Mall Customer Segmentation

Abstract

A lot of customers buy products from the mall and to generate more revenue for the mall, the authorities need to attract these customers and for this large amount of capital is required. After the advertisement, the output is only around 30-40%. Hence customer segmentation comes into the picture.

Customer Segmentation is a popular application of unsupervised learning and by using this technique we'll only focus on the potential customers (customers whose probability of buying the product is very high). With this technique, the output will drastically increase to 90-95%.

Our project aims to build clusters of customers based on their Spending Score and Annual Income. The algorithm used in this project is K-means.

Table of Contents

I.   Introduction

To make predictions and find the clusters of potential customers of the mall and thus find appropriate measures to increase the revenue of the mall is one of the prevailing applications of unsupervised learning.

For example, a group of customers have high income but their spending score (amount spent in the mall) is low so from the analysis we can convert such type of customers into potential customers (whose spending score is high) by using strategies like better advertising, accepting feedback and improving the quality of products.

To identify such customers, this project analyses and forms clusters based on different criteria which are discussed in the further sections.

II.   Dataset

The dataset name is 'Mall_Customers.csv' consists of 5 columns which are CustomerID, Gender, Age, Annual Income (k$), Spending Score (1-100) where Gender is a categorical value and rest all features are numeric.

	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

2.1 Snapshot of Dataset

The size of the dataset is (200, 5) which is 200 rows and 5 columns.

III. Proposed Method and Architecture
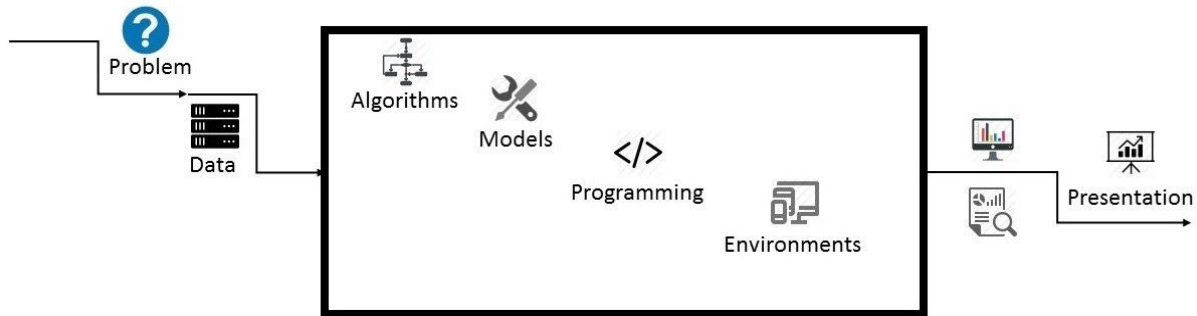
3.1 Data Science Project Architecture

Problem Statement

Customer Segmentation is a popular application of unsupervised learning. Using clustering, identify segments of customers to target the potential user base. They divide customers into groups according to common characteristics like gender, age, interests, and spending habits so they can market to each group effectively.
Use K-means clustering and also visualize the gender and age distributions. Then analyze their annual incomes and spending scores.

Data

The size of the dataset is (200, 5) which is 200 rows and 5 columns. Also the dataset does not contain any NULL or NaN values.

Algorithms

K-means algorithm is used in this project to analyze and form clusters of customers based on their income and spending score features.

Model

K-means model is used and is hyper tuned parameters like *n_clusters=5* using elbow method to find the optimal number of clusters also *init='k-means++'* to avoid random initialization traps.

Programming and Environment

Programming Language: Python 3.6
Environment (Libraries and Technologies): Numpy, Pandas, Matplotlib, Seaborn, Jupyter Notebook, Google Colab.

IV.   Methodology

The Data Science Methodology aims to answer basic questions in a prescribed sequence, that cover the five main aspects of data science projects. These aspects are:
- From Problem to Approach
- From Requirements to Collection
- From Understanding to Preparation
- From Modelling to Evaluation
- From Deployment to Feedback

In this project, the prescribed sequence is:
- Creating an approach to solve the given problem statement
- Exploring the dataset and obtaining useful insight from the same
- Cleaning the dataset by handling nan values, remove duplicate records, etc.
- Data Visualization used to obtain important information from the data
- Data Preprocessing is performed to make the data ready to fit the model this includes feature scaling, splitting the dataset into features and labels, etc.
- Model Building

V.   Implementation and Analysis

On performing data visualization on the dataset, a lot of insights were obtained such as:
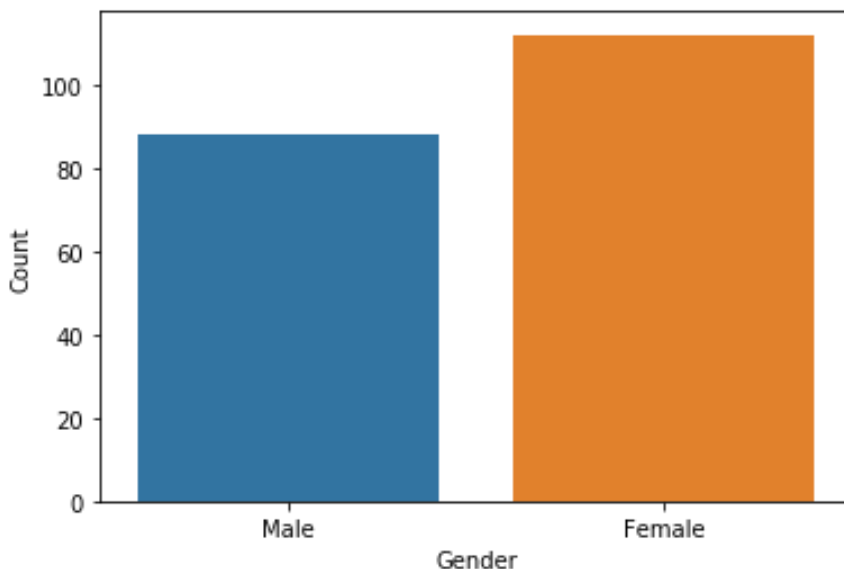
Gender Plot Analysis

From the Count plot, it is observed that the number of Female customers is more than the total number of Male customers.
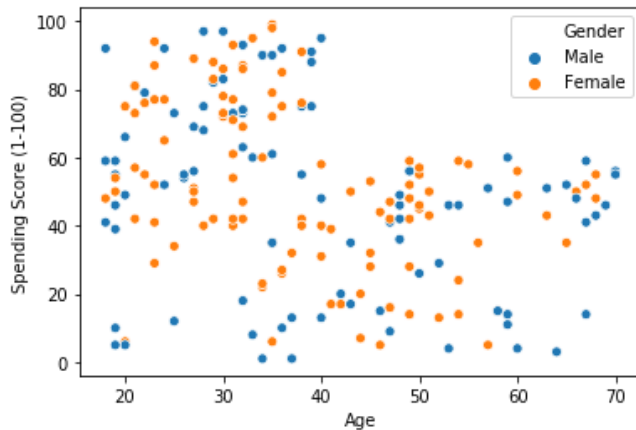
5.1 Gender Plot

Age Plot Analysis

From the Histogram it is evident that there are 3 age groups that are more frequently shop at the mall, they are: 15-22 years, 30-40 years, and 45-50 years.
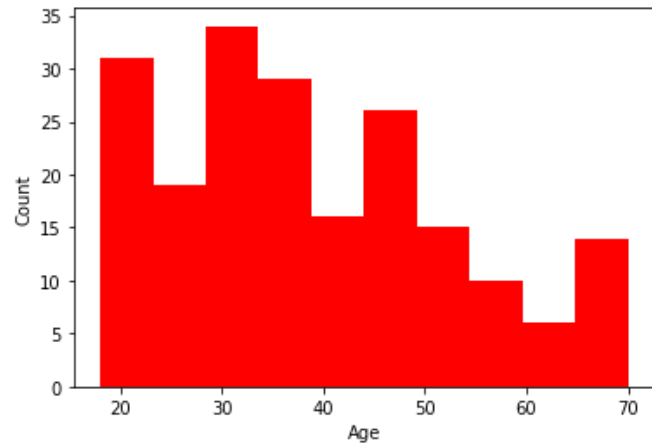

5.2 Age Plot

Age Vs Spending Score Analysis

1. From the Age Vs Spending Score plot we observe that customers whose spending score is more than 65 have their Age in the range of 15-42 years. Also from the Scatter plot it is observed that customers whose spending score is more than 65 consists of more Females than Males.
2. The customers having average spending score ie: in the range of 40-60 consists of the age group of the range 15-75 years and the count of males and females in this age group is also approximately the same.

5.3 Age Vs Spending Score

Annual Income Vs Spending Score Analysis

We observe that there are 5 clusters and can be categorized as:
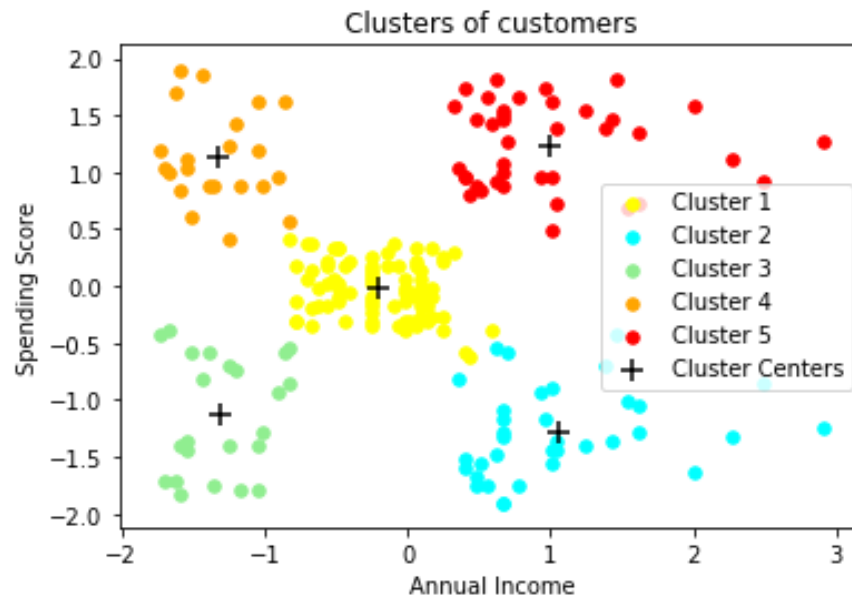
a. High Income, High Spending Score (Top Right Cluster)
b. High Income, Low Spending Score (Bottom Right Cluster)
c. Average Income, Average Spending Score (Center Cluster)
d. Low Income, High Spending Score (Top Left Cluster)
e. Low Income, Low Spending Score (Bottom Left Cluster)
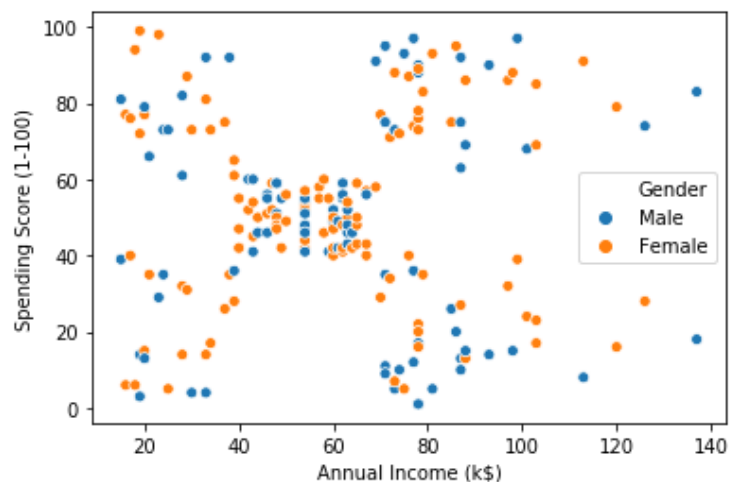
5.4 Annual Income Vs Spending Score

VI. Conclusion

For this project, the K-means algorithm is used and performs the best (with n_clusters = 5 and init = 'kmeans++'). After the clustering algorithm is applied to the dataset, this is the output.

6.1 Annual Income Vs Spending Score after Clustering

Clustering Analysis

a. High Income, High Spending Score (Cluster 5) - Target these customers by sending new product alerts which would lead to an increase in the revenue collected by the mall as they are loyal customers.

b. High Income, Low Spending Score (Cluster 2) - Target these customers by asking the feedback and advertising the product in a better way to convert them into Cluster 5 customers.

c. Average Income, Average Spending Score (Cluster 1) - May or may not target these groups of customers based on the policy of the mall.

d. Low Income, High Spending Score (Cluster 4) - Can target these set of customers by providing them with Low-cost EMI's, etc.

e. Low Income, Low Spending Score (Cluster 3) - Don't target these customers since they have less income and need to save money.