# **Project 3 Final Report: Data Engineering Track**

**Shannon Hoffman, Elisabeth Jansen, Brittany Thomas, Jose Vela**

**Data Sources**:

1. <https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E>; "Bureau of Transportation Statistics"
2. <https://github.com/ip2location/ip2location-iata-icao/blob/master/iata-icao.csv>

The topic of delayed flights and their impact on travel efficiency and satisfaction is a current and ongoing topic of interest to people for personal and professional reasons. It is broadly relevant given the number of individuals the air industry serves daily. Exploration of such data could provide insightful areas of opportunity for further process improvements related to travel as well as more informed travelers. The above data source was used to explore flight delay data from the Bureau of Transportation Statistics. Reasons for flight delays, frequency and duration of flight delays (overall and by specific airline) and possible correlations between delays and geographic region were explored for possible relationships, patterns and correlations using monthly data between the dates of January 1, 2023 and December 31, 2023.

## **Database Design**

The project used an extract, transform and load (ETL) workflow (Fig. 1) to ingest data into a SQL database. The main data set was extracted (downloaded) from the publicly available Bureau of Transportation Statistics website as a csv and named “2023\_data”. The csv was read into a pandas dataframe via Jupyter notebook (“Data\_Cleaning.ipynb”) and cleaned to remove unnecessary columns, rename columns intuitively, parse data into multiple columns (example: separate “City, State” into 2 columns instead of 1), reorder columns intuitively. Entries (rows) in which there was no flight data (n=47) or the only flight data for that carrier, airport and month were cancellations or diversions and therefore there were no arrivals (n=5) were also removed. Removal of this data was felt to have a negligible effect on data analysis. The cleaned data frame was written to a new csv file called “2023\_data\_cleaned”.

Additionally, the latitude and longitudinal coordinates for each airport represented in the “2023\_data” were exported as a json file initially and then as a csv <https://github.com/ip2location/ip2location-iata-icao/blob/master/iata-icao.csv>. The csv was read into a pandas dataframe via Jupyter notebook and cleaned to drop duplicate airport data. To aggregate the flight delay data with airport latitude and longitude coordinates, empty latitude and longitude columns were created within the existing\_df derived from the“2023\_data\_cleaned.csv” and populated with coordinates from “clean\_plots.csv” corresponding to the matching airport code by iterating through each row in the existing\_df. The new dataframe was written to a csv called “long\_lat\_test.csv”. This csv did not override the “2023\_data\_cleaned” csv as it will be used for only a subset of data analyses.

SQL was used to transform (structure) the data in an optimal format. SQL was chosen as the database of choice given its ability to handle large datasets in a consistent manner through a schema structure. Our data set consists of 22569 unique records. In addition the syntax is commonly known and commonly used for data analytics. It is easily transferable and modifiable for use in a python interface such as pandas.

Using PostgresSQL via pgAmin, a database called “flight\_data\_analysis” was created to store data from both csv files for future use. A new column in “2023\_data\_cleaned” called “flight\_data\_id” was created in the raw data csv file (outside of pandas) to facilitate incorporation of a PRIMARY KEY to tell SQL there will be unique values for every row (per T. Bogue, it was ok to add the column manually and directly using excel commands to the csv file as this was easiest) when the corresponding table is created. Two tables were created within SQL corresponding to each csv file. They are called “flight\_data” and “location”. “Airport” serves as the PRIMARY KEY in the “location” table and FOREIGN KEY in the “flight\_data” table. An ERD (Fig 2) was created to illustrate relationships between the data in the two tables.

Fig 1:

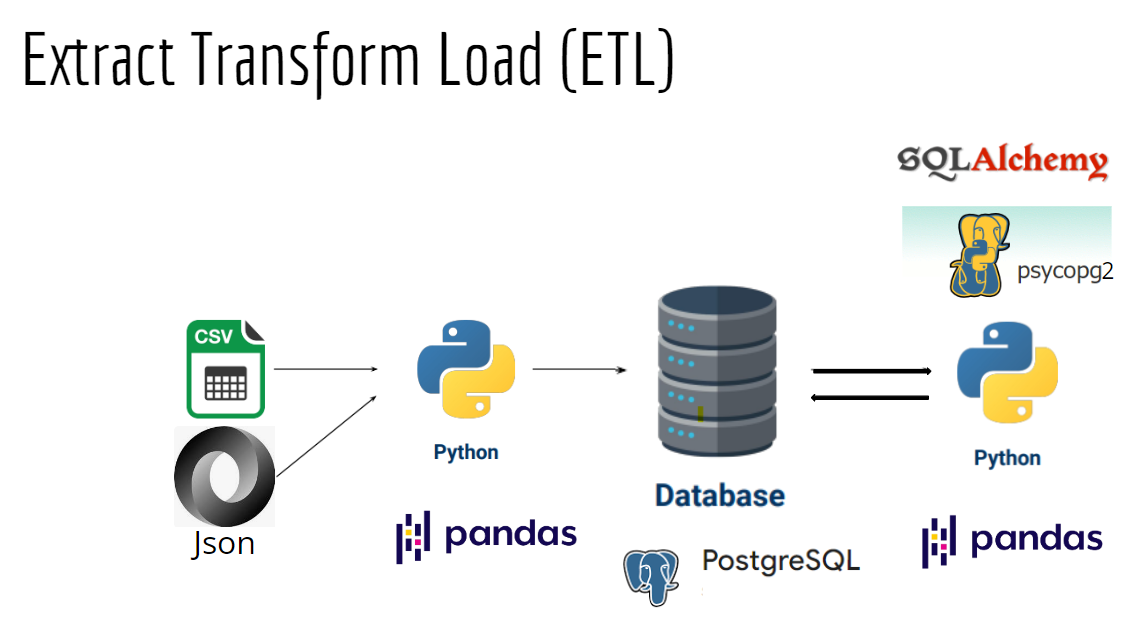
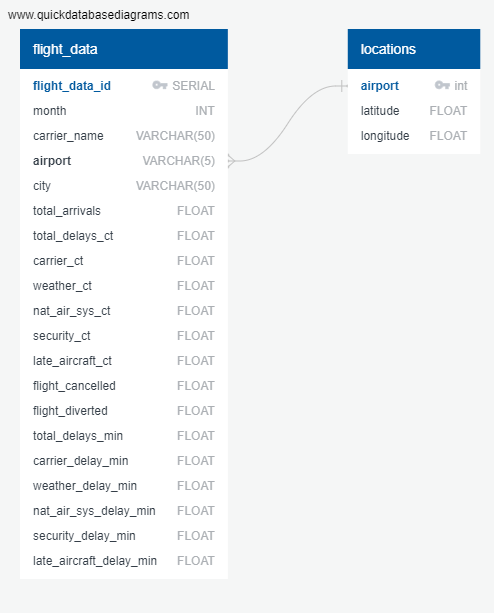


Fig 2:



## **Data and Delivery**

SQLAlchemy (ORM) and a new library called psycopg2 were used to optimize data analysis within pandas using data stored within the SQL database.

Each team member explored a subset of the variables related to flight delays as prompted by exploratory data questions and hypotheses in the project proposal. Each team member installed psycopg2 and connected to the SQL database “flight\_data\_analysis” using SQLAlchemy and psycopg2.

The new library, psycopg2, was identified to facilitate optimal communication between SQLAlchemy (provides ORM) and the database, allowing the user to use SQL syntax but display outputs and organize code in a more readable and familiar python format via pandas.

With a successful connection to an established database in PostgreSQL/PgAdmin4 users are able to complete queries utilizing SQL language. Queries can be completed through pandas with the .read\_sql function or creating a cursor with the psycopg2 and using the execute function to carry out the desired query. This library is useful for establishing a connection to databases larger or more complex than in sqlite with still allowing users to utilize the intuitive language of SQL.

Instructions for using and interacting with project data are as follows.

1. After downloading the raw data from the data sources (as “2023\_data.csv” and “clean\_plots.csv”), pandas was used to clean the data for further data exploration. See “Data\_Cleaning.ipynb” and “heat\_map\_data.ipynb” for the code used.
2. Open pgAdmin to access PostgreSQL.
3. Create a database called “flight\_data\_analysis”.
4. Create a table called “flight\_data” using the schema included in “2023\_flight\_data.sql”.
5. Create a table called “location” using the schema included in “locations\_table\_schema.sql”.
6. Import data into “flight\_data” table from “2023\_flight\_data\_cleaned.csv”. Import data into “location” table from “clean\_plots.csv”.
7. See Fig 2 above for the relationship between the data in each of the two tables. This relationship will be used to map flight delay data to pinpoints on a map of the United States using the longitude and latitude coordinates.
8. Install psycopg2 (“pip install psycopg2” in a command line).
9. Separate Jupyter notebooks were created to further clean and explore data within pandas using SQLAlchemy and psycopg2. Open each notebook to review and execute code corresponding to each analysis.

Ethical considerations of relevance to this project include the following:

* The data used for this project was not collected directly or indirectly from people, nor was it data that describes a population(s) of people. Therefore, considerations pertaining to human subjects research and analysis do not apply, nor do privacy issues.
* The data is publicly available, therefore ownership of the data is not an issue.
* Machine-learning algorithms were not used for analysis nor were they used to train a dataset used in the analysis, to the best of our knowledge. Data is collected in real time on a monthly basis and published for public use/review. Therefore the potential for introducing bias through the use of biased algorithms within this data analysis are thought to be low. However, there is a possibility that bias could exist within the data set based on ascertainment bias (e.g. certain airports are not submitting data or do not have a robust data capture system) or other forms of bias we are not aware of.
* Intent and outcomes. When performing data analysis, there is the potential to do harm unintentionally. Possibilities we considered with regards to our data exploration included the potential to perpetuate a negative stereotype about a specific airline or airport based on findings from our data exploration. Conclusions that lack statistical significance can mislead audiences and attribute generalizations about flight delays to carriers and airports unfairly. All findings from data analysis should be replicated using independent data sources or independent analyses which fall outside the scope of this project, again introducing the possibility of misleading audiences with unconfirmed findings.
* We used our best judgment to represent our data fairly and with the intent to do good, meaning to get a better understanding of flight delay patterns. In some cases, our assumptions about reasons for flight delays may be disproved which may be a positive outcome. In addition, data analysis of this sort on a larger scale could be used to identify opportunities to target specific causes or locations for process improvement and lead to better travel experiences for all.

## **Summary of Findings**

***Top 5 Analysis:*** Identifying the airlines with the highest number of delays by count is interesting, but not useful. In most “Top 5” analyses, the 5 airlines that had the most routes also had the most delays. The outliers are the interesting data points (e.g. Spirit Airlines ranks #1 in number of security delays but #9 in number of routes). A review by percentage would yield more useful information.

***Regional & Seasonal Analysis:*** the regional delays due to weather for all analyzed regions was less than 1%. While the southwest region showed over 20% of all flights were delayed, roughly only .8% of those are attributed to weather. Summer showed to be the season with the highest total delays and delays due to weather with July being the month with the highest total delays.

***Frequency vs Duration of Delays:*** There is a positive correlation between frequency (in #) of delays and total minutes delayed for a given airline, airport and month. This makes sense as more delays contribute to a longer total delay time. When graphing the distribution of frequency as the % of total flights delayed vs duration, there is a somewhat random distribution. This provides evidence against the hypothesis that airlines may be willing to endure more frequent delays if delays are shorter in duration and airlines with fewer delays may have longer delays. Lastly, when plotting % of flights delayed out of total flights against average duration of delay per flight this revealed specific airlines who perform poorly and well in both categories. This may be helpful to travelers in choosing high risk vs low risk or based on the variable they are more willing to accommodate.

## **References**

1. <https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E>; "Bureau of Transportation Statistics"
2. <https://github.com/ip2location/ip2location-iata-icao/blob/master/iata-icao.csv>
3. StackOverflow
4. Harvard Business School Online: 5 Principles of Data Ethics for Business Professionals (<https://online.hbs.edu/blog/post/data-ethics>)