ETL Project

Ana Saavedra | Analiza Asuncio | Avantika Singh | Eric Wetzel | Sahar Alaei | Shilpa Nagendra

**Extraction**–

Dates: 12/31/2019 to 5/31/2020

Obtained Data from:

1. <https://opensky-network.org/community/blog/item/6-opensky-covid-19-flight-dataset>
2. <https://ourairports.com/data/>
3. <https://ourworldindata.org/coronavirus-source-data>

We wanted to analyze flight airline data from before the closures of the pandemic and the effects while full quarantine was enforced. We found data that supported this theory not just here in the US but throughout the globe. The information was extensive and not necessarily legible as raw data. Which was great for the objectives of the project are. We layered in the COVID data for an additional angle of the data analysis.

**Transformation–**

**Data I**

As a group, we decided to remove the following data using Python and Jupyter Notebook. We felt we would not need the information for future projects, or the information in the columns did not make sense for making tables in SQL.

Removed columns:

- registration – number - type code – day

We then merged all of the clean (5) CSV files into one to then import into PostgreSQL. This information will take the database information to make the appropriate tables for the specific information that we are seeking.

A screenshot of a social media post

Description automatically generated

Figure 1. Jupiter notebook of flight data after cleaning using Python.

**Data II**

The second data set provides more information for the airports. This gives us additional information that the first set of data did not contain. We can later use this data for layering visuals like heat or weather maps. The data includes Lat, Long, City, and State and Country.

Removed columns:

A screenshot of a social media post

Description automatically generated- elevation - schedule - gps code - iata code - local\_code - home\_link - wikipedia\_link – keyword

Figure 2. Jupiter notebook of airport data after cleaning using Python.

**Data III**

A screenshot of a social media post

Description automatically generatedThis is for our COVID data. Due to the large number of columns the raw data, we just coded what information we wanted and extracted it to new and clean CSV to import into PostgreSQL. This includes a renaming function to those columns to be able to merge in with other data in PostgreSQL effectively

Figure . Jupiter notebook of COVID data after cleaning using Python.

**Load –**

**A screenshot of a cell phone

Description automatically generated**As a group, we made tables for each of the CVS that we acquired for our project. Later, we can query or join for what specifics we are looking for in the data.

Figure 4. Flight Data Table

A screenshot of a cell phone

Description automatically generated

Figure 5. Airport Data Table

**A screenshot of a cell phone

Description automatically generated**

Figure . COVID Data Table

**Summary –**

With the airport data and the flight data, we plan on joining them on the column origin. These are the queries we decided as a group to extract from the tables using PostgreSQL. By keeping a large portion of the original raw data, we will be able to obtain more precise information in the next part of the project.

1. Which airlines are part of the dataset?
2. What was the busiest day?
3. What was the slowest day?
4. What was the most popular route?
5. What was the least popular route?
6. What were the top 10 busiest airports?

We had several challenges while working on this group project. We had a hard time sharing the large files thru our GitHub repositories. There are limits to how large the data can be.