# Project 3 Proposal: Data Engineering Track

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

**Data Source**: <https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E>; "Bureau of Transportation Statistics"

## Summary

Below are the steps we will take to explore the flight delay data from the above data source. During data exploration, we will identify a new library to include in the project. Our steps, analysis and conclusions will be presented during the class group presentations on July 23, 2024.

## Proposed Process Steps

Step 1: Data between the timepoints of Jan 2023 - Dec 2023 will be downloaded as a csv file (“2023\_data.csv”) from the following source: <https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E> .

Step 2: The csv file will be read into pandas to facilitate data cleaning.

Step 3: The data will be cleaned with the following goals:

* eliminate columns and/or rows that are unnecessary for further data exploration
* identify and correct values that are not importing correctly
* determine if null values are valuable or should be eliminated
* other (e.g. rename columns, combine columns, separate columns, create new columns - delay cts as % of total)

Step 4: A new database called “2023\_flight\_data” will be created and stored in pgAdmin (SQL).

Step 5: The “2023\_data.csv” file will be imported in the newly created database in pgAdmin.

Step 6: An entity relationship diagram (ERD) will be created to demonstrate relationships between data fields in order to facilitate efficient exploration steps and strategies.

Step 7: Sql Alchemy will be used to read data from the SQL database in pgAdmin into one or more Pandas dataframes.

Step 8: Tables will be created to answer exploratory data questions (see below) calling on data from the SQL database using this resource: <https://www.datacamp.com/tutorial/how-to-use-sql-in-pandas-using-pandasql-queries> .

Step 9: The ERD will be refined to ensure accurate representation of the relationships between the data fields and tables created.

Step 10: Visualizations will be created (e.g. determine the percentage of flights delayed by region of country and create a heat map) and calculations will be performed using Pandas modules to address the Exploratory Data Questions below.

Step 11: Our proposal, process, data exploration and data analysis will be presented to the class on July 23, 2024.

## Exploratory Data Questions/Hypotheses:

Which airline carrier has the most delays due to security issues?

What month is associated with the most weather delays?

Hypothesis: There are more delays associated with weather in Seattle than in Austin.

Rank the seasons in terms of highest to lowest delays.

Rank airports with highest to lowest delays (and based on weather, late crew, security, etc).

Rank airlines with highest to lowest delays.

Identify the top 5 airlines with delays and do further in depth analysis to elicit relationships or correlations between data points (e.g. average time delay vs number of delays).

Of the top 5 airlines with the most delays, what is the range and average time of delay?

Of the top 5 airlines with the longest delays, what is the range in frequency of delays (how many)?

**Elisabeth**: season/weather vs region - correlations (month ass’d w/ most weather delays); Look at cities (Seattle vs Austin)

**Jose**: focusing on visualizing data (heat map) off of time/counts; look deeper regarding types of delays (layers)

**Shannon**: rankings – delays ranked based on different criteria (airports/airlines) – of those with highest/lowest delays, what delay type is most/least impacted

**Brittany**: - last two questions above -look at relationship between frequency of delays and length of delays