Phase 2: Feature Selection & Modelling

**Step 1 :**

**Goal:** To select meaningful features from cleaned dataset and build a clustering model to segment customers.

**Step 2 :**

# What is "Feature Selection”?

Choosing the most important and relevant columns (features) from our dataset that will actually help our model do its job well.

**Steps of feature selection:**

1)From the cleaned online store data, we selected:

1. Recency – Days since last purchase
2. Frequency – How often they purchase
3. Monetary – Total spend

We added extra features for more insights:

1. Total Quantity – Total items bought.
2. Avg\_Basket\_Size – Average number of items per invoice.
3. Avg\_Item\_Price – Average spend per item.

**Why these features?**

These capture both purchase behaviour (frequency & quantity) and value behavior

(monetary & price sensitivity).

**Step 3 :**

# Data Preparation

1. Handled missing/infinite values → replaced with 0.

2. Scaling: Scaled features using StandardScaler so all features are on the same range (important for K-Means).

3. Visualisation:Visualized distributions before & after scaling to check the effect.

**Step 4 :**

Choosing the Number of Clusters (k)

We tested “k” from 2 to 10 using:

1. **Elbow Method** → Looked for the “bend” in the WCSS curve.
2. **Silhouette Score** → Measures how well customers fit into their clusters (higher = better).

We found **k = 4** as a good balance.

**Step 5 :**

# What is "Modelling”?

Means to build a statistical model that learns from the data to perform a task and find patterns.

**Modelling with K-Means**

* Applied **K-Means** with k = 4.
* Assigned each customer to one of the four clusters.

**Step 6 :**

**Understanding the Clusters**

We created **Cluster Profiles** (average Recency, Frequency, Monetary, etc. per cluster).  
Based on the values:

* **Cluster 0:** Loyal Customers (low Recency, high Frequency, high Monetary).
* **Cluster 1:** At Risk (high Recency, lower Frequency & Monetary).
* **Cluster 2:** Best Customers (very high Frequency & Monetary).
* **Cluster 3:** Lost Customers (inactive for a long time).

**Step 7 :**

**Visualization**

* **Correlation Heatmap** → See relationships between features.
* **PCA Scatter Plot** → Visualize clusters in 2D space for easy interpretation.

**CONCLUSION:**

We now have **4 distinct customer groups** we can target with **different marketing strategies**.  
For example:

* Send **special offers** to Best Customers.
* Give **win-back campaigns** to Lost Customers.

The result is 4 clear customer segments: Best Customers, Loyal, At-Risk, and Lost, which we can target with different strategies