COSC 587 Data Analytics

Dr. Lisa Singh

Yang Chen

NetID: yc840

November 3, 2019

The Analysis of Industrial Energy Consumption

The purpose of this analysis is to predict the industrial energy consumption of a city when some attributes of the city are given.

Preprocess and Clean

The data from project 1 that is used in this analysis and does more further processing and cleaning. For industrial energy data, these attributes – city, state_abbr and num_establishments (number of industries the city) are selected from the original industrial data set, the state and city are the key to be used to locate a unique city. Then the attributes - elec_1kdollars, elec_mwh, gas_1kdollars, gas_mcf, elec_lb_ghg, gas_lb_ghg are selected from other-member dataset that attributes can impact the predication of consumption of energy in a city.

- elec_1kdollars: the expenditure of electricity.
- elec_mwh: total industrial electricity consumption.
- gas_1kdollars: the expenditure of natural gas.
- gas_mcf: total industrial natural gas consumption.
- elec_lb_ghg: the greenhouse gas produced by electricity.
- gas_lb_ghg: the greenhouse gas produced by natural gas.

Then the attributes - electricity_use_per_establishment, natural_gas_use_per_establishment and building_num are selected to calculate the mean consumption of electricity and natural gas. The total consumption of electricity/natural gas divide by the number of building that will get the mean consumption of electricity and natural gas.

- electricity_use_per_establishment: the consumption of electricity of an industry.
- natural_gas_use_per_establishment: the consumption of natural gas of an industry.
- building_num: the total number of industries.

These attributes - number_of_electricity_establishments and number_of_natural_gas_establishments are selected to multiply with the mean to determine a score for the consumption of electricity and natural gas in a city.

- number_of_electricity_establishments: the number of industries which use electricity.
- number_of_natural_gas_establishments: the number of industries which use natural gas.
- elec_score: the number of industries which use electricity that multiply with the mean to determine the weight of the electricity consumption in a city.

- gas_score: the number of industries which use natural gas that multiply with the mean to determine the weight of the natural gas consumption in a city.

All final data is stored in cleaned_energy_industrial.csv. Therefore, the data contains attribute – city, state_abbr, elec_score, gas_score, num_establishments, elec_1kdollars, elec_mwh, gas_1kdollars, gas_mcf, elec_lb_ghg, gas_lb_ghg, elec_bin_group and gas_bin_group(bin groups are from binning algorithm). Then there are some duplicate values of the same city. The reason might be the data is collected with the unit of the city. When a city contains more than one zip code, they just use the data of the whole city to fill the blank. However, some different states might have cities with the same names. Therefore, we should drop duplicated rows (with different zip). After dropping all totally same rows (except zip codes), there are also some cities containing more than one row. The reason might be these statistics data are collected based on an area of a city. We should add different rows with the same state and city attributes to get the actual energy data of the whole city. Therefore, the data does not contain missing values and duplicates anymore.

Basic Statistical Analysis and Data Cleaning Insight

The mean of each attribute:

The mean of elec_score attribute represents the mean of the weight of electricity consumption in a city. The mean of gas_score attribute represents the mean of the weight of gas consumption in a city. The mean of num_establishments attribute represents the mean of the number of establishments in a city. The mean of elec_1kdollars attribute represents the mean of electricity expenditure in a city. The mean of elec_mwh attribute represents the mean of electricity consumption in a city. The mean of gas_1kdollars attribute represents the mean of natural gas expenditure in a city. The mean of gas_mcf attribute represents the mean of gas consumption in a city. The mean of elec_lb_ghg attribute represents the mean of total greenhouse gas produced by electricity in a city. The mean of gas_lb_ghg attribute represents the mean of total greenhouse gas produced by natural gas in a city.

elec_score	1.447604e+05
gas_score	9.910556e+05
num_establishments	149.261739
elec_1kdollars	4352.988798
elec_mwh	6.057753e+04
gas_1kdollars	2297.871747
gas_mcf	4.730803e+05
elec_lb_ghg	9.113894e+07
gas_lb_ghg	5.678984e+07

The median of each attribute:

The median of elec_score attribute represents the median of the weight of electricity consumption in a city. The median of gas_score attribute represents the median of the weight of gas consumption in a city. The median of num_establishments attribute represents the median of the number of establishments in a city. The median of elec_1kdollars attribute represents the

median of electricity expenditure in a city. The median of elec_mwh attribute represents the median of electricity consumption in a city. The median of gas_1kdollars attribute represents the median of natural gas expenditure in a city. The median of gas_mcf attribute represents the median of gas consumption in a city. The median of elec_lb_ghg attribute represents the median of total greenhouse gas produced by electricity in a city. The median of gas_lb_ghg attribute represents the median of total greenhouse gas produced by natural gas in a city.

elec_score	7.649316e+03
gas_score	3.996932e+04
num_establishments	50.000000
elec_1kdollars	678.000000
elec_mwh	9.253000e+03
gas_1kdollars	327.500000
gas_mcf	5.798450e+04
elec_lb_ghg	1.349629e+07
gas_lb_ghg	6.960624e+06

The standard deviation of each attribute:

The standard deviation of elec_score attribute represents the standard deviation of the weight of electricity consumption in a city. The standard deviation of gas_score attribute represents the standard deviation of the weight of gas consumption in a city. The standard deviation of num_establishments attribute represents the standard deviation of the number of establishments in a city. The standard deviation of elec_1kdollars attribute represents the standard deviation of electricity expenditure in a city. The standard deviation of elec_mwh attribute represents the standard deviation of electricity consumption in a city. The standard deviation of gas_1kdollars attribute represents the standard deviation of natural gas expenditure in a city. The standard deviation of gas_mcf attribute represents the standard deviation of gas consumption in a city. The standard deviation of elec_lb_ghg attribute represents the standard deviation of total greenhouse gas produced by electricity in a city. The standard deviation of gas_lb_ghg attribute represents the standard deviation of total greenhouse gas produced by natural gas in a city.

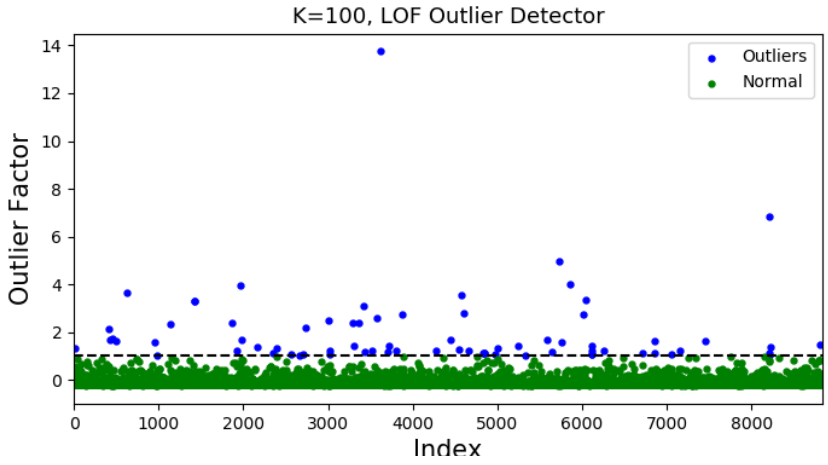
elec_score	2.575910e+06
gas_score	1.660021e+07
num_establishments	650.402945
elec_1kdollars	16844.349746
elec_mwh	2.266995e+05
gas_1kdollars	9312.090219
gas_mcf	2.388090e+06
elec_lb_ghg	3.448248e+08
gas_lb_ghg	2.866727e+08

Outlier Detection:

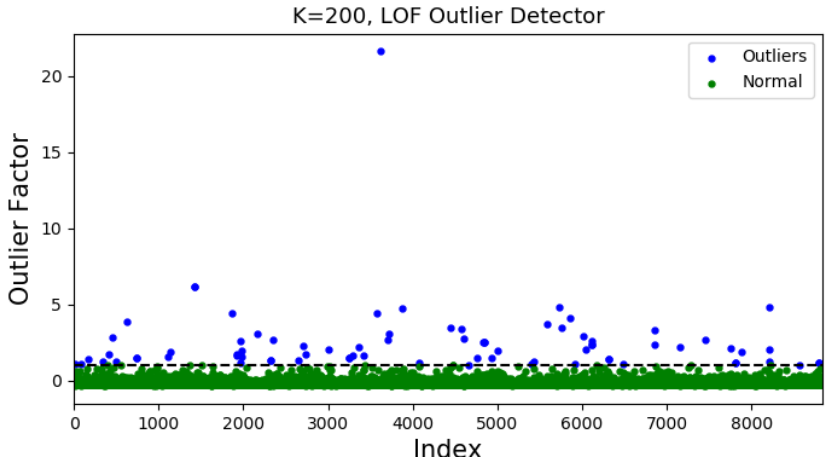
LOF algorithm is used to detect outliers in this data set. The k-values are 100, 200 and 500 are selected, and bin to k+1 bin. And the LOF is set to 1 to determine outliers. The outliers are recorded in the files outlier_100.csv, outlier_200.csv, and outlier_500.csv. I chose k-

value=200 as my value to detect outliers because the value is best for the current size of the data set and the distance of every outlier is more densely distributed in the plot. I chose to remove all outliers because the energy data cannot be assigned by personal and the size of the data set is enough to analyze.
k=100: 73 outliers
k=200: 80 outliers
k=500: 195 outliers

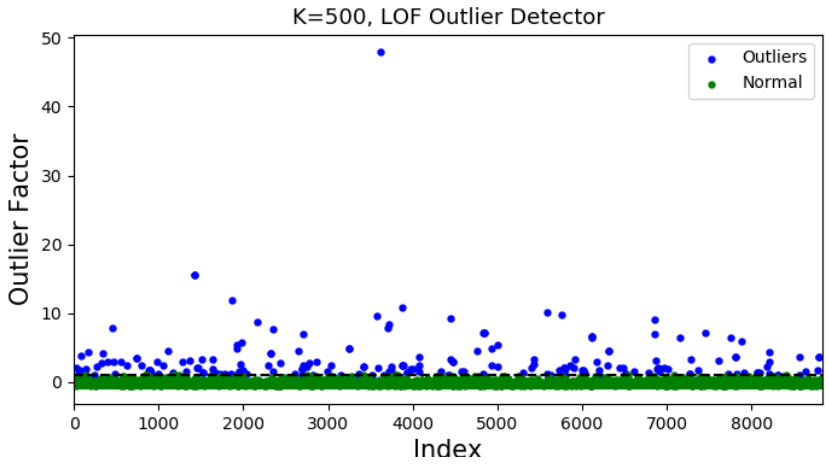
When k=100:



When k=200:



When k-500:



After LOF outlier detection and remove outliers, I also calculated the mean, median and standard derivation for the data set. The standard derivation is shrinking a lot.

Binning algorithm:

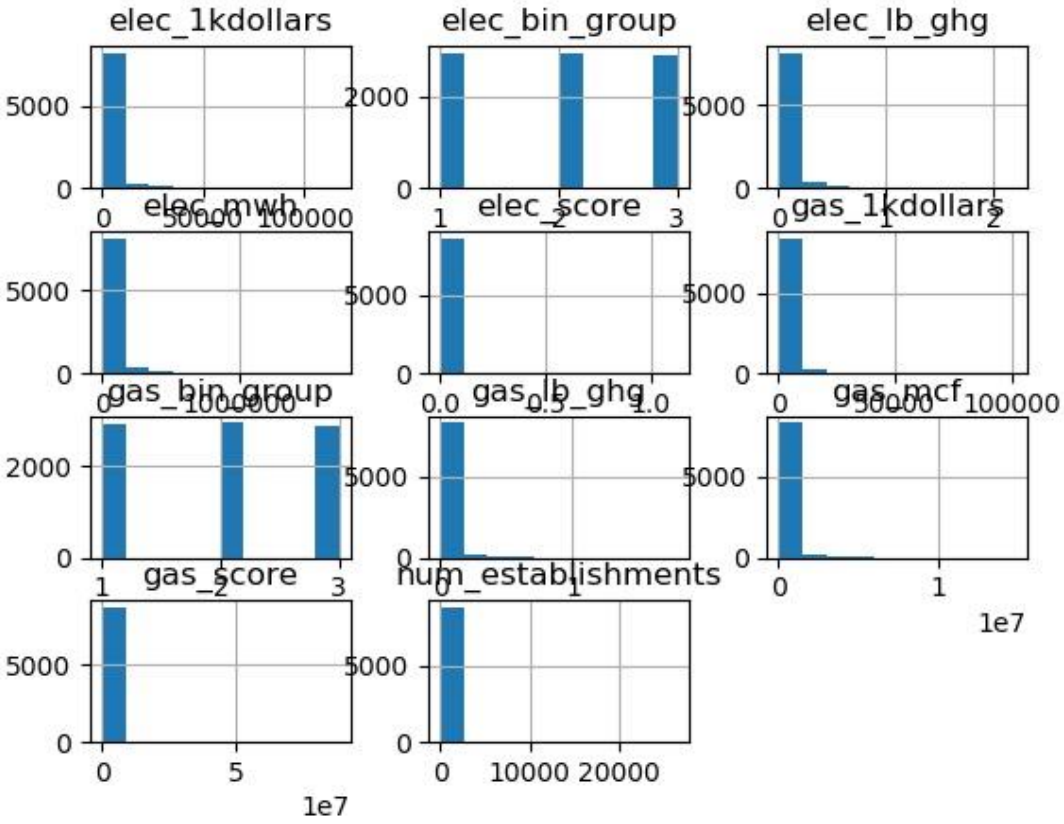
I chose the equal-depth binning algorithm to bin the data set by the attribute elec_mwh for electricity and the attribute gas_mcf for natural gas. The equal-width binning algorithm is not good for the data set because most values will be binned to the same one or two bins and many bins are empty or contain few values. However, the equal-depth binning algorithm will make all bins contain almost the same number of elements and the mean and median are good in each bin. 3 bins are determined by the equal-depth algorithm and separated the data set to three class by the consumption of electricity and natural gas in a city. Therefore, people could know the type of city in industrial energy consumption. For example, Pittsburgh, PA is a very high electricity consumption and a very high natural gas consumption city. It is a heavy industry city.

Bin Groups	Description
Bin group 1	Normal
Bin group 2	High
Bin group 3	Very High

Missing values: The data contains no missing values.

Histograms and Correlations

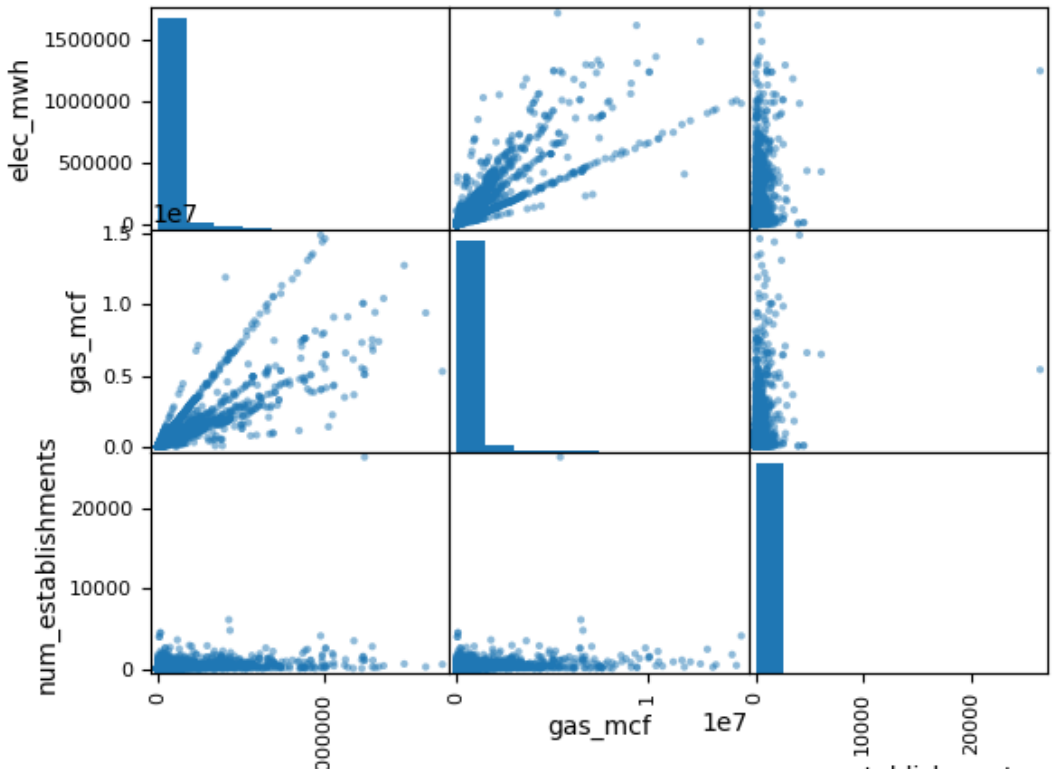
I plot the histograms containing all features. These variables all have a similar distribution and are all skewed right except bin groups. The reason is that all other features are positively correlated with elec_mwh and gas_mcf.



I identify num_establishments, elec_mwh, and gas_mcf for finding the correlation.

	elec_mwh	gas_mcf	num_establishments
elec_mwh	1.000000	0.861399	0.363124
gas_mcf	0.861399	1.000000	0.345179
num_establishments	0.363124	0.345179	1.000000

This table shows that num_establishments is weakly correlated to elec_mwh and gas_mcf. but the correlation between elec_mwh and gas_mcf is a strong positive. This is reasonable that an industry uses the most consumption of energy is either electricity or natural gas. And I also plot a set of scatter plot subplots. It also shows the same pattern as the correlation matrix.



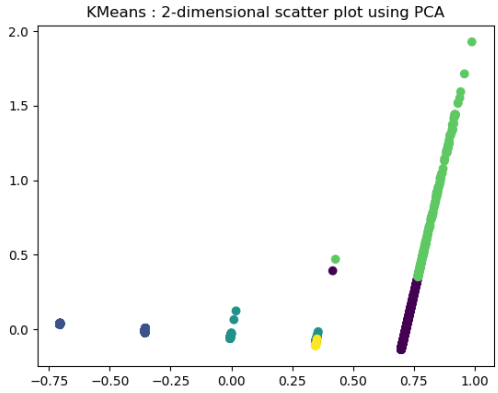
Cluster Analysis

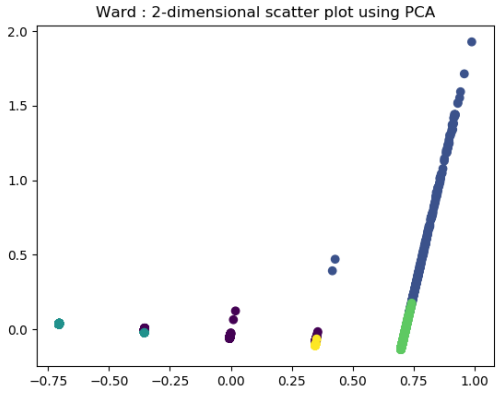
K-Means, Ward and DBScan are three cluster algorithms that are used in this analysis.

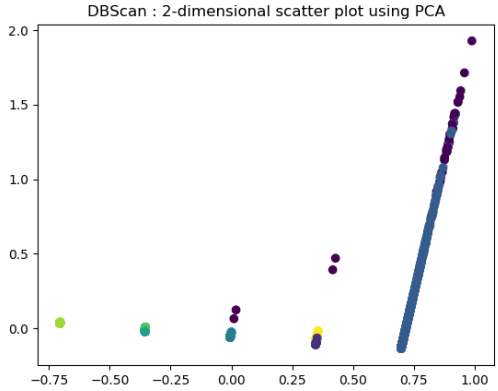
Clusters	Silhouette Score
K-Means	0.7523670413628896
Ward	0.7396858012084296
DBScan	0.8777150462943293

The Silhouette score is higher that other two clusters. That means DBScan are the best cluster for this industrial data set

Here are the PCA plots. Observing these PCA plots, the clustering results of k-means and ward algorithms are almost similar, the majority of the data set is tight. The DBScan also looks similar to other clusters but the data set is tighter than the other two clusters.







Association Rules

The attributes - elec_bin_group, elec_bin_group are selected to form the subset. Before calling the Apriori algorithm to calculate support, some transformation is necessary. Because of the bin group of electricity and gas are both numeric, they are hard to recognize in the result. Therefore, I transform the two attributes into a string type. For example,

	elec_class	gas_class
0	elec_class_high	gas_class_very_high
1	elec_class_very_high	gas_class_very_high
2	elec_class_high	gas_class_high
3	elec_class_very_high	gas_class_very_high
4	elec_class_normal	gas_class_high
5	elec_class_very_high	gas_class_very_high
6	elec_class_high	gas_class_normal
7	elec_class_high	gas_class_very_high
8	elec_class_very_high	gas_class_very_high
9	elec_class_very_high	gas_class_very_high

When min_support = 0.05

	support	itemsets
0	0.336036	(elec_class_high)
1	0.334665	(elec_class_normal)
2	0.329299	(elec_class_very_high)
3	0.336264	(gas_class_high)
4	0.334209	(gas_class_normal)
5	0.329527	(gas_class_very_high)
6	0.269125	(elec_class_high, gas_class_high)
7	0.301210	(elec_class_normal, gas_class_normal)
8	0.295387	(gas_class_very_high, elec_class_very_high)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	convictio
0	(elec_class_high)	(gas_class_high)	0.336036	0.336264	0.269125	0.800883	2.381710	0.156129	3.33340
1	(gas_class_high)	(elec_class_high)	0.336264	0.336036	0.269125	0.800340	2.381710	0.156129	3.32546
2	(elec_class_normal)	(gas_class_normal)	0.334665	0.334209	0.301210	0.900034	2.693030	0.189362	6.66018
3	(gas_class_normal)	(elec_class_normal)	0.334209	0.334665	0.301210	0.901264	2.693030	0.189362	6.73852
4	(gas_class_very_high)	(elec_class_very_high)	0.329527	0.329299	0.295387	0.896396	2.722136	0.186874	6.47372
5	(elec_class_very_high)	(gas_class_very_high)	0.329299	0.329527	0.295387	0.897018	2.722136	0.186874	6.51058

When min_support = 0.1

	support	itemsets
0	0.336036	(elec_class_high)
1	0.334665	(elec_class_normal)
2	0.329299	(elec_class_very_high)
3	0.336264	(gas_class_high)
4	0.334209	(gas_class_normal)
5	0.329527	(gas_class_very_high)
6	0.269125	(elec_class_high, gas_class_high)
7	0.301210	(elec_class_normal, gas_class_normal)
8	0.295387	(gas_class_very_high, elec_class_very_high)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	convictio
0	(elec_class_high)	(gas_class_high)	0.336036	0.336264	0.269125	0.800883	2.381710	0.156129	3.33340
1	(gas_class_high)	(elec_class_high)	0.336264	0.336036	0.269125	0.800340	2.381710	0.156129	3.32546
2	(elec_class_normal)	(gas_class_normal)	0.334665	0.334209	0.301210	0.900034	2.693030	0.189362	6.66018
3	(gas_class_normal)	(elec_class_normal)	0.334209	0.334665	0.301210	0.901264	2.693030	0.189362	6.73852

When min_support = 0.3

	support	itemsets
0	0.336036	(elec_class_high)
1	0.334665	(elec_class_normal)
2	0.329299	(elec_class_very_high)
3	0.336264	(gas_class_high)
4	0.334209	(gas_class_normal)
5	0.329527	(gas_class_very_high)
6	0.301210	(elec_class_normal, gas_class_normal)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(elec_class_normal)	(gas_class_normal)	0.334665	0.334209	0.30121	0.900034	2.69303	0.189362	6.660185
1	(gas_class_normal)	(elec_class_normal)	0.334209	0.334665	0.30121	0.901264	2.69303	0.189362	6.738527

The (elec_class_normal, gas_class_normal) are most frequent. This result is not surprising me, because the industrial usage amounts of most cities are always normal in the USA. The heavy industry city is few in the USA.

Predictive Analysis

T-test:

There is the first hypothesis - the difference between the mean of elec_bin_group and gas_bin_group attributes are not obvious.

The t-test result is:

Ttest_indResult(statistic=-0.05563937437162836, pvalue=0.955629731523744)

The p-value is greater than 0.05? True

From the result, that shows the first hypothesis is supported. Therefore, the difference between the mean of elec_bin_group and gas_bin_group attributes is not obvious. T-test is good for paired sample to test in this case.

There is second hypothesis - the difference between the mean of elec_score and gas_score attributes are not obvious.

The t-test result is:

Ttest_indResult(statistic=-16.468126914649144, pvalue=1.7708281169593017e-60)

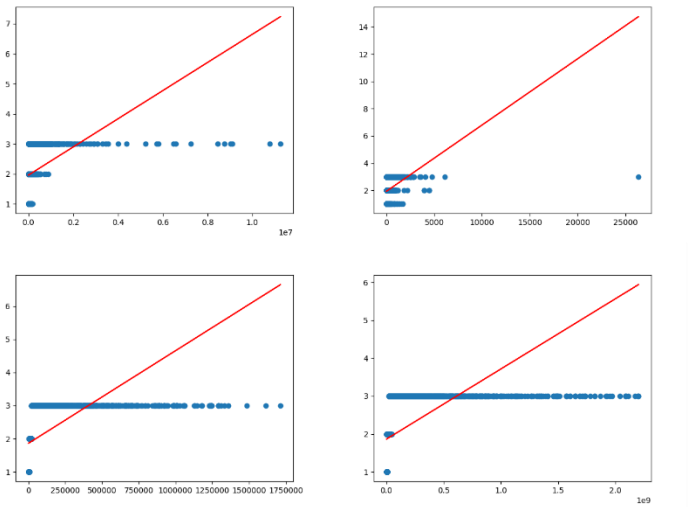
The p-value is greater than 0.05? False

From the result, that shows the second hypothesis is failed. Therefore, the difference between the mean of elec_score and gas_score attributes is obvious. The T-test is good for the paired samples to test in this case.

Linear Regression:

There is third hypothesis - the elec_score, num_establishments, elec_1kdollars, elec_mwh, and elec_lb_ghg are positively related to elec_bin_group.

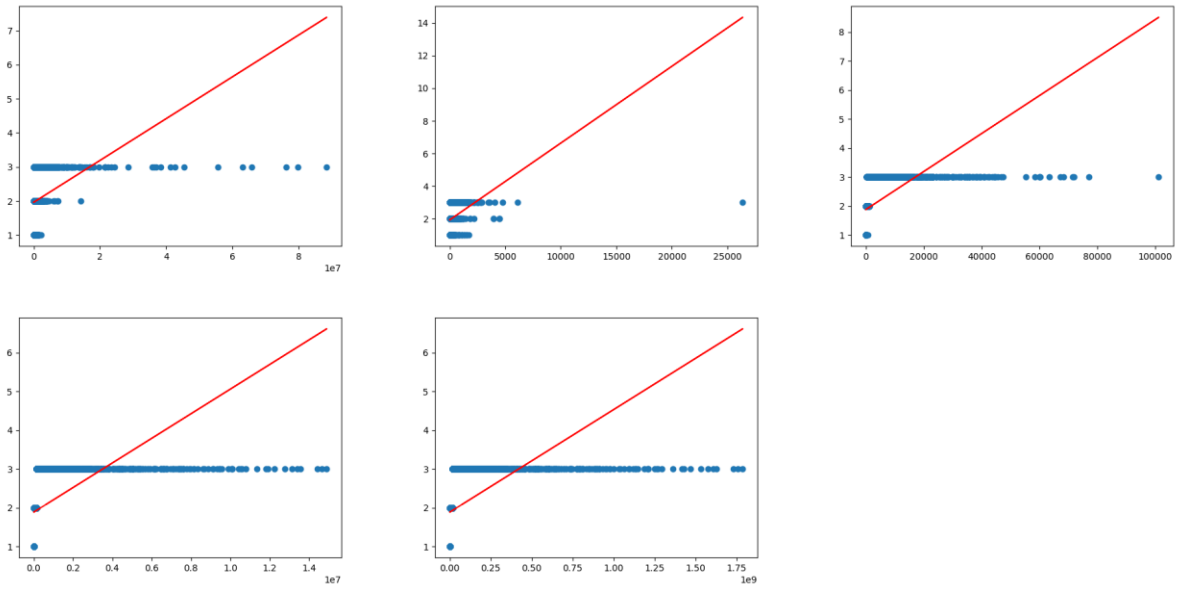
The linear regression result is:



The linear graph is not very good to present because some values impact the linear. But people notice most values are located in one horizontal line above except the num_establishments. People can know this supports the hypothesis. Therefore, the elec_score, elec_1kdollars, elec_mwh, and elec_lb_ghg are positively related to elec_bin_group. Linear regression can be easier to see the relations between features.

There is forth hypothesis - the gas_score, num_establishments, gas_1kdollars, gas_mwh, and gas_lb_ghg are positively related to gas_bin_group.

The linear regression result is:



The linear graph is not very good to present because some values impact the linear. But people notice most values are located in one horizontal line above except the num_establishments. People can know this supports the hypothesis. Therefore, the gas_score, gas_1kdollars, gas_mwh, and gas_lb_ghg are positively related to gas_bin_group. Linear regression can be easier to see the relations between features.

Classification

I separated data set to electricity subset(attributes: elec_score, num_establishments, elec_1kdollars, elec_mwh, elec_lb_ghg and elec_bin_group), and natural gas subset(the gas_score, num_establishments, gas_1kdollars, gas_mwh, gas_lb_ghg and gas_bin_group).

Electricity subset:

Evaluation of machine learning methods:

Classifiers	Average mean of accuracy	Average standard deviation of accuracy
KNN	0.920498	0.012537
CART	0.999857	0.000428
Naive Bayes	0.889095	0.014598
SVM	0.912788	0.009911
Random Forest	0.999857	0.000428

Prediction for validation data:

Classifiers	Accuracy Score
KNN	0.9212328767123288
CART	0.9994292237442922
Naive Bayes	0.884703196347032
SVM	0.9178082191780822
Random Forest	0.9994292237442922

Natural gas subset:

Evaluation of machine learning methods:

Classifiers	Average mean of accuracy	Average standard deviation of accuracy
KNN	0.999001	0.001114
CART	0.999714	0.000571
Naive Bayes	0.937767	0.007925
SVM	0.930492	0.013938
Random Forest	0.999714	0.000571

Prediction for validation data:

Classifiers	Accuracy Score
KNN	0.9994292237442922
CART	1
Naive Bayes	0.942351598173516
SVM	0.9263698630136986
Random Forest	1

According to the records of electricity and natural gas subset, the CART and Random Forest are both best for two subsets. The accuracies of CART and Random Forest are closed to 1, sometimes the accuracies can be 1. The KNN is not very bad, and the accuracy of KNN is the top 3. This also shows that the chosen attributes are strongly related with the class labels.