**EXPOSYS DATA LABS**

**DATA SCIENCE**

**CUSTOMER SEGMENTATION USING K-MEANS ALGORITHM**

**BY**

**ABDUL ZAHEER**

**Abstract:**

In the modern era, data plays a major role to analyze and improve performance. It is used to understand the relations among the factors that influence the consequence. Information is extracted from the data for better solutions. Based on the factors that influence the outcome the data is segmented and the required information is gathered. The main goal of this project is to classify the customers based on their characteristics and find suitable customers for approach. Through analyzing different groups of customers, we try to position the target clients of the company properly. On doing this we can achieve better results. Various algorithms are applied to explore the hidden patterns in the data for better decision making. In this project, customer segmentation is done by k-means which is an unsupervised learning technique.

**Table of Contents:**

Abstract 2

Introduction 2

Existing Method 3

Proposed method with Architecture 3

Architecture Diagram 4

Methodology 4

Implementation 5

Conclusion 11

**Introduction:**

Customer Segmentation is the process of making or dividing the customers into several groups based on certain criteria that are common among them. It is one of the unsupervised learning where there will be no particular reference for the dataset to divide into categories. Cluster analysis is one of the most used techniques in unsupervised learning to divide the whole dataset into the required number of segments (or parts). K-means clustering is one of the most commonly used methods for cluster analysis. In k-means clustering ‘k’ stands for the number of clusters (or segments). The minimum value of k is 2.There is a popular method called elbow method which gives the optimal value for k for the given dataset.

**Existing Method:**

The existing system uses the elbow method that is used to calculate within a cluster the sum of squared errors (WCSS) for different values of k and choose the k for which WCSS first starts to diminish. In this method, K means clustering is the segmentation of customers to get a better understanding of them which in the turn could be used to increase the revenue of the company. In this method, k means clustering is used which is the most popular clustering algorithm and usually, the first thing practitioners apply when solving clustering tasks to get an idea about the structure of the dataset.

**Proposed method with Architecture:**

We use the elbow method in the propose system that is we calculate within the cluster the sum of squared errors(WCSS) for different values of k and we plot the curve of  WCSS vs the number of clusters K for better understanding among the clusters. The elbow formation used in our system usually gives the optimum number of clusters. We made a bar plot to visualize the number of customers according to their annual income. Our system gives meaningful insights and understanding by using clustering algorithms to generate customer segments. We will use the k-means clustering algorithm to derive the optimum number of clusters and understand the underlying customer segments based on the data provided.

****

Fig a: Architecture Diagram

**Methodology:**

Clustering is an unsupervised learning method and K-means is a better- unsupervised machine learning algorithm used to divide the data set into similar groups. K-means is an iterative algorithm that iterates through the dataset to partition the dataset. This algorithm divides the dataset into k predefined subsets. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid that is the arithmetic mean of all the data points that belong to that cluster is at the minimum.

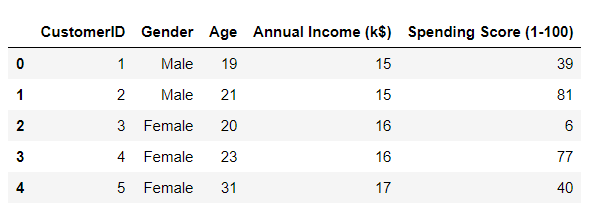
The method used in this project to find out the number of clusters is the Elbow method. The elbow method runs k-means clustering on the dataset for a set of values for k and then for each value of k the method computes an average score that is the midpoint for all clusters. Using this method we can find out the optimal number of clusters so that the diversity between the customers is reduced. Here in this project, the value of k is 5, that is the number of clusters formed from the dataset is 5.

The customers are clustered based on their age group to find out the average annual income and spending scores of different groups. Based on these clusters formed, we can understand better that which age group customers have more income and which age group customers can spend more. This type of customer segmentation is very useful in the market to attract more customers of such age groups who have more income and more spending scores.

**Implementation:**

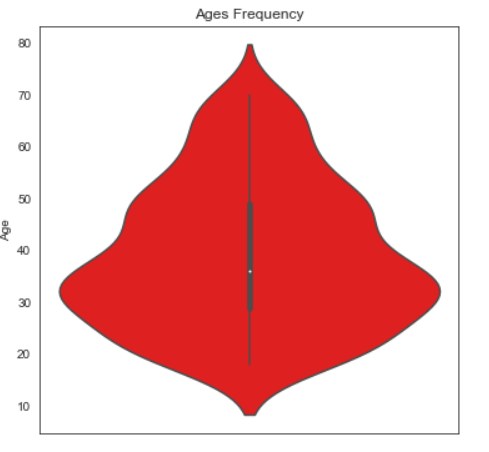
The columns present in the given dataset are Customer ID, Gender, Age, Spending score, Annual income

First, loading the given dataset-

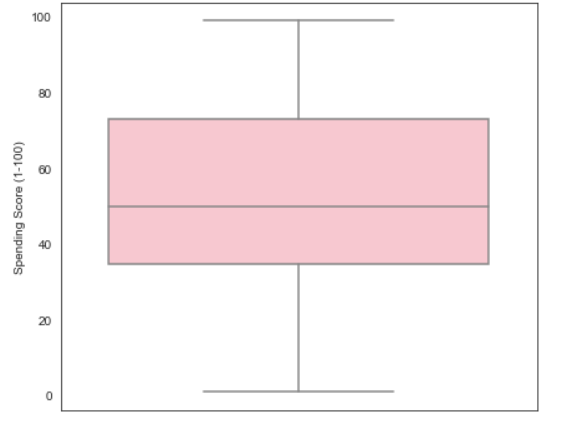


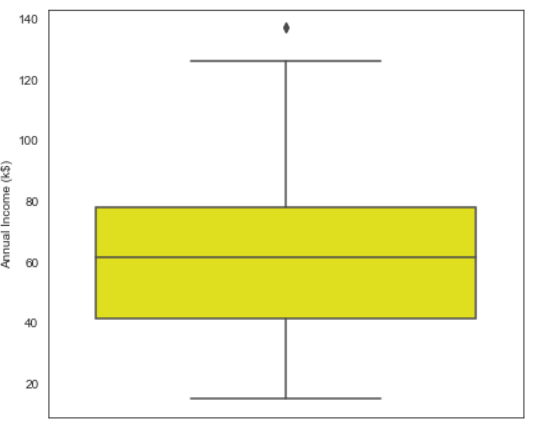
The dataset is successfully loaded.

Using the sea born library, calculate the frequency of ages by plotting the violin graph.

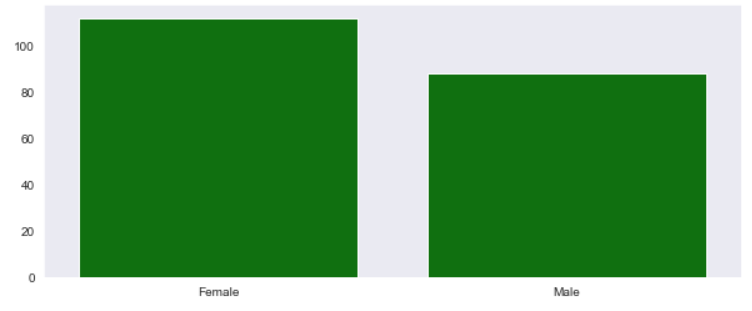


For better understanding, we made a box plot. It clearly says that the range of spending scores is more than the range of annual income.

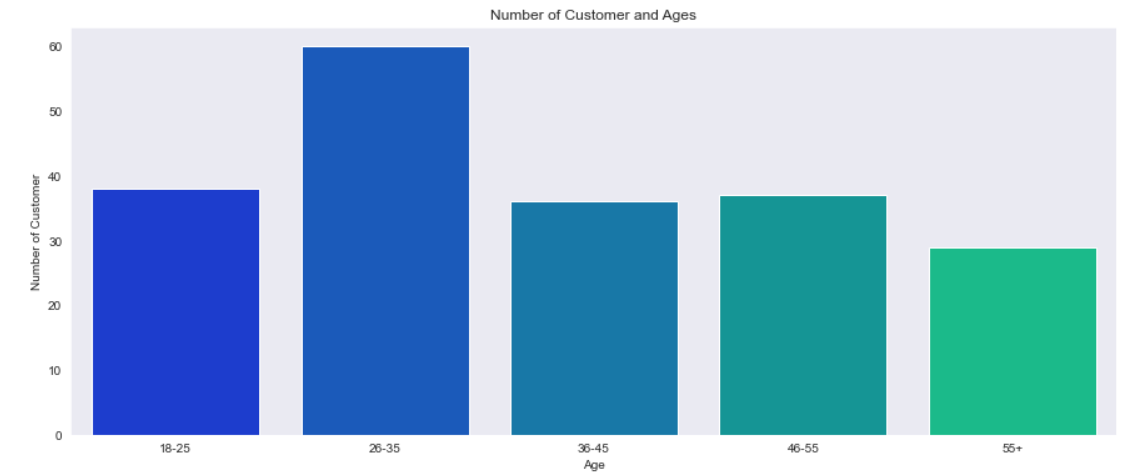




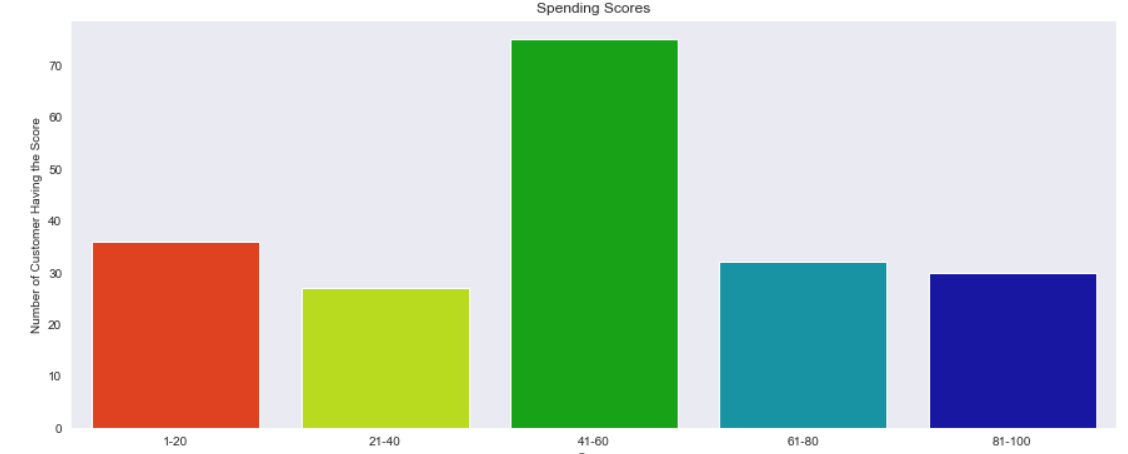
To examine the distribution of female and male population we made a bar graph. It clearly says that the female population exceeds the male population.



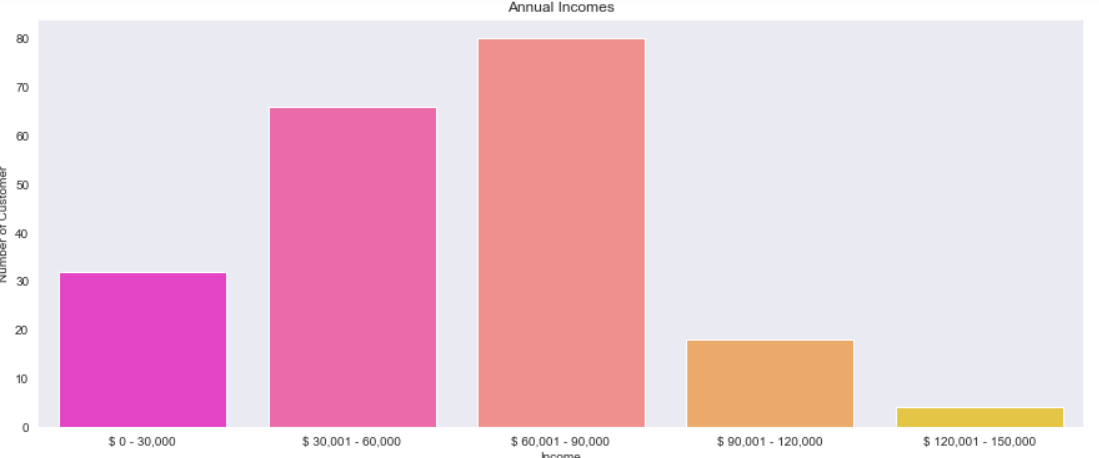
To check the distribution of the number of customers in each age group we made a box plot. It clearly says that the age group from 26-35 outweighs every other age group.



Next, we made a bar graph which visualizes the number of customers to their spending scores. The majority of customers have spending scores in the range of 41-60.



And also we made a bar graph which visualizes the number of customers to their annual income. The majority of customers have annual income in the range of 60000 and 90000.



To calculate the optimal number of clusters value we plotted a Within Cluster Sum Of Squares (WCSS measures the sum of distances of observations from their cluster centroids) against the number of clusters(K value).

                         WCSS=∑(Xi + Yi )2

where Yi is centroid for observation xi.

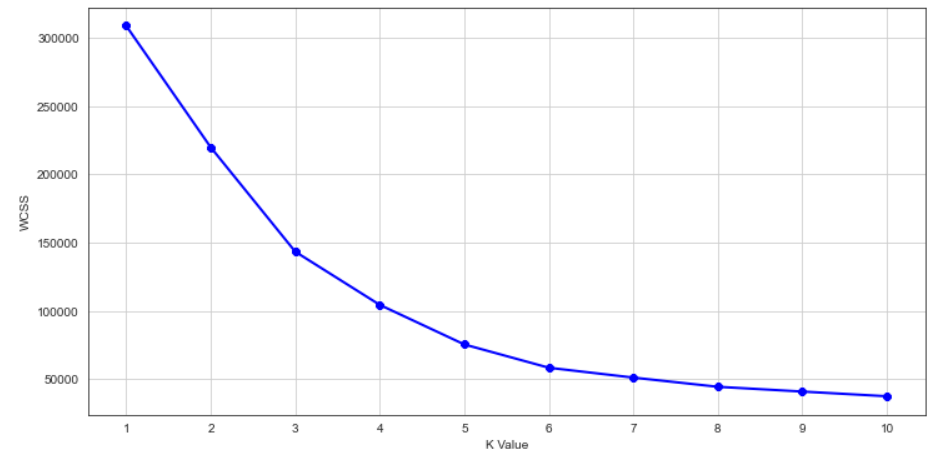
Here we use the elbow method for calculating the number of clusters i.e; K value.

Elbow method:

           Calculate the WCSS for different K values and choose the K for which WCSS first starts to diminish. The plot of WCSS versus K, visible like an elbow.

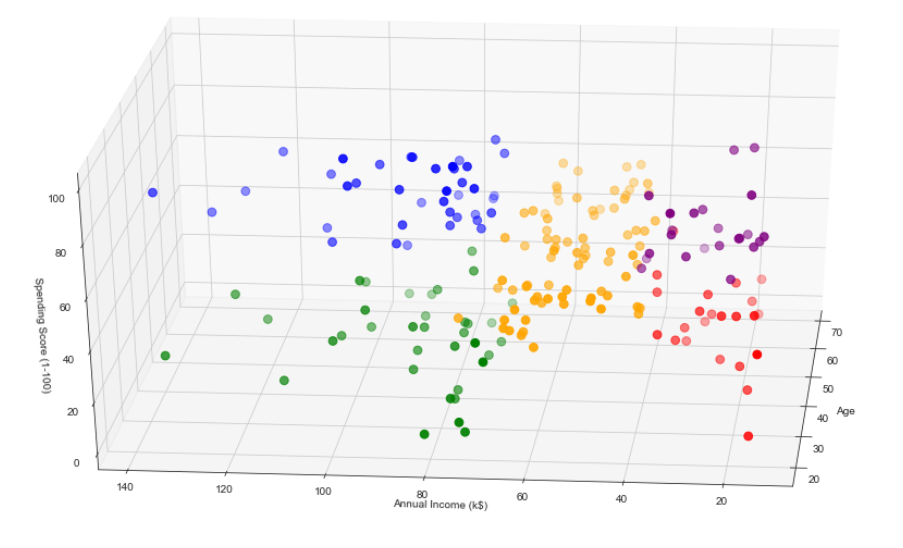
The steps can be summarized in the below steps:

1. Compute K-Means clustering for different values of K by varying K from 1 to 10 clusters.
2. For each K, calculate the within-cluster sum of square (WCSS).
3. Plot the curve of WCSS versus the number of clusters(K).
4. The location of bend in the plot is generally considered as an indicator of the appropriate number of clusters.



The optimal K value is found to be 5 using the elbow method.

Finally, we made a 3D plot to visualize the spending scores of the customers with their annual income. The data points are separated into 5 classes which are represented in different colors as shown in the 3D plot.



**Conclusion:**

For retaining arise in the progress customers must be analyzed. The patterns of the customers must be understood for attaining the best plan of action. Customer segmentation which is a process of dividing into groups based on the similarities is done. The customer segmentation is achieved by unsupervised learning methods such that inferences are drawn without label responses. The k means clustering is used to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. For optimization of result elbow method which chooses the best k value is implemented. This process helps to find target clients and improve the outcomes.