Exposys Data Labs

Customer Segmentation Model

Data Science - ML Project

By: Aman Kumar

Domain: Data Science

Institute: Birla Institute of Technology, Mesra

Abstract

Customer segmentation is a separation of a market into multiple distinct groups of consumers who share the similar characteristics. Segmentation of market is an effective way to define and meet   
customer needs. Unsupervised Machine Learning Techniques, KMeans Clustering Algorithm, Minibatch K-Means and Hierarchical Clustering are used to perform Market Basket Analysis. Market Basket Analysis is carried out to predict the target customers who can be easily converged, among all the customers. In order to allow the marketing team to plan the strategy to market the new products to the target customers which are similar to their interests.

*Keywords: Target Customers, Clusters, Unsupervised Learning, KMeans, Minibatch K-Means,   
Hierarchical Clustering Segmentation, Market Basket Analysis.*

Contents

1. Introduction
2. Unsupervised machine learning.
3. Dataset
4. Methodology
5. Code
6. Output
7. Conclusion

Introduction

Management and maintain of customer relationship have always played a vital role to provide business intelligence to organizations 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 have an interest to invest in the development of customer acquisition, maintenance and development strategies. The business intelligence has a vital role to play in allowing companies to use technical expertise to gain better customer knowledge and Programs for outreach. By using clustering techniques like k-means, customers with similar means are clustered together. Customer segmentation helps the marketing team to recognize and expose different customer segments that think differently and follow different purchasing strategies. 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   
recognition.

# Unsupervised Machine Learning

There may be many cases in which we do not have labeled data and need to find the hidden patterns from the given dataset. So, to solve such types of cases in machine learning, we need unsupervised learning techniques.

## What is Unsupervised Learning?

Unsupervised learning is a machine learning technique in which models are not supervised using training dataset. Instead, models itself find the hidden patterns and insights from the given data. It can be compared to learning which takes place in the human brain while learning new things. It can be defined as: Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision.

Unsupervised learning cannot be directly applied to a regression or classification problem because unlike supervised learning, we have the input data but no corresponding output data. The goal of unsupervised learning is to **find the underlying structure of dataset, group that data according to similarities, and represent that dataset in a compressed format.**

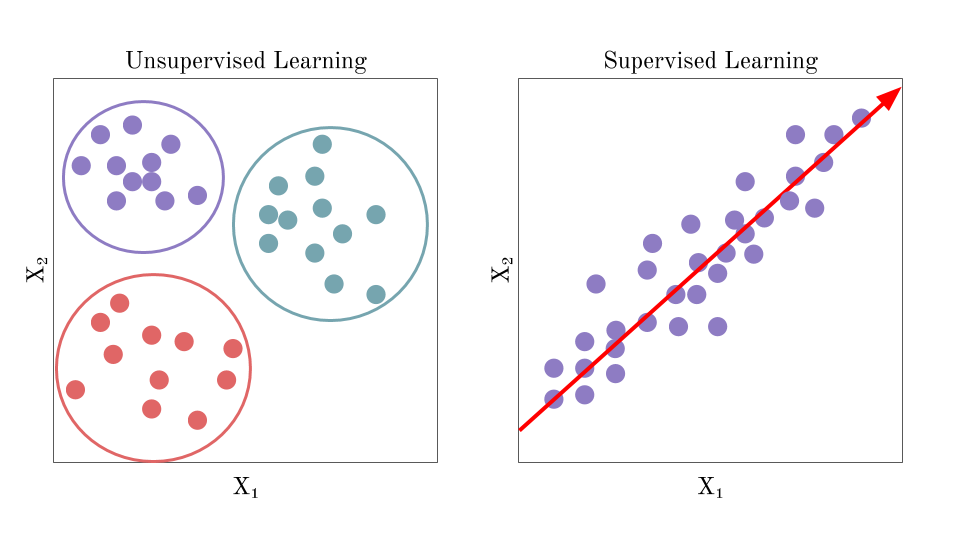
**Example:** Suppose the unsupervised learning algorithm is given an input dataset containing images of different types of cats and dogs. The algorithm is never trained upon the given dataset, which means it does not have any idea about the features of the dataset. The task of the unsupervised learning algorithm is to identify the image features on their own. Unsupervised learning algorithm will perform this task by clustering the image dataset into the groups.

## Why use Unsupervised Learning?

The importance of Unsupervised Learning:

* Unsupervised learning is helpful for finding useful insights from the data.
* Unsupervised learning is much similar as a human learns to think by their own experiences, which makes it closer to the real AI.
* Unsupervised learning works on unlabeled and uncategorized data which make unsupervised learning more important.
* In real-world, we do not always have input data with the corresponding output so to solve such cases, we need unsupervised learning.

## Working of Unsupervised Learning:

Working of unsupervised learning can be understood by the below diagram:

Here, we have taken an unlabelled input data, which means it is not categorized and corresponding outputs are also not given. Now, this unlabeled input data is fed to the machine learning model in order to train it. Firstly, it will interpret the raw data to find the hidden patterns from the data and then will apply suitable algorithms such as k-means clustering, Decision tree, etc.

Once it applies the suitable algorithm, the algorithm divides the data objects into groups according to the similarities and difference between the objects.

## Types of Unsupervised Learning Algorithm:

The unsupervised learning algorithm can be further categorized into two types of problems:

* **Clustering**: Clustering is a method of grouping the objects into clusters such that objects with most similarities remains into a group and has less or no similarities with the objects of another group. Cluster analysis finds the commonalities between the data objects and categorizes them as per the presence and absence of those commonalities.
* **Association:** An association rule is an unsupervised learning method which is used for finding the relationships between variables in the large database. It determines the set of items that occurs together in the dataset.

Association rule makes marketing strategy more effective. Such as people who buy X item (suppose a bread) are also tend to purchase Y (Butter/Jam) item. A typical Example: Market Basket Analysis.

## Unsupervised Machine Learning algorithms:

Types of Unsupervised learning algorithms:

* **K-means clustering**
* **KNN (k-nearest neighbors)**
* **Hierarchal clustering**
* **Anomaly detection**
* **Neural Networks**
* **Principle Component Analysis**
* **Independent Component Analysis**
* **Apriori algorithm**
* **Singular value decomposition**

Data Set

We are going to aim to cluster a data set that is about behaviour of the customers having credit card using many unsupervised algorithms.

**Our research question is "How many clusters can we distinguish the customers according to their transactions or behaviors?”**

1. Dataset Description- The data is gathered from Exposys Datalabs which is named as Mall\_Customers.csv Dataset. The dataset have many attributes of 200 customers.

|  |  |
| --- | --- |
| S No. | Attributes |
| 1 | CustomerID |
| 2 | Gender |
| 3 | Age |
| 4 | Annual Income |
| 5 | Spending Score (1-100) |

Methodology

**Clustering**

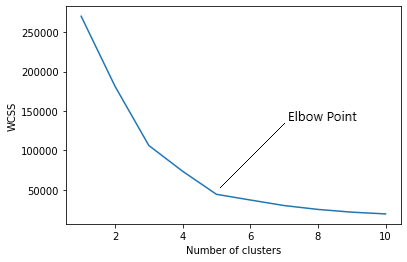
Clustering is one of the most common methods used in exploring data to obtain a clear understanding of the data structure. It can be characterized as the task of finding the subtitles and subgroups in the complete dataset. Similar data is clustered in many subgroups.

A cluster refers to a collection of aggregated data points due to some similarities. Clustering is used in Market basket analysis used to segment the customers based on their behaviors and transactions.

**K Means Clustering Algorithm**K Means Clustering is the most common and simplest Machine learning algorithm and it follows   
an iterative approach which attempts to partition the dataset into different “k” number of   
predefined and non-overlapping subgroups where each data point belongs to only one subgroup   
according to their similar qualities.



**Elbow Method**   
Elbow method is a tool used for analyzing 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 seven.



Code

port numpy as np

import pandas as pd

import matplotlib.pyplot as pl

import os

df = pd.read\_csv("/Users/aman0/Documents/Cust\_Seg/Mall\_Customers.csv")

df.info()

df.drop(["CustomerID"], axis = 1, inplace=True)

df.columns

pl.figure(figsize=(16,6))

pl.title("Age's Frequency")

pl.hist(df['Age'], bins=50)

pl.xlabel('Age')

pl.ylabel('Frequency')

pl.xticks([18,27,36,45,54,63,70])

pl.show()

pl.figure(figsize =(16,6))

pl.subplot(1,2,1)

pl.boxplot(df["Spending Score (1-100)"],notch=True, showmeans=True, meanline=True)

pl.xlabel('Spending Score')

pl.grid(linestyle = '--')

pl.subplot(1,2,2)

pl.boxplot(df["Annual Income (k$)"],notch=True, showmeans=True, meanline=True)

pl.xlabel('Annual Income (k$)')

pl.grid(linestyle = '--')

pl.show()

genders = df.Gender.value\_counts()

pl.figure(figsize=(16,6))

x= genders.index

y= genders.values

pl.bar(x,y,color =['red','blue'])

pl.show()

age18\_26 = df.Age[(df.Age <= 26) & (df.Age >= 18)]

age27\_35 = df.Age[(df.Age <= 35) & (df.Age >= 27)]

age36\_44 = df.Age[(df.Age <= 44) & (df.Age >= 36)]

age45\_53 = df.Age[(df.Age <= 53) & (df.Age >= 45)]

age53above = df.Age[df.Age >= 54]

x = ["18-26","27-35","36-44","45-53","53+"]

y = [len(age18\_26.values),len(age27\_35.values),len(age36\_44.values),len(age45\_53.values),len(age53above.values)]

pl.figure(figsize=(16,6))

pl.bar(x, y)

pl.title("Number of Customer and Ages")

pl.xlabel("Age")

pl.ylabel("Number of Customer")

pl.show()

ss1\_15 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 1) & (df["Spending Score (1-100)"] <= 15)]

ss16\_30 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 16) & (df["Spending Score (1-100)"] <= 30)]

ss31\_45 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 31) & (df["Spending Score (1-100)"] <= 45)]

ss46\_60 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 46) & (df["Spending Score (1-100)"] <= 60)]

ss61\_75 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 61) & (df["Spending Score (1-100)"] <= 75)]

ss76\_90 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 76) & (df["Spending Score (1-100)"] <= 90)]

ss91\_100 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 91) & (df["Spending Score (1-100)"] <= 100)]

ssx = ["1-15", "16-30", "31-45", "46-60", "61-75", "76-90", "91-100"]

ssy = [len(ss1\_15.values), len(ss16\_30.values), len(ss31\_45.values), len(ss46\_60.values), len(ss61\_75.values)

       , len(ss76\_90.values), len(ss91\_100.values)]

pl.figure(figsize=(16,6))

pl.bar(ssx,ssy)

pl.title("Spending Scores")

pl.xlabel("Score")

pl.ylabel("Number of Customer Having the Score")

pl.show()

ai0\_30 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 0) & (df["Annual Income (k$)"] <= 30)]

ai31\_60 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 31) & (df["Annual Income (k$)"] <= 60)]

ai61\_90 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 61) & (df["Annual Income (k$)"] <= 90)]

ai91\_120 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 91) & (df["Annual Income (k$)"] <= 120)]

ai121\_150 = df["Annual Income (k$)"][(df["Annual Income (k$)"] >= 121) & (df["Annual Income (k$)"] <= 150)]

aix = ["$ 0 - 30,000", "$ 30,001 - 60,000", "$ 60,001 - 90,000", "$ 90,001 - 120,000", "$ 120,001 - 150,000"]

aiy = [len(ai0\_30.values), len(ai31\_60.values), len(ai61\_90.values), len(ai91\_120.values), len(ai121\_150.values)]

pl.figure(figsize=(15,6))

pl.bar(aix,aiy)

pl.title("Annual Incomes")

pl.xlabel("Income")

pl.ylabel("Number of Customer")

pl.show()

from mpl\_toolkits.mplot3d import Axes3D

fig = pl.figure(figsize=(32,10))

ax = fig.add\_subplot(111, projection='3d')

ax.scatter(df.Age, df["Annual Income (k$)"], df["Spending Score (1-100)"], c='red', s=60)

ax.view\_init(30, 185)

pl.xlabel("Age")

pl.ylabel("Annual Income (k$)")

ax.set\_zlabel('Spending Score (1-100)')

pl.show()

from sklearn.cluster import KMeans

wcss = []

for k in range(1,11):

    kmeans = KMeans(n\_clusters=k, init="k-means++")

    kmeans.fit(df.iloc[:,1:])

    wcss.append(kmeans.inertia\_)

pl.figure(figsize=(16,6))

pl.grid()

pl.plot(range(1,11),wcss, linewidth=2, color="cyan", marker ="8")

pl.xlabel("K-Value")

pl.xticks(np.arange(1,11,1))

pl.ylabel("WCSS")

pl.show()

km = KMeans(n\_clusters=6)

clusters = km.fit\_predict(df.iloc[:,1:])

df["label"] = clusters

from mpl\_toolkits.mplot3d import Axes3D

import matplotlib.pyplot as pl

import numpy as np

import pandas as pd

fig = pl.figure(figsize=(20,11))

ax = fig.add\_subplot(111, projection='3d')

ax.scatter(df.Age[df.label == 0], df["Annual Income (k$)"][df.label == 0], df["Spending Score (1-100)"][df.label == 0], c='blue', s=60)

ax.scatter(df.Age[df.label == 1], df["Annual Income (k$)"][df.label == 1], df["Spending Score (1-100)"][df.label == 1], c='red', s=60)

ax.scatter(df.Age[df.label == 2], df["Annual Income (k$)"][df.label == 2], df["Spending Score (1-100)"][df.label == 2], c='green', s=60)

ax.scatter(df.Age[df.label == 3], df["Annual Income (k$)"][df.label == 3], df["Spending Score (1-100)"][df.label == 3], c='orange', s=60)

ax.scatter(df.Age[df.label == 4], df["Annual Income (k$)"][df.label == 4], df["Spending Score (1-100)"][df.label == 4], c='purple', s=60)

ax.scatter(df.Age[df.label == 5], df["Annual Income (k$)"][df.label == 5], df["Spending Score (1-100)"][df.label == 5], c='cyan', s=60)

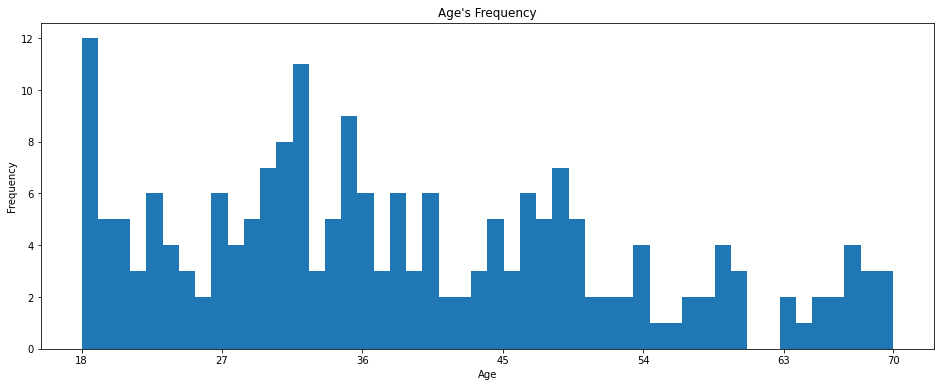
ax.view\_init(30, 185)

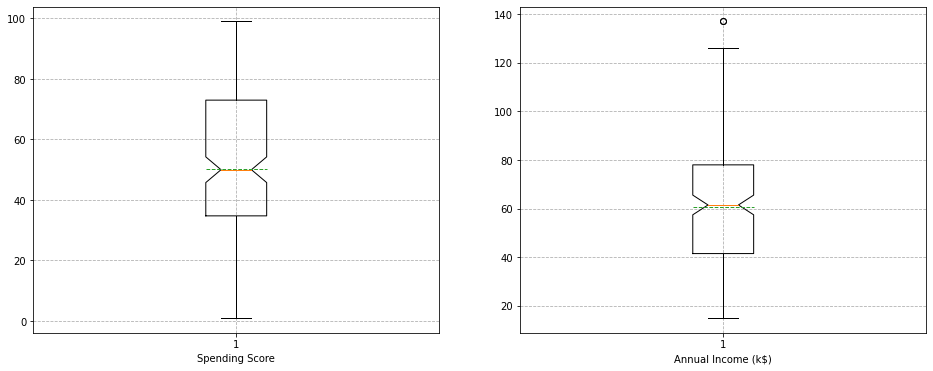
pl.xlabel("Age")

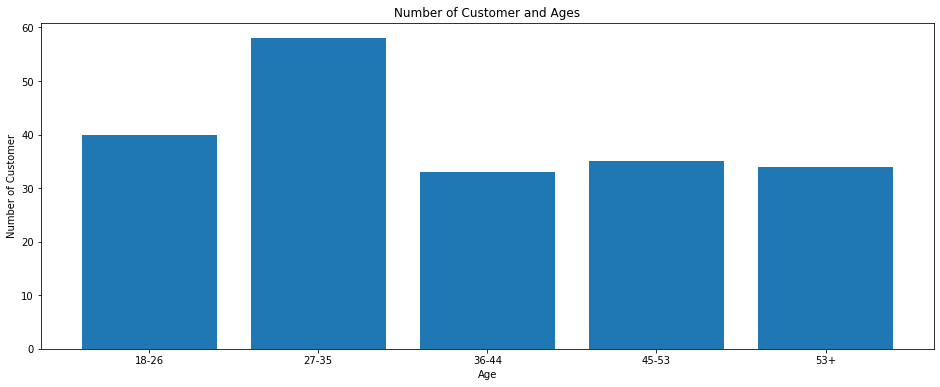
pl.ylabel("Annual Income (k$)")

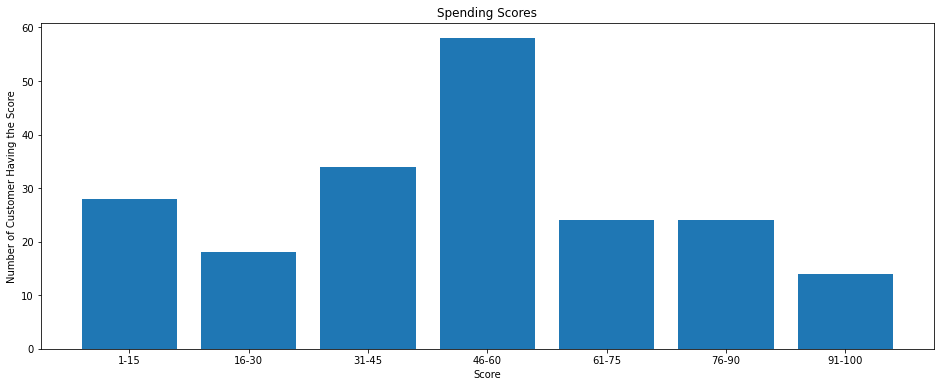
ax.set\_zlabel('Spending Score (1-100)')

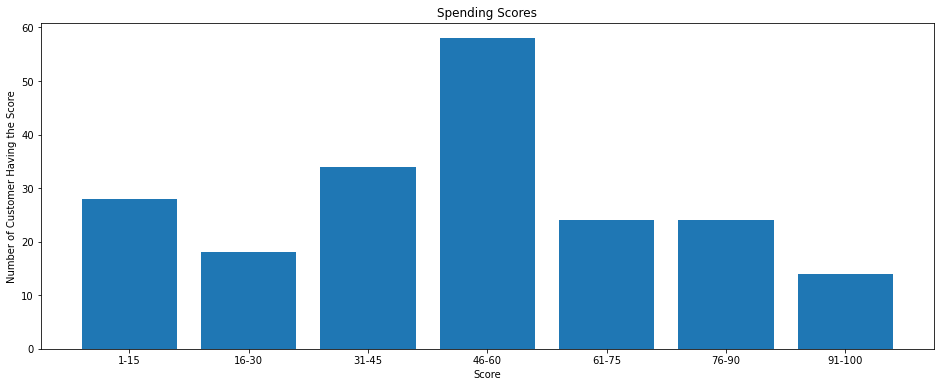
pl.show()

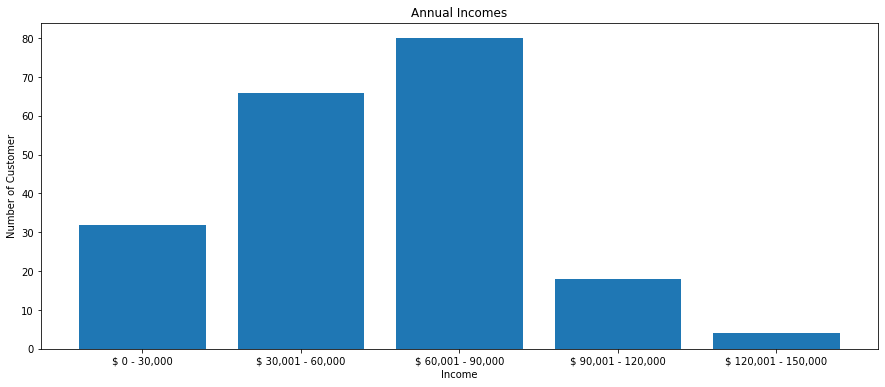
Output

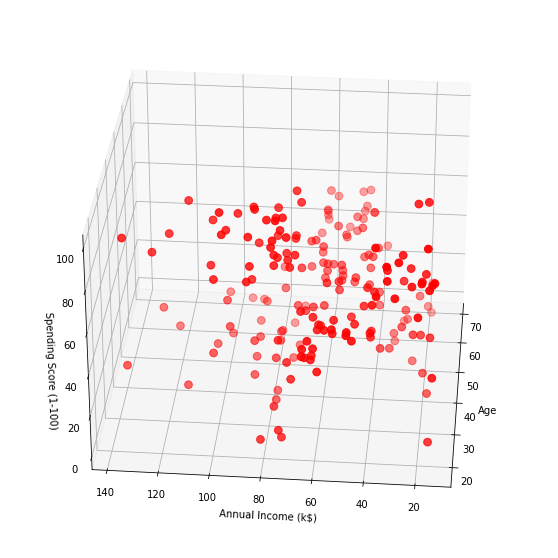


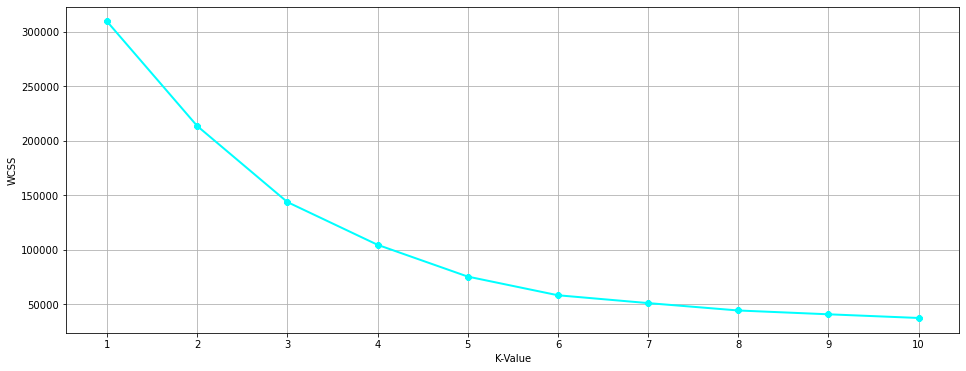




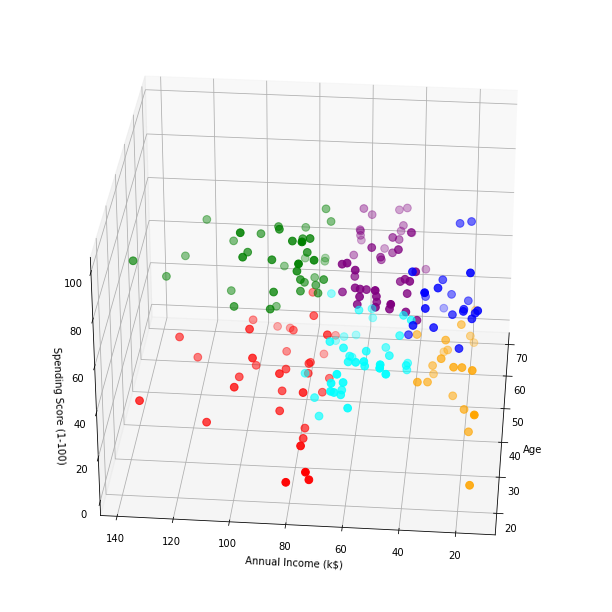








Elbow Method



K -Means Clustering

CONCLUSION

* The main aim of this project was to design and implement Customer Segmentation Using Machine Learning Methods and Performance Analysis of that methods and it has been achieved successfully.
* The proposed approach uses various methods and ensemble learning method in which K-means Clustering is used.
* The Experimental results can be used to analyze different groups of Customer dataset present.