**Weather data (NOAA):**

The weather dataset is probably the more challenging dataset to work with. There is just over 70,000 rows of data for a single year. When querying data, the stations are not associated with their state, however they do have a station code. Data looks like the following:

A screen shot of a computer

Description automatically generated

Date is the date, datatype TAVG stand for “Temperature Average”, Station “GHCND:<station\_code>” doesn’t mean a whole lot, attributes does not matter, value is the temperature. Luckily, there exists a txt file that allows us to map these station values to state codes (e.g., MN, CO, FL). Below is what it looks like:

A screen shot of a black screen

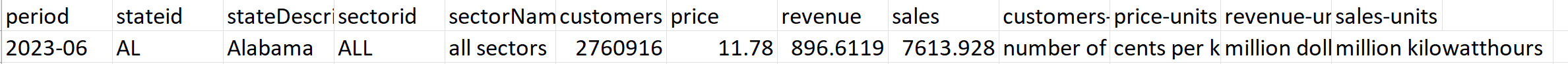
Description automatically generated

Using this txt file, it is possible to aggregate the data by states much easier in the transformation stage.

There does not appear to be any missing data, so accounting for bad data will not be something to consider.

**Electricity data (EIA):**

With around 22,000 rows for 5 years worth of electricity data, there is less data to handle, and the data itself is cleaner. Below is an example of the data:



Sales represents not the dollar amount, but the million kilowatt-hours sent out to customers. There is a very clear state id that is used to identify where the data is coming from. All the data exists, there does not appear to be any missing data. Little transformation will be needed for this dataset beyond merging it with the weather data.