ETL Project Report: Big Foot & UFOs

November 13, 2019

Ben O’Neal, George McGuire, Randall Davis

**Extract**: We extracted two CSV files. One contained big foot data from Dataworld, and the other CSV contained UFO sighting data from Kaggle. We decided on using pgAdmin4 to construct our database.

**Transform**: We had to clean up the data from the CSV files in order to successfully import the files into Jupyter Notebook by way of the Pandas dependency. Prior to doing this, we had to construct a VLOOKUP in Excel to match the lowercase abbreviated state names. By doing this, the merging of the CSV files was easier into Jupyter Notebook. When we merged the data (Big Foot was the “x” and UFO was “y”) we had to rename the columns before importing it into pgAdmin4 for the database.

**Load**: We decided to export the merged CSV data of Big Foot and UFOs into one CSV and use that for our database in pgAdmin4. Another reason we decided on pgAdmin4 is because it is a relational database and we felt the most comfortable with this method. We created a table for our database in pgAdmin4 first, then imported our new CSV into the table using the steps learned in the SQL activities from class. To ensure that the file was imported correctly and that our database would work, we used highlighted the “select \* from BF\_UFO\_Data” and clicked the lighting icon.

**Lessons Learned**: We tried to merge the CSV files on the “state” column and we encountered that the data frame was way too big to export into a CSV. It contained roughly 14 million rows, so we decided to merge on the longitude which gave us a good amount of data. We also went with this column due to the constraint on time we had to complete this project.