* **MIDDLE EAST TECHNICAL UNIVERSITY**

**Department of Statistics**

**STAT112- Introduction to Data Processing and Visualization**

**2023-2024 Fall**

# Report of Fraud Detection Project

# Instructor: Prof. Dr. Ceylan Yozgatlıgil

# Assistant: Petek Aydemir

# Submitted By: Nesrin Kaya

# Group:

# Doğa Aksan, Eda Çarkçı, Kaan Keçeci,

# Nesrin Kaya, Pelin İşgör

14.01.2024

1. **Abstract:**

This study examines the relationship between financial transactions and user profiles, focusing on a dataset for fraud detection purposes. Fraud detection typically refers to the process of detecting fraud in financial transactions or other types of interactions. Companies, financial institutions, and other organizations attempt to detect abnormal or suspicious activities by monitoring customer transactions and interactions. Fraud detection is often a process supported by automatic analysis and data mining techniques. In this process, algorithms and models attempt to identify transactions that deviate from what is normally expected, show patterns different from previous ones or violate specific rules.

Fraud detection typically involves technologies such as big data analytics, artificial intelligence, and machine learning. By using big data analytics, fraud detection data will be examined in this report. The dataset includes various columns representing specific transaction attributes. Starting from the transaction date features such as transaction type, amount, card type, location, and user age have been explored. Critical information like transaction approval status and fraud detection has also been analyzed. This study aims to understand potential relationships between user profiles and transaction features for enhancing financial security and developing fraud prevention strategies.

1. **Introduction:**

This dataset consists of 13 columns and 98 rows. Each row contains details of financial transactions, including information such as transaction type, merchant category, transaction amount, card type, and more. Demographic features like user age group and account balance are also included in each row. The dataset provides information about the approval status of transactions and a fraud label, making it valuable for analysis and modeling tasks. With its compact size, this dataset is designed for applications in financial analysis and fraud detection.

|  |  |  |
| --- | --- | --- |
| ID | Transaction identifier. | Categorical |
| Month | The month of the transaction. | Categorical |
| Year | Year of the transaction. | Categorical |
| Transaction Type | Type of transaction | Categorical |
| Merchant Category | Category of the merchant involved in the transaction. | Categorical |
| Transaction Amount | Amount of money involved in the transaction. | Numerical |
| Card Type | Type of credit/debit card used in the transaction. | Categorical |
| Location | The location where the transaction took place. | Categorical |
| Distance | Distance of the transaction location. | Numerical |
| Approval | Indicates whether the transaction has been approved or not. | Categorical |
| User’s Age | Age of the user involved in the transaction. | Numerical |
| Account Balance | The current balance in the user's account. | Numerical |
| Fraud | Indicates whether the transaction is flagged as fraudulent or not. | Categorical |

1. **Data Cleaning Steps:**

In order to understand the data and take action by using it, data should be understood successfully because if the data are understood wrongly, it may lead you to draw wrong conclusions, so to understand data correctly, some steps should be taken. These steps are:

* Removing duplicates
* Determining null values
* Controlling variable’s name (whether they are representative of values or not)
* Formatting string values (whether they have the same meaning, or problems that are caused by type errors)
* Checking uniformity of units
* Outliers

In this subsection, the steps that are used in cleaning and tidying data will be explained below. All the steps that are done in this subsection are done by using NumPy and pandas. Moreover, all the changes are done by using “inplace == True”, so all the changes are done all over the dataset.

* To examine the data easily, all the variable’s names were changed into the same format.
* All the datasets were checked whether the dataset had null variable and a duplicated row or not.

1. ID:

* This column was checked for duplicated values and null values.
* Duplicated values and null values were deleted. For instance, the 60th row was duplicated, and it has been deleted, so that the total number of rows in the dataset decreased to 98 rows.

1. MONTH:

* It was checked whether it has null values.

3. YEAR:

* The data type of the year column was changed to string type. The column type is returned to integer in order to find the index of invalid value.
* An invalid value was detected in the 36th row. Which was a number with a different number of digits, which is 20125. It was replaced with 2015.

4. TRANSACTION AMOUNT:

* After determining NA values, they were replaced with mean.

5. DISTANCE:

* After determining NA values, they were replaced with mean.

6. USER’S AGE:

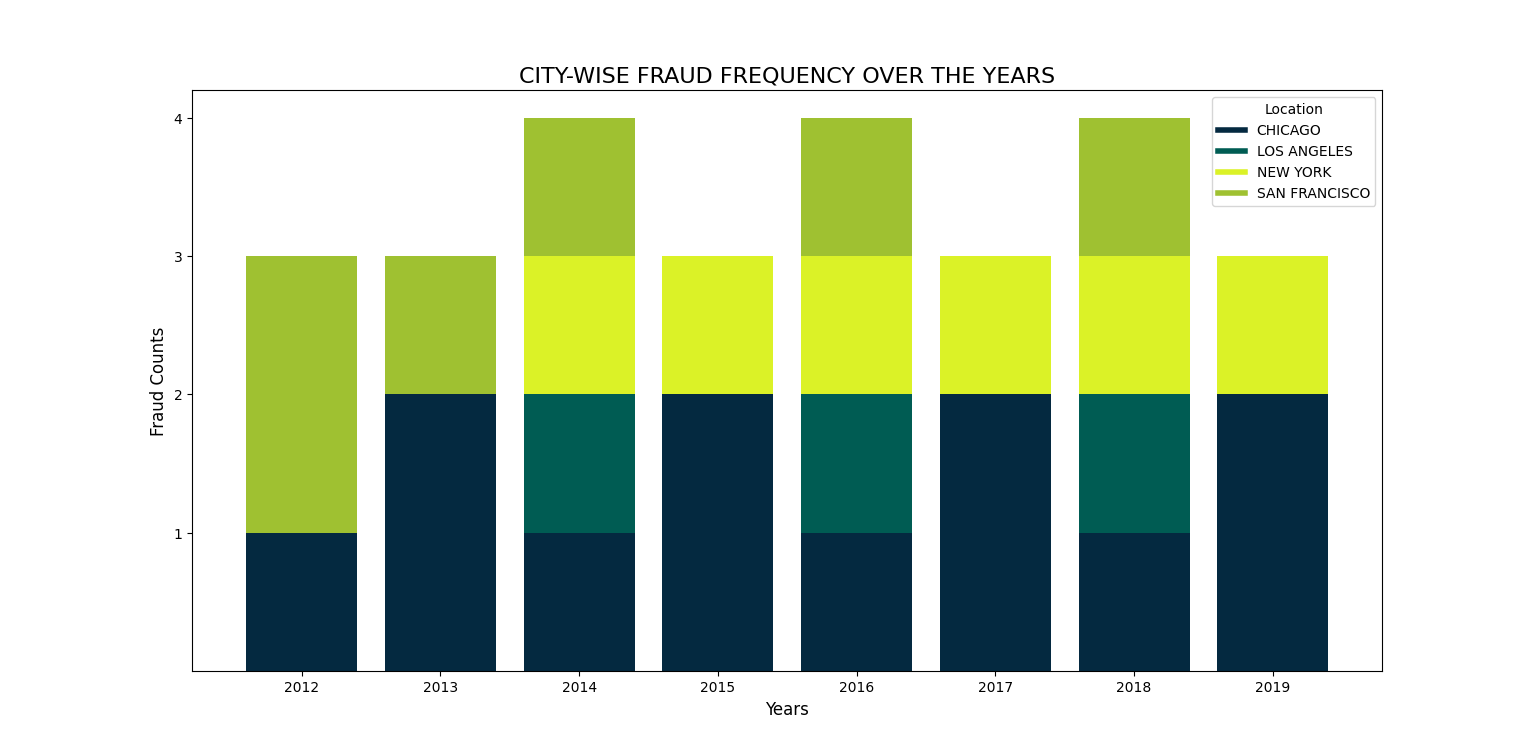
* There was a negative value, and it was fixed by turning it into positive value.

7. FRAUD:

* After determining NA values, they were replaced with mode.

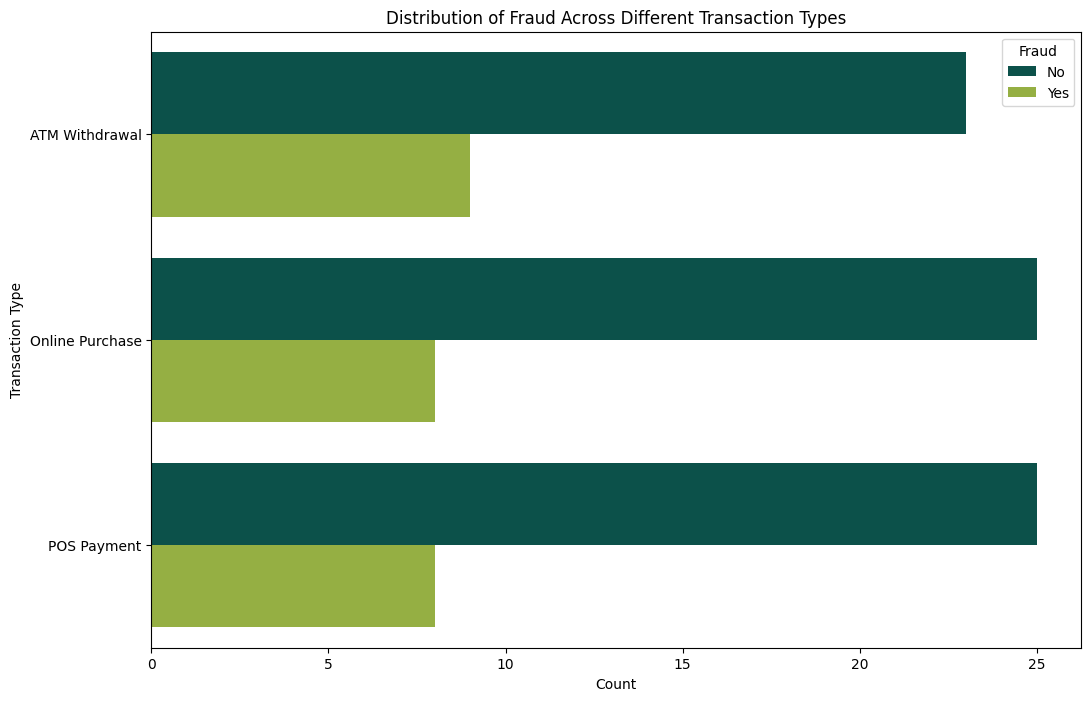
1. **Research Questions:**

**Table 1**. How has the incidence of fraud been distributed among cities throughout the years?

**Figure 1.** Stacked bar chart between fraud counts and years.

When the years are examined, in 2012 and 2013, fraud has just occurred in Chicago and San Francisco. For Chicago, it can be said that there was constant fraud between 2012 and 2019. Also, Chicago hosts fraud every year and it is seen that the number increased especially in 2013, 2015, 2017, and 2019. Fraud was first seen in Los Angeles and New York in 2014. When we look again at the years 2014, 2016, and 2018, fraud occurred in all 4 cities. In 2013, 2015, 2017, and 2019, Chicago was the place where the most fraud took place, and in these years, Chicago was accompanied only by New York. While there was no fraud in New York until 2014, there was constant fraud from this year until 2019. In San Francisco, the number of frauds decreased after 2012 but continued in the following years. Los Angeles, on the other hand, was the place with the least fraud, and only a small number of frauds were seen in 2014, 2015, and 2018. There is no fraud in 2020.

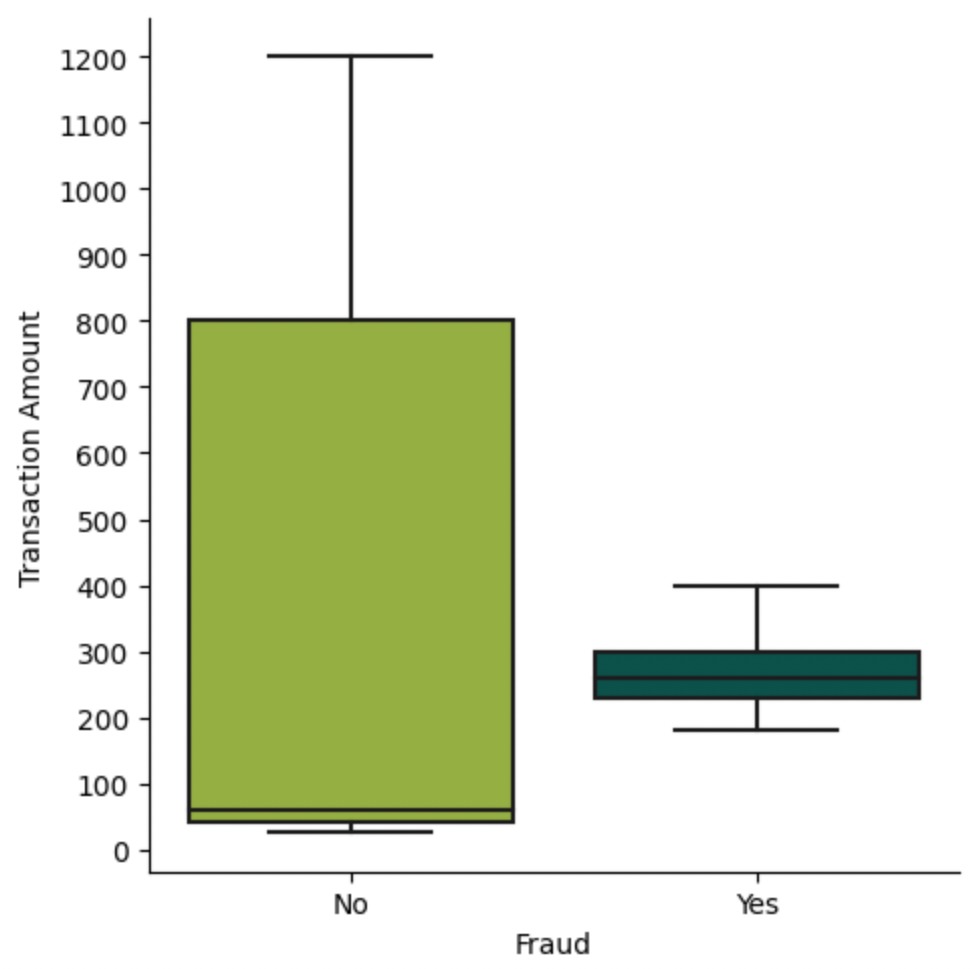
**Table 2.** How is Fraud Distributed Across Different Transaction Types?



**Figure 2.** Clustered bar chart between transaction type and fraud count.

The horizontal bar chart is designed to compare the occurrences of fraud ("Yes" and "No") across three distinct transaction types: "POS Payment," "ATM Withdrawal," and "Online Purchase." This graph aims to provide a statistical evaluation by illustrating the number of transactions falling under each fraud category for each transaction type. According to the graph for POS payments, it displays 8 transactions with fraud ("Yes") and 25 transactions without fraud ("No"). The count of transactions without fraud is higher than those with fraud, suggesting a generally low incidence of fraudulent activities in POS payments. Examining the graph for ATM withdrawals, it indicates 9 transactions with fraud ("Yes") and 23 transactions without fraud ("No"). Once again, the count of transactions without fraud is greater than those with fraud, suggesting a relatively low occurrence of fraudulent activities in ATM withdrawals. For online purchases, the graph shows 8 transactions with fraud ("Yes") and 25 transactions without fraud ("No"). Similarly, the count of transactions without fraud is higher than those with fraud, indicating a generally low prevalence of fraudulent activities in online purchases. In summary, the number of transactions with fraud is notably low across the selected three transaction types, and the count of transactions without fraud ("No" category) significantly outweighs those with fraud ("Yes" category). These observations suggest a low incidence of fraud in the specified transaction types. This information serves as a foundation for further comprehensive analysis.

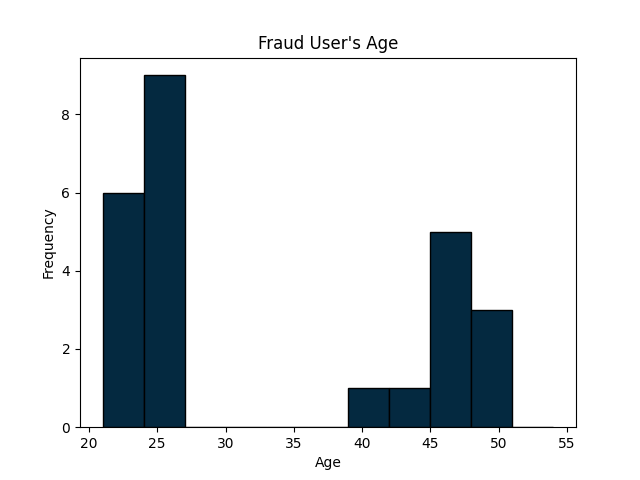
**Table 3.** How much is the expected value for detecting fraud over the observed cases?

****

**Figure 3.** Box plot between transaction amount and fraud.

First, we start by examining the relation between “Transaction Amount” and “Fraud (No)”. It has a right-skewed distribution. This can be understood by looking at whisker lengths, or the value [(Q3-Q2)>(Q2-Q1)]. The mean value of this distribution is nearly 65 units, which implies that half of the observations with no fraud are less than 65 units; therefore, the employees do not need to worry about fraud detection less than these values. Moreover, there is no outlier in our observations, which makes these values more reliable; however, the variance is higher than observations of” Fraud(Yes)”.Secondly, the relation between “Transaction Amount” and “Fraud(Yes)” is more different than the relation between ”Transaction Amount” and “Fraud(No)”.” Transaction Amount” and “Fraud(Yes)” distribution has approximately symmetrical distribution since upper whisker is slightly longer than the lower whisker. Most of the frauds are detected between 310 and 230 units, so the bank workers should give more attention to this interval when they are checking whether there is any.

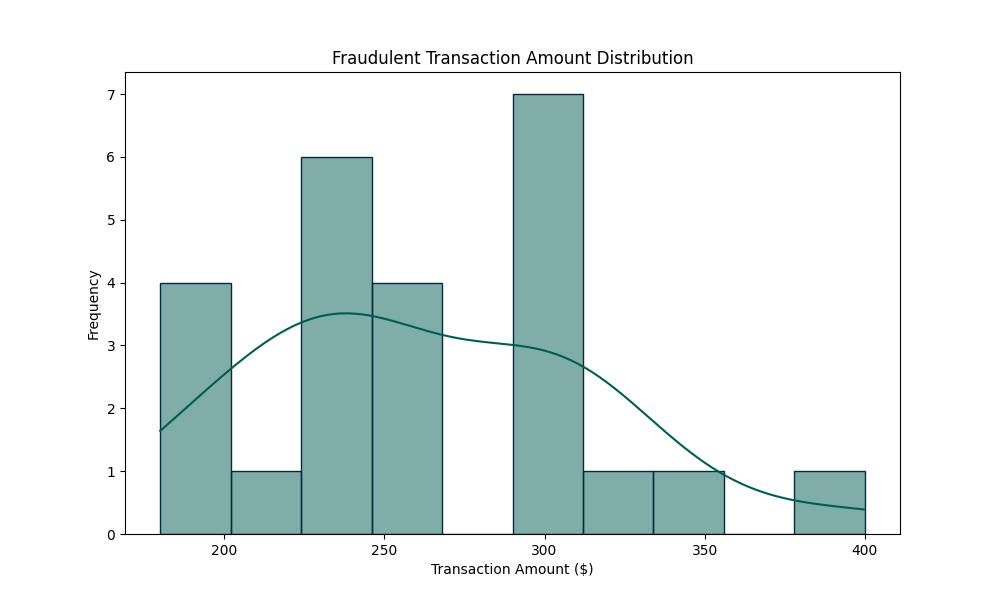
**Table 4. What is the distribution of age for fraud users?**

****

**Figure 4. Histogram of fraud users’ age.**

A histogram is an appropriate option to show the distribution of numerical data. **This histogram shows the fraud user’s age.**  According to this histogram, there is a positively skewed distribution. This can be proved with summary statistics in the dataset’s data cleaning page on GitHub. **According to this histogram, most frauds are around 25 years old, which approves the summary statistics code about the mode that is 28. Furthermore, there is nobody in the range between 30 and 35 age in this sample. While looking at this result, people who work in the bank might conduct special research or study about the age group around 25. On the other hand, there are substantial amount of fraud in around 45 age group, which means that it needs to pay attention to this group.**

**Table 5. How is the distribution of fraudulent transaction amount?**

****

**Figure 5. Histogram of fraudlent transaction amount.**

**When we analyze the distribution of transaction amount of fraud occurred purchases. It can be seen that mode of the its distribution is almost 235$, since the central tendency is shown as mean<median<mode, this graph is right(positively)-skewed, which means most of the fraud activity can be examined when the transaction amount is lower. Most of the fraud happens around 300$ and the least fraud happens around 350-400$. However, this rate of fraud seems numerous, when it is compared with the rate of fraud at 200-250$. Also, the rate of fraud at 275$ and 365$ is almost zero.**

**5. Conclusion:**In general, this study focuses on fraud detection, and which thinks a company should be careful about. Therefore, fraud has been considered from many aspects such as location, transaction amount, and even customers’ age. To conclude, in the first step the data was cleaned, and then some research questions were answered. In this research, answers to these questions were searched for by using multiple data visualization techniques. It was also supported using statistical summarization. Some important results about the research questions are listed below:

* Most of the fraud cases were observed between 230 units and 310 units with 7 frequencies of 300 units.
* While most fraud users are around 25 years and older than 45, and most of them are located in Chicago and use ATM Withdrawal.

**GitHub Link:** https://github.com/nesrinK/112.Final.Project.