Customer Segmentation | Hypothesis Testing | KMeans

1 Background

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-consumer marketing, companies often segment customers according to demographics that include: Age,Gender,Marital status,Location (urban, suburban, rural),Life stage (single, married, divorced, empty-nester, retired, etc.),Income.

Segmentation allows marketers to better tailor their marketing efforts to various audience subsets.

2 Data presentation

Context

This data set is created only for the learning purpose of the customer segmentation concepts , also known as market basket analysis . I will demonstrate this by using unsupervised ML technique (KMeans Clustering Algorithm) in the simplest form.

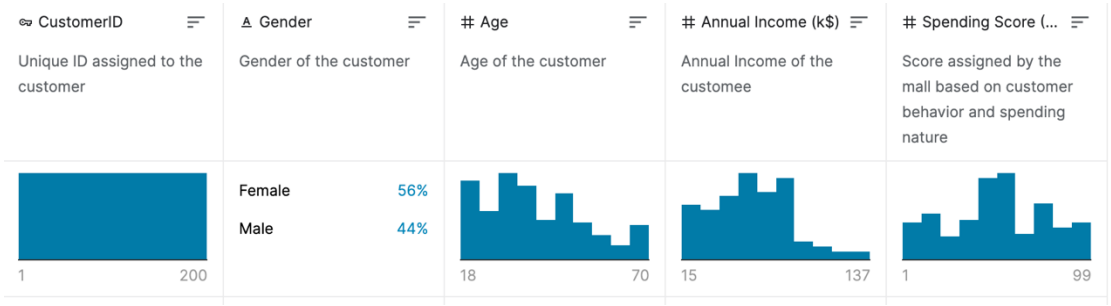
Content

You are owing a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score.

Spending Score is something you assign to the customer based on your defined parameters like customer behavior and purchasing data.

1) Mall_Customers.csv

This file contains the basic information (ID, age, gender, income, spending score) about the customers



Problem Statement

You own the mall and want to understand the customers like who can be easily converge [Target Customers] so that the sense can be given to marketing team and plan the strategy accordingly.

3 Project Objectives

By the end of this case study, you would be able to answer below questions.

1- How to achieve customer segmentation using machine learning algorithm (KMeans Clustering) in Python in simplest way.

2- Who are your target customers with whom you can start marketing strategy [easy to converse]

3- How the marketing strategy works in real world

4 Project Process

Step 1: Importing Libraries

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.set_context('notebook',font_scale=1.25)
from IPython.core.display import HTML,display
import scipy.stats
from sklearn.cluster import KMeans
from mpl_toolkits.mplot3d import Axes3D
```

Step 2: EDA exploratory data analysis

Importing Data

Basic EDA (Inference : There are 200 rows and 5 columns.)

- **Univariate Analysis**
- **Bivariate Analysis**
  - ✓ **Checking for Association between Gender and Score**

    Stating the hypothesis:

    H0: Gender and Score are independent.

    H1: Score depends on Gender.

  - ✓ **Checking for Association between Gender and Annual Income**

    Stating the hypothesis:

    H0: Gender and Annual Income are independent.

    H1: Annual Income depends on Gender.

  - ✓ **Checking for Association between Annual Income and Score**

    Stating the hypothesis:

    H0: The Annual Income and Score are not correlated.

    H1: The Annual Income and Score are correlated.

  - ✓ **Checking the Association between Age(Binned) and Spending Score**

    Stating the hypothesis:

    H0: The mean score of all Age groups is equal.

H1: At least one to mean Annual Income of Age groups differ.

- ✓ **Checking the Association between Age(Binned) and Annual Income**

  Stating the hypothesis:

  H0: The mean Annual Income of all Age groups is equal.

  H1: At least one to mean Annual Income of Age groups differ.

Step 3: Modelling-KMeans

Note:

- We can consider any number of features as an input to the Clustering Algorithm but to visualize the results at most I can consider only three features.
- Gender is the least important feature here as per the statistical tests performed so I'm not considering it.

Determining optimum value of K using the Elbow Method

- k is the parameter of KMeans Algorithm which instructs it how many clusters are needed to be formed of the given data.
- init='k-means++' ensures that initial clusters are choosen smartly and not randomly which increases the chances of convergence.