Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Md Saiful Islam Sajol
Date: 22/02/20

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	131408	129694
Average arrival delay in minutes	10.32	13.76

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

SQL Code

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

```
SELECT origin, dest,AVG(distance), ROUND (COUNT(*)/10) AS AVG_FLIGHTS_PER_YR ,
COUNT(p.seats), ROUND(AVG(arr_delay),2)

FROM fly.flights f LEFT OUTER JOIN fly.planes p

ON f.tailnum= p.tailnum

 WHERE distance >300 AND distance < 400

GROUP BY origin, dest

 HAVING AVG_FLIGHTS_PER_YR > 4999

ORDER BY COUNT(p.seats) DESC ;
```

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

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.)