**MINI PROJECT REPORT**

**ON**

**CUSTOMER SEGMENTATION USING**

**K-MEANS CLUSTERING**

(SESSION: 2023-24)

SUBMITTED BY

**YASH DHIMAN**

**UNINVERSITY ROLL NO - 2119448**

SUBMITTED TO

**DR. SATVIK VATS**

Computer Science and Engineering

Graphic Era Hill University

Dehradun, India

A logo with a mountain and a light bulb

Description automatically generated

**DEPARTMENT OF COMPUTER SCIENCE AND**

**ENGINEERING**

**TABLE OF CONTENTS:**

**1.ABSTRACT**

**2.INTRODUCTION**

**3.RELATED WORKS**

**4.DATASET OVERVIEW**

**5.METHODOLOGY**

* DATA COLLECTION AND PROCESSING
* SPLITTING THE FEATURES AND TARGET
* SPLITTING THE DATA INTO TRAINING AND TEST DATA
* MODEL TRAINING
* MODEL EVALUATION

**6.RESULT**

**7.CONCLUSION**

**8.REFERENCES**

**ABSTRACT:**

In today’s era defined by data-driven techniques, operations, enterprises confront increased competition and the complex venture of grasping the diverse range of their customer’s base. To hinder these challenges, businesses adopt the advanced techniques, prominently including machine learning algorithms such as K-means clustering using elbow plot, for the purpose of customer segmentation and classification. This research paper focusses into the application of the elbow method, a refined amplification technique, to enhance the clustering process. By utilizing the elbow method, companies are empowered and have ability and more precise segmentation.

This improvement in clustering techniques equipped by businesses and Mall owners with the capacity to derive insights from their customer base, enabling them to tailor their aspect of good customer strategies, decision-making and segmentation processes to the distinct attributes and their preferences of each customer segment and divide them into their appropriate clusters (clusters ranging from 1 to N).

Keyword*s:* Clustering algorithms, Customer Segmentation, K-Means Algorithm, Elbow Method, Visualization.

**INTRODUCTION:**

In today's combative landscape, businesses that have structured environment are predominantly depends on data extraction methodologies to plan their perception from bigger datasets that have relation with the past. Mining of data involves extracting patterns from data that bolster decision-making strategies, often in consideration was clustering techniques to group similar data. Customer segmentation or classification, a prominent aspect of marketing, divides the customer’s base into small segments based on similar characteristics such as gender, age, and spending traits. This segmentation enables personalized marketing strategies, better support in decision making, and improved management of products. This paper focuses on identifying customer segments using the K-means clustering algorithm, with the help of elbow method in determining the optimal clusters.

In the Methodology section we firstly define the definition of K-means and its process of execution we also define its advantages. The first step in the methodology section is the data collection section in which a balanced dataset was taken from Kaggle a data science application. The next section is splitting of data between training and test data. Then there is most important section which is the model training section in this section the scatterplot is being built between the two important columns of the dataset that are spending score and annual income. Then at last there is a evaluation section in which we use the formula of Within Cluster sum of Squares whose formula is defined in the methodology section.

**RELATED WORKS:**

1. In 2022 a study was done by Yash Kushwaha, Deepak Prajapati [1] that today’s world depends mainly on data analysis, of vast datasets which is most important aspect. Business Related strategies adhere to today’s era conditions, strong competition is the far most important aspect and there is a crucial need for innovation. In this phase, machine learning plays a far most important role in extracting patterns that is hidden within data for decision-making. Clustering techniques is the well-structured approach for Customer Segmentation, that helps the businesses to specifically target the customer segments. This paper focuses on the K-means algorithm for clustering, helping businesses to understand better and to get improved outcomes from their annual customers.
2. In 2022 a study was done by N. Sri Lakshmi, V. Sai Pramod, Shakeera, A. Praveen, Shaik. Kabeer [2] that mainly focusses to cluster credit card users based on their characterization of using it, aiding marketing departments in tailoring products and services. Target groups are identified using machine learning techniques and sorting algorithms. K-means clustering facilitates grouping users with similar attributes, enhancing marketing strategies. The elbow method determines the optimal number of clusters, and data visualization provides insights. Customer segmentation enables the development of customer-centered products and successful marketing strategies, unravelling hidden patterns in data for informed decision-making and targeted customer engagement in a competitive market.
3. In 2021 a study was done by Jiawei Han, Micheline Kamber, Jian Pei [3] that "Data Mining: Concepts and Techniques" elucidates the principles and tools for processing collected data, crucial for various applications. It serves as a reference for Knowledge Discovery from Data (KDD), emphasizing feasibility, usefulness, effectiveness, and scalability with large datasets. Covering data mining, preprocessing, processing, and warehousing, it explores data warehouses, OLAP, and data cube technology. Methods for mining frequent patterns, associations, correlations, classification, clustering, outlier detection, and current trends in data mining are extensively discussed.
4. In 2018 a study was done by I. S. Dhillon and D. M. Modha [4] that Unlabeled document collections pose contemporary challenges. Represented as high-dimensional, sparse vectors, text documents require specialized clustering algorithms like spherical k-means. This study empirically demonstrates fractal-like behavior in clusters and introduces concept decompositions for accurate document vector approximation, showing their efficacy compared to singular value decompositions.

**DATASET OVERVIEW:**

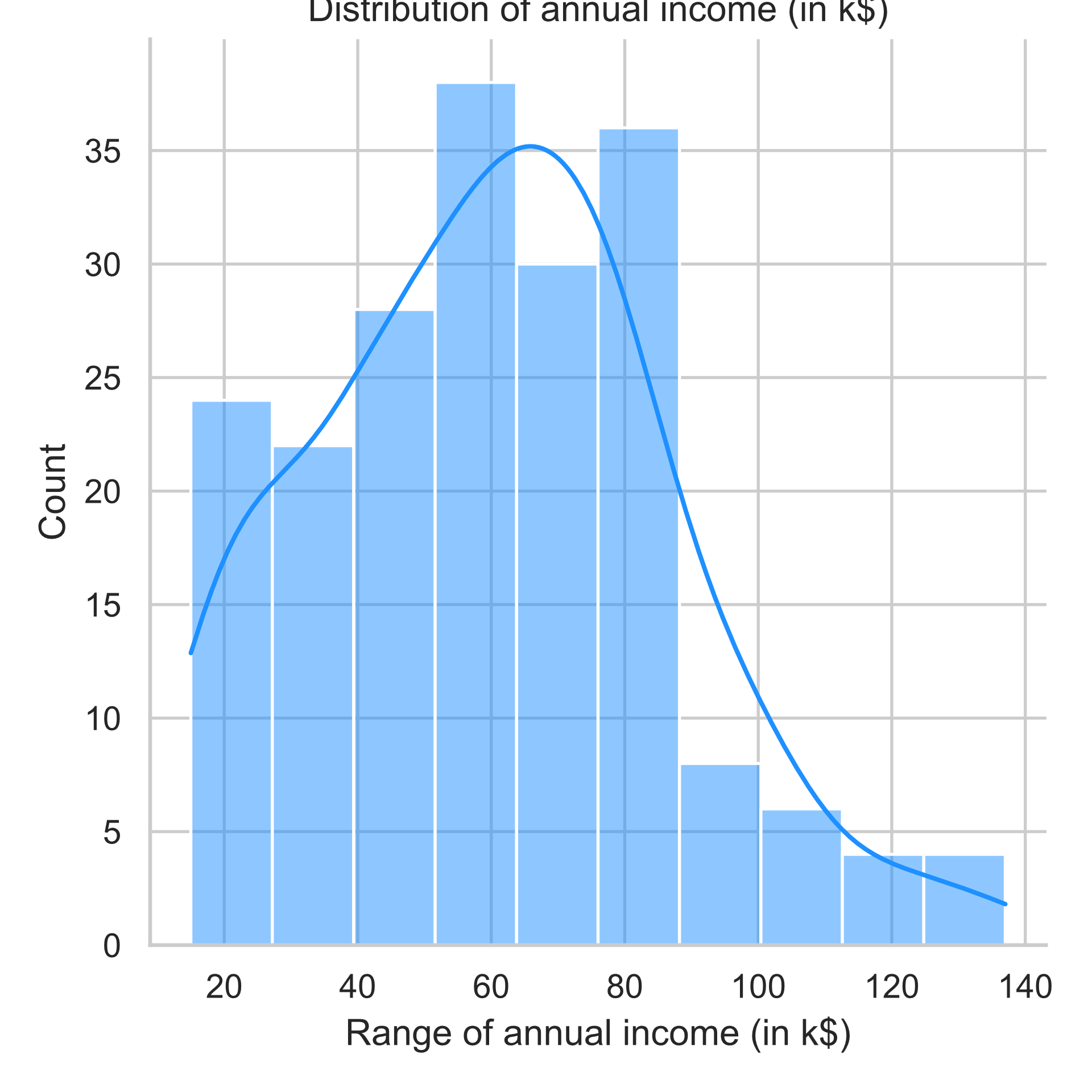
The dataset used in this study was collected from Kaggle which is a data science platform Shopping mall store is the source of dataset used in this for clustering and K-means clustering algorithm. In this we considered 200 rows which represent the single customer data. The two columns annual income and spending score are the main columns their relationship is the key to K-means clustering. The other three columns Customer-Id, Gender and Age does not play any specific role in determining the clusters. Datasets accommodate for a most important role in machine learning

**MEATHODOLOGY:**

**1.Data Collection and Processing:**

The given data set is read into a data frame (df). We will display few first rows of the data set using df.head(), to get an initial overview of the data. We will use df.corr(numeric only=True) to know the relationship between variables for correlation analysis performed on a numeric feature of data set. Representation to analyse distribution of features like annual income, age, spending score and gender. We will use the histogram and bar plots.[1]

Fig1: Distribution of Annual Income



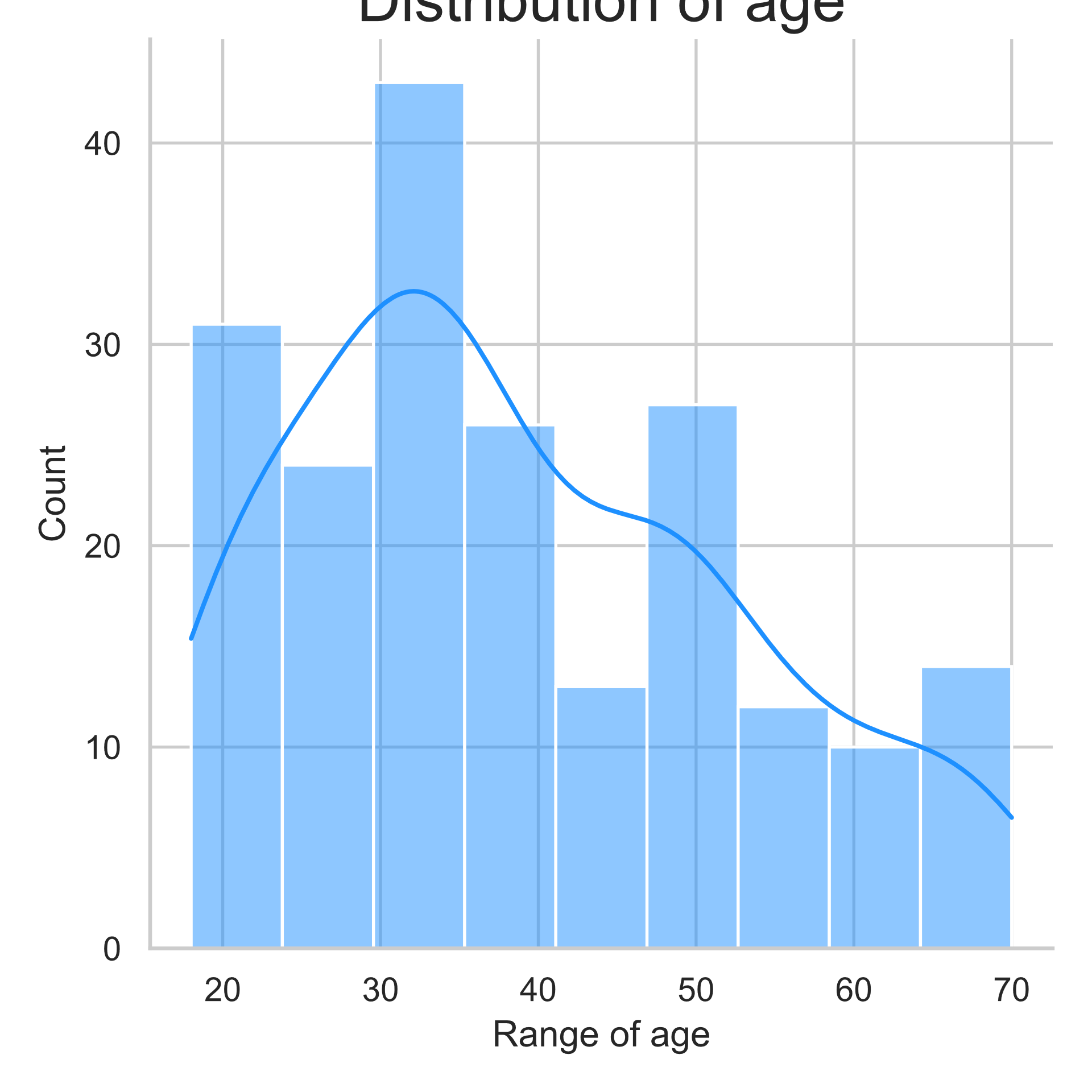
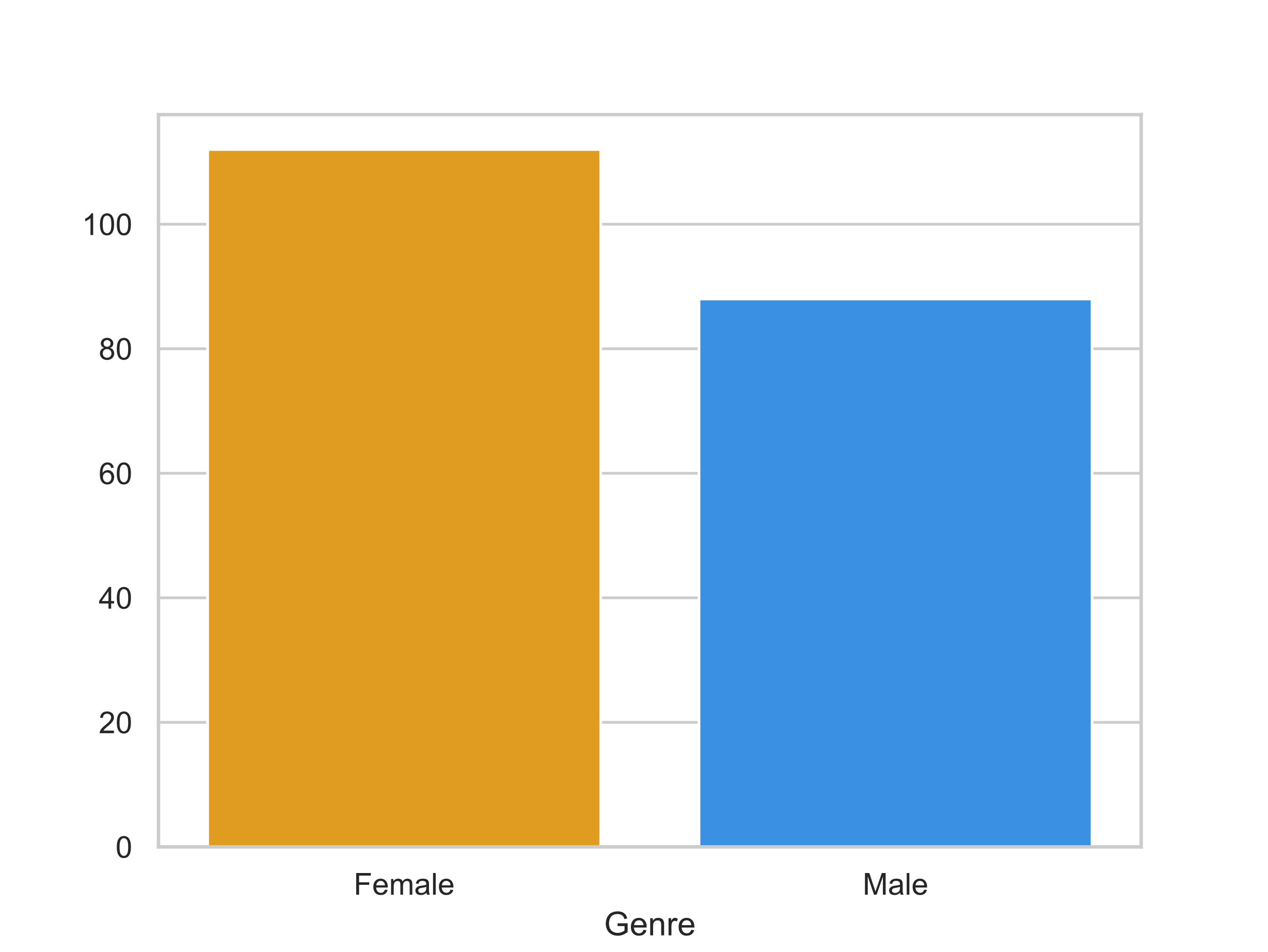
**

Fig2: Distribution of Age Fig3: Gender Analysis

**2. Splitting the Features and Target:**

Attributes like and "Annual Income (k$)" and "Spending Score (1-100)", are selected from the data set are stored in a separate data frame X. The aim variable is not precisely defined in the presented program

**3.Splitting the Data into Training and Test Data:**

There is no explicit splitting of data into training and test sets in the provided program. Although it is common step in machine learning operation to split the data for model training and evolution.

**4.Model Training:**

The K-Means algorithm is leveraged for clustering the data. Using elbow method, we can find the optimal number of clusters and for training the model, KMeans class is used from SckitLearn Library.



**5.Model Evaluation:**

The K-Means clustering was applied to the portion of mall customer data set build upon their annual income and spending score. We will determine the optimal number of clusters using the method named as Elbow method, which analyze the Within Cluster Sum of Square (WCSS) to find the critical point where supplementary clusters no longer considerably diminished the WCSS. The elbow curve analyze that 5 clusters would be suitable for the data set.

WCSS

**Where C is the cluster Centroid and d is the data point in each Cluster.**

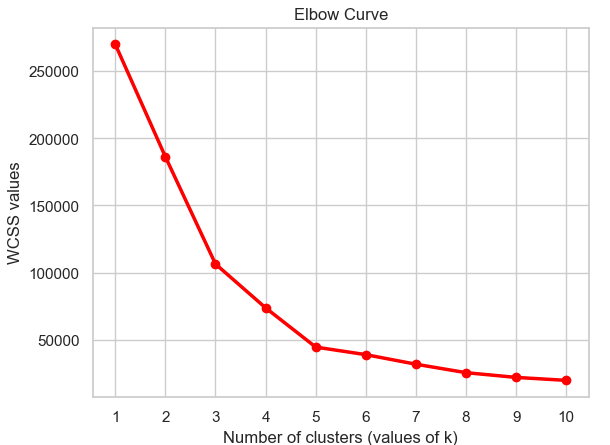


Fig 7: Elbow curve for determining optimal cluster number

**RESULT:**

**As K-Means is an unsupervised algorithm Finding the accuracy of an unsupervised learning algorithm can be challenging as there are no true labels to compare the result against. In unsupervised learning F1 and Precision score always be 1 because we are using the Unsupervised approch. If we want more precise and accurate result, we want to use the Silhouette score method of K-Means.**

To verify and evaluating the performance of the clustering algorithm we calculated F1 score and precision.  
F1 score determine the accuracy of the model while precision measure the proportion of the true positive prediction to the total number of genuine predictions.  
Representation like bar plots and tables are created to display the F1 score and precision value for each cluster.

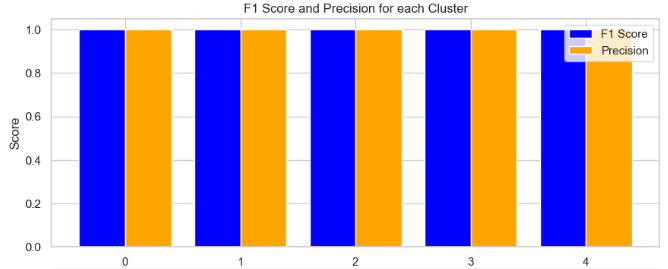


Fig 9: Bar plot for F1 score and precision for each cluster

**CONCLUSION:**

The conclusion based on the representation and visualization the following observations can be made about the clusters.

**Cluster 0:** the cluster 0 which is red coloured and is in the center signifies medium annual income and medium yearly spending score.

**Cluster 1:** the cluster 1 which is green coloured and is in the above right corner signifies high annual income and high spending score.

**Cluster 2:** the cluster 2 which is blue coloured and is in the above left corner signifies low annual income and high yearly spending score.

**Cluster 3:** the cluster 3 which is Yellow coloured and is in the lower right corner signifies high annual income but low yearly spending score.

**Cluster 4:** the cluster 4 which is voilet coloured and is in the lower left corner signifies low annual income and low yearly spending score

Here we have provided the colored scatter plot of the cluster for easy understanding.



Fig 10: Scatterplot with colors representing clusters