Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Mireya Lyons
Date: 24 August 2022

Assignment

Recommend which pair of United States airports should be connected with a high-speed passenger rail tunnel. To do this, write and run a SELECT statement to return pairs of airports that are between **300** and **400** miles apart and that had at least **5,000** (five thousand) flights per year on average *in each direction* between them. Arrange the rows to identify which one of these pairs of airports has largest total number of seats on the planes that flew between them. Your SELECT statement must return all the information required to fill in the table below.

Recommendation

I recommend the following tunnel route:

	First Direction	Second Direction
Three-letter airport code for origin	SFO	LAX
Three-letter airport code for destination	LAX	SFO
Average flight distance in miles	337	337
Average number of flights per year	14712	14540
Average annual passenger capacity	1996597	1981059
Average arrival delay in minutes	10	14

Method

I identified this route by running the following SELECT statement using **Impala** on the VM:

```
SELECT
        origin AS Origin,
        dest AS Destination,
        AVG(distance) AS Avg_Distance,
        ROUND(COUNT(flight)/10) AS Avg_Annual_Num_of_Flights,
        ROUND(SUM(seats)/10) AS Avg_Annual_Seat_Capacity,
        ROUND(AVG(arr_delay)) AS Avg_Delay
FROM flights f
LEFT OUTER JOIN planes p
ON f.tailnum = p.tailnum
WHERE 300 <= f.distance AND f.distance <= 400
GROUP BY Origin, Destination
HAVING Avg_Annual_Num_of_Flights >= 5000
ORDER BY Avg_Annual_Seat_Capacity DESC
LIMIT 10;
```

Notes

1. There was another route with a lower average number of flights and average seat capacity, the average delay between the directions was considered before recommending the tunnel route.