Exploratory Data Analysis

Overview

Exploratory Data Analysis is the finding and understanding of the underlying patterns and distributions in a dataset. EDA helps in making informed decisions about feature selection, data preprocessing, and model building. It provides a solid foundation for subsequent data analysis and machine learning tasks.

In this project eda is involved in finding patterns and in the process infer business insights from the dataset.

Data Overview

1. Datasets Used:
   (a) Customers.csv: Contains demographic and profile information such as customer ID, age, and location.
   (b) Transactions.csv: Includes transactional data, such as purchase recency, frequency, and monetary value (RFM metrics).
2. Preprocessing Steps:
   (a) Data cleaning and merging based on CustomerID.
   (b) Feature engineering to calculate:
      i. Recency: Time since last purchase.
      ii. Frequency: Number of transactions.
      iii. Monetary Value: Total spending.

Content

Dataset Used

The dataset used is from the files Customers.csv, Products.csv and Transactions.csv. The datasets are merged using inner join on ProductID and CustomerID. The dataset was structured as:

1. TransactionID 1000 non-null object
2. CustomerID 1000 non-null object
3. ProductID 1000 non-null object
4. TransactionDate 1000 non-null object
5. Quantity 1000 non-null int64
6. TotalValue 1000 non-null float64
7. Price 1000 non-null float64
8. CustomerName 1000 non-null object
9. Region 1000 non-null object
10. SignupDate 1000 non-null object

from this dataset few more features were derived:

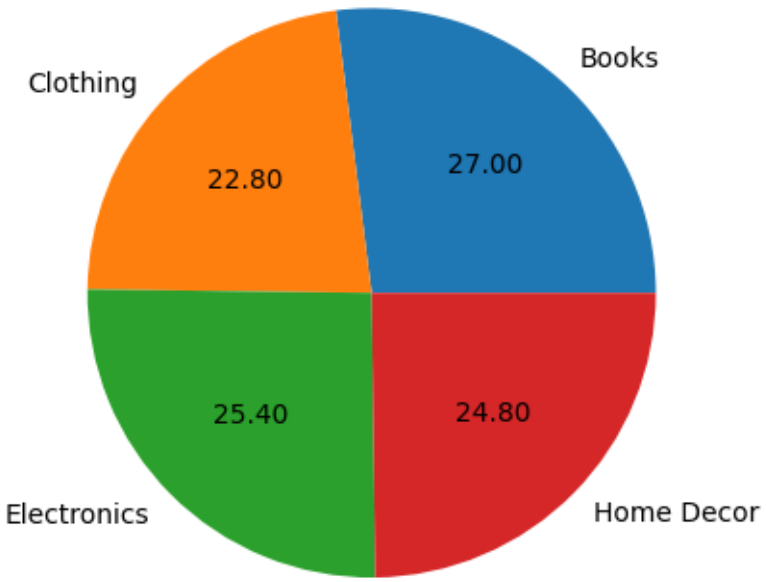
1. TransactionID 1000 non-null object
2. CustomerID 1000 non-null object
3. ProductID 1000 non-null object
4. TransactionDate 1000 non-null datetime64[ns]
5. Quantity 1000 non-null int64
6. TotalValue 1000 non-null float64
7. Price 1000 non-null float64
8. CustomerName 1000 non-null object

9. Region 1000 non-null object
10. SignupDate 1000 non-null datetime64[ns]
11. CustomerTenure 1000 non-null int64
12. TransactionFrequency 1000 non-null int64
13. Recency 1000 non-null int64

The key features considered in the clustering algorithm are Quantity, TotalValue, Price, CustomerTenure, TransactionFrequency, Recency. These features were passed through the StandardScaler.fit_transform() to standardize the data.

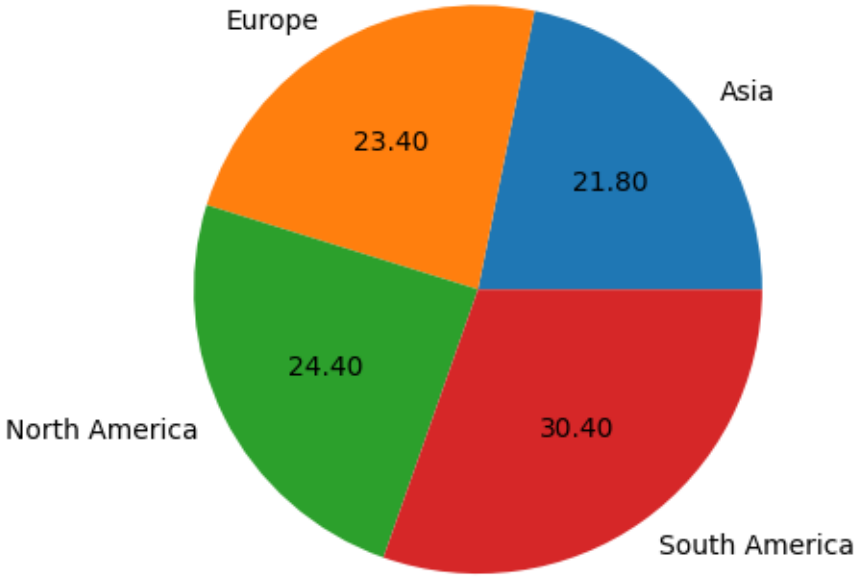
Univariate Analysis

1. Categories of Products



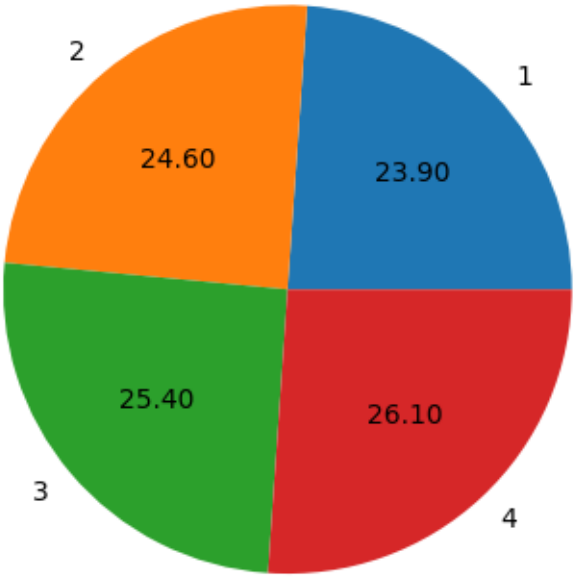
- Inference:
  - It is observed that all categories of items are bought on a similar frequency. Books is moderately sold on a larger number.

2. Categories of Regions



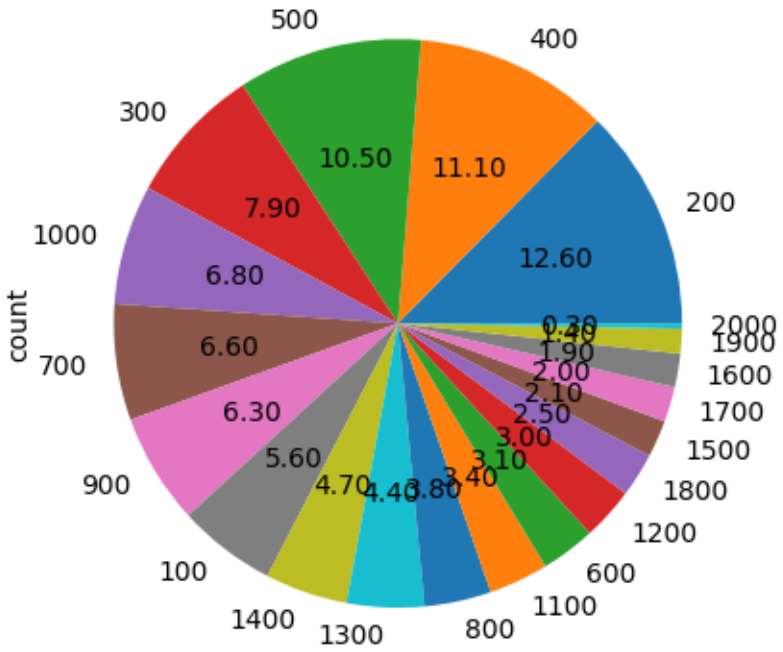
- Inference:
  - It is observed that all categories of regions have a similar frequency of customers. South America has a moderately higher frequency of customers

3. Categories of Quantity



- Inference:
  - It is observed that all categories of quantity have a similar frequency. Buying 4 items together has a frequency that is moderately higher than the rest.

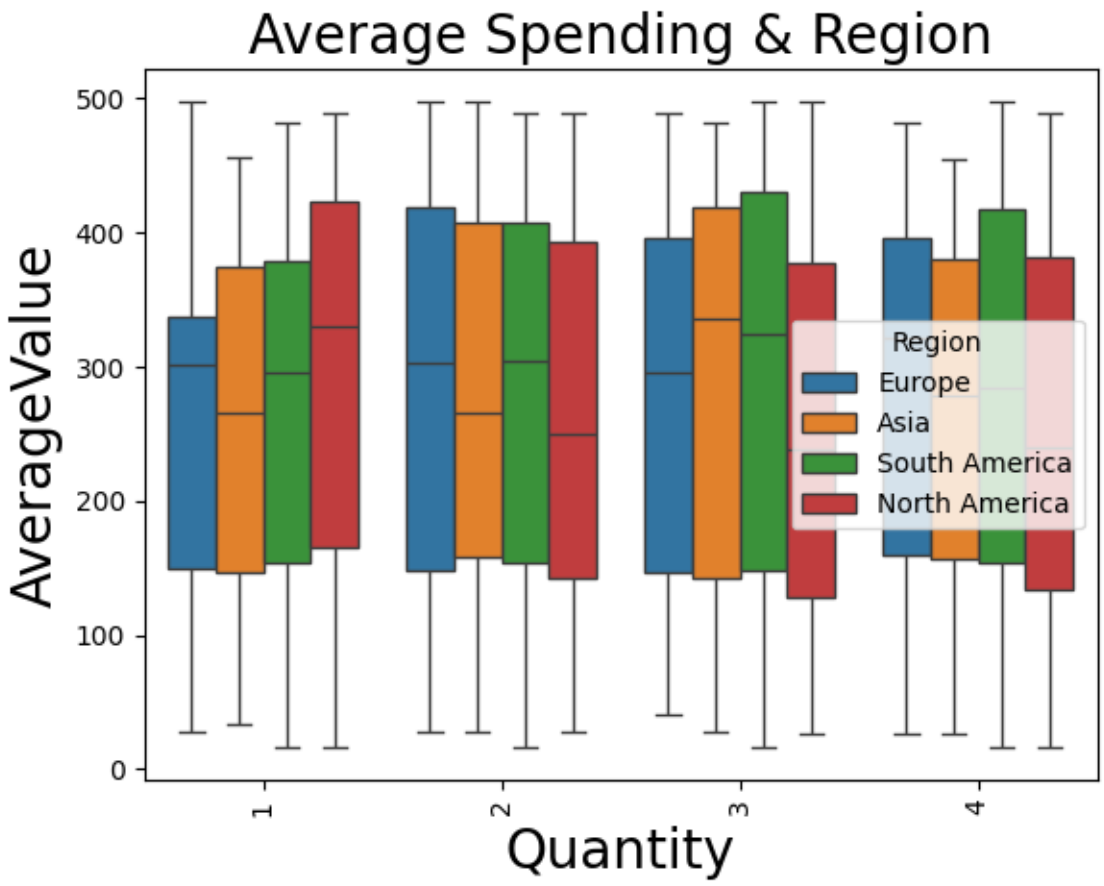
4. Categories of Total Amount Spent



- Inference:
  - It is observed that most transactions are between the range 0 and 300. This shows that most of the customers buy items that add up to this range in a single transaction.

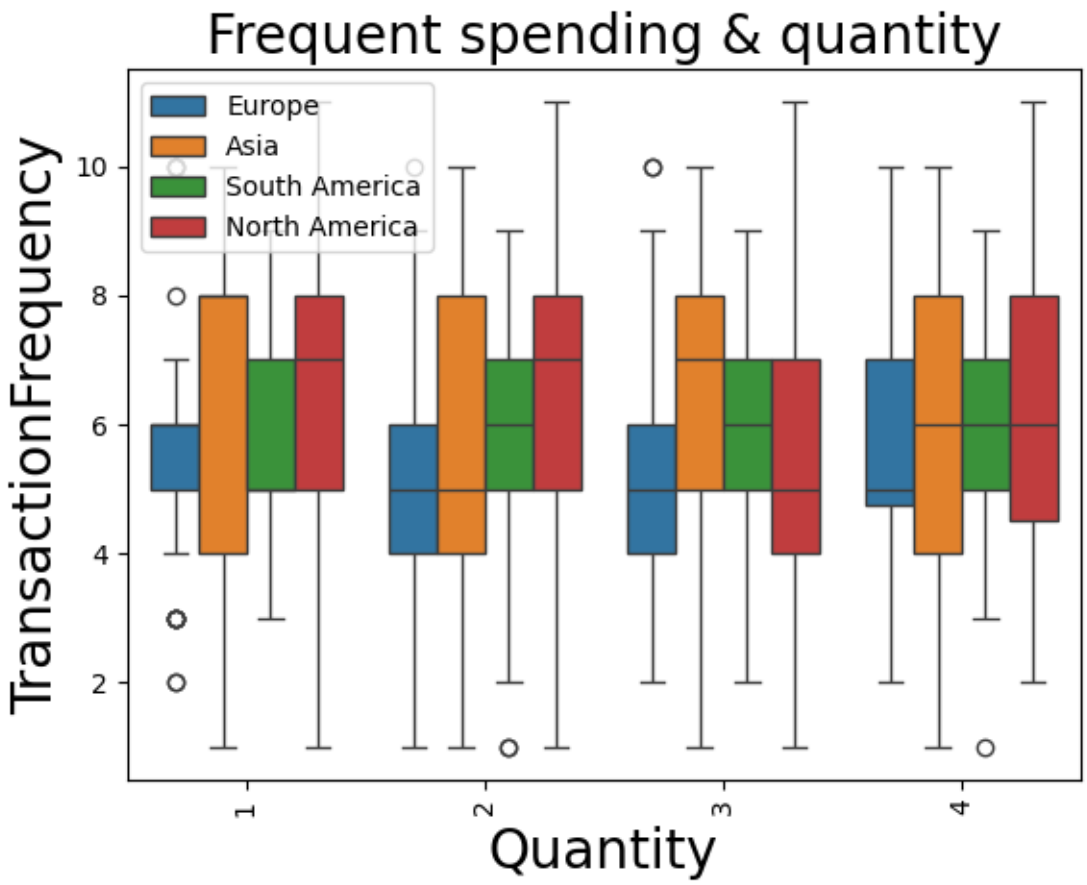
Multivariate Analysis

1. Average Spending vs Quantity in various Regions



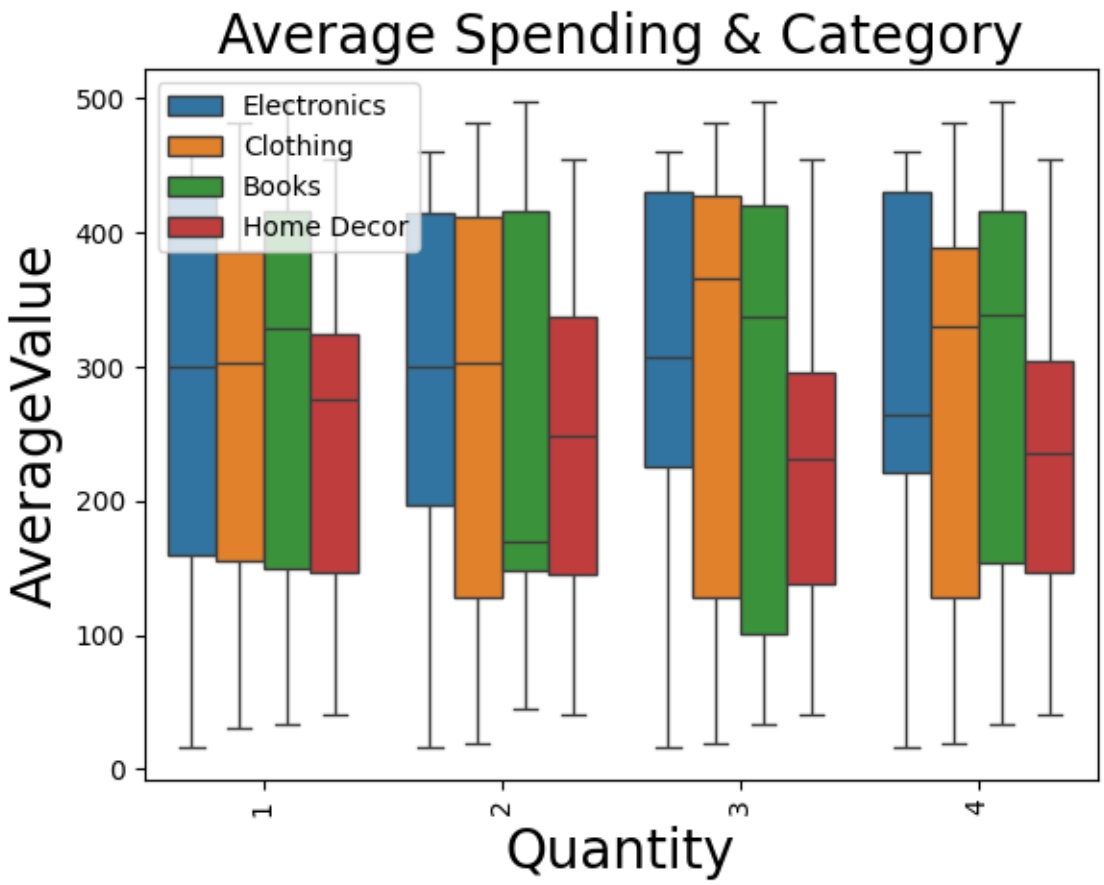
- Inference:
  - The median AverageValue increases slightly as Quantity increases.
  - There is a wide range of spending (large interquartile range), meaning some customers spend significantly more or less even for the same quantity.
  - North America and Asia have higher median spending compared to Europe and South America.
  - South America shows more variability in spending (wider interquartile range).
  - Europe has the lowest median spending, suggesting customers in this region tend to spend less on average.

2. Frequency of Spending vs Quantity for various regions



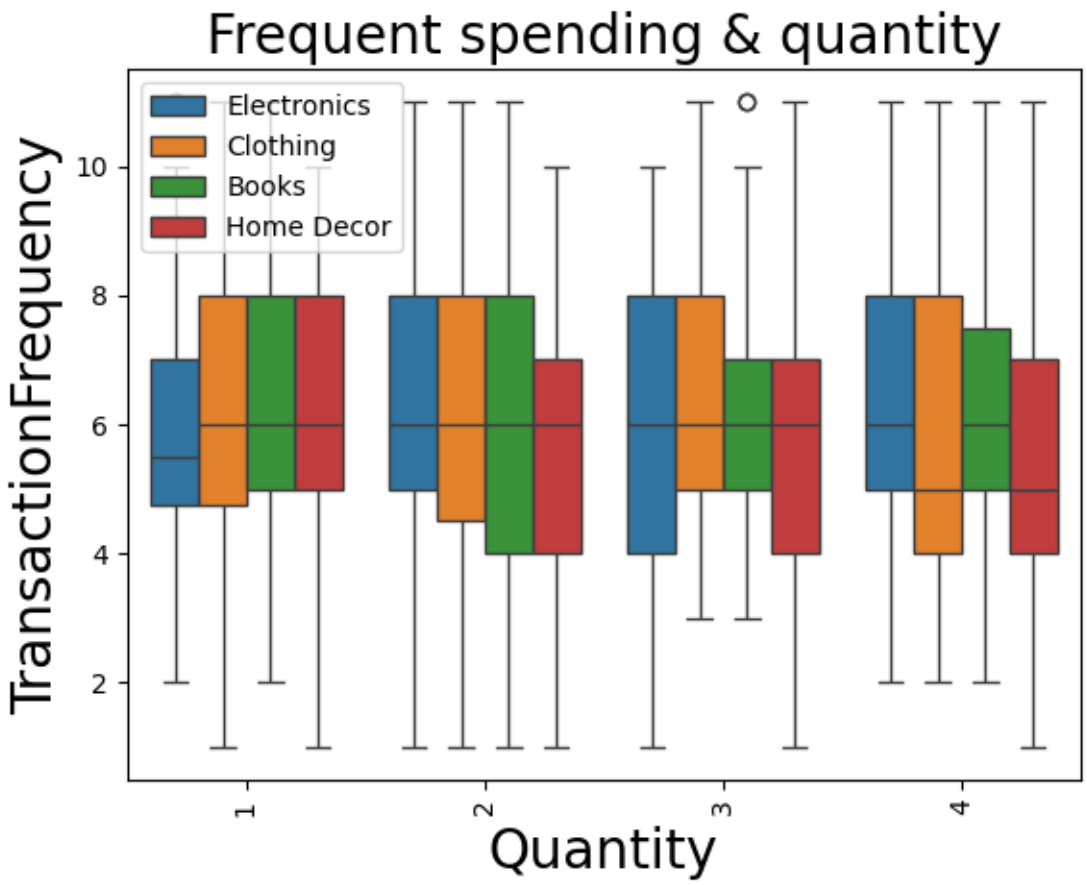
- Inference:
  - Asia and South America generally have higher median transaction frequencies compared to Europe and North America.
  - Europe has the lowest median transaction frequency across all quantities.
  - Asia exhibits a slightly tighter interquartile range (IQR), indicating more consistent transaction frequency.
  - North America and South America have wider IQRs, suggesting greater variability in transaction patterns.
  - Europe shows the widest range and includes several lower outliers, indicating less frequent purchases.

3. Average Spending vs Quantity for various product categorites



- Inference:
  - Electronics and Books have higher median spending compared to Clothing and Home Decor.
  - Home Decor has the lowest median spending, indicating that customers tend to spend less on these items.
  - Spending for Clothing and Books is more evenly distributed..
  - Across all categories, spending increases slightly with quantity, but there is significant variation.
  - Electronics and Books show the highest variability—some customers spend significantly more on a few high-value items.

4. Frequency of Spending vs Quantity for various categories of products



- Inference:
  - Electronics and Books have higher median spending compared to Clothing and Home Decor.
  - For all categories, the median transaction frequency is fairly consistent across different quantities, with some variations.
  - Electronics and Clothing show narrower interquartile ranges (IQR), indicating less variability in transaction frequency.
  - Books and Home Decor have slightly wider IQRs, indicating more variability in their transaction frequency for different quantities.
  - The distributions for Electronics and Clothing are more symmetrical, while Books and Home Decor show some skewness.

Business Insights

- North America and Asia are high-value regions targeted marketing could boost revenue further.
- Europe and South America have lower spending discounts or promotions might encourage higher spending.
- High variability in spending suggests segmentation is needed—some customers buy in bulk, while others make small purchases.
- Electronics and Books are high-value categories—promotions targeting these categories could yield high revenue.
- Home Decor might need pricing adjustments or bundled offers to increase its spending range.
- Discounts on Clothing and Home Decor could encourage higher spending, given their relatively lower median values.
- Bulk purchase incentives for Electronics and Books might attract high-spending customers.
- Ensure consistent inventory levels for Electronics and Clothing to meet customer demand.
- Focus marketing efforts on maintaining loyalty, as these are likely repeat purchases.
- Asia and South America have higher transaction frequencies and consistent customer behavior. Prioritize these regions for product launches or premium offerings.
- Europe has the lowest median transaction frequency and the widest variability. Conduct customer surveys or analyze feedback to identify barriers to frequent transactions.
- North America shows a broad range in transaction frequencies. Segment North American customers based on behavior and target them with personalized offers.