The first step in completing the ETL project was to find two datasets suitable to use. The sites explored for information were Kaggle.com and data.world. The first topic we thought about exploring was earthquakes. The data sets were very large, so we decided to try and explore UFO sightings. We found that all the data sets for UFO sightings were basically duplicates of the same set. We switched back to earthquakes. We chose data sets for earthquakes from Japan and Greece based off the high volume of earthquakes identified by the geological scientist.

The sources of data are

<https://www.kaggle.com/astefopoulos/earthquakes-in-greece-19012018>

<https://www.kaggle.com/caganseval/earthquake>

The next step was to cleanup the data using Jupyter Lab. After we imported the CSVs, the header in the dataset for Greece had to be translated from Turkish to English. The columns were renamed using pandas “.rename.” The datasets were examined for consistency and several columns were removed from each dataset. The date formats were not the same in each dataset, so they were modified using “pd.to\_datetime.”

The final database was named earthquake.db which joined the datasets from Japan and Greece. The datasets were combined by using “pd.concat.” The new data frame was exported to a csv file and then transferred to and SQLite database.

Some challenges were choosing the way the datasets were joined. We were first trying to use PgAdmin, but found concatenating the tables and loading them to SQLite through Jupyter Lab was more simple.