Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Kaushik Nagwekar
Date: 21/09/2020

(Include your name and today's date above.)

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	PHX	LAX
Three-letter airport code for destination	LAX	PHX
Average flight distance in miles	370	370
Average number of flights per year	8662	8650
Average annual passenger capacity	1219235	1210173
Average arrival delay in minutes	6	6

Method

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

```
SELECT origin AS org, dest,
round(avg(distance)) AS average_distance,
round(count(flight)/10) AS Average_flights_per_year,
round(sum(seats)/10) AS Average_Annual_seat_per_year,
round(avg(arr_delay)) AS Average_delay
FROM flights
LEFT OUTER JOIN planes
on flights.tailnum = planes.tailnum
WHERE 300 <= flights.distance AND flights.distance <= 400
GROUP BY origin, dest
HAVING Average_flights_per_year > 5000
ORDER BY average_distance DESC NULLS LAST
LIMIT 10;
```

Notes

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

I recommended the above route because of the following factors: -

1) *The average of seats per year is greater than any other routes.*
2) *There were ofcourse, other routes that were having more distance between them, but were falling short on seats.*
3) *Also, the average delay was considered before recommending the routes.*