Peer-Graded Assignment: Analyzing Big Data with SQL
Name: Satyake Bakshi
Date: 09/16/2021

(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	7054	6782
Average annual passenger capacity	1219234	1210173
Average arrival delay in minutes	308.26	308.26

(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:

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

```
SELECT origin,dest,avg(flights.arr_delay), round(sum(planes.seats)/10),
avg(distance),round((sum(flight/10)))
 as counts from fly.flights join planes on
 flights.tailnum=planes.tailnum group by
 origin,dest,planes.seats having counts>=5000
and avg(distance)>=300 and avg(distance)<=400
order by planes.seats desc, counts desc;
```

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.)