**ETL Project**

By: Austin Cole, Dustin Combs, Nicholas Burns

**Extract**

For our project we used several datasets that were a matter of public information. These datasets included:

* Plant-level data for generators.
* Additional details for generators that use solar as an energy source.
* A list of plants for solar in the United States across all sectors.

The datasets and additional information can be found at the U.S. Energy Information Administration website (<https://www.eia.gov/electricity/data/eia860/>).

Additional datasets included a csv of US Cities as well as well as a dataset from the National Oceanic and Atmospheric Agency (NOAA) that contained sunshine data for cities.

**Transform**

The first step was loading in all the CSVs and cleaning up the data, as necessary. Examples of this include averaging the sunshine percentages from the NOAA dataset and creating an “Annual Sunshine Percentage” column, shown in figure 1. This also included dropping unnecessary columns from other datasets, an example is shown in figure 2.

**Table

Description automatically generated**

**Figure 1**

Table

Description automatically generated

**Figure 2**

The step was to merge the sunshine, city, solar and plant data on columns they had in common. An example is shown below in figure 3.

Table

Description automatically generated

**Figure 3**

Next, we converted the latitude and longitude coordinates to radians to match the NOAA data to each plant, shown in figure 4.

Table

Description automatically generated

**Figure 4**

We then used the haversine formula to match the closest coordinates. Finally, we merged our final dataframes to find the annual sunshine percentage for each location.

**Load**

Next, we loaded our final, merged, output into a database. We created a new database within Postgres and created a new table with matching columns from our final dataframe. Then we connected to the database using SQLAlchemy to load the dataframe into the database. The final step was to check and make sure everything loaded in successfully. We chose this method this method because we felt it was the most effective way to work with the data. This allows us to run multiple queries based on any criteria related to our data.