**Peer-Graded Assignment:** Analyzing Big Data with SQL
**Name:** DIPTESH TARAFDAR
**Date:** 07-01-2022

## 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** | 1996597 | 1981059 |
| **Average arrival delay in minutes** | 10 | 14 |

## Method

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

```
SELECT f.origin AS Origin,
    f.dest AS Destination,
    AVG(f.distance) AS Avg_distance,
    ROUND(COUNT(f.flight)/10) AS avg_no_of_flights,
    ROUND(SUM(p.seats)/10) AS avg_annual_seat_capacity,
    ROUND(AVG(f.arr_delay)) AS avg_delay
  FROM fly.flights f
  LEFT OUTER JOIN fly.planes p
  ON f.tailnum = p.tailnum
  WHERE f.distance BETWEEN 300 AND 400
  GROUP BY f.origin,f.dest
  HAVING avg_no_of_flights>=5000
  ORDER BY avg_annual_seat_capacity DESC
```

**LIMIT 10;**

## Notes

*I recommend the above route due to the following factors -*
*1) The average number of seats per year is greater than any other routes.*
*2) There were other routes with higher distance (such as the BOS ⇔DCA route) but they fell short on the  average yearly total number of seats.*
*3) Additionally, this route had the higher average arrival delays compared to other routes*