Exploratory Data Analysis

Data Cleaning Process

The data cleaning process involves preparing three datasets—Customers.csv, Products.csv, and Transactions.csv—for analysis by resolving missing values, duplicates, and inconsistencies. Missing values in the Customers dataset (e.g., CustomerName and Region) are replaced with "Unknown," while missing Price values in the Products dataset are filled with the mean to ensure completeness. Rows with missing TotalValue in the Transactions dataset are removed to maintain accurate financial records. Duplicate rows across all datasets are eliminated to retain unique records and avoid redundancy. Date columns such as SignupDate and TransactionDate are converted to datetime format for precise time-based calculations.

The cleaned datasets are validated using the info() method to confirm the absence of missing values and ensure correct data types. Post-cleaning, the datasets are saved as new files (Cleaned_Customers.csv, Cleaned_Products.csv, and Cleaned_Transactions.csv) for further analysis. This systematic cleaning ensures the datasets are consistent, reliable, and ready for tasks such as exploratory data analysis (EDA), clustering, and modeling. By standardizing and verifying the data, this process establishes a solid foundation for generating accurate insights and meaningful outcomes.

Descriptive Statistics:

The descriptive statistics uncover notable trends in the Products, Transactions, and Customers datasets. In the Products dataset, prices range from 16.08to 497.76, with an average price of 267.55andastandarddeviationof 143.22, indicating moderate variability. Of the 100 unique products, Books and Electronics lead the categories with 26 products each, followed closely by Clothing and Home Decor. These trends suggest a stronger emphasis on Books and Electronics, potentially aligning with customer preferences or business priorities.

The Transactions dataset shows a wide range of transaction values, from 16.08to 1,991.04, with an average of $689.99, reflecting diverse purchasing behavior. In the Customers dataset, South America has the largest representation, accounting for 59 out of 200 customers, while Europe, North America, and Asia show relatively balanced participation. Notable activity includes customer C0109, who completed 11 transactions, and product P059, purchased 19 times. These insights can support data-driven strategies such as targeted marketing, customer segmentation, and optimizing inventory for high-demand products and regions.

Data Visualization and Business Insights

1. Distribution of Product Prices

Observation: Product prices range from 16to 500, with a concentration around 300− 400.
Insight: The pricing strategy caters to mid-range customers, with potential opportunities to target budget and premium buyers.

2. Distribution of Quantity Sold
Observation: Quantities sold per transaction are evenly distributed between 1 and 4 items.
Insight: Customers often purchase multiple items, suggesting opportunities for cross-selling or bulk discounts.

3. Box Plots of Product Prices and Transaction Values
Observation: Product prices and transaction values show wide ranges, with transaction values peaking around $1,000.
Insight: High transaction values indicate opportunities to target high-value customers with personalized offers.

4. Customer Distribution by Region
Observation: South America has the largest customer base, followed by Europe, North America, and Asia.
Insight: Focus on South America for marketing while exploring growth opportunities in other regions.

5. Product Distribution by Category
Observation: Books and Electronics dominate product categories, with balanced representation across all categories.
Insight: Expand inventory and promotions in Books and Electronics while diversifying Home Decor and Clothing.

6. Relationship Between Price and Quantity Sold
Observation: Quantities sold remain steady across all price ranges, including high-priced items.
Insight: High-priced products sell in bulk, highlighting the importance of balancing affordability with premium offerings.

Business Insights
Focus marketing efforts in South America and explore growth in other regions.
Cater to mid-range pricing while targeting budget and premium segments.
Promote cross-selling and bulk discounts to maximize transaction sizes.
Expand Books and Electronics offerings and target high-value customers with personalized campaigns.

Bivariate and Multivariate Analysis

1. Correlation Matrix of Transactions Data

Observation: The heatmap shows a strong positive correlation (0.72) between Price and TotalValue.
Observation: Quantity positively correlates with TotalValue (0.61), indicating higher quantities and prices drive transaction values.
Observation: Price and Quantity show no significant correlation (-0.01), suggesting independent behaviors.
Insight: Both price and quantity significantly impact transaction value but are not directly related to each other.

2. Scatter Plot of Price vs Quantity Sold
Observation: Quantity sold remains consistent (1 to 4) across different price ranges.
Insight: Higher-priced products sell in similar quantities as lower-priced products, reflecting steady demand for premium items.

3. Pair Plot of Selected Transaction Metrics
Observation: The pair plot shows linear relationships between Price, Quantity, and TotalValue.
Insight: TotalValue increases proportionally with both Price and Quantity, confirming their combined effect on transaction values.

4. Cross Tabulation between Region and Product Category
Observation: The crosstab reveals distinct product category preferences across regions.
Insight: Regional differences in product preferences can inform targeted inventory and marketing strategies.

Business Implications
Maintain diverse price points since premium products show consistent demand across quantities.
Tailor marketing efforts and inventory based on regional product preferences.
Prioritize products that contribute to high transaction values to maximize revenue potential.

Temporal Analysis - Time Series Trends

1. Daily Sales Trend
Observation: The line plot depicts daily total sales over time, revealing fluctuations in revenue throughout the year.
Insight: Sales activity is highly variable, with frequent spikes indicating peak sales days.

2. Observations
Observation: Regular periods of increased sales are observed, likely during promotions, weekends, or holidays.
Insight: Pinpointing specific dates of sales spikes can assist in optimizing marketing campaigns or inventory planning for high-demand periods.

3. Business Implications
Utilize time-series analysis to forecast future sales trends and ensure preparedness for anticipated high-demand periods.
Implement targeted promotions or campaigns on expected low-sales days to stabilize revenue.

Cohort Analysis Explanation

1. Retention Trends
Observation: The heatmap illustrates customer retention rates across cohorts over time (measured in months after the first purchase).
Insight: All cohorts start with 100% retention in the first month, with a general decline in retention in subsequent months.

2. Key Observations
Observation: Early 2024 cohorts (e.g., March and April) exhibit higher retention rates in later months (e.g., 30%-53%).
Observation: Irregular retention spikes are noted for certain months, such as May and June 2024 (e.g., 80% in June, Period 2), likely driven by successful campaigns or promotions.
Observation: Recent cohorts (e.g., late 2024) show incomplete data but suggest retention drops beyond the first month.

Business Insights

1. Improve Early Retention:

Address retention declines in the first 1-3 months post-signup.
Implement onboarding initiatives like discounts, loyalty programs, or personalized recommendations to improve early customer retention.

2. Analyze Successful Cohorts:

Investigate why certain cohorts (e.g., March and April 2024) perform better in later periods.
Recreate effective strategies, such as promotions or enhanced engagement, for new cohorts.

3. Boost Retention During Drop-Offs:

Counter retention drop-offs by re-engaging inactive customers with tailored offers or campaigns during critical decline periods.

4. Plan Based on Seasonal Trends:

Utilize months with higher retention rates (e.g., May-June 2024) for targeted campaigns and align marketing and inventory strategies accordingly.

5. Address Data Gaps in Recent CohortsL

Monitor retention patterns for recent cohorts (e.g., September-December 2024) to identify trends and formulate proactive retention strategies.

Lookalike Model - Approach and Thought Process

1. Feature Engineering
Monetary: Total spending by each customer to indicate their value to the business.
Frequency: Number of transactions made by each customer to measure engagement levels.
Recency: Days since the last transaction to capture recent activity.

2. Data Standardization
Method: Used StandardScaler to ensure Monetary, Frequency, and Recency are on the same scale.
Purpose: Guarantees that all features contribute equally to similarity calculations.

3. Weighted Features
Weights: Assigned 0.4 each to Monetary and Frequency (high priority) and 0.2 to Recency (lower priority).
Rationale: Aligns with business priorities by prioritizing high-value and frequent buyers.

4. Dimensionality Reduction
PCA: Applied Principal Component Analysis to reduce features to two components while retaining maximum variance.
Benefit: Simplifies similarity computations and improves visualization.

5. Cosine Similarity
Purpose: Measures customer similarity based on behavioral patterns in feature space.
Advantage: Ideal for high-dimensional data as it focuses on direction rather than magnitude.

6. Generating Lookalikes
Method: For each customer, identified the top 3 most similar customers using cosine similarity scores.
Output: Saved results in a CSV file (Lookalike.csv) for further analysis and application.

Business Advantages

1. Targeted Marketing:

Identifies similar customer groups for focused and effective marketing campaigns.

2. Upselling and Cross-Selling:

Recommends products popular with similar customers to increase revenue.

3. Scalability and Customization:

Adjustable feature weights allow the model to adapt to changing business priorities.

Customer Segmentation - Approach and Output

Approach:
1. Feature Engineering
Metrics Extracted: RFM metrics (Recency, Frequency, Monetary) and customer region.
Categorical Handling: Applied one-hot encoding to the 'Region' column to incorporate categorical data into the model.
2. Data Scaling
Method: Used StandardScaler to normalize all features.
Purpose: Ensures equal contribution of features during clustering.
3. Elbow Method
Objective: Determined the optimal number of clusters by identifying the "elbow point" on the inertia plot.
Result: Optimal cluster count was found to be k=4.
4. K-means Clustering
Clustering Algorithm: Implemented K-means to group customers into 4 distinct clusters.
Evaluation Metrics:
Davies-Bouldin Index: 0.4381 (indicates well-separated clusters).
Silhouette Score: 0.7124 (indicates compact and distinct clusters).
5. PCA Visualization
Dimensionality Reduction: Used PCA to reduce feature space to 2 dimensions for visualization.
Outcome: Clear cluster separation shown in a 2D scatter plot.
Output:
Results File: Clustering results saved as Customer_Segments.csv for further analysis.
Behavioral Insights: Clusters reflect distinct behavioral patterns for targeted segmentation and marketing.

Business Insights:
Targeting High-Value Clusters:

Personalize campaigns for high-value clusters to maximize revenue and customer loyalty.
Re-Engaging Low-Value Clusters:

Use promotions, discounts, or onboarding improvements to increase retention and engagement.
Adaptability Through Re-Clustering:

Periodically re-cluster to capture dynamic shifts in customer behavior and refine marketing strategies.