Lin Che Yu Frank
2017年3月19日

Big Data Analytic HW1

0556555

Used Tools:

Since the Taxi data is quite huge (10 million data per month). Weka may not be a good solution for this since the I/O overhead is more time-costly than other naive language.

I choose **python** as my processing tool to run this analytic. Also, several useful libraries were imported to accelerate my calculation, such as **pandas, pyplot.**

Data Set: Yellow trip data 2016_10

Q1.What regions have most pickups and drop-offs?

At first, I wanted to use DBSCAN to cluster the pickup and dropoff, but the 10million dataset killed my memories many times. I choose to use the data after July, since the location changed into LocationID. Then just need to calculate the occurrence for every IDs, we could get the answer for this question.

Most Pickups: ID 237 Manhattan Upper East Side South

Most Dropoffs: ID 236 Manhattan Upper East Side North

Q2.When are the peak hours and off-peak hours in taking taxi?

The same way as q1. Calculate the occurrence for every single hour. I use pandas date time format extraction to extract the day-hour for specific analytic.

Peak hour: 19 p.m.

Off-Peak hour: 5 a.m.

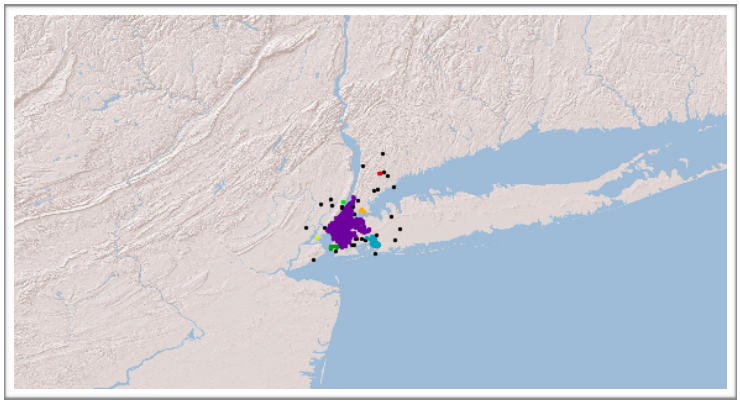
Q3.What differences exist between short and long distance trips of taking taxi?

I took the average of trip_distance as a milestone to distinguish between short and long distance trip.

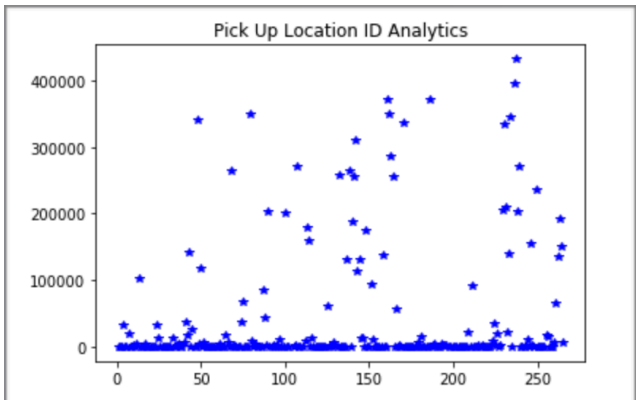
I noticed that the **average tip_amount** performs better in long trip distance. And people in long trip distance tends more to **pay in credit card.** (71% vs 65%)

Note

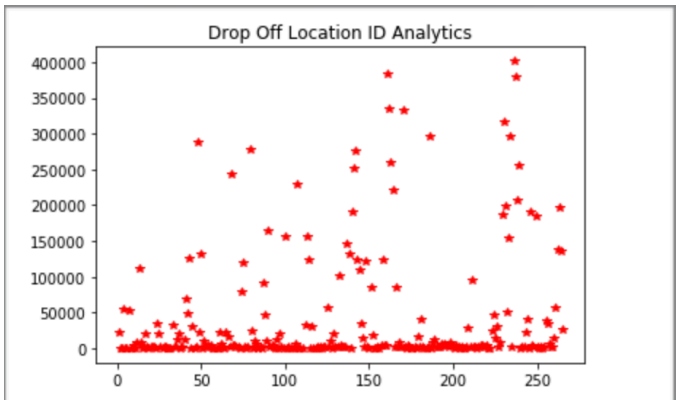
At first, I tended to use mpl_toolkit with baseman to draw the cluster on the Geography map. But the memory issue to use DBSCAN in sklearn stopped me, I can only show the 60000 points result in the end.



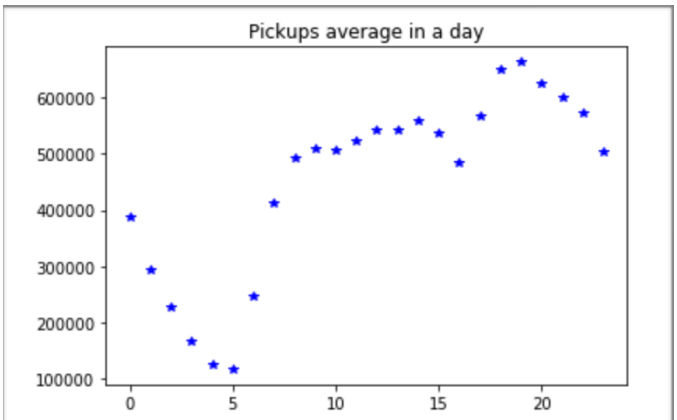
Figures.



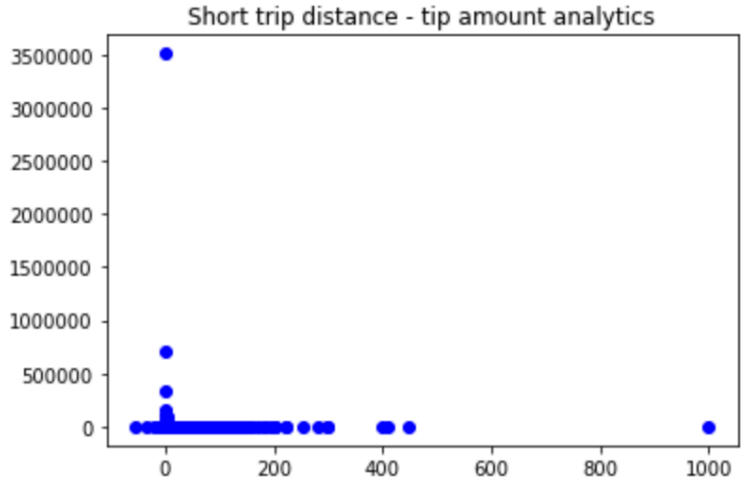
237	432244
236	396235
161	372269
186	371034
162	350427
79	350290
234	344762
48	340972
170	336389
230	335293
142	310957
163	286326
239	271255
107	269996



236	401427
161	383919
237	380132
162	334607
170	332940
230	316679
186	296656
234	295862
48	287688
79	277972
142	276897
163	260136
239	254972
141	251032

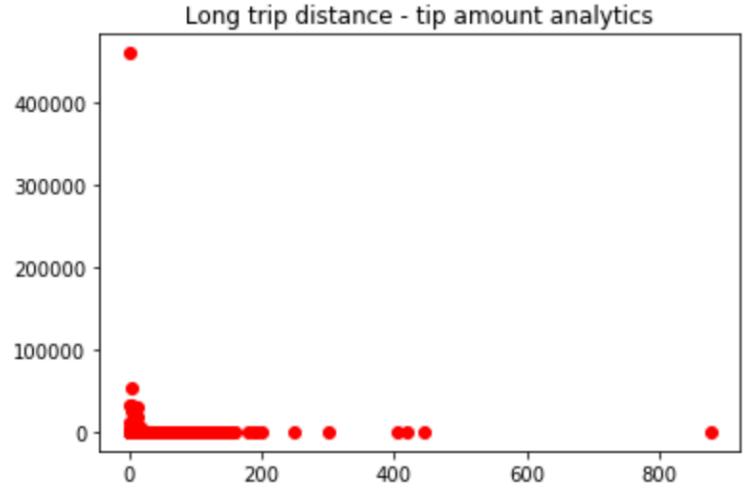


19	662858
18	651059
20	625389
21	598790
22	571744
17	567508
14	559250
13	542061
12	541814
15	536111
11	522197
9	508474
10	505199
23	503043
8	491766
16	483548
7	411946




```
The average of tips in short dist travel is 1.3513399144117206
The payment type counting is

1    6114831
2    3233286
3      33122
4      11652
Name: payment_type, dtype: int64
```




```
The average of tips in long dist travel is 4.920571037685965
The payment type counting is

1    989678
2    396069
3      7224
4      2400
5         1
Name: payment_type, dtype: int64
```