**Fraud Detection Analysis Report**

**Objective**:  
The purpose of this analysis is to explore the relationship between various transaction characteristics and fraudulent behavior. This includes examining the role of transaction amount, transaction time, geographic location, and other features in identifying potentially fraudulent transactions. The ultimate goal is to uncover patterns that can help improve the accuracy of fraud detection models and inform future fraud detection strategies.

**Data Description**

**Dataset Overview**:  
The dataset contains transaction-level data, including features such as transaction amount, time of transaction, customer and merchant locations, and a binary indicator (is\_fraud) marking fraudulent and non-fraudulent transactions. Key features include:

* amt: Transaction amount.
* hour: Time of day when the transaction occurred.
* lat and long: Geographical coordinates of the customer.
* merch\_lat and merch\_long: Geographical coordinates of the merchant.
* distance\_km: Distance between the customer and merchant.
* is\_fraud: Indicator of whether the transaction was fraudulent (1) or non-fraudulent (0).

**Methodology**

**Exploratory Data Analysis (EDA)**:  
I performed exploratory data analysis (EDA) to understand the distribution of key features and their relationships with fraudulent transactions. Key visualizations used include:

* **Boxplots**: Used to examine the distribution of transaction amounts, with a focus on log-transformed values to reduce skewness. Boxplots were chosen to better illustrate the distribution of transaction amounts, especially when using log-transformed values to minimize the effect of outliers.
* **Hexbin Plot**: Hexbin plots were used to explore the relationship between distance and fraud status by visualizing the density of transactions at varying distances.
* **Bar Charts**: Employed to examine the relationship between time of day and fraud rates.

**Statistical Analysis:**  
I conducted statistical tests to assess whether differences in transaction amounts between fraudulent and non-fraudulent transactions Ire statistically significant:

* **T-Test**: A parametric test used to compare the mean transaction amounts between fraudulent and non-fraudulent transactions.
* **Mann-Whitney U Test**: A non-parametric test used to confirm the significance of differences without assuming normality in the data distribution.

**Results and Insights**

**Transaction Amount**

* **Log-Transformed Boxplot**:
  + **Insight**: Higher transaction amounts show a stronger association with fraud. The log-transformation made this relationship clearer by reducing the impact of outliers.
  + **Result**: Fraudulent transactions tend to have higher median transaction amounts compared to non-fraudulent ones.
  + **Interpretation**: Fraudulent behavior is more likely when larger sums of money are involved, suggesting a need for closer monitoring of high-value transactions.
* **Statistical Significance**:
  + **T-Test**: The t-test comparing transaction amounts between fraudulent and non-fraudulent transactions returned a p-value of 0.0, indicating a statistically significant difference between the two groups.
  + **Mann-Whitney U Test**: The Mann-Whitney U test confirmed the significant difference in transaction amounts, with a p-value of 0.0. This supports the conclusion that fraudulent transactions are more likely to involve higher transaction amounts, even when accounting for non-normality.

**Distance from Merchant**

* **Hexbin Plot**:
  + **Insight**: Most transactions, both fraudulent and non-fraudulent, occur at short distances (close to zero). This indicates that fraudsters may often operate close to the cardholder’s location, or that transactions with zero distance could represent online purchases.
  + **Result**: No significant distinction was found between fraudulent and non-fraudulent transactions based on distance, especially with many transactions clustered near zero distance.
  + **Interpretation**: Distance from the merchant may not be a strong standalone indicator of fraud in this dataset, but it could still be useful when combined with other features.

**Time of Day**

* **Fraud Rate by Hour with Confidence Intervals**:
  + **Insight**: There are clear spikes in fraud rates during early morning hours (midnight to 2 AM) and late at night (10 PM to midnight). These time frames may reflect times when fraudsters assume cardholders are less vigilant.
  + **Result**: Fraud rates show statistically significant increases during these time periods.
  + **Interpretation**: Additional scrutiny or monitoring could be applied to transactions occurring during high-risk hours.

**Conclusion**

* **Transaction Amount**: The analysis revealed that transaction amount is a key indicator of fraud, with higher amounts significantly more likely to be associated with fraudulent behavior. Both parametric and non-parametric statistical tests confirm this finding.
* **Time of Day**: Fraud is more likely to occur during late-night and early-morning hours, indicating potential periods of higher risk.
* **Distance from Merchant**: Geographic distance did not emerge as a significant standalone factor but could still provide value when combined with other features in a predictive model.

**Recommendations**

* **Monitor High-Value Transactions**: Based on the findings, I recommend implementing stricter monitoring and fraud detection measures for high-value transactions, as they are more likely to be associated with fraud.
* **Increased Scrutiny During High-Risk Hours**: Given the elevated fraud rates during late-night and early-morning hours, it would be beneficial to apply additional scrutiny to transactions occurring during these times.
* **Further Feature Engineering**: Future work should consider combining features like distance and transaction time for more advanced predictive modeling. These features might not be strong standalone indicators but could contribute to more sophisticated fraud detection models when combined with others.