UNIVERSITY
GWALIOR • MP • INDIA
"CELEBRATING DREAMS"

Major Project

Customer Segmentation using K-Means Algorithm

Submitted to
ITM UNIVERSITY GWALIOR (M.P.)

Major PROJECT Synopsis

Subject Code: CSD0703

Submitted by

Vinayak Singh [BETN1CS19068]
Pranav Kashyap [BETN1CS19048]
Abhishek Kumar [BETN1CS19108]

Under the supervision of

Project guide

Undertaken At

Dr. Sanjeev Kumar

Dept. of Computer Science & Applications
ITM UNIVERSITY, GWALIOR (M.P.)

TABLE OF CONTENTS

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 – https://drive.google.com/file/d/1mvTZo9jLorqFqH2WzLkMeeojvRZXx_83/view?usp=share_link

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

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

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

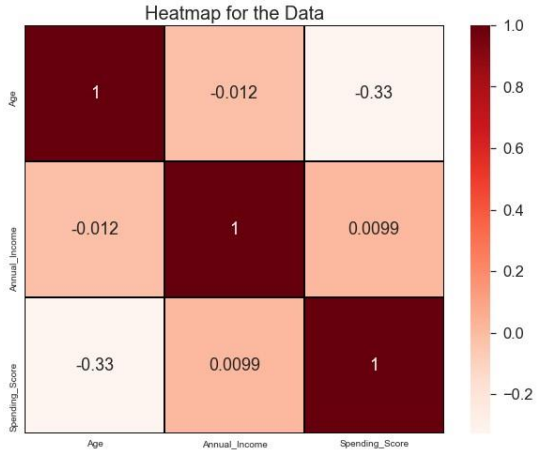
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Gender                  200 non-null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

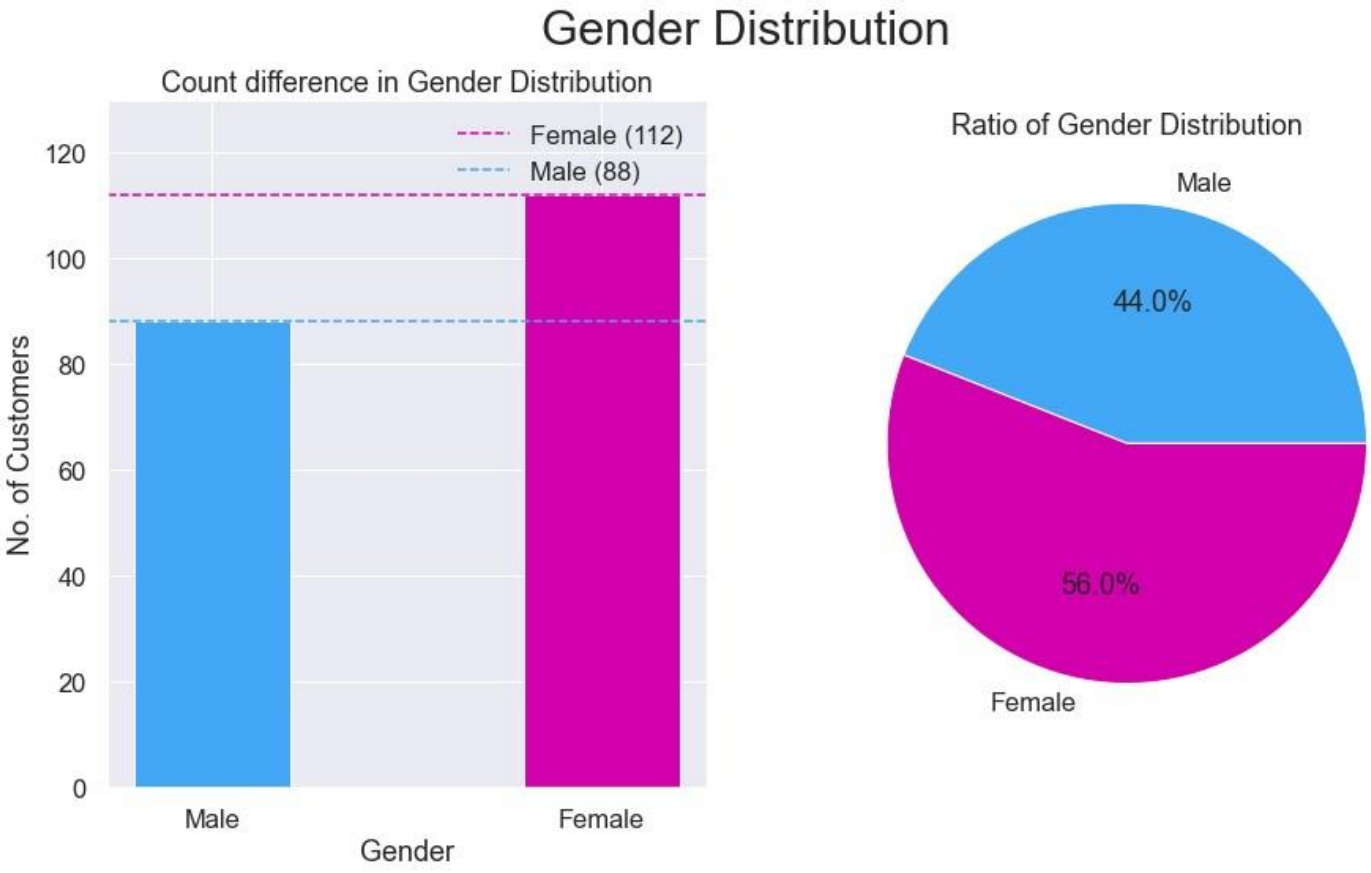
Objectives:

The objective of the project are as follows:

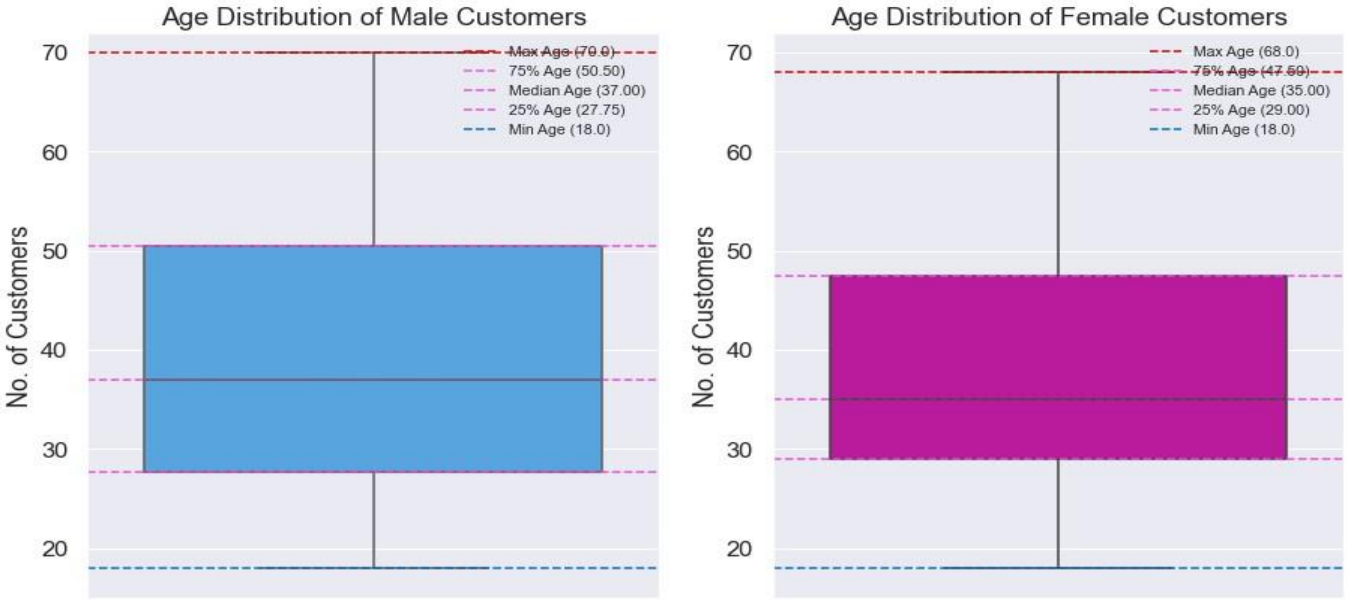
- Identify the potential customer base for selling the product.
- Implement Clustering Algorithms to group the customer base.

Exploratory Data Analysis:

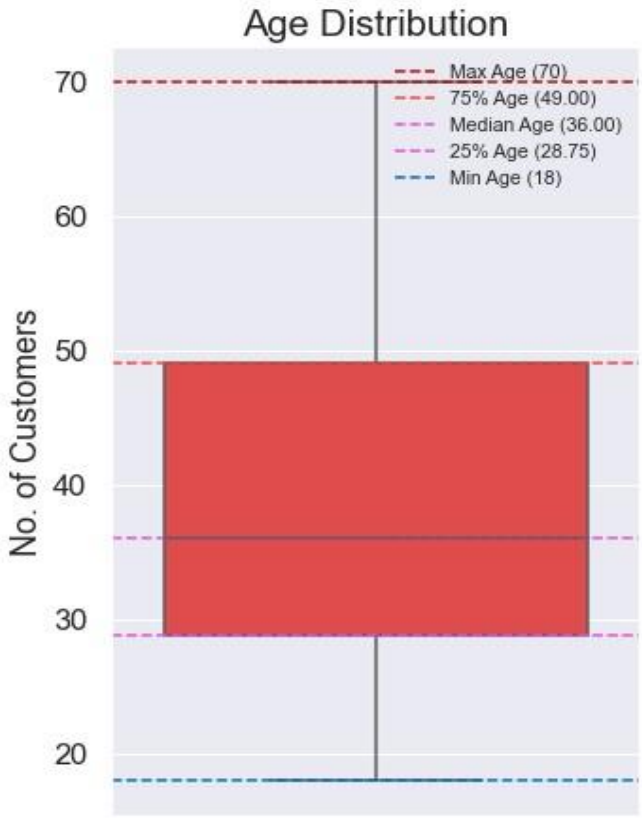
Visualization of Distribution of Males and Females:



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.

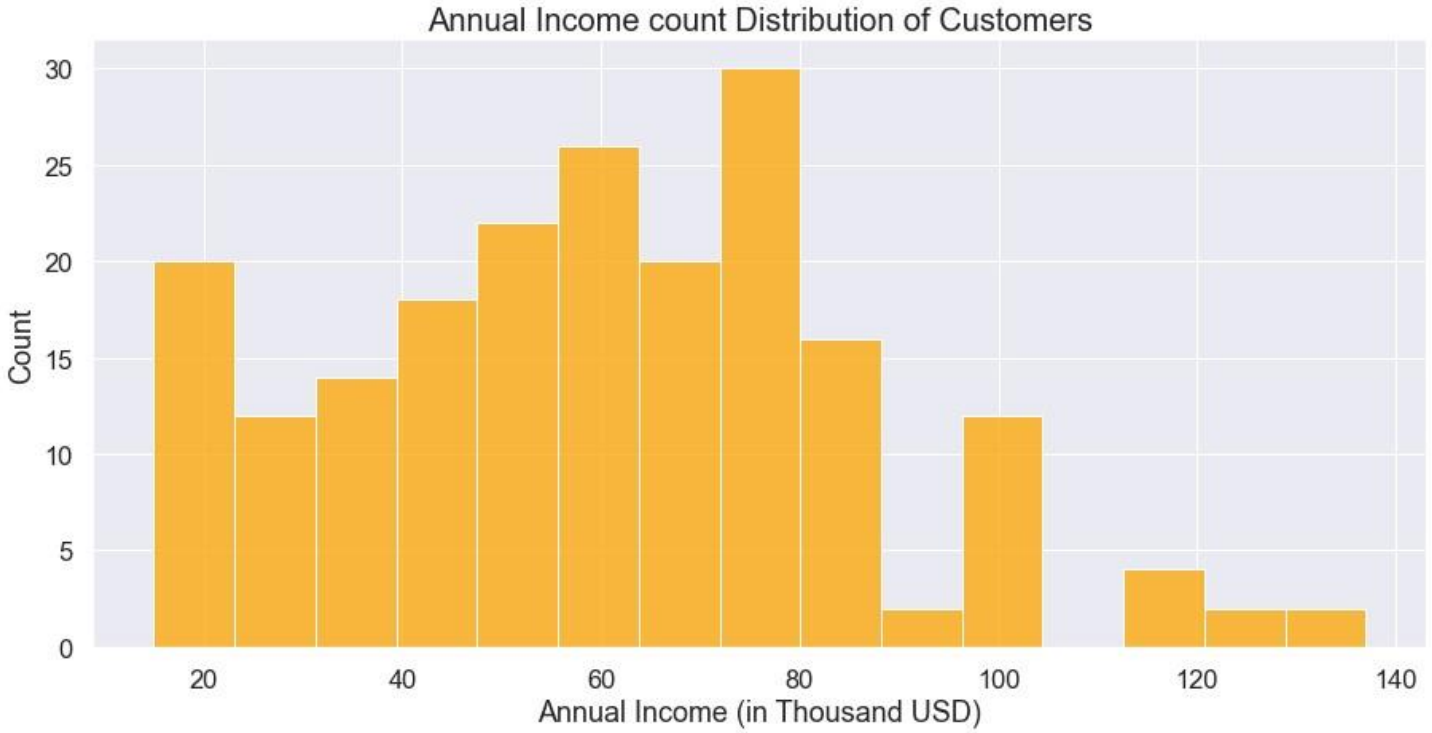


Age Analysis of Customers:

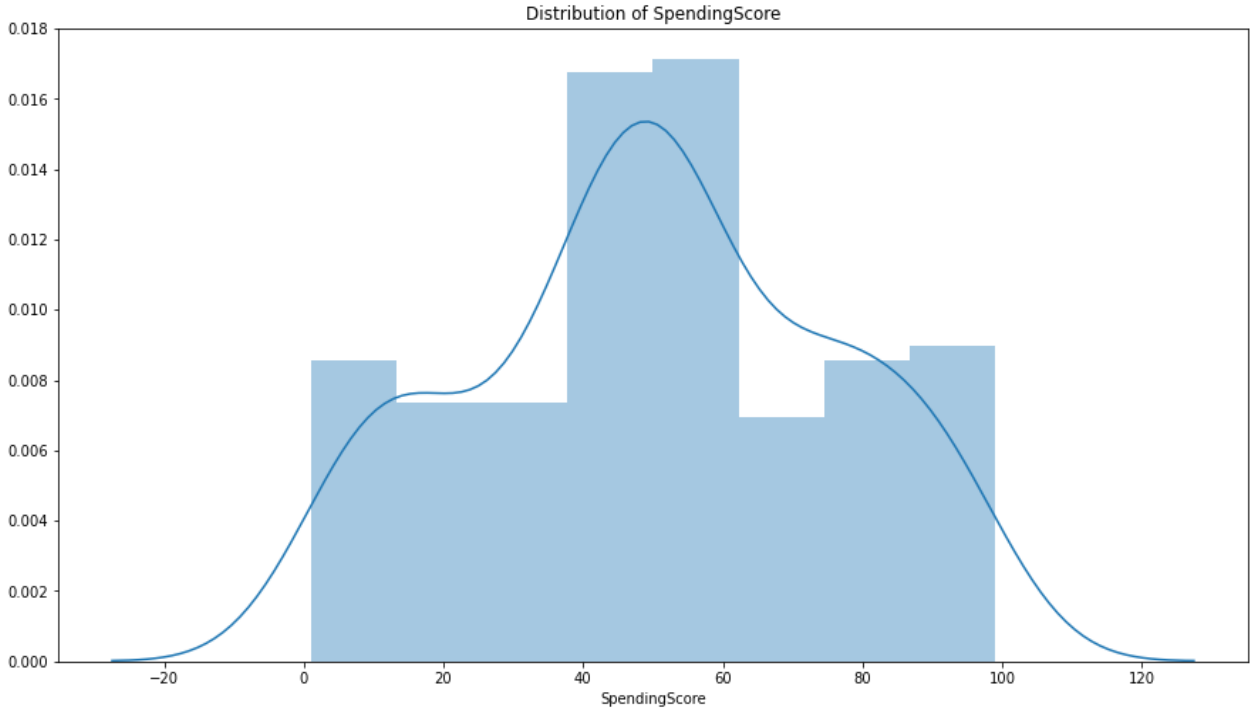


From the above boxplot, we can conclude that a large amount of ages are between 30 and 35. Min Age is 18, Max Age is 70. By comparing the age distribution of the customers, we can conclude that most of the customers were within the band between 30 to 50, where the mean is around 35 years old.

Annual Income and Spending Score Analysis:



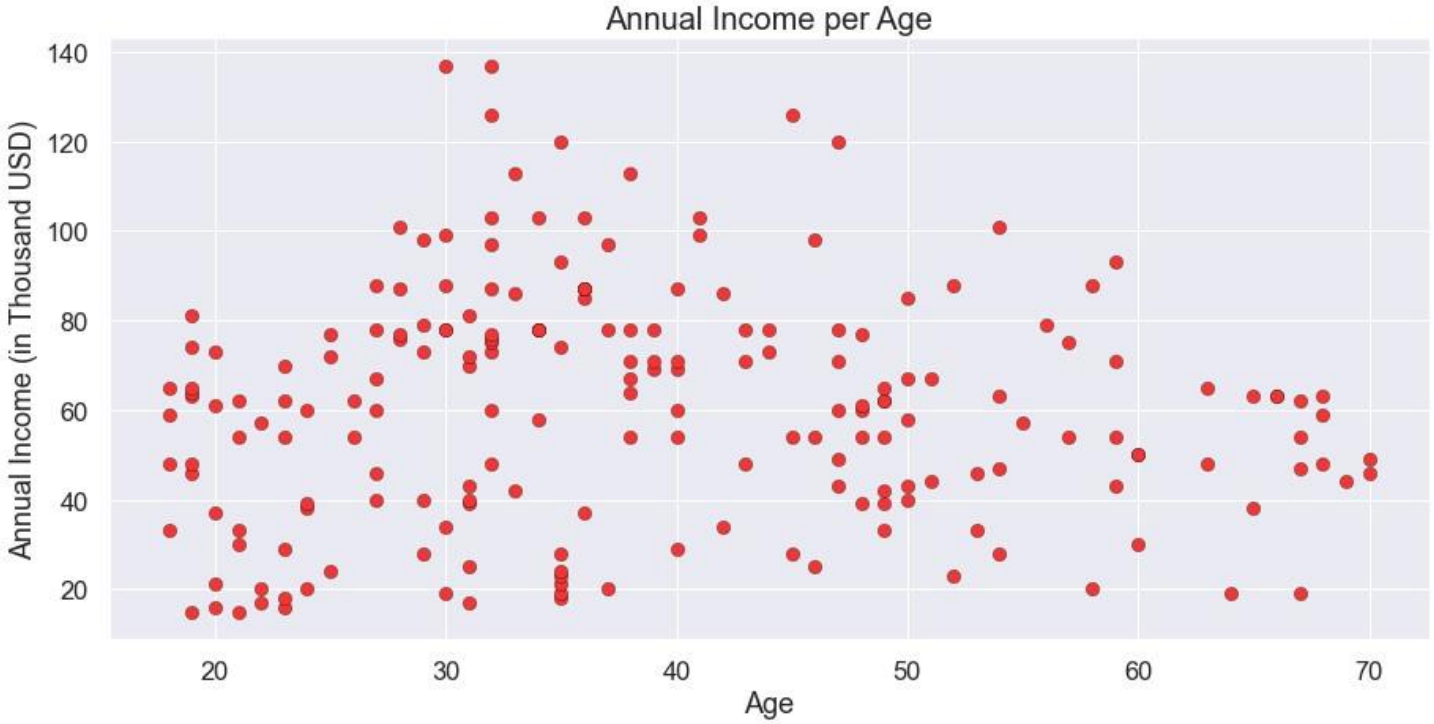
The distribution of Annal Income and Spending Score exhibited an approximation of normal distribution, with highest density around the mean of the variables. The maximum and minimum of Annual Income are 137 and 15 respectively, with the mean at 60.56. From the plot, we can see that the peak of the distribution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

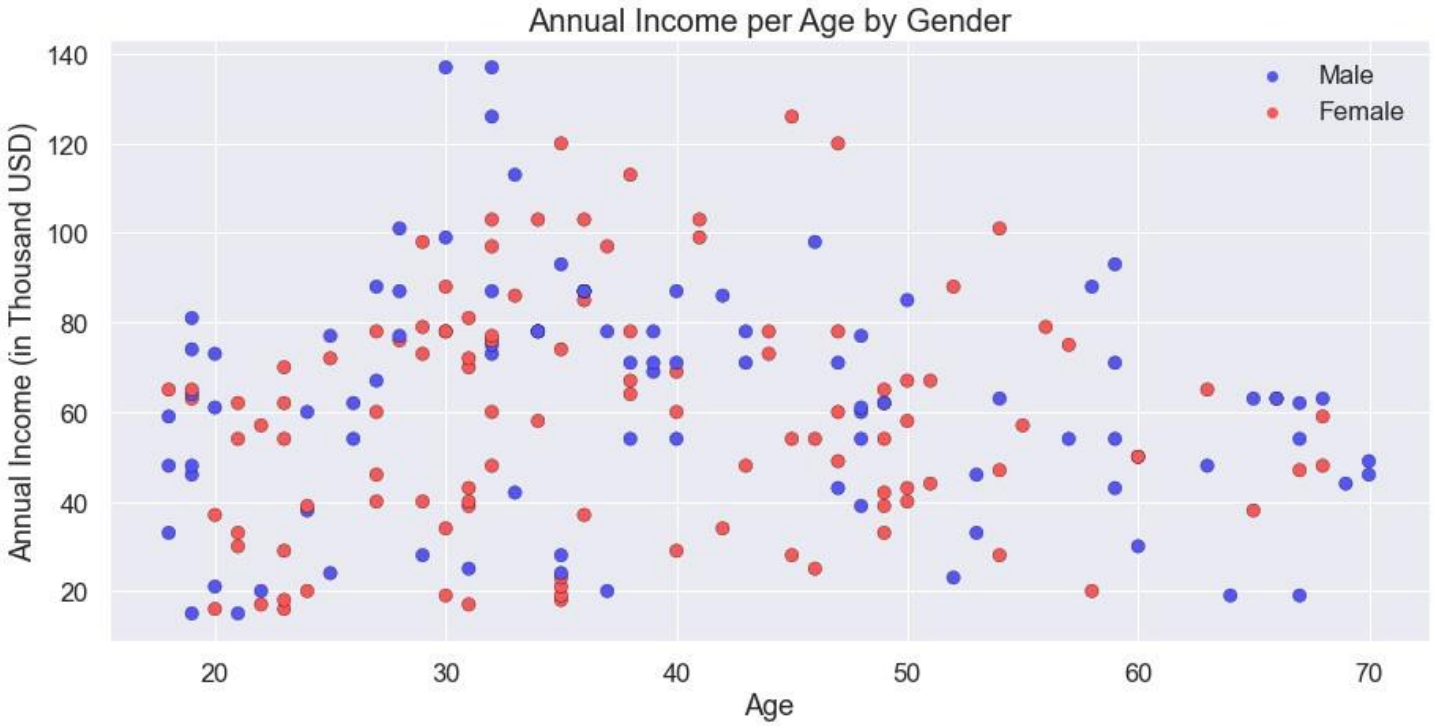


For the Spending score, the maximum and minimum are 99 and 1, while the histplot indicated that the highest number of customers have the spending score ranging from 40 to 60.

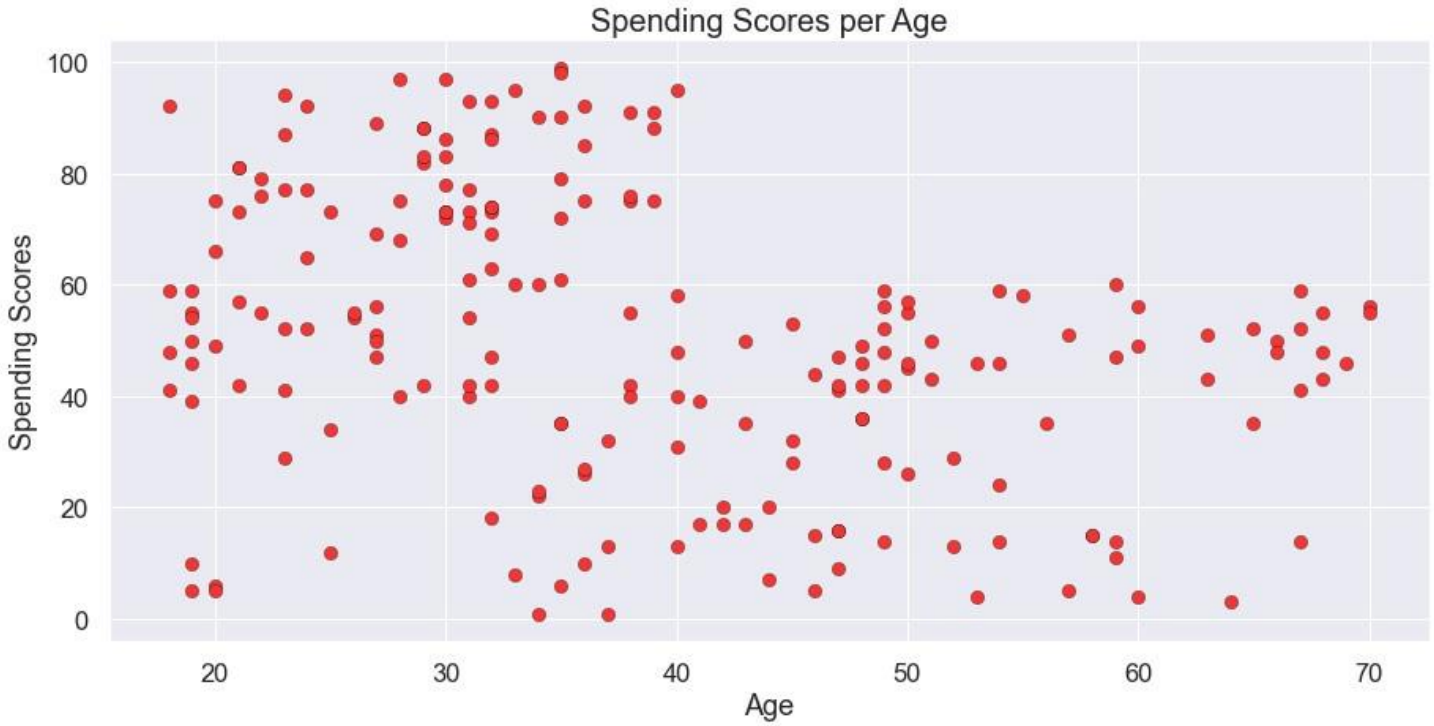
Characteristic relations:

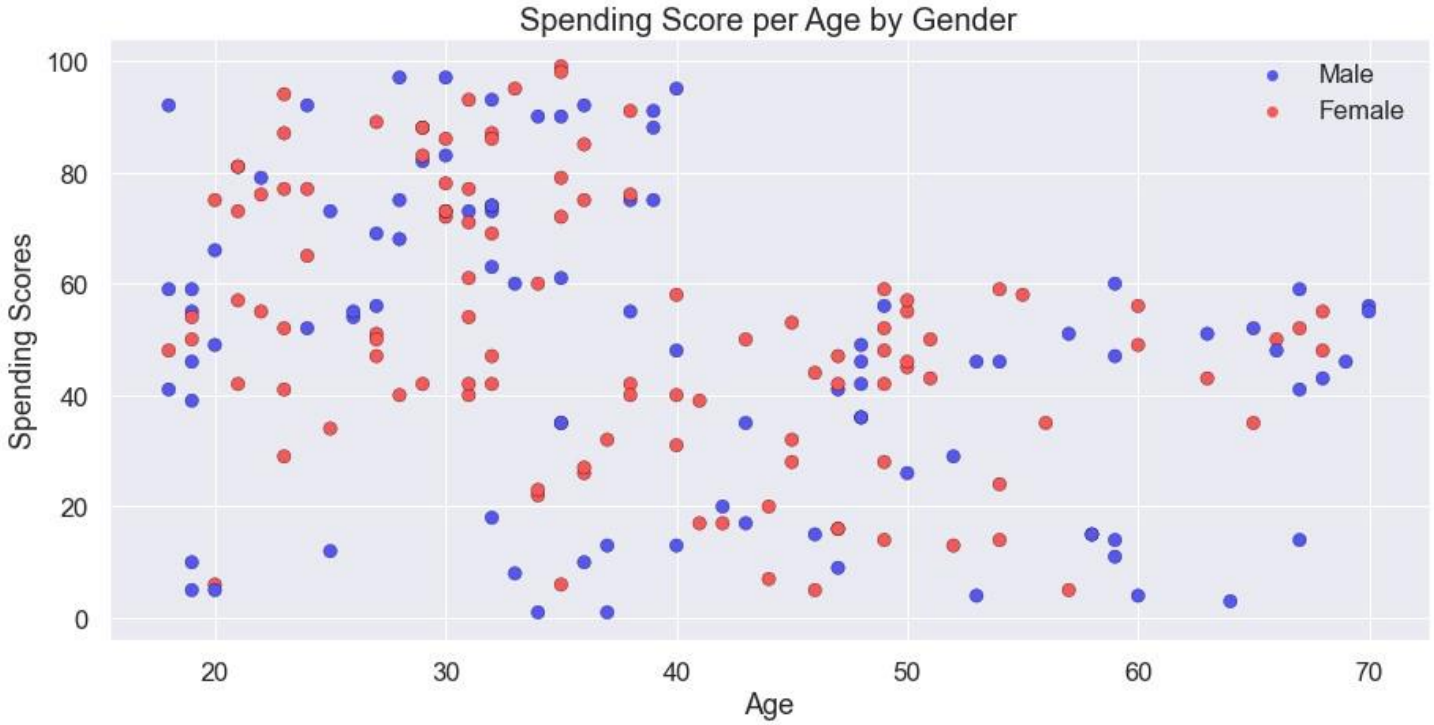
Annual Income vs Age analysis:





Spending Score vs Age analysis:


Spending Scores per Age
Spending Scores
Age
0
20
40
60
80
100
20
30
40
50
60
70


Spending Score per Age by Gender
Male
Female
Spending Scores
Age
0
20
40
60
80
100
20
30
40
50
60
70

METHODOLOGY

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 ndarray object. This encapsulates ndimensional 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 Tkinter, 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, 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.

IMPLEMENTATION

What is Clustering?

Imagine that you have a group of chocolates and liquorice 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 labelled and there is no explicit mapping of input and outputs. As such, based on the patterns present inside, clustering takes place.

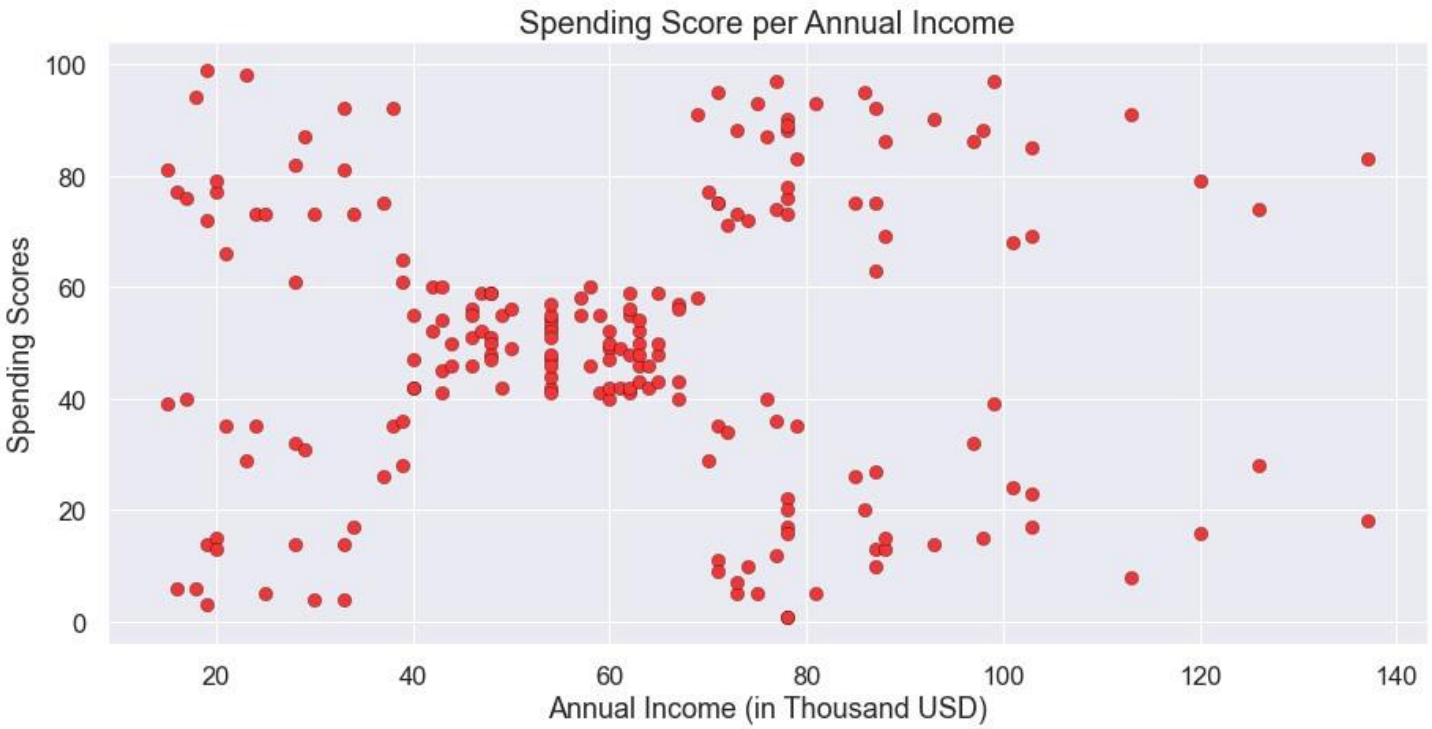
K-Means Clustering:

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.



Determining No. of Clusters Required:

The Elbow Method:

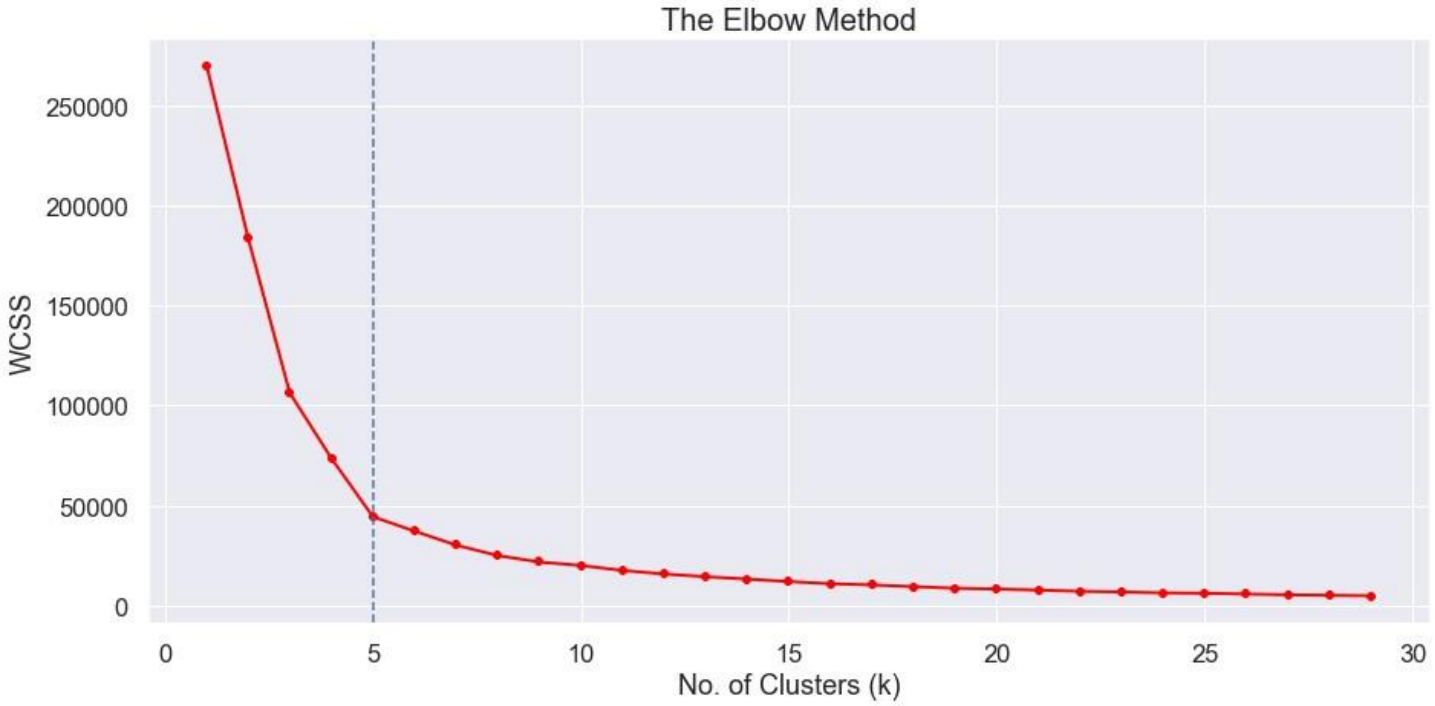
The Elbow method runs k-means clustering on the dataset for a range of values for k (say from 1-10) and then for each value of k computes an average score for all clusters. By default, the distortion score is computed, the sum of square distances from each point to its assigned center.

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 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.

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

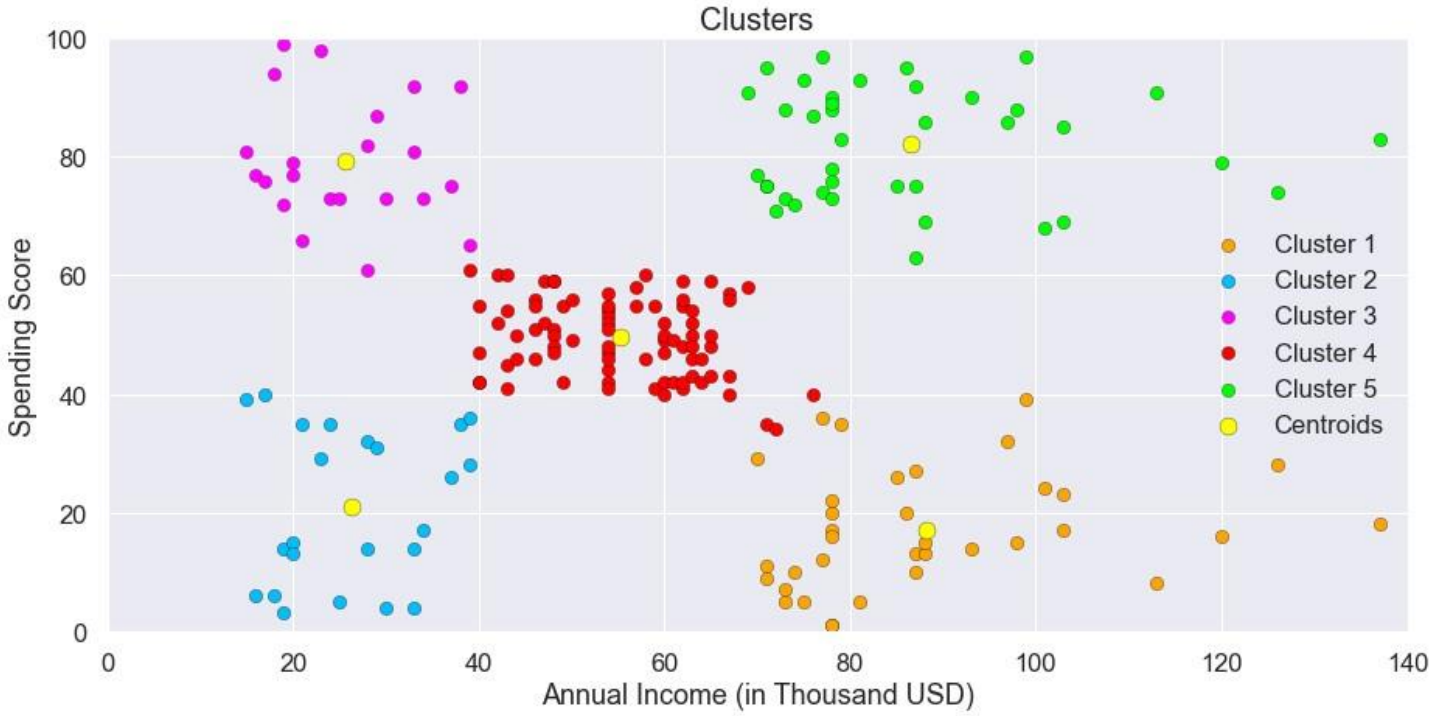
where Yi is centroid for observation Xi. The main goal is to maximize number of clusters and in limiting case each data point becomes its own cluster centroid.



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.

Clustering:

We will build the model for creating clusters from the dataset. We will use n_clusters = 5 i.e. 5 clusters as we have determined by the elbow method, which would be optimal for our dataset. We also get the centroids of the clusters by the k-means model. Visualizing the clusters will often give the fair idea about the data.



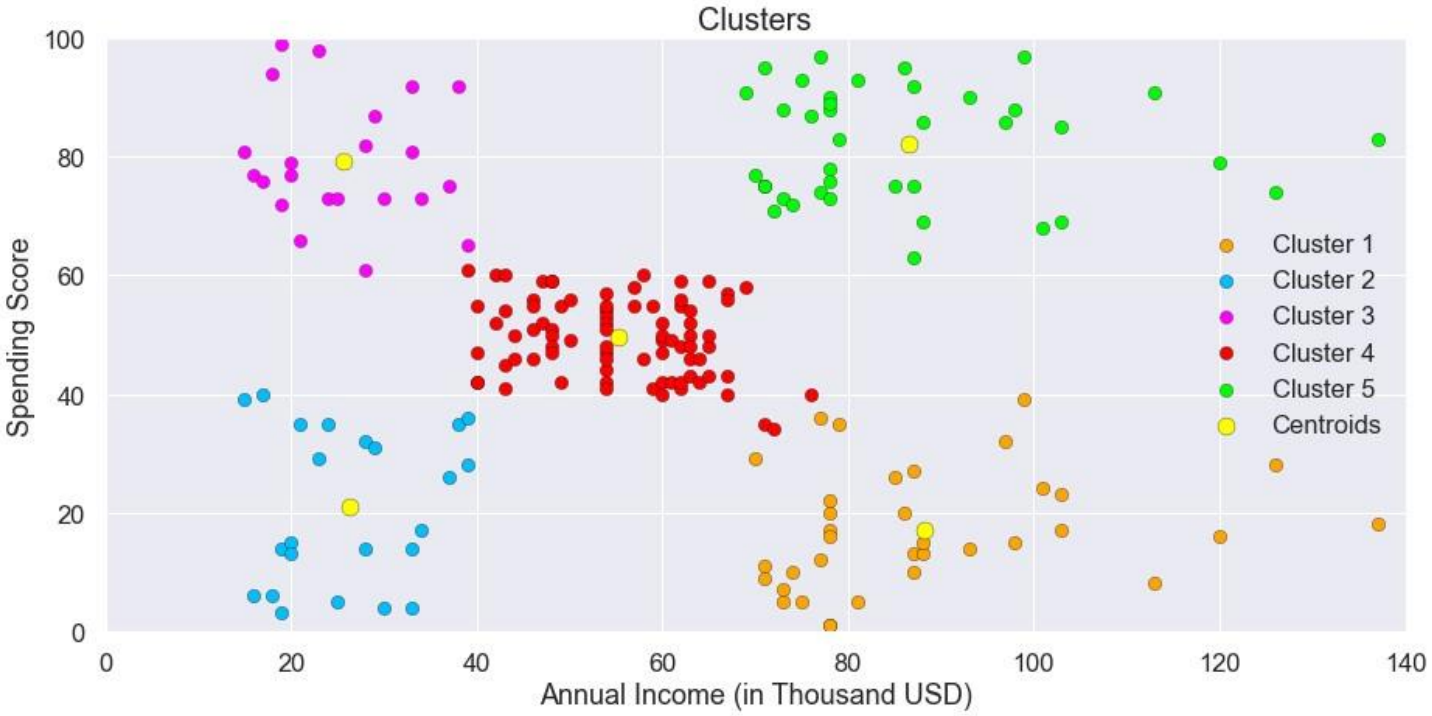
The Clusters are denoted by different colours and the centroids of the clusters is given by yellow colour. By this plot, it is easy to understand that we can divide the customers within 5 clusters of customers. Analyzing Data using the above graph becomes much easier as it gives us a visual aid for better understanding of the data. Kmeans has divided the dataset into 5 clusters based on Annual income and the spending scores of the individual customers.

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.

RESULT AND CONCLUSION

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 these 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 margin 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.

Conclusion:

Companies, Malls, super markets on Small Business Enterprises should carry out Market Basket Analysis for their business. This will enable companies to target specific groups of customers, a customer segmentation model allows for the effective allocation of marketing resources and the maximization of cross- and up-selling opportunities. When a group of customers is sent personalized messages as part of a marketing mix that is designed around their needs, it's easier for companies to send those customers special offers meant to encourage them to buy more products. Customer segmentation can also improve customer service and assist in customer loyalty and retention. As a by-product of its personalized nature, marketing materials sent out using customer segmentation tend to be more valued and appreciated by the customer who receives them as opposed to impersonal brand messaging that doesn't acknowledge purchase history or any kind of customer relationship Finally with customer segmentation Companies will stay a step ahead of competitors in specific sections of the market and identify new products that exist or potential customers could be interested in or improving products to meet customer expectations.

References

S. Dhillon and D. M. Modha, “Concept decompositions for large sparse text data using clustering,” Machine Learning, vol. 42, issue 1, pp. 143-175, 2001.

T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient K-means clustering algorithm,” IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, pp. 881-892, 2002.

MacKay and David, “An Example Inference Task: Clustering,” Information Theory, Inference and Learning Algorithms, Cambridge University Press, pp. 284-292, 2003.
Jiawei Han, Micheline Kamber, Jian Pei “Data Mining Concepts and Techniques”, Third Edition.

D. Aloise, A. Deshpande, P. Hansen, and P. Popat, “The Basis Of Market Segmentation” Euclidean sum-of-squares clustering,” Machine Learning, vol. 75, pp. 245-249, 2009.

S. Dasgupta and Y. Freund, “Random Trees for Vector Quantization,” IEEE Trans. on Information Theory, vol. 55, pp. 3229-3242, 2009.

Puwanenthiren Premkanth, ―Market Segmentation and Its Impact on Customer Satisfaction with Especial Reference to Commercial Bank of Ceylon PLC.‖ Global Journal of Management and Business Research Publisher: Global Journals Inc. (USA). 2012. Print ISSN: 0975-5853. Volume 12 Issue

https://github.com/mayursrt/customer-segmentation-using-k-means