Project Report

On

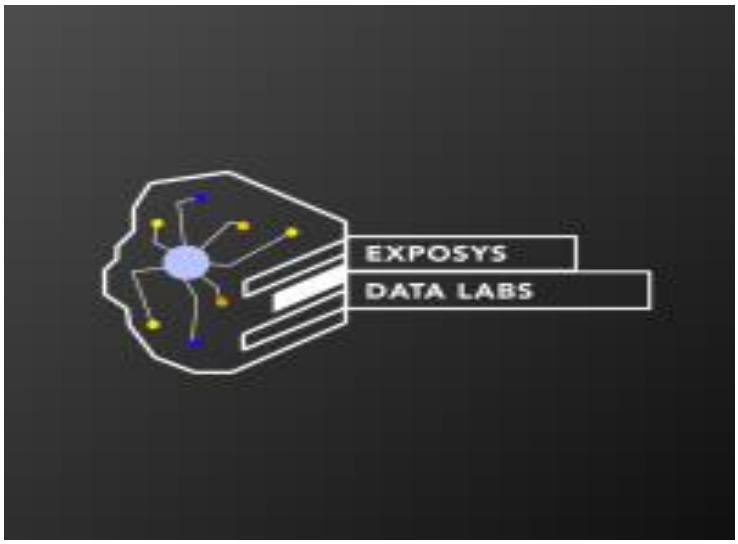
“Customer Segmentation (Using K-Means)”

Submitted by

Name: Nayan Hitesh Kacha

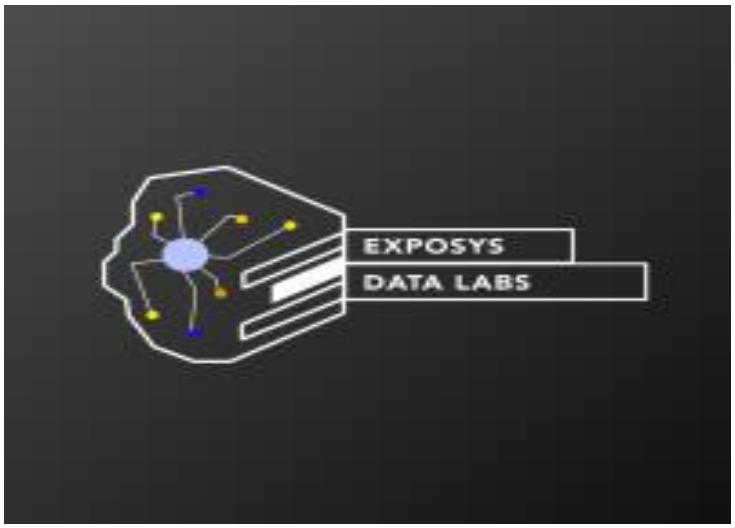
Domain: Data Science

To



EXPOSYS DATA LABS

Bengaluru-India



CERTIFICATE

This is to certify that the seminar report entitled **"Customer Segmentation"** being submitted by **Nayan Hitesh Kacha** is a record of bonafide work carried out by him/her under the supervision and guidance of **Exposys Data Labs** in partial fulfillment of the requirement for **TE (Information Technology Engineering)** in the year 2021-22

He has been successfully Done **Data Science Project report** on the topic **Customer Segmentation**. He has work in very good manner.

Keep it up!!!!

Date:

Place: Pune

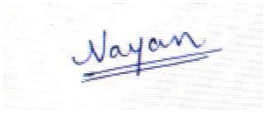
(EXPOSYS DATA LABS)

ACKNOWLEDGEMENT

I am extremely grateful and remain indebted to my Company **"EXPOSYS DATA LABS"** for being a source of inspiration and for her constant support in the Design, Implementation and Evaluation of the project. I am thankful to her for constant constructive criticism and invaluable suggestions, which benefited me a lot while developing the project on **"CUSTOMER SEGMENTATION",** He has been a constant source of inspiration and motivation for hard work, and he has been very co-operative throughout this project work. With candor and pleasure I take opportunity to express my sincere thanks and obligation to **EXPOSYS DATA LABS TEAM** Through this column, it would be my utmost pleasure to express my warm thanks to him for the encouragement, co-operation and consent without which we mightn't be able to accomplish this project.

My thanks and appreciations also go to my team members developing the project and people who have willingly helped me out with their abilities.

Finally, I gratefully acknowledge the support, encouragement & patience of my family, and as always, nothing in my life would be possible without God, Thank You!



(Nayan Kacha)

ABSTRACT

In this project, we will perform one of the most essential applications of machine learning, Customer Segmentation by using K-Means Clustering Algorithm. In this project, we will implement customer segmentation in Python. Whenever you need to find your best customer, customer segmentation is the ideal methodology. Then we will explore the data upon which we will be building our segmentation model. Also, in this data science project, we will see the descriptive analysis of our data and then implement several versions of the K-means algorithm.

Furthermore, through the data collected, we can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, we can strategize the marketing techniques more efficiently and minimize the possibility of risk to the investment.

The business done today runs on the basis of innovative ideas as there are large number of potential customers who are confounded to what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers. This is where the machine learning comes into picture, the various algorithms are applied to identify the hidden patterns in the data for better decision making. The concept of which customer segment to target is done using the customer segmentation process using the clustering technique. In this paper, the clustering algorithm used is K-means algorithm which is the partitioning. so that objects within a cluster are similar to one another and dissimilar to objects in other clusters.

CONTENTS

Title	Page no.
Certificate	2
Acknowledge	3
Abstract	4

INDEX

Sr no.	Chapters	Page no
1	INTRODUCTION 1.1. Dataset Description 1.2. Purpose 1.3. Objectives	6
2	EXISTING METHODOLOGY 2.1. Jupyter Notebook 2.2. Pandas 2.3. Numpy 2.4. Matplotlib 2.5. Scikit Learn 2.6. Seaborn	8
3	LITERATURE REVIEW	10
4	IMPLEMENTATION 3.1. What is Clustering? 3.2. K-Means Clustering 3.3. Modeling 3.4. Elbow Method	12
5	ANALYSIS 4.1. Cluster Analysis	26
6	ADVANTAGES	28
7	FUTURE SCOPE & CONCLUSION	30
8	RESULT	31
9	REFERENCE	32

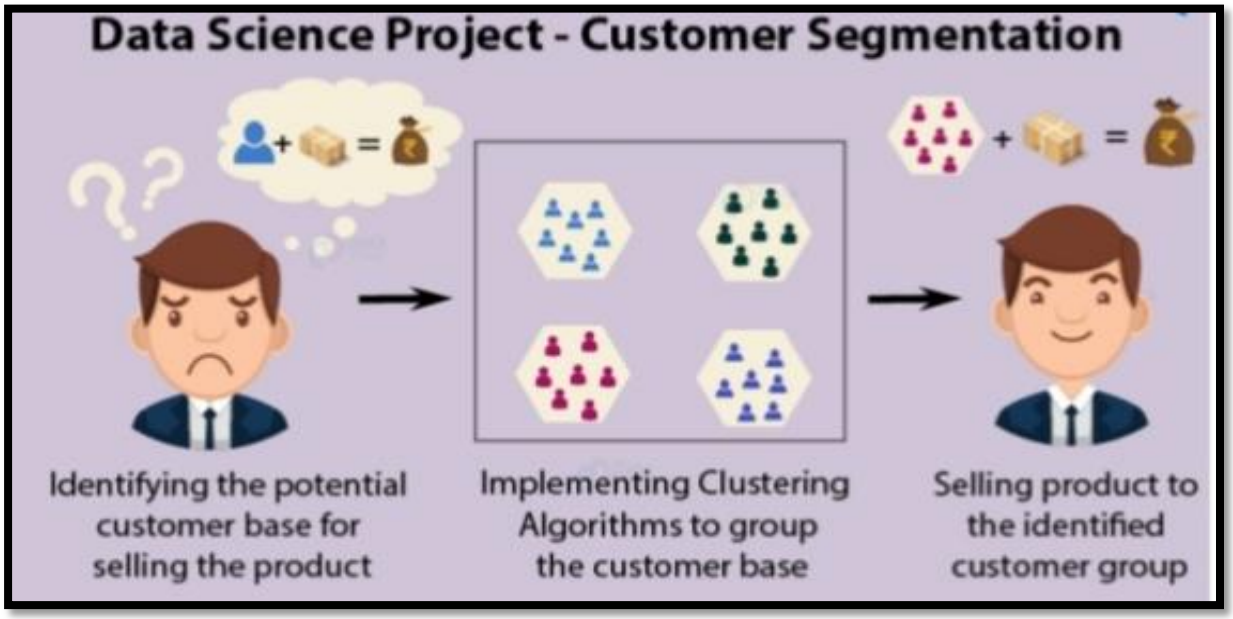
CHAPTER .1

INTRODUCTION

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. In the first step of this data science project, we will perform data exploration. We will import the essential packages required for this role and then read our data. Finally, we will go through the input data to gain necessary insights about it.

Dataset Description: Mall Customer Segmentation Data: The data is given by Exposys Data Labs. It has individual unique customer IDs, A categorical variable in the form of Gender and three columns of Age, Annual Income and Spending Score which will be our main targets to identify the patterns in the customers shopping and spending spree.

Data URL – drive.google.com/file/d/19BOhwz52NUY3dg8XErVYglctpr5sjTy4



The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decision; identification of products associated with

each customer segment and to manage the demand and supply of that product; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions.

The thrust of this paper is to identify customer segments using the data mining approach, using the partitioning algorithm called as K-means clustering algorithm. The elbow method determines the optimal clusters.

Clustering algorithms generates clusters such that within the clusters are similar based on some characteristics. Similarity is defined in terms of how close the objects are in space.

Purpose: To find the best customer, using customer segmentation methodology. To explore the data upon which building a segmentation model. Also, in this project, we will see the descriptive analysis of our data and then implement the K-means algorithm.

Objectives: The objective of the project is as follows: • Identify the potential customer base for selling the product. • Implement Clustering Algorithms to group the customer base.

Customer Segmentation & Why it Matters

At its most basic, customer segmentation (also known as market segmentation) is the division of potential customers in a given market into discrete groups. That division is based on variables and descriptors of those customers having similar enough:

1. Needs, i.e., so that a single whole product can satisfy them.

2. Buying characteristics, i.e., responses to messaging, marketing channels, and sales channels, that a single go-to-market approach can be used to sell to them competitively and economically.

CHAPTER .2

METHODOLOGY

The data set used to implement clustering and K-Means algorithm was collected from a store of shopping mall. The data set contains 5 attributes and has 200 tuples, representing the data of 200 customers. The attributes in the data set has Customer Id, gender, age, annual income (k$), spending score on the scale of (1-100).

In this project I have used Jupyter Notebook as a platform for coding.

Jupyter Notebook:

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

In our project we used following packages:

- Pandas (version: 1.1.5)
- Numpy (version: 1.19.2)
- Matplotlib (version: 3.3.2)
- Scikit Learn (version: 0.23.2)
- Seaborn (version: 0.11.1)

Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals. Its name is a play on the phrase "Python data analysis" itself. Wes McKinney started building what would become pandas at AQR Capital while he was a researcher there from 2007 to 2010

Numpy: NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more. At the core of the Numpy package, is the nd array object. This encapsulates n dimensional arrays of homogeneous data types, with many operations being performed in compiled code for performance.

Matplotlib: Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension Numpy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinker, Wxpython, Qt, or GTK+. There is also a procedural "pylab" interface based on a state machine (like OpenGL), designed to closely resemble that of MATLAB, though its use is discouraged. SciPy makes use of Matplotlib.

Scikit Learn: Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, and clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

Seaborn: Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions operate on data frames and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

CHAPTER.3

LITERATURE REVIEW

Over the years, as there is very strong competition in the business world, the organizations have to enhance their profits and business by satisfying the demands of their customers and attract new customers according to their needs. The identification of customers and satisfying the demands of each customer is a very complex and tedious task. This is because customers may be different according to their demands, tastes, and preferences and so on. Instead of "one-size-fits-all" approach, customer segmentation clusters the customers into groups sharing the same properties or behavioral characteristics. According to,[5] customer segmentation is a strategy of dividing the market into homogenous groups. The data used in customer segmentation technique that divides the customers into groups depends on various factors like, data geographical conditions, economic conditions, demographical conditions as well as behavioral patterns. The customer segmentation technique allows the business to make better use of their marketing budgets, gain a competitive edge over their rival companies, demonstrating the better knowledge of the needs of the customer. It also helps an organization in, increasing their marketing efficiency, determining new market opportunities, making better brand strategy, identifying customer's retention.

Clustering algorithms generates clusters such that within the clusters are similar based on some characteristics. Similarity is defined in terms of how close the objects are in space. K-means algorithm in one of the most popular centroid based algorithm. Suppose data set, D, contains n objects in space. Partitioning methods distribute the objects in D into k clusters, C1,...,Ck , that is, $Ci \subset D$ and $Ci \cap Cj = \emptyset$ for $(1 \leq i, j \leq k)$. A centroid-based partitioning technique uses the centroid of a cluster, Ci , to represent that cluster. Conceptually, the centroid of a cluster is its center point. The difference between an object $p \in Ci$ and ci, the representative of the cluster, is measured by dist (p,ci), where dist(x,y) is the Euclidean distance between two points x and y.

Algorithm: The k-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster. Input: k: the number of clusters, D: a data set containing n objects.

Output: A set of k clusters.

Method: (1) arbitrarily choose k objects from D as the initial cluster centers; (2) repeat (3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster; (4) update the

cluster means, that is, calculate the mean value of the objects for each cluster; (5) until no change.

How K-Means Algorithm Works

The k-means clustering algorithm works by finding like groups based on Euclidean distance, a measure of distance or similarity. The practitioner selects kk groups to cluster, and the algorithm finds the best centroids for the kk groups. The practitioner can then use those groups to determine which factors group members relate. For customers, these would be their buying preferences

K-Means clustering intends to partition n objects into k clusters in which each object belongs to the cluster with the nearest mean. This method produces exactly k different clusters of greatest possible distinction. The best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data. The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

Algorithm:

1. Clusters the data into k groups where k is predefined.

2. Select k points at random as cluster centers.

3. Assign objects to their closest cluster centre according to the Euclidean distance function.

4. Calculate the centroid or mean of all objects in each cluster.

5. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

K-Means is relatively an efficient method. However, we need to specify the number of clusters, in advance and the final results are sensitive to initialization and often terminates at a local optimum. Unfortunately there is no global theoretical method to find the optimal number of clusters. A practical approach is to compare the outcomes of multiple runs with different k and choose the best one based on a predefined criterion. In general, a large k probably decreases the error but increases the risk of over fitting.

CHAPTER.4

IMPLEMENTATION

What is Clustering?

Imagine that you have a group of chocolates and liquor ice candies. You are required to separate the two eatables. Intuitively, you are able to separate them based on their appearances. The process of segregating objects into groups based on their respective characteristics is called clustering. In clusters, the features of objects in a group are similar to other objects present in the same group.

Clustering is used in various fields like image recognition, pattern analysis, medical informatics, genomics, data compression etc. It is part of the unsupervised learning algorithm in machine learning. This is because the data-points present are not labeled and there is no explicit mapping of input and outputs. As such, based on the patterns present inside, clustering takes place.

K-Means Clustering: K-Means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

We then proceeded to perform K-means Clustering which will create different clusters to group similar spending activity based on their age and annual income. KMeans Clustering selects random values from the data and forms clusters assigned. The closest values from the centre of each cluster were taken to update the cluster and reshape the plot (just like k-NN). The closest values are based on Euclidean Distance.

Building the k-means model:

We need to visualize the data which we are going to use for the clustering. This will give us a fair idea about the data we're working on. This will give us a fair Idea and patterns about some of the data.

First we read the data from the dataset using `read_csv` from the pandas library.

```
In [2]: 1 data = pd.read_csv('data\Mall_Customers.csv')
```

Viewing the data that we imported to pandas dataframe object

```
In [3]: 1 data
```

Out[3]:

	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
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

Gathering Further information about the dataset using `info()`

In [11]:
```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual Income (k$)      200 non-null    int64
 4   Spending Score (1-100)  200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

Describing the data as basic statistics using `describe()`

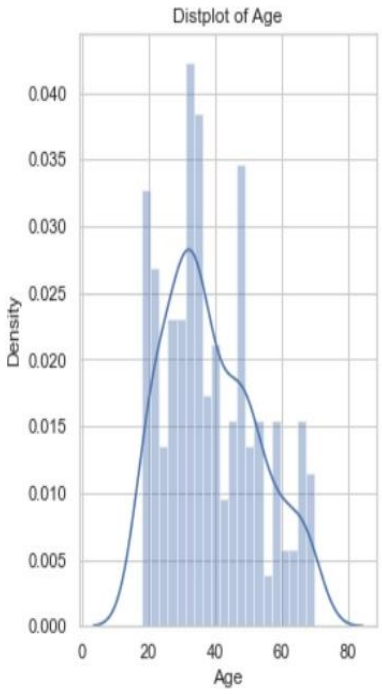
In [12]:
```
1 data.describe()
```

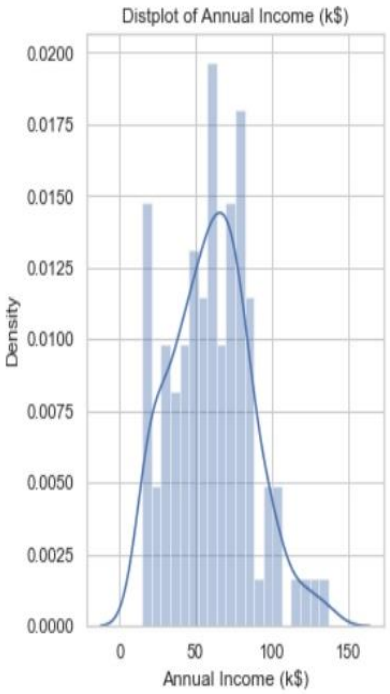
Out[12]:

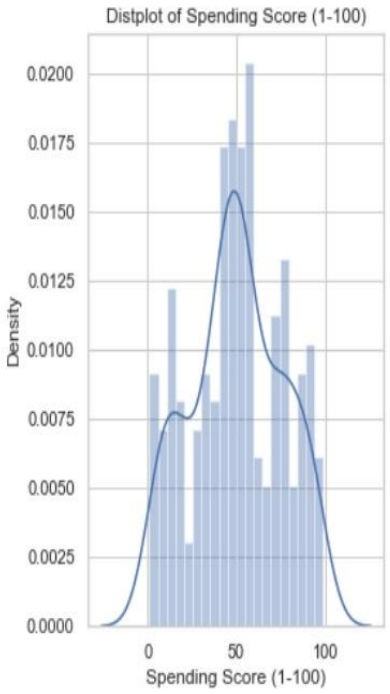
	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

EXPLORATORY DATA ANALYSIS

```
In [84]: df.dtypes

Out[84]: CustomerID                int64
         Gender                   object
         Age                       int64
         Annual Income (k$)        int64
         Spending Score (1-100)    int64
         dtype: object

In [85]: # checking the missing values
         df.isnull().sum()

Out[85]: CustomerID                0
         Gender                    0
         Age                       0
         Annual Income (k$)        0
         Spending Score (1-100)    0
         dtype: int64

In [86]: plt.figure(1, figsize=(15,6))
         n=0
         for x in ['Age','Annual Income (k$)', 'Spending Score (1-100)']:
             n+=1
             plt.subplot(1 , 3 , n)
             plt.subplots_adjust(hspace =0.5 , wspace = 0.5)
             sns.distplot(df[x] , bins = 20)
             plt.title('Distplot of {}'.format(x))
         plt.show()
```





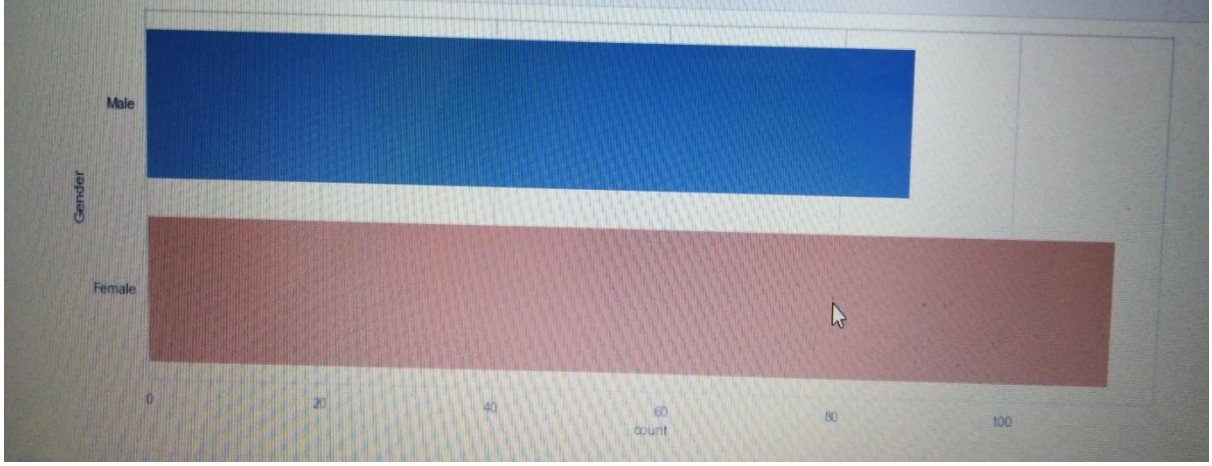


GENDER DISTRIBUTION:

- **Visualization of Distribution of Males and Females**

Ploting the No.of Gender in Dataset

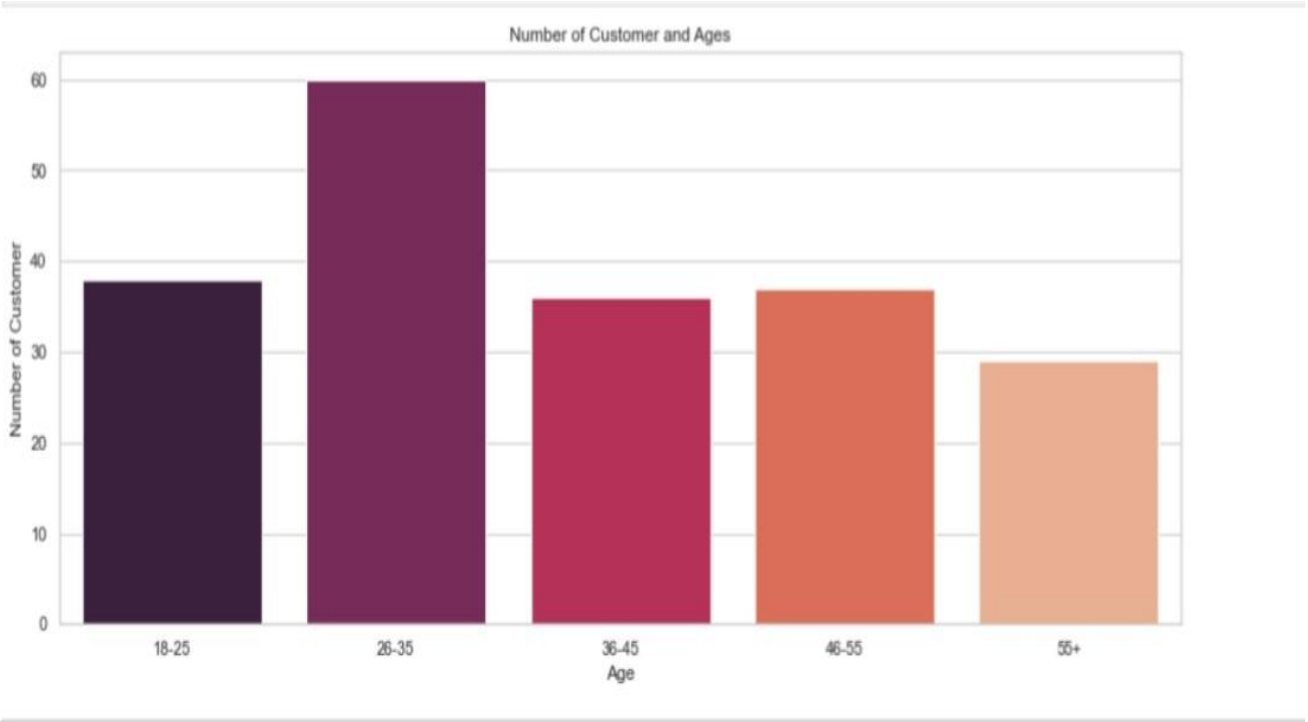
```
[87]: plt.figure(figsize=(15,5))
      sns.countplot(y='Gender',data=df)
      plt.show()
```



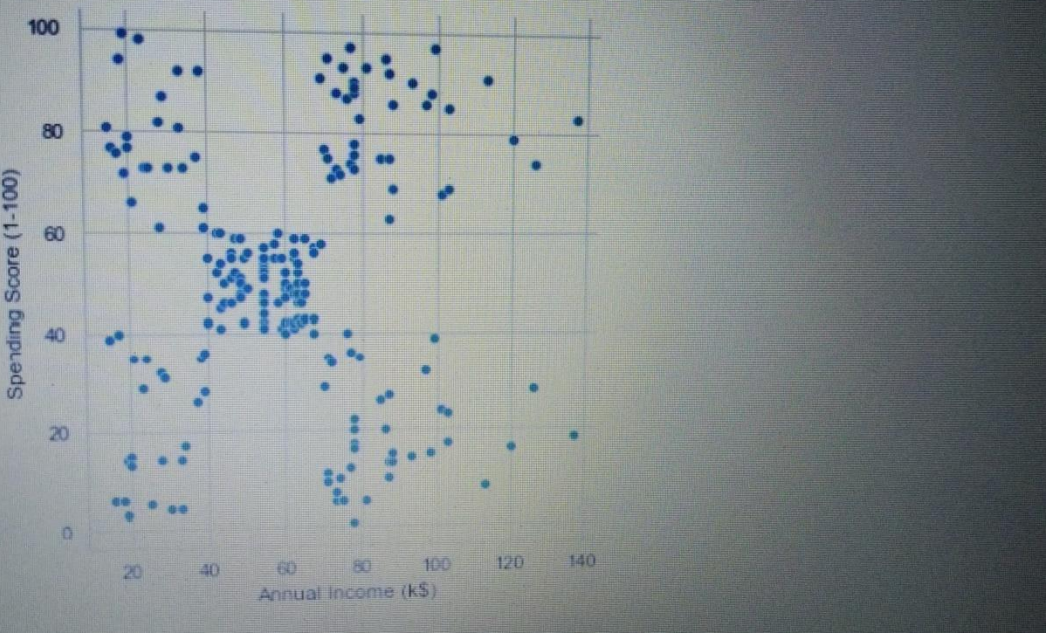
- **AGE ANALYSIS OF CUSTOMERS**

Calculating the Number of Customers and Ages in the Dataset

```
In [89]: age_18_25 = df.Age[(df.Age >=18) & (df.Age <=25)]
         age_26_35 = df.Age[(df.Age >=26) & (df.Age <=35)]
         age_36_45 = df.Age[(df.Age >=36) & (df.Age <=45)]
         age_46_55 = df.Age[(df.Age >=46) & (df.Age <=55)]
         age_55above = df.Age[df.Age >= 56]

         agex = ["18-25","26-35","36-45","46-55","55+"]
         agey = [len(age_18_25.values),len(age_26_35.values),len(age_36_45.values),len(age_46_55.values),len(age_55above.values)]

         plt.figure(figsize=(15,6))
         sns.barplot(x= agex, y= agey, palette="rocket")
         plt.title("Number of Customer and Ages")
         plt.xlabel("Age")
         plt.ylabel("Number of Customer")
         plt.show()
```



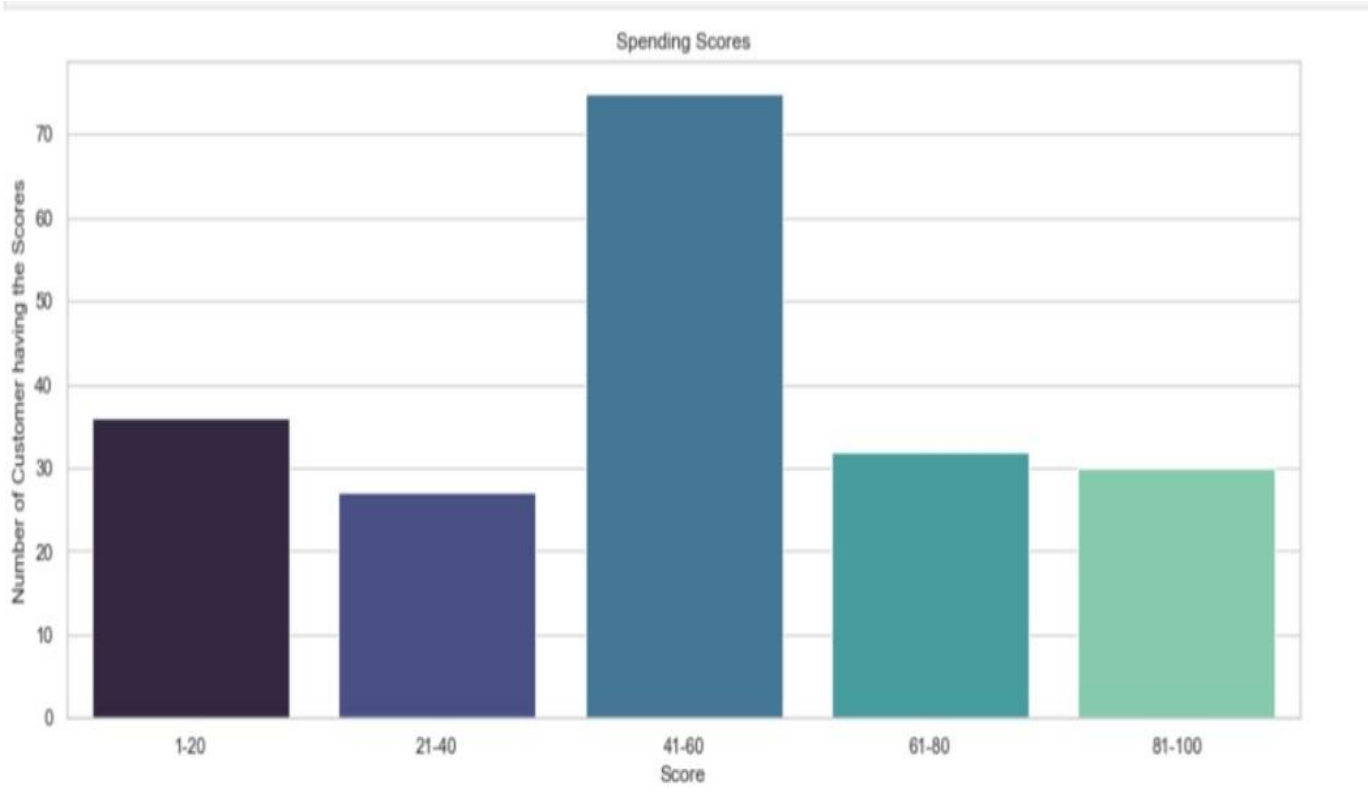

```
In [90]: sns.relplot(x="Annual Income (k$)", y="Spending Score (1-100)", data=df)
Out[90]: <seaborn.axisgrid.FacetGrid at 0x2a2c41df7c0>
```



- **SPENDING SCORE ANALYSIS :**

Calcuating the Spending Scoresin the given Dataset

```
ss_1_20 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 1) & (df["Spending Score (1-100)"] <= 20)]
ss_21_40 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 21) & (df["Spending Score (1-100)"] <= 40)]
ss_41_60 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 41) & (df["Spending Score (1-100)"] <= 60)]
ss_61_80 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 61) & (df["Spending Score (1-100)"] <= 80)]
ss_81_100 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 81) & (df["Spending Score (1-100)"] <= 100)]

ssx = ["1-20", "21-40", "41-60", "61-80", "81-100"]
ssy = [len(ss_1_20.values), len(ss_21_40.values), len(ss_41_60.values), len(ss_61_80.values), len(ss_81_100.values)]



plt.figure(figsize=(15,6))
sns.barplot(x= ssx, y= ssy, palette="mako")
plt.title("Spending Scores")
plt.xlabel("Score")
plt.ylabel("Number of Customer having the Scores")
plt.show()
```

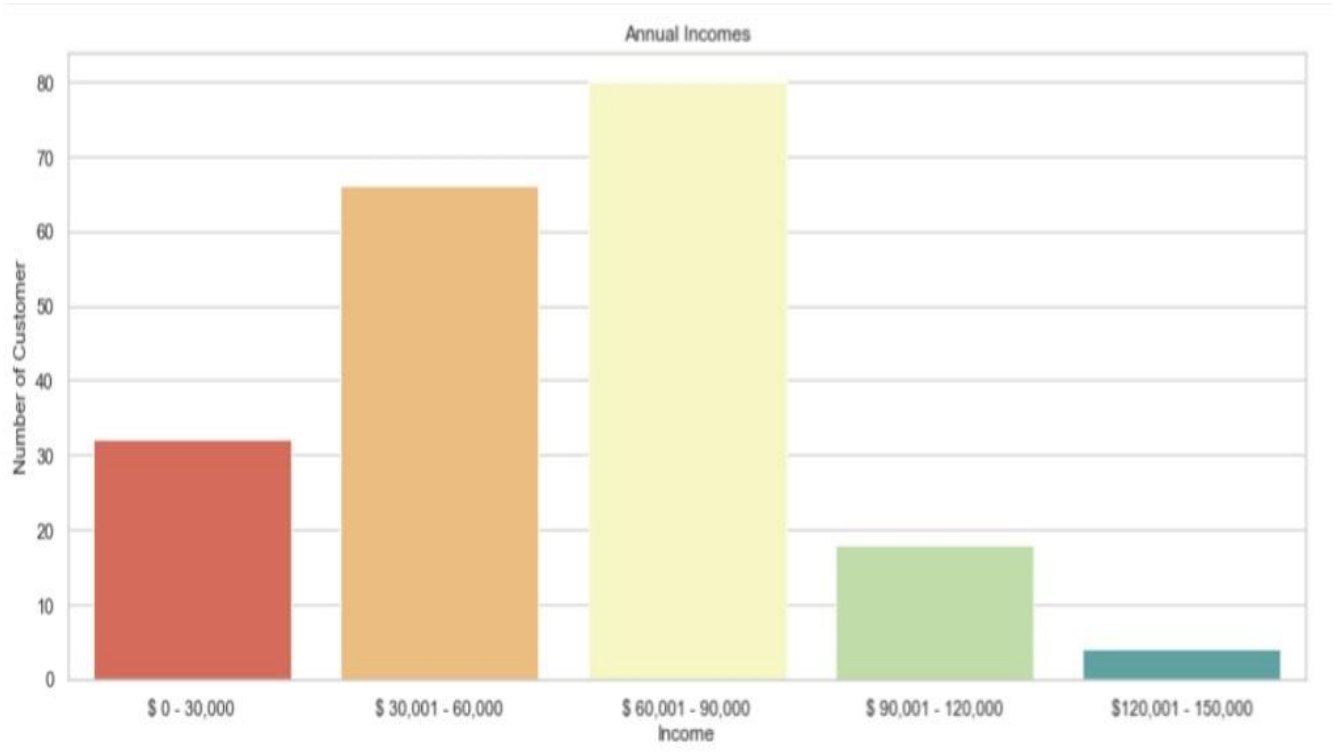


- **ANNUAL INCOME ANALYSIS:**

Calculating the Annual Incomes of the customers in the Dataset

In [92]:

```
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)]

plt.figure(figsize=(15,6))
sns.barplot(x= aix, y= aiy, palette="Spectral")
plt.title("Annual Incomes")
plt.xlabel("Income")
plt.ylabel("Number of Customer")
plt.show()
```



ELBOW METHOD:

The Elbow Method:

The Elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

When these overall metrics for each model are plotted, it is possible to visually determine the best value for k. If the line chart looks like an arm, then the "elbow" (the point of inflection on the curve) is the best value of k. The "arm" can be either up or down, but if there is a strong inflection point, it is a good indication that the underlying model fits best at that point.

We use the Elbow Method which uses Within Cluster Sum Of Squares (WCSS) against the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids which is given by the 16 below formula.

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

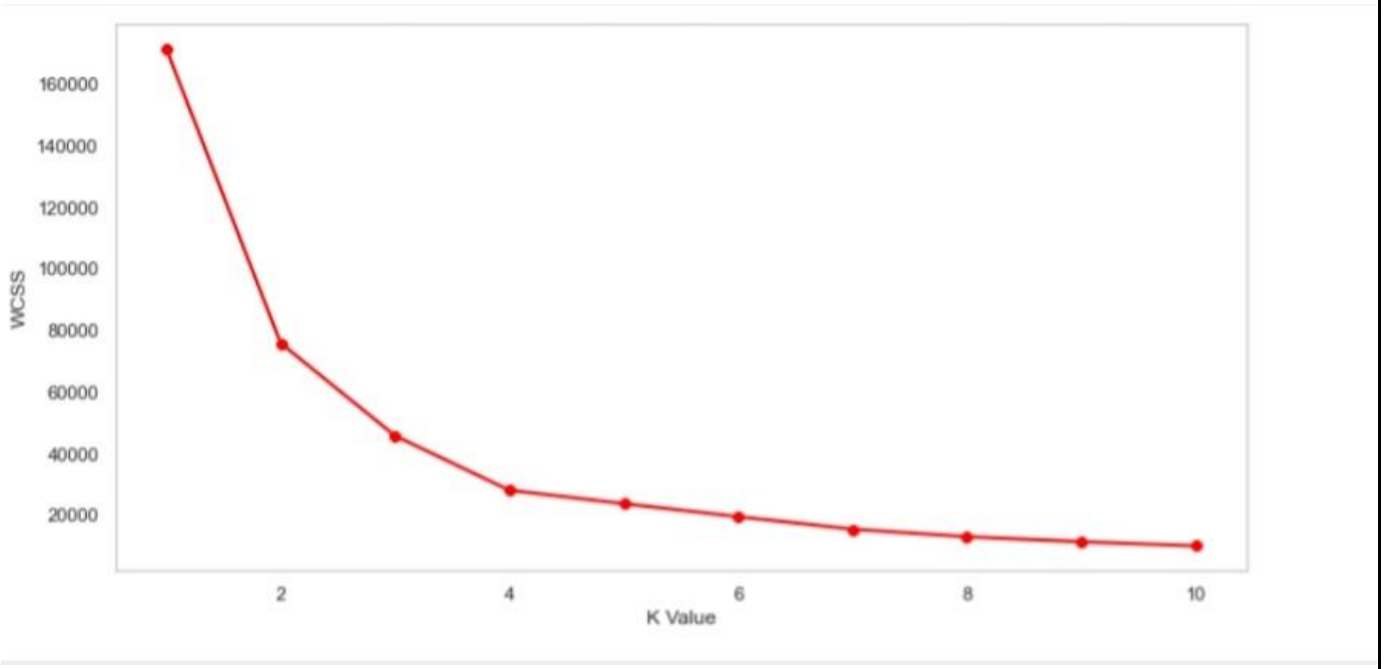
Where Yi is centroid for observation Xi. The main goal is to maximize number of clusters and in limiting case each data point becomes its own cluster centroid.

ELBOW METHOD OF AGE AND SPENDING SCORE IN THE GIVEN DATASET

In [93]:

```
X1= df.loc[:, ["Age","Spending Score (1-100)"]].values
from sklearn.cluster import KMeans
wcss = []
for k in range(1,11):
    kmeans = KMeans(n_clusters=k, init="k-means++")
    kmeans.fit(X1)
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.ylabel("WCSS")
plt.show()
```

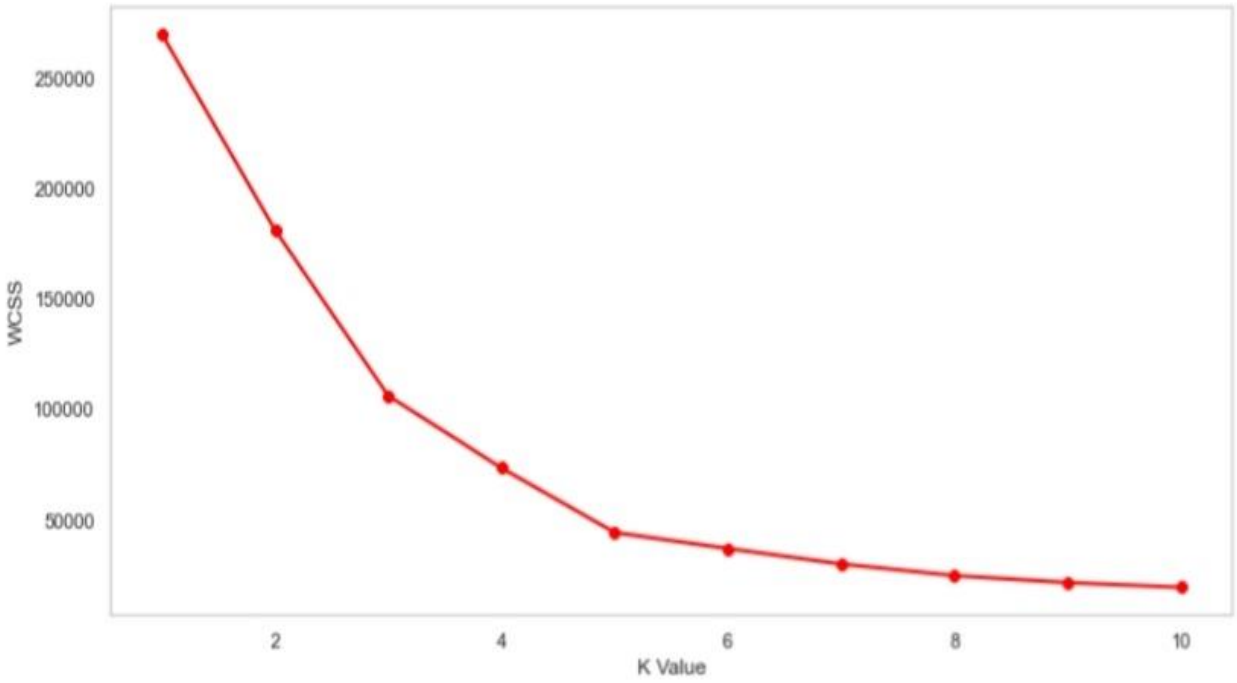
Elbow method – age, spending score



Elbow method – Annual income , Spending Score

ELBOW METHOD between Annual Income and Spending scores

```
[121...
X2=df.loc[:, ["Annual Income (k$)","Spending Score (1-100)"]].values

from sklearn.cluster import KMeans
wcss = []
for k in range(1,11):
    kmeans = KMeans(n_clusters=k, init="k-means++")
    kmeans.fit(X2)
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.ylabel("WCSS")
plt.show()
```

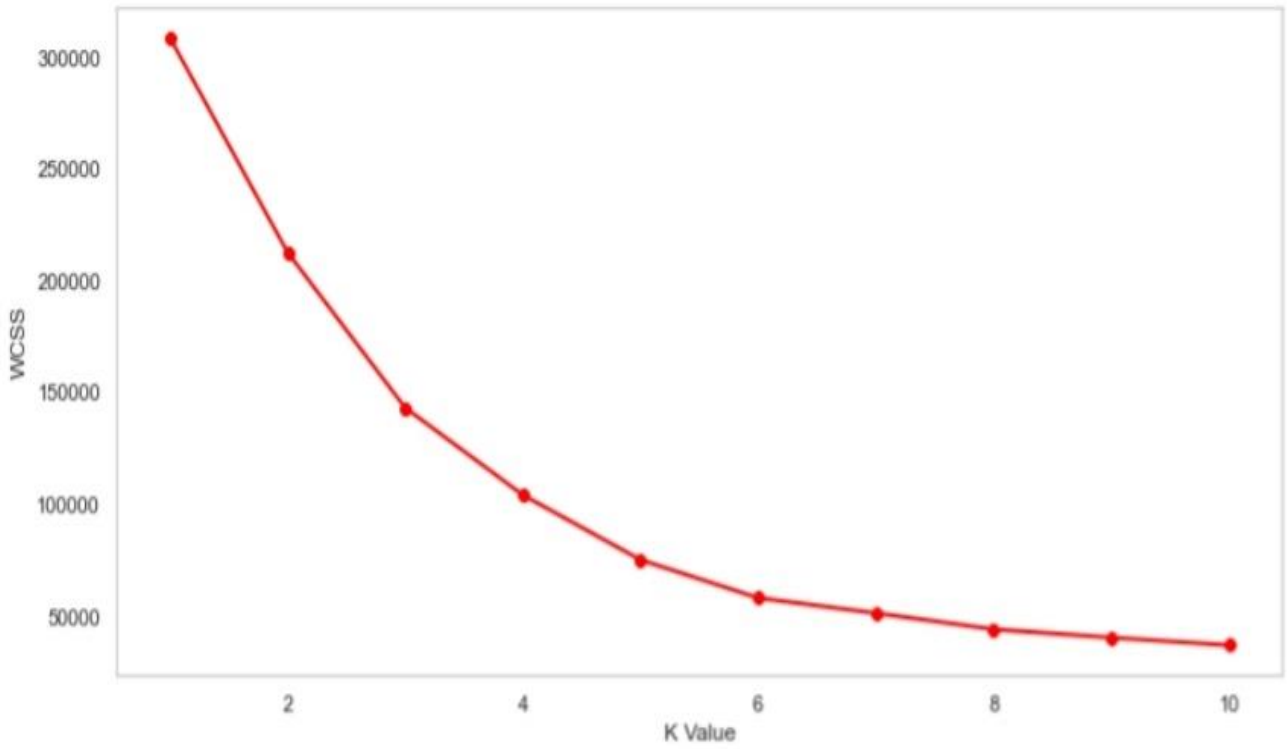


Elbow method – AGE, Annual Income, Spending Score

ELBOW METHOD FOR ALL AGE ,ANNUAL INCOME ,SPENDING SCORES.

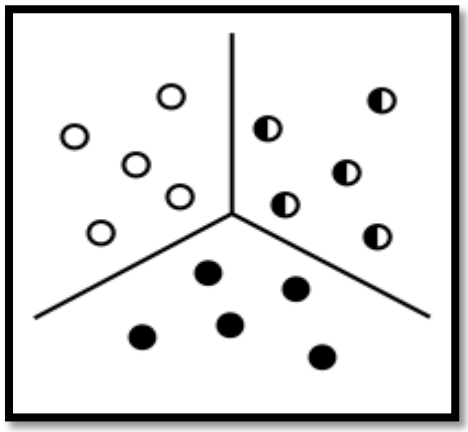
In [122...

```
X3=df.loc[:, ["Age","Annual Income (k$)","Spending Score (1-100)"]].values

wcss= []
for k in range(1,11):
    kmeans = KMeans(n_clusters=k, init="k-means++")
    kmeans.fit(X3)
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.ylabel("WCSS")
plt.show()
```



WHAT IS CLUSTERING?

The process of segregating objects into groups based on their respective characteristics is called clustering. In clusters, the features of objects in a group are similar to other objects present in the same group



K-Means Clustering

The process of segregating objects into groups based on their respective characteristics is called clustering. In clusters, the features of objects in a group are similar to other objects present in the same group

BUILDING THE K-MEANS CLUSTER MODEL:

Cluster plotting between AGE and SPENDING SCORES

```
[96]: plt.scatter(X1[:,0], X1[:,1], c=kmeans.labels_, cmap="rainbow")
      plt.scatter(kmeans.cluster_centers_[:,0] ,kmeans.cluster_centers_[:,1], color="black")
      plt.title("Clusters of Customers")
      plt.xlabel("Age")
      plt.ylabel("Spending Score(1-100)")
      plt.show()
```

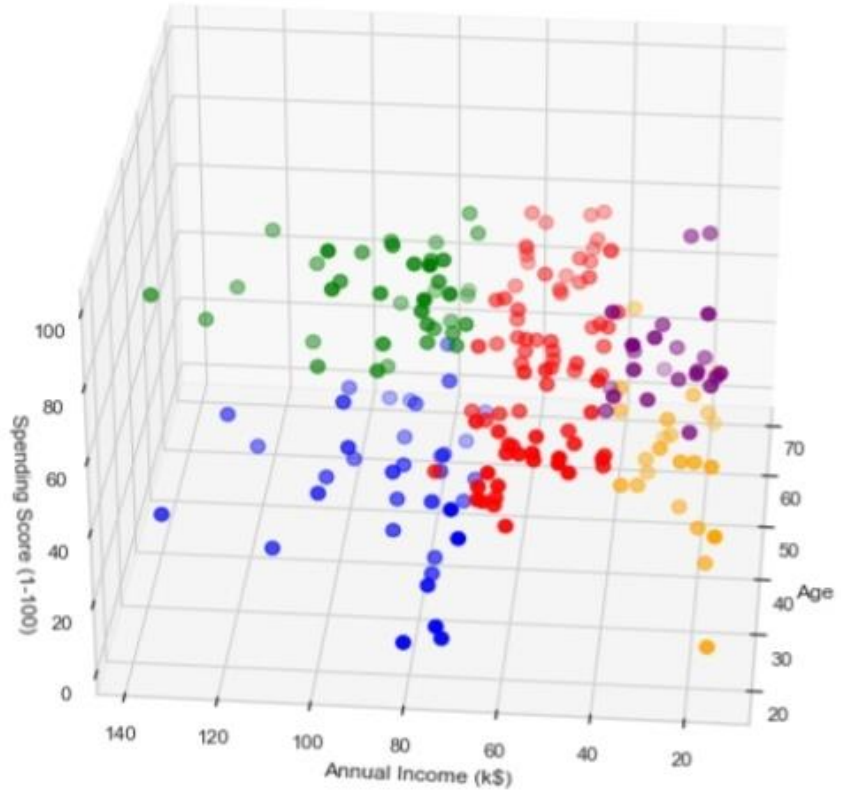
SCATTER PLOT BETWEEN ANNUAL INCOME AND SPENDING SCORES

SCATTER PLOT between Annual Income and Spending Scores

```
[10_ plt.scatter(X2[:,0], X1[:,1], c=kmeans.labels_, cmap="rainbow")
     plt.scatter(kmeans.cluster_centers_[:,0] ,kmeans.cluster_centers_[:,1], color="black")
     plt.title("Clusters of Customers")
     plt.xlabel("Annual Income (k$)")
     plt.ylabel("Spending Score(1-100)")
     plt.show()
```

3-D Plotting of Annual Income, Age, Spending Score

```
3-D PLOTTING OF ANNUAL INCOME ,AGE AND SPENDING SCORES


clusters = kmeans.fit_predict(X3)
df["label"]= clusters

from mpl_toolkits.mplot3d import Axes3D
fig = plt.figure(figsize=(20,10))
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.view_init(30, 185)

plt.xlabel("Age")
plt.ylabel("Annual Income (k$)")
ax.set_zlabel("Spending Score (1-100)")

plt.show()
```



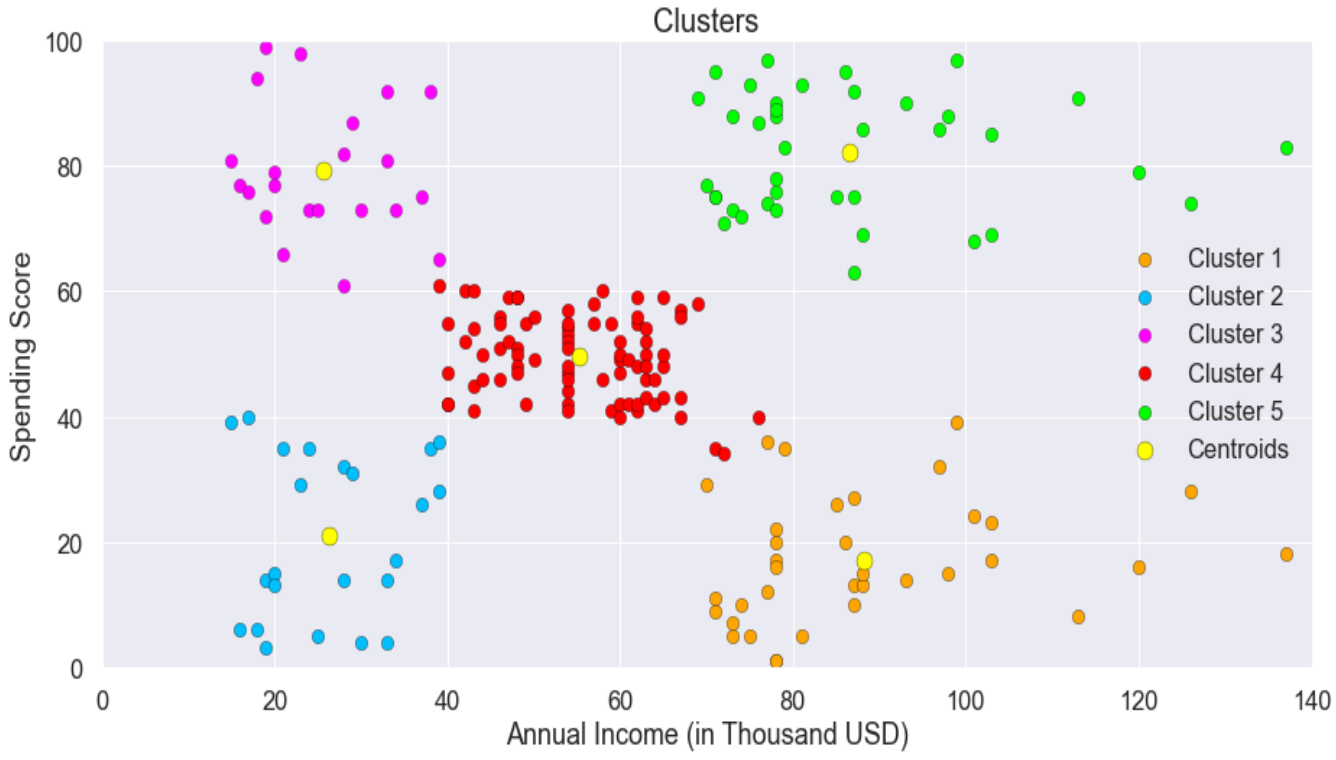
CHAPTER.5

ANALYSIS

Cluster Analysis:

The following clusters are created by the model,

1. Cluster Orange
2. Cluster Blue
3. Cluster Purple
4. Cluster Red
5. Cluster Green



1. **Cluster Orange – Balanced Customers:**

   They earn less and spend less. We can see people have low annual income and low spending scores, this is quite reasonable as people having low salaries prefer to buy less, in fact, these are the wise people who know how to spend and save money. The shops/mall will be least interested in people belonging to this cluster.

2. **Cluster Blue - Pinch Penny Customers:**

   Earning high and spending less. We see that people have high income but low spending scores, this is interesting. Maybe these are the people who are unsatisfied or unhappy by the mall's services. These can be the prime targets of the mall, as they have the potential to spend money. So, the mall authorities will try to add new facilities so that they can attract these people and can meet their needs.

3. **Cluster Purple - Normal Customer:**

   Customers are average in terms of earning and spending An Average consumer in terms of spending and Annual Income we see that people have average income and an average spending score, these people again will not be the prime targets of the shops or mall, but again they will be considered and other data analysis techniques may be used to increase their spending score.

4. **Cluster Red - Spenders**:

   This type of customers earns less but spends more Annual Income is less but spending high, so can also be treated as potential target customer we can see that people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. Maybe it's because these people are more than satisfied with the mall services. The shops/malls might not target these people that effectively but still will not lose them.

5. **Cluster Green - Target Customers:**

   Earning high and also spending high Target Customers. Annual Income High as well as Spending Score is high, so a target consumer. we see that people have high income and high spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. These people might be the regular customers of the mall and are convinced by the mall's facilities.

CHAPTER.6
ADVANTAGES

Benefits of Customer Segmentation

At the expansion stage, executing a marketing strategy without any knowledge of how your target market is segmented is akin to firing shots at a target 100 feet away — while blindfolded. The likelihood of hitting the target is a matter of luck more than anything else.

Without a deep understanding of how a company's best current customers are segmented, a business often lacks the market focus needed to allocate and spend its precious human and capital resources efficiently. Furthermore, a lack of best current customer segment focus can cause diffused go-tomarket and product development strategies that hamper a company's ability to fully engage with its target segments. Together, all of those factors can ultimately impede a company's growth.

If best current customer segmentation is done right, however, the business benefits are numerous. For example, a best current customer segmentation exercise can tangibly impact your operating results by:

1. **Improving your whole product:**
   Having a clear idea of who wants to buy your product and what they need it for will help you differentiate your company as the best solution for their individual needs. The result will be increased satisfaction and better performance against competitors. The benefits also extend beyond your core product offering, since any insights into your best customers will allow your organization to offer better customer support, professional services, and any other offerings that make up their whole product experience.
2. **Focusing your marketing message:**
   In parallel with improvements to the product, conducting a customer segmentation project can help you develop more focused marketing messages that are customized to each of your best segments, resulting in higher quality inbound interest in your product.

3. **Allowing your sales organization to pursue higher percentage opportunities:**

   By spending less time on less lucrative opportunities and more on your most successful segments, your sales team will be able to increase its win rate, cover more ground, and ultimately increase revenues.

4. **Getting higher quality revenues**:

   Not all revenue dollars are created equal. Sales into the wrong segment can be more expensive to sell and maintain, and may have a higher churn rate or lower upsell potential after the initial purchase has been made. Staying away from these types of customers and focusing on better ones will increase your margins and promote the stability of your customer base.

   Conducting best current customer segmentation research can have numerous other ancillary benefits, of course, but this guide will focus primarily on how it can impact the four cited above. The bottom line is that if you are able to sell more of your product to your most profitable customers, then you will be able to scale the business more efficiently and ensure that everything you do — from lead generation to new product development — revolves around the right things.

The two main advantages of cluster analysis over simple threshold/rulebased segmentation are –

• Practicality – it would be practically impossible to use predetermined rules to segment customers over many dimensions, and

• Homogeneity – variances within each resulting group are very small in cluster analysis, whereas rule-based segmentation typically groups customers who are actually very different from one another.

CHAPTER.7
FUTURE SCOPE & CONCLUSION

While this guide provides a step-by-step process for identifying, prioritizing, and targeting your best current customer segments, simply following it does not guarantee success. To be effective, you must prepare and plan for the various challenges and hurdles that each step may present, and always make sure to adapt your process to any new information or feedback that might change its output.

Additionally, you cannot force feed this process on your business. If the key stakeholders that will be impacted by the best current customers segmentation process do not fully buy-in, then the outputs produced from it will be relatively meaningless.

If you properly manage the best current customer segmentation process, however, the impact it can have on every part of your organization — sales, marketing, product development, customer service, etc. — is immense. Your business will possess stronger customer focus and market clarity, allowing it to scale in a far more predictable and efficient manner.

Ultimately, that means no longer needing to take on every customer that is willing to pay for your product or service, which will allow you to instead hone in on a specific subset of customers that present the most profitable opportunities and efficient use of resources. That is critical for every business, of course, but at the expansion stage, it can often be the difference between incredible success and certain failure.

CHAPTER.8

RESULT

We have explored the five segments based on customers Annual Income and Spending Score which are reportedly the best factors/attributes to determine the segments of a customer in a Mall. They include; Pinch Penny Customers, Balanced Customers, Target Customers, Spender and the normal customer. We can put Target Customers into some alerting system where SMS and emails can be sent to them on daily basis regarding the offers and discounts that they can get at the Mall; while the rest we can set once per week in a month for blast SMSs to notify them about our products.

Similarly, now we know customers behavior depending upon their Annual Income and Spending Score. There can be many marketing strategies applied for Customers on this Cluster Analysis. High income and High spending score customers are our target customers and we would always want to retain them as they give the most profit margins to our organization. High Income and Less spending score customers can be attracted with wide range of products in their life style demands and it might attract them towards the Mall Supermarket. Less Income Less Spending Score can be given extra offers and constantly sending them the offers and discounts will attract them towards spending. We can also have a cluster analysis done on what kind of products customers tend to buy and can make other marketing strategies accordingly. The data set did not have enough data to carry out more analytics on the same.

CHAPTER.9

REFERENCE

Concepts of customer segmentation

http://www.business-science.io

- https://labs.openviewpartners.com/customer-segmentation/
- Source data related to our analysis has been collected from https://github.com/mdancho84/orderSimulatoR/tree/master/data
- https://www.kaggle.com/
- https://www.r-project.org/
- https://www.rstudio.com