IMT 575 – Assignment 2 – Flights in SQL

Shreya Sabharwal

-- 1. Flights to Seattle
-- 1(a): How many flights were there from NYC airports to Seattle in 2013?

```
SELECT count(*) FROM rodriglr."table_flights.csv"
where dest = 'SEA'

--Query Result: 3885
```

-- 1(b): How many airlines fly from NYC to Seattle?

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

-- Query Result: 5
```

-- 1(c): How many unique air planes fly from NYC to Seattle?

```
SELECT count(distinct tailnum) FROM rodriglr."table_flights.csv"
where dest = 'SEA'

-- Query Result: 933
```

-- 1(d): What is the average arrival delay for flights from NYC to Seattle?

```
SELECT avg(arr_delay) FROM rodriglr."table_flights.csv"
where dest = 'SEA'

-- Query Result: -1.099
```

-- 1(e): What proportion of flights to Seattle come from each NYC airport?

```
SELECT origin, count(*)*1.0/(select count(*) from
rodriglr."table_flights.csv" where dest='SEA') as proportion
FROM rodriglr."table_flights.csv"
where dest = 'SEA'
group by origin

-- Query Result: JFK   0.534
-- EWR     0.465
```

-- 2. Flight Delays
-- 2. a) Which date has the largest average departure delay? Which date has the largest average arrival delay?

```
SELECT month, day, avg(dep_delay)
FROM rodriglr."table_flights.csv"
group by year, month, day
order by avg(dep_delay) desc
limit 1

-- Query Result:
-- month   day  avg
-- 3     8    83.6478696741854637

SELECT month, day, avg(arr_delay)
FROM rodriglr."table_flights.csv"
group by year, month, day
order by avg(arr_delay) desc
limit 1

-- Query Result
-- month   day      avg
-- 3     8    85.8621553884711779
```

-- 2. 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.)

```
SELECT month, day, count(flight) as num_flights
FROM rodriglr."table_flights.csv"
where dep_delay>0
group by year, month, day
order by num_flights desc
limit 1

-- Query Result:
-- month   day      num_flights
-- 12    23         673
```

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

```
-- autumn
select avg(avg_delay) from
(SELECT avg(dep_delay) as avg_delay, month
FROM rodriglr."table_flights.csv"
where month in (9, 10, 11)
group by month) a

-- avg 6.0946001233501496

--summer
select avg(avg_delay) from
(SELECT avg(dep_delay) as avg_delay, month
FROM rodriglr."table_flights.csv"
where month in (6, 7, 8)
group by month) a

-- avg 18.2727723593803359

--No, autumn is better than summer for flight delays for flights from NYC
```

-- 2. d) On average, how do departure delays vary over the course of a day?

```
SELECT (case when hour = 0 then 24 else hour end) as hour_1,
avg(dep_delay) as avg_delay
FROM rodriglr."table_flights.csv"
group by hour_1
order by hour_1

-- Query Result:
--hour_1   avg_delay
--1       206.7556561085972851
--2       236.2539682539682540
--3       304.7272727272727273
--4       -5.5540983606557377
--5       -4.3562932226832642
--6       -1.5218102267202899
--7       0.21472278013919379700
--8       1.09231236014715363902
--9       4.2341126461211477
--10  5.5110722974237415
--11  5.6132719004308281
--12  7.5173489765351972
--13  9.3639062036212526
--14  8.0518289693046975
--15  10.5933136589877990
--16  13.5572495053067098
--17  16.6557466309723672
--18  18.4746655479420128
--19  21.3102007951285793
--20  28.0875939616077530
--21  41.8441451346893898
--22  67.9586156381615089
--23  96.6384202453987730
--24  127.2232044198895028
```

-- Flight Delay is maximum around midnight till 3 am in the morning. Starting at 4am, the departure delays are the least.
-- The delays increase aroudnd 10pm and reach tge maximum at 3am in the morning.

--3. Velocity:
-- Which flight departing NYC in 2013 flew the fastest?

```
SELECT max((distance*1.0)/air_time) as speed, flight, carrier
FROM rodriglr."table_flights.csv"
group by flight, carrier
order by speed desc
limit 1
```

```
-- Query Result:
-- speed   flight  carrier
-- 11.72   1499    DL
```

-- Flight 1499, carrier DL has the maximum speed of 11.72 units

--4. Routine flights:
-- Which flights (i.e. carrier + flight + dest) happen every day?

```
SELECT concat(flight,'-' ,carrier,'-', dest) as fl
FROM rodriglr."table_flights.csv"
group by fl
having count(day)=365
```

```
-- Query Result:
--  fl              count
--  1783-B6-MCO      365
```

-- Flight 1783 Carrier B6 Dest MCO happens everyday

--5. Open-ended:
-- Develop one research question you can address using the nycflights2013 dataset, and that you can answer using SQL.

-- Research Question: In 2013, which flights from NYC to Seattle should a passenger consider booking for a better experience in future and which ones should they definitely avoid?

-- The question is interesting because it involves a better experience for the user and serves as a recommendation/warning for the customers who book the flight.

```
SELECT carrier, avg(dep_Delay) as avg_dep_Delay,
avg((distance*1.0)/air_time) as speed
```

```
FROM rodriglr."table_flights.csv"
where dest='SEA'
group by carrier

-- Query Results:
-- carrier avg_dep_delay        speed
-- AA 10.1000000000000000       7.2221437215783241
-- AS 5.8307475317348378        7.3946481535970275
-- B6 11.5925925925925926       7.3657413610171181
-- DL 6.9825291181364393        7.4143186276842577
-- UA 17.3215258855585831       7.3836511400628365

-- Since the speed is almost the same for all carriers, departure delay
is the one that users should look into. AS has the least departure delay
-- and UA has the maximum departure delay. Therefore, user should
consider AS for a better experience and definitely avoid UA as it has the
-- maximum average delay
```

--6. Exogenous effects:
-- We might like to understand potential causes of delays, such as weather conditions.
-- Is there any link between visibility and delay? What about temperature?

```
SELECT avg(temp) as temp, avg(visib) as visib, avg(dep_delay) as
dep_delay,
(case when dep_delay>(select avg(dep_delay) from
rodriglr."table_flights.csv") then 1 else 0 end) as is_Delayed
from
rodriglr."table_flights.csv" fl join rodriglr."table_weather.csv" wt
on fl.month = wt.month and
fl.day = wt.day and
fl.hour = wt.hour
group by is_Delayed

-- Query Results
--  temp   visib dep_delay  is_delayed
--  55.19  9.64  -2.59      0
--  60.03  9.489 59.97      1

-- For the delayed flights, average temperature is relatively higher than
that of on-time flights. In case of high heat and temperature,
-- some planes cannot take off and wait for cooler hours to take off.

-- Also, visibility is slightly lower for delayed flights. This makes
sense as low visibility due to smog or fog causes flights to delay.
-- The threshold taken for delayed flights is the average departure delay
for all flights. The mean is generally taken as the baseline for
-- comparisons as it gives a good estimate. A limitation of this is that
it may include some outliers as well.
```