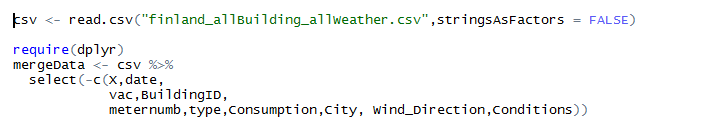
# Clustering (k-means)

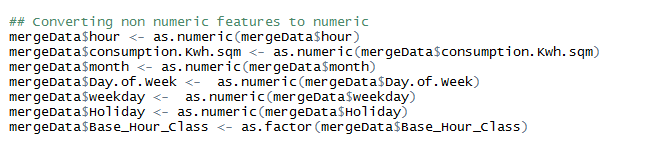
K-Means is a clustering approach that belogs to the class of unsupervised statistical learning methods. The general idea of a clustering algorithm is to partition a given dataset into distinct, exclusive clusters so that the data points in each group are quite similar to each other.

Let’s read our building records as one large dataset of around 6 million and try to find cluster’s.

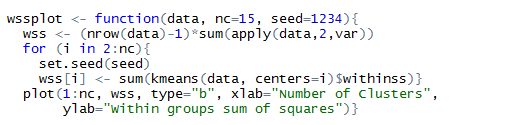
1. Reading the csv and removing features which we would not use and will not make sense in the clustering



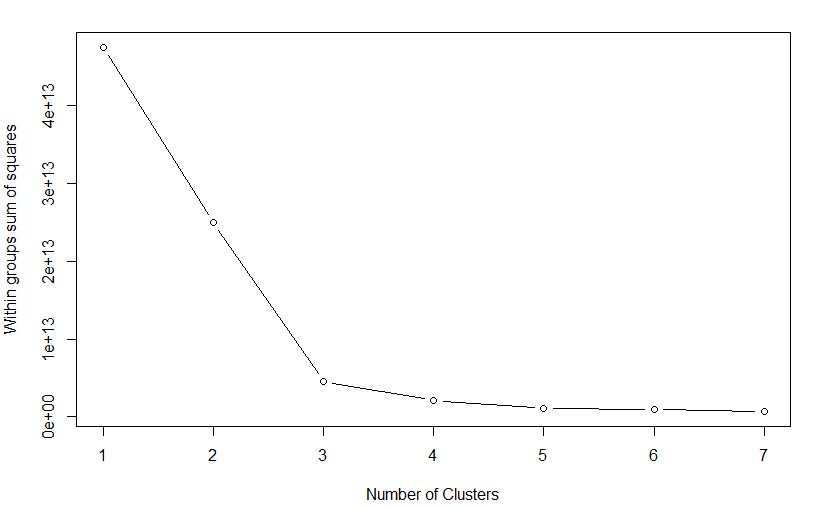
1. Converting the features to appropriate data type



1. A plot of the within groups sum of squares by number of clusters extracted can help determine the appropriate number of clusters. We will use bend graph function to check within groups sum of squares by number of clusters and deciding the optimal number of clusters







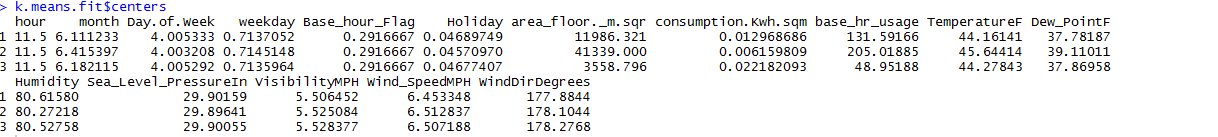
1. From the graph we can see that, within groups sum of squares is not improving after K =3
2. Using k=3, Let’s use the **kmeans** function from R base stats package. Removing Base\_Hour\_Class from the feature list. Hence mergedata[,-10]



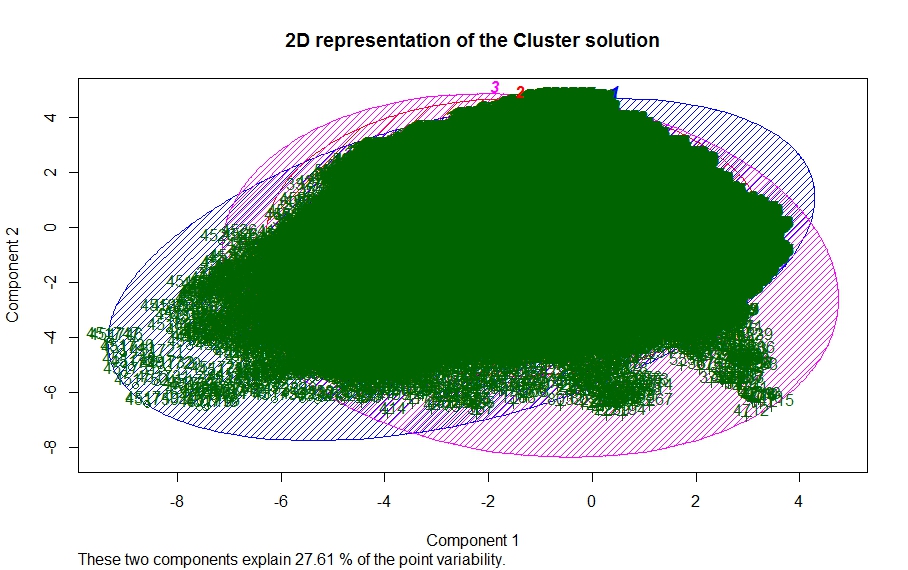
1. We can see the Within cluster sum of squares by cluster percentage of 90.5% which suggests good variance in between cluster



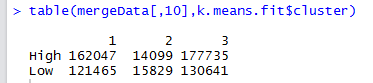
1. Further analyzing the results of our model, we can find out the centers of our clusters according to the feature



1. We can see good separation in terms of **area\_floor, consumption, base\_hr\_usage**.
2. Plotting the result in 2-D space would give some idea about out cluster. Since the data is huge, we will not able to analyze the cluster’s well



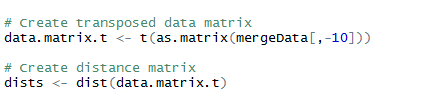
1. Plotting against our Base\_Hour\_Flag, we can see the below result



1. We can say from this interpretation is that, the Base\_Hour\_Flag is missing a cluster of records. It should be **High, Low** and maybe **Neutral (for the consumption which is equal to the base\_hour\_usage)**

# Clustering (hierarchical clustering)

1. K-means clustering requires us to specify the number of clusters, and finding the optimal number of clusters can often be hard**. Hierarchical clustering** is an alternative approach which builds a hierarchy from the bottom-up, and doesn’t require us to specify the number of clusters beforehand.
2. Hierarchical methods use a distance matrix as an input for the clustering algorithm. The choice of an appropriate metric will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another
3. We can use hclust for this. hclust requires us to provide the data in the form of a distance matrix. We can do this by using dist. By default, **the complete linkage method is used.**
4. Now since the data is huge, we will first transpose our data to a data matrix, and calculate the distance matrix

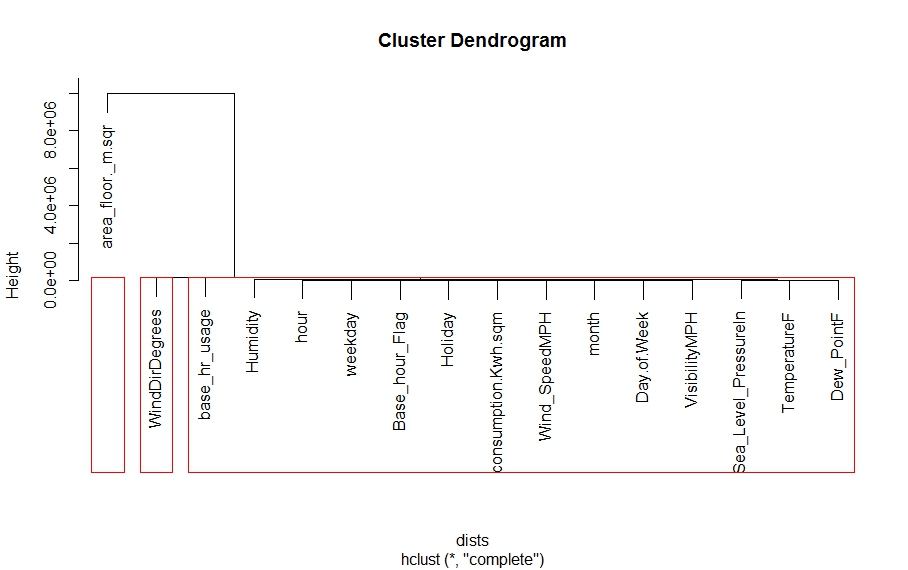


1. Now calculating the cluster

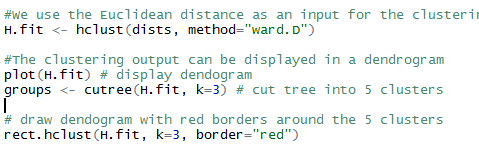


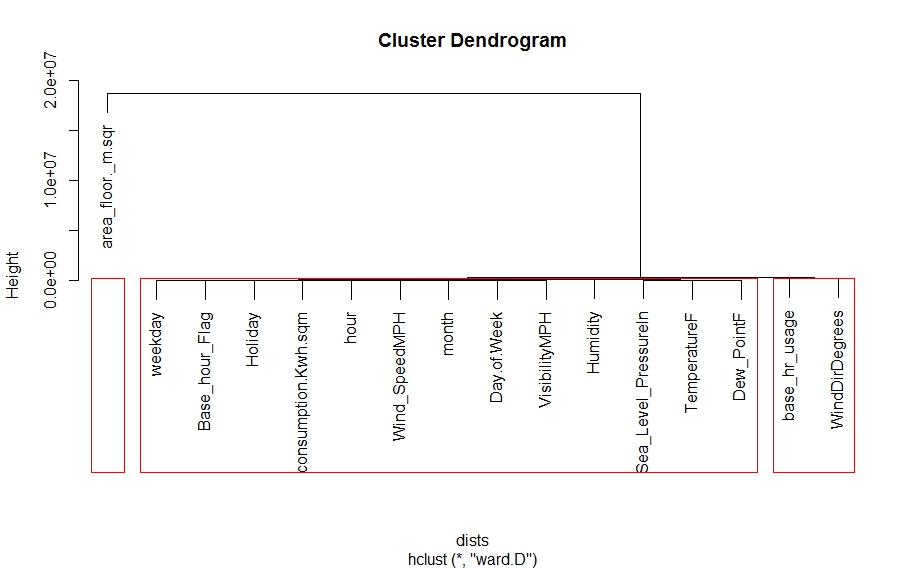
1. We can draw the dendogram and see the results





1. From the hierarchical clustering we can see that, 3 clusters are formed. With area\_floor again driving the cluster segmentation. WindDirDegrees a major contributor in a cluster and the rest for the third cluster
2. We can also use Ward’s minimum variance criterion to minimizes the total within-cluster variance and plot the result





1. A different clustering group can be analyzed from here. Base\_hr\_usage now influencing the second cluster and rest remains the same. This is pretty much in alliance with our K-means clustering algorithm