AISE3010: Assignment 3

Arnav Goyal - 251244778

March 20, 2024

Objective 1 - Data Warehouse in GCP

The first objective was completed by following the instructions given on the powerpoint slides in *Course Material* on OWL. Here are some screenshots of the results. Specifically the exported .csv files, and the view for the regional managers question.

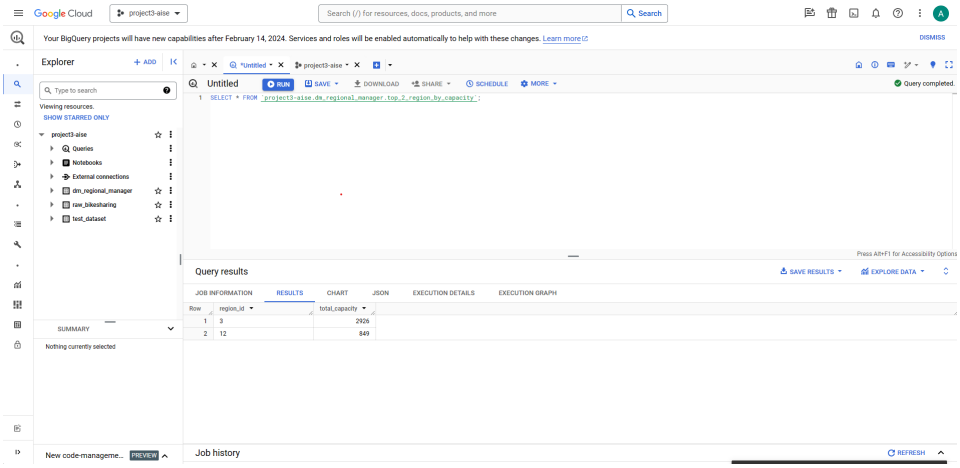

Image 1.1: Querying the created view

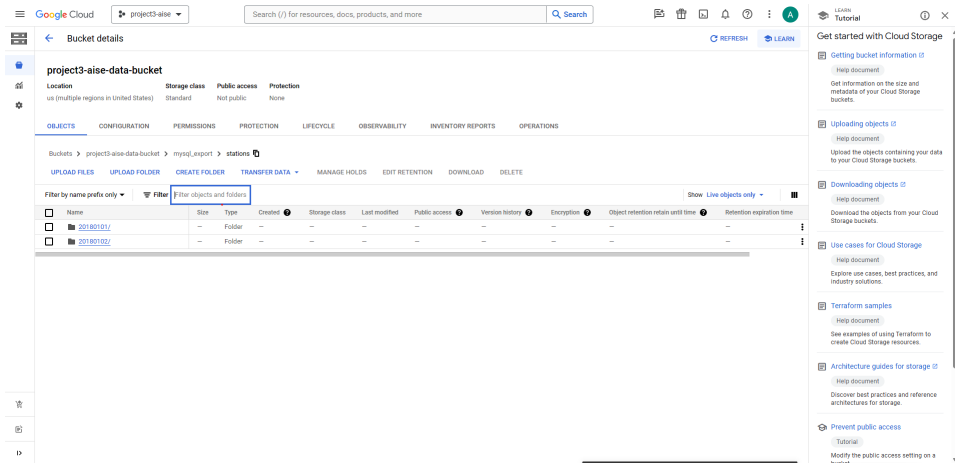

Image 1.2: The parent folders of the exported csv files

I'm unsure of what else to take screenshots of, please email *agoyal57@uwo.ca* if I need to provide any other screenshots of my work for this part.

Objective 2 - Using ML in GCP

Setting up Data Warehousing

We can access the *ulb_fraud_dataset* directly within BigQuery by accessing the following table, our SQL statement would have to look like this, Given the statement below.

```
SELECT *
FROM `bigquery-public-data.ml_datasets.ulb_fraud_detection`
LIMIT 1000;
```

We can then select the option to take these results into *Looker Studio* to perform EDA on this data, using the most un-intuitive UI I have ever had the pleasure of using! I couldnt figure out how to plot a simple histogram, thus I decided to use the other option of *Google Sheets* to plot it. We can select a column, head over to column stats, scroll down and select 'Distribution' to view the approximate distribution of values within the column. It is easy to do all of this for each column thus I'm only attaching a screenshot of the distribution from the V1 variable. The column stats for the 'Class' variable also tells us it is a categorical variable with unique values of 0 or 1.

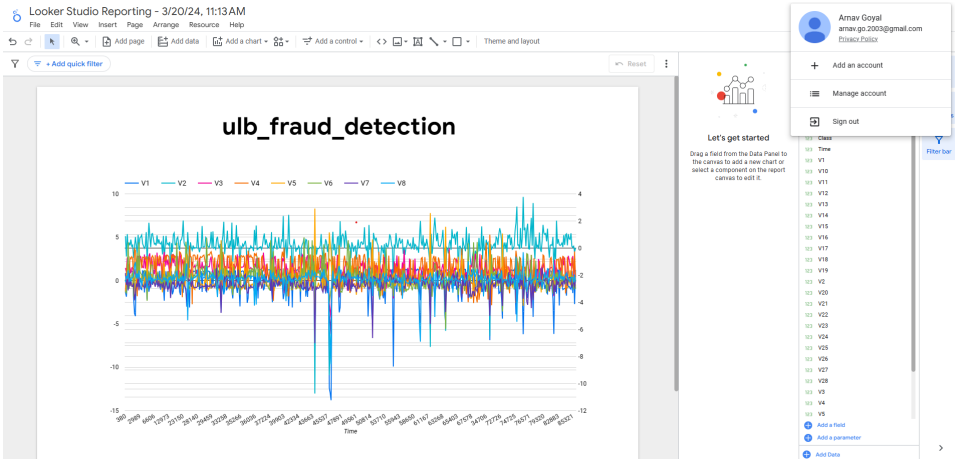

Image 2.1: Plotting V1 - V8 vs Time.

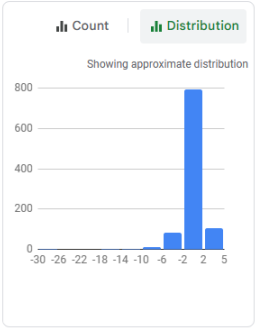

Image 2.2: Approximate Distribution of V1

Data Preprocessing & Feature Engineering

It is time to perform feature 'engineering', by picking random features (columns) from the raw dataset and saving it as a .csv file in our data bucket. This is the Python script I have run to randomly choose features from the columns available in our dataset.

```
import numpy as np

# generate a list w/ all the columns
cols = ['Time', 'Amount', 'Class']
for i in range(1,29): # 1 to 28
        colname = "V"+str(i)
        cols.append(colname)

# pick 5 random cols
chosencols = np.random.choice(cols, 5, replace=False)
```

The chosen columns end up being:

- V23
- V7
- V28
- V25
- V1

Thus we can use the following SQL Command to select these random features (columns) above, and save them to a csv and upload them manually to our data-bucket! We can also go through the process of creating a CloudSQL instance, and querying it and then deleting it. But that takes MUCH longer and costs credits.

Essentially, Run the query, then save it to Drive, then upload it to the data-bucket. I also saved this as a BigQuery Table called 'dataset_random_features' for easy access later

```
SELECT V23, V7, V28, V25, V1
FROM `bigquery-public-data.ml_datasets.ulb_fraud_detection`
ORDER BY Time;
```

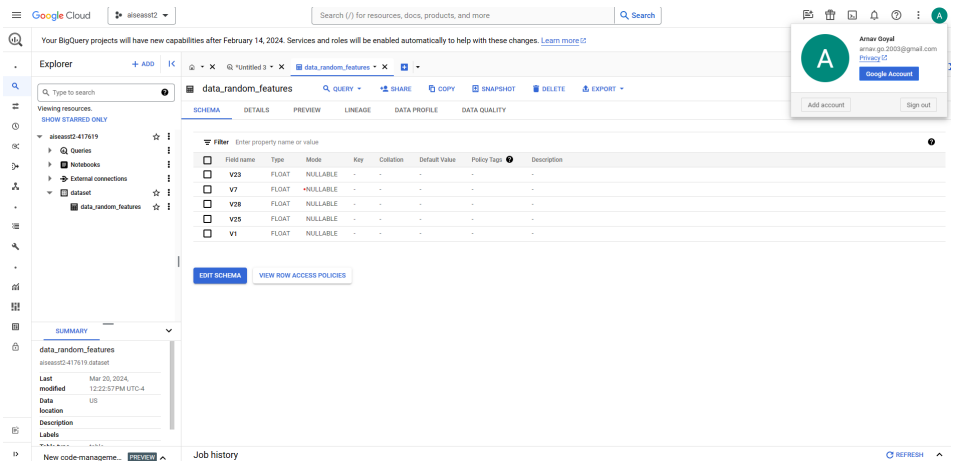

Image 2.3: The resultant table in BigQuery

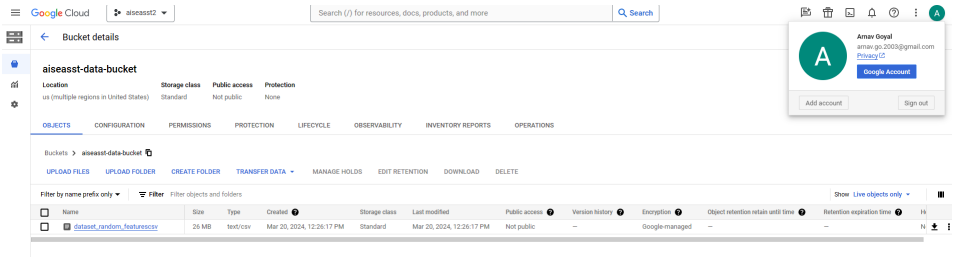

Image 2.4: The csv file in GCS