Social Media Analysis Report

Data Wrangling Final Project Fall 2024

Team Members:

Rawan Mohamed

Salma Montasser

Fayrouz Mahmoud

Omar Rayan

Supervised By:

DR. Amal Mahmoud

Table of Contents

1. Introduction:

This report describes the process followed for data wrangling and analysis on the provided dataset of social media. The main pre-processing of data, exploration, and insights derivation were carried out for this purpose. The data contains user information, such as demographics, on which platform they are engaged, and what kind of interest categories are there.

2. Data Wrangling Process

2.1. Data Loading and Initial Exploration

Overview of Dataset: The data was first loaded using pandas and explored through functions such as .head(), .info(), and .describe().

Shape and Integrity:

Dimensions: 100,000 instances and 16 features.

Data Quality: No missing values or duplicate records were found.

Initial Observations:

- The dataset has numerical and categorical columns.
- In numerical features, outliers and skewness are noticed, hence some transformations, normalizations are expected in the pre-processing.

2.2. Cleaning Steps

Dropping Irrelevant Columns: The UserID column was dropped as it had no analytical value.

Diacritic Removal: Non-standard characters in the City column were normalized using the unicodedata module for consistent analysis.

2.3. Outlier Detection and Handling

Boxplots: Boxplots were created to visualize the numeric columns and check for outliers. IQR-Based Capping: Outliers were capped using interquartile ranges so that there is robust analysis with retained significant data points.

2.4. Skewness Detection and Handling

Data Distribution: created to visualize the numeric columns and their distributions**.**

Skewness Reduction: Log transformation was applied to numeric columns to reduce skewness and normalize distributions.

2.5. Feature Engineering

- Age Calculation: Extracted age from the DOB column and grouped users into categories: Young Adult (0-30), Adult (31-50), and Senior (51+).
- Engagement Metrics: New features such as TotalFacebookEngagement, TotalInstagramEngagement, TotalTikTokEngagement, and TotalEngagement were derived using a combination of followers and time spent on platforms.
- Activity Intensity: A feature summing posts across all platforms was created to measure user activity.
- Platform Preference and Dominance: Dominance ratios for each platform were calculated to understand the share of each in the total followers.
- Preferred platform was determined based on the highest follower count.
- Interest Categorization: User interests were grouped into categories such as Entertainment, Knowledge, and Lifestyle based on predefined keywords.

2.6. Transformation and Encoding:

Scaling: The numerical features are normalized using Min-Max scaling for comparability across the variable.

LabelEncoding: LabelEncoder is used to encode categorical variables into a compatible machine learning model format.

2.7. Data Aggregation:

- Country and city columns are grouped to retrieve their numbers and get the most frequent location.
- Create a frequency table showing the number of occurrences of each age group for each preferred platform.
- Computes the sum of Entertainment, Knowledge, Lifestyle column.

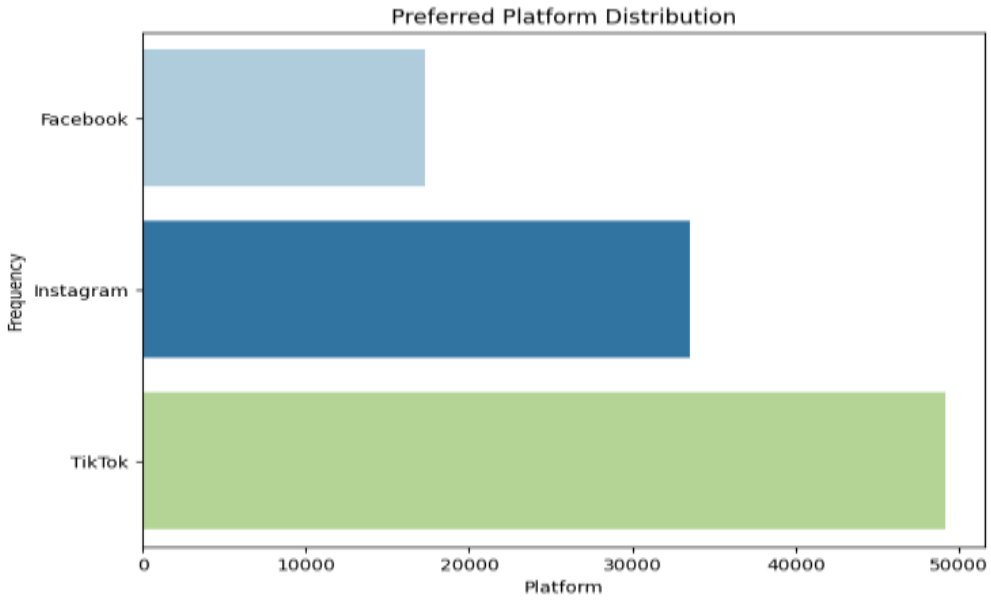
3. Data Visualization & Analysis (Python):

- **Preferred Platform Distribution**

  To analyze the distribution of users' preferred platforms, a bar chart using a count plot is utilized.

  The chart highlights the most and least preferred platforms among users, offering insights into platform popularity trends. This can help in tailoring strategies for user engagement across platforms.

  As illustrated in the chart, TikTok is the most preferred platform, followed by Instagram, then Facebook.

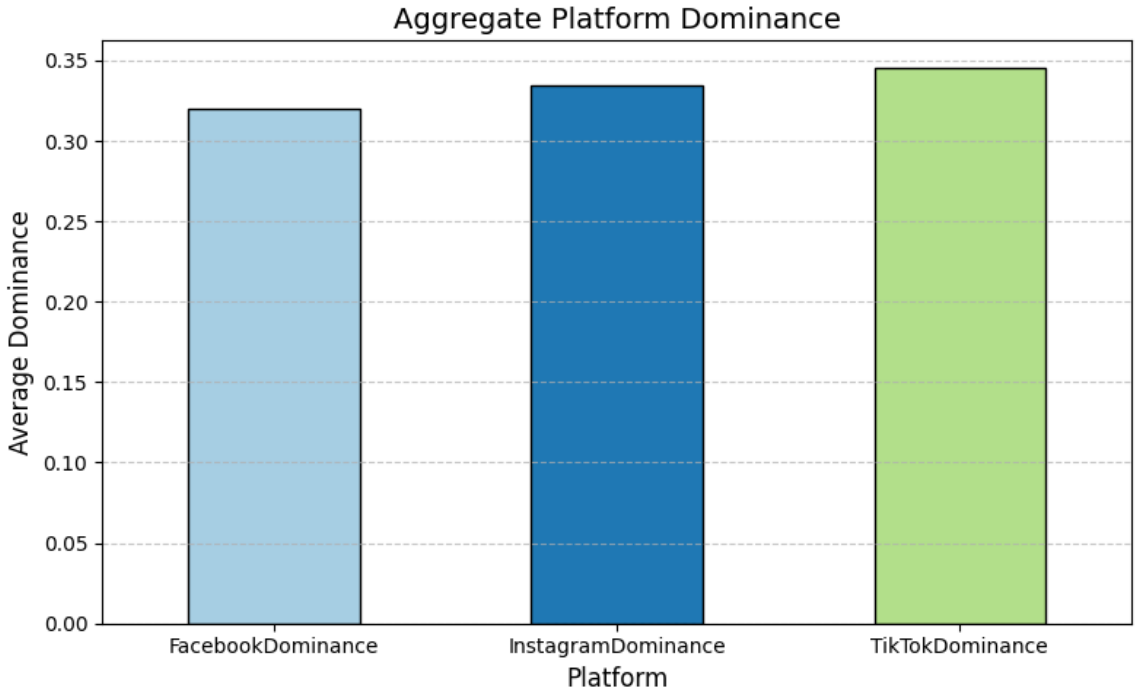


- **Dominance Across Platforms**

  This bar chart visualizes the average dominance for each social media platform Facebook, Instagram, and TikTok across all users in the dataset.

  The plot shows TikTok Dominance as the highest, followed closely by Instagram Dominance, reflecting TikTok's competitive edge while highlighting Instagram's sustained popularity. The findings indicate a close race in platform dominance, with TikTok leading, followed by Instagram and Facebook. These results align with current trends, where emerging platforms like TikTok are rapidly gaining traction,

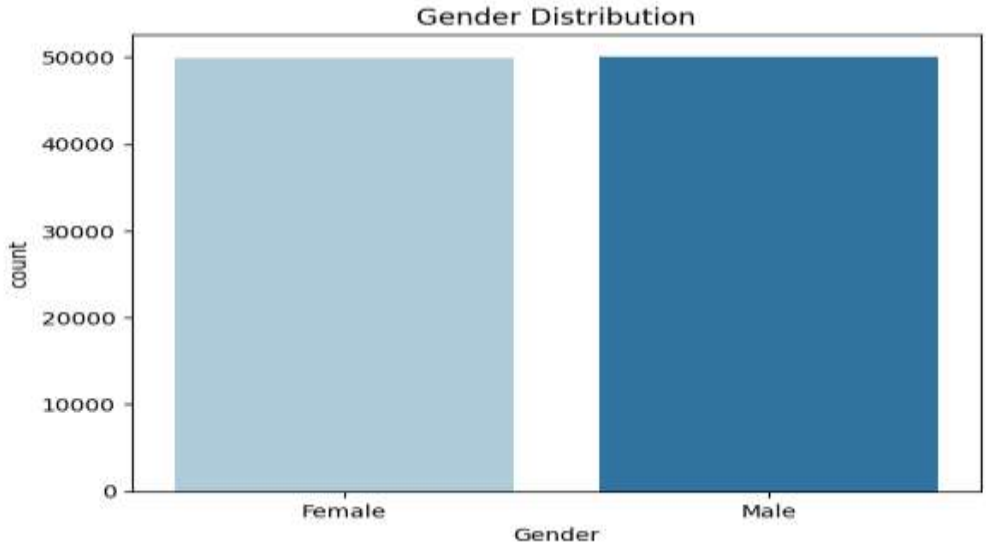
while established platforms such as Facebook and Instagram continue to maintain a significant user base.



- **Gender Distribution**

  To analyze the demographic breakdown of users by gender, a bar chart is used. The visualization displays the frequency of each gender in the dataset, providing insights into the gender composition of the users.

  The chart shows a balance between males and females in social media users.

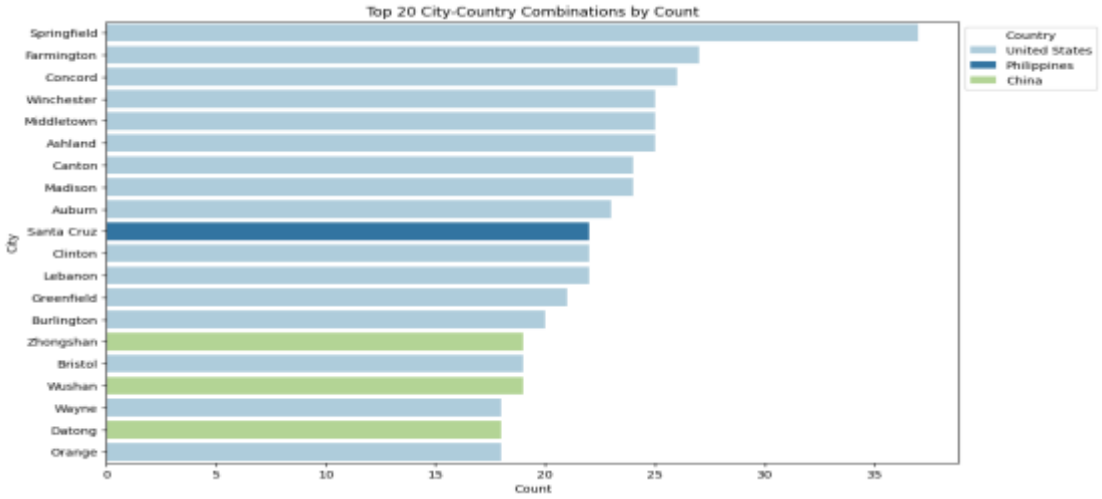


- **Top 20 City-Country Combinations by Count**

  To analyze the geographic distribution of users, a bar plot was generated to highlight the top 20 city-country combinations with the highest counts in the dataset.

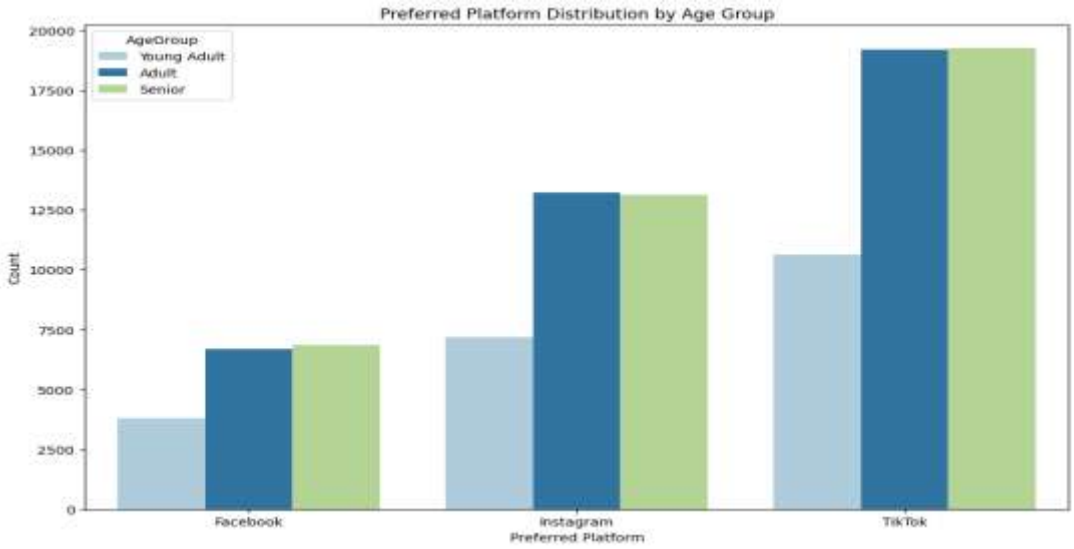
  The cities at the top of the plot indicate high user activity. The hue for each country helps identify which countries dominate the dataset or appear most frequently across cities. Countries with multiple cities in the top 20 indicate broader user engagement in those regions.

  The chart shows that the US is the country with the highest frequency and Springfield is the highest city. The US is followed by the Philippines and China.



- **Preferred Platform Distribution by Age Group**

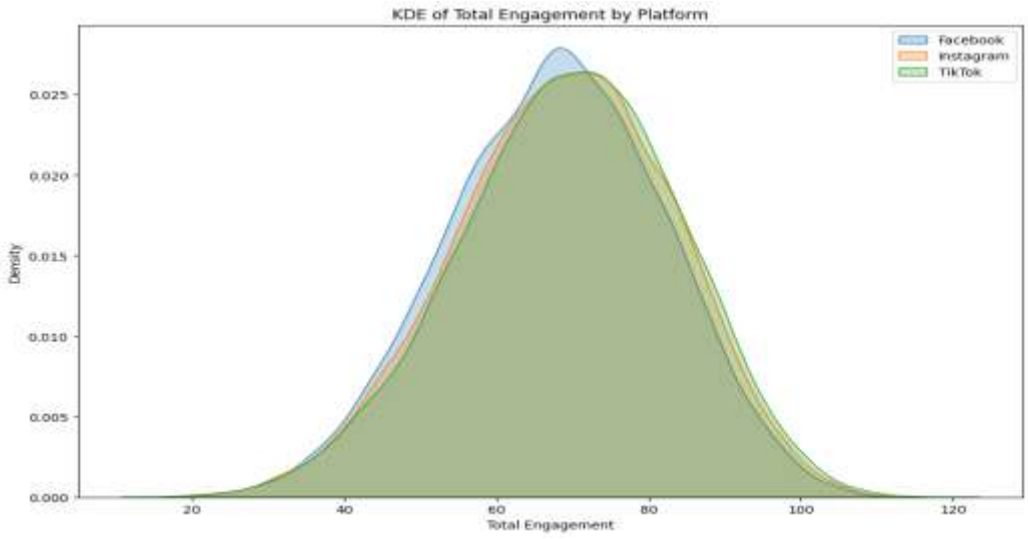
  The bar chart illustrates the distribution of Preferred Platforms across various Age Groups. TikTok leads across demographics, showcasing its global reach. Instagram maintains a balanced user base among Adults and Seniors. Facebook remains popular with older groups, with limited interest among Young Adults, reflecting broader trends where newer platforms are dominating, and traditional ones retain niche popularity.



- **KDE of Total Engagement by Platform**

  This is a Kernel Density Estimate (KDE) plot that visualizes the total engagement for three platforms: Facebook, Instagram, and TikTok. All three platforms show a similar distribution of total engagement, indicating comparable user activity levels. Facebook (blue) has a slightly higher peak density, suggesting a concentration of engagement around its average (~60-70 range). This indicates that user engagement on Facebook is more consistent.

  TikTok (green) and Instagram (orange) overlap significantly, but TikTok shows slightly broader distribution. This suggests TikTok may have more variability in engagement compared to Instagram. Engagement on all platforms spans roughly from 20 to 120, with a concentration in the 50-90 range.



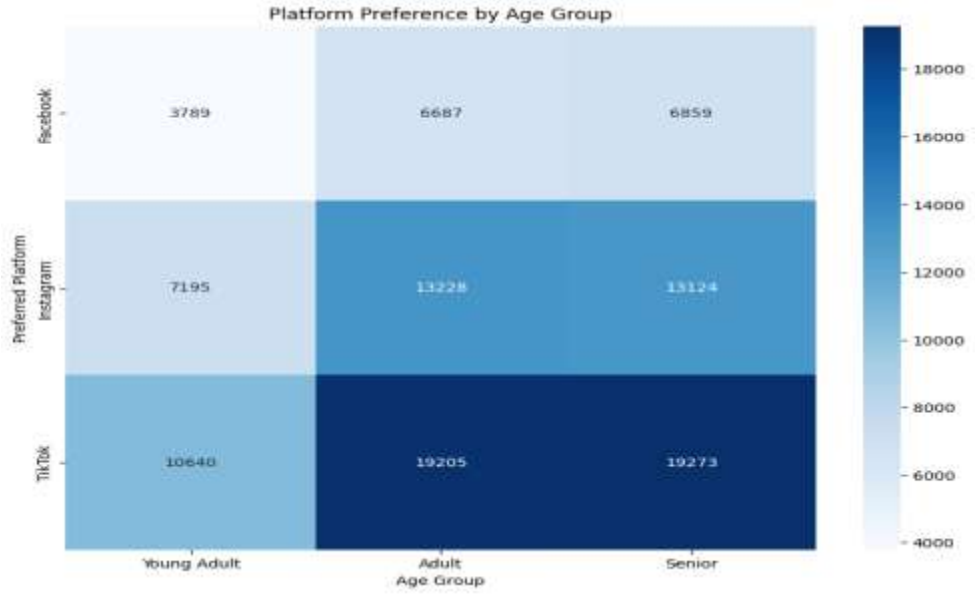
- **Platform Preference by Age Group**

  The heatmap visualizes the platform preferences across different age groups, with the intensity of color representing the frequency of preferences, annotated with integer values for clarity.

  TikTok is the most preferred platform across all age groups, with the highest numbers in adult (19,205) and Senior (19,273) categories.

  Instagram has a strong preference among adults (13,228) and Seniors (13,124).

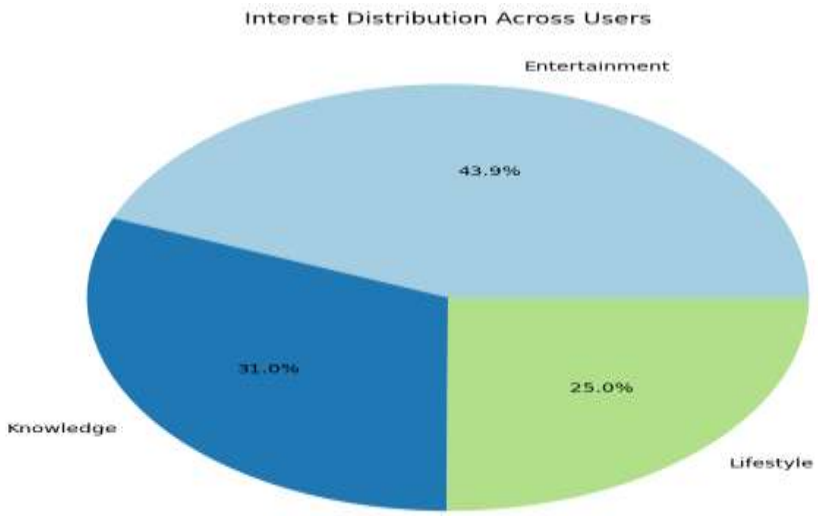
  Preference among Young Adults is significantly lower (7,195), suggesting shifting interest to other platforms like TikTok. Facebook is the least preferred platform, especially among Young Adults (3,789). However, engagement improves slightly with adults (6,687) and Seniors (6,859), indicating a trend of older age group preference.



- **Interest Distribution Across Users**

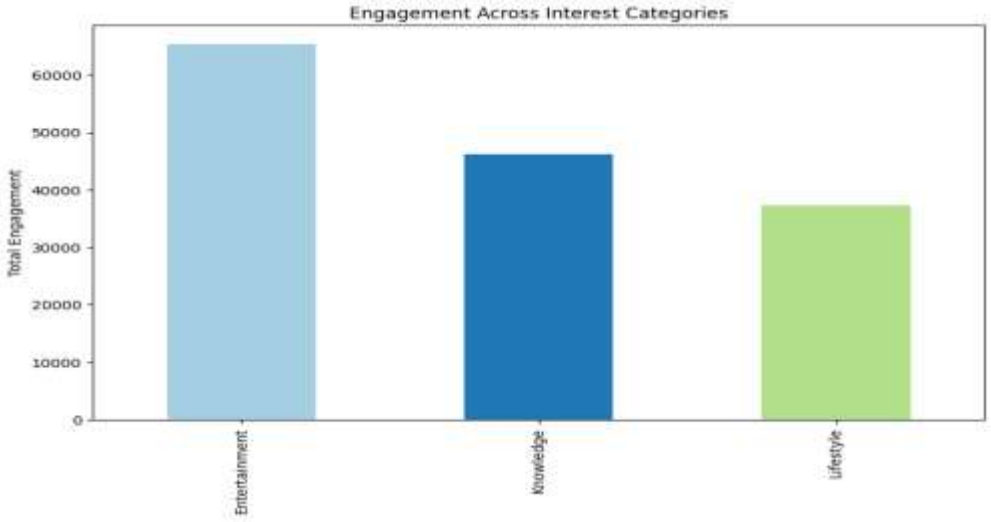
  The pie chart shows the distribution of different interests across users, with each segment representing a specific interest and labeled with its percentage share. Entertainment is the most dominant interest, accounting for 43.9% of users, indicating that most users prioritize entertainment-related content. Knowledge follows with 31%, suggesting a significant portion of users are interested in educational or informative content. Lifestyle makes up 25%, showing that a quarter

of users are focused on topics related to lifestyle, which includes areas like health, fashion, and personal development.



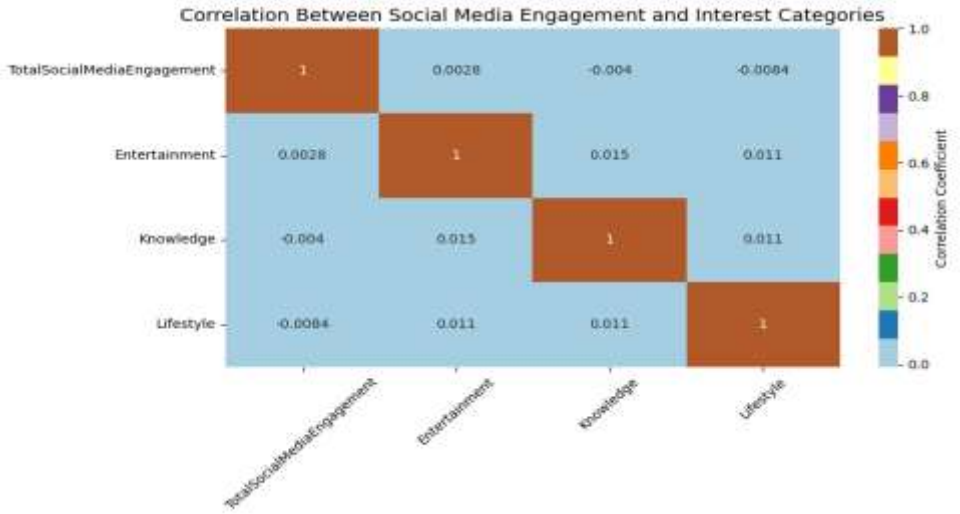
- **Engagement Across Interest Categories**

  The bar chart displays total engagement across different interest categories, with each bar representing a category, allowing for a clear comparison of engagement levels across the interests. Entertainment has the highest engagement, with slightly above 60,000, indicating it is the most popular category among users. Knowledge follows with an engagement of around 45,000, making it the second most engaging category, showing a strong interest in educational or informative content. Lifestyle has the lowest engagement, slightly above 30,000, suggesting that while it remains important, it has less attention and interaction.



- **Correlation Between Social Media Engagement and Interest Categories**

  The heatmap visualizes the correlation between total social media engagement and three interest categories. Total Social Media Engagement shows very weak correlations with all interest categories (Entertainment: 0.0028, Knowledge: -0.004, Lifestyle: -0.0084). This indicates that engagement does not depend strongly on these specific content categories. Entertainment and Knowledge have a slight positive correlation (0.015). This suggests a minor relationship where users interested in entertainment may also engage with knowledge content. Lifestyle has a weak positive correlation (0.011) with both Entertainment and Knowledge. However, the values remain near zero, implying almost no significant relationship.



4. Database and SQL Statements

4.1. Database Schema

The creation of this database involved designing a well-structured schema to store and manage social media usage data efficiently, utilizing a snowflake structure for enhanced normalization and clarity. The schema has seven interconnected tables to user interactions and behaviors across platforms. The Users table stores personal attributes such as name, gender, age, and location, while linking to tables like AgeGroups and Locations for age categorization and geographic details. The Platforms table includes social media platforms, establishing a relationship with the SocialMediaUsage table, which tracks user activities like followers, posts, average time spent, and total engagement per platform.

To analyze user interests, the EngagementCategories table identifies categories such as Entertainment, Knowledge, and Lifestyle, with user-category relationships maintained in the UserCategories table. By using a snowflake structure, the database achieves high normalization, reducing redundancy and ensuring data integrity through foreign key constraints. This design also facilitates comprehensive analysis by connecting users, platforms, locations, and engagement metrics, supporting scalability and adaptability for diverse analytical and reporting needs in social media research.

4.2. SQL Queries

1. Location by Average Engagement

The SQL query analyzes the average total social media engagement by city and country. It achieves this by joining the Users table with the Locations table to link users to their respective cities and countries. It also joins the SocialMediaUsage table to incorporate the total social media engagement data for each user. The query then calculates the average engagement per city and country using the AVG function, then sorts the results in descending order. This analysis can be useful for businesses or researchers aiming to identify regions with highly engaged social media users for targeted campaigns or further demographic studies.

Insights from the Result:

- Wenatchee, United States leads with the highest average engagement of 114.8, indicating that users from this city have highly active or engaging social media behavior compared to other locations.

- The results highlight the geographic diversity of social media usage, as represented by cities from different continents, including Europe (Germany, France), Asia (Pakistan, Iran, Uzbekistan), and North America (United States, Puerto Rico).

2. Daily Average Engagement Time by Country

The query evaluates the average time users spend on social media for each country. It joins the Users table with the Locations table to link users to their respective countries and joins the SocialMediaUsage table to include average time spent for each user's social media activity. Then groups by countries and sorts by descending order. This analysis highlights geographic patterns in user engagement time, which can provide valuable insights for marketing strategies, such as identifying regions where users are more likely to interact with long-form content or advertisements.

Insights from the Result:

- Saint Martin has the highest average engagement time of 3.83 hours, indicating that users from this country spend the longest time on social media.

- Other countries with high average engagement times include Norfolk Island and Antigua (3.5), and Macao (3.43). These regions may have cultural or technological factors that encourage prolonged social media usage.

3. Platform Popularity by Number of Users

The query analyzes the distribution of unique users across TikTok, Instagram, and Facebook. This query provides the platforms' popularity and could guide strategic decisions regarding where to focus marketing efforts.

The SocialMediaUsage table is joined with the Platforms table to link each user's activity to the platform they are using. The COUNT function calculates the number of unique users for each platform, ensuring that a user is counted only once per platform. Then the results are grouped by platform name and ordered by the number of unique users in descending order.

Insights from the Result:

- TikTok has the highest user count, with 53,114 unique users, indicating its dominance among the platforms analyzed. This highlights its popularity and significant user base compared to competitors.

- Instagram ranks second with 37,578 unique users, showcasing its strong presence and appeal, though not as extensive as TikTok.

- Facebook, with 20,480 unique users, has the smallest user base among the platforms in this dataset. This could suggest a decline in its popularity or a focus on a more specific demographic compared to the other platforms.

4. Average Followers Per Platform by Location

The SQL query retrieves the average number of followers per platform across different cities and countries. This data provides a valuable snapshot of social media behavior across platforms and regions.

The Locations table is joined with the Users table to associate users with their respective cities and countries. The SocialMediaUsage table is joined to include follower counts for each user. The Platforms table is joined to identify the specific platform associated with each entry.

Insights from the Results:

- Regional Diversity: Each platform demonstrates high follower engagement across geographically and culturally diverse cities, highlighting their global reach.

- Platform Variance: While TikTok and Instagram peak at an average of 9 followers, Facebook generally falls slightly behind with averages around 8.

- Behavioral Trends: Cities with lower averages (e.g., 7 for TikTok and 6.5 for Instagram) may indicate areas with less active users, newer adoption, or lower platform saturation.

5. Daily Usage Time by Platform

This query calculates the average daily time spent by users on each social media platform. The results are grouped by platform and ordered by average time in descending order. Facebook leads slightly, the data confirms that all three platforms maintain a robust and competitive hold on user engagement time daily.

Insights from the Results:

- Users spend the most average daily time on Facebook (2.61262 hours), slightly more than on Instagram (2.59947 hours) and TikTok (2.59931 hours).
- The marginal difference suggests that these platforms attract similar user attention daily.
- The difference between the platforms is minuscule, with only 0.01315 hours (0.79 minutes) separating Facebook and Instagram, and even less between Instagram and TikTok.
- This highlights a consistent trend of user engagement across major platforms, indicating that users divide their time almost equally among them.

6. Engagement Categories by Platform

This query determines the distribution of user engagement categories (Entertainment, Knowledge, and Lifestyle) across different social media platforms. The results are grouped by platform and category, with counts sorted first by platform and then by category count in descending order. The insights can inform marketing strategies, prioritizing TikTok for campaigns targeting larger audiences, while Instagram and Facebook may cater to more niche or loyal groups.

Insights from the Results:

- The results show a uniform distribution of users across the three categories—Entertainment, Knowledge, and Lifestyle—within each platform, indicating equal engagement across interests.
- TikTok has the largest user base (34,341 per category), followed by Instagram (24,680) and Facebook (13,828), highlighting TikTok's dominance in user engagement.

5. Tableau Visualization

Tableau is used to visualize the analysis conducted on the dataset using dashboards useful for marketing departments to determine which platform to focus on and highlight platform-specific insights, user behaviors, and engagement metrics. The key metrics to analyze are as follows:

1. **User Activity Per Platform**:
    - Total number of users on each platform.
    - AvgTime spent on platforms.
    - TotalEngagement generated per platform.
2. **Demographics of Platform Users**:
    - Gender distribution on each platform.
    - Age group distribution per platform.
3. **Performance of Content**:
    - Followers vs PostsNumber vs AvgTime across platforms.
    - AvgTime vs TotalEngagement per platform.
4. **Category Engagement Per Platform**:
    - TotalEngagement across content categories on each platform

5.1. Dashboards

5.1.1. Platform Overview Dashboard: Provide an overview of platform popularity and engagement.

1. Bar Chart - Total Users Per Platform
    - Compare the number of users across platforms to identify the most used platform.
2. Dual-Axis Chart - AvgTime and TotalEngagement
    - Visualize the proportion of male vs. female users on each platform.
3. Treemap - Engagement by Category
    - Highlight which categories perform best on specific platforms.
4. Map - Engagement by Location
    - Highlight regions with the highest engagement for each platform.

5.1.2. Audience Demographics Dashboard: Understand user demographics on each platform.

1. Stacked Bar Chart - Age Group Distribution by Platform
    - Compare the distribution of age groups for each platform in one view.
2. Stacked Bar Chart - Gender Distribution by Platform
    - Visualize the proportion of male vs. female users on each platform.
3. Bar Chart - Age Groups and AvgTime
    - Understand how different age groups interact with platforms and how much time.
4. Map - User Distribution by Location
    - Visualize where users are located geographically.

5.1.3. Engagement Efficiency Dashboard: Evaluate engagement efficiency and trends.

1. Bubble Chart - AvgTime vs. TotalEngagement
    - Analyze how time spent correlates with engagement on each platform.
2. Bubble Chart - Followers vs. PostsNumber vs. AvgTime
    - Show relationships between activity metrics on each platform.
3. Histogram - Posts Number Distribution
    - Show the frequency distribution of posts by users across platforms.

5.1.4. Demographics and Engagement Dashboard: Cross-analyze demographics with engagement.

1. Pie Chart – Engagement Share by Platform
    - Why this chart? Visualizes the share of total engagement across platforms.
2. Treemap - Age Groups and AvgTime
    - Highlight differences in time spent across age groups.
3. Treemap - Engagement by Category (Per Platform)
    - Highlight which categories perform best on specific platforms.

6. **Insights from dashboards:**

User Activity Per Platform:

- **Bar Chart - Total Users Per PlatformBar Chart:**

TikTok's Dominance: TikTok has a substantial lead in user count, suggesting it is the most popular platform among the three. This could be due to its engaging and trending content style.

Instagram's Strong Position: Instagram holds a strong position with a considerable user base, indicating its continued relevance and appeal.

Facebook's Decline: Facebook's lower user count may reflect a shift in user preferences towards more visually engaging and short-form content platforms like TikTok and Instagram

- **Dual-Axis Chart - AvgTime and TotalEngagement**
  TikTok dominates in user engagement and average time spent due to its short, engaging video format. Instagram performs strongly with high engagement driven by its visual and interactive features. Facebook lags in both metrics, indicating a shift in user preference toward more dynamic platforms like TikTok and Instagram.
- **Map - Engagement by Location**

High Engagement: The United States stands out with the highest level of engagement.

Moderate Engagement: Countries like Canada, Brazil, and Russia show a moderate level of user interaction.

Lower Engagement: Most of Europe, Africa, and parts of Asia have lower levels of engagement.

Unique Engagement: India and China show distinct engagement patterns, possibly due to their large populations and differing internet usage behaviors.

Age Group Distribution by Platform

The bar chart shows the distribution of users across three social media platforms: Facebook, Instagram, and TikTok. The users are categorized into three age groups: Young Adult, Adult, and Senior. Each bar represents a platform and is divided into segments corresponding to the age groups.

TikTok:

- Has the largest user base overall.
- The Adult age group has the highest number of users, with 812,170,320 users.

Instagram:

- Has a larger user base than Facebook but smaller than TikTok.
- Significant representation across all age groups, with a relatively even distribution.

Facebook:

- Has the smallest user base among the three platforms.
- Predominantly used by the Senior age group.

Stacked Gender distribution by Platform

The chart illustrates the distribution of users by gender across three social media platforms: Facebook, Instagram, and TikTok. Each bar is divided into sections for male and female users, creating a stacked bar effect.

Facebook:

- Male Users: A significant portion of the bar.
- Female Users: Another substantial portion, slightly less than males.

Instagram:

- Male Users: Slightly more balanced compared to Facebook.
- Female Users: A larger section than males, indicating a higher number of female users.

TikTok:

- Male Users: Similar pattern as the other platforms.
- Female Users: The largest portion of the bar, indicating the highest number of female users among the three platforms.

Bar Age Groups and Average Time

The chart displays the average time spent by different age groups: Adult, Senior, and Young Adult. The results suggest that the average time spent by users in these age groups is relatively consistent.

Adult: The average time is approximately 2.6.

Senior: The average time is also approximately 2.6.

Young Adult: The average time is similarly around 2.6.

The chart shows that all three age groups—Adult, Senior, and Young Adult—have a similar average time of approximately 2.6.

User Distribution by Location

- The United States has the highest number of users, as indicated by the darker shade of blue.
- Other countries have a significantly lower user count, as represented by the lighter shades of blue.

Followers Vs Posts Number Vs Average Time

- The largest bubble represents the account with the highest number of followers.
- The color of the bubbles indicates that the average time spent on the platform varies across different accounts.

Average time Vs Total Engagement

- The chart uses a bubble chart to represent the relationship between two variables: Average Time and Total Engagement. The size of each bubble corresponds to the Total Engagement, while the color represents the Age Group. The chart focuses on three age groups: Adult, Senior, and Young Adult.
- The largest bubble represents the age group with the highest total engagement.
- The color-coding indicates that the age groups have different levels of average time and total engagement.