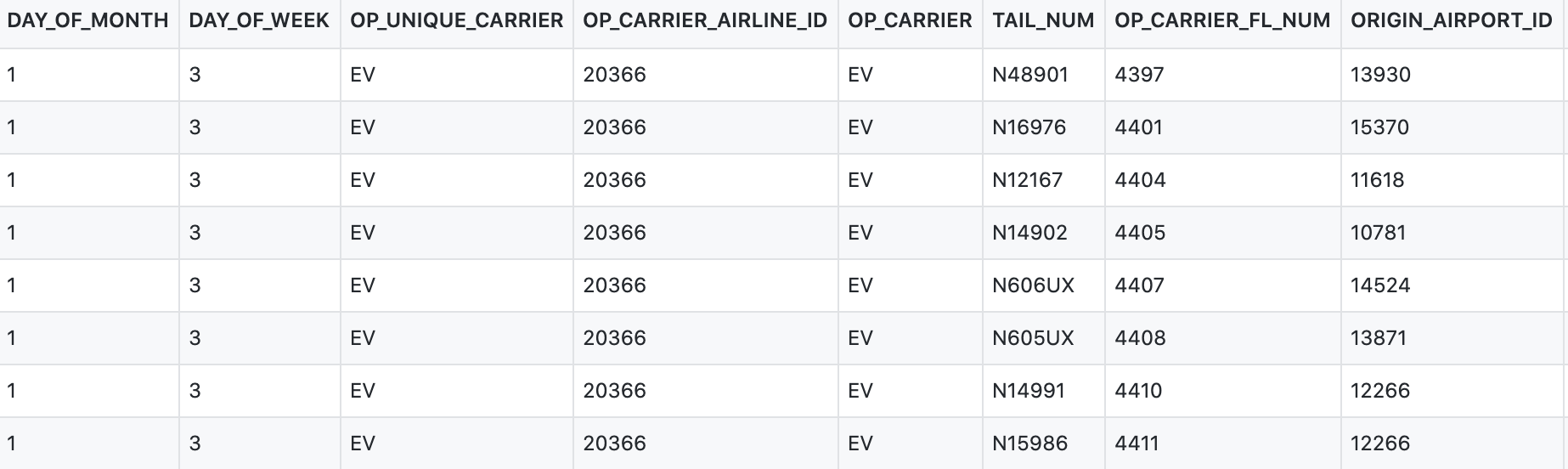
### **ETL Project**

Calvin Brown, Sam Ettenson

### Reason for Data Manipulation:

* The table created to track flight delays in January of 2020 had a lot of excellent information on it but was clearly lacking in one area. It was difficult to know what airline they were specifically talking about because the table only includes the carrier id and abbreviation of the airline. If someone was not well versed in airline id’s it probably would have been difficult for them to track which airlines had the most or least delays. Adding an airline name column would make this table a lot easier to read.

### Sources of data:

* Kaggle- January Flight Delay Predictions (<https://www.kaggle.com/divyansh22/flight-delay-prediction?select=Jan_2020_ontime.csv> ) (csv)
* Bureau of Transportation Statistics (<https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=259> )(csv)

A screenshot of a cell phone

Description automatically generated

Transformation of the data:

**Airline\_Codes:**

* Loaded csv into Pandas and manipulated data
* Renamed Columns to make data easier to read

A screenshot of a social media post

Description automatically generated

**Flight\_Data:**

* Loaded csv into Pandas

A screenshot of a cell phone

Description automatically generated

* Filter the database down to grab relevant columns. Narrowed it down from 21 columns to 8 columns and renamed them make data easier to read.

A screenshot of a social media post

Description automatically generated

* Added the Flight\_id column to set it as an index and as a unique value for the table to set as a Primary key in SQL.

A screenshot of a cell phone

Description automatically generated

### Type of final production database data is loaded into:

* Created a new database in SQL (Jan\_2020\_flight\_data\_db) with two tables to import data frames from pandas.

A screenshot of text

Description automatically generated

* A screenshot of a cell phone

  Description automatically generatedCreated the connection string to connect our jupyter notebook file with the postgreSQL database. Then we uploaded our dataframes into the tables we created in PostgreSQL.

Final table that we used in the production database:

* A screenshot of a cell phone

  Description automatically generatedJoined the tables in postgreSQL to be able to add the airline names and match them with the airline ID’s. This allowed people who are unfamiliar with the airline codes to understand what airline they pertain to.
* Final table result

A screenshot of a cell phone

Description automatically generated