**ABSTRACT**

Malls or shopping complexes are often indulged in the race to increase their customers and hence making huge profits. To achieve this task machine learning is being applied by many stores already.  
It is amazing to realize the fact that how machine learning can aid in such ambitions. The shopping complexes make use of their customers’ data and develop ML models to target the right ones. This not only increases sales but also makes the complexes task much easier.

Being a Mall owner, One always worry about how to increase their sells, on which parameter they should focus and which age group they should target.

There is a problem to give Offers and Deals to the specific customers. How they should target to the customers who are frequent to mall, having high income and recognizing their proper choice. By knowing above problems answers owner of the mall can increase their sells and serve better to the customer.

We have datasets of a Mall and we are going to predict how to increase its sells. Here we are using Unsupervised Machine Learning concept. We have some information about customers like Gender, Age, Annual Income and Spending Score. Based on these parameters we are going to apply Machine Learning Algorithm (Unsupervised Machine Learning), will do visualization and Analysis of data. After that we will find clusters in the dataset and conclude our prediction.

**TABLE OF CONTENTS**

* **Introduction**
* **Proposed method with Architecture**
* **Methodology**
* **Implementation**
* **Conclusion**

**INTRODUCTION**

There is a big mall is a specific city that has information of its customers through membership card subscription. Some of the information provided by customers includes gender, age, and estimated annual income.

The membership card is often used by subscribers to buy products at discounted rates. This means there is a purchase history of every member. With the purchase history, the mall generated a spending score of 1 – 100 for every member. This criteria for spending takes into consideration income, number of time per week a member comes to the mall and also the amount of money spend on every visit. The closer the spending score is to 1, the less the customer spends and the closer the spending score is to 100, the more the member spends.

Based on spending score and annual income of the subscribers we are going to apply Unsupervised machine learning concept and will use k-means clustering algorithm to make clusters which will group up members having approximately same spending score and Annual income. So that mall can recognize the proper customer and give them suitable deals and offers. This results increment of the Mall sells.

**ARCHITECTURE**

**Tools version**

Python : 3.7.6

seaborn : 0.10.0

matplotlib : 3.1.3

numpy : 1.18.1

pandas : 1.0.1

plotly : 4.10.0

sklearn : 0.22.1

jupyter\_notebook

The above tools and their respected versions have used in this project.

Clustering techniques reveal internally homogeneous and externally heterogeneous groups. Customers vary in terms of behaviour, needs, wants and characteristics and the main goal of clustering techniques is to identify different customer types and segment the customer base into clusters of similar profiles so that the process of target marketing can be executed more efficiently. Both, hierarchical and non-hierarchical clustering algorithms are widely used in customer segmentation, most prominent among them being K-Means and Agglomerative Hierarchical Clustering. However, here we will be focusing only on K-Means and will do rest of work on it.

A few data of a mall has been used in this project. Link is given below

<https://github.com/lavkush9051/Customer_segmentation/blob/master/Mall_Customers.csv>

**METHODOLOGY**

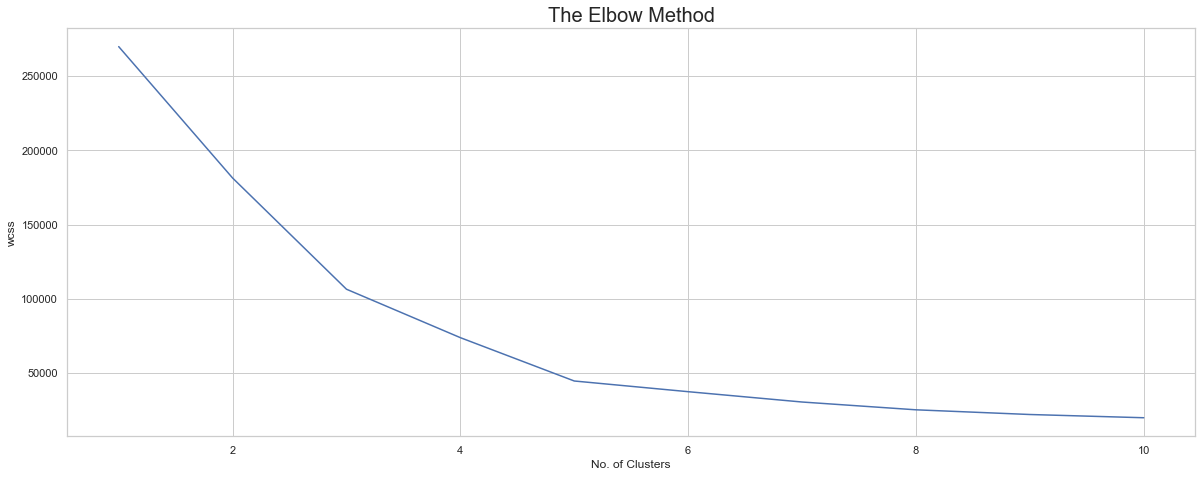
Throughout this project Jupyter notebook has been used. As our main focus on K-Means clustering algorithm we will import it from sklearn. K-Means is one of the most widely used clustering algorithms, and is simple and efficient. The aim of K-Means algorithm is to divide M points in N dimensions into K clusters (assume k centroids) fixed a priori. These centroids should be placed in a wise fashion so that the results are optimal which otherwise can differ if locations of the centroids change. So, they should be placed as far as possible from each other. Each data point is then taken and associated with the nearest centroid until no data points are pending. This way an early grouping is done and at this point, k new centroids have to be recalculated as these will be the centers of the clusters formed earlier. After having calculated these centroids, the data points are then allocated to the clusters to the nearest centroids. In this iteration, the centroids change their position stepwise until no further modifications have to be done and the location of the centroids remain intact.

The K-Means algorithm is relatively simple. The „K‟ cluster points, which will be the centroids, are placed in the space among the data points. Each data point is assigned to the centroid for which the distance is the least. After each data object has been assigned, centroids of the new groups are re-calculated. The above two steps are repeated until the movement of the centroid ceases. This means that the objective function of having the least squared error is completed and it cannot be improved further. Hence, we get K clusters as a result.

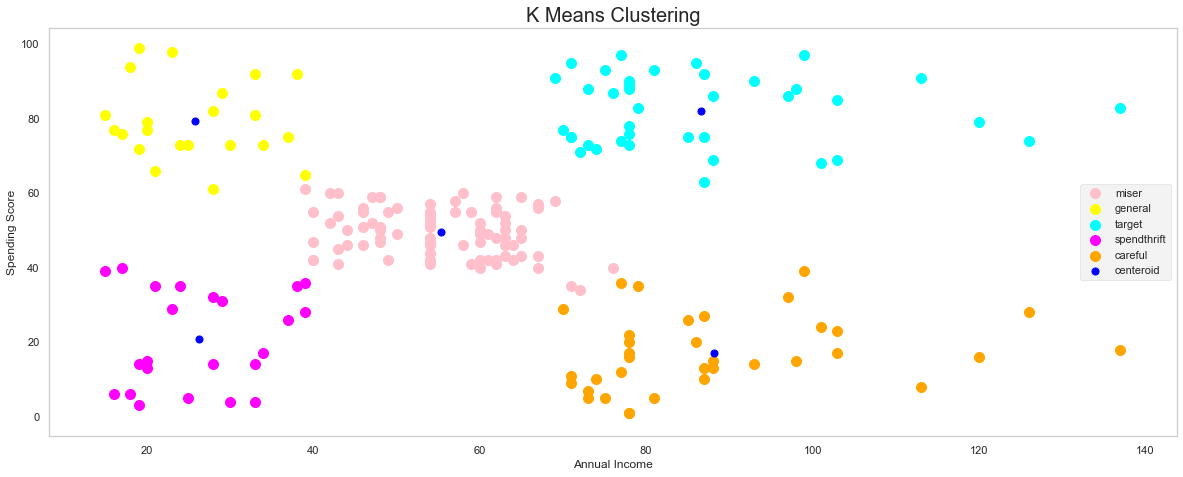
K-Means algorithm aims at minimizing an objective function, which here, is the squared-error. It is an indicator of the distance of the data points from their respective cluster centers. The process in this algorithm always terminates but the relevance or the optimal configuration cannot be guaranteed even when the condition on the objective function is met. The algorithm is also sensitive to the selection of the initial random cluster centers. That is why it runs multiple times to reduce this effect but for a large number of data points, it tends to perform very well even though it is iterative. Here we apply K-Means Clustering algorithm on a relatively small dataset and the results are depicted.

The dataset is based on customer information for a mall and has 5 attributes named Customer\_id, Gender, Age, Annual\_Income and Spending\_Score. It consists of 200 observations, each of which refers to a unique customer and the spending scores are decided and calculated by the company, based on their spending habits. Hence annual income and spending scores are the key indicators in this data. The age attribute of the customers can also be experimented with, to analyze which age group works best for a business.

Any business would always keep the monetary values of any customer as top indicators. Thus, the annual income and spending scores of the customers will be best suited for clustering. As K-Means algorithm requires the number of clusters as input, below we will use the elbow method to get the optimal number of clusters which can be formed. It works on the principal that after a certain number of „K‟ clusters, the difference in SSE (Sum of Squared Errors) starts to decrease and diminishes gradually. Here, the WCSS(Within-Cluster-Sum-of-Squared-errors) metric is used as an indicator of the same. Hence, the „K‟ value, specifies the number of clusters. In Figure 1, it can be observed that an elbow point occurs at K=5. After K=5, the difference in WCSS is not so visible. Hence, we will choose to have 5 clusters and provide the same as input to the K-Means algorithm.



Finding the optimal number of clusters using Elbow method for KMeans Clustering

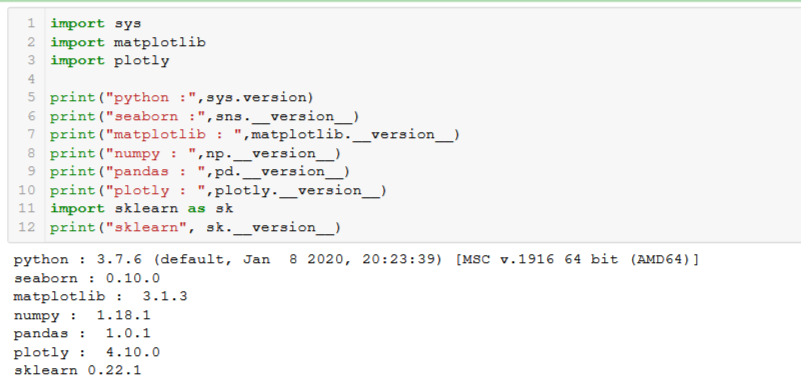


**IMPLEMENTATION**

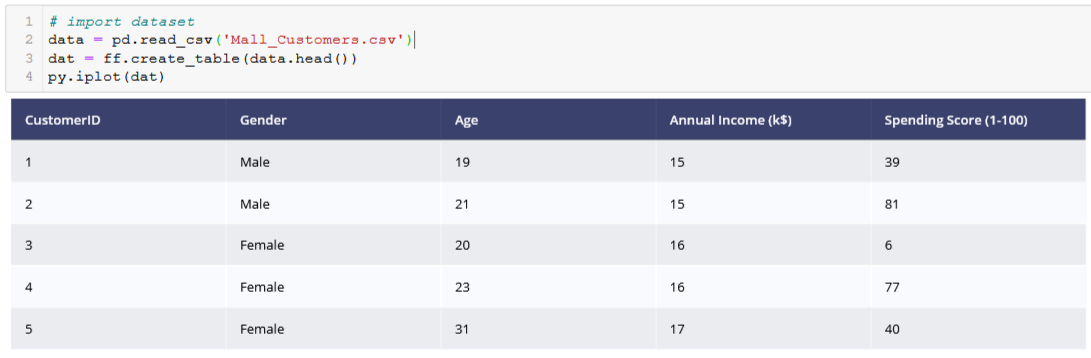
Import all the tools which was mentioned in the Architecture section



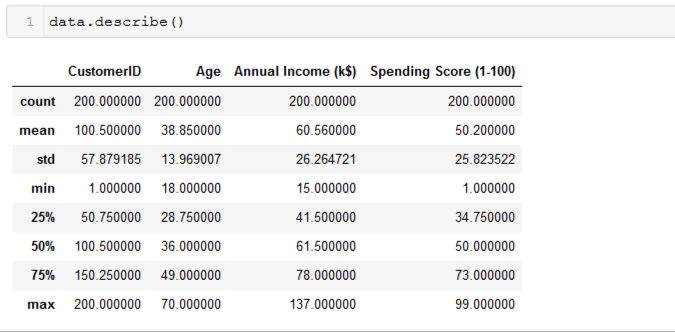
Version



Lets see how our data looks like

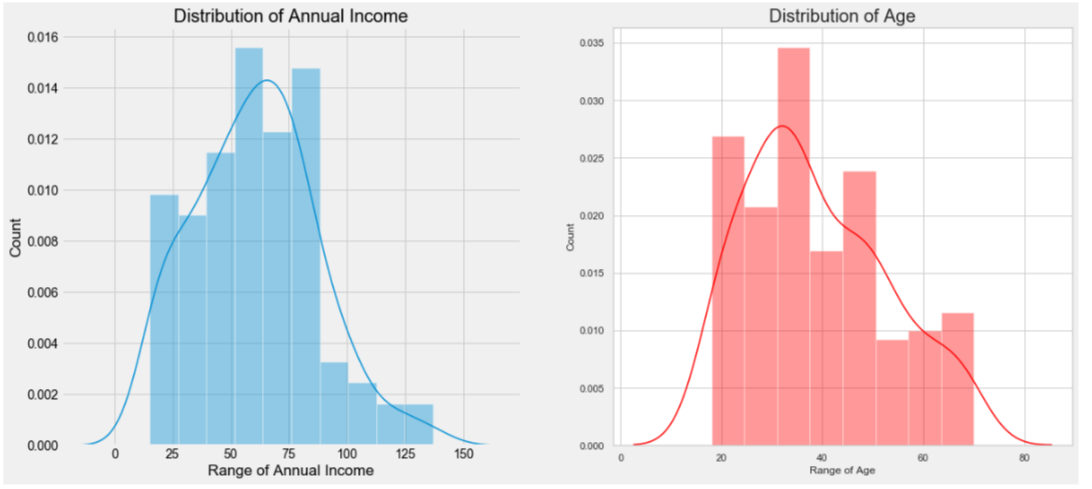
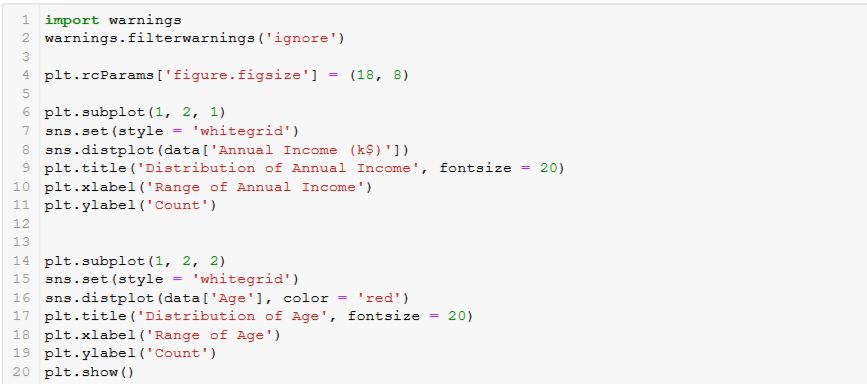


Mean, median, max, min



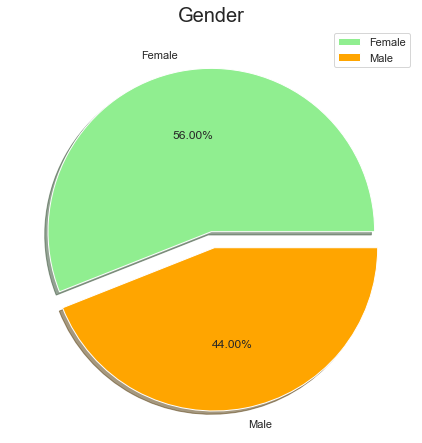
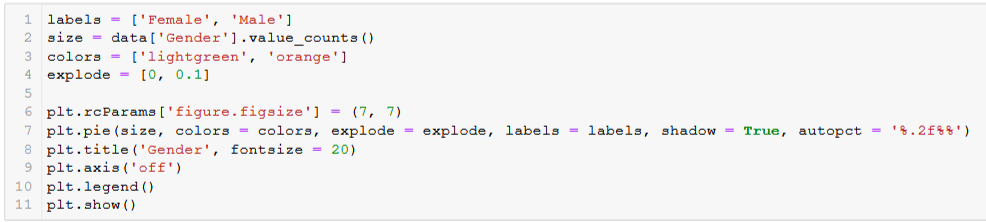
This table clearly shows that there is 200 rows and mean, median of the dataset column.

Here we have ploted distribution plot of Annual income and age.



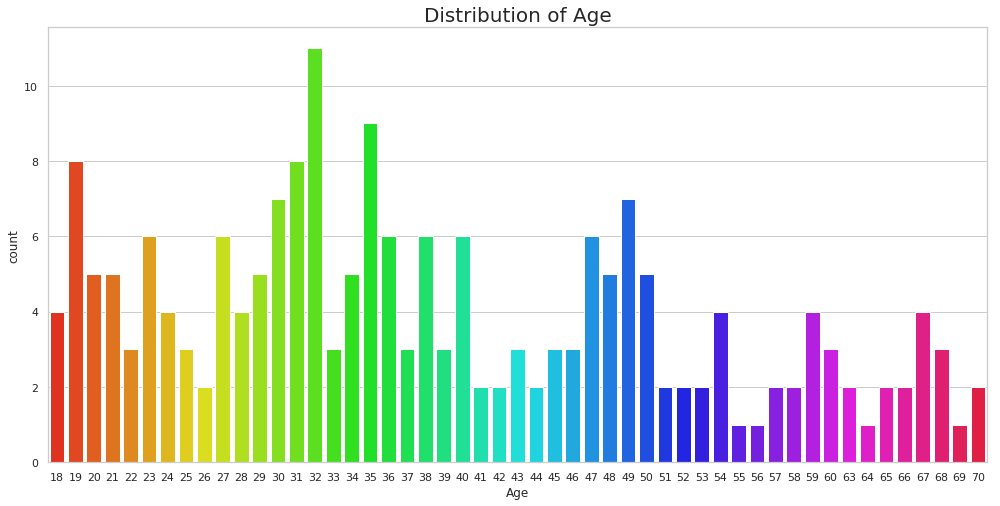
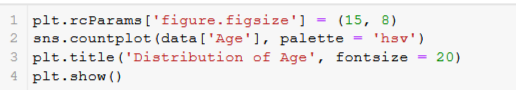
Here, In the above Plots we can see the Distribution pattern of Annual Income and Age, By looking at the plots, we can infer one thing that There are few people who earn more than 100 US Dollars. Most of the people have an earning of around 50-75 US Dollars. Also, we can say that the least Income is around 20 US Dollars. Taking inferences about the Customers. The most regular customers for the Mall has age around 30-35 years of age. Whereas the the senior citizens age group is the least frequent visitor in the Mall. Youngsters are lesser in umber as compared to the Middle aged people.

Gender distribution



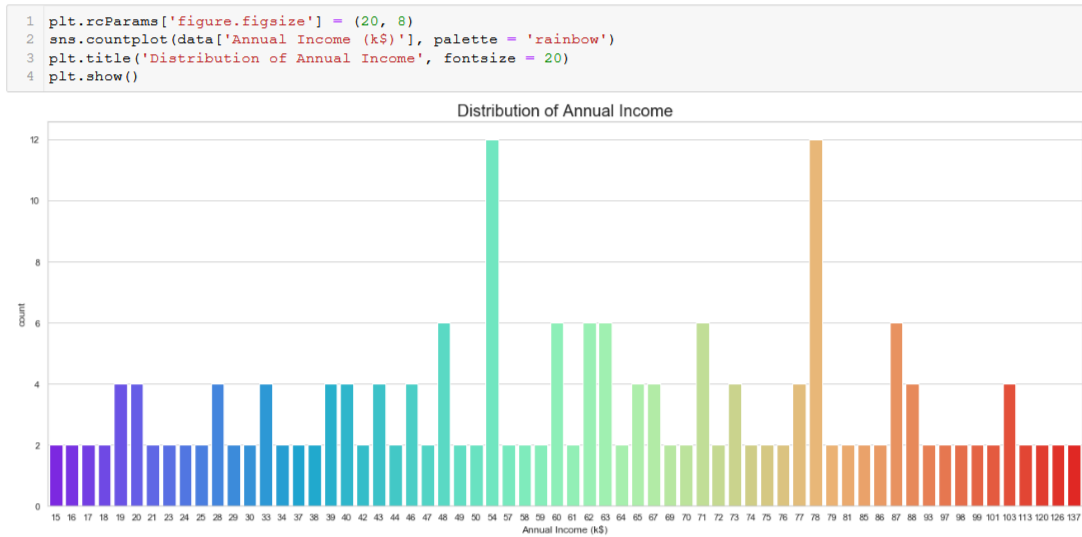
By looking at the above pie chart which explains about the distribution of Gender in the Mall Interestingly, The Females are in the lead with a share of 56% whereas the Males have a share of 44%, that's a huge gap specially when the population of Males is comparatively higher than Females.

This Graph shows a more Interactive Chart about the distribution of each Age Group in the Mall for more clearly about the Visitor's Age Group in the Mall.

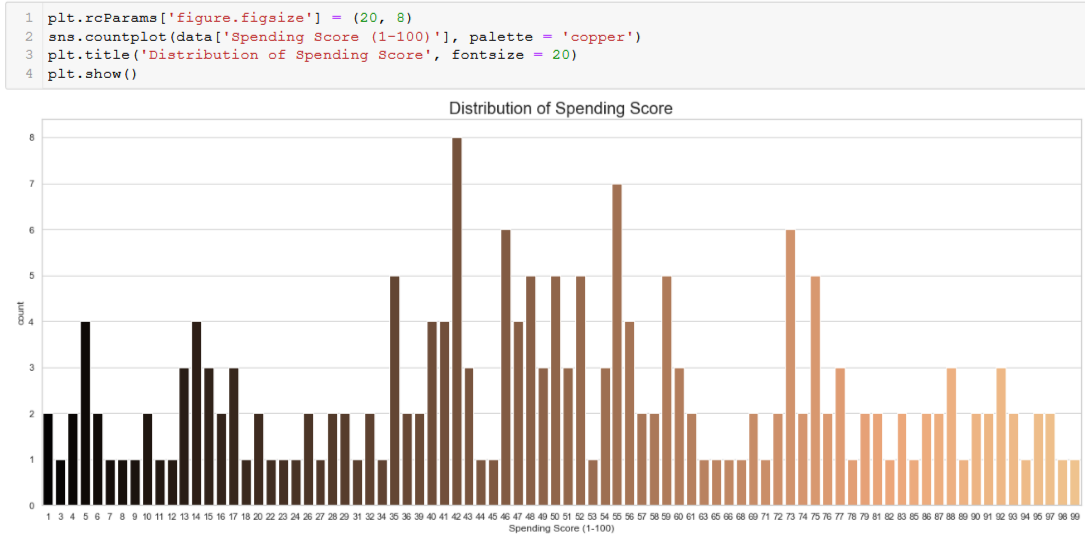


By looking at the above graph-, It can be seen that the Ages from 27 to 39 are very much frequent but there is no clear pattern, we can only find some group wise patterns such as the the older age groups are lesser frequent in comparison. Interesting Fact, There are equal no. of Visitors in the Mall for the Agee 18 and 67. People of Age 55, 56, 69, 64 are very less frequent in the Malls. People at Age 32 are the Most Frequent Visitors in the Mall.

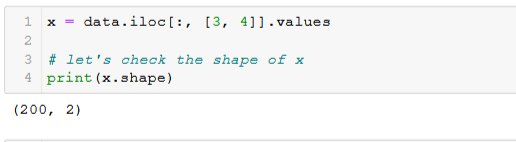
Again, This is also a chart to better explain the Distribution of Each Income level, Interesting there are customers in the mall with a very much comparable frequency with their Annual Income ranging from 15 US Dollars to 137K US Dollars. There are more Customers in the Mall who have their Annual Income as 54k US Dollars or 78 US Dollars.



This is the Most Important Chart in the perspective of Mall, as It is very Important to have some intuition and idea about the Spending Score of the Customers Visiting the Mall. On a general level, we may conclude that most of the Customers have their Spending Score in the range of 35-60. Interesting there are customers having 1 spending score also, and 99 Spending score also, Which shows that the mall caters to the variety of Customers with Varying needs and requirements available in the Mall.

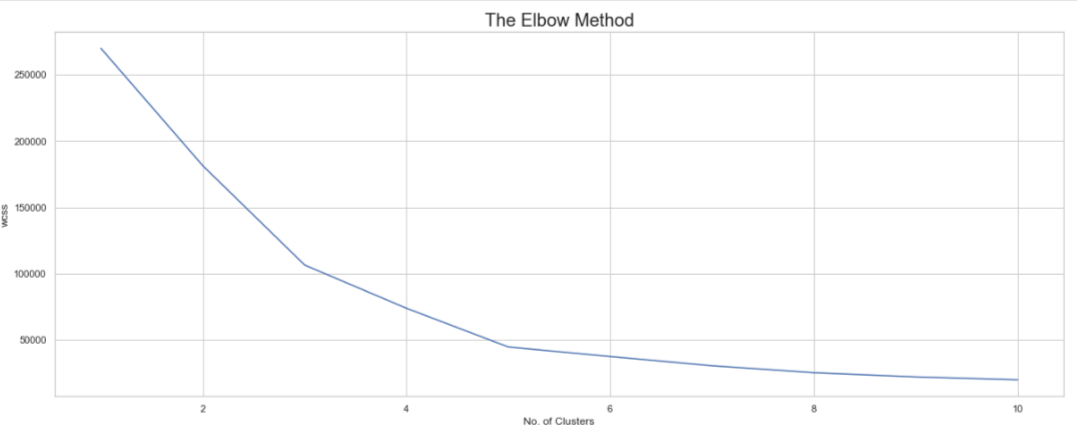
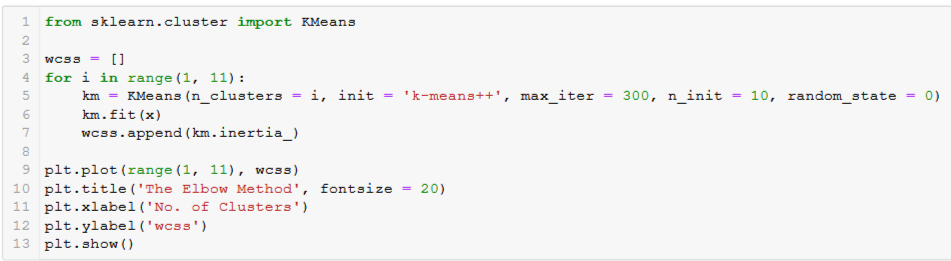


Clustering Analysis :-

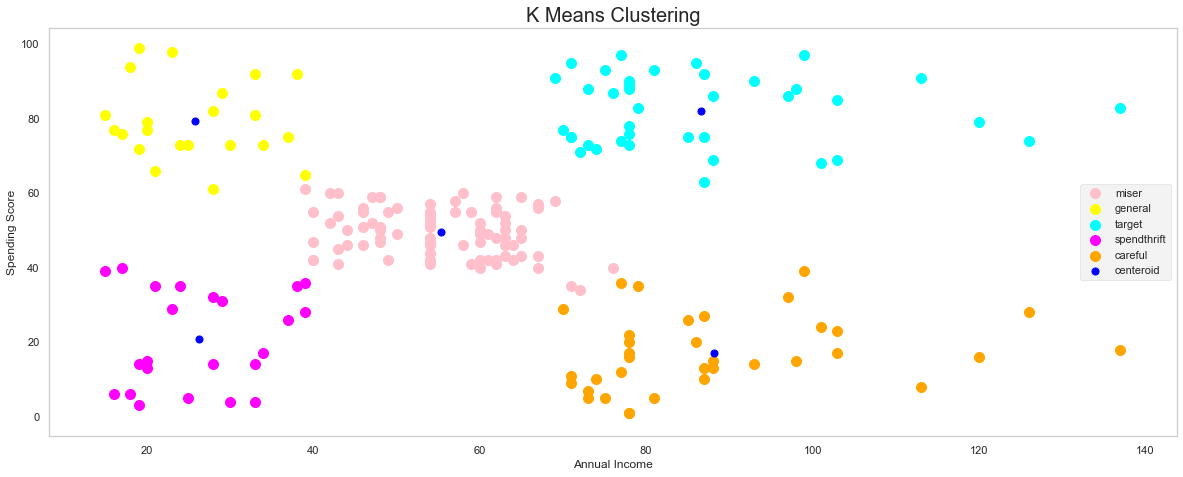
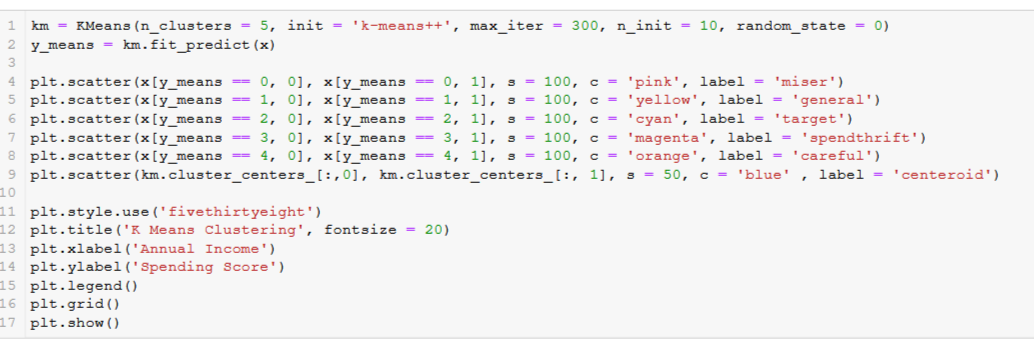


**K-means Algorithm**

The Elbow Method to find the No. of Optimal Clusters

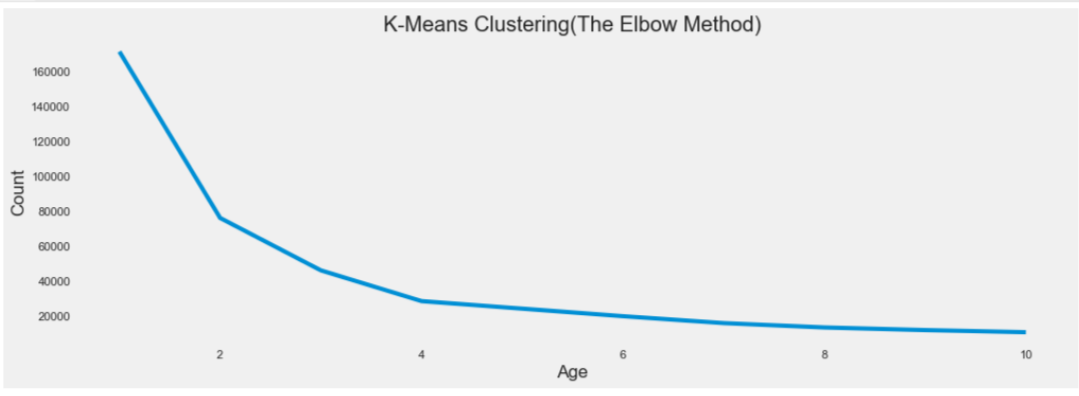
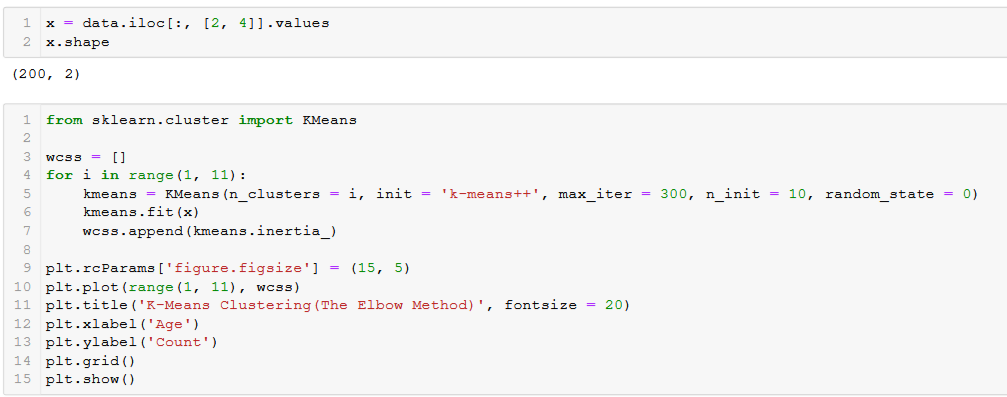


We can see from the curve that at number of cluster = 5 the curve has took the last sharp change after this point the curve has not been changes sharply so we will choose cluster = 5.

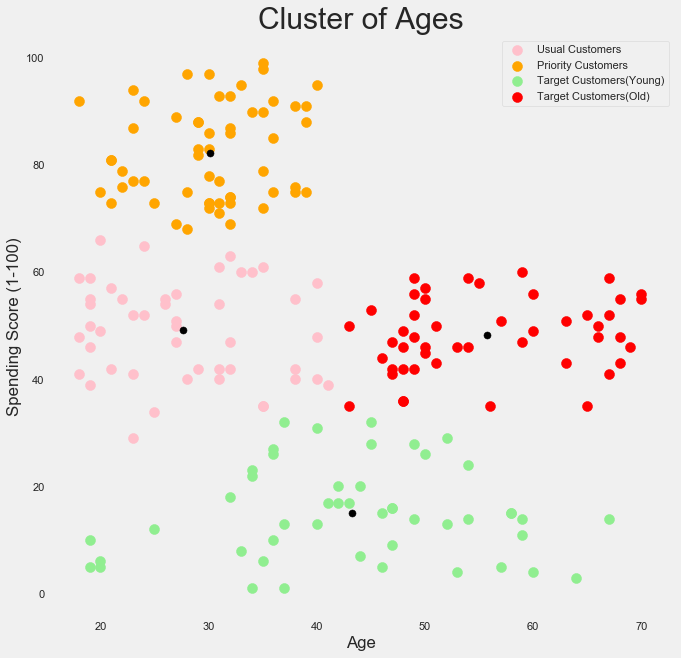
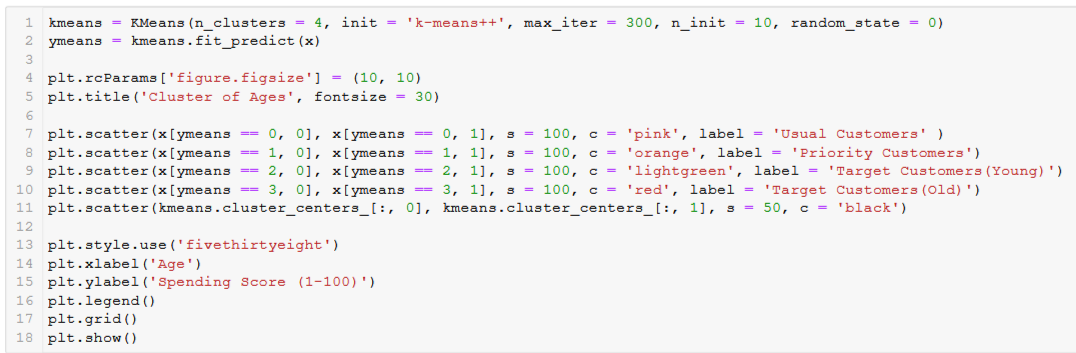


This Clustering Analysis gives us a very clear insight about the different segments of the customers in the Mall. There are clearly Five segments of Customers namely Miser, General, Target, Spendthrift, Careful based on their Annual Income and Spending Score which are reportedly the best factors/attributes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

**Clusters of Customers Based on their Ages**



**Here the last sharp change in plot at point 4 after this plot the curve is going down slowly. So we will choose 4 clusters in our Age dataset.**



According to my own intuition by looking at the above clustering plot between the age of the customers and their corresponding spending scores, I have aggregated them into 4 different categories namely Usual Customers, Priority Customers, Senior Citizen Target Customers, Young Target Customers. Then after getting the results we can accordingly make different marketing strategies and policies to optimize the spending scores of the customer in the Mall.

**CONCLUSION**

Due to increasing commercialization, consumer data is increasing exponentially. When dealing with this large magnitude of data, organizations need to make use of more efficient clustering algorithms for customer segmentation. This clustering models need to possess the capability to process this enormous data effectively. The computational speed of K-Means clustering algorithm is relatively better as compared to the hierarchical clustering algorithms as the latter require the calculation of the full proximity matrix after each iteration . K-Means clustering gives better performance for a large number of observations while hierarchical clustering has the ability to handle fewer data points.

Now we have implemented k-means clustering and saw the result. This project shows that how clustering is beneficial in grouping up same type of data. We came to know that visualization is a very important role in any project. Elbow curve is very good to find accurate no of clusters for any dataset.

However, the selection process of these suitable techniques as well as their judicious implementation could require a considerable time investment in studying and processing the data along with an adequate understanding of the goals and requirements.