ETL Write up

The purpose of this assignment was to get some practice with the ETL (extract, transform and load) process. We began by searching for datasets that shared a primary key. We decided to look for national data, and we planned on using both full names of states and state abbreviations as primary keys. We were able to locate 3 different datasets that shared these keys: one about energy production/consumption, one about homelessness rates, and a final one containing basic population data.

In order to transform our raw data into clean data that is sensitive to sql formatting, we had to perform some functions in python. After examining the csv files, we found that the energy csv had the most comprehensive list of states, containing data on territories such as Guam and Puerto Rico, so we decided that the primary key would be from there. We read the energy.csv file into pandas and had to drop the row that contained data on “United States” because it was going to interfere with our intent to make a second primary key of state codes to incorporate our data about homelessness. We also read the population data into pandas and had to drop the row on Puerto Rico out of the dataframe for the same reason. We also had to drop a few years worth of data from the homelessness csv because our other datasets didn’t contain those years.

After all bad data had been dropped, we split the energy csv data into 3 dataframes: consumption, production and price and only kept necessary columns. We also read the homelessness data into pandas and dropped unnecessary columns from the homelessness data and the population data. After all unnecessary columns and rows had been dropped, we renamed all columns into easy to read names and made sure they were all lowercase for easy manipulation in pg admin/ SQL.

We then connected to the database and loaded our dataframes directly from pandas into sql. We constructed the schema by simply defining both of our primary keys and all of our foreign keys, because pandas had conveniently made all of our tables for us already. All that was left to do was perform some joins on our data to compare the different datasets.