Social Network Analysis Documentation (PySpark on Google Colab)

Overview

For this social network analysis project, a survey was conducted to gather insights into people's usage patterns and sentiments towards social media platforms. The dataset used for the analysis contains responses from the survey participants, providing valuable information on various aspects of social media usage.

The analysis was performed using PySpark on Google Colab, leveraging the distributed computing capabilities of PySpark for efficient data processing and analysis. By utilizing Google Colab's integrated environment, the analysis was conducted collaboratively and seamlessly, enabling exploration and interpretation of the dataset to derive meaningful insights.

Code Description

The code is implemented in Python using PySpark for distributed data processing. It analyzes a dataset containing responses from a survey on social media usage.

Source Code:

```
!pip install pyspark
# Import required libraries
from pyspark.sql import SparkSession
from pyspark.sql.functions import col, explode, split
import matplotlib.pyplot as plt

# Create a SparkSession
spark = SparkSession.builder \
    .appName("SocialNetworkAnalysis") \
    .getOrCreate()

# Load  cleaned dataset into a Spark DataFrame
df = spark.read.csv("sns_addiction.csv", header=True, inferSchema=True)

# Display the first few rows of the DataFrame
print("First few rows of the DataFrame:")
df.show(5)

# Print the schema of the DataFrame
print("Schema of the DataFrame:")
df.printSchema()
```

```
# Demographic Information
# Age distribution
print("Age distribution:")
age_distribution = df.groupBy("age").count().orderBy("age")
age_distribution.show()

# Gender distribution
print("Gender distribution:")
gender_distribution = df.groupBy("gender").count().orderBy("count", ascending=False)
gender_distribution.show()

# Feelings Towards Social Media
# Count of feelings after using social media
print("Feelings after using social media:")
feelings_count = df.groupBy("feelings_after_use").count().orderBy("count",
ascending=False)
feelings_count.show()

# Count of respondents experiencing FOMO
print("Experience of FOMO (Fear of Missing Out):")
fomo_count = df.groupBy("fomo_experience").count().orderBy("count", ascending=False)
fomo_count.show()

# Count of respondents feeling addicted to social media
print("Feeling addicted to social media:")
addiction_count = df.groupBy("feel_addicted").count().orderBy("count", ascending=False)
addiction_count.show()

# Impact on Daily Life
# Count of respondents prioritizing social media over other activities
print("Prioritizing social media over other activities:")
prioritization_count = df.groupBy("prioritize_social_media").count().orderBy("count",
ascending=False)
prioritization_count.show()

# Count of respondents finding it difficult to reduce social media usage
print("Difficulty in reducing social media usage:")
difficulty_count = df.groupBy("difficulty_reducing_usage").count().orderBy("count",
ascending=False)
difficulty_count.show()

# Count of respondents experiencing negative consequences due to social media usage
print("Negative consequences due to social media usage:")
negative_consequences_count =
df.groupBy("negative_consequences").count().orderBy("count", ascending=False)
negative_consequences_count.show()

# Self-Assessment
```

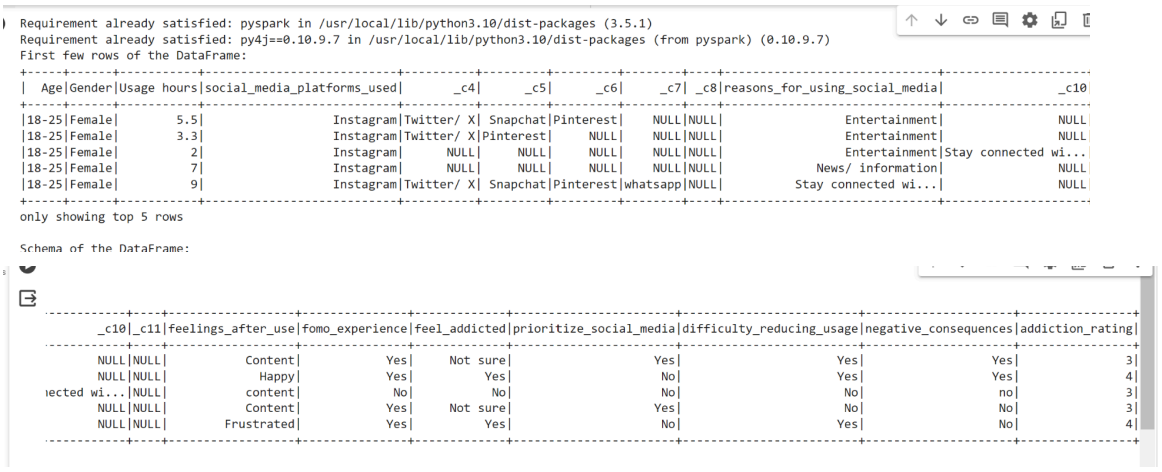
```
# Average social media addiction rating
print("Average social media addiction rating:")
average_addiction_rating = df.selectExpr("avg(addiction_rating)").collect()[0][0]
print("   ", average_addiction_rating)

# Reasons for Using Social Media
print("Reasons for using social media:")
reasons_for_using_social_media =
df.groupBy("reasons_for_using_social_media").count().orderBy("count", ascending=False)
reasons_for_using_social_media.show()

# Usage of Social Media Platforms
print("Usage of social media platforms:")
# Split the column containing multiple platform selections into an array
df_with_platforms = df.withColumn("platforms_array",
split(col("social_media_platforms_used"), ", "))
# Explode the array to create a new row for each platform
df_exploded = df_with_platforms.select(explode("platforms_array").alias("platform"))
# Count the occurrences of each platform
platform_count = df_exploded.groupBy("platform").count().orderBy("count",
ascending=False)
platform_count.show()

# Visualize Age Distribution
print("Visualizing Age Distribution...")
age_distribution_pd = age_distribution.toPandas()
plt.bar(age_distribution_pd["age"], age_distribution_pd["count"])
plt.xlabel("Age")
plt.ylabel("Count")
plt.title("Age Distribution")
plt.show()

# Visualize Gender Distribution
print("Visualizing Gender Distribution...")
gender_distribution_pd = gender_distribution.toPandas()
plt.bar(gender_distribution_pd["gender"], gender_distribution_pd["count"])
plt.xlabel("Gender")
plt.ylabel("Count")
plt.title("Gender Distribution")
plt.show()

# Visualize Feelings After Using Social Media
print("Visualizing Feelings After Using Social Media...")
feelings_count_pd = feelings_count.toPandas()
plt.bar(feelings_count_pd["feelings_after_use"], feelings_count_pd["count"])
plt.xlabel("Feelings After Using Social Media")
plt.ylabel("Count")
plt.title("Feelings After Using Social Media")
```

```
plt.xticks(rotation=45)
plt.show()

# Visualize Experience of FOMO
print("Visualizing Experience of FOMO...")
fomo_count_pd = fomo_count.toPandas()
plt.bar(fomo_count_pd["fomo_experience"], fomo_count_pd["count"])
plt.xlabel("Experience of FOMO")
plt.ylabel("Count")
plt.title("Experience of FOMO")
plt.show()

# Visualize Reasons for Using Social Media
print("Visualizing Reasons for Using Social Media...")
reasons_for_using_social_media_pd = reasons_for_using_social_media.toPandas()
plt.bar(reasons_for_using_social_media_pd["reasons_for_using_social_media"],
reasons_for_using_social_media_pd["count"])
plt.xlabel("Reasons for Using Social Media")
plt.ylabel("Count")
plt.title("Reasons for Using Social Media")
plt.xticks(rotation=90)
plt.show()

# Stop the SparkSession
spark.stop()
```

Analysis Steps

Data Loading: The code loads the cleaned dataset into a Spark DataFrame, leveraging Google Colab's integration with PySpark.

Output:



```
Requirement already satisfied: pyspark in /usr/local/lib/python3.10/dist-packages (3.5.1)
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
First few rows of the DataFrame:
+-----+------+-----------+--------------------------+----------+---------+---------+--------+----+--------------------------+--------------------+
|  Age|Gender|Usage hours|social_media_platforms_used|       _c4|      _c5|      _c6|     _c7| _c8|reasons_for_using_social_media|                _c10|
+-----+------+-----------+--------------------------+----------+---------+---------+--------+----+--------------------------+--------------------+
|18-25|Female|        5.5|                 Instagram|Twitter/ X| Snapchat|Pinterest|    NULL|NULL|             Entertainment|                NULL|
|18-25|Female|        3.3|                 Instagram|Twitter/ X|Pinterest|     NULL|    NULL|NULL|             Entertainment|                NULL|
|18-25|Female|          2|                 Instagram|      NULL|     NULL|     NULL|    NULL|NULL|             Entertainment|Stay connected wi...|
|18-25|Female|          7|                 Instagram|      NULL|     NULL|     NULL|    NULL|NULL|          News/ information|                NULL|
|18-25|Female|          9|                 Instagram|Twitter/ X| Snapchat|Pinterest|whatsapp|NULL|       Stay connected wi...|                NULL|
+-----+------+-----------+--------------------------+----------+---------+---------+--------+----+--------------------------+--------------------+
only showing top 5 rows

Schema of the DataFrame:
```

```
----------+----+-----------------+---------------+-------------+----------------------+-----------------------+---------------------+---------------+
       _c10|_c11|feelings_after_use|fomo_experience|feel_addicted|prioritize_social_media|difficulty_reducing_usage|negative_consequences|addiction_rating|
----------+----+-----------------+---------------+-------------+----------------------+-----------------------+---------------------+---------------+
       NULL|NULL|          Content|            Yes|     Not sure|                   Yes|                    Yes|                  Yes|              3|
       NULL|NULL|            Happy|            Yes|          Yes|                    No|                    Yes|                  Yes|              4|
ected wi...|NULL|          content|             No|           No|                    No|                     No|                   no|              3|
       NULL|NULL|          Content|            Yes|     Not sure|                   Yes|                     No|                   No|              3|
       NULL|NULL|       Frustrated|            Yes|          Yes|                    No|                    Yes|                   No|              4|
----------+----+-----------------+---------------+-------------+----------------------+-----------------------+---------------------+---------------+
```

Exploratory Data Analysis:

- **Demographic Information:** Analyzes the distribution of age and gender among respondents.

```
Age distribution:
+---------+-----+
|      age|count|
+---------+-----+
|  13 - 17|    1|
|    18-25|   51|
|After 40s|   15|
|   In 30s|    7|
| Late 20s|    9|
+---------+-----+

Gender distribution:
+------+-----+
|gender|count|
+------+-----+
|Female|   59|
|  Male|   24|
+------+-----+
```

- **Feelings Towards Social Media:** Investigates respondents' feelings after using social media and their experience of FOMO (Fear of Missing Out).

```
Experience of FOMO (Fear of Missing Out):
+---------------+-----+
|fomo_experience|count|
+---------------+-----+
|             No|   59|
|            Yes|   24|
+---------------+-----+
```

- **Impact on Daily Life:** Examines the prioritization of social media over other activities, difficulty in reducing social media usage, and negative consequences of social media usage.

```
+-------------+-----+

Prioritizing social media over other activities:
+----------------------+-----+
|prioritize_social_media|count|
+----------------------+-----+
|                    No|   56|
|                   Yes|   27|
+----------------------+-----+

Difficulty in reducing social media usage:
+-------------------------+-----+
|difficulty_reducing_usage|count|
+-------------------------+-----+
|                       No|   46|
|                      Yes|   37|
+-------------------------+-----+
```

```
Negative consequences due to social media usage:
+---------------------+-----+
|negative_consequences|count|
+---------------------+-----+
|                   No|   45|
|                  Yes|   37|
|                   no|    1|
+---------------------+-----+
```

- **Self-Assessment:** Calculates the average social media addiction rating reported by respondents.

```
 Average social media addiction rating:
    2.7228915662650603
```

- **Reasons for Using Social Media:** Analyzes the reasons provided by respondents for using social media.

```
Reasons for using social media:
+------------------------------+-----+
|reasons_for_using_social_media|count|
+------------------------------+-----+
|                 Entertainment|   46|
|             News/ information|   17|
|          Stay connected wi...|   12|
|           Educational purpose|    1|
|          i publish books o...|    1|
|                    depression|    1|
|          It provides a gre...|    1|
|                  Work related|    1|
|          promoting busines...|    1|
|                 entertainment|    1|
|           a mix of everything|    1|
+------------------------------+-----+
```

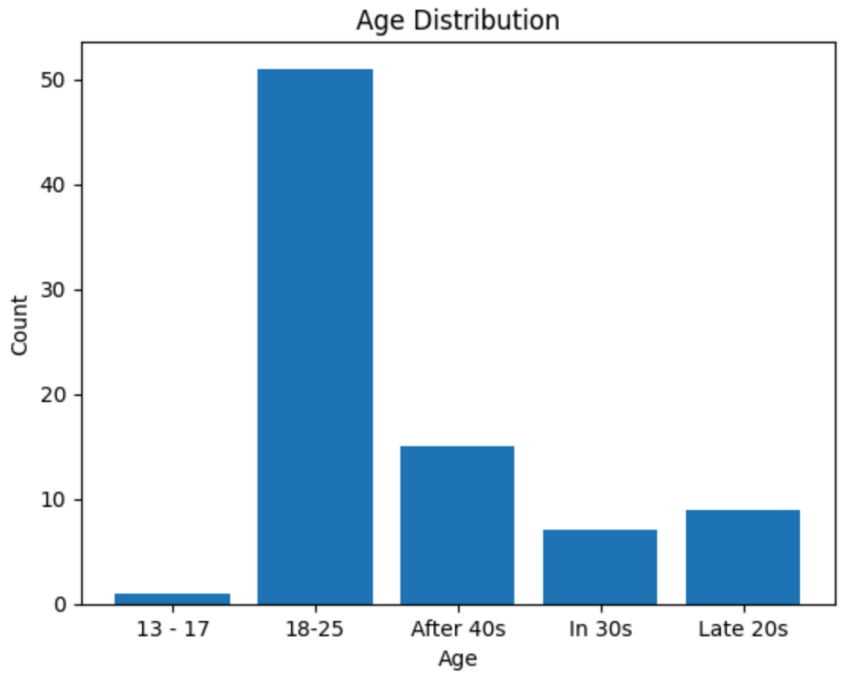
- **Usage of Social Media Platforms:** Counts the occurrences of each social media platform used by respondents.

```
Usage of social media platforms:
+----------+-----+
|  platform|count|
+----------+-----+
| Instagram|   51|
|  Facebook|   17|
|  WhatsApp|    4|
|Twitter/ X|    3|
|   YouTube|    2|
|  Snapchat|    2|
|  You tube|    1|
|   YOUTUBE|    1|
|  LinkedIn|    1|
|     Gmail|    1|
+----------+-----+
```

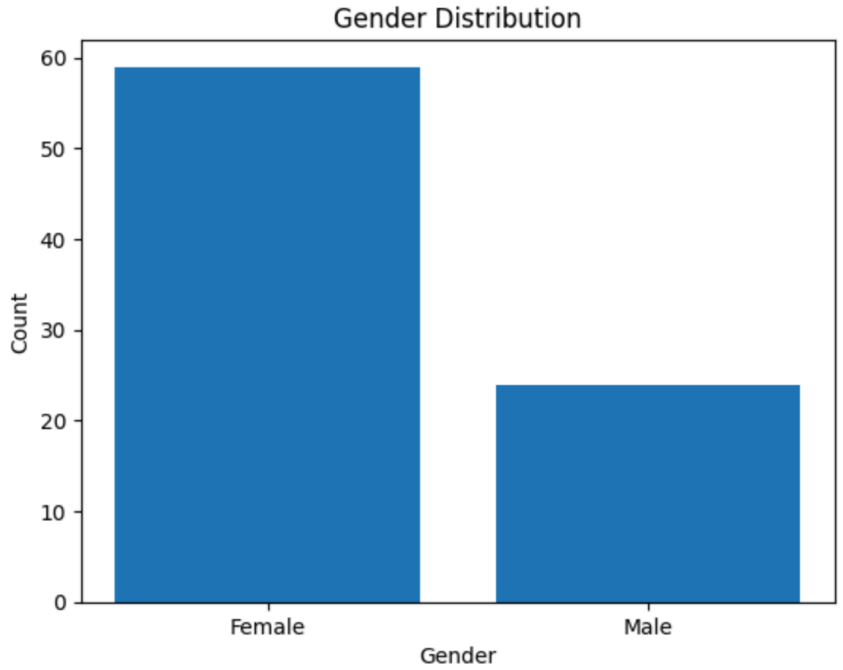
Data Visualization:

- **Age And Gender Distribution:** Visualizes the distribution of respondent ages and genders.

```
Visualizing Age Distribution...
```

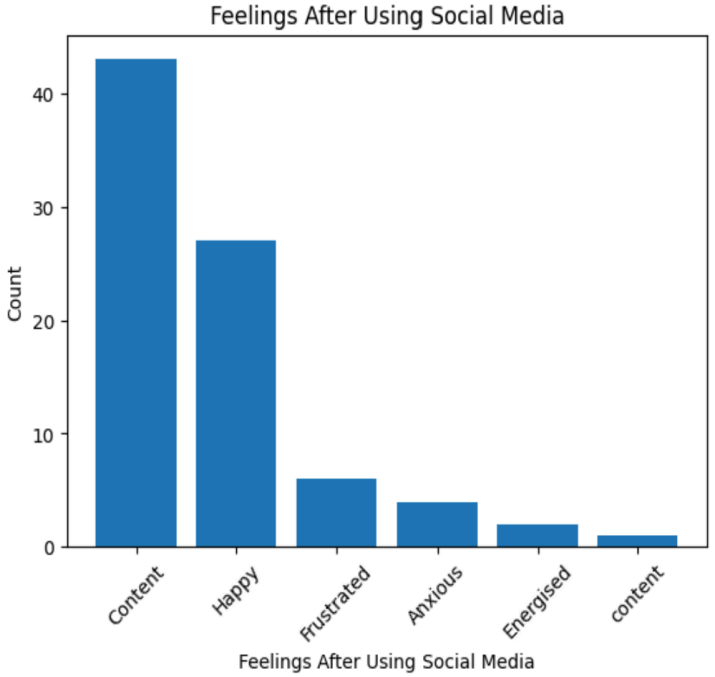



```
Visualizing Gender Distribution...
```



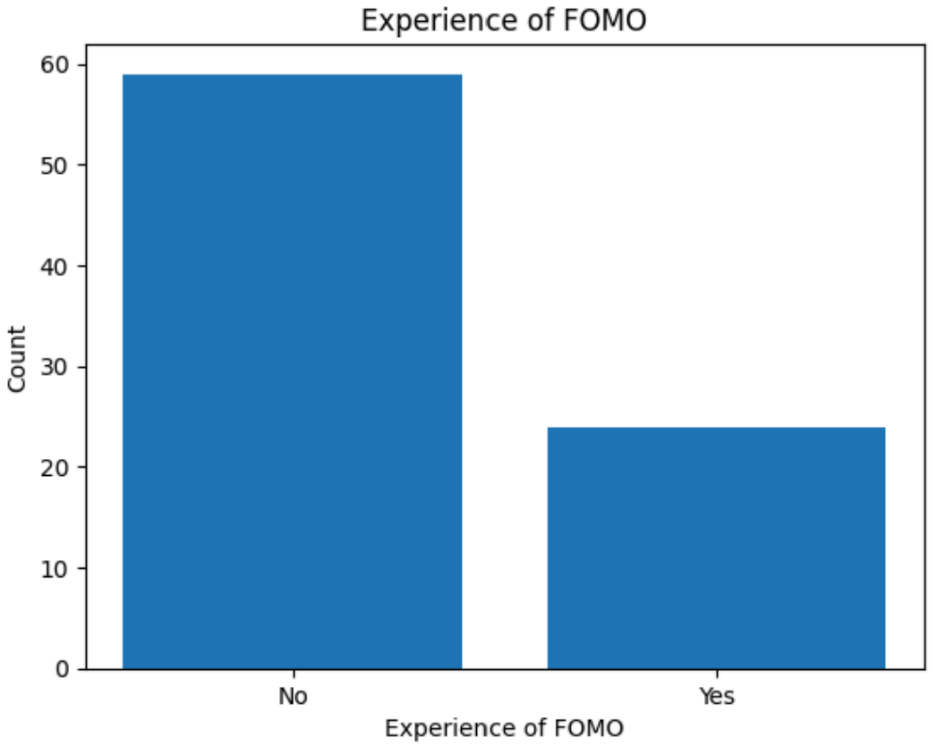
- **Feelings After Using Social Media:** Visualizes the distribution of feelings reported by respondents after using social media.

```
Visualizing Feelings After Using Social Media...
```



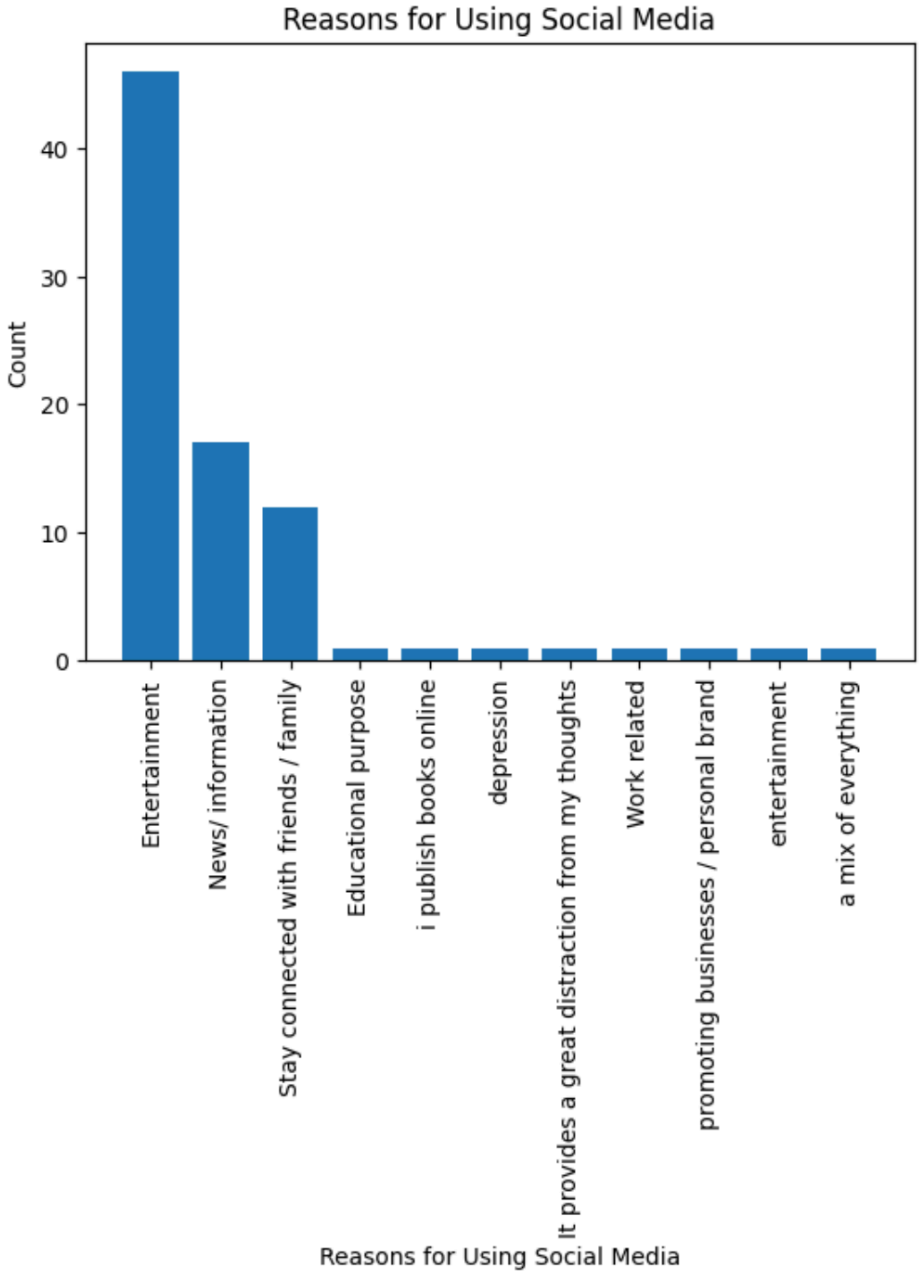
- **Experience of FOMO:** Visualizes the frequency of respondents' experience of FOMO.

```
Visualizing Experience of FOMO...
```



- **Reasons for Using Social Media:** Visualizes the reasons provided by respondents for using social media.

```
Visualizing Reasons for Using Social Media...
```



Dependencies

- PySpark (already available in Google Colab)
- Matplotlib (for data visualization)

Usage

- Open a new Google Colab notebook.
- Upload the social_network_analysis.py script or copy the code provided.
- Ensure that the dataset file is uploaded to Google Colab or accessible via a cloud storage service.
- Run the code cells in the notebook to execute the analysis.
- View the analysis results and visualizations generated in the notebook.

Conclusion

The analysis provides insights into various aspects of social media usage, leveraging the distributed computing capabilities of PySpark on Google Colab. The findings can inform decision-making in areas such as digital marketing, social media management, and mental health awareness.