UNDERSTANDING ONLINE RETAIL CUSTOMER

Customer Segmentation Using RFM

Online Retail Transaction Dataset (541,909 Record) | Python & Power BI

By Muhammad Kamil Dipinto

Business Question

In an online retail business, understanding customer purchasing behavior is critical to driving sustainable sales growth. However, treating all customers the same often leads to inefficient retention and marketing strategies. It leaves a question:

> “How can transaction data help identify which customers matter most and how they should be retained?”

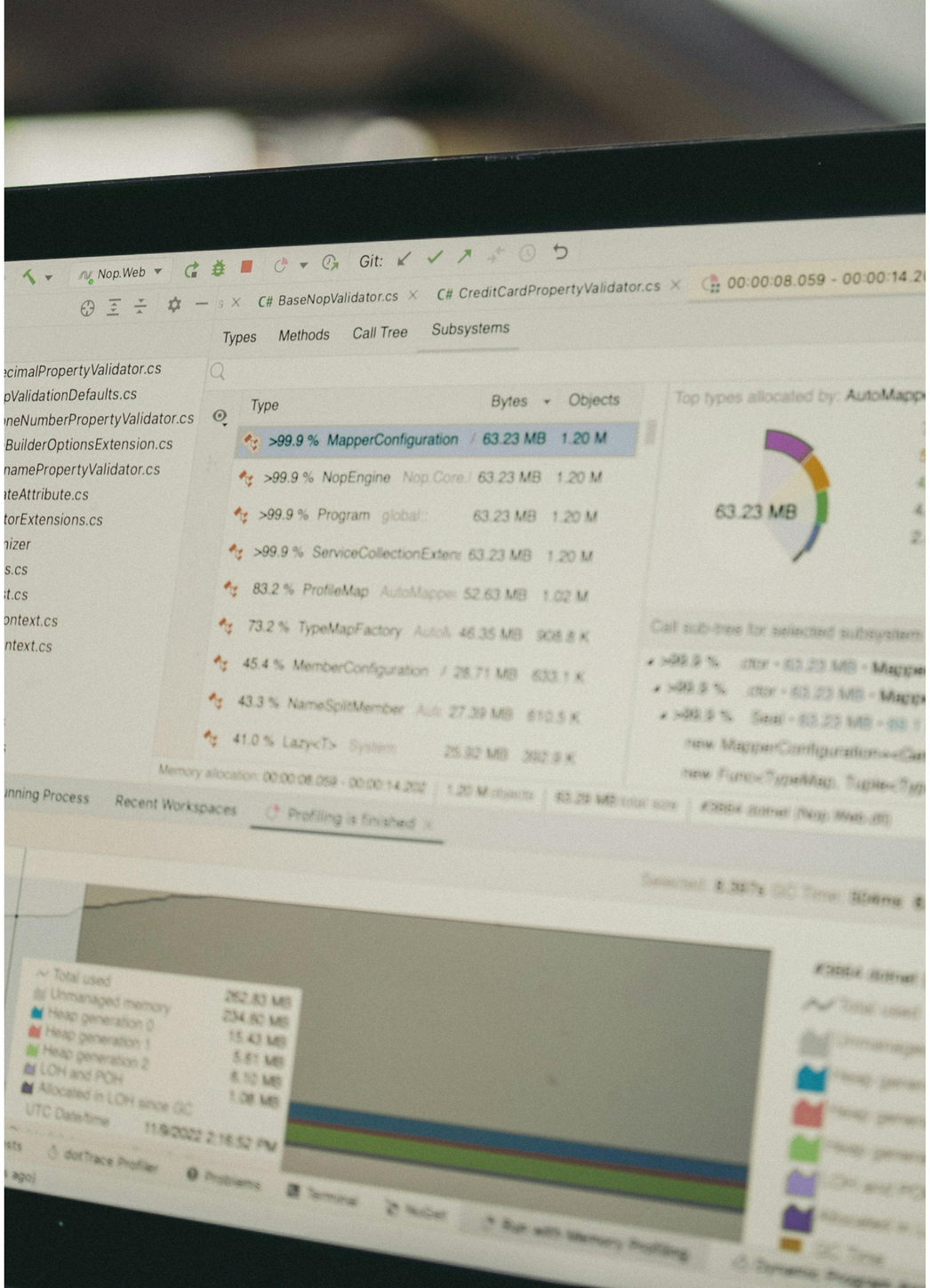
Data Overview

The analysis is based on an online retail transaction dataset containing historical purchase records. This Data contains of 541,909 transaction records, with each row represent a single purchase transaction.

Key Variable:

- **InvoiceNo** : transaction identifier
- **CustomerID** : unique customer identifier
- **InvoiceDate** : transaction timestamp
- **Quantity & UnitPrice** : purchase volume and price
- **Description** : product name
- **Country** : customer location

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



Data Cleaning (Using Python)

Before analyzing customer behavior, the transaction data was cleaned to ensure accuracy and reliability.

1

Valid transaction filtering

Dropping 2,517 non-positive value in `Quantity` and `UnitPrice`, since it is not represent valid purchasing behavior and would distort revenue-based metrics such as Monetary value.

2

Dropping Missing Customer ID

132,220 Transactions with missing `CustomerID` were removed.Records without `CustomerID` cannot be reliably assigned to any customer and are therefore excluded.

3

Date Formatting

The `InvoiceDate` column was converted from string format to datetime format to enable time-based analysis.

4

Feature Engineering

Created total transaction value by multiplying `Quantity` with `UnitPrice` and created reference date that defined as as one day after the latest transaction date. Reference date can be use at retency calculation.

After the cleaning process, the data was reduced to **397,884 data points from 541,909 data points.**

Segmentation Using RFM

Customer behavior was summarized using transaction-based metrics (RFM) to differentiate customer value and engagement.

RFM

- **Recency** – how recently a customer purchased
- **Frequency** – how often a customer purchased
- **Monetary** – how much a customer spent

Scoring

Each RFM dimension was scored on a relative scale **(0-5)** based on customer rfm distribution.

	R_Score	F_Score	M_Score
0	0	0	4
1	4	3	4
2	1	2	3
3	3	0	3
4	0	0	1

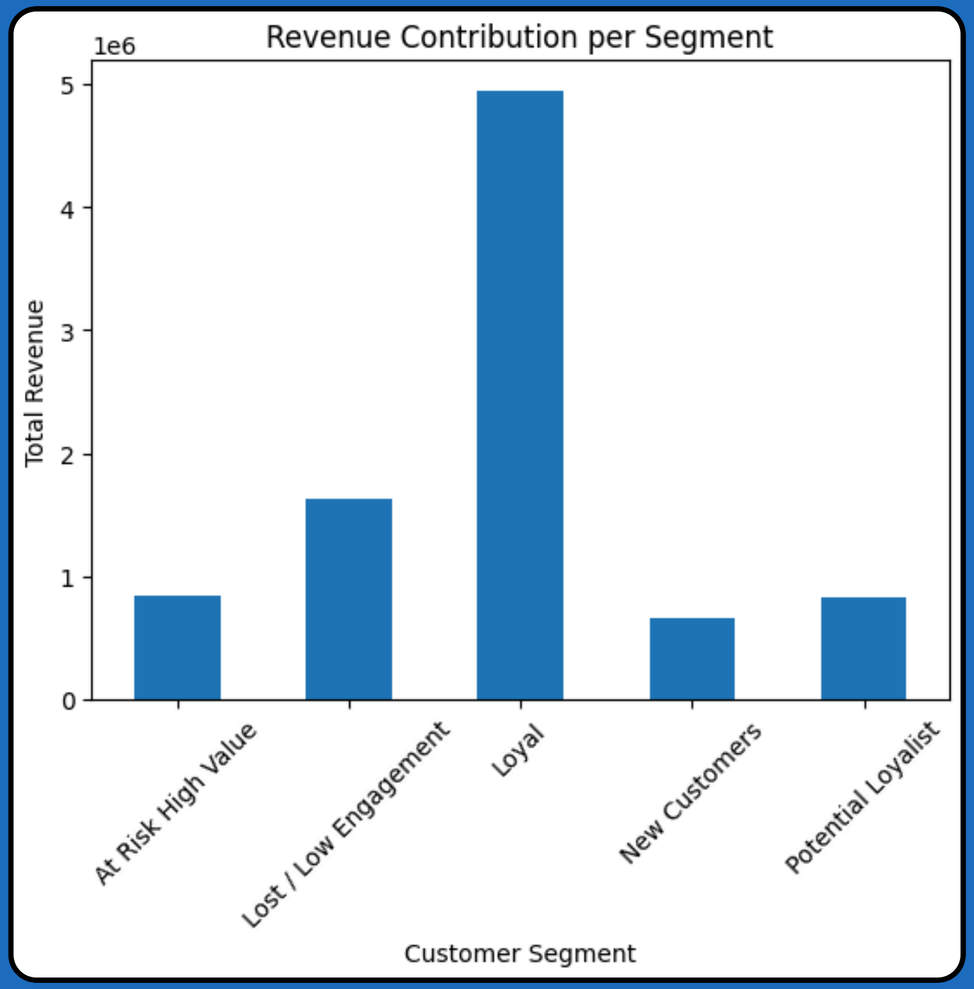
Segmentations

Customers were grouped into distinct, mutually exclusive segments:

- **Loyal**
- **Potential Loyalist**
- **New Customers**
- **At Risk High Value**
- **Lost / Low Engagement**

Customer Segmentation Overview

Segment	Description	Proportion
Loyal	High recency, frequency, and monetary value	571 (13,2%)
Potential Loyalist	Active customers with growth potential	445 (10,3%)
New Customers	Recently acquired customers	720 (16,6%)
At Risk High Value	high spenders with declining activity	340 (7,8%)
Lost / Low Engagement	inactive or low-value customers	2262 (51%)



Key Insights

A. Revenue Dependency Is Narrow

even though data shows a lot of lost customers (**>50%**) of, loyal customer is the no.1 revenue contributor by far amout.

Insight: **It means that increasing customer volume alone is unlikely to significantly improve revenue without shifting customers into higher-value segments.**

B. High-Value At Risk Customers Punch Above Their Weight

Although representing only **7.8%** of customers, this segment contributes **9.5%** of total revenue.

Insight: **This justifies focused and customized retention strategies rather than broad campaigns.**

Additional Insight

1. Top Product by Revenue

Product	Revenue
PAPER CRAFT , LITTLE BIRDIE	$168,469.6
REGENCY CAKESTAND 3 TIER	$142,592
WHITE HANGING HEART T-LIGHT HOLDER	$100,448.2
JUMBO BAG RED RETROSPOT	$85,220.8
MEDIUM CERAMIC TOP STORAGE JAR	$81,416.7

Insight:

- Top products contribute a disproportionate share of total sales
- These products likely reflect stable demand rather than one-time purchases

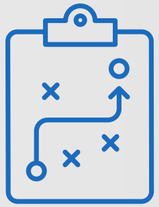
2. Top Country by Revenue

Country	Revenue
United Kingdom	$7,308,391.6
Netherlands	$285,446.3
Eire (Ireland)	$265,545.9
Germany	$228,867.1
France	$209,024.1

Insight:

- The United Kingdom dominates total revenue compared to other countries **(82%), i**ndicates potential market dependency
- Other markets contribute meaningfully but at a much smaller scale

Strategic Recommendations



1 Loyal Customer Program

Prioritize retention efforts on loyal and potential loyal customers to stabilize revenue.

- **Early access to new products**
- **Loyalty-only bundle (based on top products)**

2 Win Back High-Value Customers Early

Apply targeted and time-sensitive retention for high-value customers showing declining activity to make their comback.

- **Personalized reactivation email (based on the last purchased)**
- **Free shipping / small voucher (time-limited)**

3 Align Efforts with Revenue Hotspots

Focus retention campaigns on top-performing products and core markets.

- **Retention offer based on top products**
- **Prioritize advertise campagin on UK**

Appendix: Python Code

Check out the full project

FULL PROJECT

RFM MEASURING (FOR EACH CUSTOMER)

```
rfm = df_clean.groupby('CustomerID').agg(
    Recency=('InvoiceDate', lambda x: (reference_date - x.max()).days),
    Frequency=('InvoiceNo', 'nunique'),
    Monetary=('TotalAmount', 'sum')
).reset_index()

rfm.head()

rfm['CustomerID'] = rfm['CustomerID'].astype(str)
```

```
# RFM Sanity Check
rfm.describe()
```

	Recency	Frequency	Monetary
count	4338.000000	4338.000000	4338.000000
mean	93.059474	4.272015	2054.266460
std	100.012264	7.697998	8989.230441
min	1.000000	1.000000	3.750000
25%	18.000000	1.000000	307.415000
50%	51.000000	2.000000	674.485000
75%	142.750000	5.000000	1661.740000
max	374.000000	209.000000	280206.020000

CUSTOMER REVENUE ANALYSIS

Total Revenue Overview

```
total_revenue = rfm['Monetary'].sum()
total_customers = rfm.shape[0]
avg_revenue_per_customer = total_revenue / total_customers

print(f"Total revenue: {total_revenue}")
print(f"Total customers: {total_customers}")
print(f"Average revenue per customer: {avg_revenue_per_customer}")
```

```
Total revenue: 8911407.904
Total customers: 4338
Average revenue per customer: 2054.2664601198708
```

The dataset contains 4338 customers with a total revenue of 8,911,407.904 USD. On average, each customer contributes about 2,054.27 USD to total revenue.

Revenue Distribution

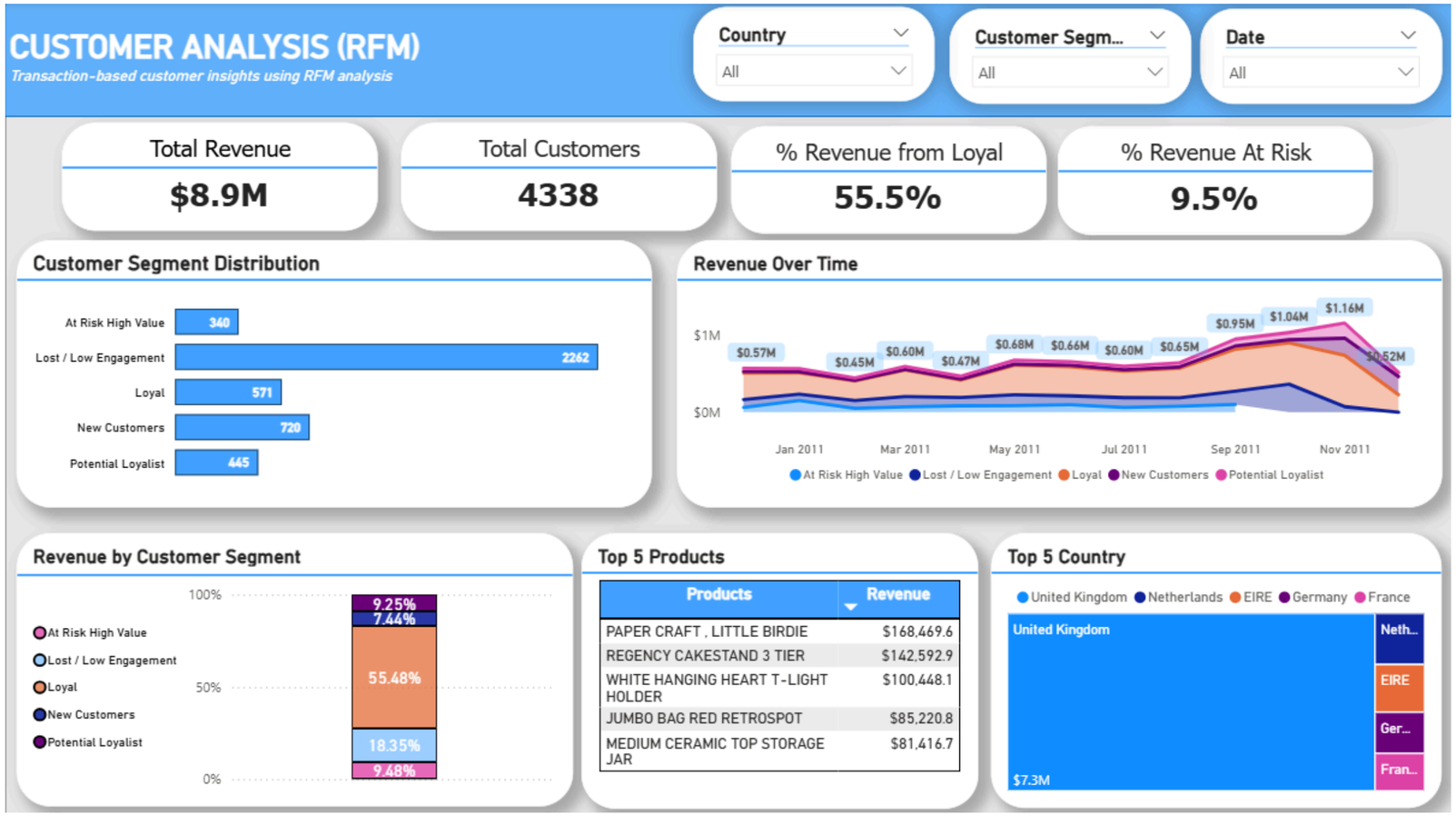
```
# Histogram
import matplotlib.pyplot as plt

plt.hist(rfm['Monetary'], bins=50)
plt.xlabel('Customer Revenue')
plt.ylabel('Number of Customers')
plt.title('Distribution of Customer Revenue')
plt.show()
```

Power BI Dashboard

An interactive Power BI Dashboard, provide a high-level view of customer value, risk, and revenue drivers to support retention-focused decisions. This dashboard also include slicer for Country, Customer segment, and Date. This Dashboard answer questions like:

- **Where is revenue concentrated across customer segments?**
- **Which customer segments contribute to revenue risk?**
- **What products and markets drive overall revenue?**



Thank you!

Muhammad Kamil Dipinto

dipintom3@gmail.com

www.linkedin.com/in/muhammadkamildipinto

Data Source:
Public dataset obtained from Kaggle
(Online Retail Dataset)

Tools:
Python (Pandas), Power BI

This project demonstrates how transaction data can be translated into actionable customer retention insights.