Report of EDA process

Dataset Overview:

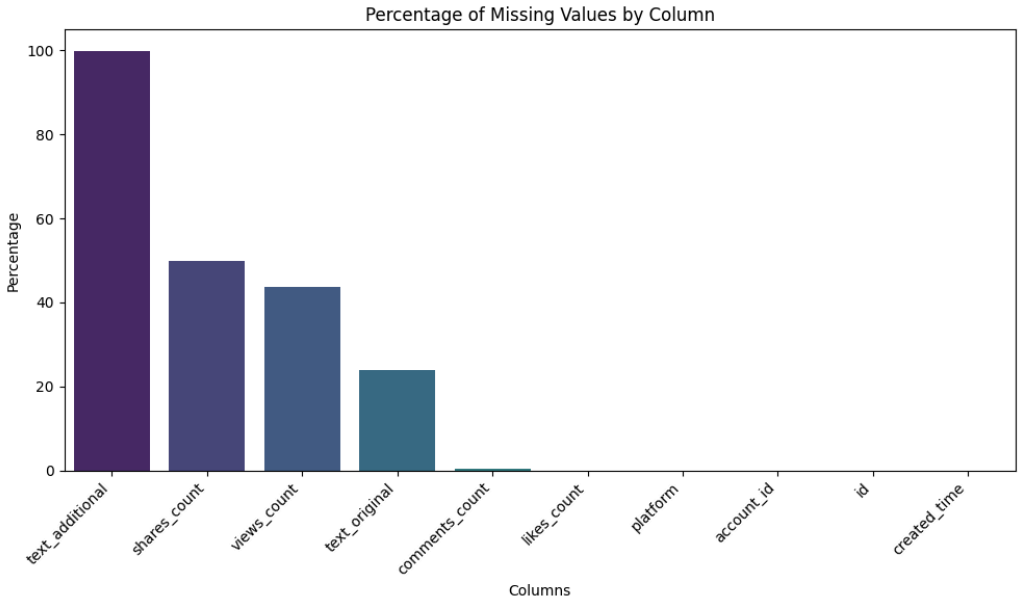
The dataset consists of 10,000 records with the following columns:

- platform: Social media platform.
- account_id: Unique identifier for the account.
- id: Post identifier.
- created_time: Timestamp when the post was created.
- text_original: Original post text.
- text_additional: Additional text, mostly missing.
- likes_count: Number of likes.
- shares_count: Number of shares.
- comments_count: Number of comments.
- views_count: Number of views.

The dataset contains 4 platforms: **Instagram, TikTok, Facebook, and YouTube.**

1. Basic Information:

Visualization for Missing Values:



Graph 1: Missing Values Bar Plot

- A **bar plot** showing the percentage of missing values per column was created to visualize the extent of missing data.
- The dataset contains 10,000 entries, and the columns include both numerical and categorical variables. The `created_time` column is correctly parsed into datetime format.
- Missing values were identified, and the percentage of missing values per column was computed.
- **Missing Values Summary:**

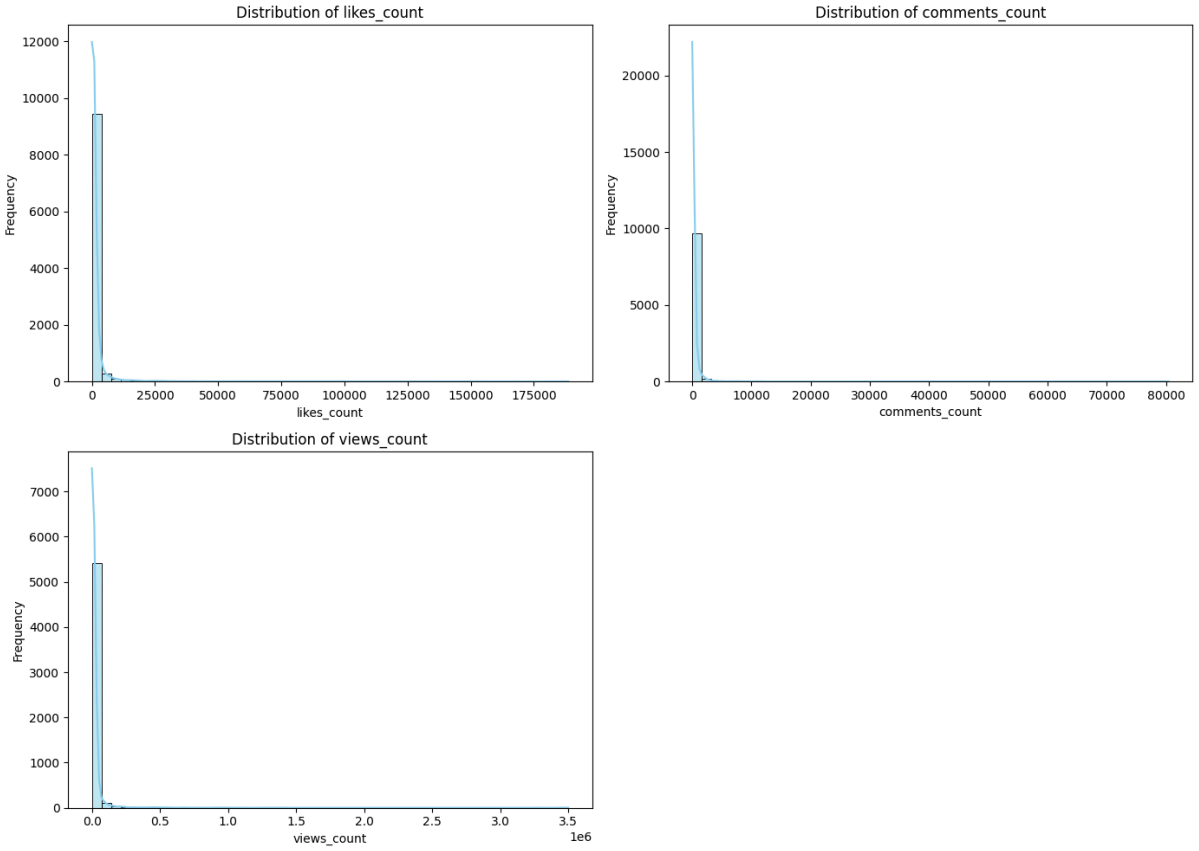
- `text_original` has 23.87% missing values.
- `text_additional` has 99.97% missing values.
- `shares_count`, `comments_count`, and `views_count` have smaller proportions of missing values.

2. Summary Statistics for Numerical Columns:

- **Likes Count:** The average number of likes per post is 1,416, with a maximum of 188,611.
- **Shares Count:** The average shares count is 79, with a maximum of 47,500.
- **Comments Count:** The average number of comments per post is 300, with a maximum of 80,415.
- **Views Count:** The average number of views per post is 17,973, with a maximum of 3.5 million views.

3. Platform Distribution:

- The dataset contains four platforms, each accounting for 25% of the records: **Instagram**, **TikTok**, **Facebook**, and **YouTube**.

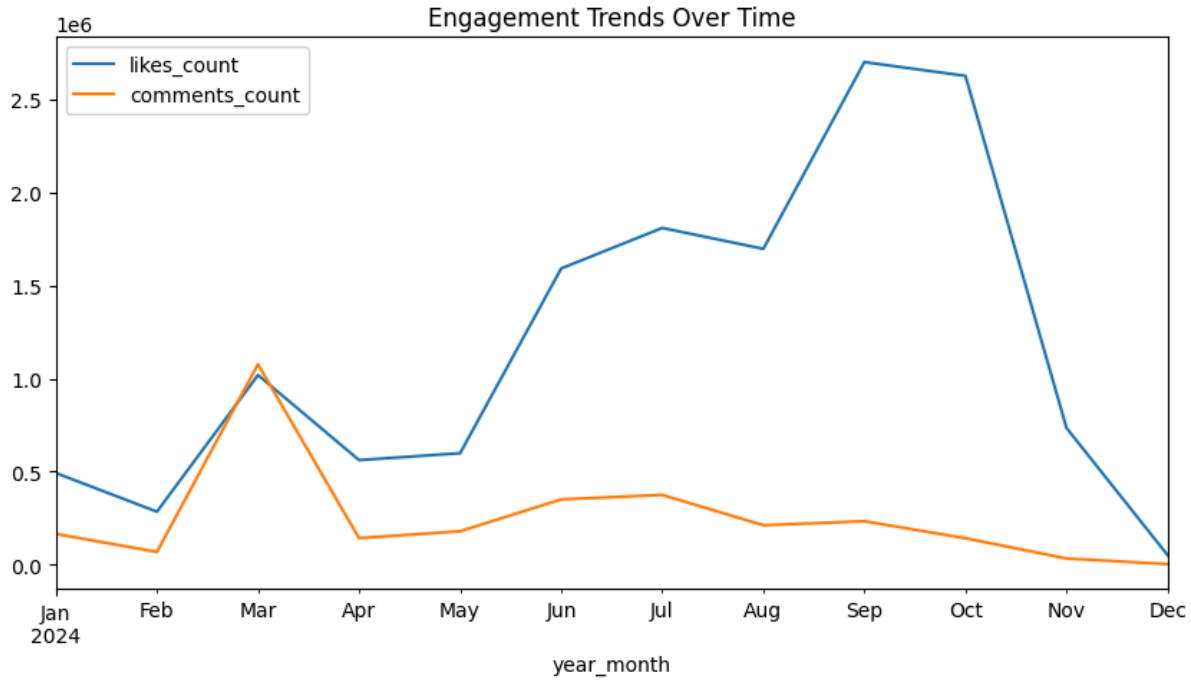


Graph 2: Platform Distribution Pie Chart

The histograms for likes_count, comments_count, and views_count show that these metrics are highly skewed, with a small number of posts receiving a large number of likes, comments, and views. This suggests the presence of outliers or a few highly engaging posts skewing the data.

4. Relationship Between Engagement Metrics:

- **Likes vs Comments:** Another scatter plot indicates that posts with higher likes also tend to have higher comments, but this relationship is more spread out.



Graph 3: Likes vs Comments

This graph visualizes the trends of likes_count and comments_count over the months of the year 2024. Here are the key observations:

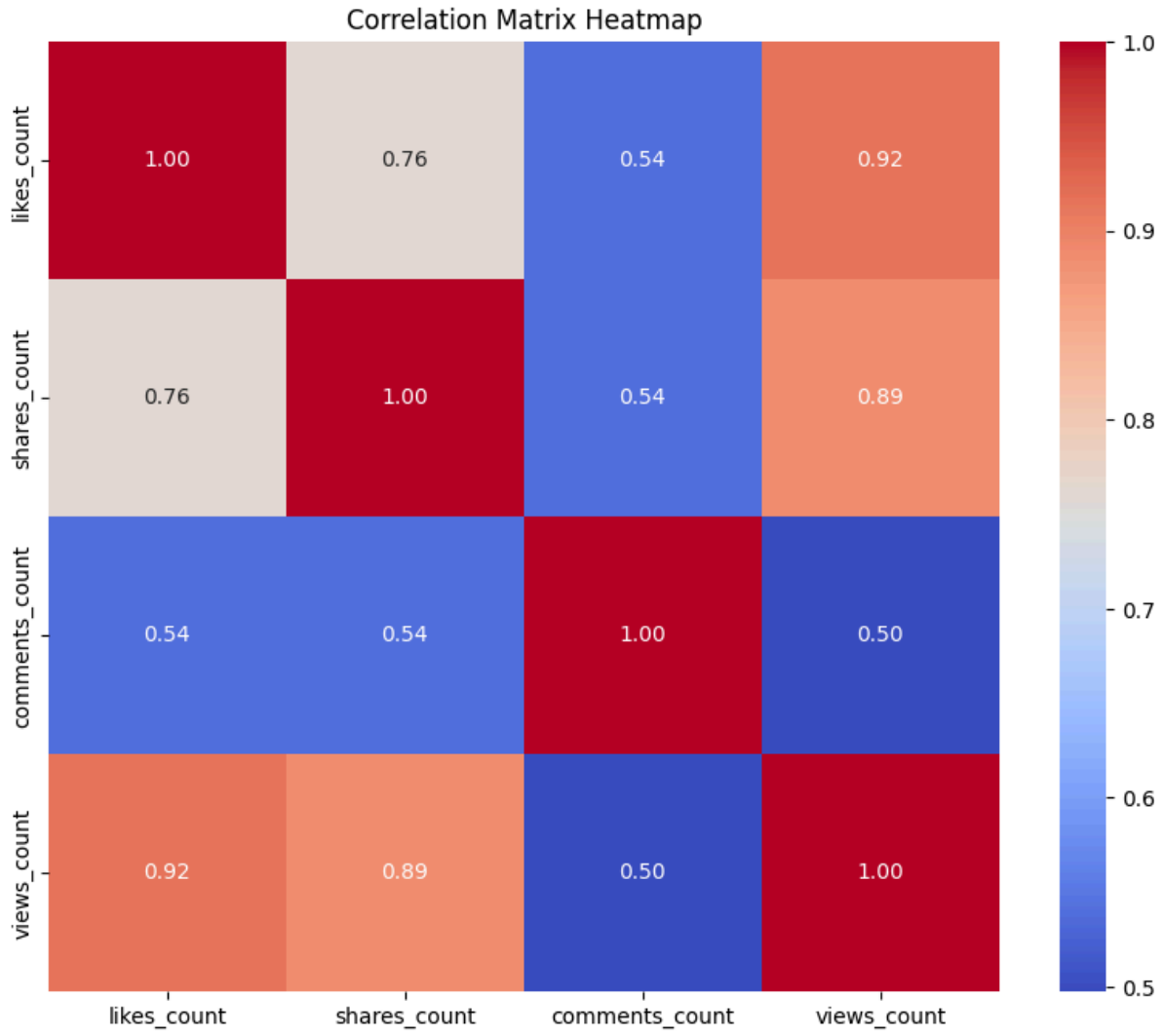
- Likes Trend (Blue Line): The number of likes saw notable spikes during the months of March and September, with a sharp decline towards the end of the year in November and December.
- Comments Trend (Orange Line): Similar to likes, comments also peaked during March and September, but the general trend shows a sharp decline after June, especially after August. This suggests that engagement (both likes and comments) was high earlier in the year but declined significantly toward the end.

Insights:

1. Peak Engagement Period: Both likes and comments were significantly higher in March and September, which might indicate specific events, campaigns, or trends that led to increased interaction during these months.
2. Decline Toward End of Year: Both metrics showed a clear decline in November and December, suggesting lower engagement towards the end of the year, possibly due to seasonal variations or changes in user behavior.
3. Possible Action: Further analysis could be done to investigate the cause of the spikes in engagement in March and September, such as particular events or campaigns that may have been running during these months.

5. Correlation Matrix:

- The **correlation matrix heatmap** illustrates the relationships between numerical columns:
    - There is a moderate positive correlation between **likes_count** and **comments_count**.
    - A strong positive correlation exists between **views_count** and **likes_count**, suggesting that posts with more views tend to get more likes.
    - Other correlations are weaker, indicating more independent relationships between the metrics.



Insert Graph 4: Correlation Matrix Heatmap

6. Key Insights:

1. **Total Records:** The dataset contains 10,000 records, representing content from four major platforms: Instagram, TikTok, Facebook, and YouTube.
2. **Engagement Metrics:**
    - **Total Likes:** 14,163,586
    - **Total Comments:** 2,984,448
    - **Total Views:** 101,029,206
    - **Average Likes per Post:** 1,416.64

    - **Average Comments per Post:** 299.79
    - **Average Views per Post:** 17,973.53
3. **Most Common Platform: Facebook** is the most common platform in the dataset, accounting for 25% of the records.
4. **Max Engagement Post:** The post with the highest number of likes was from **YouTube**, indicating that the platform might be contributing to higher engagement metrics.

Next Steps:

- **Advanced Analysis:** Further analysis could explore the impact of platform type on engagement metrics, detect trends in specific periods, or build predictive models for engagement forecasting.