YouTube Video Data Analysis Report

# Introduction

This report summarizes the analysis performed on a dataset of YouTube video statistics. The analysis includes data cleaning, exploration, and visualization, with the goal of understanding trends and relationships in the data.

# Data Description

The dataset, "videos-stats.csv", contains the following columns:

* **Title**: Title of the video.
* **Video ID**: Unique identifier for the video.
* **Published At**: Date when the video was published.
* **Keyword**: Category or topic of the video.
* **Likes**: Number of likes.
* **Comments**: Number of comments.
* **Views**: Number of views.

# Data Cleaning

The following cleaning steps were performed:

* Converted the 'Published At' column to datetime format.
* Removed rows with missing values in the 'Views', 'Likes', and 'Comments' columns.
* Ensured that the 'Views', 'Likes', and 'Comments' columns were of numeric data type.
* Removed duplicate rows.

# Data Exploration

The dataset contains 1879 entries with 8 columns. The data types of the columns are as follows:

* **Published At**: datetime64[ns]
* **Likes**: float64
* **Comments**: float64
* **Views**: float64
* **Keyword**: object

Descriptive statistics for the numerical columns are as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Column** | **Mean** | **Min** | **25%** | **50%** | **75%** | **Max** |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Likes | 170,061 | -1 | 2,672.5 | 14,787 | 60,906 | 16,445,56  0 |
| Comments | 7,863.33 | -1 | 199 | 814 | 3,377.5 | 732,818 |
| Views | 11,612,92  0 | 25 | 84,515 | 591,721 | 2,804,978 | 4,034,122,  000 |

The most frequent keywords are 'minecraft', 'asmr', and 'mrbeast', each appearing 50 times.

# Data Analysis

## Basic Statistics

After handling negative values and outliers in 'Views', 'Likes', and 'Comments', the following statistics were calculated:

* + - Mean, median, and standard deviation for each variable.
    - Interquartile range (IQR) for each variable.

## Correlation Analysis

* + - Calculated the correlation matrix to explore relationships between 'Views', 'Likes', and 'Comments'.
    - Visualized the correlations using a heatmap.
    - Created scatter plots to visualize the relationship between 'Views' and 'Likes', 'Views' and 'Comments', and 'Likes' and 'Comments'.

## Distribution Analysis

* + - Plotted histograms to visualize the distribution of 'Views', 'Likes', and 'Comments'.
    - Used kernel density estimation (KDE) to estimate the probability density function of each variable.

## Time Series Analysis

* + - Extracted the month from the 'Published At' column.
    - Calculated the total views, likes, and comments per month.
    - Visualized the monthly trends using line plots.

## Keyword Analysis

* + - Grouped the data by 'Keyword' and calculated the mean views, likes, and comments for each keyword.
    - Identified the top 10 keywords with the highest average views, likes, and comments.

## SQL Analysis

* + - Used SQL queries to analyze the data.
    - Calculated the like-to-view ratio for each video.
    - Identified the top 10 videos with the highest views and the top 10 videos with the highest like-to-view ratio.
    - Visualized the distribution of like-to-view ratios using a histogram.

# Summary

The analysis provides insights into the YouTube video dataset. Key findings include:

* Strong positive correlations between views, likes, and comments.
* Identification of top performing videos and keywords.
* Monthly trends in video engagement.

# Next Steps

Further analysis could include:

* Sentiment analysis of comments.
* Analysis of video duration and its impact on engagement.
* Predictive modeling for video views, likes, and comments.
* Exploring other factors like video category, uploader, etc.