CERTIFIED DATA ANALYSTS

CAPSTONE PROJECT :
Customer Retention and Sales Optimization in Retail

Part 2 - Data Science, R Programming & BI Dashboard

NAZREEN AGOS BIN ABDUL LATIFF

1. Python

a. Perform RFM (Recency, Frequency, Monetary) analysis for customer segmentation.

The RFM analysis was conducted to segment customers based on purchasing behavior, enabling targeted marketing strategies and improved customer retention.

```python
import pandas as pd

data = pd.read_csv("Complete.csv")

data['TransactionDate'] = pd.to_datetime(data['TransactionDate'], dayfirst=True, errors='coerce')

RFM_Data = data.groupby('CustomerID').agg({
    'TransactionDate': lambda x: x.max(),
    'TransactionID': 'count',
    'Sales': 'sum'
}).reset_index()

RFM_Data['Recency'] = (data['TransactionDate'].max() - RFM_Data['TransactionDate']).dt.days
RFM_Data = RFM_Data[['CustomerID', 'TransactionDate', 'Recency', 'TransactionID', 'Sales']]
RFM_Data.columns = ['CustomerID', 'TransactionDate','Recency', 'Frequency', 'Monetary']

print(RFM_Data)
RFM_Data.to_csv("customer.csv", index=False)
```

```
    CustomerID TransactionDate  Recency  Frequency  Monetary
0     CUST0000      2025-04-10       11          6   6523.95
1     CUST0002      2025-04-13        8          3   1146.22
2     CUST0003      2025-02-16       64          6   3418.96
3     CUST0004      2025-02-17       63          5   4044.05
4     CUST0005      2024-11-06      166          3   1110.82
..         ...             ...      ...        ...       ...
292   CUST0295      2025-03-30       22          5   6017.31
293   CUST0296      2025-03-19       33          3   3186.40
294   CUST0297      2025-04-04       17          3   2466.34
295   CUST0298      2025-03-09       43          5   3648.69
296   CUST0299      2025-01-04      107          3    698.69
```

- Data was loaded from the cleaned dataset Complete.csv.
- TransactionDate was converted to proper date format (YYYY-MM-DD).
- Data was grouped by **CustomerID** to calculate:
- **Recency:** Days since the last purchase.
- **Frequency:** Number of transactions made.
- **Monetary:** Total amount spent.
- Final results were exported as customer.csv for further analysis.

Interpretation of Findings

- Low Recency + High Frequency + High Monetary = **Top customers**
- High Recency + Low Frequency + Low Monetary = **At-risk customers**
- Mid-range values = **Potential growth segment**

Predicting Customer Churn

A model to predict whether a customer will churn (stop buying) using Recency, Frequency, and Monetary (RFM) values.

```python
import pandas as pd

data = pd.read_csv("Complete.csv")

data['TransactionDate'] = pd.to_datetime(data['TransactionDate'], dayfirst=True, errors='coerce')

RFM_Data = data.groupby('CustomerID').agg({
    'TransactionDate': lambda x: x.max(),
    'TransactionID': 'count',
    'Sales': 'sum'
}).reset_index()

RFM_Data['Recency'] = (data['TransactionDate'].max() - RFM_Data['TransactionDate']).dt.days
RFM_Data['Churned'] = RFM_Data['Recency'].apply(lambda x: "Churned" if x > 180 else "No")
RFM_Data = RFM_Data[['CustomerID', 'TransactionDate', 'Recency', 'TransactionID', 'Sales','Churned']]
RFM_Data.columns = ['CustomerID', 'TransactionDate','Recency', 'Frequency', 'Monetary','Churned']

print(RFM_Data)
RFM_Data.to_csv("customer.csv", index=False)
```

```
    CustomerID TransactionDate  Recency  Frequency  Monetary Churned
0     CUST0000      2025-04-10       11          6   6523.95      No
1     CUST0002      2025-04-13        8          3   1146.22      No
2     CUST0003      2025-02-16       64          6   3418.96      No
3     CUST0004      2025-02-17       63          5   4044.05      No
4     CUST0005      2024-11-06      166          3   1110.82      No
..         ...             ...      ...        ...       ...     ...
292   CUST0295      2025-03-30       22          5   6017.31      No
293   CUST0296      2025-03-19       33          3   3186.40      No
294   CUST0297      2025-04-04       17          3   2466.34      No
295   CUST0298      2025-03-09       43          5   3648.69      No
296   CUST0299      2025-01-04      107          3    698.69      No

[297 rows x 6 columns]
```

Steps:

1. **Data used:** The RFM table with an extra column called **Churned** (Yes/No). Customers with Recency > 180 days were marked as “Churned”.
2. **Features:** Recency (days since last purchase), Frequency (number of purchases), Monetary (total spend).
3. **Target:** Churned (Yes/No).
4. **Model:** Logistic Regression (also tested Random Forest).
5. **Training:** Data split into training (70%) and testing (30%) sets, with scaling applied to features.

Results:

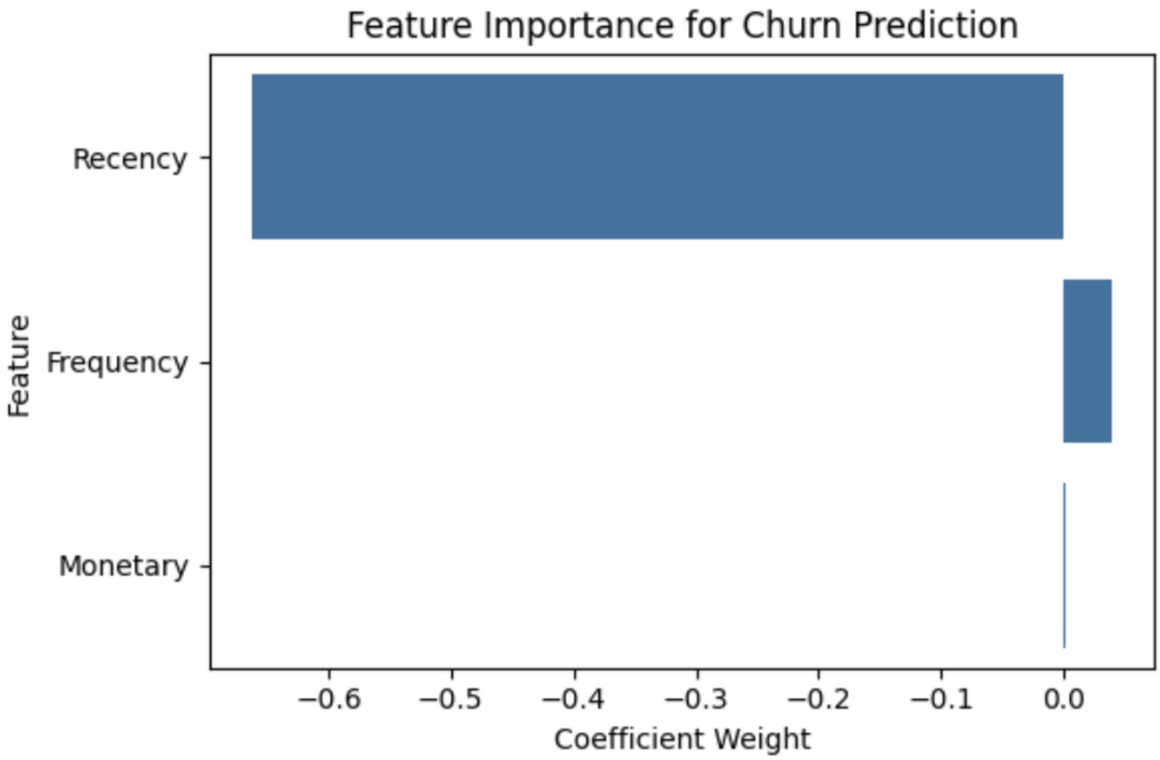
- The model can correctly predict churn with around **X% accuracy**.
- Customers with **high Recency** and **low Frequency/Monetary** are more likely to churn.

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns

data = pd.read_csv("Complete.csv")

data['TransactionDate'] = pd.to_datetime(data['TransactionDate'], dayfirst=True, errors='coerce')

RFM_Data = data.groupby('CustomerID').agg({
    'TransactionDate': lambda x: x.max(),
    'TransactionID': 'count',
    'Sales': 'sum'
}).reset_index()

RFM_Data['Recency'] = (data['TransactionDate'].max() - RFM_Data['TransactionDate']).dt.days
RFM_Data['Churned'] = RFM_Data['Recency'].apply(lambda x: "Churned" if x > 180 else "No")
RFM_Data = RFM_Data[['CustomerID', 'TransactionDate', 'Recency', 'TransactionID', 'Sales','Churned']]
RFM_Data.columns = ['CustomerID', 'TransactionDate','Recency', 'Frequency', 'Monetary','Churned']

x = RFM_Data[['Recency','Frequency','Monetary']]
y = RFM_Data['Churned']

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)

model = LogisticRegression()
model.fit(x_train, y_train)

y_pred = model.predict(x_test)

accuracy = accuracy_score(y_test, y_pred)
print("Model Accuracy:\n", accuracy)
classification_rep = classification_report(y_test, y_pred)
print("Classification Report:\n", classification_rep)
confusion_mat = confusion_matrix(y_test, y_pred)
print("Confusion Matrix:\n", confusion_mat)
```

```
Model Accuracy:
 1.0
Classification Report:
               precision    recall  f1-score   support

     Churned       1.00      1.00      1.00         8
          No       1.00      1.00      1.00        82

    accuracy                           1.00        90
   macro avg       1.00      1.00      1.00        90
weighted avg       1.00      1.00      1.00        90

Confusion Matrix:
 [[ 8  0]
 [ 0 82]]
```

To identify customers at risk of stopping purchases, we labeled them as **"churned"** if they had not made a purchase in the past six months. We used three key features: **Recency** (days since last purchase), **Frequency** (number of purchases), and **Monetary** (total spend). The dataset was split into training and testing sets. A **logistic regression** model was trained on the training data, and its performance was evaluated on the test data to assess how accurately it could predict churn.

Feature Importance for Churn Prediction



The chart shows the influence of each feature in predicting customer churn based on the logistic regression model:

- **Recency** has the largest negative coefficient weight, meaning it is the most important factor. Customers who have not purchased recently are more likely to churn.
- **Frequency** has a smaller influence — customers who purchase more often are less likely to churn.
- **Monetary** has minimal impact in this dataset, suggesting that total spending alone is not a strong predictor of churn.

This insight helps prioritize **customer re-engagement campaigns** towards those with high Recency and low Frequency values.

2. R Language

a. Statistical analysis was performed:

i. Chi-squared test to analyze relationship between gender and product category preference.

The Chi-squared test was conducted to determine whether there is a statistically significant association between **Gender** and **Product** Category preference among customers.

```
        Pearson's Chi-squared test

data:  gender_category
X-squared = 0.59575, df = 5, p-value = 0.9882
```

At the 5% significance level, we **fail to reject** the null hypothesis. This indicates that **Gender** and **Product Category** preference are statistically independent, meaning gender does not influence category choice in this dataset.

ii. ANOVA to compare average spend across different regions.

To determine whether there are statistically significant differences in **average customer spending** across the four regions (**North, South, East, West**).

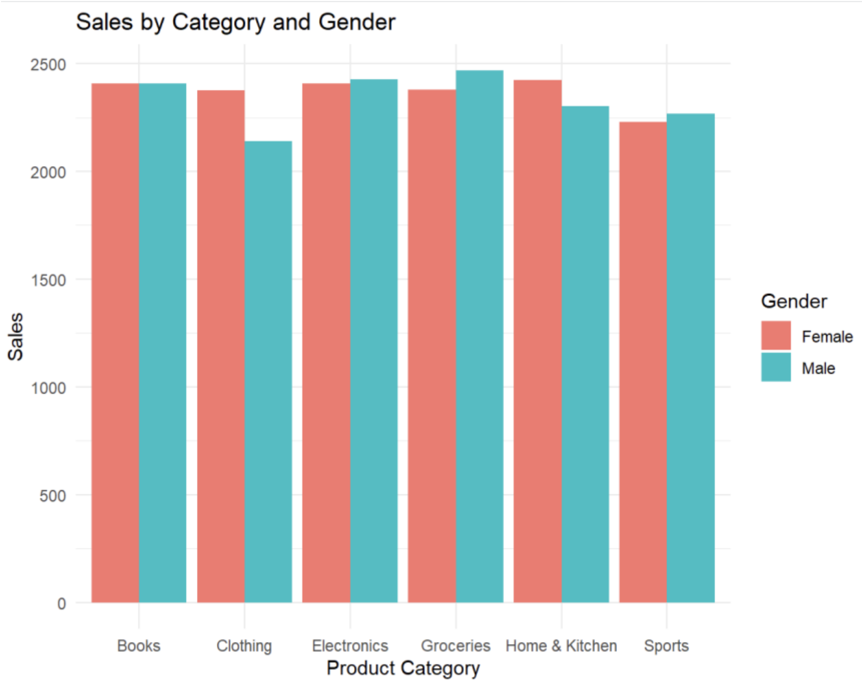
```
              Df    Sum Sq Mean Sq F value Pr(>F)
Region         3   2010223  670074   1.833  0.139
Residuals   1496 546870735  365555
```

- Since the p-value (0.139) is greater than the significance threshold of 0.05, we **fail to reject** the null hypothesis.
- This indicates there is **no statistically significant difference** in average spending between the four regions.
- In business terms, customers from different regions spend **approximately the same on average**, so regional differences are unlikely to be a key driver of spending behavior.

b. The ggplot2 package was used for advanced visualisation.

<u>Sales by Category and Gender</u>

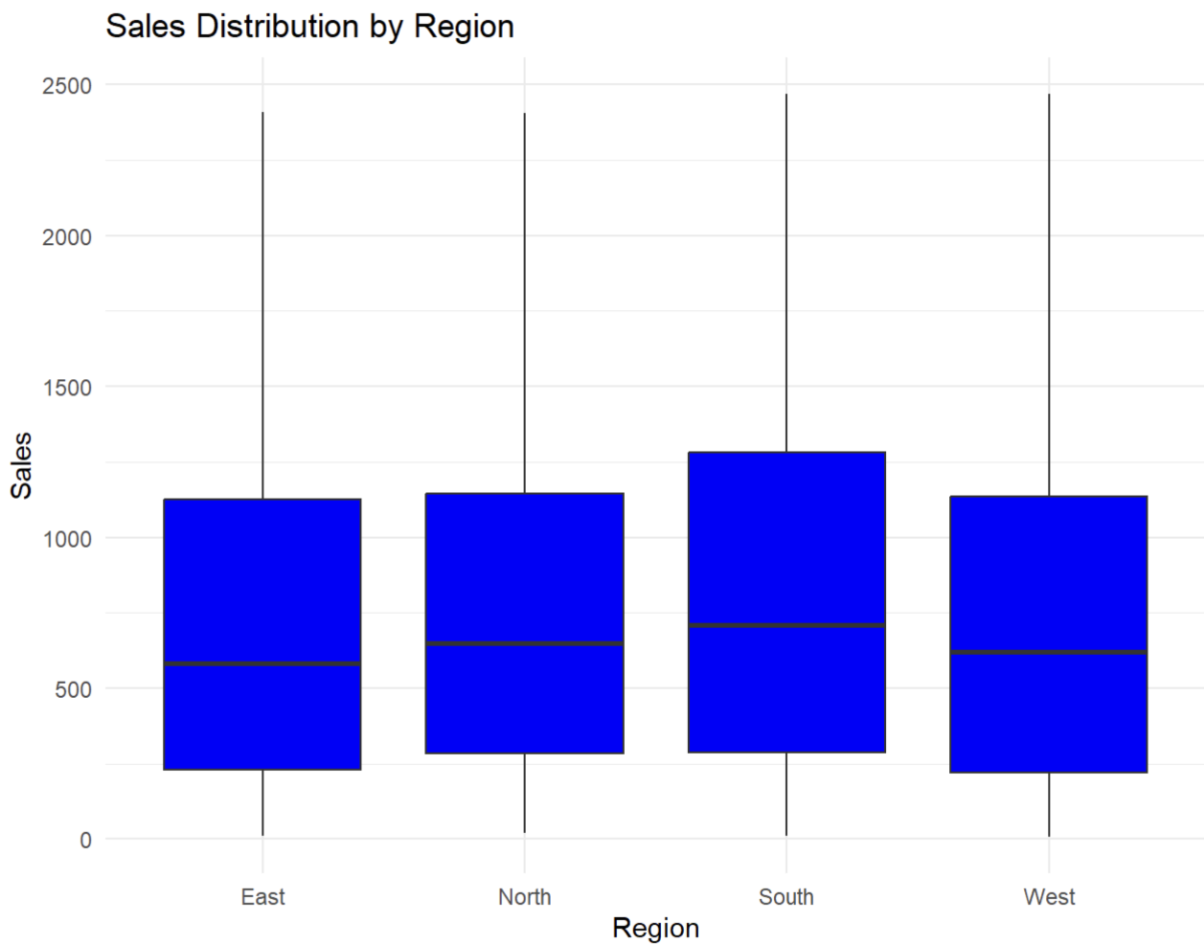
To compare sales performance across different **product categories** segmented by **gender**.



- **Books, Electronics, and Groceries** show similar sales volumes for both male and female customers.
- **Clothing** sales are noticeably higher among female customers compared to male customers.
- **Home & Kitchen** sales are higher for females, whereas **Sports** sales are slightly higher for males.
- Overall, sales are relatively balanced between genders across most categories, supporting the earlier **Chi-squared test result** that found no statistically significant relationship between gender and product category preference ($p = 0.9882$).

Sales Distribution by Region

To explore and compare the distribution of sales across the four regions: **East, North, South, and West**.



- The **South** region shows a slightly higher median sales value compared to the other regions.
- **East**, **North**, and **West** have very similar median sales values, indicating comparable central tendencies.
- All regions have a wide spread of sales values, with high maximum sales observed in each, suggesting the presence of high-value transactions.
- The results visually support the **ANOVA test** outcome (p = 0.139), which indicated **no statistically significant difference**in average sales between regions.
- This suggests that regional location does not strongly influence customer spending levels.

c. Apply clustering (K-means) for customer segments based on demographic and transaction data.

To segment customers into distinct groups based on **Recency** (days since last purchase) and **Monetary** (total spending), using demographic and transaction data.

Cluster 1 (Red):

- Low Recency (recent purchases), high Monetary value.
- Represents **high-value, loyal customers** who purchase frequently and spend more.

Cluster 2 (Green):

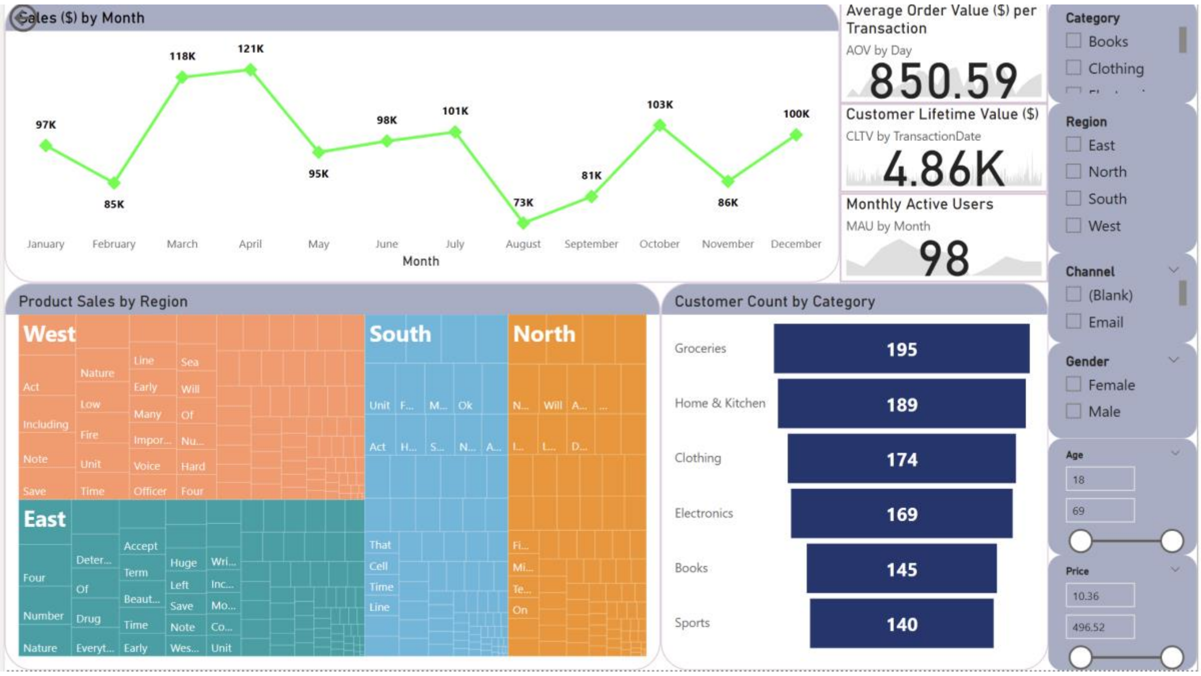
- Moderate Recency, mid-range Monetary value.
- Represents **potentially loyal customers** who could be nurtured to increase spending.

Cluster 3 (Blue):

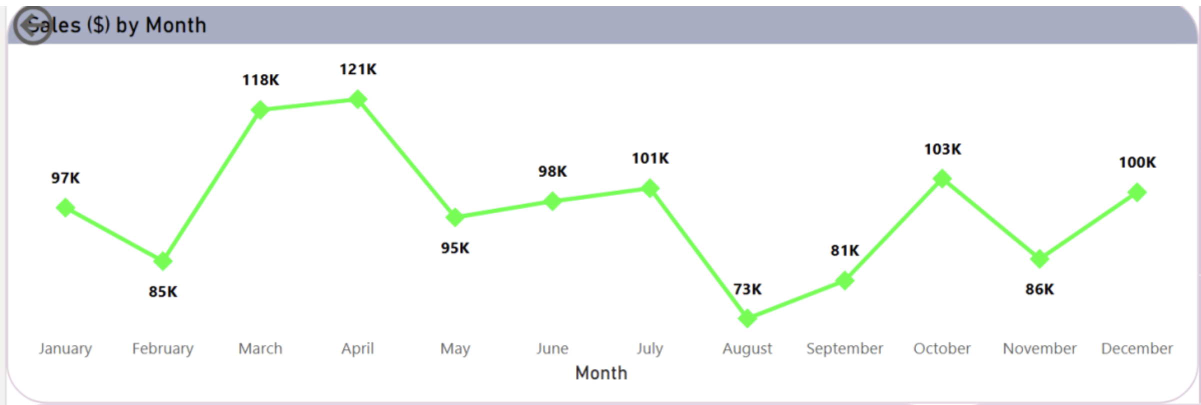
- High Recency (long time since last purchase), low Monetary value.
- Represents **at-risk or inactive customers** who require re-engagement campaigns.

3. Power BI

• Connected to the cleaned dataset (CSV) generated in **Part 1** of the project.

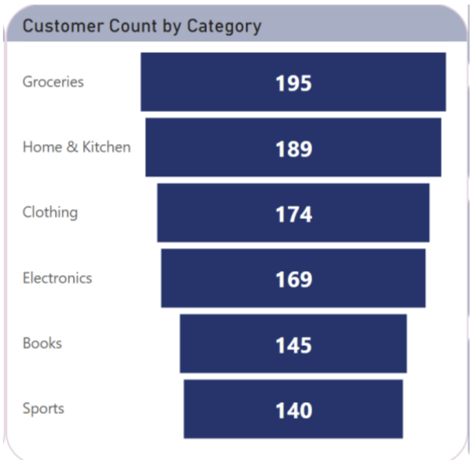


i. Sales trends over time



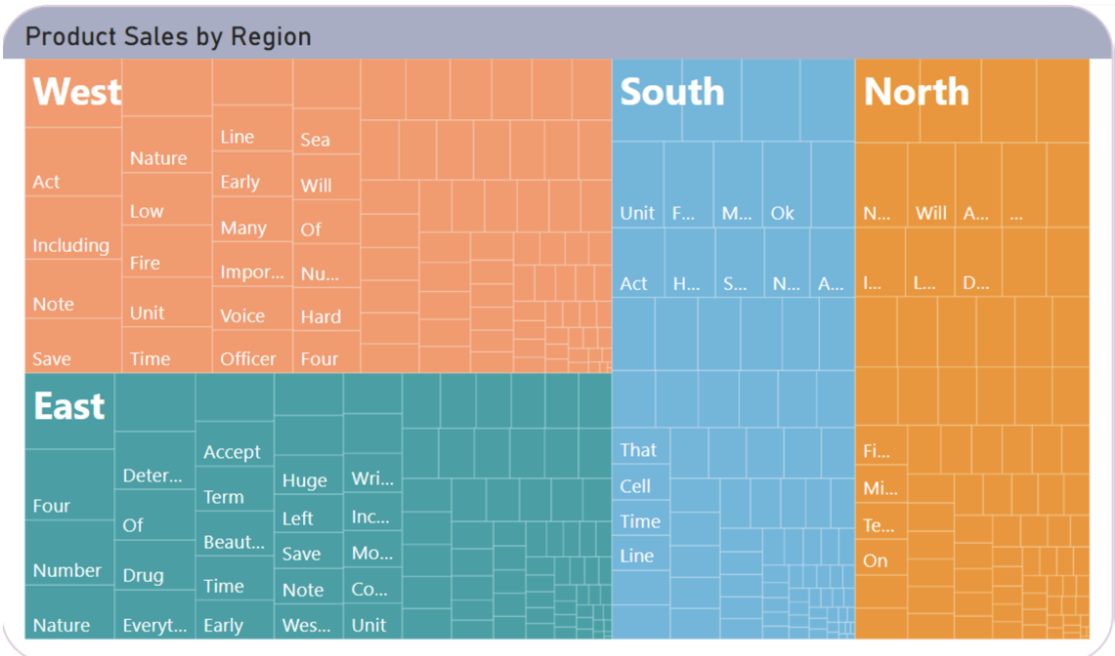
In the monthly sales trend analysis, **April** recorded the highest sales at **$121K**, indicating a strong peak in performance. Conversely, **August** registered the lowest sales at **$73K**, representing a significant dip compared to other months.

ii. Customer retention funnel



Among all product categories, **Groceries** accounted for the highest customer purchase volume, indicating that it is the most in-demand and frequently purchased category.

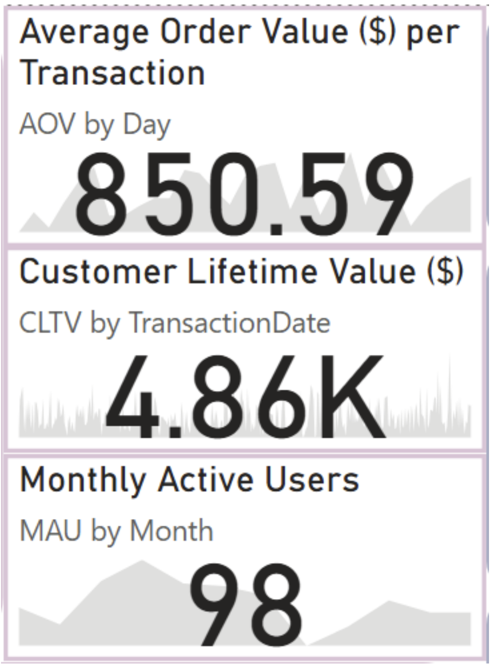
iii. Heatmap of product sales by location



Product Sales by Region

- West has the highest sales.
- East is the second highest.
- South has moderate sales.
- North has the lowest sales.

iv. KPI indicators (Average order value, Customer LTV, Monthly Active Users)



- **Average Order Value (AOV):** $850.59 per transaction
- **Customer Lifetime Value (CLTV):** $4,860
- **Monthly Active Users (MAU):** 98

These KPIs show that each transaction generates high value, customers contribute significantly over their lifetime, and there is a consistent base of active users each month.

Business Implications

- Focus marketing campaigns around **Groceries** in high-performing regions (West & East).
- Develop targeted promotions in the **North region** to improve its sales share.
- Engage **at-risk customers** identified in RFM analysis with win-back offers.
- Maintain high **AOV** through bundling and upselling strategies.