**Project Report**

**Customer Segmentation**

**Abstract:**

Customer segmentation and pattern extraction is one of the key aspects of business decision support system. In order to grow the business intelligently in competitive market, identification of potential customer should be done timely. This paper proposes an approach for determining target customers and also discover their associative buying patterns using K-means clustering algorithm. After identification of targeted customers and their associative buying pattern, the business managers take the strategic profitable decisions accordingly.

**Table of Contents:**

1. **Introduction**

* **What is customer segmentation?**
* **Why customer segmentation?**

1. **Methodology**

* **Packages used**
* **Exploratory Data Analysis**
* **K-means Clustering**
* **Elbow method**

1. **Implementation**

* **Data pre-processing**
* **EDA**
* **Gender Distribution**
* **Uni-variate Analysis**
* **Bi-variate Analysis**
* **K-Means Clustering**

1. **Conclusion**

**Introduction :**

Customer segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately.

In business-to-business marketing, a company might segment customers according to a wide range of factors, including:

* Industry
* Number of employees
* Products previously purchased from the company
* Location

In business-to-consumer marketing, companies often segment customers according to demographics that include:

* Age
* Gender
* Annual income
* Spending score
* Life stage (single, married, divorced, empty-nester, retired, etc.)

Segmentation allows marketers to better tailor their marketing efforts to various audience subsets. Those efforts can relate to both communications and product development. Specifically, segmentation helps a company:

* Create and communicate targeted marketing messages that will resonate with specific groups of customers, but not with others (who will receive messages tailored to their needs and interests, instead).
* Identify ways to improve products or new product or service opportunities.
* Establish better customer relationships.
* Test pricing options.
* Focus on the most profitable customers.
* Improve customer service.
* Upsell and cross-sell other products and services.

**Methodology:**

In this project we used dataset of mall customers having details of 200 customers. We used Jupyter Notebook for coding. We used different packages for our purpose.

* Numpy
* Pandas
* Seaborn
* Matplolib
* Sklearn

**Exploratory Data analysis (EDA):**

Exploratory Data Analysis is a process of examining or understanding the data and extracting insights or main characteristics of the data. EDA is generally classified into two methods, i.e. graphical analysis and non-graphical analysis.

EDA is very essential because it is a good practice to first understand the problem statement and the various relationships between the data features before getting your hands dirty.

Technically, the primary motive of EDA is to

* Examine the data distribution
* Handling missing values of the dataset (a most common issue with every dataset)
* Handling the outliers
* Removing duplicate data
* Encoding the categorical variables
* Normalizing and Scaling

**Kmeans clustering.**

We use K-means clustering Algorithm for segmentation of customers. 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. K-Means 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

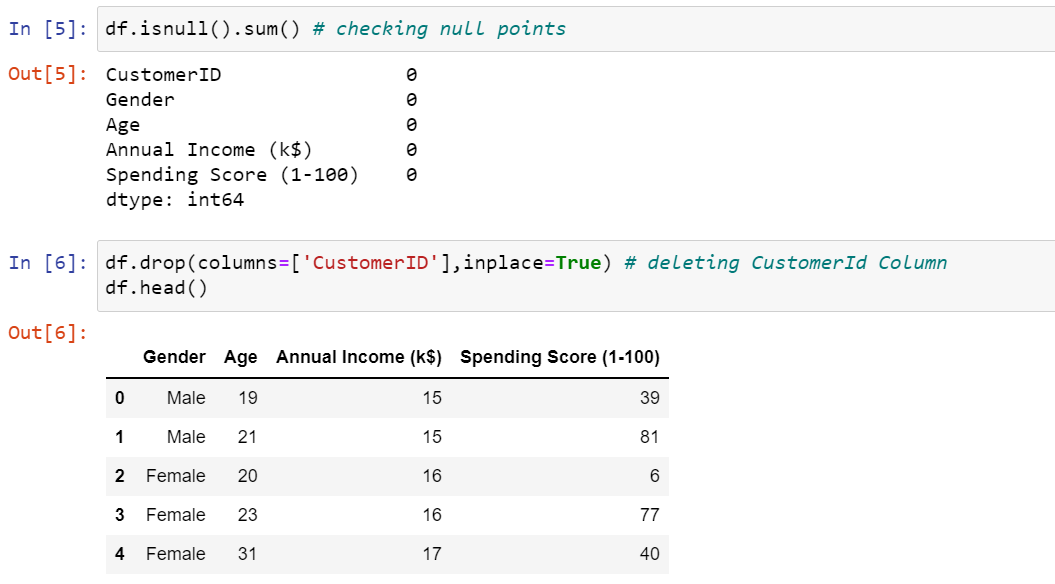
**Elbow method**

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The **Elbow Method** is one of the most popular methods to determine this optimal value of k.  
We now demonstrate the given method using the K-Means clustering technique using the **Sklearn** library of python.

**Implementation:**

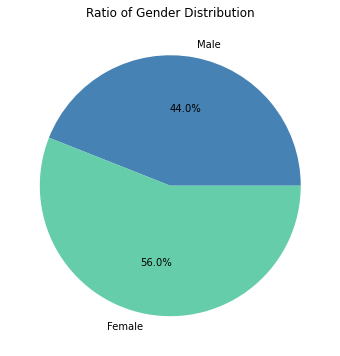
**Data pre-processing:**

We will process the data by checking for null values in the data set and removing the unnecessary column “*Customer-Id”*.



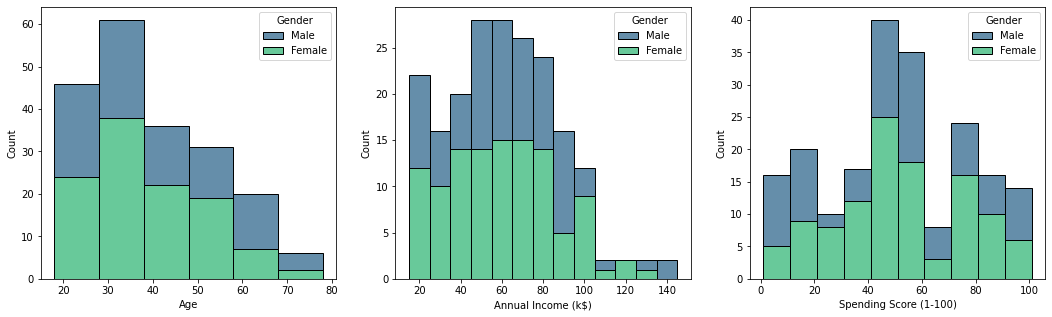
**Gender Distribution Analysis:**

**1.Male and Female distribution:**



* Female Customers are 6% more compared to Male customers.
* 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.

**2. Age Annual income and spending score:**



#### Age

* Female dominate in 20-50 (years old) age group

#### Annual Income

* Most of the customers earn 55k-80k annual income (Female dominate)

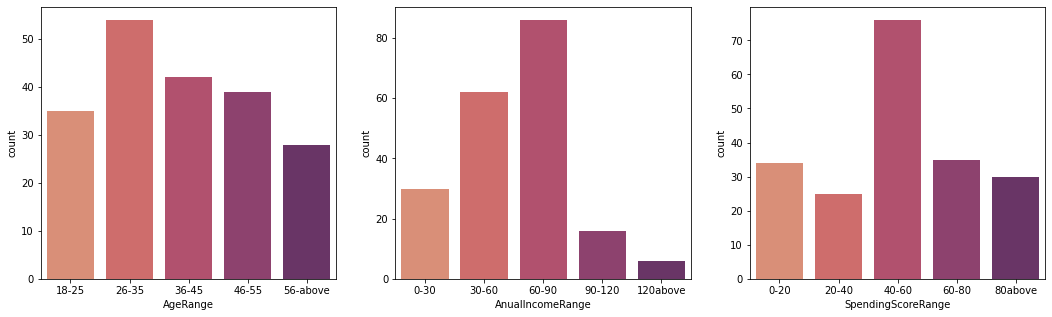
#### Spending Score

* Male Dominate in spending score 0-20 and female dominate in spending score 20-100
* Most Female customers have spendingscore between 40 and 60 .
* Women tend to spend more than men

**Exploratory Data Analysis:**

we perform EDA univariately as well as bi-variately.

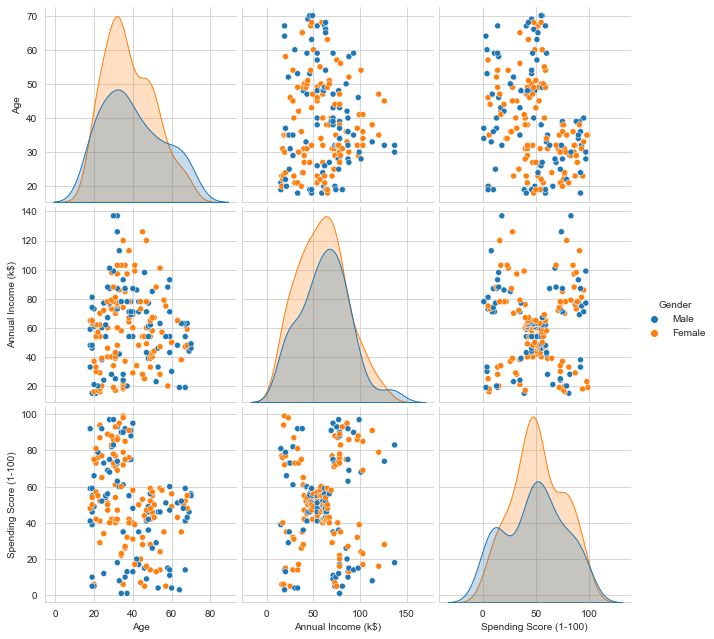
**Uni-Variate Analysis:**



* Most customers have the age range of 26-35 years
* Most customers have an Annual income range of 60-90 $k
* Most customers have a spending score range of 40-60

**Bi-Variate Analysis:**

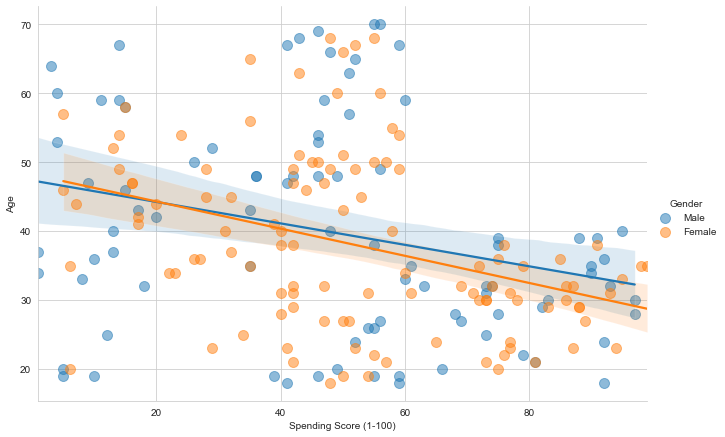
**1. Pair plot**



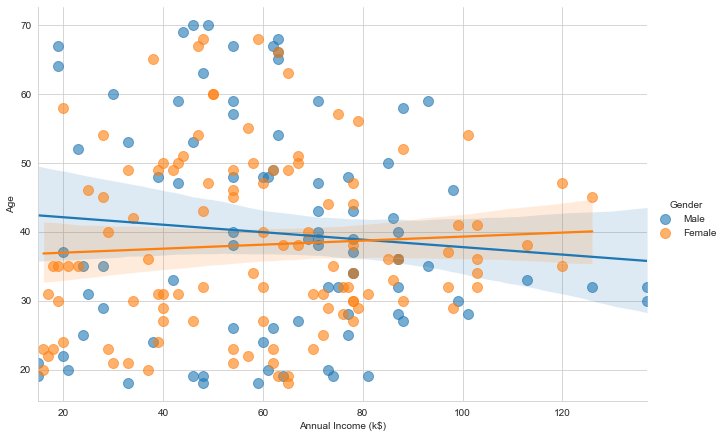
* Age, Income and Spending Score features are multimodal distributions
* Annual Income vs Spending Score form dense regions in the plot

**2**.**Linear Regression:**

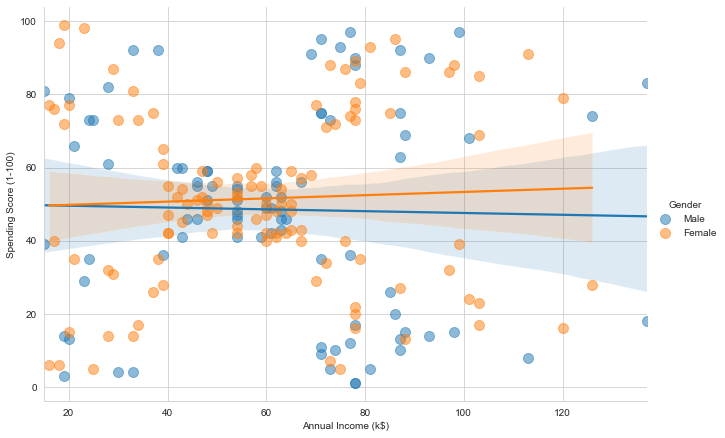
**Age vs Spending Score**

****

**Age vs Annual Income**

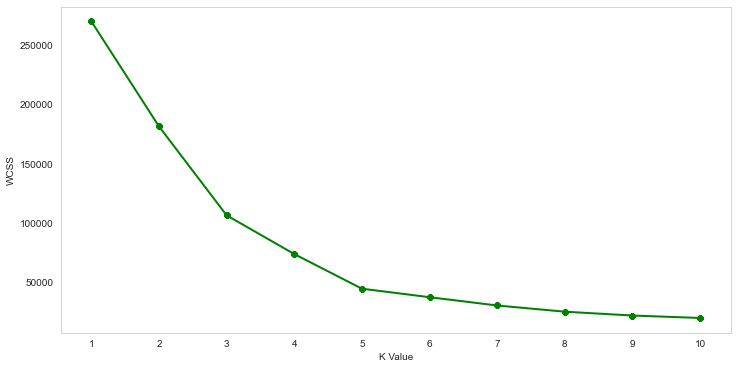
****

**Spending Score vs Annual Income**

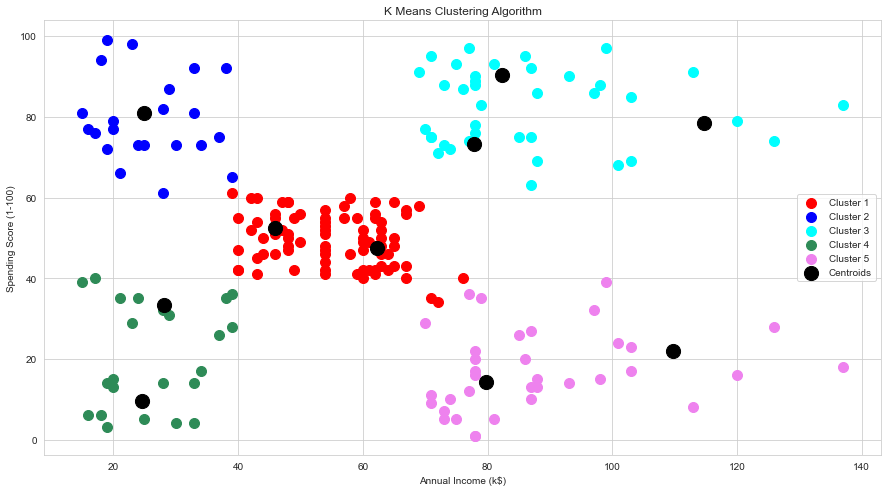
****

##### spendingscore and annual income are our main feautures for performing clustering on Mall Customers

**K-Means Clustering:**

**Elbow method**

**K-means Clustering**

****

### **Conclusion**

### **Insights**

* Here we have performed Kmeans clustering between Annual Income and Spending Score.
* There are 5 differente groups, each group having specific characteristics.
* There are costumers with high, medium and low annual income and with high, medium and low spending score. For each one, the administrators could plan different actions.

### **Clusters**

* CLUSTER 1 - Medium Income(40-70 k) and Medium SpendingScore(40-60)
* CLUSTER 2 - Low Income(20-40 k) and High SpendingScore(60-100)
* CLUSTER 3 - High Income(70-140 k) and High SpendingScore(60-100)
* CLUSTER 4 - Low Income(20-40 k) and Low SpendingScore(0-40)
* CLUSTER 5 - High Income(70-140 k) and Low SpendingScore(0-40)

###### Most of the customers belong to CLUSTER 1 having medium Annual Income and medium SpendingScore