**EXTRACT:**

Original data sources:

“NASA Astronauts 1959-Present”, Kaggle.com, CSV file

“Space walking”, Kaggle.com, CSV file

“https://www.archives.gov/research/alic/reference/space-timeline.html”, scraped HTML

**TRANSFORM:**

Using Python and Jupyter Notebook we read-in the Space walking CSV file. This file was straightforward regarding transformation. The data had EVA#’s in each row which correspond to that Extravehicular Activity. There were several lines which had an EVA # but did not provide meaningful data in the other fields. These were missions which were initiated but did not make it past the beginning stage. An example was an air lock which wouldn’t open. Most other columns would be blank with a duration of 0:00. We decided to remove these data points as they wouldn’t contribute to analysis later.

The web scraping we also used Jupyter Notebook in along with Beautiful Soup and Pandas. Inspecting the HTML we were able to mine the sections which would be useful for our data. There was a year header which we wanted to gather in addition to facts for missions during each year. We looped through both the years and the facts to create a list of each. These lists needed to have items removed. We were able to loop through the lists with .remove until left with only the points we needed. Once we had matching length lists, we imported them into a Pandas data frame.

We wanted to pull the astronauts education so we downloaded a separate data set from Kaggle that had their undergrad and graduate degrees. We pulled two data frames that gave the most popular undergraduate and graduate degrees. We then had to figure out how to merge these degrees to the first csv file. We noticed the main csv file only had first and last name for the crews and the education data set had a middle initial. On the education csv, we split the name to get the first and last name with no middle initial. The main file, we then created two columns for the undergraduate and graduate degrees. We then had to loop through the main csv crew column to get the name(s) of all crew members for each mission. At the same time, we had to also loop through the education file name column to match the name to the crew and add each degree for the corresponding astronauts. If the astronaut wasn’t listed on the education csv or a degree wasn’t listed, we added a “Not Listed” string. We also used that same string for all blank and nan education cells.

**LOAD:**

Our final database was SQLite where we had two tables. One was the year and the facts for each year. The other was the majority of the data with the astronauts, background and spacewalk missions. We chose SQLite as it was a relational database and it was relatively lightweight given our analysis would be relatively light as well. We had also planned on using a Flask app to present the data on a website. The SQLite database seemed the most straightforward way to integrate with Flask.