

# Fraud Detection Analysis

Saloni Bonde

Using Data Mining  
Techniques

01 - Introduction

02 - Data Visualization

03 - Models

04 - Conclusions



# 01 - Introduction

- Financial fraud costs billions annually.
- Fraud detection is critical for both financial institutions and consumers.
- In 2016, global losses from fraudulent transactions exceeded \$16 billion.

*Our goal:  
To classify  
transactions  
as legitimate  
or fraudulent*

# Dataset Overview

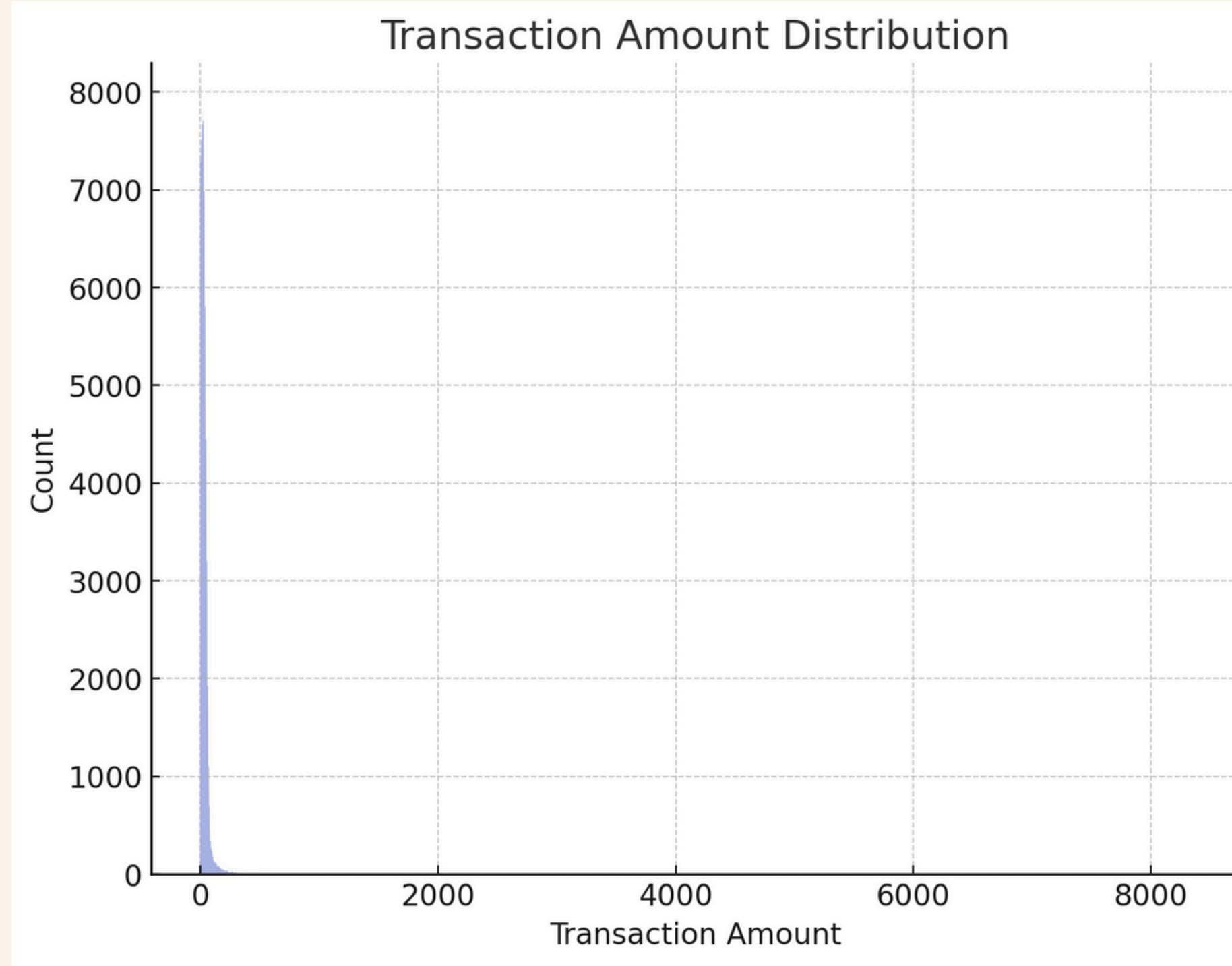
The dataset consists of 594,643 transactions with various attributes such as:

- Customer information (age, gender)
- Merchant information (ZIP code, merchant category)
- Transaction details (amount, time step)

The target variable: Fraudulent (1) or Legitimate (0)



# 02 - Data Visualization

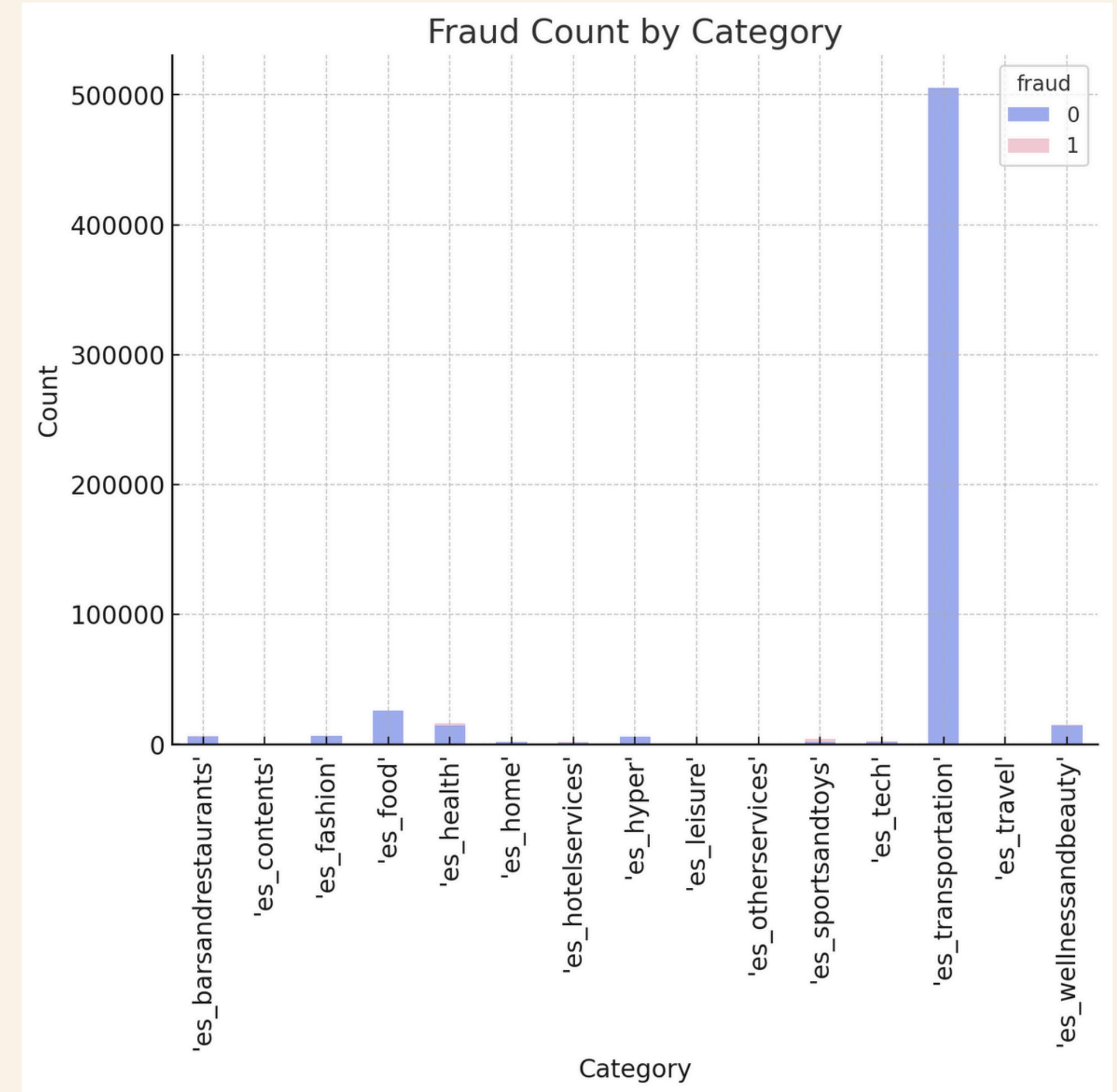


*Distribution of  
Transaction  
Amounts*

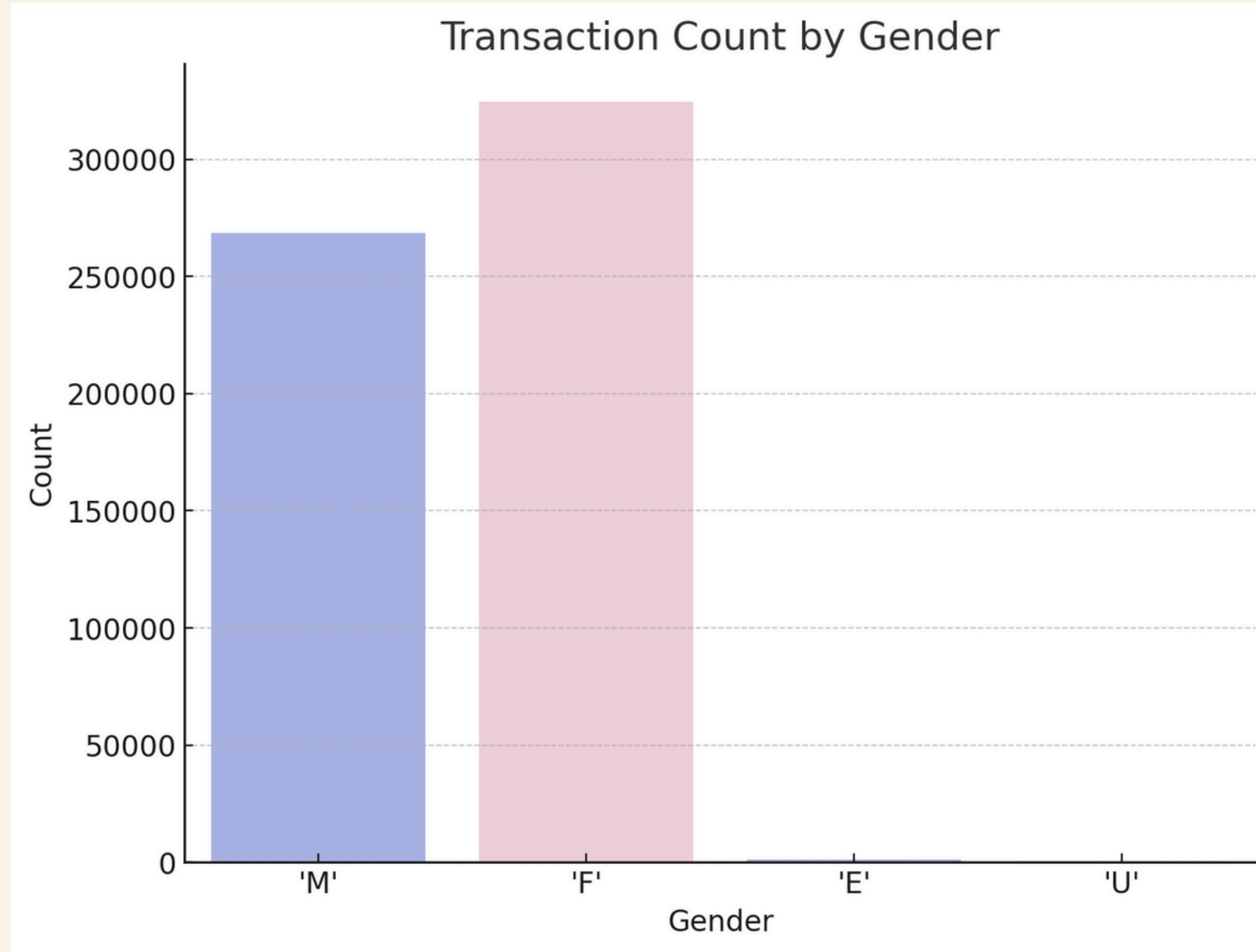


# Fraud Distribution by Category

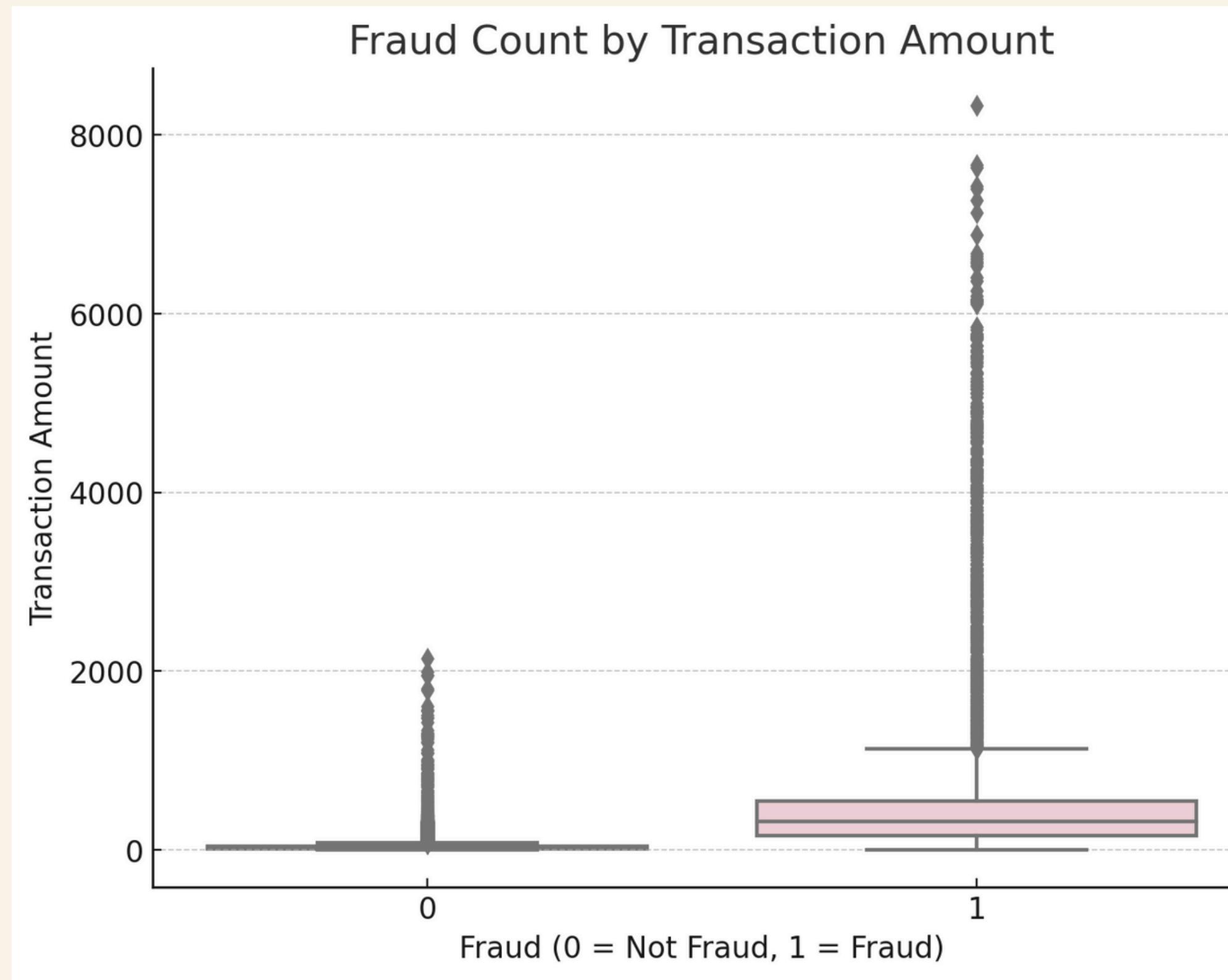
*Visualizes the count of fraudulent and non-fraudulent transactions by merchant category.*



# Transaction Amount by Gender

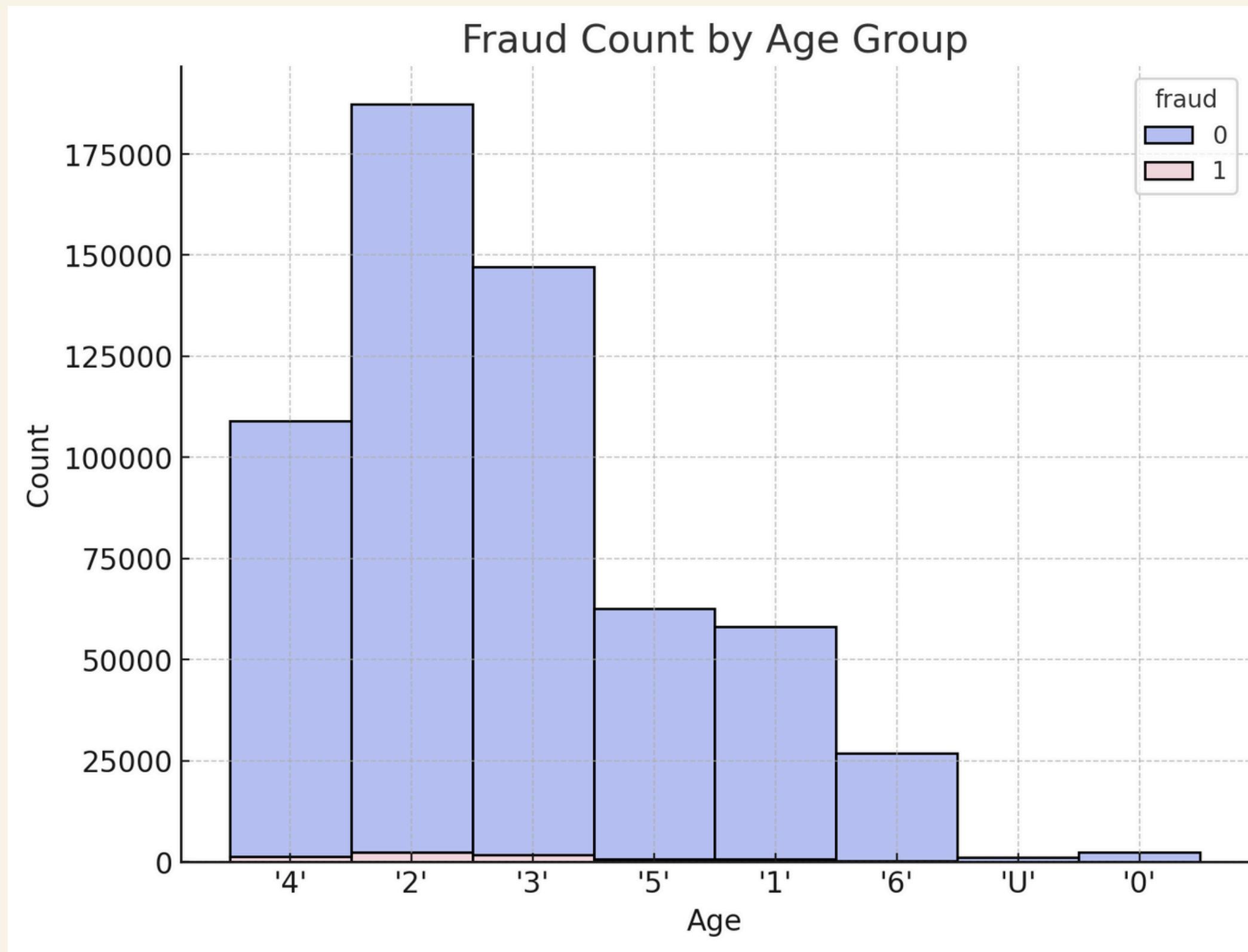


# Fraud Distribution by Transaction Amount



Boxplot  
showing how  
transaction  
amounts  
differ  
between  
fraudulent  
and non-  
fraudulent  
transactions.

# Fraud Count by Age Group



Count plot depicting fraud distribution across different age groups.

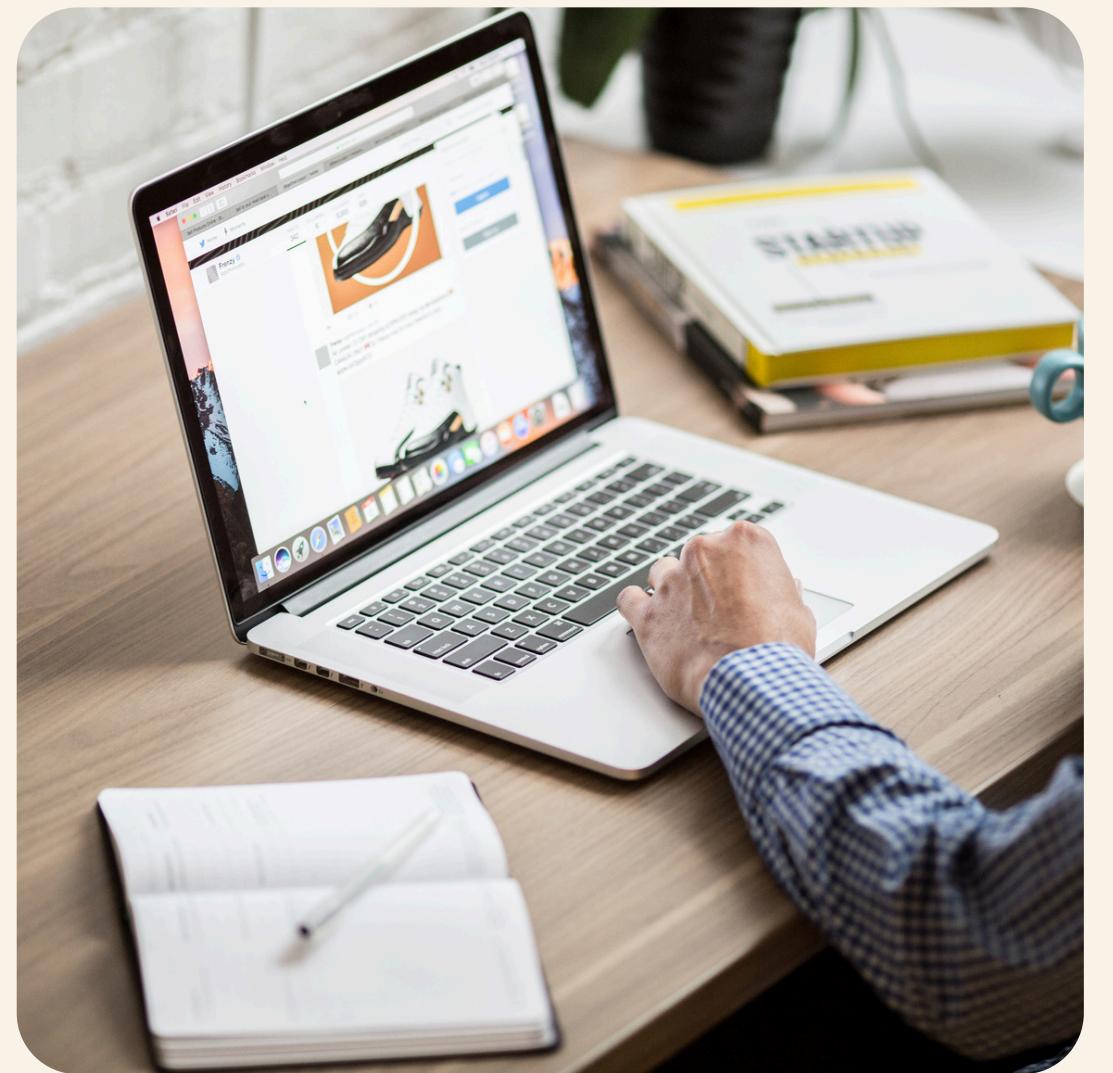
# 03 - Models

## Models Used for Fraud Detection

- a. Naive Bayes
- b. Logistic Regression
- c. Neural Networks
- d. Decision Trees
- e. Random Forest



Alba Castro



# Data Preprocessing

## Data Cleaning:

- Removed special characters from categorical columns
- Categorical variables were encoded using one-hot encoding

## Standardization:

- Standardized numerical features to improve model performance

## Handling Imbalance:

- Used SMOTE to address class imbalance (very few fraudulent transactions)

# Evaluation Metrics

- **Precision:** How many predicted fraud cases were correct?  
(High precision = fewer false positives)
- **Recall:** How many actual fraud cases were correctly identified?  
(High recall = fewer false negatives)
- **False Positives (FP):** Legitimate transactions classified as fraudulent.
- **False Negatives (FN):** Fraudulent transactions classified as legitimate.



# Evaluation

Model	Precision	Recall	FP	FN
Naive Bayes	0.1957	0.9993	2	11695
Logistic Regression	0.8887	0.7236	796	261
Neural Network	0.8721	0.7624	684	322
Decision Tree	0.9027	0.7212	803	224
Random Forest	0.8845	0.7462	1096	421

Model  
Performance  
Comparison

# Key Insights

- **Naive Bayes:** High recall but impractically low precision. Good at identifying fraud but too many false positives.
- **Logistic Regression:** Balanced performance but slightly lower recall.
- **Neural Networks:** Excellent balance between precision and recall, making it reliable for reducing both FP and FN.
- **Decision Tree:** High precision, useful for identifying fraud accurately, but slightly lower recall.
- **Random Forest:** Robust model with balanced performance, fewer false negatives than decision trees.

## 04 - Conclusions

*How to Use These Findings in Business*

### **Fraud Detection Automation:**

- Implement the chosen machine learning models (e.g., Neural Networks, Random Forest) into real-time transaction monitoring systems.
- Automate fraud detection, flagging suspicious transactions for further review before they are processed.

### **Customer Trust & Experience:**

- Reduce false positives to avoid unnecessary transaction holds or customer service calls.
- Improve customer satisfaction by minimizing fraud-related disruptions.

# *How to Use These Findings in Business*

## **Cost Savings:**

- By minimizing both false positives and negatives, reduce losses from fraudulent activities.
- Decrease manual transaction review costs by utilizing high-precision models to reduce the volume of false alerts.

## **Risk Mitigation:**

- Leverage the model insights to enhance fraud prevention strategies, identify high-risk customers or transactions, and adjust security measures proactively.

# Thanks