Analyzing User Engagement with News Articles Using A/B Testing

Project Overview and Objective:

In this project, news articles are extracted from NewsAPI across three categories: Health, Technology, and Entertainment. The objective is to evaluate how user interaction with these articles is influenced by their time of publication.

To achieve this, the articles are divided into two groups based on their publication times:

- Group A: Articles published between 6:00 AM and 12:00 PM
- Group B: Articles published between 6:00 PM and 12:00 AM

Simulated engagement metrics such as views and likes are generated for each article. Using these metrics, along with article features such as title length, publish time, and category, statistical and visualization techniques were used to identify patterns and insights regarding user behaviour and content performance across time window.

Methodology:

Data Extraction

News articles across multiple categories in the US were extracted using the NewsAPI. Due to the limitations of free access, a maximum of 100 articles per category could be retrieved in a single request.

Dataset Description

Columns:

- **Source:** The news organization that published the article.

- **Author:** The author of the article.
- **Title:** The headline of the article.
- **Category:** The news category (Health, Technology, or Entertainment).
- **Description:** A brief summary of the article.
- **URL:** Link to the full article.
- **URLToImage:** Link to an image related to the article.
- **PublishedAt:** The date and time the article was published.
- **Content:** The main body/content of the article.

Data Preparation

After retrieving and loading the articles, the dataset was cleaned and prepared for analysis:

- **Handling Missing Values:** Some columns contained missing values, which were addressed using appropriate techniques.
- **Duplicates:** No duplicate records were found in the dataset.
- **Standardization:** Inconsistencies in formatting and input were corrected to ensure uniformity.
- **Data Type Conversion:** The "PublishedAt" column was converted from an *object* to a *datetime* format, as it represents the date and time of publication. ‘Source’ and ''Category’ were also changed from *object* to *category* data type.
- **Hour Extraction:** The publication time was extracted in 24-hour format from the "PublishedAt" column and stored in a new column named "Hour of Publication".
- **Time-Based Grouping:** Articles were categorized into two groups: Group A (articles published between 6:00 AM and 12:00 PM) and Group B (articles published between 6:00 PM and 12:00 AM).

- **Article Length:** The length of each article was calculated to facilitate content analysis.
- **Synthetic Engagement Data:** Since the dataset lacked user interaction metrics, simulated values for views and likes were generated for analysis purposes**.**
- **Filter Data in Relevant Groups:** The dataset was filtered to include only articles published within the Group A and Group B time windows.

Analysis And Observation

Descriptive Statistics

To better understand the data and grab insights, descriptive statistic was employed:

- Summarized engagement metrics and article count using pandas functions like .info(), .describe() and .groupby()
- Calculated aggregates like **mean, std, median, min, max, and count** using pandas
- Identified categories with the highest interaction levels

Group Comparisons

- Grouped articles by **source, category, and publication time**
- Compared **Group A vs Group B** in terms of:
    - Number of publications
    - Average engagement (likes, views)
    - Top performing categories

Correlation Analysis

- Checked relationships between **article count, likes, and views**
- Used correlation heatmaps and scatterplots for visualization
- Scatterplot to view relationship between **likes** and **views**

Statistical Testing

Performed a **two-sample t-test** to determine whether there was a **statistically significant difference** in engagement metrics between Group A and Group B.

- **Null Hypothesis (H_0):** There is no significant difference in engagement between Group A and Group B.
- **Alternative Hypothesis (H_1):** There is a significant difference in engagement between Group A and Group B.

Data Visualization

Used matplotlib and seaborn to create visual insights, including:

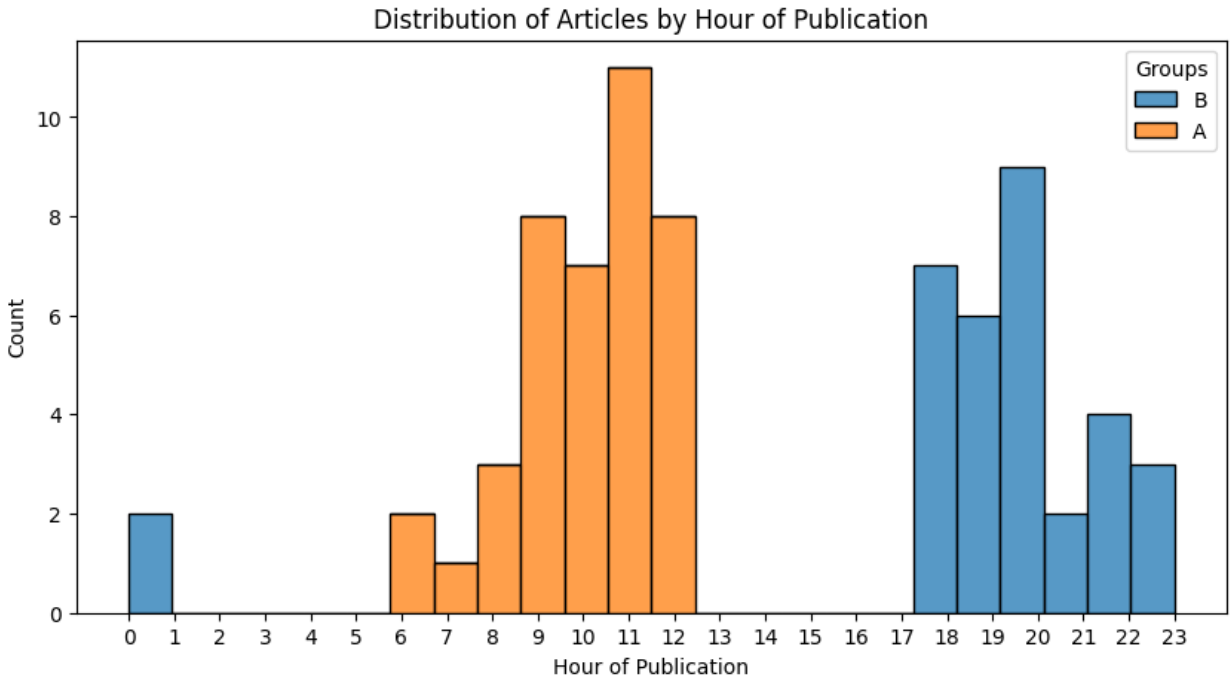
- Bar plots for top-performing categories and sources, engagement across groups
- Histogram to show the distribution of articles by their hour of publication
- Heatmaps for correlation analysis
- Boxplots to visualize outliers and engagement distributions

Insights & Findings

Time Distribution

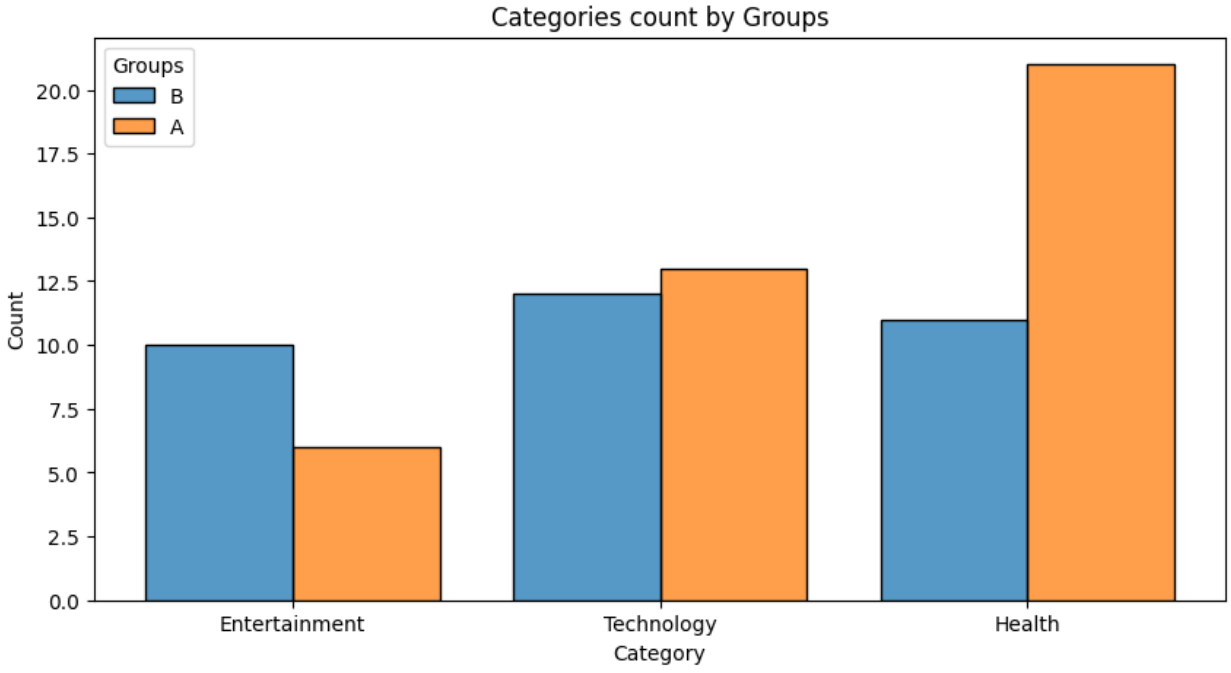
A histogram was used to visualize how articles are distributed across the 24-hour clock. The plot clearly shows the number of articles published within the two defined time windows: **Group A:** 6:00 AM – 12:00 PM and **Group B:** 6:00 PM – 12:00 AM

The visualization confirms that articles were correctly assigned to their respective groups and highlights that **more articles were published in the morning** (Group A) compared to the evening (Group B).



A second histogram shows which category dominated the peak posting period. Since **Group A (6:00 AM – 12:00 PM)** had the highest number of articles published, this visualization focuses on the categories contributing most to that window.

The results show that the **Health** category had the **highest number of posts** during this period, making it the leading contributor to the morning publication peak. Other categories had significantly fewer articles in comparison.

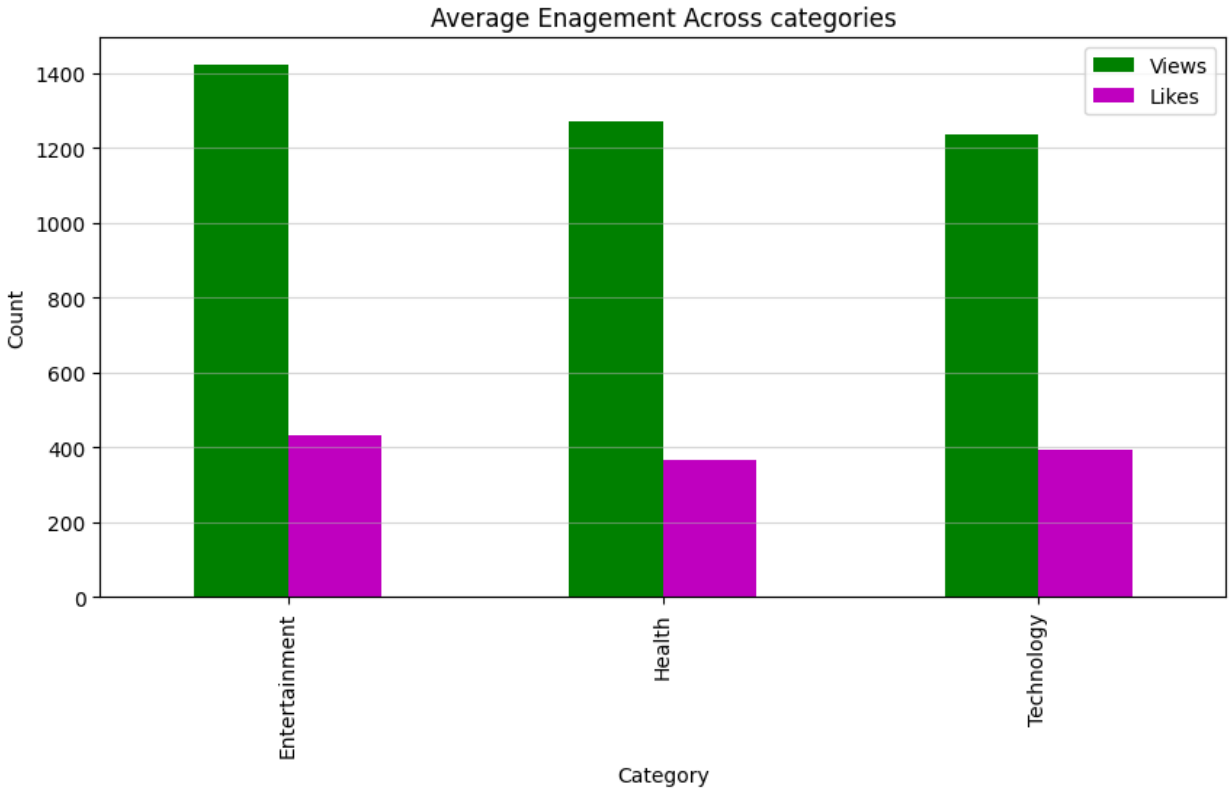


Engagement by categories:

The bar chart blow compares average engagement across categories using **views** and **likes**. It reveals that **Entertainment** articles received the **highest engagement** compared to other categories. This suggests that audiences are more likely to interact with entertainment related content compared to other categories.

Health articles had the second highest number of views but received fewer likes, indicating that although they draw a lot of readers, they tend to generate less engagement(likes).

Technology articles showed close view counts to Health but achieved **slightly higher likes**, indicating better audience interaction than Health despite fewer reads.

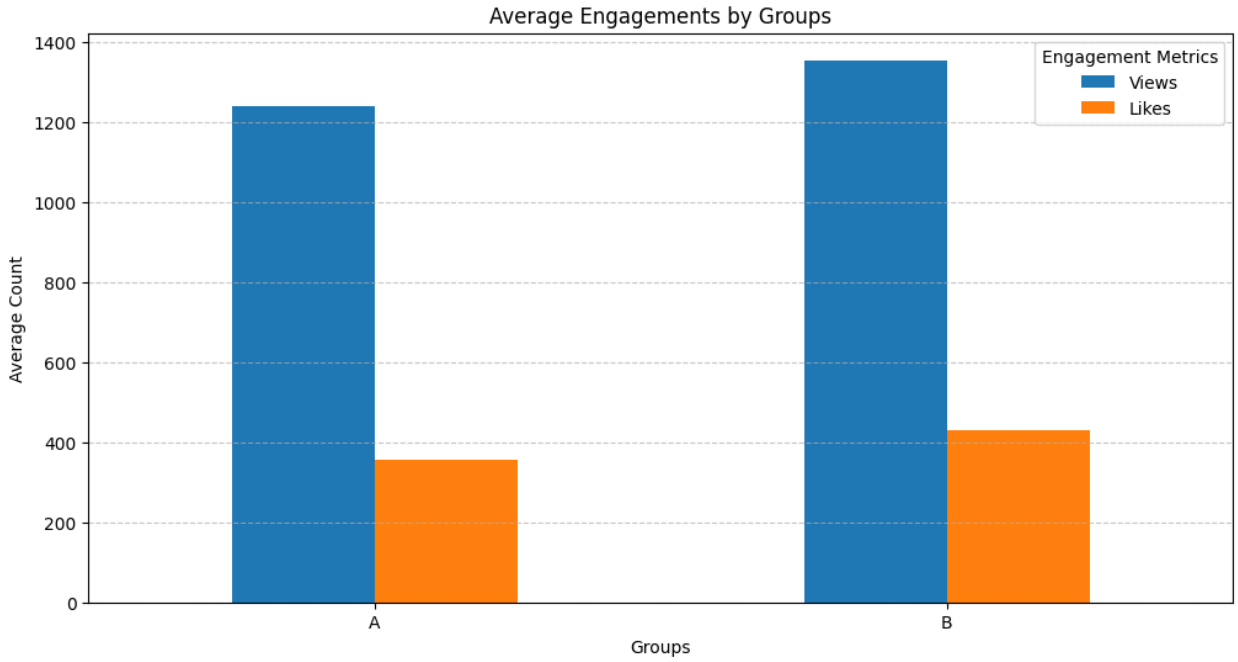


Engagement by Groups

The bar chart below illustrates the **average engagement** for articles across the two publication time groups (**Group A** and **Group B**).

- **Group B (6 PM – 12 AM)** recorded **higher engagement levels** overall, with **1,355 average views** and **430 average likes**.
- **Group A (6 AM – 12 PM)** had slightly lower engagement, **1,240 average views** and **357 average likes**.

This indicates that **articles published in the evening (Group B)** tend to attract **more user interaction** in terms of both **views** and **likes**, compared to those published in the morning.



Correlation Analysis

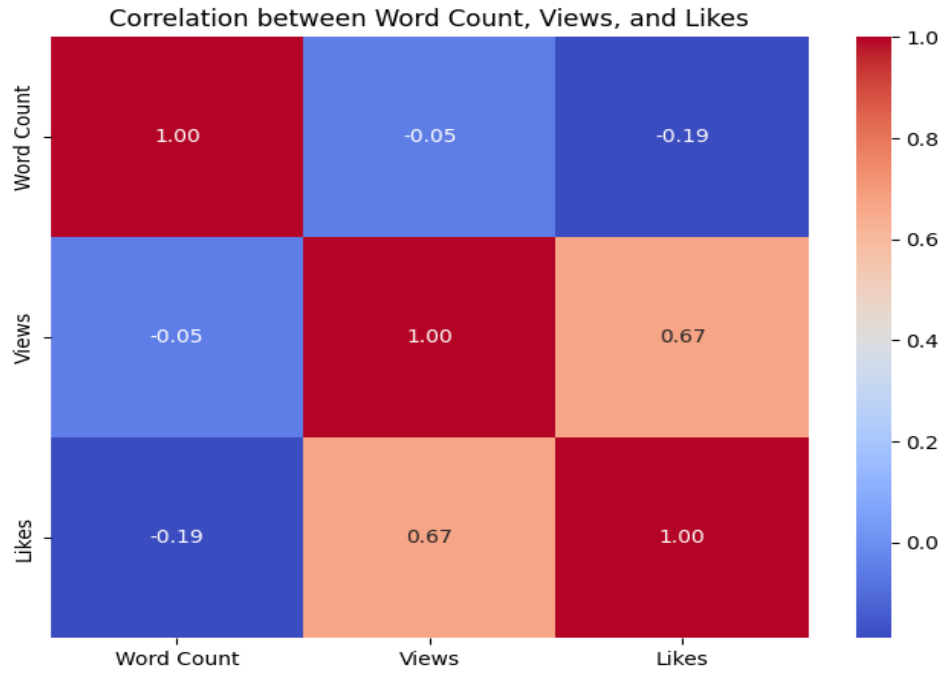
The heatplot below shows how **Word Count**, **Views**, and **Likes** are related:

Word Count & Views had **-0.05** as correlation coefficient. This shows a very weak **negative** correlation, interpreting that article length barely affects number of views.

Word Count & Likes had **-0.19** as correlation coefficient. This is a weak negative correlation, showing that article length is not a strong predictor for likes.

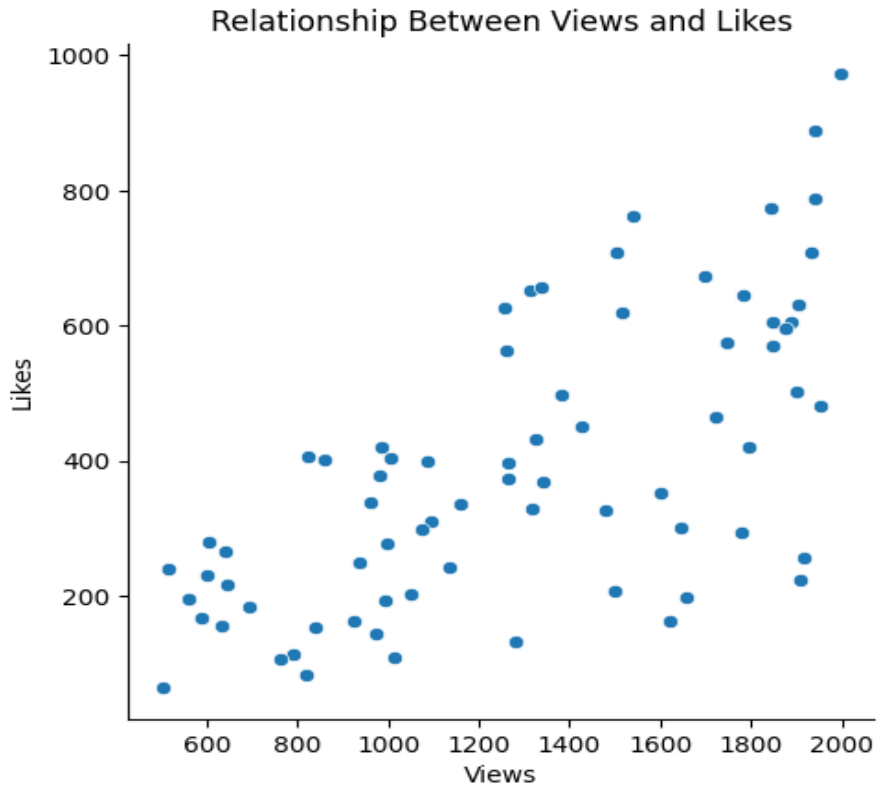
Views & Likes had **0.67** had a moderate positive correlation, interpreting that articles with more views generally get more likes

Overall, engagement seems to be driven more by **views** rather than article length, with a clear link between **views** and **likes**.



Along with that is a scatter plot showing the relationship between Views and Likes. There is a clear upward trend showing that articles with more views get more likes, suggesting visibility drives user engagement.

However, the spread of points also shows a little variability. Not every highly viewed article gets a proportional number of likes, which indicates that content quality and relevance still play a role.



Statistical Test

Since Group A and Group B are two independent groups, an **independent two-sample t-test** was conducted to determine whether there is a **significant difference** in **user engagement** (**views** and **likes**) between the two groups.

Hypotheses

- **Null Hypotheses (H_0):** There is **no significant difference** in engagement between **Group A** and **Group B**.
- **Alternative Hypotheses (H_1):** There **is a significant difference** in engagement between **Group A** and **Group B**.

Results

- **Views:** The **p-value** is **0.2877**. Since the p-value is greater than the significance level (**alpha = 0.05**), we **fail to reject the null hypothesis**. This indicates that

there is **no statistically significant difference** in the average number of views between **Group A** and **Group B**.

- **Likes:** The p-value is **0.1457**. Similarly, because the p-value is greater than **alpha = 0.05**, we **fail to reject the null hypothesis**. This suggests that there is **no statistically significant difference** in the average number of likes between **Group A** and **Group B**.

The results suggest that **time of publication** (morning vs evening) **does not have a statistically significant impact** on **user engagement**. While **Group B** visually appears to have slightly higher likes and views, this difference is **not strong enough** to be considered significant at the **95% confidence level**.

Conclusion

The analysis was conducted to evaluate how user interaction with these articles is influenced by their time of publication. Based on the statistical test results, there is no significant difference in engagement between posts published in the morning (Group A) and those published in the evening (Group B) at a 95% confidence level.

While Group B (evening posts) showed a slightly higher average in likes and views visually, the difference is not statistically significant. This suggests that time of publication alone may not be a strong determinant of user engagement. Other factors, such as content quality, relevance, headlines, author popularity and source, are likely to have a greater influence on how users interact with posts.