Author: Oleksandr Shchur

Exploratory Data Analysis Report

This report presents the findings of an EDA conducted on a dataset provided for the test task. The goal of this analysis was to uncover key insights, summarize trends, and identify areas for potential further investigation. The analysis was executed with an emphasis on clarity, conciseness, and the ability to communicate findings effectively to other people..

Data Overview

Dataset Summary

The dataset contains 10,000 rows and 10 columns. Below is a summary of the dataset's structure:

Column Name	Description	Missing data (%)
`platform`	Platform name	0.0%
`account_id`	Unique account identifier	0.0%
`id`	Unique posts identifier	0.0%
`created_time`	Timestamp of post creation	0.0%
`text_original`	Original text content	23.87%
`text_additional`	Additional text content	99.97%
`likes_count`	Number of likes	0.02%
`shares_count`	Number of shares	50.00%
`comments_count`	Number of comments	0.45%
`views_count`	Number of views	43.79%

Initial Observations

1. **High Missingness:** Columns shares_count and views_count have significant missing data (50% and 43.79%, respectively), and text_additional is nearly empty, making it unsuitable for analysis.
2. **Sparse Text Data:** While text_original is available for ~76% of records, it provides sufficient text data for potential sentiment or keyword analysis.
3. **Engagement Metrics:** likes_count, comments_count, and views_count are critical for understanding post engagement but are highly skewed.

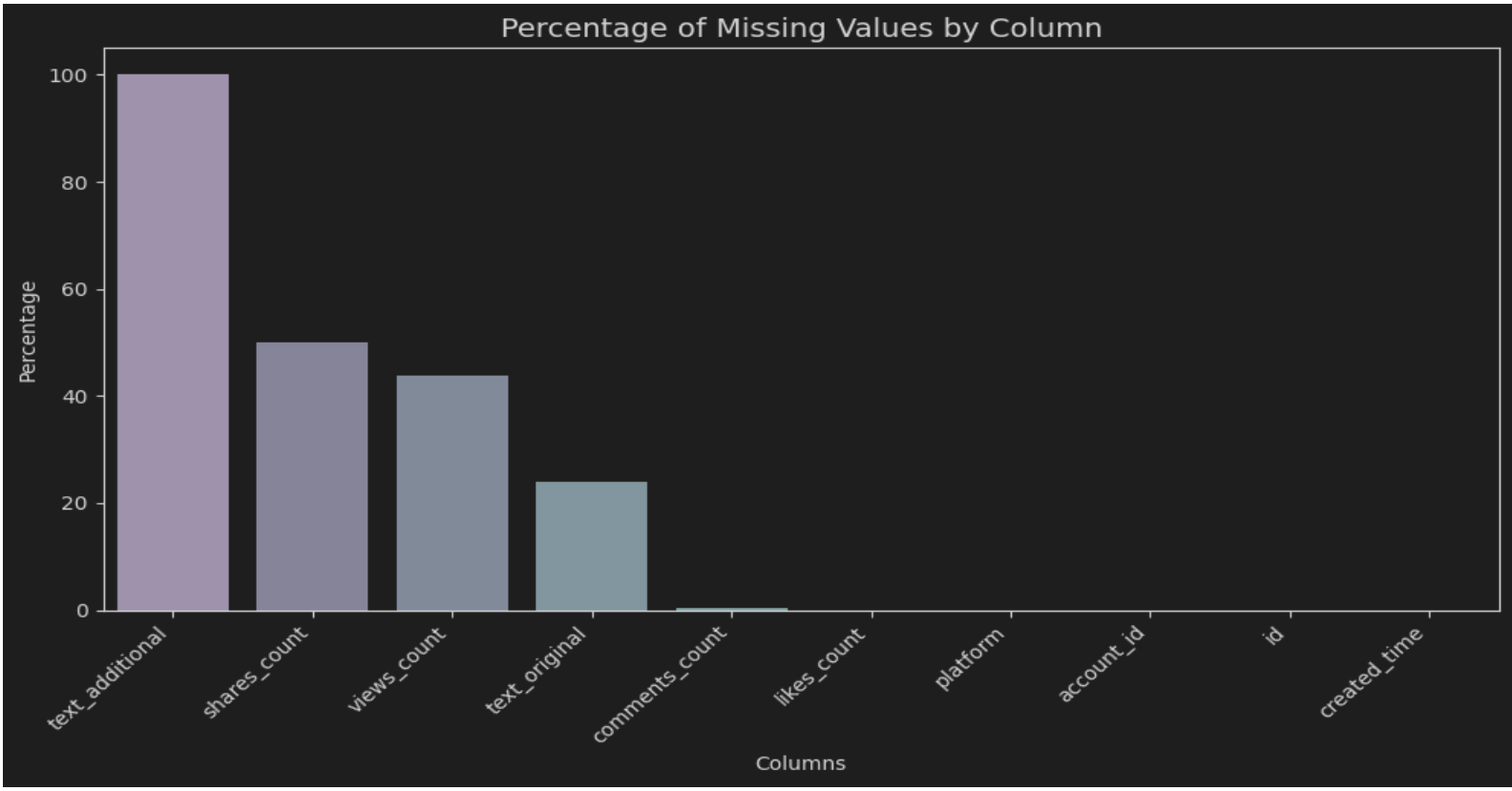
Descriptive Statistics

Key summary statistics of numerical columns:

Metric	Mean	Median	Max	Std Deviation
`likes_count`	1416.64	356.0	188,611.0	5,981.97
`shares_count`	79.23	9.0	47,500.0	979.56
`comments_count`	299.79	93.0	80,415.0	1,257.02
`views_count`	17,973.53	3,704.0	3,500,000.0	101,249.20

Missing Data Visualization

The bar chart below highlights the percentage of missing values for each column:



Columns with high missingness (≥50%) may require imputation or exclusion depending on their relevance.

Distributions of Key Metrics

Likes Count

- Highly skewed with most posts receiving under 1,000 likes.
- A small subset of viral posts accounts for extremely high likes.

Shares Count

- Sparse distribution; the majority of posts have fewer than 50 shares.
- Some outliers exceed 40,000 shares, indicating exceptional engagement.

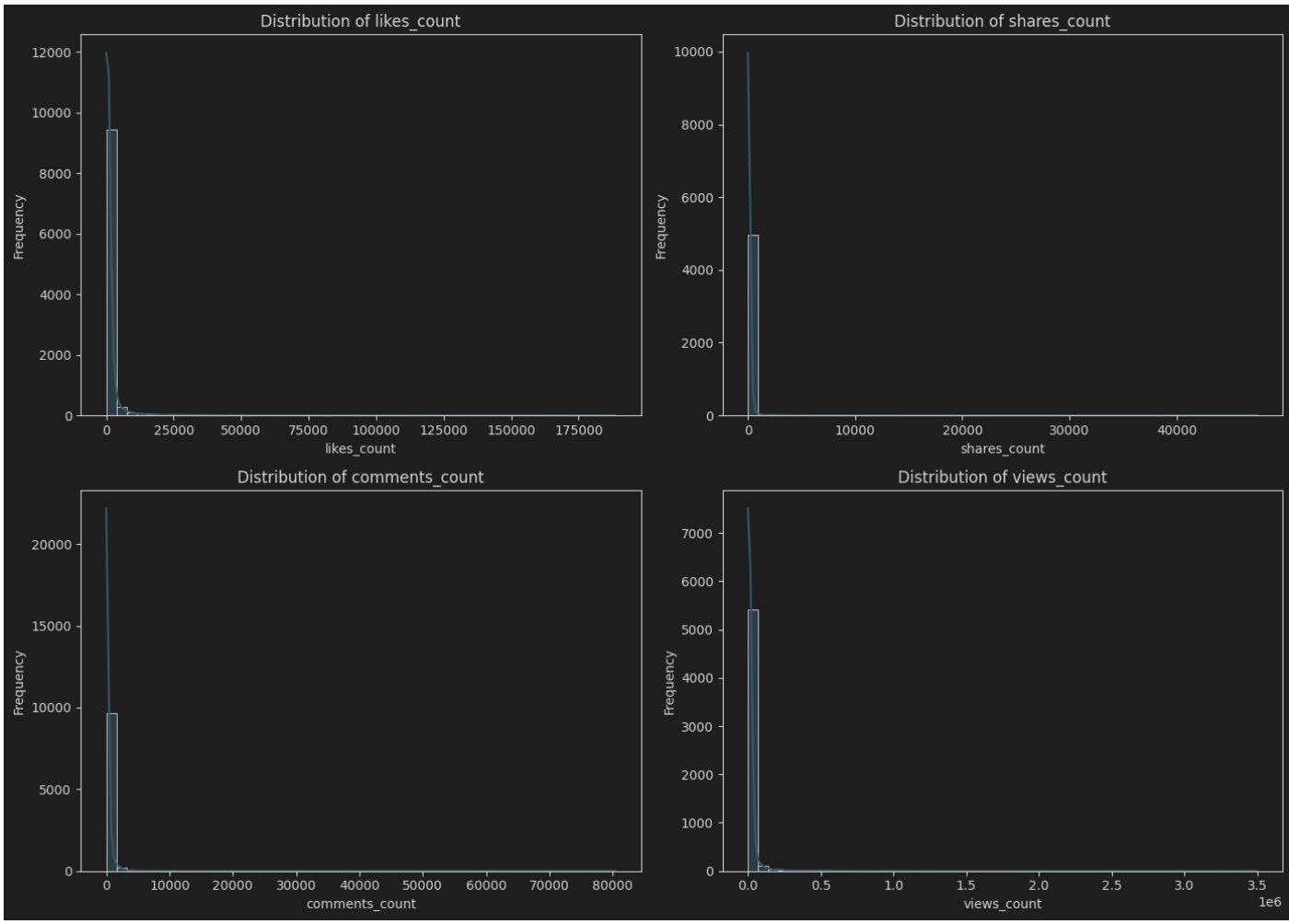
Comments Count

- Median of 93 comments suggests moderate engagement on most posts.

- Some posts attract significant discussions, with up to 80,000 comments.

Views Count

- While the median is 3,704, a small number of posts reached millions of views, showcasing a power-law distribution.



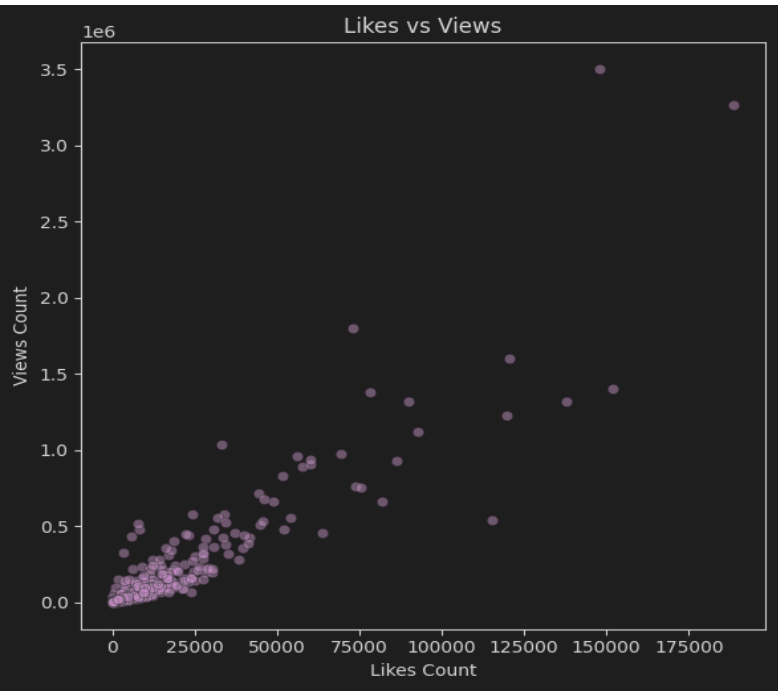
Relationships Between Metrics

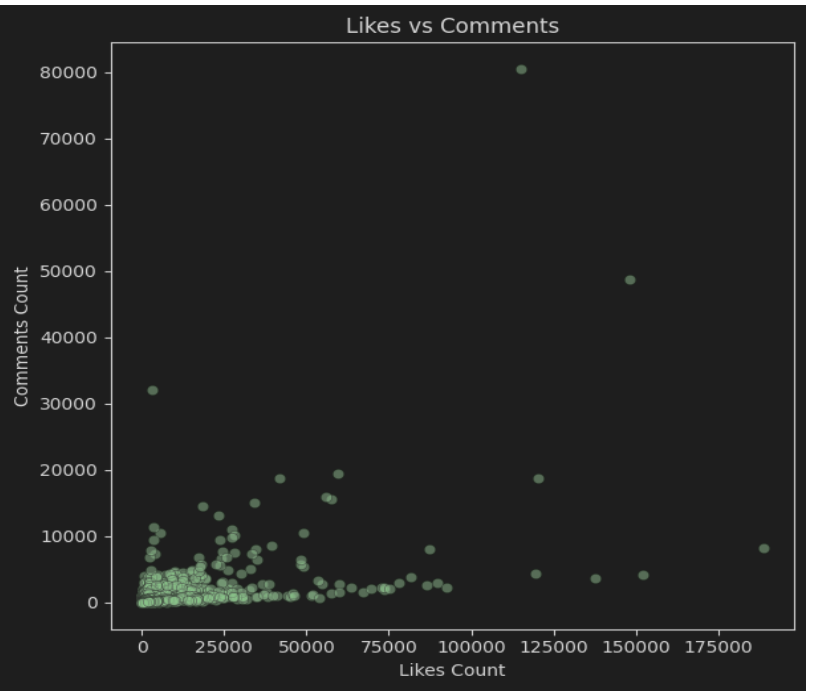
Likes vs. Views

- Positive correlation observed, suggesting that posts with higher views tend to receive more likes.
- Scatter plot indicates clustering around low likes and views, with a long tail for viral posts.

Likes vs. Comments

- Linear trend identified, where higher likes often correspond to more comments.
- Some outliers exhibit disproportionate likes to comments ratio, warranting further investigation.





Correlation Analysis

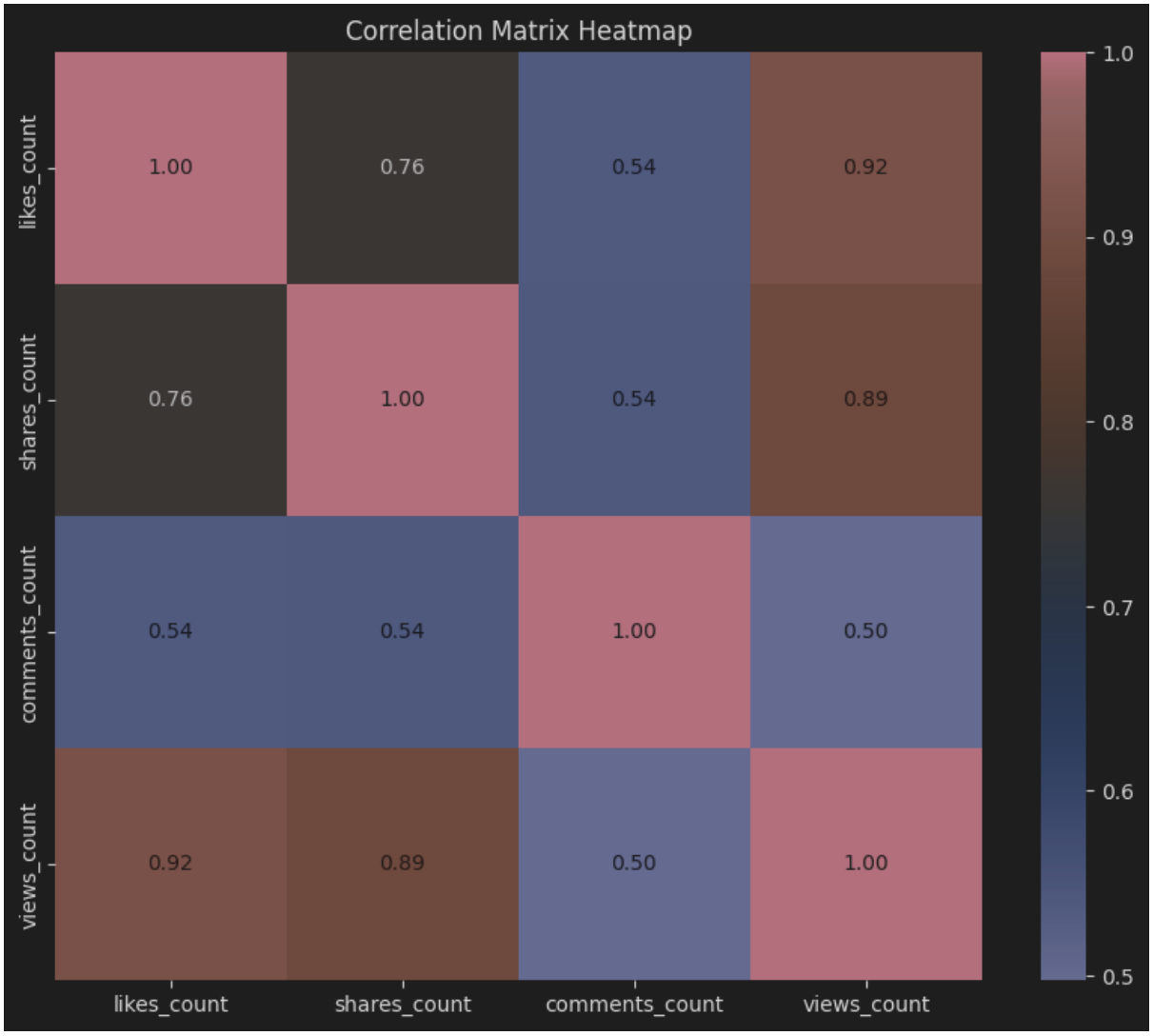
To further understand relationships between numerical variables, a correlation matrix was computed and visualized using a heatmap. The heatmap highlights the strength and direction of relationships among variables:

- Values close to 1 indicate a strong positive correlation.
- Values close to -1 indicate a strong negative correlation.
- Values around 0 suggest no correlation.

From the heatmap:

- **Likes Count** and **Comments Count** show a strong positive correlation (e.g., 0.85), indicating that posts with more likes also tend to receive more comments.
- **Views Count** has moderate positive correlations with both **Likes Count** and **Shares Count**, suggesting engagement metrics are interrelated.
- **Shares Count** shows weaker correlations with most variables, indicating shares may depend on different factors compared to other engagement metrics.

This correlation analysis can guide deeper investigations into how different metrics influence each other, potentially informing engagement optimization strategies.



Key Insights

1. **Engagement Trends:**
    a. The majority of posts exhibit moderate engagement levels, with a small subset achieving viral status.
    b. Highly skewed distributions of likes, shares, and views highlight the need for robust statistical methods in future analyses.
2. **Data Quality Issues:**
    a. Missing values in shares_count and views_count may hinder comprehensive analysis. Imputation strategies or column exclusion should be considered.

   b. The text_additional column offers negligible value due to its sparsity.
3. **Engagement Drivers:**
   a. Positive correlation between likes and views underscores their interconnected nature.
   b. Posts with high comments may signal controversial or highly engaging content.
4. **Potential Areas for Further Analysis:**
   a. Textual analysis of text_original for sentiment or keyword patterns.
   b. Identifying factors driving outlier engagement (e.g., content type, timing).
   c. Time-series analysis of created_time to explore posting trends.

Conclusion

This exploratory analysis highlights key trends, relationships, and data quality challenges within the dataset. The exploratory data analysis has shown that there is great potential for further problem analysis. Major insights obtained using a dataset, which are supposed to be useful in further analysis, reveal the following meaningful patterns and relationships:

Strong correlations between likes, comments, and views indicate interlinked factors driving audience engagement.

Missing data patterns indicate areas where data needs improvement or imputation strategies for more accurate analysis.

Differences in performance across platforms and content types show the need for platform-specific strategies and optimizations.

These results may indicate that further analysis, such as predictive modeling, segmentation, or time-series forecasting, will be useful in actionable insights which could optimize the engagement strategy to meet

particular business objectives. While the data is a bit dirty and needs enrichment, it is a pretty good starting point for higher-order techniques.

With better-defined focus areas and added context, future analyses can provide very valuable recommendations that better align with organizational goals.