**CUSTOMER SEGMENTATION PROJECT REPORT**

1. **Problem statement**

Use K-Means clustering to segment customers based on behavioral and demographic data to enable targeted marketing strategies.

1. **Dataset description**

**Selected dataset:** [Online Retail dataset-Kaggle](https://www.kaggle.com/datasets/vijayuv/onlineretail)

The **Online Retail Dataset** is a real-world transactional dataset from a **UK-based e-commerce store** that sells household goods (mostly gifts and stationery). It contains **actual invoice-level purchase records** between **December 2010 and December 2011**.

**Key Features:-**

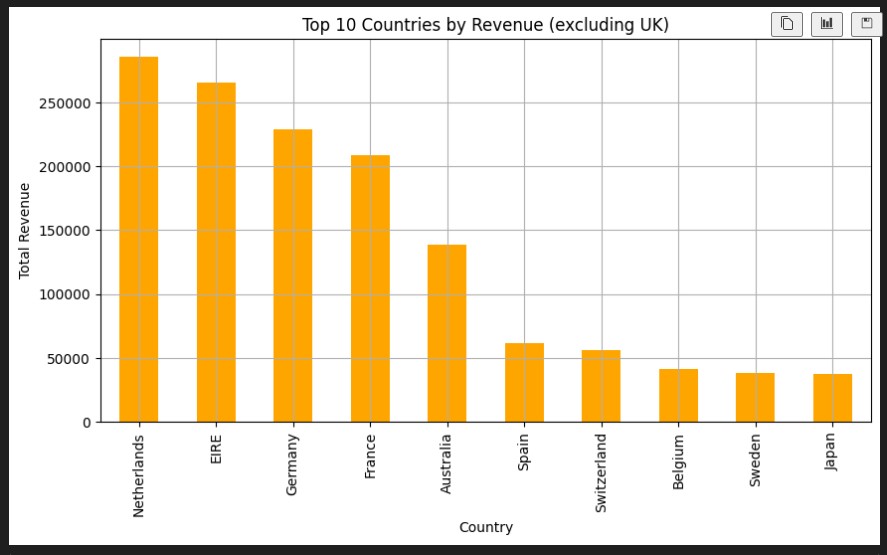
| **Column** | **Description** |
| --- | --- |
| InvoiceNo | Unique invoice number (can start with 'C' if canceled) |
| StockCode | Product code |
| Description | Name of the product |
| Quantity | Quantity of product purchased |
| InvoiceDate | Date and time of the invoice |
| UnitPrice | Price per product |
| CustomerID | Unique identifier for each customer |
| Country | Country of the customer |

**Behaviorial data:- ‘**Quantity’ , ‘InvoiceNo’ , ‘InvoiceDate’ , ‘ UnitPrice’ , ‘CustomerID’ columns show behaviorial data. These columns allow us to build **RFM (Recency, Frequency, Monetary)** values, which form the **core of customer behavior analysis**.

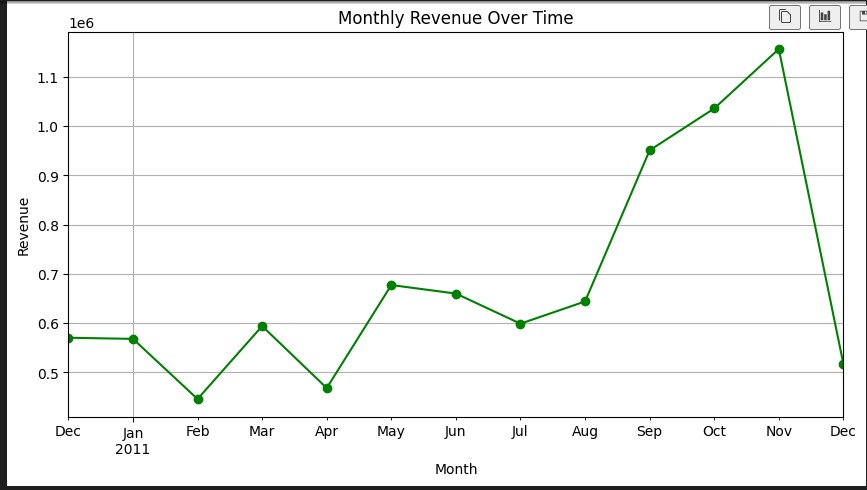
**Demographic data:-** ‘Country’ column show demographic information about customer.

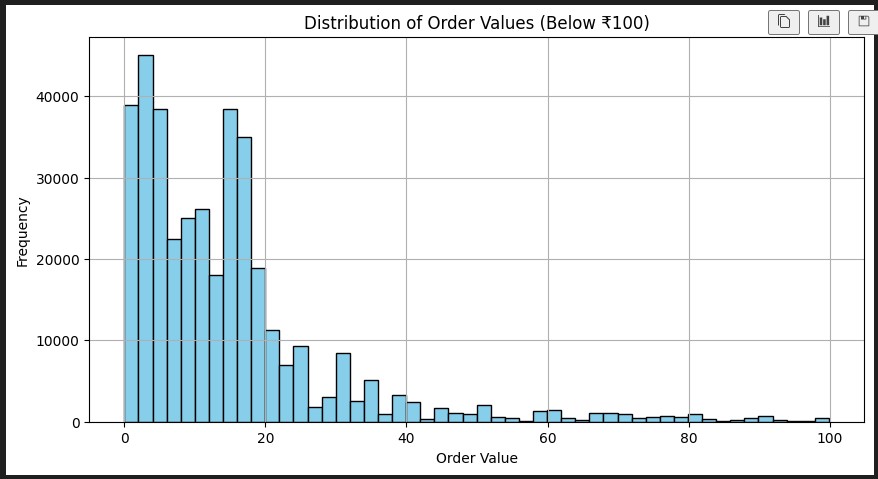
1. **Data Cleaning**

* Understanding the dataset structure and datatypes(eg**. Dataset has 541908 rows**). Observe that ‘CustomerID’ has missing values also ‘Quantity’ and ‘UnitPrice’ has outiliers.
* Dropped rows with missing values, also remove duplicates records from dataset and check the dimension after these steps.
* Create ‘**TotalPrice**’ column by multiplying ‘**Quantity**’ and ‘**UnitPrice**’.

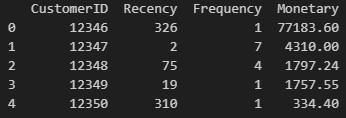
1. **Exploratory Data Analysis**
2. The **United Kingdom** dominates with over **96% of all orders**, confirming it as the **primary market** (***349203*** orders).
3. Other European countries such as Germany, France, and Ireland show notable activity and could be explored for targeted marketing campaigns.(Germany-***9025*** , France-***8326*** , EIRE-***7226*** , Spain-***2479*** Orders).
4. Total Revenue by Countris(Excluding UK) :- ****
5. Top 10 selling products:-

****

1. Monthly revenue over time:- **Q4 (Sep–Nov)** is the most profitable period, possibly due to **seasonal shopping trends or holidays.**
   * 1. ****
2. Distribution of order values:- A large number of orders fall between **₹0 and ₹20**, As the order value increases beyond ₹20, **frequency sharply drops**.



1. **RFM Table Formation**



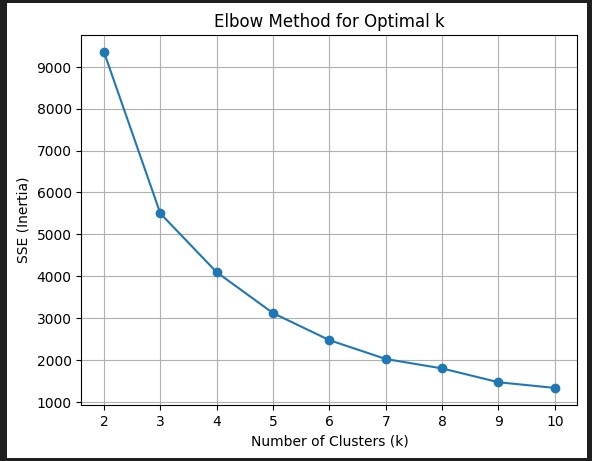
* **Customer 12347** is the most engaged with a recent purchase (Recency = 2), high purchase frequency (7), and moderate spending.
* **Customer 12346** is a **high spender** (₹77k+) but hasn’t purchased recently (Recency = 326) — possibly a lost or dormant customer.
* **Customer 12350** has low frequency and spending — likely a low-value or one-time customer.

1. **Feature Scaling**

Since RFM values have different units and ranges, **StandardScaler** was used to normalize the data:

* Converts values to a standard normal distribution (mean = 0, std = 1).
* Ensures that **no feature dominates** due to scale differences in clustering or machine learning models.

1. **Elbow method (To find optimal no. of Clusters)**

****

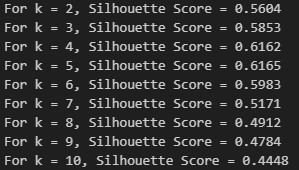
* + - * The plot is a **line graph** where:
        + **X-axis:** Number of clusters (k)
        + **Y-axis:** SSE (Inertia)
      * There's a **sharp drop from k=2 to k=4**, and then the curve starts to **flatten out**.
      * The **"elbow point"** appears around **k = 4**, indicating the **optimal number of clusters**.

1. **Silhouette score ( Evaluating cluster quality )**

The **Silhouette Score** measures how well-separated the clusters are.

**Range:** from -1 to +1

* +1 – well defined clusters
* 0 – overlapping clusters
* -1 – misclassified point



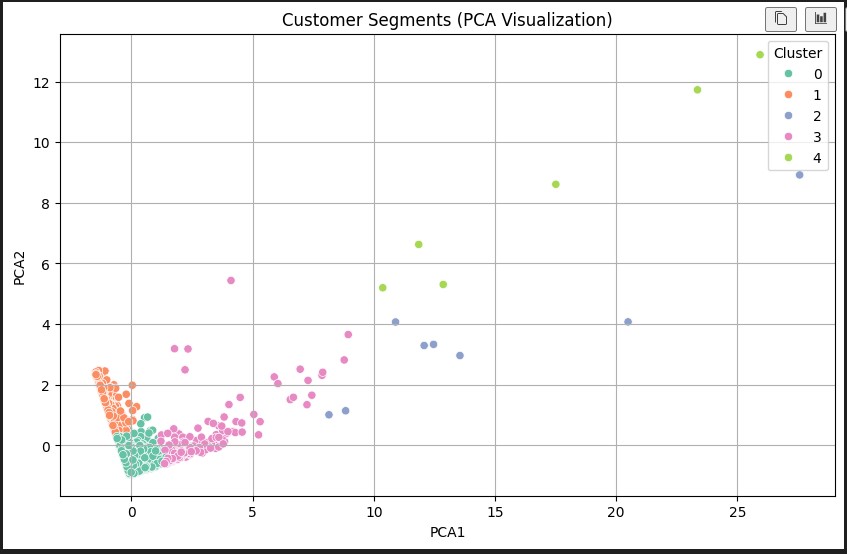
* From the output:
  + Scores increase from **k=2 to k=5**, peaking at **k=5 (score = 0.6165)**.
  + Beyond k=5, the score steadily **declines**, indicating reduced clustering quality.
* Thus, **k=5** is the optimal number of clusters based on silhouette analysis, offering the best balance of cohesion and separation.

1. **Kmeans Clustering**

The **K-Means algorithm** was applied to the **normalized RFM data** (rfm\_scaled) to segment customers into different groups based on their purchasing behavior. This process groups customers into segments that behave similarly, helping businesses tailor marketing strategies to different customer types.

1. **Dimensionality reduction using PCA ( Cluster Visualization )**

To visualize customer segments in 2D space, we applied **Principal Component Analysis (PCA)** to the scaled 3D RFM data. PCA helps reduce high-dimensional data into fewer components while retaining most of the variance. PCA helps **visually separate clusters** by projecting them onto **2 principal components (PCA1 and PCA2)**.

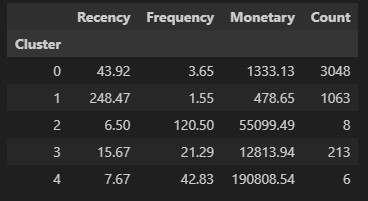
****

The plot shows customer segments projected onto two principal components: **PCA1 (x-axis)** and **PCA2 (y-axis)**. Each color represents a unique cluster obtained from KMeans clustering with **k = 5**.

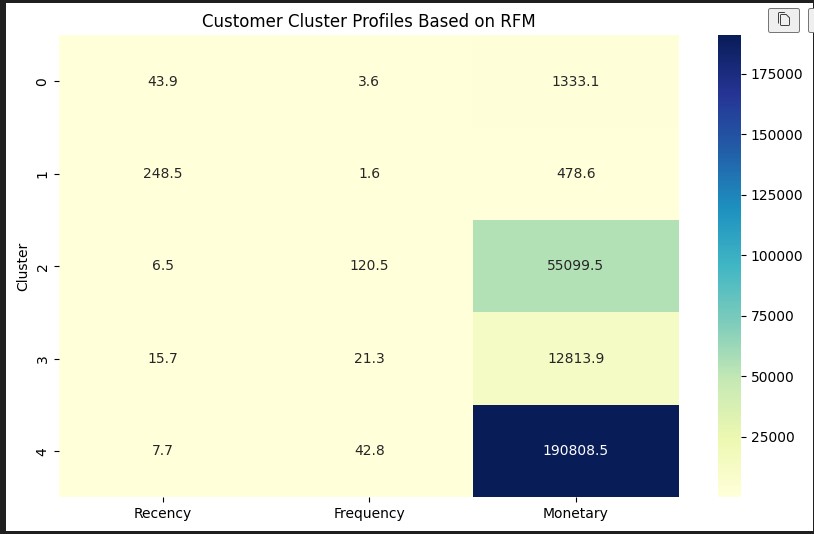
1. **Cluster 0 (Green):** These are likely **high-value outliers** — customers who might purchase infrequently but spend significantly when they do. (Possibly **VIP** customers or **seasonal** big spenders.)
2. **Cluster 1 (Orange):** These customers are **very similar to each other** — likely **frequent buyers with moderate spending**. (Represents your **core customer base** with regular activity.)
3. **Cluster 2 (Blue):** This group shows moderate diversity — could be **occasional buyers** with average spend.
4. **Cluster 3 (Pink):** These are **low-value or inactive customers** — probably recent customers or those who haven’t made significant purchases.
5. **Cluster 4 (Yellow):** A very **diverse and scattered group**, likely **anomalies** or **niche customers**. (Could represent business buyers, gift shoppers, or inconsistent patterns.)
6. **Final Profiling of Clusters**

Adding cluster labels back to the original (scaled) RFM data , then group by cluster to analyze:

* **Recency:** How recently a customer made a purchase (Lower is better).
* **Frequency:** How often they purchase (Higher is better).
* **Monetory:** How much they spend (Higher is better).
* **Count:** Number of customer in that cluster.



1. **0 : Largest group** (3048 customers). Moderate recency, low frequency, low spending. Likely **inactive or low-value** customers.
2. **1 :** Very **old recency**, lowest frequency and spending. These are **churned or lost customers** (need re-engagement).
3. **2 :** Extremely **high frequency** and **high spending**, very recent. These are **top VIP customers** (only 8 customers).
4. **3 :** Good frequency and spending, recent activity. These are **loyal and active customers** worth retaining and upselling.
5. **4 : Very recent**, very high spenders with decent frequency. Likely **new but premium** customers (6 in total), high-value targets.
6. **Visualization of cluster profiles**

****

 **Cluster 0**: Low frequency (3.6), low spenders (~1333), moderate recency → **Occasional buyers**

 **Cluster 1**: Very high recency (248.5), lowest frequency (1.6), very low spend (~478) → **Dormant/Churned customers**

 **Cluster 2**: Very low recency (6.5), highest frequency (120.5), high spend (~55K) → **Most Loyal & Active**

 **Cluster 3**: Recent (15.7), good frequency (21.3), decent spend (~12.8K) → **Loyal Customers**

 **Cluster 4**: Very recent (7.7), high frequency (42.8), **highest spend (~190K)** → **Top Premium Customers**

1. **Cluster-Based Targeted marketing strategies**
2. **Occasional Buyers(cluster 0): (**Low-Mid Value | 3,048 customers)

* **Behavior:** Moderate recency, low frequency, low monetary.
* **Goal:** Increase engagement and spending.
* **Strategies:**
  + Offer **combo deals** or **limited-time discounts** to encourage repeat purchases.
  + Use **email reminders** with personalized recommendations.
  + Introduce a **points-based loyalty program** to build frequency.
  + **Product Recommendations:** Based on browsing history.

1. **Dormant/Churned(cluster 1):** (Low Value | 1,063 customers)

* **Behavior:** High recency (inactive), very low frequency and spending.
* **Goal:** Re-engage or win back lost customers.
* **Strategies:**
* Launch a **“We Miss You” reactivation campaign** with a strong incentive (e.g., ₹200 off).
* Run **exit surveys** to understand drop-offs and improve offerings.
* Provide **low-barrier re-entry deals**, such as free shipping or free trials.
* **Flash Sales:** Create urgency through time-limited offers.

1. **Most loyal & active(cluster 2):** (Top-Tier | 8 customers)

* **Behavior:** Very recent, extremely frequent, very high spenders.
* **Goal:** Retain and reward loyalty.
* **Strategies:**
* Offer **exclusive benefits**: early access to sales, loyalty tiers, or luxury packaging.
* **Upsell & Cross-sell:** Suggest premium or related products.
* Encourage **referrals** with high-value rewards.
* **Surprise Rewards:** Gift cards, handwritten thank-you notes.

1. **Loyal customers(cluster 3):** (Mid-High Value | 213 customers)

* **Behavior:** Recent activity, moderate frequency, good spending.
* **Goal:** Strengthen loyalty and encourage upselling.
* **Strategies:**
* **Frequency Boosters:** Loyalty stamps, buy X get 1 free.
* **Conversion Discounts:** Encourage upgrading purchases.
* **Seasonal Promotions:** Target holidays and festivals.
* **Content Engagement:** Product tips, user stories, newsletters.

1. **Top Premium Customer(cluster 4):** (Elite VIPs | 6 customers)

* **Behavior:** Very recent, moderately frequent, **highest monetary value**.
* **Goal:** Retain and deepen emotional connection.
* **Strategies:**
* Offer **ultra-personalized experiences** (e.g., birthday gifts, handwritten notes).
* Invite to **VIP-only events** or provide **lifetime value discounts**.
* Treat as **brand advocates** — promote referrals and testimonials.
* **Personalized Offers:** Early access to premium launches.