Big Data Analytic Techniques and Applications
Homework1 report

310551158 林鑫伯

development environment

Operating System: Ubuntu 20.04 LTS
Memory Size: 16 GiB
CPU: Intel core i7-6700
tools: pandas

Questions

1. What regions have the most pickups? What are the top-5 regions with the most pickups and drop-offs (pickups and drop-offs should be counted separately)?

Region definition: I concatenated the longitude and latitude info as the an identical location, due to wired values read from shapefile.

As shown in the following pictures, the region with most pickups is (W73.97°, N40.76°).
Top 5 regions with most pickups and drop-offs are also shown in the following pictures.

```
# count passenger count
     start_location  Passenger_Count
5269    -73.9740.76          4382779
5352    -73.9840.76          4021265
5428    -73.9940.75          3885514
5351    -73.9840.75          3441813
5429    -73.9940.76          3273840
...             ...              ...
6522    -74.1341.17                1
6523    -74.1341.19                1
6525    -74.1341.24                1
6526    -74.1341.28                1
9590     -75.041.36                1
```

Pic.1. top 5 regions with most pickups

```
# count passenger count
     End_location  Passenger_Count
5323  -73.9840.76          3973530
5251  -73.9740.76          3871929
5404  -73.9940.75          3621350
5322  -73.9840.75          3368477
5405  -73.9940.76          2710925
...           ...              ...
7276   -74.240.25                1
7277   -74.240.27                1
7278   -74.240.29                1
1760  -73.4940.88                1
9563   -75.041.12                1
```

Pic.2. top 5 regions with most drop-offs

2. When are the peak hours and off-peak hours for taking a taxi?
   - hint: You can count the number of pickups in different hours of day.

As we can see in the following picture, 19pm-22pm are the peak hours and 3pm-6pm are the off-peak hours.

```
# count peak hour
    Hour   counts
19    19  4693107
18    18  4487196
20    20  4249192
21    21  4105992
22    22  4046984
17    17  3775191
23    23  3553683
15    15  3472391
14    14  3384601
16    16  3233192
13    13  3227037
12    12  3209858
8      8  2984342
9      9  2979752
11    11  2902430
0      0  2881474
10    10  2785061
7      7  2233217
1      1  2111596
2      2  1583416
3      3  1196927
6      6  1196072
4      4   843948
5      5   594417
```

Pic.3. time-passenger_count table

3. What are the differences between big and small total amounts when taking a taxi?
   ■ hint: First, you should define what big and small total amounts are. And then, you should point out the difference between them. You should at least observe the results of Q1 and Q2.

The boundary of big and small total amount is defined as if it is larger than 60. If total amount is larger or equals to 60, noted as 1. Otherwise, noted as 0.

Time-Amount relation

As shown in the following picture, big total amount events tend to happen in 14pm-17pm, which are not peak hours according to my answer to question no. 2. However, more small amount events happen in peak hours. That's the difference I found in time-amount relation.

```
# count total amount
    Total_Amt  Hour  hour_count
24        1.0     0        2980
36        1.0    12        4072
27        1.0     3        1767
28        1.0     4        2509
29        1.0     5        3932
30        1.0     6        4332
31        1.0     7        3920
32        1.0     8        3255
33        1.0     9        3111
34        1.0    10        3092
35        1.0    11        3500
37        1.0    13        4910
25        1.0     1        2190
38        1.0    14        6863
39        1.0    15        7945
40        1.0    16        7067
41        1.0    17        5455
42        1.0    18        4508
43        1.0    19        3534
44        1.0    20        3422
45        1.0    21        3729
46        1.0    22        3897
26        1.0     2        1696
47        1.0    23        3609
```

```
1         0.0     1     1170478
23        0.0    23     1973060
2         0.0     2      878423
3         0.0     3      662594
4         0.0     4      478317
5         0.0     5      369437
6         0.0     6      786767
7         0.0     7     1449921
8         0.0     8     1908789
9         0.0     9     1904516
10        0.0    10     1737225
11        0.0    11     1776085
12        0.0    12     1938270
13        0.0    13     1929249
14        0.0    14     2020146
15        0.0    15     2045644
16        0.0    16     1892677
17        0.0    17     2212425
18        0.0    18     2642456
19        0.0    19     2727155
20        0.0    20     2456016
21        0.0    21     2334167
22        0.0    22     2263769
0         0.0     0     1593628
```

Pic. 4. time-amount relation table