EXECUTIVE SUMMARY

In this project, I have tackled the most used transportation for fuel delivery, relation of ash content to fuels, and choosing the best method for analyzing and choosing the value of clusters. At the end of this project, I have concluded that there are numerous materials that comes with extracting, pricing, and delivering fuels, therefore, my recommendations about power generation in the US is to begin the transition to more renewable energy sources such as solar, wind, and hydropower. Though it may cost more initially, the sustainability and reliability it could bring is worth it. Second would be to analyze the long-term impact these fuels are causing to our environment, for example, coal is derived from mines and sooner or later it will affect humans too. Extreme deforestation is needed to clear vast amounts of land to gain access to seams, not to mention soil erosion will start to show as the support and stability of these lands are reduced. Finally, burning coal is a major factor in climate change as it releases carbon dioxide and other greenhouse gases. It is vital to begin the practice of sustainability as early as possible to eliminate unchangeable situations in the future.

INTRODUCTION

The fuel_receipts_costs_eia923 dataset states the records of fuel deliveries to different power plants. As mentioned from the site where it was derived, there are multiple instances of deliveries to the same plant therefore it is vital to understand that all rows could mean just one place.

For my R code, I have utilized the simple elimination of columns 15,18, and 20 since those columns have significant missing values. Including the rowID column, there was a total of 21 columns but after removing the blank ones, we were left with 18.

The dataset has a total of 663,572 observations in which it is inclusive of integers, characters, and numerical. To ensure that we have the correct attributes, I used the str() function to have those outputs and determine which rows have what since it could come in handy when coding for clusters or categorizing variables.

Finally, I used only 2% of the dataset and partitioned it into training and test sets. 70% of the sampled data was for the former while the latter was for testing.

PROBLEM STATEMENTS

Question 1: Which type of transportation is mostly used for delivering different fuels? (Coal, Gas, and Oil)

Question 2: What information is revealed by clustering? Use any and all variables to describe and identify the clusters.

Question 3: How should the value for the number of clusters be chosen?

ANALYSIS AND DISCUSSION

Question 1

To find out the most used type of transportation for delivering fuels, we use a bar plot to visually differentiate. I have compared the three types of fuels and their transportation differences. Starting with coal, delivery via railroad was the most prevalent, almost reaching 3,000 out of the 9,000+ observations sampled while delivery via river and truck are a tie at around 1,000. Second, gas was mostly entirely via pipeline. For example, according to the Energy Information Administration (EIA) website, most U.S. natural gases (up to 99%) come from Canada and is much more in demand during wintertime when it is much more needed to heat up homes and are nearly all delivered by pipeline. Finally, oil is mostly delivered by truck which can be concluded from the plot amounting to roughly about 1,000 deliveries out of the 9,000+ observations.

Question 2

By using the k-means clustering, it is revealed that in the naked eye, a smaller k number would have worked better. There is a congested area in the middle that tells us that the points which are more similar than we think should have been grouped together which results from a lower k, but since k = 4, they had to be dispersed to others. Nevertheless, this cluster includes 1,000 observations focusing on fuels received in units and their ash content. Eliminating the negative values since that does not occur in products, we see that there seems to be no correlation of the amount of delivered fuels to its ash content. For example, looking at ash content with at least 2, we see that it could either be 0 – 3+. The cluster it's in (blue) is also scattered throughout the place which tells us that the number of fuels received varies therefore the amount of ash content will also vary.

Question 3

To determine the value for the number of clusters, we have a few options. First, I utilized the elbow method. In the method, we see that the bend is at 3. Next, I also used the silhouette method to see if it will yield the same result. As we can see, according to this procedure, the optimal number of clusters should be 3, which was the same as our previous result and theory of 3 and a smaller number than 4. In my opinion, the latter method should be chosen as the primary way of selecting the number of clusters since it is much more straightforward, and the user does not have to rely on mere observations like how the elbow method is interpreted.

Upon re-doing the clustering with k=3, we now have a better view of the centroids and points belonging to much more similar characteristics. We can conclude that the highest fuel received has the lowest ash content, while the lowest fuel received has the highest ash content.

CONCLUSION

Beginning with the type of transportation used to deliver fuels, it seems that the most used method would be the pipeline and that is almost, if not all, natural gas. As many people begin to be more aware of climate change and potential harmful environmental effects of these fuels, I assume that in the next 15-20 years, coal, gas, and oil might not be prevail as often as they do now as electric technology is starting to take over.

Adding on to the issue of using biofuels are the amount of mercury, sulfur, and ash in them. As seen in the cluster which only uses the ash contents in relation to fuel received in units, it is evident that the higher ash content, the lesser the fuel delivery. My thoughts are that high ash content has lesser matter in fuels therefore resulting in a reduction of fast ignition. Having less fuel will require the delivery company to send more to fulfill their clients' needs.

Finally, in the next 30 years, I see the trend as sustainability, so I do not expect much demand from all three of these fuels as much. There will also be less damage to our natural resources such as mines and we will be able to preserve our environment better.