Big Data Analytics Techniques and Applications

Homework 1

Name: 陳以瑄 Student ID:109705001

1. I think I am lucky that I haven't encounter any serious "Big Data" issue in analyzing the NYC dataset. The only problem I face is that, it is pretty time consuming when I try to read parquet files or to modify the value in columns.

2. Descriptions of the scale of data, analytical tools, and spec of the platform you use.
   - Scale of data :

     43168922 rows × 20 columns in total
   - Analytical tools:

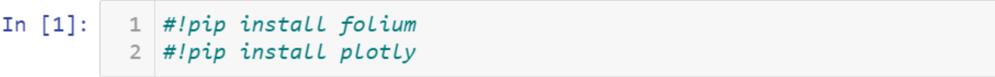
     python pandas
   - Visualize tools:

     Folium, for drawing map

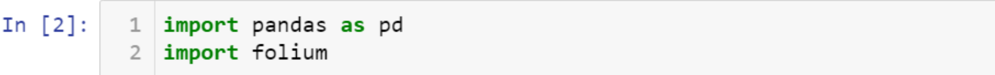
     Plotly, for bar chart and pie chart

3. Manipulation steps of data analysis tools

   Since I use Jupyter notebook, I guess I don't need to show steps in detail. But it is important to run these two cell.



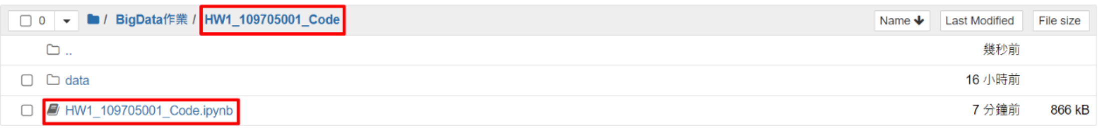
```
In [1]:  #!pip install folium
          #!pip install plotly
```



```
In [2]:  import pandas as pd
          import folium
```

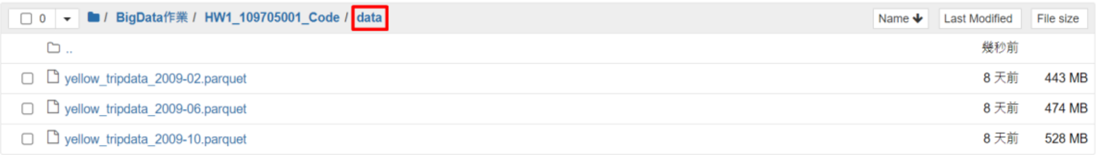
By installing folium and plotly, you can successfully import the modules I use in this homework.

And make sure you put the code outside the folder call "data", and save the parquet files in the "data" folder.



/ BigData作業 / HW1_109705001_Code

..	幾秒前	
□ data	16 小時前	
□ HW1_109705001_Code.ipynb	7 分鐘前	866 kB



/ BigData作業 / HW1_109705001_Code / data

..	幾秒前	
□ yellow_tripdata_2009-02.parquet	8 天前	443 MB
□ yellow_tripdata_2009-06.parquet	8 天前	474 MB
□ yellow_tripdata_2009-10.parquet	8 天前	528 MB

4. Descriptions of how you solve each question in detail.

Q1. 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)?

- Def of the region:

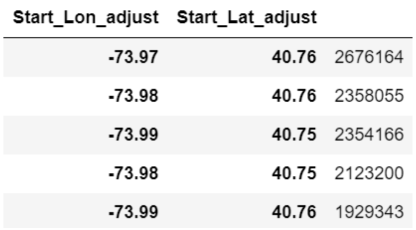
  I divide the area into squares, the width is 0.01 degree longitude and the height is 0.01 latitude.

- How I solve

  1. Round the value in columns “Start_Lon”, “Start_Lat”, “End_Lon” and “End_Lat” to 2 decimals by using round() function in pandas. Save the rounded value in columns “Start_Lon_adjust”, “Start_Lat_adjust”, “End_Lon_adjust” and “End_Lat_adjust”
  2. Using groupby() function in pandas to group the rows by “Start_Lon_adjust”, “Start_Lat_adjust” to count the pickups in each region. And use “End_Lon_adjust”, “End _Lat_adjust” to count the drop-offs in each region
  3. Sort the value to get the top 5 regions.

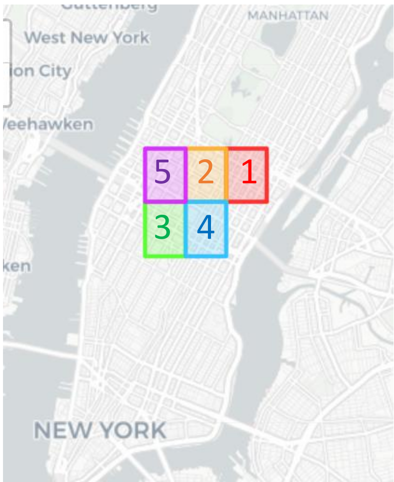
- Result

  1. For pickups, I got the longitude and latitude of the center in these 5 regions (Table 1). Then I visualize the result on map with folium (Figure 1), and find out that all of them are in Manhattan.



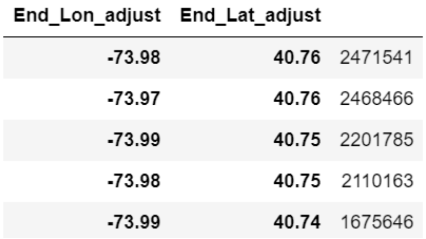
Start_Lon_adjust	Start_Lat_adjust	
-73.97	40.76	2676164
-73.98	40.76	2358055
-73.99	40.75	2354166
-73.98	40.75	2123200
-73.99	40.76	1929343

▲ Table 1. Coordinate of most pickups of all data



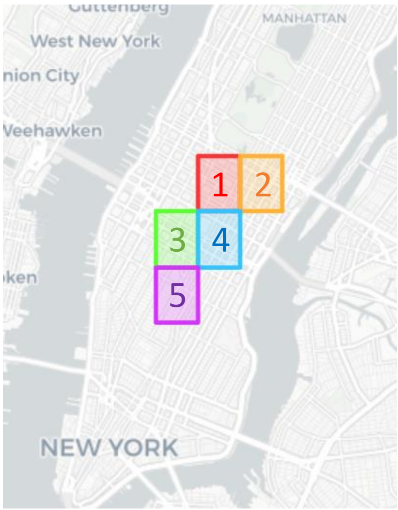
▲ Figure 1. The most pickup regions of all data on map

2. For drop-offs, I got the longitude and latitude of the center in these 5 regions (Table 2). Then I visualize the result on map(Figure 2), and find out that although the coordinates are somehow different from the most pickups, they still are all in Manhattan.



End_Lon_adjust	End_Lat_adjust	
-73.98	40.76	2471541
-73.97	40.76	2468466
-73.99	40.75	2201785
-73.98	40.75	2110163
-73.99	40.74	1675646

▲ Table 2. Coordinate of most drop-offs of all data



▲ Figure 2. The most drop-off regions of all data on map

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

- Def of the peak hours and off-peak hours:

  Peak hours: pickup times is greater than 0.8 quantile

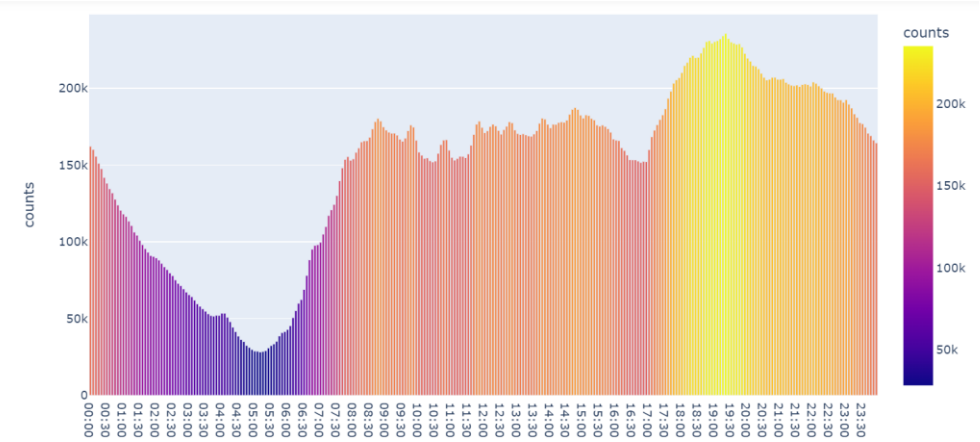
  Off-peak hours: pickup times is less than 0.2 quantile

- How I solve:

  1. Extract hour and minute from datetime that recorded in column “'Trip_Pickup_DateTime'”. And round time to the nearest 5 minute, then save the rounded value in column “Time_adjust”.
  2. Group the rows by “Time_adjust” to count the total number of pickups of each five minutes during the three months.
  3. Count the 0.8 quantile to set the bound of peak, and count the 0.2 quantile to set the bound of off-peak.
  4. Filter time that the pickup numbers are higher than the bound of peak, record the earliest as peak-start, and the last one as peak-end. Then filter time that the pickup numbers are less than the bound of off-peak, record the earliest as off-peak -start, and the last one as off-peak -end.

- Result:

As the Figure 3 shows, the busiest time is about 19:30, while the least busy time is near 5:00.

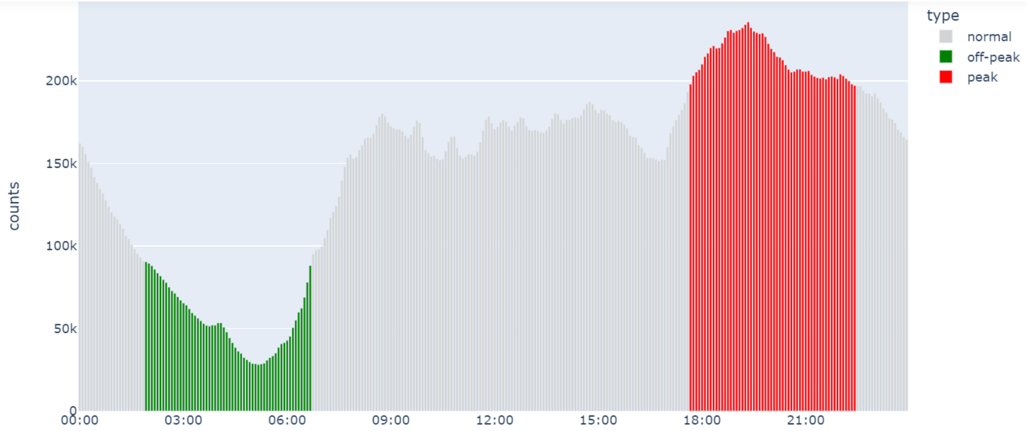


▲ Figure 3. Pickups in each five minutes for all data

I also divide the time into peak and off-peak by the definition I mention above.

1. Peak hours, pickups more than 0.8 quantile, is 17:40 to 22:25.
2. Off-peak hours, pickup less than 0.2 quantile, is 1:55 to 6:40.

As the Figure 4 shows, the red area is peak hours, while the green region is off-peak time.

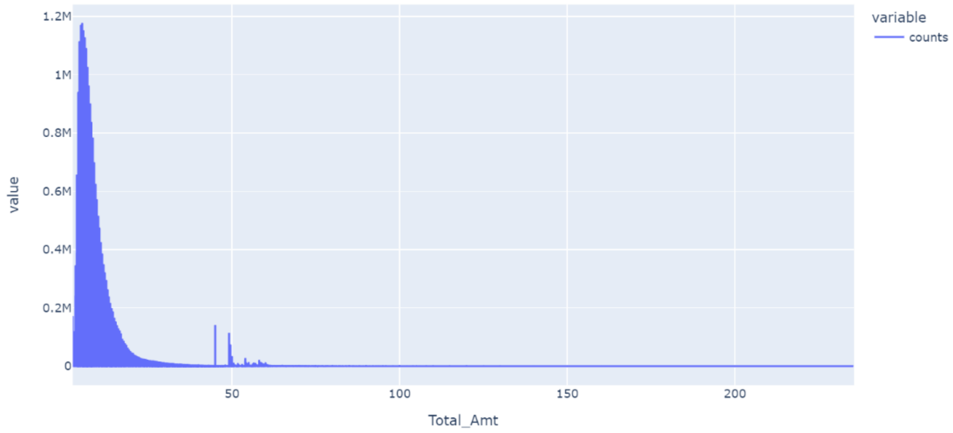


▲ Figure 4. Peak hours and off-peak hours for all data

Q3: What are the differences between big and small total amounts when taking a taxi?

- Def of big and small total amounts:

  As the Figure 5 shows, the distribution of total amounts is skew-right.



▲ Figure 5. Distribution of total amounts

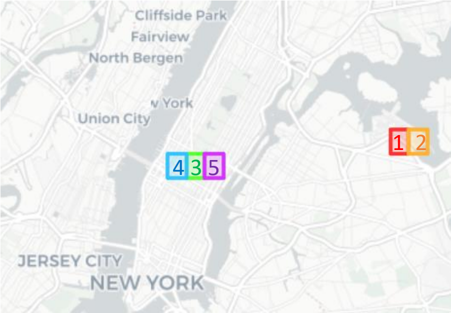
  So I sort the data by total amounts and choose the highest 10% data as big amounts and the lowest 10% data as small amounts.

- How I solve:
  1. I sort the data by the value of "Total_Amt" in ascending order.
  2. Since it has 43168922 rows in total, I set the value of "Total_Amt" in the 4316892th row, $4.6, as the boundary of small total amounts. And set the value of "Total_Amt" in the 38852030th row, $18.59, as the boundary of big total amounts.
  3. Extract the rows that "Total_Amt" is less than $4.6 as small total amounts' data, and save as new dataframe "Small_Total_Amounts" ; select the rows that "Total_Amt" is larger than $18.59 as big total amounts' data, and save as "Big_Total_Amounts".
  4. For the most pickups and drop-offs, I use the same steps in Q1.
  5. For the peak hours and off-peak hours, I use almost the same steps in Q2. However, since the peak time in this question is not continued, I list all the hours rather than showing the start time and end time.
  6. For the payment types analysis:
     i. Since the origin categories in column "Payment_Type" contains "CASH" and "Cash", I modified the value "CASH" to "Cash", so that I can merge these two as the same group. I also do the same thing on "CREDIT" and "Credit".
     ii. Group the data by "Payment_Type" and count the percentage.

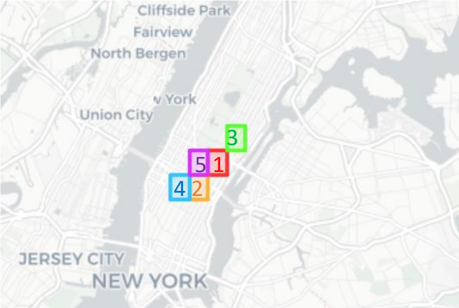
- Result:

1. The most pickups:

   Figure 6 shows the most pickups of big total amounts, and Figure 7 shows the top 5 pickups of small total amounts. The difference is that, for big total amounts, the top 2 regions are in Queens and the rest of all are in Manhattan; while all of the 5 regions of small total amounts are in Manhattan.



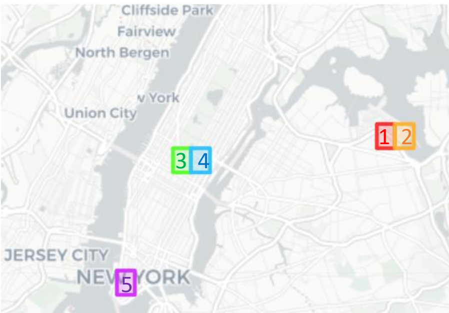
▲ Figure 6. The most pickup regions of big total amounts data on map



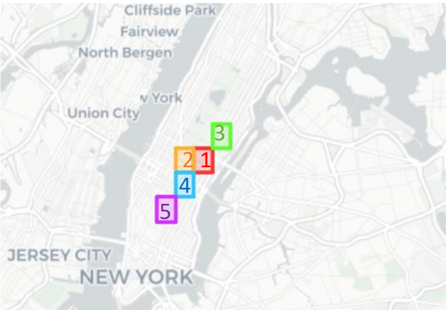
▲ Figure 7. The most pickup regions of small total amounts data on map

2. The most drop-off:

   Figure 8 on the left shows the most pickups of big total amounts, and Figure 9 shows the top 5 pickups of small total amounts. The difference is that, for big total amounts the regions are sparse, 2 of them are in Queens, 2 of them are in Middle Manhattan, 1 of them is in South Manhattan; while all of the 5 regions of small total amounts are in Middle Manhattan.



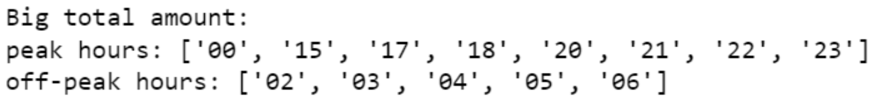
▲ Figure 8. The most drop-off regions of big total amounts data on map



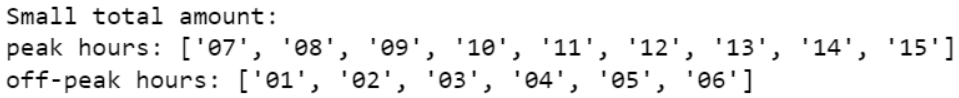
▲ Figure 9. The most drop-off regions of small total amounts data on map

3. The peak and off-peak:

The result shows below.



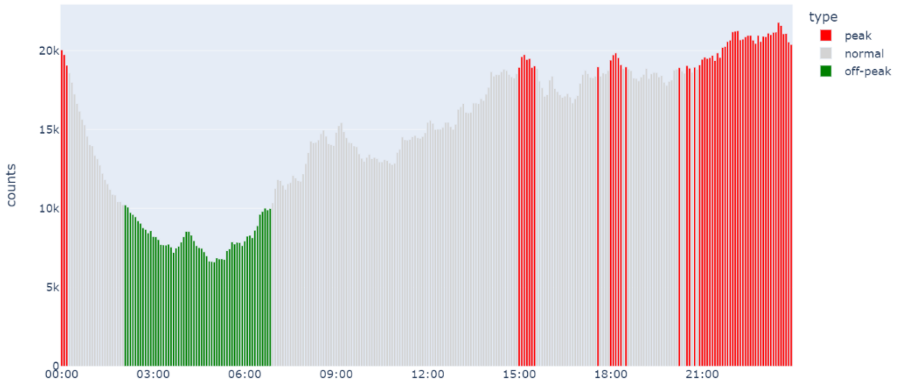
```
Big total amount:
peak hours: ['00', '15', '17', '18', '20', '21', '22', '23']
off-peak hours: ['02', '03', '04', '05', '06']
```



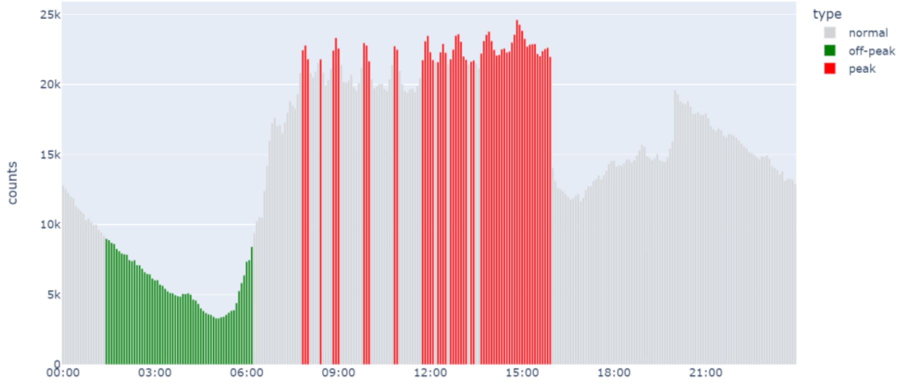
```
Small total amount:
peak hours: ['07', '08', '09', '10', '11', '12', '13', '14', '15']
off-peak hours: ['01', '02', '03', '04', '05', '06']
```

There is a significant different in peak hours, as for big total amounts, the peak hours is mainly at night; while for small total amounts, the peak hours is in morning and noon. However, the off-peak hours seems the same, both contains 2 a.m. to 6 a.m.

Figure 10 shows the peak hours and off peak hours of big total amounts; Figure 11 shows the peak hours and off peak hours of small total amounts.



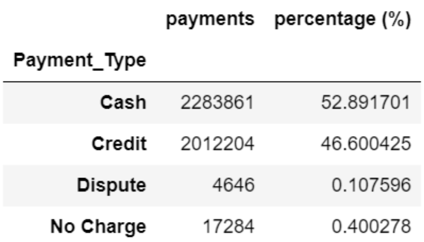
▲ Figure 10. Peak hours/ Off-peak hours for big total amounts data



▲ Figure 11. Peak hours/ Off-peak hours for small total amounts data

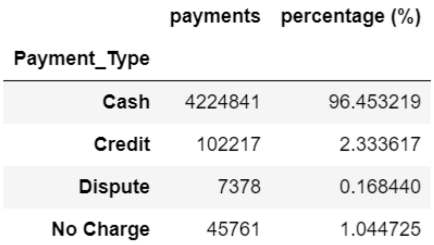
4. The payment types:

GTable3 and Figure 12 show the payment type proportion in big total amounts, Table4 and Figure 13 show the result of small total amounts. For the big total amounts, people tend to pay either in cash or credit; however, for the small total amounts, most people choose to pay in cash.



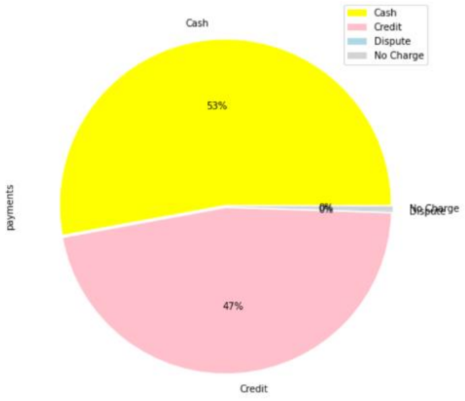
	payments	percentage (%)
Payment_Type		
Cash	2283861	52.891701
Credit	2012204	46.600425
Dispute	4646	0.107596
No Charge	17284	0.400278

▲ Table 3. The payment types proportion for big total amounts

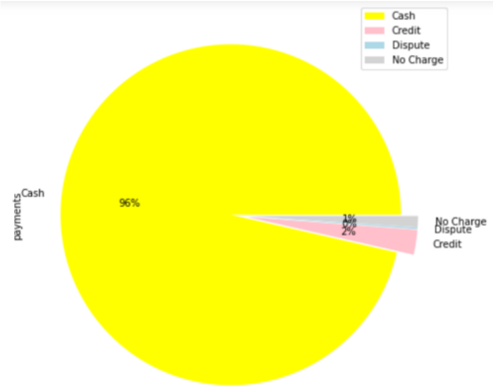


	payments	percentage (%)
Payment_Type		
Cash	4224841	96.453219
Credit	102217	2.333617
Dispute	7378	0.168440
No Charge	45761	1.044725

▲ Table 4. The payment types proportion for small total amounts



▲ Figure 12. The pie chart of payment types proportion for big total amounts



▲ Figure 13. The pie chart of payment types proportion for small total amounts