Capstone Project 1: Milestone Report

Title: Data Visualization for later Prediction of electricity rates (cost/Kwh)

Author: Elizabeth Izarra

General Problem:¶

Electricity has a very dynamic market price since it is a commodity that is essential for daily life and non-storable where generation and demand must be continuously balanced. This is turn make it dependable of the weather conditions.

Data Sources:¶

- U.S. Energy Information Administration (EIA)
- National Oceanic and Atmospheric Administration NOAA

Data Acquisition and Data Wrangling:¶

The data has to be gotten from different web sources. Each source provides APIs which have to be explored to get the required data for the project.

The data acquisition and data wrangling was divided in the following parts:

1. U.S. Energy Information Administration (EIA): API Exploration, data acquisition and data wrangling in order to get a single view data set of electricity prices, demand, etc by state per month in a year.

   Findings:

   - Each variable series has to be independently fetched per State through APIs.

   Approach:

   - All the data by state corresponding to the same variable was fetched in a loop and concatenated while adding the state information. It was nested in a loop corresponding to all the variables of interest where after fetching each variable, merged them in a single view.
   - The rows that do not correspond to one State were deleted.

2. National Oceanic and Atmospheric Administration (NOAA): API Exploration, data acquisition and data wrangling of data in order to get a single view data set of temperatures by state per month in a year.

   Findings:

- Global Monthly Summary series can be fetched per state through APIs. It provides summary data from each station at a State into a time range. Some States have around 350 stations. Nevertheless, the API maximum limit is 1000 records per fetch.

Approach:

- A couple of loops where nested to fetch data per state per month to cope with the fetching limit. The first loop was used to get the data from all the stations in a State in a month. It was then grouped by date and variables to get the mean of all the stations in the state during that month. The group was unstacked to be able to concatenate with the information of the following month while adding the corresponding State information. This was nested in a per-state loop.

3. Merging EIA and NOAA Datasets of one year data

Findings:

- The columns ('date' and 'iso3166'/'State') to be used to merge in a single view both data sets had different formats.

Approach:

- The ISO3166 acronyms for US states with its correspondent states where searched and placed in a csv file.
- On the NOAA temperature data set, a column with the corresponding ISO3166 codes was added, as well as, 'date' was formatted to "YYYY-MM-DD'.
- On the EIA data set, the date was formatted to "YYYY-MM-DD''
- Cleaning functions were defined for both EIA and NOAA data sets.
- Merging of both data sets was done to get a single view of all the data of interest.

4. Retrieving Data from 2001 to 2018:

Findings:

- NOAA server disconnect after certain time.
- EIA does not have data available through API previous to 2001

Approach:

- Getting the data Year by Year (From 2001 to 2018):
- The three previous steps (1.- EIA date set acquisition per month by state in a year, 2.- NOAA data set Acquisition per month by state in a year, and 3.- Merging of EIA and NOAA data sets) were nested in a loop to get 18 years of data and save them in individuals .csv files per year.
- Getting all years of data in a single view:
  All the yearly .csv file were appended to get all the data in a single view and saved into a file.
  Files generated:
  eia_YYYY_YYYY.csv - one file per EIA year fetched
  noaa_YYYY_YYY.csv - one file per NOAA year fetched
  data_all_YYYY_YYYY.csv - one file per year fetched with EIA and NOAA data merged

AllData_1.csv - File with all the data_all_YYYY_YYYY.csv files combined

5. Data Processing

5.1. Checking for missing information:

Findings:

- All States were missing "number of accounts' per month for years early to 2008. In addition the state of Alaska also missed the 'monthly customer accounts' in the year 2016.

Approach:

- A .csv file was found available in EIA website with the mean annual "number of accounts" for each state. This file was downloaded and used as a reference for later interpolation of the "number of accounts", if needed.

5.2. Checking for apparent wrong data:

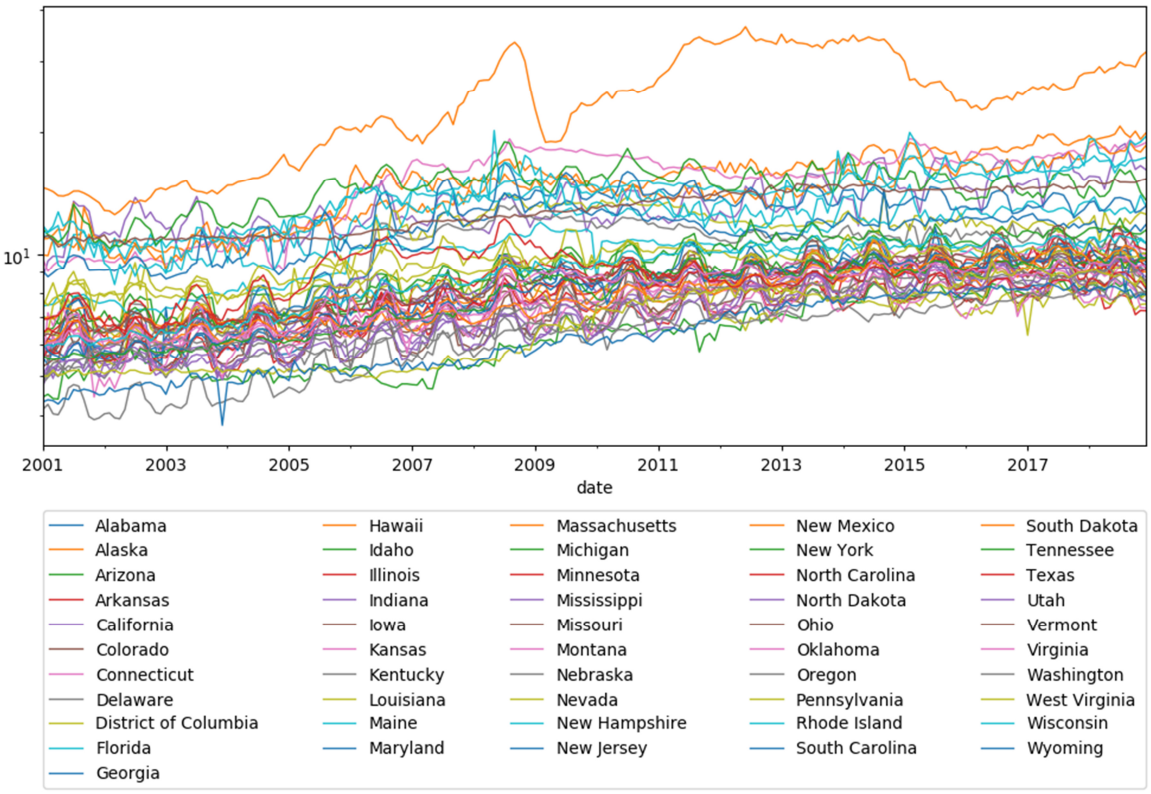
Findings:

- All variables were plotted per State. Graphs show an apparent regular pattern along states with exception of 'Net generation' where 'District of Columbia' presents some negative values. Also, Retail Electricity Price of "District of Columbia presents" presents an irregular patter. (See Figure 1 as an example)

Approach:

- Further analysis will be done before deciding if dropping 'District of Columbia' Data

Figure 1. Average retail of Electricity (cents per kilowatt-hour) – per State



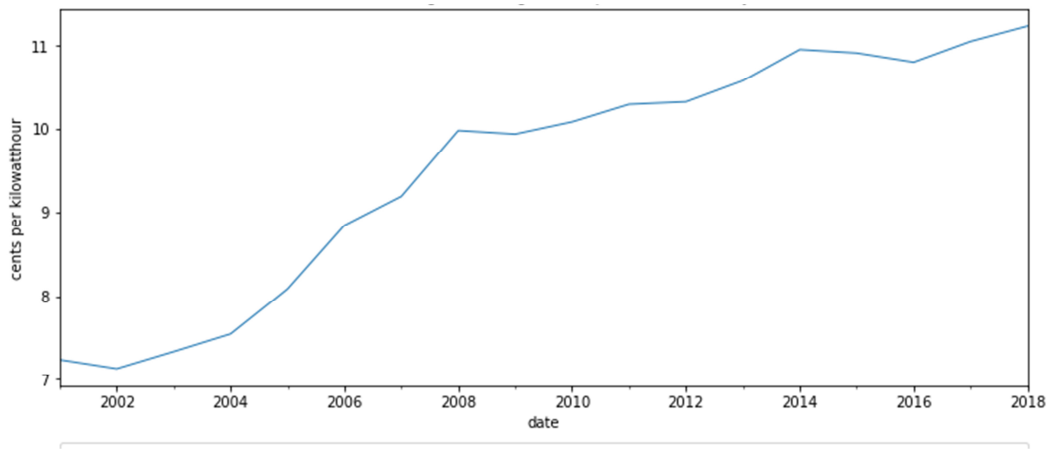
Data Visualization¶

Different data visualizations were done in order to identify possible trends, seasonality and dependencies.

Behavior of electricity prices along the years¶

It has been a steady increase in annual aggregated average of electricity price in USA over the last 18 years.

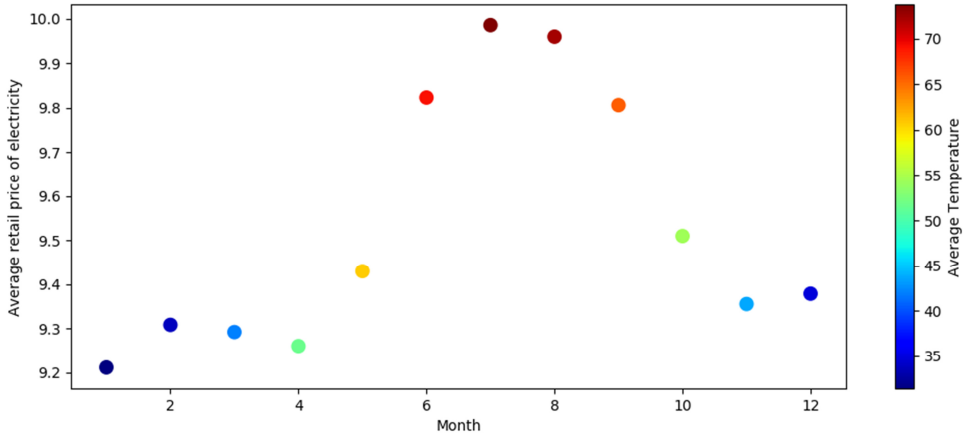
Figure 2. USA Annual Aggregated Average of retail price of Electricity



Relationship between Electricity price and Temperature¶

Agregated data from 2001 to 2018 of average electricity price in USA per month is presented below. The months of higher temperature present higher prices. A seasonality or correlation between Electricity Price and Temperature is identified

Figure 3. USA Monthly aggregated Average of retail price of Electricity (cents per kilowatt-hour)



Relationship between Electricity Demand (Retail sales of electricity), Electricity Net Generation, price and revenue¶

Figure 4. USA Monthly Aggregated balance between generated and sales electricity

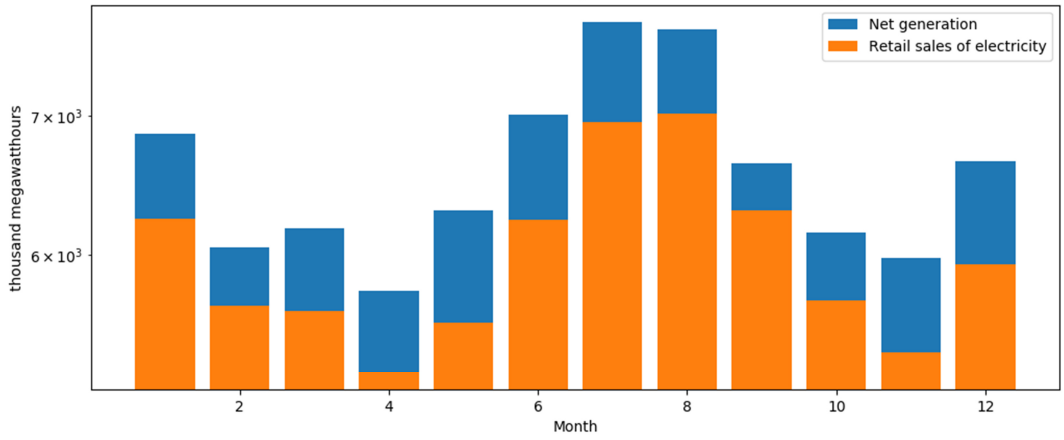
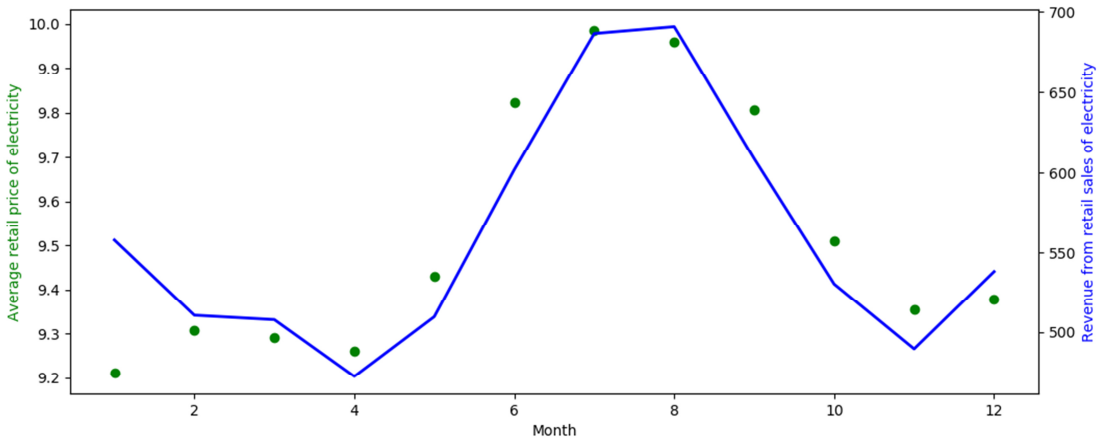


Figure 5. USA Monthly Aggregated average of Electricity Price (cents per kilowatt-hour) and Revenue ($Millions)



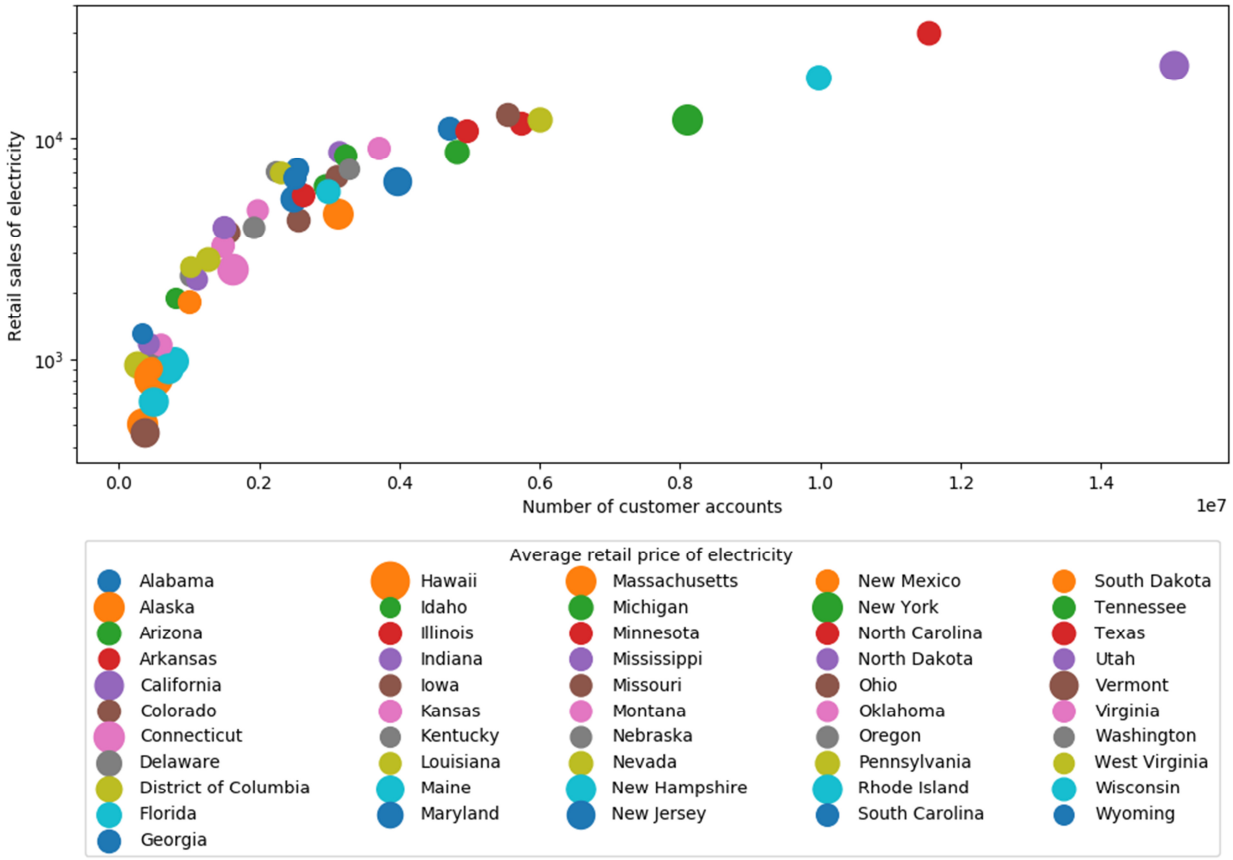
Looking at the two graphs above, we can appreciate:
1.- Revenue is higher when prices are higher.
2.- Consecutive months with similar temperatures/demands as Dec-Jan and July-Aug, the prices go down a bit the second month but revenue stay or increase.
Conclusion: We can infer that appropriate demand prediction lead to lower prices and higher revenue.

Relationship between Number of customer Account and Electricity Demand (Retail sales of electricity)¶

As expected, there is a relationship between the number of customer accounts and demand (Retail sales of Electricity).

The size of the bubble represents the aggregated average of retail price of electricity. At this point it is not clear the relationship between number of customer accounts and price, if any. But certainly there is a relationship between number of accounts and Net generation. Therefore, number of accounts is a feature that has to be consider in forecasting.

Figure 6. Annual Aggregated Average of Retail sales of Electricity (million kilowatt-hours) and Number of Customer Accounts

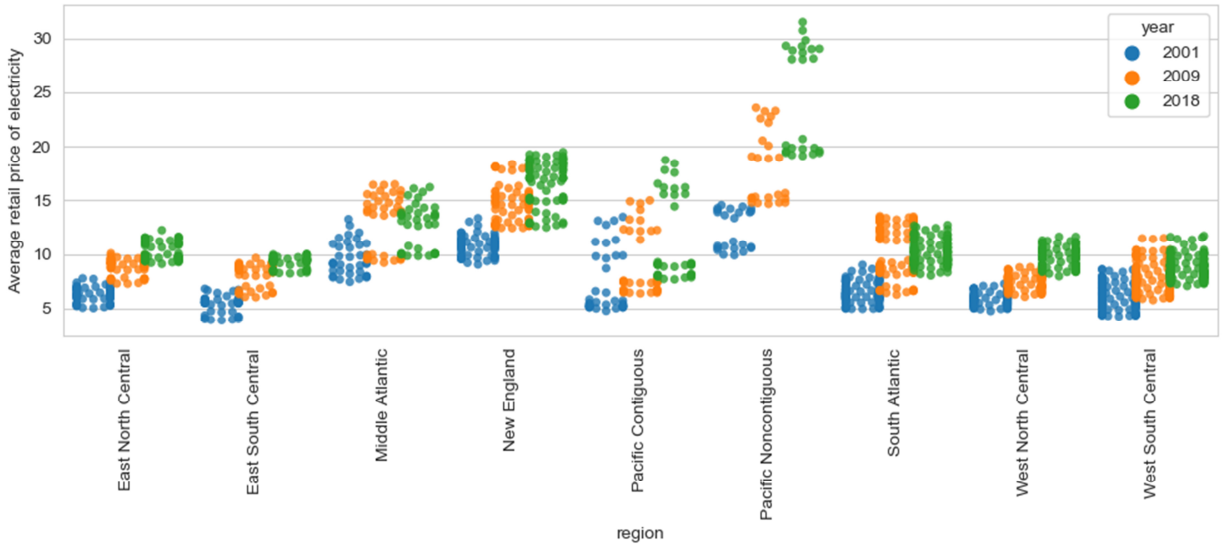


Exploratory Data Analysis (Inferential Statistics)

In Figure 1Figure 2 was visualized that the average retail price of electricity has increased over the last years. A deeper analysis is done by comparing data from three different years into the period 2001-2018 per region. For this analysis the years 2001, 2009 and 2018 were selected for comparison purposes. The figure below shows the percentiles (boxplot) and distribution of the average retail price of electricity per region.

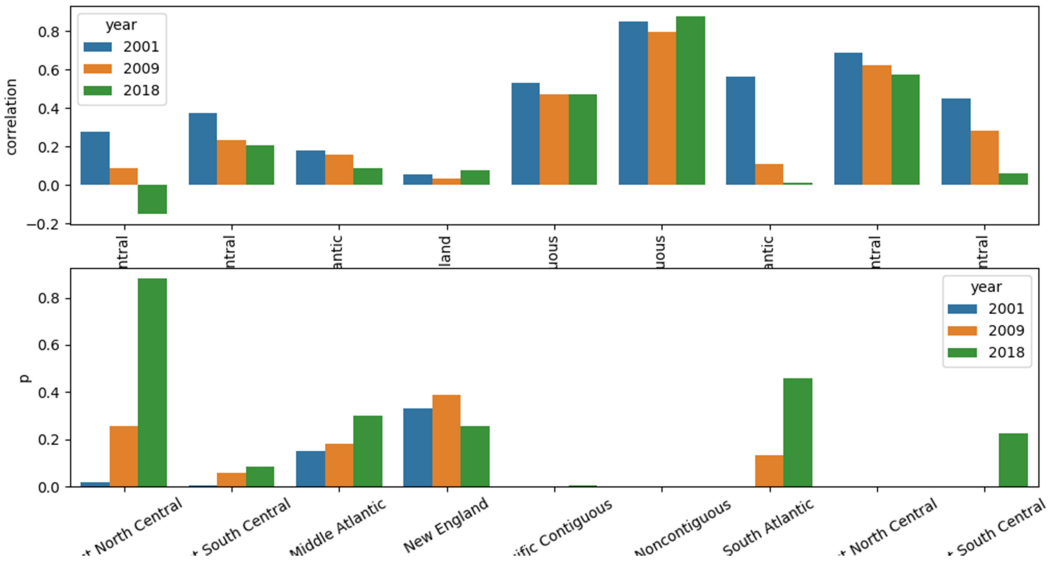
It is observed that there is a tendency in electricity price increase. No remarkable outliers were identified. The regions with higher average price variance are the pacific continuous and non-continuous regions.

Figure 7. Analysis of Average Retail Price of Electricity by Region



A correlation between average temperature and price was visualized in Figure 3. Therefore a fast calculation of the aggregated USA temperatures (TMAX, TAVG and TMIN) was done, finding that TMIN (Minimum temperature) has the higher correlation. In this part, the correlation between TMIN and average retail price of electricity was quantified, as well as, the probability that such correlation might appear as shown below. The probability was found by bootstrapping the data per region per year.

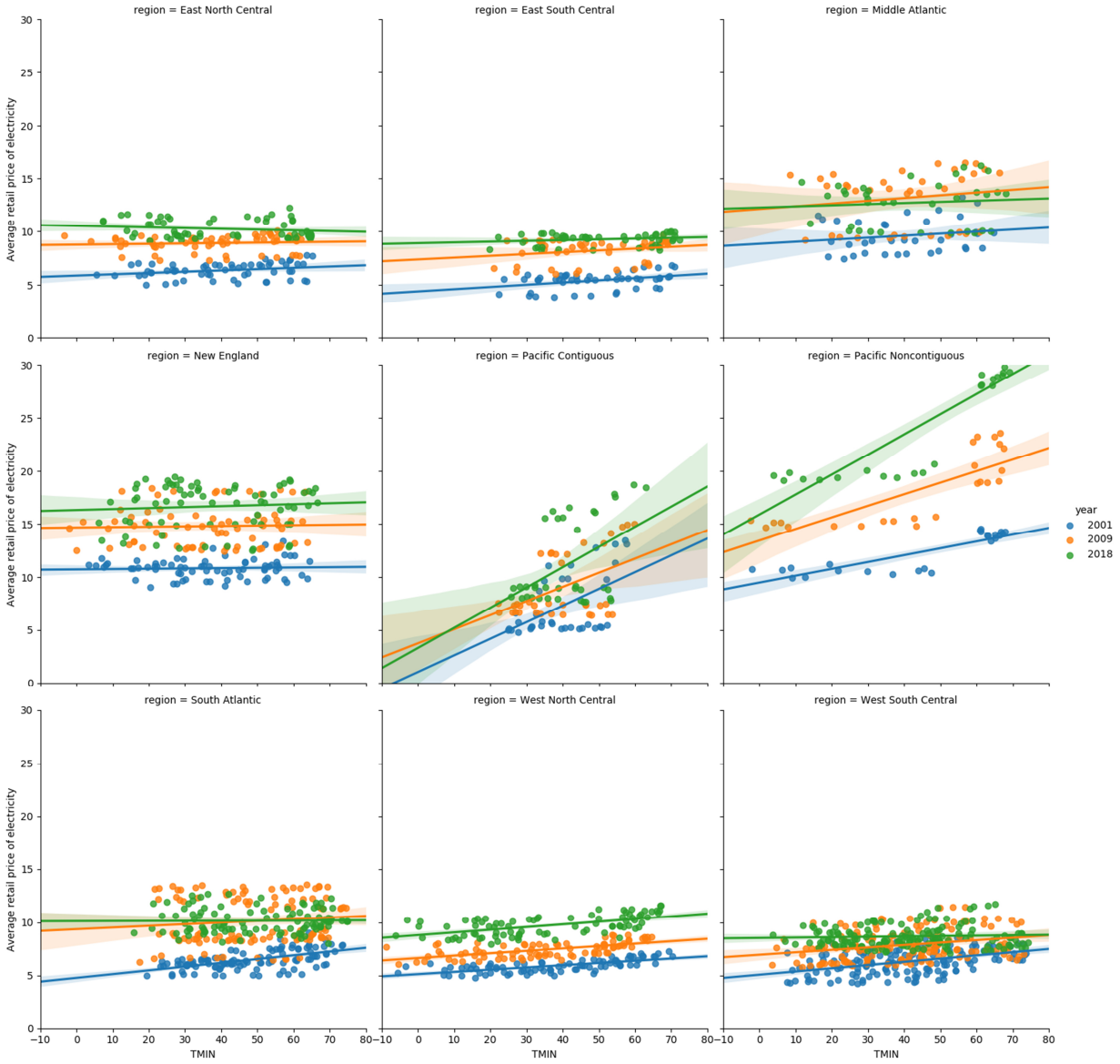
Figure 8. Correlation and probability of occurrence of such correlation per region



The findings indicate that low correlation between TMIN and Retail Price is something that is likely to happen but high correlation does not seen to be very likely.

Even though the best suitable model and features will be determined later with neural network, linear regression by using ‘seaborn’ and numpy were performed by region per year as shown below.

Figure 9. Linear regression by using 'Seaborn plot' of Average Retail Price of Electricity and TMIN



The

	region	year	rmse
0	East North Central	2001	0.662384
1	East North Central	2009	0.726542
2	East North Central	2018	0.828491
3	East South Central	2001	0.710874
4	East South Central	2009	1.012675
5	East South Central	2018	0.516796
6	Middle Atlantic	2001	1.516029
7	Middle Atlantic	2009	2.586864
8	Middle Atlantic	2018	2.034192
9	New England	2001	0.870373
10	New England	2009	1.848804
11	New England	2018	2.049188
12	Pacific Contiguous	2001	2.439026
13	Pacific Contiguous	2009	2.632352
14	Pacific Contiguous	2018	3.427269
15	Pacific Noncontiguous	2001	0.942331
16	Pacific Noncontiguous	2009	2.006429
17	Pacific Noncontiguous	2018	2.306196
18	South Atlantic	2001	0.738421
19	South Atlantic	2009	2.247671
20	South Atlantic	2018	1.103539
21	West North Central	2001	0.421232
22	West North Central	2009	0.541789
23	West North Central	2018	0.748887
24	West South Central	2001	1.012611
25	West South Central	2009	1.327254
26	West South Central	2018	0.996461