Big Data Analytics Techniques and Applications

Homework 2

Name: 陳以瑄 Student ID:109705001

1. Descriptions of platform you use.
   - Pyspark via Google Colab
2. Descriptions of how you solve each question in detail.

   Q1: Find the maximal delays (you should consider both ArrDelay and DepDelay) for each month of 2008.
   - Source code

```
1 df8 = df8.withColumn("TotalDelay", df8.ArrDelay + df8.DepDelay)
2 df_max_delay = pd.DataFrame(columns=['Month', 'TotalDelay'])
3 for i in range (1,13):
4     df8_month = df8.filter(df8.Month == i)
5     max_total_delay = df8_month.agg({"TotalDelay": "max"}).collect()[0][0]
6     print(f"Max total delay in {i}th month : {max_total_delay}")
7     df_max_delay = pd.concat([df_max_delay,\
8                               pd.DataFrame([{'Month':i,'TotalDelay':max_total_delay}])],\
9                               axis=0, ignore_index=True)
10 df_max_delay
```

   - How I solve
     1. Using “withColumn” function (line 1) to store the sum of “ArrDelay” and “DepDelay” in a new column
     2. Using “filter” function (line 4) to select the data of specific month
     3. Using “agg({“col”: “max”})” function (line 5) to find the maximal value of “TotalDelay” column.
     4. Safe the data in a table
   - Result

Month	TotalDelay
1	2800
2	4918
3	2980
4	4920
5	3903
6	3417
7	3028
8	2726
9	3135
10	2761
11	2594
12	3252

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

- Source code

```
df=[df0, df1, df2, df3, df4, df5]
df_sec_delay = pd.DataFrame(columns=['Year', 'SecDelay'])
for i in range (6):
    security_delay = df[i].filter((df[i].SecurityDelay.isNotNull()) & (df[i].SecurityDelay> 0)).count()
    print(f"Security Delay in 200{i} : {security_delay}")
    df_sec_delay = pd.concat([df_sec_delay,\
                              pd.DataFrame([{'Year':2000+i, 'SecDelay':security_delay}])],\
                              axis=0, ignore_index=True)
df_sec_delay
```

- How I solve:
    1. Using “filter” function (line 4) to select the row that “SecurityDelay” is neither null nor 0
    2. Using “count” function to calculate the number of rows
    3. Safe the data in a table
- Result

    The total number of flights were delayed caused by security between 2000 ~ 2005 is 18525.

    However, the data in 2000 ~ 2002 are missing. I use “groupby” function and find out that all of the values in “SecurityDelay”column are null.

```
df0.groupby('SecurityDelay').count().show()
+-------------+-------+
|SecurityDelay|  count|
+-------------+-------+
|         null|5683047|
+-------------+-------+
```

```
df1.groupby('SecurityDelay').count().show()
+-------------+-------+
|SecurityDelay|  count|
+-------------+-------+
|         null|5967780|
+-------------+-------+
```

```
df2.groupby('SecurityDelay').count().show()
+-------------+-------+
|SecurityDelay|  count|
+-------------+-------+
|         null|5271359|
+-------------+-------+
```

So actually the total number is the sum from 2003~2005, and the counting of each year is

Year	SecDelay
2000	0
2001	0
2002	0
2003	3740
2004	8158
2005	6627

Q3: List Top 5 airports which occur delays most and least in 2007. (Please show the IATA airport code)

- Source code

  Part1

```
df7_arr=df7.filter((df7.ArrDelay.isNotNull()))                                  #真的有飛機來  不管有沒有delay
df7_arr_delay=df7.filter((df7.ArrDelay.isNotNull())  &  (df7.ArrDelay>  0))
df7_dep=df7.filter((df7.DepDelay.isNotNull()))
df7_dep_delay=df7.filter((df7.DepDelay.isNotNull())  &  (df7.DepDelay>  0))

arr_count  =  df7_arr.groupby('Dest').count().withColumnRenamed('Dest',  'arr_Dest')\
                                             .withColumnRenamed('count',  'arr_count')
arr_delay_count  =  df7_arr_delay.groupby('Dest').count().withColumnRenamed('Dest',  'arr_delay_Dest')\
                                                    .withColumnRenamed('count',  'arr_delay_count')
dep_count  =  df7_dep.groupby('Origin').count().withColumnRenamed('Origin',  'dep_Origin')\
                                               .withColumnRenamed('count',  'dep_count')
dep_delay_count  =  df7_dep_delay.groupby('Origin').count().withColumnRenamed('Origin',  'dep_delay_Origin')\
                                                   .withColumnRenamed('count',  'dep_delay_count')

arr  =  arr_count.join(arr_delay_count,  arr_count.arr_Dest  ==  arr_delay_count.arr_delay_Dest,  how  =  'left')\
              .na.fill(value=0)
arr  =  arr.withColumn("arrive_delay",  0  +  arr.arr_delay_count)\
           .drop("arr_delay_Dest").drop('arr_count').drop('arr_delay_count')
dep  =  dep_count.join(dep_delay_count,  dep_count.dep_Origin  ==  dep_delay_count.dep_delay_Origin,  how  =  'left')\
              .na.fill(value=0)
dep  =  dep.withColumn("departure_delay",  0  +  dep.dep_delay_count)\
           .drop("dep_delay_Origin").drop('dep_count').drop('dep_delay_count')
```

  Part2

```
total  =  arr.join(dep,  arr.arr_Dest  ==  dep.dep_Origin,  'outer')
total  =  total.withColumn("total_delay",  total.arrive_delay+  total.departure_delay)\
               .drop("dep_Origin").withColumnRenamed('arr_Dest',  'IATA').sort("total_delay")

total_top  =  total.tail(5)
df_total_top  =  pd.DataFrame(total_top,  columns  =['IATA','arrive_delay','departure_delay','total_delay'])
df_total_top  =  df_total_top.sort_values(by='total_delay',  ascending=False)
df_total_top=df_total_top.reset_index()
df_total_top=df_total_top.drop(columns={"index"})
total_least  =  total.head(5)
df_total_least  =  pd.DataFrame(total_least,  columns  =['IATA','arrive_delay','departure_delay','total_delay'])

print("Top  5  airports  which  occur  delays  most  in  2007")
display(df_total_top)
print("Top  5  airports  which  occur  delays  least  in  2007")
display(df_total_least)
```

- How I solve

  1. For ArrDelay
     I. I use “filter” function (Part 1, line 2)to select the row that “ArrDelay” is neither null nor less than 0, which means these flight have arrival delay.
     II. Since “ArrDelay” means the delay happened when arriving to the destination, I use “groupby” function (Part 1, line 8) to group the rows based on the “Dest” column. And save the result in “arr_delay_count”
     III. To make sure that the airport is “did have flight arrive and none of them delayed” not “did not have any flight arrive so it could not have any delay record ”, I record the airport that

did have flight arrive, by filtering “ArrDelay” is not null (Part 1, line 1), and grouping them based on “Dest” (Part 1, line 6) . Then, save the result in “arr _count”

IV. Merge the “arr_count” and “arr_delay_count” together by the left join method (Part1, line15).And save the number of arrival delay in column “arrive_delay”(Part1, line17). Since it is left-joined, all the airport on this table did have flight arrive. So that we can make sure the number 0 in “arrive_delay” means “did have flight arrive and none of them delayed”. And save the result in “arr”.

2. For DepDelay
   I. Similarly to ArrDelay.
   II. The only different is that, since “DepDelay” means the delay happened when departure from origin, I use “groupby” function (Part 1, line 10) to group the rows based on the “Origin” column.
   III. Save the final result in “dep”.
3. Merge “arr” and “dep” together in “total” (Part 2, line 1).
4. Calculate the sum of “arrive_delay” and “departure_delay” for each airport in the new column “total_delay” (Part 2, line 2).
5. Sort the data by “total_delay” in ascending order(Part 2, line 3).
6. Using “tail” function (Part 2, line 5) to extract the last 5 rows, which has the largest number of total delay.
7. Using “head” function (Part 2, line 8) to extract the last 5 rows, which has the smallest number of total delay.
8. Safe the data in a table

- Result
  1. The Top 5 airports which occur delays most:
     ATL, ORD, DFW, DEN, LAX
  2. The Top 5 airports which occur delays least:
     GLH, MKC, ISO, PIR, EAU

Top 5 airports which occur delays most in 2007

	IATA	arrive_delay	departure_delay	total_delay
0	ATL	186911	206118	393029
1	ORD	177054	183984	361038
2	DFW	134824	135433	270257
3	DEN	110417	109839	220256
4	LAX	109643	101585	211228

Top 5 airports which occur delays least in 2007

	IATA	arrive_delay	departure_delay	total_delay
0	GLH	2	0	2
1	MKC	1	1	2
2	ISO	3	2	5
3	PIR	2	4	6
4	EAU	20	12	32