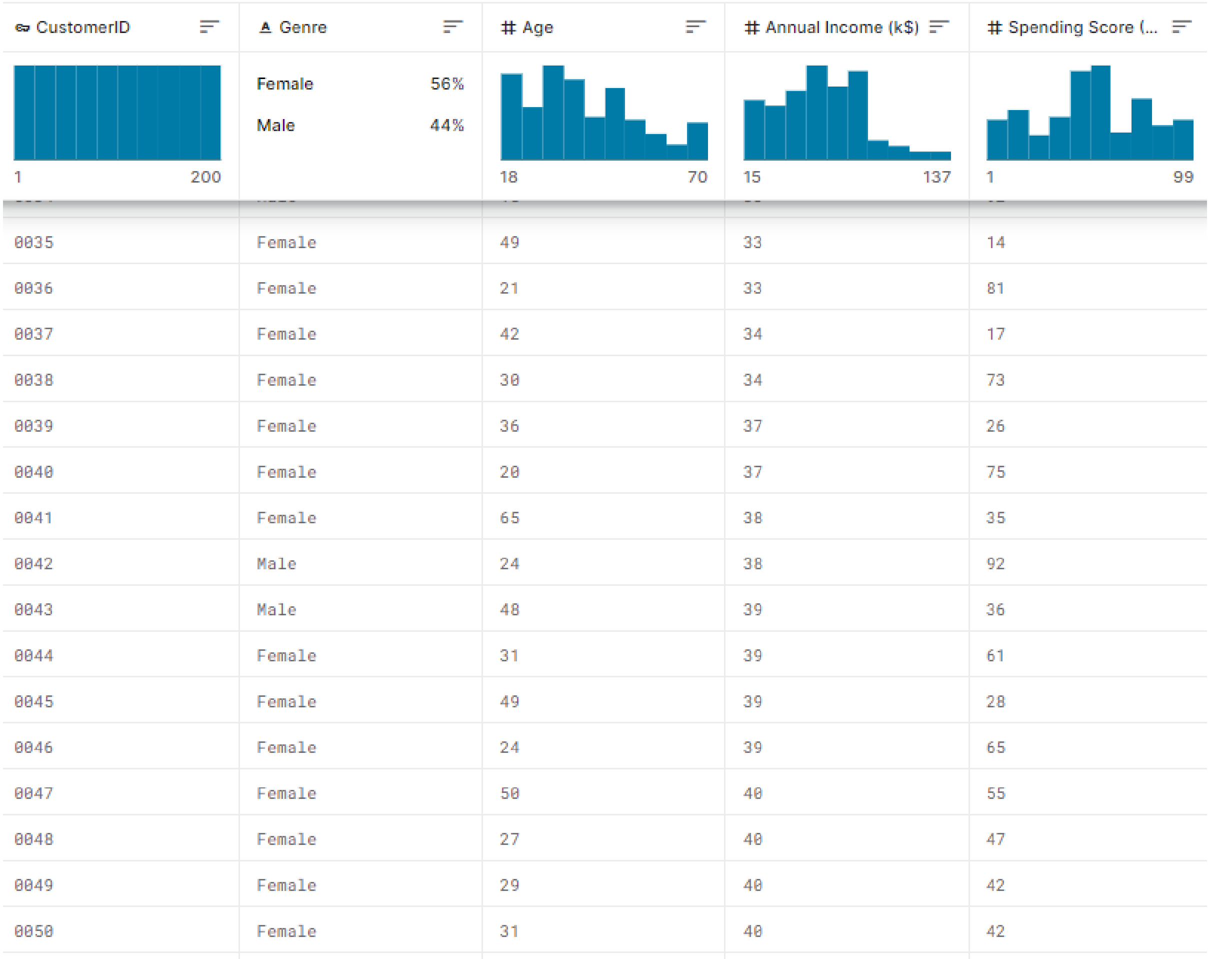
# CUSTOMER SEGEMENTATION
# PHASE 4

## INTRODUCTION

The problem is to implement data science techniques to segment customers based on their behaviour, preferences, and demographic attributes. The goal is to enable businesses to personalize marketing strategies and enhance customer satisfaction. This project involves data collection, data preprocessing, feature engineering, clustering algorithms, visualization, and interpretation of results. Customer Segmentation is the process of dividing a company's customers into groups based on common characteristics so companies can market to each group effectively and appropriately.

In this phase the development of customer segmentation is going to be done.

## GIVEN DATA SET



| CustomerID | Genre | # Age | # Annual Income (k$) | # Spending Score (... |
|---|---|---|---|---|
| 1 – 200 | Female 56%<br>Male 44% | 18 – 70 | 15 – 137 | 1 – 99 |
| 0035 | Female | 49 | 33 | 14 |
| 0036 | Female | 21 | 33 | 81 |
| 0037 | Female | 42 | 34 | 17 |
| 0038 | Female | 30 | 34 | 73 |
| 0039 | Female | 36 | 37 | 26 |
| 0040 | Female | 20 | 37 | 75 |
| 0041 | Female | 65 | 38 | 35 |
| 0042 | Male | 24 | 38 | 92 |
| 0043 | Male | 48 | 39 | 36 |
| 0044 | Female | 31 | 39 | 61 |
| 0045 | Female | 49 | 39 | 28 |
| 0046 | Female | 24 | 39 | 65 |
| 0047 | Female | 50 | 40 | 55 |
| 0048 | Female | 27 | 40 | 47 |
| 0049 | Female | 29 | 40 | 42 |
| 0050 | Female | 31 | 40 | 42 |

There exist 5*5 columns in the data set that is using in this project.

## OVERVIEW OF THE PROCESS

The following is an overview of the process of building a house price prediction model by feature selection, model training, and evaluation:

1. Prepare the data: This includes cleaning the data, removing outliers, and handling missing values.

2. Perform feature selection: This can be done using a variety of methods, such as correlation analysis, information gain, and recursive feature elimination.

3. Train the model: There are many different machine learning algorithms that can be used for house price prediction. Some popular choices include linear regression, random forests, and gradient boosting machines.

4. Evaluate the model: This can be done by calculating the mean squared error (MSE) or the root mean squared error (RMSE) of the model's predictions on the held-out test set.

5. Deploy the model: Once the model has been evaluated and found to be performing well, it can be deployed to production.

## PROCEDURE

- Identify the target variable
- Explore the data
- Remove redundant features
- Remove irrelevant features

## FEATURE SELECTION

### In[1] : Listing all input files into directory :

```
import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

### In[2] : Renaming the columns :

```
df=pd.read_csv('/kaggle/input/mall-customers/Mall_Customers.csv')

df.rename(columns={'Genre':'Gender'},inplace=True)
df.head()
```

### In[3] : Checking Missing Values :

```
               df.isnull().sum()
```

```
Out[3]:

CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

## MODEL TRAINING

Model Training is the process of teaching a machine learning model to do customer segmentation. It involves feeding the model historical data and features. The model then learns the relationships between these customers and spending scores.

1) KMeans :

```
X1 = df.loc[:,["Age","Spending Score (1-100)"]].values
from sklearn.cluster import KMeans
wcss=[]
for k in range(1,11):
kmeans = KMeans(n_clusters = k, init = "k-means++")
kmeans.fit(X1)
wcss.append(kmeans.inertia_)
plt.figure(figsize =( 12,6))
plt.grid()
plt.plot(range(1,11),wcss,linewidth=2,color="red",marker="8")
plt.xlabel("K Value")
plt.ylabel("WCSS")
plt.show()
```

2) Hierarchial Clustering :

```
   train_path = "/kaggle/input/mall-customers/Mall_Customers.csv"
   train_data = pd.read_csv(train_path)
   hierarchical_cluster = AgglomerativeClustering(n_clusters=2, affinity='
euclidean', linkage='ward')
    labels = hierarchical_cluster.fit_predict(train_data)

    xs = train_data[:,0]
    ys = train_data[:,1]
    plt.scatter(xs,ys,c=labels,alpha=0.5)
```

## MODEL EVALUATION

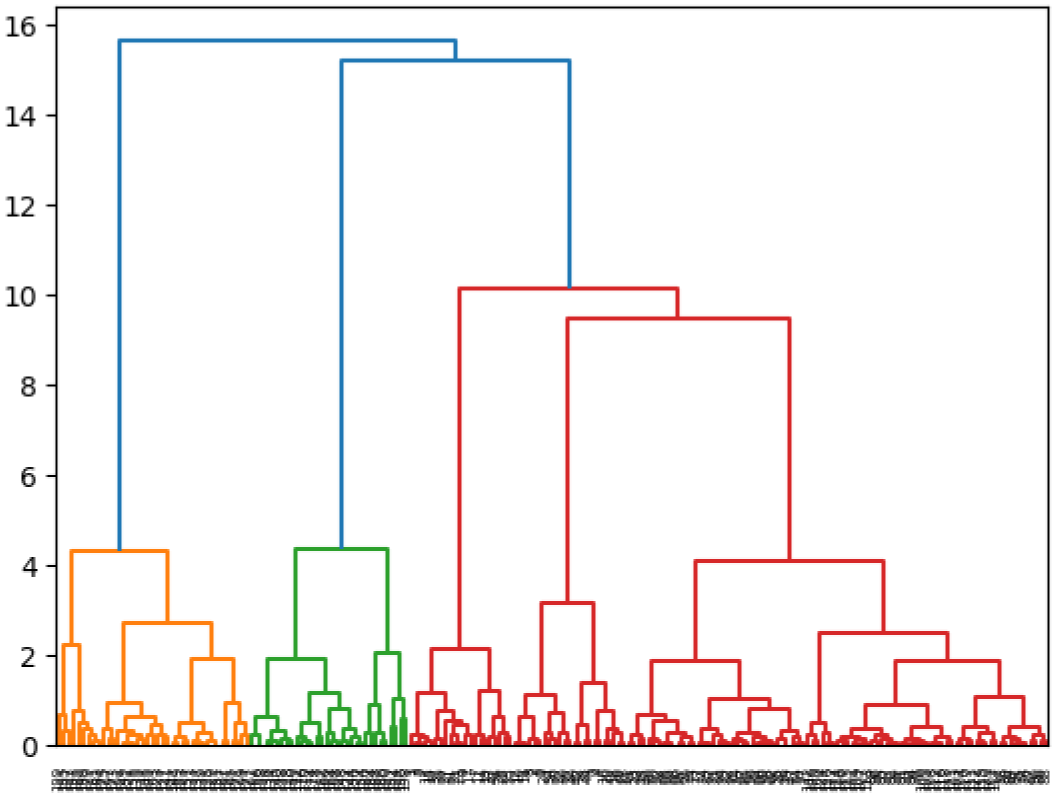
Model Evaluation is the process of assessing the performance of a machine learning model on unseen data. This is important to ensure that the model will generalize to the new data. There are number of metrics to evaluate the customer segmentation model which listed below.,

- Mean Squared Error
- Root Mean Squared Error
- Mean Absolute Errror
- R-Squared
- Bias
- Variance
- Interpretability

In[1] :

```
linkage_data = linkage(train_data, method='ward', metric='euclidean')
dendrogram(linkage_data)
plt.show()
```

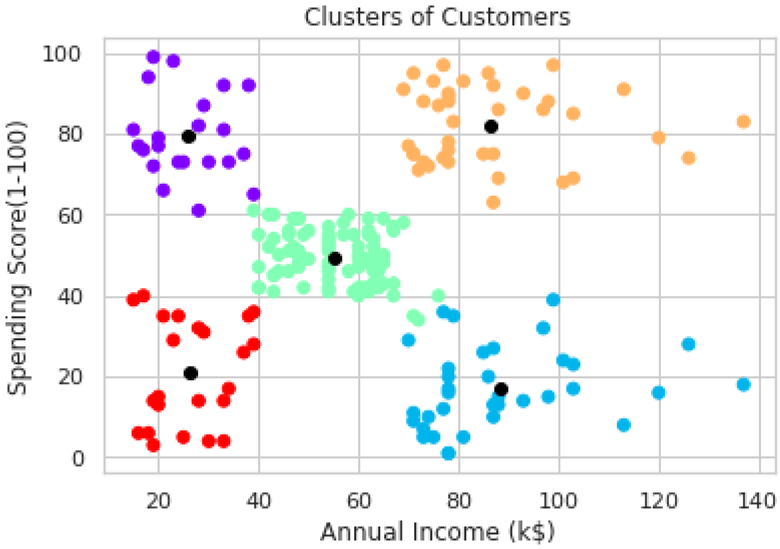
Out[1] :



In[2] :

```
plt.scatter(X2[:,0],X1[:,1],c=kmeans.labels_,cmap='rainbow')
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],color
='black')
plt.title('Clusters of Customers')
plt.xlabel('Annual Income (k$)')
plt.ylabel('Spending Score(1-100)')
plt.show
```

Out[2] :



## MODEL COMPARISON

Model Comparison is the tool enhanced to provide systematic representation on the relationship between annual income and spending score.

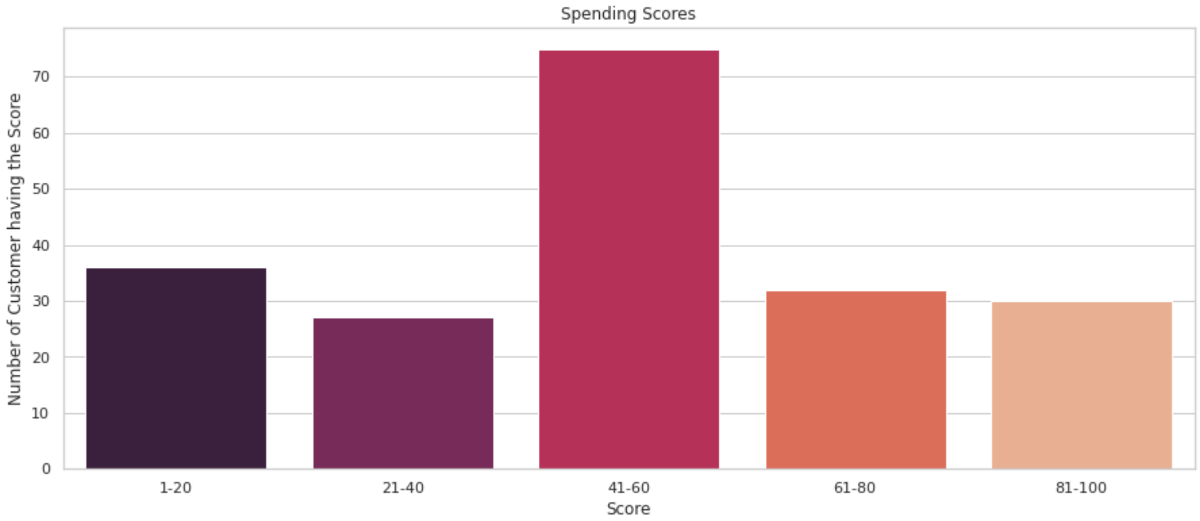
In[1] :

```
ss_1_20 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 1)
& (df["Spending Score (1-100)"] <= 20)]
ss_21_40 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 21
) & (df["Spending Score (1-100)"] <= 40)]
ss_41_60 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 41
) & (df["Spending Score (1-100)"] <= 60)]
ss_61_80 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 61
) & (df["Spending Score (1-100)"] <= 80)]
ss_81_100 = df["Spending Score (1-100)"][(df["Spending Score (1-100)"] >= 8
1) & (df["Spending Score (1-100)"] <= 100)]

ssx= ["1-20","21-40","41-60","61-80","81-100"]
ssy=[len(ss_1_20.values),len(ss_21_40.values),len(ss_41_60.values),len(ss_6
1_80.values),len(ss_81_100.values)]

plt.figure(figsize=(15,6))
sns.barplot(x=ssx,y=ssy, palette="rocket")
plt.title("Spending Scores")
plt.xlabel("Score")
plt.ylabel("Number of Customer having the Score")
plt.show()
```

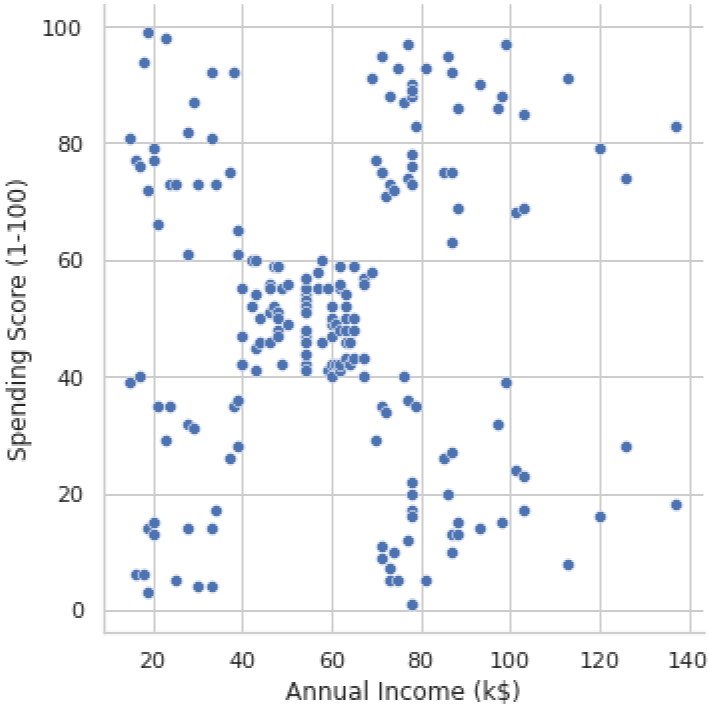
Out[1] :



In[2] :

```
sns.relplot(x="Annual Income (k$)",y = "Spending Score (1-100)",data=df)
```

Out[2] :



## MODEL ANALYSIS

Once evaluation of model gets completes, the analysis of model’s predictions can be started to identify any patterns or biases. This will help to understand the strengths and weakness of the model and to improve it.

## FEATURE ENGINEERING

Feature Engineering is a crucial aspect of building a customer segmentation model using machine learning. It involves creating new features, transforming the existing ones and selecting the most relevant variable to improve the model. Here are some feature engineering ideas for customer segmentation:

- Data Preparation & Cleansing
- Handling missing values
- Transformation of Categorical Features
- Numerical Representation
- Zoning Information
- Accessibility Features
- Time-related features

## CONCLUSION

The process of customer segmentation ensures that your brand is customer-centric and helps you serve them better. It boosts conversions, brings your marketing efforts to fruition, and also helps build everlasting customer relationships. The strategies discussed here will help you organize your segments, but after you have them in place, continue to monitor and make sure your product is still valuable to the groups. The key to successful customer segmentation is the constant research it entails to ensure your brand and product stay relevant and indispensable. The benefits of this will positively impact on the brand's revenue. Some of the benefits are.,

- ➢ Drive-up retention rates
- ➢ Increase Visibility
- ➢ Drive sales with well-timed discounts
- ➢ Bring old customers back
- ➢ Deliver a most personalized experience

The customer segmentation project using machine learning has yielded valuable insights that can significantly benefit our business. Through the application of advanced algorithms and data analysis techniques, we have successfully divided our customer base into distinct segments based on various attributes such as behaviour, demographics, and preferences.