ETL Report

Tremayne Watterson & Kelly Hull

**Project goal**: To perform ETL on multiple datasets to create a larger database on which thorough analysis could be performed.

**Selected topic**: Urbanization levels and UFO sightings in the United States

For this ETL project, we wanted to have a little fun, and gave ourselves a bit of a challenge. We decided to pick two unique datasets that at their face may have nothing in common, but potentially after some analysis may prove to have some strong correlation.

**Extracting the Data**:

Datasets utilized:

* [Urbanization Census Tract](https://data.world/fivethirtyeight/urbanization-index/workspace/file?filename=urbanization-census-tract.csv) (maintained by [FiveThirtyEight](https://data.world/fivethirtyeight/urbanization-index/workspace/file?filename=urbanization-census-tract.csv) on [data.world](https://data.world/fivethirtyeight/urbanization-index/workspace/file?filename=urbanization-census-tract.csv)) – dataset breaking down the “urbanization index” of a census tract as a numerical value, with higher numbers meaning a higher level of urbanization, basedo n the avarege number of people living within a five-mile radius of a given resident according to 5-year American Community Survey data from 2017. This dataset was originally created for the article “[How Urban Or Rural Is Your State? And What Does That Mean For The 2020 Election?](https://fivethirtyeight.com/features/how-urban-or-rural-is-your-state-and-what-does-that-mean-for-the-2020-election/)” by Nathaniel Rakich, also for FiveThirtyEight.
* [UFO Sightings](https://www.kaggle.com/NUFORC/ufo-sightings) (maintained by the [National UFO Reporting Center](https://www.kaggle.com/NUFORC/ufo-sightings) on [Kaggle.com](https://www.kaggle.com/NUFORC/ufo-sightings)) – rich dataset documenting UFO sightings in North America, Great Britain, Denmark, and Australia since from 1910-2013.
* [State Population Totals: 2010-2019](https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html) (maintained by [the US Census Bureau on census.gov](https://www.census.gov/en.html)) – dataset which estimates the resident population for the United States, regions, states, and Puerto Rico from 2010-2019. While the FiveThirtyEight dataset does contain some population data, we wanted to ensure that we had the option of breaking our dataset down cleanly by state, which the FiveThirtyEight dataset could not do on its own.

All three datasets were fairly straightforward .csv files, so extraction simply involved downloading the files and adding them to our shared Github repository.

**Transforming the data**:

We were strategic in choosing robust, well-organized datasets with columns that named the states/territories for which the data was measured, which provided us a common column on which to join any of our tables. [We transformed all of our data in a jupyter notebook document](https://github.com/kehull/ETL-project/blob/main/Final_ETL_Python.ipynb) using Python, specifically the Pandas and NumPy packages.

The first step of cleaning our data meant standardizing the format in which our state names were presented within the tables. For this task, we discovered a function written by GitHub user [rogerallen](https://gist.github.com/rogerallen/1583593) called [us\_state\_abbrev.py](https://gist.github.com/rogerallen/1583593) which allowed us to perform this transformation seamlessly.

During the process of working with our data, we found that certain locations (such as Puerto Rico and Guam) didn’t appear in each dataset. To keep our data consistent, we chose to limit our data to only the 50 United States and Washington D.C.

Another hurdle in normalizing our data was ensuring the timeframes in which our data was collected was relevant. As demonstrated above, our datasets represented three different timeframes. We chose to use the mean of the estimated state population from the Census Bureau data rather than select a single year, and we chose to limit our UFO sighting data to only data collected from 2010 to the end of the dataset. These slightly mismatched timeframes are an unfortunate limitation of our database; however, we believe that use of analysis outside the scope of this ETL project would demonstrate that the combined data is still robust enough to be meaningful.

In preparing to transform the data, we imagined the sorts of questions we might ask of the data if we were analysts on this project, and we quickly realized that having quantitative categories for the urbanization index would be helpful for binning data. We chose to split urbanization into four categories based on their numerical score- Very Rural, Rural, Urban, and Very Urban.

In the spirit of useful categorical data, we also created a table of the states and Washington D.C. broken down by their region (such as the south or the northwest), as well as the smaller divisions within those regions.

With both datasets having more than just the actual urbanization index, or ufo count of ufo sightings, our first step was to cut down the data to something a little more to our liking. Since we wanted to compare information by state, we needed to be able to have the datasets merge on the state column, and there in lied our first problem. The UFO sightings data set had each of the states, but only had the abbreviation for each state. While the Urban index had the opposite problem. To keep things simple, we decided to change the Urban index into their abbreviations. Thankfully we found some code where someone had already typed out each of the state’s acronyms matched up to the full name. We then created a new column in the data set and dropped the old one. We then created an additional column called “Urban Scale”. This had each of the states cut into 4 different quartiles based on their Urban Index score. They were Very Rural, Rural, Urban, and Very Urban.

For the UFO dataset, we had to shorten the time frame by quite a bit. Since it contained data from the entire past century, we trimmed down the time period to better match up with the data from the Urbanization Index. Since FiveThirtyEight took data from the 2017 American community survey, we cut the UFO data set to only data from the year 2010 on. In order to do this, we took the datetime category of the data set and made the formatting a bit cleaner to better be able to drop all data from before the year 2010. After renaming all the columns to make them a bit easier to understand, we had the data sets prepared to be better merged.

**Loading data**:

Loading the data involved the use of a fairly straightforward [SQL schema](https://github.com/kehull/ETL-project/blob/main/schema.sql). We were able to demonstrate proof of concept as far as our tables’ ability to be joined and transformed with [example queries](https://github.com/kehull/ETL-project/blob/main/query.sql). This dataset is now incredibly robust and well-crafted for our goal as stated at the start of this document.