**Project Assignment 2**

**Name: Dandan Wang**

**NetID: dw862**

The **goal** of this part is to predict the residential energy consumption levels of a city when some attributes of the city are given.

* **Preprocess**

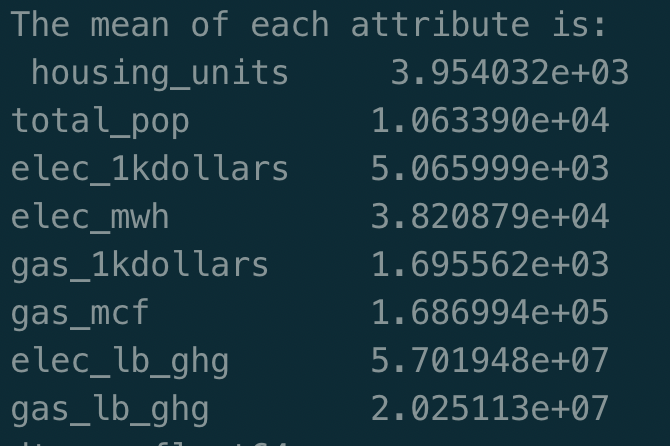
Before analysis, it is necessary to extract the useful attributes for analysis. For residential energy data, choose ‘state\_id’, ‘city’, ‘housing\_units’, ‘total\_pop’ (total population), ‘elec\_mwh’ (total residential electricity consumption), ‘gas\_mcf’ (total residential natural gas consumption), ‘elec\_1kdollars’ (the expenditure of electricity), ‘gas\_1kdollars’ (the expenditure of natural gas), ‘elec\_lb\_ghg’ (greenhouse gas produced by electricity) and ‘elec\_lb\_ghg’ (greenhouse gas produced by natural gas). The state, city attributes are used to locate a unique city. The amount of housings, population are both relevant to the electricity and gas consumptions. Besides, according to the expenditures in electricity (elec\_1kdollars) and gas (gas\_1kdollars), it is easy to derive the unit-price of electricity and natural gas in different areas. The unit-prices might have impacts on the electricity and natural gas consumptions as well. Because people might prefer to use the cheaper energy. In addition, the elec\_lb\_ghg and gas\_lb\_ghg attributes are also kept because some people might prefer to use cleaner energy.

* **Clean**

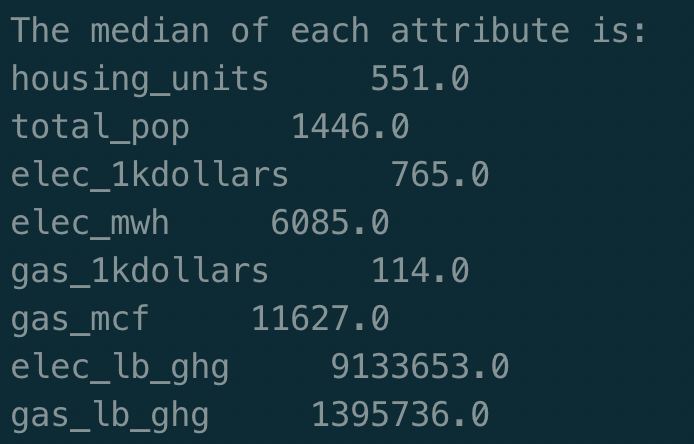
Reanalyzing the cleaned data in project 1, there are some duplicate values of the same city. The reason might be the data is collected with the unit of city. When a city contains more than one zip code, they just use the data of the whole city to fill the blank. However, some different states might have cities with the same names. Therefore, it is reasonable to 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. Different rows with same state and city attributes should be added to get the actual energy data of the whole city. The processed data is stored in “cleaned\_residential.csv”.

* **Basic Statistical Analysis and data cleaning insight.**

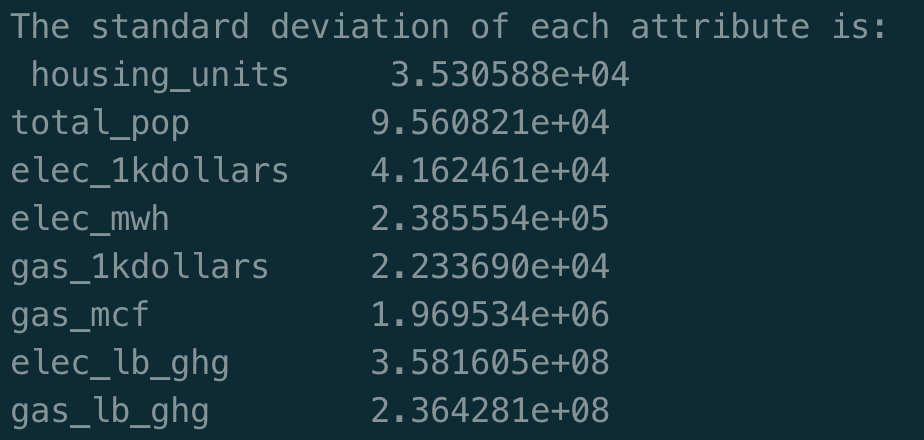
The mean of housing\_units attribute represents the average amount of housings in a city. The mean of total\_pop attribute represents the average population in a city. The mean of elec\_1kdollars attribute represents the average electricity expenditure in a city. The mean of elec\_mwh attribute represents the average electricity consumption in a city. The mean of gas\_1kdollars attribute represents the average natural gas expenditure in a city. The mean of gas\_mcf attribute represents the average gas consumption in a city. The mean of elec\_lb\_ghg attribute represents the average amount of greenhouse gas produced by electricity in a city. The mean of gas\_lb\_ghg attribute represents the average amount of greenhouse gas produced by natural gas in a city.



The median of housing\_units attribute represents the median of total housings in a city. The median of total\_pop attribute represents the median of population 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 i 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 meidan 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.



The standard deviation of ousing\_units attribute represents the discrete degree of total housings in a city. The standard deviation of total\_pop attribute represents the discrete degree of population. The standard deviation of elec\_1kdollars attribute represents the discrete degree of electricity expenditure. The standard deviation of elec\_mwh attribute represents the discrete degree of electricity consumptions. The standard deviation of gas\_1kdollars attribute represents the discrete degree of natural gas expenditure. The standard deviation of gas\_mcf attribute represents the discrete degree of gas consumption. The standard deviation of elec\_lb\_ghg attribute represents the discrete degree of total greenhouse gas produced by electricity. The standard deviation of gas\_lb\_ghg attribute represents the discrete degree of total greenhouse gas produced by natural gas.



* **Detect Outliers**

Although these cities have different scales, it is not reasonable for a city to have totally different values from most cities. The types of ‘housing\_units’, ‘total\_pop’, ‘elec\_1kdollars’, ‘elec\_mwh’, ‘gas\_1kdollars’, ‘gas\_mcf’, ‘elec\_lb\_ghg’, ‘gas\_lb\_ghg’ are all continuous numeric. Therefore, it is obvious to use a simple binning method to select the outliers firstly. Divide the range of each attribute into 100 even bins and output the count of each bin. Please refer to output.txt to get the detailed output. The distributions of all attributes have a same pattern: most attributes gather at the the 1st bins. Thus, the data in other bins should be regarded as outliers. And summing up the counter in outlier bins can calculate the percentage of outliers of each attribute. Please refer to output.txt to get the percentage. The percentage shows that the percentage of each attribute is pretty low. Use the maximal value of non-outliers to replace the values of outliers. The outliers would be classified as the very high-level energy consumption in this way. It is reasonable that these outlier have much more energy consumptions than other cities. The processed data by binning method is stored into “binned\_residential.csv”.

* **Missing Values**

The missing values have already been cleaned in Project 1. Replaced the missing values with its neighbors. Because the neighbors in the raw data are also close to each other in reality. Therefore, the scale, energy consumption and expenditure should be similar as well. It is reasonable to use the mean of neighbors to fill the missing value. For the raw data, the fraction of missing values in each column is pretty low.

housing\_units 0.0148%

total\_pop 0.0162%

elec\_mwh 0.0148%

gas\_mcf 0.0148%

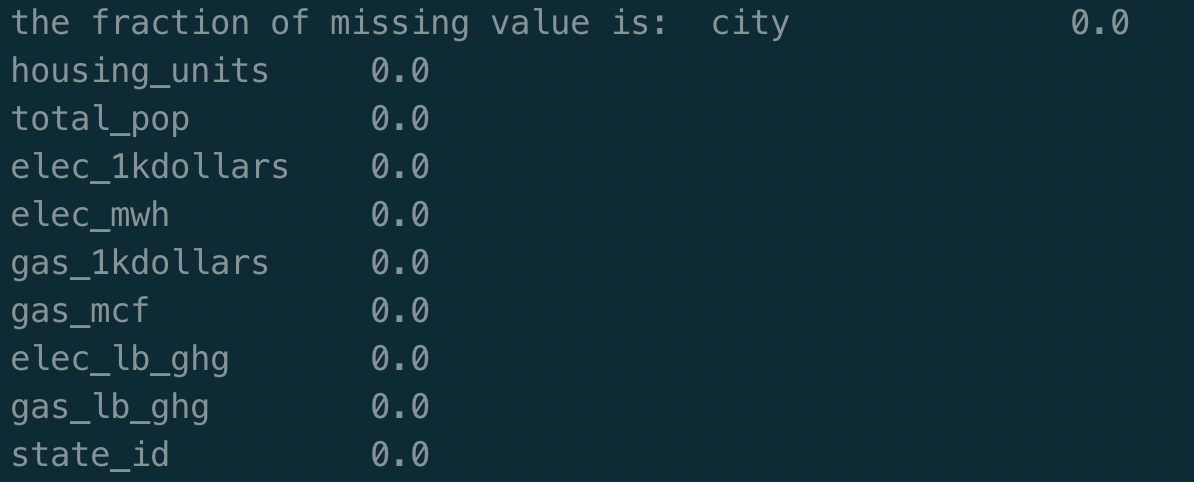
elec\_1kdollars 0.0148%

elec\_1kdollars 0.0148%

gas\_lb\_ghg 0.0148%

elec\_lb\_ghg 0.0148%

The number of rows in the raw data is 67732. There are only tens of rows containing missing values. Of course, deleting these rows directly is also ok. They have few impacts on the final results. Besides, a function named “findMissing” is designed to detect the fraction of missing values in the data frame. The output shows there is no missing value in the current data frame.



* **Bin Data**

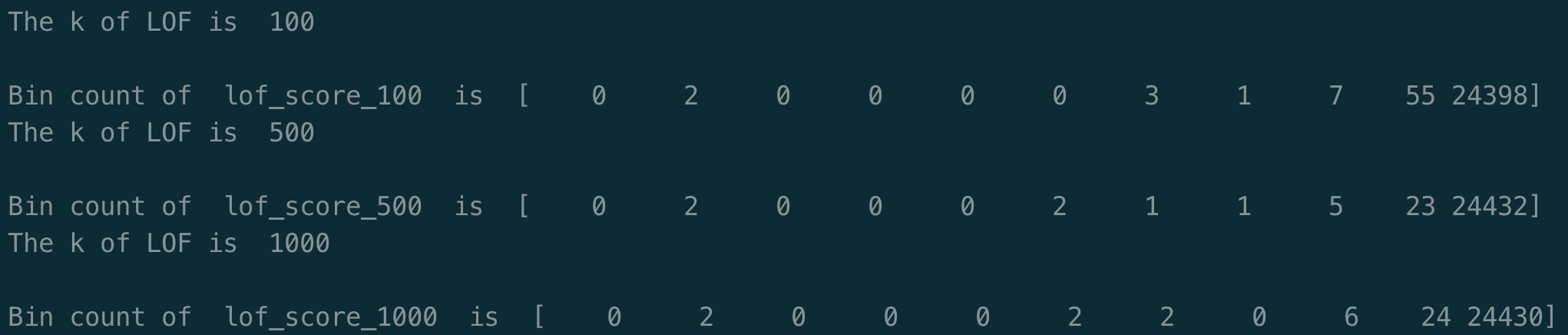
Bin two attributes – ‘elec\_mwh’ and ‘gas\_mcf’. The reason for binning them is intuitive. Because the labels of energy consumption are based on the electricity consumptions and the natural gas consumptions. And binning them into three even bins also gets the degrees of energy consumptions. For example, the city whose elec\_degree is 1 has normal electricity consumption level. There are 3 degrees for electricity and natural gas consumptions individually.

|  |  |
| --- | --- |
| degree | representation |
| 1 | Normal |
| 2 | High |
| 3 | Very High |

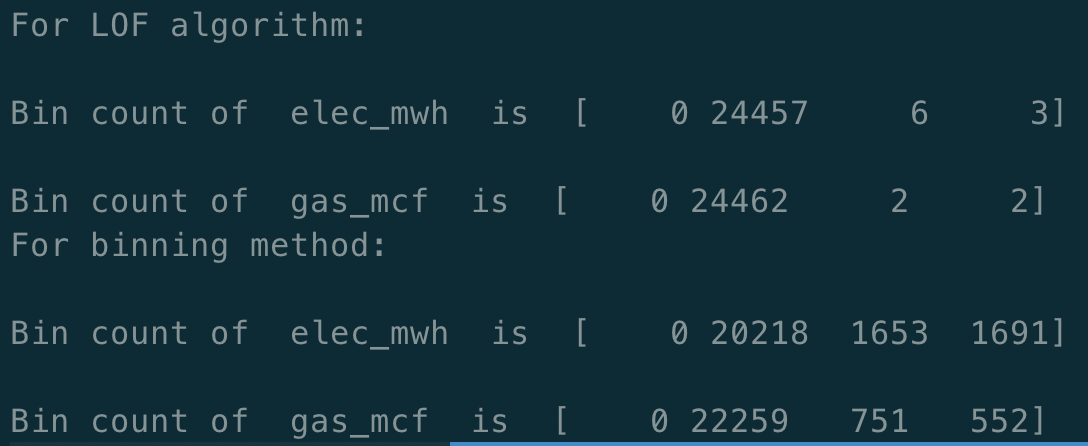
According to the degree, we can know the type of one city in residential energy consumption. For example, Suffield is a normal electricity consumption and a normal natural gas consumption city. The new data is stored in “degreed\_residential.csv”.

* **Additional Part for CS students**

Use 3 different k values - [100, 500, 1000] and bin the scores into 10 bins. For different k, the major bin is always the 10th bin. However, the count of data located in the 10th bin changes a little. The 10th bin contains maximal negative outlier factor whose absolute value is smallest. Therefore, data in 10th bin are normal data not outliers. We should abandon data located in other bins. The processed data is stored in “lof\_residential.csv”. The fraction of outliers is really low. I drop all possible outliers based on the three outlier factor attributes.

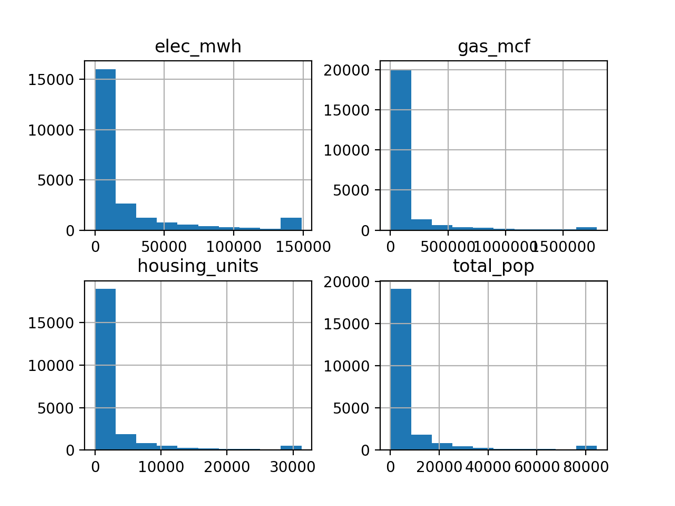


To compare above binning method with LOF, bin the two attributes ‘elec\_mwh’ and ‘gas\_mcf’ into 3 bins and print the counts of all bins. It is obvious that binning method is more suitable for this data set.



* **Histograms and Correlations**

The histograms contains 4 variables – ‘housing\_units’, ‘total\_pop’, ’ ‘elec\_mwh’, ‘gas\_mcf’. These variables all have similar distribution and are all skewed right. The reason is that ‘housing\_units’ and ‘total\_pop’ are positive correlated with ‘elec\_mwh’ and ‘gas\_mcf’.

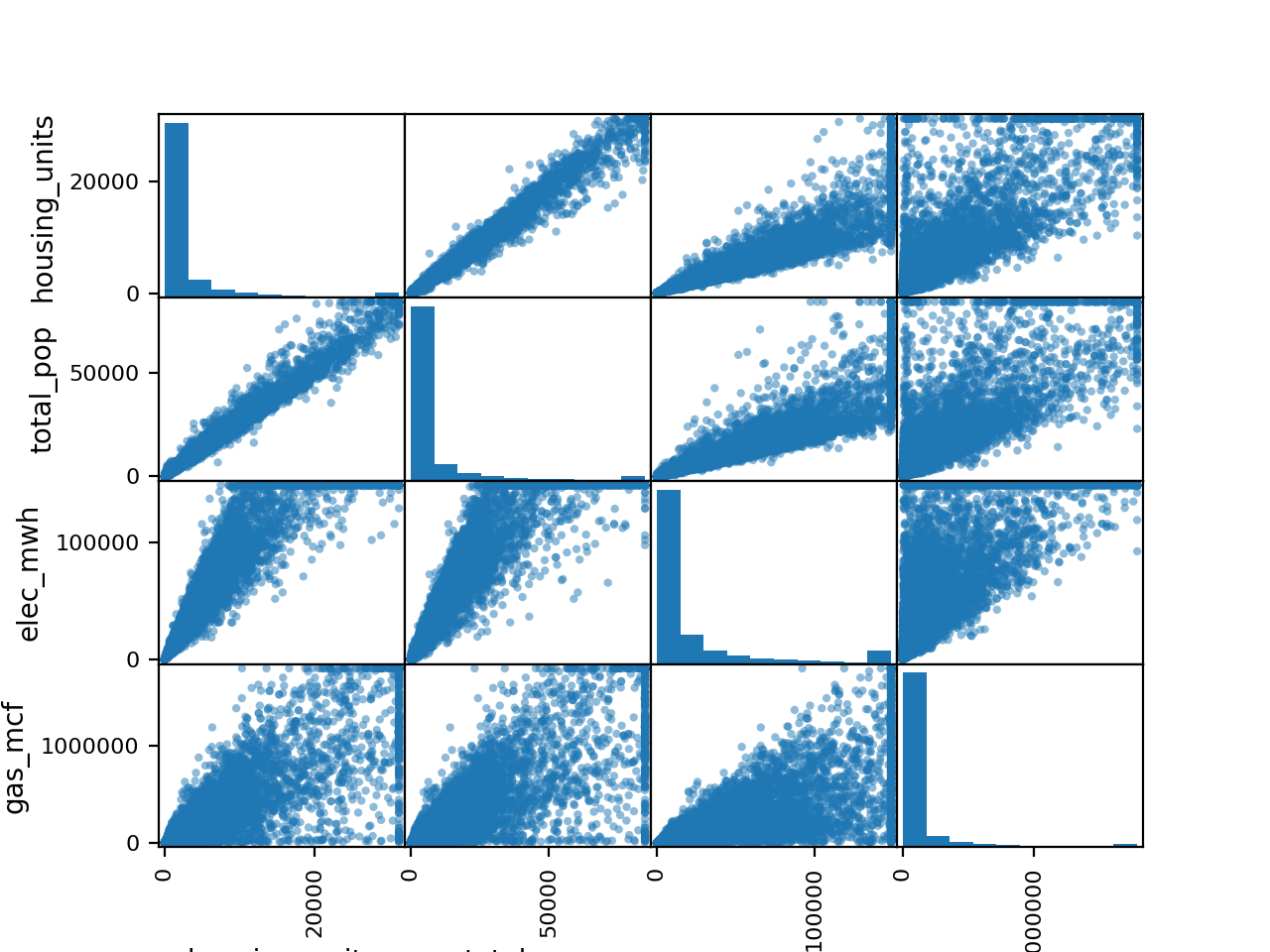


Here is the output table of correlations between the 4 variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | housing\_units | total\_pop | elec\_mwh | gas\_mcf |
| housing\_units | 1.000000 | 0.994423 | 0.923064 | 0.884154 |
| total\_pop | 0.994423 | 1.000000 | 0.917678 | 0.877276 |
| elec\_mwh | 0.923064 | 0.917678 | 1.000000 | 0.796833 |
| gas\_mcf | 0.884154 | 0.877276 | 0.796833 | 1.000000 |

This table shows that ‘housing\_units’ and ‘total\_pop’ are both strong positive correlated to ‘elec\_mwh’ and ‘gas\_mcf’. This means the amount of housing and population are strong positive correlated to energy consumptions. Besides, the correlation between housing and population is also strong positive. This is obvious. However, the correlation between ‘elec\_mwh’ and ‘gas\_mcf’ is not as strong as others. This is reasonable that a city whose electricity consumption is high level may not be a city whose natural gas consumption is also high level.

Here is a set of scatterplot subplots. It also shows the same patter as the correlation matrix.

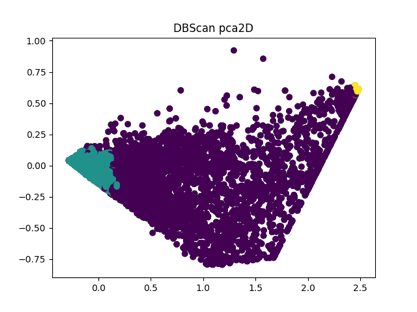
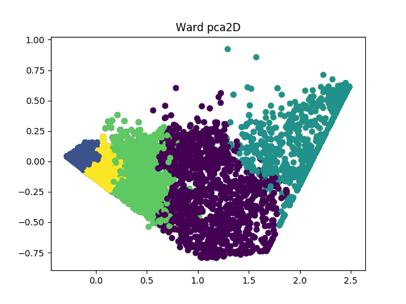
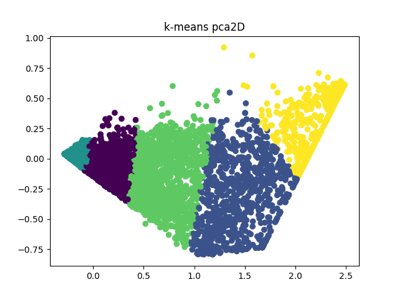


* **Cluster Analysis**

Here is the clustering result. According to silhouette score, DBSCAN is closer to 1. It performs best. While k-means performs best according to calinski harabasz score.

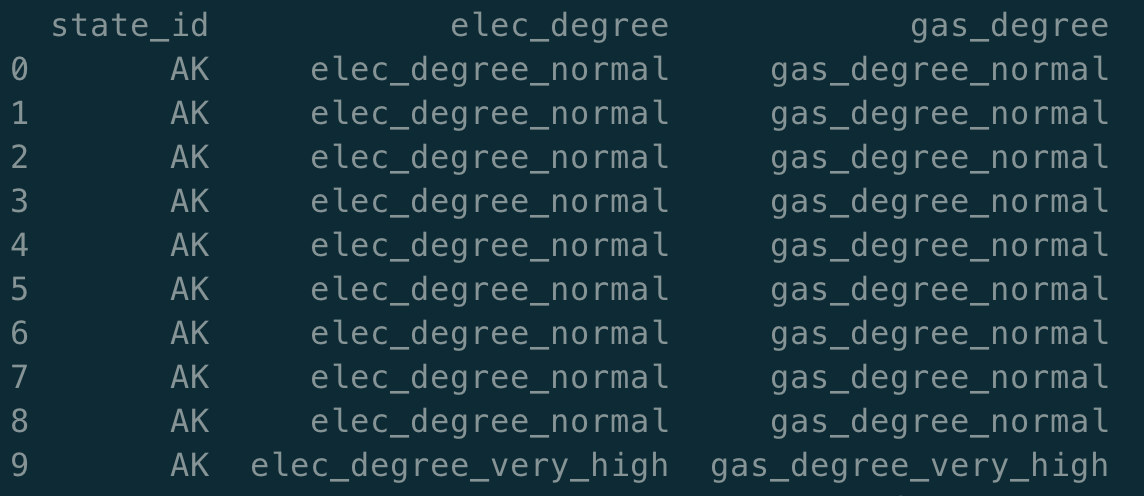
|  |  |  |
| --- | --- | --- |
|  | silhouette\_score | calinski harabasz score |
| k-means | 0.6680704850340499 | 61496.10131248855. |
| Ward | 0.6195685793092233 | 52046.17918418183. |
| dbscan | 0.7344738624534686 | 25606.67619663513. |

Here is the PCA plots. Observing these PCA plots, the clustering results of k-means and ward algorithm are similar except the yellow and dark blue points. The majority of data set is tight. The number of clusters becomes fewer by using dbscan algorithm. The plots show that k-means is more suitable for this data set.

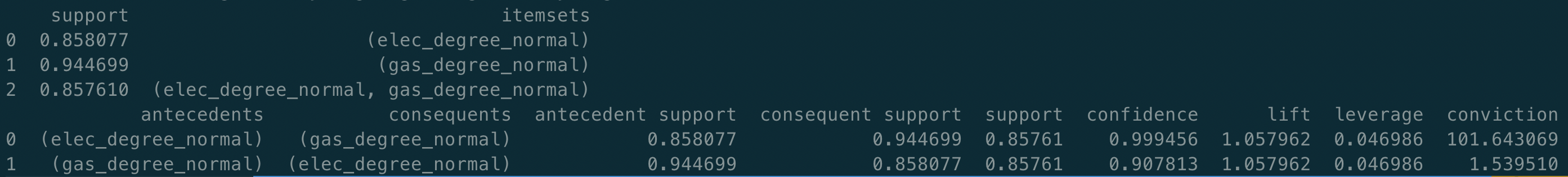


* **Association Rules**

The subset contains 3 columns – ‘state\_id’, ‘elec\_degree’, ‘gas\_degree’. Before calling Apriori algorithm to calculate support, some transformation is necessary. Because the degree of electricity and gas are both 1 to 5. If use the raw subset to calculate, the algorithm would regard two attributes same. Therefore, it is necessary to transform the two attributes into string type. For example,

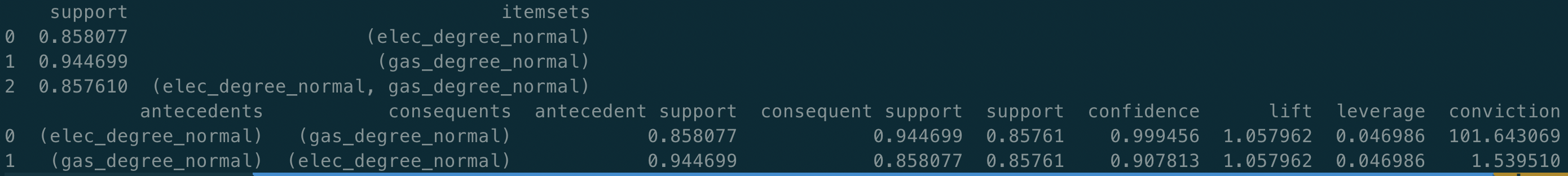


Use three support levels – 0.6, 0.1, 0.05. When support = 0.6 and confidence = 0.6, the output is:

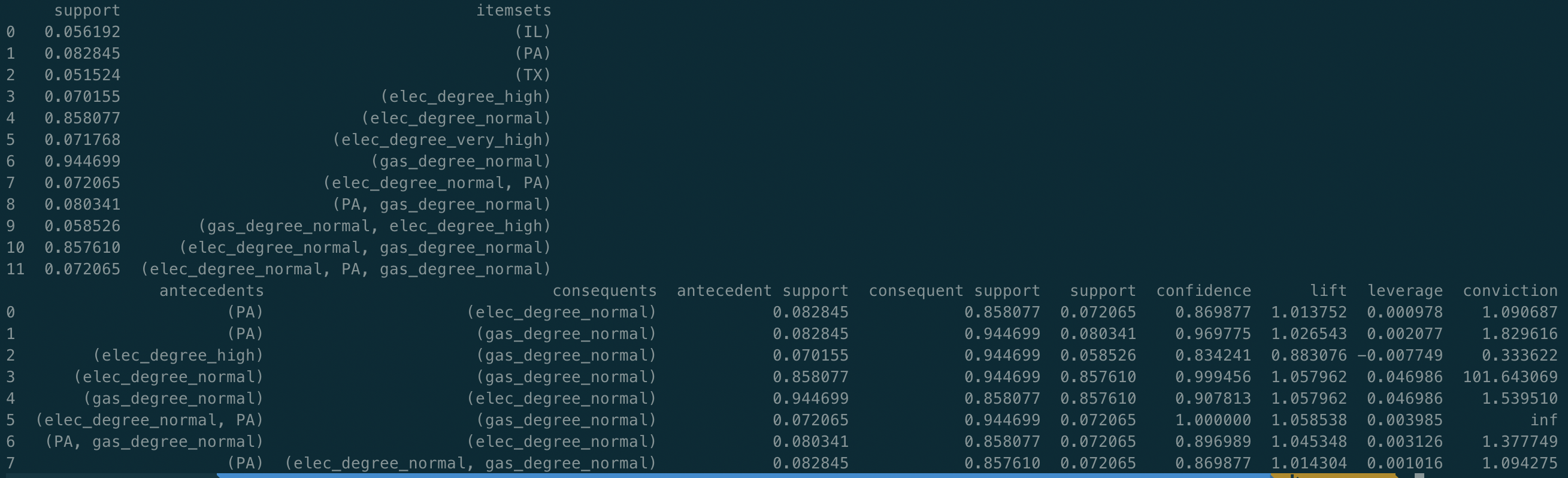


There are only three rules. We can conclude that the degrees of energy consumptions (both electricity and natural gas) of most cities are very low level. Besides, cities who consumed little electricity also consume little natural gas.

When support = 0.1 and confidence = 0.6, the output is same with above:



When support = 0.05 and confidence = 0.6, the output is:



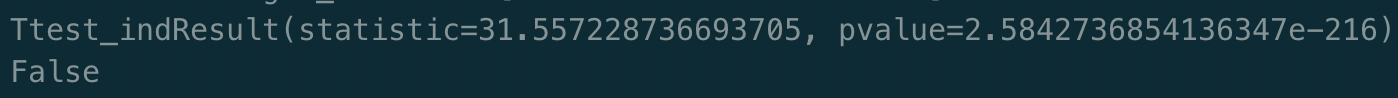
In addition to above three rules, there are eight more rules. We can conclude that most cities in Pennsylvania state consume normal-level energy resources. Many cities consume high-level and very high-level electricity. Many cities which consume normal-level natural gas might consume high-level electricity. And there are more cities in Pennsylvania, Illinois and Texas states.

The most frequent pattern is ‘gas\_degree\_very\_low’. This means that the natural gas consumptions of most cities are at normal level. It is reasonable.

* **Predictive Analysis**

**t-test**

Firstly, a hypothesis – the difference between the means of ‘elec\_degree’ and ‘gas\_degree’ attributes is not obvious. Here is the output of t-test:



The first line is the return value of t-test statistical containing t-statistics and p-value. The second line is the return value of a condition statement (pvalue > 0.05?). It is False. The original hypothesis fails. Therefore, the difference between degrees of electricity and natural gas is obvious. It represents the distribution of electricity degrees is different from the distribution of natural gas degrees.

**Linear Regression**

Another hypothesis – the ‘housing\_units’ and ‘total\_pop’ are positive related with ‘elec\_degree’. The fit coefficient is 0.7770601547487651. The hypothesis is true.

To predict the class of data, ‘state\_id’, ‘city’, ‘elec\_mwh’, ‘gas\_mcf’ attributes are all dropped and ‘elec\_degree’ and ‘gas\_degree’ are selected as the class columns. The model becomes to predict the energy consumption levels of a city given some attributes including ‘housing\_units’, ‘total\_pop’, ‘elec\_1kdollar’, ‘gas\_1kdollar’, ‘elec\_lb\_ghg’ and ‘gas\_lb\_ghg’. Here are the evaluation and prediction results with different method.

Evaluation of machine learning methods (electricity degree):

|  |  |  |
| --- | --- | --- |
|  | Average mean of accuracy | Average standard deviation of accuracy |
| Decision Tree | 0.982705 | 0.002644 |
| KNN | 0.962120 | 0.003793 |
| Naïve Bayes | 0.664545 | 0.011824 |
| SVM | 0.966470 | 0.004020 |
| Random Forest | 0.985145 | 0.003104 |

Prediction (electricity degree):

|  |  |
| --- | --- |
|  | Accuracy Score |
| Decision Tree | 0.9853596435391471 |
| KNN | 0.966687884574581 |
| Naïve Bayes | 0.6619987269255252 |
| SVM | 0.9652026310205813 |
| Random Forest | 0.987481434330575 |

Evaluation of machine learning methods (natural gas degree):

|  |  |  |
| --- | --- | --- |
|  | Average mean of accuracy | Average standard deviation of accuracy |
| Decision Tree | 0.999947 | 0.000159 |
| KNN | 0.998833 | 0.000849 |
| Naïve Bayes | 0.763489 | 0.012315 |
| SVM | 0.999151 | 0.000794 |
| Random Forest | 0.999947 | 0.000159 |

Prediction (natural gas degree):

|  |  |
| --- | --- |
|  | Accuracy Score |
| Decision Tree | 0.9997878209208572 |
| KNN | 0.9985147464460005 |
| Naïve Bayes | 0.7695735200509229 |
| SVM | 0.9987269255251432 |
| Random Forest | 0.9997878209208572 |

According to above accuracies, all methods except Naïve Bayes perform very well. And the prediction accuracy could be as high as 99.9%. This also shows that the chosen attributes are strong related with the class labels.