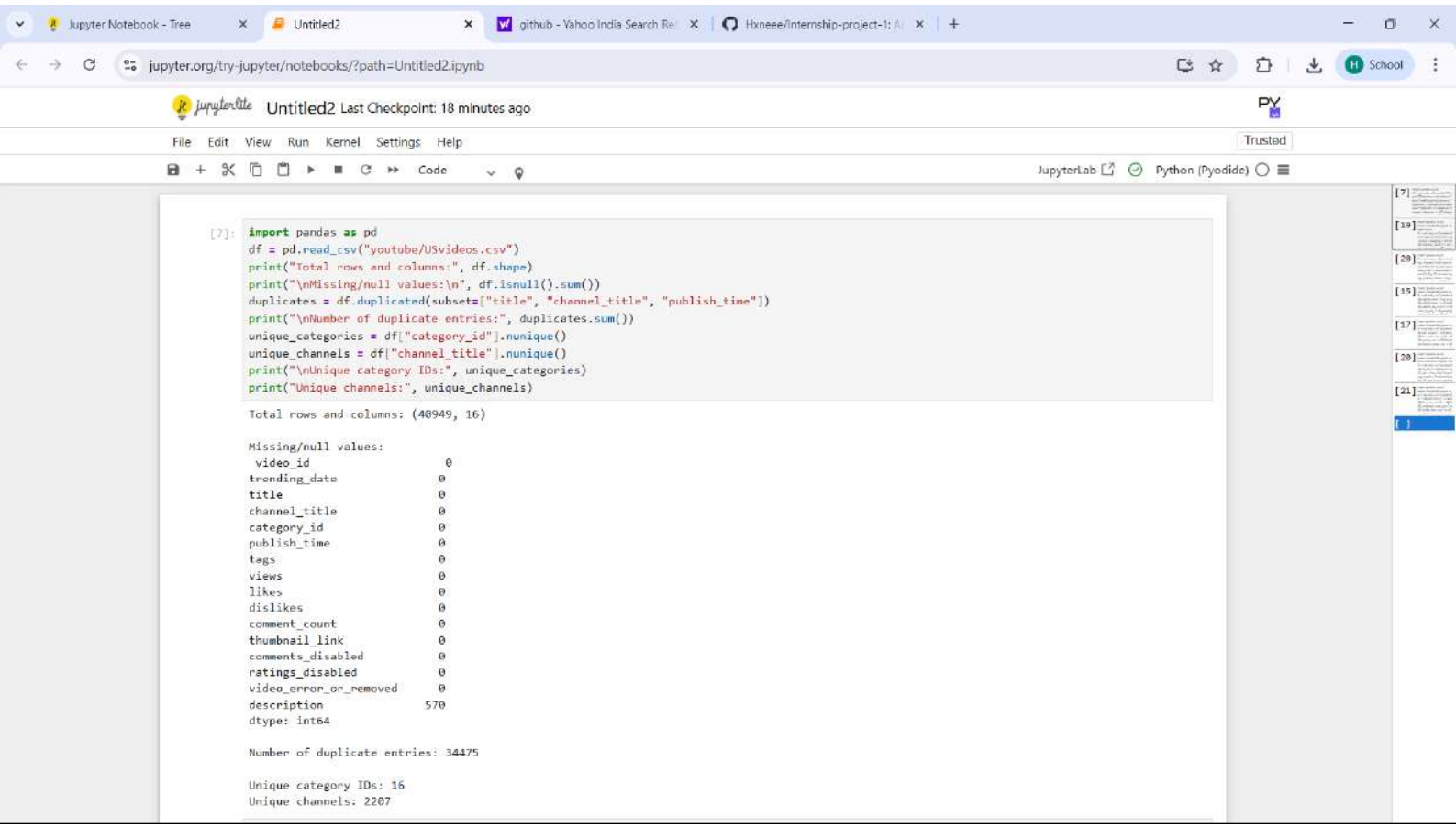

Jupyter Notebook - Tree
Untitled2
github - Yahoo India Search Re
Hxneee/Internship-project-1: A
jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb
School
jupyterlite
Untitled2 Last Checkpoint: 18 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python (Pyodide)
[7]:
import pandas as pd
df = pd.read_csv("youtube/USvideos.csv")
print("Total rows and columns:", df.shape)
print("\nMissing/null values:\n", df.isnull().sum())
duplicates = df.duplicated(subset=["title", "channel_title", "publish_time"])
print("\nNumber of duplicate entries:", duplicates.sum())
unique_categories = df["category_id"].nunique()
unique_channels = df["channel_title"].nunique()
print("\nUnique category IDs:", unique_categories)
print("Unique channels:", unique_channels)
Total rows and columns: (40949, 16)
Missing/null values:
video_id 0
trending_date 0
title 0
channel_title 0
category_id 0
publish_time 0
tags 0
views 0
likes 0
dislikes 0
comment_count 0
thumbnail_link 0
comments_disabled 0
ratings_disabled 0
video_error_or_removed 0
description 570
dtype: int64
Number of duplicate entries: 34475
Unique category IDs: 16
Unique channels: 2207

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

School

jupyterlite Untitled2 Last Checkpoint: 18 minutes ago

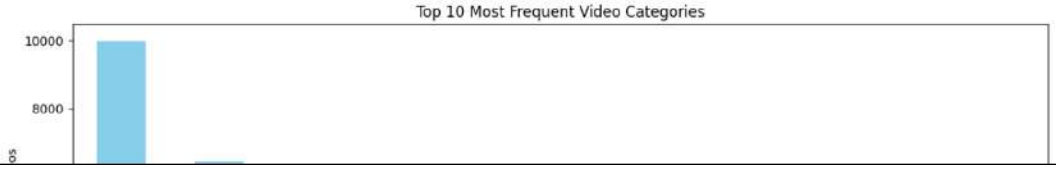
File Edit View Run Kernel Settings Help

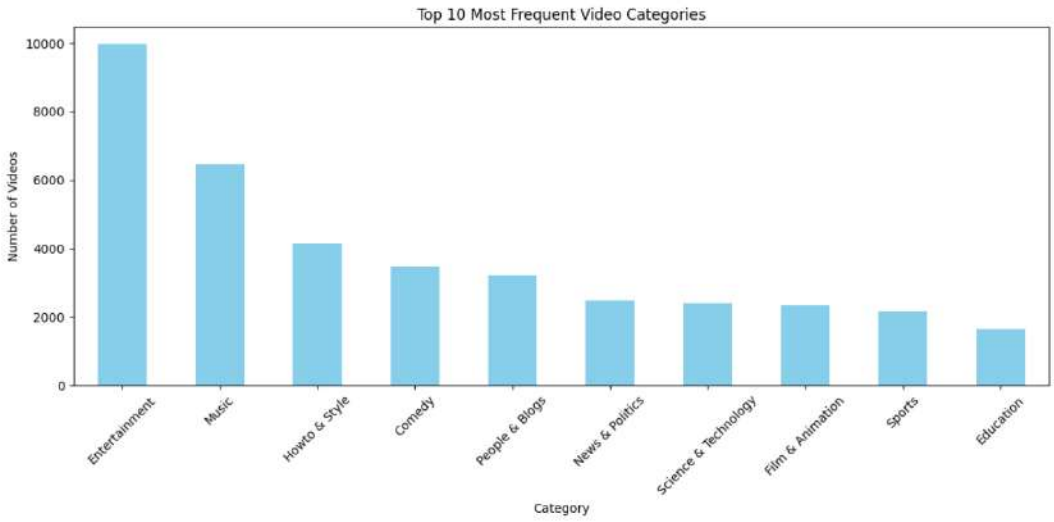
Trusted

Code

JupyterLab | Python (Pyodide)

[19]:
```
import pandas as pd
import matplotlib.pyplot as plt
import json
df = pd.read_csv("youtube/USvideos.csv")
with open("youtube/US_category_id.json") as f:data = json.load(f)
category_mapping = {int(item["id"]): item["snippet"]["title"] for item in data["items"]}
df["category_name"] = df["category_id"].map(category_mapping)
top_categories = df["category_name"].value_counts().head(10)
plt.figure(figsize=(12, 6))
top_categories.plot(kind="bar", color="skyblue")
plt.title("Top 10 Most Frequent Video Categories")
plt.xlabel("Category")
plt.ylabel("Number of Videos")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
avg_metrics = df.groupby("category_name")[["views", "likes", "comment_count"]].mean()
print("Average views, likes, and comments per category:\n", avg_metrics.round(2))
df["like_ratio"] = df["likes"] / df["views"]
df["comment_ratio"] = df["comment_count"] / df["views"]
engagement = df.groupby("category_name")[["like_ratio", "comment_ratio"]].mean()
print("\nCategory with highest like/view ratio:", engagement["like_ratio"].idxmax())
print("Category with highest comment/view ratio:", engagement["comment_ratio"].idxmax())
df["dislike_ratio"] = df["dislikes"] / (df["likes"] + df["dislikes"])
dislike_ratios = df.groupby("category_name")["dislike_ratio"].mean()
print("\nCategory with highest dislike percentage:", dislike_ratios.idxmax())
```





Average views, likes, and comments per category:

category_name	views	likes	comment_count
Autos & Vehicles	1355965.41	11056.40	2042.83
Comedy	1480308.42	62582.22	6521.72
Education	712940.82	29745.03	3286.38
Entertainment	2067883.20	53243.33	7383.23
Film & Animation	3106250.20	70787.84	7627.74
Gaming	2620830.63	84502.18	18042.49

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

JupyterLite Untitled2 Last Checkpoint: 19 minutes ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab | Python (Pyodide)

Category

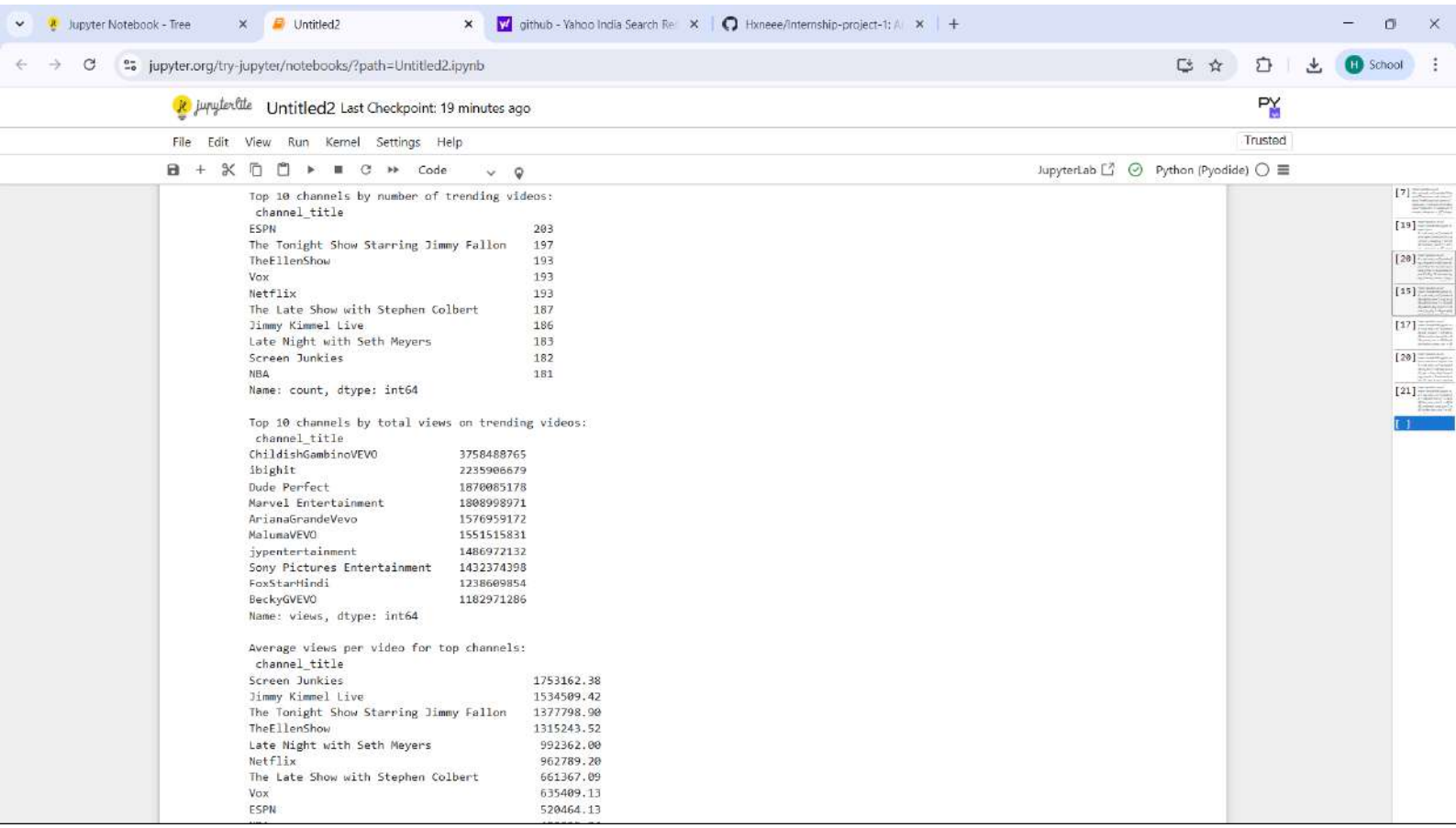
```
Average views, likes, and comments per category:
                          views      likes  comment_count
category_name
Autos & Vehicles       1355965.41   11056.40        2042.83
Comedy                 1480308.42   62582.22        6521.72
Education               712940.82   29745.03        3286.38
Entertainment          2067883.20   53243.33        7383.23
Film & Animation       3106250.20   70787.84        7627.74
Gaming                 2620830.63   84502.18       18042.49
Howto & Style           983730.12   39286.08        5583.59
Music                  6201003.12  218918.20       19359.76
News & Politics         592587.74    7298.36        2428.40
Nonprofits & Activism  2963884.07  259923.61       84364.86
People & Blogs         1531835.43   58135.83        7719.01
Pets & Animals          831143.47   21055.11        2892.07
Science & Technology   1452626.75   34374.28        4993.72
Shows                   903527.33   18993.67        1668.72
Sports                 2025969.03   45363.94        5148.19
Travel & Events         854619.61   12030.46        2267.44

Category with highest like/view ratio: Music
Category with highest comment/view ratio: Gaming

Category with highest dislike percentage: News & Politics
```

[20]:

```
import pandas as pd
df = pd.read_csv("youtube/USvideos.csv")
top_channels = df['channel_title'].value_counts().head(10)
print("Top 10 channels by number of trending videos:\n", top_channels)
total_views = df.groupby('channel_title')['views'].sum().sort_values(ascending=False).head(10)
print("\nTop 10 channels by total views on trending videos:\n", total_views)
top_channel_names = top_channels.index
average_views = df[df['channel_title'].isin(top_channel_names)].groupby('channel_title')['views'].mean().sort_values(ascending=False)
print("\nAverage views per video for top channels:\n", average_views.round(2))
```


Jupyter Notebook - Tree
Untitled2
github - Yahoo India Search Re
Hxneee/Internship-project-1: A
jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb
School
jupyterlite
Untitled2 Last Checkpoint: 19 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python (Pyodide)
Top 10 channels by number of trending videos:
channel_title
ESPN 203
The Tonight Show Starring Jimmy Fallon 197
TheEllenShow 193
Vox 193
Netflix 193
The Late Show with Stephen Colbert 187
Jimmy Kimmel Live 186
Late Night with Seth Meyers 183
Screen Junkies 182
NBA 181
Name: count, dtype: int64
Top 10 channels by total views on trending videos:
channel_title
ChildishGambinoVEVO 3758488765
1bighit 2235906679
Dude Perfect 1870085178
Marvel Entertainment 1808998971
ArianaGrandeVevo 1576959172
MalumaVEVO 1551515831
jypentertainment 1486972132
Sony Pictures Entertainment 1432374398
FoxStarHindi 1238609854
BeckyGVEVO 1182971286
Name: views, dtype: int64
Average views per video for top channels:
channel_title
Screen Junkies 1753162.38
Jimmy Kimmel Live 1534509.42
The Tonight Show Starring Jimmy Fallon 1377798.90
TheEllenShow 1315243.52
Late Night with Seth Meyers 992362.00
Netflix 962789.20
The Late Show with Stephen Colbert 661367.09
Vox 635409.13
ESPN 520464.13

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

School

jupyterlite Untitled2 Last Checkpoint: 19 minutes ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab | Python (Pyodide)

```
NBA                                     400025.24
Name: views, dtype: float64
```

[15]:

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("youtube/USvideos.csv")
df['publish_time'] = pd.to_datetime(df['publish_time'])
df['publish_hour'] = df['publish_time'].dt.hour
df['publish_day_name'] = df['publish_time'].dt.day_name()
views_by_day = df.groupby('publish_day_name')['views'].mean().reindex(['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Su
plt.figure(figsize=(10, 5))
plt.bar(views_by_day.index, views_by_day.values, color='skyblue')
plt.title("Average Views by Day of the Week")
plt.ylabel("Average Views")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
views_by_hour = df.groupby('publish_hour')['views'].mean()
plt.figure(figsize=(10, 5))
plt.plot(views_by_hour.index, views_by_hour.values, marker='o')
plt.title("Average Views by Hour of the Day")
plt.xlabel("Hour (0-23)")
plt.ylabel("Average Views")
plt.grid(True)
plt.tight_layout()
plt.show()
evening_views = df[df['publish_hour'].between(17, 21)]['views'].mean()
other_views = df[~df['publish_hour'].between(17, 21)]['views'].mean()
print(f" Average views for evening uploads (17-21h): {evening_views:,.0f}")
print(f" Average views for non-evening uploads: {other_views:,.0f}")
df['like_ratio'] = df['likes'] / df['views']
engagement_by_hour = df.groupby('publish_hour')['like_ratio'].mean()
plt.figure(figsize=(10, 5))
plt.plot(engagement_by_hour.index, engagement_by_hour.values, marker='o', color='orange')
plt.title("Like-to-View Ratio by Hour of Publish")
plt.xlabel("Hour (0-23)")
plt.ylabel("Average Like/View Ratio")
```

[7] [19] [20] [15] [17] [20] [21] [ ]

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

School

jupyterlite Untitled2 Last Checkpoint: 19 minutes ago

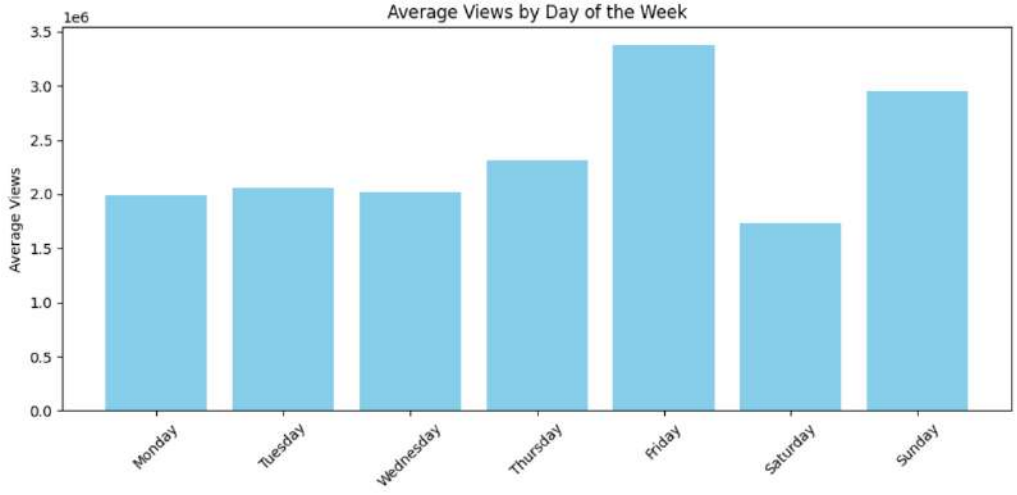
File Edit View Run Kernel Settings Help

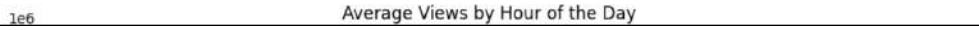
Trusted

Code

JupyterLab | Python (Pyodide)

```
plt.title("Like-to-View Ratio by Hour of Publish")
plt.xlabel("Hour (0-23)")
plt.ylabel("Average Like/View Ratio")
plt.grid(True)
plt.tight_layout()
plt.show()
```






Jupyter Notebook - Tree
Untitled2
github - Yahoo India Search Re
Hxneee/Internship-project-1: A
jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb
School
jupyterlite
Untitled2 Last Checkpoint: 19 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python (Pyodide)
0.0
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Average Views by Hour of the Day
1e6
Average Views
Hour (0-23)
Average views for evening uploads (17-21h): 1,760,363
Average views for non-evening uploads: 2,630,965
Like-to-View Ratio by Hour of Publish

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A |

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

School

jupyterlite Untitled2 Last Checkpoint: 19 minutes ago

File Edit View Run Kernel Settings Help

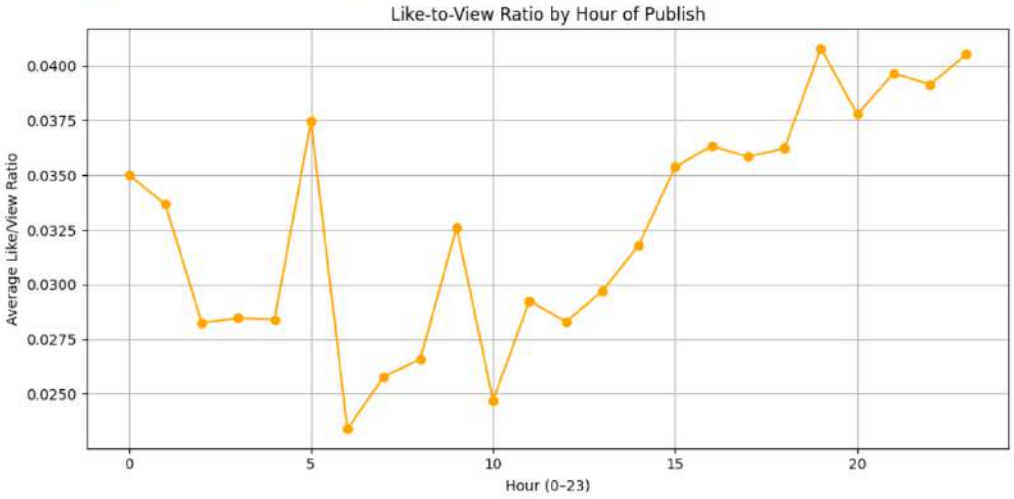
Trusted

Code

JupyterLab | Python (Pyodide)

Hour (0-23)

```
Average views for evening uploads (17-21h): 1,760,363
Average views for non-evening uploads: 2,630,965
```



[17]:

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("youtube/USvideos.csv")
df['title_length'] = df['title'].astype(str).apply(len)
df['description_length'] = df['description'].astype(str).apply(len)
title_views_corr = df['title_length'].corr(df['views'])
```

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

School

jupyterlite Untitled2 Last Checkpoint: 19 minutes ago

File Edit View Run Kernel Settings Help

Trusted

Code

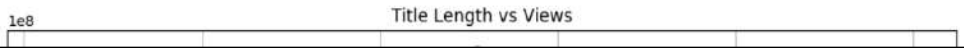
JupyterLab Python (Pyodide)

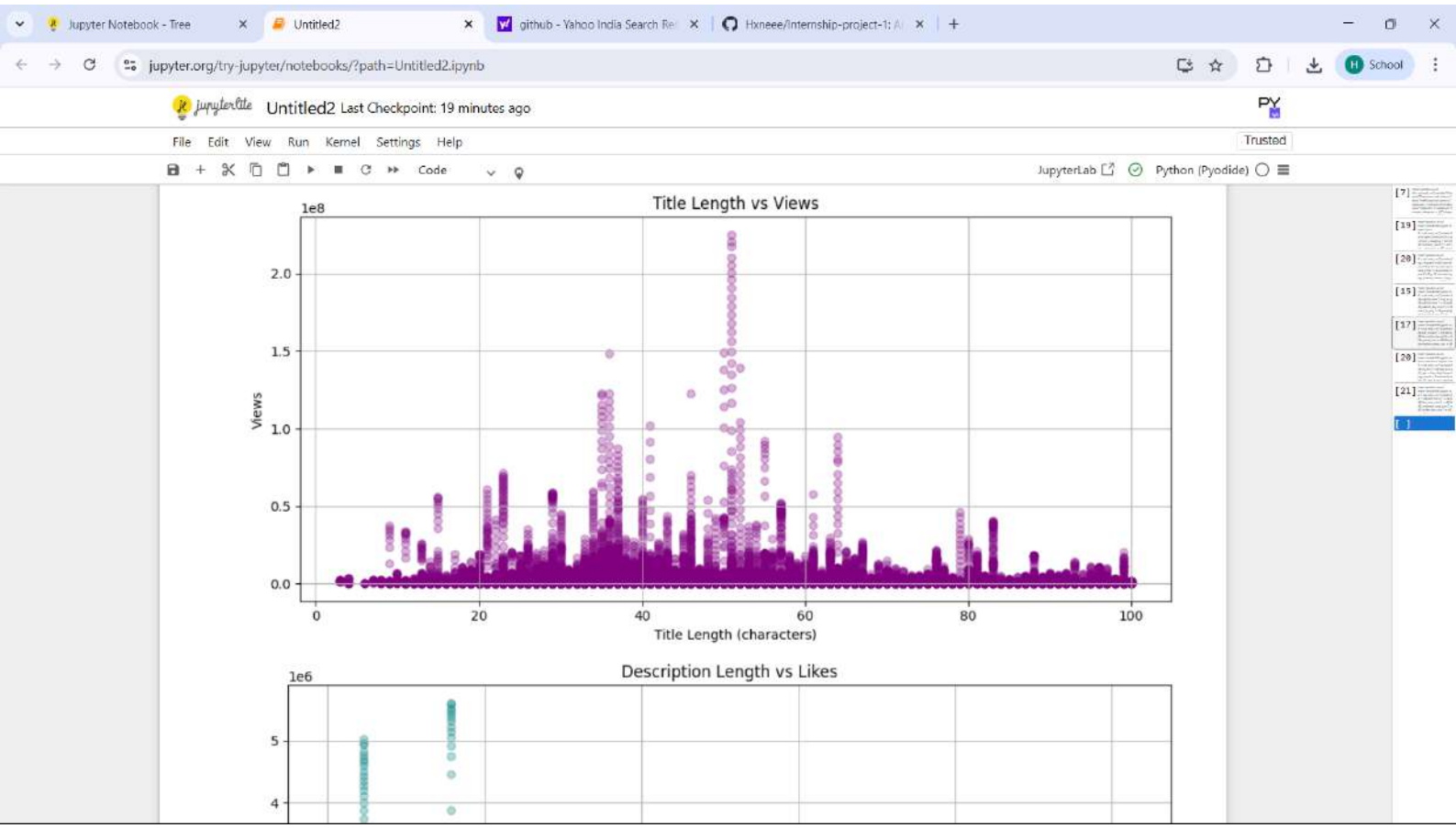
[17]:
```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("youtube/USvideos.csv")
df['title_length'] = df['title'].astype(str).apply(len)
df['description_length'] = df['description'].astype(str).apply(len)
title_views_corr = df['title_length'].corr(df['views'])
description_views_corr = df['description_length'].corr(df['views'])
description_likes_corr = df['description_length'].corr(df['likes'])

print("Correlation between title length and views:", round(title_views_corr, 3))
print("Correlation between description length and views:", round(description_views_corr, 3))
print("Correlation between description length and likes:", round(description_likes_corr, 3))
plt.figure(figsize=(10, 5))
plt.scatter(df['title_length'], df['views'], alpha=0.3, color='purple')
plt.title("Title Length vs Views")
plt.xlabel("Title Length (characters)")
plt.ylabel("Views")
plt.grid(True)
plt.tight_layout()
plt.show()

plt.figure(figsize=(10, 5))
plt.scatter(df['description_length'], df['likes'], alpha=0.3, color='teal')
plt.title("Description Length vs Likes")
plt.xlabel("Description Length (characters)")
plt.ylabel("Likes")
plt.grid(True)
plt.tight_layout()
plt.show()
```

```
Correlation between title length and views: -0.036
Correlation between description length and views: -0.014
Correlation between description length and likes: -0.014
```




Jupyter Notebook - Tree
Untitled2
github - Yahoo India Search Re
Hxneee/Internship-project-1: A
jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb
School
jupyterlite
Untitled2 Last Checkpoint: 19 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python (Pyodide)
Title Length vs Views
1e8
Views
Title Length (characters)
Description Length vs Likes
1e6

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

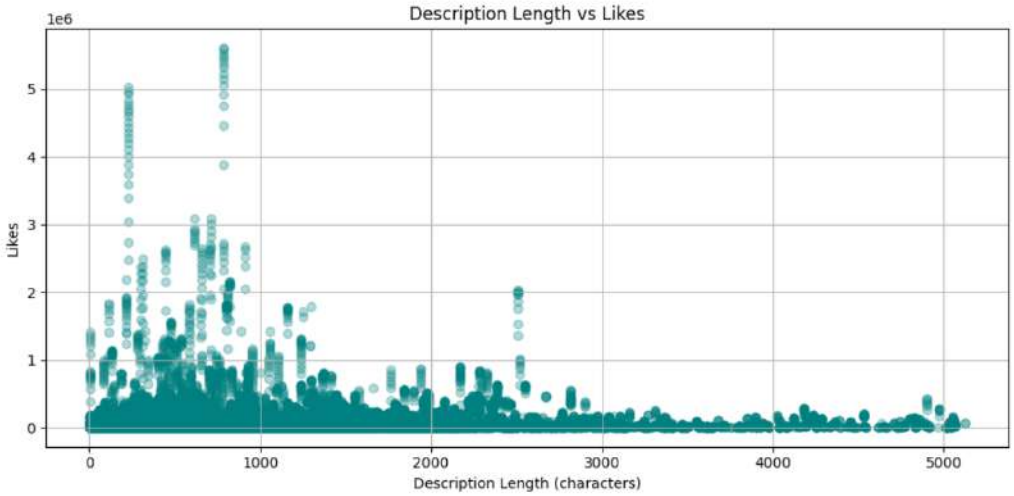
jupyterlite Untitled2 Last Checkpoint: 19 minutes ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab | Python (Pyodide)



[20]:
```
import pandas as pd
import matplotlib.pyplot as plt
from collections import Counter
df = pd.read_csv("youtube/USvideos.csv")
df['tag_list'] = df['tags'].astype(str).apply(lambda x: x.replace('"', '').split('|'))
all_tags = [tag.strip().lower() for sublist in df['tag_list'] for tag in sublist if tag.lower() != '[none]']
```

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

School

jupyterlite Untitled2 Last Checkpoint: 19 minutes ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab Python (Pyodide)

[20]:

```
import pandas as pd
import matplotlib.pyplot as plt
from collections import Counter
df = pd.read_csv("youtube/USvideos.csv")
df['tag_list'] = df['tags'].astype(str).apply(lambda x: x.replace('"', '').split('|'))
all_tags = [tag.strip().lower() for sublist in df['tag_list'] for tag in sublist if tag.lower() != '[none]']
tag_counts = Counter(all_tags)
top_50_tags = tag_counts.most_common(50)
tags, counts = zip(*top_50_tags[:20])
plt.figure(figsize=(12, 6))
plt.bar(tags, counts, color='skyblue')
plt.title("Top 20 Most Common Tags")
plt.ylabel("Frequency")
plt.xticks(rotation=75)
plt.tight_layout()
plt.show()
title_words = ' '.join(df['title'].astype(str).tolist()).lower().split()
title_word_counts = Counter(title_words)
top_20_title_words = title_word_counts.most_common(20)
words, freqs = zip(*top_20_title_words)
plt.figure(figsize=(12, 6))
plt.bar(words, freqs, color='lightcoral')
plt.title("Top 20 Words in Video Titles")
plt.ylabel("Frequency")
plt.xticks(rotation=75)
plt.tight_layout()
plt.show()

top_10_percent = df[df['views'] >= df['views'].quantile(0.9)]
top_tags_high_views = [tag.strip().lower() for sublist in top_10_percent['tag_list'] for tag in sublist if tag.lower() != '[none]']
top_tags_counter = Counter(top_tags_high_views)
top_20_tags_high_views = top_tags_counter.most_common(20)

print("Top 20 Tags Among Top 10% Videos by Views:")
for tag, count in top_20_tags_high_views:
    print(f"{tag}: {count}")
```

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

jupyterlite Untitled2 Last Checkpoint: 20 minutes ago

File Edit View Run Kernel Settings Help

Trusted

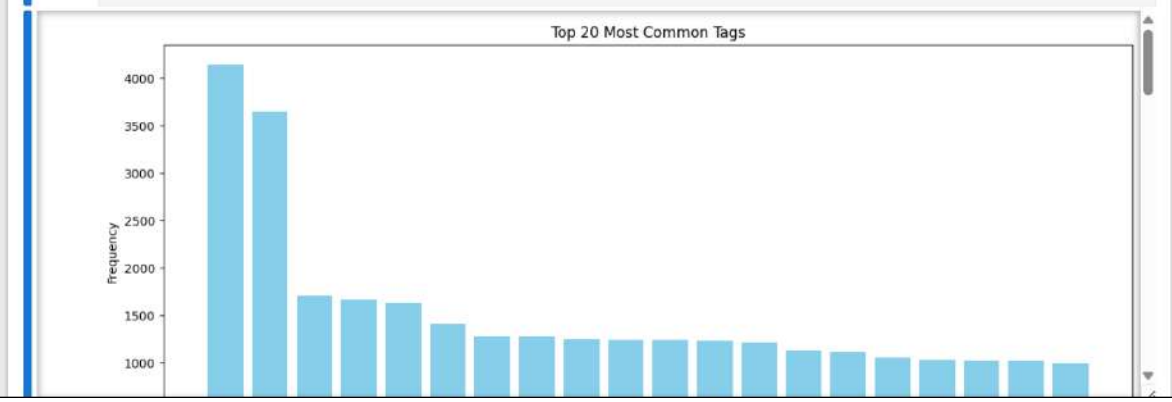
Code

JupyterLab | Python (Pyodide)

```
    print(f"{tag}: {count}")


df['num_tags'] = df['tag_list'].apply(len)
plt.figure(figsize=(10, 5))
plt.scatter(df['num_tags'], df['views'], alpha=0.3, color='green')
plt.title("Number of Tags vs Views")
plt.xlabel("Number of Tags")
plt.ylabel("Views")
plt.grid(True)
plt.tight_layout()
plt.show()

correlation = df['num_tags'].corr(df['views'])
print(f"\n Correlation between number of tags and views: {correlation:.3f}")
```



Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

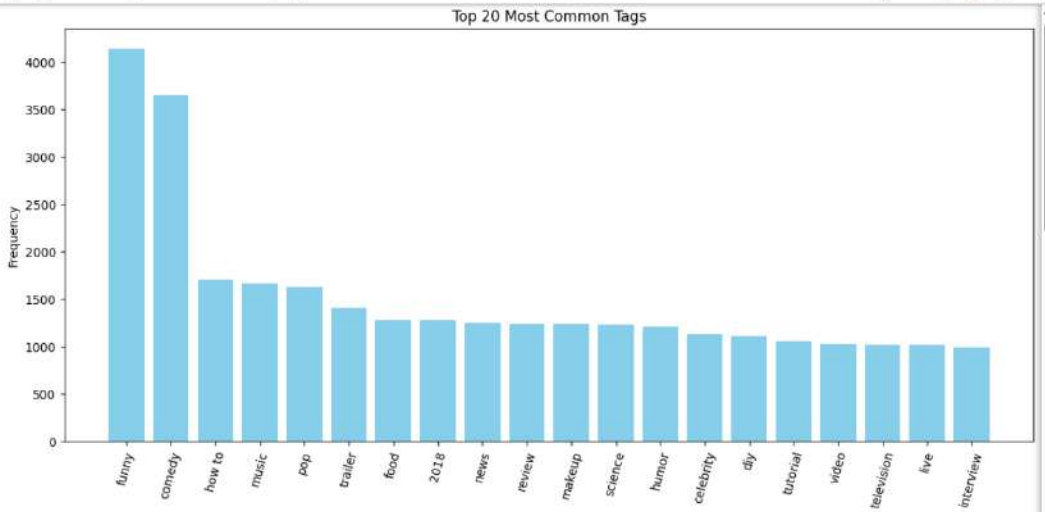
jupyterlite Untitled2 Last Checkpoint: 20 minutes ago

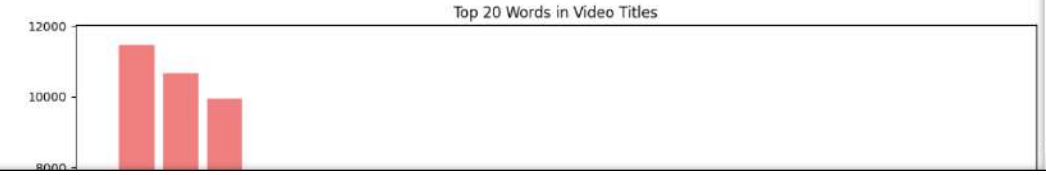
File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab | Python (Pyodide)





Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

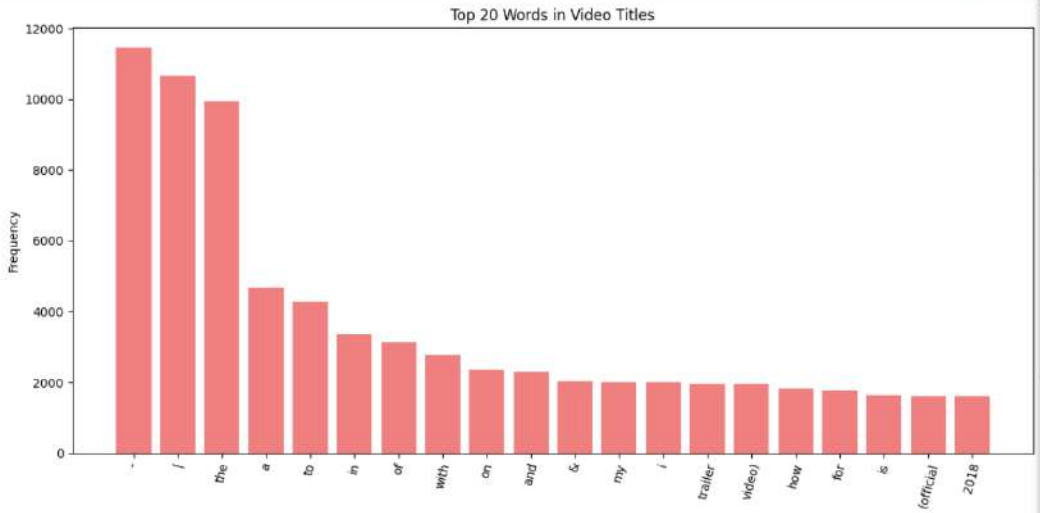
jupyterlite Untitled2 Last Checkpoint: 20 minutes ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab | Python (Pyodide)




```
Top 20 Tags Among Top 10% Videos by Views:
pop: 513
trailer: 322
funny: 278
comedy: 272
official: 213
animation: 189
family friendly: 183
bottle flip: 182
```

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

jupyterlite Untitled2 Last Checkpoint: 21 minutes ago

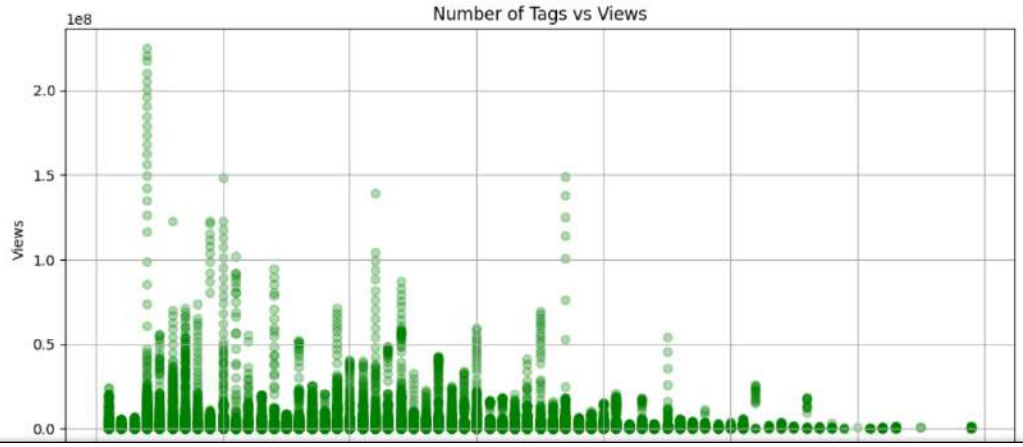
File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab Python (Pyodide)

```
family friendly: 183
bottle flip: 182
music video: 178
marvel: 174
family: 174
rap: 165
music: 165
records: 163
movie: 163
clean: 139
fidget spinners: 139
official video: 136
football: 135
rca records label: 134
```



Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

School

jupyterlite Untitled2 Last Checkpoint: 21 minutes ago

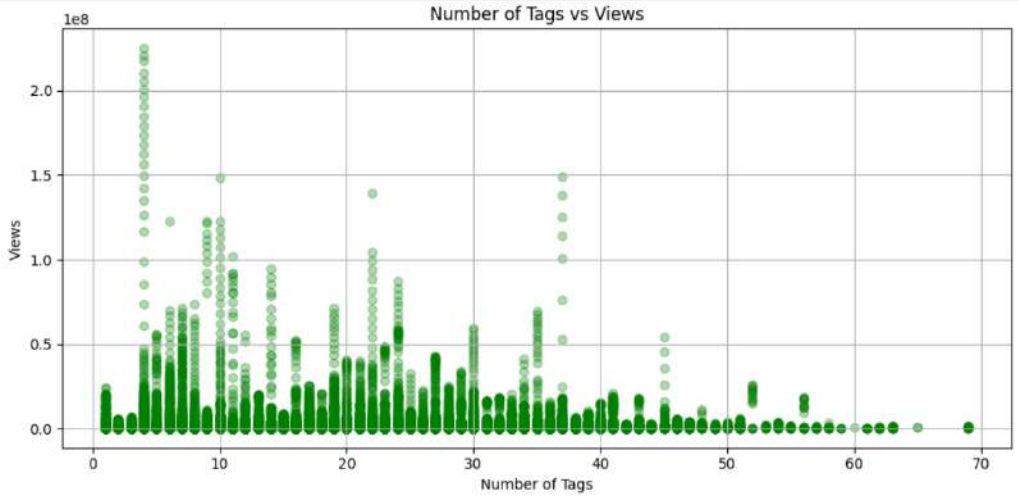
File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab Python (Pyodide)

```
print(f"\n Correlation between number of tags and views: {correlation:.3f}")
```




```
Correlation between number of tags and views: -0.029
```

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

School

jupyterlite Untitled2 Last Checkpoint: 21 minutes ago

File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab Python (Pyodide)

[21]:

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv("youtube/USvideos.csv")
df = df[(df['views'] > 0) & (df['likes'] > 0)]
df["like_view_ratio"] = df["likes"] / df["views"]
df["comment_view_ratio"] = df["comment_count"] / df["views"]
df["dislike_like_ratio"] = df["dislikes"] / df["likes"]
print("📊 Average Likes-to-Views Ratio:", df["like_view_ratio"].mean())
print("💬 Average Comments-to-Views Ratio:", df["comment_view_ratio"].mean())
print("👎 Average Dislikes-to-Likes Ratio:", df["dislike_like_ratio"].mean())
corr_likes = df["views"].corr(df["likes"])
corr_comments = df["views"].corr(df["comment_count"])
print(f"\n Correlation between views and likes: {corr_likes:.2f}")
print(f" Correlation between views and comments: {corr_comments:.2f}")

plt.figure(figsize=(8, 5))
plt.scatter(df["views"], df["likes"], alpha=0.3, color='blue')
plt.title(f"Views vs Likes (Correlation: {corr_likes:.2f})")
plt.xlabel("Views")
plt.ylabel("Likes")
plt.xscale("log")
plt.yscale("log")
plt.grid(True)
plt.tight_layout()
plt.show()

plt.figure(figsize=(8, 5))
plt.scatter(df["views"], df["comment_count"], alpha=0.3, color='green')
plt.title(f"Views vs Comments (Correlation: {corr_comments:.2f})")
plt.xlabel("Views")
plt.ylabel("Comments")
plt.xscale("log")
plt.yscale("log")
plt.grid(True)
plt.tight_layout()
plt.show()
```

Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

jupyterlite Untitled2 Last Checkpoint: 21 minutes ago

File Edit View Run Kernel Settings Help

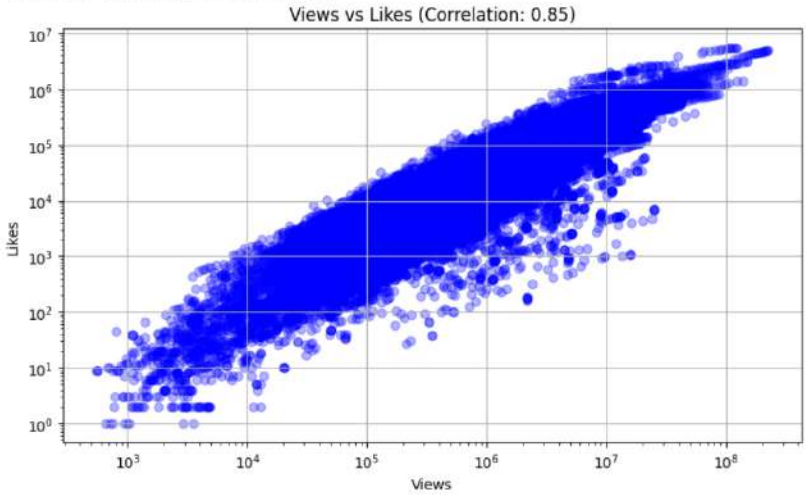
Trusted

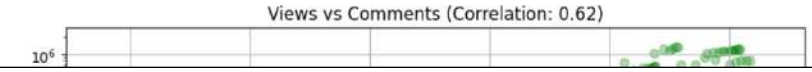
Code

JupyterLab | Python (Pyodide)

```
Average Likes-to-Views Ratio: 0.03455813287276936
Average Comments-to-Views Ratio: 0.00446419802582360l
Average Dislikes-to-Likes Ratio: 0.10676917522764086

Correlation between views and likes: 0.85
Correlation between views and comments: 0.62
```





Jupyter Notebook - Tree | Untitled2 | github - Yahoo India Search Re | Hxneee/Internship-project-1: A

jupyter.org/try-jupyter/notebooks/?path=Untitled2.ipynb

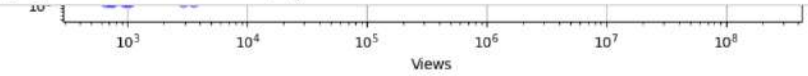
jupyterlite Untitled2 Last Checkpoint: 21 minutes ago

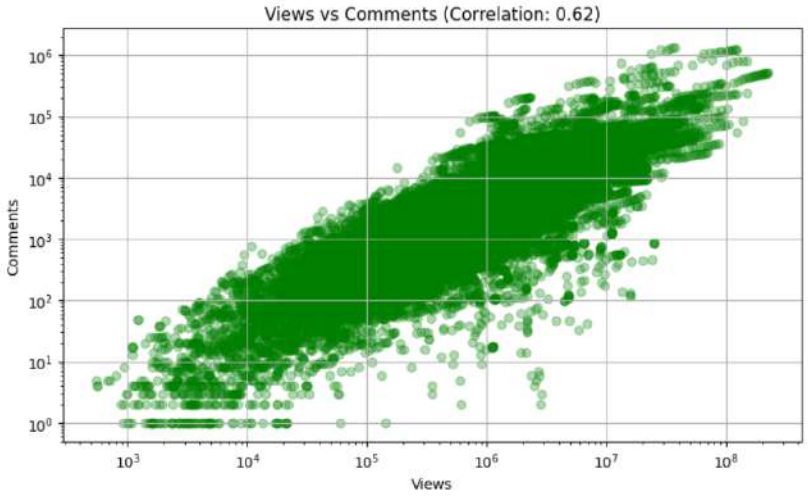
File Edit View Run Kernel Settings Help

Trusted

Code

JupyterLab Python (Pyodide)





[ ]: