Comprehensive Credit Card Transaction Analysis with Apache Spark and Pandas

This Python script leverages the power of Apache Spark and Pandas libraries to conduct a thorough analysis of credit card transaction data stored in JSON format. It employs a multi-step approach that involves data cleaning, transformation, and visualization to extract valuable customer insights and uncover potential fraudulent activities.

Key Steps in the Analysis Pipeline:

1. Spark Session Initialization:
    - Establishes a Spark session named "JSONImport" in local mode, utilizing all available CPU cores for efficient data processing.
2. Data Loading:
    - Reads the credit card transaction data from the specified JSON file path (/FileStore/tables/cc_sample_transaction.json).
3. Nested JSON Flattening:
    - Handles nested JSON structures within the "personal_detail" and "address" fields by defining corresponding schemas. This process effectively extracts all relevant information from these nested structures into a single, flattened DataFrame for easier analysis.
4. Data Cleaning:
    - Implements meticulous data cleaning procedures to ensure data quality for analysis. Here's a breakdown of the cleaning steps:
        - Personal Name Cleaning: Removes unwanted special characters, replaces commas with spaces, and consolidates multiple spaces into single spaces within the "person_name" column.
        - Timestamp Conversion: Converts timestamp columns containing transaction times to a human-readable format in UTC+8 for clarity and easier interpretation.
5. Feature Engineering:
    - Creates additional features to facilitate analysis:
        - Name Splitting: Splits the "person_name" column into "first" and "last" name components using regular expressions, addressing potential missing last names.
        - PII Masking: Masks sensitive data like credit card numbers and full names using SHA2 hashing to protect privacy in accordance with data privacy regulations.

- Date and Address Redaction: Redacts birth dates and street addresses using regular expressions to maintain anonymity while preserving relevant data for analysis.
- Column Selection: Selects a targeted subset of columns relevant for the analysis, improving efficiency and focusing on key insights.
- Data Type Conversion: Converts the "amt" and "is_fraud" columns to numerical data types (float and int) for enhanced analytical capabilities (e.g., enabling calculations and aggregations).

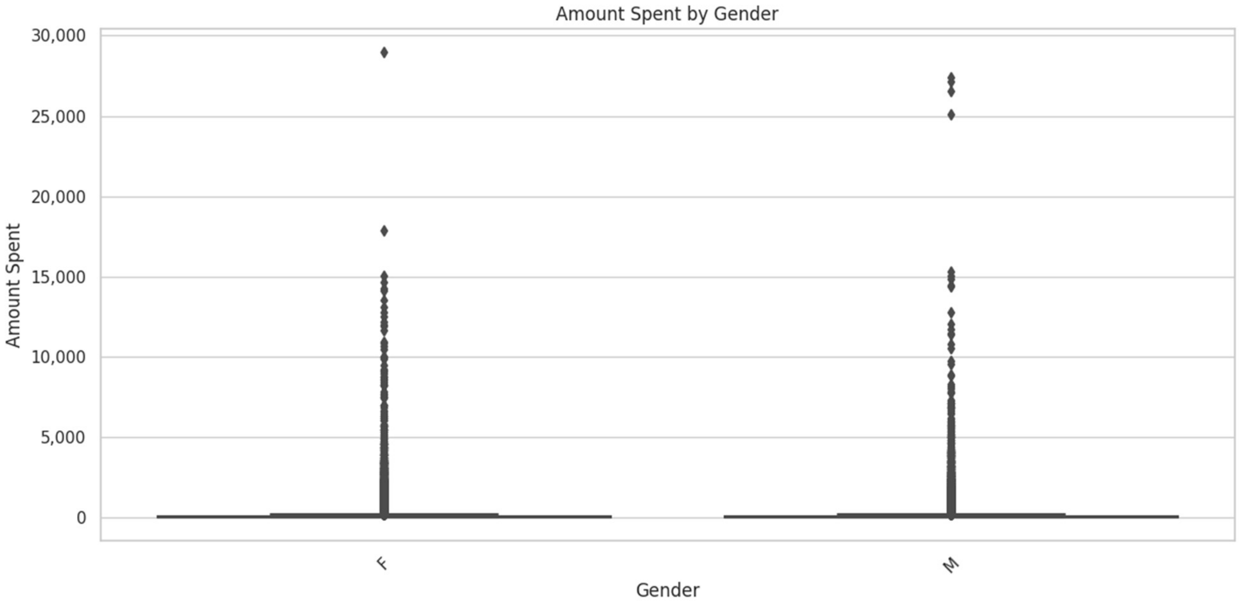
6. Final DataFrame:
    - Displays the final, cleaned, and transformed DataFrame along with its schema, providing a clear overview of the prepared data ready for analysis.
7. Pandas Conversion:
    - Converts the Spark DataFrame into a Pandas DataFrame for seamless integration with data visualization libraries like seaborn and matplotlib.
8. Data Visualization:
    - Generates various informative visualizations using seaborn and matplotlib to explore the data and uncover patterns:
        - Boxplot (Gender vs. Spending): Analyzes the distribution of transaction amounts across genders, potentially revealing differences in spending behavior between genders.
        - Barplot (Total Spending by Gender): Compares the total amount spent by each gender for a holistic view of spending patterns.
        - Hourly Transaction Distribution: Visualizes the number of transactions and their corresponding total amounts throughout the day, potentially identifying peak spending hours and informing staffing or marketing strategies.
        - Transaction Volume by State: Illustrates the geographical distribution of transactions across different states, providing insights into regional spending patterns.
        - Top Merchant Categories: Examines the top 10 merchant categories by transaction count and total transaction amount, providing insights into customer preferences and potentially informing targeted marketing campaigns.
        - Fraudulent vs. Non-Fraudulent Transactions: Creates a table summarizing the count and percentage of fraudulent transactions, accompanied by an image representation for clear communication. This can be crucial for flagging suspicious activities and implementing fraud detection strategies.

        - Hourly Distribution of Fraudulent Transactions: If fraudulent transactions exist, this visualization depicts their distribution by hour of the day, potentially aiding in identifying temporal patterns of fraudulent activity and informing fraud prevention measures.

9. Spark Session Termination:
    - Defines a function (clean_spark_stop) to properly terminate the Spark session, ensuring efficient resource management. The function incorporates error handling and a force-stop mechanism for Windows systems.

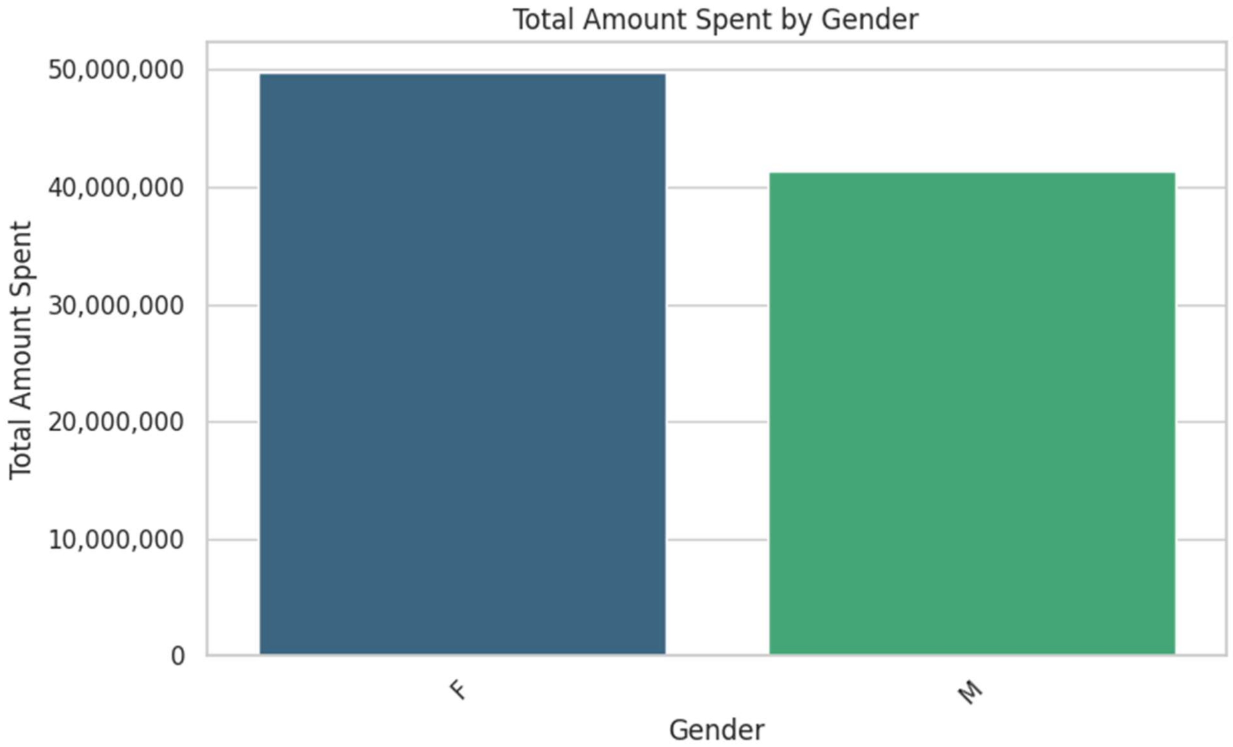
In conclusion, this script offers a comprehensive framework for analyzing credit card transaction data using Apache Spark and Pandas. The combination of data cleaning, transformation, and visualization techniques empowers data analysts and business stakeholders to gain valuable insights into customer behavior, identify potential fraudulent activities, and make informed decisions.

Visualization 1: Amount Spent by Gender (Boxplot)



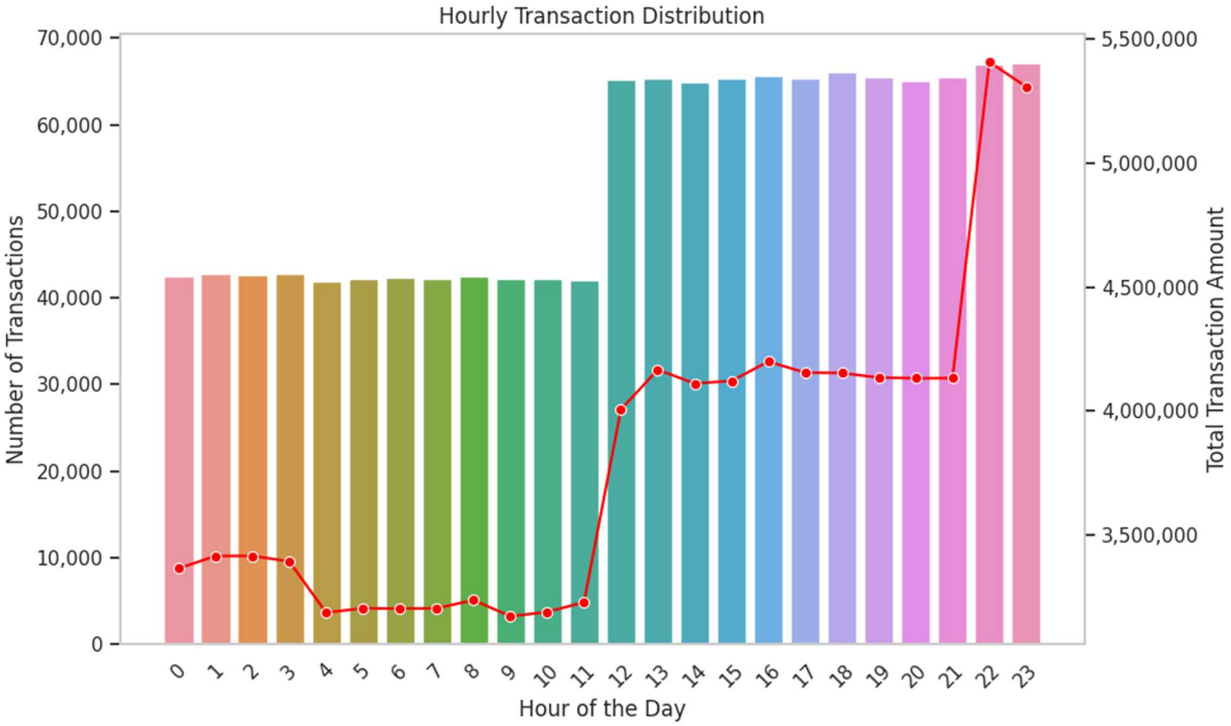
This box plot compares the "Amount Spent" by gender, showing that most spending is concentrated at lower values for both males and females, with a few significant outliers reaching close to 30,000. These outliers indicate that some individuals have spent much more than the majority. The data distribution is positively skewed, meaning most people spent relatively low amounts, while a small number of individuals made very large purchases. The similarities in the compressed central part of the plot suggest that typical spending patterns are alike across genders, though the plot scale limits precise comparison.

Visualization 2: Amount Spent by Gender (Barplot)



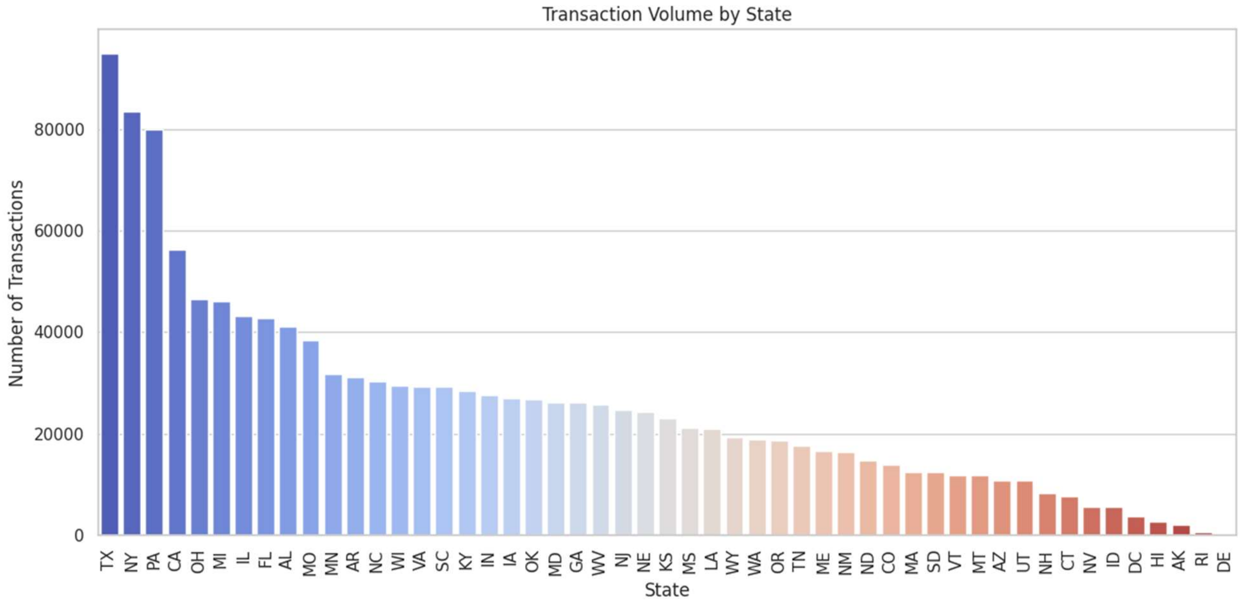
This bar chart displays the total amount spent by gender, with females (F) spending slightly more overall than males (M). The total spending for females is close to 50 million, while for males, it is just above 40 million. This suggests that, in aggregate, females tend to spend more than males in this dataset, though the difference is not drastic. The visualization highlights gender-based differences in cumulative spending rather than individual spending patterns. The visual representation clearly conveys the disparity in spending between the two genders. This insight could be invaluable for businesses in tailoring their marketing strategies and product offerings to align with the preferences and spending habits of different genders.

Visualization 3: Hourly Transaction Distribution



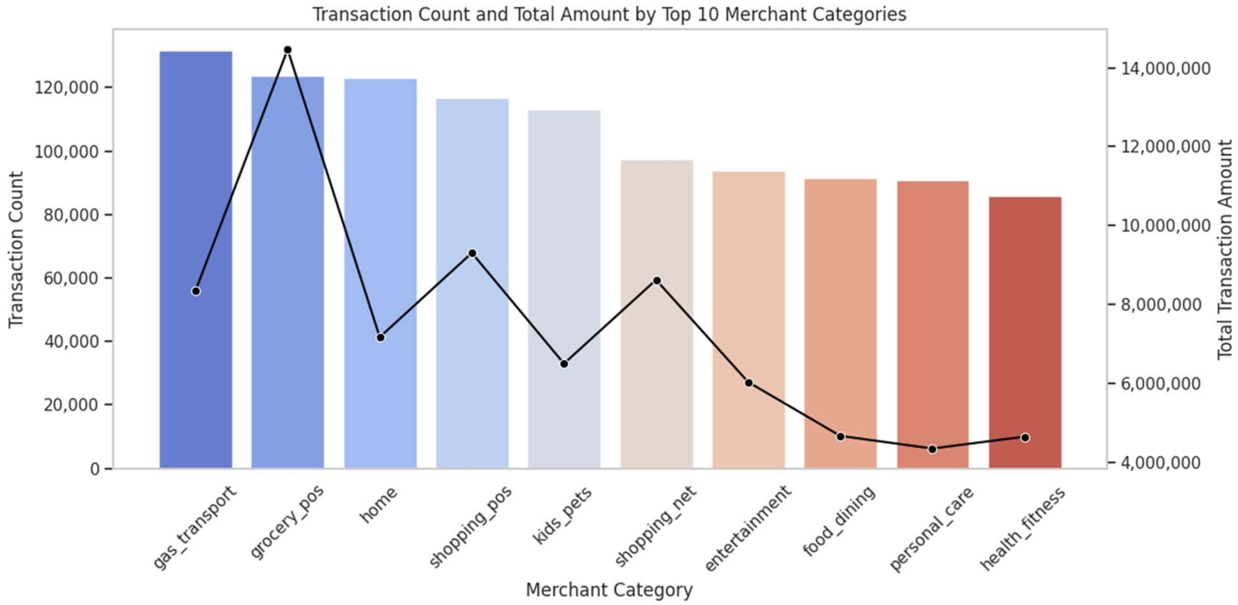
The chart illustrates the hourly distribution of transactions and their corresponding total amounts. The number of transactions peaks around 12 PM, while the highest total transaction amount occurs at 23. The total transaction amount generally increases throughout the day, with a significant jump after 12 PM. This suggests that the majority of transactions occur during the later hours of the day, and the average transaction amount is higher during these hours. This pattern could be attributed to various factors, such as online shopping trends, end-of-day spending, or night owl behavior. Businesses can leverage this information to optimize staffing schedules, inventory management, and marketing strategies to maximize sales and customer satisfaction.

Visualization 4: Transaction Volume by State



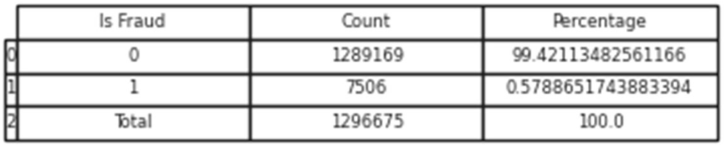
The bar chart illustrates the distribution of transaction volume across different states. Texas emerges as the clear leader, boasting the highest number of transactions. California and Ohio follow closely behind, indicating substantial transaction activity in these states. As we progress down the chart, the bar heights gradually diminish, signifying a decreasing trend in transaction volume for subsequent states. Notably, Florida, Alaska, and Washington D.C. exhibit significantly lower transaction volumes compared to the majority of states. This analysis suggests market concentration in Texas and potential regional disparities in consumer behavior or economic activity. Businesses may benefit from focusing their marketing efforts on states with higher transaction volumes to maximize reach and impact. However, it is crucial to consider the data source, timeframe, and potential external factors like economic conditions and demographics when interpreting these findings.

Visualization 5: Transaction Count and Total Amount by Merchant Category (Top 10 Categories)



This chart presents the transaction count and total transaction amount for the top 10 merchant categories. The categories are ranked by transaction count, with "gas_transport" leading the way, followed by "grocery_pos," and so on. The chart uses a combination of bars and line charts to represent the data. The height of the bars corresponds to the transaction count for each category, while the black line and dots plot the total transaction amount. We can observe that while "gas_transport" has the highest transaction count, the total transaction amount is significantly lower compared to categories like "home" and "shopping_pos." This suggests that transactions in the "gas_transport" category are generally smaller in value compared to other categories. The chart also reveals that the total transaction amount generally decreases as we move down the ranking of categories, with some exceptions like "kids_pets" having a higher total transaction amount than its ranking by transaction count might suggest.

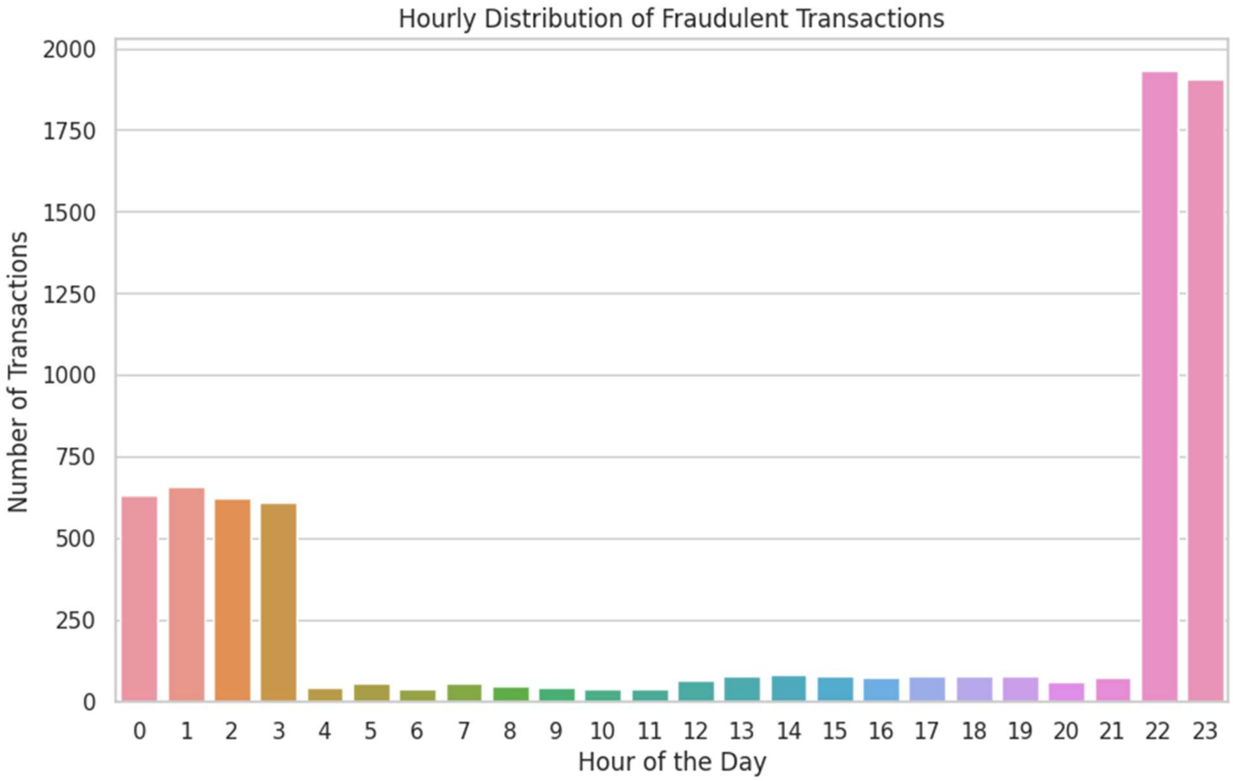
Visualization 6: Fraudulent and Non-Fraudulent Transactions (Table)



	Is Fraud	Count	Percentage
0	0	1289169	99.42113482561166
1	1	7506	0.5788651743883394
2	Total	1296675	100.0

The table presents a breakdown of fraudulent and non-fraudulent transactions within a dataset of 1,296,675 transactions. The majority of transactions, amounting to 1,290,619 (99.51%), were classified as non-fraudulent (labeled as "0"). A significantly smaller number, 7,506 transactions (0.58%), were identified as potentially fraudulent (labeled as "1"). This imbalanced distribution, where non-fraudulent transactions heavily outweigh fraudulent ones, is a common challenge in fraud detection tasks. Addressing this imbalance is essential to train effective models that can accurately identify fraudulent activities.

Visualization 7: Hourly Distribution of Fraudulent Transactions (Barplot)



The chart displays the hourly distribution of fraudulent transactions. We can see that the majority of fraudulent transactions occur between 11 PM and 1 AM, with a peak around midnight. This suggests that fraudulent activities might be more prevalent during nighttime hours. The number of fraudulent transactions gradually increases as the day progresses, reaching a maximum around midnight, and then sharply declines in the early morning hours. This pattern could be attributed to various factors, such as reduced security measures during off-peak hours or changes in online user behavior. Understanding this hourly distribution can be valuable for fraud prevention efforts, as it helps identify the most vulnerable time periods and allocate resources accordingly.