‘

PROJECT REPORT

Customer Segmentation using Kmeans

A PROJECT REPORT

submitted by

Nishad Chaoji

College – MIT Manipal
Branch – Computer and communication engineering

ACKNOWLEDGEMENT

At the very outset, I would like to give the first honors to the Almighty, who gave me the wisdom and knowledge to complete this project. We also express our gratitude to Advisor and Project Guide for providing us with adequate facilities, ways and means by which I was able to complete the project.

ABSTRACT

Today's business run based on such innovation having ability to captivate the customers with the products, but with such a large raft of products leave the customers confused about what to buy and what to not and the companies are confused about what section of customers to target to sell their products. This is where machine learning comes into play, various algorithms are applied to find out the hidden patterns in the data so that better decisions are made in the future. The process of segmenting the customers with similar behaviors into the same segment and with different patterns into different segments is called customer segmentation. A python program has been developed and the program has been trained by applying standard scaler onto a dataset having two features of 200 training sample taken from local retail shop. The dataset contains their Customer ID, Age, Gender, Annual Income and spending score. Segmentation is done based on the given data.

Table of Contents

INTRODUCTION

Management of customer relationship have always played a vital role to provide the organizations with strategy to build, manage and develop valuable long term customer relationships. The importance of treating customers as an organizations main asset is increasing in value in present day and era. Organizations nowadays must invest in the development of customer acquisition, maintenance and development strategies. This gives the businesses strategies that play vital role in gaining better customer knowledge and Programs for outreach. By using techniques like k-means clustering, customers with similar means are clustered together. This helps the marketing team to study and recognize different customer segments that think differently and make strategies that will attract the customers. Customer segmentation helps in figuring out the customers who vary in terms of preferences, expectations, desires and attributes. The main purpose of performing customer segmentation is to group people, who have similar interest so that the marketing team can converge in an effective marketing plan. Clustering is an iterative process of knowledge discovery from vast amounts of raw and unorganized data. Clustering is a type of exploratory data mining that is used in many applications, such as machine learning, classification and pattern and recognition.

EXISTING METHODS

Clustering Algorithm is one of the most popular algorithms used for customer segmentation

• Process of grouping a set of physical or abstract objects into clusters (example: customer, product etc.)

• A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters

• Similarity is calculated based distance between point

• Common distance measure is Euclidian distance

Types of clustering

1) Hierarchical Clustering

It is a method of unsupervised machine learning clustering where it begins with a pre-defined top to bottom hierarchy of clusters. It then proceeds to perform a decomposition of the data objects based on this hierarchy, hence obtaining the clusters. This method follows two approaches based on the direction of progress, i.e., whether it is the top-down or bottom-up flow of creating clusters. These are Divisive Approach and the Agglomerative Approach respectively.

2) Centroid Based Clustering

It is considered as one of the simplest clustering algorithms, yet the most effective way of creating clusters and assigning data points to it. The intuition behind centroid based clustering is that a cluster is characterized and represented by a central vector and data points that are near these vectors are assigned to the respective clusters.

3) Density Based Clustering

If one looks into the previous two methods that we discussed, one will observe that both hierarchical and centroid based algorithms are dependent on a distance (similarity/proximity) metric. The very definition of a cluster is based on this metric. Density-based clustering methods take density into consideration instead of distances. Clusters are considered as the densest region in a data space, which is separated by regions of lower object density, and it is defined as a maximal set of connected points.

4) Distribution-Based Clustering

Distribution-based clustering creates, and groups data points based on their likely hood of belonging to the same probability distribution (Gaussian, Binomial etc.) in the data.

Proposed Method with architecture

Here, for this project we are going to segment the customers using K-means algorithm(Centroid based clustering). K-Means clustering is an unsupervised machine learning algorithm that divides the given data into the given number of clusters. Here, the “K” is the given number of predefined clusters, that need to be created. This algorithm takes raw data as an input and divides the given dataset into clusters and the process is repeated until the best clusters are found.

We are going to form clusters depending on 4 major segmentation factors given in the data set i.e. Gender, Age, Annual income, Spending score

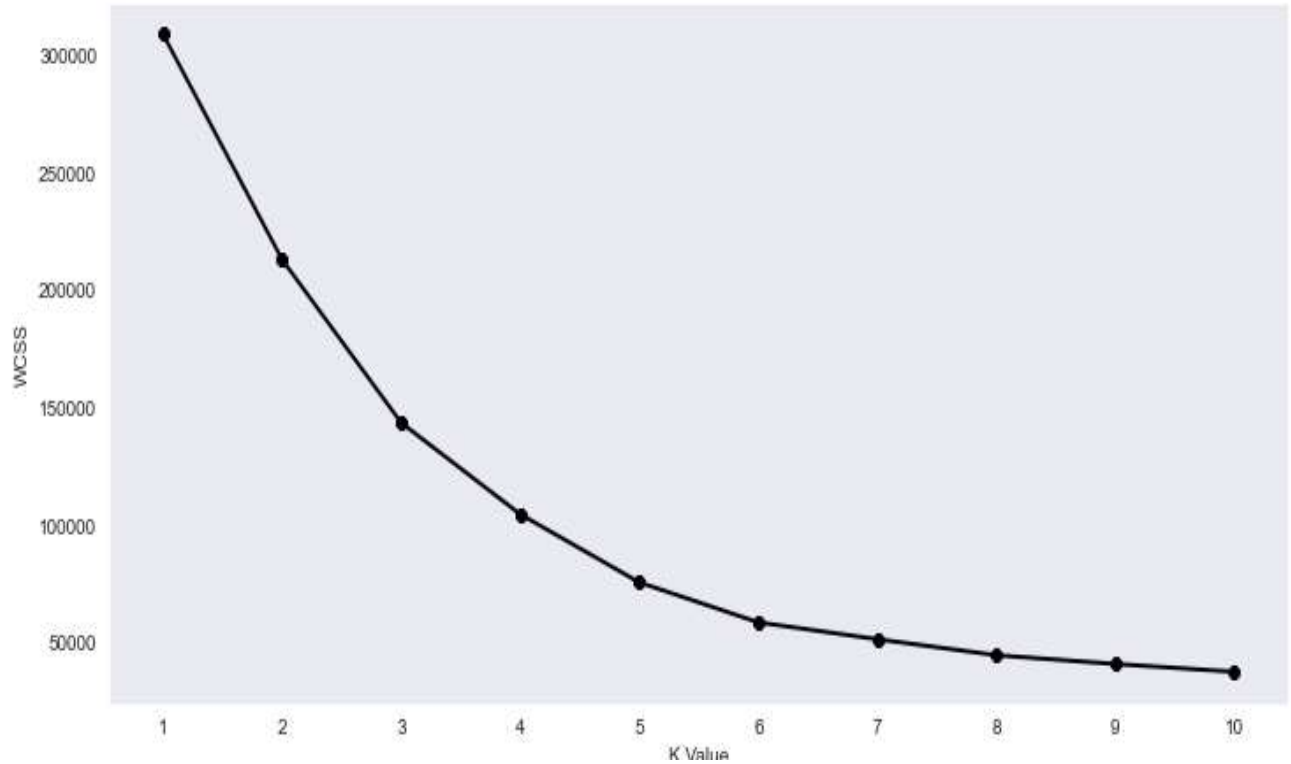
In this Algorithm we first must find the no of clusters which is found by using the elbow method where a plot is made b/w the value of K vs WCSS (Within sum of squares) .

Elbow Method

Elbow method is a tool used for analysing the clusters formed from our dataset and helps to

interpret the appropriate number of optimal clusters in dataset. From this method the optimal

number of clusters for our dataset is found to be five



Once we have the value of K or no of clusters we can simply Implement the Kmeans algorithm on the given data set and get the clusters

METHODOLOGY

Libraries used –

1) Pandas version 1.3.1
2) Numpy version 1.21.1
3) Matplotlib version 3.4.2
4) Sckit-learn version 0.24.2
5) Seaborn version 0.11.1

Clustering Algorithm – K-means Algorithm

1. Specify number of clusters i.e., *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e., assignment of data points to clusters isn't changing.

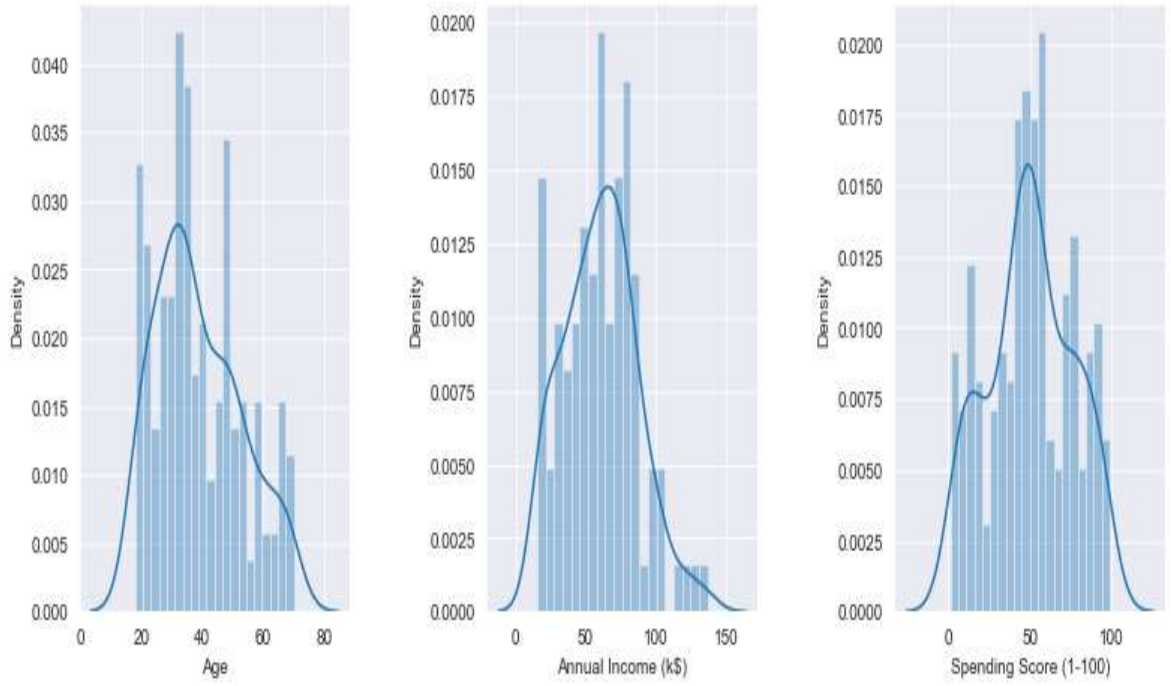
Dataset – Customer ID, Age, Gender, Annual Income and Spending Score.

IMPLEMENTATION

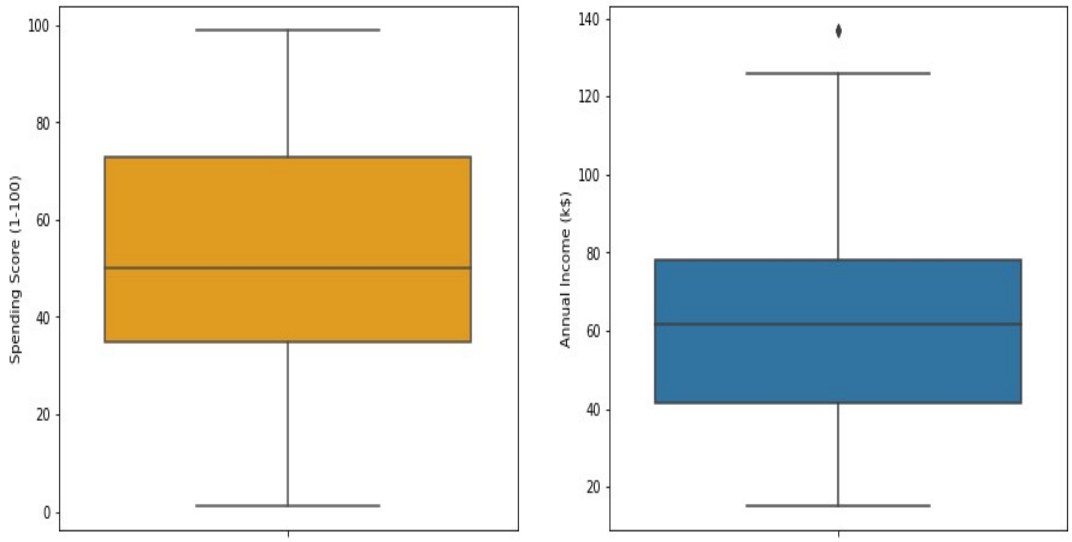
I started with loading all the libraries and dependencies and then imported the data. The columns in the dataset are customer id, gender, age, income and spending score.

I dropped the id column as that does not seem relevant to the context as we do not need the customer ID for segmentation.

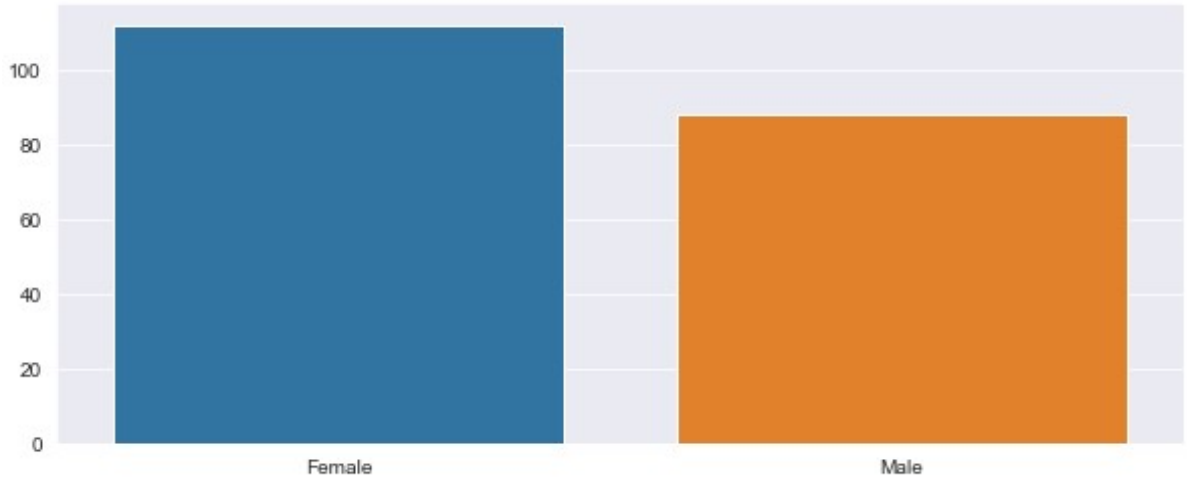
I plotted the Distplots for Age, Spending score and Annual income



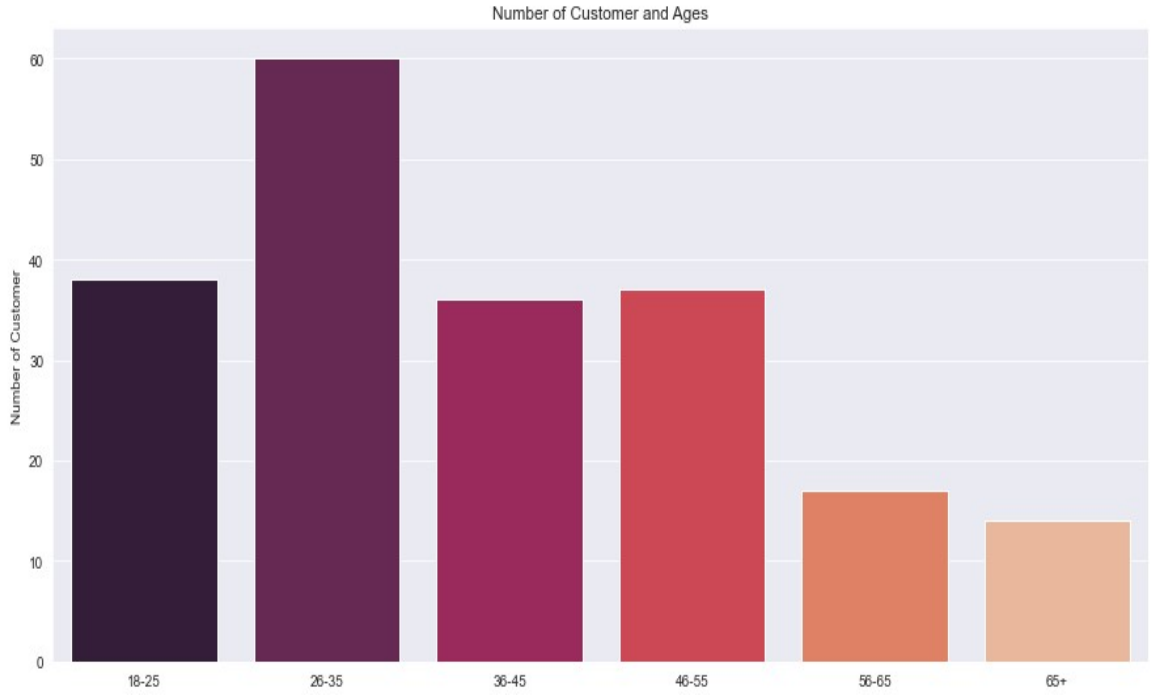
Next, I made a box plot of spending score and the annual income. This to better visualize the distribution range of the given dataset. The range of spending score is clearly more than the annual income range.



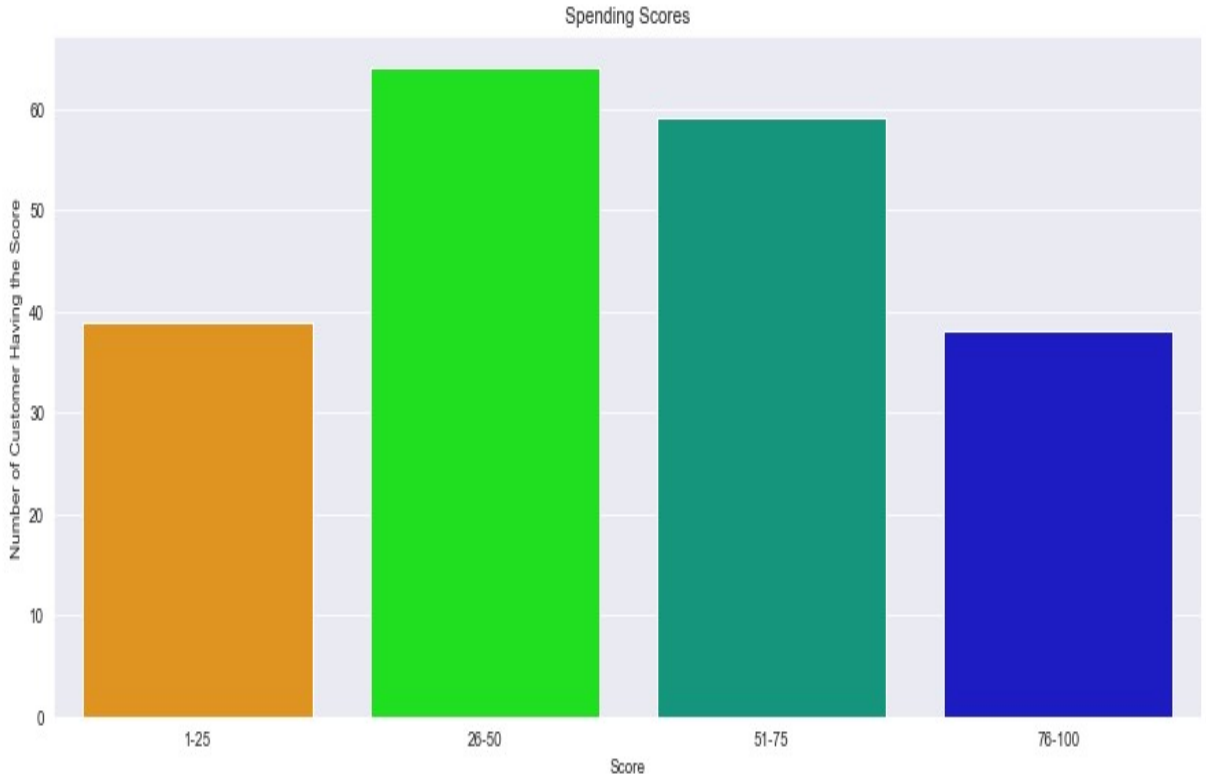
Next, I made a bar plot to check the distribution of male and female population. From the results the female population clearly outweighs the male counterpart.



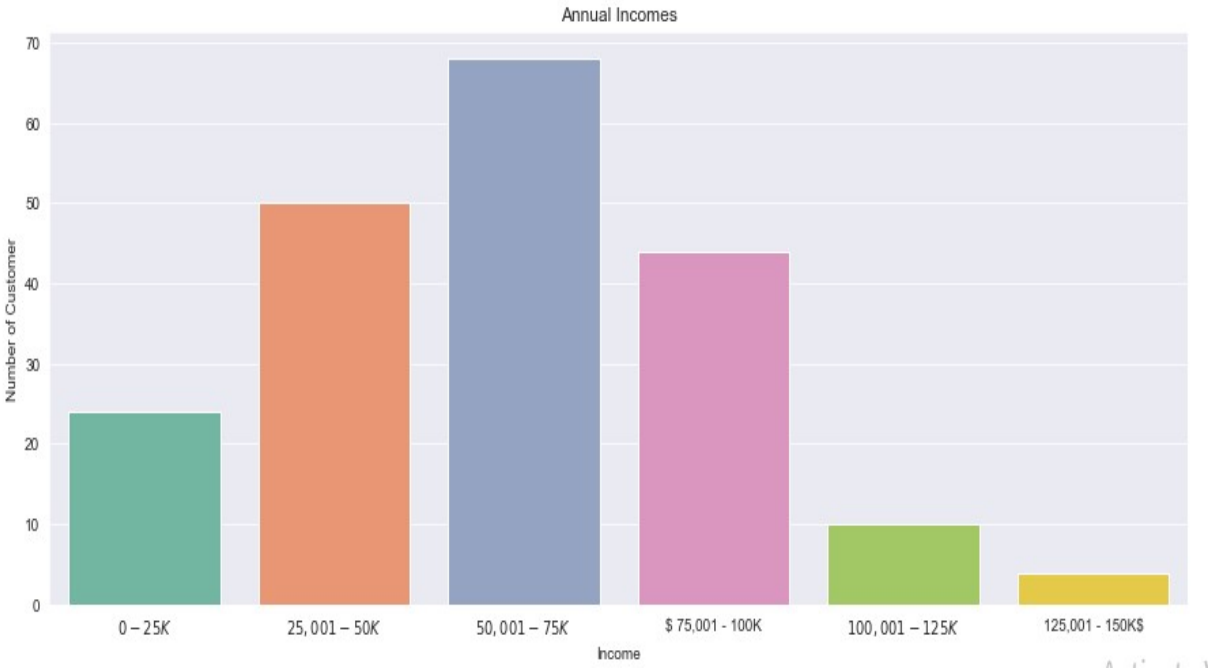
Next, I made a bar plot to check the distribution of number of customers in each age group. Clearly the age group of 26-35 has the highest no of customers.



Next we make a bar plot to visualize the number of customers according to their spending scores. Most of the customers have spending score in the range 26-50 closely followed by 51-75.



Next, I made a bar plot to visualize the number of customers according to their annual income. Many of the customers have annual income in the range 50,001 to 75,000.



Next, I plotted WCSS against K Value to find out the optimal value of K. WCSS measures sum of distances of observations from their cluster centroids which is given by the below formula.

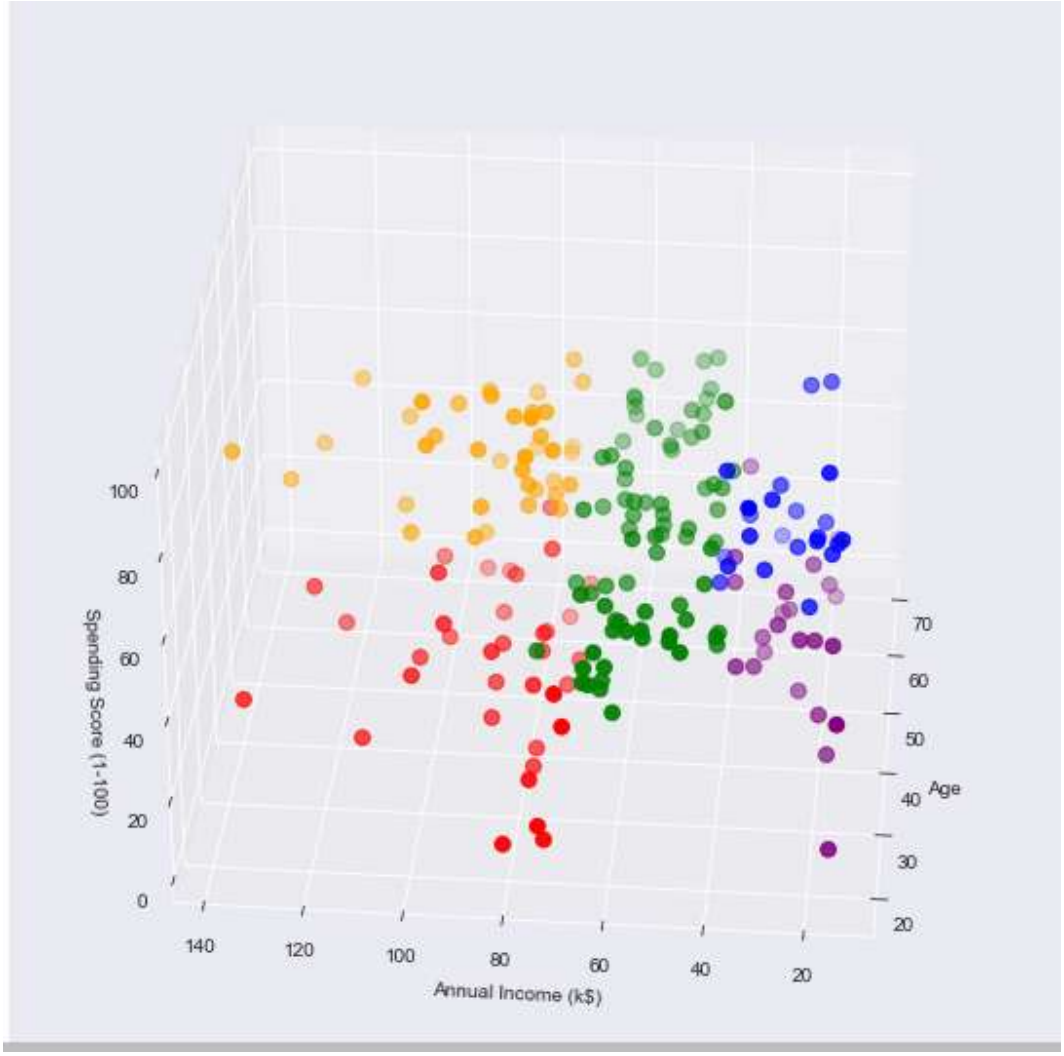
WCSS – Within clusters sum of squares

K- No of clusters

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

where *Yi* is centroid for observation *Xi*. The main goal here is to maximize number of clusters to the point where each clusters centroids doesn't move anymore . We calculated the Value of K using the elbow method

Lastly, I made a 3D plot to visualize the spending score 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

Cluster 1 (Blue Colour) -> earning low but spending high

Cluster 2 (Red Colour) -> Earn more and spend less

Cluster 3 (Orange Colour) -> earning high and spending high [TARGET Audience]

Cluster 4 (Purple Colour) -> earning less and spend less

Cluster 5 (Green Colour) -> Average in terms of spending and Earning

Learning Outcomes

1) What is Data Analysis and why is it so Important

In today's world.

2) what is customer segmentation and why is it important

3) How machine learning helps with Customer

Segmentation

4) Practical Implementation of K-means Algorithm