**ETL Report: Funatics & Lunatics**

Group 4: Benjamin Hawkins, Brittney Oleniacz, David Seger, Andrew Walker

1. Extract

Dataset of UFO sightings was chosen due to the novelty of the topic. Looking at the UFO data, California contributed the most occurrences of UFO sightings. The UFO data provided criteria for the second dataset by limiting the study to a geographic location. Thus, the Los Angeles Crime & Arrest Data was chosen. The raw data for both datasets was acquired by downloading CSV files from Kaggle.com. The CSV file for the LA Crime data was 723 MB, containing nearly 2 million rows of 26 columns. The file of UFO data was 28 MB with over 80,000 rows with 11 columns.

1. Transform

Jupyter Notebook was used to clean, filter, and transform the datasets. First, the number of columns needed to be reduced in both datasets. The LA crime data included 26 columns, many of which were found to be irrelevant, excessive, and unusable to the project. Thus, 23 columns were eliminated, leaving “Date Occurred”, ”Area Name”, and “Crime Code Description” columns. The UFO data was reduced to six columns, keeping “datetime”, “city”, “state”, “shape”, “duration (seconds)”, and “comments”.

Next, we worked to decrease the number of rows in each data set. For the UFO sightings, first those recorded outside of California were eliminated. Then, rows containing NaNs/missing data were dropped, effectively reducing the dataset from over 80,000 rows to 9404 rows. As for the LA Crime data, there was no NaN/missing data in the three columns remaining, thus the dataset still had nearly 2 million rows.

The only way to further reduce the number of rows was to restrict the data to the overlapping dates between the two datasets. To do so, each dataset required transforming their own, uniquely-formatted date-time columns to date-only datetime objects and then applied our own formatting to the date to ensure consistency between datasets.. The LA Crime data provided the “start date” and the UFO Sighting Data provided the end date. After the cleanings, filtering, transforming and formatting, we were left with 2718 rows × 6 columns for the UFO Sighting data and 867296 rows × 3 columns for the LA Crime Data.

1. Load

The final database used: PostgreSQL

- Two tables created, UFO and Crime, with table schema queries

- PGAdmin was used for ease of coding/familiarity through Jupyter Notebook

- Also picked for practical application (can easily demonstrate joins between tables on particular dates/periods of time)