Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Bryce Quintana
Date: April 6, 2022

(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	LAX	SFO
Three-letter airport code for destination	SFO	LAX
Average flight distance in miles	337	337
Average number of flights per year	14540	14712
Average annual passenger capacity	1981058	1996597
Average arrival delay in minutes	13.76	10.32

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

Method

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

```
SELECT f.origin, f.dest,
round(AVG(f.distance), 0) AS avg_dist,
round(COUNT(*)/10, 0) AS avg_flights,
round(SUM(p.seats/10), 0) AS avg_capacity,
round(AVG(f.arr_delay), 2) AS avg_arr_delay

FROM flights AS f
LEFT OUTER JOIN planes AS p
ON f.tailnum = p.tailnum
```

```
WHERE f.distance >= 300 AND f.distance <= 400

GROUP BY f.origin, f.dest

HAVING avg_flights >= 5000

ORDER BY avg_capacity 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.)