Bennett Jackson
ISTA 322

Final Project Report

For this final project, my overall goal was to examine the relationship between COVID-19 cases and domestic flights. To do this I used five CSV files and a REST API to get all of the data needed to analyze this relationship. Looking at the CSV files, four of them contained COVID-19 data. This data included the date, county, state, fips, cases, and deaths reported from COVID-19. The other CSV that was used was a CSV that contained airport data from around the globe. This data included an ID, an identification number, a type of airport, a name, longitude, and latitude, the elevation of the airport, the continent, the iso country, the iso region, the municipality, the service schedule, the GPS code, the IATA code, the local code, a home link, a Wikipedia link, and keywords. Out of all the information that was present in this CSV not much of it was relevant to my project. Finally, to get the actual flight data from historic flights the OpenSky Network REST API was used. The COVID-19 data from the New York Times provided a cumulative number of cases and deaths per county, per day. The airport data taken from OurAirports provided the relevant airport codes needed to determine where the flight's departure and arrival were from. Finally, the OpenSky Networks API provided information on the departure and arrival location of a flight as well as a callsign for the plane and other flight-related statistics that were not relevant to the project. For this project, I decided to host all of the CSV files in my Google Drive to ensure they were stable. Below are the links to where I found the data, to see the raw CSV files those links are in the notebook that was submitted.

Links:
OpenSky Network: https://opensky-network.org/
New York Times COVID data: https://github.com/nytimes/covid-19-data/tree/master
OurAirports data: https://ourairports.com/data/

Throughout this project, some challenges arose when cleaning the data. The first challenge was the size of the data. To ensure that accurate assumptions were made, all of the COVID-19 data over the years were needed. Additionally, another challenge with cleaning the airport data was determining what data was important and what data wasn't important since there were very technical pieces of data present in the CSV. The ultimate challenge when it came to this project was the REST API. Since OpenSky Network was the only free API that would offer historical data it needed to be used, but the request took a very long time and a time frame of only two hours could be in each request. This meant that getting all of the flight data for every day in the COVID-19 data would take almost 15 hours. This leads to only focusing on the morning flight rush from the hours of 6 AM to 10 AM as this is when there is a peak in flights during days. Additionally, only a few selected days were analyzed for their flight data to help reduce the time it takes to run the notebook. To try and speed up the process the two API requests for each were parallelized. This is a flaw with the API not being optimized but it had to be used since it was the only one available.

When integrating the data, five different data frames were created that all help separate information. The first data frame created was to hold all of the aggregated COVID-19 data. The

second data frame held each state's ID number and abbreviation. The third data frame held all of the data that related to airport codes. The fourth data frame held the date and total number of domestic US flights. Finally, the fifth data frame held the specific flight information that was pulled from the API. Overall, all of these data frames are related to each other so queries could be run on them. In Figure 1 below the database schema can be seen with all of its relations. The red star in the schema is the primary key in the table and the blue arrows are the relationships between the tables.

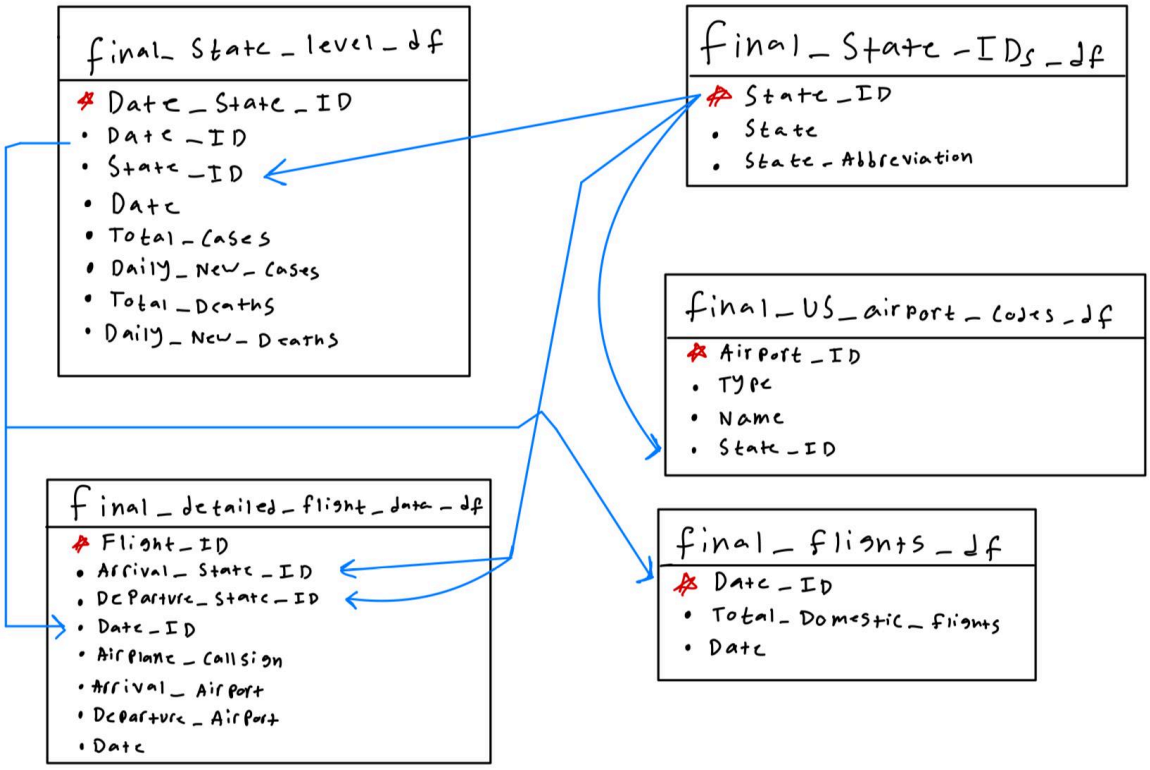
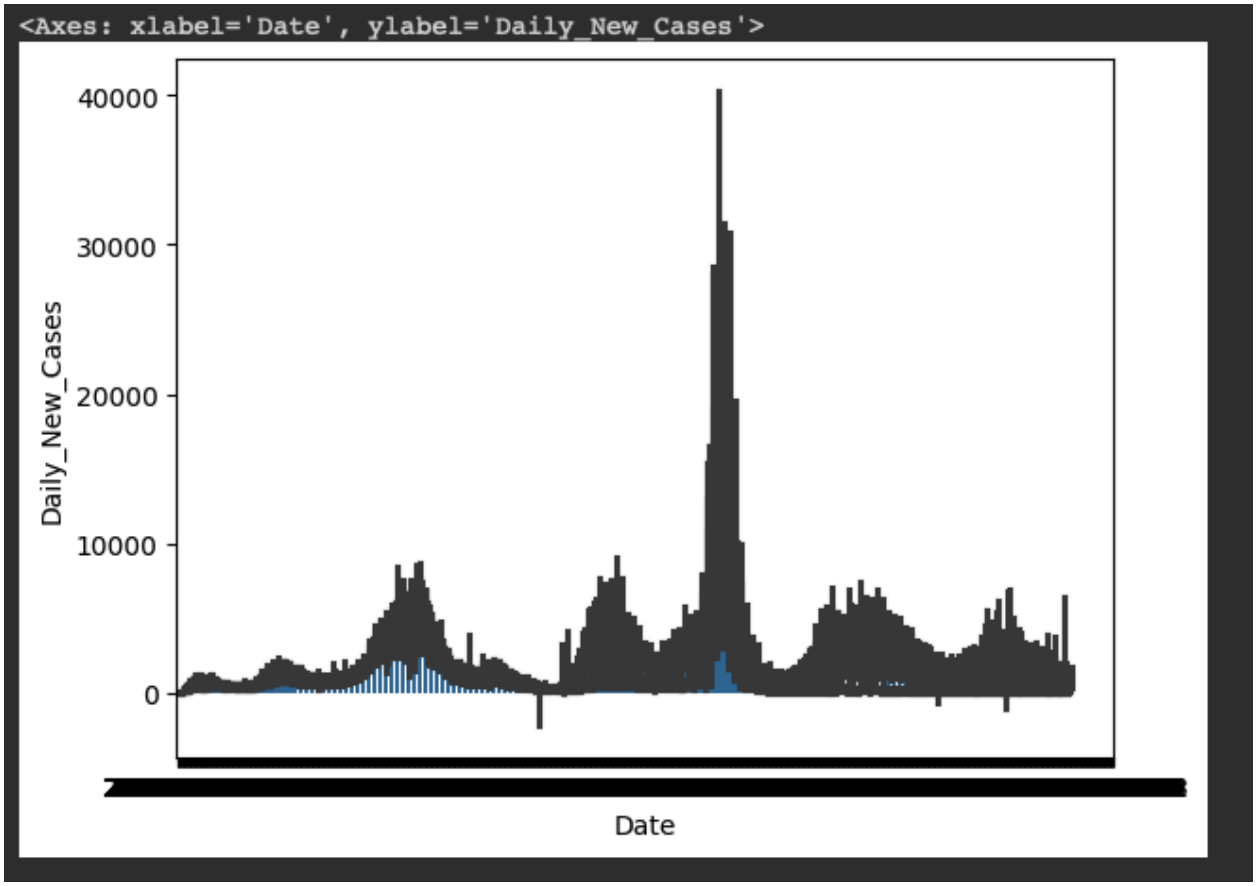
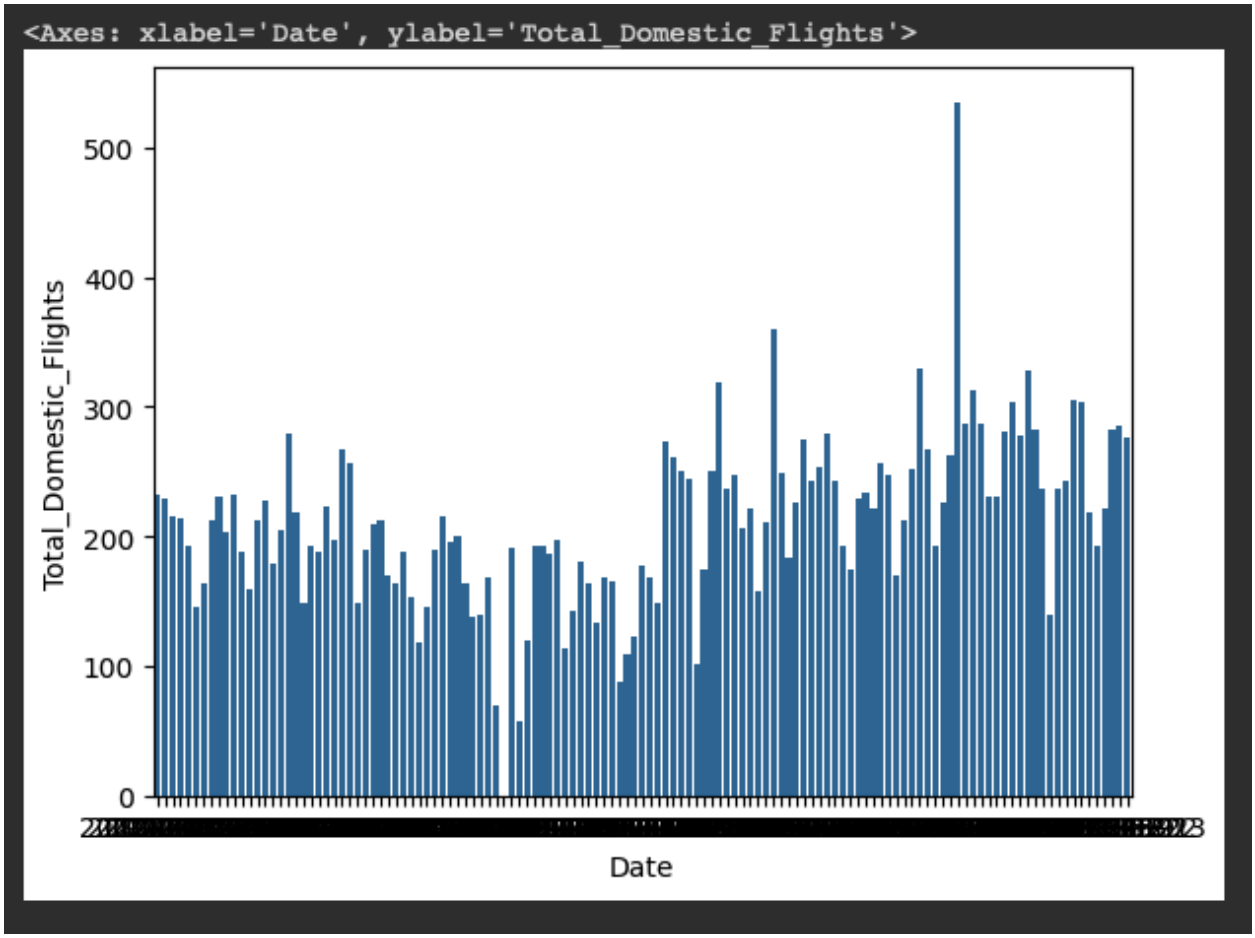


Figure 1. The database schema for this final project.

The main goal of this notebook was to determine if there was any relation between COVID-19 cases and the number of domestic flights. Looking at the two graphs below, even though there was incomplete data it appears that there is a relationship and that the overall flights did decrease during COVID-19. Graph 1 shows the total number of new COVID-19 cases daily. Graph 2 displays the number of flights per day.



Graph 1. Daily new COVID cases



Graph 2. Daily domestic flights

Moving onto the SQL queries that were run to make inferences about the data, five different queries were made. Below will be an image of each query and the explanation that goes with the query.

In this query it is looking at the total number of flights per state compared to the total number of COVID cases per state. This shows the use case of determining which states didn't care as much about covid and which states cared about COVID. This query states that California had the most number of flights where main had the least amount of flights. This can also be seen in the vast difference between the total number of COVID cases. This shows that Even though the COVID cases kept on rising the flight volume didn't change much.

```
query = """SELECT s.State, COUNT(DISTINCT f.Flight_ID) AS Total_Flights, SUM(c.Total_Cases) AS Total_Cases
FROM final_detailed_flight_data_df f
JOIN final_US_airport_codes_df a ON f.Arrival_Airport = a.Airport_ID
JOIN final_state_IDs_df s ON a.State_ID = s.State_ID
JOIN final_state_level_df c ON s.State_ID = c.State_ID AND f.Date_ID = c.Date_ID
GROUP BY s.State
ORDER BY Total_Cases DESC;"""
```

Figure 2. A SQL query made for this project.

This query, explores the effects of COVID-19 on flights that were within state lines. This query shows that the flights started to die down around the COVID-19 peaks. This data is skewed though as there were issues with the API that didn't allow for a full set of data to be collected from the API. Even though good data point ranges were selected from the beginning and end of COVID as well as around the peaks there was still a lot of data missing.

```
query = """SELECT a.Date, s.State AS Arrival_State_Name, s.State AS Departure_State_Name, a.Total_Domestic_Flights, c.Total_Cases
FROM final_flights_df a
JOIN final_detailed_flight_data_df f ON a.Date_ID = f.Date_ID
JOIN final_state_level_df c ON a.Date_ID = c.Date_ID
JOIN final_state_IDs_df s ON f.Arrival_State_ID = s.State_ID
WHERE f.Arrival_State_ID = f.Departure_State_ID
ORDER BY a.Date;"""
```

Figure 3. A SQL query made for this project.

In this query it looked at similar information to the first query except it looked at the total death count for each state. It can generally be seen that the more amount of flights per state the more amount of deaths there were. This means that the more amount of flights there were the greater the spread of COVID. It can be inferred that the more travel there was to and from a state also increased the amount of deaths.

```
query = """SELECT s.State, COUNT(DISTINCT f.Flight_ID) AS Total_Flights, SUM(c.Total_Deaths) AS Total_Deaths
FROM final_detailed_flight_data_df f
JOIN  final_US_airport_codes_df a ON f.Arrival_Airport = a.Airport_ID
JOIN final_state_IDs_df s ON a.State_ID = s.State_ID
JOIN final_state_level_df c ON s.State_ID = c.State_ID AND f.Date_ID = c.Date_ID
GROUP BY s.State
ORDER BY Total_Deaths DESC; """
```

Figure 4. A SQL query made for this project.

This query looks at the flights on more of a state level. The state chosen for this was Arizona. This query looked at the total number of flights compared to the daily number of cases in the state of Arizona. This data wasn't as determinant at a state level since there weren't a lot of morning flights coming into the state overall even at the start and end of COVID. It's important to note that this isn't the full dataset as the API wasn't very efficient at extracting the data.

```
query = """SELECT f.Date, COUNT(f.Flight_ID) AS Daily_Flights_to_Arizona, c.Daily_New_Cases AS New_Covid_Cases_in_Arizona
FROM final_detailed_flight_data_df f
JOIN final_state_level_df c ON f.Date_ID = c.Date_ID AND c.State_ID = '2'
WHERE f.Arrival_State_ID = '2'
GROUP BY f.Date, c.Daily_New_Cases
ORDER BY  f.Date;"""
```

Figure 5. A SQL query made for this project.

In this query, it is aggregating the data monthly to look at the full span of COVID and morning flights. From this query, it can be seen that the major COVID peaks the flights morning flights started to decrease but once the new cases started decreasing the flights started increasing again.

```
query = """SELECT  EXTRACT(YEAR FROM f.Date) AS Year, EXTRACT(MONTH FROM f.Date)
AS Month, SUM(f.Total_Domestic_Flights) AS Total_Flights, SUM(c.Daily_New_Cases)
AS Total_New_Covid_Cases
FROM final_flights_df f
JOIN final_state_level_df c ON f.Date_ID = c.Date_ID
GROUP BY Year, Month
ORDER BY Year, Month;
"""
```

Figure 6. A SQL query made for this project.

Overall, the data hinted that due to the increase in daily COVID-19 cases morning flights started to decrease. Unfortunately, due to issues with the API, the data is inconclusive. In the future if I have the computing power to run all of the API requests for all 1,000+ days to get more conclusive evidence for if the hints that are in the data are accurate. Overall, after completing this data engineering project and applying the concepts learned in class to perform the full ETL process, I am more confident in my skills as a data engineer.