Double-click (or enter) to edit

```
from google.colab import drive
drive.mount('/content/drive')
```

    Drive already mounted at /content/drive; to attempt to forcibly remount, c:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In this project, we conducted a comparative analysis of two popular subreddits, namely **NBA** and **Politics**. The aim was to explore and compare various features of the posts within these communities. Our analysis focused on the following key features:

1. **Score**: The score of the post, indicating its popularity based on upvotes and downvotes.
2. **Number** of Comments: The total number of comments on each post.
3. **Created UTC**: The UTC timestamp when each post was created.
4. **Author**: The username of the post's author.
5. **Is Self-Post**: A boolean value indicating whether the post is a self-post or a link to external content.

```
nba_file = '/content/drive/My Drive/Reddit/nba_data.csv'
nba_df = pd.read_csv(nba_file)
nba_df
```

	score	num_comments	created_utc	author	is_self
0	2	0	1.702972e+09	nutelamitbutter	False
1	5	15	1.702971e+09	ooCrimson	True
2	1	17	1.702970e+09	EGarrett	True
3	0	6	1.702970e+09	Substantial_Raise700	True
4	17	9	1.702969e+09	xpillindaass	True
...	...	...	...	...	...
976	927	356	1.702447e+09	Quality_Cucumber	True
977	139	25	1.702447e+09	Chris_Paul_3	True
978	54	61	1.702446e+09	CyborgAlgoInvestor	False
979	2305	639	1.702445e+09	nonamesleftadmin	False
980	330	49	1.702445e+09	swapan_99	True

981 rows × 5 columns

```
politics_file = '/content/drive/My Drive/Reddit/politics_data.csv'
politics_df = pd.read_csv(politics_file)
politics_df
```

	score	num_comments	created_utc	author	is_self
0	0	10	1.702968e+09	JaneDoeThought	False
1	15	7	1.702967e+09	marji80	False
2	7	17	1.702966e+09	Affectionate-Cup5202	False
3	21	4	1.702966e+09	tta2013	False
4	10	5	1.702964e+09	ParadeSit	False
...	...	...	...	...	...
988	447	118	1.702250e+09	nutritionvegan	False
989	2805	224	1.702248e+09	FloridaGirlNikki	False
990	3978	296	1.702246e+09	m0i0k0e0	False
991	0	87	1.702246e+09	redditor01020	False
992	1263	237	1.702246e+09	OwnBattle8805	False

993 rows × 5 columns

To provide a comprehensive overview of the communities, we analyzed the number of submissions in both the NBA and Politics subreddits during the year 2023. Our analysis revealed that the number of submissions in the NBA and Politics subreddits during 2023 was approximately the same. This suggests that both communities experienced a similar level of posting activity and maintained comparable levels of user engagement throughout the year.

```
nba_submissions = nba_df.shape[0]
politics_submissions = politics_df.shape[0]

bar_positions = np.arange(2)
bar_width = 0.35

fig, ax = plt.subplots(figsize=(8, 6))

bars = ax.bar(bar_positions, [nba_submissions, politics_submissions], bar_widt

ax.set_xticks(bar_positions)
ax.set_xticklabels(['NBA', 'Politics'])

ax.set_xlabel('Community')
ax.set_ylabel('Number of Submissions')
ax.set_title('Comparison of Number of Submissions\nbetween NBA and Politics Co
```
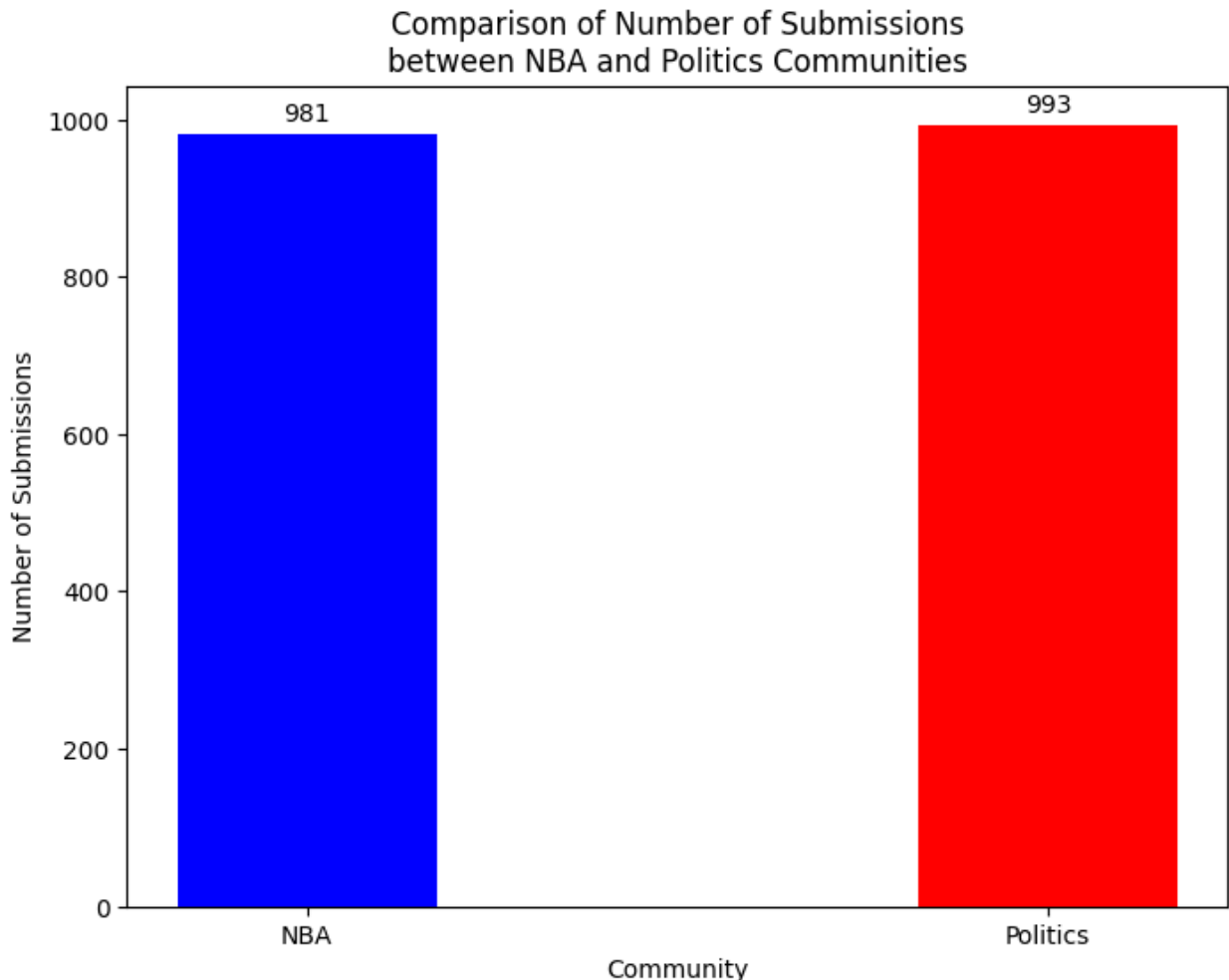
```
for bar in bars:
    height = bar.get_height()
    ax.annotate(height,
                xy=(bar.get_x() + bar.get_width() / 2, height),
                xytext=(0, 3),
                textcoords="offset points",
                ha='center',
                va='bottom')

plt.show()
```


Score Analysis

The score of a post serves as an indicator of its popularity within the community. We examined the average scores of posts in the NBA and Politics subreddits to assess the relative popularity of posts in each community.

Our analysis revealed that the average scores of posts in the Politics subreddit were consistently higher compared to those in the NBA subreddit. This suggests that posts in the Politics community tend to receive more upvotes and positive engagement, indicating a higher level of popularity within the community.

Number of Comments Analysis

The number of comments on a post provides insights into the engagement and discussion
generated by the community. We analyzed the average number of comments for posts in the
NBA and Politics subreddits to understand the level of interaction and participation within
each community.

Interestingly, we found that the average number of comments on posts in the Politics
subreddit was significantly higher than those in the NBA subreddit. This indicates that the
Politics community fosters more active discussions and engagement among its members,
resulting in a greater number of comments on average.

```
nba_comments = nba_df['num_comments']
politics_comments = politics_df['num_comments']
nba_score = nba_df['score']
politics_score = politics_df['score']

fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(12, 6))


community_labels = ['NBA', 'Politics']
comment_means = [nba_comments.mean(), politics_comments.mean()]
colors = ['blue', 'red']

bars1 = ax1.bar(community_labels, comment_means, color=colors)

ax1.set_xlabel('Community')
ax1.set_ylabel('Average Number of Comments')
ax1.set_title('Comparison of Average Number of Comments\nbetween NBA and Polit


for bar in bars1:
    yval = bar.get_height()
    ax1.text(bar.get_x() + bar.get_width() / 2, yval, f'{yval:.2f}', ha='cente


score_means = [nba_score.mean(), politics_score.mean()]

bars2 = ax2.bar(community_labels, score_means, color=colors)

ax2.set_xlabel('Community')
ax2.set_ylabel('Average Score')
ax2.set_title('Comparison of Average Score\nbetween NBA and Politics Communiti


for bar in bars2:
    yval = bar.get_height()
    ax2.text(bar.get_x() + bar.get_width() / 2, yval, f'{yval:.2f}', ha='cente

plt.subplots_adjust(wspace=0.4)

plt.show()
```
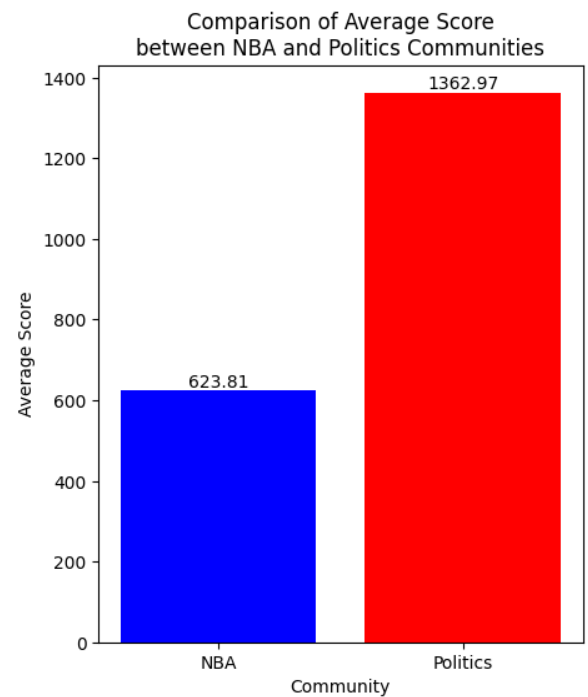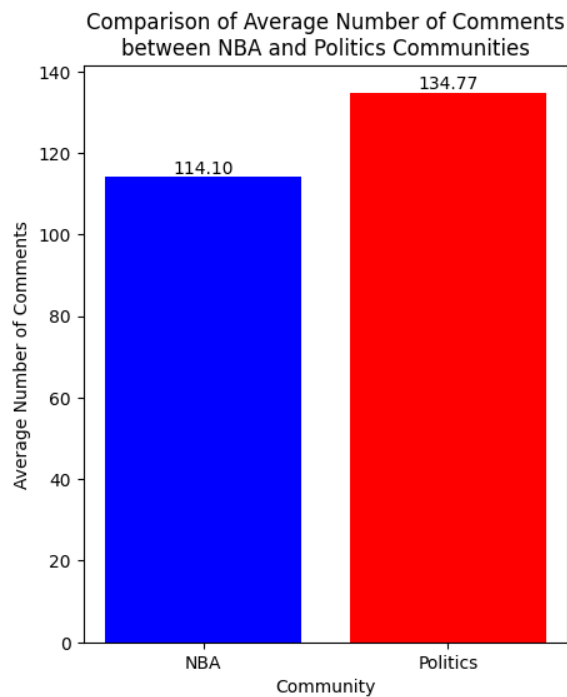
```
ptt.show()
```


Creation Timestamp Analysis

By examining the post frequency in both the NBA and Politics subreddits, we can identify trends and patterns in the posting behavior of each community.

Post Frequency by Hour Analysis

Analyzing the post frequency by hour provides insights into the specific time periods when users are most active within each subreddit. By examining the distribution of posts over the hours of the day, we can identify peak engagement hours and understand the posting patterns of the NBA and Politics communities.

Upon analyzing the post frequency by hour for both subreddits, we observed interesting patterns:

- NBA Subreddit: The post frequency in the NBA subreddit indicates that the peak engagement hours occur between 0-5 am. This suggests that users in the NBA community are particularly active during the late night and early morning hours. It is worth noting that this observation aligns with the time when NBA games are typically played in different time zones, leading to discussions, reactions, and updates being

played in different time zones, leading to discussions, reactions, and updates being shared during these hours.

- Politics Subreddit: In contrast, the post frequency in the Politics subreddit reveals a different pattern. The peak engagement hours for politics-related discussions and posts occur between 15-20 (3-8 pm). This indicates that users in the Politics community are most active during the late afternoon and early evening. This pattern could be attributed to users catching up on news, political events, or engaging in discussions after their work or daily activities.

These findings highlight the distinct engagement patterns within the NBA and Politics subreddits. The NBA subreddit experiences heightened activity during the late night and early morning hours, potentially driven by real-time game events, while the Politics subreddit sees increased engagement during the late afternoon and early evening, likely influenced by the availability of news updates and users' free time after work.

```python
nba_df['created_utc'] = pd.to_datetime(nba_df['created_utc'], unit='s')
politics_df['created_utc'] = pd.to_datetime(politics_df['created_utc'], unit='

nba_bins = nba_df['created_utc'].dt.hour
politics_bins = politics_df['created_utc'].dt.hour

fig, axs = plt.subplots(2, 1, figsize=(10, 8))
axs[0].hist(nba_bins, bins=24, alpha=0.5, label='NBA')
axs[1].hist(politics_bins, bins=24, alpha=0.5, label='Politics')
axs[0].set_title('NBA Post Frequency by Hour')
axs[1].set_title('Politics Post Frequency by Hour')
axs[0].legend()
axs[1].legend()
plt.tight_layout()
plt.show()
```
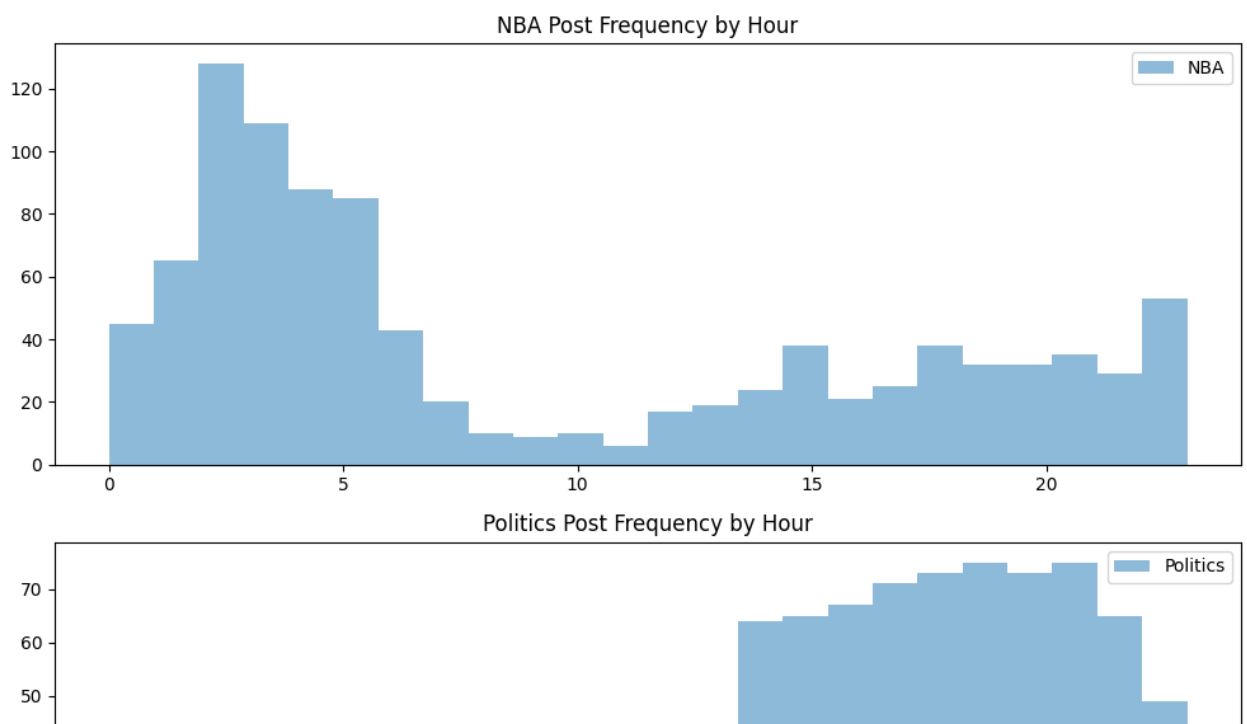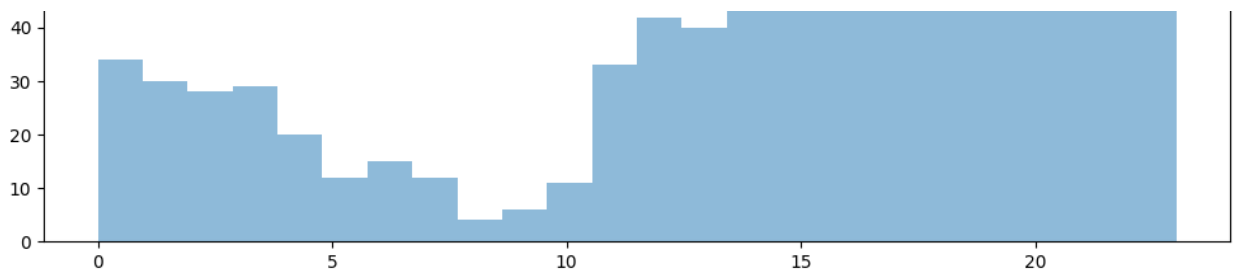
Analysis of Unique Contributors and Average Posts per Author

In this analysis, we examined the number of unique contributors and the average number of posts per author in the NBA and Politics subreddits. The following observations were made:

- NBA Subreddit: The NBA subreddit has a total of 448 unique contributors. This indicates that 448 different authors have contributed posts to the subreddit during the analyzed period. On average, each author in the NBA subreddit has made approximately 2.19 posts.

- Politics Subreddit: The Politics subreddit, on the other hand, has 414 unique contributors. This suggests that 414 different authors have contributed posts to the subreddit. The average number of posts per author in the Politics subreddit is approximately 2.4.

These findings highlight the level of participation and contribution within each subreddit. Both the NBA and Politics communities have a substantial number of unique contributors, indicating an active and diverse user base.

```
nba_unique_contributors = nba_df['author'].nunique()
politics_unique_contributors = politics_df['author'].nunique()

nba_avg_posts_per_author = len(nba_df) / nba_unique_contributors
politics_avg_posts_per_author = len(politics_df) / politics_unique_contributor

categories = ['NBA', 'Politics']
unique_contributors = [nba_unique_contributors, politics_unique_contributors]
```

```
avg_posts_per_author = [nba_avg_posts_per_author, politics_avg_posts_per_autho

fig, ax = plt.subplots()
ax.bar(categories, unique_contributors, color=['#0077B6', '#FF5733'])

for i, num in enumerate(unique_contributors):
    ax.text(i, num, str(num), ha='center', va='bottom', fontweight='bold')

for i, avg in enumerate(avg_posts_per_author):
    ax.annotate(f'Avg: {avg:.2f}', xy=(i, avg), xytext=(0, 10), ha='center', t

plt.xlabel('Community')
plt.ylabel('Number of Unique Contributors')
plt.title('Number of Unique Contributors and Average Posts per Author')

plt.show()
```
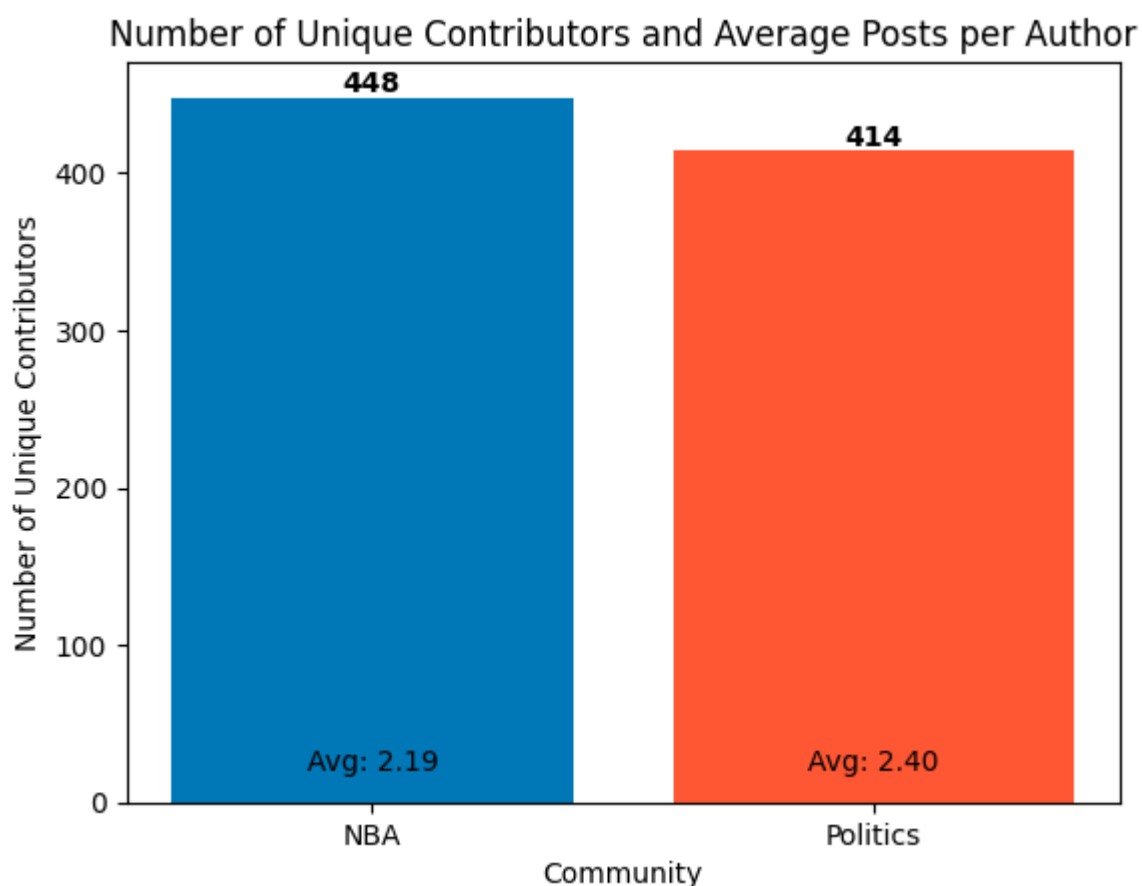

Comparison of Engagement Metrics: Self-posts vs. External Links

In this analysis, we compared the engagement metrics of self-posts and external links within the NBA and Politics subreddits. Self-posts are text-based posts created directly on Reddit, containing discussions, questions, personal experiences, or other textual content. On the other hand, external links include URLs to external websites or resources.

The following observations were made regarding the engagement metrics:

- Average Number of Comments: NBA Subreddit: Self-posts in the NBA subreddit receive an average of approximately 101.56 comments, whereas external links receive an

average of around 128.54 comments. This indicates that external links tend to generate slightly higher engagement in terms of the number of comments. Politics Subreddit: Self-posts in the Politics subreddit receive an average of about 31.25 comments, while external links receive an average of approximately 135.19 comments. Here, external links exhibit significantly higher engagement in terms of comments compared to self-posts.

- Average Score: NBA Subreddit: Self-posts in the NBA subreddit have an average score of approximately 382.47, while external links have an average score of around 901.67. This indicates that external links generally receive higher scores, suggesting greater approval and appreciation from the community. Politics Subreddit: Self-posts in the Politics subreddit have an average score of about 71.25, whereas external links have an average score of approximately 1368.20. Similar to the NBA subreddit, external links in the Politics subreddit demonstrate significantly higher scores compared to self-posts.

These findings suggest that external links, on average, tend to generate higher engagement in terms of both comments and scores compared to self-posts in both the NBA and Politics subreddits. External links attract more discussion and receive greater approval from the community.

It is important to consider the nature of the content and its format when analyzing engagement metrics. Self-posts, being text-based and created directly on Reddit, may foster in-depth discussions and personal experiences. On the other hand, external links provide access to external resources and websites, which can generate broader discussions and attract community interest.

```
nba_self_posts = nba_df[nba_df['is_self']]
politics_self_posts = politics_df[politics_df['is_self']]

nba_avg_comments_self = nba_self_posts['num_comments'].mean()
politics_avg_comments_self = politics_self_posts['num_comments'].mean()

nba_avg_score_self = nba_self_posts['score'].mean()
politics_avg_score_self = politics_self_posts['score'].mean()

nba_external_links = nba_df[~nba_df['is_self']]
politics_external_links = politics_df[~politics_df['is_self']]

nba_avg_comments_external = nba_external_links['num_comments'].mean()
politics_avg_comments_external = politics_external_links['num_comments'].mean(

nba_avg_score_external = nba_external_links['score'].mean()
politics_avg_score_external = politics_external_links['score'].mean()

bar_width = 0.35

plt.figure(figsize=(12, 7))
```

```
categories = ['NBA', 'Politics']
engagement_levels = ['Comments', 'Score']

bar_positions_1 = np.arange(len(categories))
bar_positions_2 = bar_positions_1 + bar_width

plt.subplot(1, 2, 1)
plt.bar(bar_positions_1, [nba_avg_comments_self, politics_avg_comments_self],
plt.bar(bar_positions_2, [nba_avg_comments_external, politics_avg_comments_ext
plt.xlabel('Community')
plt.ylabel('Average Number of Comments')
plt.title('Comparison of Comments between Self-posts and External Links')
plt.xticks(bar_positions_1 + bar_width/2, categories)
plt.legend()

for i, value in enumerate([nba_avg_comments_self, politics_avg_comments_self])
    plt.text(i, value, str(round(value, 2)), ha='center', va='bottom', color='

for i, value in enumerate([nba_avg_comments_external, politics_avg_comments_ex
    plt.text(i + bar_width, value, str(round(value, 2)), ha='center', va='bott

plt.subplot(1, 2, 2)
plt.bar(bar_positions_1, [nba_avg_score_self, politics_avg_score_self], bar_wi
plt.bar(bar_positions_2, [nba_avg_score_external, politics_avg_score_external]
plt.xlabel('Community')
plt.ylabel('Average Score')
plt.title('Comparison of Scores between Self-posts and External Links')
plt.xticks(bar_positions_1 + bar_width/2, categories)
plt.legend()

for i, value in enumerate([nba_avg_score_self, politics_avg_score_self]):
    plt.text(i, value, str(round(value, 2)), ha='center', va='bottom', color='

for i, value in enumerate([nba_avg_score_external, politics_avg_score_external
    plt.text(i + bar_width, value, str(round(value, 2)), ha='center', va='bott

plt.tight_layout()
plt.show()
```