Horizon DE Coding Assessment

By: Edwin Zamudio

I provided the final answer in the Excel sheet. This document contains some notes while completing the assessment!

Before starting the assessment, I did some exploring on the data to make sure we had some of my assumptions clear!

- All HCP's and HCO's are unique and there is no conflicting data

Some choices of packages that I used pandas and Duckdb to run efficient SQL queries on the data from the excel sheet. I could have used pandas transformation as well but I am familiar with SQL as I wrote a lot of it in my previous role!

Pgeocode was used to compute the distance between 2 zip codes. Requests was used to download the excel sheet.

Question 1: What is the number of distinct specialties per territory?

For this question, HCP's have specialties so I assumed that we wanted to use the HCP table. From here I used the zipcode for each HCP and joined it to the zip to ter table to get the territory for each HCP, then I could use count distinct and a group by to answer the question.

I could have used an INNER join instead of a left join since every zipcode has a corresponding zipcode!

Optimization:
If we don't want the exact number but a rough estimate if the data was huge, we could use approx_count_distinct or similar function.

Manual Validation:
Check that there is at least 6 unique values in the HCP table
Check one resulting row has the same territory for npi and ic
74105629 -> 08857 -> EAST
840809 -> 08859 -> EAST

Question 2: What is the straight-line distance (km) for each hcp with a RHEUMATOLOGY or NEPHROLOGY specialty to the closest IC (infusion center) in their territory?

For this question, I did some research on how to get the distance between 2 zipcodes. I initially wrote the distance function and calculation to question 2 using python but then I remembered I could introduce user defined functions into duckdb SQL engine!

For this question, we want to filter the data down for HCP's and HCO first which is what my CTE's are doing. From there I joined the 2 CTE's on territory, this would result in multiple HCO's and HCP's being join similar to a cross join, then reducing down the data to show the closest distance in km using a group by and MIN with the UDF i created to calculate the distance between 2 zipcodes!

Validation:
There are 99 rows returned so there should be 99 HCP with RHEUMATOLOGY or NEPHROLOGY.

Question 3: For each IC, what are the closest 3 HCP's with a RHEUMATOLOGY or NEPHROLOGY specialty in the ICs territory?

For this question, we could use the data from question 2 since it is already correctly filtered and we have the smallest distance between HCP and IC;s within the same territory. I added the closest IC column to question 2 so we could see the ID for the closest Infusion centers ID to a given HCP.

From here we can use the ROW_NUMBER window function to get the top 3 closest HCP's for each IC by grouping by infusion_center and filter on ROW_NUMBER < 3.

Question 4: All HCP's within 100 miles for each HOSP (hospital) only if they are not closer to another HOSP.

For this question, my first thoughts were to break this question down into 2 parts. The first part gets the closest hospital for every HCP.

To do this, I created a table that calculates the distance between an HCO and HCP using zip codes and external UDF that we created in question 2. This query takes a 3+ minutes to run since it uses a cross join and requires running python code so I stored it in a duckdb file.

Using this table, I created a CTE to get the closest hospital for each HCP using the min_by function and a filter on 100 km. From this CTE, I aggregated a list of every HCP that is closest to a given hospital.

Final Thoughts:

I did not do any validation on the pgeocode distance function!

Sources:
DuckDB documentation: https://duckdb.org/docs/
Pgeocode: https://pgeocode.readthedocs.io/en/latest/generated/pgeocode.GeoDistance.html
Stackoverflow