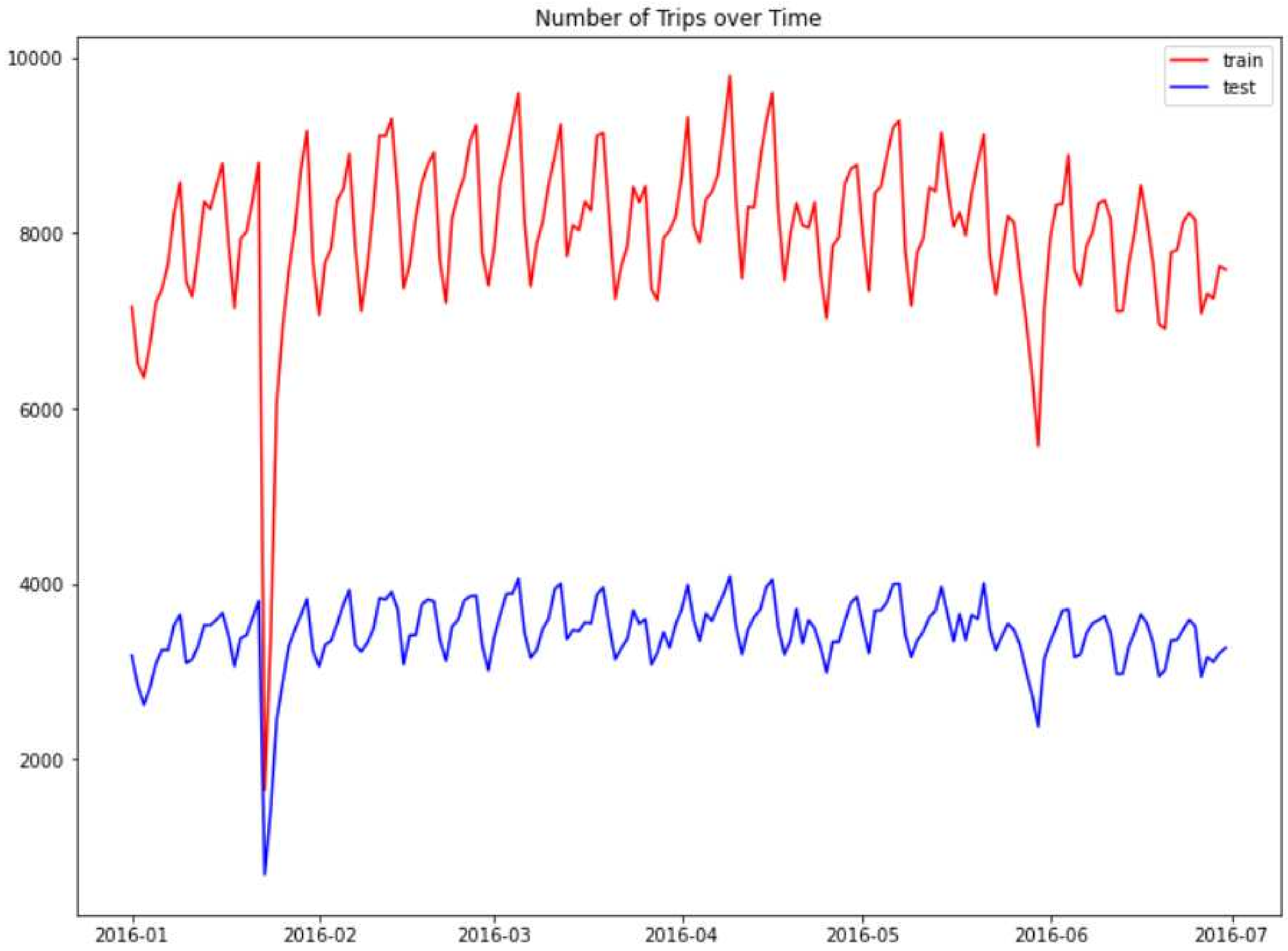
- passenger_count:
- 

```
1    1033540
2     210318
5      78088
3      59896
6      48333
4      28404
0         60
7          3
9          1
8          1
Name: passenger_count, dtype: int64
```

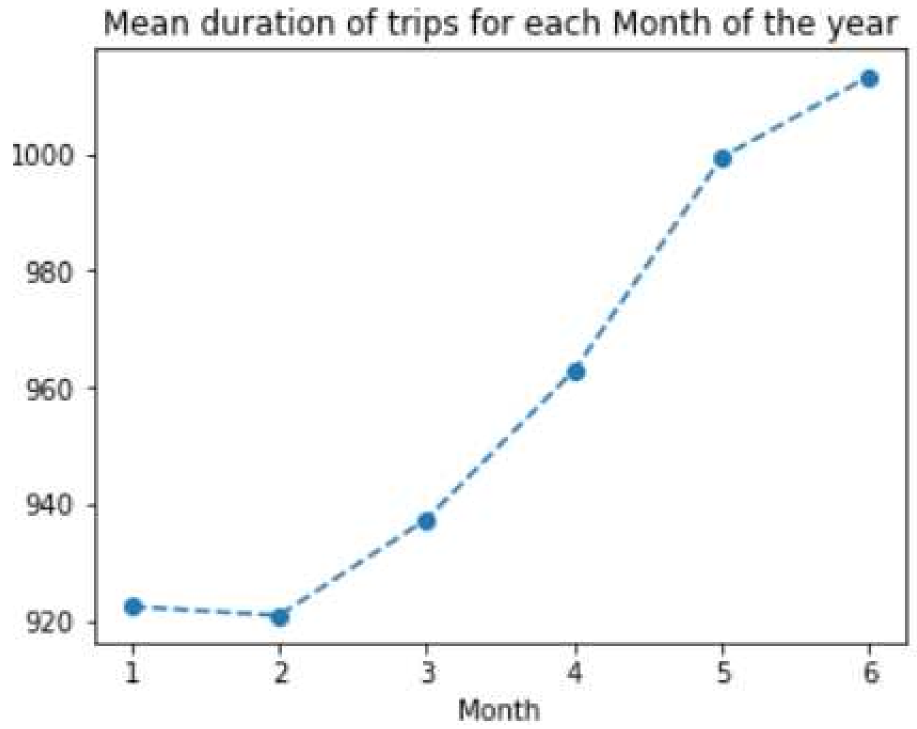
Complex Attributes

- dropoff_datetime:
  - Only exist in test.csv
  - check dropoff_datetime-pickup_datetime == trip_duration
  - drop

- pickup_datetime:
  -

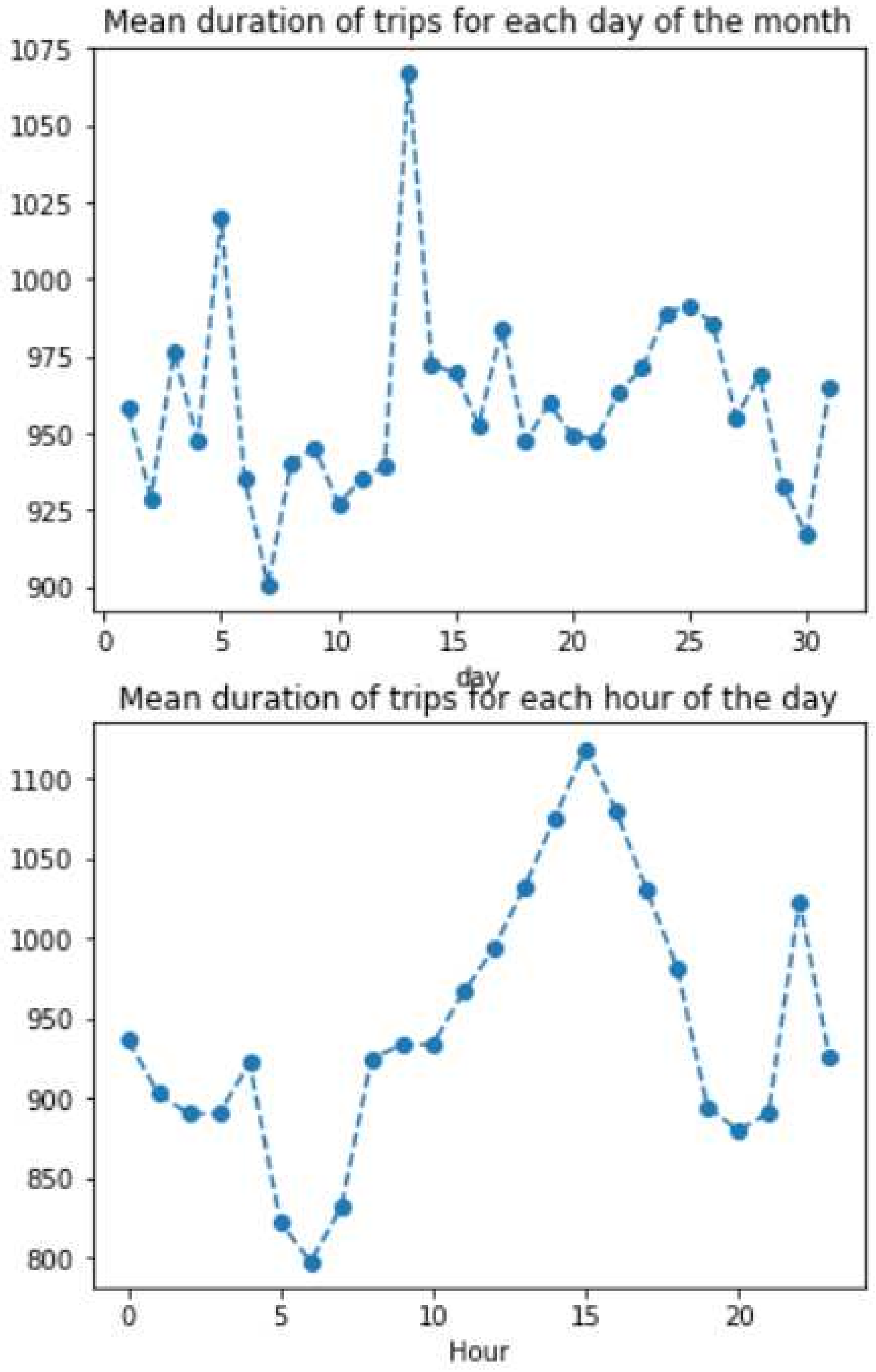


- pickup_datetime:
  - Divide it into year, month, day, hour
  - The year is all 2016,so drop
  -

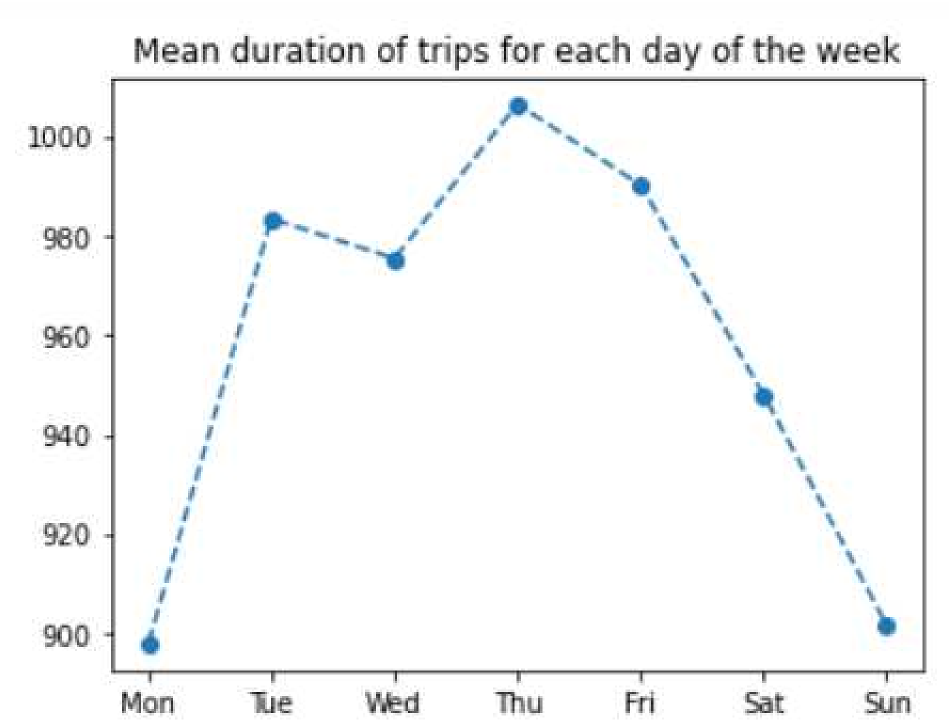


  - Since only six monthes in train and test(1,2,3,4,5,6),convert the type of month into string for one-hot encoding.

- pickup_datetime:
  - ◆

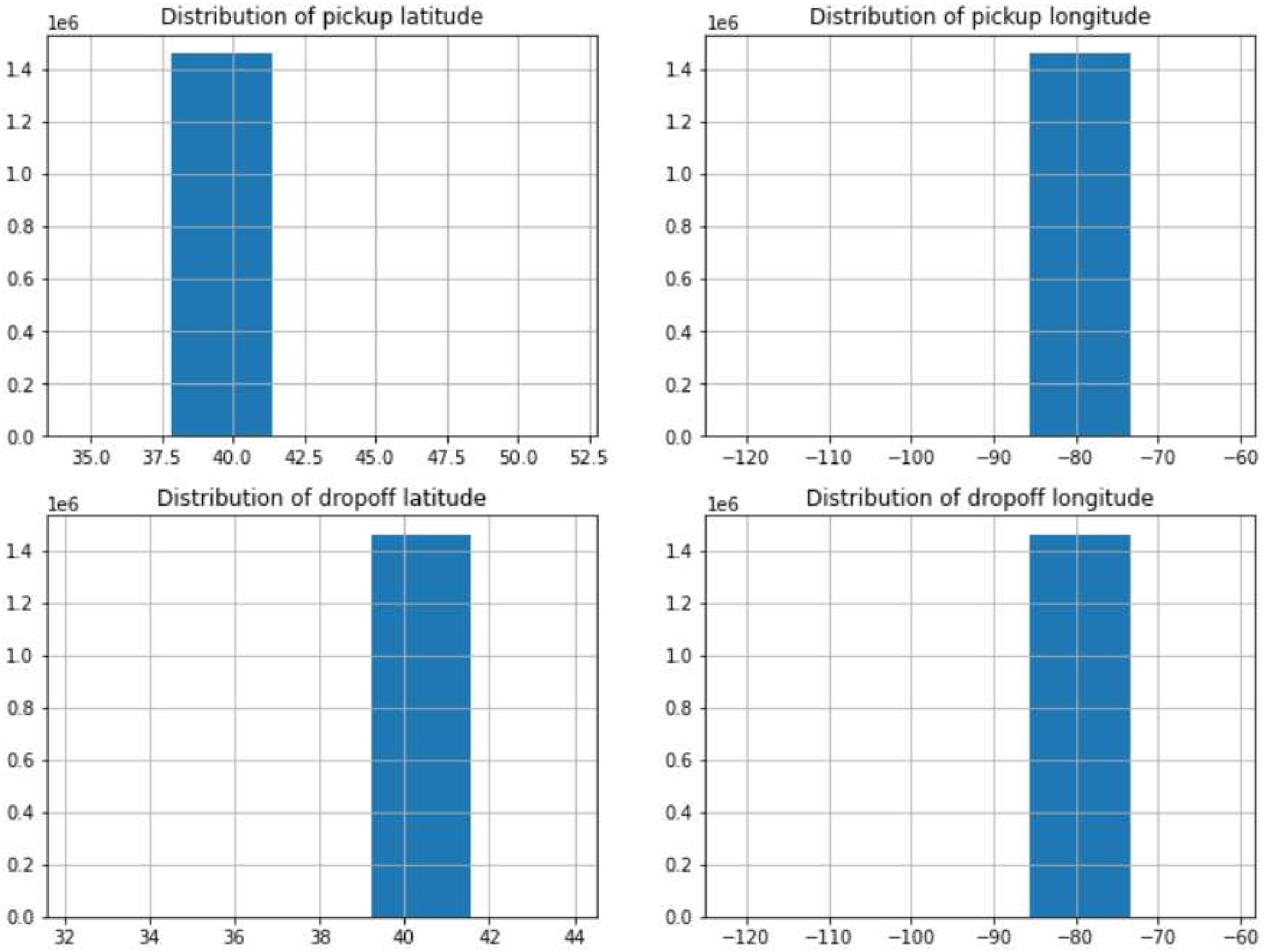


  - ◆
    - It can be seen from the figure that the day of the month will affect the taxi trip turation.
    - However,if we use one-hot enconding directily,it will cause the disaster of dimensionality.
    - In term of the trip turation,we divide it into three categories.
    - a:5,13 b:7,30 c:other
    - The same process for the hour.
    - x:0.1.2.3.4.8.9.10.19.20.21.23
      y:11.12.13.17.18.22
      z:14.15.16
      w:5,6,7

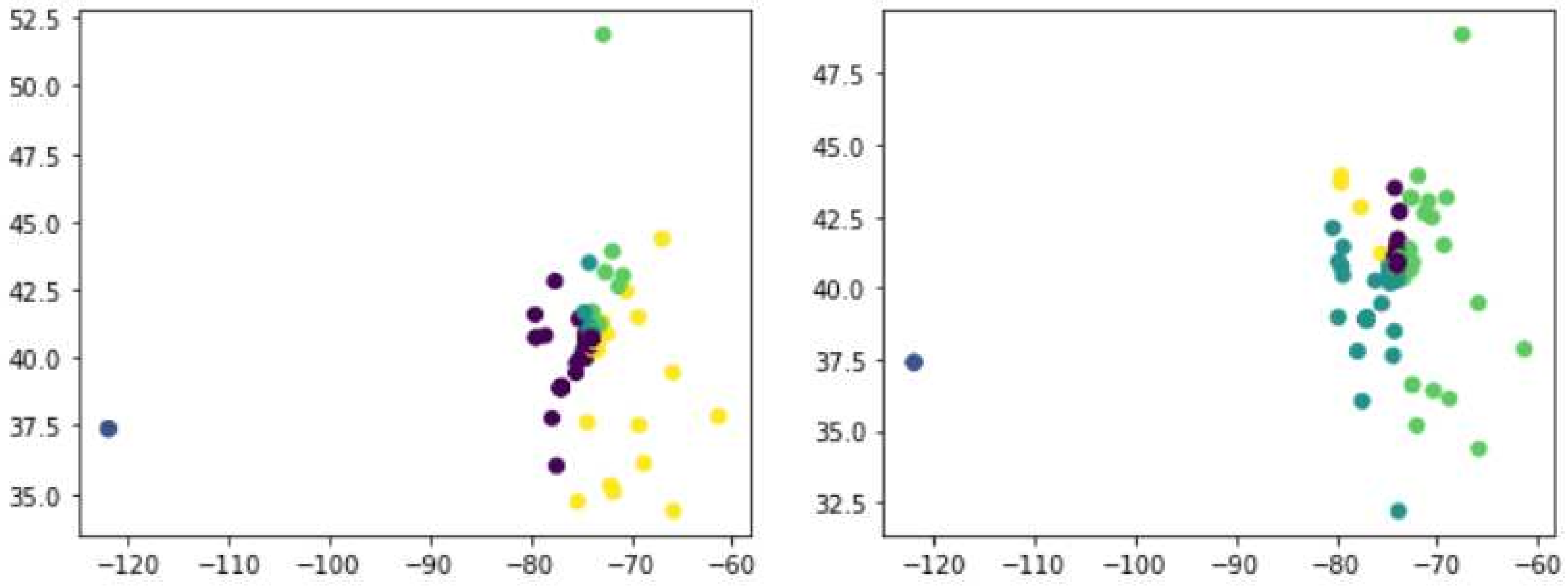
- pickup_datetime:
  - Derive weekday from the date.
  - 

  - Obviously,weekday affects the trip duration,so we add the attribute for the data.

- latitude, longitude:
  - duration = distance / speed
  -

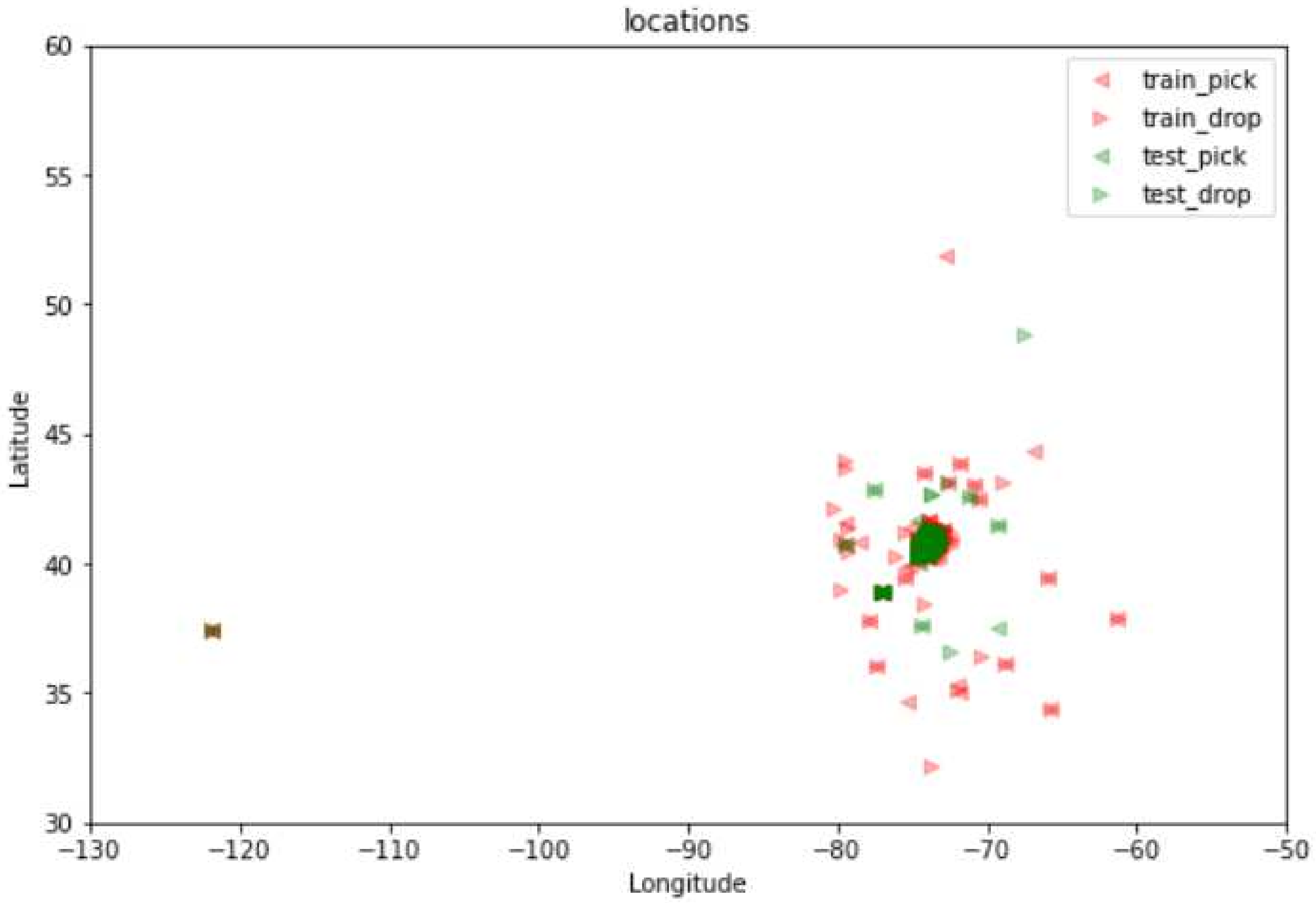


- distance:
  - Euclidean distance
  - Manhattan distance

- speed:
  - In term of pickup_latitude,pickup_longitude,dropoff_longitude,drop_latitude,divide data into 5 clusters by KMeans
  - 

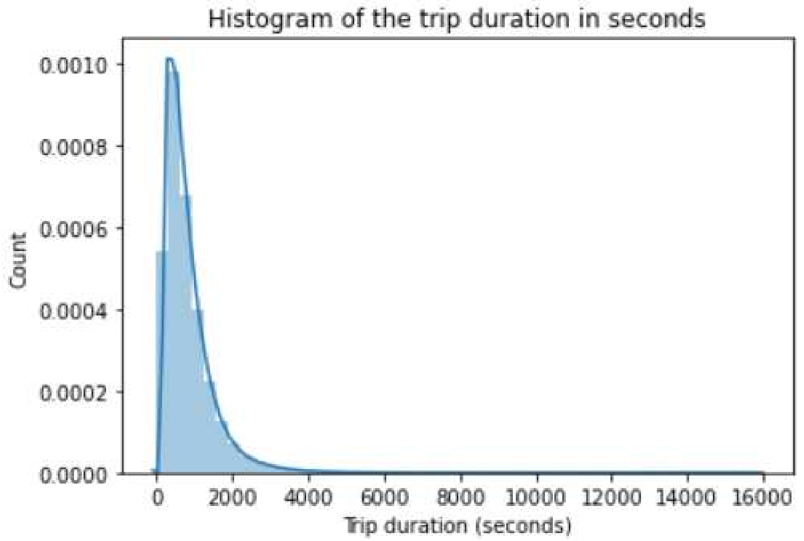
  - Direction

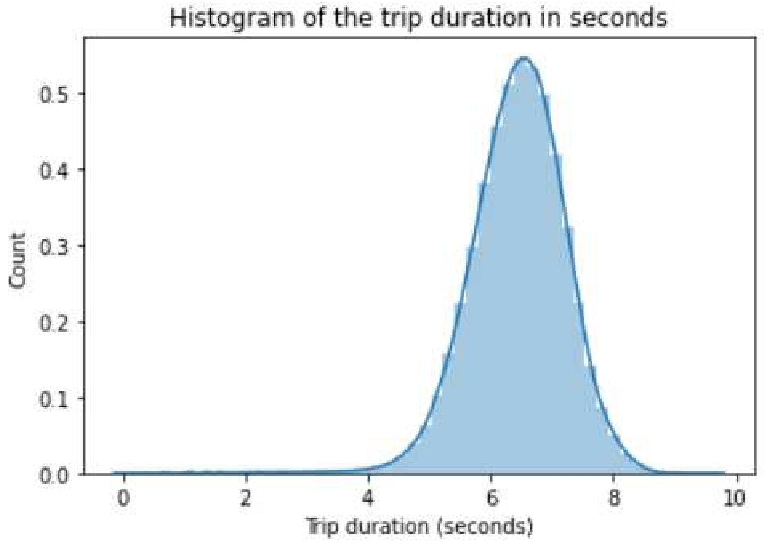
- outliers:
  -



  - From Google,we know New York City longitude vary from -74.03 to -73.75,and latitude vary from 40.63 to 40.85.
  - In term of this, we drop outliers

Labels

- trip_duration:
  - 

  - Through power conversion, it is more like Gaussian distribution.
  - 

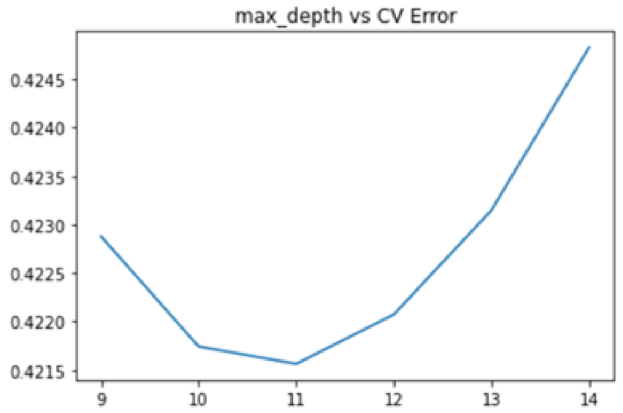
- outliers:
  - Trip duration varies from 1 second to 3526282 second,and there are some outliers.
  - We keep the trip duration between mean-3*std and mean+3*std.(approximately 99%)

Model selection

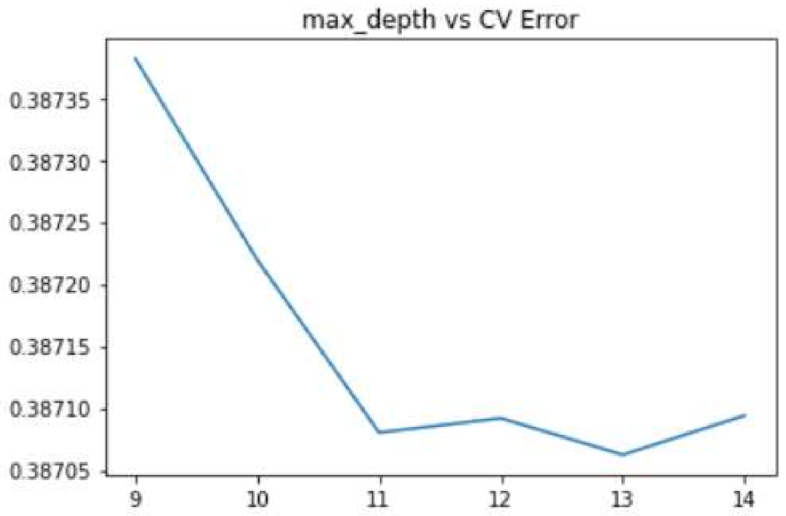
Models

- Models:
  - Ridge
  - Bagging
  - Boosting
  - RandomForest
  - Lightgbm
  - Xgboost

- ## Xgboost

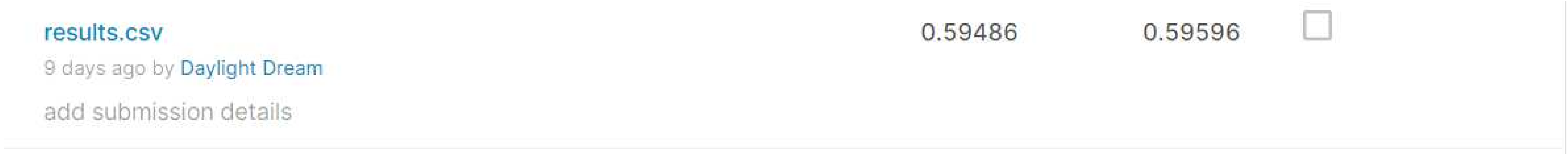


- ## Lightgbm



Results

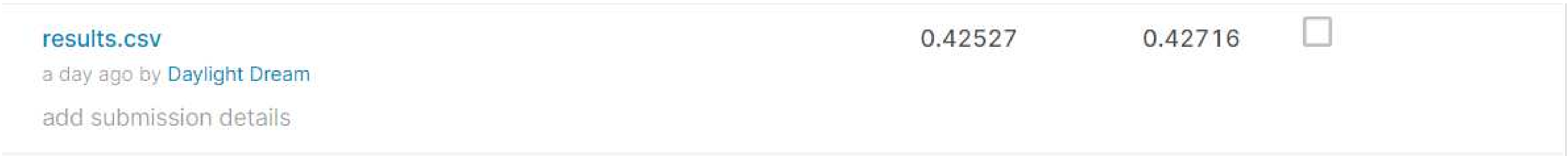
Results



results.csv 0.59486 0.59596 □

9 days ago by Daylight Dream

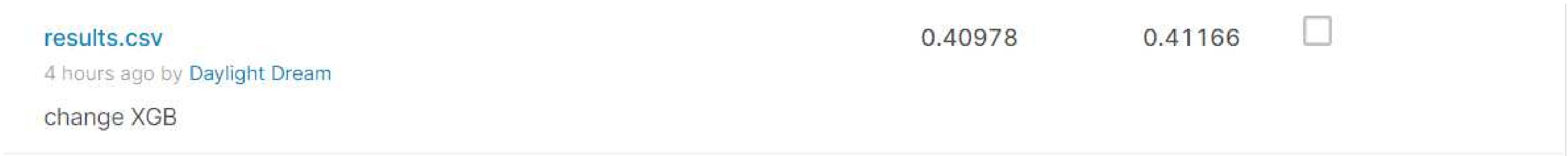
add submission details



results.csv 0.42527 0.42716 □

a day ago by Daylight Dream

add submission details



results.csv 0.40978 0.41166 □

4 hours ago by Daylight Dream

change XGB

Contact Information

Undergraduate Siyi Yu
School of Computer Science and Technology
Jilin University, China

✉ YUSIYICSAT@163.COM