IMT 575

Flights in SQL

Prem Shah

1. Flights to Seattle:

a. How many flights were there from NYC airports to Seattle in 2013?

```
SELECT count(*)

FROM rodriglr."table_flights.csv"
WHERE dest = 'SEA'
```

ANS: 3885

b. How many airlines fly from NYC to Seattle?

```
SELECT count(DISTINCT carrier)
FROM rodriglr."table_flights.csv"
WHERE dest = 'SEA'
```

ANS: 5

c. How many unique air planes fly from NYC to Seattle?

```
SELECT count(DISTINCT tailnum) AS 'No. of unique airlines'
FROM rodriglr."table_flights.csv"
WHERE dest = 'SEA'
```

ANS: 933

d. What is the average arrival delay for flights from NYC to Seattle?

```
SELECT AVG(arr_delay) AS "Average_arrival_delay"
FROM rodriglr."flights.csv"
WHERE dest='SEA'
```

ANS: -1.0990990

e. What proportion of flights to Seattle come from each NYC airport?

```
SELECT (count(dest)*1.0/
        (SELECT COUNT(*)
         FROM rodriglr."flights.csv"
         WHERE dest = 'SEA')) as Percentage_of_flights, origin
FROM rodriglr."flights.csv"
WHERE dest = 'SEA'
GROUP BY origin
```

ANS:

percentage_of_flights	origin
0.5341055341055341055	JFK
0.4658944658944658944	EWR

2. Flights Delays

a. Which date has the largest average departure delay? Which date has the largest average arrival delay?

```
SELECT year, month, day, AVG (arr_delay) AS "average_arrival_delay"
FROM rodriglr."flights.csv"
GROUP BY year,month,day
ORDER BY Average_arrival_delay DESC
LIMIT 1

SELECT year, month, day, AVG (dep_delay) AS "average_departure_delay"
FROM rodriglr."flights.csv"
GROUP BY year,month,day
ORDER BY Average_departure_delay DESC
LIMIT 1
```

Average arrival delay:

year	month	day	average_arrival_delay
2013	3	8	85.8621553884711779

Average departure delay:

year	month	day	average_departure_delay
2013	3	8	83.6478696741854637

b. What was the worst day to fly out of NYC in 2013 if you dislike delayed flights? (This one is less straightforward in SQL than you may expect.)

I tried to execute this query in two ways. One is maximum delay in a particular year, month and day only considering delay > 0.

The other way was to select delay_per_flight

```
SELECT year,  month, day ,( AVG (dep_delay)* 1.0 / (COUNT(flight))) as delay_per_flight
FROM rodriglr."flights.csv"
WHERE dep_delay >0
GROUP BY year,month,day
ORDER BY delay_per_flight DESC
LIMIT 1
```

year	month	day	delay_per_flight
2013	9	12	0.27133987555165702696

```
SELECT year,  month, day ,AVG (dep_delay) as delay_per_flight
FROM rodriglr."flights.csv"
WHERE dep_delay >0
GROUP BY year,month,day
ORDER BY ratio DESC
LIMIT 1
```

year	month	day	delay_per_flight
2013	9	12	103.6518324607329843

Both of the results point to 12th September.

c. Is Autumn (September, October, November) worse than Summer (June, July, August) for flight delays for flights from NYC?

From the below results, you can see that Summer flight delays are almost three times than Autumn flight delays.

```
SELECT AVG (dep_delay) AS "Avg. Autumn Monthly Dep Delay"
    FROM rodriglr."flights.csv"
    WHERE month IN (9,10,11)

SELECT AVG (dep_delay) AS "Avg. Monthly Dep Delay"
    FROM rodriglr."flights.csv"
    WHERE month IN (6,7,8)
```

Avg. Summer Monthly Dep Delay
18.2058746612143978

Avg. Autumn Monthly Dep Delay
6.0976161939006526

d. On average, how do departure delays vary over the course of a day? You can compute the average delay by hour of day, such that your result will have 24 records (be careful -- there are records with hour 0 and hour 24. Consider lumping these together, or justify any other solution you come up with.) No need to plot the results.

```
SELECT AVG (dep_delay) AS "Average_Hourly_Delay", (CASE WHEN hour = 24 THEN 0 ELSE hour END)
as av_hour

    FROM rodriglr."flights.csv"
    GROUP BY av_hour
```

Average_Hourly_Delay	av_hour
127.2232044198895028	0
206.7556561085972851	1
236.2539682539682540	2
304.7272727272727273	3
-5.5540983606557377	4
-4.3562932226832642	5
-1.5218102267202899	6
0.21472278013919379700	7
1.09231236014715363902	8
4.2341126461211477	9
5.5110722974237415	10
5.6132719004308281	11
7.5173489765351972	12
9.3639062036212526	13
8.0518289693046975	14
10.5933136589877990	15
13.5572495053067098	16
16.6557466309723672	17
18.4746655479420128	18
21.3102007951285793	19
28.0875939616077530	20
41.8441451346893898	21
67.9586156381615089	22
96.6384202453987730	23

From the above results, you can see that the delay is high in early mornings and late nights.

3. Velocity

Which flight departing NYC in 2013 flew the fastest

From the below results, you can see flight 1499 has the highest speed

```
SELECT year, month, day, flight, tailnum, (distance* 60.0/air_time) as mph_speed
    FROM rodriglr."flights.csv"
    ORDER BY mph_speed DESC
    LIMIT 1
```

year	month	day	flight	tailnum	mph_speed
2013	5	25	1499	N666DN	703.3846153846153846

4. Routine Flights:

Which flights (i.e. carrier + flight + dest) happen every day?

Here, I see the total distinct combination number of days and month the flight has flown. If it is 365, the flight has flown on every single day of the year 2013.

```
SELECT carrier, flight,dest , COUNT(DISTINCT (CONCAT (day, '-', month, '-', year))) AS
newdate
FROM rodriglr."table_flights.csv"
GROUP BY carrier,flight,dest
ORDER BY newdate DESC
LIMIT 1
```

carrier	flight	dest	newdate
B6	1783	MCO	365

5. Open Ended Research Question

For flights from New York to Seattle, which airlines have the best performance in terms of delays?

This question is interesting because it helps people who do not like flight delays and who want the best flight in terms of the least arrival and departure delay.

As you can see from the above plots, AS (Alaska Airlines) has the least average departure delay and the least average arrival delay as well while United Airlines (UA) has the highest average departure delay for flights to Seattle. B6 has the highest average arrival delay for flights from NYC to Seattle. Hence, people who want departures & arrivals on time should choose Alaska Airlines. This inference have one problem which we did not consider is that since Alaska Airlines is based out of Seattle, hence might have more flights in this route which might have resulted in the lower

departure delay. But as we can see from the below table, that is not the case. Hence we can safely say that Alaska Airlines has a good track record of arriving and departing on time.

```
SELECT ROUND(AVG(dep_delay),2) as avg_dep_delay,
ROUND(AVG(arr_delay),2) as avg_arr_delay,
carrier
FROM rodriglr."flights.csv" f
WHERE dest='SEA'
group by carrier
ORDER by avg_arr_delay, avg_dep_delay DESC

SELECT COUNT(tailnum),
carrier
FROM rodriglr."flights.csv" f
WHERE dest='SEA'
group by carrier
ORDER by COUNT(tailnum) ASC

SELECT ROUND(((AVG(dep_delay) + AVG(arr_delay))/2/COUNT(tailnum)),7)  as
avg_total_delay_per_flight,
ROUND(AVG(dep_delay),2) as average_dep_delay,
Round(AVG(arr_delay),2) as average_arr_delay,
COUNT(tailnum) as total_flights,
carrier
FROM rodriglr."flights.csv" f
WHERE dest='SEA'
group by carrier
ORDER by avg_total_delay_per_flight ASC
```

avg_dep_delay	avg_arr_delay	carrier
5.83	-9.93	AS
6.98	-5.89	DL
10.10	-1.48	AA
17.32	5.83	UA
11.59	7.72	B6

count	carrier
360	AA
513	B6
709	AS
1101	UA
1202	DL

avg_total_delay_per_flight	average_dep_delay	average_arr_delay	total_flights	carrier
-0.0028915	5.83	-9.93	709	AS
0.0004561	6.98	-5.89	1202	DL
0.0105127	17.32	5.83	1101	UA
0.0119792	10.10	-1.48	360	AA
0.0188244	11.59	7.72	513	B6

6. **Exogenous Effects**

   Is there any link between visibility and delay? What about temperature?

```
SELECT ROUND(AVG(dep_delay),2) as avg_delay,
ROUND(AVG(visib),2) as avg_visib,
ROUND(AVG(temp),2) as avg_temp,
flight, tailnum
FROM rodriglr."flights.csv" f
INNER JOIN rodriglr."weather.csv" w
ON f.year = w.year AND
f.month = w.month AND
f.day = w.day AND
f.hour = w.hour
WHERE dep_delay > 60
group by flight,tailnum
ORDER by avg_delay ASC


SELECT ROUND(AVG(dep_delay),2) as avg_delay,
ROUND(AVG(visib),2) as avg_visib,
ROUND(AVG(temp),2) as avg_temp,
f.month, f.day
FROM rodriglr."flights.csv" f
INNER JOIN rodriglr."weather.csv" w
ON f.year = w.year AND
f.month = w.month AND
f.day = w.day AND
f.hour = w.hour
WHERE dep_delay BETWEEN -15 and 0
group by f.month, f.day
ORDER by avg_delay DESC
```

I define delayed as anything more than 60 minutes delay and on time as anything with delay between -15 and 0.

avg_delay	avg_visib	avg_temp	month	day
-2.91	9.86	28.31	12	18
-2.92	9.49	62.34	5	24
-3.00	8.13	63.28	5	8
-3.00	9.50	77.89	7	23
-3.00	2.47	33.43	3	8
-3.10	9.45	31.02	12	19
-3.18	10.00	44.73	1	31
-3.24	8.46	25.77	12	17
-3.24	8.98	34.59	2	8
-3.27	8.85	59.46	12	23
-3.35	9.77	31.81	2	11
-3.39	10.00	76.52	6	28
-3.41	10.00	80.96	7	11
-3.42	10.00	66.50	12	22

avg_delay	avg_visib	avg_temp	month	day
201.15	9.00	44.59	1	9
180.03	8.36	82.22	9	12
178.59	10.00	65.55	4	10
175.78	9.77	85.43	7	10
172.08	8.36	78.19	9	2
168.65	9.44	76.19	7	28
167.62	10.00	44.63	1	10
166.66	9.44	81.54	7	22
165.11	8.06	27.08	12	14
164.27	9.25	77.84	5	23
162.23	5.34	56.64	12	5
161.30	10.00	91.05	7	7
159.98	9.54	80.76	6	27
159.73	10.00	87.30	6	24
158.96	9.14	38.84	3	8
157.65	9.51	69.26	4	19

From the above results, I do not see any particular trends between visibility, temperature and delay.