Big Data Analytics Techniques and Applications

HW1_0556562_陳鴻君

Data Source

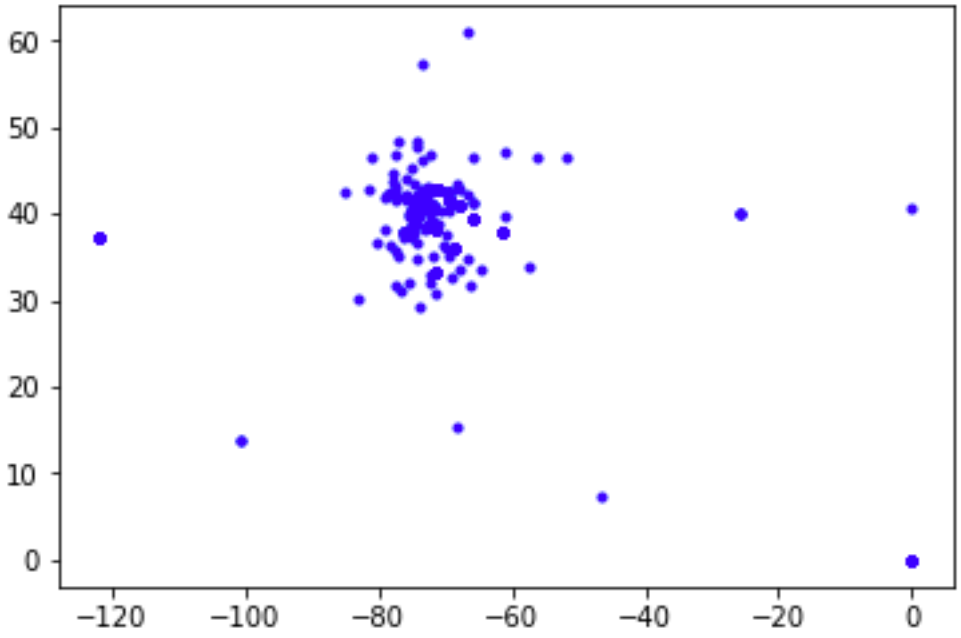
In this project, the NYC yellow taxi data was used. On the website of NYC government, they provided years of taxi datas separated monthly. I used the yellow taxi data in January and August 2016 in New York City for this project.

Tools

Originally, I decided to use Weka (e.g., It's a useful machine learning tool.) as my tool for analyzing big data in this project. But I figured out that the document of NYC yellow taxi data was too large to import. It caused IO overhead for Weka. So, I used iPythonNotebook in Python 3 for this project.

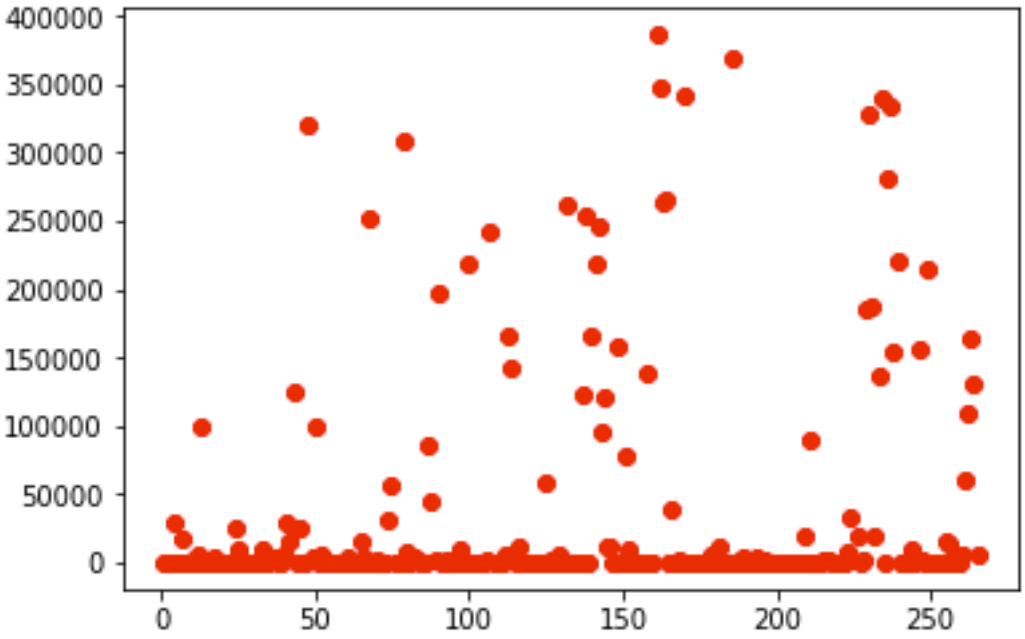
Q1: What regions have most pickups and drop-off?

I found out that there were two types of taxi datas on NYC gov website. The main difference is at position expression. One was in latitude and longitude (e.g., January) ; the other was in location ID (e.g., August). I tried to plot all pickups latitude and longitude in January on map (shown in below).

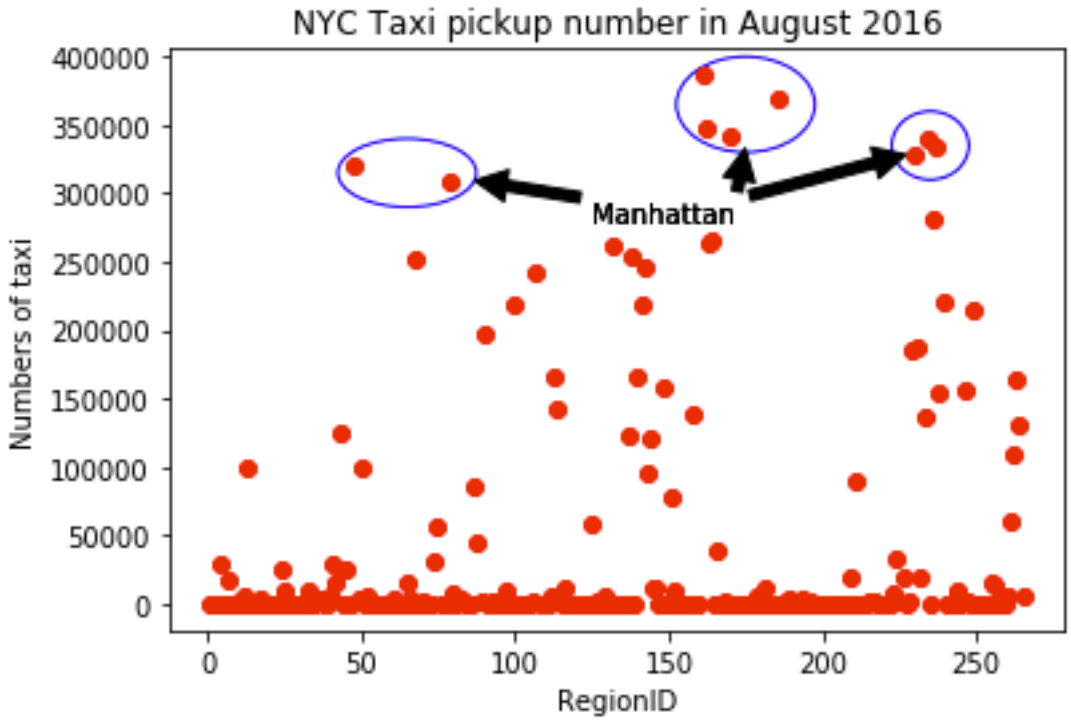


But it was useless, because I didn't know what region it was, even I used cluster to separate group by group.

So I used August's data, such expressed as location ID. I counted all taxi pickups by region, and created figure as below.



The x-axis was location ID, and the y-axis was numbers of taxi datas. Then I came out that the regions of high-number taxis are all Manhattan region.



Next step, I wondered the regions corresponded their location ID. Therefore I looked up “taxi+_zone_lookup.csv”, which restore location ID and corresponded region. I summarized all the region that yellow taxi can serve. And also I statistically summed up the location ID belonged to a region.

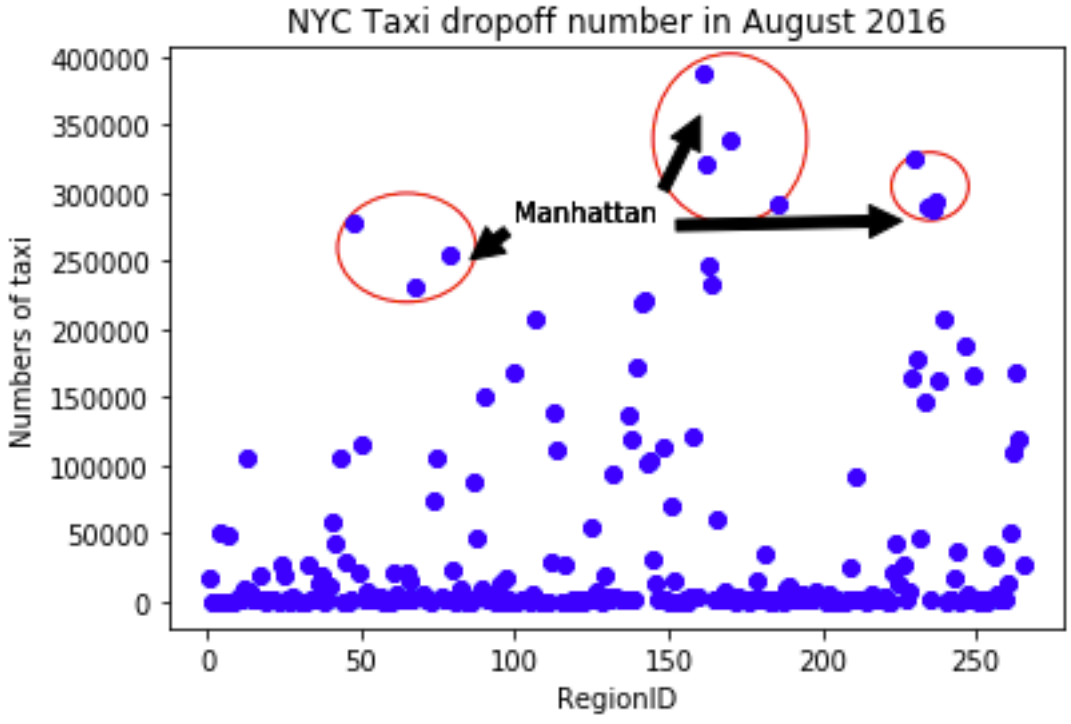
```
Borough
Bronx             43
Brooklyn          61
EWR                1
Manhattan         69
Queens            69
Staten Island     20
Unknown            2
Name: LocationID, dtype: int64
```

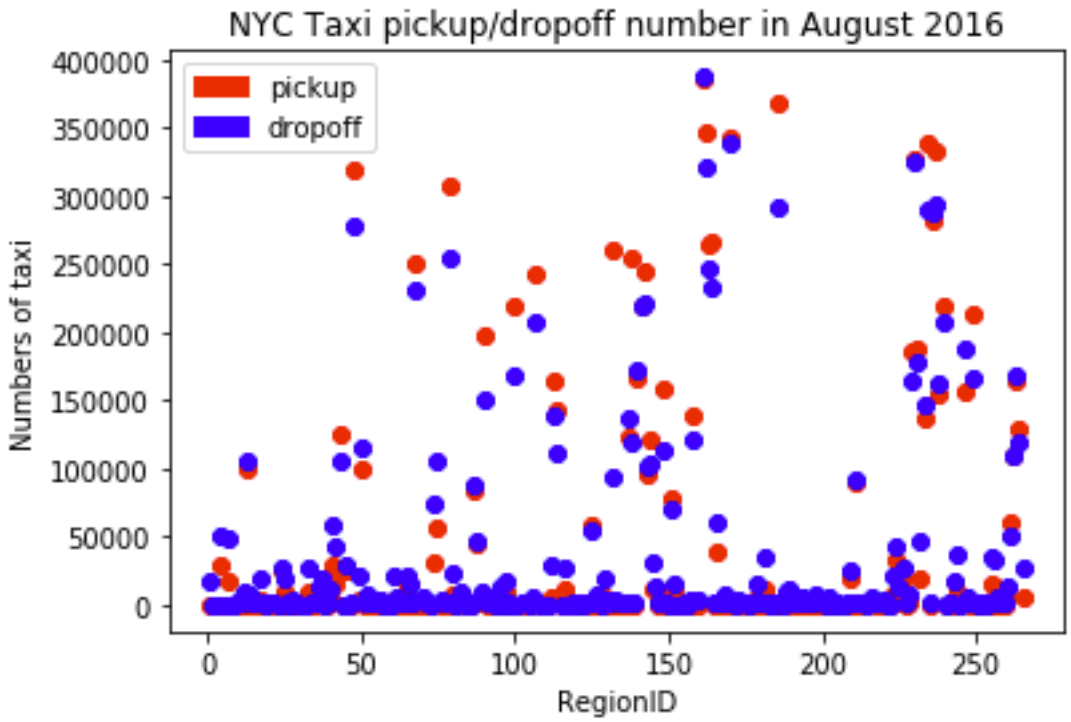
By the way, I can merge location ID with regions. Below was the result of what regions had the most pickups.

```
Counter({'Bronx': 8180,
         'Brooklyn': 156022,
         'EWR': 747,
         'Manhattan': 9010878,
         'Queens': 629670,
         'Staten Island': 633,
         'Unknown': 136133})
```

Observably, Manhattan region had the most pickups among all region that yellow taxi can serve.

I did the same steps to drop-off. There were some difference between pickups and drop-off. But the result was the same the Manhattan region had the most drop-off.





As the result of drop-off numbers, Manhattan had the most drop-off than other regions in NYC.

```
Counter({'Bronx': 65624,
         'Brooklyn': 508142,
         'EWR': 18173,
         'Manhattan': 8662716,
         'Queens': 538287,
         'Staten Island': 2758,
         'Unknown': 146563})
```

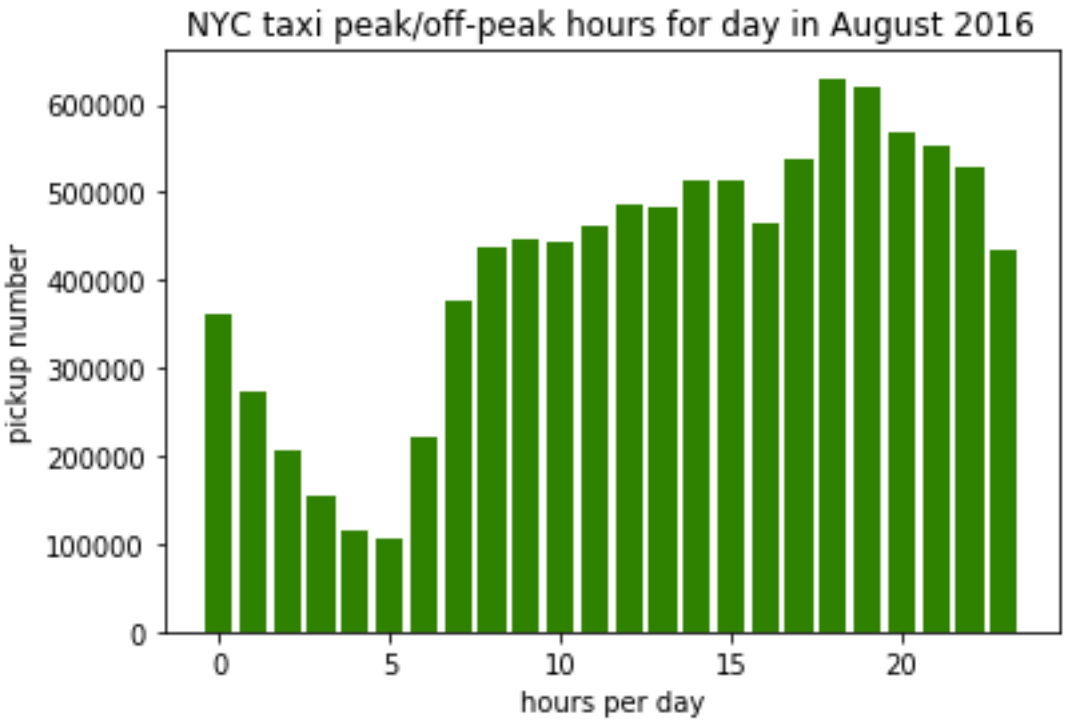
Summary: either pickup or drop-off, Manhattan had the most yellow taxi service.

Q2: When are the peak hours and off-peak hours of taking taxi?

By counting pickups in 24 hours everyday, we can earn pickup ratio aka peak hour and vice versa. Following this theory, I did statistics as below.

```
18    629963
19    621356
20    568738
21    554244
17    538746
22    527841
15    513577
14    512740
12    484830
13    483598
16    465822
11    461285
9     447070
10    443958
8     438189
23    435347
7     377731
0     360555
1     272368
6     222803
2     207505
3     153648
4     115593
5     104756
Name: hour, dtype: int64
```

As result, 18 o'clock was the peak hour, and 5 o'clock was the off-peak hour of days in August. I made this information into a bar-chart.



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

According to this question, I defined the mean of trip distance as the threshold of distance. Taxi trip who was far from mean was seen to be long trip and vice versa.

Why I used mean as threshold? I thought the special case in big data analysis wouldn't make any influence. The noise would be divide by others which was large number. So it didn't matter.

In this case, the mean of trip distance in August 2016 was 3.9484142775140905 km.

After calculating, there were 7998710 short distance trips and 1943553 long distance trips. So short distance trips were 80.45% and long distance trips were 19.55%.

About the payment type, there were 6 types: Credit card, Cash, No charge, Dispute, Unknown, Voided trip.

There were 68.12% paid by credit card and 31.24% paid by cash in long distance trips. On the other hand, there were 63.24% paid by credit card and 36.20% paid by cash in short distance trips. So, in this case, more passengers who took long distance trips preferred pay by credit card than who took short distance trips.

About the fare, it was $29.89 in average when taking long distance trips, but $9.21 when taking short distance trips. It showed that fare of long distance trips cost more than fare of short distance trips.

About the tip, it was $3.90 in average for long distance trips, but $1.22 for short distance trips. It showed that costumers would give more tip when the taxi driver had a long drive.