UNDERSTANDING CLUSTER ANALYSIS

USING CUSTOMER SEGMENTATION CASE STUDY

1. INTRODUCTION

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group than those in other groups. In simple words, the aim is to segregate groups with similar traits and assign them into clusters. The goal of performing a cluster analysis is to sort different objects or data points into groups in a manner that the degree of association between two objects is high if they belong to the same group, and low if they belong to different groups.

Let's understand this with an example. Suppose, the head of a rental store wish to understand preferences of his costumers to scale up his business. it is not possible for him to look at details of each costumer and devise a unique business strategy for each one of them. But what he can do is to cluster all of your costumers into say 10 groups based on their purchasing habits and use a separate strategy for customers in each of these 10 groups. And this is what we call clustering.

2. OBJECTIVE

In our assignment we take a dataset which is created for the learning purpose of the customer segmentation concepts. **We will use this dataset to understand two major cluster analysis(clustering) algorithms in the simplest form.** Through a membership cards of supermarket mall, it has some basic data about its customers. Now the owner of the mall wants to understand the customers, like who can be easily converge [Target Customers] so that the sense can be given to marketing team and plan the strategy accordingly to maximize profit.

3. SOURCE OF DATA

From the membership card of that particular mall the company can derived the customers information such as customerID, age, gender, customer's annual income and their spending score (1-100). You can find the dataset and the details in the link given below.

https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python

4. PRE-PROCESSING

We distribute the whole study into three parts:

- Part I consists of loading the dataset.
- Part II consists of the exploratory analysis of the data.
- Part III consists of implementing Cluster Analysis of the data.

4.1. LOADING THE DATASET

Here we have used Python on Jupyter Notebook for implementation and visualization.

```
In [9]: data=pd.read_csv('C:/Users/aritr/ProjectDatasets/Mall_Customers.csv')
```

```
In [42]: data.head(5)
```

Out[42]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

```
In [43]: data.tail(5)
```

Out[43]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

As we can see, the dataset consists of **5 rows**: *'CustomerID', 'Gender', 'Age', 'Annual Income(k$)', 'Spending Score (1-100)'*. Also, we can see that it has a total of **200 records** (*customerID 0 to customerID199*). Here, *data.head(5)* shows the record of first 5 customer membership card record and *data.tail(5)* shows the record of last 5 customer membership card record.

After loading the dataset, we head towards analyzing all the attributes in the dataset to see if we can find any abnormalities before cluster analysis.

4.2. EXPLORATORY DATA ANALYSIS

It is a good practice to understand the data first and try to gather as many insights from it. Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. In simple words, EDA is all about making sense of data in hand, before getting them dirty with it.

4.2.1. GENDER DISTRIBUTION

```
In [14]: plt.rcParams['figure.figsize']=(7,7)
         piedata=data['Gender'].value_counts()
         plt.pie(piedata,labels=['Female','Male'],autopct='%1.1f%%')
         plt.title('Gender Distribution',fontsize=22)
         plt.show()
```

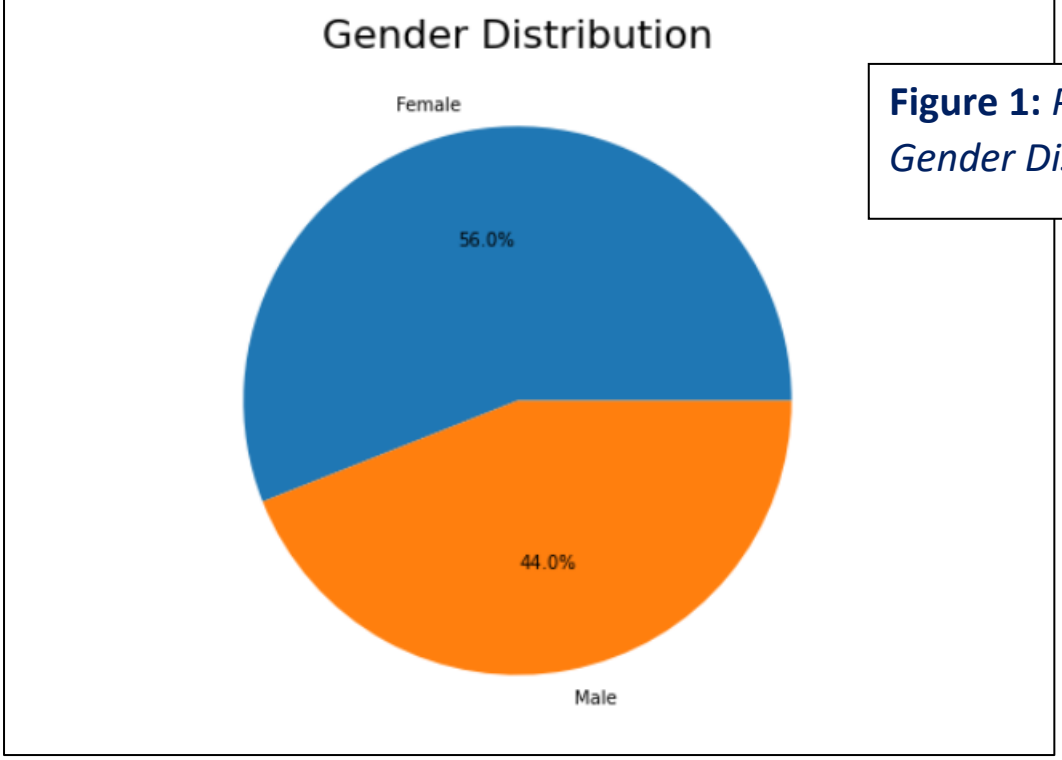


Figure 1: *Pie Chart Showing Gender Distribution*

From the pie chart in *Figure 1*, we can see that the **56.0% is female** and the remaining **44.0% is male**. From this, we get a clear picture that the **population of female is more than that of male in this particular data set**. That's a huge gap specially when the population of Males is comparatively higher than Females.

4.2.2. AGE DISTRIBUTION

```
In [15]: plt.rcParams['figure.figsize']=(25,6)
         plt.subplot(1,2,2)
         sns.set(style='whitegrid')
         sns.distplot(data['Age'],color='g')
         plt.title('Age Distribution', fontsize = 22)
         plt.xlabel('Age', fontsize=16)
         plt.ylabel('Count', fontsize=16)

Out[15]: Text(0, 0.5, 'Count')
```

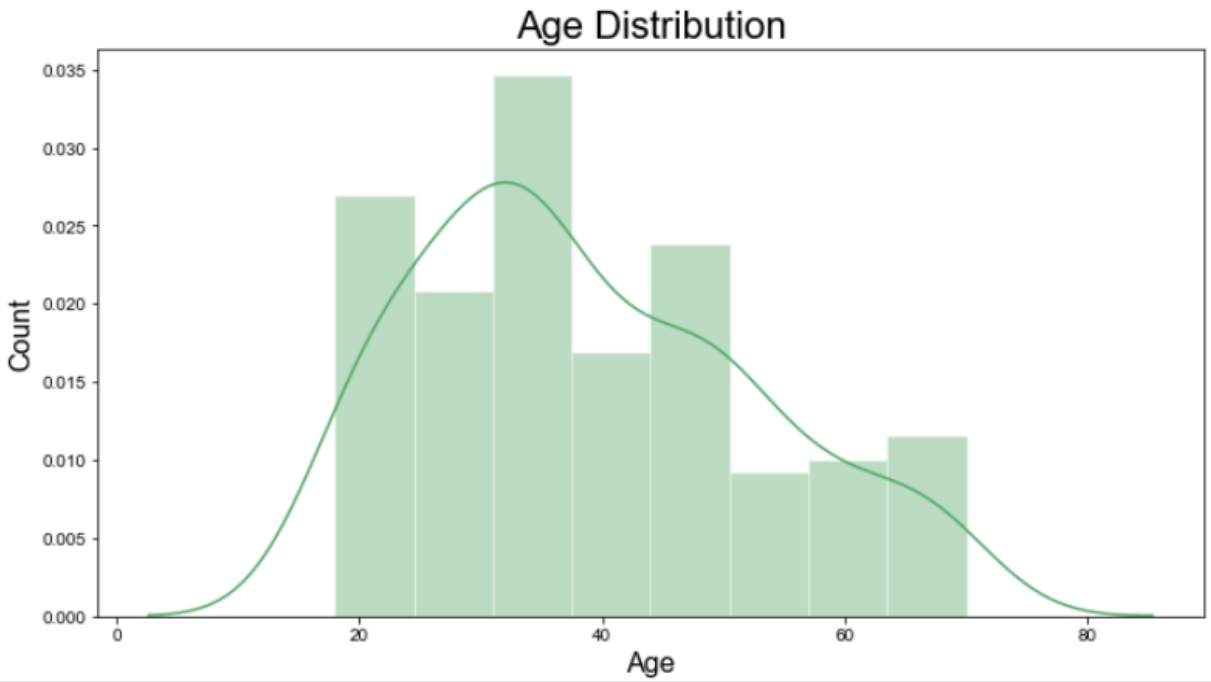


Figure 2: *Distribution Plot for Age*

While going through the graph in *Figure 2*, we can analyse that **most of the consumers are between the age of 20 and 40**. Highest count is about 0.035, which is approximately between the age of 28 and 35. And the **lowest count is 0.010, around between the age of 50 to 58**. From this we could say that the majority of consumers are the people belonging to the Early Adulthood.

Thus, we can concentrate on the fact that there are big differences in the number of people visiting the mall from a particular age group.

4.2.3. ANNUAL INCOME DISTRIBUTION

```
In [16]: plt.rcParams['figure.figsize'] = (25, 6)

         plt.subplot(1, 2, 1)
         sns.set(style = 'whitegrid')
         sns.distplot(data['Annual Income (k$)'])
         plt.title('Annual Income Distribution', fontsize = 22)
         plt.xlabel('Range of Annual Income', fontsize=16)
         plt.ylabel('Count', fontsize=16)

Out[16]: Text(0, 0.5, 'Count')
```

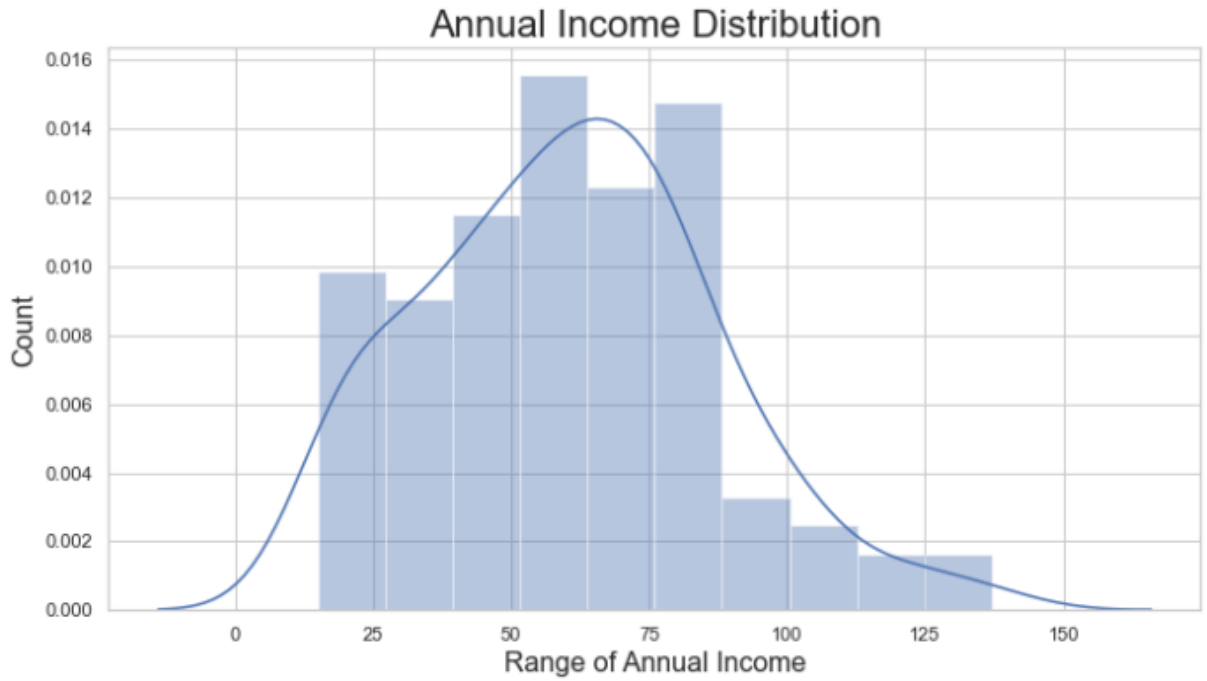


Figure 3: *Distribution Plot for Annual Income of Customers*

In *Figure 3*, we can observe that **most of the consumer's annual income is approximately between 50k$ and 60k$** with count around 0.016. Only count nearly to 0.002 has high annual income around 112.5k$ to 137.5k$. Mainstream consumers have annual income between 20k$ to 87.5k$.

Here, we can see how the **people earning from 50k$ to 87.5k$ are highest** in number to visit the mall. It definitely plays an important role as to which class of people generate more profit and also on which class to uplift.

4.2.4. SPENDING SCORE DISTRIBUTION

```
In [17]: plt.rcParams['figure.figsize']=(35,15)
         plt.subplot(2,3,3)
         sns.set(style='whitegrid')
         sns.distplot(data['Spending Score (1-100)'],color='r')
         plt.title('Spending Score Distribution', fontsize = 22)
         plt.xlabel('Spending Score', fontsize=16)
         plt.ylabel('Count', fontsize=16)

Out[17]: Text(0, 0.5, 'Count')
```

Figure 4: *Distribution Plot for Spending Score of Customers*

As per the *Figure 4*, the highest count is near to 0.018 with spending score approximately between 50 and 62. The lowest count is around 0.007 with spending score between 62 and 76. **The highest count of the consumers have spending scores between 38 to 62**. The highest spending score is between 84 to nearly 100.And the lowest spending score is between 2 to 14 with a count above 0.008.

Thus, because of the variation in the Spending Score distribution plot, it shows that the mall caters to the variety of customers with varying needs and requirements available in the mall.

5. CLUSTER ANALYSIS

Cluster analysis or Clustering is a form of data mining in which observations are divided into different groups that share common characteristics. The purpose of cluster analysis is to construct groups while ensuring the following property: **within a group** the observations must be as **similar** as possible, while observations belonging to **different groups** must be as **different** as possible.

There are two main types of clustering technique that we see every day:

K-Means Clustering and **Agglomerative Hierarchical Clustering.**

Now, we will see how to implement both of these clustering techniques using this dataset one by one. But firstly, we need to **create a Pandas DataFrame** which is used to store the content of our feature columns. Feature columns here, are the column which are considered for the execution of algorithms.

For this Dataset, we take '*Annual Income (k$)*' and '*Spending Score (1-100)*' as the two feature columns to work with.

5.1 K-MEANS CLUSTERING

This method aims at **partitioning n observations into k clusters** in which each observation belongs to the cluster with the closest average, serving as a prototype of the cluster. The K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids. K-means clustering is an extensively used technique for data cluster analysis. Furthermore, it **delivers training results quickly**.

The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

We will see how K-Means Algorithm works step wise:

- First, we determine the optimum number of clusters using a method called Elbow Method.
- Then we take for eg. K random points as centroids.
- Now, assign all points to the closest cluster centroid.
- Recompute the centroid of newly formed clusters until you predict the final result.

5.1.1. ELBOW METHOD

```
In [25]: wcss=[]  #sum of squared dist b/w each cluster
         for i in range(1,10):
             k=KMeans(n_clusters=i,init='k-means++',max_iter=300,n_init =10,random_state=0)
             k.fit(x)
             wcss.append(k.inertia_)

         plt.rcParams['figure.figsize']=(8,5)
         plt.plot(range(1,10),wcss)
         plt.title('Elbow Method to determine no. of clusters',fontsize=22)
         plt.xlabel('No. of cluster',fontsize=14)
         plt.ylabel('sum of squared dist. b/w each point' ,fontsize=14)
Out[25]: Text(0, 0.5, 'sum of squared dist. b/w each point')
```

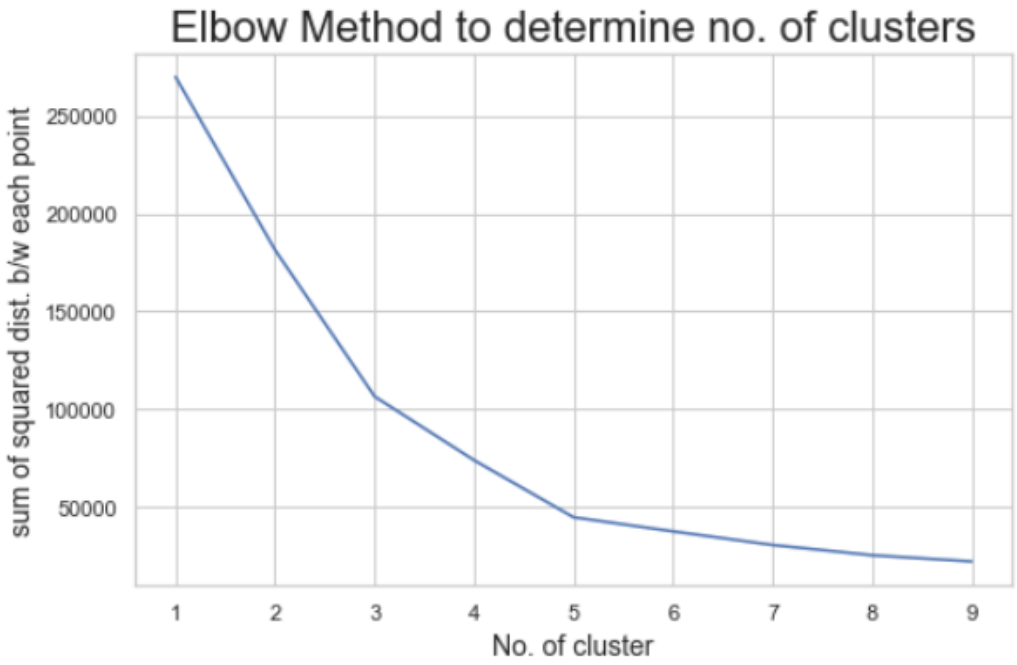


Figure 5: *Implementing Elbow Method*

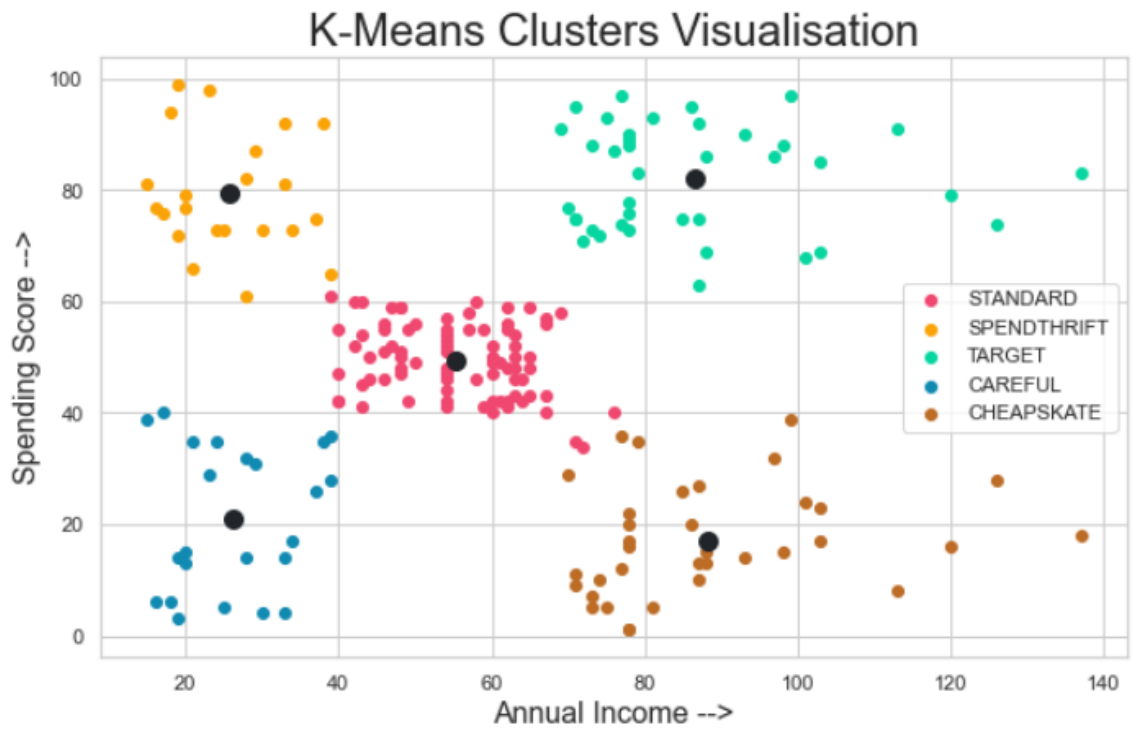
The elbow method **helps to choose the optimum value of 'k' (number of clusters)** by fitting the model with a range of values of 'k'. Here we would be using a 2-dimensional data set but the elbow method holds for any multivariate data set. We also define cost function of K-means clustering as 'Epsilon' or 'wcss' which is sum of squares of distance between data points and respective centroid of cluster to which the data point belongs. We expect the cost function to decrease with number of iterations (clusters).

In *Figure 5*, there is a sharp elbow in the graph of explained variation versus clusters: **decreasing rapidly until 5** (under-fitting region), and then **decreases slowly after 5** (over-fitting region). Therefore, the **optimum number of clusters are 5** according to the graph.

5.1.2. VISUALISING K-MEANS CLUSTERS

```
In [33]: plt.rcParams['figure.figsize']=(10,6)
         plt.scatter(x[clus==0,0],x[clus==0,1],c='#ef476f',label='STANDARD')
         plt.scatter(x[clus==1,0],x[clus==1,1],c='#faa307',label='SPENDTHRIFT')
         plt.scatter(x[clus==2,0],x[clus==2,1],c='#06d6a0',label='TARGET')
         plt.scatter(x[clus==3,0],x[clus==3,1],c='#118ab2',label='CAREFUL')
         plt.scatter(x[clus==4,0],x[clus==4,1],c='#bc6c25',label='CHEAPSKATE')
         plt.scatter(centers[:,0],centers[:,1],s=100,c='#212529')
         plt.title('K-Means Clusters Visualisation',fontsize=24)
         plt.xlabel('Annual Income -->',fontsize=16)
         plt.ylabel('Spending Score -->',fontsize=16)
         plt.legend()
         plt.show()
```

Figure 6: *K-Means Clusters Visualisation*



In *Figure 6*, we plot a Scatter graph to show the clusters and how they are separated from each other. We get five clusters and depending on their ways of spending money we have named them as Standard, Spendthrift, Target, Careful and Cheapskate.

A '**Standard**' customer has average spending score and monthly income. A '**Spendthrift**' spends more recklessly even though the monthly income is too low.

A '**Target**' customer spends more and their monthly income is also high. A '**Careful**' customer has low income and their spending score is also less. Lastly, '**Cheapskate**' is those whose monthly income is high but their spending score is low.

From this, we get an overall idea about the income and expenditure of customers belonging to different cluster. According to the clustering, we get the inference that the '*Target*' is our targeted customer group. They have high income so they are able and willing to purchase.

5.2. Agglomerative Hierarchical Clustering

The difference with the partition by *k*-means is that for hierarchical clustering, the number of classes is **not** specified in advance. Hierarchical clustering will help to determine the optimal number of clusters. This is of two types:

a. Divisive Hierarchical clustering Technique: Here, we consider all the data points as a single cluster and in each iteration, we **separate the data points** from the cluster which are **not** similar. Each data point which is separated is considered as an individual cluster.

b. *Agglomerative Hierarchical clustering Technique:* In this technique, initially each data point is considered as an individual cluster. At each iteration, the **similar clusters merge** with other clusters until one cluster or K clusters are formed. The Hierarchical clustering Technique can be visualized using a **Dendrogram.**

Among these two types of Hierarchical clustering Technique we will use *Agglomerative Hierarchical clustering Technique* as it is simpler than the Divisive Hierarchical clustering Technique and is most commonly used method which **requires no prior assumption** and uses the analysis of variance to calculate distances among the clusters.

Also, we will see how the Hierarchical clustering works step by step:

- We have made each data point a cluster.
- The two closest clusters are considered and we made them into one cluster.
- The previous step is repeated until there is only one cluster.

5.2.1. DENDROGRAM

We can use a dendrogram to visualize the history of groupings and figure out the optimal number of clusters.

```
In [35]: import scipy.cluster.hierarchy as sc
         den=sc.dendrogram(sc.linkage(x,method='ward'))
         #plt.rcParams['figure.figsize']=(16,8)

         plt.title('Dendrogram to determine no. of Clusters',fontsize=28)
         plt.xlabel('Customers', fontsize=20)
         plt.ylabel('Distance', fontsize=20)

Out[35]: Text(0, 0.5, 'Distance')
```

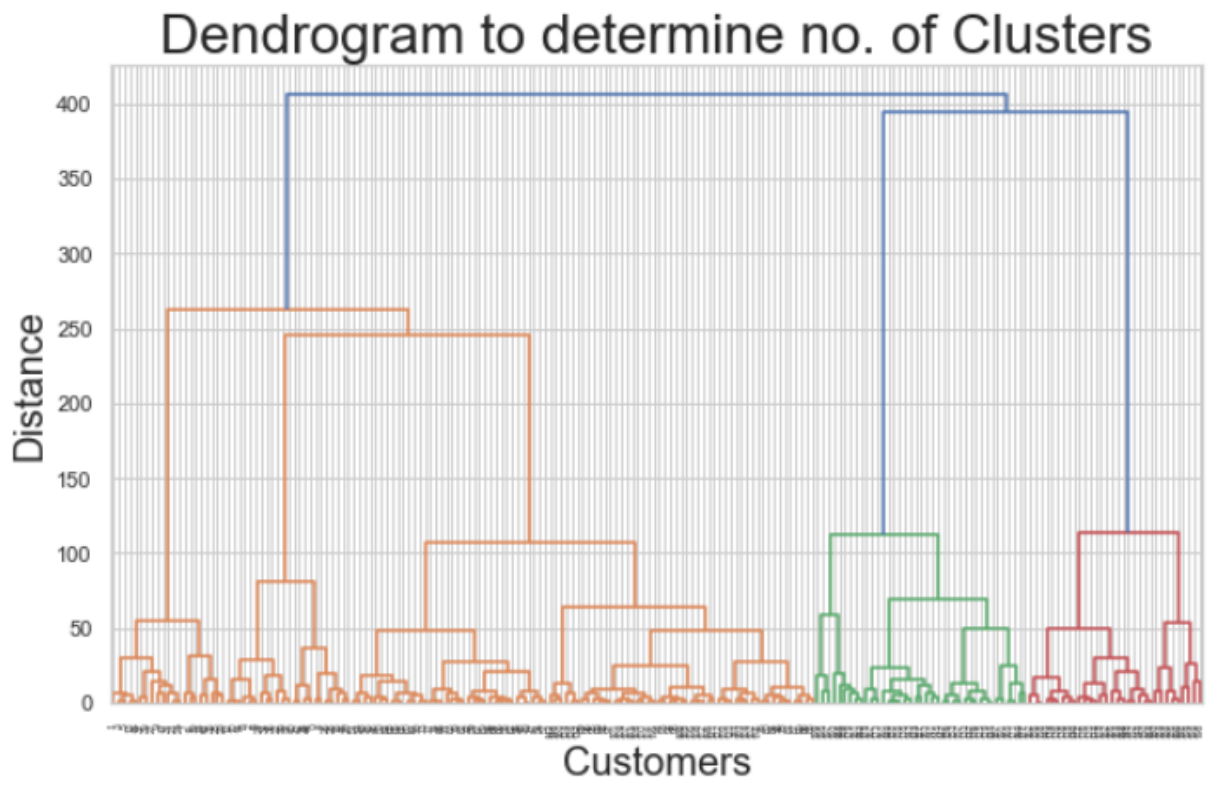


Figure 7: *A dendrogram to determine no. of clusters.*

While working on the dendrogram, we have determined the largest vertical distance that does not intersect any of the other clusters. A horizontal line has been drawn at both the extremities. The optimal number of clusters is equal to the number of vertical lines going through the horizontal line.

Looking at the dendrogram, the highest vertical distance that does not intersect with any clusters will be the one with a horizontal line through distance=150. The **5 vertical lines cross the threshold** given in the dendrogram and the **optimal number of clusters is 5.**

5.1.2. VISUALISING HIERARCHICAL CLUSTERS

```
In [41]: plt.rcParams['figure.figsize']=(10,6)
plt.scatter(x[hclus==0,0],x[hclus==0,1],c='#ef476f',label='CHEAPSKATE')
plt.scatter(x[hclus==1,0],x[hclus==1,1],c='#faa307',label='STANDARD')
plt.scatter(x[hclus==2,0],x[hclus==2,1],c='#06d6a0',label='TARGET')
plt.scatter(x[hclus==3,0],x[hclus==3,1],c='#118ab2',label='SPENDTHRIFT')
plt.scatter(x[hclus==4,0],x[hclus==4,1],c='#bc6c25',label='CAREFUL')
plt.scatter(centers[:,0],centers[:,1],s=100,c='#212529')
plt.title('Hierarchical Clustering Visualisation',fontsize=24)
plt.xlabel('Annual Income -->',fontsize=16)
plt.ylabel('Spending Score -->',fontsize=16)
plt.legend()
plt.show()
```

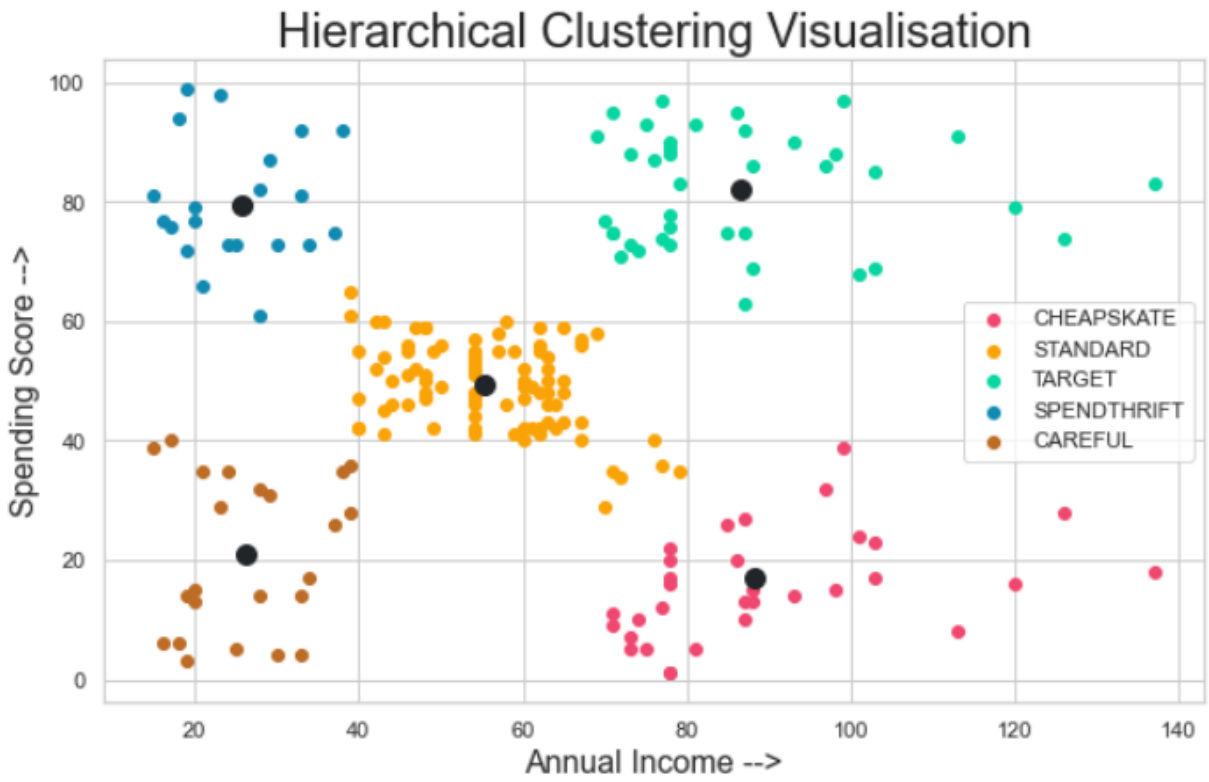


Figure 7: *Hierarchical Clusters Visualisation*

As in *Figure 7*, we have made a scattered graph to visualise the 5 clusters that we got from the Hierarchical clustering. There are 5 clusters in the above graph that give the concept about the customers' annual income and the behaviour of how they are spending money for which we have taken spending score and annual income from the dataset. **The clusters thus produced are similar to that of the K-Means Clusters** and thus we will also describe them as **Cheapskate, Standard, Target, Spendthrift and Careful.**

6. RESULTS

This Clustering Analysis gives us a very clear insight about the different segments of the customers in the Mall. There are clearly Five segments of Customers namely **Standard, Spendthrift, Target, Careful and Cheapskate** based on their Annual Income and Spending Score which are reportedly the best factors/attributes to determine the segments of a customer in a Mall.

7. DISCUSSION

In this Dataset, **Annual Income and Spending Score are reportedly the best factors** to determine the segments of a customer in a mall. But on the other hand, **we can also take Age and Spending Score to determine segments** of customers in a mall and marketing department can work on that too.

We clearly have 5 segments of customers:

TARGET **-** The cluster with **high Monthly income and high Spending score** can be called as Target group. This group is more important than others because these people are generally the heavy investors and working on this group of customers will surely rake up your profit percentages. But, as per *Figure 3*, we can see they are **lower in numbers**. So, if we make improvements for this group it might attract more people like this which will mean better profit percentages.

CHEAPSKATE **-** The cluster with **high income but low spending score** can be called as Cheapskate group. These types of customers **spend their money very carefully** and tend to managing their money before spending them.

SPENDTHRIFT **-** The cluster with **low income and high spending score** can be called as Spendthrift group. These are people who **tend to purchase what they feel is good**, before checking on their own monetary condition.

CAREFUL **–** The cluster with **low income and low spending score** can be called as Careful group. They **tend to purchase only what they need** and keep a close look on their monetary status at all times. They seem to be very sensible with the way they spend their money.

STANDARD **–** The cluster with **average income as well as average spending score** can be called as Standard group. This group is very important and should be taken

care of them as in *figure 3*, we can see people of this group as **most likely to visit the mall**. This group **attracts the highest population of people** as much as almost five times more than the Target group. Thus, it's always profitable to see to their needs as this group is another group which derives high profit margin.

Thus, we can say, the cluster analysis was successful as the clusters are well defined for the marketing team to work on. Our main objective was to thus understand clustering and its implementation in real life problems.

8. CONCLUSION

In this case study with the help of the dataset, we got 5 different clusters and understood cluster analysis using different clustering approaches. With this study case that we have worked on and from the methods, result and discussion we understood how clustering works.

The clustering technique can be very handy when it comes to unlabelled data. **Since most of the data in the real-world is unlabelled and annotating the data has higher costs, clustering techniques can be used to label unlabelled data.**

We will see how both the clustering methods i.e., K-Means clustering and Hierarchical clustering helped in customer segmentation analysis.

8.1. K-Means Clustering

In the K-Means clustering, we easily get to **find the number of clusters by applying the elbow method**. K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of K means is to **group data points into distinct non-overlapping subgroups**. One of the major applications of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.

8.2. Hierarchical Clustering

We have used mainly Agglomerative Hierarchical clustering Technique for finding the optimal number of clusters. Therefore, to find the number of clusters **dendrogram is created to breakdown the whole scattered plot into clusters so**

that we can define each cluster. Hierarchical clustering is a highly useful unsupervised clustering algorithm that can be utilized in the business.

Even though both the clustering has different methods and approaches to find the number of clusters, they **essentially convey the same information**. Through clustering, we also can understand the behaviour of the customers which may help a company to run its business accordingly.

9. BIBLIOGRAPHY

- Dataset from Kaggle- https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python
- Wikipedia- https://en.wikipedia.org/
- Towards data science- https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
- Scikit-learn guide- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- Analytics Vidhya- https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/
- Towards data science- https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec
- Scikit-learn guide- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
- https://stackabuse.com/hierarchical-clustering-with-python-and-scikit-learn/
- Full Code link:
  In .html format- https://drive.google.com/file/d/1dzU5aeQ-X7ikVrQ4PHaOxJP_YA5MJ-Wp/view?usp=sharing

  In .ipynb format- https://drive.google.com/file/d/1zzqI_KttO8p1OMxPvO8vfbBHZU8-uoJ_/view?usp=sharing

#Group Assignment by- Group A

- **ARITRA MAZUMDAR**
  MCA 1ST YEAR (*2ND SEM*)
  REG NO.: 20352207

- **ANUSHREE DATTA**
  MCA 1ST YEAR(*2ND SEM*)
  REG NO.: 20352205

- **MEENU M BIJU**
  MCA 1ST YEAR (*2ND SEM*)
  REG NO.: 20352218

- **SOUVIK DUTTA**
  MCA 1ST YEAR (*2ND SEM*)
  REG NO.: 20352236