Financial Fraud Detection: Anomaly Detection in Credit Card Transactions

Business Introduction:

Welcome to SecureGuard Financial Solutions, a leading provider of innovative fraud detection analysis for the financial industry. At SecureGuard, we specialize in developing cutting-edge solutions to safeguard our clients' assets and protect against fraudulent activities in real-time.

Our mission at SecureGuard is to provide comprehensive fraud detection services that empower financial institutions to maintain the trust and confidence of their customers while minimizing financial losses due to fraudulent transactions.

With the increasing prevalence of online transactions and digital payments, the risk of fraudulent activity has become a significant concern for financial institutions worldwide. From unauthorized credit card transactions to identity theft, the threat of fraud poses a substantial risk to both financial institutions and their customers.

That's where SecureGuard comes in. By analyzing patterns and anomalies in transaction data, our system can swiftly flag potentially fraudulent transactions, enabling financial institutions to take immediate action to mitigate the risk.

Problem Statement:

Design a fraud detection system to identify fraudulent credit card transactions in real-time. The system should detect suspicious activities such as unauthorized transactions, unusual spending patterns, and fraudulent card usage to prevent financial losses for both cardholders and financial institutions.

Excel Tasks:

1. **Data Exploration:**
    - Create a statistical summary on amt and city_pop
    - Plot a histogram on amt
    - Create a report showing number of Frauds by gender and category
    - Show the top 3 states by the highest number of transactions

2. **Data Analysis in Excel:**
    - Is there a correlation between transaction amount and city population?
    - How many fraudulent transactions occurred in each category?
    - How does the average transaction amount vary between different job roles?

SQL Tasks:

3. **Data Loading:**

- Create a schema named **finance**, set **finance** as the default schema, and create tables with cc_data.csv and location_data.csv

4. **Data Exploration with SQL:**

    - Calculate the total number of transactions in the cc_data table
    - Identify the top 10 most frequent merchants in the cc_data table
    - Find the average transaction amount for each category of transactions in the cc_data table
    - Determine the number of fraudulent transactions and the percentage of total transactions that they represent
    - Join the cc_data and location_data tables to identify the latitude and longitude of each transaction
    - Identify the city with the highest population in the location_data table
    - Find the earliest and latest transaction dates in the cc_data table

5. **Using Data Aggregation with SQL:**

    - What is the total amount spent across all transactions in the cc_data table?
    - How many transactions occurred in each category in the cc_data table?
    - What is the average transaction amount for each gender in the cc_data table?
    - Which day of the week has the highest average transaction amount in the cc_data table?

Python Tasks (including EDA):

6. **Exploratory Data Analysis (EDA) with Python:**

    - What are the dimensions (number of rows and columns) of the dataset?
    - How many unique values are there in each categorical variable?
    - What is the distribution of numerical variables in the dataset?
    - Are there any missing values in the dataset? If so, how should they be handled?
    - What are the summary statistics (mean, median, min, max, etc.) for numerical variables?
    - Is there any correlation between numerical variables? If so, how strong is the correlation?
    - How does the distribution of an amt differ across is_fraud categories?
    - Are there any outliers in the city_pop and amt?
    - Are there any trends or patterns in the data over time (if applicable)?
    - How does the target variable (if available) distribute across different categories?

- Are there any unusual or unexpected values in the dataset that require further investigation?
- Are there any potential data entry errors or inconsistencies in the dataset?
- How does the distribution of numerical variables vary between different groups or segments of the dataset?
- What are the top factors that influence the target variable, if applicable?
- Write an analysis report on performing exploratory data analysis (EDA) using Python in the context of building a fraud detection system for the financial industry.

PowerBI Tasks:

7. **Interactive Dashboard Design in PowerBI:**
    - Show the amount spent by different genders on different categories through a box and whisker plot.
        - Open Power BI Desktop and click on **"Get data"** > **"Text/CSV",** then upload the **cc_data.csv** file.
        - Click on the ellipsis (...) in the Visualizations pane and select "**Get more visuals**".
        - Search for "**Box and Whisker Chart by MAQ Software**" and click **"Add**".
        - Click on the added Box and Whisker visual to insert it into the canvas.
        - Drag **Amt** to the **"Values"** field well.
        - Drag **Category** to the **"Axis"** field well.
        - Drag **Gender** to the "**Legend**" field well.
        - Turn on **Data Labels** from the **Format pane** for better readability.

    - Show the geographical distribution of transactions using latitude and longitude coordinates
        - Click on the **Map visual** from the Visualizations pane.
        - In **Data view**, set the data category of **Lat** to "**Latitude**" and **Long** to "**Longitude**".
        - Drag **Lat** to **Latitude** and **Long** to **Longitude** in the map fields.
        - Drag **Amt** to Size and Color saturation or **Tooltips** for better visual analysis.
        - Adjust the bubble size and color gradient from the Format pane.

    - Use Power BI's map to visualize the locations of fraudulent transactions.
        - Insert another Map visual from the Visualizations pane.
        - Ensure **Lat** is categorized as **Latitude** and **Long** as **Longitude**.
        - Drag **Lat** to **Latitude** and **Long** to **Longitude**.
        - Drag **Is Fraud** to **Legend** or **Color**.
        - Optionally, add a filter to show only fraudulent transactions by setting **Is Fraud = 1**

- Create a time series analysis line chart showing the trend of monthly transaction amounts transactions.
    - Click on the **Line Chart visual** in the Visualizations pane.
    - Drag **trans_date_trans_time** to the **Axis** field.
    - Right-click on the Axis and select "**Date Hierarchy**" > keep only **"Month".**
    - Drag **Amt** to **Values** and set its aggregation to "**Count**".
    - Turn on Data Labels from the Format pane

- Create a calculated field for inflation-adjusted transaction amounts.
    - Click on "**Modeling**" tab and select "**New column**".
    - Name the column **Inflation_Adj_Amt** and use the formula**:** **Inflation_Adj_Amt = [Amt] * (1 + 0.02) ^ (YEAR(TODAY()) - YEAR([trans_date_trans_time]))**
    - Create another column to extract week number **WeekNum = WEEKNUM([Trans Date Trans Time])**
    - Insert a **Line Chart** and drag **WeekNum** to Axis.
    - Drag **Inflation_Adj_Amt** to **Values** and enable Data Labels.

- Create a dashboard with all the visualizations.
    - Click the "**+**" icon to add a **new report page**
    - **Copy** and paste the visualization on the new report page.
    - Right click on the Visualization then click on copy the visual.
    - Then paste it on the new report page.
    - Add the following visuals to the canvas: **Box and Whisker, Transaction Map, Fraud Map, Monthly Trend, and Inflation Line Chart**.
    - Resize and arrange visuals for clarity and balance.
    - Add Slicer visuals for filters like Gender, Category, and Is Fraud. Add Text boxes to label and organize your dashboard sections.