LAB MANUAL
PART A
(PART A: TO BE REFFERED BY STUDENTS)

Experiment No-04

A.1 Aim:
Exploratory Data Analysis and visualization of Social Media Data for business.

Lab Objective	To understand the fundamental concepts of social media networks
Lab Outcome	Collect, monitor, store and track social media data

A-2 Prerequisite
Data Mining, Data Analytics

A.3 OutCome
Students will able to perform exploratory data analysis and visualization on the chosen social media data.

A.4 Theory:

What is Exploratory Data Analysis?
We can define exploratory data analysis as the essential data investigation process before the formal analysis to spot patterns and anomalies, discover trends, and test hypotheses with summary statistics and visualizations. It gives an idea about the data we will be digging deep into while analyzing. It aids in formulating how we can handle data during analysis, like choosing models, handling outliers, deciding model accuracy parameters, etc. Visualization helps to infer insights easily from massive datasets.

Need for visualizing data:

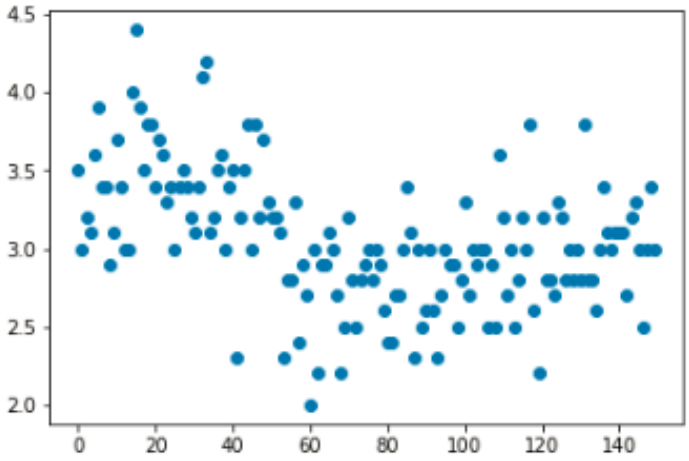
- Understand the trends and patterns of data
- Analyze the frequency and other such characteristics of data
- Know the distribution of the variables in the data.
- Visualize the relationship that may exist between different variables

The number of variables of interest featured by the data classifies it as **univariate, bivariate, or multivariate**. For example, if the data features only one variable of interest then it is a univariate

data. Further, based on the characteristics of data, it can be classified as **categorical/discrete** and **continuous** data.

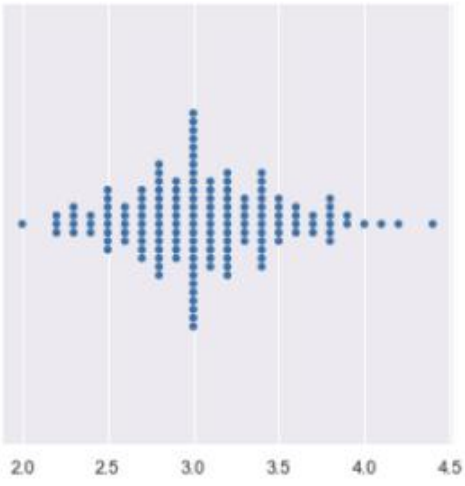
Types of Exploratory Data Analysis

1. Univariate Plots
Univariate plots show the frequency or the distribution shape of a variable.



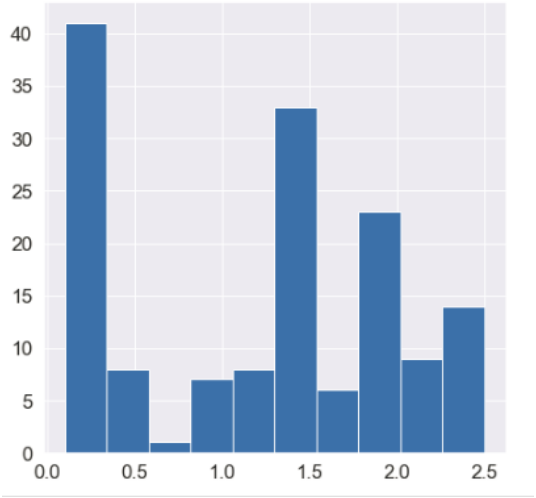
2. **Swarm Plot**

The swarm-plot, similar to a strip-plot, provides a visualization technique for univariate data to view the spread of values in a continuous variable. The swarm-plot spreads out the data points of the variable automatically to avoid overlap and hence provides a better visual overview of the data.



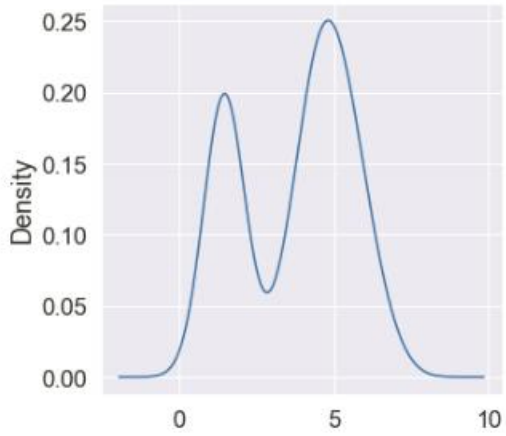
2. Histograms

Histograms are two-dimensional plots in which the x-axis divide into a range of numerical bins or time intervals. The y-axis shows the frequency values, which are counts of occurrences of values for each bin. Bar graphs have gaps between the bars to indicate that they compare distinct groups, but there are no gaps in histograms. Hence, they tell us if the distribution is left/positively skew (most of the data falls to the right side), right/negatively skewed (most of the data falls to the left side), bi-modal (graphs having two distinct peaks), normal (perfectly symmetrical without skew), or uniform (almost all the bins have similar frequency).



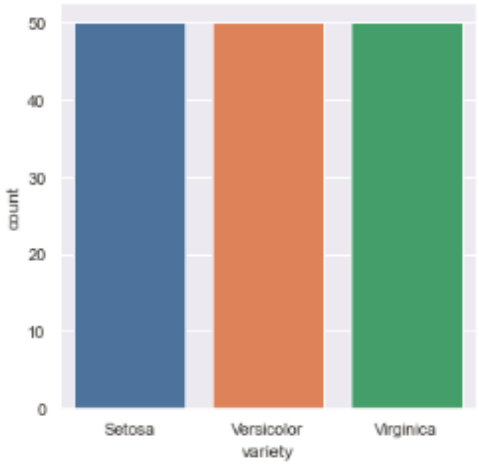
Density Plots:

A density plot is like a smoother version of a histogram. Generally, the kernel density estimate is used in density plots to show the probability density function of the variable. A continuous curve, which is the kernel is drawn to generate a smooth density estimation for the whole data.



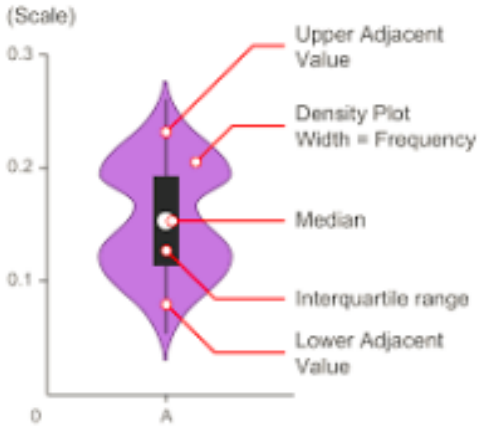
Bar Graphs

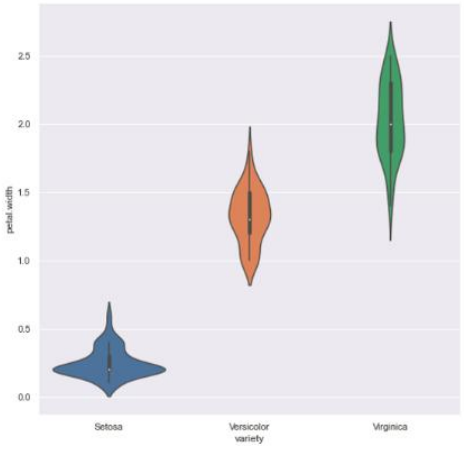
Bar charts can be used to compare nominal or ordinal data. They are helpful for recognizing trends.



Violin Plots:

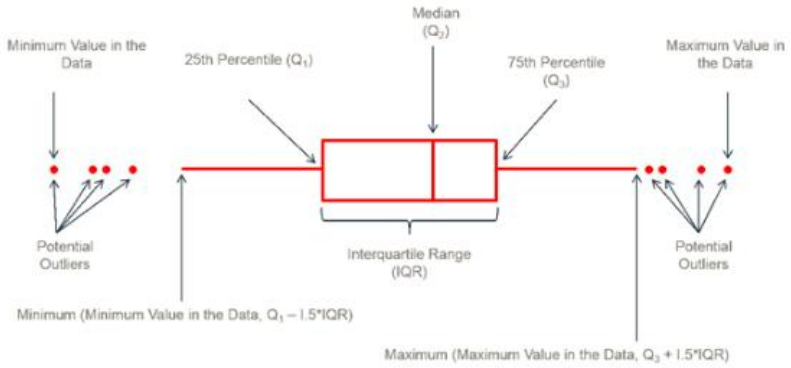
The Violin plot is very much similar to a box plot, with the addition of a rotated kernel density plot on each side. It shows the distribution of quantitative data across several levels of one (or more) categorical variables such that those distributions can be compared.

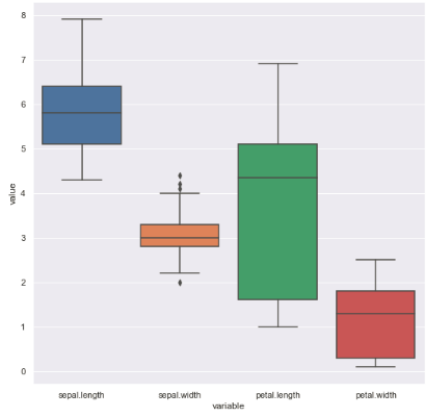




Box Plots

These charts show the distribution of values along an axis. Rectangular boxes are used in order to bucket the data, giving us an idea of how the data points are spread out. These boxes are also called quartiles which represent a quarter of a data set. Boxes can be drawn vertically or horizontally. Box plots are suitable for identifying outliers. The below figure shows the structure of a box plot.





Heat Maps

For instance, correlation heat maps show the interrelationship between variables—areas as shaded as per the data's values. So, colour differences can easily spot similar and different values and make sense of the data variation. They are usually helpful when you have a large amount of data. They are used during A/B testing to see which parts of a web page are accessed by users on a website.

PART B
(PART B: TO BE COMPLETED BY STUDENTS)

(Students must submit the soft copy as per following segments within two hours of the practical. The soft copy must be uploaded on the Blackboard or emailed to the concerned lab in charge faculties at the end of the practical in case the there is no Black board access available)

Roll. No.: C5	Name: Tavishaa Jaiswal
Class: BE-Comps	Batch: C1
Date of Experiment:	Date of Submission:
Grade:	

B.1.Study the fundamentals of social media platform and implement data cleaning, pre-processing, filtering and storing social media data for business:

Social media is a dynamic platform for tracking trends, understanding audience behavior, and refining business strategies. This experiment leverages web scraping, data processing, and visualization to analyze “2024 Olympics” discussions, extracting valuable insights on engagement and trending topics. By interpreting these patterns, businesses can optimize marketing, anticipate consumer interests, and maintain a competitive edge in an ever-evolving digital landscape.

B.2 Input and Output:
Input:

exp4.py

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import os

# Define the existing directory where CSV is stored
save_path = r"C:/Users/91702/OneDrive/Desktop/SMA"

# Load dataset (Ensure the file exists in the specified path)
file_path = os.path.join(save_path, "olympics_2024_data_20250217_214839.csv")
df = pd.read_csv(file_path)

# Display the first few rows
print("Dataset Preview:")
print(df.head())

# Convert 'created_utc' to datetime format (if applicable)
df['created_utc'] = pd.to_datetime(df['created_utc'], errors='coerce')

# Check for missing values
print("\nMissing Values:")
print(df.isnull().sum())

# Fill missing values (if needed)
df.fillna("", inplace=True)

# Descriptive statistics
print("\nDescriptive Statistics:")
print(df.describe())

# Engagement Level Analysis
print("\nEngagement Level Distribution:")
print(df['engagement_level'].value_counts())

# Category Distribution
print("\nCategory Distribution:")
print(df['category'].value_counts())
```

main ⊗ 0 ⚠ 0 Tavishaa (3 minutes ago)

exp4.py

```python
# --- Visualization Starts Here ---
# 1 Bar Chart: Post Categories Distribution
plt.figure(figsize=(10,6))
df['category'].value_counts().plot(kind='bar', color='skyblue', edgecolor='black')
plt.title("Post Categories Distribution")
plt.xlabel("Category")
plt.ylabel("Number of Posts")
plt.xticks(rotation=45)
plt.savefig(os.path.join(save_path, "bar_chart.png"))
plt.show()

# 2 Scatter Plot: Score vs. Number of Comments (Engagement Analysis)
plt.figure(figsize=(8,5))
sns.scatterplot(data=df, x='score', y='num_comments', hue='engagement_level', palette="viridis")
plt.title("Score vs. Number of Comments")
plt.xlabel("Score")
plt.ylabel("Number of Comments")
plt.savefig(os.path.join(save_path, "scatter_plot.png"))
plt.show()

# 3 Box Plot: Scores across different categories
plt.figure(figsize=(12,6))
sns.boxplot(data=df, x='category', y='score', palette="Set3", showfliers=True)
plt.title("Scores Across Different Categories")
plt.xlabel("Category")
plt.ylabel("Score")
plt.xticks(rotation=30, ha="right")
plt.grid(axis='y', linestyle="--", alpha=0.7)
plt.savefig(os.path.join(save_path, "box_plot.png"))
plt.show()

# Save the cleaned data
cleaned_csv_path = os.path.join(save_path, "cleaned_olympics_data.csv")
df.to_csv(cleaned_csv_path, index=False)

print(f"\n EDA and Visualization Completed. Graphs saved in: {save_path}")
print(f"Processed data saved as '{cleaned_csv_path}'.")
```

main ⊗ 0 ⚠ 0 — Tavishaa (7 minutes ago)

Output:

```
PROBLEMS  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS                                                    powershell

PS C:\Users\91702\OneDrive\Desktop\SMA> python exp4.py
Dataset Preview:
                                    title  score  num_comments  ...              author         created_utc                                                url
0  Imane Khelif poses with her Gold medal after t...  53002          6613  ...           cmaia1503  2024-08-13 19:39:17  https://www.reddit.com/r/pics/comments/1er8uik...
1  USA, a country so obsessed with guns, has so f...  76712          7347  ...           krnranger  2024-07-29 22:26:29  https://www.reddit.com/r/interestingasfuck/com...
2         Most controversial pic from olympics 2024  37934          8146  ...  SumneOndHakbekalva  2024-07-28 21:09:21  https://www.reddit.com/r/pics/comments/1eeaa5m...
3  Talent Like This Is What Should Have Been In T...  34068          1070  ...           ourearsan  2024-10-14 18:33:49  https://www.reddit.com/r/nextfuckinglevel/comm...
4  USA and China tie for most gold medals in the ...  25120          1997  ...         CrispyMiner  2024-08-11 23:20:10  https://www.reddit.com/r/news/comments/1epqvq1...

[5 rows x 9 columns]

Missing Values:
title             0
score             0
num_comments      0
engagement_level  0
category          0
subreddit         0
author            0
created_utc       0
url               0
dtype: int64

Descriptive Statistics:
              score  num_comments                   created_utc
count     69.000000     69.000000                            69
mean   24045.260870   2334.956522  2024-07-29 18:17:35.304347392
min       13.000000      8.000000           2020-05-08 09:53:54
25%     8888.000000    559.000000           2024-07-28 08:26:49
50%    18521.000000   1058.000000           2024-08-08 00:33:16
75%    34068.000000   2033.000000           2024-08-13 20:18:25
max    85105.000000  28213.000000           2025-02-06 05:17:04
std    19993.456801   4762.775388                           NaN

Engagement Level Distribution:
engagement_level
High      66
Medium     2
Low        1
Name: count, dtype: int64
```

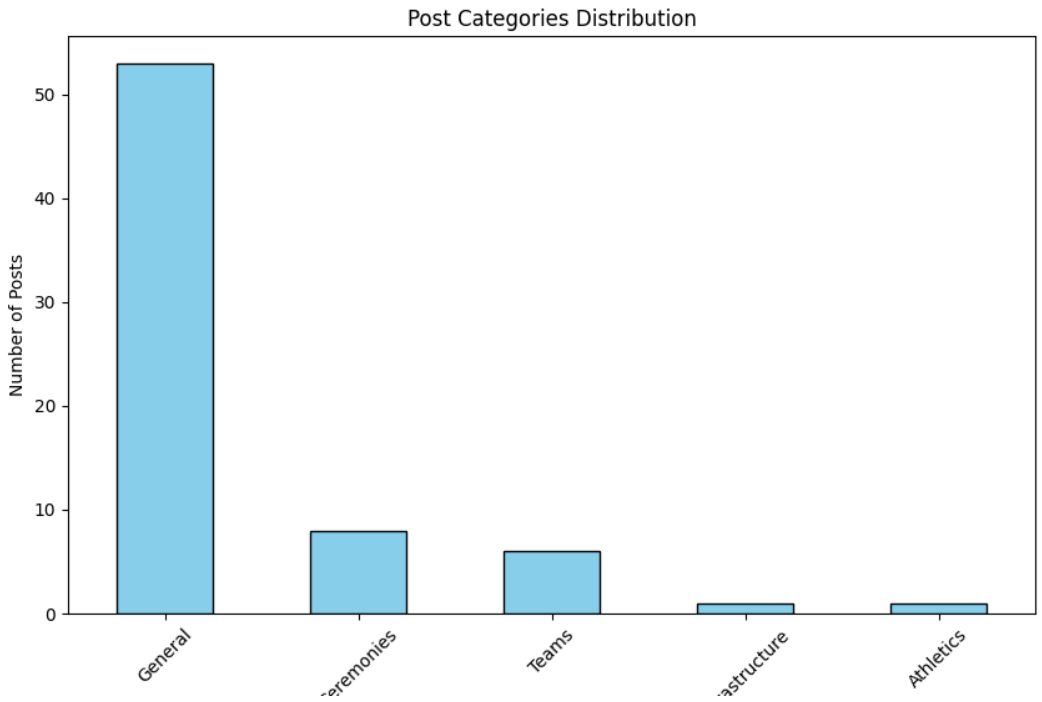
main ⊗ 0 ⚠ 0 — Tavishaa (7 minutes ago) — Ln 53, Col 11 — Spaces: 4 — UTF-8 — CRLF — {} Python — 3.11.2 64-bit

```
PROBLEMS   OUTPUT   DEBUG CONSOLE   TERMINAL   PORTS

subreddit          0
author             0
created_utc        0
url                0
dtype: int64

Descriptive Statistics:
              score  num_comments                   created_utc
count     69.000000     69.000000                            69
mean   24045.260870   2334.956522  2024-07-29 18:17:35.304347392
min       13.000000      8.000000           2020-05-08 09:53:54
25%     8888.000000    559.000000           2024-07-28 08:26:49
50%    18521.000000   1058.000000           2024-08-08 00:33:16
75%    34068.000000   2033.000000           2024-08-13 20:18:25
max    85105.000000  28213.000000           2025-02-06 05:17:04
std    19993.456801   4762.775388                           NaN

Engagement Level Distribution:
engagement_level
High      66
Medium     2
Low        1
Name: count, dtype: int64

Category Distribution:
category
General          53
Ceremonies        8
Teams             6
Infrastructure    1
Athletics         1
Name: count, dtype: int64
C:\Users\91702\OneDrive\Desktop\SMA\exp4.py:61: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.boxplot(data=df, x='category', y='score', palette="Set3", showfliers=True)

 EDA and Visualization Completed. Graphs saved in: C:/Users/91702/OneDrive/Desktop/SMA
Processed data saved as 'C:/Users/91702/OneDrive/Desktop/SMA\cleaned_olympics_data.csv'.
PS C:\Users\91702\OneDrive\Desktop\SMA>
```

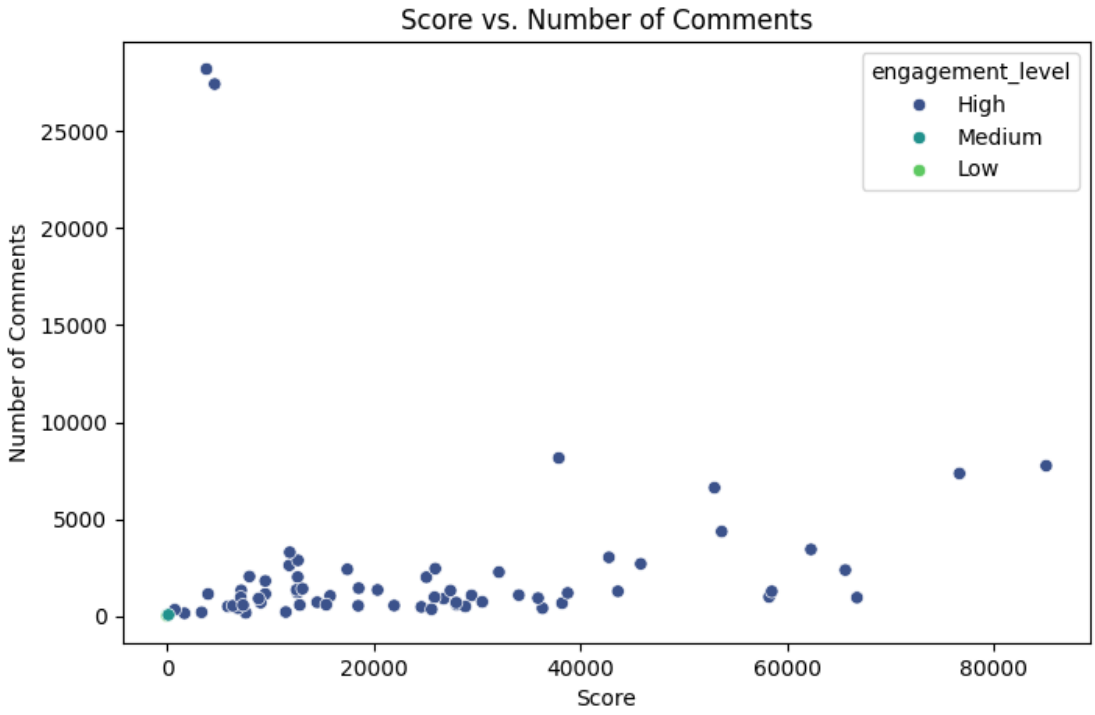
main ⊗ 0 ⚠ 0 Tavishaa (7 minutes ago) Ln 53, Col 11 Spaces: 4 UTF-8 CRLF {} Python

Cleaned data: https://github.com/Tavishaa/SMA-analysis/blob/main/cleaned_olympics_data.csv

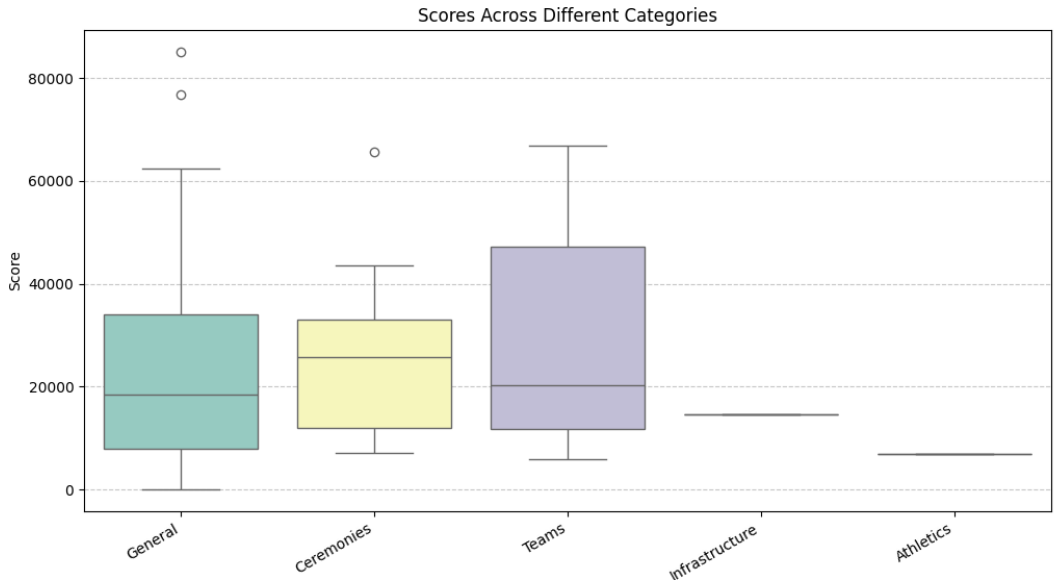
Bar Chart:



Scatter Plot:



Box Plot:



B.3 Observations and learning:
The analysis revealed that topics like "Ceremonies" and "Athletics" attracted the most attention, with engagement levels closely tied to post scores and comment activity. Refining data categorization improved accuracy, ensuring that all key themes were represented. Effective data processing and visualization provided a clear picture of audience behavior, helping businesses interpret trends, detect outliers, and optimize engagement strategies.

B.4 Conclusion:
This experiment demonstrates how social media data can be transformed into valuable business intelligence. By structuring and analyzing online interactions, companies can track public interest, refine engagement strategies, and make data-driven decisions to stay ahead in a competitive digital landscape.

B.5 Question of Curiosity

(To be answered by student based on the practical performed and learning/observations)

Q1. What is EDA? Explain its importance?

Ans: (EDA) is the process of examining, summarizing, and visualizing data to uncover patterns, detect anomalies, and gain insights before applying predictive models.

- Detects patterns, trends, and anomalies in data.
- Enhances data-driven decision-making.
- Improves data quality through cleaning and structuring.
- Helps in selecting suitable models for further analysis.
- Optimizes business strategies by understanding user behavior.

Q2. What is the importance of visualization?

Ans: Visualization helps in converting complex data into easily understandable insights, making analysis more effective and decision-making faster.

- Makes data interpretation intuitive and accessible.
- Highlights trends, patterns, and relationships clearly.
- Speeds up data-driven decision-making.
- Identifies anomalies and outliers effectively.
- Enhances storytelling and presentation of insights.

Q3. Explain the steps involved in EDA?

Ans: Exploratory Data Analysis (EDA) follows a structured approach to understand data, identify patterns, and prepare it for further analysis.

1. **Data Collection:** Gather raw data from various sources.
2. **Data Cleaning:** Handle missing values, remove duplicates, and fix inconsistencies.
3. **Data Transformation:** Convert raw data into a structured format for analysis.
4. **Univariate Analysis:** Examine individual variables using histograms and box plots.
5. **Bivariate & Multivariate Analysis:** Identify relationships between multiple variables using scatter plots and heatmaps.
6. **Feature Engineering:** Create new meaningful variables to improve analysis.
7. **Visualization:** Use charts and graphs to interpret and communicate insights effectively.