Greeshma Sai Sree Duggireddy
grdugg

AIR TRAFFIC PASSENGER STATISTICS WITH BIG DATA CONCEPTS AND VISUALIZATIONS

Course Project

INTRODUCTION

This hands-on course project for big data is based on the dataset – 'Air Traffic Passenger Statistics'. I downloaded this dataset from the suggested website – data.gov. The data.gov is a website that has data directly obtained from trusted sources. Air traffic passenger statistics are obtained directly from the San Francisco International Airport. This dataset consists of data from 2005 to June 2022.

For this project, I want to implement a data pipeline where I will download, transform, summarize, and visualize the data. To do this, I will use the Google Cloud Platform - GCP and the tableau desktop app.

BACKGROUND

San Francisco is a city in California that is a commercial, financial and cultural center for Northern California. Moreover, San Francisco has close proximity to Silicon Valley which is considered the giant tech center of the world, which automatically makes it one of the wealthiest cities in the United States [2]. For these reasons, I wanted to analyze the air traffic statistics of the San Francisco Airport from the year 2005 to the present. Since the year 2005, there have been a lot of technological developments, but we have also encountered a global pandemic. To understand and analyze these effects on airplane passengers, this data has been chosen.

The dataset consists of the attributes - Activity_Period, Operating_Airline, Operating_Airline_IATA_Code, Published_Airline, Published_Airline_IATA_Code, GEO_Summary, GEO_Region, Activity_Type_Code, Price_Category_Code, Terminal, Boarding_Area, and Passenger_Count. There are 24968 entries of data. The timeline of the data is from January 2005 to June 2022[1].

I have downloaded this dataset from the website - data.gov, uploaded it to the GCP platform, analyzed and summarized the data, and finally visualized it on the tableau desktop.

METHODOLOGY

I searched for datasets in data.gov to fit my data pipeline model. I looked at a lot of datasets, and I felt that there is a purpose for this dataset and that this dataset would fit my outlined model. After I downloaded the dataset, I started to set up my environment for implementing the pipeline. The first step was to retrieve the coupon code that my professor gave us access to. I retrieved the

coupon using my IU mail and I successfully logged in to my GCP. After this, I downloaded the tableau desktop with my IU mail to visualize the data.

I created a project named - 'Project Grdugg'. I wanted to store the data that is downloaded in the google cloud platform. In one of the course module assignments in GCP, I stored the data in a GCP bucket. I did the same to store my dataset as well. In my project, I clicked on the navigation menu and selected cloud storage > Buckets. I clicked on 'CREATE BUCKET' with my customized options. After I created the bucket named - 'airlinesdatagrdugg', I uploaded the dataset - 'Air Traffic Passenger Statistics' onto the bucket. I can now access the dataset to perform any operations within GCP.

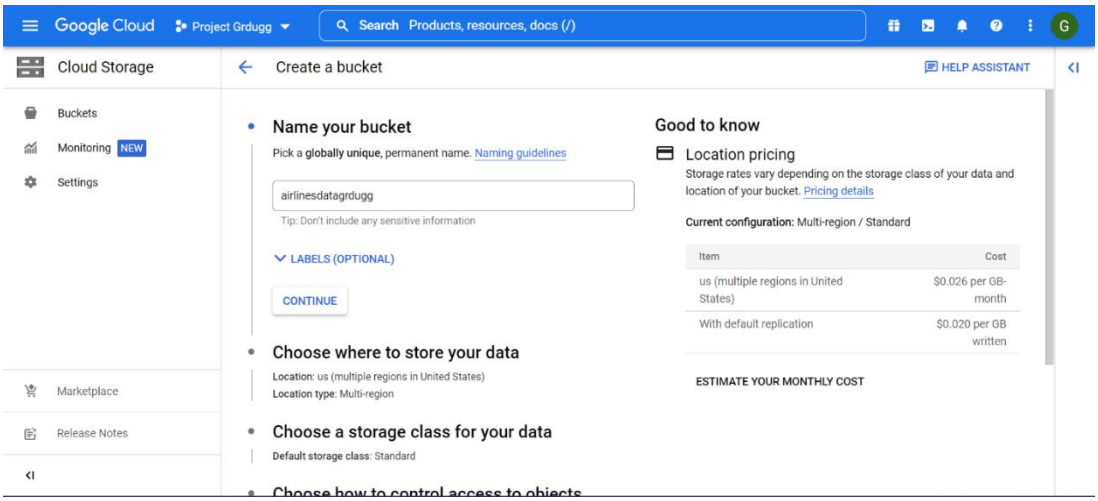

Fig 1: The creation of the bucket

Now that I can access my airlines dataset throughout GCP, I will now use BigQuery feature in GCP to work with SQL. From the Navigation menu, I selected on BigQuery and then on SQL workspace.

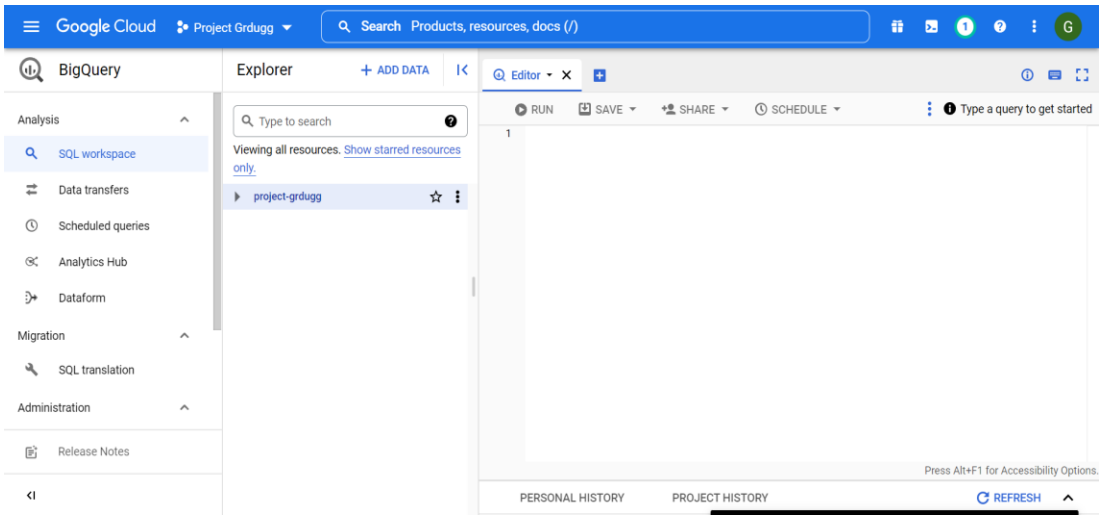

Fig 2: The SQL workspace in BigQuery

From the above screenshot, we can see the SQL workspace in BigQuery. Through the 'ADD DATA' option, I will connect my GCP bucket and BigQuery. I uploaded the dataset that is in the GCP Bucket on to the BigQuery as a dataset table to execute SQL Queries on it.

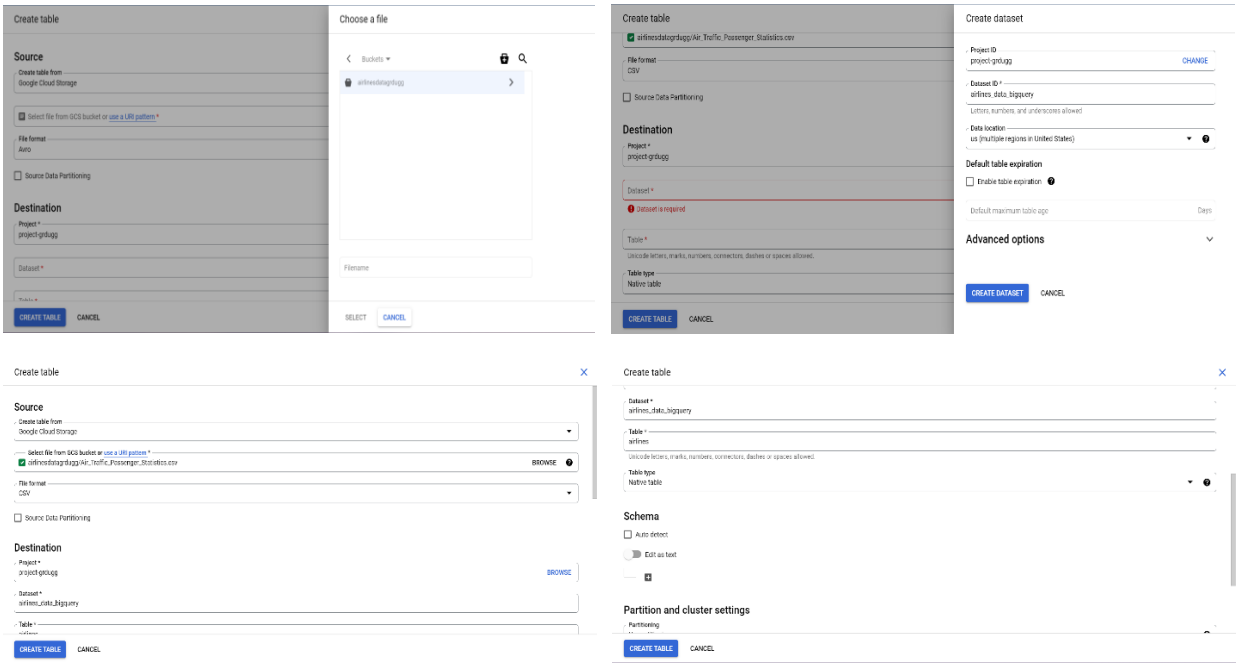

Fig 3: The above screenshots correspond to connecting the data in the bucket to BigQuery in GCP

The above screenshots correspond to the process of creating a dataset table in BigQuery from the GCP bucket. After the creation of the table, I checked if I could access the data.

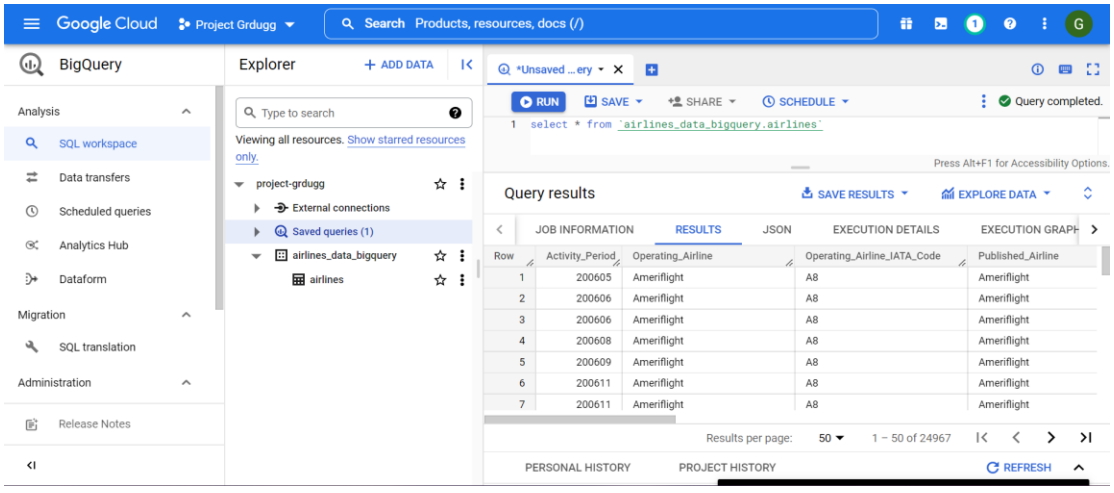

Fig 4: Airlines Table data in BigQuery

With the help of SQL Queries, I cleaned the data. I removed the rows which had null values so that data quality in ensured.

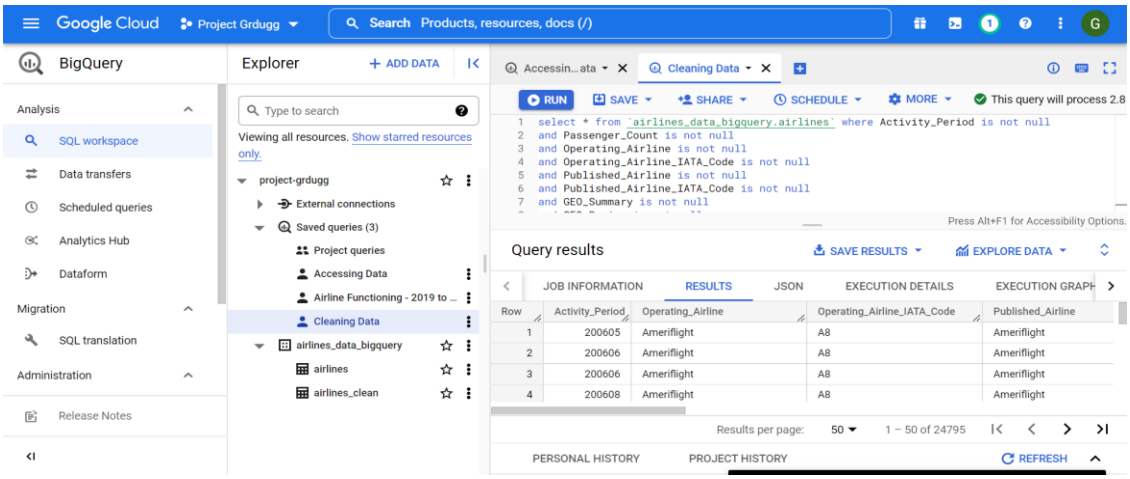

Fig 5: Cleaning the Data

I wrote a couple of other SQL Queries to analyze the data. After executing the query, I saved the results as a BigQuery Table. For visualization, I connected Tableau Desktop with BigQuery Server. After connecting the data for tableau from BigQuery, I visualized the data to get clear insights from it.

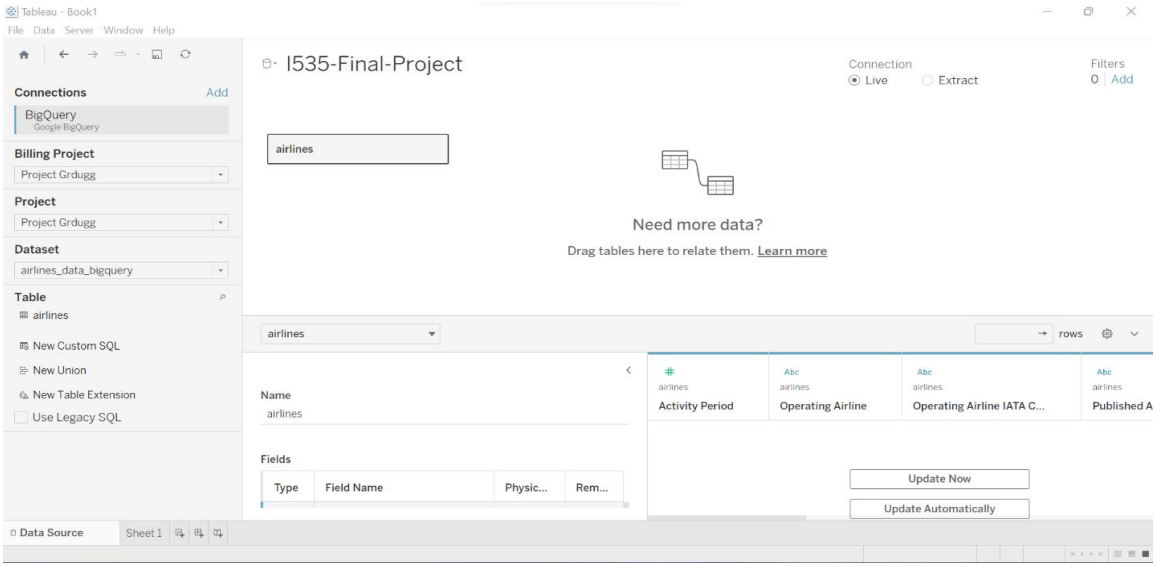

Fig 6: Connected Tableau with BigQuery Server

RESULTS

I wanted to understand the trends of airlines in two periods – one is when the tech industry in Silicon Valley was growing and the other is when the global pandemic was hit. To understand these trends, I wrote SQL queries to summarize and segregate this data.

In my first query, I wanted data grouped by domestic and international terminals. The timeframe that I considered here is from July 2005 to December 2018. I wanted to check trends right before a year of the global pandemic when all circumstances were well.

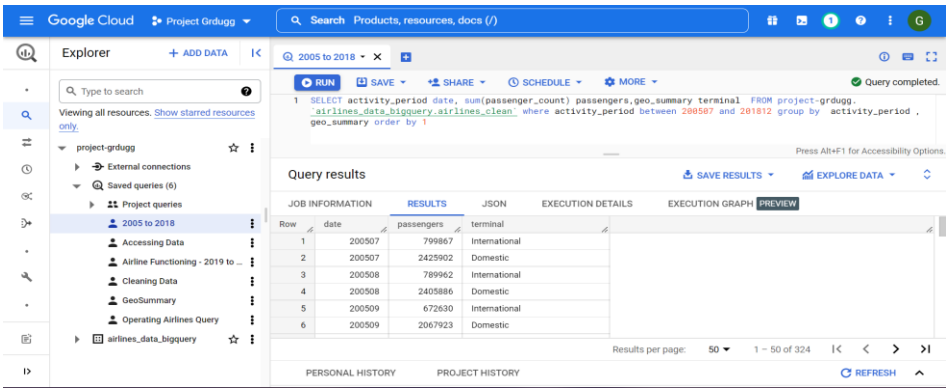

Fig 7: Query – Airline Functioning from July 2005 to December 2018

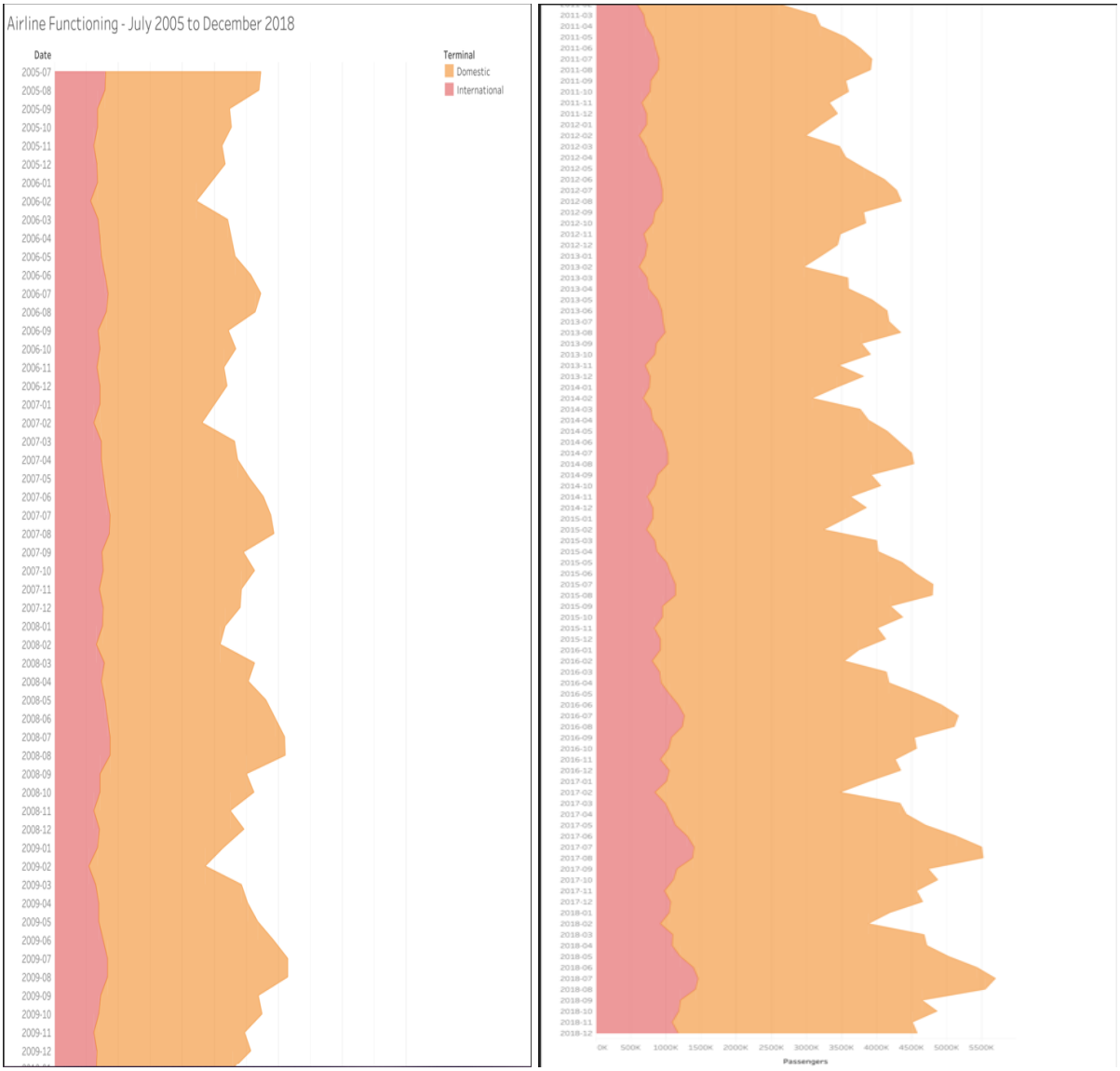

Fig 8: Airline Functioning from July 2005 to December 2018

In my second query, the timeframe that I considered here is from January 2019 to June 2022. I wanted to analyze the airline trends during the time of COVID-19.

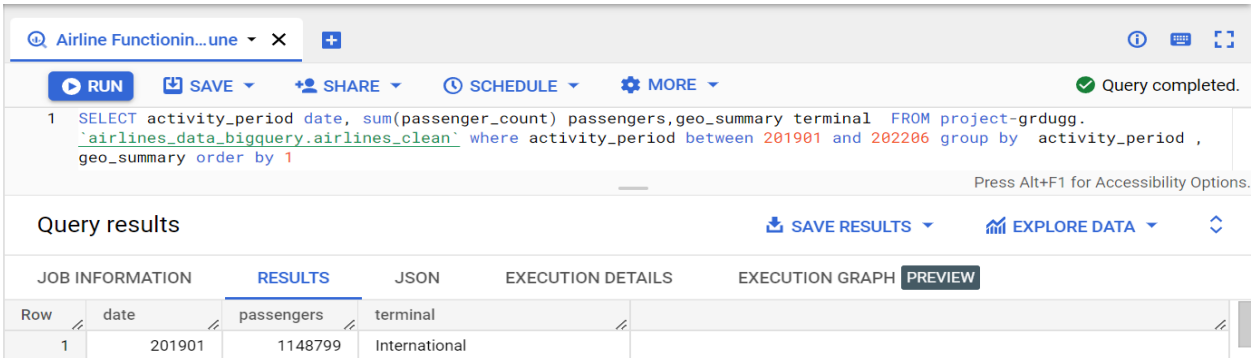

Fig 9: Query - Airline Functioning from January 2019 to June 2022

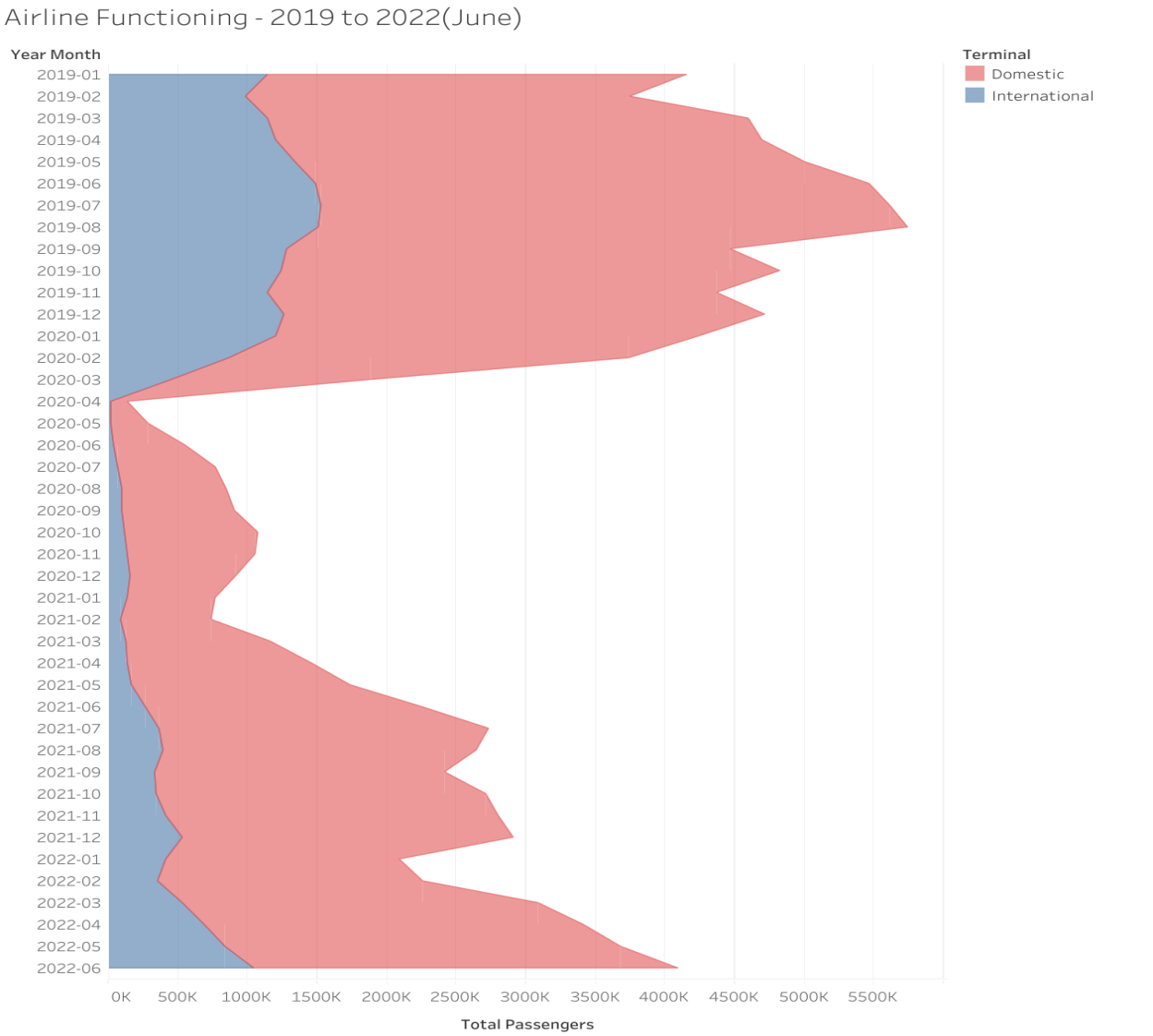

Fig 10: Airline Functioning from January 2019 to June 2022

In my third query, the timeframe that I considered here is from January 2019 to June 2022. I wanted to analyze the operating airlines' trends during the time of COVID-19.

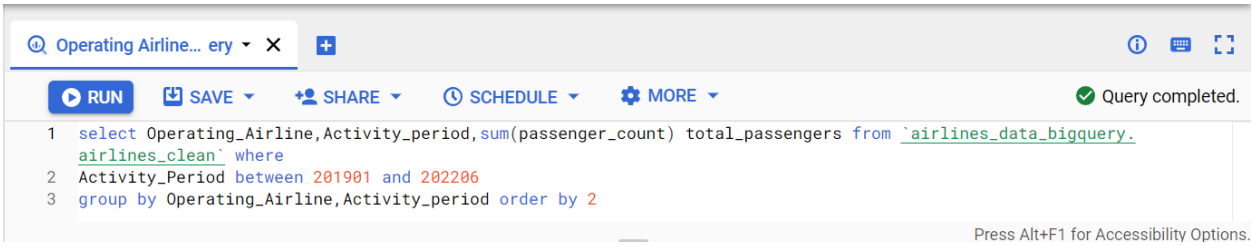

Fig 11: Query - Operating Airlines from January 2019 to June 2022

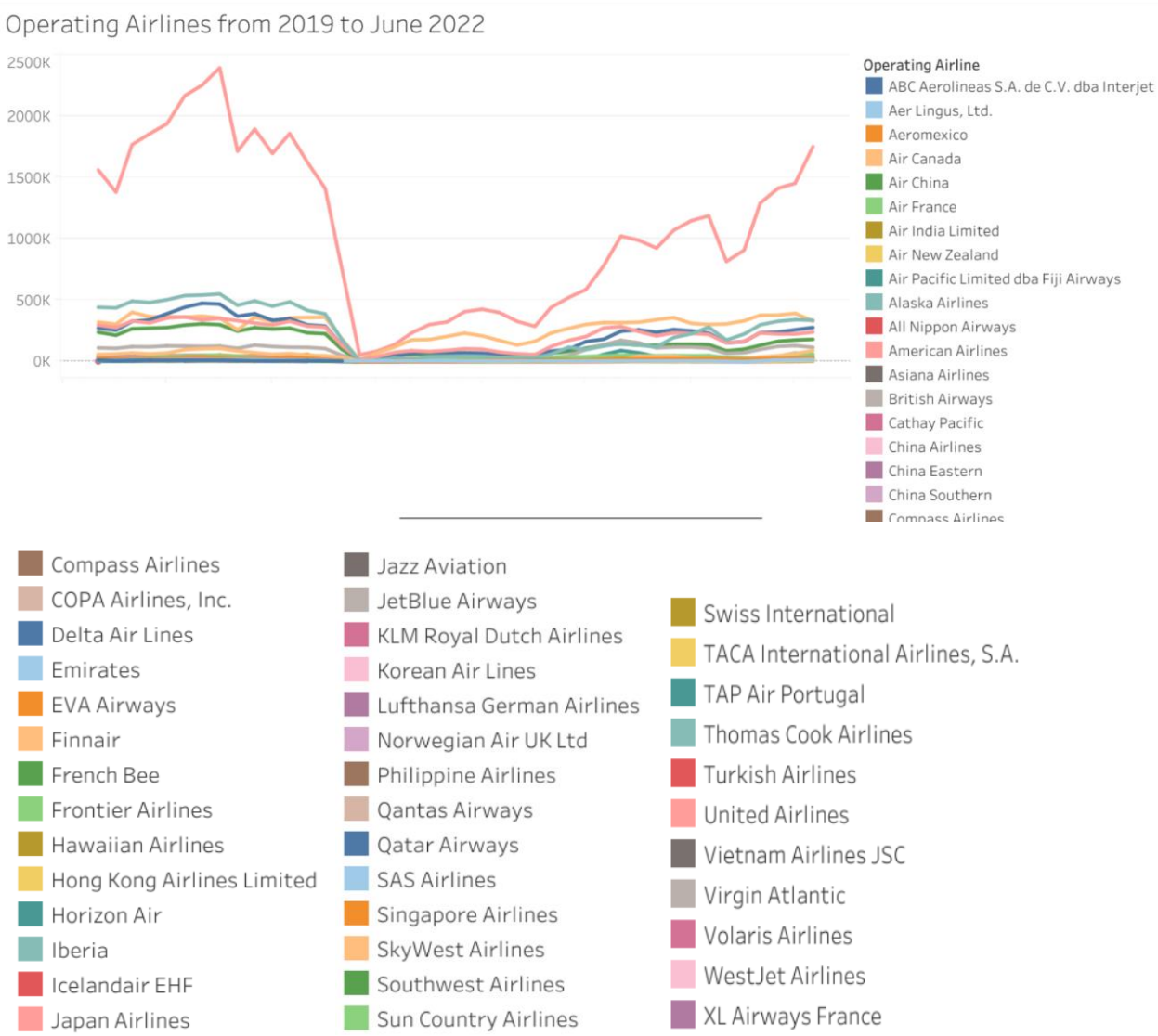

Fig 12: Operating Airlines from January 2019 to June 2022

In my fourth query, the timeframe that I considered here is from July 2005 to June 2022. I wanted to analyze the trends of total passengers and arrivals to different parts of the world from San Francisco.

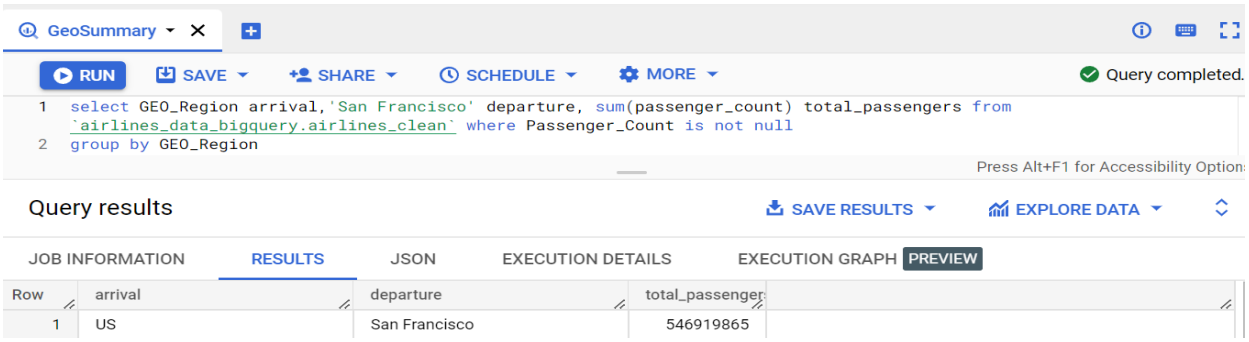

Fig 13: Query - Transit of passengers through San Francisco

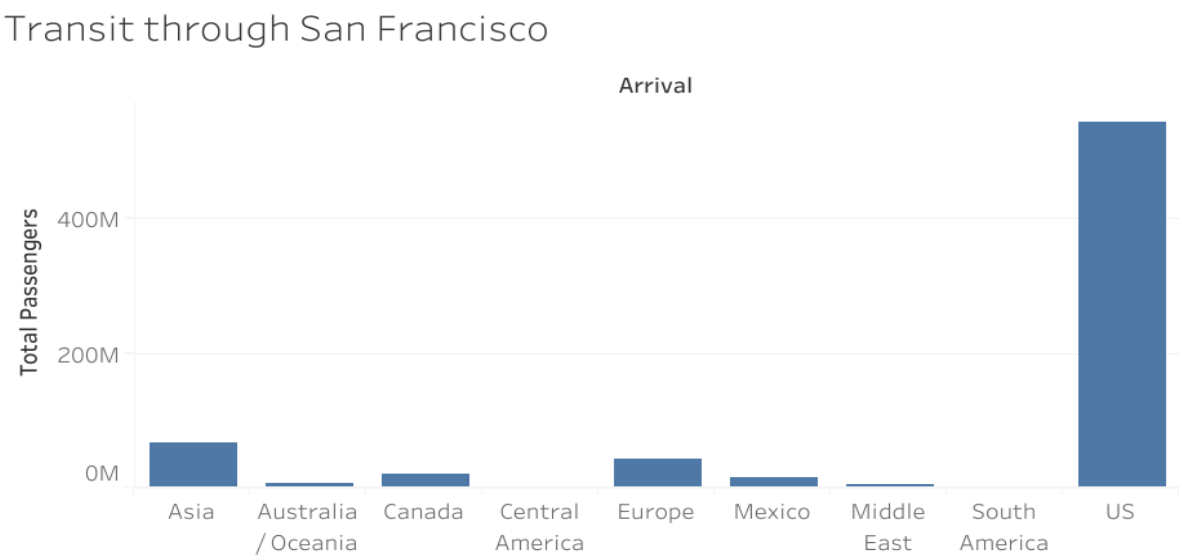

Fig 14: Transit of passengers through San Francisco including US

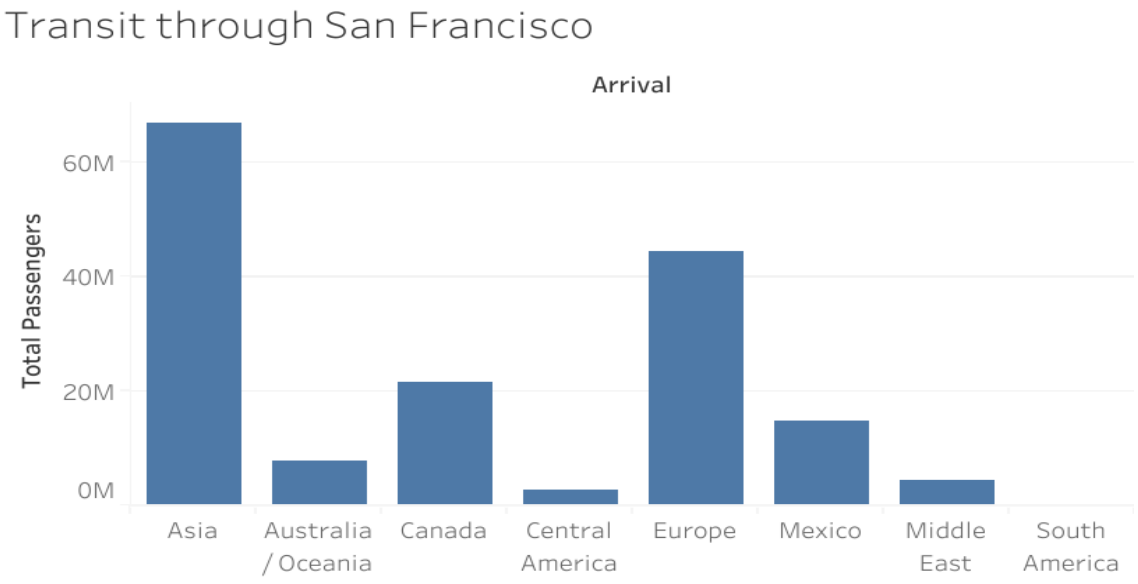

Fig 15: Transit of passengers through San Francisco excluding US

DISCUSSION

INTERPRETATION OF RESULTS:

I have visualized the parts of air traffic passengers' statistics. After analyzing the graphs and visualizations, I have interpreted the following:

In the Fig 8, the number of passengers flown on domestic and foreign flights over time is depicted in this area graph. The higher the area on this area graph, the more passengers there were. We can observe that the number of domestic and international air journeys has been steadily rising over time. When compared to international flights, there have never been more passengers on domestic flights.

In Fig 10, the area graph displays the number of passengers flown on domestic and international carriers concurrently. We can observe that between 2020 March and 2020 April, there was a significant decrease in both local and international flights due to the global pandemic. An all-time low was recorded in April 2020. It was the time when the COID-19 pandemic was newly emerging. Passengers on domestic flights have dominated over those on foreign flights. Overall, it is evident from this graph and the one above that there was a significant influence on airlines throughout this period. And it is quickly recovering.

In Fig 12, the line graph shows the various airline services from 2019 to 2022. It is clear that customers are more inclined to fly with United airlines. Additionally, some airlines, including Aer Lingus and ABc, have not yet recovered from the effects of COVID. We can see that all airlines experienced all-time lows in April 2020.

In Fig 14 and 15, the bar graphs show the total number of passengers and the countries they traveled to and from San Francisco airport over the years. We can see that majority of the passengers traveled within the country – the US. The international travels were most frequently from Asia, Europe, and Canada. It is evident that there are almost no passengers traveling to or from South America.

EMPLOYED TECHNOLOGIES AND SKILLS FROM THE COURSE:

The skills and tools that I learned from this course, were of great help in executing this project. In the class assignments, I learned about the GCP features like storing the data in buckets and the usage of the BigQuery tool. I have used these concepts as key execution elements in my project. With the knowledge of data lifecycles and pipelines, I built a data pipeline model for my project. The data pipelines' detailed steps are outlined below.

The Data Pipeline outlined for the project –

1. Download: The dataset is downloaded from the data source – 'data.gov'. After downloading it, the data is stored in a GCP Bucket to use across all GCP Features.

2. Transform: The dataset in the Bucket is then connected with BigQuery and transformed into a Table to perform data processing on it.
3. Quality: After transforming the dataset into an SQL table, the data is checked for quality. If there are any null values in the dataset, they are removed. Through SQL queries, the data is cleaned, and the results are stored as a separate SQL table in BigQuery.
4. Summarize: After performing all the data preprocessing steps, the data is summarized based on the requirements of the project. SQL queries can be executed for the desired output. These outputs can again be saved as separate tables in BigQuery.
5. Visualize: For visualizing the data, Tableau desktop is used. Now that we have all the data saved in BigQuery as dataset tables, Tableau desktop is connected to Google BigQuery through its 'connect to data' feature. Tableau can access all the datasets available in our project – BigQuery. With these datasets, visualizations are obtained.

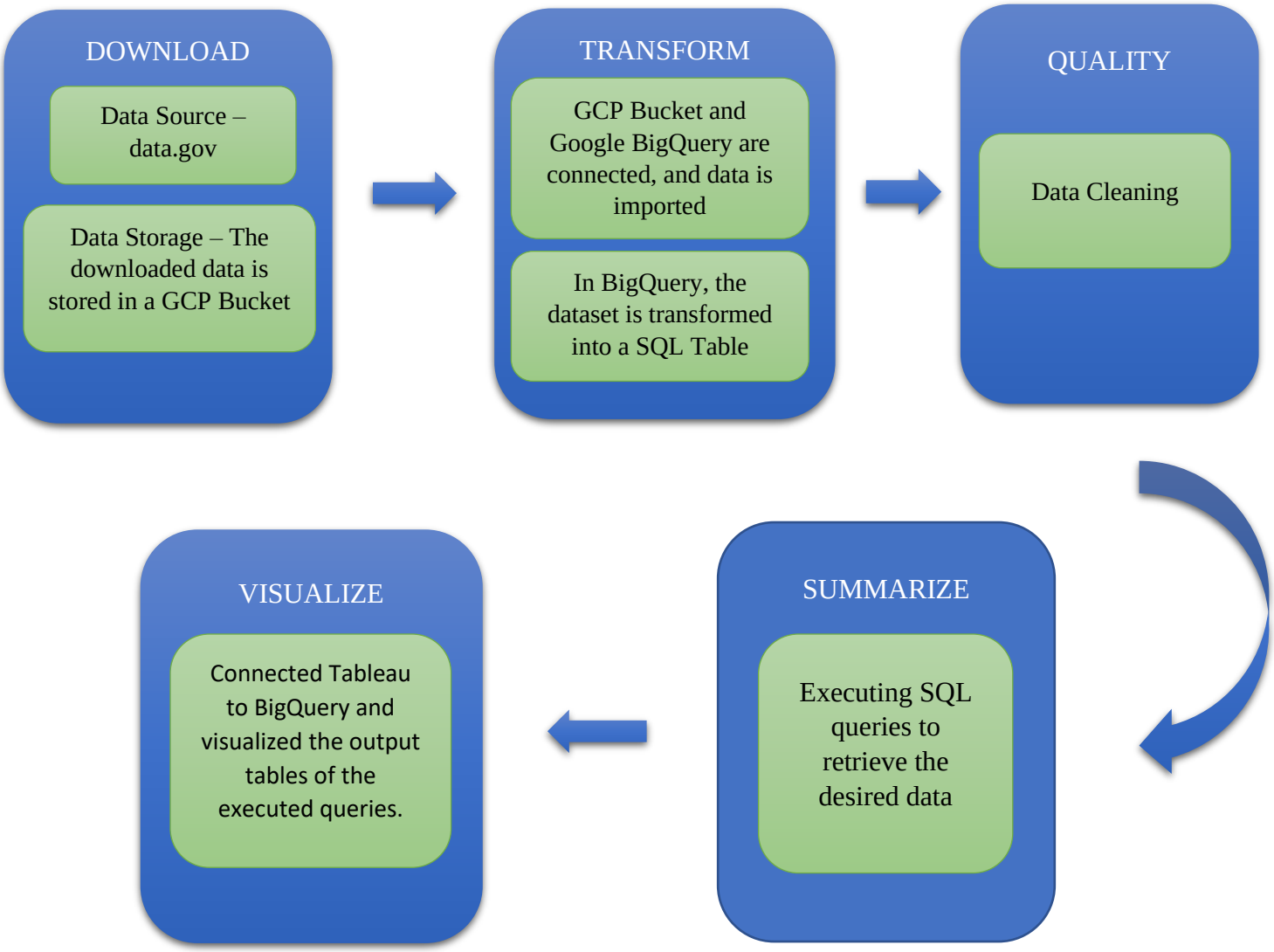

Fig 16: Data Pipeline Model (*drew in word)

BARRIERS AND FAILURES:

I had a hard time choosing the dataset for my project as I did not see the other datasets fit my pipeline model. While working on GCP, my initial plan was to execute the project using PostgreSQL, but I faced difficulties while importing datasets. Therefore, I changed the plan and executed it in BigQuery which was taught in the class assignments.

CONCLUSION:

I have implemented the data pipeline successfully. I have used GCP and Tableau tools to execute this project. I have executed various concepts that I learned from the course in the project. The building of a data pipeline and the usage of GCP buckets, Google BigQuery, and Tableau are the concepts that I used in this project. The queries and visualizations have helped me understand and interpret the data trends over the years. As mentioned in the background, a part of the airline traffic in San Francisco airport is because of the tech industry residing close to the city. Although the global pandemic has caused a dent in airline traffic, the operations of airlines are getting back into the business.

REFERENCES:

[1] "Air Traffic Passenger Statistics." Catalog, Publisher Data.sfgov.org, 4 Nov. 2022, https://catalog.data.gov/dataset/air-traffic-passenger-statistics.

[2] "San Francisco." Wikipedia, Wikimedia Foundation, 21 Nov. 2022, https://en.wikipedia.org/wiki/San_Francisco.

[3] Canvas Material.

[4] Qwiklabs. "Rent-A-VM to Process Earthquake Data: Google Cloud Skills Boost." Qwiklabs, https://www.cloudskillsboost.google/focuses/1846?catalog_rank=%7B%22rank%22%3A1%2C%22num_filters%22%3A0%2C%22has_search%22%3Atrue%7D&parent=catalog&search_id=7008005.

[5] Qwiklabs. "Ingesting New Datasets into Bigquery: Google Cloud Skills Boost." Qwiklabs, https://www.cloudskillsboost.google/focuses/3692?catalog_rank=%7B%22rank%22%3A5%2C%22num_filters%22%3A0%2C%22has_search%22%3Atrue%7D&parent=catalog&search_id=14163071.