Project 1: Airlines

Michał Lubczyński

01.04.2022

a) What was the average arrival delay?

```sql
SELECT avg(arr_delay_new) as avg_delay
FROM "Flight_delays"
```

avg_delay
15.91152

b) What was the maximum arrival delay?

```sql
SELECT max(arr_delay_new)
FROM "Flight_delays"
```

max
1895

c) Which flight had the biggest delay in arrival?

```sql
SELECT carrier,
       origin_city_name,
       dest_city_name,
       fl_date,
       arr_delay_new
FROM "Flight_delays"
WHERE arr_delay_new IN (SELECT MAX(arr_delay_new)
                          FROM "Flight_delays")
```

carrier	origin_city_name	dest_city_name	fl_date	arr_delay_new
AA	Kona, HI	Los Angeles, CA	2017-07-26	1895

d) Which days of the week are the worst to travel?

```sql
SELECT weekday_name,
       AVG(arr_delay_new) AS avg_delay
FROM "Weekdays"
   JOIN "Flight_delays" ON weekday_id = day_of_week
GROUP BY weekday_name
ORDER BY avg_delay DESC
```

weekday__name	avg__delay
Friday	20.80747
Monday	18.04801
Wednesday	16.10514
Thursday	15.64696
Saturday	15.21876
Tuesday	12.88056
Sunday	12.77606

e) Which airlines flying from San Francisco (SFO) have the lowest arrival delay?

```sql
SELECT AL.airline_name,
       avg(FD.arr_delay_new) AS avg_delay
FROM "Flight_delays" FD
    JOIN "Airlines" AL ON FD.airline_id = AL.airline_id
WHERE FD.airline_id IN (SELECT airline_id
                        FROM "Flight_delays"
                        WHERE origin = 'SFO')
GROUP BY airline_name
ORDER BY avg_delay DESC
```

airline__name	avg__delay
JetBlue Airways: B6	28.841148
Frontier Airlines Inc.: F9	18.980300
American Airlines Inc.: AA	18.375314
United Air Lines Inc.: UA	16.950403
SkyWest Airlines Inc.: OO	16.808273
Virgin America: VX	13.964467
Southwest Airlines Co.: WN	13.823983
Delta Air Lines Inc.: DL	12.258788
Alaska Airlines Inc.: AS	7.453927
Hawaiian Airlines Inc.: HA	4.202719

f) What proportion of airlines has regular delays, i.e., their flight has an average delay of at least 10 minutes?

```sql
SELECT (count(DISTINCT FD2.airline_id)*1.0)
       /count(DISTINCT FD1.airline_id) AS late_proportion
FROM "Flight_delays" FD1
    JOIN "Flight_delays" FD2 ON FD1.fl_num = FD2.fl_num
    WHERE FD2.airline_id IN (SELECT airline_id
                             FROM "Flight_delays"
                             GROUP BY airline_id
                             HAVING AVG(arr_delay_new)>10)
```

late__proportion
0.8333333

g) How do departure delays affect arrival delays?
Pearson correlation coefficient between departure delay time and arrival delay time.

```sql
SELECT ((SUM(dep_delay_new * arr_delay_new)
    - (SUM(dep_delay_new) * SUM(arr_delay_new)) / COUNT(*)
    )
  )
  / (SQRT(SUM(dep_delay_new * dep_delay_new)
        - (SUM(dep_delay_new) * SUM (dep_delay_new)) / COUNT(*)
        )
    * SQRT(SUM(arr_delay_new * arr_delay_new)
          - (SUM(arr_delay_new) * SUM(arr_delay_new)) / COUNT(*)
          )
    ) AS Pearsons_r
FROM "Flight_delays"
```

pearsons_r
0.9717058

h) Which airline had the largest increase in average arrival delay during the last week of the month, i.e., between July 1-23 and July 24-31?

```sql
SELECT AL.airline_name,
       avg(FD1.arr_delay_new)
       -avg(FD2.arr_delay_new) AS delay_increase
FROM "Flight_delays" FD1
    JOIN "Flight_delays" FD2 ON FD1.airline_id = FD2.airline_id
    AND FD2.day_of_month<24 AND FD1.day_of_month>=24
    JOIN "Airlines" AL ON FD1.airline_id = AL.airline_id
GROUP BY AL.airline_name
ORDER BY delay_increase DESC
LIMIT 1
```

airline_name	delay_increase
Southwest Airlines Co.: WN	0.584763

i) Which airlines fly both SFO -> PDX (Portland) and SFO -> EUG (Eugene)?

```sql
SELECT DISTINCT airline_name
FROM "Airlines" AL
    Left JOIN "Flight_delays" FD ON AL.airline_id = FD.airline_id
    AND FD.dest ='EUG'
WHERE FD.origin = 'SFO' AND AL.airline_id IN (SELECT airline_id
                                              FROM "Flight_delays"
                                              WHERE dest = 'PDX')
```

airline_name
SkyWest Airlines Inc.: OO
United Air Lines Inc.: UA

j) What is the fastest way to get from Chicago to Stanford, assuming departure after 2pm local time?

```sql
SELECT origin,
       dest,
       avg(arr_delay_new) AS avg_delay
FROM "Flight_delays"
WHERE origin IN ('MDW','ORD')
AND dest IN ('SFO','SJC','OAK')
AND crs_dep_time > 1400
GROUP BY origin,
       dest
ORDER BY avg_delay desc
```

origin	dest	avg_delay
ORD	SFO	22.19253
MDW	SFO	19.85714
MDW	SJC	17.20000
ORD	SJC	14.81111
MDW	OAK	12.12903