Report on Customer Behaviour Analysis

1. Data Preparation

- Loaded the CSV file from an AWS S3 bucket using Apache Spark.
- Imported all necessary libraries and executed the script to initiate the data processing workflow.

2. Data Cleaning

- Corrected column names and structure.
- Addressed missing values appropriately.
- Detected and managed outliers.
- Performed feature engineering to create new meaningful variables.
- Utilized AWS S3 as the storage location for the CSV data source.

3. Exploratory Data Analysis (EDA)

- Conducted in-depth exploratory analysis to uncover data patterns, relationships, and initial insights.

4. Temporal Purchase Analysis

- Analyzed purchasing behavior across different times of the day, days of the week, and months to identify time-based trends.

5. Customer Demographics vs Purchase Frequency

- Analyzed the relationship between Age, Gender, Income and purchase frequency to identify demographic-driven behavior.

6. Weekdays vs Weekends Purchase Comparison

- Compared customer purchasing behavior on weekdays versus weekends to highlight variations in demand.

7. Commonly Bought Together Products

- Identified products frequently bought together to support cross-selling and recommendation strategies.

8. Product Performance Analysis

- Assessed product sales volume and revenue contribution to identify high and low-performing items.

9. Top Products by Quantity

- Highlighted the products with the highest sales volume.

10. Distribution of Purchases by State

- Analyzed state-wise purchasing patterns to identify high-demand regions.

11. Price vs Product Quantity

- Examined the relationship between product price and purchase quantity.

12. Spending KPIs

- Calculated the average spend per customer as a key performance indicator.

13. Average Spending by Purchase Type

- Analyzed average spending patterns across purchase types (e.g., online vs offline, payment methods).

14. Top 10 High-Engagement Customers

- Identified the most frequent and highest-value customers to improve loyalty programs.

15. Seasonal Trends in Purchases

- Studied seasonal fluctuations in purchasing behavior and their impact on revenues.

16. Customer Location vs Purchasing Behavior

- Explored how geographic factors influenced purchase preferences.

17. Data Preparation for RFM Analysis

- Transformed the RFM dataset by scaling Recency, Frequency, and Monetary values.

18. Elbow Curve Analysis

- Plotted the Elbow Curve to determine the optimal number of clusters for K-Means clustering.

19. K-Means Clustering Model

- Implemented the K-Means algorithm using the optimal number of clusters.

20. RFM Relationships Visualization

- Created pair plots to explore correlations and distributions among RFM features.

21. Customer Segmentation

- Grouped customers into clusters using scaled RFM features to support targeted marketing.

22. Cluster Distribution by Income

- Analyzed income distribution across clusters.

23. Purchase Frequency vs Recency

- Explored the relationship between how often and how recently customers purchased.

24. Top Categories by Cluster

- Identified leading product categories across different customer clusters.

25. Sales by Weekdays

- Analyzed weekday sales performance to identify peak business days.

26. Top 10 Products by Revenue

- Ranked the top revenue-generating products.

27. Revenue by State

- Assessed revenue generation across different states.

28. Repeat Purchase Patterns

- Studied repurchase behavior and customer loyalty trends.

29. Fraud Detection

- Reviewed anomalies in transactions to flag potential fraudulent activity.

30. Demand Fluctuations by Product Category

- Monitored variations in demand across categories for inventory optimization.

31. Bulk Buying Patterns

- Examined how bulk purchasing contributed to revenue and impacted supply chain efficiency.

32. Analytical Queries using MRJob

- Implemented MRJob for distributed data processing:
- • Query 1: Top 10 products by purchase frequency.
- • Query 2: State-wise revenue aggregation.
- • Query 3: Identification of repeat customers with above-average purchase frequency.
- These queries ensured efficient parallel processing and validated insights from PySpark analysis.

33. Conclusion & Recommendations

- This project provided a comprehensive understanding of customer purchasing behavior using PySpark and MRJob.
- Key insights:
- • Revenue varies significantly by season, state, and customer demographics.
- • High-value customers contribute disproportionately to revenue.
- • Fraud detection mechanisms need to be integrated for anomaly monitoring.
- Recommendations:
- • Implement targeted marketing campaigns by cluster and demographics.
- • Focus retention efforts on high-value and repeat customers.
- • Use product bundling strategies for commonly purchased products.
- • Strengthen fraud detection systems and supply chain planning based on demand fluctuations.