Final Project Report

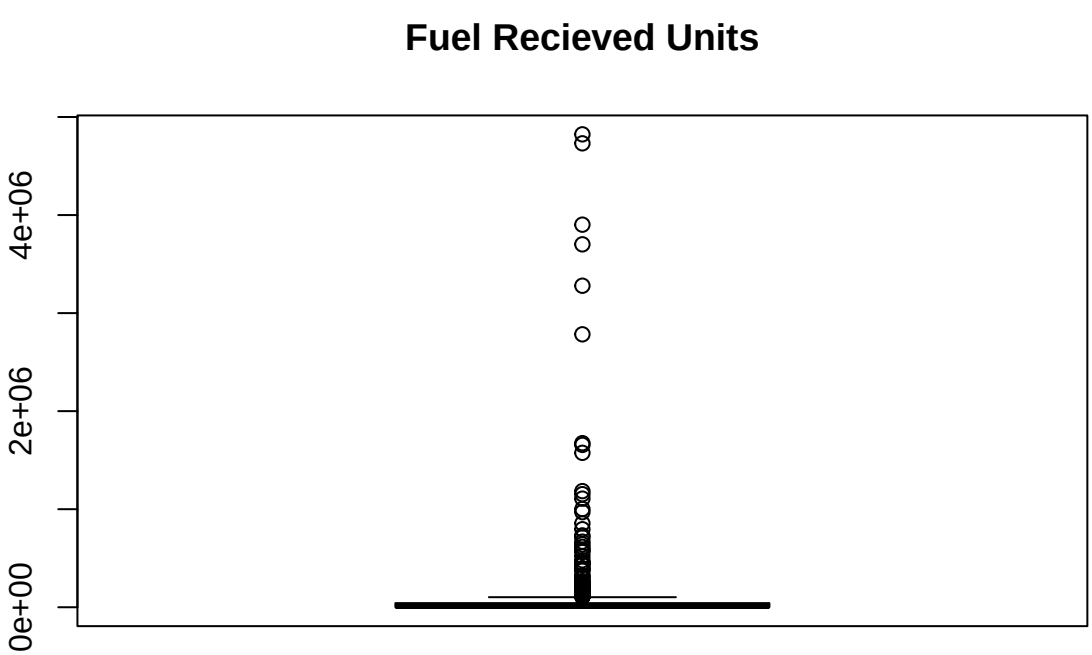
Jacob Joy

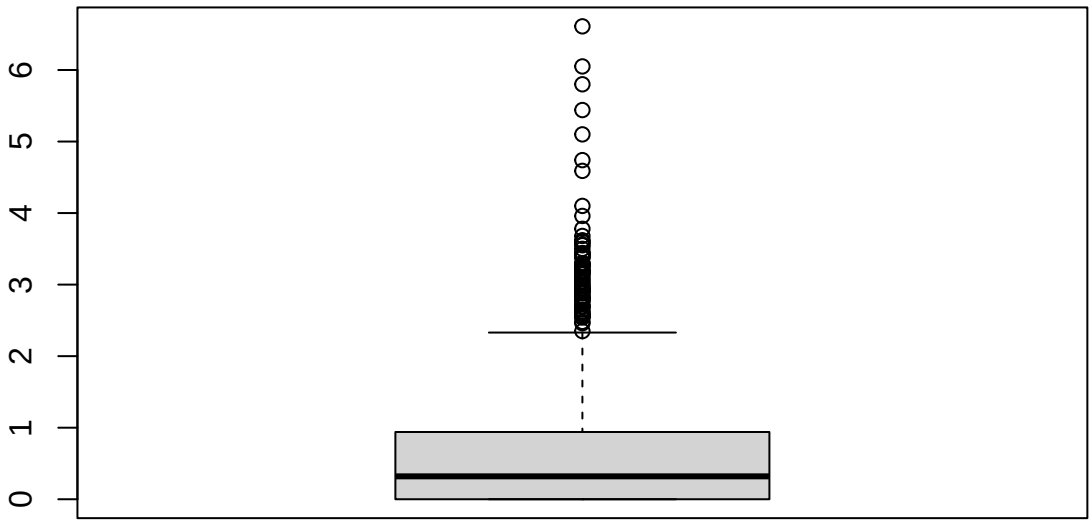
Executive Summary

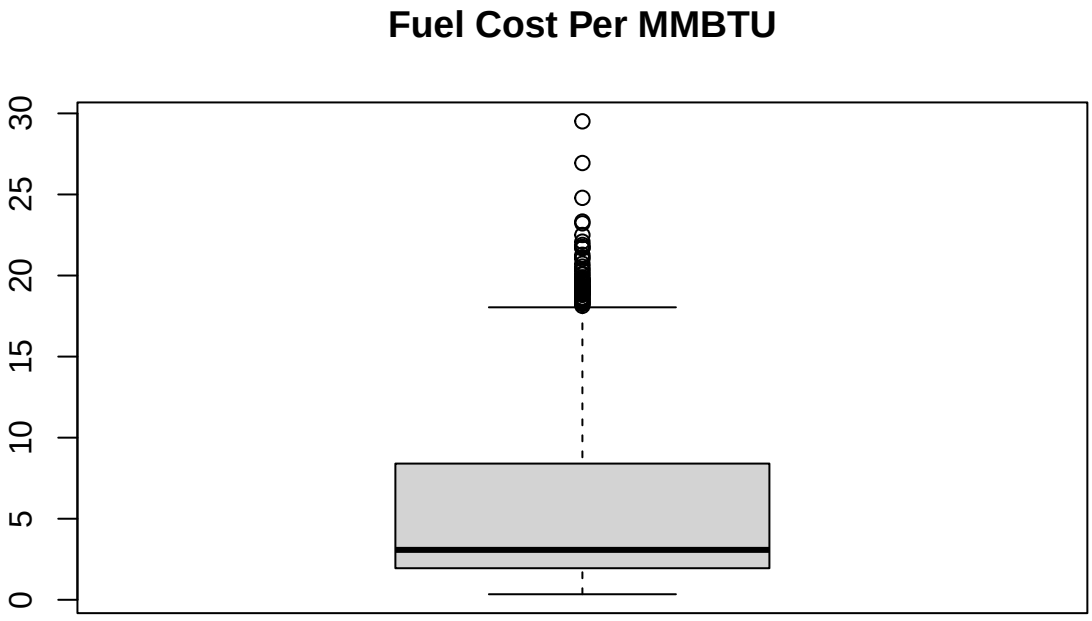
This report suggests that Oil may be more efficient then Gas with similar amounts of pollutants but the cost makes it prohibitive to use. Gas is not used as much due to the higher cost and lower output. Coal is the cheapest option with the highest efficiency but the highest in pollutants. These three fuel sources have distinct characteristics which lead to highly accurate clustering and classification.
The U.S. relies heavily on coal for its energy needs and in this data set did not mention any other types of energy other then fossil fuels. In order to promote the usage of additional energy the costs for oil and gas would need to be a considerate amount less then coal. As energy companies are more concerned with the highest output at the lowest cost instead of what is best for the environment.

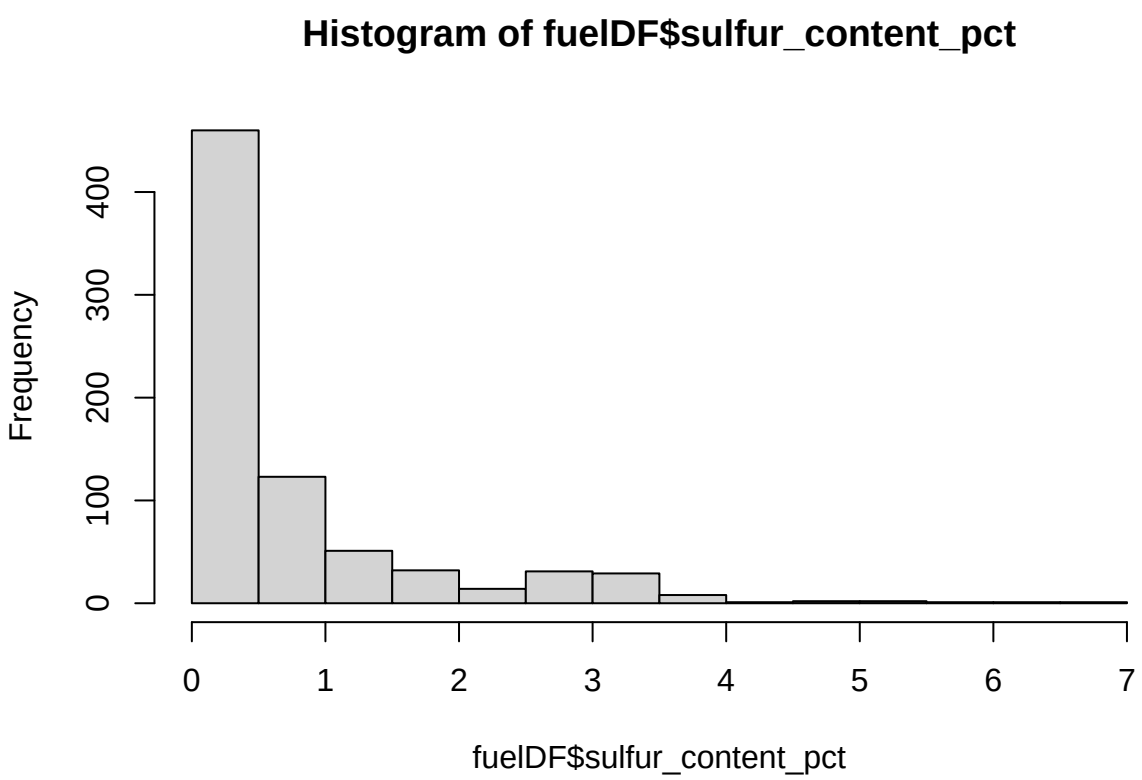
Introduction

After importing the provided data set, the seed value was set to the last 4 digits of my student number. Next, the data structure was looked at to determine the types of variables. There are two variables of type char that will need to be changed to factors if they are to be used. Next, a summary of statistics was performed and showed the potential of outliers as well as how different the scales are between the variables. Next, the data was checked for missing values. No missing values were identified. Dummy Variables were created for the fuel_type_code_pudl as this was a requested value to be used in the clustering analysis. Next, boxplots were created to identity outliers. No outliers were identified in fuel_mbtu_per_unit or ash_content_pct.

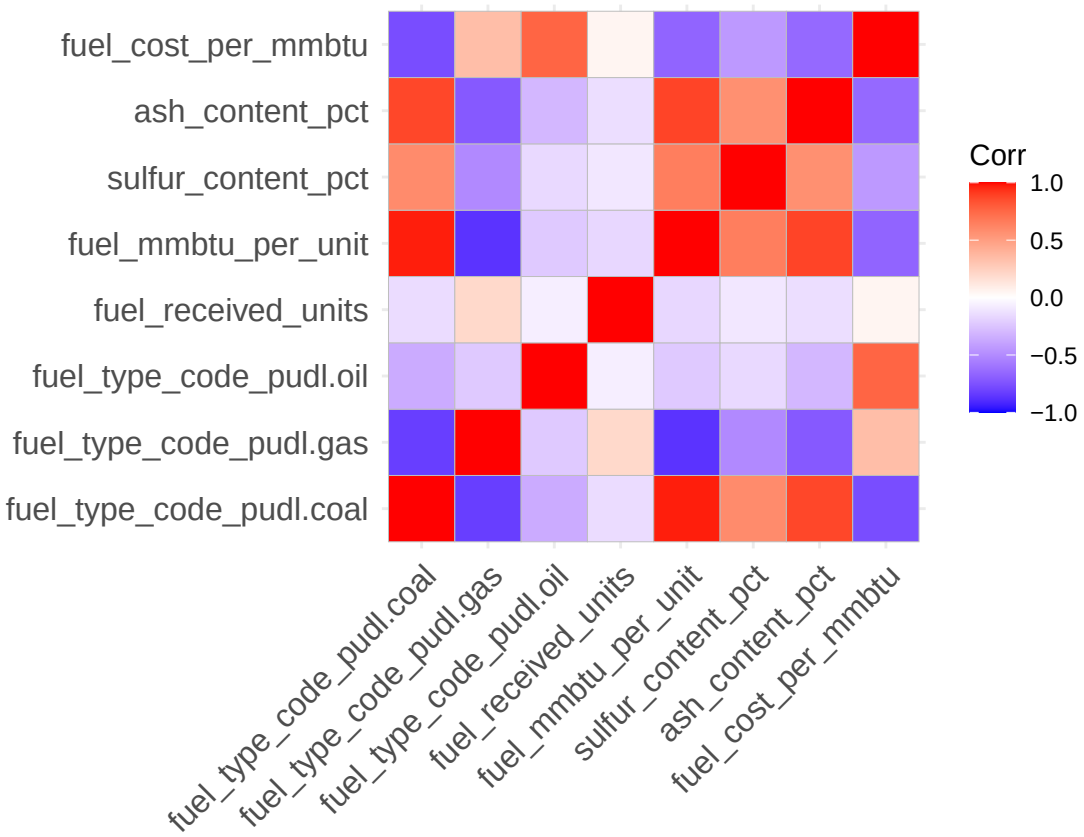

Fuel Recieved Units
0e+00
2e+06
4e+06


Sulfur Content Pct
0
1
2
3
4
5
6


Fuel Cost Per MMBTU
0
5
10
15
20
25
30



There are some outliers in fuel_received_units, sulfur_content_pct, and fuel_cost_per_mmbtu. Upon investigating each variable sulfur_content_pct was kept because it is a percent and makes sense that some types of fuels would have higher sulfur content, this variable is to important to leave out. fuel_cost_per_mmbtu was also left alone as it is a average price and it is not uncommon for outliers to appear in pricing data. Further investigation was done on fuel_received_units, this variable had wide ranges of values and is described in the data as a value associated with either tons, barrels, or Mcf. Having mixed units inside the same variable can lead to issues and lease accuracy. Assuming that each of these units align with a individual energy source summary data was looked at and there still appeared to be wide ranges throughout. At this point the value was kept intake and was discovered in a correlation matrix to have very small correlation that this variable was removed.



The final preprocessing step taken was to remove the char fuel_type_code as this was now captured in 3 dummy variables. Once this value was removed the data was normalized using z-score.

What could this data be used for?

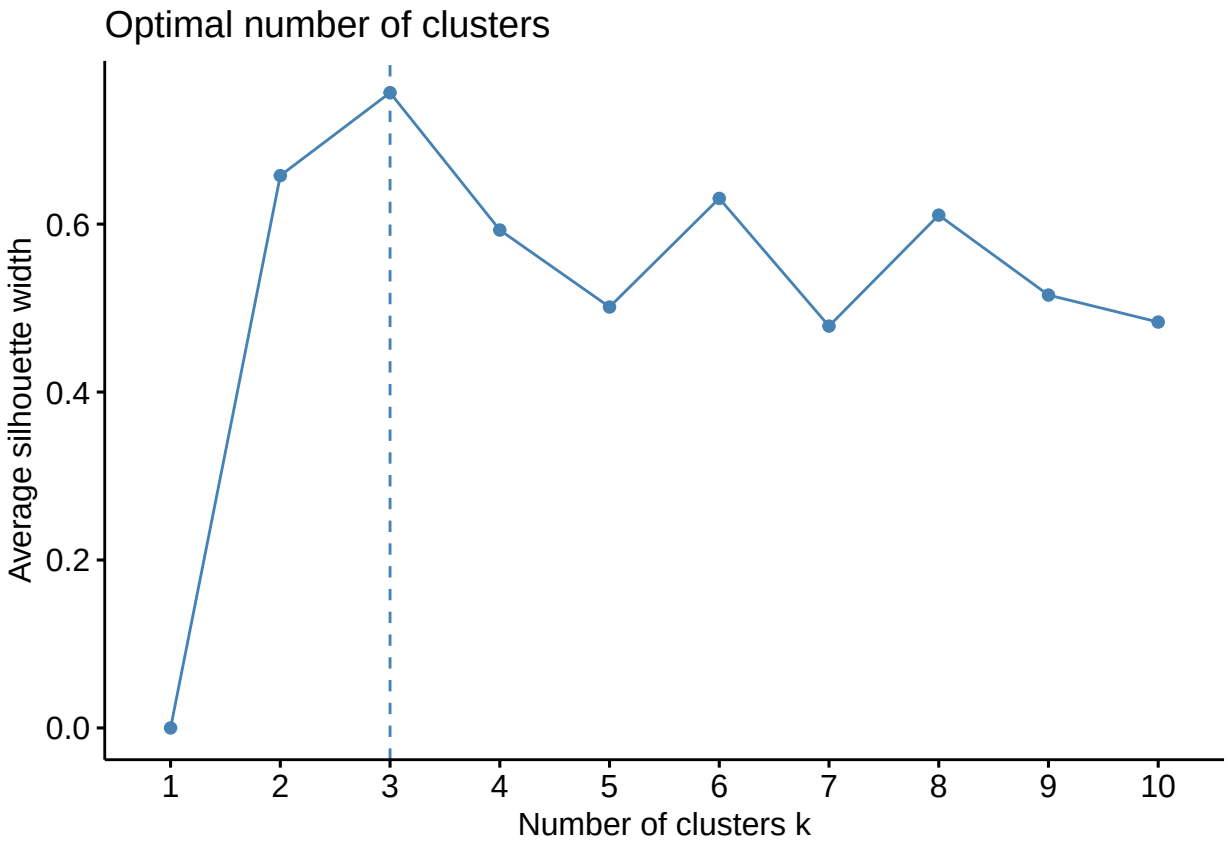
This data seems best used for identifying price vs output vs type vs pollution. If the energy plants were left in the data we could use this data to identify biggest polluters and who produces the most energy vs cost/pollution.

Clustering Analysis and Discussion

Three types of clustering were used to conduct analysis on the data set.

K-means Clustering

First, K-Means Clustering was used. Based on the instruction to include the fuel type, I felt that a good value for k would be 3. This also confirmed that the fuel_received_units should be removed from our clustering data as this would be useful if we were trying to answer questions on the energy plants. Why my assumption was to use 3 for k WSS and Silhouette tests were performed with silhouette indicating 3 for the value of k and WSS showing a value of 2 at the elbow. 2 was not select and 3 was used as the graph indicated there may be multiple values for k and a value of 2 could potentially lead to simplistic results that do not provide much information.



What benefit is there to including a categorical value in the cluster analysis?

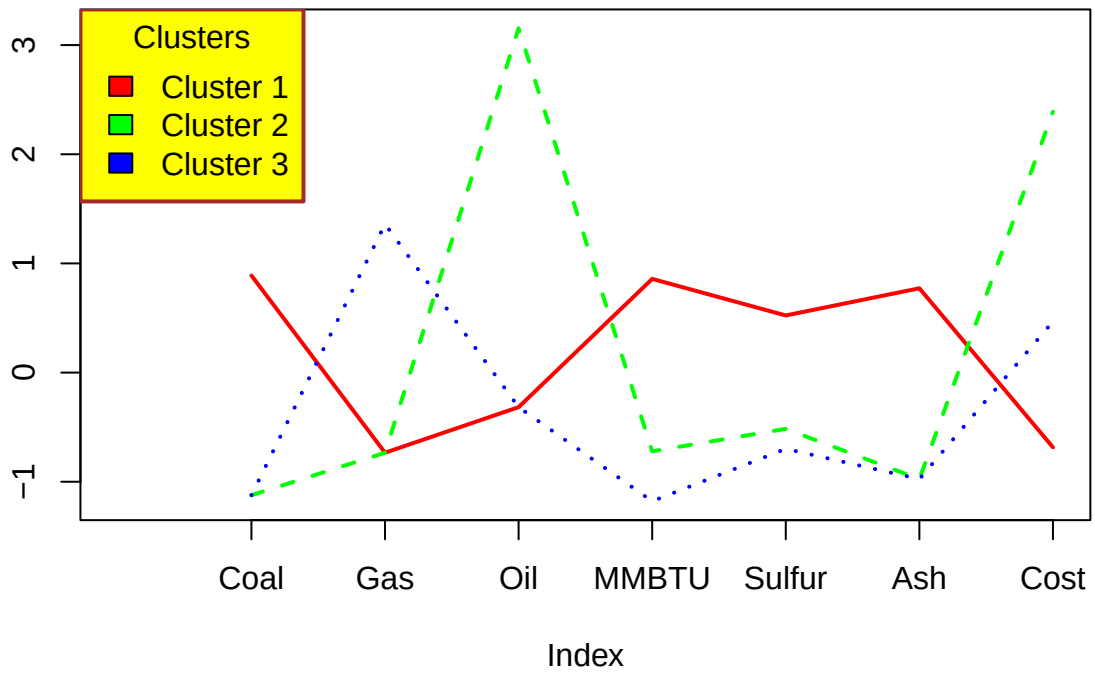
While the categorical value will help create distinct clusters we may be able to find trends within the clusters themeselves that can prove useful. In order to do this the value of K would need to be increased. However, for K-means 3 clusters was choosen. Additional details on amount of clusters and results are discussed bellow.

What distance measure should be used?

The decision was made to use euclidean for distance, Gowers and Jacquard were also considered and investigated. However, the results using these distances/similarities did not appear to provide results that could be interpreted meaningfully. Looking at the clusters generated by K-means we can see 3 distinct clusters and each of them is equal to oil, gas, or coal. By counting the number of each in our data set and comparing this to the size of each cluster there is an exact match. This suggests that K-means may not be as useful for discovering unknown patterns. This also shows that including the categorical variable is dominating the other variables.

What do each cluster represent?

Looking at the graph of centroids it is determined that Cluster 1 is coal, cluster 3 is gas, and cluster 2 is oil. We can also determine that coal costs the least compared to the amount of energy released but is highest in pollutants of both ash and sulfur. Oil produced the next highest amount of mmbtu but costs the most. Lastly, cluster 3 which is gas has the lowest mmbtu and a cost in the middle. Both cluster 2 and 3 have similar sulfur and ash content.



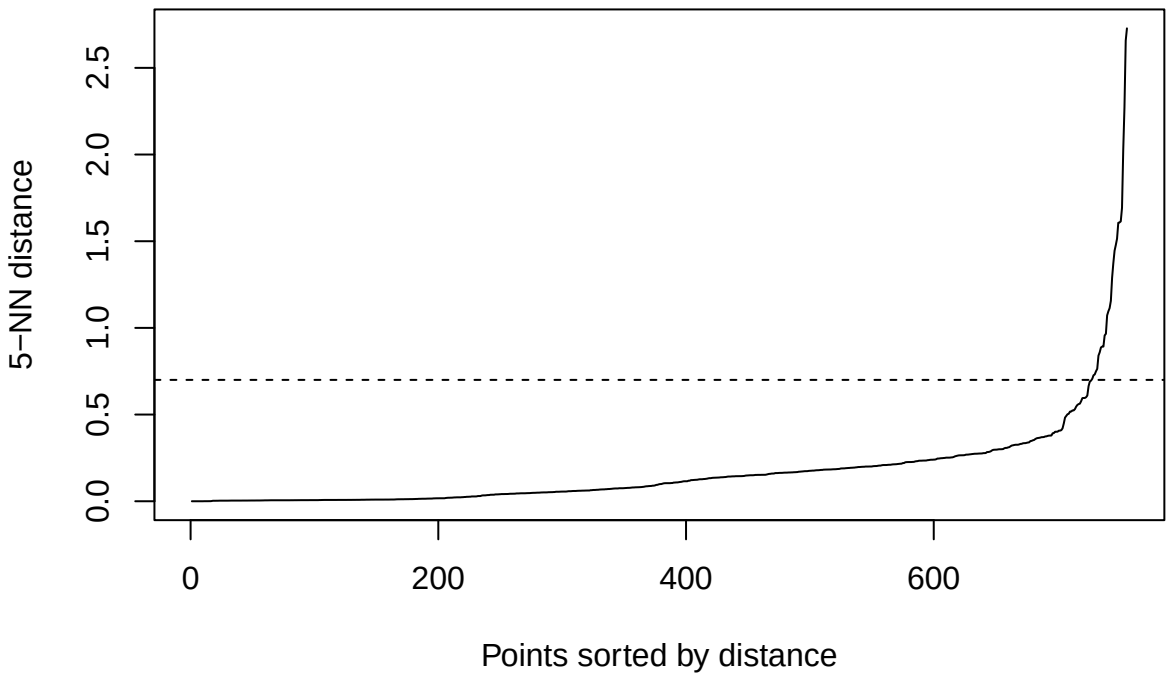
What fuel will provide the least pollution at the lowest cost?

This can be answered by looking at cluster 2 and 3 which are similar outside of price. This is seen by both clusters being on the left side of the plot. Looking back at the centroid graph it the cheapest option with the least pollution would be Gas(cluster 3), but this also puts out the least amount of energy. However, this would be a good starting point to determine actual pricing and look into additional variables that are not part of the data set. Such as location, time of year, political climate, etc.

DBSCAN

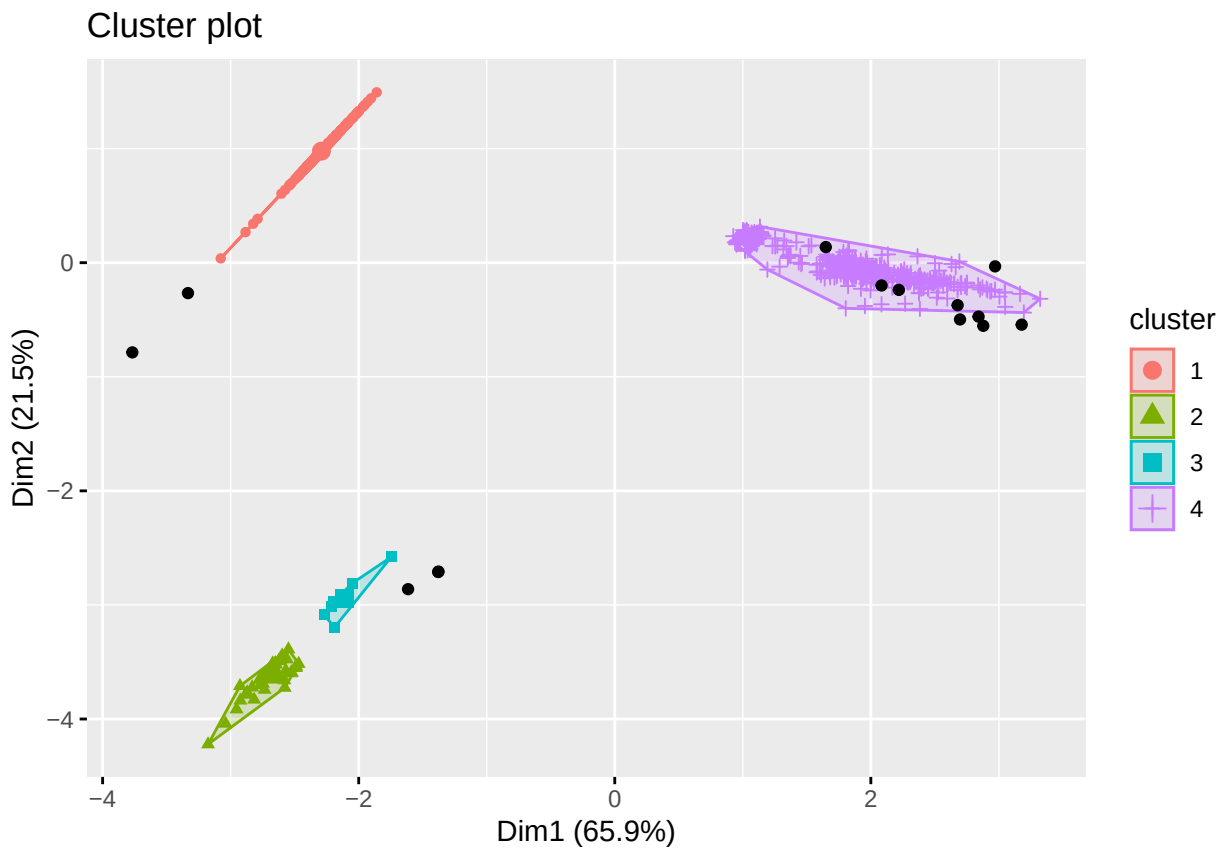
How will the value for minpts and epsilon be selected?

Using the same data as K-Means DSCAN was conducted. First, a value for epsilon was determined based of KNN and a value of 5 for k. 5 was chosen because of a heuristic rule for selecting min points of (features + 1). I considered there being only 4 features even though we have 7 because the dummyVars all represent the same category. In doing this I selected a value of 0.7 for epsilon as this appears to be the point on the plotted KNN distances to be where distance starts to increase the most. Minpoints is 5 as it should be the same value as what was selected in performing the KNN test.



Why did DBSCAN create 4 clusters instead of 3?

The results of the DBSCAN created 4 clusters and some noise points. These 4 clusters appear similar to the results in K-Means. However, the bottom left cluster that we had in K-Means is split into 2 in DBSCAN. This split of the oil cluster suggests that there is enough difference consider two types of oil. While we did not include the plants that produced the energy the data points are from multiple locations this could mean that there are at least 6 observations that are based on more then just them being oil. An additional note which could be slightly seen in k-means is that Cluster 1 in DBSCAN which is assumed to be gas is linear this seems accurate as from the energy industry on downside to using gas all the time is the variability in its costs. By changing the amount of minpoints the amount of clusters can change. If it was important that only 3 clusters where to be identified this could be achieved by increasing the minpoints to 6 but would also identify more noise.



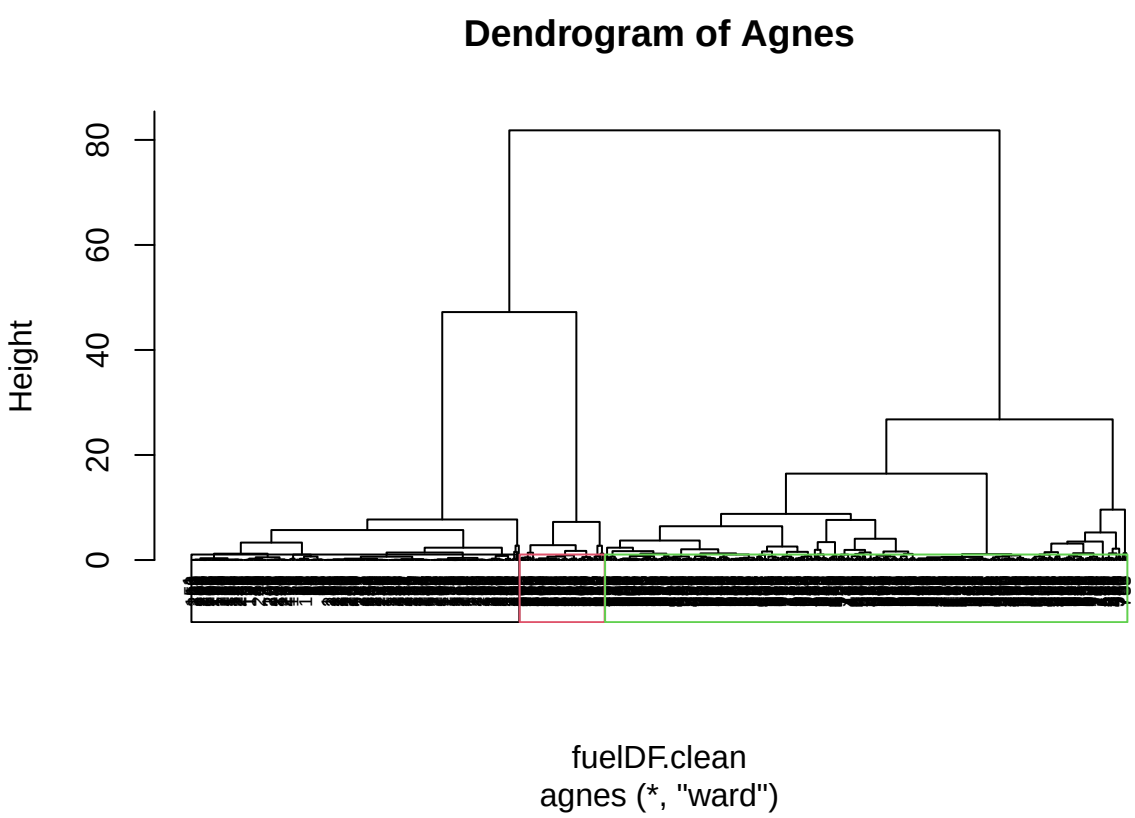
Hierarchical Clustering

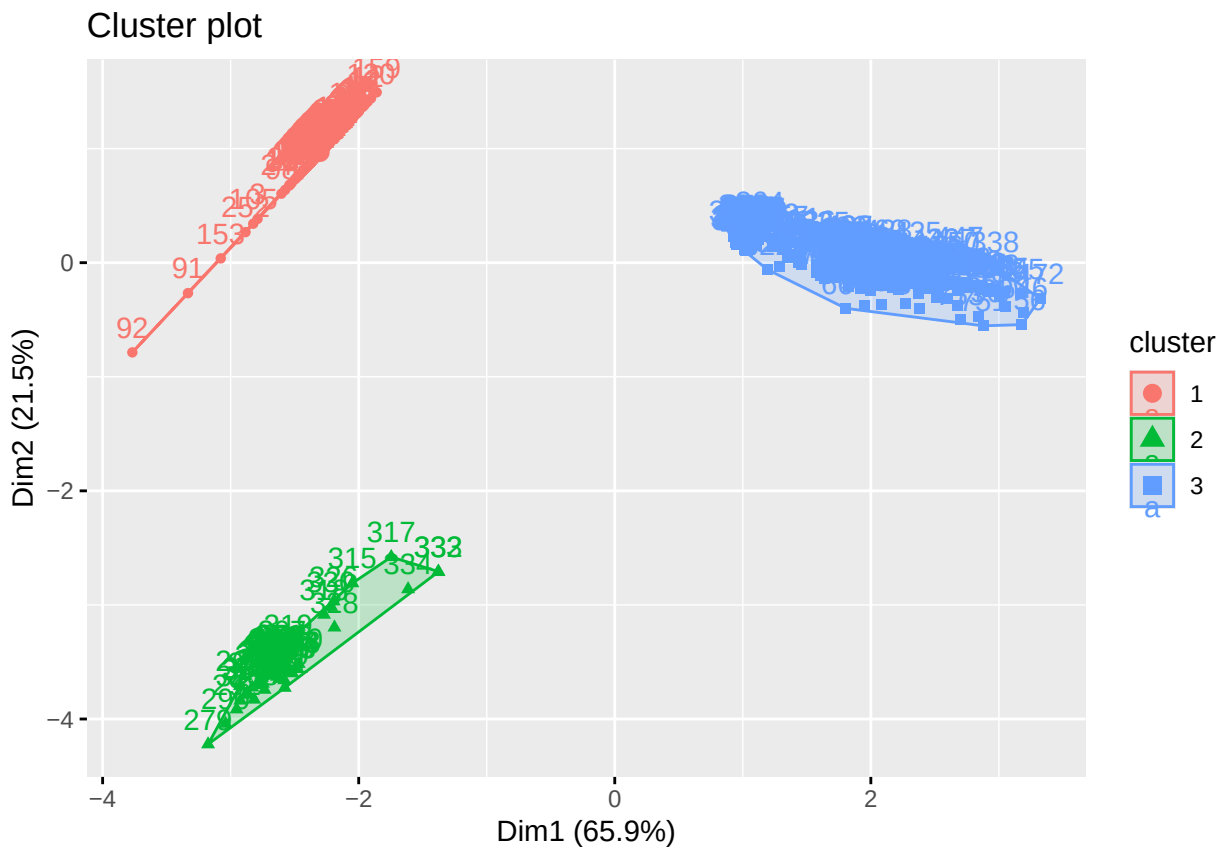
What type of Hierarchical Clustering provided the highest coefficent?

The third type of clustering conducted was Hierarchical. Both AGNES and DIANA where conducted as well as the different methods to perform AGNES, it was determined to use AGNES with the Ward method as it produced the highest coefficient at 0.99887.

Why select 3 clusters instead of what appears to be 5 clusters in the dendogram?

From the AGNES dendogram and the plotted clusters we can see a larger vertical line for one cluster indicating it has more differences then the others. Next, there is a shorter line that splits into two groups with almost equal length lines indicating that the second split has similar properties. This can be confirmed in the visual plot of the clusters that again is similar to our other clustering methods. However, while 3 is being used for our cut line, investigation was done into the cluster furthest to the right which is assumed to be coal. From the dendogram it appears that there are many more variations and a cut line of 5 was looked into. Using 5 for the number of clusters did highlight a difference but there is much overlap and no clear distinction that could just be attributed to the hierarchy process.


Dendrogram of Agnes
Height
0
20
40
60
80
fuelDF.clean
agnes (*, "ward")



Classification Analysis and Discussion

Two types of classification were used in this analysis. First, the data was partitioned into 70% training data, 20% validation data, and 10% test data. Each of these partitions were stratified based on the fuel_type_code_pudl. As with the clustering the fuel_received_units has been left out. Note: The partitioning was completed on non-normalized data and the data was then normalized to conduct KNN and left alone to perform naive Bayes.

KNN

Using the normalized partitions and the training and validation data a loop was created to try values of k 1 - 15 to find the best value of k to use. The results showed that values 1,2,3 or 4 would have 100% accuracy.

What value of k should be used when there is a multi-way tie?

I decided to select the value of 1 for k.I came to this value from not only the test conducted but based on the very distinct clusters that were created in the cluster analysis. Once the value of k was determined the testing data and validation data were combined and a model was created to use the test data. A confusion matrix was created which showed 100% accuracy.

```
## [1] "Confusion Matrix of predictions using model on test data KNN"
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction coal gas oil
##       coal  100   0   0
##       gas     0  61   1
##       oil     0   2  15
##
## Overall Statistics
##
##                Accuracy : 0.9832
##                  95% CI : (0.9518, 0.9965)
##     No Information Rate : 0.5587
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9699
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: coal Class: gas Class: oil
## Sensitivity               1.0000     0.9683    0.93750
## Specificity               1.0000     0.9914    0.98773
## Pos Pred Value            1.0000     0.9839    0.88235
## Neg Pred Value            1.0000     0.9829    0.99383
## Prevalence                0.5587     0.3520    0.08939
## Detection Rate            0.5587     0.3408    0.08380
## Detection Prevalence      0.5587     0.3464    0.09497
## Balanced Accuracy         1.0000     0.9798    0.96262
```

Naive Bayes

Using Naive Bayes a model was created and tested against the validation partition. After viewing the confusion matrix which showed 100% accuracy a new modle was created using both the training and validation partitions and used with the testing data. Another confusion matrix was created indicating a 100% accuracy.

```
## [1] "Confusion Matrix of predictions using model on test data NB"
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction coal gas oil
##       coal  100   0   0
##       gas     0  63   0
##       oil     0   0  16
##
## Overall Statistics
##
##                Accuracy : 1
##                  95% CI : (0.9796, 1)
##     No Information Rate : 0.5587
##     P-Value [Acc > NIR] : < 2.2e-16
##
```

```
##                  Kappa : 1
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: coal Class: gas Class: oil
## Sensitivity               1.0000      1.000    1.00000
## Specificity               1.0000      1.000    1.00000
## Pos Pred Value            1.0000      1.000    1.00000
## Neg Pred Value            1.0000      1.000    1.00000
## Prevalence                0.5587      0.352    0.08939
## Detection Rate            0.5587      0.352    0.08939
## Detection Prevalence      0.5587      0.352    0.08939
## Balanced Accuracy         1.0000      1.000    1.00000
```

Comparison

What does both methods showing 100% accuracy indicate?

100% accuracy indicates that the selected variables are able to classify an energy source extremely well. I did want to confirm that while I was coming up with 100% accuracy for both models would this always hold true? I removed my seed value and ran the code a few times and did have some instances of less then 100% however they were few.

Conclusion

In conclusion I am assuming that this data may be more useful in determining associations around the different energy produces instead of the actual energy source. From the results of the clustering and classification it would seem to suggest that telling Oil, Gas, and coal apart is easier to accomplish then other items. I think back to the wine assignments and could see and find small differences that lead to more overlap which makes sense because of the complexities in taste vs the amount of heat and pollutants that are produced from fossil fuels. It would be interesting to see the changes if additional fuels and data were added to the data set.