ETL Report.

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. After some sleuthing through data world and Kaggle we found our perfect pair. For the first, we chose FiveThirtyEights urbanization index. This was created for an article the post did in April leading up to the November 2020 election. For the second, we chose a report from the National UFO reporting Center on UFO sightings over the past century. The report has the time, date and even shape of different UFO sightings across the past hundred years. While the two data sets may seem to be completely different, both have similar ideas at heart. For FiveThirtyEight’s dataset, they wanted to see if a state’s level of urbanization, had any correlation to how a state or census tract voted in 2016. And for the UFO report, the researches wanted to see if any areas of the country are more likely to have higher numbers of UFO sightings. With our two datasets selected we began the actual ETL process.

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.