**📦 Customer Segmentation Project – Subscription Box Business**

*Finding patterns in customer behavior using RFM and clustering*

**👋 About the Project**

In this project, I dove into customer transaction data from a **subscription box business** to better understand their buying behavior. The goal was simple: **segment customers** based on how recently they bought something, how often they shop, and how much they spend. Once we identify the different types of customers, the business can create more personalized strategies — like who to reward, who to re-engage, and who needs more attention.

**🔍 What Data Did I Use?**

The dataset had basic transaction details, and I engineered the following key features using Python:

* **Recency** – How many days since the customer’s last order
* **Frequency** – Total number of purchases
* **Monetary Value (MntTotal)** – Total amount spent

I also cleaned and prepared the data by removing outliers and normalizing the features before jumping into modeling.

**📊 Step-by-Step Breakdown**

**1. RFM Calculation**

* Created a table of customers with their **Recency**, **Frequency**, and **MntTotal**
* Observed large variations in spending, so we transformed the monetary values using a log scale to reduce skewness

**2. Clustering with K-Means**

* Used the **Elbow Method** and **Silhouette Score** to find the ideal number of clusters (settled on **5**)
* Ran K-Means and grouped the customers into clusters based on their RFM values

**🎯 Segment Summary**

After clustering, I analyzed the average RFM scores for each group and assigned them intuitive labels:

| **Cluster** | **Recency ↓** | **Frequency ↑** | **MntTotal ↑** | **Interpretation** |
| --- | --- | --- | --- | --- |
| **0** | 72.5 (Old) | 21.6 (High) | 1092.9 (High) | *"Big Spenders (At Risk)"* |
| **1** | 24.5 (Recent) | 8.7 (Low-Med) | 109.7 (Low) | *"New Customers"* |
| **2** | 22.6 (Recent) | 21.8 (High) | 1044.8 (High) | *"Champions"* |
| **3** | 74.6 (Old) | 9.1 (Low) | 120.0 (Low) | *"Lost Customers"* |

Each segment represents a unique behavior pattern. For instance:

* **Champions** shop often, recently, and spend well — our most valuable customers.
* **Big Spenders** aren’t recent buyers but have a high overall value — time to re-engage them.
* **Lost Customers** haven’t been active and haven’t spent much — maybe try a win-back campaign.

**📈 Visualizations**

To better see how these groups behave, I created:

* A **3D scatter plot** showing Recency, Frequency, and MntTotal for each customer
* A version of the plot with just **cluster centers**, then another with **all data points**
* Applied **log transformation** to smooth out the skew in spending

<p align="center"> <img src="path-to-3d-plot.png" width="600"/> </p>

These visuals helped bring the segmentation to life and made it easier to spot trends.

**🔎 Key Takeaways**

* Only **2%** of customers were truly lost — good sign
* Around **15%** were **At Risk** – worth focusing on for retention
* The segmentation revealed actionable insights for targeting, rewards, and re-engagement strategies