Perpay Data Scientist Candidate Problem Set

Revised: 2020-02

public.itinerary

row_id	customer_id	flight_id
1	1	1
2	2	2
3	1	5
4	3	2
5	4	3
6	2	8
7	6	9
8	7	10
9	8	10
10	8	11
11	9	9
12	9	12

public.customer

customer_id	customer_first_name	customer_last_name
1	Douglas	Hartree
2	Paul	Dirac
3	Marie	Curie
4	Rosalind	Franklin
5	NULL	NULL
6	Joseph-Loius	Lagrange
7	Hedy	Lamarr
8	Neils	Bohr
9	Margret	Hamilton

public.flight

flight_id	airline	destination	cancelled	actual_departure_time	actual_arrival_time	projected_departure_time	projected_arrival_time
1	Liberty Airlines	Philadelphia	FALSE	06/14/19 06:30 AM	06/14/19 08:35 AM	06/14/19 05:30 AM	06/14/19 07:30 AM
2	Salmon Air	Salt Lake City	FALSE	06/14/19 08:30 AM	06/14/19 12:20 PM	06/14/19 08:00 AM	06/14/19 12:00 PM
3	Costal Airlines	Portland	FALSE	06/14/19 09:30 AM	06/14/19 10:45 AM	06/14/19 09:30 AM	06/14/19 11:15 AM
5	Air National	New York	FALSE	06/14/19 08:30 AM	06/14/19 11:00 AM	06/14/19 08:30 AM	06/14/19 11:00 AM
8	Liberty Airlines	Geneva	FALSE	06/14/19 02:00 PM	06/14/19 04:00 PM	06/14/19 02:00 PM	06/14/19 04:00 PM
9	Costal Airlines	Los Alamos	FALSE	06/14/19 01:30 PM	06/14/19 06:00 PM	06/14/19 12:45 PM	06/14/19 05:00 PM
10	Eagle Airlines	London	TRUE	NULL	NULL	06/14/19 03:00 PM	06/14/19 05:00 PM
11	Eagle Airlines	London	FALSE	06/14/19 05:00 PM	06/14/19 07:50 PM	06/14/19 04:45 PM	06/14/19 08:00 PM
12	Salmon Air	Boston	FALSE	06/14/19 05:45 PM	06/14/19 10:00 PM	06/14/19 05:30 PM	06/14/19 09:30 PM

Question 1: SQL

Using the itinerary, customer, and flight tables, please provide a SQL statement and solution to the following problems.

1) What is the average number of customers per flight?
2) What is the final destination for each customer?
3) How many people missed their connections due to a previously late flight?

Question 2 Math & Statistics

a) You have 10,000+ variously sized polyhedrons, all plotted individually in a cartesian space. Discuss a possible methodology that could be used to remove duplicate polyhedrons and be reasonably assured you have a final set of unique structures.

b) You have an n-dimensional surface that represents the solution space of your problem. A dense sampling of this space is not computationally feasible. How can the global minimum of this surface be determined? How can you be sure you are not on a meta-stable state?

Question 3: Business Insight

You are hired as a data scientist by a general contractor that constructs townhome communities all over Pennsylvania. These communities are of various sizes as well as various price and quality.

a) What are some considerations to be made when determining a location for the construction of a new townhome community?

b) Incorporating these considerations, how would you approach designing a model that would output the current viability of constructing a townhome community at a location? How would you approach designing a forecast for the future demand of new townhome communities in Pennsylvania?

c) How could the contractor use these models to improve their business model?