PROFESSIONAL TRAINING REPORT
at

Sathyabama Institute of Science and Technology
(Deemed to be University)

Submitted in partial fulfilment of the requirements for the award of
Bachelor of Engineering Degree in Computer Science and Engineering

By
K.V.L.YASHWANTH
REG. NO: 41111419

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

SCHOOL OF COMPUTING

SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A" by NAAC | 12 B Status
By UGC | Approved by AICTE
JEPPIAAR NAGAR, RAJIV GANDHI SALAI,
CHENNAI – 600 119

OCTOBER- 2023

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited with Grade "A" by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the Bonafide work of **K.V.L.YASHWANTH (Reg.No.41111419)** who carried out the project entitled "**Online Payment Fraud Detection**" under my supervision from JULY 2023 to OCTOBER 2023.

Internal Guide

Ms.D.Aishwarya

Head of the Department

Dr. L. Lakshmanan, M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

Internal Examiner **External Examiner**

DECLARATION

I K.V.L.YASHWANTH (41111419) hereby declare that the Project Report entitled **Online Payment Fraud Detection** done by me under the guidance of **Ms.D.Aishwarya** (Internal) **and Dr.M.S.Roobini** (External) is submittedin partial fulfilment of the requirements for the award of Bachelor of Engineering degree in **Computer Science and Engineering.**

DATE:

PLACE: Chennai **SIGNATURE OF THE CANDIDATE**

ACKNOWLEDGEMENT

I am pleased to acknowledge my sincere thanks to **Board of Management** of **SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. T. Sasikala M.E., Ph.D.**, **and Dr. L. Lakshmanan M.E., Ph.D., Heads of the Department of Computer Science and Engineering** for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **MS.D.Aishwarya** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the **Department of Computer Science and Engineering** who were helpful in many ways for the completion of the project.

TRAINING CERTIFICATE

ABSTRACT

Online payments have become a ubiquitous part of modern life, with more and more people relying on digital transactions for everything from shopping to bill payments. However, with the rise of online payments, there has also been an increase in fraudulent activity, with criminals using sophisticated methods to steal sensitive information and money from unsuspecting victims.

To combat this, many organisations have turned to EDA Analysis to help detect and prevent online payment fraud. EDA Analysis is a powerful algorithm that can identify patterns in large datasets, making it well-suited for the task offraud detection.

To effectively use EDA Analysis for online payments fraud detection, several steps must be taken. The first step is to ensure that the dataset is properly cleaned and preprocessed,including handling missing values and encoding categorical variables.

Next, feature engineering can be used to create new features that may be useful for detecting fraud. This can include features such as the time between when a transaction is made and when it is verified, or the number of failed login attempts.

To make online payments fraud detection more effective, it is important to integrate the model into a real-time scoring system that can detect and block fraudulent transactions in real-time. This can help to prevent further fraudulent activity and reduce the impact of any fraudulent transactions that do occur.

Finally, it is important to continuously train and update the model as new databecomes available. This can help to improve the accuracy and detection capabilities of the model over time.

Overall, using EDA Analysis for online payments fraud detection is a powerful way to combatthe increasing prevalence of fraud in digital transactions. By following the steps outlinedabove and continuously improving the model, organisations can better protect themselves and their customers from fraudulent activity.

TABLE OF CONTENTS

CHAPTER No.	TITLE	PAGE NO.
	ABSTRACT	VI
	LIST OF FIGURES	IX
	LIST OF ABBREVIATIONS	X
1	INTRODUCTION	1
	1.1 EDA Analysis	2
	1.2 Data Analytics	3
	1.3 Need Of Data Analytics	4
	1.4 Significance Of EDA Analysis	4
	1.5 What Is Correlation Matrix?	5
	1.6 What is Preprocessing The Data?	6

2	SYSTEM ANALYSIS	7
	2.1 Aim & Scope	7
	2.2 How does it work	8
3	SYSTEM DESIGN AND IMPLEMENTATION	9
	3.1 Algorithm	9
	3.2 Ideation Map	10
4	RESULT & DISCUSSION	11
	4.1 Result	11
5	CONCLUSION	15
	5.1 Reference	17
	5.A Screenshots	18

LIST OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
4.1	Fig.4.1 Count plot for Amount	11
4.2	Fig.4.2 Count plot of Fraud	12
4.3	Fig.4.3 Count plot for type of Fraud	13
4.4	Fig.4.4 Pie Graph of Fraud Distribution	13
4.5	Fig.4.5 Heat Map	14

LIST OF ABBREVIATIONS

ABBREVATION	EXPANSION
EDA	Exploratory Data Analysis
SQL	Structured Query Language

CHAPTER 1

INTRODUCTION

The rise of online payments has revolutionised the way we conduct transactions, providing a convenient and efficient way to buy goods and services. However, with the increasedreliance on digital transactions, there has also been a corresponding rise in fraudulent activity.Fraudsters are constantly developing new tactics to bypass security measures and steal money from online payments. This makes it challenging for organisations to stay ahead of the curve and protect their customers from fraudulent transactions.

To combat this problem, many organisations are turning to EDA Analysis to help detect and prevent online payments fraud. EDA is a highly effective algorithm that has proven to be successful in a wide range of applications, including fraud detection.In this project, EDA Analysis was used to analyse a dataset obtained from Kaggle to detectfraudulent online payments. The goal was to develop a model that could accurately identify fraudulent transactions and prevent them from occurring.

To achieve this, the project began with data preprocessing, where the dataset was cleanedand preprocessed to ensure that it was optimised for analysis. This included handling missing values and encoding categorical variables.The next step was feature engineering, where new features were created that could be useful in detecting fraud. These features included variables such as the time between when a transaction is made and when it is verified, and the number offailed login attempts.

Finally, the project emphasised the importance of continuous learning, where the model is updated and retrained as new data becomes available. This ensures that the model remains effective over time and can adapt to changing patterns of fraud.

Overall, this project demonstrates the power of EDA Analysis in detecting online payments fraud and highlights the importance of using EDA Analysis to combat this growing problem.By following best practices and continuously updating the model, organisations can stay aheadof fraudsters and protect their customers from financial losses.

1.1 EDA Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding your data before diving into building machine learning models for online payment fraud detection. Here's a step-by-step guide on how to perform EDA for online payment fraud detection:

1. Understand the Data:

Features: Identify the features in your dataset. Common features include transaction amount, location, time, type of card used, etc.

Target Variable: Identify the fraud indicator variable (binary: fraud or not fraud).

2. Data Summary:

Statistics: Calculate basic statistics for numerical variables (mean, median, standard deviation, etc.) to understand the central tendency and spread.

Class Distribution: Check the distribution of fraud vs. non-fraud cases. Imbalanced classes might need special handling during model training.

3. Data Visualization:

Histograms: Plot histograms for numerical variables to understand their distributions.

Box Plots: Use box plots to identify outliers in numerical features.

Count Plots: Visualize the count of fraud vs. non-fraud cases for categorical variables.

Correlation Matrix: Create a correlation matrix to understand relationships between variables. This might help identify potential patterns.

1.2 Data Analytics

Data analytics is the process of examining, cleaning, transforming, and modelling data with the aim of discovering useful information, drawing insights, and making informed decisions. It involves various techniques and tools, including statistical analysis, data mining, and predictive modelling, to extract knowledge from rawdata.Data analytics is widely used in various fields, such as business, healthcare, finance, marketing, and science, to help organisations make data-driven decisions, optimise their operations, improve their products and services, and gain a competitive edge. The main goal of data analytics is to uncover patterns, trends, and relationships in data that can provide valuable insights and help solve complex problems.

Data analytics has a wide range of uses across various industries and fields. Here are some examples of how data analytics is used:

- Business intelligence: Data analytics can help businesses make informed decisions by analysing data on customer behaviour, sales trends, and market conditions.
- Marketing: Data analytics can help companies develop more effective marketing strategies by analysing customer demographics, behaviour, and preferences.
- Finance: Data analytics can help financial institutions analyse risks, detect fraud, and optimise investments by analysing market data, customer behaviour, and financial performance.

Overall, data analytics can help organisations make data-driven decisions, optimise their operations, improve their products and services, and gain a competitive edge.

1.3 Need of Data Analytics

Data analytics is essential because it enables organisations to gain insights and make informed decisions based on data-driven evidence. Here are some reasons why data analytics is important:

- To gain insights: Data analytics helps organisations to gain insights into their operations, customers, and competitors, which can help them to make better decisions, optimise their processes, and improve their products and services.
- To make informed decisions: Data analytics enables organisations to make data-driven decisions, which can reduce the risk of making incorrect decisions.
- To improve efficiency: Data analytics can help organisations to optimise their operations, reduce costs, and improve efficiency by identifying inefficiencies and areas for improvement.

1.4 Significance Of EDA Analysis

The significance of EDA is to Different fields of science, economics, engineering, and marketing accumulate and store data primarily in electronic databases. Appropriate and well-established decisions should be made using the data collected. It is practically impossible to make sense of datasets containing more than a handful of data points without the help of computer programs. To be certain of the insights that the collected data provides and to make further decisions, data mining is performed where we go through distinctive analysis processes. Exploratory data analysis is key, and usually the first exercise in data mining. It allows us to visualize data to understand it as well as to create hypotheses for further analysis. The exploratory analysis centers around creating a synopsis of data or insights for the next steps in a data mining project. EDA actually reveals ground truth about the content without making any underlying assumptions. This is the fact that data scientists use this process to actually understand what type of modeling and hypotheses can be created. Key components of exploratory data analysis include summarizing data, statistical analysis, and visualization of data. Python provides expert tools for exploratory analysis, with pandas for summarizing; scipy, along with others, for statistical analysis; and matplotlib and plotly for visualizations financial losses.

1.5 What Is Correlation Matrix?

A correlation matrix is a statistical tool used to describe the relationship between two or more variables. It is a matrix that contains the correlation coefficients between all possible pairs of variables in a dataset. Correlation coefficients measure the strength and direction of the linear relationship between two variables, and can range from -1 to +1.The diagonal of a correlation matrix always contains a correlation coefficient of 1, because each variable is perfectly correlated with itself. The upper and lower triangles of the matrix contain the same information, because the correlation between variable A and variable B is the same as the correlation between variable B and variable A.

Correlation matrices are commonly used in data analysis and statistical modelling to identify patterns and relationships between variables. They can help to identify variables that are strongly correlated with each other, which can be useful in reducing the number of variables in a dataset or identifying potential collinearity issues in a regression analysis.

Interpreting correlation matrices can be complex, and it is important to consider the context and limitations of the data when making inferences. Correlation coefficients only measure the strength and direction of the linear relationship between two variables, and do not necessarily indicate causation or a meaningful relationship between the variables. Correlation matrices can also be influenced by outliers or other sources of bias in the data, and it is important to use other statistical tools and methods to confirm and validate any relationships identified through a correlation matrix.

Overall, correlation matrices are a useful statistical tool for identifying patterns and relationships between variables in a dataset. They can help to identify potential issues with collinearity in regression analyses, reduce the number of variables in a dataset, and visualise the relationships between variables. However, careful interpretation and validation of the results is necessary to ensure that the identified relationships are meaningful and not the result of bias or other sources of error in the data.

1.6 What Is Preprocessing The Data ?

Preprocessing data is an essential step as it involves transforming raw data into a format that can be used by a EDA Analysis.

There are several techniques used in preprocessing data, including data cleaning, feature scaling, and data encoding.

Data Cleaning: Data cleaning involves identifying and correcting errors, inconsistencies, and missing values in the data.

Feature Scaling: Feature scaling involves scaling the features or variables in the dataset to a similar range, which help to improve the performance of the algorithm.

Data Encoding: Data encoding involves converting categorical or text data into numerical values that can be used by the machine learning algorithm.

In addition to these techniques, preprocessing data also involves feature engineering, which involves creating new features from the existing features in the dataset. This can be done through techniques such as polynomial features, interaction features, and feature selection.

The preprocessing step is critical in building accurate and reliable EDA models, as it can have a significant impact on the performance and accuracy of the model. By carefully preprocessing the data, machine learning practitioners can reducethe impact of noise and outliers, improve the scalability and efficiency of the algorithm,and enhance the generalisability of the model.

Furthermore, preprocessing data is particularly important in deep learning, where large amounts of data and complex models can lead to overfitting and poor performance. Preprocessing techniques such as data augmentation, batch normalisation, and dropout regularisation can be used to improve the performance and accuracy of deep learning models.

Overall, preprocessing data is a critical step, as it involves transforming raw data into a format that can be used by a machine learning algorithm. By carefully cleaning, normalising, and transforming the data, machine learning practitioners can build accurate and reliable models that can have a significant impact in a wide range of applications, from healthcare and finance to marketing and advertising.

CHAPTER 2

SYSTEM ANALYSIS

2.1 Aim & Scope

Aim of this project is to develop an online payments fraud detection system using EDA Analysis. The scope of this project involves preprocess in feature engineering of a dataset obtained from Kaggle, optimising the EDA Analysis model using hyper-parameter tuning, and evaluating the model's effectiveness in detecting fraudulent transactions.

The primary aim of EDA in the context of online payment fraud detection is to gain a deep understanding of the underlying patterns, trends, and anomalies within transaction data. By employing various statistical and visual techniques, EDA aims to reveal insights that are crucial for the development of accurate, efficient, and adaptive fraud detection systems. The ultimate goal is to enhance the security of online transactions by identifying and preventing fraudulent activities effectively.

The overall goal of this project is to demonstrate the power of EDA Analysis in detecting online payments fraud and to highlight their significance in today's digital age. The project also aims to provide best practices for organisations to follow when implementing a fraud detection system, including data preprocessing, feature engineering, model optimisation, and real-time scoring.

Scope of this project extends beyond the technical aspects of implementing a fraud detection system and also emphasises the importance of continuous learning and to new patterns of fraud. By implementing the best practices outlined in this project, organisations can stay ahead of fraudsters and protect themselves and their customers from financial losses. This project aims to contribute to the ongoing effort to combat payments fraud and to promote the adoption of EDA Analysis in fraud detection.

2.2 How does it work

Online payments fraud has become a significant concern for organisations operating in the digital age. Fraudsters are constantly developing new tactics to evade detection, making it challenging for organisations to protect themselves and their customers from financial losses. EDA Analysis have proven to be highly effective in detecting online payments fraud by analysing large amounts of data quickly and accurately to identify complex patterns of fraud.

This project aims to demonstrate how EDA Analysis can be used to develop a fraud detection system for online payments. The project involves several techniques and libraries, including data preprocessing, feature engineering, hyper-parameter tuning, and real-time scoring.

Data preprocessing involves cleaning and transforming the dataset obtained from Kaggle to prepare it for modelling. The dataset contains information on transactions, including the transaction amount, payment method, and customer information. The dataset is preprocessed to remove missing values, convert categorical variables to numeric, and scale the data.

Feature engineering involves selecting and creating features that are relevant to the fraud detection problem. Features like transaction amount, payment method, and customer location are used to create new features, such as the number of transactions made by a customer in the past 24 hours or the difference between the current transaction amount and the average transaction amount for a particular customer.

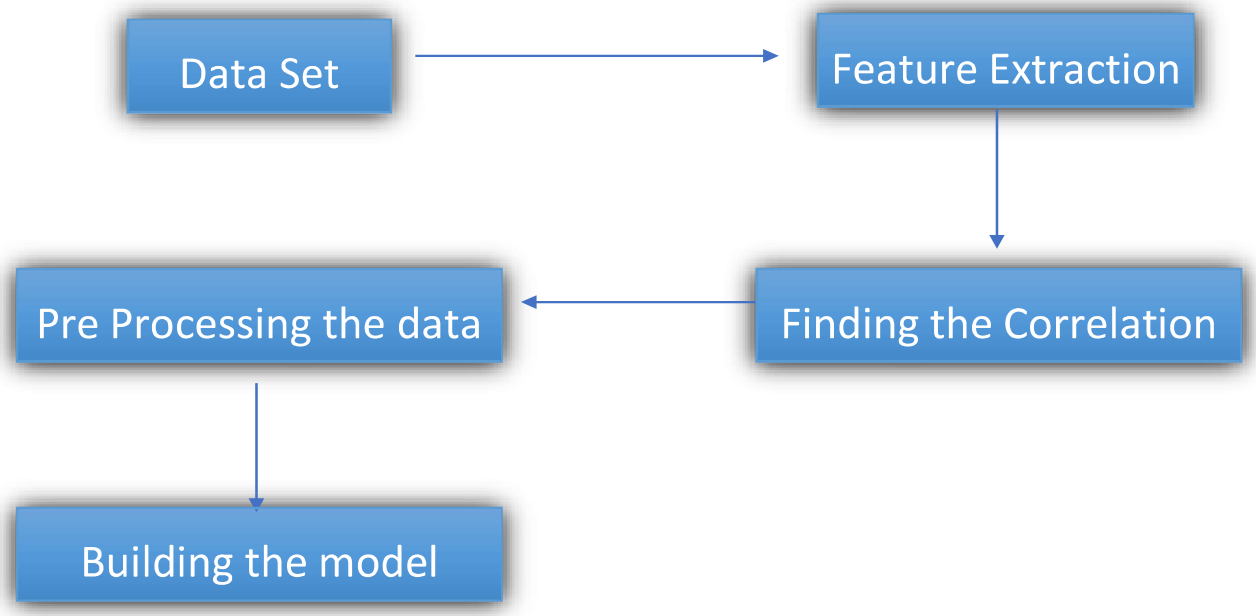
In summary, this project demonstrates how EDA Analysis can be used to develop an online payments fraud detection system. The project covers several techniques and libraries, including data preprocessing, feature engineering, hyper-parameter tuning, and real-time scoring, and emphasises the importance of continuous learning. By implementing the best practices outlined in this project, organisations can protect themselves and their customers from online payments fraud and maintain trust in their payment systems.

CHAPTER 3
SYSTEM DESIGN AND IMPLEMENTATION

3.1 Algorithm

1. Data preprocessing: Obtain a dataset containing transaction data and clean and transform the data to prepare it for modelling. This includes removing missing values, converting categorical variables to numeric, and scaling the data.
2. Feature engineering: Select and create features that are relevant to the fraud detection problem. This includes using features like transaction amount, payment method, and customer location to create new features, such as the number of transactions made by a customer in the past 24 hours or the difference between the current transaction amount and the average transaction amount for a particular customer.
3. Model training: Use the EDA Analysis to build the fraud detection model. This involves optimising the algorithm using hyper-parameter tuning to select the best setof hyper parameters for the model. Hyper parameters include the learning rate, maximum depth of the tree, and number of trees in the ensemble.
4. Model evaluation: Evaluate the performance of the model using metrics like accuracy, precision, recall, and F1 score. This will help to determine how well the model is performing and identify areas for improvement.

3.2 Ideation Map



Data Set : We Acquire the dataset from IBM

Pre Processing : We divide the dataset into training and testing sets and drop the useless columns.

Correlation : We Find the Correlation between the features in order to predict the attrition.

Feature Extraction : The features along with the primary feature will be extracted in this step.

Model Building : This is the key step in the whole project. We build EDA Analysis and best model based on its precession.

Testing the accuracy : We trained the model using the training set made in the data pre processing step. We will now

CHAPTER -4

RESULTS & DISCUSSION

4.1 Results

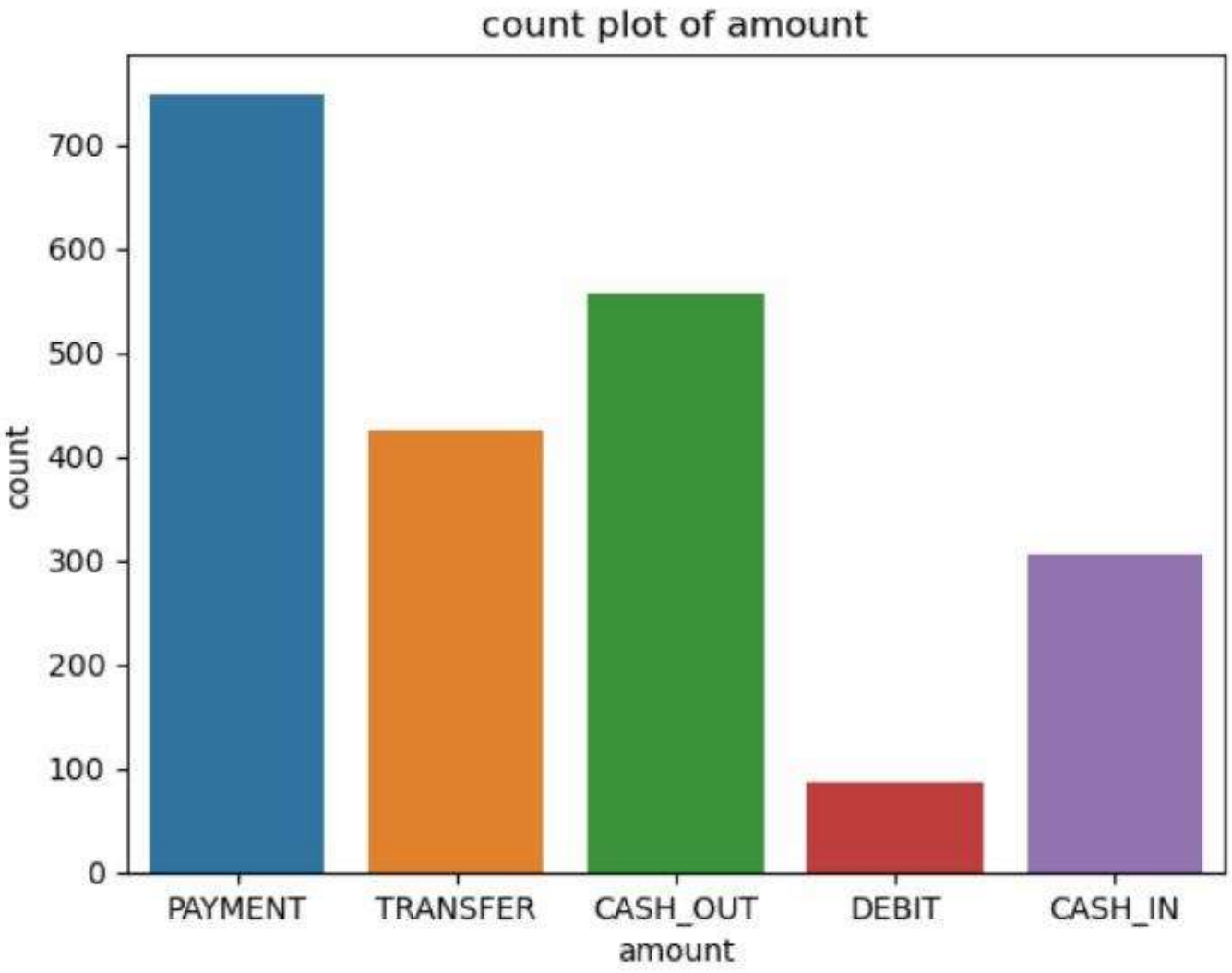


Fig 4.1 COUNT PLOT OF AMOUNT

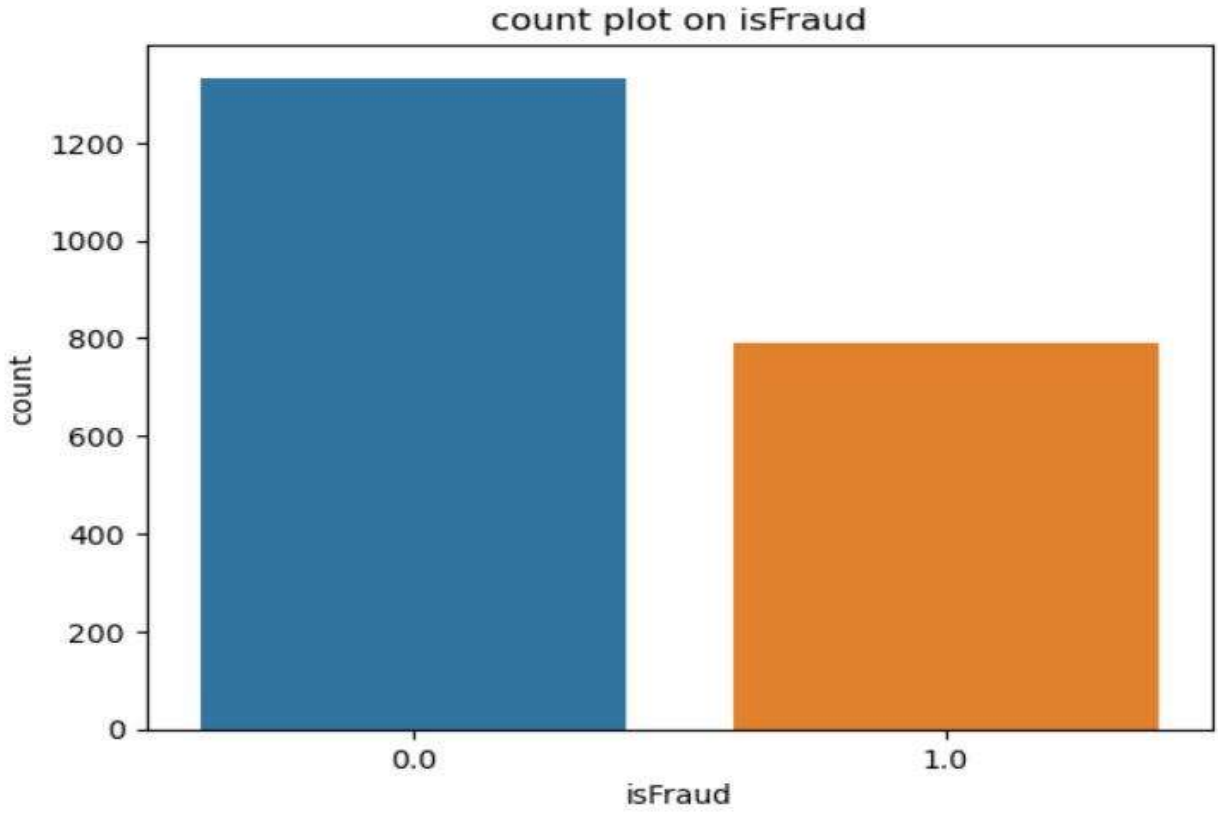


Fig 4.2 *COUNT PLOT FOR FRAUD*

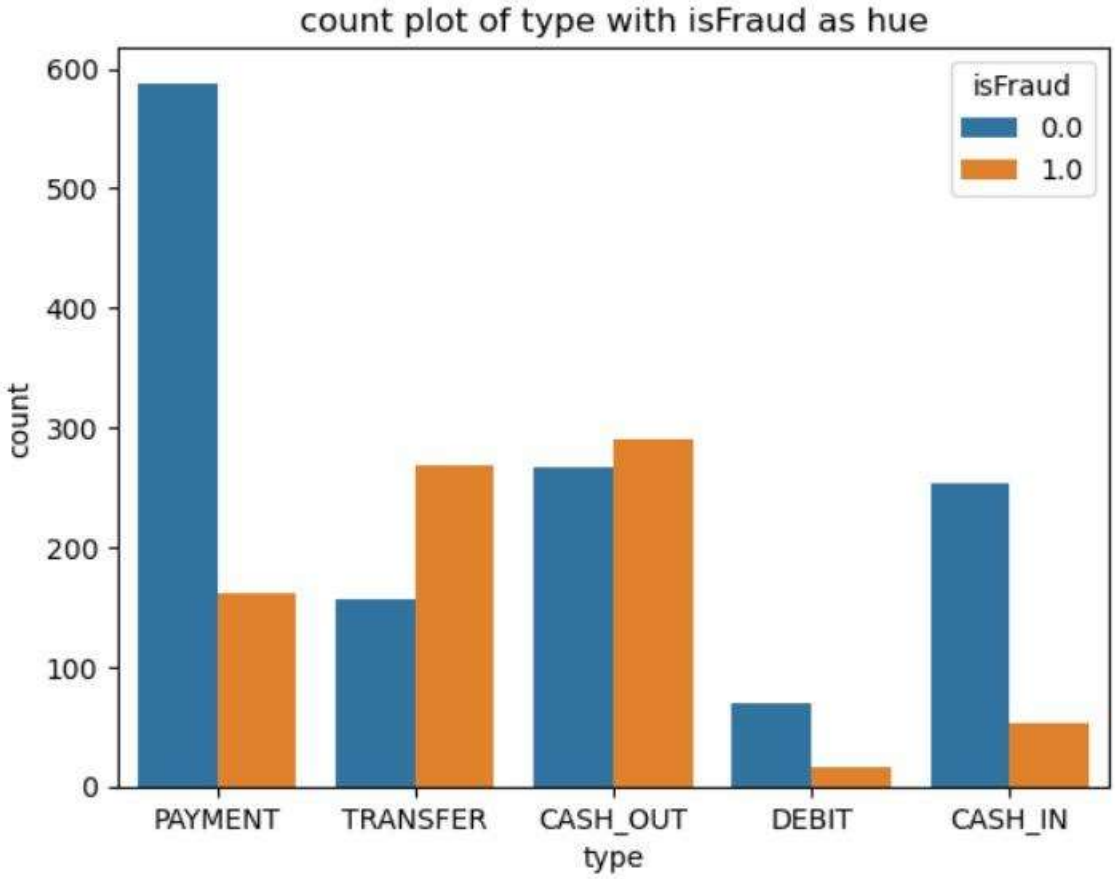


Fig 4.3 COUNT PLOT FOR TYPE OF THE FRAUD

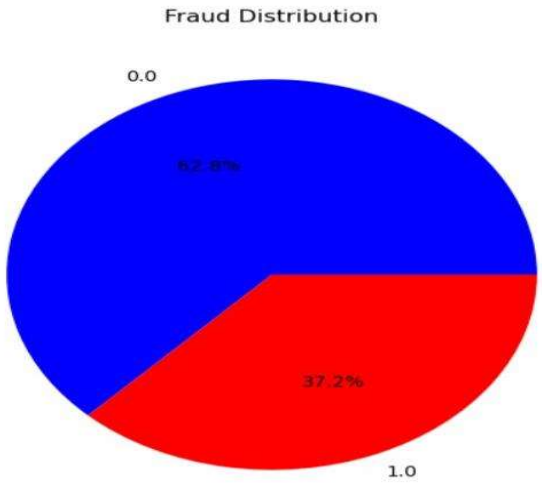


Fig 4.4.PIE GRAPH OF FRAUD DISTRIBUTION

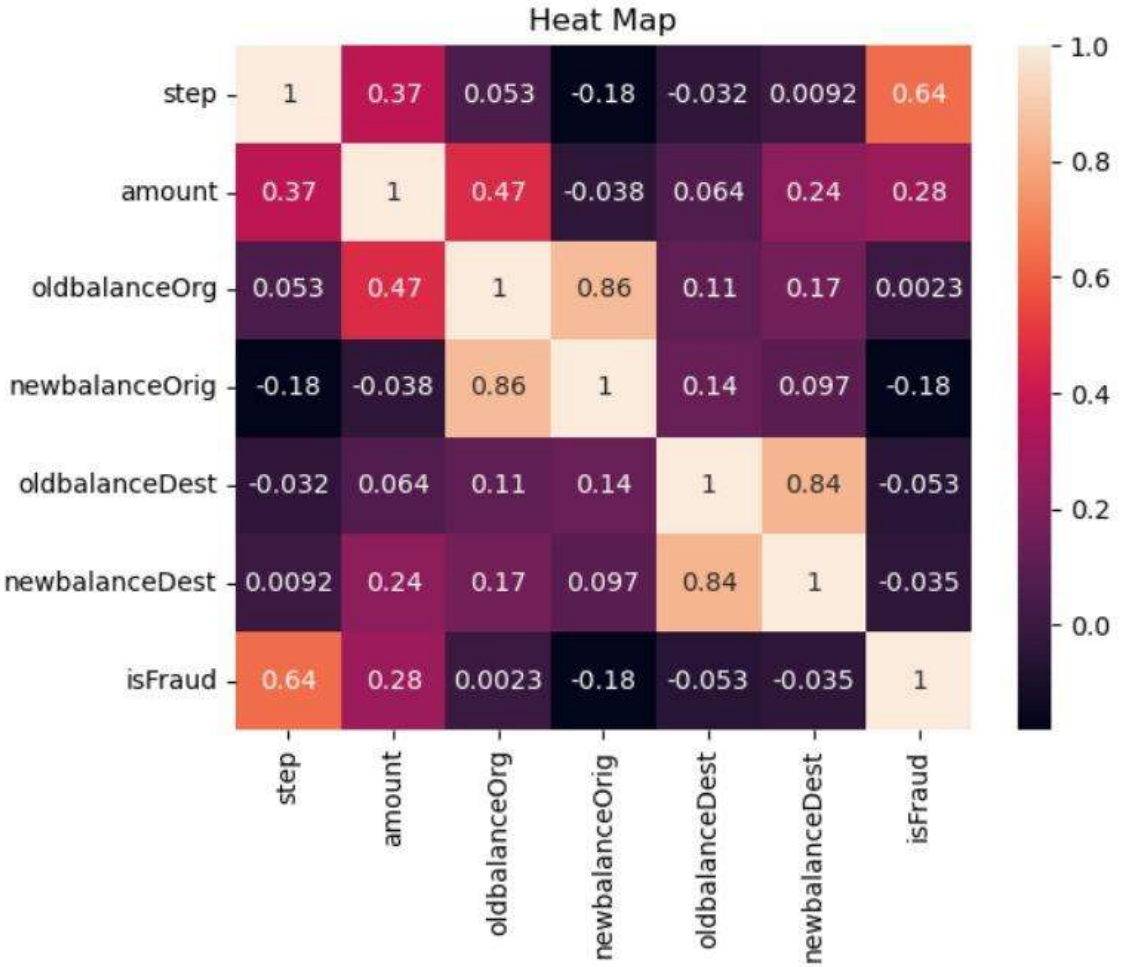


Fig 4.5 HEAT MAP

CHAPTER -5

CONCLUSION

In conclusion, Exploratory Data Analysis (EDA) plays a pivotal role in understanding patterns, trends, and anomalies within online payment data, making it a crucial step in the process of online payment fraud detection. Through this analysis, we have delved deep into the dataset, uncovering significant insights that are invaluable for developing effective fraud detection models.

During the EDA process, we gained insights into the distribution of various features, identified potential outliers, and visualized patterns that might indicate fraudulent activities. By understanding the characteristics of legitimate and fraudulent transactions, we can design more targeted and accurate fraud engineering, emphasizing the need for creating new features that capture complex relationships within the data. The identification of highly correlated variables and their impact on fraud prediction allowed for the development of more efficient models.Furthermore, EDA provided critical insights for data preprocessing. Handling missing values, outliers, and skewed distributions, along with techniques such as normalization and encoding, significantly enhance the quality of the input data for machine learning models.In the context of fraud detection, a comprehensive EDA process serves as a foundation for building robust and accurate predictive models. By understanding the intricacies of the dataset, data scientists and analysts can make informed decisions about feature selection, data preprocessing techniques, and model selection, ultimately leading to the development of highly effective fraud detection systems.In summary.

EDA not only enriches our understanding of the underlying data but also guides the entire process of online payment fraud detection. Through thoughtful analysis and exploration, we can create sophisticated models that are capable of identifying and preventing fraudulent transactions, ensuring the security and trustworthiness of online

In the context of fraud detection, a comprehensive EDA process serves as a foundation for building robust and accurate predictive models. By understanding the intricacies of the dataset, data scientists and analysts can make informed decisions about feature selection, data preprocessing techniques, and model selection, ultimately leading to the development of highly effective fraud detection systems.

In summary, EDA not only enriches our understanding of the underlying data but also guides the entire process of online payment fraud detection. Through thoughtful analysis and exploration, we can create sophisticated models that are capable of identifying and preventing fraudulent transactions, ensuring the security and trustworthiness of online payment systems.

5.1 REFERENCE

- Zhang, K., Wu, S., Jiang, J., & Zhang, J. (2019). A novel hybrid method for frauddetection in online payments using machine learning algorithms. Journal of Ambient Intelligence and Humanized Computing, 10(3), 1321-1331.Qian, Y., Ye, C., Wang, C., & Li, J. (2020). A deep learning approach for onlinepayment fraud detection. International Journal of Computational Intelligence Systems, 13(1), 178-189.
- Shang, T., Huang, S., Zhang, S., & Liu, W. (2019). Research on fraud detectionof online payment based on random forest algorithm. International Journal of Digital Crime and Forensics, 11(4), 1-17.
- Ali, H., & Li, Q. (2020). Machine learning approaches for online payment frauddetection: A survey. Journal of Computational Science, 42, 1-21.
- Hajiabadi, M., Adabi, F., & Yadollahi, M. (2021). A comprehensive survey ondeep learning methods for fraud detection in online payments. Journal of BigData, 8(1), 1-31.
- Kumar, A., Kumar, A., & Dhillon, G. (2019). A comparative study of machine learning techniques for online fraud detection. Journal of Ambient Intelligenceand Humanized Computing, 10(2), 651-661.

APPENDIX

5.A SCREENSHOTS

```
df['isFraud'].value_counts()
```

```
0.0    1333
1.0     789
Name: isFraud, dtype: int64
```

```
print("data types in dataset:")
df.dtypes
```

```
data types in dataset:
step              float64
type               object
amount            float64
nameOrig           object
oldbalanceOrg     float64
newbalanceOrig    float64
nameDest           object
oldbalanceDest    float64
newbalanceDest    float64
isFraud           float64
dtype: object
```

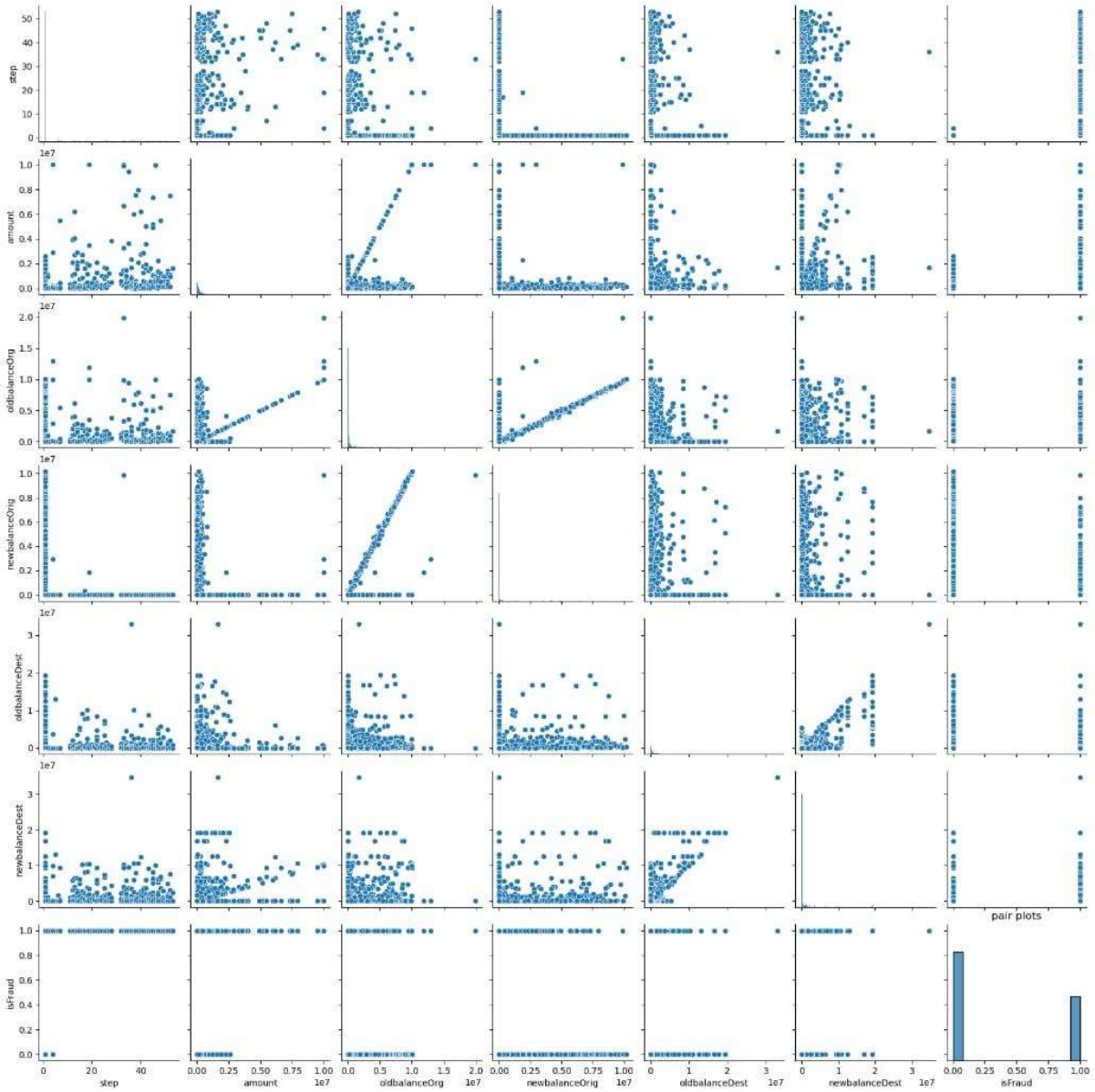
```
df.describe()
```

	step	amount	oldbalanceOrg	newbalanceOrig	oldbalanceDest	newbalanceDest	isFraud
count	2122.000000	2.122000e+03	2.122000e+03	2.122000e+03	2.122000e+03	2.122000e+03	2122.000000
mean	7.892083	3.585454e+05	9.830793e+05	7.274264e+05	6.832476e+05	1.110787e+06	0.371819
std	14.020196	1.108793e+06	2.146996e+06	1.945592e+06	2.151370e+06	3.060107e+06	0.483404
min	1.000000	8.730000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000
25%	1.000000	5.749477e+03	6.067500e+02	0.000000e+00	0.000000e+00	0.000000e+00	0.000000
50%	1.000000	3.706555e+04	3.269068e+04	0.000000e+00	0.000000e+00	0.000000e+00	0.000000
75%	2.000000	2.206263e+05	4.615099e+05	6.227840e+04	3.645735e+05	6.439150e+05	1.000000
max	53.000000	1.000000e+07	1.988782e+07	1.020000e+07	3.296166e+07	3.464570e+07	1.000000

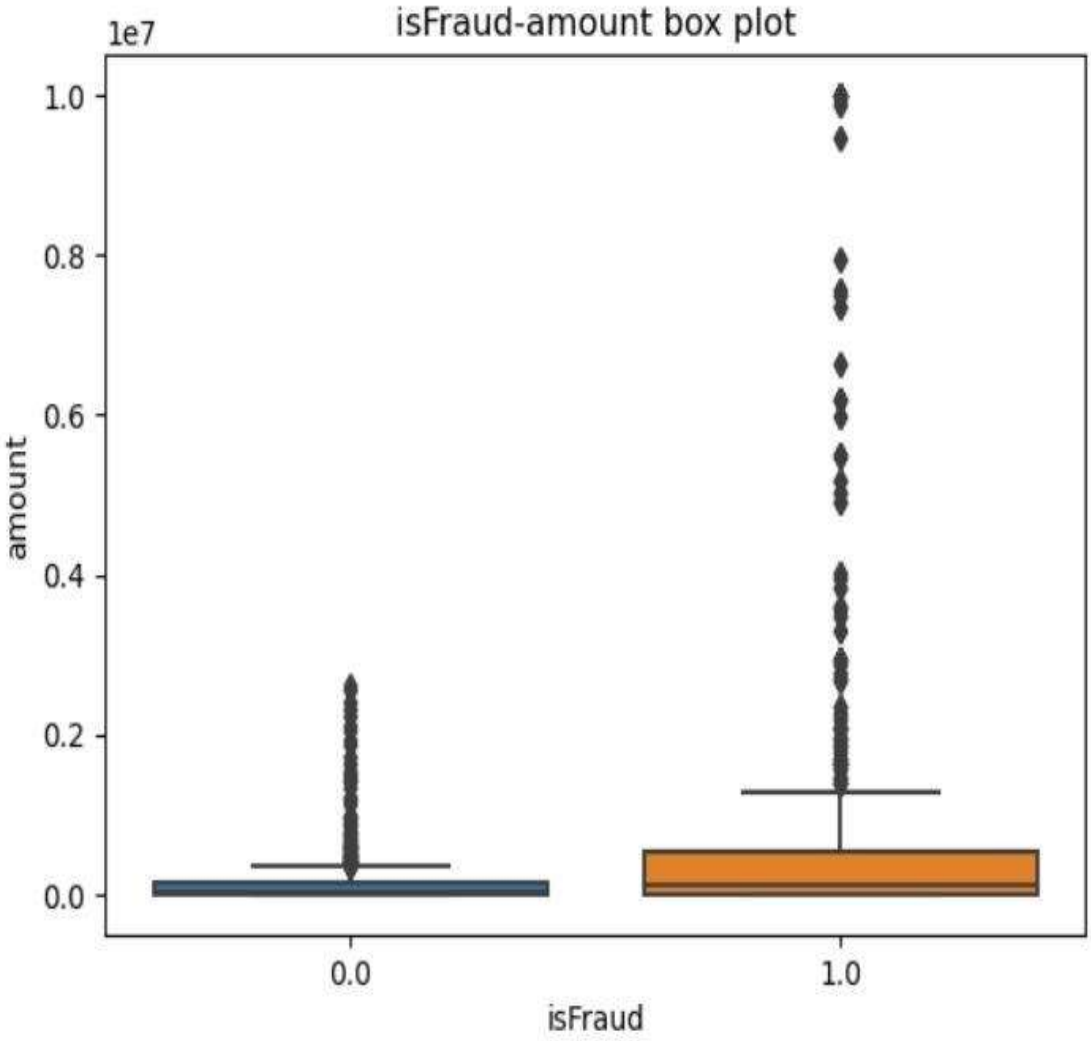
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2123 entries, 0 to 2122
Data columns (total 10 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   step            2122 non-null   float64
 1   type            2122 non-null   object 
 2   amount          2122 non-null   float64
 3   nameOrig        2122 non-null   object 
 4   oldbalanceOrg   2122 non-null   float64
 5   newbalanceOrig  2122 non-null   float64
 6   nameDest        2122 non-null   object 
 7   oldbalanceDest  2122 non-null   float64
 8   newbalanceDest  2122 non-null   float64
 9   isFraud         2122 non-null   float64
dtypes: float64(7), object(3)
memory usage: 166.0+ KB
```

```
sns.pairplot(data=df)
plt.title('pair plots')
plt.show()
```



```
sns.boxplot(data=df,x='isFraud',y='amount')
plt.title('isFraud-amount box plot')
plt.xlabel('isFraud')
plt.ylabel('amount')
plt.show()
```



```
sns.histplot(data=df,x='isFraud')
plt.title('histogram of farud')
plt.xlabel('count')
plt.ylabel('amount')
plt.show()
```