



ITM UNIVERSITY GWALIOR, MP, INDIA
www.itmuniversity.ac.in

1ST RANKING
in state consecutively in last
5 years

Program: B.Tech, Semester-8, Fourth Year

CSD0804: Major Project -II

Customer Segmentation using K-Means Algorithm

Submitted By:

Vinayak Singh [BETN1CS19068]

Pranav Kashyap [BETN1CS19048]

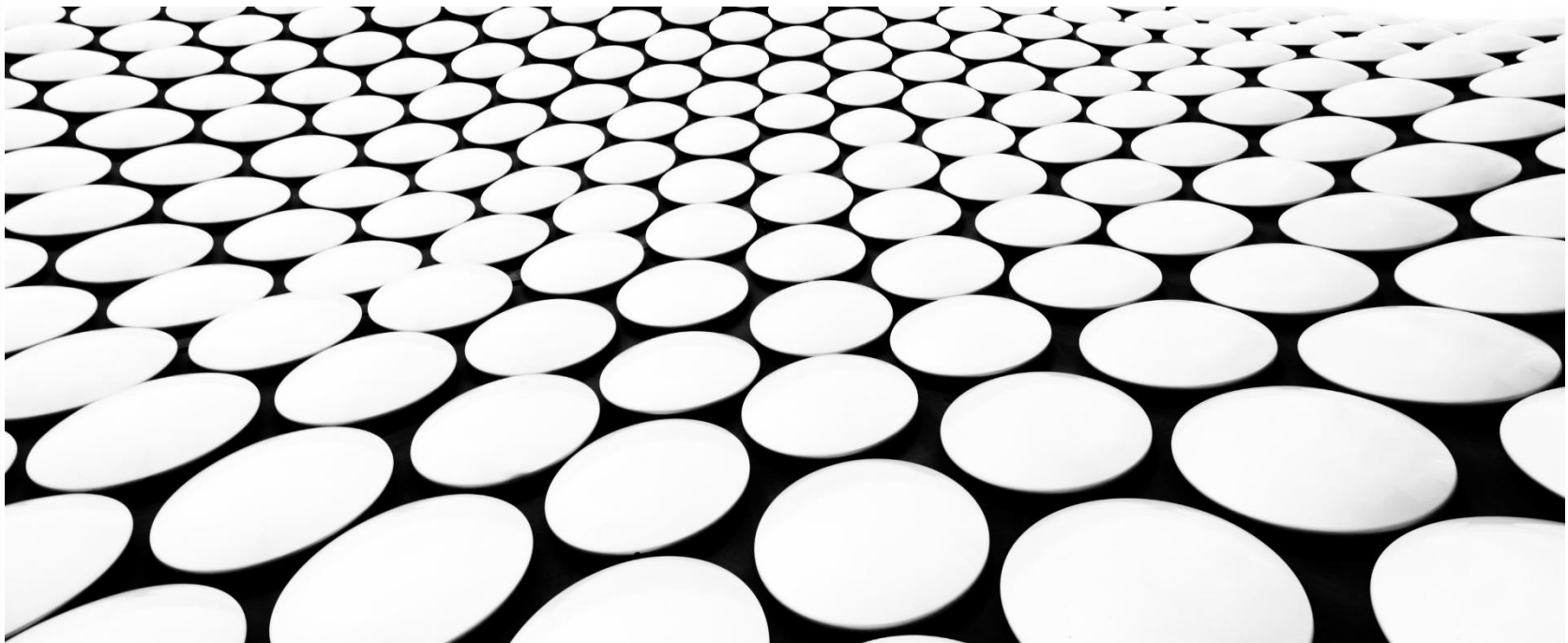
Abhishek Kumar [BETN1CS19108]



ITM UNIVERSITY GWALIOR, MP, INDIA
www.itmuniversity.ac.in

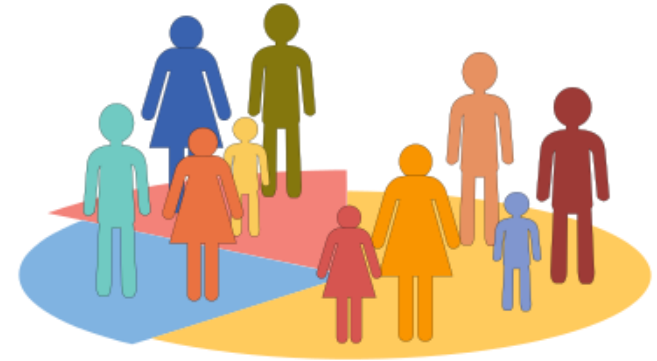
1ST RANKING
in state consecutively in last
5 years

Customer Segmentation (using K-Means)



What is Customer segmentation?

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.





Dataset Description:

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 .



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()

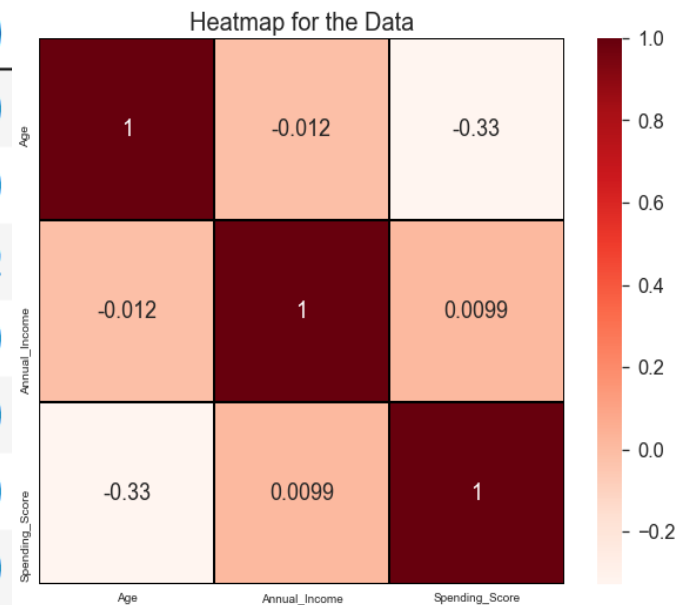
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   CustomerID                           200 non-null    int64
1   Gender                               200 non-null    object
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





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.

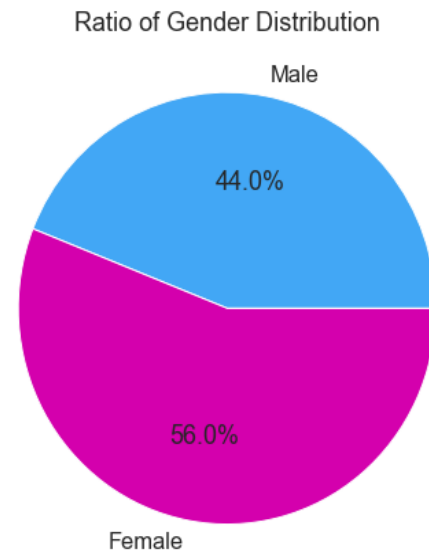
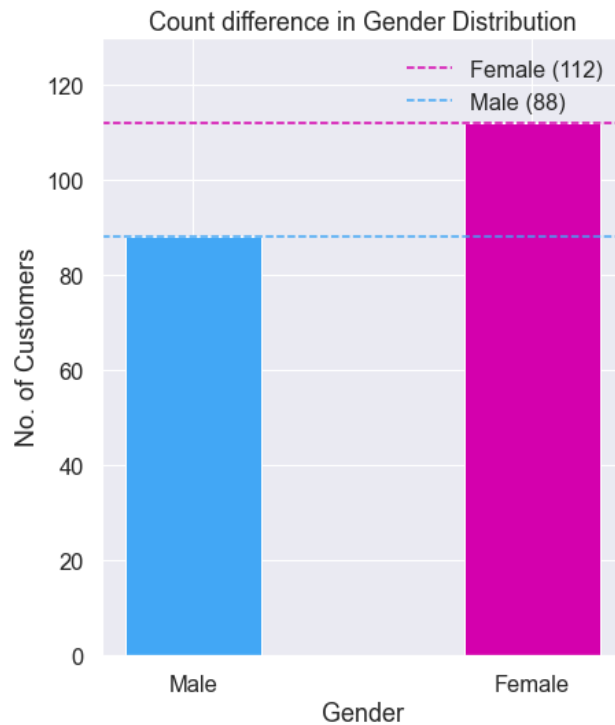


OBJECTIVE

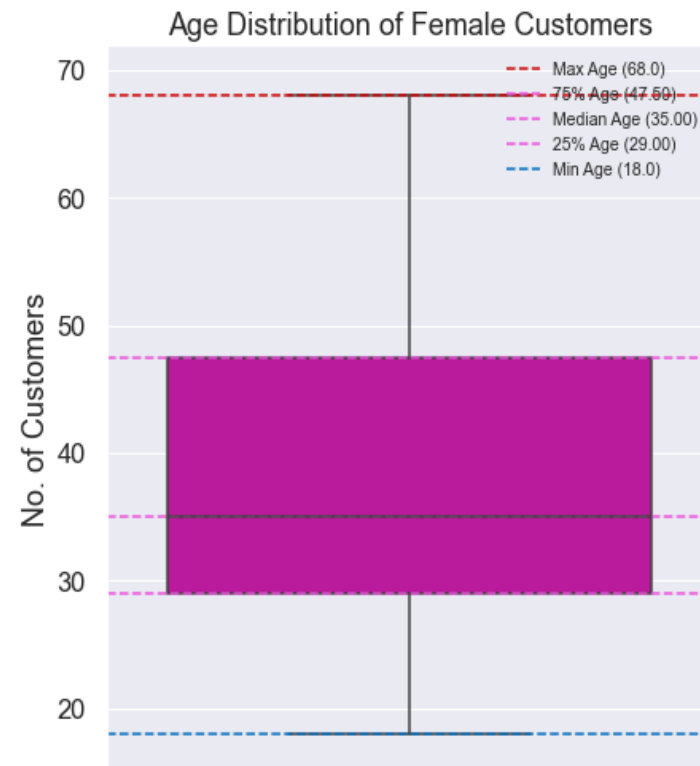
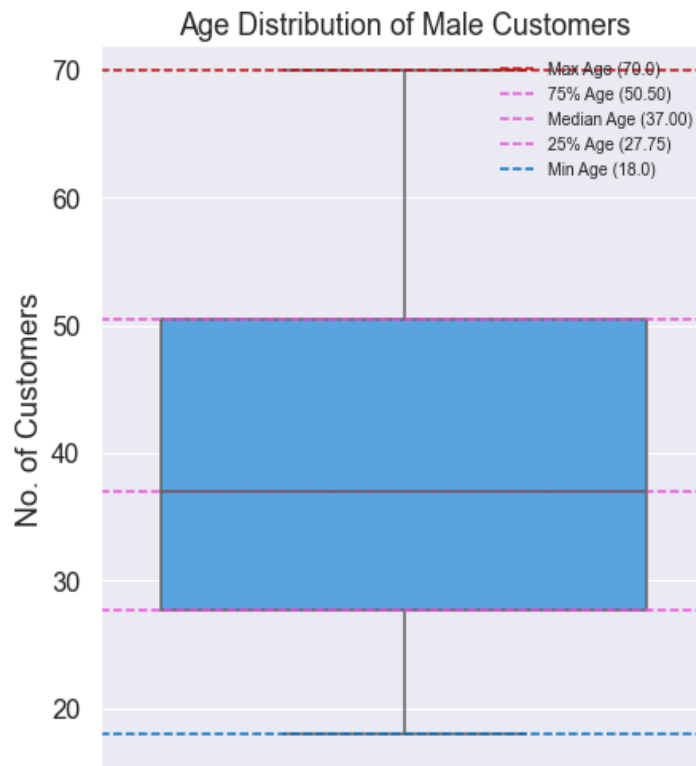
The objective of the project are as follows:
Identify the potential customer base for selling the product.
Implement Clustering Algorithms to group the customer base.

EXPLORATORY DATA ANALYSIS

Visualization of Distribution of Males and Females: Gender Distribution

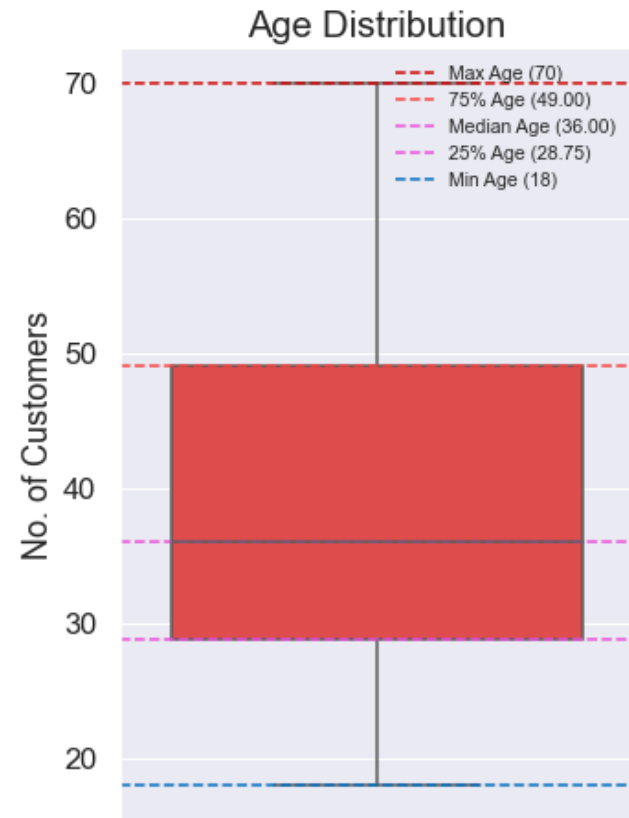


From the above graphs, we observe that the number of females(112) is higher than the males(88). The Ratio of Gender population is 56% Females and 44% Males. By this we can say that majority of the customers that visit the mall are Females.



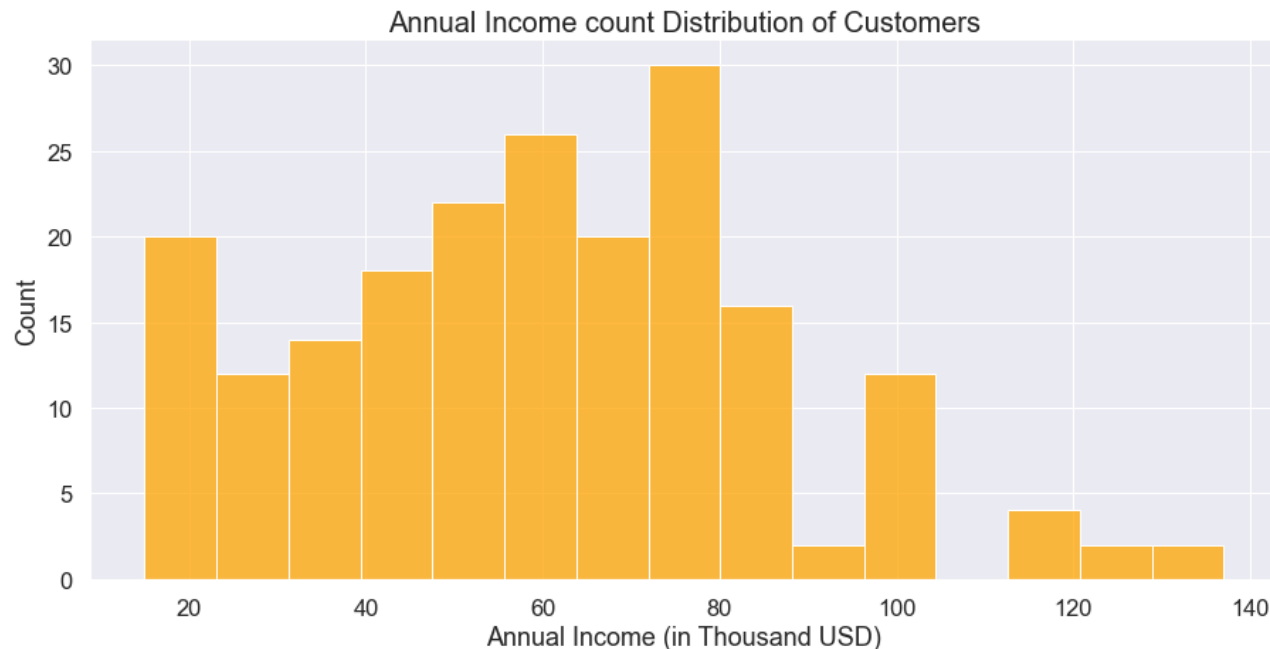
AGE ANALYSIS OF CUSTOMERS

From the above boxplot, we can conclude that a large amount of ages are between 30 and 35. Min Age is 18, Max Age is 70. By comparing the age distribution of the customers, we can conclude that most of the customers were within the band between 30 to 50, where the mean is around 35 years old.

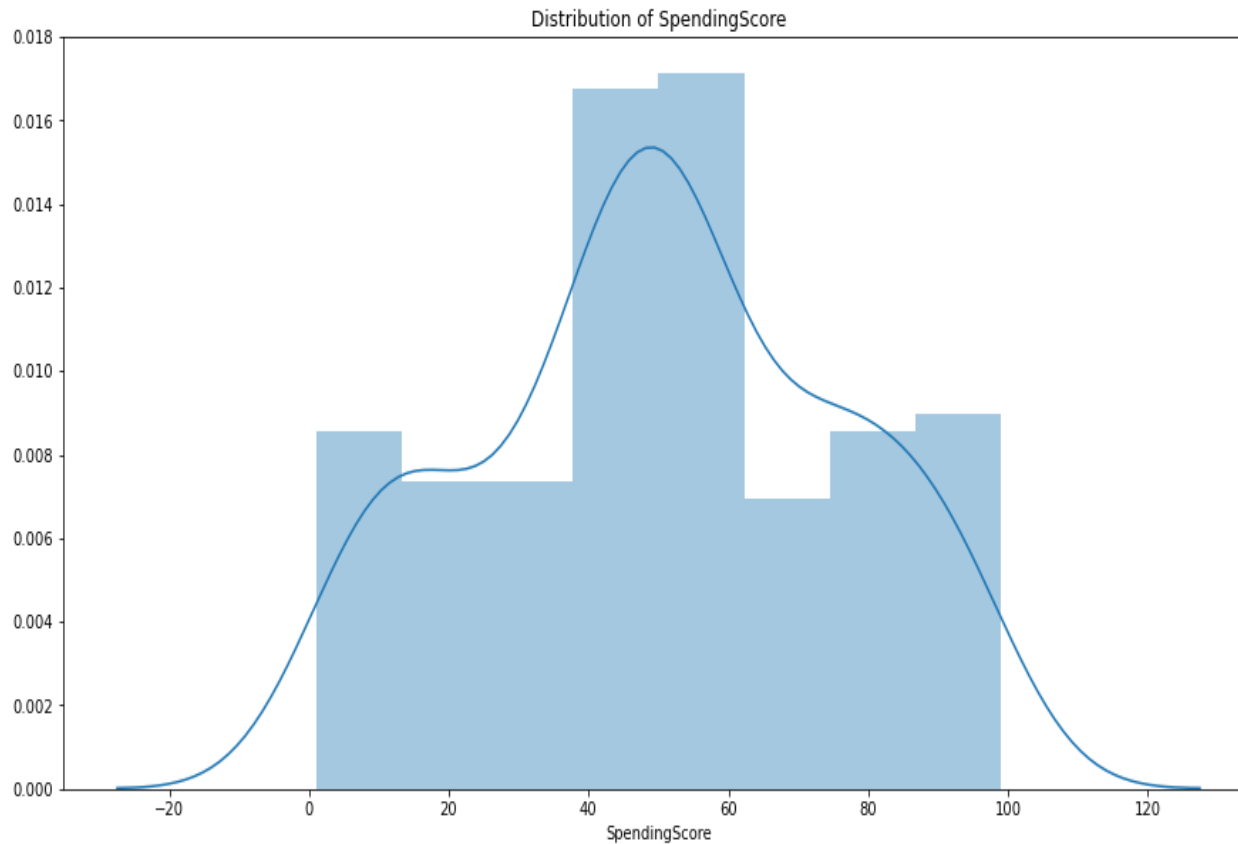


Annual Income and Spending Score Analysis:

The distribution of Annual Income and Spending Score exhibited an approximation of normal distribution, with highest density around the mean of the variables. The maximum and minimum of Annual Income are 137 and 15 respectively, with the mean at 60.56. From the plot, we can see that the peak of the distribution fell in the region of 60 to 75.



For the Spending score, the maximum and minimum are 99 and 1, while the histplot indicated that the highest number of customers have the spending score ranging from 40 to 60.



CHARACTERISTIC RELATIONS

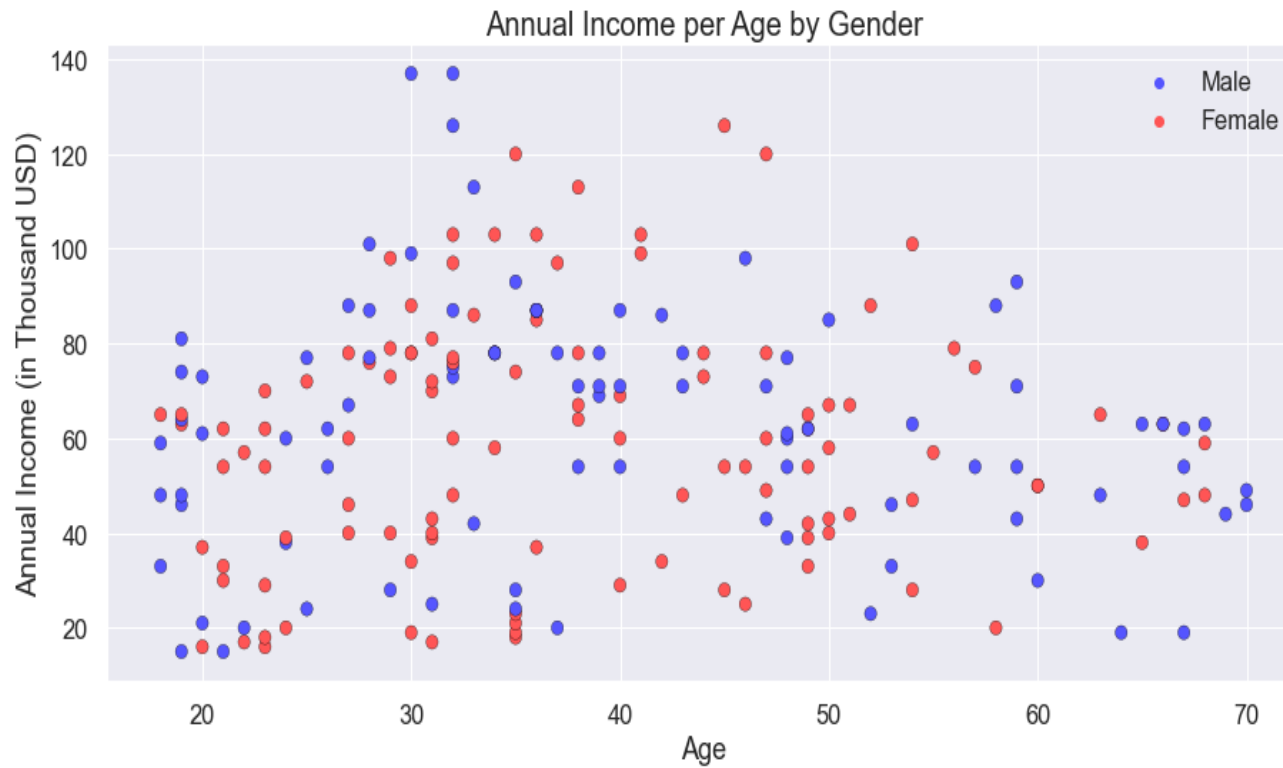
Annual Income vs Age analysis:





ITM UNIVERSITY GWALIOR, MP, INDIA
www.itmuniversity.ac.in

1ST RANKING
in state consecutively in last
5 years





ITM UNIVERSITY GWALIOR, MP, INDIA
www.itmuniversity.ac.in

1ST RANKING
in state consecutively in last
5 years

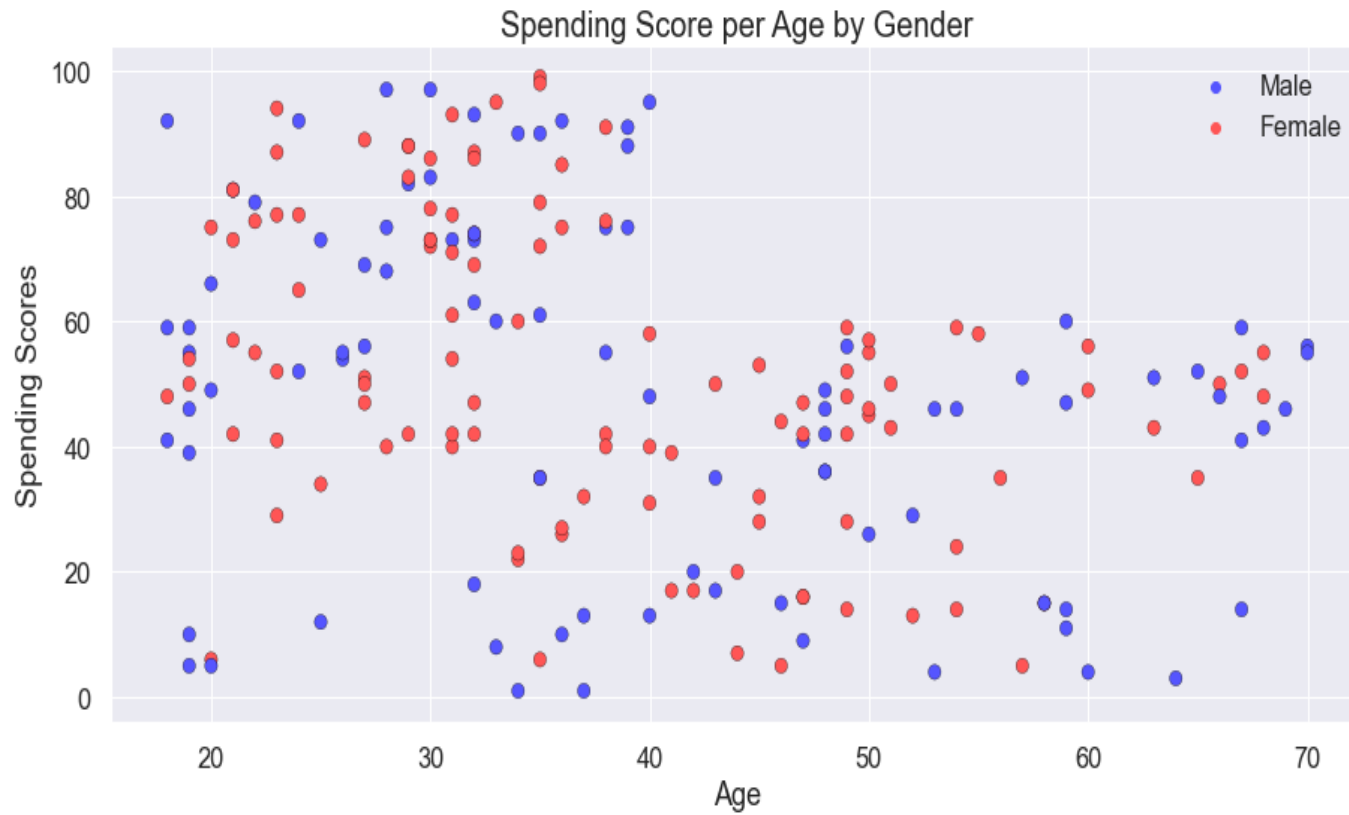
SPENDING SCORE VR AGE ANALYSIS





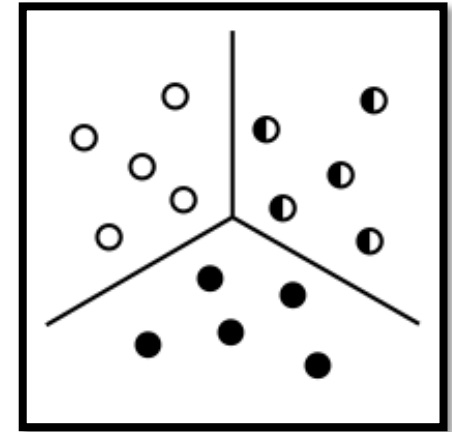
ITM UNIVERSITY GWALIOR, MP, INDIA
www.itmuniversity.ac.in

1ST RANKING
in state consecutively in last
5 years



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.

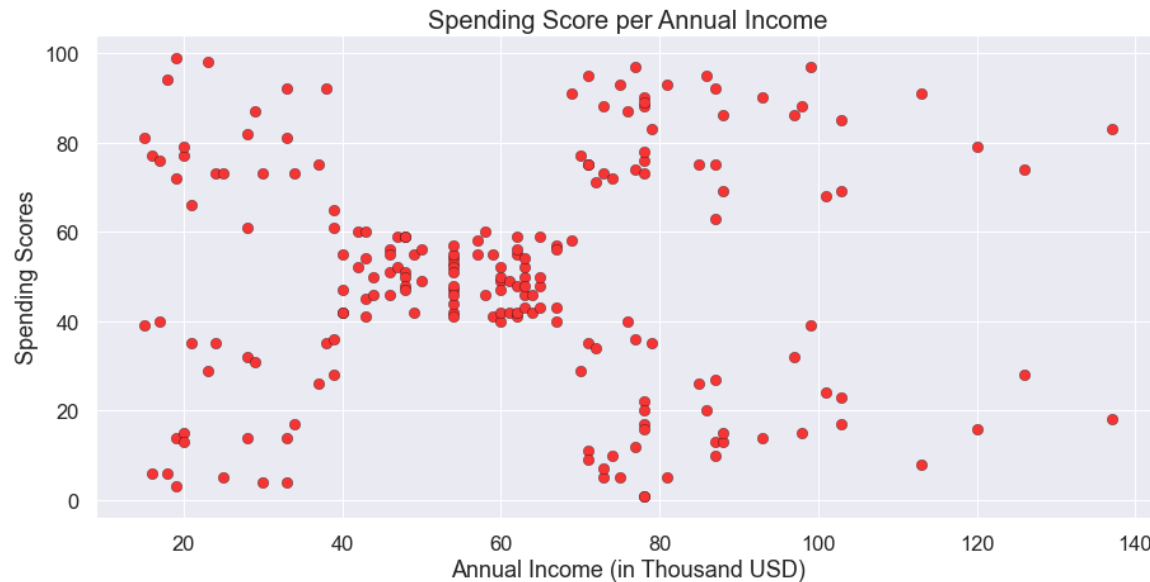


K-MEANS CLUSTERING

- 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 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.



The elbow method

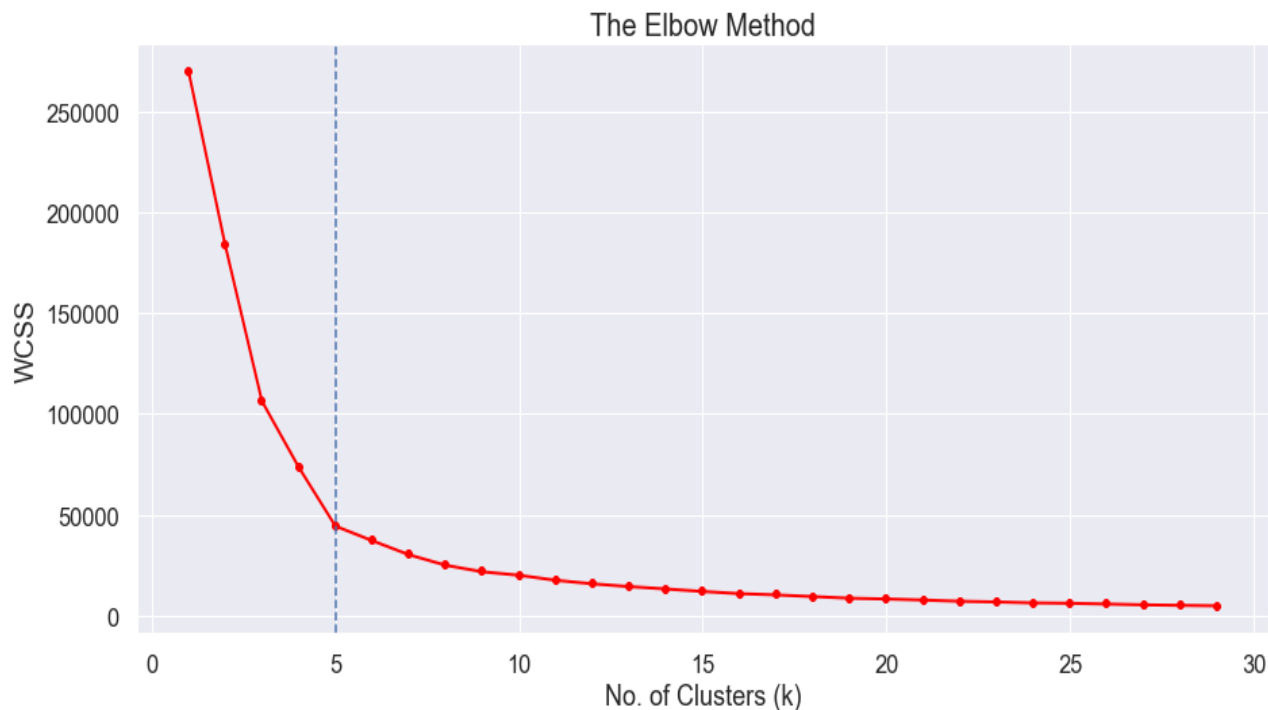
The Elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

We use the Elbow Method which uses Within Cluster Sum Of Squares (WCSS) against the 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 below formula.

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

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

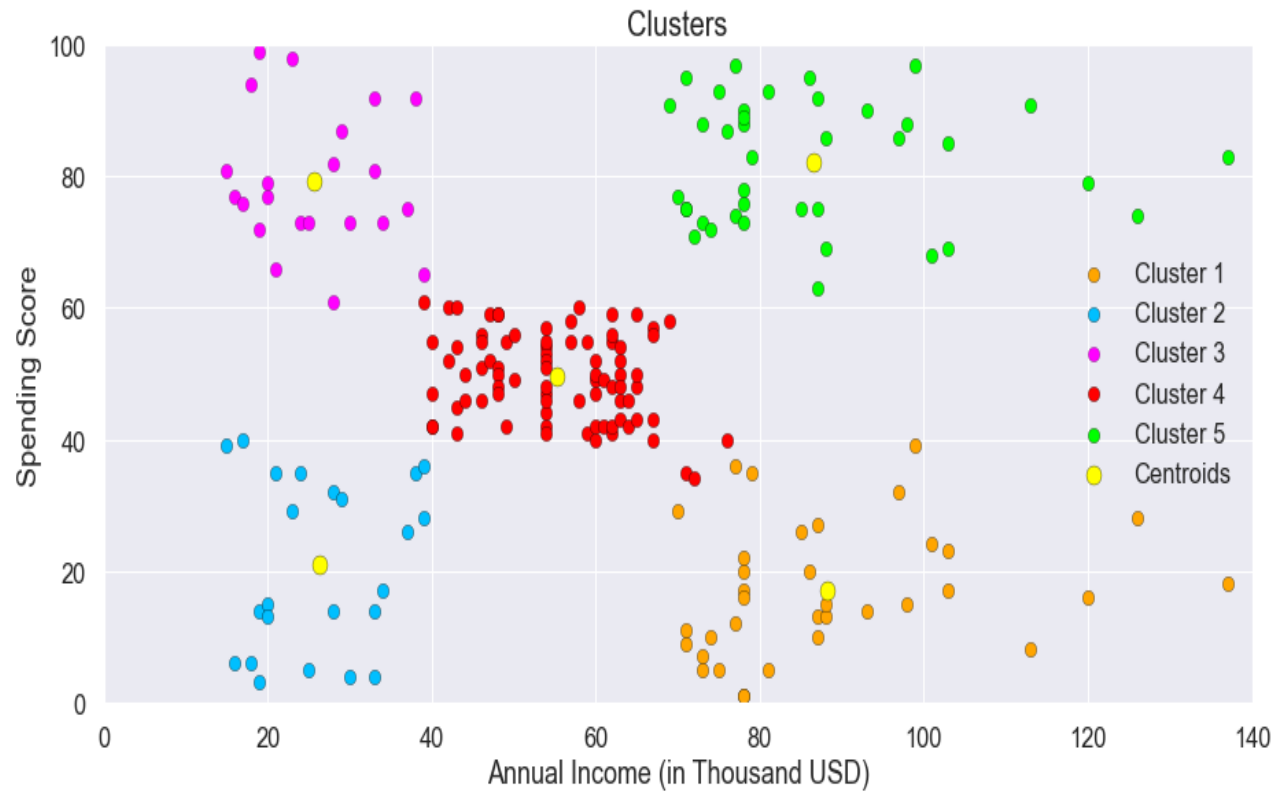
It is clear, that the optimal number of clusters for our data are 5, as the slope of the curve is not steep enough after it. When we observe this curve, we see that last elbow comes at $k = 5$, it would be difficult to visualize the elbow if we choose the higher range.



Cluster analysis :

The following
clusters are
created by the
model,

Cluster Orange
Cluster Blue
Cluster Purple
Cluster Red
Cluster Green





CLUSTER ORANGE – BALANCED CUSTOMERS:

They earn less and spend less. We can see people have low annual income and low spending scores, 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.

CLUSTER BLUE – PITCH PENNY CUSTOMERS:

Earning high and spending less.. 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 need

CLUSTER PURPLE – NORMAL CUSTOMERS:

Customers are average in terms of earning and spending, 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.

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. The shops/malls might not target these people that effectively but still will not lose them.

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. These people might be the regular customers of the mall and are convinced by the mall's facilities.



ITM UNIVERSITY GWALIOR, MP, INDIA
www.itmuniversity.ac.in

1ST RANKING
in state consecutively in last
5 years

Thank You!