# MACHINE LEARNING
# ASSIGNMENT - 4

ID:008458886
Vamsi Krishna

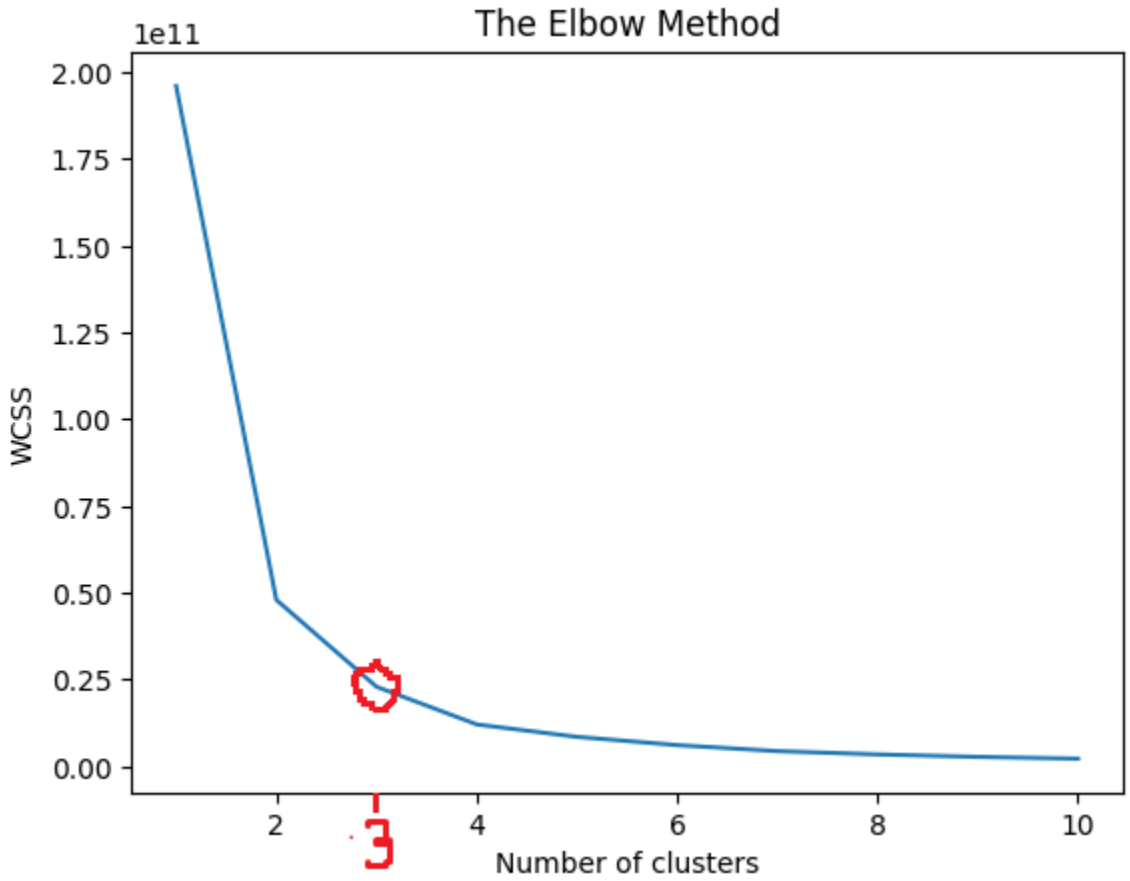
## Overview of the Dataset:

I'm considering **Health Insurance Data** to apply **K-Means Clustering** and **Hierarchical Clustering**. In that I'm considering **BMI** and **Insurance Charges** data column values to enforce above methods of clustering.

### 1. K-Means Clustering:

K-Means Clustering is an Unsupervised Learning Algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of predefined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

**Elbow Method:**



# MACHINE LEARNING
# ASSIGNMENT - 4

ID:008458886
Vamsi Krishna

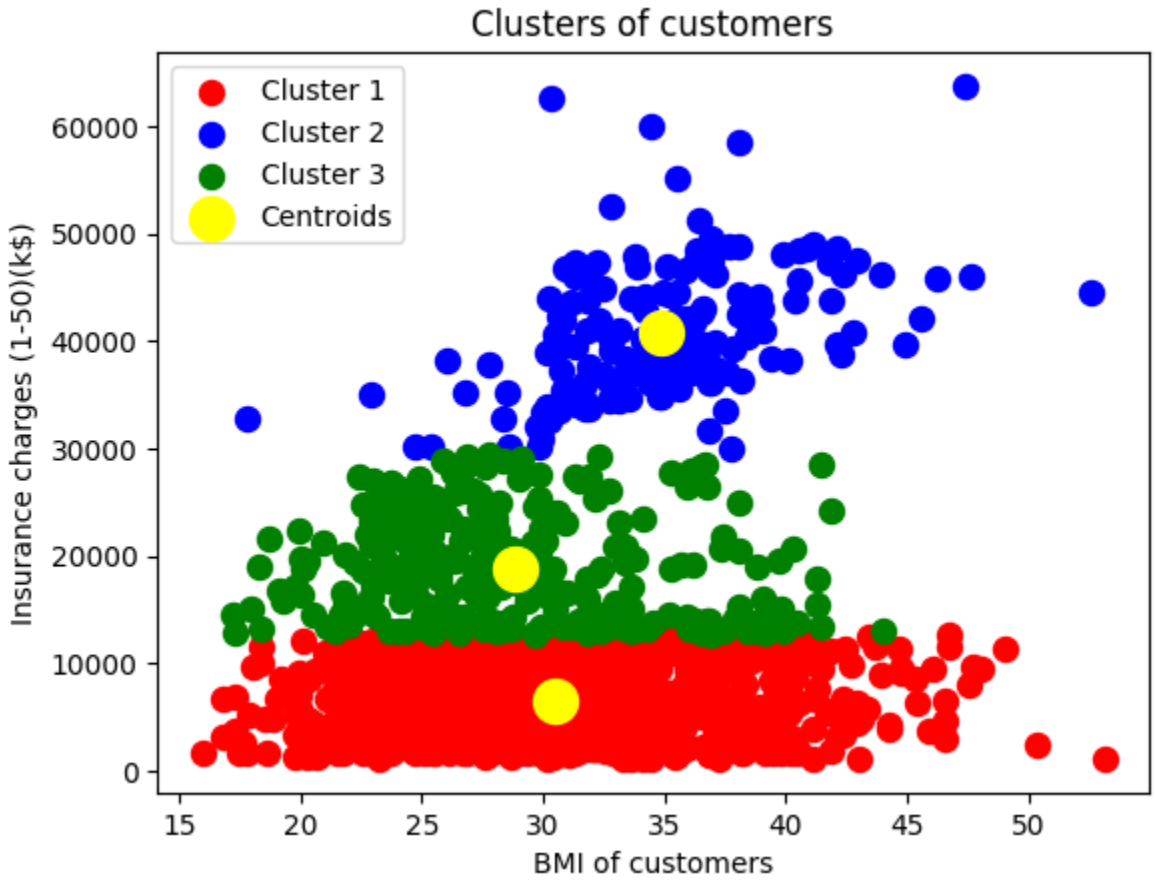
I am using the Elbow method to find the optimal number clusters. This method uses the concept of **WCSS** (Within Cluster Sum of Squares).

If we observe a sharp bed(above figure) , the sharp point of bend or a point of the plot looks like an arm, then that point is considered as the best value of K.

So, we are considering the 3 as the best value from the Elbow method which we can consider as the optimal number of clusters.

**Visualizing the clusters:**

Once we get to know the optimal number of clusters from the Elbow Method then we will visualize the clusters like in the figure below.



# MACHINE LEARNING
# ASSIGNMENT - 4

ID:008458886
Vamsi Krishna

**Observation:**

The above figure clearly shows that we have 3 different clusters with different colors. As we know that, the clusters are formed between two parameters of the dataset, **BMI** and **Insurance Charges**.

Customers in **Cluster 1**, characterized by higher BMI values, exhibit a trend of paying higher insurance charges. This group can be labeled as **Careless**.

Customers in **Cluster 2,** characterized by lesser BMI values than customers in **Cluster 1**, exhibit a trend of paying lesser insurance charges than customers in **Cluster 1**. This group can be labeled as **Less Careless**.

Customers in **Cluster 3,** characterized by lesser BMI values than customers in **Cluster 1** and **2**, exhibit a trend of paying lesser insurance charges than customers in **Cluster 1** and **2**. This group can be labeled as **Careful**.

**2. Hierarchical Clustering:**

Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as hierarchical cluster analysis or HCA.

In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the dendrogram.
Sometimes the results of K-means clustering and hierarchical clustering may look similar, but they both differ depending on how they work. As there is no requirement to predetermine the number of clusters as we did in the K-Means algorithm.

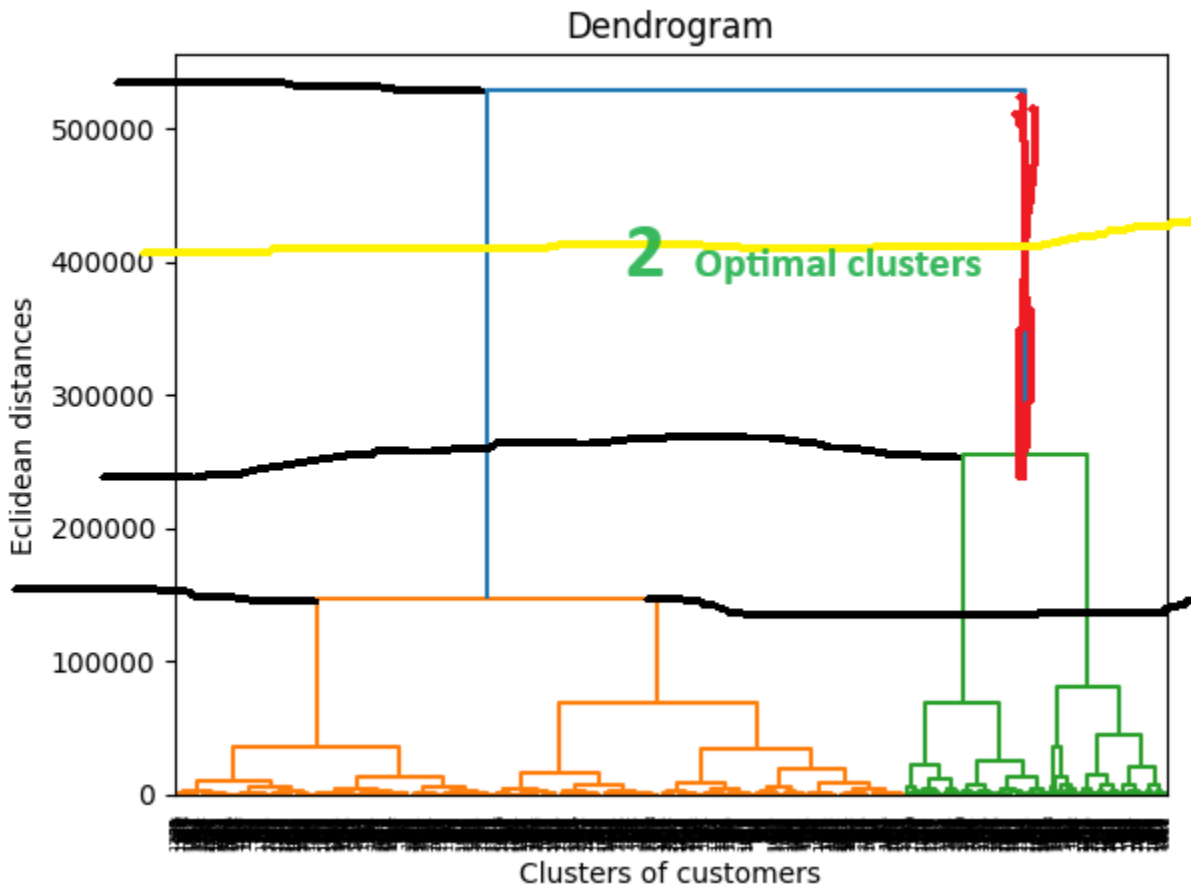
The hierarchical clustering technique has two approaches:

# MACHINE LEARNING
# ASSIGNMENT - 4

ID:008458886
Vamsi Krishna

1. **Agglomerative**: Agglomerative is a bottom-up approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
2. **Divisive**: Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

We are using an **Agglomerative approach** and we will create a **Dendrogram** to get to know about the optimal number of clusters.




The **Dendrogram** is a tree-like structure that is mainly used to store each step as a memory that the HC algorithm performs. In the dendrogram plot, the Y-axis shows the Euclidean distances between the data points, and the x-axis shows all the data points of the given dataset.

# MACHINE LEARNING
# ASSIGNMENT - 4

ID:008458886
Vamsi Krishna

The above graph shows that we can consider 2 clusters to perform Hierarchical Clustering. Because the vertical line that's marked in red is highest and it has 2 clusters.

**Visualizing the Clusters:**

Once we get to know the optimal number of clusters from the Dendrogram then we will visualize the clusters like in the figure below.


**Observation:**

By referring to the above observation(K-Means) and above figure, we can observe that Cluster1(Careless) and Cluster 2(Less Careless) Customers from K-Means Observation become **Cluster1** of Customers in **Hierarchical** when the

# MACHINE LEARNING
# ASSIGNMENT - 4

ID:008458886
Vamsi Krishna

cluster value is 2. **Cluster2** customers here in **Hierarchical** are considered as **Careful**.