Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Kyle Hollands
Date: November 21st, 2020

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 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 of these pairs of airports has the largest 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 the destination	LAX	SFO
The average flight distance in miles	337	337
The average number of flights per year	14712	14540
The average annual passenger capacity	1996597	1981059
The average arrival delay in minutes	147765	194572

(Replace AAA and BBB with the actual airport codes, and fill in all the table cells.)

Method

(Fill in the blank to indicate whether you used Hive or Impala, and fill in the SQL query.)

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

```
SELECT origin AS Origin,
    dest as Destination,
    ROUND(AVG(f.distance)) AS Average_Distance,
    ROUND(COUNT(f.flight) / 10) AS Annual_Average_Number_of_Flights,
    ROUND(SUM(p.seats) / 10) AS Annual_Average_Seat_Capacity,
    ROUND(SUM(f.arr_delay) / 10) As Annual_Average_Arrival_Delay
  FROM fly.flights f LEFT OUTER JOIN fly.planes p
  ON f.tailnum = p.tailnum --Join the planes table by unique ID tailnum to the flights table by unique ID tailnum.
    WHERE f.distance >= 300 AND distance <= 400 --Filter results that are between 300 and 400 miles.
    GROUP BY Origin, Destination
    HAVING Annual_Average_Number_of_Flights >= 5000 --Filter results over 5000 average number of flights annually.
    ORDER BY Annual_Average_Seat_Capacity DESC
    LIMIT 10;
```

Notes

(This section is optional. You may use it to describe your process, add details or caveats, explain your interpretations, or express any further analysis that you performed.)

When I initially began creating the SQL statement, a broad overview encompassed a large portion of the data. From that point on, I began to define it in a more focused way, as I introduced specific requirements outlined by the project specifications.

Like how I approach writing scripts in VBA or Python, this allowed for a more structured approach instead of attempting to accommodate all requirements in the initial analysis.