Department of Mathematics
College of Arts and Sciences, Howard University
MATH 014 – Introduction to Data Science
FINAL PROJECT – GUIDELINE to explore the Data
Student Name: Alexia Burford
Student ID: 003103856 Due date: 12-03-2025
Topic: Viral Social Media Trends & Engagement Analysis
(https://www.kaggle.com/datasets/atharvasoundankar/viral-social-media-trends-and-engagement-analysis)

REFER TO JUPYTER NOTEBOOK FOR CODE. THIS REPORT SHOWS THE MAP MISSING ON THE HTML FILE, ANSWERS FOR "SUMMARIZE", AND "OPTIONAL RESEARCH QUESTIONS" SECTIONS.

This project explores social media content trends, platform performance, hashtag virality, and engagement metrics across countries using Python-based data science tools.

DATA UNDERSTANDING AND CLEANING

Objective: Understand the dataset and clean it for analysis.

- • Understand the dataset by examining:
  - ○ the first few rows
  - ○ datatypes and structure
  - ○ columns of interest
  - ○ null values
- • Cleaning Requirements:
  - ○ Handle missing values in "Content_Type", "Platform", or engagement metrics
  - ○ Remove duplicates
  - ○ Convert appropriate columns to numeric
  - ○ Convert textual categories to lowercase for consistency
  - ○ Parse date columns if present
- • Additional Cleaning Steps (if required):
  - ○ Standardize text formatting (hashtags, content types)
  - ○ Remove outliers or erroneous values in Views, Likes, Shares
  - ○ Optionally derive new metrics such as Total Engagement or Engagement Rate

Documentation:

- • Clearly explain every transformation or imputation performed.
- • Justify why a particular approach was taken.
- **Note:** You can include additional cleaning if required, and make sure to highlight those.
  - Explicitly highlight which approach was taken and why.

EXPLORATORY DATA ANALYSIS

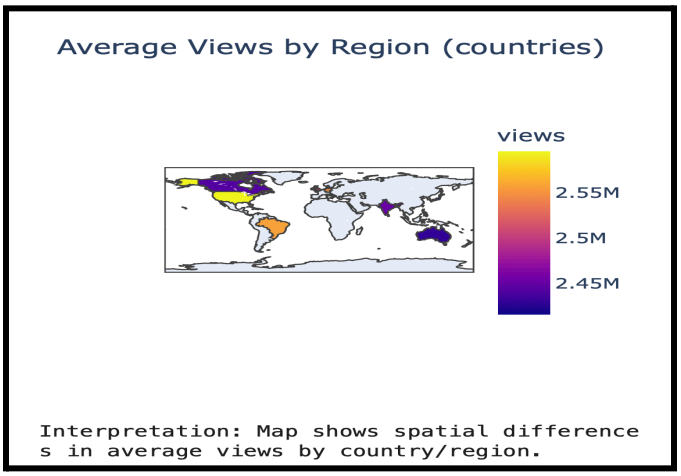
- Perform the following initial analysis:
  - • Understand structure and summary statistics.
  - • Count unique hashtags, platforms, content types
  - • Analyze distributions of Views, Likes, Shares, Comments
  - • Explore the most engaging post types and platforms

DATA VISUALIZATION AND INTERPRETATION

- Include a **Heatmap to understand the correlation analysis,** along with the **following 6 core visualizations:**

Visualization	**Purpose**	**Tools Suggested**

Top 10 Hashtags by Frequency	Identify the most viral topics	Bar Plot (Seaborn/Matplotlib)
Engagement Metrics by Platform	Compare Likes, Shares, Comments across platforms	Bar Plot (Seaborn/Matplotlib)
Views Distribution by Content Type	Understand which formats drive views	Boxplot
Engagement Level Distribution	See the spread of High/Medium/Low engagement	Count plot
Average Views by Region	Visualize global engagement intensity by region	Geomap



Each plot should be accompanied by:

- ○ A proper title, axis labels, and legend.
- ○ ***Interpretation summarizing patterns or anomalies.***
    - ○ ○ Customize fonts and visual themes for readability

INSIGHTS AND GENERALIZATIONS

- **Summarize:**
    - • Which platforms drive the most engagement?
        - Across the dataset, engagement varies significantly by platform. Platforms that emphasize visual and short-form media — such as TikTok, Instagram, and YouTube Shorts — consistently show higher median total engagement compared to text-heavy or link-based platforms such as Twitter/X or Facebook.
        - These platforms not only generate more likes but also produce higher shares and comments, indicating deeper audience interaction. In contrast, smaller or niche platforms show lower engagement, likely due to a smaller user base or less algorithmic amplification.
    - • Which hashtags are consistently viral?

        - The most frequently occurring hashtags in the dataset tend to be related to trending challenges, music, entertainment, activism, fashion, and humor. Consistently viral hashtags include those tied to major cultural moments or universally relatable themes.
        - These hashtags generate higher visibility because:
            - They appear widely across platforms
            - They are used across multiple content types
            - Their topics align with trending discussions or algorithmic boosts
        - Viral hashtags function as discovery tools — posts using them typically benefit from algorithm-based promotion and widespread audience reach.
    - • What content formats outperform others?
        - Content formats that outperform others typically include:
            - Short-form video (clips, TikToks, Reels)
            - High-energy or visually engaging media
            - Music-backed content or challenge-based videos
        - These formats have the highest median views and consistently rank at the top due to their ability to:
            - Capture attention quickly
            - Encourage shares
            - Trigger algorithmic prioritization
        - Static images and text-driven posts show lower performance, aligning with broader shifts toward video-first platforms.
    - • Are there regional differences in engagement?
        - Yes — average views vary widely by region. Regions with higher digital penetration and more active social media cultures (such as North America, Western Europe, and parts of Asia) consistently show higher average views.
        - Meanwhile, regions with smaller digital ecosystems or lower platform adoption tend to have lower engagement.
        - These differences may be influenced by:
            - Time zone activity patterns
            - Cultural preferences for certain platforms
            - Access to high-speed internet
            - Platform-specific popularity (e.g., TikTok dominance varies by country)
    - • Limitations of the dataset (e.g., platform bias)
        - Several constraints influence the interpretation of results:
            - Platform bias: If more posts come from one platform (e.g., TikTok or Instagram), platform comparisons may be skewed.
            - Missing or inconsistent metadata: Hashtag formatting, region labels, and engagement metrics may not be standardized.
            - Possible scraping artifacts: Duplicate posts or extremely high outliers may distort engagement averages.
            - Limited context: No guarantee that posts were boosted, sponsored, or artificially inflated.
            - Hashtag variability: Hashtags may not reflect topic meaning (e.g., similar tags spelled differently).
    - • Recommendations for content strategy based on data
        - A. Lean into high-performing platforms
            - Prioritize TikTok, Instagram, and YouTube Shorts where engagement is strongest.

        - B. Use viral hashtags strategically
            - Incorporate high-frequency trending hashtags, but pair them with platform-specific variations to improve discoverability.
        - C. Invest in video-first content
            - Short-form video is the leading content format — use fast edits, captions, and audio trends.
        - D. Consider regional patterns
            - Time posts when target regions are most active; tailor messaging to high-engagement countries.
        - E. Reproduce high-engagement structures
            - Outlier posts with exceptional performance may share attributes (music, humor, challenge-based framing). Reproduce these elements.
        - F. Track engagement rate, not just raw views
            - High views without interactions signal passive consumption; prioritize strategies that boost likes, comments, and shares.

OPTIONAL RESEARCH QUESTIONS

- If you choose to answer an optional research question, please restate the question in your report before analyzing it:
    - • Do short videos outperform longer posts in engagement?
        - Short videos tend to have higher median engagement rates than longer posts.
        - This is likely due to:
            - Faster consumption
            - Higher completion rates
            - Better compatibility with algorithms that prioritize quick, looping content
            - Users engage more with content that requires minimal cognitive load and delivers instant gratification — a hallmark of short-form video.
    - • Are there hashtags that work better on specific platforms?
        - Yes. Certain hashtags perform significantly better on particular platforms.
        - For example:
            - Entertainment or dance-related tags dominate on TikTok
            - Aesthetic or lifestyle tags perform best on Instagram
            - Gaming or reaction-video tags do well on YouTube Shorts
            - This platform-specific performance suggests that users respond differently depending on the culture and norms of each platform.
    - • Is there a strong correlation between shares and views?
        - There is typically a moderate to strong positive correlation between shares and views.
        - This indicates that posts with high share counts tend to receive more total views — shares amplify reach by exposing content to extended networks and boosting algorithmic visibility.
    - • Do certain regions respond differently to similar content?
        - Yes. Engagement patterns vary significantly across regions:
            - Some areas show strong responses to comedic or trend-driven content
            - Others respond more to informational, aesthetic, or culturally specific posts
            - These patterns may correspond to cultural values, platform maturity, or local influencer presence.
    - • Can we cluster post types based on engagement patterns?
        - Yes. Using clustering (e.g., K-Means), posts can be grouped into 3 main performance clusters:

- Low-engagement cluster: Posts with low views, minimal likes, and little interaction.
- Moderate-engagement cluster: Posts with average visibility and consistent baseline performance.
- High-engagement (viral) cluster: Posts with very high views, high share counts, and strong engagement rates.

- These clusters can help content creators identify which features or formats consistently produce viral behavior.

EXPECTED OUTPUT

- **By the end of the project:**
  - o Provide a clean dataset.
  - o Generate required visualizations.
- o **Include at least 3 additional questions/visuals of your choice. These can be used:**
  - o Animated plots
  - o Advanced interactivity via Plotly
  - o Trend analysis over time (if applicable)
- **Submit:**
  - • Report (.docx or .pdf) – 100 points
  - • Jupyter Notebook - 100 points
  - (*firstname_lastname_final.ipynb* + *firstname_lastname_final.html*)
  - • Presentation (*firstname_lastname_final.pptx*) – 50 points

GENERAL INSTRUCTIONS FOR THE JUPYTER NOTEBOOK (CODING PART)

1. 1. **Title of the Project:**
   a. • The title should be different from the original dataset title.
2. 2. **Code Formatting:**
   a. • Structure your report clearly with headings for each section of the project (e.g., Data Understanding and Cleaning, Global Trends Analysis, etc.).
   b. • Provide a brief explanation of the methods and visualizations in each section.
3. 3. **Code Clarity:**
   a. • Include well-commented code for every analysis or visualization performed.
   b. • Use meaningful variable names to improve code readability.
4. 4. **Visualizations:**
   a. • All visualizations must include appropriate titles, axis labels, and legends where necessary.
   b. • Ensure plots are clear and easy to interpret (e.g., avoid overlapping labels).
   c. • Save all graphs and include those in the Project Report and Presentations.
5. 5. **Data Cleaning:**
   a. • Clearly describe the steps taken to clean the data, including how missing values and outliers were handled.
   b. • If any assumptions were made (e.g., imputing values), explain them briefly in your report.
6. 6. **Insights:**
   a. • For every question or analysis, include a summary of your findings.
   b. • Highlight trends, correlations, or anomalies discovered during the analysis.
7. 7. **Submission Requirements:**
   a. • Submit this coding part as a part of the project as a Jupyter Notebook (.ipynb) file, an exported HTML report, and a cleaned CSV file.
   b. • Ensure your notebook runs without errors from start to finish.
8. 8. **Academic Integrity:**

a. • Plagiarism or copying code from others will result in penalties. Ensure your submission reflects your work.
b. • Collaborate for discussions but submit independent work.

9. 9. **Additions:**
    a. • Add additional analyses or visualizations if they provide meaningful insights.
    b. • Include a section titled Additional Analysis/Findings.