DATA ANALYSIS USING K-MEANS CLUSTERING

IBM-312 Assignment

Dawar Husain (18115030) Inzamam (18115042) Keshav Dixit (18115050)
Kshitij Bithel (18115053) Pranjal Mangal (18115070)

ABSTRACT

Malls are often indulged in the race to increase their customers and make profits. To this end, they often collect customer data and apply various machine learning techniques to improve their sales by targeting the right audience.

Clustering is the process of making a group of abstract objects into classes of similar objects. Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.

We use the mall customers dataset and apply k-means clustering to partition the data into clusters. After forming clusters, we can target the customers belonging to a certain cluster separately.

INTRODUCTION

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data. The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity.

We are using K-means clustering to identify different clusters that share similar characteristics in customers' dataset. This will help us develop customized marketing campaigns targeting different groups of customers and allow us to choose appropriate distribution strategy among those groups.

The mall customers dataset contains the following features:

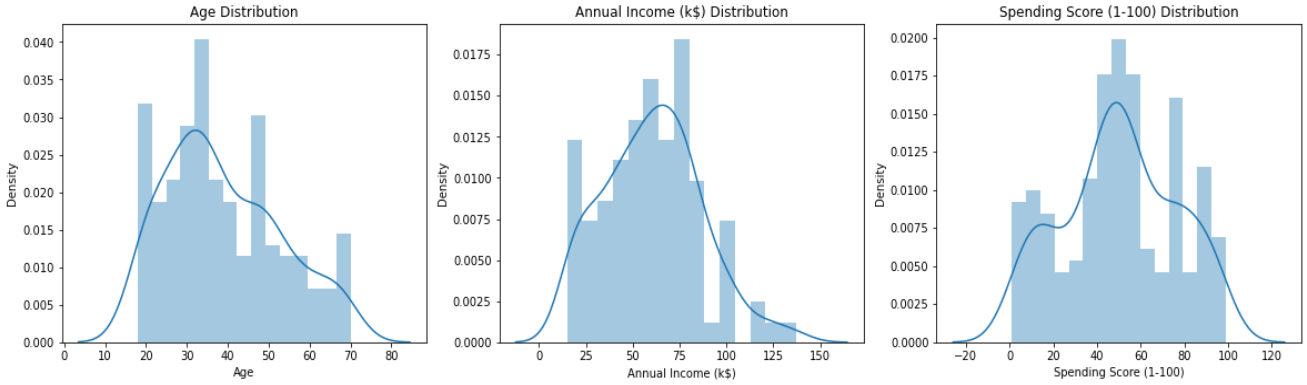
1. Customer id
2. Gender (male or female)
3. Age
4. Annual income (expressed in k$)
5. Spending score (metric measure customer expenditure, lies from 1 to 100)

DATA VISUALISATION

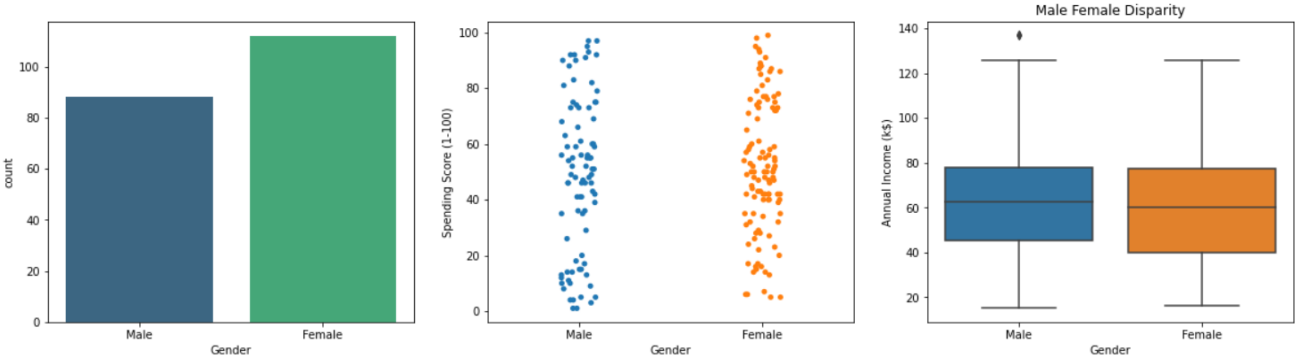
This data contains 200 customers.For each Customer, we have Age, Gender Annual income expressed in k$, spending score (1-100).

	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

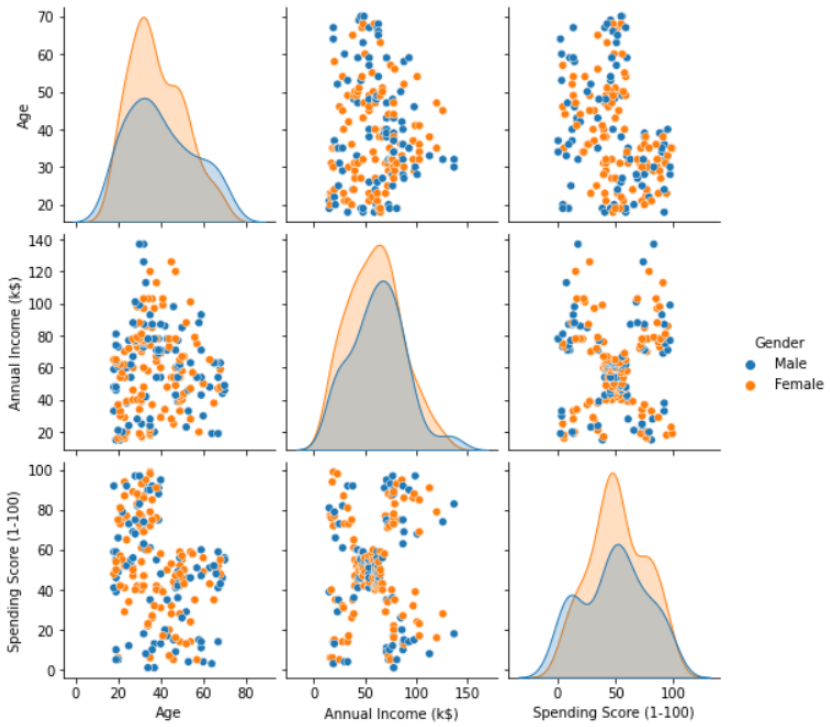
There are some basic observations which we get from data and is pictorially represented such as:
We can check the amount of customers coming in an age group, also we can see the amount of people who have the Annual income in the corresponding range. Another thing which can be analysed is the amount of people which are willing to spend in a particular range of amounts.



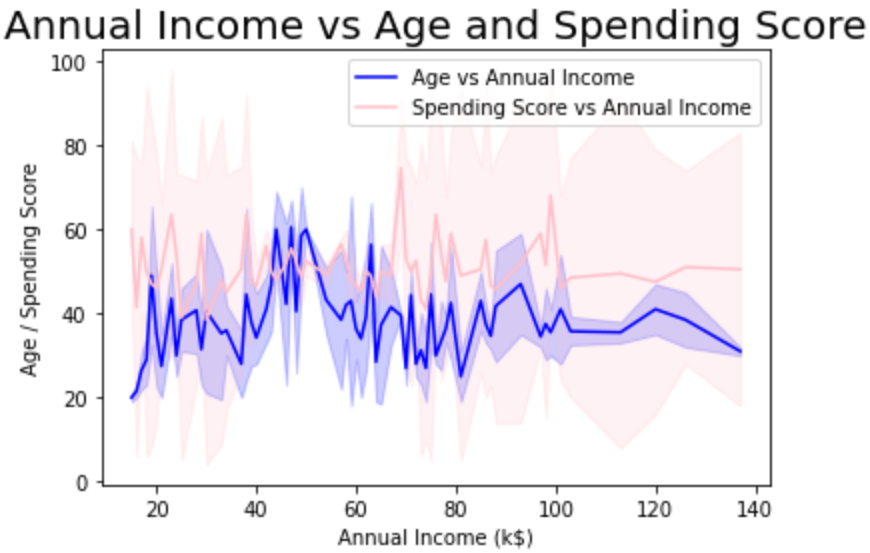
Further this data can also be used to check how many people who come to the mall are male and how much are female. We can also see Gender wise distribution of the spending score in the data.



Male Female disparity is not present here as well. i.e. the mall is situated in a locality where males and females are not biased.
There is not much correlation in the scatterplots which means our variables are independent.



There is a very low correlation in this plot where people in certain annual income intervals have a high standard deviation in spending score.



MODEL EXPLANATION

K-means

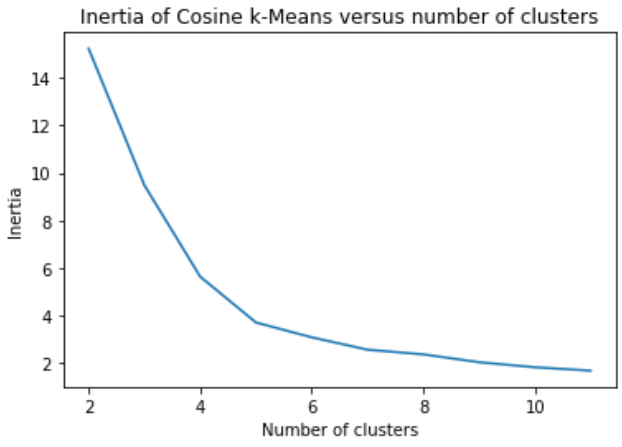
In this Model we use the K-means clustering algorithm. The basic algorithm is as follows:

1. First we specify the number of clusters K.
2. Then we initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement
3. We keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.

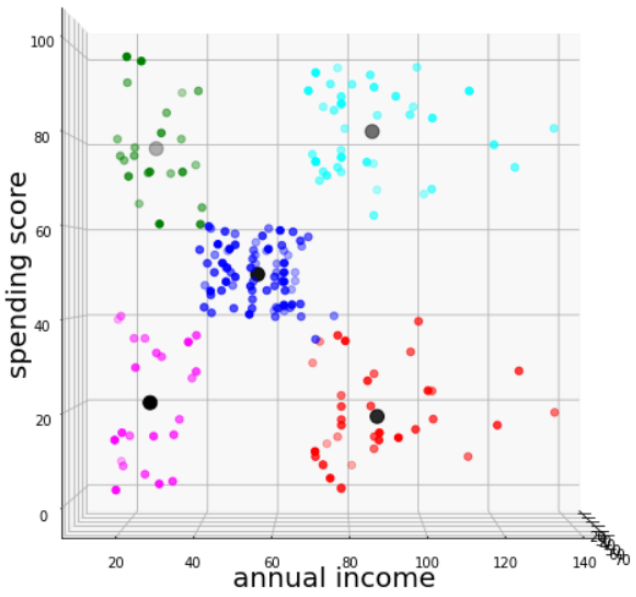
In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set. The method consists of plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

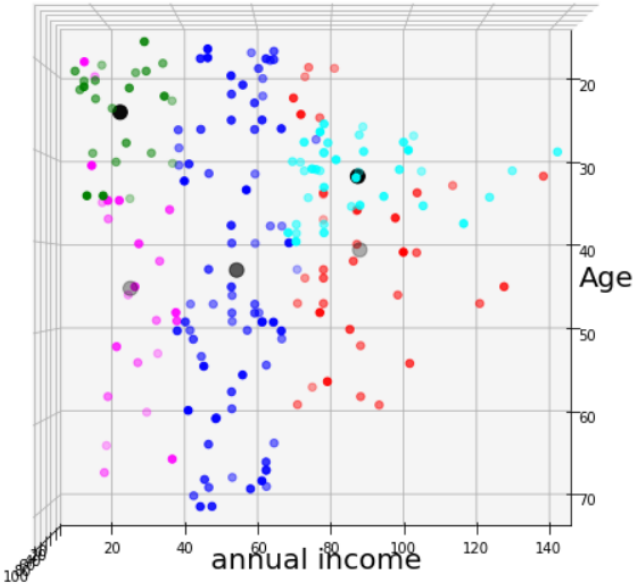
First we have Normalized all three Parameters of the data set we have. Then we vary the number of clusters (K) from 2 to 11 and for each K we fit the data into the number of clusters. Using the algo-method, we calculate the inertia for each K. Then we plot inertia vs number of clusters and find the location of the elbow of the curve to give the appropriate number of clusters.

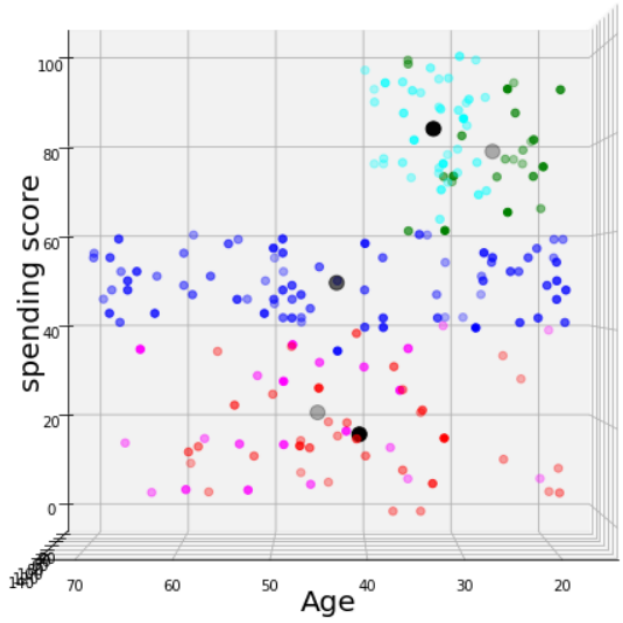
Here we used 5 and 6 numbers of clusters for which we plotted the data but we found that for the number of clusters equal to 6 there was too much overlapping of clusters. So we chose to go with 5 clusters.

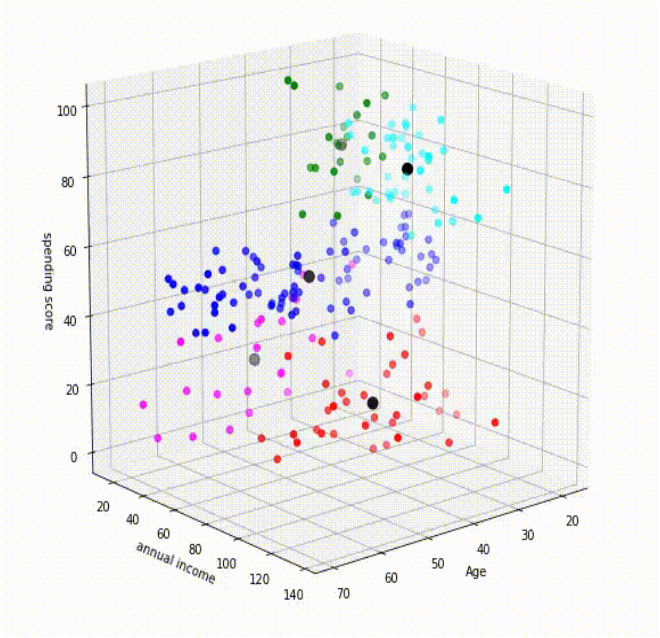


Then we performed K-means clustering using K = 5. And after fitting the data we received these results







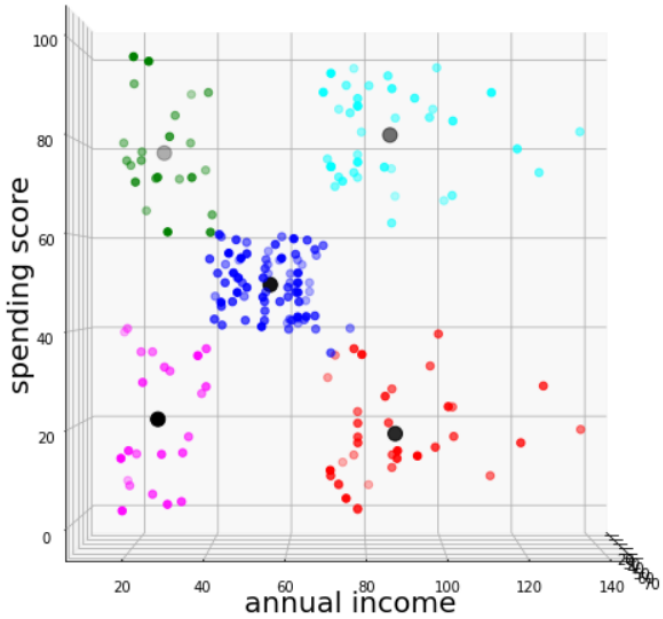


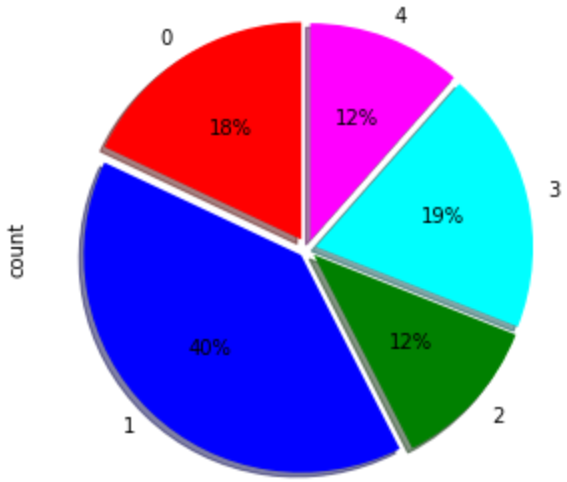
PCA

As we found that the dependency of the data on tha age was not very profound so we used Principal Component Analysis for dimensionality reduction. We reduced the feature list from three to two. Then we performed K-Means clustering on the obtained Feature set and found that the clusters formed are quite similar to the one obtained from spending score and annual income.

CONCLUSION

Using K-means clustering the mall customers in the dataset can be grouped into 5 different cluster groups based on their spending score and annual income.





Green Cluster

It represents the set which has less income and high spending scores. It comprises quite less percentage(12%) of people. This means that there are not many people around the mall who have a low income and still try to live a luxurious life. So following the work ethics we should not target them and should not tempt them to spend more.

Blue Cluster

This set represents the people who have average annual income and average Spending Score. These people comprise the highest percentage of the set(40%). This means that the mall is situated in a locality and not very famous to attract rich customers. Hence, products of particular interest to this section should be kept to generate a regular income from these people. We can encourage these people to spend more so that they try to improve their annual income.

Cyan Cluster

It represents the set which has high income and high spending scores. Since they are spending the most and have high income we should target them more. Proper luxurious spaces can be designed with extra comfort and talented executives for this particular section of people to get most out of their pockets since they comprise the second highest in the count(19%).

Magenta Cluster

This cluster consists of people with relatively low expenditure and low income. Visitors in this section are limited(12%) which means that the mall is a bit expensive and not affordable at all.

Red Cluster

It represents the set which has high income but low spending scores. The super rich section of the society having the lowest average spending score should be a bit astonishing to the mall administration. The reasons may be poor appropriate sized parking spaces, hygiene or security. Despite having high income they are spending less so we should take feedback from these customers and bring their recommended products in malls and provide attractive discounts.

REFERENCES

[1] https://www.udemy.com/course/machinelearning/
[2] https://matplotlib.org/stable/gallery/mplot3d/scatter3d.html
[3] https://matplotlib.org/
[4] https://pandas.pydata.org/docs/
[5] https://seaborn.pydata.org/introduction.html
[6] https://scikit-learn.org/stable/modules/clustering.html
[7] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.normalize.html
[8]https://en.wikipedia.org/wiki/Elbow_method_(clustering)#:~:text=In%20cluster%20analysis%2C%20the%20elbow,number%20of%20clusters%20to%20use.