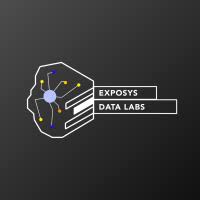
**INTERNSHIP FROM:**



**PROJECT REPORT**

Customer Segmentation using K-Means

Submitted by:

Suyash Kumar(Delhi Technical Campus)

Ahana Drall(Delhi Technical Campus)

A B S T R A C T

With the increase in number of people interested in buying new and varied products, and also the increase in number of companies providing the same, the concept of customer segmentation is gaining interest of many businesses. Customer Segmentation includes dividing the customers into groups based on their features. Studying the demographics, behaviours, and geological location etc. of the customers can help businesses to choose their audience in an effective way. It also helps them to cater to the needs of different segments of customers in a more efficient way.

The proposed model uses K-Means Clustering to divide the customers based on their Gender, Age, Annual Income, and the Spending Scores. The results of this model may be used by companies to serve each cluster of customers in a personalised way.

T A B L E O F C O N T E N T S

1. Introduction
2. Existing Method
3. Proposed Method with Architecture
4. Methodology
5. Implementation
6. Conclusion
7. I N T R O D U C T I O N

Customer Segmentation based on various important features has been proven to provide better results for the companies. This area is being researched into for quite a long time now. Researchers have worked on models with complex mathematical computations, some have made solutions based on neural networks[2], and others have tried using spatial data for customer segmentation[3]. All these methods came out to be effective for the purpose.

From so much research in this area, it is very clear that money is not the only factor which influences customers’ choices for buying or not buying a product. The use of Demographics data, as done by Namvar et. al. [4], can be a helpful step in the direction of success. The Age, Gender, Profession etc. of a person can tell a lot about the products that he can be interested in.

The data including Recency of purchase, Frequency of purchase and total monetary, as used by Namvar et. al. [4] and Marcus [1] has been proven to be effective as well.

The proposed model tries to segment customers based on their Age, Gender, Annual Income and Spending Score. This data describes a customer in quite a good way. For example, if a customer belongs to a high income and high spending score class, it can be figured that he can be of much importance to the company as compared to a person with low income and low spending score. Also, the company can analyse the needs and behaviours of the people of different segments and work towards providing each segment with the best product.

1. E X I S T I N G M E T H O D S

Many researchers have contributed to this field and used various methods to get effective results. The RFM has gained an enormous popularity among businesses.[1] RFM basically represents the data about Recency of purchase, Frequency of purchase and the Monetary value. If analysed, this data can be used to describe the most important features of the customers. The customer with high frequency of purchase can be targeted in a way that we would continue to buy the products as the same frequency(or may be more), the customer with low frequency of purchase can be targeted in another way.

The concept of Customer Value Matrix was also introduced[1]. In this method, the customers are given scores and based on these, the customers can be allotted different segments. The idea of Customer Value Matrix was to use RFM for small businesses as well. But the disadvantage of this method was that despite being a very simple concept, the outputs achieved were very complex to be studied by everyone and it was eve more complex to implement them.

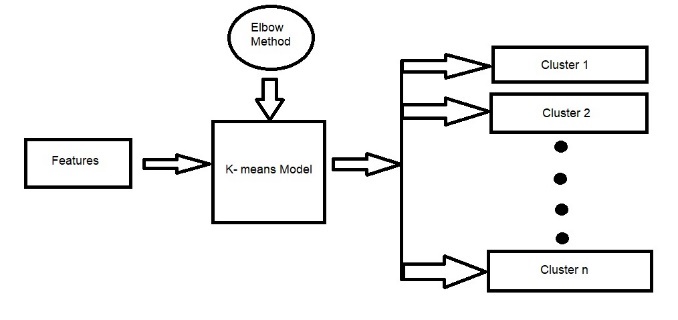
A Neural network based approach is applied by Yan et.al. [2], unlike the other neural networks approaches, they have first segmented the customers based on their behaviours and after that, combined it with the static attributes. This output is given as input to the neural network and the customers are then classified. The behavioural attributes that are used for clustering the customers are: Trends data, monetary, purchase frequency and standard variance. The trends data basically represents the trends in customer transaction records. The monetary represents the consumption amount of the customer. The third attribute i.e. purchase frequency, is used to analyse the purchasing habits of the customers. Some customers have low purchase frequency as compared to others and hence, contributing differently to the company. The standard variance represents the fluctuation in customers’ purchase trend. Next, the classification is done after the feature selection step.

The technique of spatially enabled segmentation task[3] has also been researched on. The customer attributes used in this work are sales information of the customers, non-spatial attributes including demographic data about a customer’s workplace, demographic data about the customers’ residential area etc. and the third attribute is the spatial predicates. Initially, a number of spatial attributes are selected and then, a filtration process is applied to find out the most relevant features. Rough Set clustering is used along with the filtration algorithm, to segment the customers into clusters.

Namvar et.al [4] describes in their work, that the use of demographics data, RFM and LTV is very common method for customer segmentation. So, in their work, they combined the above mentioned three variables in such a way that it provides better results than the other approaches. A three step approach is applied where in the first step, the examination of the demographics and the RFM has been done. In the second step, LTV is used to choose the most efficient method of segmentation. And third, a rank based approach is used to compare these clusters.

1. P R O P O S E D M O D E L

The proposed model is a three step model. First, Visualising the trends between various features. Second, making segments of customers using K-Means Clustering algorithm. And finally, making conclusions using the clustered data.



The dataset that has been used is the Mall\_Customers dataset. The features that are used in this model are Gender, Age, Annual Income, Spending Score. These features are analysed against each other and insights are recorded.

The features are given to the K-Means model not all at a time, but gradually.

At first, they are given as input to the model in pairs and then, three of the features are given to the K-Means model to obtain a 3D clustering.

The number of clusters is specified by ‘The Elbow Method’. Elbow method is a plot between the sum of squared errors and the number of clusters. The elbow of this curve corresponds to the number of optimal clusters.

The output of the K-Means model is the clustered data points based on the features provided to the model.

These clusters are then analysed and decisions can be made by the companies towards providing personalised treatment to all segments.

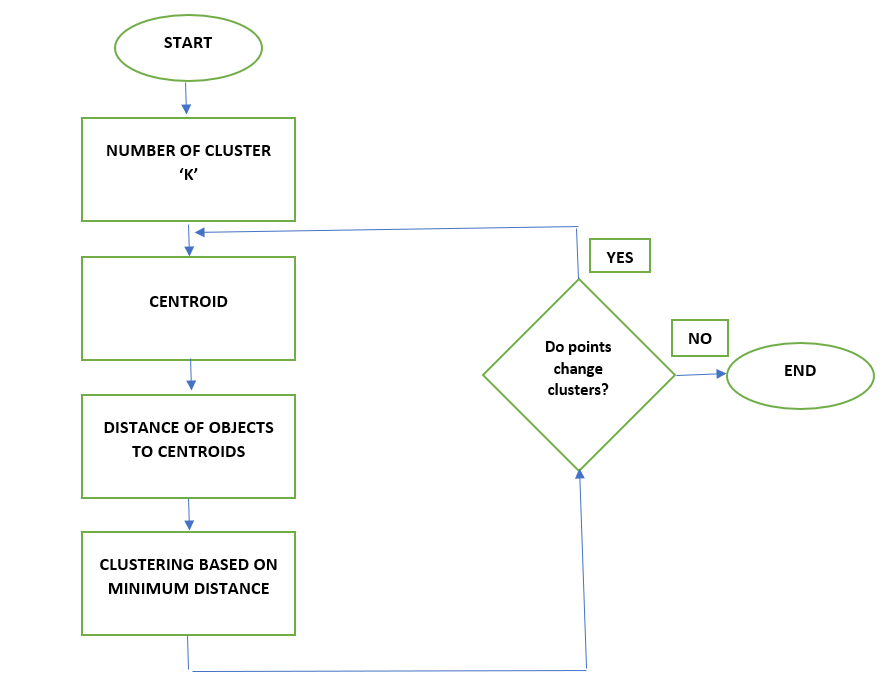
1. M E T H O D O L O G Y

The algorithm used for this model is K-Means clustering algorithm. This is basically a clustering algorithm which aims at segmenting n data points into k clusters in a way that each observation belongs to the cluster with nearest means.

4.1 K-MEANS CLUSTERING: Imagine having data points which you want to divide into three clusters. Initially, choose three random distinct data points. These are the initial cluster centroids. Now calculate the euclidien distance of each data point(sample) to these selected distinct data points. Assign each sample to the nearest centroid and make clusters.

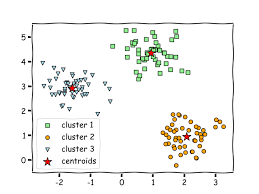
Now since we have our initial three clusters, we calculate the mean of each cluster and recalculate the distance of each data point from these means. Re-cluster these data points according to the nearest clusters. This will repeat till the time, there is no change in the assignment of data points to the clusters, i.e. the data points do not change their clusters.

The flow chart for K-Means is given below for better understanding.



Basically, clustering is about minimizing the distance between the points in a cluster and maximizing the distance between the clusters.

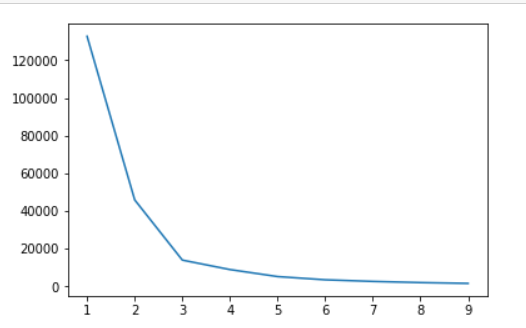
We have successfully classified our data into three clusters.



4.2 THE ELBOW METHOD: in the above section, we assumed the value of ‘K’ to be 3. ‘K’ here is the number of clusters. The elbow method is a method which specifies the best number of ‘K’ for clustering.

The elbow method is a graph between the WCSS and the number of clusters. WCSS is the within cluster sum of squares, i.e. sum of distances from each data point to the centroid added with the sum of distances from each data point in another cluster to their cluster centroid.

If we minimize the WCSS, we have reached the optimal clustering solution.



It can be seen from the graph that the WCSS is decreasing at a very fast speed for less number of clusters. But then, for k=3; the speed of decrement is almost negligible.

This point is referred to as ‘the elbow’. And the number of clusters corresponding to this elbow is take as the value for n\_clusters for K-Means model.

The data points in the proposed model have been clustered based on the above mentioned methodology.

1. I M P L E M E N T A T I O N

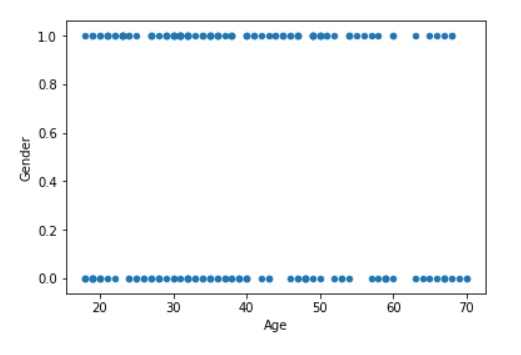
Jupyter Notebook has been used for the complete project. The libraries that have been used for the same are:

1. Pandas version 1.0.1: this is a python library that is used for handling and manipulating data.
2. Matplotlib version 3.1.3: this is a python library that is used for plotting.
3. Sklearn version 0.22.1: Sklearn is a python library for machine learning. It features regression, classification and much more.

The data set used for this project is Mall\_Customers. The Gender column in this dataset has been modified. ‘Male’ is represented by 0, and ‘Female’ is represented by 1.

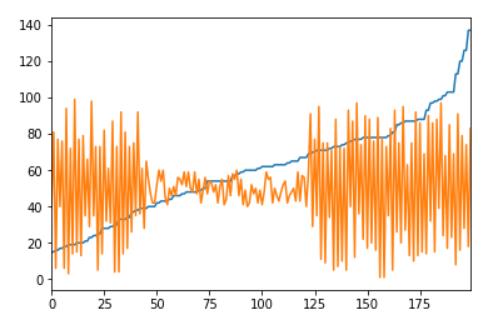
The three steps in which the project is implemented are given below:

* 1. VISUALISING THE DATA
     1. Age v/s Gender



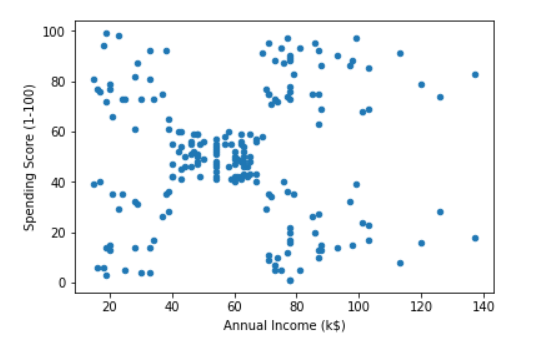
The scatter plot for the Age v/s Gender shows the distribution of Males and Females based on Age.

* + 1. Annual Income v/s Spending Score



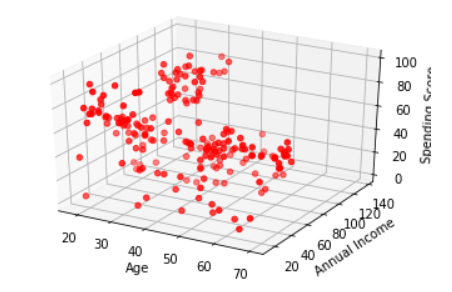
This is the line plot between Annual Income and Spending Score. The blue lines shows the annual income whose highest value is 137 k$. the orange plot shows the ranges of the spending score.

To understand this further, we made a scatter plot for the same.



In this scatter plot, we can clearly see that there Is a group of customers who have low incomes and low spending score, low income high spending score, high income low spending score, high income and high spending score. Also there is a group of people who come in the moderate class. But, we will not reach any conclusion yet, we will confirm this from our K-Means model in further sections.

* + 1. Age v/s Annual Income v/s Spending Score

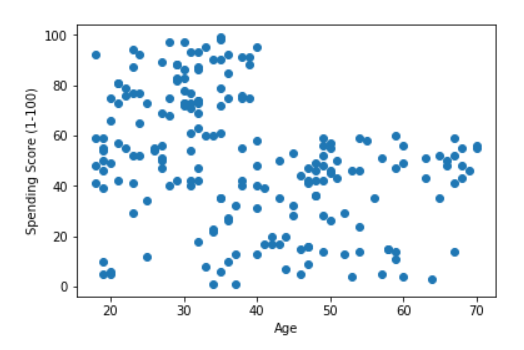


This is a 3D scatter plot which plots the data points based on not two, but three features. We can make insights out of this visualisation. We can see that there are less people with low annual income as compared to people with high annual income. Also, a group of people can be clearly seen who are younger age grouped and have high incomes and high spending scores. For clearer segmentation, we will look at K-means clustering in the next section.

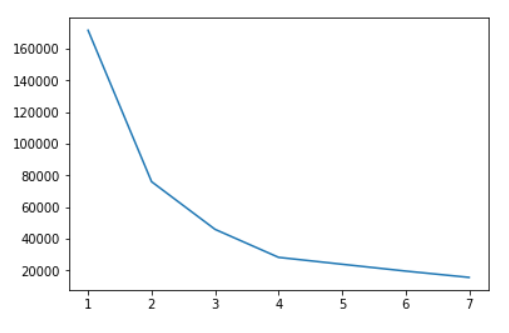
* 1. CLUSTERING USING K-MEANS

In this section, the use of K-Means clustering algorithm has been discussed.

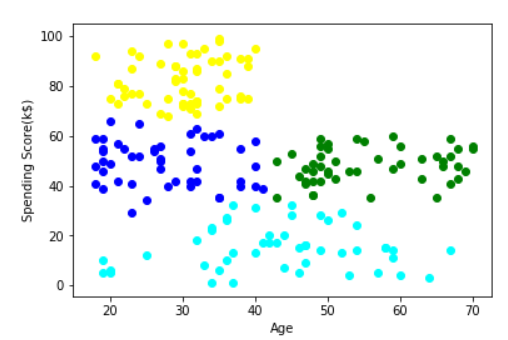
* + 1. Clustering based on Age and Spending Score.



This is a scatter plot of Age v/s Spending Scores. The elbow curve given below has given us the perfect number of clusters for segmenting the customers based on their age and spending score.



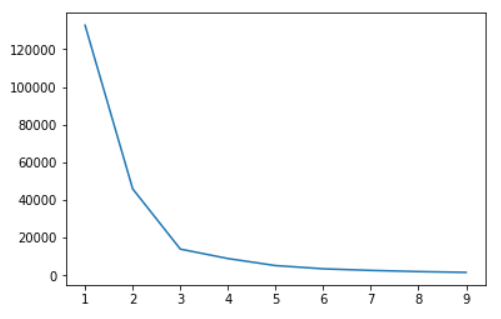
This elbow curve has its elbow at 4. So the number of clusters is 4. Next we make a K-Means model with n\_clusters=4. The model gives us the following output.



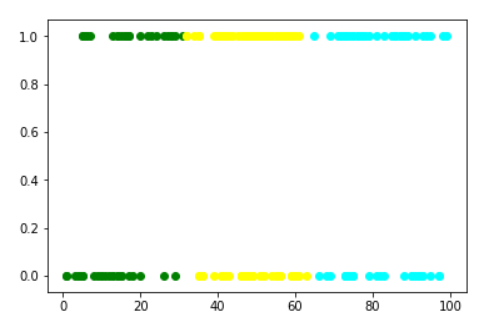
Here, we can clearly see that the cyan cluster is of the people with very less spending scores. The yellow cluster is of young people who spend quite much. The blue cluster is of the customers who have moderate spending score whereas, the green cluster is of the older people who spend somewhat more than other customers in their age group.

* + 1. Clustering based on Gender and Spending Score



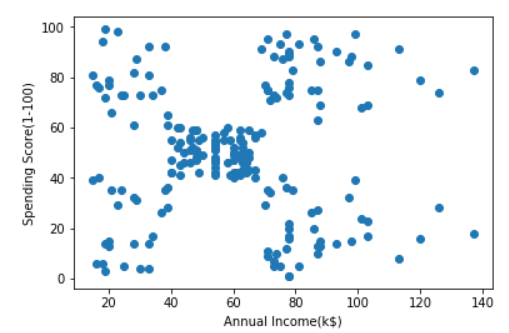


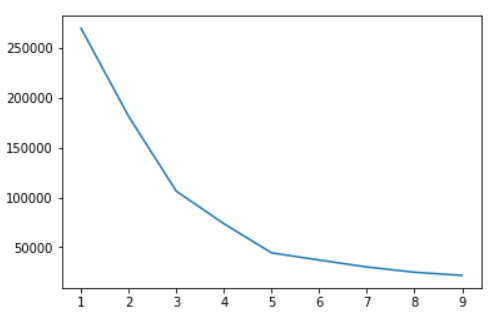
The elbow curve here shows the number of clusters should be 3 and hence, n\_clusters has been given the value 3, in the K-Means model.



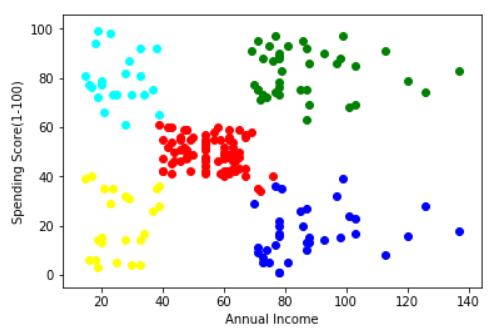
The output has only one significant insight, i.e. the number of females with high spending score is a bit more than the number of males spending high.

* + 1. Clustering based on Annual Income and Spending Score





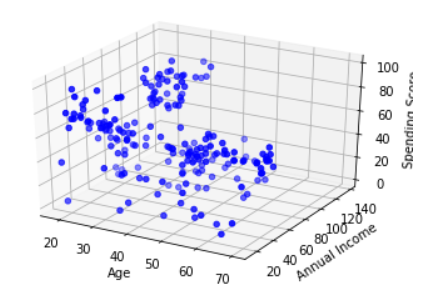
Number of clusters can be clearly seen as 5 from the elbow curve. Segmenting our data points into 5 clusters gives the following output.



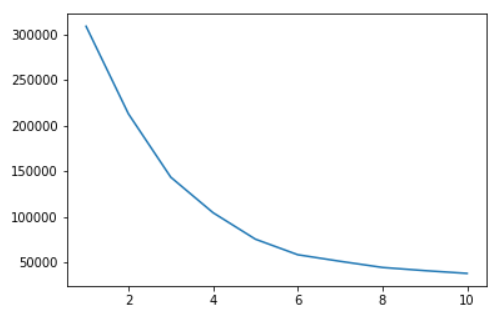
Here, the cyan cluster is a cluster of people with low annual income and high spending scores. The yellow cluster is of the people with low annual income and low spending score. The green cluster is of the people with high income and high spending score and the blue cluster is of the people with high annual income but low spending scores.

Hence we have successfully clustered customers based on their annual income and spending score.

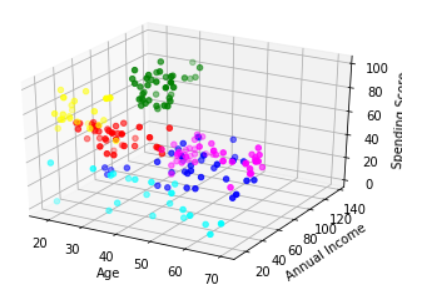
* + 1. Clustering based on Age and Annual Income and Spending Score



The above is a 3D plot between Age, Annual Income and Spending Score. Next we have used K-Means algorithm to cluster this 3D plotted data.



The value of K being 6, the K-Means model has n\_clusters=6. The segmented data can be seen in the scatter plot below.



Here, the yellow and the green cluster is of the customers who have very high spending scores and are from younger age groups. The cyan cluster is of the customers with very low spending scores. The red cluster is of the people with moderate annual income and moderate spending scores. Also, the pink cluster is of the people with high annual income but low spending score.

* 1. ANALYSING THE CLUSTERS

Analysing the cluster in section 5.2.1. we can figure different groups of people having similar properties in each group. In section 5.2.2. we can see how more number of females spend more than males. In section 5.2.3. we can see five clusters of people in which there are people who have low incomes but still they spend very much, also the people who have high incomes and high spending scores. These customers contribute more to the enterprise as compared to other customers. the enterprise should take intense care of these customers. in section 5.2.4. we segment customers based on their age, annual income and spending scores. This segmentation basically combines all the above clustering. From this distribution, many decisions can be taken. The people with very low spending score contribute negligibly to the enterprise whereas the people with younger age groups and high spending scores contribute more in the enterprise. The younger customers with low income also spend more and hence, the company should keep up their trust with the product. There are customers who have high incomes but still low spending scores. The companies should take feedbacks from these customers and should try to analyse as to why they spend so less, even after have very high incomes.

1. C O N C L U S I O N

We have successfully segmented the customers based on various features and made insights out of the clusters. The enterprise can use this data to give different treatment to each of the segments, based on their annual incomes, spending scores, gender and ages.

1. R E F E R E N C E S:
2. Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. Journal of consumer marketing.
3. Yan, X. B., & Li, Y. J. (2006, February). Customer segmentation based on neural network with clustering technique. In Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Madrid, Spain (pp. 265-268).
4. Fan, B., & Zhang, P. (2009). Spatially enabled customer segmentation using a data classification method with uncertain predicates. Decision Support Systems, 47(4), 343-353.
5. Namvar, M., Khakabimamaghani, S., & Gholamian, M. R. (2011). An approach to optimised customer segmentation and profiling using RFM, LTV, and demographic features. International Journal of Electronic Customer Relationship Management, 5(3-4), 220-235.