Final Project

Customer Segmentation and Purchase Patterns in Online Retail

Dana Nunez
30/05/25

Dataset Overview

Purpose: Study consumer behavior, purchase frequency, and patterns to support marketing strategies.

Name: Online Retail

Main Region: United Kingdom and other European countries

Time Period: December 2010

Volume: Over 500,000 transaction records

Customer data: CustomerID

Invoice details: InvoiceNo, InvoiceDate, Country

Product info: Description, Quantity, UnitPrice

```
Shape of dataset: (541909, 8)
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Problem Statement & Objectives

Problem Statement

How can we identify different types of customers based on their purchasing behavior to help the business take action?

Project Goals

- Segment customers using RFM (Recency, Frequency, Monetary) metrics
- Apply unsupervised learning (K-Means) to group customers by patterns
- Discover high-value vs low-engagement customers
- Analyze trends in purchasing and revenue
- Create an interactive dashboard for insights and decision-making

Methodology

Data Cleaning & Preparation

- Removed duplicates and nulls
- Filtered non-positive values and canceled transactions
- Calculated revenue per transaction

RFM Feature Engineering

Recency: Days since last purchase

Frequency: Number of transactions

Monetary: Total amount spent

Unsupervised Learning (K-Means Clustering)

Normalized RFM features

Used Elbow Method to find optimal K=3

Assigned customer segments based on cluster membership

```
Number of duplicate rows: 5268
```

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	541909.000000	541909	541909.000000	406829.000000
mean	9.552250	2011-07-04 13:34:57.156386048	4.611114	15287.690570
min	-80995.000000	2010-12-01 08:26:00	-11062.060000	12346.000000
25%	1.000000	2011-03-28 11:34:00	1.250000	13953.000000
50%	3.000000	2011-07-19 17:17:00	2.080000	15152.000000
75%	10.000000	2011-10-19 11:27:00	4.130000	16791.000000
max	80995.000000	2011-12-09 12:50:00	38970.000000	18287.000000
std	218.081158	NaN	96.759853	1713.600303

	Quantity	InvoiceDate	UnitPrice	CustomerID
count	392732.000000	392732	392732.000000	392732.000000
mean	13.153718	2011-07-10 19:15:24.576301568	3.125596	15287.734822
min	1.000000	2010-12-01 08:26:00	0.000000	12346.000000
25%	2.000000	2011-04-07 11:12:00	1.250000	13955.000000
50%	6.000000	2011-07-31 12:02:00	1.950000	15150.000000
75%	12.000000	2011-10-20 12:53:00	3.750000	16791.000000
max	80995.000000	2011-12-09 12:50:00	8142.750000	18287.000000
std	181.588420	NaN	22.240725	1713.567773

Clustering Results & Customer Segments

K-Means with K=3 revealed 3 customer profiles:

Cluster	Recency	Frequency	Monetary	Insight
0	Low	Low	High	High-value buyers, worth retaining
1	Low	Low	Low	Inactive or one-time customers
2	Medium	Low	Medium	Occasional buyers, potential for growth

Key Observations

- Most customers purchase infrequently
- Cluster 0 customers spend more despite low frequency
- Visualizations:
  - RFM scatterplots helped identify patterns
  - Clusters differ mainly in **Monetary** value

Interactive Dashboard: Streamlit

Clustering Results & Customer Segments

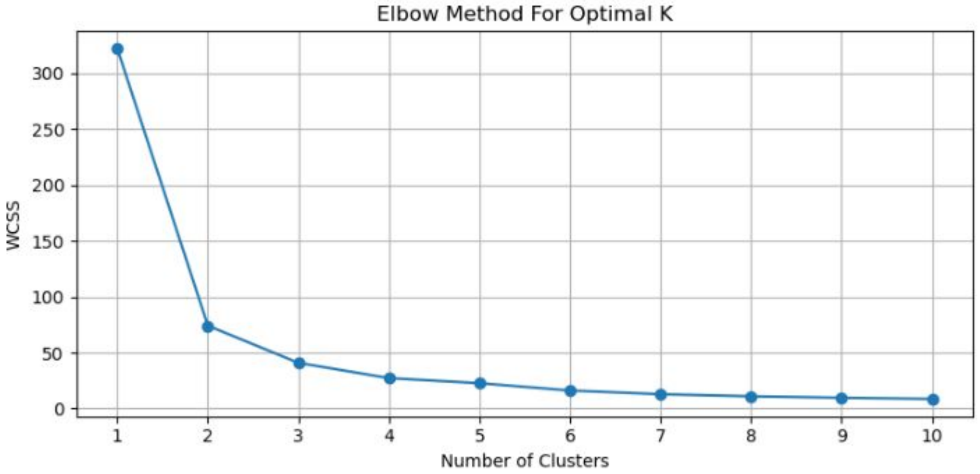
Key Insights

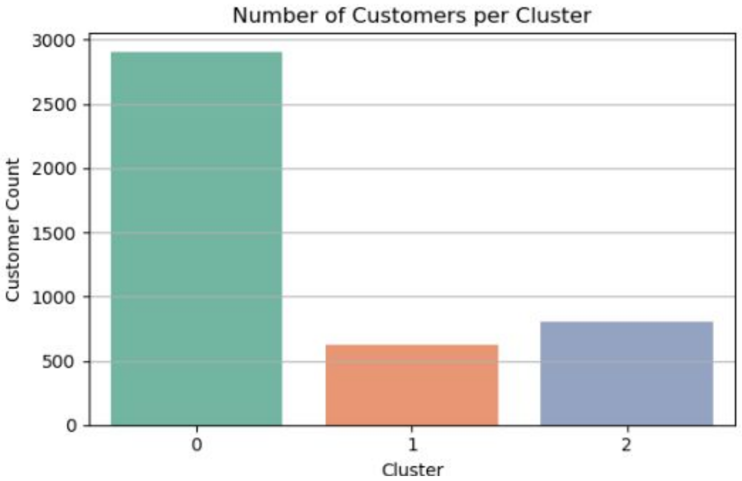
- ➔ Most customers buy infrequently (low frequency)
- ➔ A small segment contributes disproportionately high revenue (Cluster 0)
- ➔ Many purchases are recent, showing current engagement
- ➔ Returns and incomplete records influence sales trends at month-end

Recommendations

- ➢ Focus **retention strategies** on Cluster 0 (high spenders)
- ➢ Design **reactivation campaigns** for Cluster 1 (low spenders)
- ➢ Offer **personalized incentives** to boost frequency in Cluster 2
- ➢ Monitor **monthly sales trends** to anticipate inventory and marketing needs

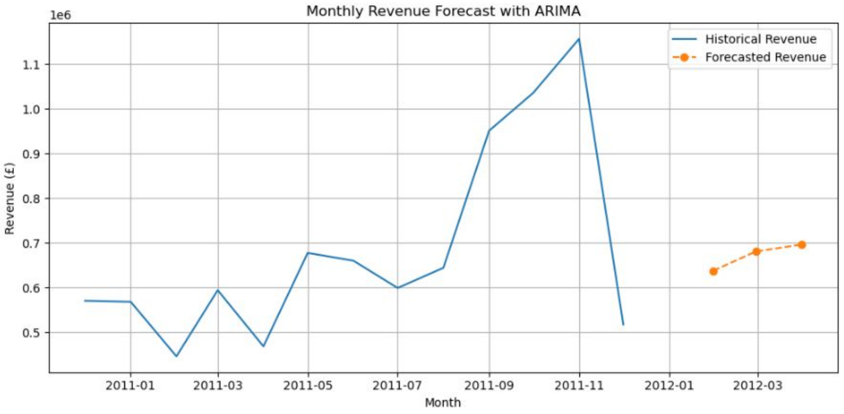
Visualizations - Clustering





Visualizations - Time Series

Forecasted monthly revenue using ARIMA. No clear seasonality observed; sharp drop due to incomplete data in the last month.



Conclusion

- The project successfully applied **RFM analysis** and **unsupervised machine learning (K-Means)** to segment customers and uncover valuable purchasing patterns.
- **Cluster 0** revealed a high-value group with low recency and high monetary scores — ideal for retention strategies.
- The **data cleaning process** was deeply tailored to the business case, removing irrelevant transactions (like cancellations) and focusing only on **active, revenue-generating purchases**.

Key Learnings

★ Applied **unsupervised clustering** (K-Means) to segment customers.

★ Strengthened skills in **data cleaning** and business-focused analysis.

★ Focused on **data preprocessing** tailored to the RFM-based segmentation.

Thank you