CREDIT CARD FRAUD DETECTION

PHASE 4 PROJECT SUBMISSION

TOPIC: START BUILDING THE CREDIT CARD DETECTION PROJECT BY LOADING AND PRE-PROCESSING THE DATASET

Introduction:

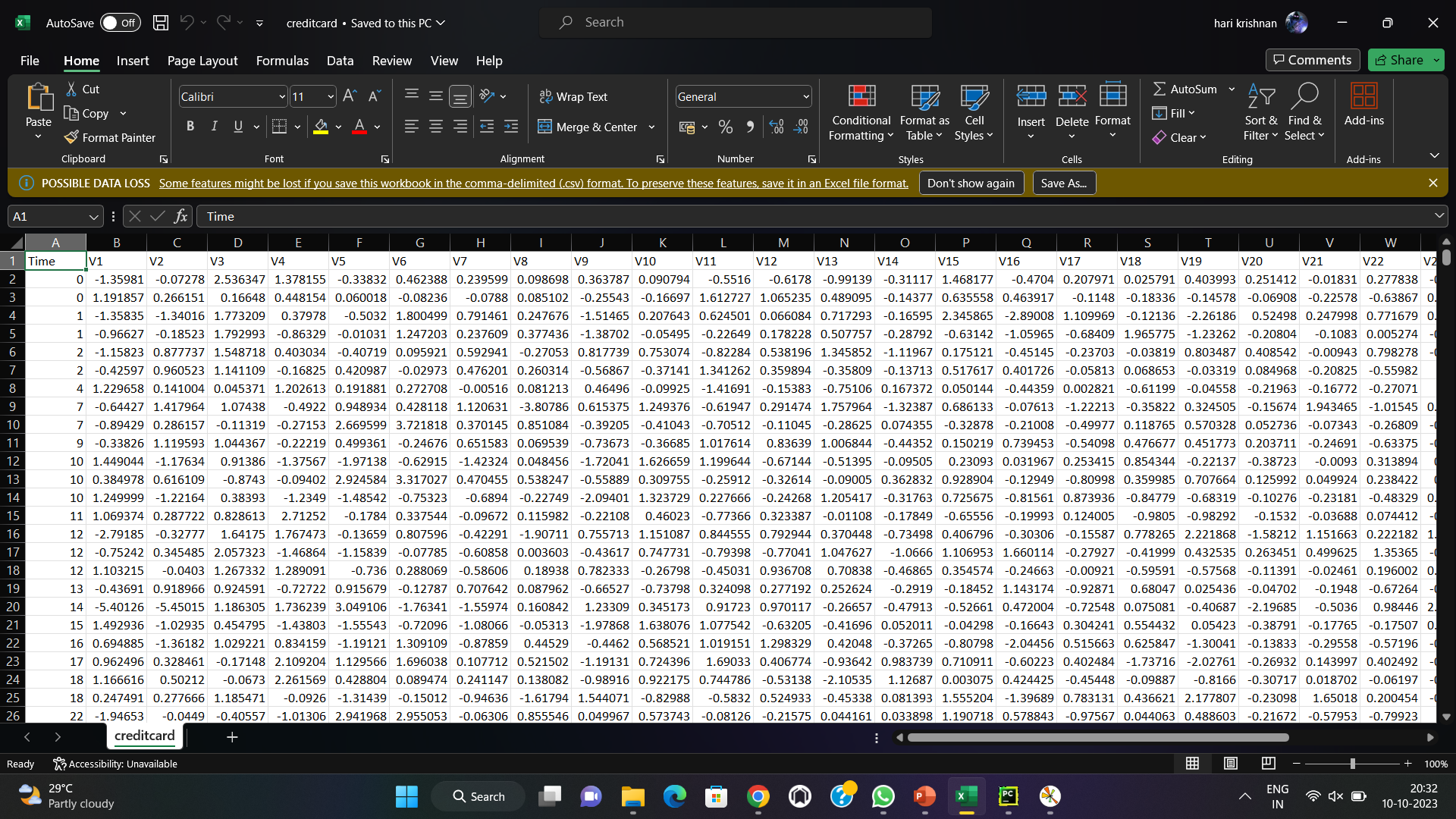
* **Data Collection**: Financial institutions collect a vast amount of data related to card transactions, including transaction amount, location, date and time, merchant information, and more. This data forms the basis for fraud detection.
* **Machine Learning and Artificial Intelligence**: Machine learning algorithms and AI models play a pivotal role in analyzing the data. They are trained to identify patterns, anomalies, and trends associated with legitimate and fraudulent transactions. Common algorithms include decision trees, neural networks, logistic regression, and random forests.
* **Rules-Based Systems**: In addition to machine learning, rules-based systems are used to set predefined rules that can trigger alerts or actions when certain conditions are met. These rules may include transaction amount limits, transaction frequency thresholds, and location-based checks.
* **Behavioral Analysis**: By monitoring a cardholder's behavior over time, financial institutions can establish a baseline of normal activity. Deviations from this baseline, such as an unusual transaction in a different location or an unusually large purchase, can raise red flags.
* **Real-time Monitoring**: Many systems operate in real-time, continuously assessing incoming transactions as they occur. This allows for immediate responses to potential fraud, such as flagging a transaction for manual review or blocking a card.
* **Historical Data Analysis**: Historical transaction data is invaluable for identifying recurring fraud patterns. By analyzing past incidents, financial institutions can refine their fraud detection models and rules.
* **Machine Learning Model Training**: Regular updates and retraining of machine learning models are essential to adapt to evolving fraud tactics. New data helps models become more accurate in recognizing fraudulent behavior.
* **Customer Verification**: Multi-factor authentication, including something the cardholder knows (e.g., PIN), something the cardholder has (e.g., the card itself), and something the cardholder is (e.g., biometrics), can provide an extra layer of security.
* **Collaboration**: Financial institutions often collaborate with industry organizations and share fraud intelligence to stay ahead of emerging threats and trends.
* **Customer Alerts**: Providing cardholders with alerts for suspicious activity via text messages, emails, or mobile apps allows them to quickly report unauthorized transactions.

Dataset:

The dataset for the given credit card fraud detection project can be downloaded from the link given below.

<https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/code>

The given dataset is picturized as below:



Loading and pre-processing of dataset:

Loading and pre-processing the dataset is an important first step in building up any machine learning model. However ,it is especially important for credit card fraud detection dataset are complex and noisy.

1.LOADING THE DATASET:

* To load a dataset for credit card fraud detection, you can follow these steps:
* as Pandas, **Choose a Dataset**: There are several publicly available datasets for credit card fraud detection, which you can use for research or model development. One of the most commonly used datasets is the Credit Card Fraud Detection dataset available on Kaggle. You can download this dataset or use other reputable sources like the UCI Machine Learning Repository.
* **Import Python Libraries**: You'll typically use Python for data analysis and machine learning. Import libraries such NumPy, and Scikit-Learn to manipulate and work with the dataset.

PROGRAM:

Since we are given two datasets, we are going to load both of these datasets separately.

import pandas as pd

import numpy as np

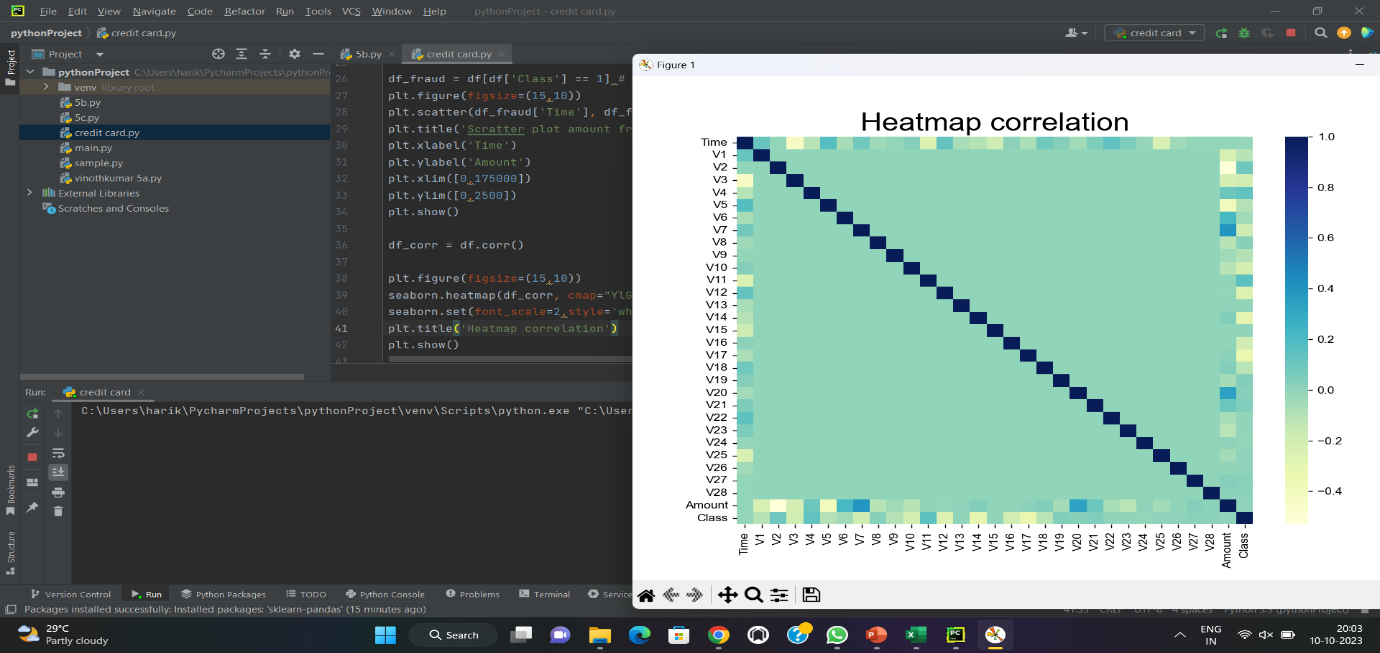
data = pd.read\_csv('filename.csv')

from sklearn.model\_selection import train\_test\_split

X = data.drop('Class', axis=1)

y = data['Class']

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_ size=0.3, random\_state=42)

Output:

2.PRE-PROCESSING THE DATASET:

Preprocessing the data in a credit card fraud detection dataset is a crucial step to ensure that the data is ready for machine learning. The goal of preprocessing is to clean and prepare the data, making it suitable for training a model that can effectively detect fraudulent transactions. Here are the common preprocessing steps for credit card fraud detection:

import pandas as pd

import numpy as np

*# Scikit-learn library: For SVM*

from sklearn import preprocessing

from sklearn.metrics import confusion\_matrix

from sklearn import svm

df\_fraud = df[df['Class'] == 1] *# Recovery of fraud data*

plt.figure(figsize=(15,10))

plt.scatter(df\_fraud['Time'], df\_fraud['Amount']) *# Display fraud amounts according to their time*

plt.title('Scratter plot amount fraud')

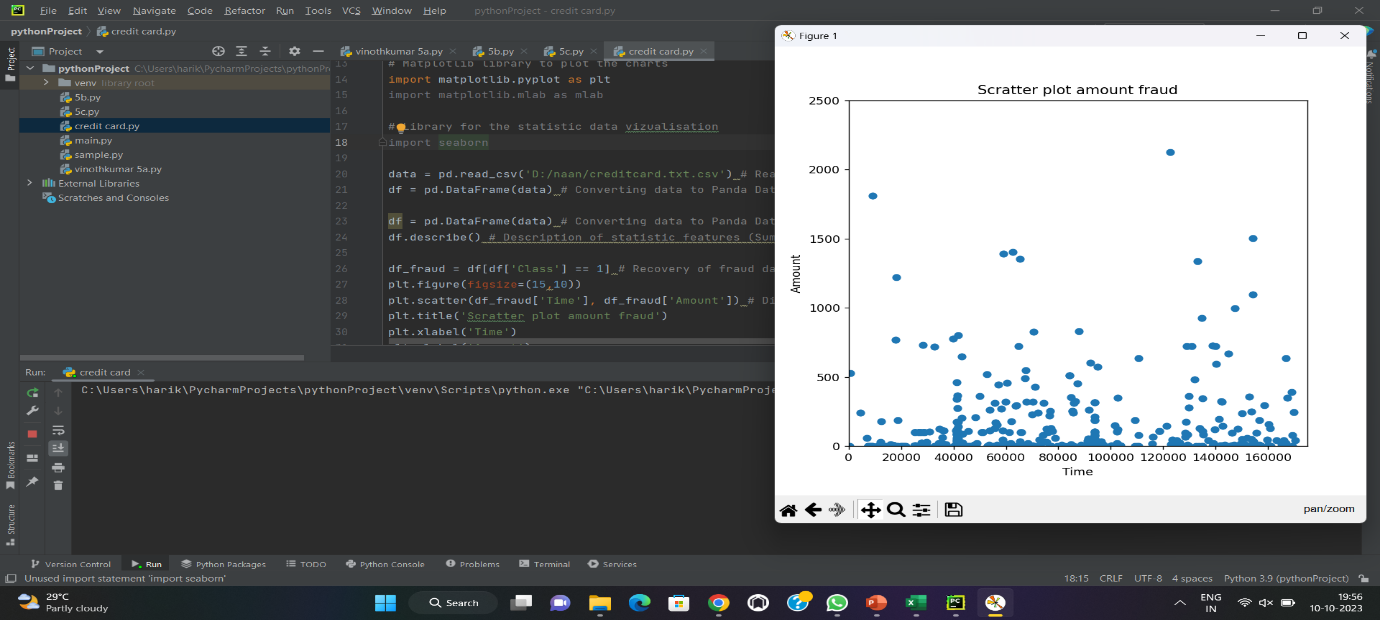
plt.xlabel('Time')

plt.ylabel('Amount')

plt.xlim([0,175000])

plt.ylim([0,2500])

plt.show()

OUTPUT: 

***Basic Exploratory Analysis:***  

Basic exploratory data analysis (EDA) is an essential initial step in data analysis that involves summarizing, visualizing, and understanding the main characteristics of your dataset.

exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

***Program:***

import pandas as pd

import matplotlib.pyplot as plt

df=pd.read\_csv("country\_vaccinations.csv",encoding="unicode\_escape")

print(df.head())

print(df.describe())

print(df['people\_vaccinated'].value\_counts())

df.hist()

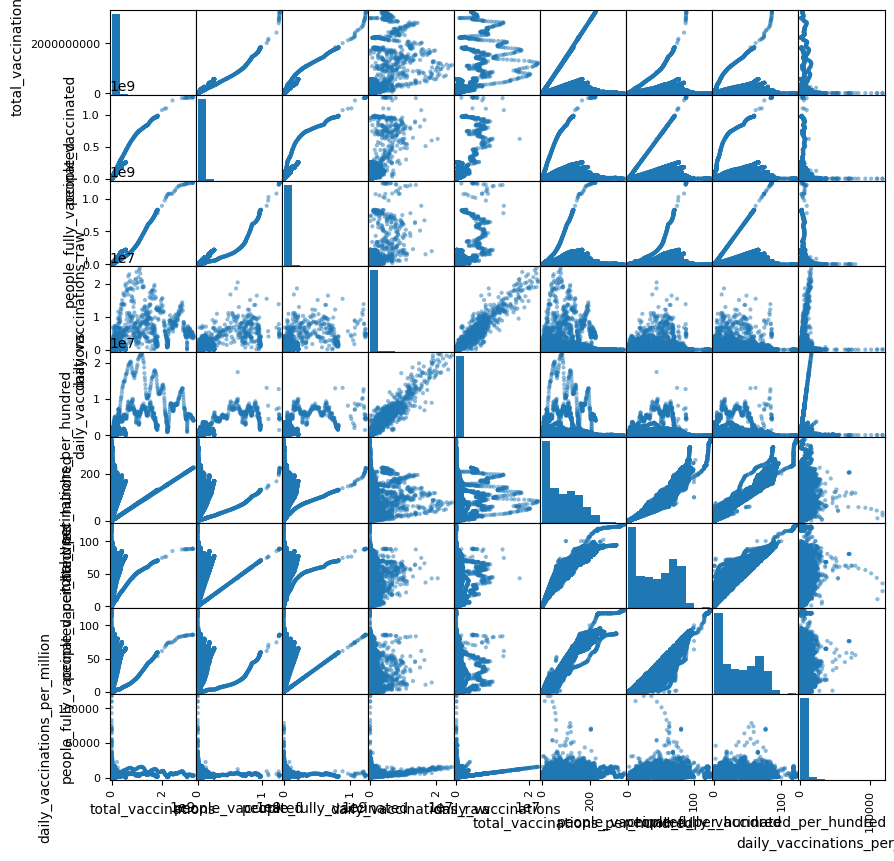
plt.show()

from pandas.plotting import scatter\_matrix

scatter\_matrix(df, figsize=(10, 10))

plt.show()

***Output:***



**Statistical Analysis:**

      Statistical analysis is a systematic approach to understanding data through the application of mathematical and statistical techniques. It plays a crucial role in making sense of complex information, identifying patterns, and drawing meaningful insights. The process typically begins with data collection, followed by data cleaning and preprocessing to ensure data quality. Descriptive statistics are employed to provide an initial summary of the dataset, revealing central tendencies and variability. Inferential statistics, on the other hand, are used to make predictions and test hypotheses about the population from which the data was collected. This branch of analysis encompasses a wide array of methods, including hypothesis testing, regression analysis, and analysis of variance, among others. Statistical analysis is a cornerstone in fields ranging from science and business to healthcare and social sciences, aiding in decision-making, problem-solving, and evidence-based reasoning.

***Program:***

 import pandas as pd

import numpy as np

from scipy import stats

df = pd.read\_csv("merged\_dataset.csv")

df=df.tail(10)

summary = df.describe()

group1\_data = df['total\_vaccinations\_x']

group2\_data = df['people\_vaccinated']

t\_statistic, p\_value = stats.ttest\_ind(group1\_data, group2\_data)

correlation\_coefficient = df['people\_fully\_vaccinated'].corr(df['daily\_vaccinations\_raw'])

print("Descriptive Statistics:")

print(summary)

print("\nT-Test Results:")

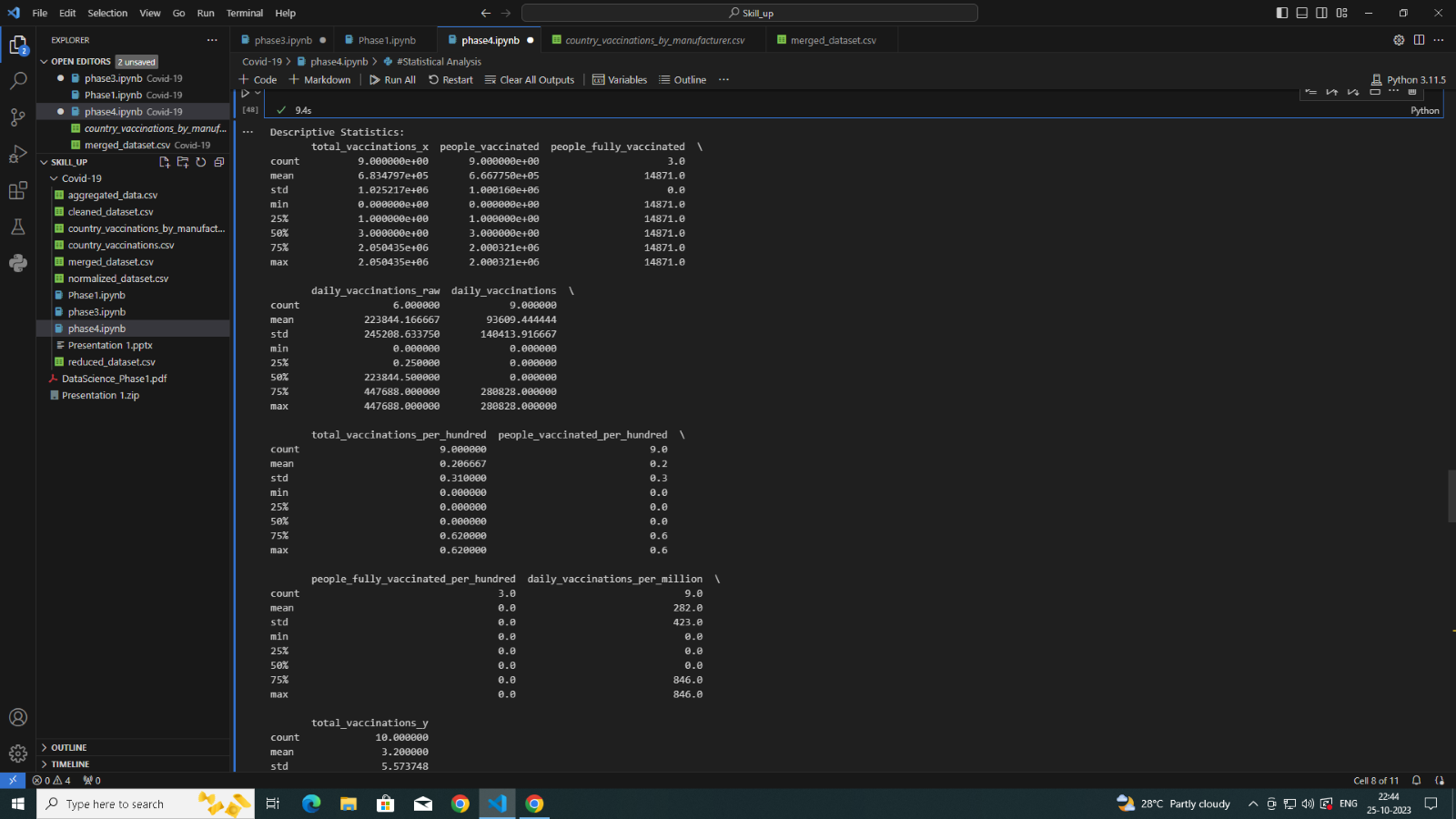
print(f"T-Statistic: {t\_statistic}")

print(f"P-Value: {p\_value}")

print("\nPearson Correlation Coefficient:")

print(correlation\_coefficient)

***Output:***



***Data Visualization:***

Data visualization is a powerful method of representing complex data in a visual and understandable format. It involves the creation of charts, graphs, and diagrams to illustrate patterns, trends, and relationships within the data. By presenting information visually, data visualization makes it easier for individuals to comprehend and interpret large datasets. Data visualizations can take many forms, from simple bar charts and pie graphs to intricate heat maps and interactive dashboards. They are particularly useful for identifying outliers, correlations, and data distributions, making data more accessible and actionable. In an era of data abundance, data visualization has become an indispensable tool for turning data into knowledge and communicating findings effectively to both experts and non-experts.

***Program:***

import pandas as pd

import matplotlib.pyplot as plt

df = pd.read\_csv("country\_vaccinations.csv")

df=df.tail(50)

categories = df["date"]

values = df["total\_vaccinations"]

plt.figure(figsize=(10, 6))

plt.bar(categories, values)

plt.xlabel("Categories")

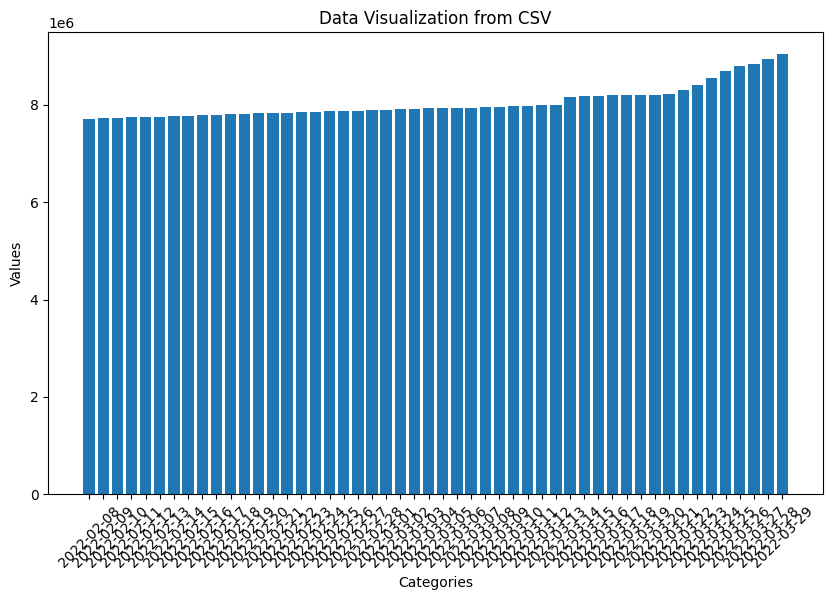
plt.ylabel("Values")

plt.title("Data Visualization from CSV")

plt.xticks(rotation=45)

plt.show()

***Output:***



***Program2:***

import matplotlib.pyplot as plt

from mpl\_toolkits.mplot3d import Axes3D

from sklearn.datasets import load\_iris

df = pd.read\_csv("country\_vaccinations\_by\_manufacturer.csv")

fig = plt.figure()

ax = fig.add\_subplot(111, projection='3d')

scatter = ax.scatter(X[:, 0], X[:, 1], X[:, 2], c=y, cmap=plt.cm.Set1)

ax.set\_xlabel('total\_vaccinations')

ax.set\_ylabel('people\_vaccinated')

ax.set\_zlabel('people\_vaccinated\_per\_100')

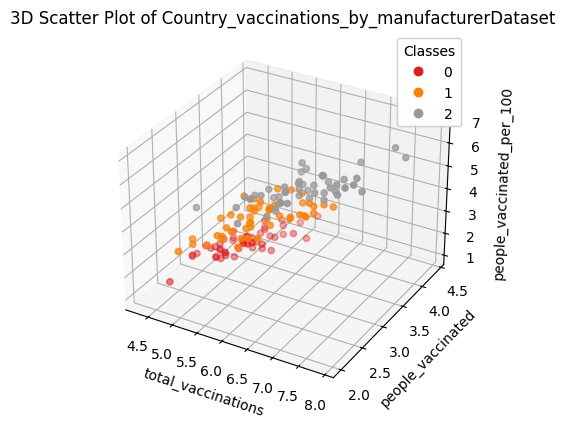
legend = ax.legend(\*scatter.legend\_elements(), title="Classes")

ax.add\_artist(legend)

ax.set\_title('3D Scatter Plot of Country\_vaccinations\_by\_manufacturerDataset')

plt.show()

***Output:***



***Program 3:***

import numpy as np

import pandas as pd

import seaborn as sns

import matplotlib as plt

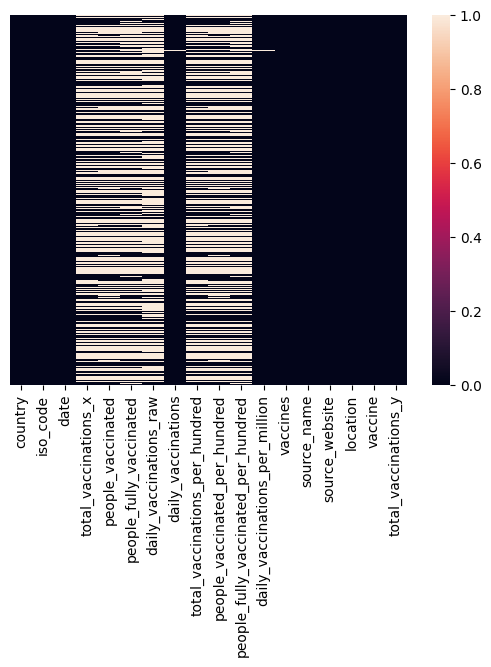
df=pd.read\_csv("merged\_dataset.csv",encoding="unicode\_escape")

missing\_values=["N/a","na",np.nan]

df=pd.read\_csv("merged\_dataset.csv",na\_values=missing\_values,encoding="unicode\_escape")

print(sns.heatmap(df.isnull(),yticklabels=False))

***Output:***



CONCLUSION:

In conclusion, credit card fraud detection is a critical aspect of financial security in today's digital age. Fraudulent activities can result in significant financial losses, damage to a financial institution's reputation, and harm to cardholders. To combat this threat effectively, credit card fraud detection systems rely on data analysis, machine learning, and various preprocessing steps to prepare the data for modeling. Here are the key takeaways:

\*\*Importance of Credit Card Fraud Detection\*\*: Credit card fraud poses a substantial risk to financial institutions and cardholders, necessitating advanced fraud detection methods.

* \*\*Data Preprocessing\*\*: Preprocessing the dataset is a vital step, involving data cleaning, handling class imbalance, scaling, feature engineering, encoding, dimensionality reduction, and splitting the data into training and testing sets.