Internship

On

CUSTOMER SEGMENTATION USING K-MEANS ALGORITHM

EXPOSYS DATA LABS

Profile: DATA SCIENCE

Submitted By

BATHINI MONISH KAMAL

Department of Metallurgical Engineering

IIT (B.H.U) VARANASI

November 2021

Abstract

In the modern era, data plays a significant role in analyzing and improving 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 the approach. Through analyzing different groups of customers, we try to position the target clients of the company properly. By 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

Chapter 1

Introduction to Customer Segmentation

1.1)What is Customer Segmentation?

Customer segmentation is the practice of dividing a company's customers into groups that reflect similarity among customers in each group. The goal of segmenting customers is to decide how to relate to customers in each segment in order to maximize the value of each customer to the business.

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.

1.2)The Importance of Customer Segmentation

Customer segmentation has the potential to allow marketers to address each customer in the most effective way. Using the large amount of data available on customers (and potential customers), a customer segmentation analysis allows

marketers to identify discrete groups of customers with a high degree of accuracy based on demographic, behavioral and other indicators.

Since the marketer's goal is usually to maximize the value (revenue and/or profit) from each customer, it is critical to know in advance how any particular marketing action will influence the customer. Ideally, such "action-centric" customer segmentation will not focus on the short-term value of a marketing action, but rather the long-term customer lifetime value (CLV) impact that such a marketing action will have. Thus, it is necessary to group, or segment, customers according to their CLV.

1.3)CLV-Focused Customer Segmentation

Of course, it is always easier to make assumptions and use "gut feelings" to define rules which will segment customers into logical groupings, e.g., customers who came from a particular source, who live in a particular location or who bought a particular product/service. However, these high-level categorizations will seldom lead to the desired results.

It is obvious that some customers will spend more than others during their relationship with a company. The best customers will spend a lot for many years. Good customers will spend modestly over a long period of time, or will spend a lot over a short period of time. Others won't spend too much and/or won't stick around too long.

The right approach to segmentation analysis is to segment customers into groups based on predictions regarding their total future value to the company, with the goal of addressing each group (or individual) in the way most likely to maximize that future, or lifetime, value.

1.4)Customer Segmentation Models

Accurate customer segmentation involves tracking dynamic changes, and frequently updating new data. Although segmenting customers according to their CLV is the recommended approach, there are many types of customer segmentation models. Some of the more common types are segmentation via

cluster analysis, RFM segmentation, and longevity. Some marketers might even combine one or more segmentation models in order to reach their goals.

No matter the types of segmentation models marketers decide to use, they all require marketers to create groupings of customers to serve as a first step in segmenting the customer base. Usually this will result in marketers having a series of tiers for each type of segmentation model. Marketers can then mix different tiers across models to create more defined segments. For example, mixing the highest tier of customers based on an RFM model and combining it with a low longevity tier will result in marketers having a segment of highly active, newly acquired customers.

1.5)Customer Segmentation and Machine Learning

An additional approach to customer segmentation is leveraging machine learning algorithms to discover new segments. Different to marketer-designed segmentation models, as the ones described above, machine learning customer segmentation allows advanced algorithms to surface insights and groupings that marketers might find difficulty discovering on their own.

Furthermore, marketers that create a feedback loop between the segmentation model and campaign results will have ever improving customer segments. In these cases, the machine learning model will be not only able to refine its definition of segments, but also be able to identify if a specific subset of the segment is outperforming the rest, optimizing marketing performance.

1.6)The Optimove Approach to Customer Segmentation

Most frequently, the methods marketers use for segmentation take the form of hard-coded rules based on experience and assumptions. More sophisticated approaches use mathematical models to analyze large amounts of data to group customers with similar data sets into particular segments. However, these

approaches ignore a critical component of accurate customer segmentation: how do customers migrate from one segment to another over time?

Optimove uses all available data and employs sophisticated clustering models to perform highly accurate segmentation. In fact, this technology actually results in large numbers of finely sliced micro-segments. However, the “secret sauce” of Optimove’s segmentation is a focus on the dynamic nature of customer behavior. In other words, Optimove continuously recalculates the segmentation of every customer and tracks how customers move from one micro-segment to another over time.

In other words, most companies view segmentation as a method of clustering similar customers together at a given point in time, but they completely disregard the path or route that each customer has taken to reach his or her present segment. By analyzing customers based on their movement among segments over time, Optimove achieves far more accurate segmentation than any other known method.

Furthermore, the combination of this dynamic segmentation approach with Optimove’s ability to create extremely homogenous and compact micro-segments results in an unparalleled degree of customer segmentation accuracy. When factoring in Optimove’s core focus on customer lifetime value in all calculations, it is easy to see why Optimove’s ability to predict the response of every customer to any marketing action is a generation ahead of any other marketing action optimization solution.

Chapter 2

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.

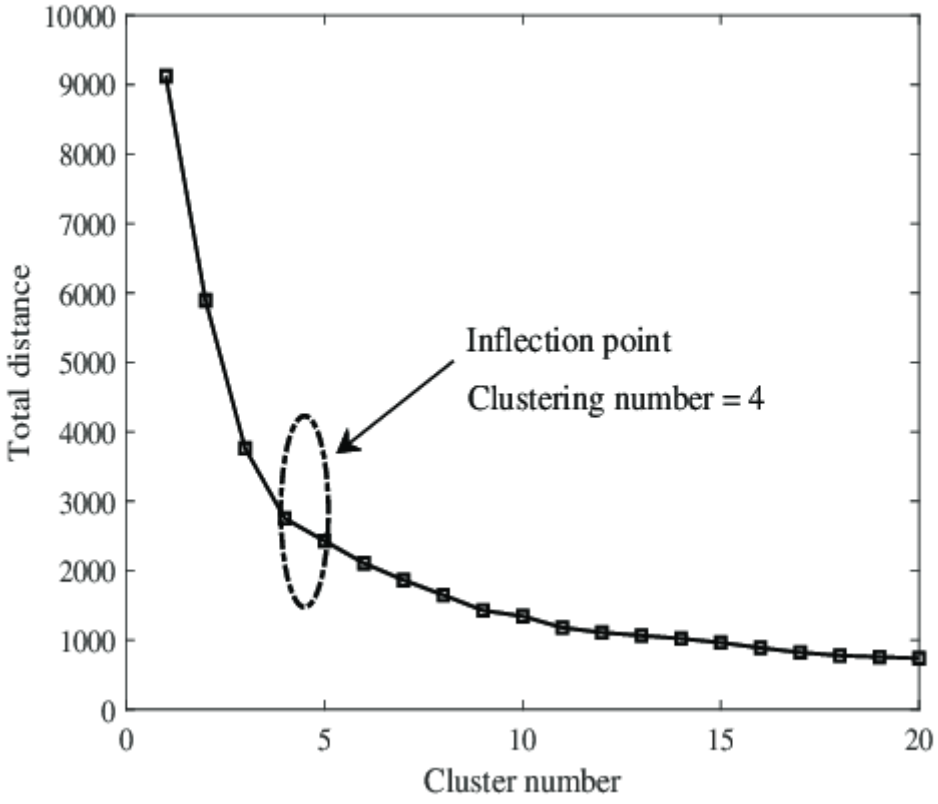


Fig: Elbow Method

Chapter 3

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.

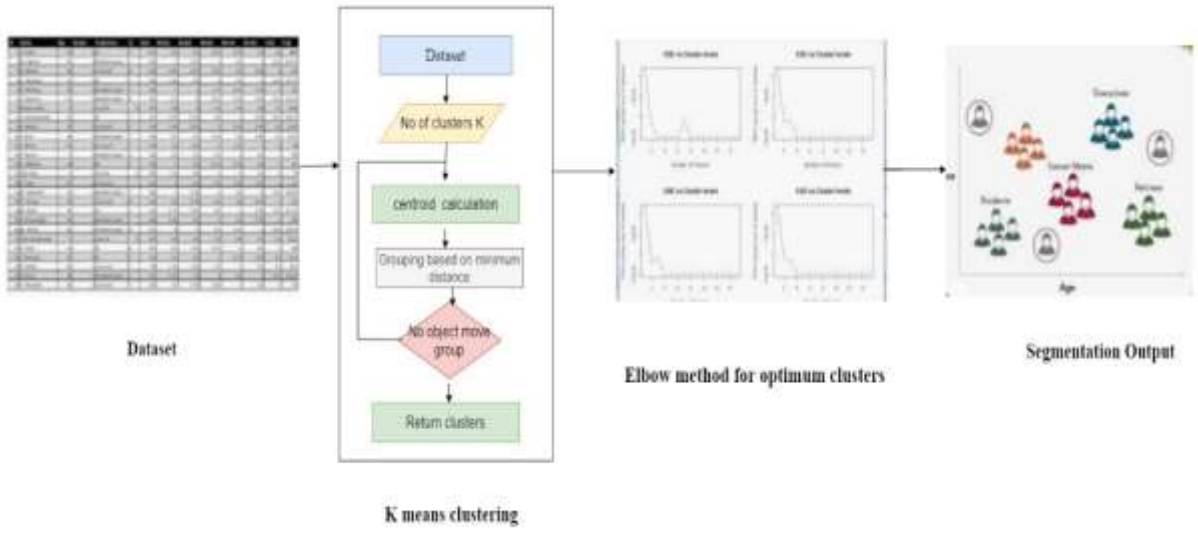


Fig :Architecture Diagram

Chapter 4

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 4 that is the number of clusters formed from the dataset is 4.

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.

Chapter 5

Implementation

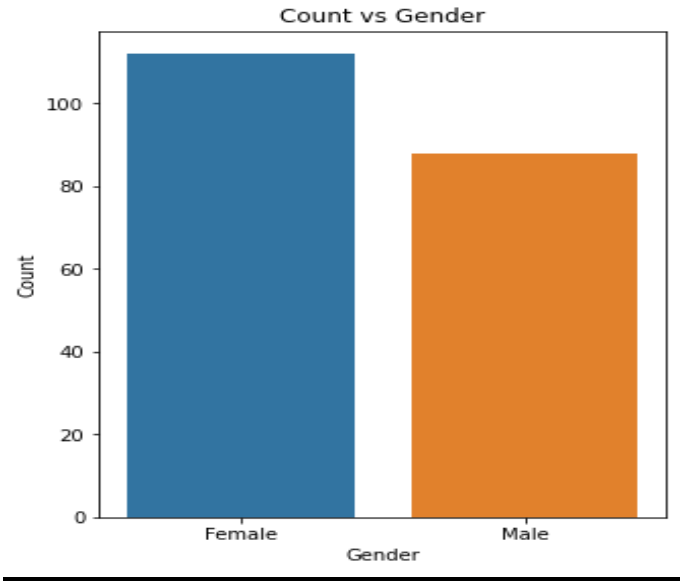
The columns present in the given dataset are Customer ID, Gender, Age, Spending score, Annual income

First, loading the given dataset-

	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

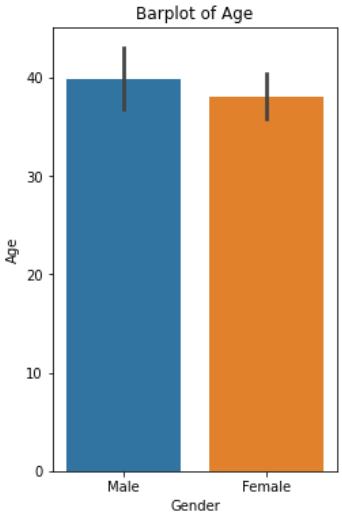
The dataset is successfully loaded.

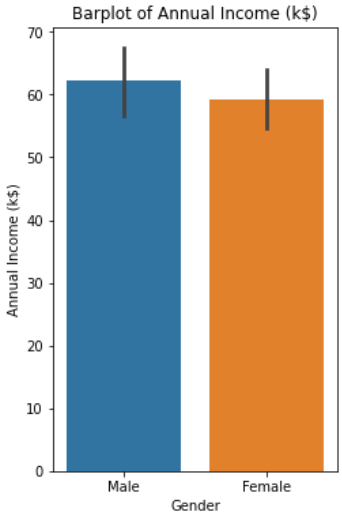
Using the sea born library, calculate the frequency of ages by plotting the bar graph.

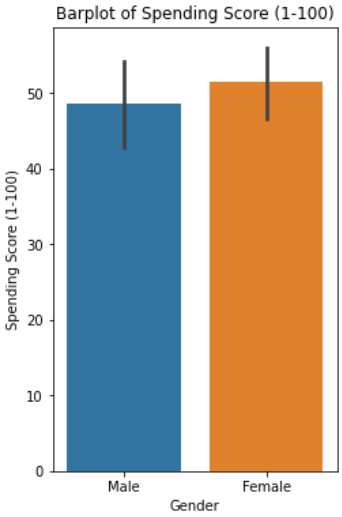


To examine the distribution of female and male population we made a bar graph as shown in the above diagram. It clearly says that the female population exceeds the male population.

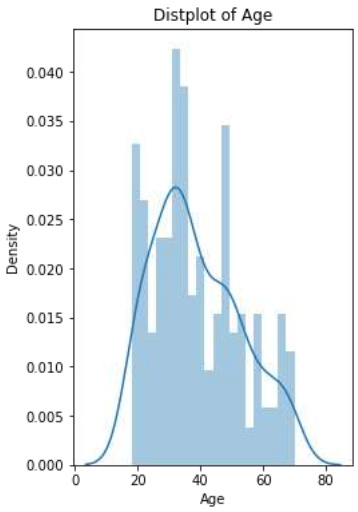
Extending the above graph we plotted graphs between “Age vs Gender”, “Annual Income vs Gender” and “Spending Score vs Gender” for understanding ratio in which the male and female are divided based on each category as shown in the diagram below. It clearly says that, except in case of Spending score, the number of female customers are lesser than male customers.

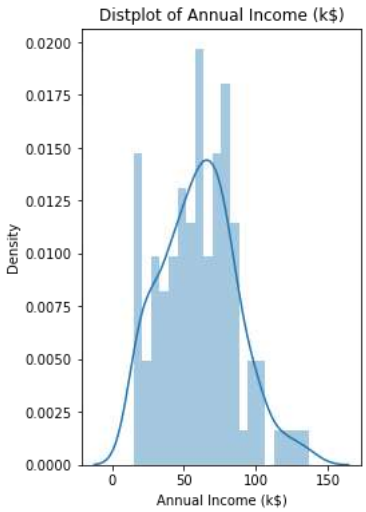


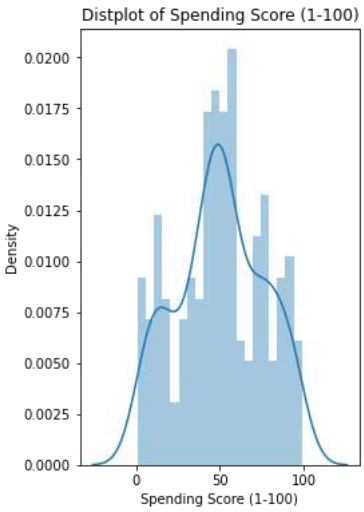




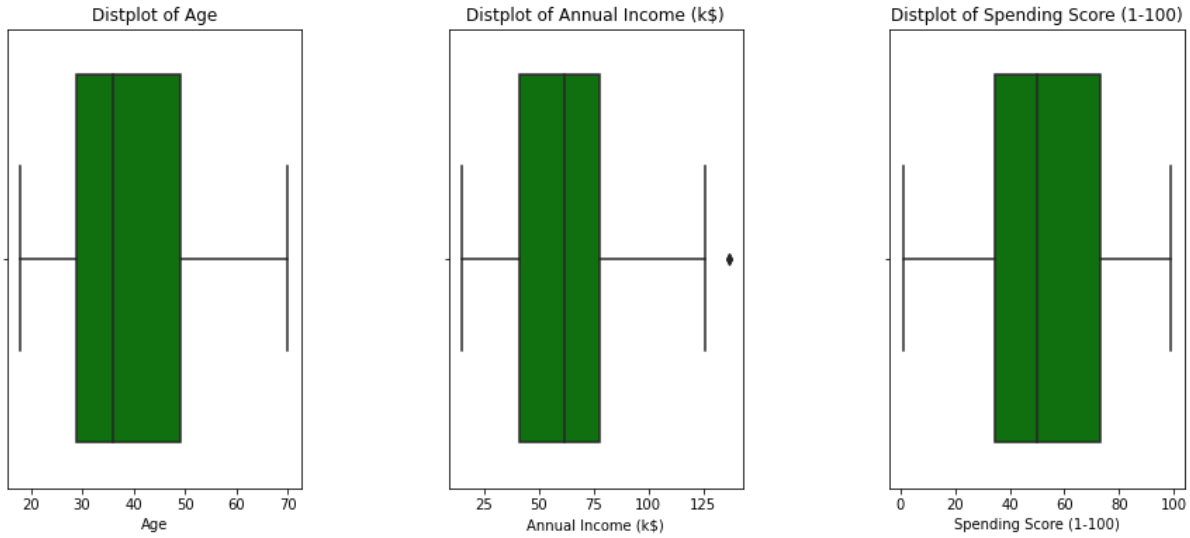
Now we made a distplot to understand the desity at which more customers are concentrated as shown in the diagram below.



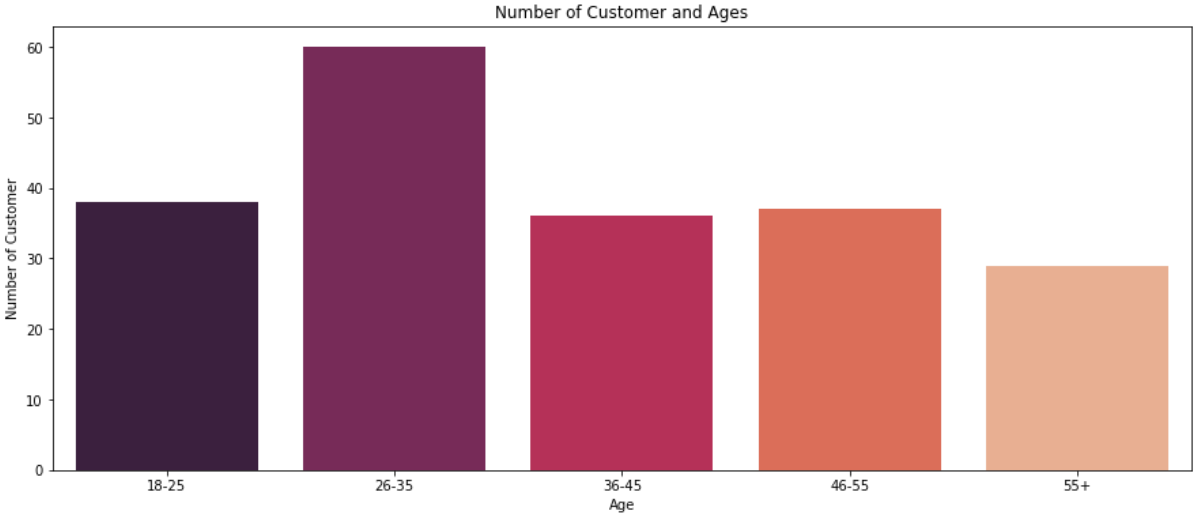




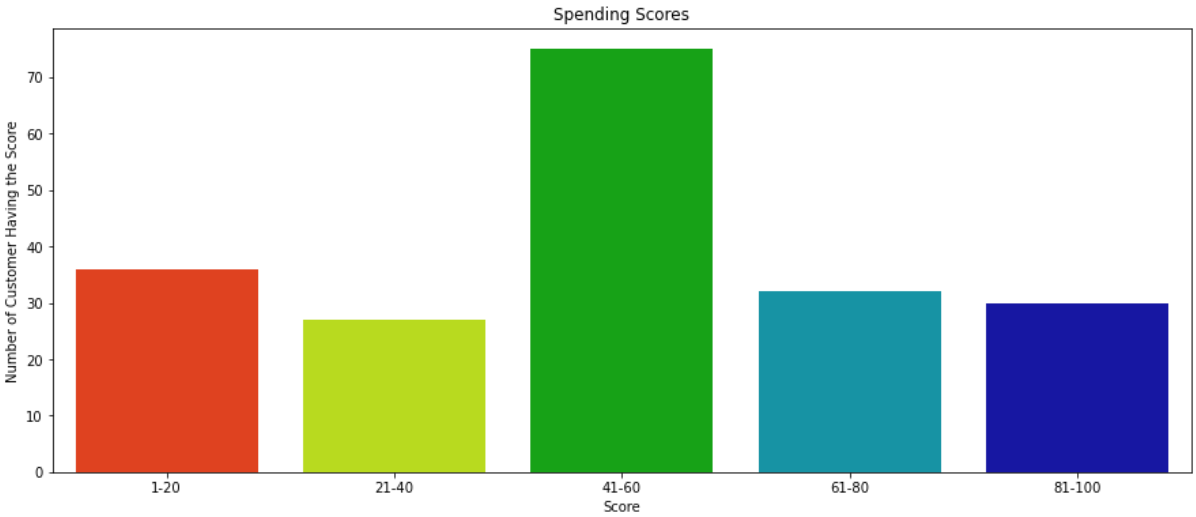
For better understanding, we made a box plot. It clearly says that the range of spending scores is more than the range of annual income.



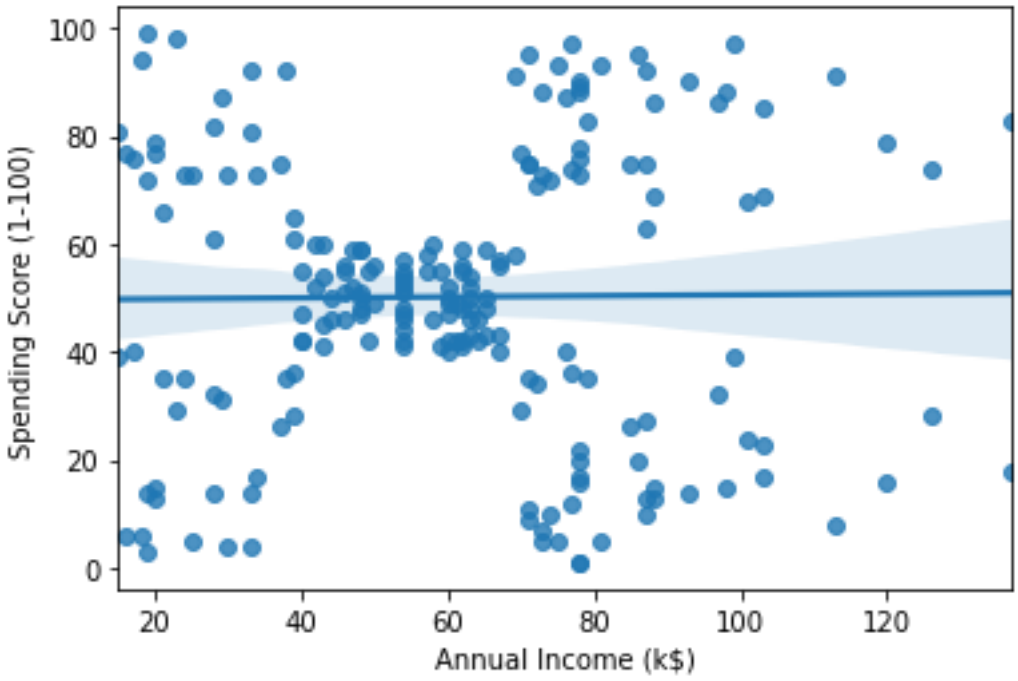
To check the distribution of the number of customers in each age group we made a box plot as shown in the below diagram. It clearly says that the age group from 26-35 outweighs every other age group.



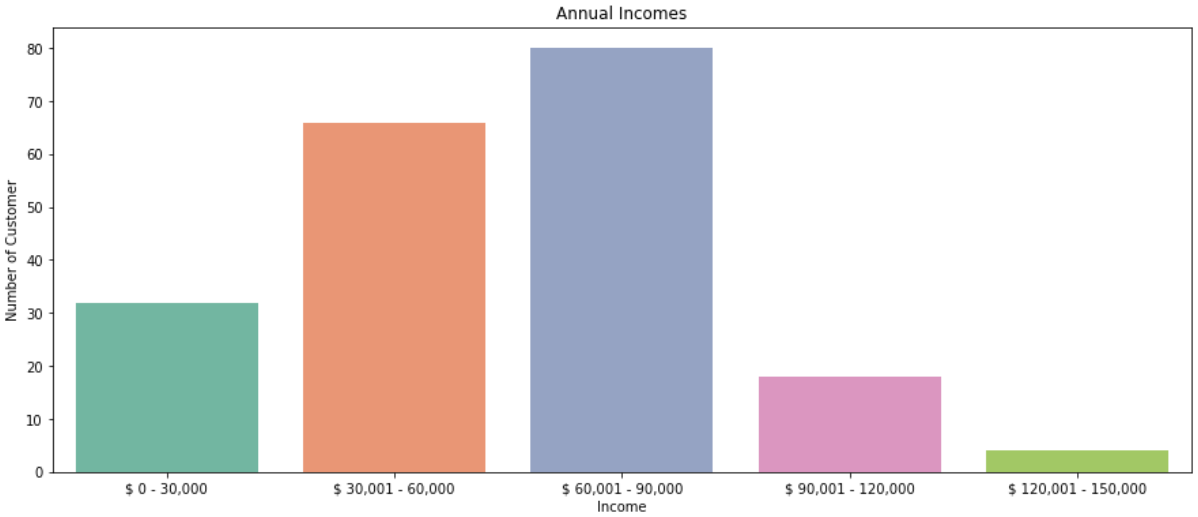
Next, we made a bar graph which visualizes the number of customers to their spending scores as shown in the below diagram. The majority of customers have spending scores in the range of 41-60.



To make it more clear, we made a regression plot as shown in the diagram below which clearly states that the customers with annual income of between 40k$ and 60k$ have a very typical spending score which lies somewhere between 40 and 60.



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



Calculation of clusters

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).

$$WCSS=\sum(X_i + Y_i)^2$$

where Yi is centroid for observation xi.

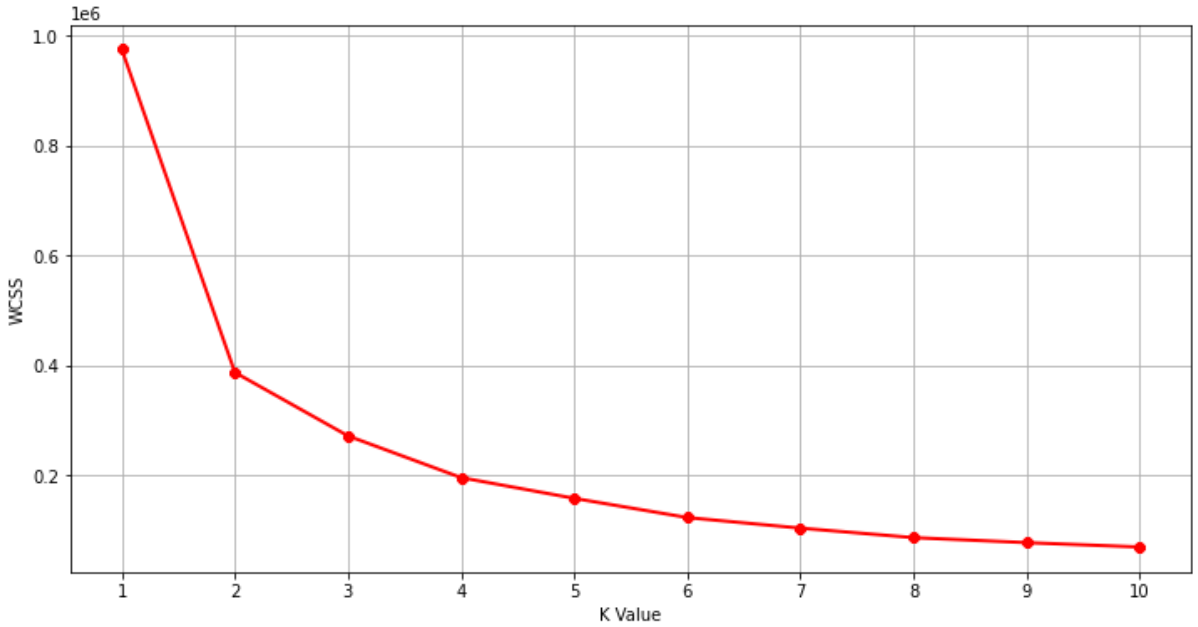
Here we use the elbow method for calculating the number of clusters i.e; K value.

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:

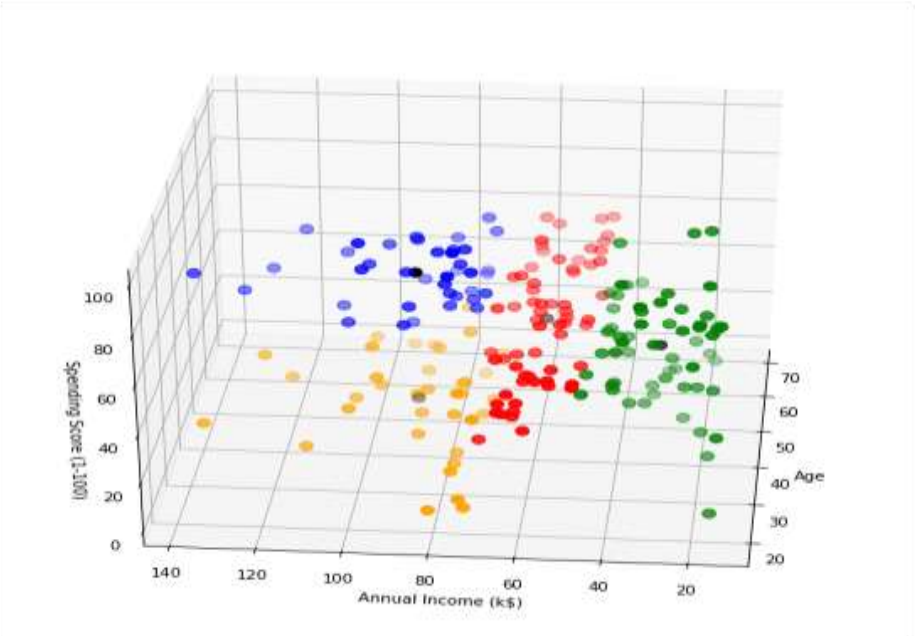
1. Compute K-Means clustering for different values of K by varying K from 1 to 10 clusters.
2. For each K, calculate the within-cluster sum of square (WCSS).
3. Plot the curve of WCSS versus the number of clusters(K).

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 4 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 4 classes which are represented in different colors as shown in the 3D plot.



Conclusion

For retaining arise in 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 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.