Tweets and User Engagement

Introduction:

**Statement of Purpose**: This analysis explores Twitter data to extract actionable insights for content creation and engagement strategy.

**Background Information**: The Twitter dataset offers a comprehensive view of tweets, user behaviors, and influential factors like Klout scores.

**Target Audience**: Tailored for researchers, analysts, and social media strategists aiming to optimize their Twitter presence.

Problem Statement:

**Goals and Objectives:** Unravel patterns in engagement, Klout score impact, sentiment, temporal trends, geographic influences, language dynamics, and gender participation.

**Scope and Limitations:** This analysis focuses on the specific dataset. It's important to acknowledge potential limitations, such as:

* Data Representativeness: The dataset may not fully represent all Twitter activity due to sampling methods and timeframe.
* Data Quality: Missing values or errors in the data could affect the accuracy of results.
* Analytical Biases: Choices made during analysis might influence conclusions drawn.

Data Source and Collection:

**Data Source**:

* The dataset, titled "Tweets and User Engagement," was acquired from the public repository Kaggle: <https://www.kaggle.com/datasets/thedevastator/tweets-and-user-engagement>

**Description of the Dataset**:

* Contains 100,000 records of tweets and related information.
* Features:
  + TweetID (unique identifier)
  + Weekday, Hour, Day (temporal information)
  + Lang (language of the tweet)
  + IsReshare (indicates if a tweet is a retweet)
  + Reach, RetweetCount, Likes (engagement metrics)
  + Klout (user influence score)
  + Sentiment (sentiment analysis score)
  + Text (content of the tweet)
  + LocationID, UserID (links to user and location details)
  + Country, State, StateCode, City (geographical information)
  + Gender (gender of the user)
  + Sentiment Category (categorical sentiment label)
  + Language Name (language name in plain text)

**Data Collection Methods**:

* The dataset was collected and published on Kaggle by a third party.
* Collection involved merging three tables (Tweet data, Location, and Gender) using XLOOKUP in Excel.
* Ethical Considerations:
  + The dataset is publicly available, suggesting no major ethical concerns.
  + However, it's essential to respect Twitter's terms of service and user privacy when analyzing this data.

**Data Preprocessing**:

1. Merging Tables: Three separate tables (tweets, locations, and genders) were merged into a single table using XLOOKUP functions in Excel.
2. Handling Missing Values: One missing value in the Gender column was imputed with the most common value (Male).
3. Data Type Check: Data types were primarily left as "General" to allow Excel to interpret them appropriately.
4. Addressing Inconsistencies: Minor inconsistencies in country names were corrected for consistency.
5. Removing Unnecessary Columns: Irrelevant columns were removed to streamline the analysis.

**Sample Size and Characteristics**:

* The dataset includes 100,000 tweets, a relatively large sample.
* The dataset's size and variety of features suggest it can provide valuable insights into Twitter user behavior and engagement patterns.

Data Cleaning:

**Step 1 : Table Merging**

We have three separate table. Tweet data, Location and Gender.

We need to merge all three table and make single table. Following the cleaning steps:

1. Save as the CSV file in xlxs format.

2. Merge the "Tweet" and "Location" tables using the common column "LocationID" with the XLOOKUP formula.

* Formula: =XLOOKUP(N2,$W$1:$W$6290,$W$1:$AA$6290)

3. Fill the formula in all records using copy and paste special.

- Copy the formula (Ctrl+C).

- Select the range of cells where you want to paste the formula.

- Right-click on the selected range and choose "Paste Special."

- In the ‘Paste Special’ dialog box, choose "Formulas."

- Click OK.

- Copy the value of XLOOKUP and Paste it as Values.

4. Delete the ‘Location’ table.

5. Merging Tweet table with Gender table. With below XLOOKUP formula and rest process will remain same.

=XLOOKUP(O2,$AB$1:$AB$100001,$AB$1:$AC$100001)

**Steps 2: Handle the missing values**

1. Navigate to the ‘Home’ tab -> Find & Select -> Go to Special -> select ‘Blanks’ -> click ‘ok’
   * Observation: 1 cell is blank in column 'Gender'
2. Impute missing values with the most common value (Male). For that we need to find out the most count value by Gender category.
3. Checking the Distribution of Gender Values
   * Unique values of the Gender column : ‘=UNIQUE(T1:T57484)’
   * Count of each Gender values : ‘=COUNTIF($T$1:$T$100001,V4)’
   * Result :

|  |  |
| --- | --- |
| Gender | Count |
| Female | 9878 |
| Male | 51688 |
| Unisex | 8994 |
| Unknown | 29440 |

4. Since the most common Gender is Male, we are filling the blank cell with the value ‘Male’.

**Step 3 : Data Type Check**

Most of features are in Text or Number so we let the datatype as ‘General’.

"General" format in Excel interprets and displays data based on content.

**Step 4 : Removing Duplicates**

Data -> Data Tools -> Remove Duplicates -> Select All -> ok

No Duplicate values found.

**Step 5 : Addressing Inconsistencies**

Identified and merged four inconsistent country names for consistency.

Bolivia | Bolivia (Plurinational State of)

Venezuela | Venezuela (Bolivarian Republic of)

Taiwan | Taiwan (Province of China)

Viet Nam | Vietnam

**Step 6 : Removed Unnecessary Columns**

Now we are ready to make the analysis.

**Descriptive Analysis & EDA**

* We’ll focus on summarizing and characterizing the overall data set.
* Also aims to uncover hidden patterns, trends, and anomalies in the data.

1. Engagement Metrics Analysis:

Analyzing **retweets, likes, and reach** to quantify popularity, assess user influence, and identify outliers.

**1.1 Summary statistics**

Navigate to Data -> Analysis -> Data Analysis -> Descriptive Statistics -> select Input Range, check Summary Statistics -> ok

Also get the first and third percentile to get clear idea about values range.

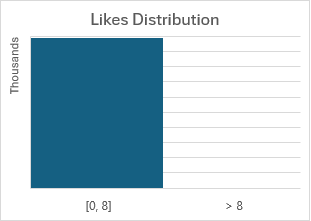
1st Quartile (25%) : =QUARTILE.INC(Data!I2:I100001,1)

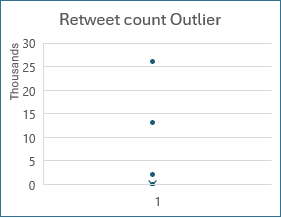
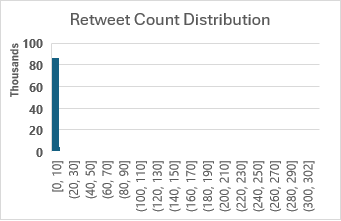
3rd Quartile (75%): =QUARTILE.INC(Data!I2:I100001,3)

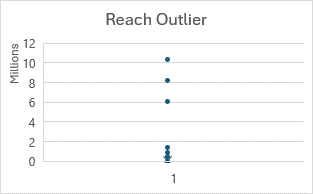
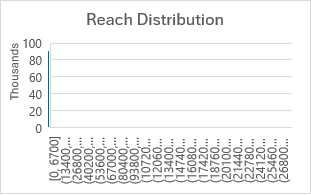
Below are the findings:

|  |  |  |  |
| --- | --- | --- | --- |
| *Statistics* | *RetweetCount* | *Likes* | *Reach* |
| Mean | 8.05275 | 0.15277 | 8542.39598 |
| Minimum | 0 | 0 | 0 |
| 25% | 0 | 0 | 151 |
| Median | 0 | 0 | 448.5 |
| 75% | 3 | 0 | 1496 |
| Maximum | 26127 | 133 | 10342452 |
| Mode | 0 | 0 | 4 |
| Standard Error | 0.309471479 | 0.008170164 | 280.4000026 |
| Standard Deviation | 97.86347449 | 2.583632729 | 88670.26642 |
| Sample Variance | 9577.25964 | 6.675158079 | 7862416147 |
| Kurtosis | 54039.45909 | 638.8127394 | 4868.899236 |
| Skewness | 214.9690573 | 22.45129288 | 57.80700402 |
| Range | 26127 | 133 | 10342452 |
| Sum | 805275 | 15277 | 854239598 |
| Count | 100000 | 100000 | 100000 |

**Visualization** : Histogram for value distribution and Boxplot to check outliers in values.







**Observations and Interpretations**:

Overall Engagement:

* Retweets: On average, tweets receive 8 retweets, but the distribution is highly skewed, with a median of 0 and a maximum of 26k. This suggests that a few tweets get significantly more retweets than most.
* Likes: Likes are even less common, with an average of only 0.15 per tweet. This suggests that users are more likely to retweet content than to actively like it.
* Reach: Third quartile reach is 1,496. However, the wide range (0 to 10M) suggests that reach varies greatly across tweets.
* In summary, the data suggests a skewed distribution with a majority of posts having low engagement, but with some outliers that significantly impact the mean and distribution shape. The high standard deviations and ranges highlight the need for further investigation into the factors influencing engagement disparities.

**1.2 Correlations between engagement metrics**

What is the relationship between reach, retweets, and likes on Twitter?

Formula : =CORREL(H2:H100001,I2:I100001)

Result:

|  |  |  |
| --- | --- | --- |
| Correlation Between Two Feature | | |
| Reach & Retweet | Reach & Likes | Retweet & Likes |
| 0.013426755 | 0.328170509 | 0.017383319 |

* The correlations between engagement features are relatively weak, indicating that they are not strongly related to each other.

**2. Klout Score Analysis:** Influence and Reputation

Quick insights into Twitter users' online influence.

What is the distribution of Klout scores among Twitter users in the dataset?

How do different users' Klout scores correlate with their engagement metrics?

**2.1. Summary Statistics:**

Navigate to Data -> Analysis -> Data Analysis -> Descriptive Statistics -> select Input Range, check Summary Statistics -> ok

Also get the first and third percentile to get clear idea about values range.

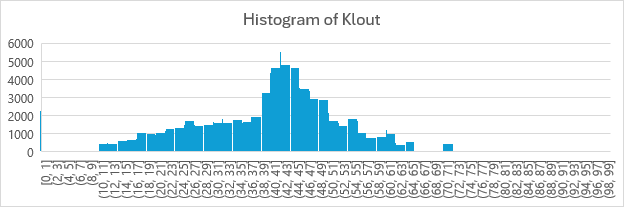
1st Quartile (25%) : =QUARTILE.INC(Data!K1:K100001,1)

3rd Quartile (75%): =QUARTILE.INC(Data!K1:K100001,3)

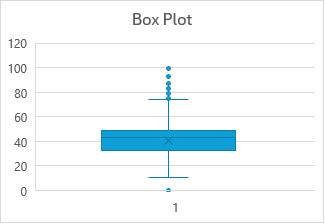
Below are the findings:

|  |  |
| --- | --- |
| *Klout* | |
| Mean | 40.38926 |
| Minimum | 0 |
| 25% | 32 |
| Median | 43 |
| 75% | 49 |
| Maximum | 99 |
| Standard Error | 0.043122439 |
| Mode | 43 |
| Standard Deviation | 13.63651261 |
| Sample Variance | 185.9544762 |
| Kurtosis | 0.752372747 |
| Skewness | -0.53435224 |
| Range | 99 |
| Sum | 4038926 |
| Count | 100000 |

**Distribution:** histogram to visualize the distribution of Klout scores.



**Outliers in the Klout scores**



**Observations and Interpretations :**

Overall Score:

* The average Klout score is 40.39, indicating a moderate level of perceived influence within the Twitter community.
* The score distribution is slightly right-skewed, with more users falling in the lower range than the higher range.

Interpretations:

* The data suggests that most tweets in the analyzed dataset belong to users with moderate levels of Klout score, while a small portion comes from more highly influential users.
* This distribution reflects the nature of Twitter itself, where a large number of ordinary users participate alongside a smaller group of well-known personalities or thought leaders.
* Focusing on content quality and engagement strategies can help users with lower Klout scores gain further influence and reach a wider audience.

**2.2 Klout Correlation with Engagement Metrics**

helps understand the relationships between a user's influence (Klout) and their engagement metrics on Twitter.

Formula : =CORREL(Data!K1:K100001,Data!H1:H100001)

Result :

|  |  |  |
| --- | --- | --- |
| Correlation | | |
| Kluot & Reach | Klout & Retweet | Klout & Likes |
| 0.196436222 | -0.021488449 | 0.135158534 |

* Klout score weakly linked to engagement. Klout score doesn't directly translate to a significant increase in these engagement metrics.

**3. Sentiment Analysis**

3.1. **Sentiment Metrics & Distribution:**

How are sentiments distributed among the tweets?

Summary statistics:

|  |  |
| --- | --- |
| *Sentiment Summary Statistics* | |
| Mean | 0.380921427 |
| Minimum | -6 |
| 25% | 0 |
| Median | 0 |
| 75% | 0.666667 |
| Maximum | 7.333333 |
| Mode | 0 |
| Standard Deviation | 1.046559317 |
| Standard Error | 0.003309511 |
| Sample Variance | 1.095286404 |
| Kurtosis | 3.202644949 |
| Skewness | 0.677057175 |
| Range | 13.333333 |
| Sum | 38092.14268 |
| Count | 100000 |

**Sentiment Distribution :**

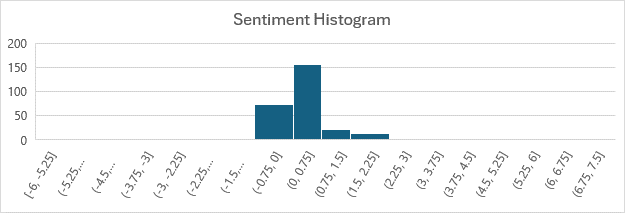
Sentiments : =UNIQUE(L2:L100001)

Counts : =COUNTIF(L2:L100001,W3)

Percentage : =B47:B355/B356\*100

Result :

|  |  |  |
| --- | --- | --- |
| **Sentiments** | **Counts** | **Percentage** |
| 0 | 67424 | 67% |
| 1 | 8771 | 9% |
| 2 | 8587 | 9% |
| 3 | 3945 | 4% |
| -1 | 1770 | 2% |
| -2 | 1729 | 2% |
| .. | … | … |
| 0.1088 | 1 | 0% |
| 0.136351 | 1 | 0% |
| **Total :** | **100000** | **100%** |



**Observation** : The majority of sentiments are neutral, but there is notable diversity, including both positive and negative expressions. The positive skewness and positive kurtosis suggest a distribution with a tendency toward positive sentiment and a concentration around the neutral point.

**3.2 Sentiment Category :**

What is the distribution of tweet sentiment categories (Positive, Neutral, Negative) in this dataset?

Created new column to bring the Sentiments number range in three categories. Positive, Neutral and Negative sentiments.

Formula : =IF(L2>0,"Positive",IF(L2=0,"Neutral","Negative"))

Not let’s analyze the tweet sentiment category.

Below are the Pivot Table and Pie Chart:

Rate = =E3/$E$6\*100

Result :

|  |  |  |
| --- | --- | --- |
| Sentiment Category | Count | Rate |
| Neutral | 67424 | 67% |
| Positive | 27024 | 27% |
| Negative | 5552 | 6% |
| **Total:** | **100000** | **100%** |

Interpretation: Content received favorable response with limited criticism.

**3.3. Weekly Sentiment Breakdown: Likes and Retweets on Twitter**

|  |  |  |
| --- | --- | --- |
| **Row Labels** | **Sum of Likes** | **Sum of RetweetCount** |
| **Negative** |  |  |
| Sunday | 43 | 1534 |
| Monday | 42 | 1872 |
| Tuesday | 39 | 3195 |
| Wednesday | 39 | 3242 |
| Thursday | 31 | 17079 |
| Friday | 0 | 4249 |
| Saturday | 0 | 2182 |
| **Neutral** |  |  |
| Sunday | 412 | 45883 |
| Monday | 1538 | 61255 |
| Tuesday | 1995 | 84615 |
| Wednesday | 1859 | 122520 |
| Thursday | 2058 | 81110 |
| Friday | 1492 | 75227 |
| Saturday | 943 | 39144 |
| **Positive** |  |  |
| Sunday | 112 | 17059 |
| Monday | 995 | 45701 |
| Tuesday | 668 | 53376 |
| Wednesday | 991 | 40480 |
| Thursday | 1025 | 40150 |
| Friday | 815 | 39902 |
| Saturday | 180 | 25500 |
| **Grand Total** | **15277** | **805275** |

Interpretation:

* Midweek buzz: Engagement peaks on Wednesdays and Thursdays.
* Weekends see a general drop in engagement for all sentiment categories.
* Users prioritize neutral content for engagement, potentially valuing informative or factual updates.

**4. Temporal Engagement Analysis**

studying patterns, trends, or behaviors over time

4.1. **Engagement by Hours and Weeks:**

* What are the engagement levels based on the hour of the day and day of the week?

Pivot tables: Twitter engagement Patterns – Retweet, Likes, Reach by Weeks and Hours

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sum of RetweetCount** | **Weekdays** |  |  |  |  |  |  |  |
| **Hours** | **Sunday** | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** | **Saturday** | **Grand Total** |
| 0 | 1257 | 2773 | 5432 | 7054 | 5599 | 6469 | 2342 | 30926 |
| 1 | 2557 | 3061 | 3915 | 8199 | 5167 | 4632 | 2796 | 30327 |
| 2 | 3255 | 3133 | 6411 | 7816 | 3007 | 4082 | 3743 | 31447 |
| … | … | … | … | … | … | … | … | … |
| 23 | 768 | 6277 | 7671 | 4384 | 5615 | 3041 | 2029 | 29785 |
| **Grand Total** | **64476** | **108828** | **141186** | **166242** | **138339** | **119378** | **66826** | **805275** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sum of Likes** | **Weekdays** |  |  |  |  |  |  |  |
| **Hours** | **Sunday** | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** | **Saturday** | **Grand Total** |
| 0 | 0 | 0 | 0 | 24 | 36 | 0 | 16 | 76 |
| 1 | 4 | 2 | 55 | 41 | 35 | 9 | 0 | 146 |
| 2 | 0 | 0 | 0 | 2 | 2 | 1 | 0 | 5 |
| … | … | … | … | … | … | … | … | … |
| 23 | 0 | 1 | 18 | 0 | 0 | 20 | 0 | 39 |
| **Grand Total** | **567** | **2575** | **2702** | **2889** | **3114** | **2307** | **1123** | **15277** |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Sum of Reach** | **Weekdays** |  |  |  |  |  |  |  |
| **Hours** | **Sunday** | **Monday** | **Tuesday** | **Wednesday** | **Thursday** | **Friday** | **Saturday** | **Grand Total** |
| 0 | 1497348 | 668988 | 1159896 | 10421448 | 2535179 | 1101894 | 2130701 | 19515454 |
| 1 | 1652216 | 3100442 | 13156678 | 4397665 | 5462148 | 3345734 | 931513 | 32046396 |
| 2 | 529537 | 652558 | 1284951 | 4313315 | 8972361 | 2699728 | 496698 | 18949148 |
| … | ... | ... | ... | ... | ... | ... | ... | ... |
| 23 | 821999 | 2258698 | 3599428 | 1899430 | 1205343 | 1044367 | 506300 | 11335565 |
| **Grand Total** | **33136465** | **132413017** | **171650687** | **160320752** | **199253504** | **117588530** | **39876643** | **854239598** |

Visual Charts :

Overall Patterns:

* Weekdays dominance: Peak reach, likes, and retweets from 10am-4pm on weekdays, with Wednesday on top.
* Thursdays intrigue: Potential content surge around 6pm, retweets especially.
* Weekends slump: Engagement metrics drop considerably on Saturdays and Sundays, with the lowest levels typically observed on Sundays.

4.2. **Engagement by Hours:**

Are there specific times when tweets tend to receive higher engagement?

Pivot tables :

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Hours** | **Sum of RetweetCount** |  | **Hours** | **Sum of Likes** |  | **Hours** | **Sum of Reach** |
| 0 | 30926 |  | 0 | 76 |  | 0 | 19515454 |
| 1 | 30327 |  | 1 | 146 |  | 1 | 32046396 |
| 2 | 31447 |  | 2 | 5 |  | 2 | 18949148 |
| ... | ... |  | ... | ... |  | ... | ... |
| 23 | 29785 |  | 23 | 39 |  | 23 | 11335565 |
| **Grand Total** | **805275** |  | **Grand Total** | **15277** |  | **Grand Total** | **854239598** |

**Observation:**

* Midday Magic: 10am-4pm peaks for likes, retweets, and reach.
* Morning Buzz: 7am-9am sees additional buzz, especially retweets.
* Retweets peak midday (10am-4pm) with notable surges in the morning (7am-9am) and early evening (7pm), suggesting prime times for content visibility and sharing.
* Likes peak midday (10am-4pm), with sunrise and sunset dips. Weekend likes plummet.
* Reach peaks prominently during midday hours (9am-4pm), with a notable early morning rise around 7am.

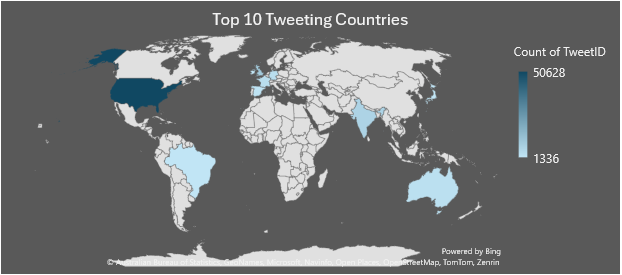
5. Geographic Analysis

How does engagement vary based on the geographic location of the tweets (country, city)?

**5.1. Engagement by Country/City:**

What is the distribution of tweets/retweet across different countries and cities?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Top 10 Tweeting Countries | |  |  | Top 10 Cites |  |  |
| **Country** | **Count of TweetID** | **Percentage** |  | **City** | **Count of TweetID** | **Percentage** |
| United States | 50628 | 60.5% |  | Seattle | 7004 | 24.4% |
| United Kingdom | 9862 | 11.8% |  | London | 5328 | 18.6% |
| India | 6850 | 8.2% |  | San Francisco | 4136 | 14.4% |
| France | 3648 | 4.4% |  | Paris | 2694 | 9.4% |
| Germany | 3532 | 4.2% |  | New York City | 2652 | 9.2% |
| Australia | 2937 | 3.5% |  | Boston | 1587 | 5.5% |
| Spain | 2092 | 2.5% |  | Bangalore | 1427 | 5.0% |
| Ireland | 1444 | 1.7% |  | Chicago | 1357 | 4.7% |
| Brazil | 1402 | 1.7% |  | Sydney | 1285 | 4.5% |
| Japan | 1336 | 1.6% |  | Pune | 1212 | 4.2% |
| **Grand Total** | **83731** | **100.0%** |  | **Grand Total** | **28682** | **100.0%** |



**Retweet by Country and City :**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Country** | **Sum of RetweetCount** | **Percentage** |  | **City** | **Sum of RetweetCount** | **Percentage** |
| USA | 359366 | 53% |  | Seattle | 69374 | 28% |
| India | 98503 | 15% |  | London | 32976 | 13% |
| UK | 60468 | 9% |  | Paris | 27794 | 11% |
| France | 39712 | 6% |  | Bexley | 26127 | 11% |
| Germany | 31996 | 5% |  | San Francisco | 24569 | 10% |
| Australia | 25561 | 4% |  | Bangalore | 17783 | 7% |
| Spain | 20932 | 3% |  | Tomball | 15347 | 6% |
| Brazil | 14603 | 2% |  | Beijing | 11811 | 5% |
| China | 12803 | 2% |  | New York City | 11533 | 5% |
| Netherlands | 10117 | 2% |  | Sydney | 11321 | 5% |
| **Grand Total** | **674061** | **100%** |  | **Grand Total** | **248635** | **100%** |



**Geographic Reach and Influence:**

* US Dominates: United States leads in all metrics - tweets, retweets, Klout score, and positive sentiment - solidifying its Twitter dominance.
* UK Support: United Kingdom follows consistently in second or third place for most metrics, indicating strong engagement.
* India Rising: India shines in retweets and sentiment, suggesting high user engagement and potentially positive content.
* Other Players: France, Germany, and Australia maintain a steady presence, while Brazil and Japan show potential for growth.

**City Spotlight:**

* Seattle Soars: Seattle takes the lead in both tweet volume and retweets, establishing itself as a major Twitter hub.
* London Calling: London ranks second in tweets and third in retweets, demonstrating significant reach and influence.
* Paris Buzz: Paris ranks high in retweets and sentiment, suggesting engaging content and positive reception.
* Emerging Cities: Bangalore, Bexley, San Francisco, and Tomball show promise with notable retweet activity.

5.2. **Influencer by Country:**

How are Klout scores distributed across countries?

|  |  |  |
| --- | --- | --- |
| Top 10 Klout-Ranked Countries | |  |
| **Country** | **Sum of Klout** | **Percentage** |
| United States | 2130885 | 62% |
| United Kingdom | 395766 | 12% |
| India | 253946 | 7% |
| Germany | 152108 | 4% |
| France | 148980 | 4% |
| Australia | 104057 | 3% |
| Spain | 81890 | 2% |
| Ireland | 55651 | 2% |
| Japan | 50822 | 1% |
| Brazil | 47780 | 1% |
| **Grand Total** | **3421885** | **100%** |
| overall influence of Twitter users from different countries | | |

**Klout Scores:**

* US Expertise: United States holds the highest Klout score, implying influential users and valuable content.
* UK Authority: United Kingdom follows closely, suggesting a community of knowledgeable and impactful users.
* Global Knowledge: India, Germany, and France showcase respectable Klout scores, indicating a presence of informed users.

**5.3. Sentiments Across Top 10 Nations:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | |  | |
| **Count of Sentiment Category** | | **Column Labels** | |  | |  |  |
| **Country** | | **Negative** | | **Neutral** | | **Positive** | **Grand Total** |
| United States | | 2782 | | 34410 | | 13436 | 50628 |
| United Kingdom | | 779 | | 6159 | | 2924 | 9862 |
| India | | 307 | | 4626 | | 1917 | 6850 |
| Australia | | 210 | | 1891 | | 836 | 2937 |
| Germany | | 158 | | 2444 | | 930 | 3532 |
| France | | 132 | | 2312 | | 1204 | 3648 |
| Japan | | 89 | | 952 | | 295 | 1336 |
| Spain | | 68 | | 1436 | | 588 | 2092 |
| Ireland | | 62 | | 1020 | | 362 | 1444 |
| Brazil | | 40 | | 902 | | 460 | 1402 |
| **Grand Total** | | **4627** | | **56152** | | **22952** | **83731** |

**Observation:**

* US Positivity: United States leads in positive sentiment, potentially due to a focus on upbeat content or a larger user base.
* Mixed Emotions: Most countries exhibit a mix of positive, neutral, and negative sentiments, suggesting diverse content and opinions.
* India's Engagement: India's high retweet volume alongside neutral and positive sentiment suggests engaging content that sparks conversation.

6. Language Analysis

What is the distribution of languages in the dataset?

|  |  |  |
| --- | --- | --- |
| Top 10 Languages on Twitter | |  |
| **Language** | **Count of Language Name** | **Rate** |
| English | 91886 | 93.0% |
| Spanish | 1700 | 1.7% |
| French | 1085 | 1.1% |
| Turkish | 877 | 0.9% |
| Undetermined | 861 | 0.9% |
| Japanese | 671 | 0.7% |
| German | 563 | 0.6% |
| Indonesian | 416 | 0.4% |
| Tagalog | 370 | 0.4% |
| Korean | 330 | 0.3% |
| **Grand Total** | **98759** | **100%** |

In which countries are different languages used the most?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Count of Language** | **Column Labels** | |  |  |  |  |
| **Country** | **English** | **French** | **Spanish** | **Japanese** | **German** | **Grand Total** |
| United States | 60% | 0.1% | 0.2% | 0.0% | 0.1% | 61% |
| United Kingdom | 12% | 0.0% | 0.0% | 0.0% | 0.0% | 12% |
| India | 8% | 0.0% | 0.0% | 0.0% | 0.0% | 8% |
| France | 3% | 1.1% | 0.0% | 0.0% | 0.0% | 4% |
| Germany | 4% | 0.0% | 0.0% | 0.0% | 0.5% | 4% |
| Australia | 3% | 0.0% | 0.0% | 0.0% | 0.0% | 3% |
| Spain | 2% | 0.0% | 0.6% | 0.0% | 0.0% | 2% |
| Ireland | 2% | 0.0% | 0.0% | 0.0% | 0.0% | 2% |
| Japan | 1% | 0.0% | 0.0% | 0.8% | 0.0% | 2% |
| Brazil | 1% | 0.0% | 0.0% | 0.0% | 0.0% | 1% |
| **Grand Total** | **97%** | **1.2%** | **0.8%** | **0.8%** | **0.6%** | **100%** |

How is sentiment distributed across different languages?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Tweet Sentiment Spectrum | |  |  |  |  |  |
| **Count of TweetID** | **Column Labels** |  |  |  |  |  |
| **Sentiment** | **English** | **French** | **Japanese** | **Spanish** | **Turkish** | **Grand Total** |
| Negative | 5% | 0% | 0% | 0% | 0% | 6% |
| Neutral | 64% | 1% | 1% | 1% | 1% | 67% |
| Positive | 26% | 0% | 0% | 1% | 0% | 28% |
| **Grand Total** | **95%** | **1%** | **1%** | **2%** | **1%** | **100%** |

**Observation :**

Global Tongue: English dominates with 93%, followed by Spanish at 1.7%.

Regional Voices: English dominates in most countries, but local languages thrive in France, Germany, Spain, and Japan.

Sentiment Shifts: predominant use of English, with minimal representation of other languages across sentiments.

7. Gender Analysis

**7.1 Gender Engagement Analysis Summary**

How is gender distributed in terms of engagements and Klout scores?

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gender Overview | |  | Reach by Gender | |  | Retweet by Gender | |  | Gender-wise Klout | |
| **Gender** | **Count of Gender** |  | **Gender** | **Sum of Reach** |  | **Gender** | **Sum of Retweet** |  | **Gender** | **Sum of Klout** |
| Female | 9878 |  | Female | 37201091 |  | Female | 59888 |  | Female | 387839 |
| Male | 51688 |  | Male | 162077476 |  | Male | 513704 |  | Male | 2042084 |
| Unisex | 8994 |  | Unisex | 33740186 |  | Unisex | 63384 |  | Unisex | 366566 |
| Unknown | 29440 |  | Unknown | 621220845 |  | Unknown | 168299 |  | Unknown | 1242437 |
| **Grand Total** | **100000** |  | **Grand Total** | **854239598** |  | **Grand Total** | **805275** |  | **Grand Total** | **4038926** |

**7.2 Sentiment and Geographic Analysis by Gender**

What is the distribution of gender sentiments, and how does gender vary across different countries?

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Count of Gender** | **Sentiments** | | | **Grand Total** |  | **Count of Gender** | **Countries** | | | | | **Grand Total** |
| **Gender** | **Negative** | **Neutral** | **Positive** |  | **Gender** | **US** | **UK** | **India** | **France** | **Germany** |
| Female | 5% | 70% | 24% | 100% |  | Female | 4977 | 1116 | 646 | 165 | 228 | 7132 |
| Male | 6% | 65% | 29% | 100% |  | Male | 24041 | 5609 | 4090 | 2724 | 1751 | 38215 |
| Unisex | 6% | 70% | 24% | 100% |  | Unisex | 5780 | 713 | 361 | 148 | 111 | 7113 |
| Unknown | 5% | 70% | 25% | 100% |  | Unknown | 15830 | 2424 | 1753 | 611 | 1442 | 22060 |
| **Grand Total** | **6%** | **67%** | **27%** | **100%** |  | **Grand Total** | **50628** | **9862** | **6850** | **3648** | **3532** | **74520** |

**Twitter Gender: A Quick Look**

* M:F: Mostly male (52% vs 10%), with big "Unknown" and "Unisex" chunks.
* Reach & Retweets: Males rule, spreading content further and sparking more buzz.
* Klout Kings: Males hold the influence crown, deemed more expert and impactful.
* Sentiment: Everyone's mostly neutral (65-70%), with similar positive and negative vibes across genders.
* Country Shifts: Gender gap varies, US widest, India more balanced.

**7.3 Hourly and Weekly Gender Activity**

Do different genders exhibit different patterns in daily tweeting activity?

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Tweet : Gender Breakdown by Hour | | | | |  |  | Genderwise Retweet Activity Throughout the Day | | | | | |  | Weekly Gender Share Analysis | | | |  | |  |  |
| **Count of Gender** | **Hours** | | | | **Grand Total** |  | **Sum of Retweet** | **Hours** | | | | **Grand Total** |  | **Count of Gender** | **Weeks** | | | | | | **Grand Total** |
| **Gender** | **0** | **1** | **...** | **23** |  | **Gender** | **0** | **1** | **...** | **23** |  |  | **Gender** | **Sunday** | **Monday** | **…** | | **Saturday** | |  |
| Female | 3% | 3% | ... | 3% | 100% |  | Female | 3% | 4% | ... | 4% | 100% |  | Female | 9% | 16% | … | | 9% | | 100% |
| Male | 3% | 3% | ... | 3% | 100% |  | Male | 4% | 4% | ... | 4% | 100% |  | Male | 7% | 15% | … | | 7% | | 100% |
| Unisex | 3% | 3% | ... | 3% | 100% |  | Unisex | 2% | 3% | ... | 2% | 100% |  | Unisex | 7% | 16% | … | | 9% | | 100% |
| Unknown | 3% | 3% | ... | 2% | 100% |  | Unknown | 4% | 3% | ... | 3% | 100% |  | Unknown | 7% | 15% | … | | 7% | | 100% |
| **Grand Total** | **3%** | **3%** | **...** | **3%** | **100%** |  | **Grand Total** | **4%** | **4%** | **...** | **4%** | **100%** |  | **Grand Total** | **7%** | **15%** | **…** | | **7%** | | **100%** |

**Observation :**

1. Gender by Hour: Uniform engagement, peaking at 6-7 AM and 3-4 PM.

2. Retweet Activity by Gender: Similar patterns, males more active in the evening.

3. Weekly Gender Share: Weekdays see higher engagement. Males slightly more active on weekdays.

4. Overall: Diverse gender representation.

Limitations and Assumptions

Data Quality and Availability:

* Data Representativeness: This analysis is based on a specific dataset, which may not be representative of all Twitter activity. The source, collection methods, and timeframe could bias the results.
* Missing Data: The analysis may have excluded tweets with missing values or incomplete information, potentially impacting the accuracy of conclusions.
* Measurement Errors: Metrics like Klout score or sentiment analysis algorithms might have inherent biases or inaccuracies, affecting the interpretations.

Analysis Assumptions:

* Stationarity: The analysis assumes trends and patterns observed in the data are consistent over time, which may not hold true for dynamic platforms like Twitter.
* Causality: Correlations between metrics don't necessarily imply causation. Further analysis is needed to determine the cause-and-effect relationships.
* Feature Engineering: Choices made during feature selection and processing can influence the conclusions drawn.

Potential Biases and Limitations:

* Sampling Bias: The data collection process might have inadvertently favored certain types of users or content, introducing bias into the analysis.
* Language Bias: The dominance of English in the data may mask trends or sentiments in other languages.
* Social Desirability Bias: Users might misrepresent their opinions or behavior in public platforms, impacting sentiment analysis.
* Algorithmic Bias: The tools used for analysis, like sentiment algorithms, might have inherent biases that skew the results.

Limited data, analysis choices, and potential biases might skew results. Consider these insights specific to this dataset, not absolute Twitter truths.

Conclusion

• Most tweets get low engagement, but a few stand out.

• Users have moderate Klout and mostly express positive sentiment.

• Weekday mornings and afternoons see peak engagement in the US.

• Gender gap exists, with males leading in activity and influence.

Recommendations

• Craft high-quality content and post during peak times.

• Collaborate with relevant influencers and tailor content for diverse audiences.

• Deep-dive into sentiment analysis and retweet drivers.

• Experiment with language targeting and A/B testing.

• Monitor Klout scores and influencer performance.

Github Repository

Repo Link :

<https://github.com/abhishekmishra8/Excel-Project-Hotel-Booking-Cancellation-Data-Analysis>