# Social Media Engagement Analytics

## 1. Project Overview

In this project, I analyzed a social media engagement dataset (social_media_engagement_5000.csv) to understand user behavior and content performance. The goal was to identify patterns in likes, comments, shares, impressions, watch time, engagement rate, sentiment, and other factors. The analysis helps provide insights into content strategy and audience interaction.

The dataset contains user and post metadata such as user_id, age, gender, country, post_id, post_type, post_category, engagement metrics, device type, and sentiment.

---

## 2. Data Import & Setup

I started by importing the dataset using **Pandas** and inspected the info() to understand columns and data types.

I converted the posted_at column to datetime format to enable time- based analysis (like daily engagement trends).

- Loaded CSV into a DataFrame.
- Checked data types and converted posted_at to datetime.
- Verified data with describe() and initial exploration.

---

## 3. Data Cleaning

### a. Handling Missing Values

- I used isnull() and isna() to detect missing values.
- Numerical columns like age,likes, comments, and shares were filled with median values.
- Categorical columns like gender, and sentiment filled with mode.

### b. Duplicates & Incorrect Values

- There is no duplicates.

### c. Data Formatting

- Ensured columns like likes, comments, and shares had realistic values (non- negative).
- The gender data was already clean. I just capitalized the text to keep the format consistent for analysis and charts.

### c. Feature Cleaning

- Extracted hashtag_count from the hashtags column.
- Cleaned sentiment labels into categories: negative, neutral, positive.

---

## 4. Data Exploration & Wrangling

- Checked dataset structure using `head()`, `tail()`, `shape`, and `columns`
- Verified data types and overall info using `info()` and `dtypes`
- Generated summary statistics with `describe()` to understand numeric data
- Analyzed categorical columns using `value_counts()`, `unique()`, and `nunique()`
- Created a correlation matrix for numeric engagement metrics

Used `groupby()` to calculate:

- Average likes by post type
- Impressions by country
- Merge/Join was **not performed** as the project uses a **single dataset** and no additional data sources were provided.
- **Engagement score** using likes, comments, and shares
- **Log-transformed engagement score** to handle skewness and reduce the impact of extreme values
- **Hashtag count** extracted from the hashtags column

Used `groupby()` to summarize engagement metrics by:

Post type

- Country
- Sentiment

---

## 5. Statistical Analysis

Computed **mean, median, and mode** to understand central tendency

Calculated **standard deviation and variance** to measure data spread

Analyzed **percentiles** to understand distribution and extreme values

Checked **skewness and kurtosis** (optional) to understand data distribution shape

---

## 6. Visualization

Created multiple visualizations using **Matplotlib, Seaborn, and Plotly** to understand engagement patterns clearly

### Matplotlib Plots

- Scatter plot to analyze **likes vs impressions**
- Line chart to show **daily engagement trends**
- Bar chart for **number of posts by category**
- Pie chart to visualize **gender distribution**
- Histogram to understand **age distribution**
- Box plot to analyze **engagement rate and detect outliers**

### Seaborn Plots

- Count plot to show **post type distribution**
- Bar plot to compare **average likes by category**
- Violin plot to analyze **followers vs sentiment**
- Pair plot to explore **relationships between numeric features**
- Heatmap to visualize **correlation between engagement metrics**
- Swarm plot to compare **engagement rate across device types**

### Plotly (Interactive)

Created interactive charts (line/bar/scatter) to explore trends dynamically

---

## 7. Final Insights

Based on the analysis and visualizations:

### Content Performance

- **Video posts** had the highest average engagement rate.
- **Tutorial and entertainment categories** received the most likes.
- Users from **USA, UK, and India** showed the highest engagement.

### User Trends

- Engagement was highest among users aged **25–35**.
- **Verified accounts** consistently outperformed non- verified accounts in engagement.

### Behavioral Insights

- Posts between **6 PM and 9 PM** got the highest impressions.
- **Mobile users** spent more time watching content than desktop users.

### Sentiment Analysis

- **Positive sentiment posts** had the highest engagement.
- Negative sentiment posts had lower median engagement and wider variation; neutral posts performed moderately.

---

## 8. Conclusion

The project successfully cleaned and analyzed the dataset using Python, visualized key patterns, and generated actionable insights. The findings can help content creators focus on formats and times that maximize engagement and understand audience behavior effectively.