\* \*\*E\*\*xtract:

The original data sets were found from Kaggle at the following link:

<https://www.kaggle.com/usdot/flight-delays#flights.csv>

The data sets were in Comma Separated Value (CSV)format. These have been successfully read into Postgres by importing each of these databases into individual tables, after making certain transformations in the Jupyter Notebook.

\* \*\*T\*\*ransform:

The flights data set contained multiple columns with several null values. Those columns have been dropped. Also, few rows in the flights dataset contained numeric origin airport and destination airport codes, that were not found in the airport dataset. Those columns have been not read while reading the dataset into pandas. Flights ID column has been created with the index row number values to serve as a Primary key for the flights data set. Modflight dataset is the modified dataset which has been created from the flights dataset, with the above mentioned two transformations, along with certain column name changes.

\* \*\*L\*\*oad: the final database, tables/collections, and why this was chosen.

The final flights database contains three tables: airports , airlines and modflights. Each table has an individual and separate primary key. Airports and Airlines tables are related to the modflights table, as shown through the ERD diagram. These relationships thereby make these datasets relational and hence, Postgres database has been chosen.