**IDEAL POWER PLANT RECOMMENDATION IN USA**

**Fundamentals of Machine Learning**

**MIS 64060-001**

**Professor Student**

**Murali Shanker Pavan Chaitanya B**

**INDEX:**

* + **PROBLEM**
    - **PROBLEM STATEMENT.**
  + **INTRODUCTION.**
    - **DESCRIPTION.**
    - **DATA DESCRIPTION AND DATA CLEANING.**
  + **TECHNIQUES.**
    - **FINDINGS**
  + **CONCLUSION**
  + **EXECUTIVE SUMMARY**
  + **APPENDIX**
    - **REFENCES**.

**Problem:**

An open-source data processing pipeline called Public Utility Data Liberation (PUDL) makes US energy data more accessible. The US government and other public entities provide them with an availability of material regarding the energy production. **The main objective of the problem is to analyze data (EIA-923) from the Public Utility Data Liberation (PUDL) project, which contains all the information about the types of materials (coal, natural gas, and petroleum) used to produce electricity in relation to the quantity of heat that results in environmental pollution, and to provide recommendations based on the findings regarding power generation in the USA.**

**Introduction**:

We all understand that there are four essential needs in life, but we are unaware of the fact that power has also grown to be significant in addition to these four needs. The USA is a developed nation, and in this research, we will determine the optimal power generation method using data from the Public Utility Data Liberation (PUDL) Project (EIA-923).

**Data Description and Data Cleaning**:

I have taken set.seed(2875)

For the provided data, I started by taking 2% of the total data and then removing the columns that weren't necessary (report data, data maturity), computed the missing values by first converting them into factors and then using the missforest(), then portioned the data into 75%, which produced some outliers, and computed the outliers before scaling the necessary variables that we need for our project.

The variables I've selected are

fuel\_recieved\_units : tells about how much fuel is received .

fuel\_mmbtu\_per\_units : tells about how much heat is generated.

fuel\_cost\_per\_mmbtu : tells about how much is the cost for the fuel heat produced(Main Variable that I am focusing on).

ash\_content\_pct : tells about the ash content

mercury\_content\_pct : tells about the mercury content

sulphur\_content\_pct. : tells about the sulphur content

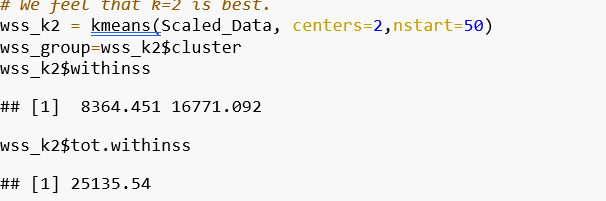
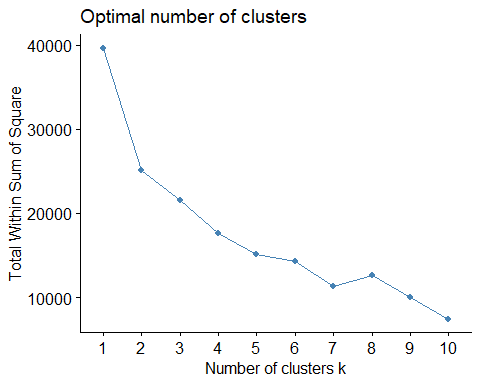
**Techniques:**

**Kmeans**

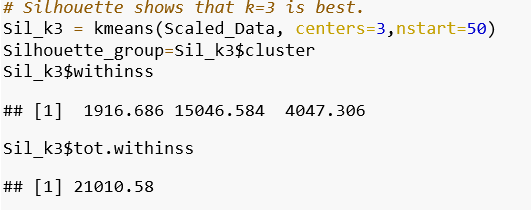
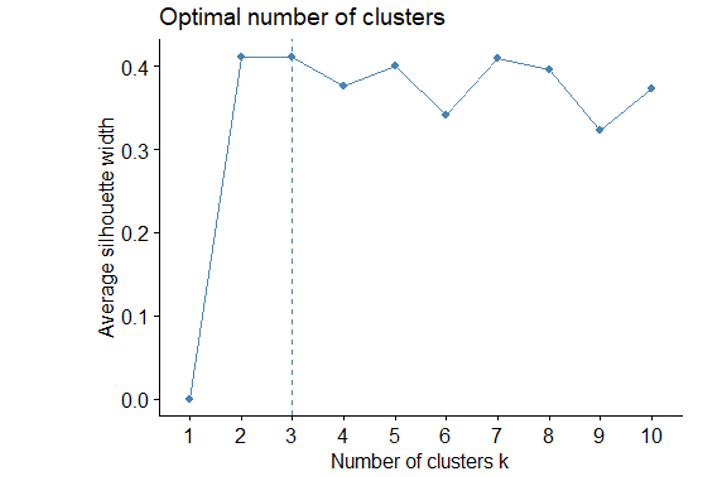
To achieve the goal of making it easier to identify the clusters and recommend to the usa government, our approach uses both K-means and Hierarchical Clustering. This approach provides both a cluster plot (from K-means) and cluster dendrogram (from Hierarchical clustering).

Using this Kmeans, we can determine the ideal "K”. We can further our analysis of the clusters by using this K as a reference. In order to determine which K to employ for the project data, we are evaluating the data using two methods that are contained in Kmeans: the WSS(ELBOW) method and the SILHOUETTE method.

**WSS (Elbow Method): K=2 Calculations**



**Silhouette Method: K=3 Calculations**



Our main purpose of finding K is to find the Total Within Sum of Squares with

Elbow method we can see total within sum of squares is 25135.54 and

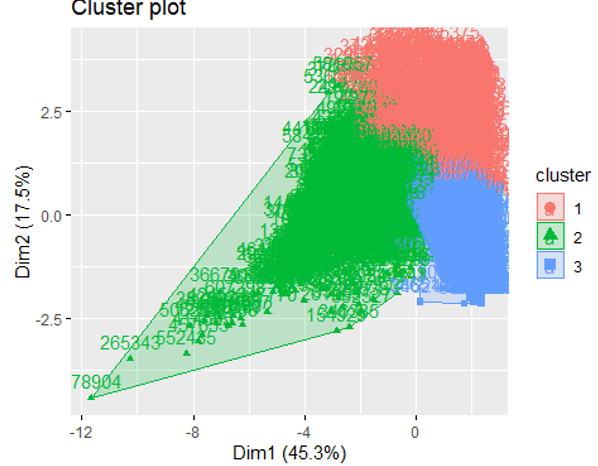
Silhouette method we can see total within sum of squares is 21010.58.

So, we are considering k=3 as our final ‘K’. Consequently, it is the number of clusters that is most ideal.

**Plotting the Clusters:**

**Cluster Plot:**

Now that we have determined the most optimal number of clusters, we can create a cluster plot using K-means to better illustrate the clustering. The graph below shows our 3 clusters obtained using K-means clustering.

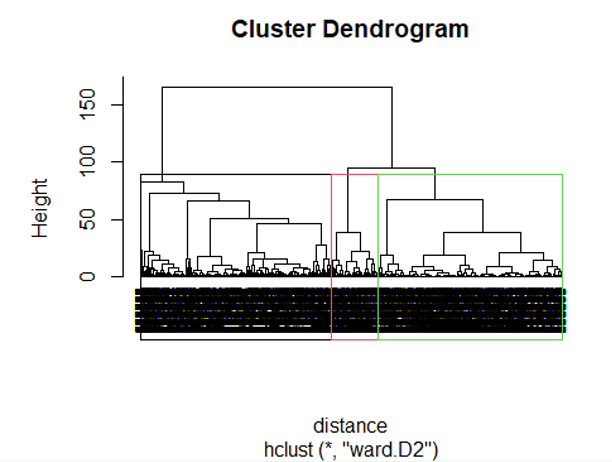


The three clusters above are generated from our sampled Data. These clusters were created using all numerical data from the dataset.

**Hierarchical Clustering**

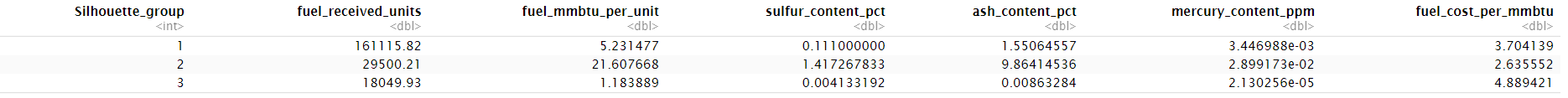
Because we are using a different method of clustering to show characteristics, we want to revalidate our optimal number of clusters for the new clustering algorithm. This was achieved by creating an Cluster dendrogram that illustrates our clusters in a tree format using the Ward’s Metric.

**Cluster Dendrogram:**



Now that we have confirmed that three clusters will be optimal for Hierarchical clustering, we can make our interpretations of the clusters formed.

**Conclusion:**

****

Since the silhouette method produced 3 clusters, it is now time to examine the clusters:

Because the amounts of sulfur, mercury, and ash in the given data are less than 0.01, we will ignore them when interpreting the results.

**Cluster 1:**

The fuel received by the power plants in this cluster, which is higher than that of the other two clusters at 161115.82, has a heat content of 5.231477, which is excellent. With respect to fuel received and heat content, the fuel cost per mmbtu is likewise quite good (3.704139). Given all the relevant considerations, including those listed below, this Cluster is the one that the US Government should be advised to pursue (fuel recieved,heat content,fuel cost per mmbtu).

**Cluster 2:**

The fuel received by the power plants in this cluster, which is somewhat more than Cluster 3 but not Cluster 1, is 29500.21. Compared to the other 3 clusters, their fuel has a heat content that is 21.607668 times higher. The fuel costs less per mmbtu (2.635552) than all three of the formed clusters.

Due to the fuel's mmbtu per unit, this cluster is also not one that is favored for recommendation to the US Government.

**Cluster 3:**

The fuel received by the power plants in this cluster, which is low in comparison to other units, is 18049.93. Due to the poor fuel supply, their fuel's heat content (fuel mmbtu) is also low (1.183889). The fuel costs more per mmbtu (4.889421) than any of the three generated clusters.

Due to the high cost of fuel per mmbtu, this Cluster is not a favoured one to suggest to the US government.

**We may infer that Cluster 1 is the best cluster from the mentioned three Clusters.**

**Executive Summary:**

After analysing the data in the EIA-923 dataset, we discovered that there are unnecessary and missing data present. After computing, we had an idea of what is the best 'K' (K=3) by seeing both the Wss method and silhouette method and by computing the total within sum of squares. We then plotted both the cluster plot and cluster dendrogram for visually representing the clusters. We then individually analysed each cluster and concluded that the power plants present in cluster 1 are the best.

**Appendix:**

<https://www.statology.org/remove-outliers-r/>

<http://tiantiy.people.clemson.edu/blog/2019/MissingData/random_forest.html>

<https://stackoverflow.com/questions/54332680/multiple-imputation-in-r-using-missforest-on-categorical-variables>

<https://www.eia.gov/energyexplained/us-energy-facts/>

<https://catalyst.coop/pudl/>

<https://stats.stackexchange.com/questions/483222/how-to-use-missforest-in-r-for-test-data-imputation>