Alfido Tech – Data Analytics Internship

Customer Behavior Analysis Report

Name: Yash Ajit Dokhe

Role: Data Analytics Intern

Tools: Python, Pandas, NumPy, Matplotlib, Seaborn

Platform: Google Collab

---

1. Introduction / Problem Statement

The objective of this project is to analyze customer transaction data to understand purchasing behavior, segment customers using data-driven techniques, analyze retention and churn risk, and provide actionable business recommendations to improve customer engagement.

---

2. Dataset Overview

- **Source:** Kaggle – Customer Behavior Analysis Dataset
- **Records:** 135,000+ transaction records
- **Key Columns:**
  Customer ID, Purchase Date, Product Category, Quantity,
  Total Purchase Amount, Payment Method, Returns, Age, Gender, Churn

Two datasets were provided. After cross-dataset validation, it was observed that the same Customer ID had different customer names across datasets. To avoid incorrect assumptions and data inconsistency, only the raw transaction dataset was used for analysis.

	Customer ID	Purchase Date	Product Category	Product Price	Quantity	Total Purchase Amount	Payment Method	Customer Age	Returns	Customer Name	Age	Gender	Churn
0	44605	2023-05-03 21:30:02	Home	177	1	2427	PayPal	31	1.0	John Rivera	31	Female	0
1	44605	2021-05-16 13:57:44	Electronics	174	3	2448	PayPal	31	1.0	John Rivera	31	Female	0
2	44605	2020-07-13 06:16:57	Books	413	1	2345	Credit Card	31	1.0	John Rivera	31	Female	0
3	44605	2023-01-17 13:14:36	Electronics	396	3	937	Cash	31	0.0	John Rivera	31	Female	0
4	44605	2021-05-01 11:29:27	Books	259	4	2598	PayPal	31	1.0	John Rivera	31	Female	0

Figure 1: Preview of the raw transaction dataset showing key columns and structure.

---

3. Tools & Technologies Used

- Python
- Pandas & NumPy (Data manipulation)
- Matplotlib & Seaborn (Visualization)
- Google Colab (Development environment)

---

4. Data Cleaning & Preprocessing

The following steps were performed to clean the data:

- Checked missing values and handled them appropriately
- Filled missing values in the Returns column with 0
- Converted Purchase Date to datetime format
- Removed duplicate records
- Performed final sanity checks

These steps ensured the dataset was accurate and ready for analysis.

	0
Customer ID	0
Purchase Date	0
Product Category	0
Product Price	0
Quantity	0
Total Purchase Amount	0
Payment Method	0
Customer Age	0
Returns	47382
Customer Name	0
Age	0
Gender	0
Churn	0

Figure 2: Missing value analysis and data cleaning results.

```
RangeIndex: 250000 entries, 0 to 249999
Data columns (total 13 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Customer ID             250000 non-null  int64
 1   Purchase Date           250000 non-null  datetime64[ns]
 2   Product Category        250000 non-null  object
 3   Product Price           250000 non-null  int64
 4   Quantity                250000 non-null  int64
 5   Total Purchase Amount   250000 non-null  int64
 6   Payment Method          250000 non-null  object
 7   Customer Age            250000 non-null  int64
 8   Returns                 250000 non-null  float64
 9   Customer Name           250000 non-null  object
 10  Age                     250000 non-null  int64
 11  Gender                  250000 non-null  object
 12  Churn                   250000 non-null  int64
```

Figure 3: Dataset summary after cleaning and preprocessing.

---

5. Feature Engineering (Customer-Level Dataset)

Transaction-level data was aggregated to create a customer-level dataset. The following features were engineered:

- Total number of orders
- Total quantity purchased
- Total spending
- Average order value
- Last purchase date
- Total returns
- Churn indicator
- Age and Gender

Additionally, **Recency (in days)** was calculated using the difference between the most recent purchase date in the dataset and each customer's last purchase date.

	Customer ID	Total_Order	Total_quantity	Total_spent	Avg_Order_value	Last_purchase_date	Total_return	Churn	Age	Gender
0	1	3	15	6290	2096.666667	2022-11-29 06:48:25	0.0	0	67	Female
1	2	6	18	16481	2746.833333	2023-07-03 17:26:19	4.0	0	42	Female
2	3	4	15	9423	2355.750000	2023-02-03 03:58:07	0.0	0	31	Male
3	4	5	19	7826	1565.200000	2022-06-29 03:41:09	3.0	0	37	Male
4	5	5	13	9769	1953.800000	2022-07-16 04:08:09	3.0	0	24	Female

Figure 4: Customer-level feature engineered dataset (customer_df).

	Customer ID	Total_Order	Total_quantity	Total_spent	Avg_Order_value	Last_purchase_date	Total_return	Churn	Age	Gender	Recency_days
0	1	3	15	6290	2096.666667	2022-11-29 06:48:25	0.0	0	67	Female	288
1	2	6	18	16481	2746.833333	2023-07-03 17:26:19	4.0	0	42	Female	72
2	3	4	15	9423	2355.750000	2023-02-03 03:58:07	0.0	0	31	Male	222
3	4	5	19	7826	1565.200000	2022-06-29 03:41:09	3.0	0	37	Male	441
4	5	5	13	9769	1953.800000	2022-07-16 04:08:09	3.0	0	24	Female	424

Figure 5: Calculation of customer recency based on last purchase date.

---

6. Customer Segmentation using RFM Analysis

Customers were segmented using the **RFM (Recency, Frequency, Monetary)** model:

- **Recency:** Days since last purchase
- **Frequency:** Total number of orders
- **Monetary:** Total spending

Each metric was divided into quartiles and scored. The combined RFM score was used to segment customers into:

- High-Value Customers
- Mid-Value Customers
- Low-Value Customers

	Customer ID	Total_Order	Total_quantity	Total_spent	Avg_Order_value	Last_purchase_date	Total_return	Churn	Age	Gender	Recency_days	R_score	F_score	M_score	RFM_score	Segment
0	1	3	15	6290	2096.666667	2022-11-29 06:48:25	0.0	0	67	Female	288	2	1	1	4	Low-Value
1	2	6	18	16481	2746.833333	2023-07-03 17:26:19	4.0	0	42	Female	72	4	3	3	10	High-Value
2	3	4	15	9423	2355.750000	2023-02-03 03:58:07	0.0	0	31	Male	222	2	2	2	6	Mid-Value
3	4	5	19	7826	1565.200000	2022-06-29 03:41:09	3.0	0	37	Male	441	1	2	1	4	Low-Value
4	5	5	13	9769	1953.800000	2022-07-16 04:08:09	3.0	0	24	Female	424	1	2	2	5	Low-Value

Figure 6: RFM score calculation and customer segmentation output.

---

7. Segment Profiling

Each customer segment was profiled using the following metrics:

- Total customers
- Average recency days

- Average number of orders
- Average spending
- Total revenue contribution

This profiling helped compare behavior across segments.

	Segment	total_customers	avg_recency_days	avg_orders	avg_spent	total_revenue
0	High-Value	12916	101.862264	7.685197	21690.672190	280156722
1	Low-Value	13837	499.143890	2.825468	7006.799089	96953079
2	Mid-Value	22908	207.607342	4.873494	13280.797014	304236498

Figure 7: Segment-wise customer profiling and revenue contribution.

---

8. Visualization of Purchase Patterns & Retention Trends

The following visualizations were created:

- Average number of orders per customer segment
- Average spending per segment
- Distribution of recency days
- Average recency days by customer segment

These visualizations highlight loyalty, engagement, and churn risk.

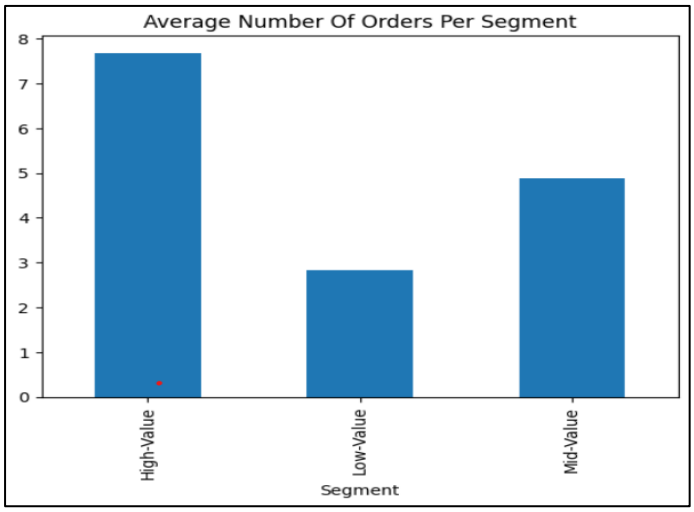

Figure 8: Average number of orders across customer segments

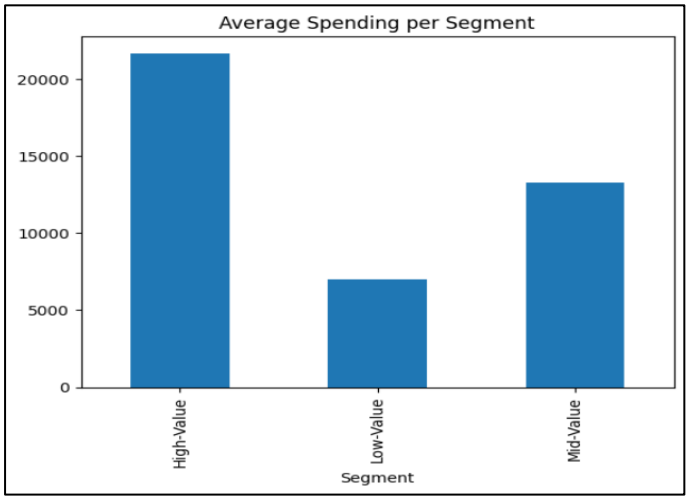

Figure 9: Average spending per customer segment.

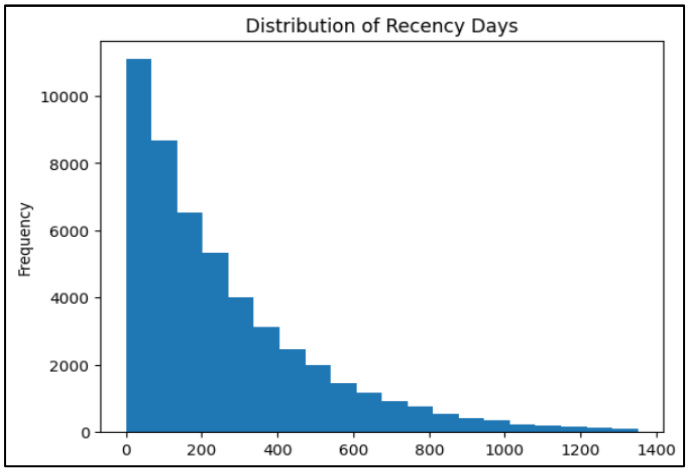

Figure 10: Distribution of customer recency (days since last purchase).

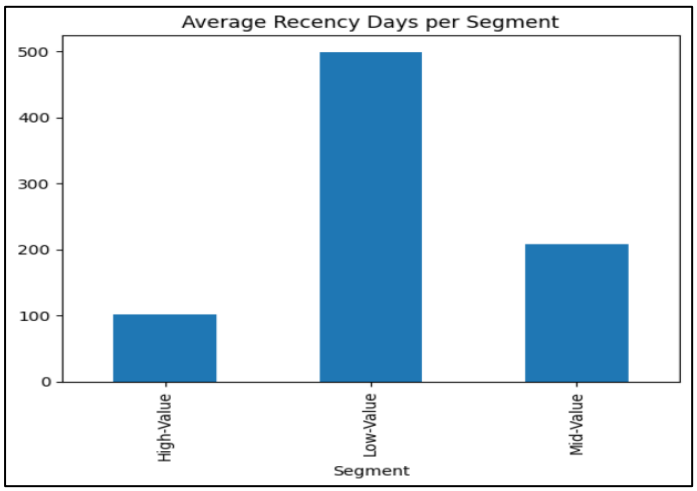

Figure 11: Average recency days by customer segment.

9. Key Insights

- High-Value customers generate a significant portion of total revenue.
- Mid-Value customers represent the largest group and show strong growth potential.
- Low-Value customers have high inactivity and churn risk.
- Recency and frequency are strong indicators of customer loyalty.
- Segment-based analysis provides more meaningful insights than overall averages.

10. Actionable Recommendations

1. Retain High-Value customers through loyalty programs and personalized offers.
2. Convert Mid-Value customers into High-Value customers using targeted promotions.
3. Re-engage Low-Value customers with reminder campaigns and discounts.
4. Implement segment-based marketing strategies instead of one-size-fits-all approaches.
5. Monitor customer activity regularly to identify churn risk early.

11. Conclusion

This project successfully analyzed customer behavior and segmented customers using RFM analysis. The insights generated can help Alfido Tech improve customer engagement, reduce churn, and increase revenue through data-driven decision-making.