Big Data Analytics Techniques and Applications

Homework 1

Analyzing NYC Taxi Data

309709022 陳政廷

- Description

  1. 原始資料大小：41,859,906 筆(約 7.2 GB)

  2. 分析工具：Python, Spark, Matplotlib

  3. 分析平台：Jupyter Notebook

- Manipulation Step

  1. 安裝好 Spark 環境並載入相關套件

  2. 讀入後將三個月的資料進行合併

  3. 將資料當中相對不合理的數值(e.g.經度應大於 0)與相同的字串(e.g.CASH = Cash)進行處理，經處理後資料剩下 41,250,115 筆

  4. 接下來便依序根據問題所述進行相關操作與計算

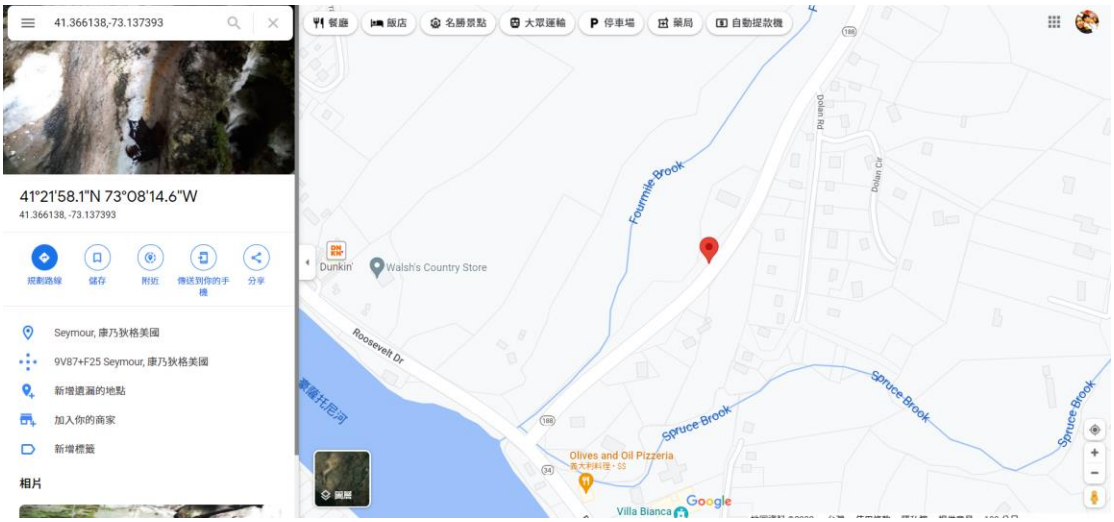
- Q1

  1. 操作：由於經度與緯度結合在一起方才是一個地標，故我們在此同時 groupby('Start_Lon','Start_Lat')並計算車次數量(多少 index)，經由大到小排序後變得出此題答案。

  2. 回答：

Start_Lon	Start_Lat	count
-73.137393000000003	41.366137999999999	54541
-73.951818000000003	40.733696999999999	5197
-73.988536999999994	40.698458000000002	2286
-73.989039000000005	40.75808	1336
-73.945795000000004	40.778751999999997	1331
-73.989053999999996	40.758087000000003	1136
-73.937512999999996	40.758152000000003	1042
-73.952003000000005	40.733759999999997	859
-73.991084999999998	40.733310000000003	827
-73.989047999999997	40.758083999999997	774

此表即為最常搭車之地點排名，我們可利用 google map 將經緯度輸入後得出當地地標，如下圖所示即是美國最常搭計程車之位置。



下車地點操作方式同上，下表為前 10 名 。

```
+------------------+-----------------+-----+
|End_Lon           |End_Lat          |count|
+------------------+-----------------+-----+
|-73.137393000000003|41.366137999999999|45724|
|-73.951818000000003|40.733696999999999|5197 |
|-73.988536999999994|40.698458000000002|2286 |
|-73.989039000000005|40.75808          |1336 |
|-73.945795000000004|40.778751999999997|1331 |
|-73.989053999999996|40.758087000000003|1136 |
|-73.937512999999996|40.758152000000003|1044 |
|-73.952003000000005|40.733759999999997|859  |
|-73.991084999999998|40.733310000000003|828  |
|-73.989047999999997|40.758083999999997|774  |
+------------------+-----------------+-----+
only showing top 10 rows
```

從中可發現最常搭車地點與最常下車地點相同，推測為人口稠密且為交通樞紐之位置。

- Q2

1. 操作：巔峰搭車與離峰搭車時間可透過擷取'Trip_Pickup_DateTime'的時間資訊，便可得出 3 個月來每天 24 小時之搭車數量，並同樣透過 groupby 與 sort 得出結果

2. 回答：從下表可發現，人群主要搭車的時間位於晚上 7 點到 10 點最為密集，而半夜 4、5 點為休憩時間，故較少人搭乘

```
+------+-------+
|pickup|count  |
+------+-------+
|19    |2730956|
|18    |2647195|
|20    |2459694|
|21    |2338148|
|22    |2267936|
|17    |2218048|
|15    |2053659|
|14    |2027168|
|23    |1976869|
|12    |1942527|
|13    |1934311|
|8     |1912203|
|9     |1907802|
|16    |1899858|
|11    |1779785|
|10    |1740495|
|0     |1596750|
|7     |1453983|
|1     |1172730|
|2     |880165 |
|6     |791182 |
|3     |664392 |
...
|4     |480862 |
|5     |373397 |
+------+-------+
```

- Q3

1. 操作：首先計算出'Total_Amt'之欄位平均，以平均之上定義為

big_total_amount，之下為 small_total_amount，再觀察其在類別欄位'Payment_Type'之表現(其中'Payment_Type_New'為將同義字處理過後之結果，後續將會採用該欄位進行運算)，透過 Spark 當中的 when 與 groupby 進行實踐，最後利用 matplotlib 進行繪圖。

2. 回答：從下圖可發現支付金額較大（big_total_amount）的人通常會採用信用卡居多，而小額支付(small_total_amount)大多使用金錢，反映出人們的用錢習慣