Credit Card Spending Habits in India

A report for the Mini Project

Submitted by

Name : MRUDULA A P

Student ID: 230450

AICTE Internship ID: INDASCCE258369

AICTE ID : STU654bbe5c938d81699462748

Data Science Internship Programme

PRAGYASHAL PRIVATE LIMITED

Contents

Dataset Name and Metadata

Dataset Name: **Credit Card Spending Habits in India**

File: **Credit card transactions - India - Simple.csv** [Link]

This dataset contains insights into credit card transactions made in India, offering a comprehensive look at the spending habits of Indians across the nation.The table represents the data dictionary which contains the features and their description.

Column name	Description
City	The city in which the transaction took place. (String)
Date	The date of the transaction. (Date)
Card Type	The type of credit card used for the transaction. (String)
Exp Type	The type of expense associated with the transaction. (String)
Gender	The gender of the cardholder. (String)
Amount	The amount of the transaction. (Number)

Dataset Background

This dataset contains insights into a collection of credit card transactions made in India, offering a comprehensive look at the spending habits of Indians across the nation. From the Gender and Card type used to carry out each transaction, to which city saw the highest amount of spending and even what kind of expenses were made, this dataset paints an overall picture about how money is being spent in India today.

Overview of dataset

	index	City	Date	Card Type	Exp Type	Gender	Amount
0	0	Delhi, India	29-Oct-14	Gold	Bills	F	82475
1	1	Greater Mumbai, India	22-Aug-14	Platinum	Bills	F	32555
2	2	Bengaluru, India	27-Aug-14	Silver	Bills	F	101738
3	3	Greater Mumbai, India	12-Apr-14	Signature	Bills	F	123424
4	4	Bengaluru, India	5-May-15	Gold	Bills	F	171574

Description of Dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26052 entries, 0 to 26051
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   index       26052 non-null  int64 
 1   City        26052 non-null  object
 2   Date        26052 non-null  object
 3   Card Type   26052 non-null  object
 4   Exp Type    26052 non-null  object
 5   Gender      26052 non-null  object
 6   Amount      26052 non-null  int64 
dtypes: int64(2), object(5)
memory usage: 1.4+ MB
```

- The dataset has 26052 rows and 7 features.
- The features are of object and integer datatype.
- There are no missing values in the dataset.

Hypothesis

From the categorical features, we observe that:

- Bengaluru is most frequent in transactions.
- Silver is the most used card Category type.
- Most of the transactions are done in food Category.
- Female Does the most no of transactions.

From the numerical column (Amount), we observe that:

- The amount spent is ranging from 1k to 998k.
    - 25% of transactions amount lies between 1005 to 77k.
    - 50 % of transactions lies between 77k to 228k.
    - 25 % of transaction lies between 228k to 998k.

Notebook link

https://github.com/mrudulamadhavan/infyskill_internship/blob/main/Mini_Project_1.ipynb

Functions used in Project

The various functions used for detailed exploratory analysis using Pandas library is:

- read_csv(): This function is used to retrieve data from CSV files in the form of a dataframe.
- head(): This function is used to return the top n (5 by default) values of a data frame or series.
- info(): This method is used to generate the summary of the dataframe, this will include info about columns with their names, their datatypes, and missing values.
- dtypes(): This method returns a Series with the data type of each column.
- shape(): It returns a tuple representing the dimensionality of the Pandas dataframe.
- describe(): Returns descriptive statistics about the data like mean, minimum, maximum, standard deviation, etc.
- nunique(): Returns the number of unique values in the column
- unique(): It returns all the unique values in a particular column.
- isna(): Returns dataframe/series with bool values. Missing values gets mapped to True and non-missing gets mapped to False.
- value_counts(): Returns the counts of the unique values in a series or from a dataframe's column
- nlargest(): Used to get n largest values from a data frame or a series.
- drop(): Used to drop rows/columns from a dataframe.
- astype(): This method is used to cast pandas object to a specified dtype.
- set_index(): This method is used to set a List, Series or Data frame as an index of a Data Frame.
- reset_index(): This method is used to reset the index of a Data Frame.

Conclusions

The choice of approaches to solve a problem based on data depends on the nature of the problem, the type of data available, and the specific goals of the analysis.

Here are some common approaches and techniques employed in EDA:

1. **Descriptive Statistics:**
    - **Summary Statistics:** Calculate and analyze measures such as mean, median, mode, standard deviation, range, and percentiles to understand the central tendency and dispersion of the data.
    - **Frequency Distribution:** Create histograms, bar charts, or pie charts to visualize the distribution of categorical and numerical variables.
2. **Data Cleaning:**

- **Handling Missing Values:** Identify and address missing data through techniques like imputation or removal.
- **Outlier Detection:** Detect and handle outliers that may skew the analysis. This can involve visualization techniques or statistical methods.

3. **Univariate Analysis:**
    - **Histograms and Box Plots:** Visualize the distribution of individual variables to identify patterns, skewness, or potential outliers.
    - **Kernel Density Plots:** Estimate the probability density function of a variable.
4. **Bivariate Analysis:**
    - **Scatter Plots:** Explore relationships between two variables to identify patterns or correlations.
    - **Correlation Analysis:** Calculate correlation coefficients (e.g., Pearson, Spearman) to quantify the strength and direction of relationships.
5. **Multivariate Analysis:**
    - **Pair Plots:** Visualize relationships between multiple variables simultaneously.
    - **Heatmaps:** Display the correlation matrix to identify patterns and relationships among multiple variables.

Observations from Dataset

- The topmost 3 Cities having highest no of transactions are Bengaluru, Greater Mumbai and Ahmedabad
- The bottommost 3 Cities having lowest no of transactions are Ali Rajpur, Bagaha and Changanassery.
- Silver is the highest used card & Gold is the least used card type.
- Food has the highest no of transactions whereas Travel has the lowest no of transactions.
- Females has the most no of transactions than Males.
- Silver has the highest overall amount while gold is least contribution to total amount.
- Bills has the highest overall amount while travel is least contribution to total amount.
- Female has the highest overall amount than men who contributes to total amount
- Only in Fuel expense subcategory no of transactions dominates for men.
- All subcategories in Card type dominated by women.
- Silver card holder dominates in all sub categories except Travel.
- Gold card holder has most transactions in travel sector
- Food is the most frequent category in all card types & travel being the least

---