Experiment No.4
To perform exploratory data analysis and visualization of Social media data for business.
Date of Performance: 04/02/2025
Date of Submission: 11/02/2025

Aim :To perform exploratory data analysis and visualization of Social media data for business.

Objective : Uncover insights from social media data, including user behavior, sentiment, and trends, to inform strategic business decisions through visualization.

Theory :
Exploratory Data Analysis (EDA) and visualization of social media data for business involves understanding the structure, patterns, and trends within the data to derive actionable insights and inform decision-making processes.
Data Understanding: EDA begins with understanding the characteristics of the social media data. This includes examining the types of data available (text, numerical, categorical), the range of values, and the structure of the dataset (number of rows and columns).
Descriptive Statistics: Descriptive statistics such as mean, median, mode, standard deviation, and quartiles provide initial insights into the central tendency, dispersion, and distribution of numerical variables in the social media data.
Data Quality Assessment: EDA involves assessing the quality of the social media data by identifying and addressing data quality issues such as missing values, duplicates, outliers, and inconsistencies. Ensuring data quality is essential for reliable analysis and decision-making.
Visualization Techniques: Visualization plays a crucial role in EDA as it allows analysts to visually explore and communicate insights from the social media data. Common visualization techniques include:

- Histograms and Density Plots: Visualize the distribution of numerical variables.
- Bar Charts: Display the frequency distribution of categorical variables.
- Line Plots: Show trends and patterns over time.
- Scatter Plots: Explore relationships between two numerical variables.
- Heatmaps: Visualize correlations between variables.
- Word Clouds: Highlight frequently occurring words or phrases in text data.
- Geospatial Maps: Display geographical patterns and distributions.
- Network Graphs: Visualize relationships and connections between entities such as users, hashtags, or topics.

Temporal Analysis: Social media data often includes timestamps, allowing for temporal analysis. EDA may involve analyzing trends, seasonality, and patterns over time to understand how user behavior evolves and responds to external events or marketing campaigns.
Sentiment Analysis: EDA can include sentiment analysis to understand the sentiment (positive, negative, neutral) of social media posts or comments related to the business. Visualizing sentiment trends over time helps in gauging customer satisfaction and identifying areas for improvement.

User Engagement and Influence: Analyzing user engagement metrics such as likes, shares, comments, and followers can provide insights into user behavior and preferences. EDA may involve identifying influential users or content that drives high engagement.
Competitor Analysis: EDA can extend to analyzing competitors' activities and performance on social media platforms. Comparing key metrics such as follower growth, engagement rates, and sentiment helps in benchmarking against competitors and identifying competitive advantages.
Audience Segmentation: EDA may involve segmenting social media users based on demographic, geographic, or behavioral characteristics. Understanding different user segments allows businesses to tailor their marketing strategies and messaging accordingly.
Insight Generation: The ultimate goal of EDA is to generate actionable insights that drive business decisions. By exploring and visualizing social media data, businesses can identify opportunities, mitigate risks, optimize marketing campaigns, and enhance customer experiences.

- **Descriptive Statistics:**

Descriptive statistics can be defined as the measures that summarize a given data, and these measures can be broken down further

1. Measure of central tendency
2. Measure of spread/dispersion
3. Measure of symmetry/shape

Measure of Central Tendency

Measure of central tendency is used to describe the middle/ centre value of the data. Mean, Median, Mode are measures of central tendency.

1. Mean

- Mean is the average value of the dataset.
- Mean is calculated by adding all values in the dataset divided by the number of values in the dataset.
- We can calculate the mean for only numerical variables.

2. Median

- The Median is the middle number in the dataset.
- Median is the best measure when we have outliers.

3. Mode

The mode is used to find the common number in the dataset.

Measure of spread

- The measure of spread/dispersion is used to describe how data is spread. It also describes the **variability** of the dataset.
- **Standard Deviation, Variance, Range, IQR,** are used to describe the measure of spread/dispersion

- The measure of spread can be shown in graphs like **boxplot**.

1.Variance

- Variance is used to describe how far each number in the dataset is from the mean.
- Formula to calculate population variance

$$\sigma^2 = \frac{\sum(x-\mu)^2}{N}$$

2.Standard Deviation

- Standard Deviation is the measure of the spread of data from the mean.
- Standard deviation is the square root of variance.
- More the standard deviation, more the spread.

3.Range

- The range is the difference between the largest number and the smallest number.
- Larger the range, the more the dispersion.

4. Interquartile range (IQR)

- Quartiles describe the spread of data by breaking into quarters. The median exactly divides the data into two parts.

- **Q1(Lower quartile)** is the middle value in the first half of the sorted dataset.
- **Q2**– is the median value
- **Q3 (Upper quartile)** is the middle value in the second half of the sorted dataset

- The interquartile range is the difference between the 75th percentile(Q3) and the 25th percentile(Q1).
- 50% of data fall within this range.

Boxplot is used to describe how the data is distributed in the dataset. This graph represents five- point summary (minimum, maximum, median, lower quartile, and upper quartile) and is used to identify **outliers**.

- whiskers — denote the spread of data
- box— represents the IQR- 50% of data lies within this range.

Measure of shape

1. Skewness

Skewness, which is the measure of the symmetry, or lack of it, for a real-valued random variable about its mean. The skewness value can be positive, negative, or undefined. In a perfectly symmetrical distribution, the mean, the median, and the mode will all have the same value.

2. *Kurtosis*

Kurtosis provides a measurement about the extremities (i.e. tails) of the distribution of data, and therefore provides an indication of the presence of outliers. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution. That is, data sets with high kurtosis tend to have heavy tails, or outliers. Data sets with low kurtosis tend to have light tails, or lack of outliers.

Program:

```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import skew, kurtosis

# Load sample social media data
data = pd.DataFrame({
    'user_id': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    'followers_count': [120, 250, 400, 180, 550, 140, 80, 320, 450, 850],
    'likes': [12, 20, 15, 30, 35, 55, 50, 65, 110, 85],
    'shares': [6, 4, 8, 10, 7, 14, 12, 11, 18, 20],
    'comments': [2, 3, 4, 5, 6, 7, 3, 4, 7, 12],
    'engagement_rate': [0.12, 0.18, 0.22, 0.20, 0.32, 0.28, 0.30, 0.38, 0.42, 0.55],
    'sentiment_score': [0.85, 0.55, 0.75, 0.95, 0.65, 0.45, 0.65, 0.85, 0.95, 0.75]
})
```

```
# Display the first few rows of the data
print("First 5 rows of data:")
print(data.head())
```

```
First 5 rows of data:
   user_id  followers_count  likes  shares  comments  engagement_rate  \
0        1              120     12       6         2             0.12
1        2              250     20       4         3             0.18
2        3              400     15       8         4             0.22
3        4              180     30      10         5             0.20
4        5              550     35       7         6             0.32

   sentiment_score
0             0.85
1             0.55
2             0.75
3             0.95
4             0.65
```

```
# Descriptive Statistics
print("\nDescriptive Statistics:")
print(data.describe())
```

```
Descriptive Statistics:
        user_id  followers_count       likes     shares   comments  \
count  10.00000        10.000000   10.000000  10.000000  10.000000
mean    5.50000       334.000000   47.700000  11.000000   5.300000
std     3.02765       237.963583   32.000174   5.163978   2.907844
min     1.00000        80.000000   12.000000   4.000000   2.000000
25%     3.25000       150.000000   22.500000   7.250000   3.250000
50%     5.50000       285.000000   42.500000  10.500000   4.500000
75%     7.75000       437.500000   62.500000  13.500000   6.750000
max    10.00000       850.000000  110.000000  20.000000  12.000000

       engagement_rate  sentiment_score
count        10.000000        10.000000
mean          0.297000         0.740000
std           0.127893         0.166333
min           0.120000         0.450000
25%           0.205000         0.650000
50%           0.290000         0.750000
75%           0.365000         0.850000
max           0.550000         0.950000
```

```
# Central Tendency - Mean, Median, Mode
print("\nMean of Followers Count:", data['followers_count'].mean())
print("Median of Followers Count:", data['followers_count'].median())
print("Mode of Followers Count:", data['followers_count'].mode()[0])
```

```
Mean of Followers Count: 334.0
Median of Followers Count: 285.0
Mode of Followers Count: 80
```

```
# Measure of Spread - Standard Deviation, Variance, Range
print("\nStandard Deviation of Followers Count:", data['followers_count'].std())
print("Variance of Followers Count:", data['followers_count'].var())
print("Range of Followers Count:", data['followers_count'].max() -
data['followers_count'].min())
```

```
Standard Deviation of Followers Count: 237.9635826479898
Variance of Followers Count: 56626.666666666664
Range of Followers Count: 770
```

```
# Interquartile Range (IQR)
Q1 = data['followers_count'].quantile(0.25)
Q3 = data['followers_count'].quantile(0.75)
IQR = Q3 - Q1
print("\nInterquartile Range (IQR) of Followers Count:", IQR)
```

```
Interquartile Range (IQR) of Followers Count: 287.5
```

```
# Skewness and Kurtosis
print("\nSkewness of Followers Count:", skew(data['followers_count']))
print("Kurtosis of Followers Count:", kurtosis(data['followers_count']))
```

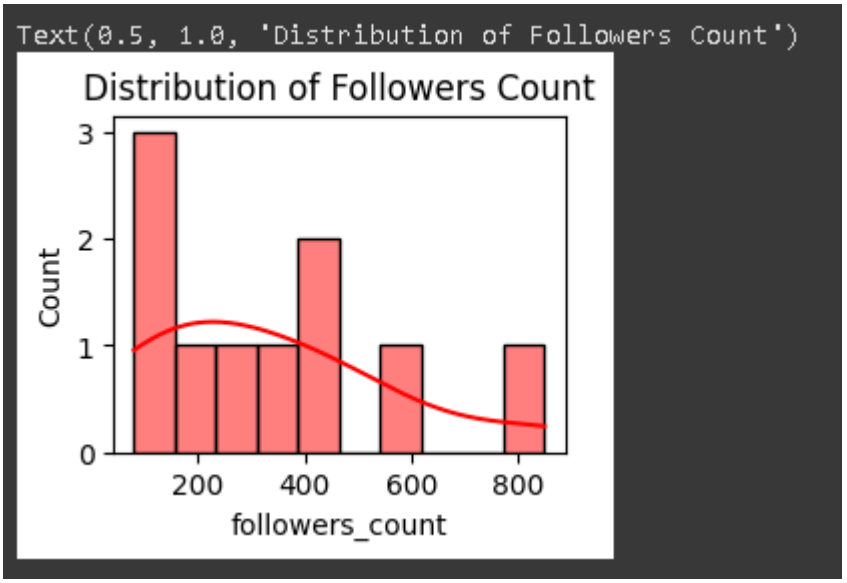
```
Skewness of Followers Count: 0.9697997469537556
Kurtosis of Followers Count: 0.1400488665646651
```

```
# Data Quality - Checking for Missing Values
print("\nMissing Values in each column:")
print(data.isnull().sum())
```

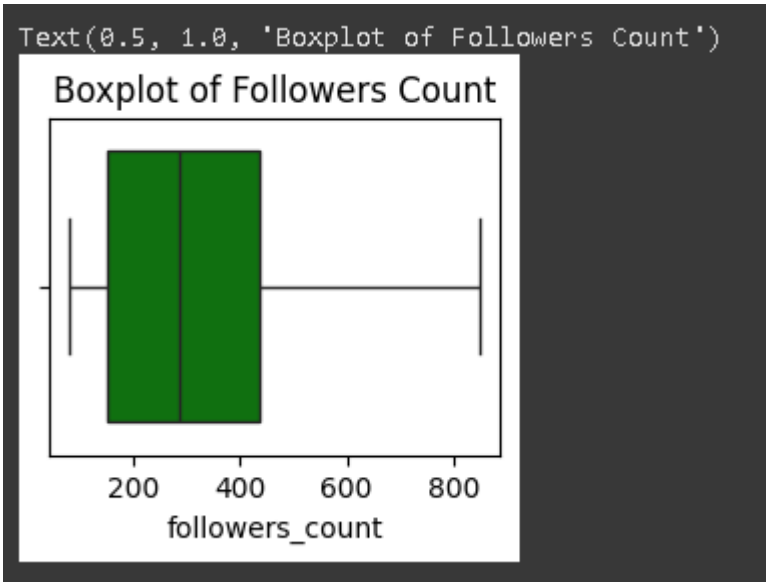
```
Missing Values in each column:
user_id            0
followers_count    0
likes              0
shares             0
comments           0
engagement_rate    0
sentiment_score    0
dtype: int64
```

```
# Visualizations
```

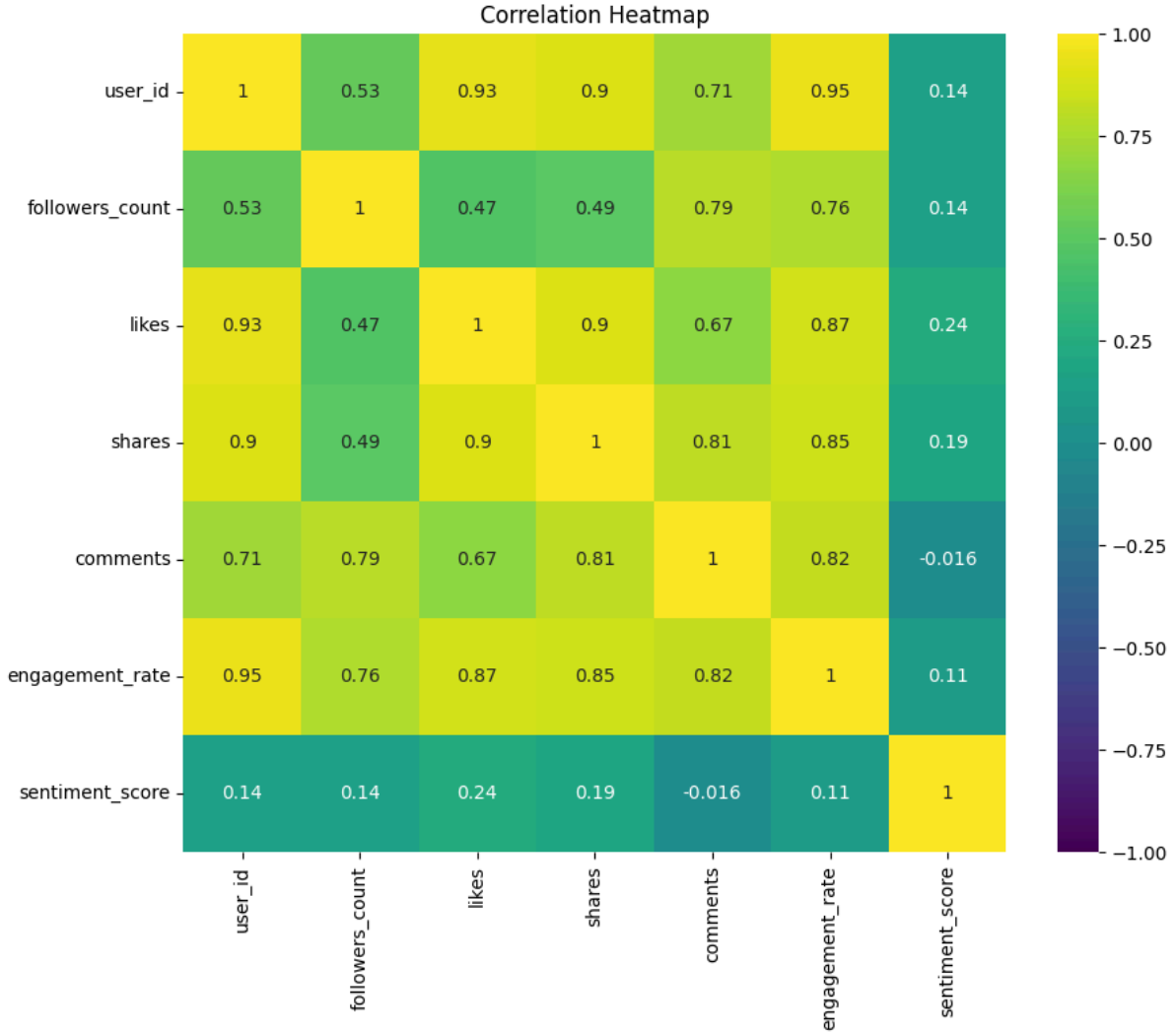
```
plt.figure(figsize=(14, 8))

# Histogram of Followers Count
plt.subplot(2, 2, 1)
sns.histplot(data['followers_count'], kde=True, color='orange', bins=10)
plt.title('Distribution of Followers Count')
```




```
# Boxplot of Followers Count
plt.subplot(2, 2, 2)
sns.boxplot(x=data['followers_count'], color='purple')
plt.title('Boxplot of Followers Count')
```




```
# Heatmap of correlations
plt.figure(figsize=(10, 8))
sns.heatmap(data.corr(), annot=True, cmap='viridis', vmin=-1, vmax=1)
plt.title('Correlation Heatmap')
plt.show()
```



Conclusion:

This experiment demonstrated the importance of exploratory data analysis (EDA) in social media data. By analyzing key metrics like followers, likes, shares, and sentiment scores, we identified trends and patterns. Visualizations such as histograms, boxplots, and heatmaps provided insights into data distributions and relationships. EDA helps businesses make data-driven decisions by optimizing engagement strategies and understanding customer sentiment.