King Saud University

College of Computer Science and Information

Department of Information Technology

Data Science & Artificial Intelligence Track

جامعة
الملك سعود
King Saud University
1957

Logbook Entry:

Google Trends API Data Analysis for Saudi Arabia (2020-2024)

Group#1

Supervised By:

Dr. Khulood Alyahya

2024-2025

1446-1447

- ## Introduction

Primary Objective:

The main goal of collecting Google Trends data for Saudi Arabia is to analyze the interest in social media and mental health across the country over specific time periods between 2020 and 2024. This analysis will help to understand how interest in social media evolves and its potential impact on mental health in Saudi Arabia over time.

Secondary Objectives:

- Trend Analysis Over Time:

We aim to determine whether there is an increase or decrease in interest in social media and mental health over the years in Saudi Arabia. This will help identify any emerging patterns or changes in public concern related to these topics.

- Drawing Conclusions About the Potential Impact of Social Media:

Based on the data, we will be able to form preliminary conclusions about the relationship between increased social media usage and the level of interest in mental health in the country. This may offer insights into how digital engagement correlates with mental health awareness or concern within the Saudi cultural context.

- ## Data Collection

- Source of Dataset:

The dataset was sourced from Google Trends, accessible via the following link: [Google Trends.]
The data was retrieved programmatically using the 'pytrends' library, which provides a Python interface to interact with the Google Trends API.

- Data Collection steps:

1. Importing Necessary Libraries
   - from pytrends.request import TrendReq:

   We imported TrendReq from pytrends, allowing us to access Google Trends data. This is the main object used to send search queries to Google Trends.
   - import "pandas as pd":

   We imported Pandas, a popular data analysis library, to organize and manipulate the data using DataFrames for easy analysis.
   - import "time":

We imported time to manage time-based functions like delaying execution between requests to avoid overwhelming Google with too many queries at once.

- import "random":

We imported random to introduce random delays between requests, ensuring that our requests appear natural and avoid triggering any rate limits.

- from pytrends.exceptions import ResponseError:

We imported ResponseError to handle potential errors from Google Trends, allowing us to retry requests or manage failures gracefully.

2. **Setting up Google Trends Request:**

We are initializing a `TrendReq` object from the pytrends library with specific parameters:

- hl='en-US': Sets the language to English (US).
- tz=360: Sets the time zone to UTC+6 hours.
- timeout=(10, 25): Sets timeouts for connection (10 seconds) and data retrieval (25 seconds).

This setup prepares the object for sending queries to Google Trends with specified language, time zone, and timeout settings.

3. **Defining Keyword Groups:**

- Keyword Group 1: Social media, Mental health, Depression and social media, Anxiety and social media

- Keyword Group 2: Insomnia and social media, Stress and social media, Addiction and social media
- Keyword Group 3: Instagram, Twitter, Platform X, Facebook
- Keyword Group 4: Snapchat, TikTok, LinkedIn, YouTube, WhatsApp

This section of the code organizes the keywords into four distinct groups. These keywords will be used for Google Trends data analysis to track public interest in social media platforms and their potential impact on mental health.

The keywords were divided into two groups for several practical reasons:

- Reduce API request load:

By splitting the keywords into smaller groups, we can perform multiple API calls without overloading the system, ensuring smooth data retrieval.

- Data Accuracy:

By splitting the terms, we can ensure more accurate and focused data collection. Querying too many terms at once can dilute the relevance of the data, especially when comparing distinct platforms or topics.

This separation makes it easier to conduct a detailed analysis of trends for both social media platforms and mental health topics without overwhelming the system.

4. **Specifying Time Ranges for Each Year:**
   The dictionary years specifies the time range for each year from 2020 to 2024 because the Google Trends API does not allow for direct yearly data retrieval. Instead, the API requires precise date ranges that include specific months and days. By defining each year with exact start and end dates (from January 1st to December 31st), we can simulate yearly data collection.
   This step was necessary because the API only accepts data requests with monthly and daily granularity, not by year. Therefore, creating these specific date ranges ensures that we can retrieve data for an entire year without gaps.
5. **Function to Fetch Google Trends Data:**
   The 'fetch_data_with_retry' function retrieves Google Trends data while implementing a retry mechanism. It attempts to fetch data multiple times if an error occurs, particularly focusing on handling situations where the rate limit is exceeded.
   The main goal of this function is to ensure reliable data retrieval from Google Trends, even when facing rate limit errors. By retrying requests with increasing wait times, it enhances the robustness of the data collection process, allowing for smoother operations when querying frequent or large sets of keywords.
6. **Function to fetch data for a country:**
   The fetch_data_for_country function is designed to retrieve Google Trends data for a specified country(SA) and a set of keyword groups over multiple years. It aggregates the results into a single DataFrame, allowing for straightforward analysis of trends related to the specified keywords.
   The main goal of this function is to systematically collect and organize Google Trends data for a given country across different years and keywords.
7. **Fetch data from all keyword groups:**
   The code snippet initializes an empty DataFrame named all_data and then populates it by concatenating the results from multiple calls to the fetch_data_for_country function. Each call retrieves Google Trends

data for a specified country (SA) and a different group of keywords (keywords_group1, keywords_group2, keywords_group3, and keywords_group4).

The purpose of this code is to aggregate Google Trends data related to various keyword groups into a single DataFrame (all_data).This unified dataset facilitates easier exploration and comparison of the trends over time.

8. **Exporting to CSV**
9. **Data Organization and Preparation:**
   When we exported the CSV file, we noticed the need for further organization and arrangement of the data, so we took the following steps:
   Data Cleaning and Organization Steps:
   - Merging "Twitter" and "Platform X" Columns Reason: To eliminate redundancy caused by Twitter's rebranding to "X," both columns were merged into Twitter(X) for consolidated data.
   - Dropping Original Columns Reason: After merging, the separate "Twitter" and "Platform X" columns were removed to maintain a clean dataset.
   - Dropping the "date" Column Reason: Since we focus on yearly data with the "Year" column, the more granular "date" column was removed to simplify the dataset.
   - Reordering Columns Reason: Columns were rearranged to place time-related data first, enhancing readability and facilitating analysis.

   Objective:
- Eliminate Redundancy: Combine similar data for clarity.
- Simplify Data: Remove unnecessary columns.
- Improve Organization: Create a structured dataset for easier analysis.

• Advanced exploratory data analysis (EDA)

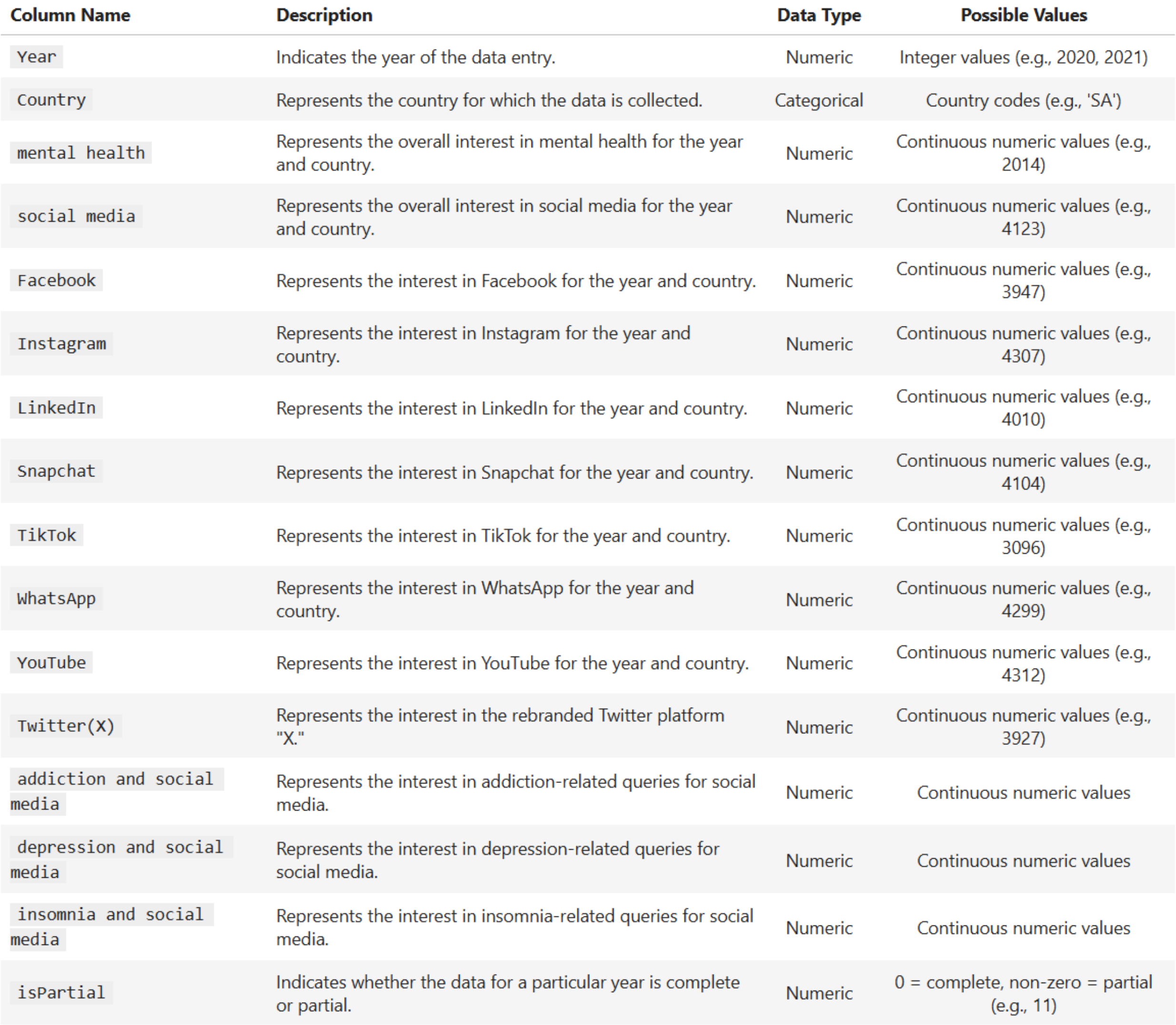
At this stage of my analysis, I will focus on exploring the data I collected from Google Trends through Exploratory Data Analysis (EDA). My goal is to gain a deep understanding of the structure, quality, and content of the data, which will help me identify patterns and relationships between variables. This step is crucial for properly preparing the data before I conduct more advanced analyses.

- o Analysis Steps:
  - ▪ Structure Analysis: I will explore the overall shape of the Google Trends dataset, including the number of rows and columns, and examine the data types used in each column, such as time series and trend values.

- Quality Analysis: I will check for any missing or duplicated values in the Google Trends data and address any unwanted entries that may affect my analysis.
- Content Analysis: I will dive deep into the values and variables within the Google Trends dataset, examining the relationships between them using statistical analysis and graphical visualizations to illustrate patterns and trends.

1. Structure Investigation

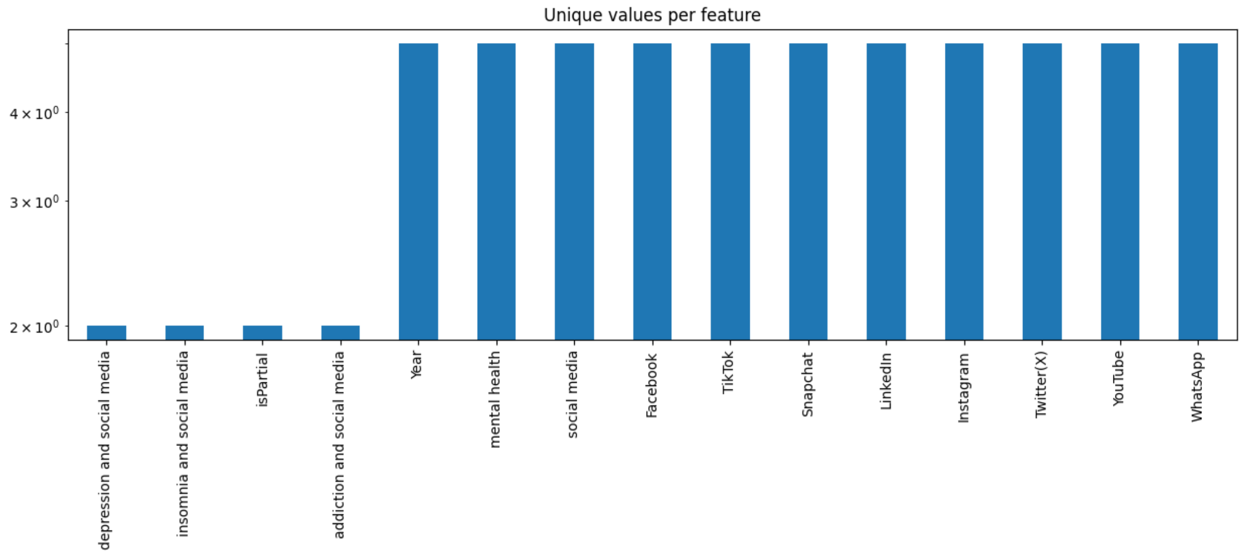
We have 5 rows and 16 Columns, See details for each column in this table:



Column Name	Description	Data Type	Possible Values
`Year`	Indicates the year of the data entry.	Numeric	Integer values (e.g., 2020, 2021)
`Country`	Represents the country for which the data is collected.	Categorical	Country codes (e.g., 'SA')
`mental health`	Represents the overall interest in mental health for the year and country.	Numeric	Continuous numeric values (e.g., 2014)
`social media`	Represents the overall interest in social media for the year and country.	Numeric	Continuous numeric values (e.g., 4123)
`Facebook`	Represents the interest in Facebook for the year and country.	Numeric	Continuous numeric values (e.g., 3947)
`Instagram`	Represents the interest in Instagram for the year and country.	Numeric	Continuous numeric values (e.g., 4307)
`LinkedIn`	Represents the interest in LinkedIn for the year and country.	Numeric	Continuous numeric values (e.g., 4010)
`Snapchat`	Represents the interest in Snapchat for the year and country.	Numeric	Continuous numeric values (e.g., 4104)
`TikTok`	Represents the interest in TikTok for the year and country.	Numeric	Continuous numeric values (e.g., 3096)
`WhatsApp`	Represents the interest in WhatsApp for the year and country.	Numeric	Continuous numeric values (e.g., 4299)
`YouTube`	Represents the interest in YouTube for the year and country.	Numeric	Continuous numeric values (e.g., 4312)
`Twitter(X)`	Represents the interest in the rebranded Twitter platform "X."	Numeric	Continuous numeric values (e.g., 3927)
`addiction and social media`	Represents the interest in addiction-related queries for social media.	Numeric	Continuous numeric values
`depression and social media`	Represents the interest in depression-related queries for social media.	Numeric	Continuous numeric values
`insomnia and social media`	Represents the interest in insomnia-related queries for social media.	Numeric	Continuous numeric values
`isPartial`	Indicates whether the data for a particular year is complete or partial.	Numeric	0 = complete, non-zero = partial (e.g., 11)

1.1. Structure of Numerical Features

Here, we examined the numeric attributes in more detail. More specifically, we looked at the number of unique values for each of these attributes.



The following analysis explores the unique values in the dataset features and categorizes them accordingly.

- Number of Unique Values:

The bar chart displays the number of unique values for each feature in the dataset. This helps us understand the diversity of the data and its distribution across various features.

- Features with High Unique Values:

Features such as year contain a very large number of unique values, indicating that they represent continuous temporal data, encompassing different years.

- Features with Limited Unique Values:

Features like addiction and social media, depression and social media, and insomnia and social media have a limited number of unique values; however, they are not binary data in the traditional sense. Instead, these features may represent metrics related to the impact of social media on feelings of addiction, depression, and insomnia, which are continuous metrics that can take on various values reflecting the intensity of these feelings.

- Classification of Features Based on Unique Values:
  - Continuous Data: All displayed features, including those related to addiction and depression, can be considered continuous data as they may reflect varying levels of impact or interaction with social media.
  - Temporal Data: Features such as year indicate temporal data, allowing for trend analysis over time.

1.2 Conclusion of Structure Investigation

This dataset consists of 5 samples and 16 features, where each sample represents an individual case with values across the features.

Key Findings:

- Data Types:
  - There are 13 columns of float64 type (numerical values with decimals).
  - 2 columns are of int64 type (integer values).
  - 1 column is of object type, representing categorical data for the country ("SA" for Saudi Arabia).

- Numerical Features:
  - Features like **year** have a high number of unique values, indicating they are continuous temporal data.
  - Features such as **addiction and social media**, **depression and social media**, and **insomnia and social media** have limited unique values. These metrics capture varying levels of impact from social media, rather than being binary.

- Classification:
  - All features are classified as continuous data, reflecting varying levels of impact.
  - year features are classified as temporal data, suitable for trend analysis over time.

This investigation provides insights into the data's diversity and distribution, serving as a foundation for further analysis.

2. Quality Investigation(Data Cleaning)

Before analyzing the specific content of our Google Trends data, it's important to assess the overall dataset quality. This includes checking for duplicates, missing values, and any unwanted entries or potential recording errors to ensure data integrity.

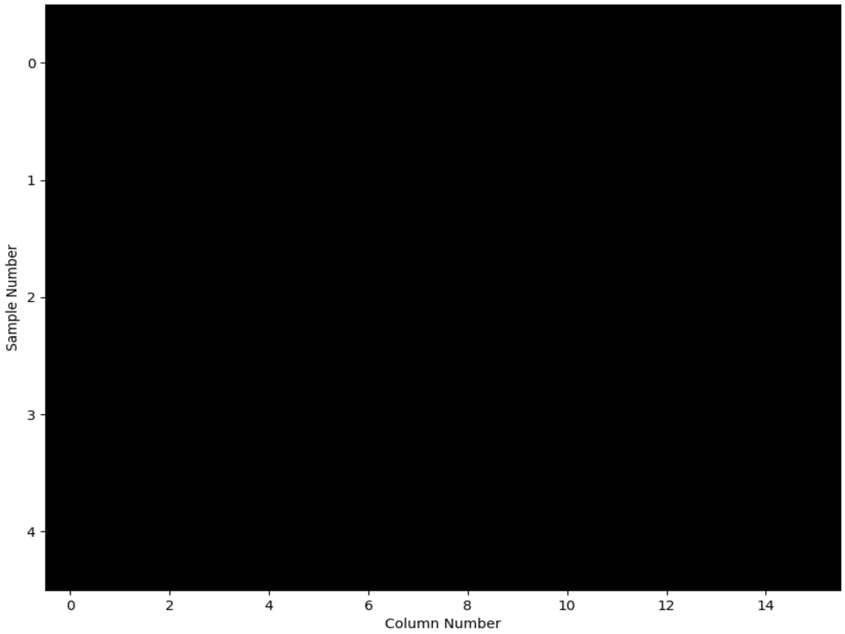
2.1. Duplicates

We do not need to study duplicates in Google Trends data because is research focuses on the frequency of searches related to specific topics. The primary objective is to assess public interest in these subjects over time. Since the data is collected based on specific temporal and geographical criteria, the presence of duplicates would not contribute meaningfully to understanding this interest. In fact, analyzing duplicates could obscure the insights derived from the frequency of searches. Therefore, we can rely on the search counts without the necessity of examining duplicates.

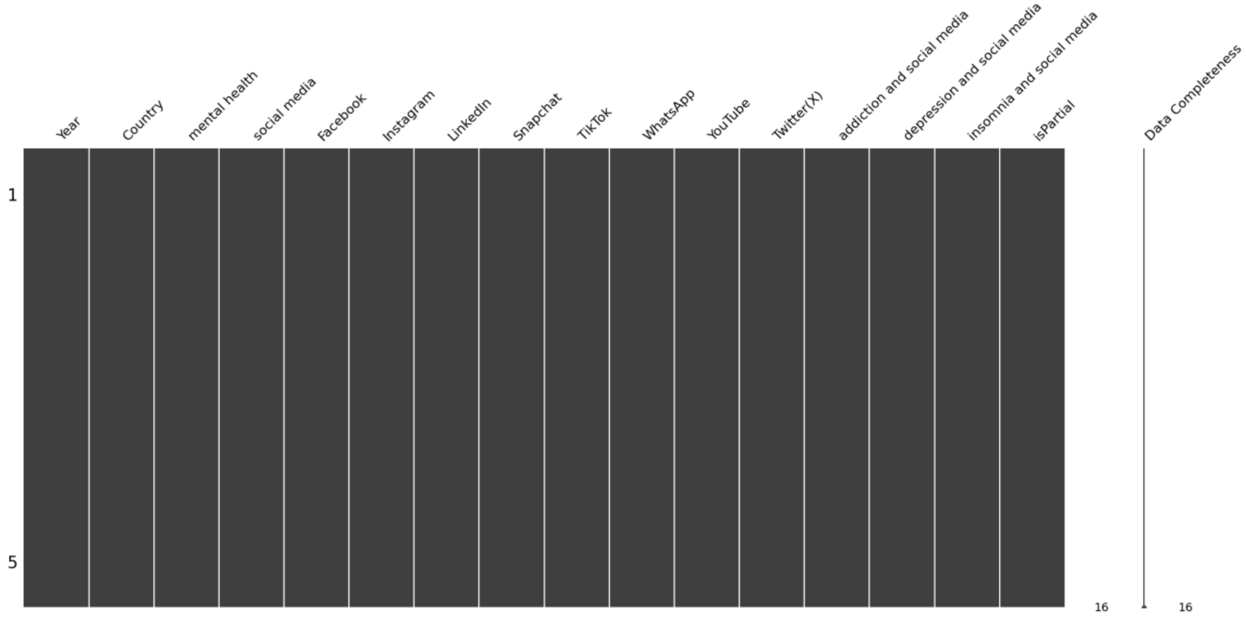
2.2. Missing Values

An important aspect of quality assessment is the examination of missing values within the dataset. While it is typical for datasets to have some incomplete entries, our objective is to pinpoint any considerable voids in the data. Specifically, we aim to detect samples or features that exhibit a substantial number of missing values. This analysis is crucial in understanding the overall completeness of the data retrieved from Google Trends and ensuring the reliability of our findings regarding public interest in the examined topics.

2.2.1. Per sample



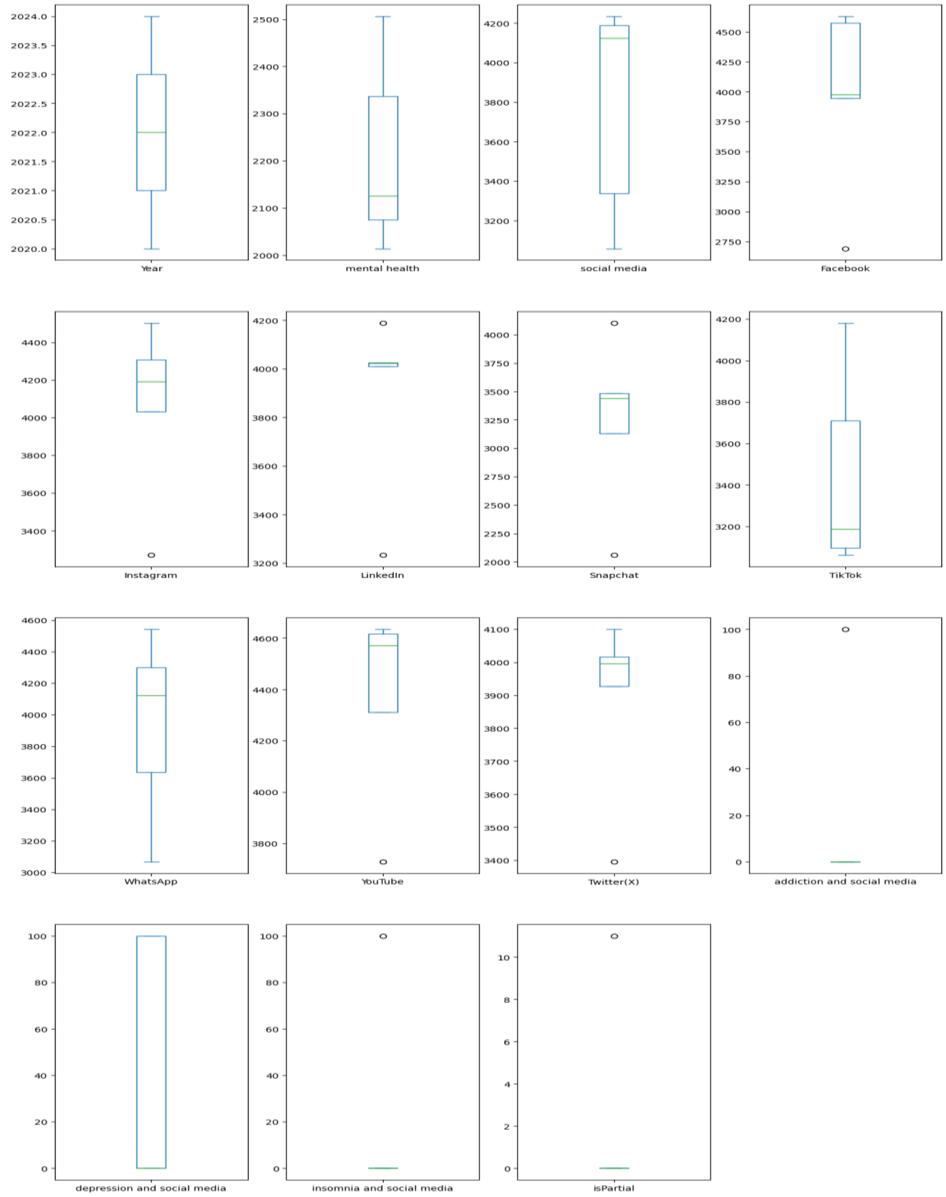
The figure displays each sample (row) of the data against each column (feature) in the dataset. The black color indicates the absence of missing values, while the white color signifies their presence. Since there are no missing values in the data, the entire figure appears in black, indicating that all values are available and valid in the dataset.
To make sure, we reviewed another, more accurate drawing:



The array from the missingno library clearly shows the completeness of the dataset. Since each column is represented without any gaps, we can be sure that the dataset is free of missing values.

2.3.1. Numerical Features

In this section, the numerical features in the dataset are analyzed to identify any unwanted entries or recording errors. The goal is to explore the overall distribution of the numerical data and detect any unusual or outlier values that might affect the final analysis. Box plots are used to understand the range of values, identify outliers, and ensure the quality of the numerical data being used.



The box plot visualizations represent the distribution of the numerical data in your Google Trends dataset across multiple features, with a specific focus on detecting outliers and understanding the spread of values.

- Year: show a uniform spread from 2020 to 2024. There are no visible outliers, indicating that the data spans across these years without anomalies.

- Mental Health & Social Media: The box plot for social media shows a wider range of values compared to mental health, indicating more variability in searches related to social media. Both have no extreme outliers.
- Social Media Platforms (Facebook, Instagram, etc.):
- Facebook and YouTube: Both have outliers, indicating certain periods where search interest was significantly higher or lower than the usual distribution.
- Instagram, LinkedIn, Snapchat: All display outliers, meaning that there were some unusually low search results compared to the overall dataset.
- TikTok and WhatsApp: These two platforms have a broad distribution but no significant outliers.

- Twitter(X): Shows a smaller range of variability with an outlier, suggesting one or two data points significantly deviated from the norm.

- Mental Health & Social Media Topics:

- Addiction, Depression, and Insomnia related to Social Media: All three of these show extremely low values, with a few scattered outliers, which might represent specific peaks in search interest.
- The range is compressed toward the bottom, indicating that these topics generally did not generate much search interest, but with occasional spikes that stand out as outliers.

- IsPartial: Also shows an outlier. This feature might be related to incomplete data retrievals from the Google Trends API, hence the outlier.

Key Takeaways:

- Outliers: Many of the social media platforms (e.g., Instagram, LinkedIn, YouTube) show outliers, which could correspond to events that caused spikes or drops in search interest for these platforms.
- Range: Features like "social media" and "TikTok" show a wide range of values, meaning searches related to them were more variable compared to more focused terms like "mental health" or the specific conditions (addiction, depression).
- Unusual Patterns: For addiction, depression, and insomnia, search interest was mostly low with a few spikes. These may need closer investigation to understand the cause of the outliers.

2.3.2 Show that five-number summary

The five-figure summary provides a valuable overview of the data and reveals important patterns, so we explored it for further understanding and detail. The results are as follows:

Year:

- **Range**: 2020 to 2024.
- **Mean**: 2022, indicating that the data is centered around recent years.
- **Standard Deviation**: Low (1.58), suggesting that the years are close to the mean.

Mental Health:

- **Range**: 2014 to 2506, reflecting a wide range of mental health measures.
- **Mean**: 2211.6, indicating a general upward trend over the years.
- **Standard Deviation**: Moderate (204.59), showing variability in the data.

Social Media:

- **Range**: 3059 to 4234.
- **Mean**: 3788.2, indicating increasing interest in social media over the years.
- **Standard Deviation**: High (548.70), reflecting fluctuating interest levels.

Platform-Specific Data (Facebook, Instagram, LinkedIn, etc.):

- Each platform shows an increasing trend in engagement, with averages being higher than their respective lows.
- **Facebook**: Values range from 2692 to 4629, with a mean of 3963.6.
- **Instagram**: Range from 3271 to 4502, with a mean of 4060.6.
- **LinkedIn**: Shows a lower range but also demonstrates an increasing trend over the years.

Depression, Insomnia, Addiction, and Social Media:

- **Range**: 0 to 100, with a median of 20 for addiction and 40 for depression.
- The majority of mental health outcomes (depression, insomnia, anxiety) show zero values, with one or two values reaching 100, indicating data outages. This suggests that a significant number of respondents did not report issues related to depression linked to social media.

IsPartial:

- **Range**: 0 to 11, with a median of 2.2.
- This indicates the dataset contains incomplete data, particularly for the year 2024.

Key Insights and Relationships

- **General Trends**: There has been a general increase in mental health metrics and social media use over the years, suggesting a relationship between increased social media use and mental health metrics.
- **Addiction and Depression for Insomnia and Social Media**: Low medians (20 for addiction and 40 for depression) combined with high extremes (100) indicate significant variability. Most years recorded zero values, highlighting challenges in identifying these issues.

2.3.3 Handling outliers

Based on the previous results, it becomes clear to us that there are extreme values that we must take action towards, so we have further identified them and then analyzed and processed them as follows:

Columns with Long Phrases:

Columns such as 'addiction and social media', 'insomnia and social media', and 'depression and social media' exhibit different behavior compared to other columns. Most of these columns show few or no outliers, suggesting that the API may struggle to handle long phrases effectively or process them accurately.

Columns with Outliers:

Columns related to platforms like 'Facebook', 'Instagram', and 'YouTube' demonstrate significant variations in interest over time. These fluctuations are valuable as they indicate shifts in user behavior and reflect the impact of major events or updates on these platforms.

Column isPartial:

The outlier in this column pertains to the year 2024, indicating that the data for this period is incomplete.

Based on these findings, we took the following actions:

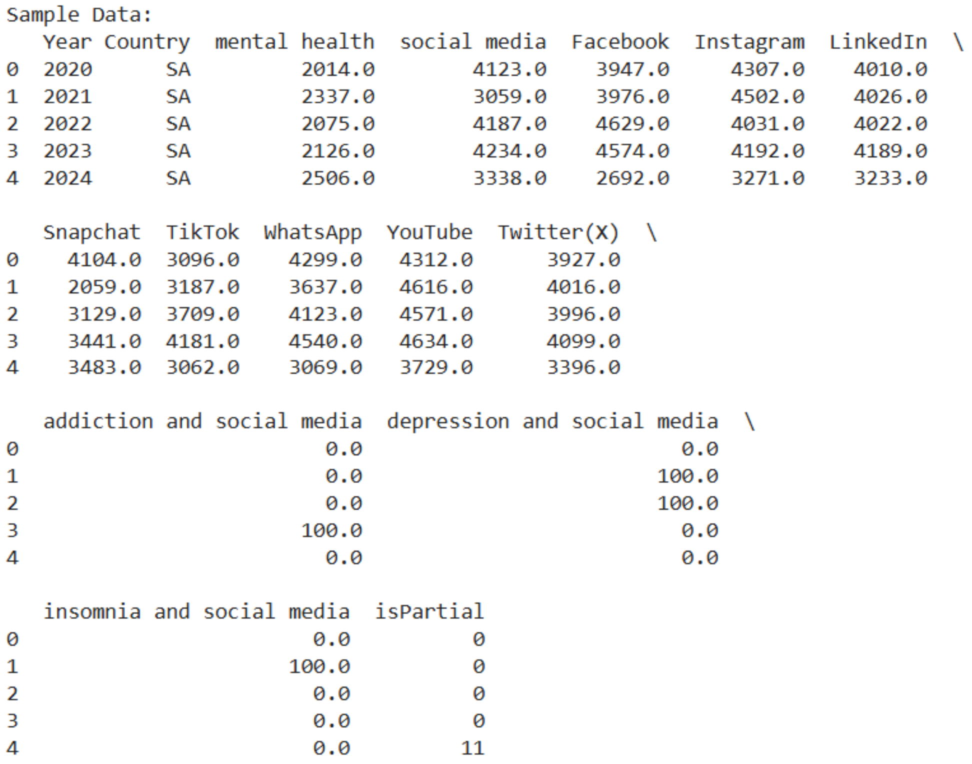
- **Remove columns with long phrases:** We removed columns such as “Addiction and Social Media” and “Insomnia and Social Media,” as they did not provide clear value for analysis due to API limitations in processing them.

- **Retain columns with useful outliers**: We retained columns with significant outliers, especially those related to social media platforms (such as Facebook and Instagram), as they can provide valuable insights into user behavior.

- **Title Partial as a special case:** We realized that the outlier in 2024 was the result of incomplete data, so we decided to leave it out and make sure to account for it during the final analysis.

Quick look at data before and after cleaning:

Before Cleaning:



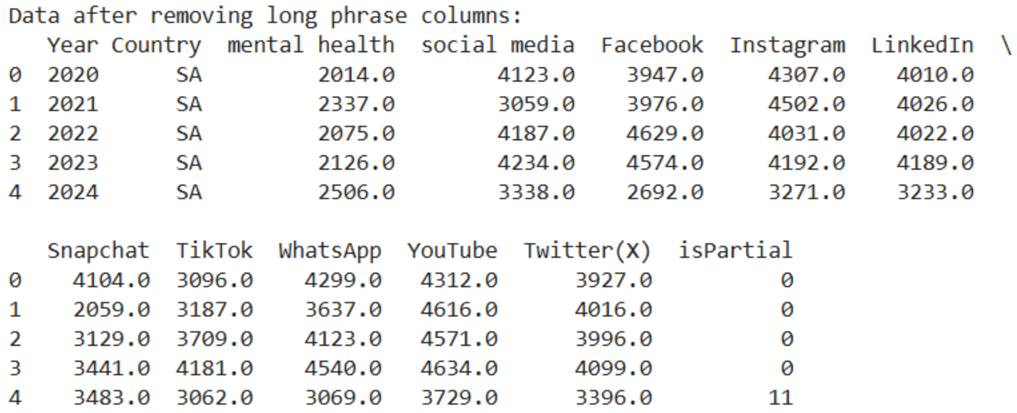
```
Sample Data:
   Year Country  mental health  social media  Facebook  Instagram  LinkedIn  \
0  2020      SA         2014.0        4123.0    3947.0     4307.0    4010.0
1  2021      SA         2337.0        3059.0    3976.0     4502.0    4026.0
2  2022      SA         2075.0        4187.0    4629.0     4031.0    4022.0
3  2023      SA         2126.0        4234.0    4574.0     4192.0    4189.0
4  2024      SA         2506.0        3338.0    2692.0     3271.0    3233.0

   Snapchat  TikTok  WhatsApp  YouTube  Twitter(X)  \
0    4104.0  3096.0    4299.0   4312.0      3927.0
1    2059.0  3187.0    3637.0   4616.0      4016.0
2    3129.0  3709.0    4123.0   4571.0      3996.0
3    3441.0  4181.0    4540.0   4634.0      4099.0
4    3483.0  3062.0    3069.0   3729.0      3396.0

   addiction and social media  depression and social media  \
0                         0.0                          0.0
1                         0.0                        100.0
2                         0.0                        100.0
3                       100.0                          0.0
4                         0.0                          0.0

   insomnia and social media  isPartial
0                        0.0          0
1                      100.0          0
2                        0.0          0
3                        0.0          0
4                        0.0         11
```

After Cleaning:



```
Data after removing long phrase columns:
   Year Country  mental health  social media  Facebook  Instagram  LinkedIn  \
0  2020      SA         2014.0        4123.0    3947.0     4307.0    4010.0
1  2021      SA         2337.0        3059.0    3976.0     4502.0    4026.0
2  2022      SA         2075.0        4187.0    4629.0     4031.0    4022.0
3  2023      SA         2126.0        4234.0    4574.0     4192.0    4189.0
4  2024      SA         2506.0        3338.0    2692.0     3271.0    3233.0

   Snapchat  TikTok  WhatsApp  YouTube  Twitter(X)  isPartial
0    4104.0  3096.0    4299.0   4312.0      3927.0          0
1    2059.0  3187.0    3637.0   4616.0      4016.0          0
2    3129.0  3709.0    4123.0   4571.0      3996.0          0
3    3441.0  4181.0    4540.0   4634.0      4099.0          0
4    3483.0  3062.0    3069.0   3729.0      3396.0         11
```

3. Content Investigation

Up until now we only looked at the general structure and quality of the dataset. Let's now go a step further and take a look at the actual content. an investigation would be done feature by feature.

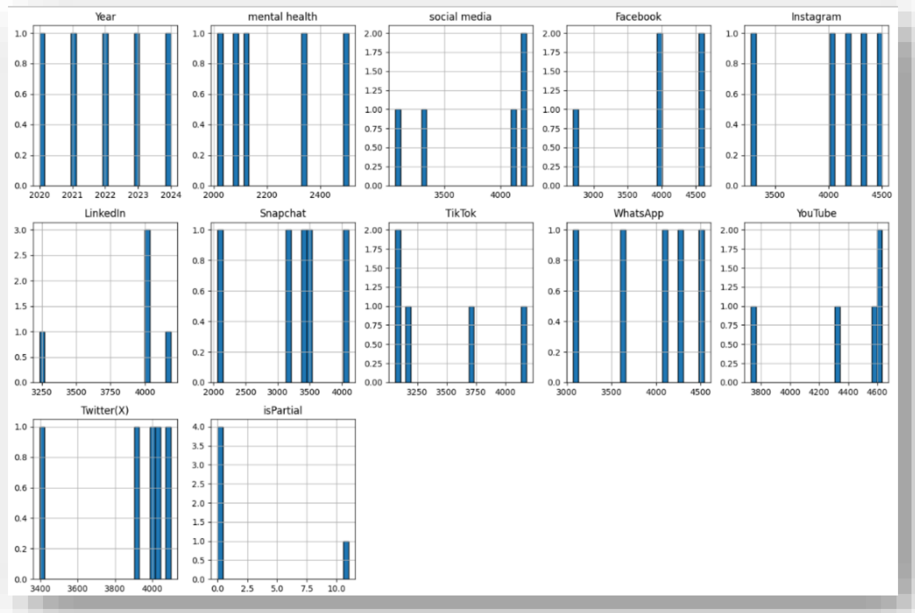
Content investigation is a vital step in data analysis, as it allows us to understand the feature distributions within our dataset and how these features interact with one another. This understanding leads to deeper insights into the dataset, making the results more interpretable and meaningful.

we will explore three different approaches which are Feature distribution, Feature patterns and Feature relationships that can give us a very quick overview of the content stored in each feature and how they relate.

3.1 Feature distribution

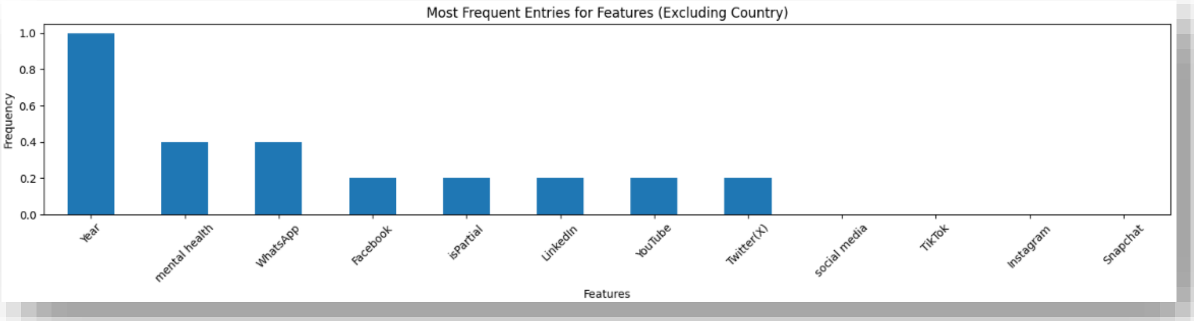
To gain a better understanding of our dataset and to provide useful insights for data cleaning and feature transformation, we first examined the value distribution of each feature using histograms. Next, we calculated the mode for each feature (excluding the 'Country' column) to identify the most common values. A bar plot was then used to represent the top features with the highest frequency of singular value content.

Libraries Used:
matplotlib.pyplot: Used for visualizing data in the form of histograms and bar charts
pandas: Used for data manipulation, handling, and computing the mode (most frequent values) for the dataset.



Looking at the charts presented, we can find the follows:

- **Year**: Data is evenly distributed across 2020 to 2024, indicating consistent collection.
- **Mental Health**: The distribution is consistent, showing stable interest or research over the years.
- **Social media**: Displays variation, with ups and downs possibly tied to specific events or social behavior changes.
- **Facebook**: Shows stability with some data clustering, suggesting relatively steady usage with occasional fluctuations.
- **Instagram**: The data is concentrated around a specific range, indicating steady interest with minor usage changes.
- **LinkedIn**: Greater variability suggests fluctuating interest, potentially influenced by economic or platform changes.
- **Snapchat**: Exhibits a wide distribution, hinting at varying levels of interest and usage over time.
- **TikTok**: Relatively stable distribution with slight growth, suggesting consistent interest over the period.
- **WhatsApp**: Shows stability with narrow value clusters, indicating generally steady interest or usage.
- **YouTube**: Greater variation in values suggests fluctuating interest, possibly linked to trends or events.
- **Twitter(X):** Displays significant variability, likely due to impactful political or social events.
- **isPartial**: Indicates incomplete data for 2024, showing a jump in values due to ongoing data collection.



Looking at the chart presented, we can find the follows:

- **Year:** Consistency is observed, with all entries reflecting the same value, as expected for the dataset covering the years 2020 to 2024.

- **Mental Health & WhatsApp**: Both show a frequency of 40%, indicating some commonality in data, but with noticeable variability in interest.

- **Facebook, LinkedIn, YouTube, Twitter (X):** Each has a 20% frequency for the most common value, highlighting fluctuations in user interest over the years.

- **Social Media, Instagram, Snapchat, TikTok**: These features have diverse values with no dominant trend, reflecting varying levels of interest.

3.2 Feature distribution

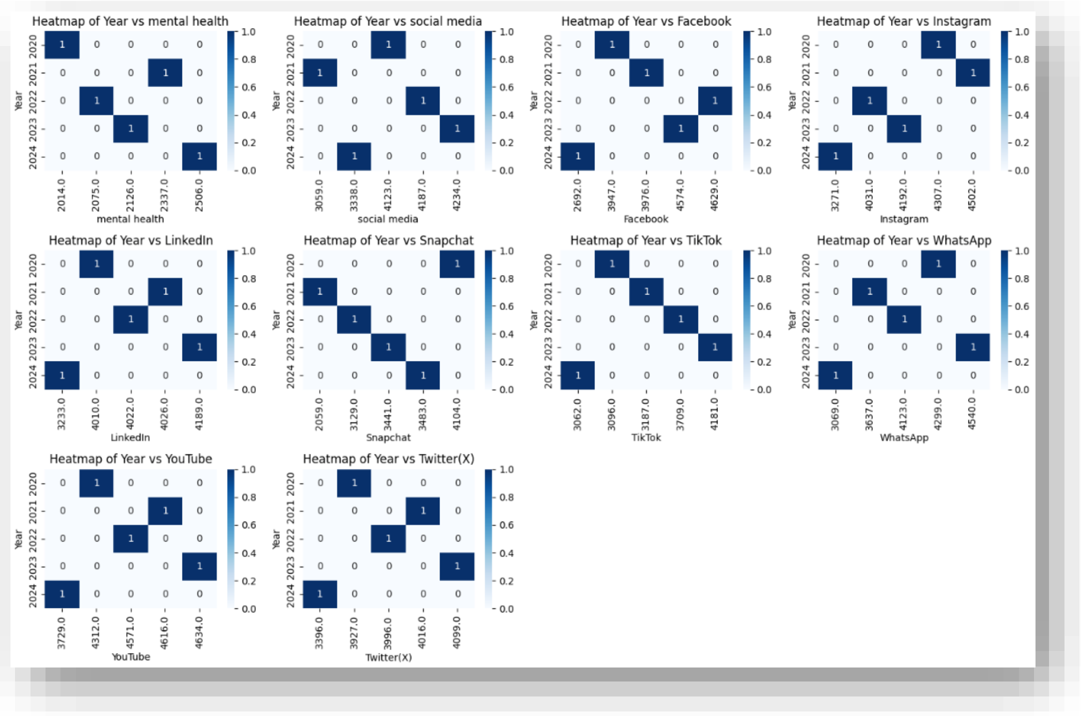
We conducted this analysis to investigate feature-specific patterns and identify particular relationships between features, which will enhance our understanding of the dataset.

Initially, we created heatmaps to visualize the relationship between each feature and the year, allowing us to identify trends, explore the strength and nature of these relationships, and facilitate comparative analysis.

I would like to point out that we have excluded the “isPartial” column because it is only to clarify that the 2024 data is incomplete since it has not been completed yet, so it will not help us in our current goal.

Libraries Used:

- **Pandas**: This library was used for data manipulation, specifically to create cross-tabulations that summarize the frequency of occurrences between the years and the different social media topics.
- **Seaborn**: A data visualization library built on top of Matplotlib, Seaborn was employed to create the heatmaps.
- **Matplotlib**: This library served as the foundational plotting library for Python, utilized for creating the figure and subplots that house the heatmaps.

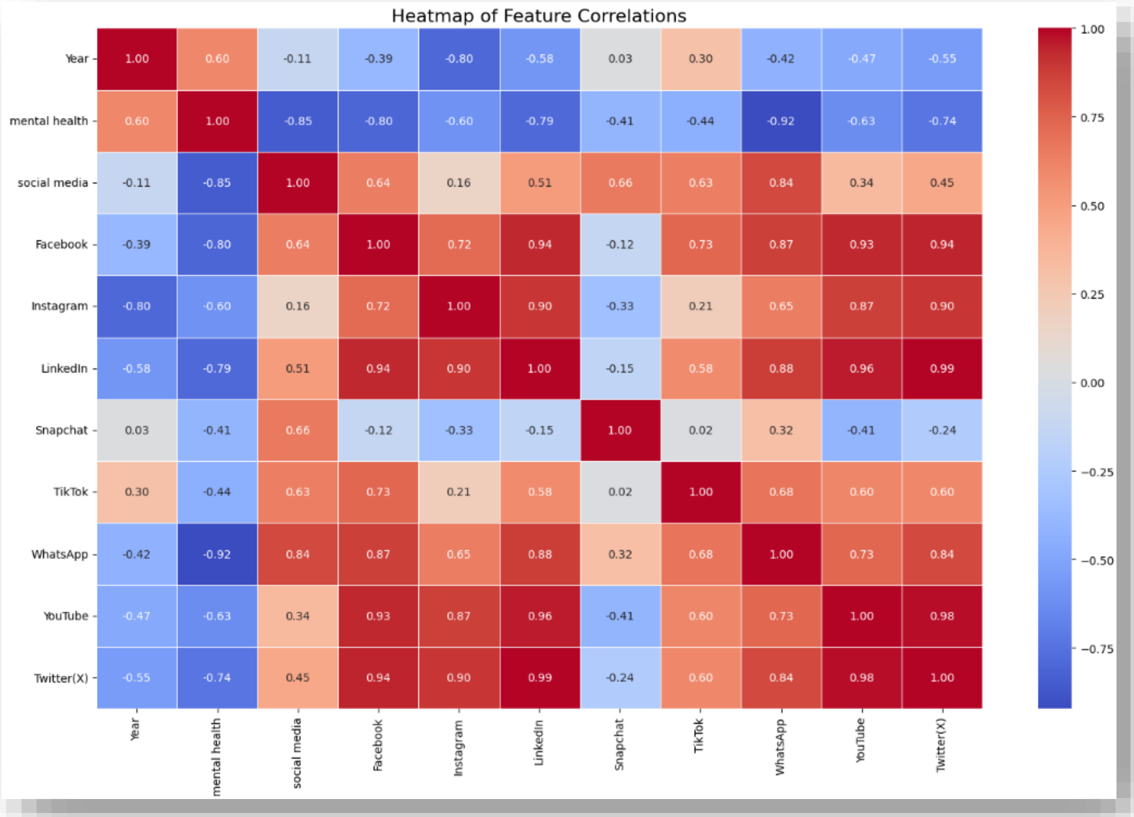


Looking at the chart presented, we can find the following:

- There is an almost positive association between the year and mental health as the rate of searching for mental health increases over time and this indicates a growing awareness of mental issues in recent years.
- Searches for Facebook rose in earlier years but declined in more recent years.
- Searches for Instagram showed a downward trend as the years progressed.
- Snapchat searches dropped in 2021 but have surged significantly in recent years.
- TikTok searches increased from 2020 to 2023, followed by a decline in the last year.
- We found no consistent relationship between the frequency of searches for "social media," "LinkedIn," "WhatsApp," "YouTube," and "Twitter" over the years.

After visualizing the relationships between each feature and the year, we then examined the relationships between the features themselves. We first verified that the type of API_data that do not contain country column is discrete to choose the appropriate chart for us and we found that all of them are discrete and do not contain any continuous data, so we used the heatmap as a suitable option for discrete type data.

I should also note here that we excluded the “isPartial” column because it is just to clarify that the 2024 data is incomplete as it is not complete yet, so it will not help us in our current goal.



Looking at the chart presented, we can find the following:

- There is a strong correlation between Twitter, LinkedIn, Facebook, Instagram, YouTube and WhatsApp, this could indicate that users tend to engage with multiple platforms simultaneously
- There is a negative correlation between mental health and social media and all its platforms.
- There is a weak correlation between social media and all platforms except WhatsApp, which is considered a strong correlation

3.3 Feature relationships

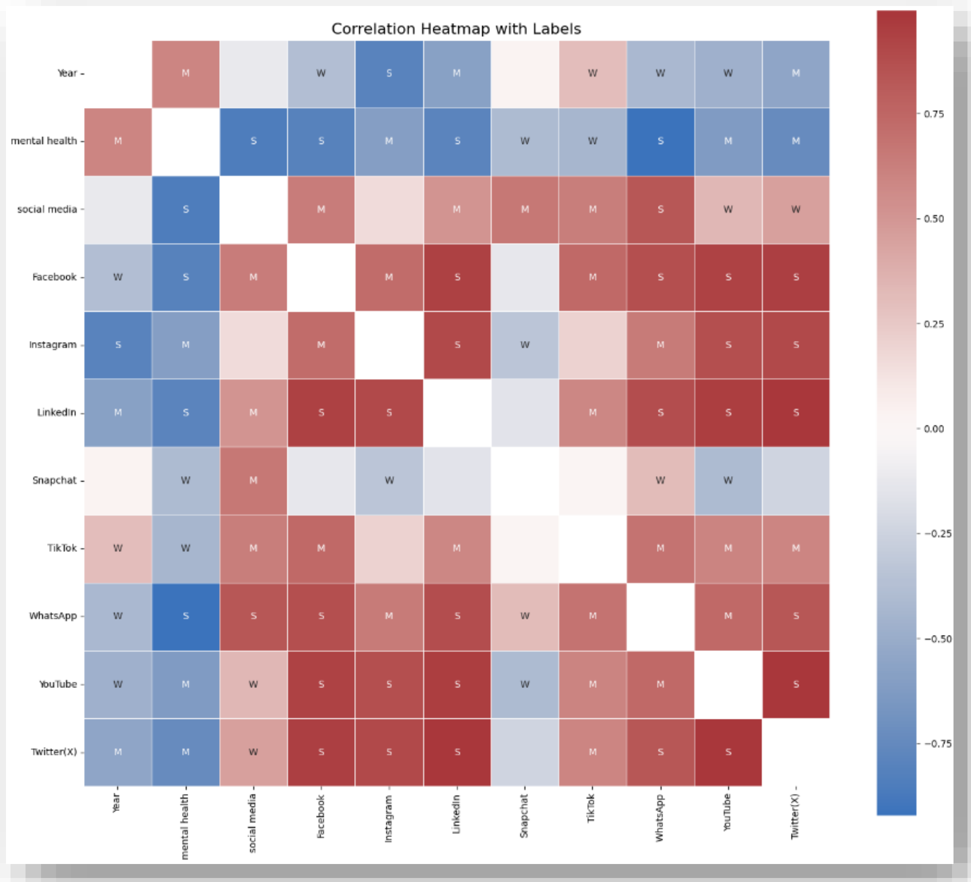
Investigating how features are related to each other helps us to understand their associations or dependencies, and understanding this is crucial when building machine learning models, as highly correlated features can lead to multicollinearity, which affects the accuracy of the model. It also helps in feature selection by identifying the features that contribute the most to the desired outcome.

Here we also used a heat map as we did in section 3.2 but with a slight difference, where we used specific labels to indicate the strength of the correlation where S stands for strong correlation (absolute value greater than 0.75), M for medium correlation (absolute value between 0.5 and 0.75) ,W for weak correlation (absolute value between 0.25 and 0.5) and An empty string for correlations below 0.25, which are not labeled.

The column and index named isPartial are removed because it does not provide meaningful information for the correlation analysis.

Libraries Used:

- **NumPy:** Used to handle numerical operations and create labels for the heatmap.
- **Matplotlib**: Used to create the figure and layout for the heatmap visualization.
- **Seaborn**: A data visualization library based on Matplotlib, used here to plot the heatmap and make it more visually appealing.



This heatmap is similar to the heatmap mentioned in section 3.2.2 and the results are the same, but the difference here is that this heatmap is coded with letters that indicate the correlation between the features to facilitate understanding and see the relationships more clearly

We also ordered the correlations between different features to make it easier to identify significant relationships. By displaying correlations from the most negative to the most positive, the approach highlights which pairs of features have the strongest interactions, both inverse and direct. This organization helps prioritize analysis, and discover patterns.

we found this order with this results :

- **WhatsApp vs. Mental Health (-0.921408):** there is a significant inverse relationship between WhatsApp and mental health, indicating that increased searches for WhatsApp are associated with a decline in mental health-related searches.
- **Twitter(X) vs. LinkedIn (0.994743), Twitter(X) vs. YouTube (0.981474) and LinkedIn vs. Facebook (0.937804):** Features like Twitter(X) and LinkedIn, as well as Twitter(X) and YouTube, exhibit very strong positive correlations, meaning that searches for these platforms often increase or decrease together. Additionally, LinkedIn and Facebook also show a high positive correlation, suggesting that their search trends are closely linked.

3.4 Conclusion of Interesting findings:

- There is an almost positive association between the year and mental health as the rate of searching for mental health increases over time and this indicates a growing awareness of mental issues in recent years.
- Searches for Facebook and Instagram showed a downward trend as the years progressed.
- Snapchat searches dropped in 2021 but have surged significantly in recent years.
- TikTok searches increased from 2020 to 2023, followed by a decline this year, possibly because this year's data is incomplete.
- There is a strong correlation between Twitter, LinkedIn, Facebook, Instagram, YouTube and WhatsApp, this could indicate that users tend to engage with multiple platforms simultaneously.
- There is a negative correlation between mental health and social media and all its platforms, especially Facebook, LinkedIn and WhatsApp.
- Twitter and LinkedIn, as well as Twitter and YouTube, exhibit very strong positive correlations, meaning that searches for these platforms often increase or decrease together. Additionally, LinkedIn and Facebook also show a high positive correlation, suggesting that their search trends are closely linked.

4. Data Biases

- **Selection Bias:**

  The data collected from Google Trends reflects only the search behavior of users who use Google. This introduces a bias, as it does not represent the entire population or those who use other search engines or platforms. The trends may not accurately reflect the behavior of individuals who directly engage with social media without searching for related topics.

- **API Limitations on Phrase Length:**

  Long phrases like "addiction and social media" or "insomnia and social media" may not be processed accurately by Google Trends API. This presents a bias, as shorter and more commonly searched terms are better represented, which may skew the analysis towards those terms.

- **Incomplete Data:**

  The presence of incomplete data, especially for the year 2024, introduces a bias towards previous years that have more complete data. This may affect the analysis of recent trends.

- **Social Media Behavior vs. Search Behavior:**

  There is an inherent bias in relying on search behavior to infer social media use and trends related to mental health. People may not always search for what they are experiencing (such as mental health issues), leading to underrepresentation or distortion of the actual impact of social media on mental health.

- **Decision made:**

1. **Handling Incomplete Data (2024)**: We addressed the incomplete data for the year 2024 by ensuring that our analysis took into consideration the `isPartial` column. This column indicated that data for 2024 was not fully available, and we handled it accordingly to avoid skewing the final results.
2. **Random Wait Times**: To avoid exceeding Google Trends’ rate limits and encountering API errors, we implemented random wait times between data retrieval requests. This ensured smoother data collection and helped us avoid disruptions in the process.
3. **Keyword Grouping**: To optimize data retrieval, we divided the keywords into four distinct groups. This was necessary to ensure the API could handle the requests efficiently and avoid errors.
4. **Yearly Timeframes**: Since Google Trends only allows data retrieval based on specific date ranges (down to months and days), we manually defined the yearly ranges (e.g., '2020-01-01 to 2020-12-31') to capture trends over entire years. This allowed us to compare year-on-year trends for each keyword.
5. **Data Merging and Cleanup**:
   - **Combining Columns**: We merged the Twitter and Platform X columns into a single column named `Twitter(X)` since they represent the same platform. This step was crucial to avoid redundancy and streamline our analysis.
   - **Handling Outliers**: During data cleaning, we identified outliers in several columns, which indicated significant deviations in search interest. We decided to retain these outliers in our dataset as they provided insights into unusual spikes or drops in interest related to social media and mental health. Furthermore, we decided to initially remove the keywords with long phrases. This decision was made because Google Trends struggled to process these keywords accurately, leading to zero search interest for most years. This action aimed to improve the overall quality and reliability of our data analysis.
6. **Language Restriction**: Due to Google Trends' inability to effectively recognize Arabic keywords, we decided to limit our analysis to English keywords only. This decision ensured that we could gather more accurate and reliable data on public interest in social media and mental health topics.
7. **Column Reordering**: For better organization and clarity, we reordered the columns in our final dataset to prioritize key variables.

- ## Challenges:

1. **Rate-Limiting Errors in Google Trends API:**
   - **Description:** While trying to extract large amounts of data from Google Trends API, I encountered rate-limiting issues. These errors occur when the API restricts the number of allowed queries in a short period of time.
   - **How it was handled:**
     - I implemented a retry mechanism with random time intervals between queries to avoid exceeding the rate limit. This involved inserting a random sleep period after each set of queries to prevent the system from detecting rapid consecutive requests.
     - Additionally, I divided the time periods for each query logically, ensuring that the system wasn't overloaded with too many requests at once.
2. **Splitting Keywords into Groups:**
   - **Challenge:** Google Trends cannot process a large number of keywords at once, and handling this large volume of data in a single request could lead to system overload, impacting speed and efficiency.
   - **Solution:** I split the keywords into smaller groups to reduce the load on the API. This improved system response time and reduced the likelihood of errors. Grouping keywords allowed for smoother data collection with appropriate wait times, which helped manage API usage more efficiently.
3. **Choosing the Right Keywords:**
   - **Description:** One of the main challenges was identifying the most representative and relevant keywords related to the research topic (social media and mental health) within the Saudi context. General keywords may not always reflect the specific interests of the local population.
   - **How it was handled:**
     - I conducted preliminary research to choose keywords that reflect general public interests, including major social media platforms (Facebook, Instagram, TikTok), along with specific terms like "mental health," "anxiety," and "addiction."
4. **Handling the Rebranding of Twitter to X:**
   - **Description:** I encountered a challenge in handling the rebranding of Twitter to X in 2023, as it required merging data related to both names.
   - **How it was handled:**
     - To address this, I merged the data associated with both "Twitter" and "X" into a single column labeled `Twitter(X)`, ensuring that no data was duplicated or lost due to the rebranding. This way, all relevant insights remained consistent across the years.