**Context**

The business context is a Power Company-seeking meaningful recommendations that can increase its revenue by offering innovative power plans to its customers. Hence, the analysis aims to find clusters of consumers that can benefit from the Power Company’s innovative power plans. Understanding the data set is vital in this context. The dataset contains data on a household’s global apparent power, global active power, global reactive power, voltage, global intensity, and three submetering variables.

Submetering, one corresponds to the kitchen, containing mainly a dishwasher, an oven, and a microwave. Submetering two corresponds to the laundry room containing a washing machine, a tumble drier, a refrigerator, and a light. Submetering three corresponds to an electric water heater and an air conditioner. Global apparent power is the total flowing power or the sum of global active power and global reactive power. Global active power is the power consumed by appliances other than the appliances mapped to sub-meters. Global reactive power is the power that travels back and forth between load and line without any usage.

**Data Preparation**

With an understanding of the data, data preparation can begin. First, the dataset is checked for missing values. No, missing values are found in this dataset. Then a boxplot of each attribute reveals potential inconsistent data or outliers. Because our method of clustering is k-means, it is vital to remove the outliers to create high-quality clusters with high intra-class similarity and low inter-cluster closeness. Outliers are removed from all attributes except the sub-meter attributes. This is because once the outliers are removed from those attributes, the k-means algorithm cannot handle NA values that are produced when attempting to scale the data in the sub-meter attributes once the outliers are removed.

**Clustering Model**

Now that the data has been prepared for clustering, several k-means models are implemented. Initially, clustering is attempted on the entire dataset. However, these clusters proved challenging to interpret because of the large selection of attributes. Therefore, the data is reduced in two cases. The first reduction reduced the dataset to the global apparent power, voltage, and intensity attributes. These attributes are selected to identify clusters of consumers with large power consumption. Six clusters are selected using the “elbow method” to maximize the information gained from each cluster. The second reduction reduced the dataset to sub-meter one, sub-meter two, and sub-meter three attributes. These attributes are selected to identify consumers with similar sub-meter power consumption clusters. Using the “elbow method,” four clusters are chosen to maximize the information gained from each cluster.

**Results and Interpretations**

Once the k-means algorithms are implemented, tables and visualizations are generated to interpret the results. In the first case, two clusters have noticeably higher global apparent power than the other clusters. Those clusters are four and six, containing 150 and 156 observations, respectively. The Power Company can target these two clusters with innovative methods to reduce their global apparent power and, as a result, gain new customers, increasing revenue. The other case implements a k-means model on the sub-meter attributes. The results showed that cluster one has a higher mean for submetering than the other clusters. Cluster three has the highest mean for sub-metering two, and cluster two has the highest mean for sub-metering three. Cluster one will unlikely increase revenue because cluster one has 15 observations. So, efforts should be made to offer innovative plans to clusters two and three. Results from cluster two indicate that the household’s kitchen produces the most power compared to other clusters. And households in cluster three have water heaters and air conditioning units that consume lots of power. If offering innovative plans to both clusters requires large amounts of resources and is not possible, efforts should be focused on providing an innovative to households in cluster two because 714 households could potentially benefit. Thus, the plan will likely yield more revenue if offered to more households.

**R Code**

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Cluster Visualizations**

Chart

Description automatically generated

Text

Description automatically generated

-------------------------------------------------------------------------------------------------------------------------------

Chart, line chart

Description automatically generated

Text, letter

Description automatically generated