## Big Data Analytics Techniques and Applications Homework 2

309709022 陳政廷

Q1: Find the maximal delays (you should consider both ArrDelay and DepDelay) for each month of 2007.

A: 首先查看 2007 年資料集之 column type,發現"ArrDelay"與"DepDelay"的資料屬性出現問題,故優先轉為 IntegerType,完成轉換後由於此題為探討上述兩欄位之最大 delay 時間,故分別根據"Month"進行 groupby 後分別找出 2007 年每個月"ArrDelay"與"DepDelay"之最大值,如圖 1、圖 2 所示。

+	+			
Month max(ArrDelay)				
++				
1	1426			
2	1359			
3	1564			
4	1402			
5	1429			
6	1351			
7	1386			
8	1472			
9	1665			
10	2598			
11	1146			
12	1942			
+	·+			

+  Month	  max(DepDelay)
+	++
1	1406
2	1340
3	1547
4	1415
5	1416
6	1360
7	1369
8	1449
9	1689
10	2601
11	1137
12	1956
+	·+

圖 2

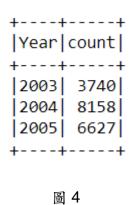
可發現在"ArrDelay"與"DepDelay"兩欄位中,均為 10 月會發生最長的 Delay,且明顯大於其他幾個月,故若今天為航空公司主管應深入探討其原因 (搭乘人數過多?或是當年 10 月可能發生什麼事件導致此結果)以改善客戶的搭乘品質。

接下來,學生分別統計"ArrDelay"與"DepDelay"兩欄位於每月當中的最大值,如圖 3 所示。

Month ma	x(ArrDelay)	max(DepDelay)	Maximal_Delay
1 1	1426	1406	1426
i -i			!
2	1359	1340	1359
3	1564	1547	1564
4	1402	1415	1415
5	1429	1416	1429
6	1351	1360	1360
7	1386	1369	1386
8	1472	1449	1472
9	1665	1689	1689
10	2598	2601	2601
11	1146	1137	1146
12	1942	1956	1956
++		++	++

Q2: How many flights were delayed caused by security between 2000 ~ 2005? Please show the counting for each year.

A: 首先應先檢查各年度是否有缺失或 null 優先將其排除,排除後由於此題要探討因 security 造成的 delay,故排除"SecurityDelay"欄位中為"NA"、0後,便可根據上述欄位計算每年發生該種 delay 的次數,由於 2000~2002 均為 NA 故最後剩下 2003~2005 之資料,如圖 4 所示。



Q3: List Top 5 airports which occur delays most and least in 2008. (Please show the IATA airport code)

A: 因此提要探討在 2008 年最常發生 delay 的機場,故分為出發("Origin"欄位) 與抵達("Dest"欄位)進行探討,接著分別針對"ArrDelay"與"DepDelay"兩欄位篩 選其大於 0 之值留存(提早到達或準時到達不在 delay 之考量範圍),接著便可 分別 groupby 出發("Origin"欄位)與抵達("Dest"欄位)並計算其次數,即可分別得 到 delay 次數最多與最少的前五名出發機場與抵達機場,如圖 5、圖 6 所示。

```
|Dest| count|
+----+
ATL 187243
ORD 151871
DFW 119817
DEN 102681
| LAX | 93508 |
+----+
only showing top 5 rows
+----+
|Dest|count|
+----+
TUP
       2
| PIR|
       4
BJI
       12
| INL|
       17
SUX 35
+----+
only showing top 5 rows
       圖 5
+----+
|Origin| count|
+----+
ATL 175017
   ORD 159427
DFW 127749
DEN 104414
| LAX| 87258|
+----+
only showing top 5 rows
+----+
|Origin|count|
+----+
   INL
         1
   TUP
         1
   PUB
         2
   PIR
         3
```

+----+

only showing top 5 rows

4

BJI

+----+