## Introduction:

With a dramatic increase in commercial buildings' energy consumption over the past years, consequences such as energy shortages and carbon emissions threatening our living environment have been common to us. Therefore, building energy efficiency can provide solutions to address those issues and mitigate the negative impact of the high demand energy consumption. It is important to identify buildings that are energy-inefficient to reduce the carbon footprint and rising prices of energy. The given dataset contains information about buildings' monthly consumption of energy and their characteristics. I applied machine learning techniques such as clustering to group unlabelled data and a decision tree regression to estimate a benchmark for the energy consumption.

## Methodology:

In order to decide which buildings with particular characteristics are energy efficient or inefficient, we need to specify the energy consumption threshold for those buildings. Since not all buildings are similar, we need to stratify them into groups of similar features and estimate a benchmark for each category. Then, we could compare the buildings' actual energy consumption from the dataset with the benchmark to label them as either efficient or inefficient in terms of the energy usage.

With the use of machine learning techniques, I went along with the following methodology:

1. Pre-process data by dropping unnecessary columns and binarizing categorical variables
2. Cluster data into certain groups using the agglomerative hierarchical clustering approach
3. Utilize a decision tree regression model for each cluster to predict the consumption energy values
4. Take the average of those predicted values for each cluster as its benchmark
5. Compare the benchmark with the buildings' actual energy consumption assigned to the cluster

**1**. Firstly, I removed some features from the dataset because they are repetitive. For instance, there were columns such as *country, longitude* and *latitude*. Since *latitude* and *longitude* represent a specific location of a certain building, dropping the attribute *country* can reduce multicollinearity between features. Moreover, *latitude* and *longitude* are numerical variables. Hence, it would be easier to perform clustering. Another circumstance is when values of one attribute are in the subset of other attribute values such as features *PropertyType* and *PropertySubType*. I kept the latter variable since it has more values and contains all values from the first one. Furthermore, I created a new variable *month*, which is derived from the column *PeriodLabel.* I would like to have information about seasonal energy consumptions in terms of months instead. Having reduced the dimension of the dataset, I binarized categorical variables for clustering purposes and removed some resulting dummy variables NULL.

**2**. Clustering helps to group data in a certain way that the data in the same category are more intrinsically similar than those in other groups. To juxtapose one building energy efficiency with others, they need to be in the same category. There are many clustering methods, but I selected the agglomerative hierarchical clustering.

Hierarchical clustering starts with clusters containing one data point each (singleton clusters) and then merge two closest clusters at each step until one cluster is left. The whole process produces a dendrogram. Compared to other clustering methods such as k-means, the number of clusters is not required to be specified at the beginning while clustering data. We can navigate through the tree (dendrogram) to look

for the number that makes the most sense to our application. In addition, the hierarchical clustering has fewer assumptions about the distribution of data unlike k-means. Below is the dendrogram of our data produced by the hierarchical clustering.

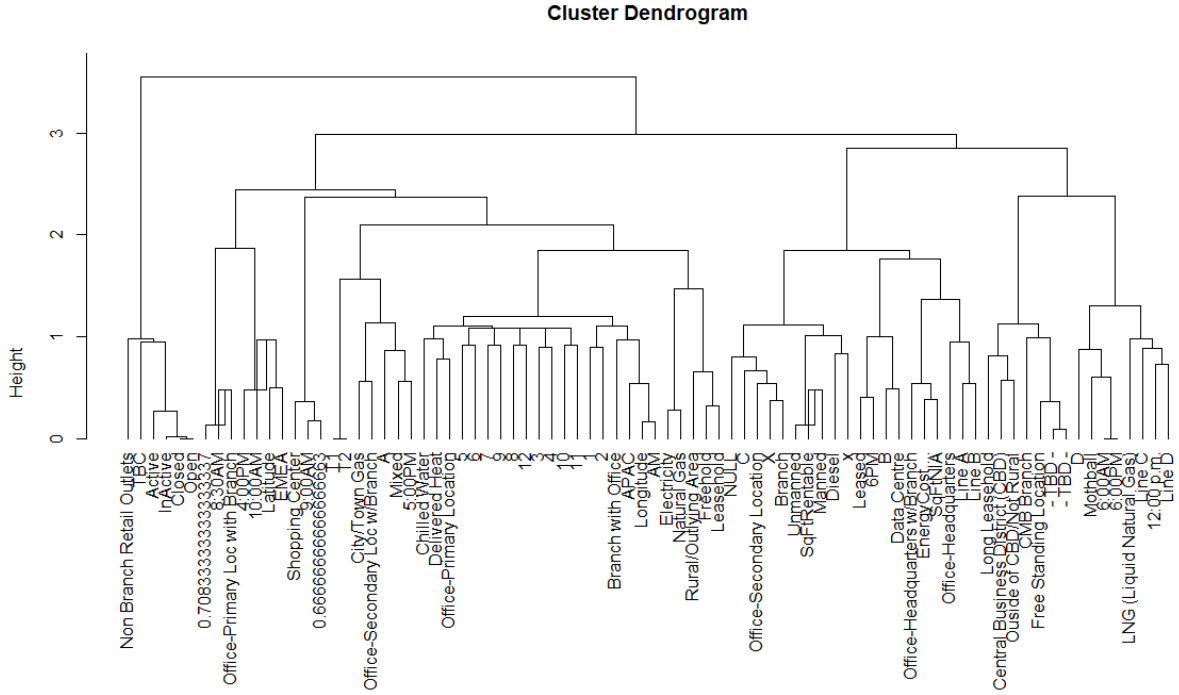


*Fig 1. Resulted Cluster Dendrogram*

As shown in Fig 1, the variables originated from the same node of the tree are linked. Even though there is a connection between two attributes, it does not infer whether the relationship is positive or negative. I have used the function *hclustvar* from *clustofvar* library in R because it deals better with categorical variables. To identify an appropriate number of clusters, I decided to utilize the stability function. The function is based on a bootstrap approach that evaluates the stability of partitions received from a hierarchy of variables. It uses a defined clustering method, in this case *hclustvar*, to apply to B bootstrap samples of n rows. The partitions of 2 to m clusters resulted from the B bootstrap hierarchies are compared with the initial hierarchy's partitions. From those partitions, we can calculate averaged Rand indices (agreement between partitions). The larger the index is, the higher the agreement is between the two partitions. Usually more B samples would be better because they give more convergent Rand index results. However, I chose B=20 bootstrap samples to be included in my model due to the computational limitation of my computer.

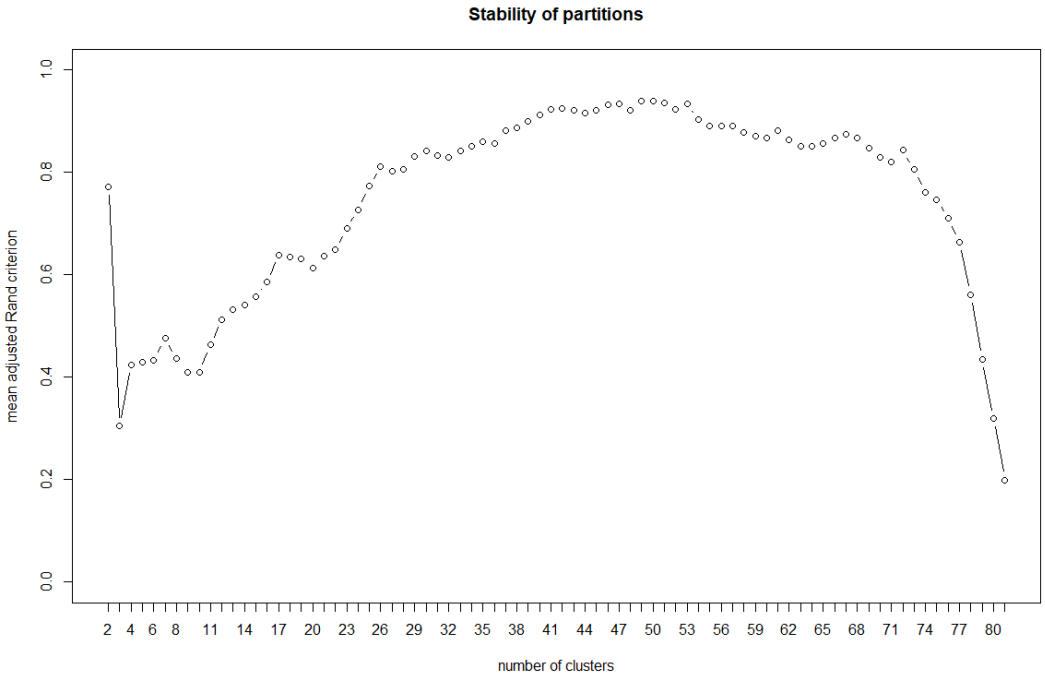


*Fig 2. Stability of Partitions based on B=20*

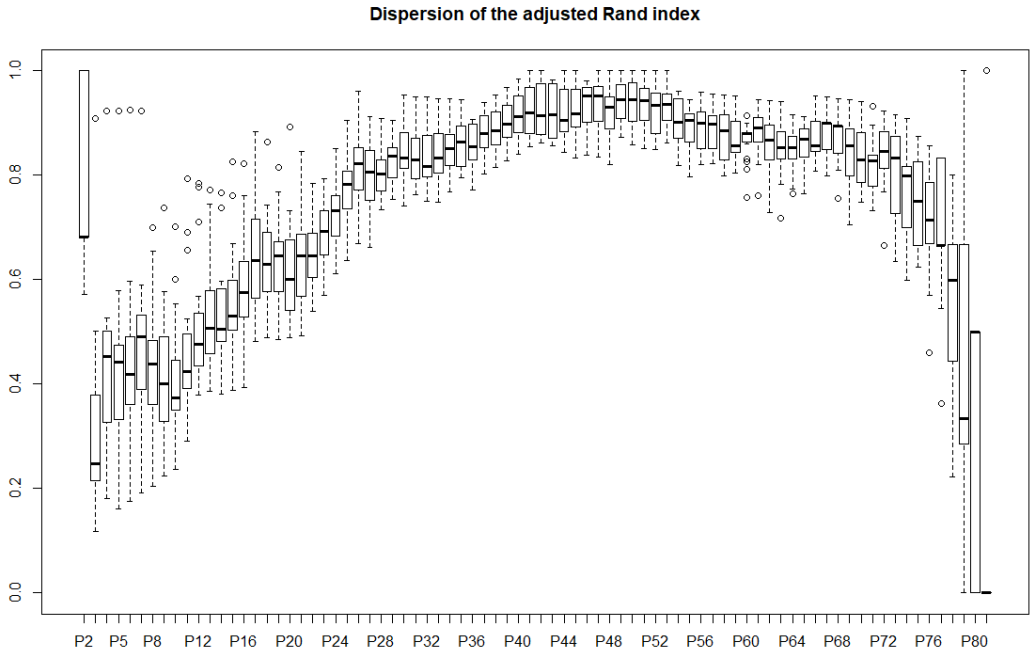


*Fig 3. Dispersion of The Adjusted Rand Index*

It might seem from Fig 2 and Fig 3 that 46 clusters should be selected since they give the best mean adjusted Rand index and the deviation from the mean is small. But, some groups from the 46 clusters have only one attribute variable. As far as I am concerned, we need at least two variables in each cluster to determine the energy consumption. Otherwise, it would be a simple linear relationship between that variable and the energy consumption.

Moreover, anything above 46 clusters would produce more clusters with one variable. Instead, I looked at the range of the number of clusters between 18 and 46 because they have an index measure of at least 0.8 roughly. It turns out that 28 clusters are appropriate because all clusters have more than one variable and the gain in cohesion, which is a percentage of homogeneity by partition, is close to 0.6. Hence, I have cut the dendrogram into 28 clusters.

**3**. Having clustered the data into 28 groups, I have created a decision tree regression for each cluster. The decision tree creates a non-linear relationship between variables and copes quite well with missing values. I have built 28 cluster models via the tree wherein the dependent variable is the buildings' monthly energy consumption and the features are building's characteristics assigned to a particular cluster.

| **RMSE** | **Rsquared** | **MAE** | **Cluster** |
|---|---|---|---|
| 4026.648 | 0.936268 | 2577.885 | cluster4 |
| NA | 0.011965 | NA | cluster13 |
| 100039.9 | 0.761707 | 43800.96 | cluster12 |
| 25117.87 | 0.328683 | 17406.35 | cluster19 |
| NA | 0.288155 | NA | cluster2 |
| NA | 0.112394 | NA | cluster20 |
| NA | 0.345057 | NA | cluster10 |
| 98158.51 | 0.255218 | 78282.88 | cluster15 |
| 4406.787 | 0.394366 | 3255.75 | cluster3 |
| NA | 0.968936 | NA | cluster14 |
| NA | 0.604505 | NA | cluster8 |
| 136444.2 | 0.301043 | 106951.9 | cluster7 |
| 223584.7 | 0.983739 | 134251.8 | cluster1 |

*Table 1. Decision Tree Regression Metrics*

The table is an incomplete snapshot of metric results produced by each cluster decision tree regression. Cluster 4 seems to have a good decision tree model since R^2 is high and the error is relatively smaller than other models. The missing error values in remaining clusters are due to the fact that the energy consumption variable has some missing data. There are more missing metric information in other clusters 23, 25 or in 26. If we dropped rows in our dataset that have any missing information in a cell, the dataset would be small.

**4**. Then, I have predicted buildings' energy consumption and take the average of the forecasted values as a benchmark for buildings clustered in the same group.

**5**. Finally, I have labelled a building as energy inefficient if its energy consumption is greater than the benchmark. Otherwise, the data is recorded as efficient. I have saved my results in *output.csv* file and performed additional visualisation graphs in Tableau to present results of energy efficiency.

## Results

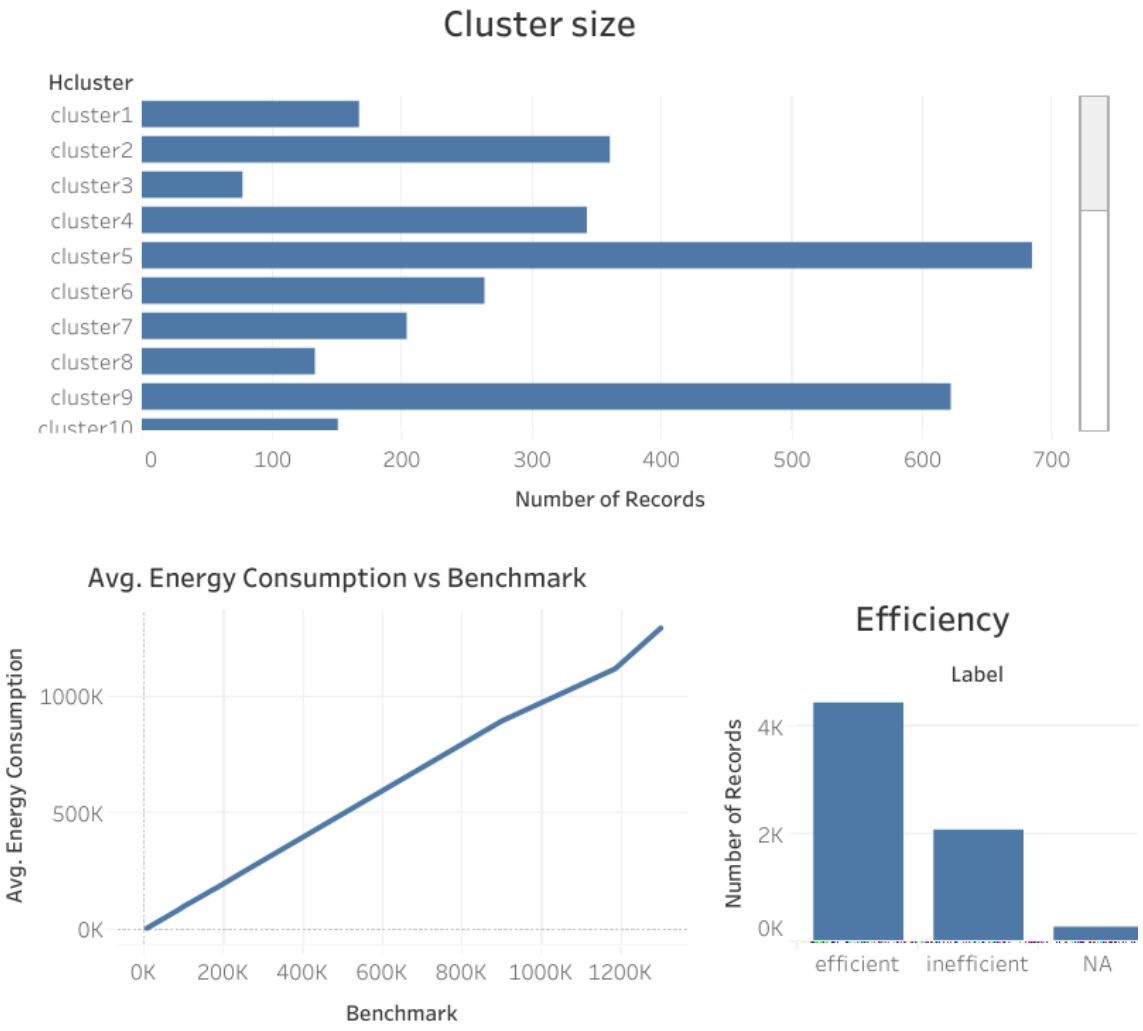


*Fig 4. Dashboard Visualization 1*

As shown above, each cluster has a distinct size. *Cluster 5* has the most building members whereas *cluster 18* possesses the smallest volume. The average of buildings' actual energy consumption is almost perfectly and linearly correlated with the calculated benchmark. Since the benchmark is the mean of the predicted energy consumption, this means that discrepancy between actual and forecasted energy consumption values is small. This reinforces the strengths of our machine learning methods. Overall, there are more buildings that are energy efficient than inefficient ones. Some of the buildings (*NA*) could not be identified because they have missing information regarding the energy consumption in the given dataset.


*Fig 5. Dashboard Visualization 2*

Regarding the seasonal trend, we can see that buildings consume more energy in the first quarter and at the end of the year. This is not an anomaly since more energy is used for heating in Winter. Additionally, the proportions of the efficient and inefficient energy consumption stay almost the same throughout the year. From the second graph in Fig 5, it is easy to notice that various property types utilize the amount of energy differently. Branches seem to manage and balance their energy consumption well compared with others that do not.

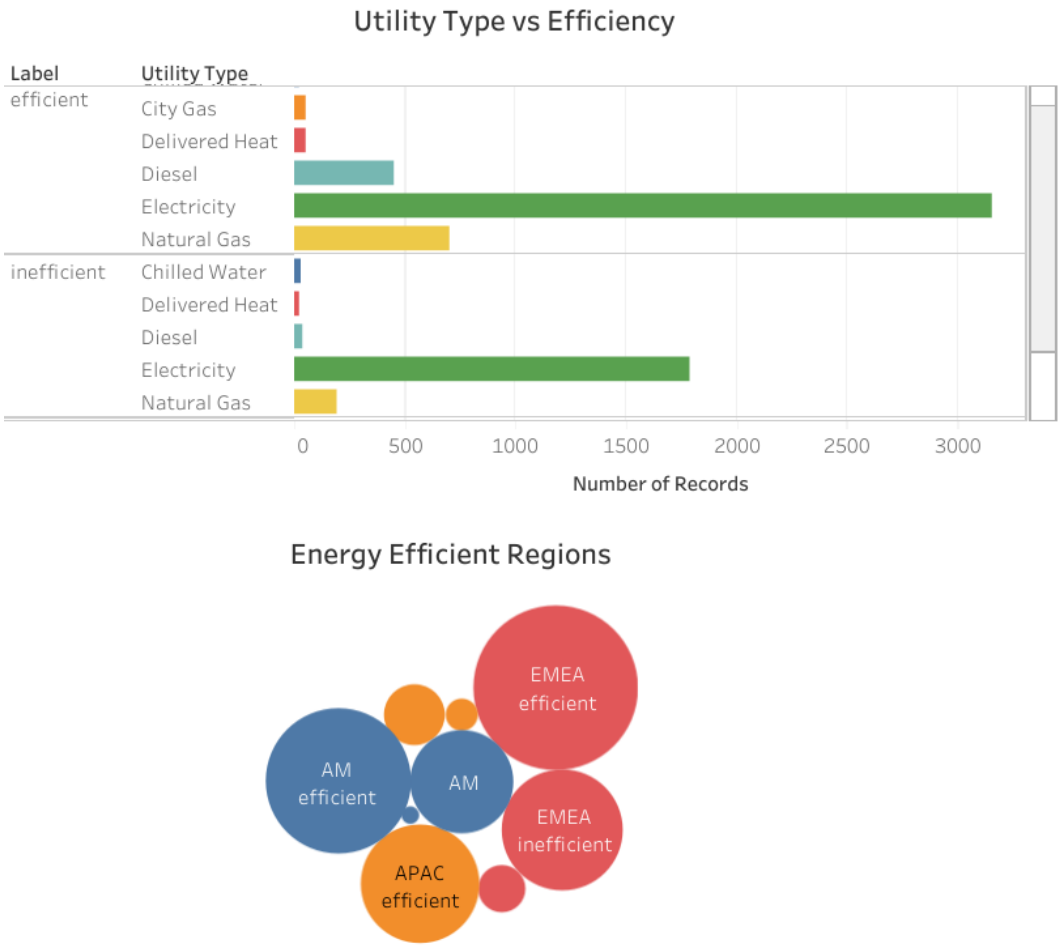


*Fig 6. Dashboard Visualization 3*

When it comes to a utility type, electricity is the most common product of energy. Though a lot of buildings use electricity more efficiently, there is still a significant amount of waste. By looking at the proportions of the efficient to inefficient energy usage, commercial buildings with a natural gas or diesel manage well their power consumption. Fortunately, all regions tend to consume energy wisely. However, there is still a plenty of room for those regions to improve and reduce the wasteful amount of energy in their commercial buildings.