Data Engineering Project: Toronto Covid-19

Introduction:

This project was created as part of DataTalks' Data Engineering zoomcamp final assignment. The tools used are:

- Cloud: GCP
- IaC: Terraform
- Workflow Orchestration: Prefect
- Data Warehouse: BigQuery
- Transformation: DBT

I really enjoyed learning about data engineering as it has help me grow my programming skills as well as my knowledge about various tools. Working on this project helped me improve my debugging skills as well as self-learning from docs. I am thankful to the instructors of this zoomcamp as well as the Slack community for helping me whenver I got stuck. This was my third time trying out a zaamcamp by DataTalks and I am glad I was able to work till the end as the previous two times I had to drop out due to academic commitments. Thank you 😃

Problem Description:

I have chosen the COVID-19 cases in Toronto dataset for this project as the pandemic is still ongoing and the results from the analysis is something everyone can learn from to help each other. The dataset is updated on a weekly basis.

The first case of COVID-19 in Toronto was reported in January 2020 and since then the virus is been monitored along with its mutation and the kind of experience the patients of various demographics have gone through.
The focus of this project is to identify which groups had the most number of cases as well as to identify the most common source of infection and other information. The analysis performed can help in protecting the vulnerable groups as well as understand how to restrict the spread of the virus.

Data Description:

The final dataset after the transformation has the following columns:

- Assigned_ID: "A unique ID assigned to cases by Toronto Public Health for the purpose of posting to Open Data, to allow for tracking of specific cases." (Source)
- Age_Group: Age of the person at the time they got infected
- Client_Gender: Gender of the person reported by themselves
- Neighbourhood: Neighbourhoods in Toronto
- Postal_District: The first 3 characters of postal code
- Outbreak_Associated: Outbreaks associated with COVID-19
- Classification: Is the case confirmed to be COVID-19 case or it's just a possibility
  = Source: Source from where COVID-19 was possibly acquired
- Episode_Date: The earliest date the virus was acquired

- Reported Date: The date the case was reported on to Toronto Public Health
- Delay_in_Reporting: Number of days between Episode_Date and Reported_Date
- Outcome: Describes if the patient died, recovered or still has the virus
- Ever_Hospitalized: Cases that were hospitalized due to COVID-19
- Ever_in_ICU: Cases admitted to ICU due to the virus
- Ever_Intubated: Cases that had a tube inserted for ventilation due to COVID-19
  All the columns starting with "Ever_" include cases that are currently hospitalized, deceased or discharged. The Delay_in_Reporting column was not provided in the original dataset; it had to be created.

Replication: