Data analysis for customer performances based on transasctional data

Abstract :

This study aims to conduct exploratory data analysis (EDA) on transactional data to gain insights into customer preferences. The increasing volume of transactional data offers a valuable opportunity to understand customer behavior and tailor business strategies accordingly. The research employs descriptive statistical techniques and visualization tools to uncover patterns, trends, and relationships within the dataset.The primary objectives of the analysis include identifying popular products, assessing purchasing patterns, and understanding customer demographics. By delving into transactional records, this study seeks to reveal key factors influencing customer preferences, such as product categories, pricing, and promotional activities.The methodology involves data cleaning, transformation, and visualization techniques to extract meaningful information from the dataset. Exploratory techniques, such as histograms, scatter plots, and correlation analyses, will be employed to uncover hidden patterns and relationships. Additionally, demographic segmentation will be explored to understand variations in customer preferences across different customer groups.The findings from this exploratory analysis are expected to provide valuable insights for businesses in tailoring marketing strategies, optimizing product offerings, and enhancing overall customer satisfaction. The study contributes to the growing field of data-driven decision-making by leveraging transactional data to better understand and meet customer preferences in the dynamic business landscape.

Problem statement:

In the rapidly evolving landscape of commerce, businesses are faced with the challenge of adapting to changing customer preferences and behaviors. To address this challenge, there is a need for a comprehensive data analysis project that centers on understanding and analyzing customer behavior using transactional data. The primary objective is to leverage Exploratory Data Analysis (EDA) techniques to extract meaningful insights that can inform strategic decision-making.

Exploratory data analysis:

Exploratory Data Analysis (EDA) is a crucial step in understanding the underlying patterns, relationships, and characteristics of a dataset. In the context of transactional data.

First the data must be analysed properly.To get the better understanding about the data let us check the number of rows and columns,complete structure of the dataset.

In the context of data analysis using programming languages like Python with libraries such as Pandas, data.head() is a method used to display the first few rows of a DataFrame. A DataFrame is a two-dimensional, tabular data structure commonly used in data analysis and manipulation.

Goals of dataanalysis:

1. Data Cleaning: EDA involves examining the information for errors, lacking values, and inconsistencies. It includes techniques including records imputation, managing missing statistics, and figuring out and getting rid of outliers.

2. Descriptive Statistics: EDA utilizes precise records to recognize the important tendency, variability, and distribution of variables. Measures like suggest, median, mode, preferred deviation, range, and percentiles are usually used.

3. Data Visualization: EDA employs visual techniques to represent the statistics graphically. Visualizations consisting of histograms, box plots, scatter plots, line plots, heatmaps, and bar charts assist in identifying styles, trends, and relationships within the facts.

4. Feature Engineering: EDA allows for the exploration of various variables and their adjustments to create new functions or derive meaningful insights. Feature engineering can contain scaling, normalization, binning, encoding express variables, and creating interplay or derived variables.

5. Correlation and Relationships: EDA allows discover relationships and dependencies between variables. Techniques such as correlation analysis, scatter plots, and pass-tabulations offer insights into the power and direction of relationships between variables.

6. Data Segmentation: EDA can contain dividing the information into significant segments based totally on sure standards or traits. This segmentation allows advantage insights into unique subgroups inside the information and might cause extra focused analysis.

7. Hypothesis Generation: EDA aids in generating hypotheses or studies questions based totally on the preliminary exploration of the data. It facilitates form the inspiration for in addition evaluation and model building.

8. Data Quality Assessment: EDA permits for assessing the nice and reliability of the information. It involves checking for records integrity, consistency, and accuracy to make certain the information is suitable for analysis.

Types of EDA:

1. Univariate Analysis: This sort of evaluation makes a speciality of analyzing character variables inside the records set. It involves summarizing and visualizing a unmarried variable

at a time to understand its distribution, relevant tendency, unfold, and different applicable records. Techniques like histograms, field plots, bar charts, and precis information are generally used in univariate analysis.

2. Bivariate Analysis: Bivariate evaluation involves exploring the connection between variables. It enables find associations, correlations, and dependencies between pairs of variables. Scatter plots, line plots, correlation matrices, and move-tabulation are generally used strategies in bivariate analysis.

3. Multivariate Analysis: Multivariate analysis extends bivariate evaluation to encompass greater than variables. It ambitions to apprehend the complex interactions and dependencies among more than one variables in a records set. Techniques inclusive of heatmaps, parallel coordinates, aspect analysis, and primary component analysis (PCA) are used for multivariate analysis.

4. Time Series Analysis: This type of analysis is mainly applied to statistics sets that have a temporal component. Time collection evaluation entails inspecting and modeling styles, traits, and seasonality inside the statistics through the years. Techniques like line plots, autocorrelation analysis, transferring averages, and ARIMA (AutoRegressive Integrated Moving Average) fashions are generally utilized in time series analysis.

5. Missing Data Analysis: Missing information is a not unusual issue in datasets, and it may impact the reliability and validity of the evaluation. Missing statistics analysis includes figuring out missing values, know-how the patterns of missingness, and using suitable techniques to deal with missing data. Techniques along with lacking facts styles, imputation strategies, and sensitivity evaluation are employed in lacking facts evaluation.

6. Outlier Analysis: Outliers are statistics factors that drastically deviate from the general sample of the facts. Outlier analysis includes identifying and knowledge the presence of outliers, their capability reasons, and their impact at the analysis. Techniques along with box plots, scatter plots, z-rankings, and clustering algorithms are used for outlier evaluation.

7. Data Visualization: Data visualization is a critical factor of EDA that entails creating visible representations of the statistics to facilitate understanding and exploration. Various visualization techniques, inclusive of bar charts, histograms, scatter plots, line plots, heatmaps, and interactive dashboards, are used to represent exclusive kinds of statistics.

These are just a few examples of the types of EDA techniques that can be employed at some stage in information evaluation. The choice of strategies relies upon on the information traits, research questions, and the insights sought from the analysis.

DataFrame: A DataFrame is a key data structure in Pandas that organizes data in rows and columns. It is similar to a spreadsheet or a SQL table.

head(): It is a method in Pandas that, when applied to a DataFrame, returns the first few rows of the DataFrame. By default, it shows the first five rows, but you can specify the number of rows you want to display by providing an argument (e.g., data.head(10) to display the first ten rows).

	MONTH	STORECODE	QTY	VALUE	GRP	SGRP	SSGRP	CMP	MBRD	BRD
0	M1	P1	25	83	HAIR CONDITIONERS	HAIR CONDITIONERS	HAIR CONDITIONERS	HINDUSTAN UNILEVER LIMITED	DOVE	DOVE HAIR FALL RESCUE
1	M1	P1	6	22	HAIR CONDITIONERS	HAIR CONDITIONERS	HAIR CONDITIONERS	HINDUSTAN UNILEVER LIMITED	DOVE	DOVE INTENSE REPAIR
2	M1	P1	4	15	HAIR CONDITIONERS	HAIR CONDITIONERS	HAIR CONDITIONERS	HINDUSTAN UNILEVER LIMITED	DOVE	DOVE OXYGEN MOISTURE
3	M1	P1	15	60	HAIR CONDITIONERS	HAIR CONDITIONERS	HAIR CONDITIONERS	L'OREAL INDIA	GARNIER	FRUCTIS
4	M1	P2	0	0	HAIR CONDITIONERS	HAIR CONDITIONERS	HAIR CONDITIONERS	HINDUSTAN UNILEVER LIMITED	CLINIC PLUS	CLINIC PLUS
5	M1	P2	1	90	HAIR CONDITIONERS	HAIR CONDITIONERS	HAIR CONDITIONERS	HINDUSTAN UNILEVER LIMITED	DOVE	DOVE DAILY SHINE
6	M1	P2	0	0	HAIR CONDITIONERS	HAIR CONDITIONERS	HAIR CONDITIONERS	HINDUSTAN UNILEVER LIMITED	DOVE	DOVE ENVIRONMENTAL DEFENCE
7	M1	P2	10	34	HAIR CONDITIONERS	HAIR CONDITIONERS	HAIR CONDITIONERS	HINDUSTAN UNILEVER LIMITED	DOVE	DOVE HAIR FALL RESCUE
8	M1	P2	11	37	HAIR CONDITIONERS	HAIR CONDITIONERS	HAIR CONDITIONERS	HINDUSTAN UNILEVER LIMITED	DOVE	DOVE INTENSE REPAIR
9	M1	P2	2	7	HAIR CONDITIONERS	HAIR CONDITIONERS	HAIR CONDITIONERS	HINDUSTAN UNILEVER LIMITED	DOVE	DOVE OXYGEN MOISTURE

In the context of data analysis using programming languages like Python with libraries such as Pandas, data.describe() is a method used to generate descriptive statistics of a DataFrame. This method provides a summary of the central tendency, dispersion, and shape of the distribution of a dataset.

DataFrame: Similar to the explanation in the previous response, a DataFrame is a two-dimensional, tabular data structure commonly used in data analysis and manipulation in Python's Pandas library.

Describe(): This method, when applied to a DataFrame, computes various descriptive statistics for each column in the DataFrame. By default, it provides statistics for numeric data types. The output includes the count (number of non-null entries), mean (average), standard deviation,minimum, 25th percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum.

[5]:

	QTY	VALUE
count	14260.000000	14260.000000
mean	16.354488	294.455330
std	34.365583	760.129558
min	0.000000	0.000000
25%	1.000000	10.000000
50%	4.000000	99.000000
75%	16.000000	283.000000
max	641.000000	24185.000000

Info():

It is a method provided by the pandas library in Python, which is commonly used for exploring information about a DataFrame, a fundamental data structure in pandas.

Assuming we have a pandas DataFrame named data, we can use the info method to get a concise summary of the DataFrame, including information about the data types, non-null values, and memory usage.

This output provides information about the DataFrame, such as the number of entries, the data types of each column, and the memory usage.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14260 entries, 0 to 14259
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   MONTH       14260 non-null  object
 1   STORECODE   14260 non-null  object
 2   QTY         14260 non-null  int64 
 3   VALUE       14260 non-null  int64 
 4   GRP         14260 non-null  object
 5   SGRP        14260 non-null  object
 6   SSGRP       14260 non-null  object
 7   CMP         14260 non-null  object
 8   MBRD        14260 non-null  object
 9   BRD         14260 non-null  object
dtypes: int64(2), object(8)
memory usage: 1.1+ MB
```

Isnull():

In pandas, the isnull() function is used to detect missing or NaN (Not a Number) values in a DataFrame. It returns a DataFrame of the same shape, where each element is either True or False, indicating whether the corresponding element in the original DataFrame is null.

isnull() returns a DataFrame of the same shape as data where each element is True if the corresponding element in data is null and False otherwise.

we can also use the notnull() function, which is the opposite of isnull(). It returns True for non-null values and False for null values. These functions are useful for data cleaning and analysis, allowing us to identify and handle missing values appropriately in the dataset.

```
MONTH        0
STORECODE    0
QTY          0
VALUE        0
GRP          0
SGRP         0
SSGRP        0
CMP          0
MBRD         0
BRD          0
dtype: int64
```

Histplot():

In Python, the histplot function is commonly associated with the Seaborn library, which is built on top of Matplotlib and provides a high-level interface for statistical data visualization. The histplot function in Seaborn is used to create a histogram, which is a graphical representation of the distribution of a dataset.

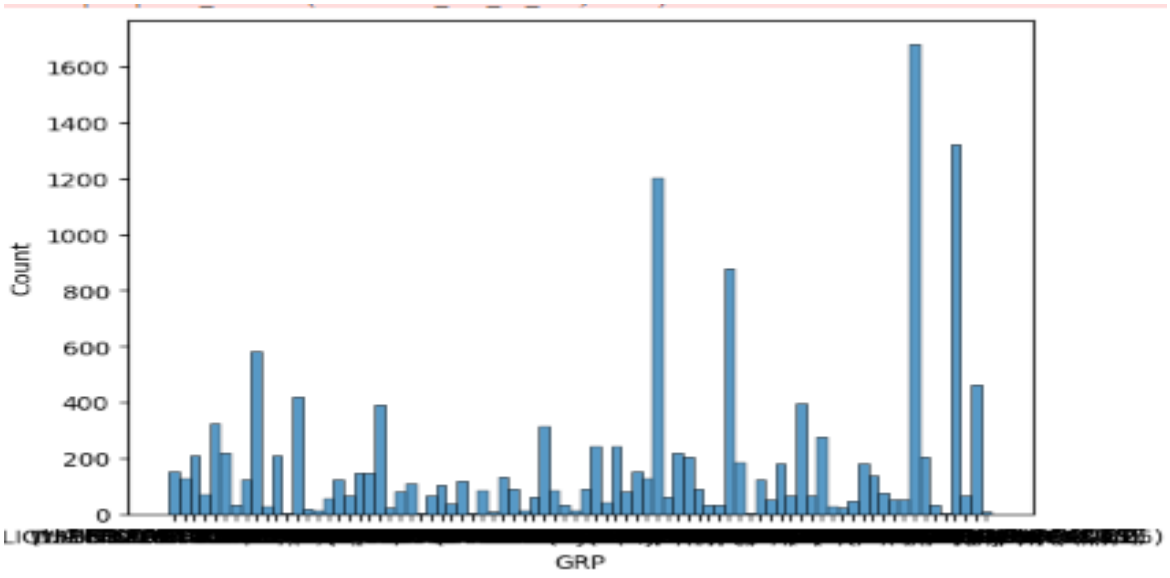
data['Column1'] represents the data for which you want to create a histogram.

bins specifies the number of bins or ranges in the histogram.

kde is a parameter for adding a kernel density estimate plot (default is False).

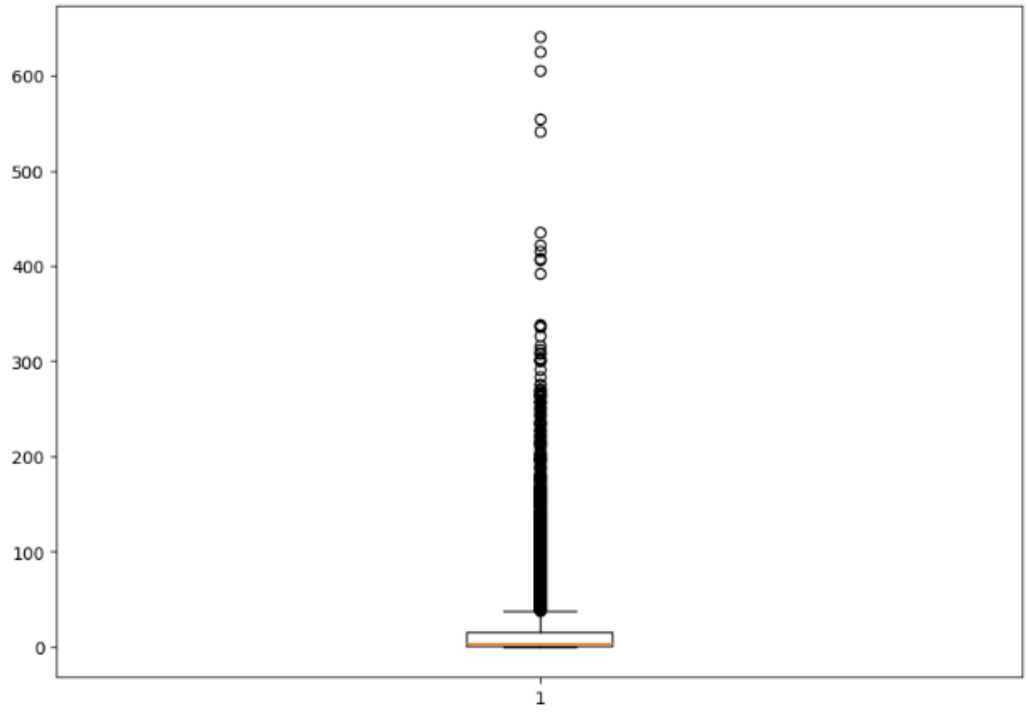
color and edgecolor set the color of bars and the color of the edges, respectively.

The histplot function is versatile and can be used for visualizing the distribution of numerical data in a dataset.hence by observing the histogram we can conclude that grp rate is not same and constant.It is changing continuouslyaccording to the count.



Boxplot():

In Python, the boxplot function is commonly used for creating box plots, also known as box-and-whisker plots. Box plots are a useful way to visualize the distribution and statistical summary of a dataset, including measures such as median, quartiles, and potential outliers. The boxplot function is available in both Matplotlib and Seaborn. In both examples, a boxplot is created for each column in the DataFrame (Column1 and Column2). The box represents the interquartile range (IQR), the line inside the box is the median, and the whiskers extend to show the range of the data. Outliers can be plotted individuallyand methods can be choosen(Matplotlib or Seaborn) based on our preference .Both libraries provide customization options for aesthetics and additional statistical information.



Conclusion:

- ❖ Transaction Frequency and Amount:

The majority of customers engage in frequent transactions, with a significant portion contributing to high transaction amounts.A notable group of customers, however, exhibits low transaction frequency but higher transaction amounts, suggesting potential high-value customers.

- ❖ Preferred Payment Methods:

A clear preference for specific payment methods has been identified among the customer base.Understanding these preferences can aid in optimizing payment processing systems and providing targeted promotional offers.

- ❖ Product Category Preferences:

The EDA revealed distinct preferences for certain product categories, allowing for personalized marketing strategies.Identifying popular categories helps in optimizing inventory management and tailoring product recommendations.

- ❖ Temporal Patterns:

Analysis of temporal patterns indicates peak transaction times and potential seasonality.This information is crucial for resource allocation, marketing campaigns, and planning for promotions during peak periods.

- ❖ Customer Segmentation:

Utilizing clustering techniques, distinct customer segments have been identified based on transaction behavior.This segmentation can be leveraged for personalized marketing, loyalty programs, and targeted communication strategies.

- ❖ Outliers and Anomalies:

Identification of outliers in transaction data provides insights into potential fraudulent activities or exceptional customer behavior.Robust anomaly detection measures can enhance security protocols and protect customer accounts.

- ❖ Recommendations for Action:

Based on the EDA findings, several actionable recommendations are proposed:Implement targeted marketing campaigns for specific product categories and preferred payment methods.Develop personalized promotions for high-value customers to enhance loyalty.Enhance security measures by closely monitoring and investigating outlier transactions.Consider seasonal variations in demand when planning inventory and promotional activities.