Analysis of Customer Transactions

**Milestone 1**

**Instructions before the project execution**

1. Read the entire document before beginning
2. Write down the understanding
3. PLAN properly start to end the path of working on this project
4. Do not use any code generation tools for completion
5. In the project review, questions about any of the above can be asked, including code, logic and the reasoning.
6. Hence it is advised to do the project independently.
7. Document the project execution from start to finish
8. Document the code, the logic, the reasoning and the challenges faced during the project

**Project Overview**

The goal of this project is to give you hands-on experience in applying Python-based data science techniques using libraries like Pandas, NumPy, Matplotlib, and Seaborn.

With a dataset of 100,000 financial transactions, you will explore various aspects of

a. Data manipulation,

b. Cleaning,

c. Visualization with Python, NumPy, pandas, matplotlib, seaborn and plotly to achieve the results.

The dataset is a collection of 100,000 financial transactions, with each transaction represented by a unique identifier, date(time), amount, and category.

The objective is to analyse the data to answer the following questions

1. What is the total amount spent in each category?
2. The Account type most used
3. The type of transactions most made
4. The average amount spent in each account type
5. The total amount spent in each city?
6. What can be predicted?

**Steps taken to accomplish task**

1. Imported data from the CSV file
2. Done Cleaning and preparation
3. Tried changing data types and added new columns as required
4. Grouped and aggregated data
5. Descriptive Statistics
6. Customer Transaction Frequency
7. City-wise Transaction Amount
8. Transaction Amount by Account Type
9. Visualized the data using excel charts used in presentation
10. Tried visualizing data in matplotlib, seaborn and plotly
11. Used various types of charts
12. Correlation Analysis by Transaction Amount & Customer ID
13. As the dates were not in correct format tried generating random dates,

Considering the data is for last two years[2022,2023] as mentioned.

**Graphs to consider**

1. ***Bar and Horizontal Bar Plots:***

Good for general comparison and when categories (customer IDs) are numerous.

1. ***Pie Chart:***

Effective for showing proportions but can become cluttered with many categories.

1. ***Line Plot:***

Useful for ordered or time-series data.

1. ***Scatter Plot:***

Good for spotting outliers or patterns in individual customer behaviours.

1. ***Box and Violin Plots:***

Ideal for understanding the distribution of transaction amounts and identifying outliers.

1. ***Histogram:***

Helps understand the frequency of various transaction amounts.

1. ***Stacked Bar Chart:***

Useful for comparing multiple variables per customer.

**Some predicting possibilities**

1. **Customer Spending Patterns:**
   1. Analyse transaction amounts by account type to identify trends in spending.
   2. For instance, you could determine if customers are spending more from their checking accounts versus savings.
2. **Account Health Assessment:**
   1. Predict the likelihood of customers overdrawing their checking accounts based on their transaction history and patterns in spending.
3. **Segmentation of Customers:**
   1. Group customers by their transaction behaviours or account types to identify distinct segments (e.g., high spenders, savers, or those relying heavily on credit).
4. **Credit Utilization:**
   1. Analyse credit usage over time to predict future borrowing behaviours.
   2. You could determine the average credit utilization rate and identify customers at risk of exceeding their limits.
5. **Geographic Insights:**
   1. Analyse how transaction behaviours vary by city.
   2. This can reveal regional trends or economic conditions affecting spending and saving.

**Observations**

From the bar graph showing the average transaction amounts by account type, you can make several predictions and insights:

1. **Customer Behaviour:**

Checking accounts have the highest average transaction amounts, suggesting that customers use these accounts more frequently for larger transactions.

This could indicate that checking accounts are the primary accounts for daily expenses and bill payments.

1. **Account Usage:**

Savings accounts have the lowest average transaction amounts, which might imply that these accounts are used primarily for storing money rather than frequent transactions.

1. **Marketing Strategies:**

Financial institutions could focus marketing efforts on promoting checking accounts for everyday use and savings accounts for long-term financial goals.

1. **Product Development:**

Banks might consider developing new features or products tailored to the usage patterns of each account type. For example, offering rewards or incentives for higher transaction volumes in checking accounts.

1. **Risk Management:**

Understanding transaction patterns can help in identifying potential risks, such as overdrafts in checking accounts or low engagement in savings accounts.

Q. Why the average transaction amount for the Checking account might be negative.

This could happen for a few reasons:

1. **Overdrafts:**

If the account frequently goes into overdraft, the average transaction amount could be negative.

1. **Fees:**

High fees associated with the account could result in a net negative average transaction amount.

1. **Refunds or Reversals:**

If there are more refunds or transaction reversals than deposits, this could also lead to a negative average.

**Observations from Charts visualized from Excel & Using Python and libraries:**

**Descriptive Statistics**

This chart represents descriptive statistics for transaction amounts, such as count, mean, standard deviation, min, max, and various percentiles.

Insights:

1. The count of transactions is significantly larger than other statistical measures, indicating a high volume of data.
2. The mean, standard deviation, and other statistics show relatively small values compared to the total count.
3. There is a stark contrast between the max value and lower percentiles, possibly indicating the presence of outliers or large transactions.

**Total Transaction Amount by Customer ID**

This chart shows the transaction amounts for different customer IDs.

Insights:

1. There is a wide fluctuation in transaction amounts across customer IDs, ranging from large positive to large negative values.
2. It suggests a high variability, implying that different customers have vastly different transaction behaviours.
3. Some IDs have significantly high transaction amounts, while others show negative values, possibly indicating refunds or chargebacks.

**Total Transaction Amount by City**

This chart represents transaction amounts categorized by city.

Insights:

1. Los Angeles and New York show the highest transaction amounts, with Los Angeles leading slightly.
2. Other cities like Phoenix and Chicago display much lower amounts, and Houston shows negative amounts, potentially indicating losses or refunds.
3. The colour gradient highlights transaction amounts, with higher values in yellow and lower values in darker shades.

**Overall Analysis:**

1. The data contains large volumes of transactions with high variability in transaction amounts, both across individual customers and cities.
2. Certain cities and customers contribute significantly to the total transaction amounts, while others may involve more refunds or lower levels of activity.
3. The presence of negative transaction amounts in some cities and customer data suggests possible financial returns or chargebacks that may require further investigation.
4. This analysis can help in understanding the overall financial activity and the impact of different customers and locations on the transaction metrics.

**References Used:**

For pandas

1. Previous Codes
2. [User Guide — pandas 2.2.3 documentation (pydata.org)](https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html)

For matplotlib

1. Previous Codes
2. [API Reference — Matplotlib 3.9.2 documentation](https://matplotlib.org/stable/api/)

For seaborn

1. Previous Codes
2. [Example gallery — seaborn 0.13.2 documentation (pydata.org)](https://seaborn.pydata.org/examples/index.html)

For plotly

1. [Plotly Python Graphing Library](https://plotly.com/python/)

For trendline in excel

1. [Choosing the best trendline for your data - Microsoft Support](https://support.microsoft.com/en-us/office/choosing-the-best-trendline-for-your-data-1bb3c9e7-0280-45b5-9ab0-d0c93161daa8)

Project review & guidance:

Kiran VVN