## **EXPERIMENT NO. 4**

**Aim:** Exploratory Data Analysis and visualization of Social Media Data for business.

**Theory:**

**Exploratory Data Analysis (EDA)** is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. An EDA is a thorough examination meant to uncover the underlying structure of a data set and is important for a company because it exposes trends, patterns, and relationships that are not readily apparent.

**The four types of EDA are:**

1. Univariate non-graphical,
2. Multivariate non-graphical,
3. Univariate graphical,
4. Multivariate graphical.

**Techniques and Tools:**

There are a number of tools that are useful for EDA, but EDA is characterized more by the attitude taken than by particular techniques.

**Typical graphical techniques used in EDA are:**

* Box plot
* Histogram
* Multi-vari chart
* Run chart
* Pareto chart
* Scatter plot (2D/3D)
* Stem-and-leaf plot
* Parallel coordinates
* Odds ratio
* Heat map
* Bar chart
* Horizon graph
* Dimensionality reduction:
  + Multidimensional scaling
  + Principal component analysis (PCA)
  + Multilinear PCA
* Iconography of correlations

**Code:**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.preprocessing import LabelEncoder

data = pd.read\_csv('Stroke\_Prediction\_Indians.csv')

non\_numeric\_cols = data.select\_dtypes(include=['object']).columns

label\_encoders = {}

for col in non\_numeric\_cols:

    le = LabelEncoder()

    data[col] = le.fit\_transform(data[col])

    label\_encoders[col] = le

print(data.describe())

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

sns.histplot(data['Age'], kde=True)

plt.title('Distribution of Age')  # Add title

plt.xlabel('Age')  # Add x-axis label

plt.ylabel('Frequency')  # Add y-axis label

plt.show()

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

sns.boxplot(x='Hypertension', y='Age', data=data)  # Box plot of 'bmi' column by 'gender'

plt.title('BMI by Hypertension')  # Add title

plt.xlabel('Age')  # Add x-axis label

plt.ylabel('BMI')  # Add y-axis label

plt.show()

corr\_matrix = data.corr()

sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm')

plt.title('Correlation Matrix')

plt.show()

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

sns.scatterplot(x='Work Type', y='Age', data=data)  # Scatter plot of 'avg\_glucose\_level' vs 'age'

plt.title('Glucose Level vs Age')

plt.xlabel('Average Glucose Level')

plt.ylabel('Age')

plt.show()

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

sns.boxplot(x='Gender', y='Age', data=data)  # Box plot of 'age' column by 'stroke' status

plt.title('Age by Stroke Status')  # Add title

plt.xlabel('Stroke Status')  # Add x-axis label

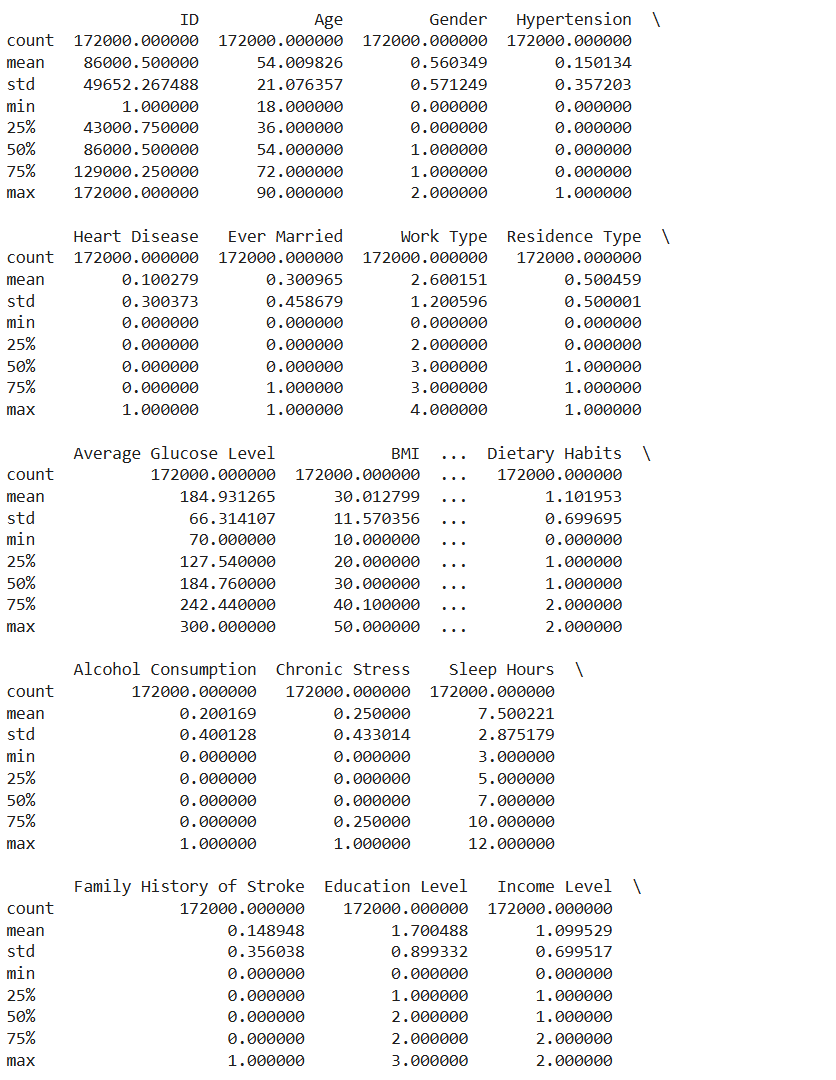
plt.ylabel('Age')  # Add y-axis label

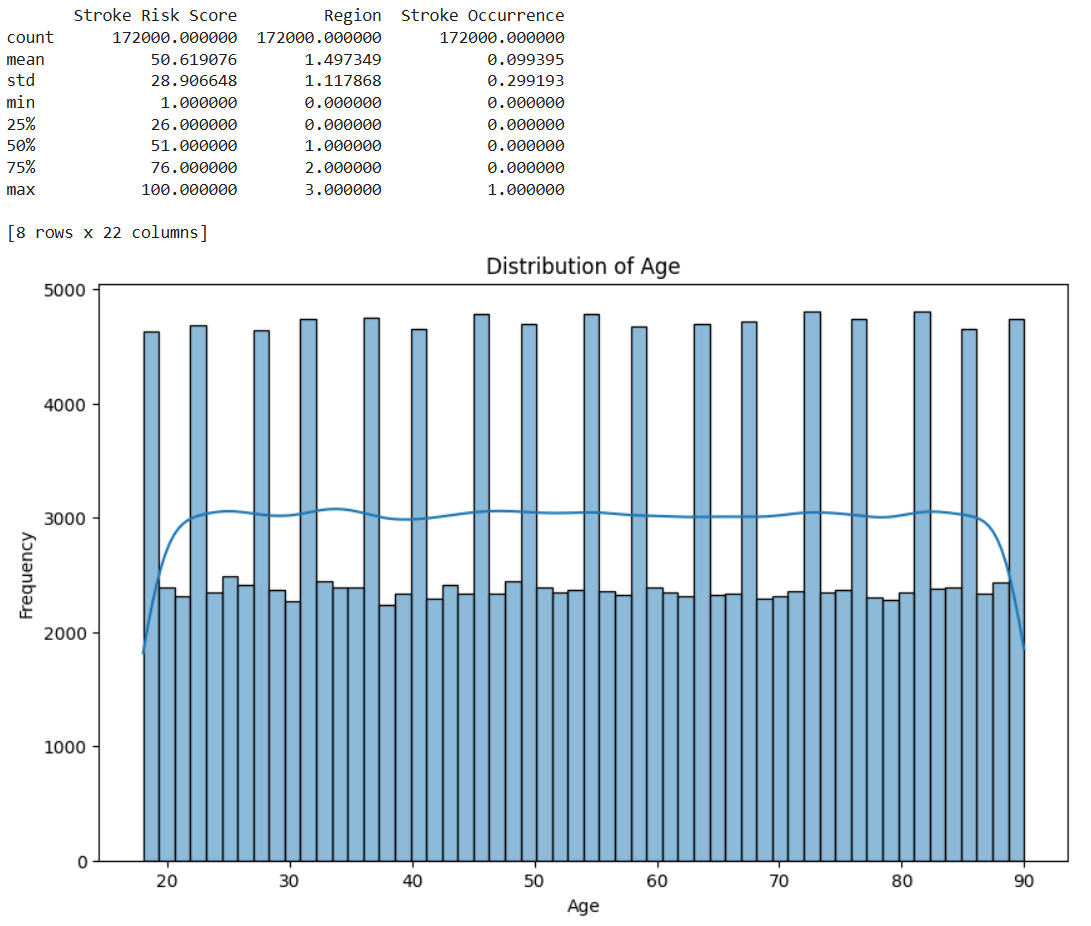
plt.show()

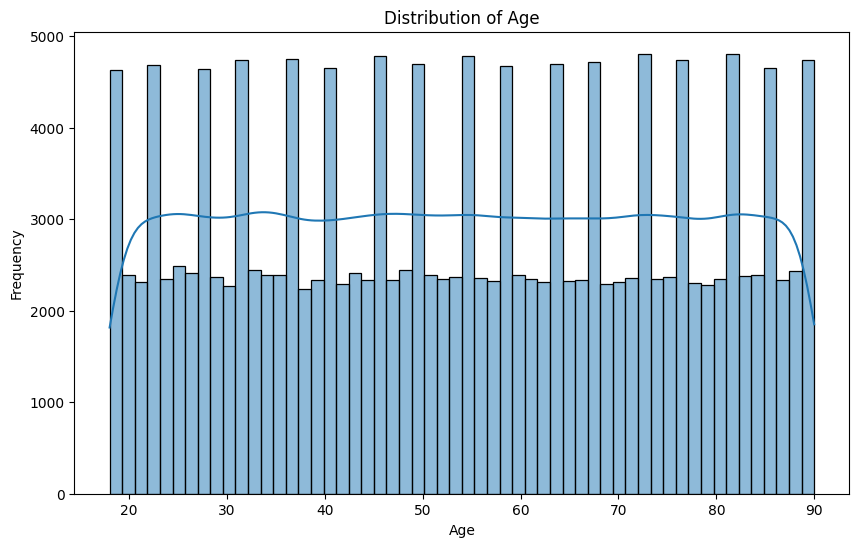
sns.pairplot(data)

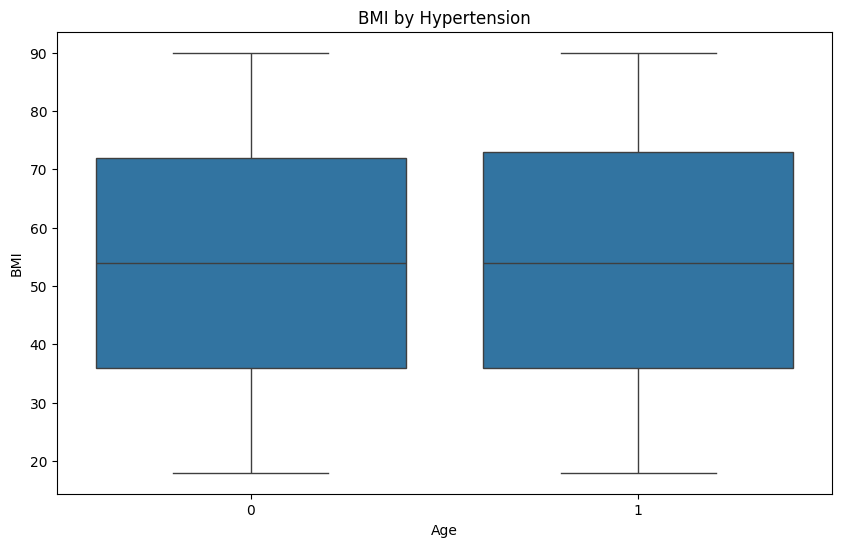
plt.show()

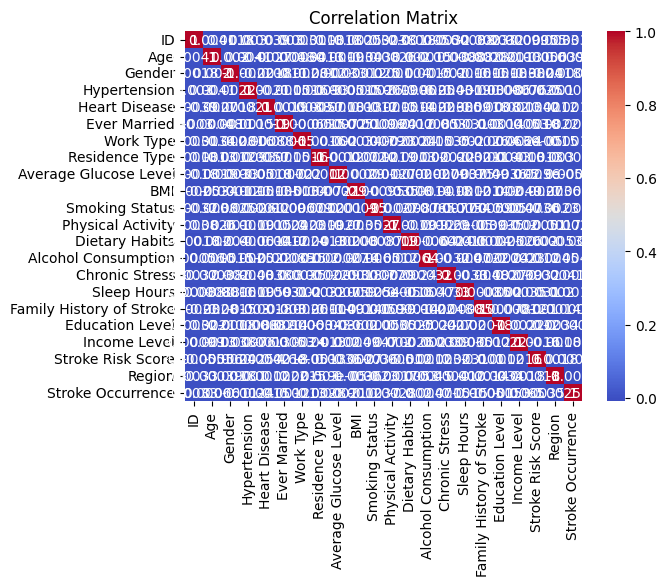
**Output:**

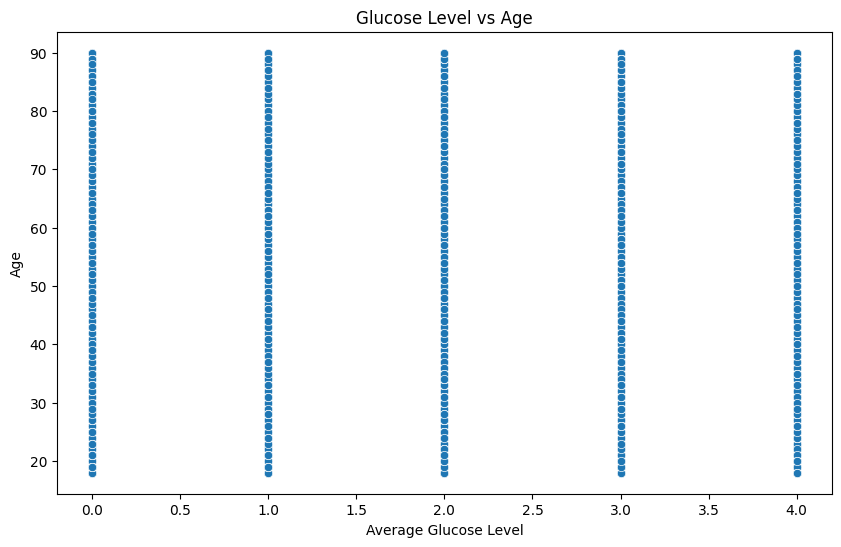


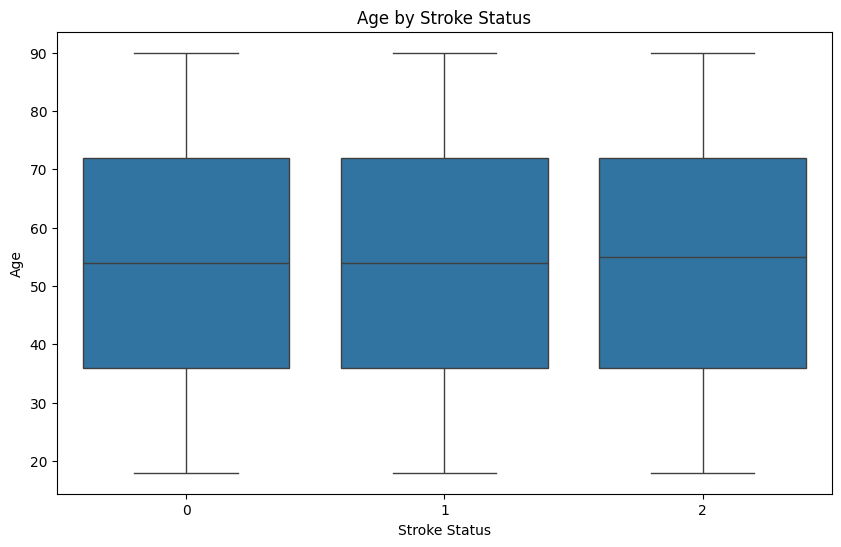


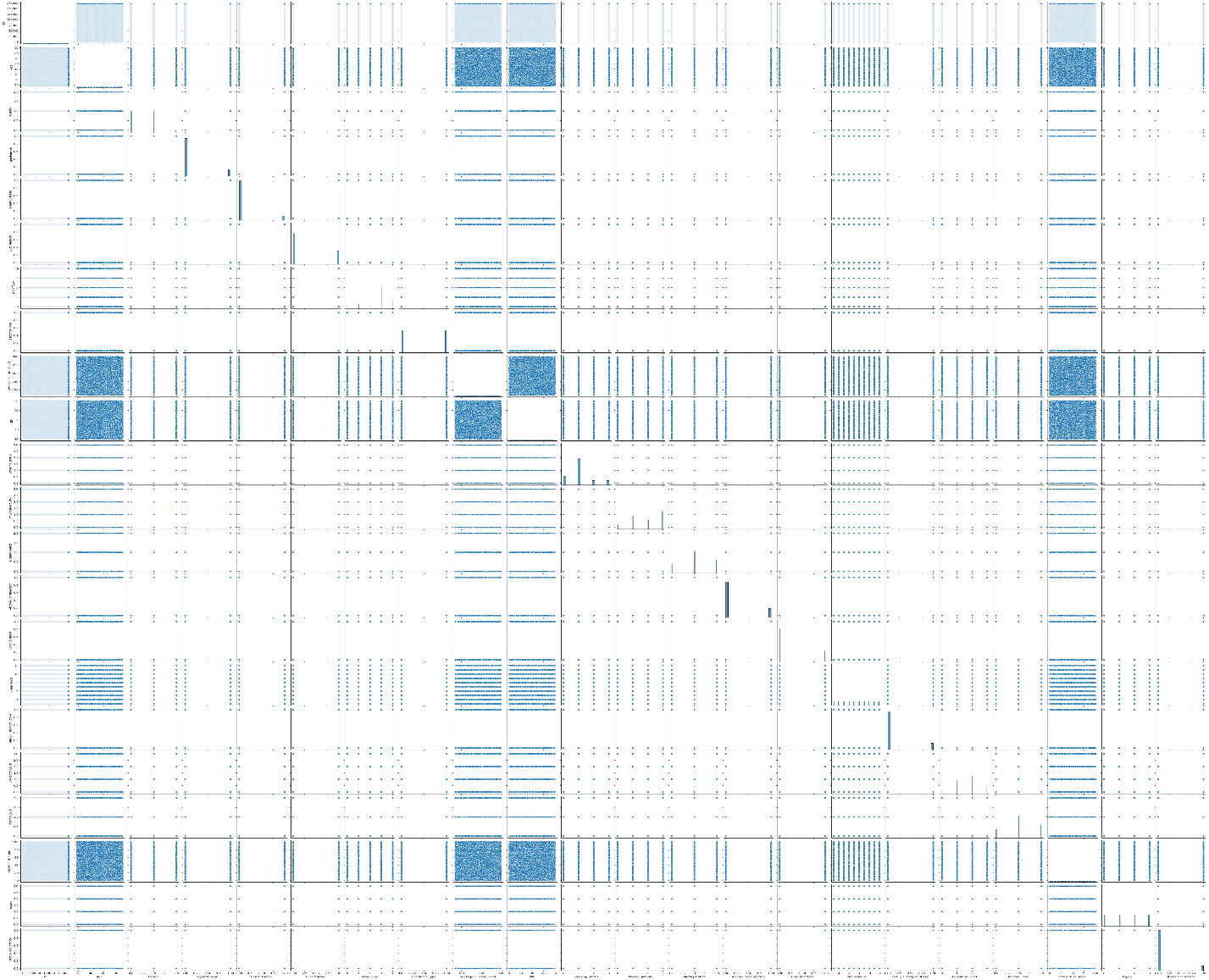












**Conclusion:**

Thus, we have successfully implemented Exploratory Data Analysis and visualization of social media data for business.