**PROJECT TITLE - CUSTOMER SEGMENTATION**

**TEAM MEMBER**

**210621104029 – DHARUNRAJ N**

**APPLIED DATA SCIENCE**

**Phase-3 Development Part 1**

# 1.Introduction

The main goal of this project is to use advanced techniques in data science to group customers based on how they behave, what they like, and information about them like their age or location. This grouping strategy is really important because it helps businesses create special marketing plans for each group of customers. This can make customers happier and more likely to keep coming back to the business. The project has several important steps like getting the data, cleaning it up, making new useful information from the data, using different ways to group the customers, showing the data in pictures, and understanding what the groups mean.

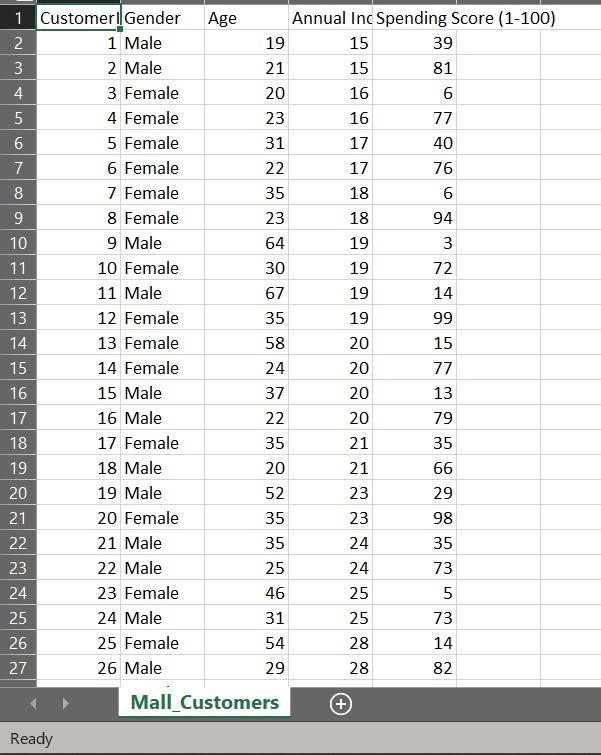
## 2.Dataset

The data set is collected from Kaggle which is a leading collaborative data science platform.

Dataset Link :<https://www.kaggle.com/datasets/akram24/mall-customers>

All the data are stored in the Mall\_Customers.csv (Comma Separated Values) format, which is used to store the data efficiently.

Using pandas, we can use the .csv format for Data Processing and Manipulation



## The columns that are present are

1. CustomerID
2. Gender
3. Age
4. Annual Income
5. Spending Score

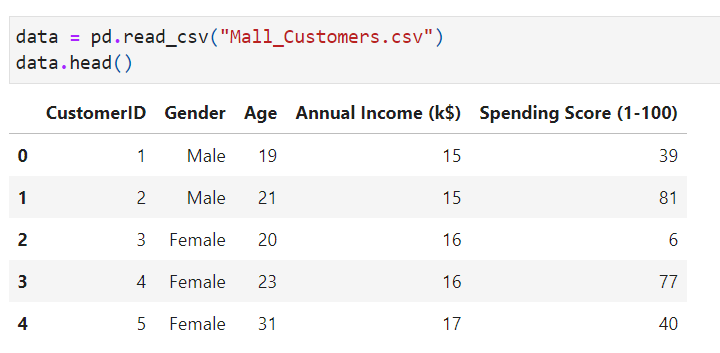
CustomerID only contains values that are useless, so we will not be using it.

We will use all the other columns except CustomerID for Exploratory Data Analysis (EDA)

For clustering, we will be using Annual Income and Spending Score column.

## 3. Loading the Dataset

## We will load the dataset named “Mall\_Customers.csv” inside our python notebook using the pd.read\_csv() method.

**4. Pre-processing**

Data pre-processing is an important step in the data science process. It refers to the cleaning, transforming, and integrating of data in order to make it ready for analysis

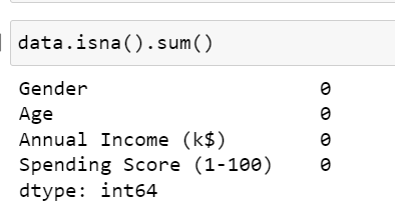
**Removing Useless Column**



We will remove the “CustomerID” column from our data. Because the values of this column are useless.

**Checking for Missing Values**

Having missing values in dataset will affect the machine learning model’s performance. So, we need to make sure that our dataset doesn’t contains any form of missing values



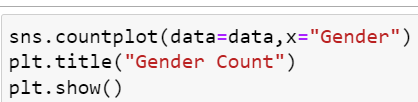
We can say our data set doesn’t contain any missing values. After this we can perform exploratory data analysis for our data

**5. Exploratory Data Analysis**

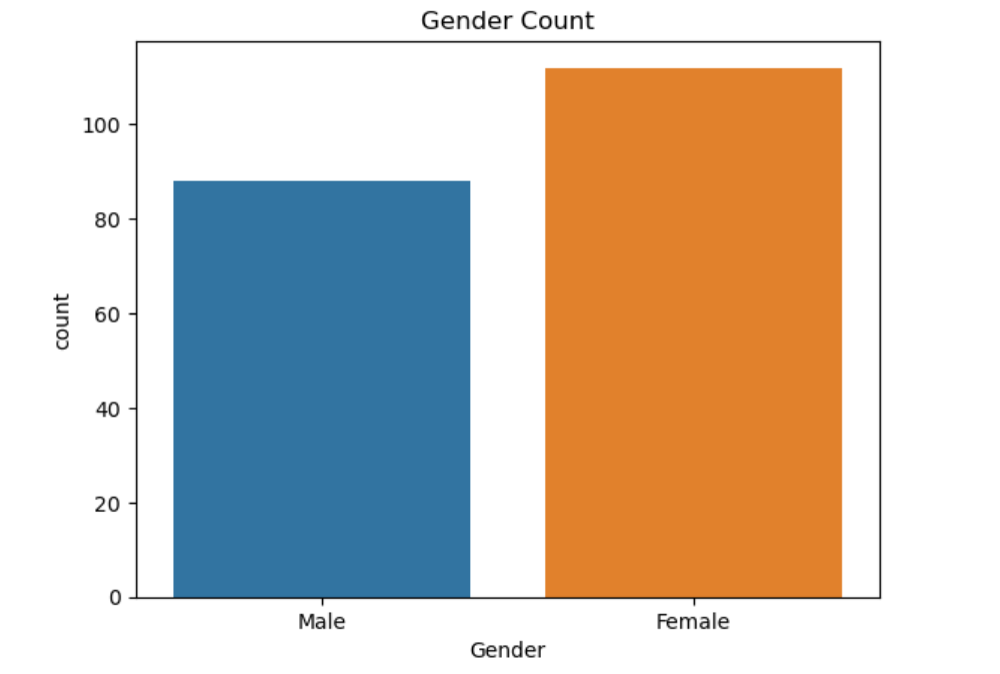
Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods

We will perform data analysis for our given data

**Age**

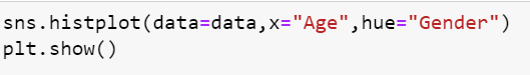
****

We will use a categorical plot to define the count of each gender

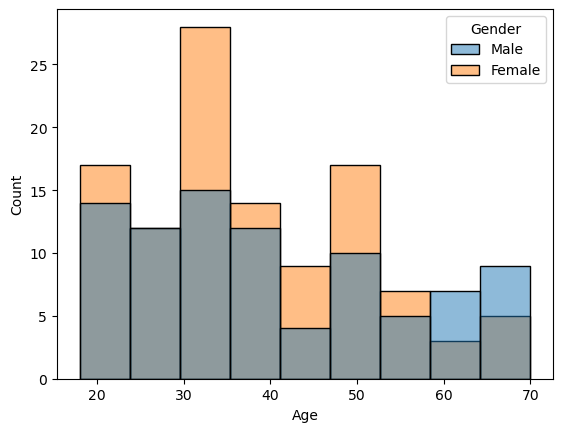


It seems there are more "Female" customers than "Male" customers

**Age with Gender**

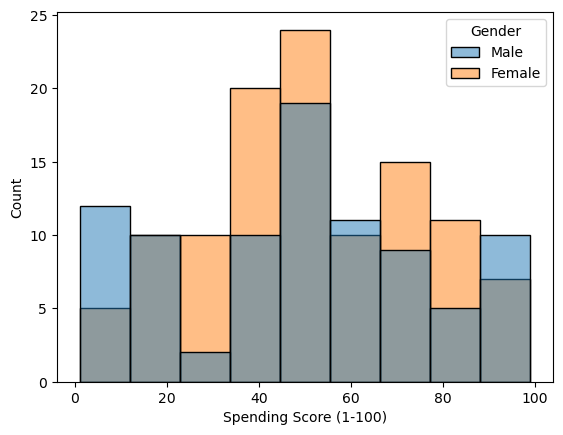
****

We will use a histogram plot to find the distribution of age within certain intervals and for each gender

****

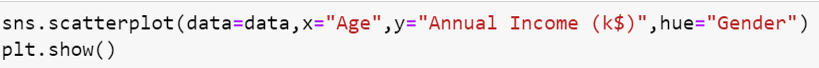
**Spending score with Gender**

****

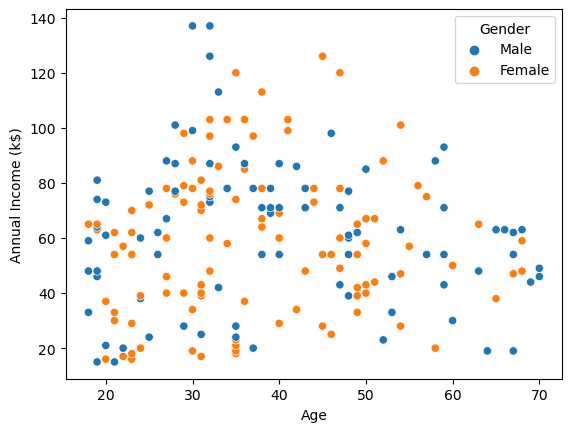
****

"Female" customers have most of the 70-100 spending score

**Age and Income**

****

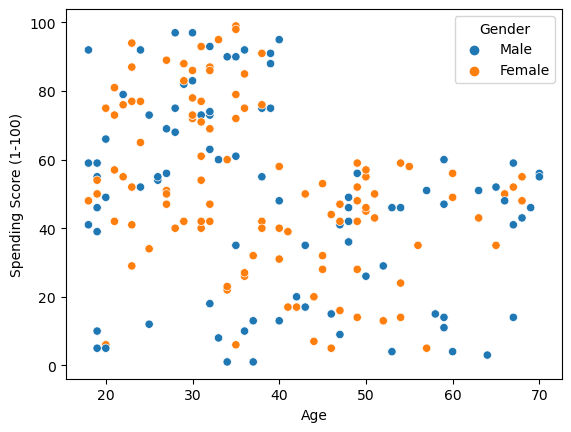
We will use a scatterplot to plot Age against Annual Income

****

People with age less than 30 and greater than 60 has low "Annual Income"

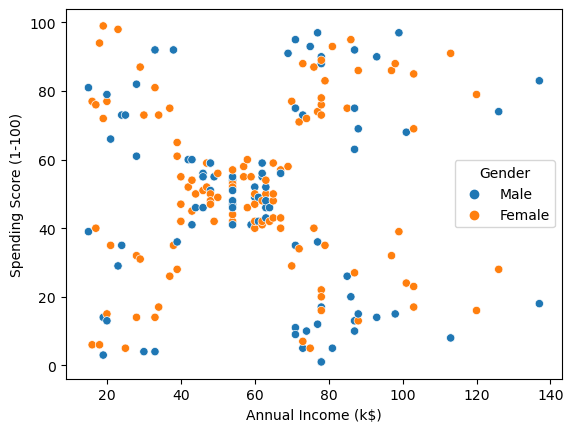
**Age and Spending Score**





**Annual Income and Spending Score**

****

****

There is definitely a pattern in we can segment the given customers based on Annual Income and Spending Score. We will explore it using KMeans in the next phase