ANALYSIS OF POWER GENERATION IN US

A FINAL PROJECT REPORT WAS SUBMITTED

TO

KENT STATE UNIVERSITY

MASTER OF SCIENCE

IN

BUSINESS ANALYTICS

SUBMITTED BY

DODDA VENU (811224121)

VDODDA@KENT.EDU

INSTRUCTOR

Dr. MURALI SHANKER

PROFESSOR

DEPARTMENT OF MANAGEMENT INFORMATION SYSTEMS

KENT STATE UNIVERSITY

NOMENCLATURE

mmbtu	:	Metric Million British Thermal Unit
Btu	:	British Thermal Unit
Mcf	:	One thousand cubic feet
Ppm	:	Parts per million

INTRODUCTION

The analysis of the power generation in the United States was done by the k means clustering. The US government and other public entities provide electric utility reports with a vast amount of information, which is the source of the data utilized in the clustering. This contains annual, monthly, and even hourly information on fuel consumed, electricity produced, operating costs, usage trends at power plants, and pollutants. Unfortunately, a lot of this data is not made available in formats that are machine-readable, ready-to-use, and well-documented. The Public Utility Data Liberation (PUDL) takes that information and makes it publicly usable, by cleaning, standardizing, and cross-linking utility data from different sources in a single database. As a result, users can now spend more time on data analysis and less time on data preparation.

PROBLEM STATEMENT

After a brief data analysis, I address one question for the project that I would solve by using k means clustering.

The operating costs of power plants mainly depend on the amount of fuel used for generating power. The power plants used fuel, not only for power generation they also used for refining emissions of pollutants in the plants to reduce environmental hazards.

The project's main theme is how effectively we can comprehend and suggest solutions to generate power with lower operating costs in the US by performing some analysis and segmentation of the data.

DATA DESCRIPTION

The data was taken from the PUDL website which consists of 608,565 rows and 30+ variables.

Data cleaning:

For our project, some of the columns are not required so I removed some of the columns. Later, several variables have significant missing values so I just removed all those rows from the dataset.

Finally, I came up with the following variables.

1. plant_id_eia: It is the six-digit facility identification number given to each of the plants.
2. fuel_received_units: Quantity of fuel received in tons, barrel, or Mcf.
3. fuel_mmbtu_per_unit : Heat content of the fuel in millions of Btus per physical unit.
4. sulfur_content_pct : Sulfur content percentage by weight to the nearest 0.01 percent.
5. ash_content_pct : Ash content percentage by weight to the nearest 0.1 percent.

6. mercury_content_ppm : Mercury content in parts per million (ppm) to the nearest 0.001 ppm.
7. fuel_cost_per_mmbtu : Average fuel cost per mmBTU of heat content in nominal USD.
8. **chlorine_content_ppm:** Chlorine content in parts per million (ppm) to the nearest 0.001 ppm
9. **moisture_content_pct:** Moisture content percentage by weight to the nearest 0.1 percent.

Data preparation:

For standardizing the data, I just normalize the data by using StandardScaler. I have taken 2% of the data used from the dataset as a sample. Later I used K means cluster to examine the data.

ANALYSIS & DISCUSSION

K means Clustering:

K-means separates the collection of data items into distinct subgroups (clusters), where each data item refers to a single subset.

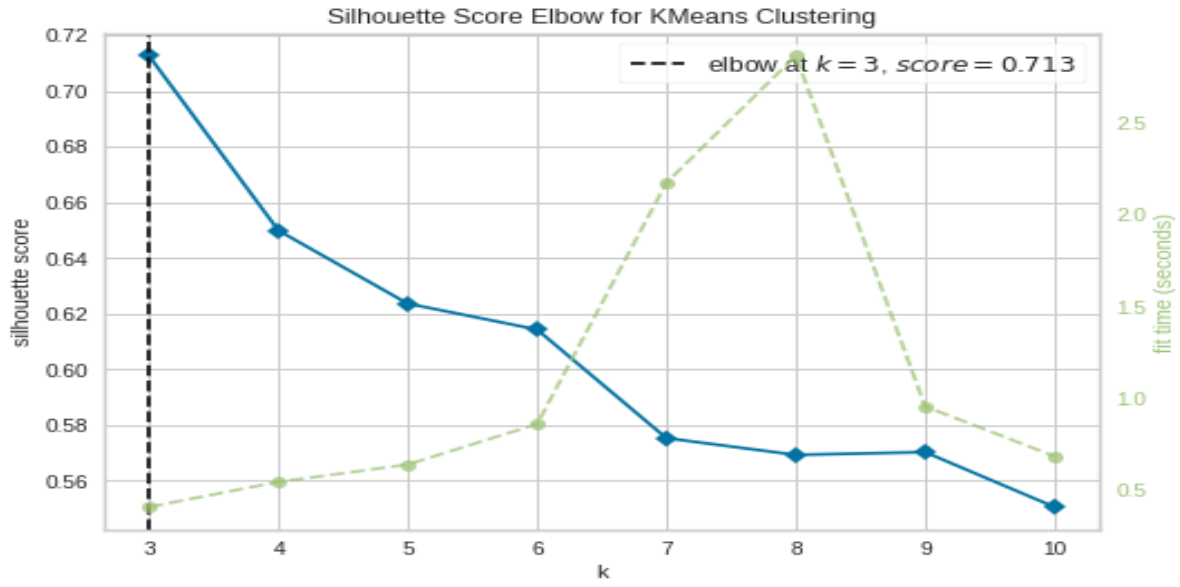


Fig: Silhouette Score Elbow graph

For our analysis, I choose k means clustering, In the way of finding optimum clusters to perform our task I used silhouette score elbow methods. I got k=3 from the graph shown above. So that I will come up with 3 clusters for my project you can see in the below graph.

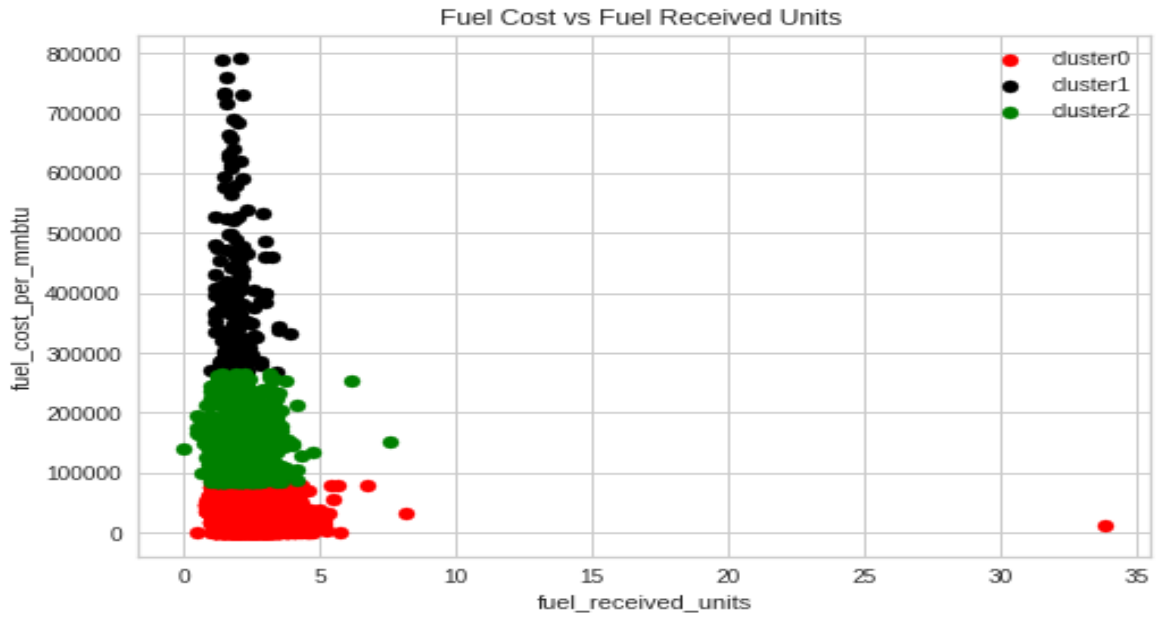


Fig: Clusters Formation

The clusters divide the data uniformly and each cluster contains the power plants which use all fuel types such as coal, Petroleum, natural gas, Petroleum Coke, etc.

The main agenda is to find which cluster consists of low fuel costs and Low emission of pollutants. So that we can predict which power plants generate power with lower operating costs in the US.

Cluster 0:

```
[ ] columns[cluster_ids==0].describe()
```

	mine_id_pudl	mine_id_pudl_label	fuel_received_units	fuel_mmbtu_per_unit	sulfur_content_pct	ash_content_pct	mercury_content_ppm	fuel_cost_per_mmbtu	moisture_content_pct	chlorine_content_ppm
count	7174.000000	7174.000000	7174.000000	7174.000000	7174.000000	7174.000000	7174.000000	7174.000000	7174.000000	7174.000000
mean	1419.423195	1419.423195	26923.319626	21.140057	1.303814	8.024749	0.014444	2.308040	15.754456	49.165877
std	1630.205120	1630.205120	19786.575436	3.407660	1.208145	3.605003	0.034610	0.813037	10.340817	246.032489
min	11.000000	11.000000	10.000000	10.300000	0.120000	0.000000	0.000000	0.603000	0.000000	0.000000
25%	21.000000	21.000000	12321.000000	17.703250	0.270000	5.100000	0.000000	1.828000	6.790000	0.000000
50%	221.000000	221.000000	21401.000000	22.420500	0.790000	7.900000	0.000000	2.193000	12.190000	0.000000
75%	2797.000000	2797.000000	39624.750000	24.061500	2.570000	9.680000	0.000000	2.678000	26.710000	0.000000
max	4557.000000	4557.000000	77916.000000	28.000000	6.070000	56.000000	0.400000	38.889000	42.400000	3043.000000

Table: Cluster 0 Data Prediction

Cluster 1:

```
[ ] columns[cluster_ids==1].describe()
```

	mine_id_pudl	mine_id_pudl_label	fuel_received_units	fuel_mmbtu_per_unit	sulfur_content_pct	ash_content_pct	mercury_content_ppm	fuel_cost_per_mmbtu	moisture_content_pct	chlorine_content_ppm
count	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000	301.000000
mean	733.202658	733.202658	355441.083056	17.572362	0.704784	8.668605	0.014475	2.038302	23.055748	9.926910
std	1268.477419	1268.477419	120783.077418	3.130710	0.827801	5.568907	0.033127	0.636997	10.794917	100.257675
min	19.000000	19.000000	242936.000000	10.100000	0.180000	4.100000	0.000000	0.983000	0.000000	0.000000
25%	20.000000	20.000000	269677.000000	16.710000	0.250000	4.700000	0.000000	1.666000	13.800000	0.000000
50%	127.000000	127.000000	318447.000000	17.590000	0.360000	5.700000	0.000000	1.940000	27.160000	0.000000
75%	633.000000	633.000000	409882.000000	18.363000	0.800000	10.700000	0.000000	2.262000	29.700000	0.000000
max	4399.000000	4399.000000	935695.000000	26.450000	4.160000	30.700000	0.118000	6.366000	37.620000	1600.000000

Table: Cluster 1 Data Prediction

Cluster 2:

```
[ ] columns[cluster_ids==2].describe()
```

	mine_id_pudl	mine_id_pudl_label	fuel_received_units	fuel_mmbtu_per_unit	sulfur_content_pct	ash_content_pct	mercury_content_ppm	fuel_cost_per_mmbtu	moisture_content_pct	chlorine_content_ppm
count	1634.000000	1634.000000	1634.000000	1634.000000	1634.000000	1634.000000	1634.000000	1634.000000	1634.000000	1634.000000
mean	806.654223	806.654223	128945.067319	19.268250	1.022564	7.208556	0.013974	2.070557	20.248213	26.212362
std	1329.786664	1329.786664	41386.693318	3.091573	1.176196	4.038887	0.033901	0.642604	10.005091	189.670865
min	18.000000	18.000000	77990.000000	9.848000	0.170000	3.300000	0.000000	0.539000	0.000000	0.000000
25%	20.000000	20.000000	95097.250000	17.230750	0.240000	4.800000	0.000000	1.671000	11.500000	0.000000
50%	42.000000	42.000000	118149.000000	17.860500	0.340000	5.600000	0.000000	2.034500	26.095000	0.000000
75%	703.500000	703.500000	156630.000000	22.243500	1.640000	8.600000	0.000000	2.362000	27.985000	0.000000
max	4522.000000	4522.000000	242126.000000	28.280000	4.510000	48.500000	0.230000	8.519000	40.000000	3121.000000

Table: Cluster 2 Data Prediction

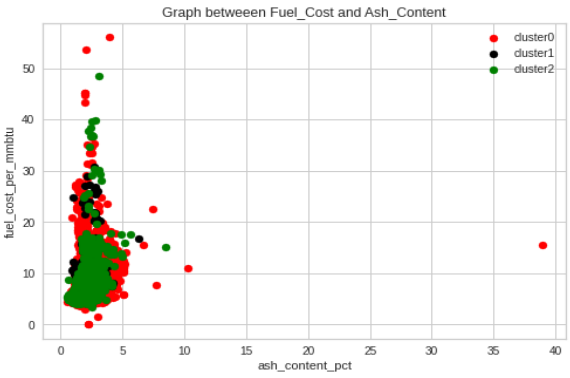
From the above three tables, we can say that

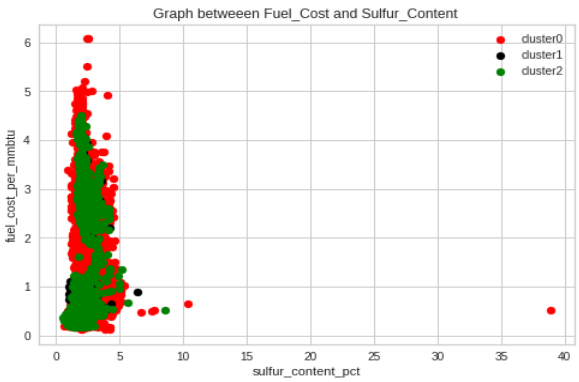
Cluster 2 powerplants generate power with lower emission of pollutants but they are spending more money on fuel costs than other powerplants in clusters 0 and 1.

Cluster 0 power plants generate power with moderate emission of pollutants but they are spending money on fuel costs moderately.

Cluster 1 power plants generate power with High emissions of pollutants but they are spending less money on fuel costs than cluster 0 and cluster 2.

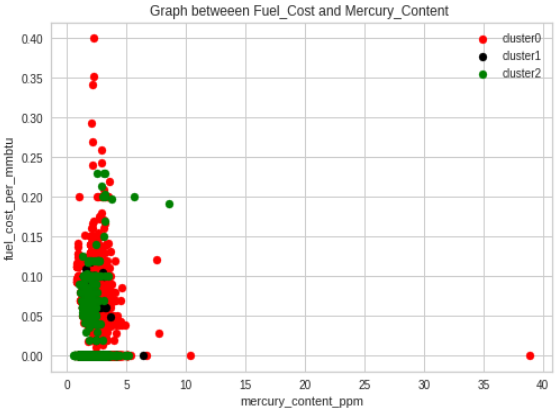
Plotting the Graphs based on the cluster values:

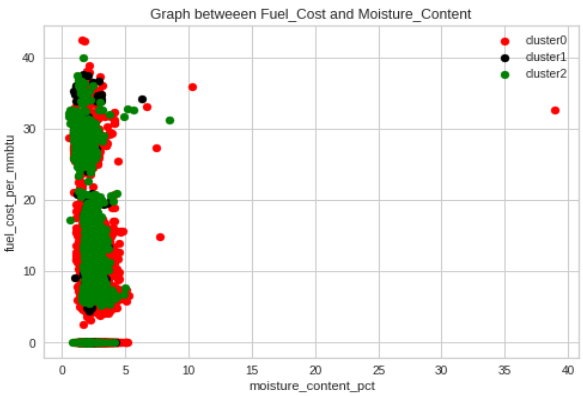




From Fuel_Cost and Ash_Content graph cluster 2 emits a lower amount of ash content than cluster 0 and cluster 1.

From Fuel_Cost and Sulfur_Content graph cluster 1 emits a lower amount of sulfur content than cluster 0 and cluster 2

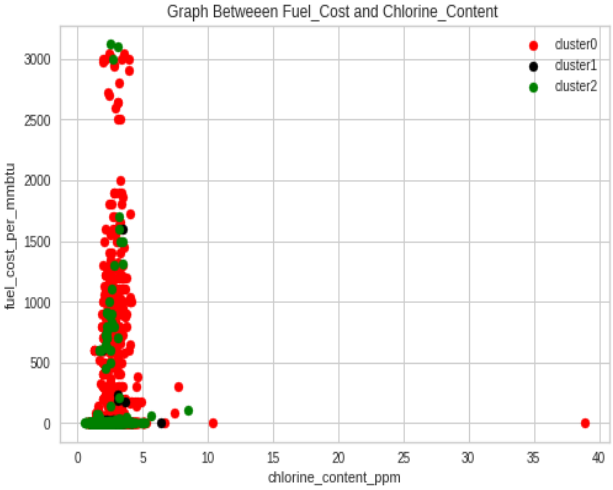




From Fuel Cost and Mercury_Content graph

cluster 2 emits lower amount of mercury content than cluster 0 and cluster 1.

From fuel Cost and Moisture Content graph
cluster 0 emits a lower amount of moisture content than cluster 1 and cluster 2.



From Fuel_Cost and Chlorine_Content graph cluster 1 emmits lower amount of chlorine content than the cluster 0 and cluster 2.

CONCLUSION

The power plants in Cluster 2 produce power with reduced pollution emissions, but they also incur higher fuel costs than the power plants in Clusters 0 and 1. Cluster 0 power plants produce power with moderate pollution emissions, but they incur moderate fuel costs. Compared to clusters 0 and 2, cluster 1 power plants produce power with higher pollution outputs but lower fuel prices.

I concluded that power plants use extra fuel to reduce the amount of pollution they emit into the atmosphere. By using additional fuel to purify the pollutants.

Among all Cluster 2 gives the optimum results because power plants in that cluster produce power while emitting fewer pollutants into the atmosphere.

Consequently, an increase in fuel cost will result in an increase in the Operating costs of the power plants.

EXECUTIVE SUMMARY

In this project, I could observe the below results

The variation in operating costs of power plants mainly depends on fuel costs over the years.

To address this issue, I have used K means clustering to draw the observation about fuel costs over the pollutants. From these observations, I can say that power plants use additional fuel to refine the pollutants while processing the power generation. From the K means clustering, Cluster 2 produces better outcomes than the others since the power plants in Cluster 2 emit less pollution, but they also have higher fuel costs than the power plants in Clusters 0 and 1. Cluster 0 power plants emit low levels of pollution while incurring modest fuel expenditures. In comparison to clusters 0 and 2, cluster 1 power plants produce more pollution but at cheaper

fuel prices. I came to the conclusion that power plants use more fuel in order to limit the quantity of pollution they emit into the environment. By utilizing more fuel to clean up the pollutants.

I believe that power plants will implement new strategies to limit pollutant emissions. And consistently recommends using less polluting fuel to generate power.

References:

1. https://catalyst.coop/pudl/
2. https://www.researchgate.net/project/EIA-Electricity-Demand-Data
3. https://nccleantech.ncsu.edu/2022/07/26/the-public-utility-data-liberation-pudl-project-adds-clean-energy-standards-with-dsire/