**ABSTRACT**

This project aims at doing a market basket analysis using hierarchical clustering. In this project I have grouped customers based on their income and how much they spend on the products.

5 clusters were plotted.

The group of customers that the business would have to focus on more would be in cluster 1. These customers buy less in spite of having relatively high annual income. Thus, the company would need to improve on certain products or customer services, so that those customers are satisfied.

**OBJECTIVE**

The project aims at exploring how hierarchical clustering, a type of unsupervised learning, can be used to give insights to businesses on their performance.

**INTRODUCTION**

Hierarchical clustering is of 2 types – agglomerative and divisive. Agglomerative clustering is a bottom – up approach of making clusters; at first considers each data point as a separate cluster and iteratively merges the closest pairs of clusters until all the data points are part of a single cluster, whereas in divisive, we consider all datapoints as a single cluster and divide it until each point is a single cluster.

In this project, agglomerative hierarchical clustering has been used.

Unlike K-means clustering, in hierarchical clustering we don’t need to specify the number of clusters it has to be separated into in the beginning.

We decide the number of clusters required from the dendrogram, which is plotted based on the Euclidean distance between clusters, which can be found in a number of ways; the distance between centroids, or the closest points, etc.

I have used the ward method, which focusses on variance minimisation.

**METHODOLOGY**

Firstly, the dataset was imported. Since focus was on the annual income and spending score, the remaining columns with irrelevant data were deleted.

Using the library scipy.cluster.hierarchy, the dendrogram was plotted, based on the Euclidian distance between the points. This was done to find out the best number of clusters for optimal results, using the ward method, which checks for points that can be clustered with minimum increase in variance within the cluster. From this diagram it was clear that it was best to make 5 clusters.

Keeping these insights in mind, 5 clusters were made.

**CODE**

#import libraries

import numpy as np

import matplotlib.pyplot as plt

import pandas as pd

# import dataset

dataset = pd.read\_csv("Mall\_Customers.csv")

dataset = dataset.drop(['CustomerID', 'Genre', 'Age'], axis=1)

x = dataset.iloc[:,:].values

# Dendogram to decide optimal number of clusters

import scipy.cluster.hierarchy as sch

dendrogram = sch.dendrogram(sch.linkage(x, method = 'ward'))

plt.title("DENDROGRAM")

plt.xlabel("Customers")

plt.ylabel("Euclidean distance")

plt.show()

# Train the model

from sklearn.cluster import AgglomerativeClustering

clustering = AgglomerativeClustering(n\_clusters = 5)

y\_hc = clustering.fit\_predict(x)

# Visualising clusters

plt.scatter(x[y\_hc == 0, 0], x[y\_hc == 0, 1], c = "red", label = 'Cluster 1')

plt.scatter(x[y\_hc == 1, 0], x[y\_hc == 1, 1], c = "green", label = 'Cluster 2')

plt.scatter(x[y\_hc == 2, 0], x[y\_hc == 2, 1], c = "teal", label = 'Cluster 3')

plt.scatter(x[y\_hc == 3, 0], x[y\_hc == 3, 1], c = "violet", label = 'Cluster 4')

plt.scatter(x[y\_hc == 4, 0], x[y\_hc == 4, 1], c = "brown", label = 'Cluster 5')

plt.title("CLUSTER OF CUSTOMERS")

plt.xlabel("Annual income")

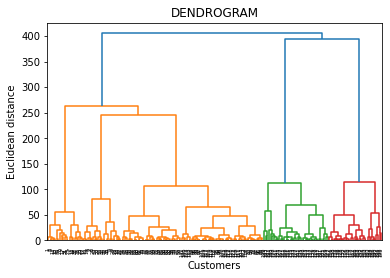
plt.ylabel("Spending score")

plt.legend()

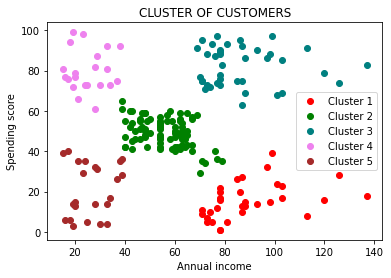
plt.show()

**CONCLUSION**

Based on the dendrogram, I decided that 5 clusters would show an ideal analysis of the data. The dendrogram is shown on the next page.



The clusters plotted were as follows:



This plot can be interpreted as follows:

* Customers in Cluster 1 have a high annual income compared to the others, but spend low on products. This may be because they are not satisfied with the products or services, so care must be taken to satisfy them.
* People in Cluster 2 have a moderate income and spend reasonably on the products.
* Customers in cluster 3 have high annual income and spend a lot, which means they are satisfied. Businesses should ensure that these customers remain satisfied in the future as well.
* People in cluster 4 have relatively low annual income, but spend a lot on the products, which means that they are very satisfied and their satisfaction should be ensured in the future as well.
* People in the 5th Cluster have low annual income and hence spend less on their products.