Synopsis

Title : Credit Card Transactions Data

1. Overview

The dataset represents a collection of synthetic credit card transaction data, containing 10,000 records. Each transaction includes various attributes such as card type, transaction amount, merchant category, transaction type, location, and a flag indicating whether the transaction is fraudulent. The data was generated to simulate real-world credit card transactions and can be used for various types of data analysis, including fraud detection, spending patterns, and merchant behavior analysis.

TOOLS:

1. Data Wrangling and Analysis:

- **Pandas** (Python) – For cleaning, manipulating, and analyzing the dataset.
- **NumPy** (Python) – For numerical operations and statistical calculations.

2. Data Visualization:

- **Matplotlib** (Python) – For creating basic plots like histograms, bar charts, and line charts.
- **Seaborn** (Python) – For advanced statistical visualizations like heatmaps and pairplots.
- **Plotly** (Python) – For interactive visualizations.

3. Time-Series Analysis:

- **Statsmodels** (Python) – For time-series analysis, including trend detection and forecasting.
- **Prophet** (Python/R) – For easy-to-use time-series forecasting.

4. Geospatial Analysis:

- **Geopandas** (Python) – For spatial data analysis and map-based visualizations.
- **Folium** (Python) – For creating interactive maps.

5. Statistical Analysis:

- **SciPy** (Python) – For performing statistical tests and hypothesis testing.

6. Database and Storage:

- **SQL** – For querying and managing the dataset in databases.
- **SQLite** (Python) – For local, file-based database management.

7. Reporting & Dashboarding:

- **Jupyter Notebooks** (Python) – For interactive data analysis and reporting.
- **Tableau/Power BI** – For creating visual dashboards and reports.

Objective:

The primary objective of this project is to analyze a synthetic dataset of 10,000 credit card transactions in order to uncover insights into consumer spending behavior, merchant activity, and fraudulent transaction detection. Specifically, the goals are:

1. **Fraud Detection**:
   - Develop and evaluate methods for identifying fraudulent transactions based on transaction attributes such as card type, amount, merchant category, and location.
   - Identify patterns or anomalies in fraudulent transactions to improve detection accuracy.

2. **Spending Pattern Analysis**:
   - Investigate consumer spending behavior across various merchant categories, card types, and locations.
   - Analyze average transaction amounts, trends in high-value purchases, and the frequency of purchases versus refunds.

3. **Merchant Behavior Analysis**:
   - Explore spending patterns within different merchant categories to identify consumer preferences, such as the dominance of Electronics, Groceries, and Restaurants.
   - Assess the performance and distribution of transactions across various merchant categories and locations.

4. **Card Type Usage**:
   - Examine the relationship between card types (Visa, MasterCard, Amex) and transaction characteristics, including transaction amounts, merchant categories, and fraud incidence.

5. **Location-Based Insights**:
   - Conduct an analysis of spending behavior across different geographic locations to identify regional trends in consumer purchases and merchant activity.

6. **Time-Series Trends**:
   - Track transaction data over time to uncover patterns related to peak spending periods, seasonal trends, or temporal increases in fraudulent activity.

Through these analyses, the project aims to provide actionable insights for improving fraud detection systems, optimizing merchant strategies, and enhancing customer experience based on spending behavior trends.

2. Data Structure

- **Transaction ID**: A unique identifier for each transaction.
- **Card Type**: The type of credit card used for the transaction (e.g., Visa, MasterCard, Amex).
- **Transaction Date**: The date and time of the transaction.
- **Merchant Category**: The type of merchant where the transaction took place (e.g., Electronics, Restaurant, Groceries).
- **Amount**: The amount of money spent in the transaction (in USD).
- **Transaction Type**: Whether the transaction was a "Purchase" or a "Refund".
- **Location**: The city where the transaction took place.
- **Fraudulent Flag**: A flag indicating whether the transaction was fraudulent ("Yes" or "No").
- **Merchant ID**: A unique identifier for the merchant involved in the transaction.

3. Descriptive Statistics

- **Transaction Amounts**:
  - Range: $10 to $500 (randomly generated within this range).
  - Average Transaction Amount: Approximately **$200**.
  - Standard Deviation: Varies depending on the distribution, but we expect higher variability due to random fluctuations in the amount.

- **Transaction Types**:
  - **Purchase**: 98% of the transactions are purchases.
  - **Refund**: 2% of transactions are refunds, randomly distributed across different card types and merchant categories.

- **Card Types**:
  - **Visa**: 34% of transactions.
  - **MasterCard**: 33% of transactions.
  - **Amex**: 33% of transactions.

- **Merchant Categories**:
  - **Electronics**: 20% of the transactions.
  - **Restaurant**: 20% of the transactions.
  - **Groceries**: 20% of the transactions.
  - **Clothing**: 10% of the transactions.
  - **Toys**: 10% of the transactions.
  - **Books**: 10% of the transactions.

- **Fraudulent Transactions**:
  - **Fraudulent**: 2% of the transactions are flagged as fraudulent, representing approximately 200 transactions.
  - **Non-Fraudulent**: 98% of the transactions are not fraudulent.

- **Location Distribution**:
  - The transactions are evenly distributed across several U.S. cities, including major locations like **New York**, **Los Angeles**, **Chicago**, **San Francisco**, and **Austin**.
  - No location shows a significant skew, as it is randomly chosen from a fixed list of cities.

4. Key Insights

- **Card Usage**: All three card types (Visa, MasterCard, and Amex) are used roughly equally in the dataset. This reflects a balanced distribution and could help in analyzing trends by card type.

- **Merchant Behavior**:
    - **Electronics, Restaurants, and Groceries** dominate merchant categories, suggesting these are common transaction types in consumer spending.
    - The other categories like **Clothing**, **Toys**, and **Books** represent smaller, niche markets.

- **Transaction Amount Analysis**:
    - The average transaction value is around $200, with some high-value purchases (close to $500), likely representing more significant purchases in categories like Electronics.
    - Refund transactions have a lower average value than regular purchases, as expected in refund scenarios.

- **Fraudulent Transactions**:
    - Approximately 2% of the transactions are flagged as fraudulent, which is a relatively low rate but still significant enough to warrant attention.
    - Fraudulent transactions could be analyzed for patterns such as high-value amounts, specific merchant categories, or particular card types.

5. Possible Areas for Further Analysis

- **Fraud Detection**:
    - Identify trends in fraudulent transactions (e.g., high transaction amounts, specific locations, certain card types, or merchant categories). Fraudulent transactions may exhibit patterns that differ from legitimate transactions, and machine learning models could be trained to predict fraudulent transactions based on these features.

- **Spending Patterns**:
    - Investigate spending behavior across different **merchant categories** or **locations** to identify trends in consumer habits (e.g., higher spending in certain categories or cities).

- **Card Type Behavior**:
  - Analyze whether specific card types are more likely to be used for certain merchant categories or transaction amounts. For example, are **Amex** cards typically used for high-value purchases at Electronics merchants?

- **Location-Based Analysis**:
  - Explore spending trends based on **location**. Are certain cities seeing higher spending in particular categories? This could help identify regional preferences and behaviors.

- **Time-Series Analysis**:
  - Track transactions over time to identify patterns such as peak transaction times, seasonal spikes in spending, or increases in fraudulent activity during certain periods.

6. Conclusion

This synthetic credit card transaction dataset provides a diverse set of features for analysis. While the data is entirely fictional, it mirrors real-world transaction behavior, making it useful for practice with data analysis and machine learning techniques. Whether the goal is fraud detection, consumer spending analysis, or merchant performance, this dataset provides a rich source of information for building predictive models or identifying trends.