Big Data Analytics Techniques and Applications _ HW2

310712009 楊家碩 (Nick Yang) GMBA

Questions:

- Q1: Find the maximal delays (you should consider both ArrDelay and DepDelay) for each month of 2007.
- Q2: How many flights were delayed caused by security between 2000 ~ 2005? Please show the counting for each year.
- Q3: List Top 5 airports which occur delays most and least in 2008. (Please show the IATA airport code)
- Anything else worth mentioning (e.g. other valuable observations, or difficulties encountered in this work and how you resolve them).

Ans:

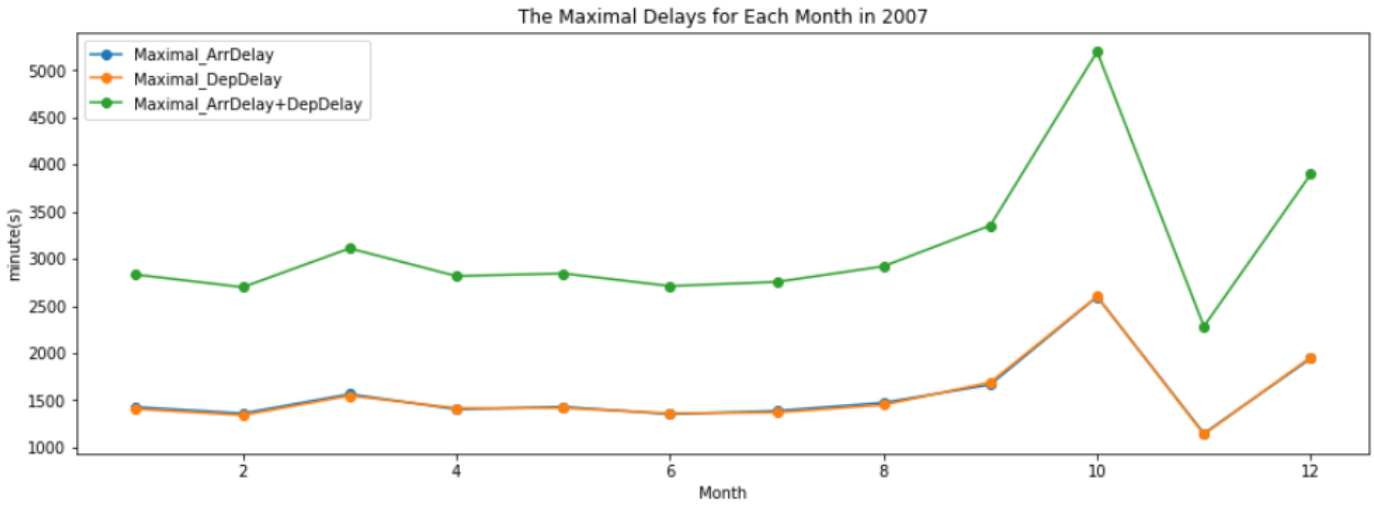
I use both Pyspark and Python to finish the homework, learn their difference in the process, and find the usefulness of Pyspark in dealing with big data.

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

From the table and the figure shown above, we can see that in 2007,

1. The largest arrival delay happens in October when the enormous departure delay occurs.
2. The smallest arrival delay and the most negligible departure delay occurred in November.
3. The maximal delays in January to September and November are smaller than the maximal delays in October and December.

	Month	Maximal_ArrDelay	Maximal_DepDelay	Maximal_ArrDelay+DepDelay
0	1	1426.0	1406.0	2832.0
1	2	1359.0	1340.0	2699.0
2	3	1564.0	1547.0	3111.0
3	4	1402.0	1415.0	2817.0
4	5	1429.0	1416.0	2845.0
5	6	1351.0	1360.0	2711.0
6	7	1386.0	1369.0	2755.0
7	8	1472.0	1449.0	2921.0
8	9	1665.0	1689.0	3354.0
9	10	2598.0	2601.0	5199.0
10	11	1146.0	1137.0	2283.0
11	12	1942.0	1956.0	3898.0

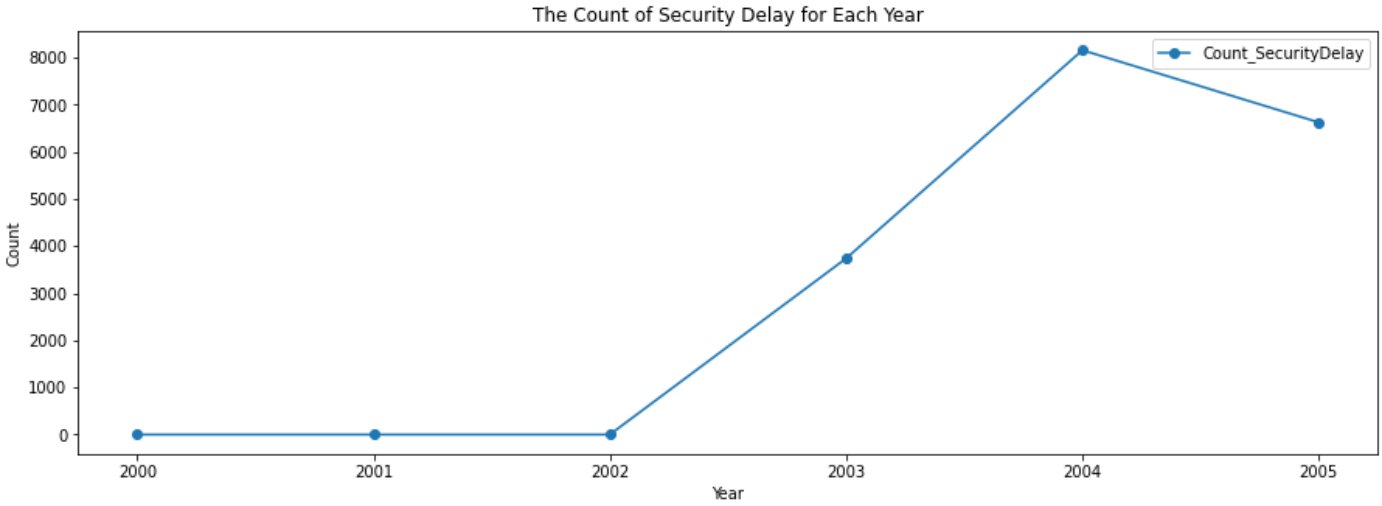


Q2: How many flights were delayed caused by security between 2000 ~ 2005? Please show the counting for each year.

From the table and the figure above, we can find that

1. There are no security delays from 2000 to 2002 in-flight records. However, I think this may occur due to the missing data or because they did not record if a delayed flight was caused by security before 2003.
2. There were more and more security delays from 2003 to 2004, which may be partially driven by the increase in the total number of flights and the increasing security policy.
3. As we can see on the graph, the Year 2004 conducted the most count of security delays than the other year.

	Year	Count_SecurityDelay
0	2000	0
1	2001	0
2	2002	0
3	2003	3740
4	2004	8158
5	2005	6627



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

From the table and the figure above, we can find that

- The top 5 airports that occurs delay the most times are the same for departure delay and arrival delay in 2008, which are

    1. ATL - William B Hartsfield-Atlanta Intl Airport in Atlanta
    2. ORD - Chicago O'Hare International Airport in Chicago
    3. DFW - Dallas-Fort Worth International Airport in Dallas-Fort Worth
    4. DEN - Denver Intl Airport in Denver
    5. LAX - Los Angeles International in Los Angeles

- The top 5 airports that occurs the least times of departure delay and arrival delay in 2008, which are
    - The Top 5 Airports that Occur Departure Delays Least Frequently in 2008

    1. PUB - Pueblo Memorial Airport in Colorado
    2. PIR - Pierre Regional Airport in South Dakota
    3. TUP - Tupelo Regional Airport in Mississippi
    4. INL - Falls International Airport in Minnesota
    5. BJI - Bemidji Regional Airport in Minnesota

    - The Top 5 Airports that Occur Arrival Delays least Frequently in 2008

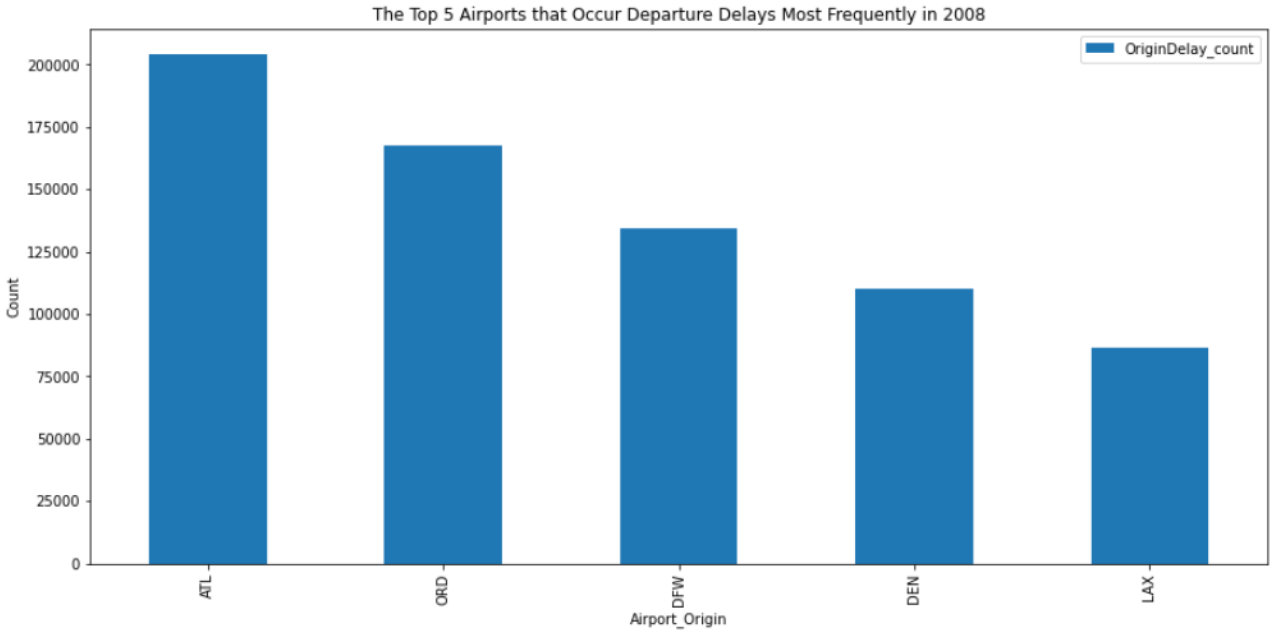
    1. OGD - Ogden-Hinckley Airport in Utah
    2. CYS - Cheyenne Regional Airport in Wyoming
    3. TUP - Tupelo Regional Airport in Mississippi

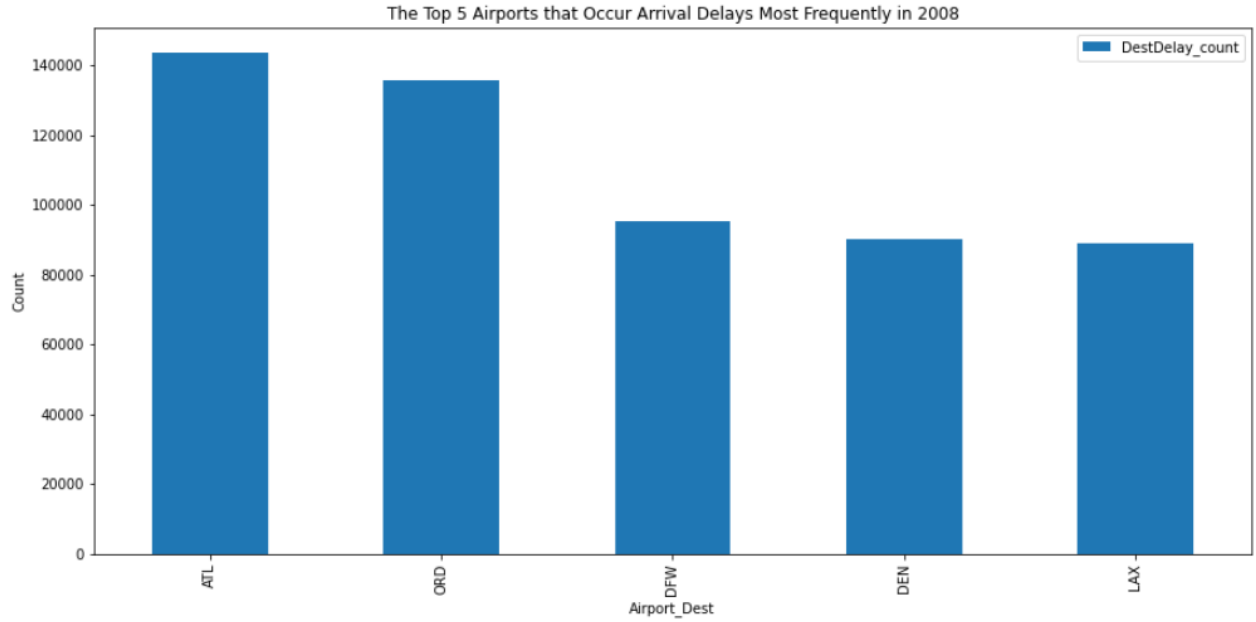
4. PIR - Pierre Regional Airport in South Dakota
5. BJI - Bemidji Regional Airport in Minnesota

The list of Top 5 Airports that Occur Departure Delays (OriginDelay_count) **Most** Frequently in 2008 and the list of Top 5 Airports that Occur Arrival Delays (DestDelay_count) **Most** Frequently in 2008

Origin	OriginDelay_count
ATL	204157
ORD	167536
DFW	134254
DEN	110344
LAX	86615

Dest	DestDelay_count
ATL	143629
ORD	135780
DFW	95262
DEN	89988
LAX	89000

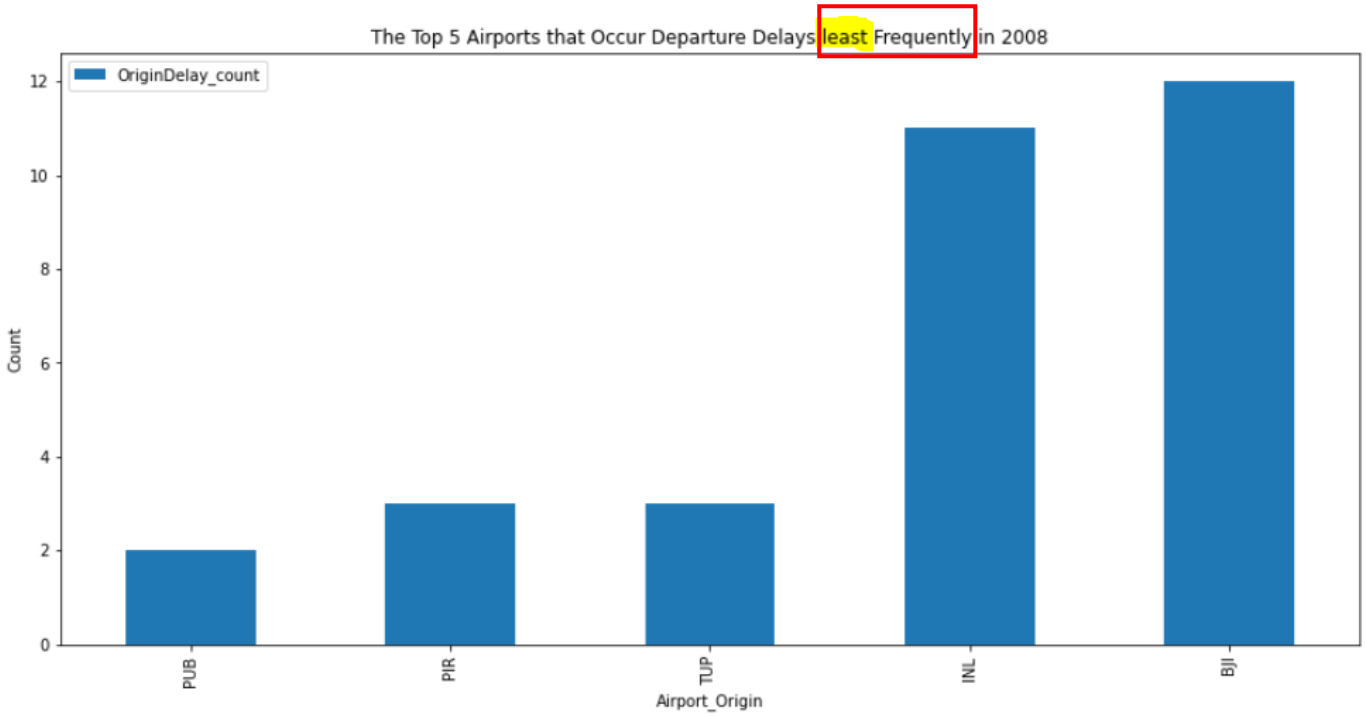


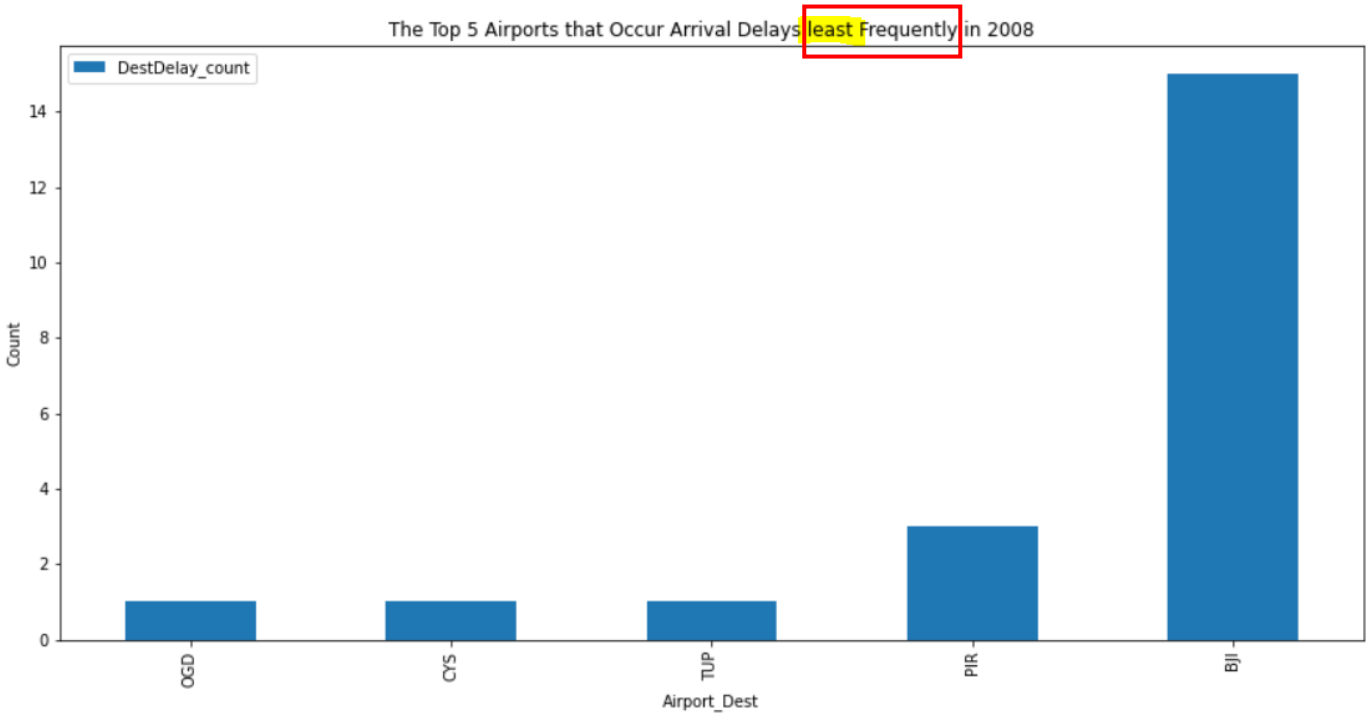


The list of Top 5 Airports that Occur Departure Delays (OriginDelay_count) **least** Frequently in 2008 and the list of Top 5 Airports that Occur Arrival Delays (DestDelay_count) **least** Frequently in 2008

Airport_Origin	OriginDelay_count
PUB	2
PIR	3
TUP	3
INL	11
BJI	12

	Airport_Dest	DestDelay_count
212	OGD	1
289	CYS	1
243	TUP	1
303	PIR	3
297	BJI	15





Anything else worth mentioning (e.g. other valuable observations, or difficulties encountered in this work and how you resolve them).

In this homework, I spend a lot of time on the Pyspark. To understand how to use it, what is the logic, what kind of functions and library we need to include in order to finish the homework.

I learned how to set up the Pyspark in the beginning and got the value from the column to do the calculation and classification to find out the answer I wanted. There are many differences between Python and Pyspark when doing data processing. Still, I was able to overcome the difficulties and challenges to finish and learn this beneficial technique.

I start to notice the advantages of Pyspark, like in-memory computation and swift processing, but I also notice that sometimes Pyspark is slow than other tools like Scala, and when it comes to expressing a problem in MapReduce fashion, sometimes it's difficult. In conclusion, I still think Pyspark is a potent and valuable tool for dealing with big data. I did learn a lot from this homework.