Twitter Data

# **Introduction**

The purpose of this report is to analyze and decode a decade's worth of social media data from one of Playhouse Communication's high-profile clients. The data, which spans across four major social media platforms - Facebook, Instagram, LinkedIn, and Twitter, provides a unique opportunity to gain deep insights into the client's digital engagement patterns.

Playhouse Communication, a leading digital marketing agency in Nigeria, has been managing the client's social media platforms for the past 10 years. This report aims to transform the raw data from these platforms into actionable insights that could redefine the future of digital marketing for the client.

The data has been provided in both .csv and .xls formats, allowing for comprehensive analysis using various data processing and analytics tools. The goal is not just to understand past trends but also to predict future ones, thereby providing the client with a strategic advantage in their digital marketing efforts.

This isn't just a hackathon; it's an opportunity to shape the future of digital engagement. By leveraging this exclusive social media data, we aim to strike gold with game-changing insights that could potentially revolutionise digital marketing strategies for Playhouse Communication and its clients.

# **Methodology**

The methodology for analysing the social media data involves several steps:

1. **Data Cleaning**: The first step in the process is data cleaning. This involves handling missing values, removing unnecessary columns, and converting data types where necessary. For instance, any columns with more than 50% missing values are dropped from the data.

2 **Exploratory Data Analysis (EDA):** This step involves understanding the distribution of data and relationships between different variables. For example, using time\_series\_analysis to plot various metrics over time.

3. **Topic Modeling using Latent Dirichlet Allocation (LDA)**: LDA is used to discover the main topics that occur in the collection of social media posts. This unsupervised machine learning technique assumes that each document (or post) is a mix of various topics, and each topic is a collection of words. By looking at which words often appear together, LDA can figure out the possible topics and how much each document belongs to each topic.

4. **Named Entity Recognition (NER) and Part-of-Speech Tagging (POS)**: These Natural Language Processing (NLP) techniques are used to extract more information from the text data. NER can tell us which specific entities (like people, places, or organisations) are being mentioned in the posts, while POS tagging helps us understand the context better by knowing the grammatical role of each word.

5. **Data Visualization**: The last step involves visualising the results of our analysis. This could be as simple as bar plots or pie charts showing the distribution of posts across different topics, or more complex visualisations like word clouds that highlight the most frequently mentioned entities in each topic.

The tools used for this analysis include Python for data cleaning and analysis, various libraries like pandas for data manipulation, matplotlib and plotly for data visualization, and sklearn and nltk for machine learning and NLP tasks.

# **Findings**

x\_data.info()  
  
<class 'pandas.core.frame.DataFrame'>

RangeIndex: 8529 entries, 0 to 8528

Columns: 147 entries, Date to Tags

dtypes: float64(119), int64(1), object(27)

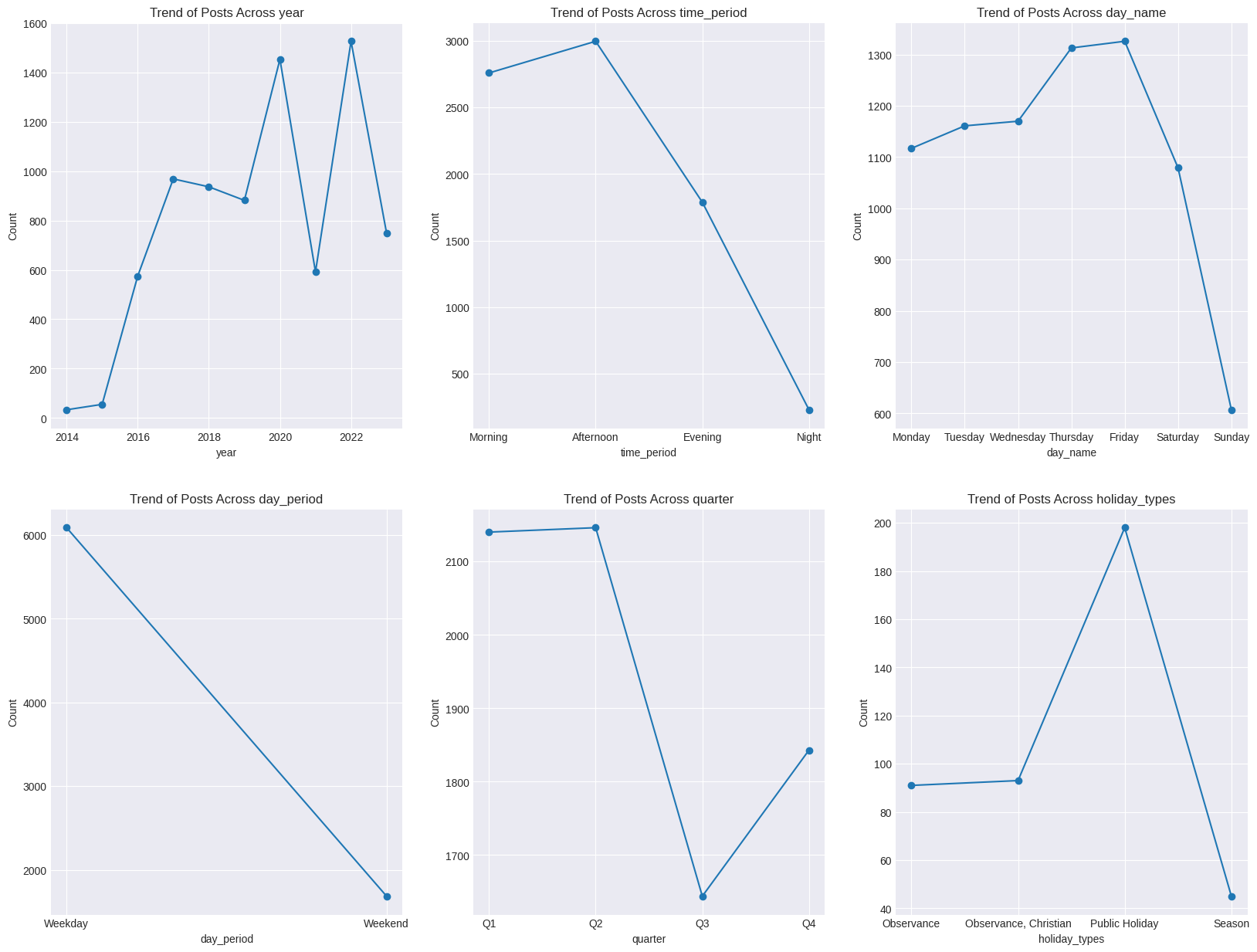
memory usage: 9.6+ MB

Impressions: This is the number of times that a tweet has been seen on Twitter. Impressions are non-unique, so if one user sees the same tweet three times, that tweet will have three impressions.

Organic Impressions: These are impressions that come from unpaid distribution. In other words, these are the views your tweet gets from your followers and people who find your tweet through Twitter’s algorithms, hashtags, or other organic means.

Potential Reach: This is an estimate of the total number of people who could potentially see your tweet. It’s calculated based on the number of followers you have plus the followers of anyone who retweets your post. It’s important to note that this is just a potential number and doesn’t guarantee that all these people have seen your tweet.

Count of Posts against each feature



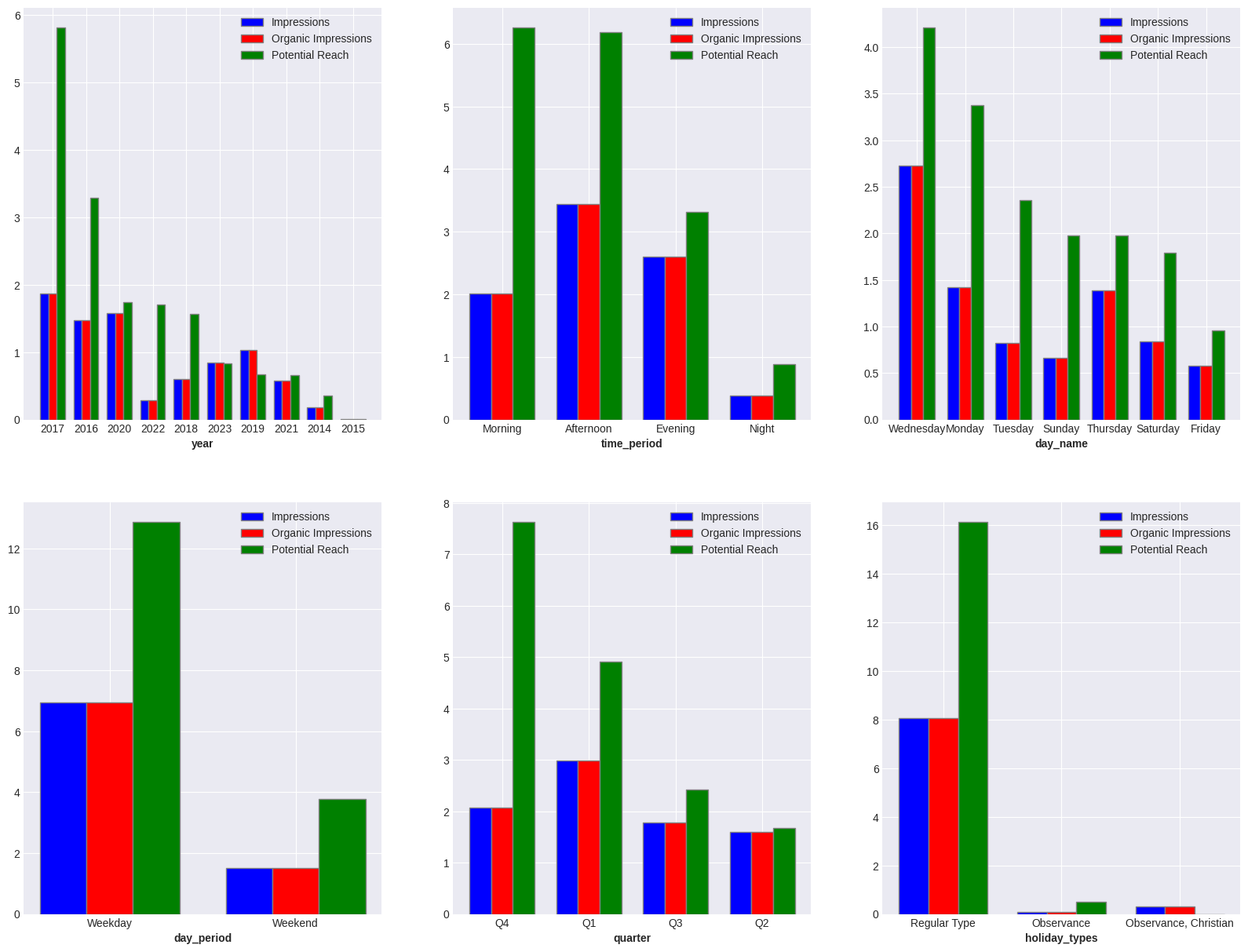
cleaned\_impressions\_reach[impression\_reach\_cols].describe().T

cleaned\_impressions\_reach[impression\_reach\_cols].describe().T

|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Impressions** | 7773.0 | 3216.237489 | 3339.175577 | 0.0 | 1308.0 | 2462.0 | 3825.0 | 77765.0 |
| **Organic Impressions** | 7773.0 | 3216.237489 | 3339.175577 | 0.0 | 1308.0 | 2462.0 | 3825.0 | 77765.0 |
| **Potential Reach** | 7773.0 | 284877.816673 | 106183.426405 | 0.0 | 226192.0 | 305303.0 | 322187.0 | 2194275.0 |

It seems like we have a lot of Outliers in our Dataset, let's see what is the number of outliers we have  
  
last\_quantile = cleaned\_impressions\_reach['Potential Reach'].quantile(0.99)

df = cleaned\_impressions\_reach[cleaned\_impressions\_reach['Potential Reach'] > last\_quantile]

  
**Yearly Performance**: The year 2017 was the most successful in terms of impressions, organic impressions, and potential reach. However, there has been a decline in these metrics in the following years. This could indicate a need to revisit the content strategy or audience engagement tactics used in 2017.

**Potential vs. Actual Reach**: While the potential reach is high for certain years like 2018 and 2022, the actual impressions (both total and organic) are not as high. This suggests that while the tweets have the potential to reach a large audience, they may not be getting seen or interacted with as much as they could be.

**Recent Performance**: The years 2022 and 2023 have seen lower numbers compared to previous years. This could be due to various factors such as changes in Twitter’s algorithm, shifts in follower behavior, or changes in the content being posted.

**Time of Day Trends**: The afternoon time period has the highest number of impressions and organic impressions, followed by evening and morning. However, the morning time period has the highest potential reach. This suggests that while more tweets are seen and interacted with in the afternoon, the tweets posted in the morning have a wider audience.

**Day of the Week Performance**: The highest number of impressions and organic impressions occur on Wednesdays, followed by Thursdays and Mondays. This suggests that these might be the best days to post for maximum visibility.

**Potential Reach**: Despite having lower impressions and organic impressions, Saturdays rank third in terms of potential reach. This indicates that tweets on Saturdays have the potential to reach a larger audience, even if they aren’t currently doing so.

**Underperforming Days**: Fridays and Sundays have the lowest impressions, organic impressions, and potential reach. This could suggest that these days are less active for your Twitter audience.

**Quarterly Performance**: The first quarter (Q1) of the year has the highest number of impressions and organic impressions. This suggests that this might be the best period to post for maximum visibility.

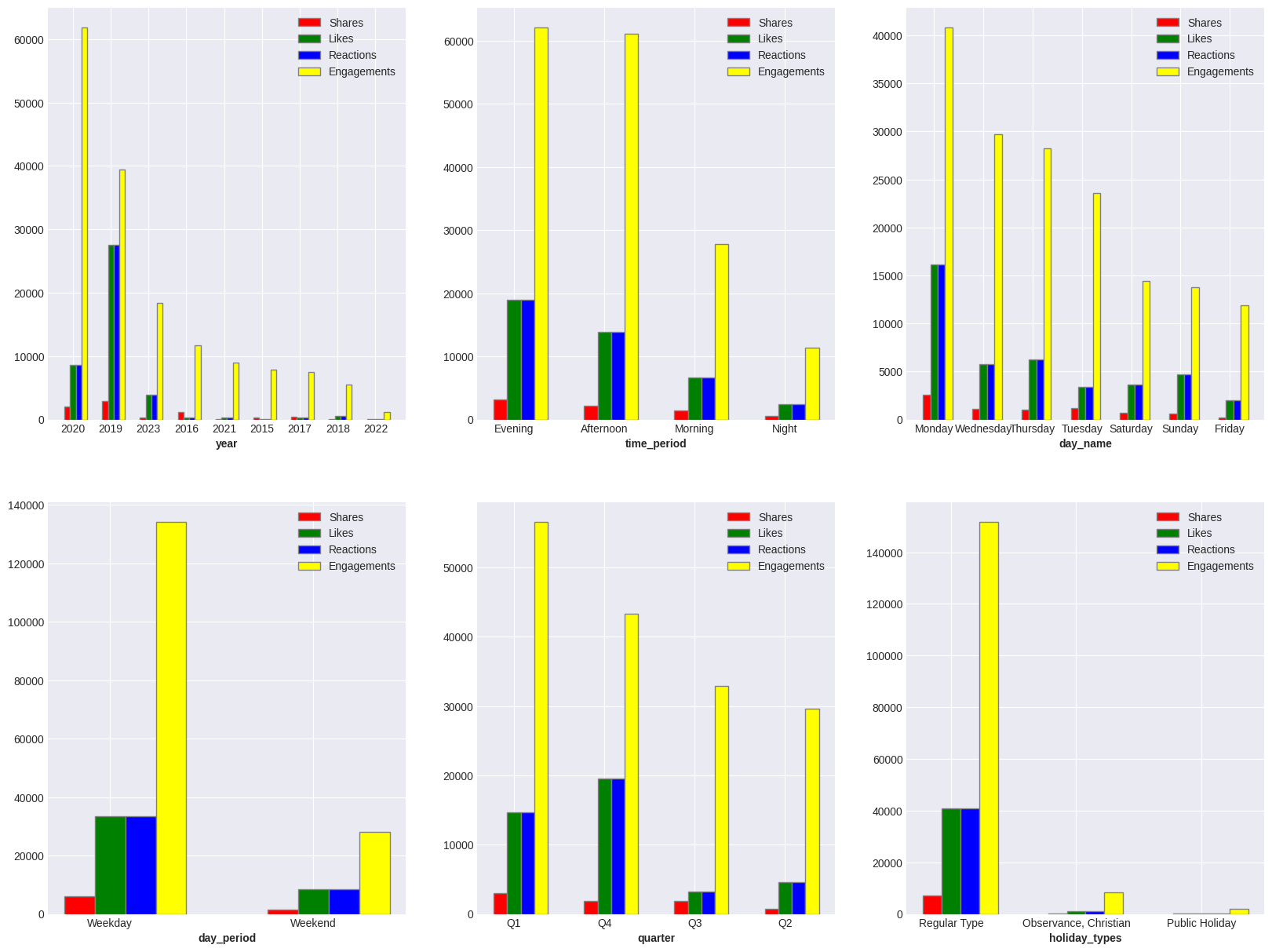
**Potential Reach**: Despite having fewer impressions and organic impressions, the fourth quarter (Q4) has the highest potential reach. This indicates that tweets in Q4 have the potential to reach a larger audience, even if they aren’t currently doing so.

**Underperforming Quarters**: The second (Q2) and third quarters (Q3) have the lowest impressions, organic impressions, and potential reach. This could suggest that these periods are less active for your Twitter audience.

|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Engagements** | 7773.0 | 111.031777 | 306.665224 | 0.0 | 29.0 | 57.0 | 109.0 | 14972.0 |
| **Reactions** | 7773.0 | 20.387881 | 171.944868 | 0.0 | 5.0 | 10.0 | 17.0 | 13528.0 |
| **Likes** | 7773.0 | 20.387881 | 171.944868 | 0.0 | 5.0 | 10.0 | 17.0 | 13528.0 |
| **Shares** | 7773.0 | 8.390969 | 21.688060 | 0.0 | 2.0 | 4.0 | 10.0 | 1135.0 |

last\_quantile = cleaned\_impressions\_reach['Engagements'].quantile(0.99)

df = cleaned\_impressions\_reach[cleaned\_impressions\_reach['Engagements'] > last\_quantile]

  
  
**Yearly Engagement**: The year 2020 had the highest number of engagements, reactions, likes, and shares, followed by 2019 and 2023. This suggests that these years had the most interactive and engaging content.

**Engagement vs Reach**: While the year 2017 had the highest potential reach and impressions, it ranks lower in terms of engagements (likes, shares, reactions). This could indicate that while the content posted in 2017 reached a large audience, it did not engage the audience as effectively as the content posted in other years.

**Recent Performance**: The years 2022 and 2023 have seen lower numbers compared to previous years. This could be due to various factors such as changes in Twitter’s algorithm, shifts in follower behavior, or changes in the content being posted.

**Engagement vs Reach by Time of Day**: The evening and afternoon periods have the highest engagements (likes, shares, reactions), while the morning period has the highest potential reach. This suggests that while more tweets are interacted with in the evening and afternoon, the tweets posted in the morning have a wider audience.

**Underperforming Time Period**: The night period has the lowest engagements and potential reach. This could suggest that this time is less active for your Twitter audience.

Comparing these insights with the top 1% posts with high impressions and potential reach, we can see some differences:

* While the morning period has the highest potential reach, it ranks third in terms of engagements. This could indicate that while the content posted in the morning reaches a large audience, it does not engage the audience as effectively as content posted in other time periods.
* Conversely, while the evening and afternoon periods have high engagement rates, they do not have as high a potential reach as the morning period. This suggests that while these time periods are good for engaging with your current audience, they might not be as effective for reaching new audiences.

**Day of the Week Performance**: The highest number of engagements (likes, shares, reactions) occur on Monday, followed by Wednesday and Thursday. However, in terms of potential reach and impressions, Wednesday ranks the highest, followed by Thursday and Saturday. This suggests that while more tweets are interacted with on Monday, the tweets posted on Wednesday have a wider audience.

**Underperforming Days**: Friday has the lowest engagements, potential reach, and impressions. This could suggest that Friday is less active for your Twitter audience.

Comparing these insights with the top 1% posts with high impressions and potential reach, we can see some differences:

* While Monday has the highest engagements, it ranks fifth in terms of potential reach and impressions. This could indicate that while the content posted on Monday engages the audience effectively, it does not reach as wide an audience as content posted on other days.
* Conversely, while Wednesday has high potential reach and impressions, it ranks second in terms of engagements. This suggests that while Wednesday’s content reaches a large audience, it does not engage the audience as effectively as content posted on Monday.

**Quarterly Engagement**: The first quarter (Q1) of the year has the highest number of engagements (likes, shares, reactions), followed by the fourth quarter (Q4). This suggests that these might be the best periods to post for maximum engagement.

**Engagement vs Reach**: While Q1 has the highest engagements, it ranks second in terms of potential reach and impressions. This could indicate that while the content posted in Q1 engages the audience effectively, it does not reach as wide an audience as content posted in Q4.

**Underperforming Quarters**: The second (Q2) and third quarters (Q3) have the lowest engagements and potential reach. This could suggest that these periods are less active for your Twitter audience.

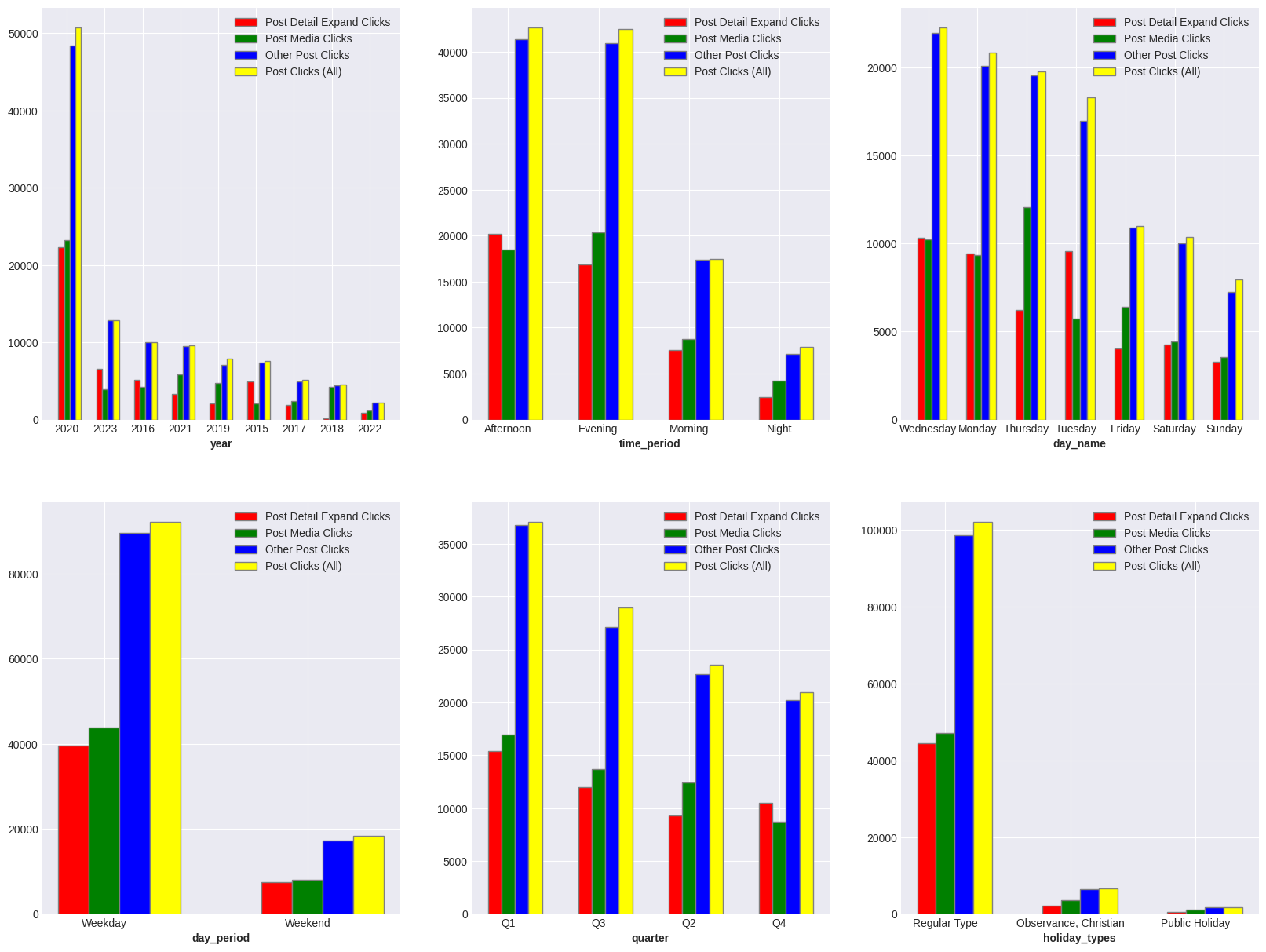
Comparing these insights with the top 1% posts with high impressions and potential reach, we can see some differences:

While Q4 has the highest potential reach and impressions, it ranks second in terms of engagements. This suggests that while Q4’s content reaches a large audience, it does not engage the audience as effectively as content posted in Q1.

Conversely, while Q1 has high engagement rates, it does not have as high a potential reach as Q4. This suggests that while Q1 is good for engaging with your current audience, it might not be as effective for reaching new audiences.  
  
cleaned\_impressions\_reach[click\_cols].describe().T

|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Other Post Clicks** | 7773.0 | 69.601312 | 199.003158 | 0.0 | 15.0 | 32.0 | 67.0 | 9083.0 |
| **Post Clicks (All)** | 7773.0 | 72.855526 | 202.260342 | 0.0 | 17.0 | 34.0 | 71.0 | 9124.0 |
| **Post Media Clicks** | 7773.0 | 38.827608 | 99.699418 | 0.0 | 5.0 | 16.0 | 37.0 | 3698.0 |
| **Post Detail Expand Clicks** | 7773.0 | 24.449762 | 106.224266 | 0.0 | 4.0 | 9.0 | 21.0 | 4776.0 |

last\_quantile = cleaned\_impressions\_reach['Post Clicks (All)'].quantile(0.99)

df = cleaned\_impressions\_reach[cleaned\_impressions\_reach['Post Clicks (All)'] > last\_quantile]  
  
  
**Yearly Clicks vs Engagements**: The year 2020 had the highest number of post clicks (all types) and engagements (likes, shares, reactions). This suggests that this year had the most interactive and engaging content.

**Clicks vs Engagements**: While the year 2020 had the highest number of post clicks and engagements, there are differences in the rankings of other years. For example, 2019 ranks second in terms of engagements but fourth in terms of post clicks. This could indicate that while the content posted in 2019 engaged the audience effectively, it did not result in as many clicks.

**Underperforming Years**: The year 2022 has the lowest number of post clicks and engagements. This could suggest that this year was less active for your Twitter audience.

Comparing these insights with top 1% Engagements, Reactions , Likes and Shares, we can see some differences:

* While 2020 ranks highest in both engagements and post clicks, other years show different trends. For example, while 2023 ranks third in terms of engagements, it ranks second in terms of post clicks. This suggests that while the content posted in 2023 engaged the audience effectively, it resulted in more clicks compared to other years with similar engagement levels.
* Conversely, while 2019 ranks second in terms of engagements, it ranks fourth in terms of post clicks. This suggests that while the content posted in 2019 engaged the audience effectively, it did not result in as many clicks.

**Time of Day Performance**: The evening and afternoon periods have the highest number of post clicks (all types) and engagements (likes, shares, reactions). However, the morning period ranks third in both categories. This suggests that while more tweets are interacted with in the evening and afternoon, the tweets posted in the morning also receive a fair amount of interaction.

**Clicks vs Engagements**: While the afternoon period has the highest number of post clicks, it ranks second in terms of engagements. This could indicate that while the content posted in the afternoon results in more clicks, it does not engage the audience as effectively as content posted in the evening.

**Underperforming Time Period**: The night period has the lowest number of post clicks and engagements. This could suggest that this time is less active for your Twitter audience.

Comparing these insights with top 1% Engagements, Reactions , Likes and Shares, we can see some differences:

* While the evening period ranks highest in both engagements and post clicks, other time periods show different trends. For example, while the morning period ranks third in terms of engagements, it ranks third in terms of post clicks as well. This suggests that while the content posted in the morning engages the audience effectively, it results in a similar amount of clicks.
* Conversely, while the afternoon period ranks second in terms of engagements, it ranks first in terms of post clicks. This suggests that while the content posted in the afternoon engages the audience effectively, it results in more clicks compared to other time periods.

**Day of the Week Performance**: The highest number of post clicks (all types) occur on Wednesday, followed by Monday and Thursday. However, in terms of engagements (likes, shares, reactions), Monday ranks the highest, followed by Wednesday and Thursday. This suggests that while more tweets are interacted with on Monday, the tweets posted on Wednesday result in more clicks.

**Clicks vs Engagements**: While Wednesday has the highest number of post clicks, it ranks second in terms of engagements. This could indicate that while the content posted on Wednesday results in more clicks, it does not engage the audience as effectively as content posted on Monday.

**Underperforming Days**: Friday has the lowest number of post clicks and ranks last in terms of engagements. This could suggest that Friday is less active for your Twitter audience.

Comparing these insights with the top 1% posts with high impressions and potential reach, we can see some differences:

* While Monday ranks highest in both engagements and post clicks, other days show different trends. For example, while Tuesday ranks fourth in terms of engagements, it ranks fourth in terms of post clicks as well. This suggests that while the content posted on Tuesday engages the audience effectively, it results in a similar amount of clicks.
* Conversely, while Wednesday ranks second in terms of engagements, it ranks first in terms of post clicks. This suggests that while the content posted on Wednesday engages the audience effectively, it results in more clicks compared to other days.

**Quarterly Clicks vs Engagements**: The first quarter (Q1) of the year has the highest number of post clicks (all types) and engagements (likes, shares, reactions). However, the second quarter (Q2) ranks fourth in both categories. This suggests that while more tweets are interacted with in Q1, the tweets posted in Q2 receive less interaction.

**Clicks vs Engagements**: While Q1 has the highest number of post clicks and engagements, other quarters show different trends. For example, while Q3 ranks second in terms of post clicks, it ranks third in terms of engagements. This could indicate that while the content posted in Q3 results in more clicks, it does not engage the audience as effectively as content posted in other quarters.

**Underperforming Quarters**: The fourth quarter (Q4) has the lowest number of post clicks but ranks second in terms of engagements. This could suggest that this period is less active for your Twitter audience in terms of clicking on posts, but they are more engaged with the content.

Comparing these insights with the top 1% posts with high impressions and potential reach, we can see some differences:

* While Q1 ranks highest in both engagements and post clicks, other quarters show different trends. For example, while Q4 ranks second in terms of engagements, it ranks last in terms of post clicks. This suggests that while the content posted in Q4 engages the audience effectively, it does not result in as many clicks.
* Conversely, while Q3 ranks third in terms of engagements, it ranks second in terms of post clicks. This suggests that while the content posted in Q3 engages the audience effectively, it results in more clicks compared to other quarters.

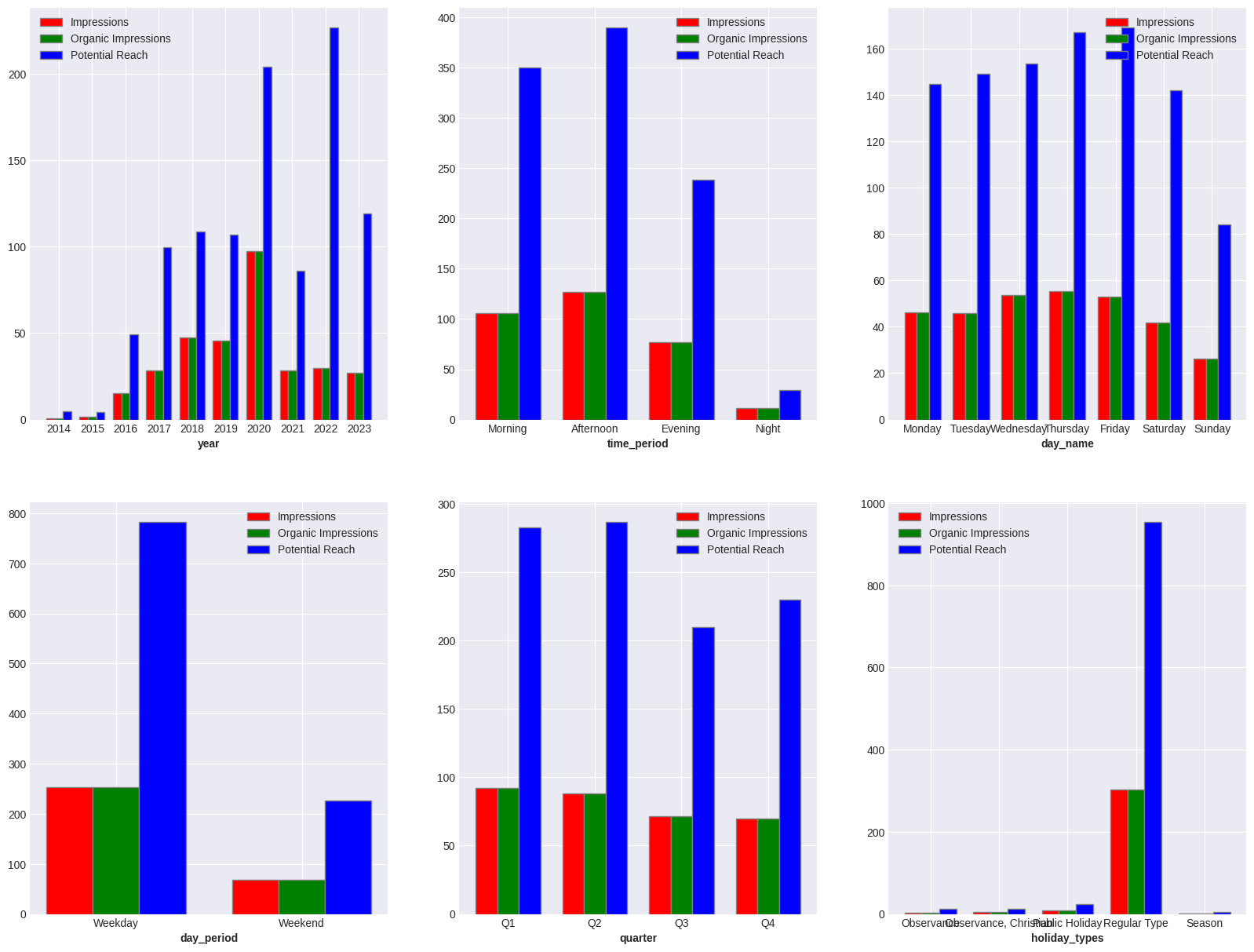
**Using the Whole Dataset**

**Yearly Performance**: The year 2017 had the highest potential reach among the top 1% of posts with the highest potential reach. However, when considering the whole dataset, the year 2022 had the highest potential reach. This suggests that while a few posts in 2017 reached a large audience, overall, more people were reached in 2022.

**High Performing Posts vs Overall Performance**: There seems to be a difference in trends when comparing the top 1% of posts with the highest potential reach and the overall dataset. For example, while 2017 ranks first in terms of potential reach for the top 1% of posts, it ranks seventh when considering all posts. This could indicate that a small number of highly successful posts in 2017 skewed the results for that year.

**Underperforming Years**: The year 2015 has the lowest potential reach both among the top 1% of posts and in the overall dataset. This could suggest that this year was less active for your Twitter audience.

As for further research, it might be interesting to explore what made the top-performing posts so successful. Were there any common themes or characteristics among these posts? Additionally, analyzing changes in Twitter’s user base or algorithms over the years could provide more context for these trends.

  
  
**Time of Day Performance**: For the top 1% of posts with the highest potential reach, the morning period has the highest potential reach. However, when considering the whole dataset, the afternoon period has the highest potential reach. This suggests that while a few posts in the morning reached a large audience, overall, more people were reached in the afternoon.

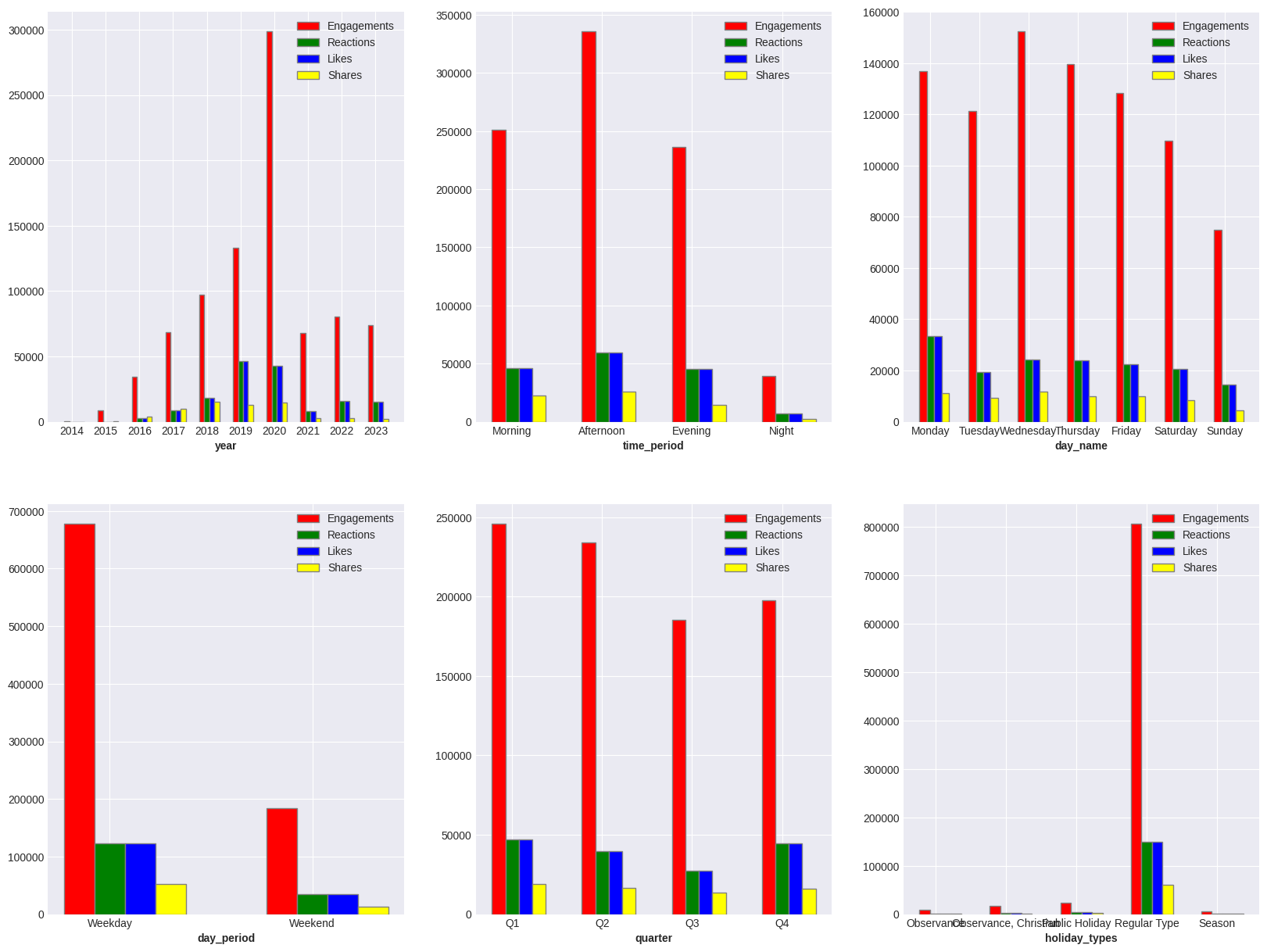
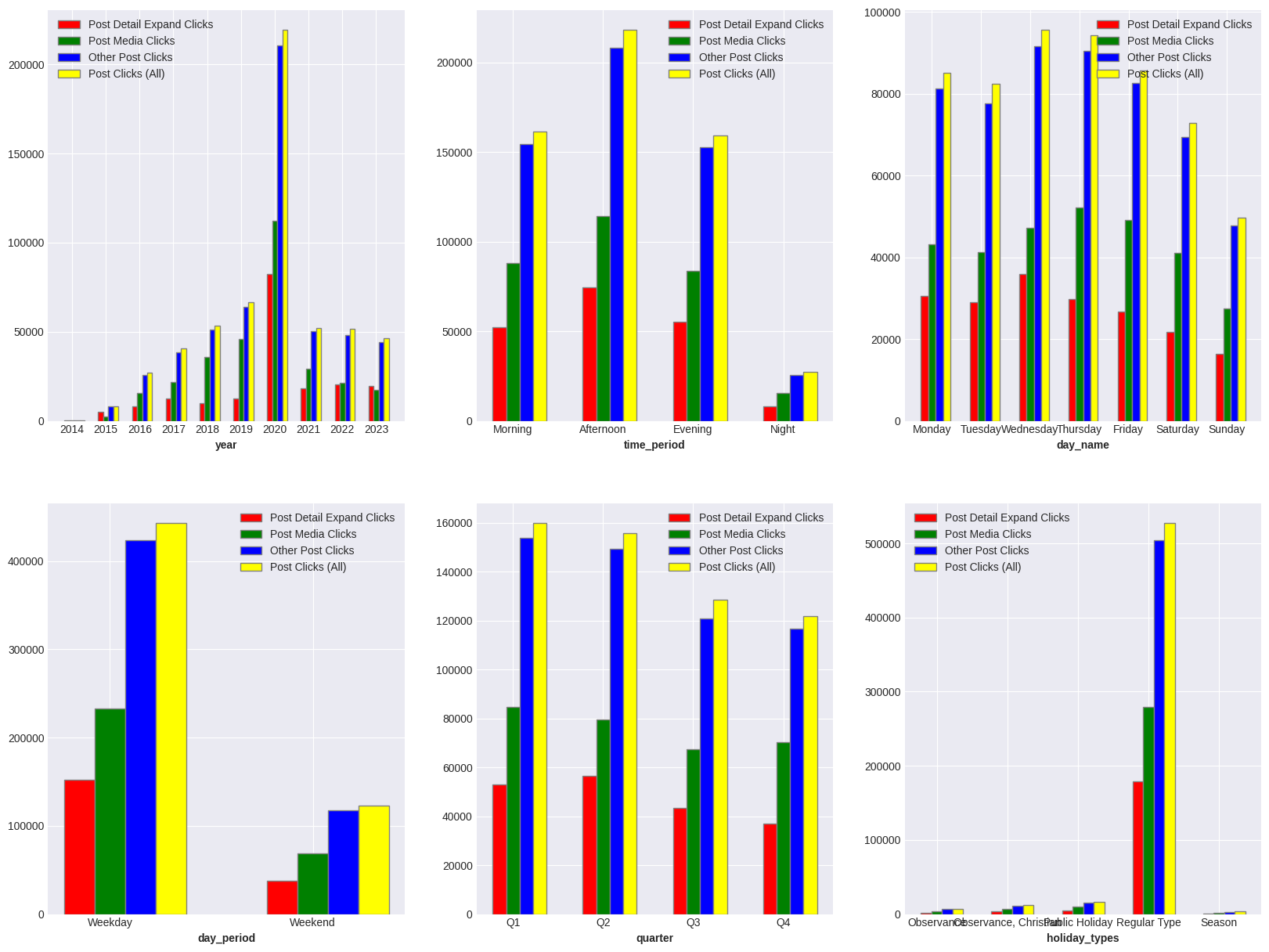
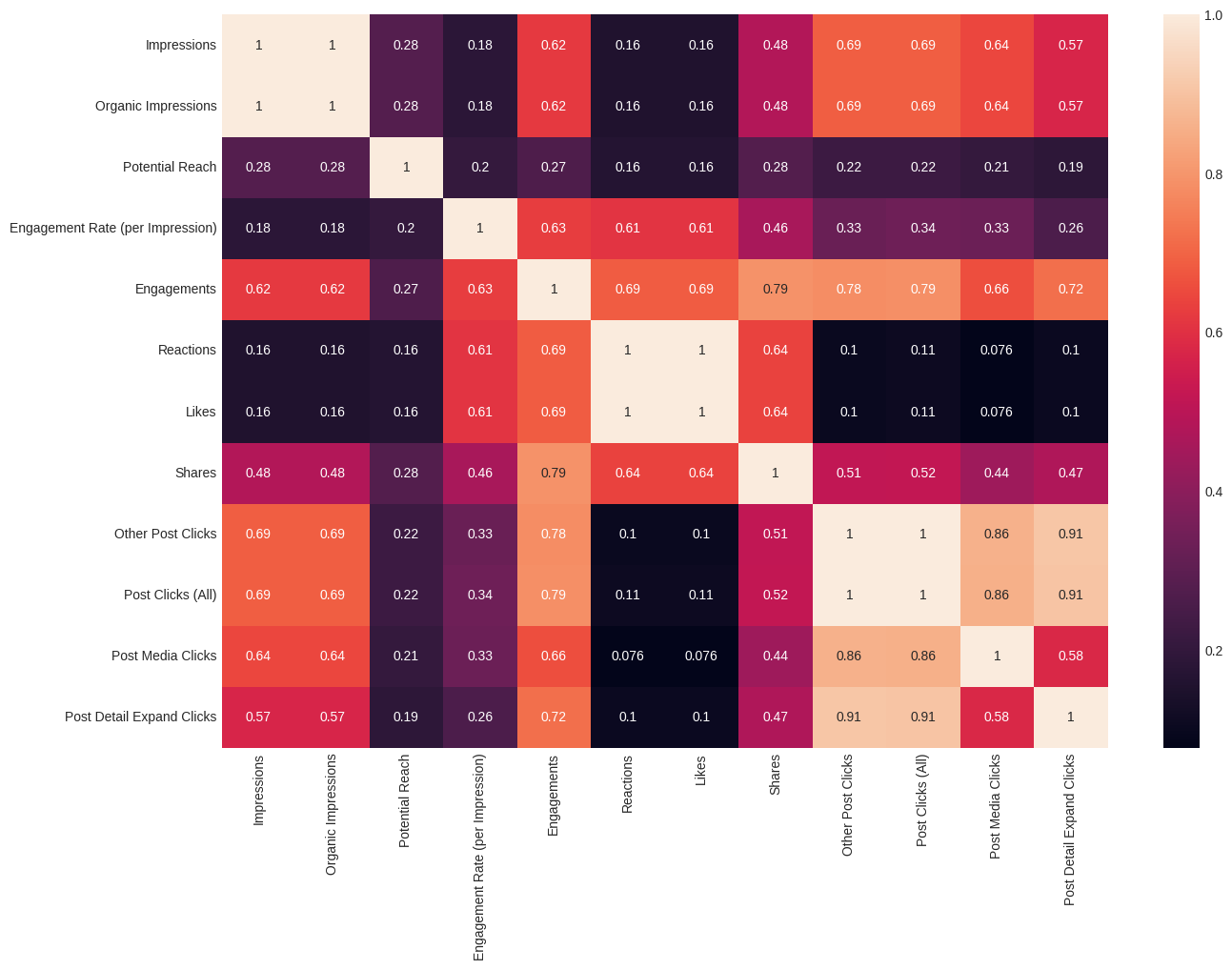
**High Performing Posts vs Overall Performance**: There seems to be a difference in trends when comparing the top 1% of posts with the highest potential reach and the overall dataset. For example, while the morning period ranks first in terms of potential reach for the top 1% of posts, it ranks second when considering all posts. This could indicate that a small number of highly successful posts in the morning skewed the results for that time period.

**Underperforming Time Period**: The night period has the lowest potential reach both among the top 1% of posts and in the overall dataset. This could suggest that this time is less active for your Twitter audience.

**Day of the Week Performance**: For the top 1% of posts with the highest potential reach, Wednesday has the highest potential reach. However, when considering the whole dataset, Friday has the highest potential reach. This suggests that while a few posts on Wednesday reached a large audience, overall, more people were reached on Friday.

**High Performing Posts vs Overall Performance**: There seems to be a difference in trends when comparing the top 1% of posts with the highest potential reach and the overall dataset. For example, while Wednesday ranks first in terms of potential reach for the top 1% of posts, it ranks third when considering all posts. This could indicate that a small number of highly successful posts on Wednesday skewed the results for that day.

**Underperforming Days**: Sunday has the lowest potential reach both among the top 1% of posts and in the overall dataset. This could suggest that this day is less active for your Twitter audience.

  
  
  
**Heatmap of the Correlations in our Dataset**  
  
  
**Impressions and Organic Impressions**: These two metrics are perfectly correlated with a correlation coefficient of 1. This suggests that every time an impression is made, it’s an organic impression.

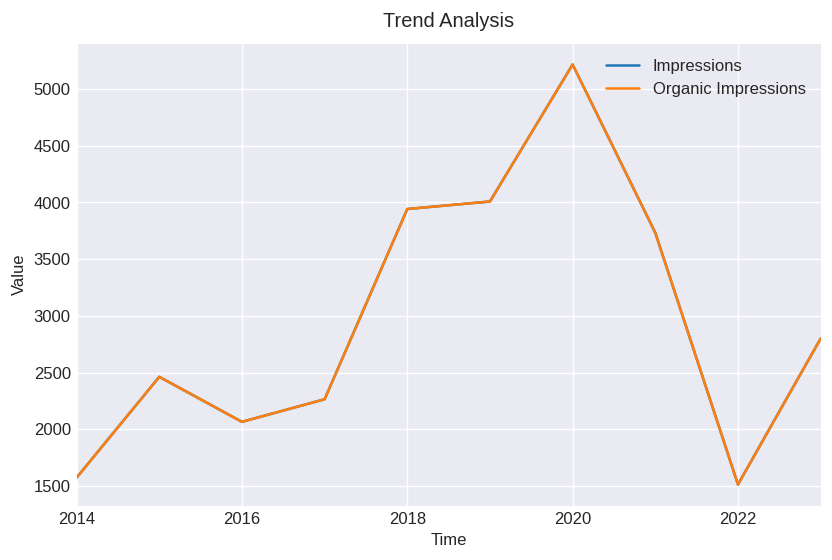
**Engagements and Other Metrics**: Engagements have a strong positive correlation with other post clicks (0.78) and post clicks (all) (0.79). This indicates that as engagements increase, the number of post clicks also tends to increase.

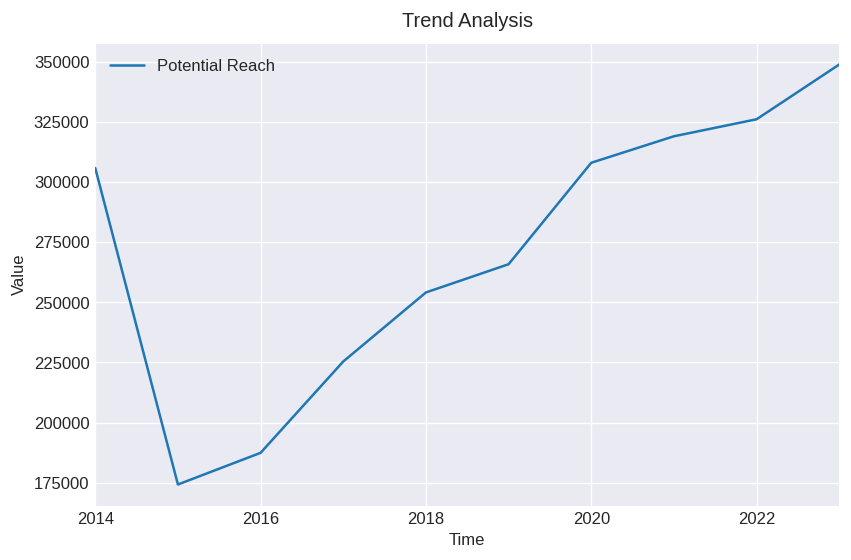
**Post Clicks (All) and Other Post Clicks**: These two metrics also have a very strong positive correlation (0.99), suggesting that other post clicks make up a large portion of all post clicks.

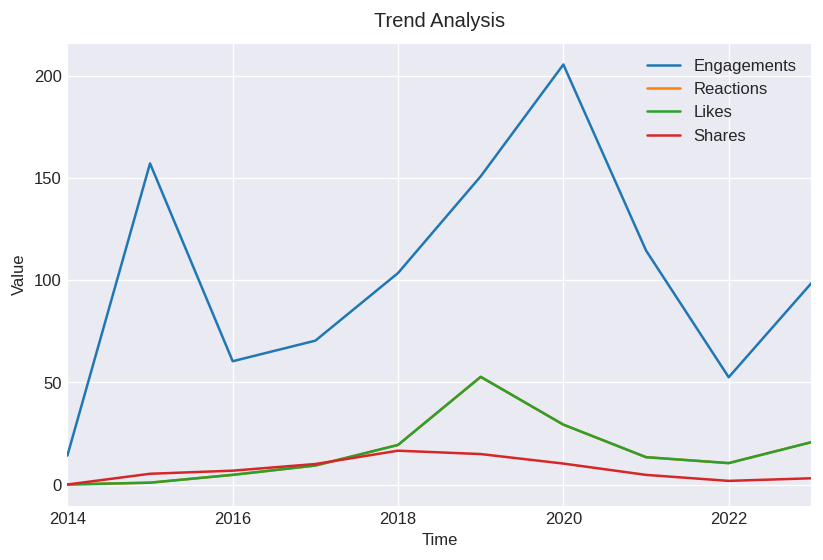
**Potential Reach and Other Metrics**: Potential reach has a moderate positive correlation with impressions and organic impressions (both 0.28), indicating that as the potential reach of a post increases, the number of impressions tends to increase as well.

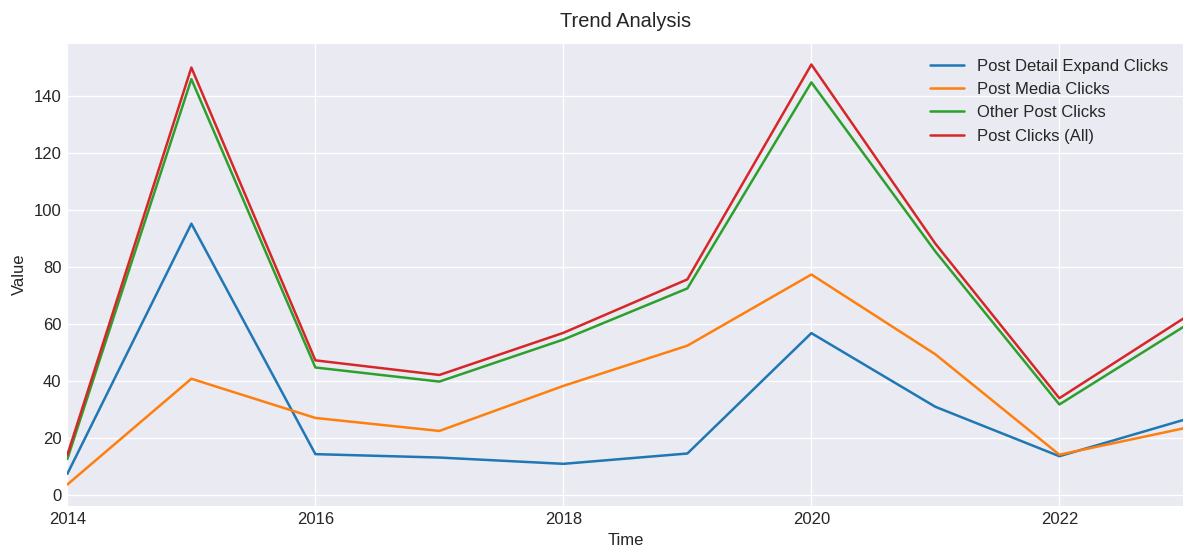
**Observing Trends in our Dataset**

We can observe the following Trends of Impressions/Organic Impressions over the Years

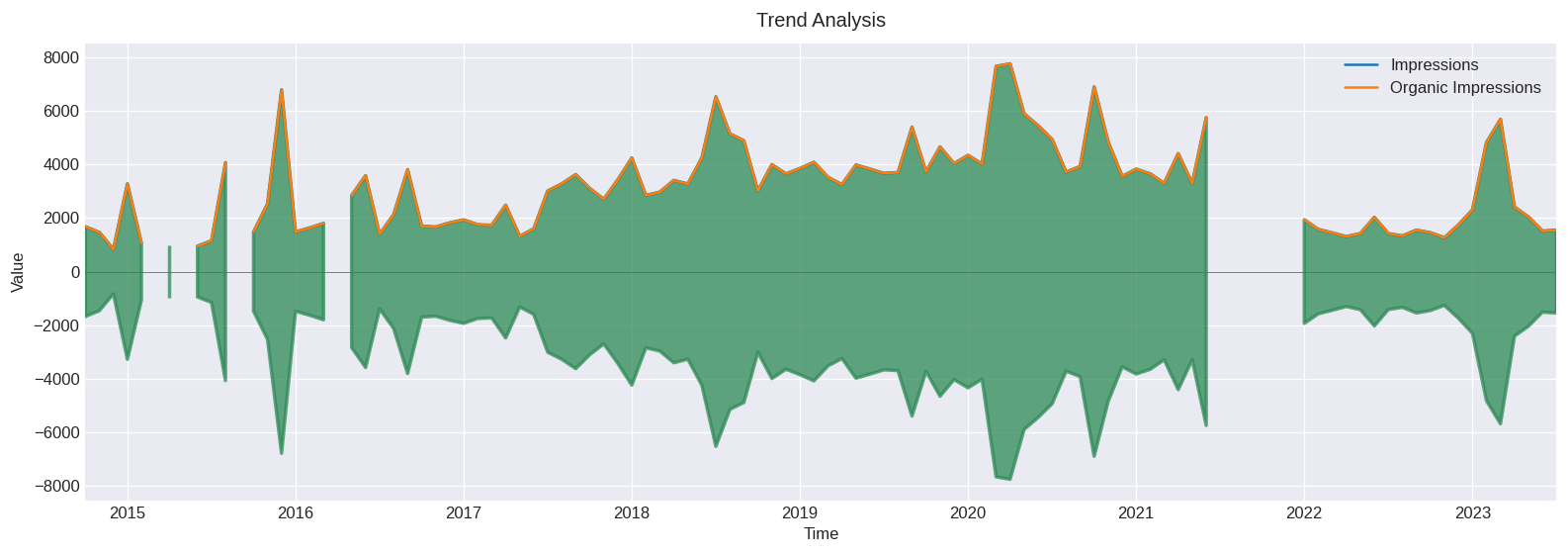
  
We can observe that the average value of Impressions/Organic Impressions rises from 2018 to 2020 and then drops in 2022

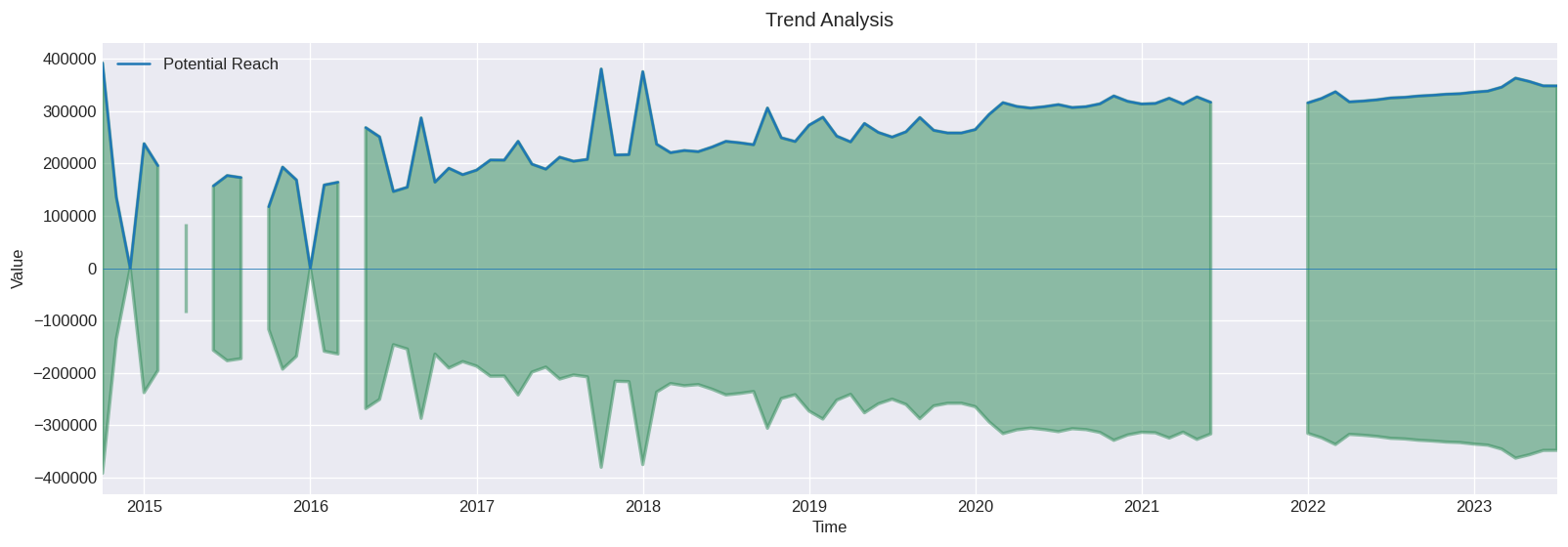
  
But here we can observe that the average value of Potential Reach rises from 2015 and keeps rising steadily over the years

  
The Trend of Engagements just like the Trend of Impressions/Organic Impressions rises till 2020 but then there is a sharp drop in 2022

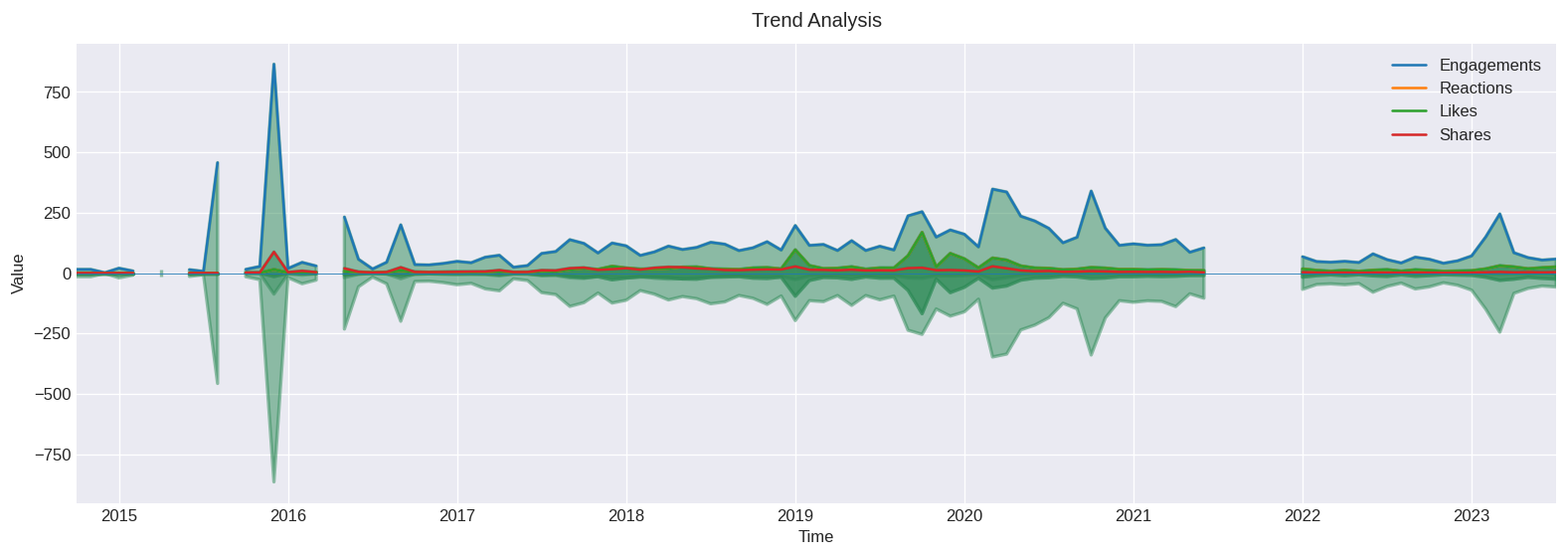
  
Engagements have a strong positive correlation with other post clicks (0.78) and post clicks (all) (0.79). This indicates that as engagements increase, the number of post clicks also tends to increase. This can be observed in the above Trend Analysis

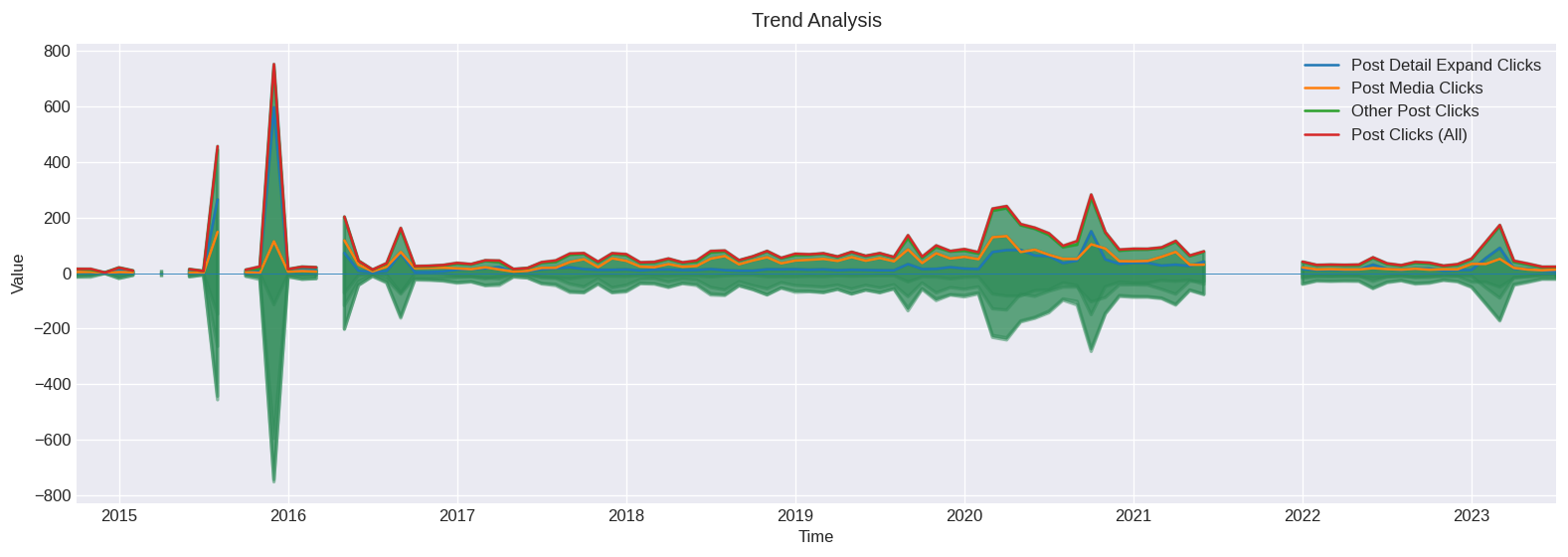
Let us view the trend of average number of Impressions over the months

  
We can see that though it increases per month, it fluctuates, it goes up and down, we see the highest values at 2015-11, 2020-03, 2020-08, 2021-07,2023-02

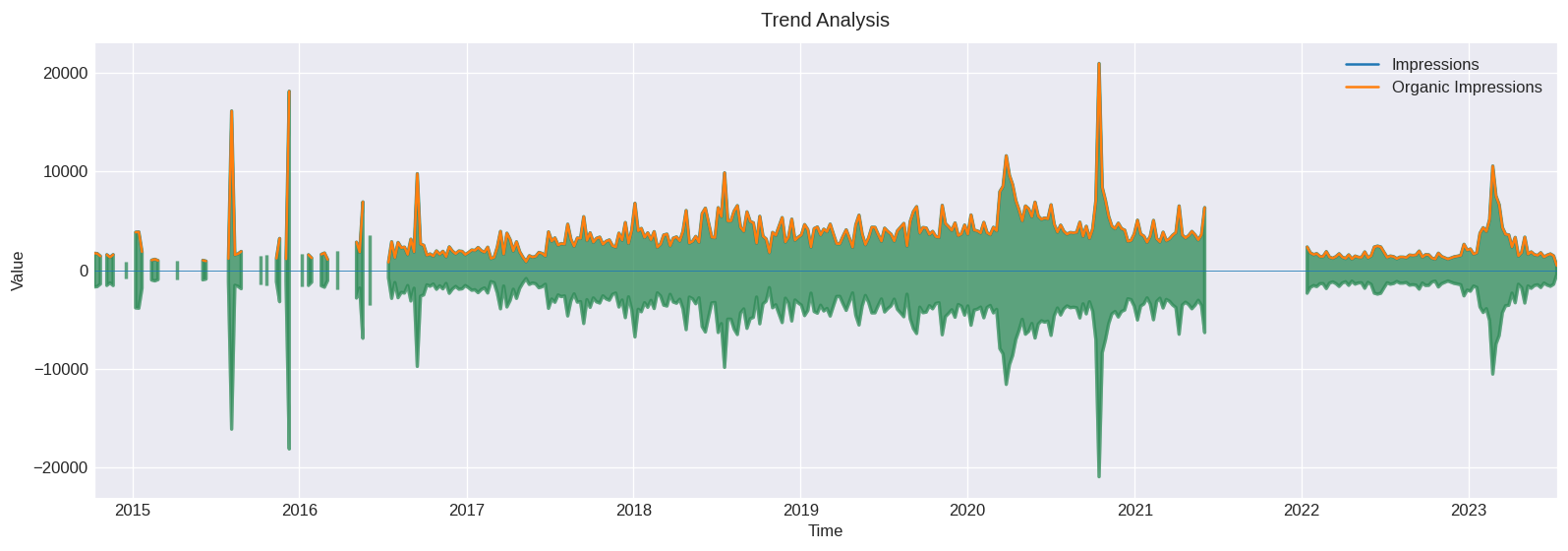


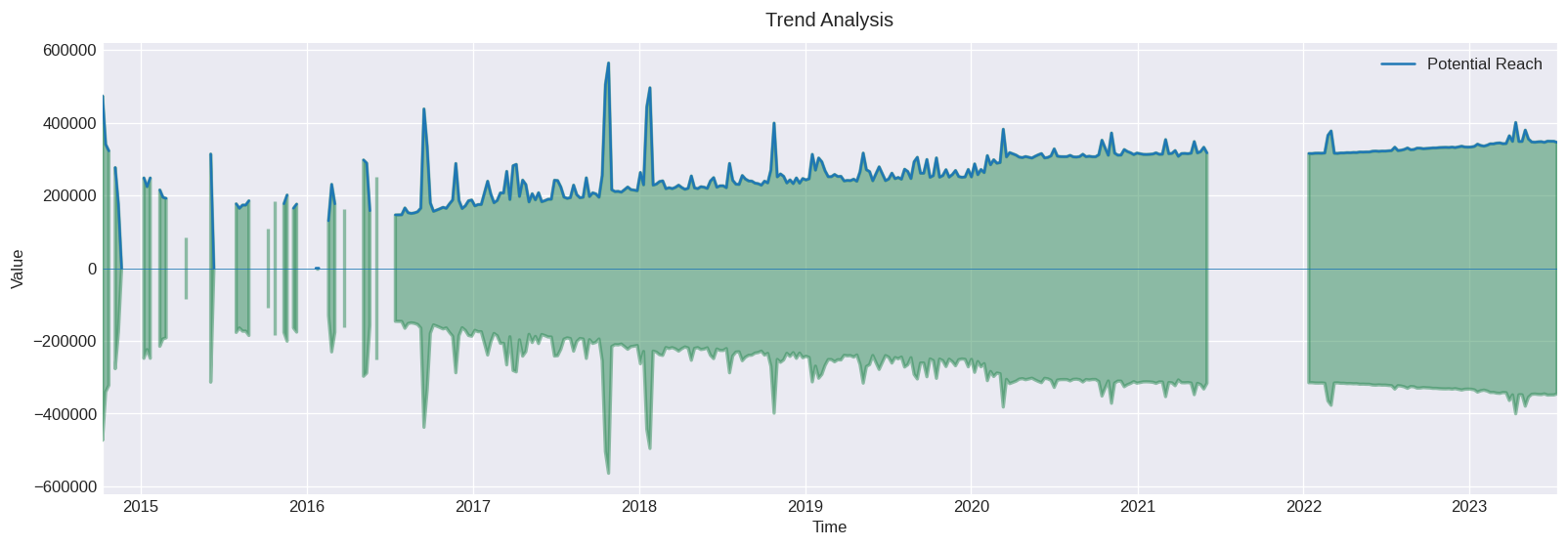
We can see the average value of Potential Reach across the Months, it increases but also fluctuates but not as much as Impressions/Organic Impressions did, we can see the highest values at 2017-09, 2018-01

  
We can see the average values of Engagements across the Months, it increases but fluctuates just as wildly as Impressions/Organic Impressions did, we can also see how correlated it is, the highest points for Impressions/Organic Impressions are also the highest Points for Engagements across the Months

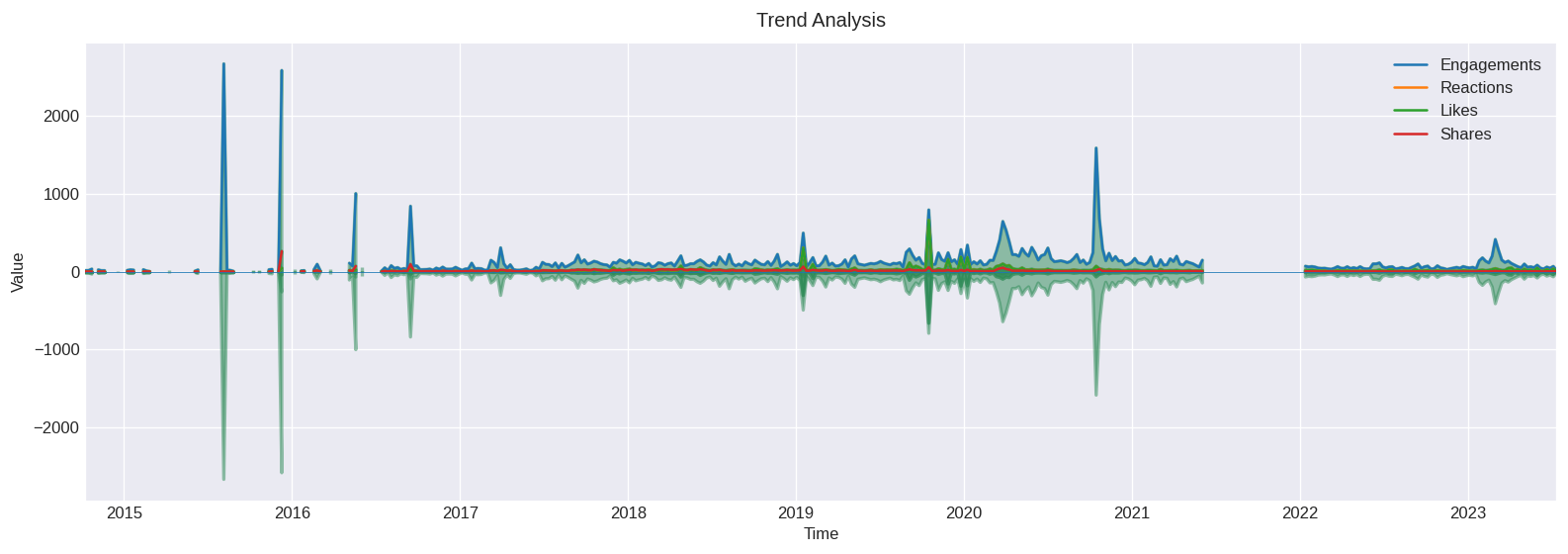
  
We can see the average values of our Post Media Clicks, Other Post Clicks, Post Clicks(All) across the months, and just as we see **Engagements and Other Metrics**: Engagements have a strong positive correlation with other post clicks (0.78) and post clicks (all) (0.79). This indicates that as engagements increase, the number of post clicks also tends to increase. This correlation has also been proven here.

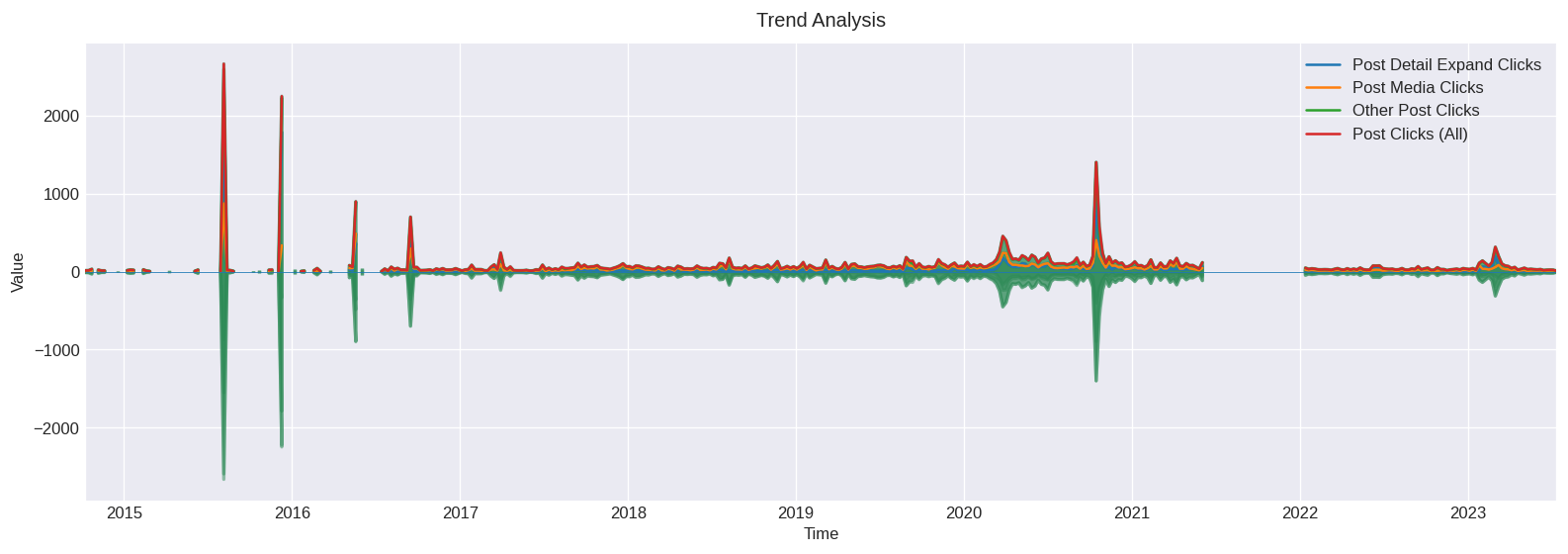
Diving Deeper

  
We now see the average values of Impressions/Organic Impressions across the week, and as it increases we can also see that majority of our posts are between the period 2017-2021, we see sharp spikes at 2015-05,2015-09, 2020-08

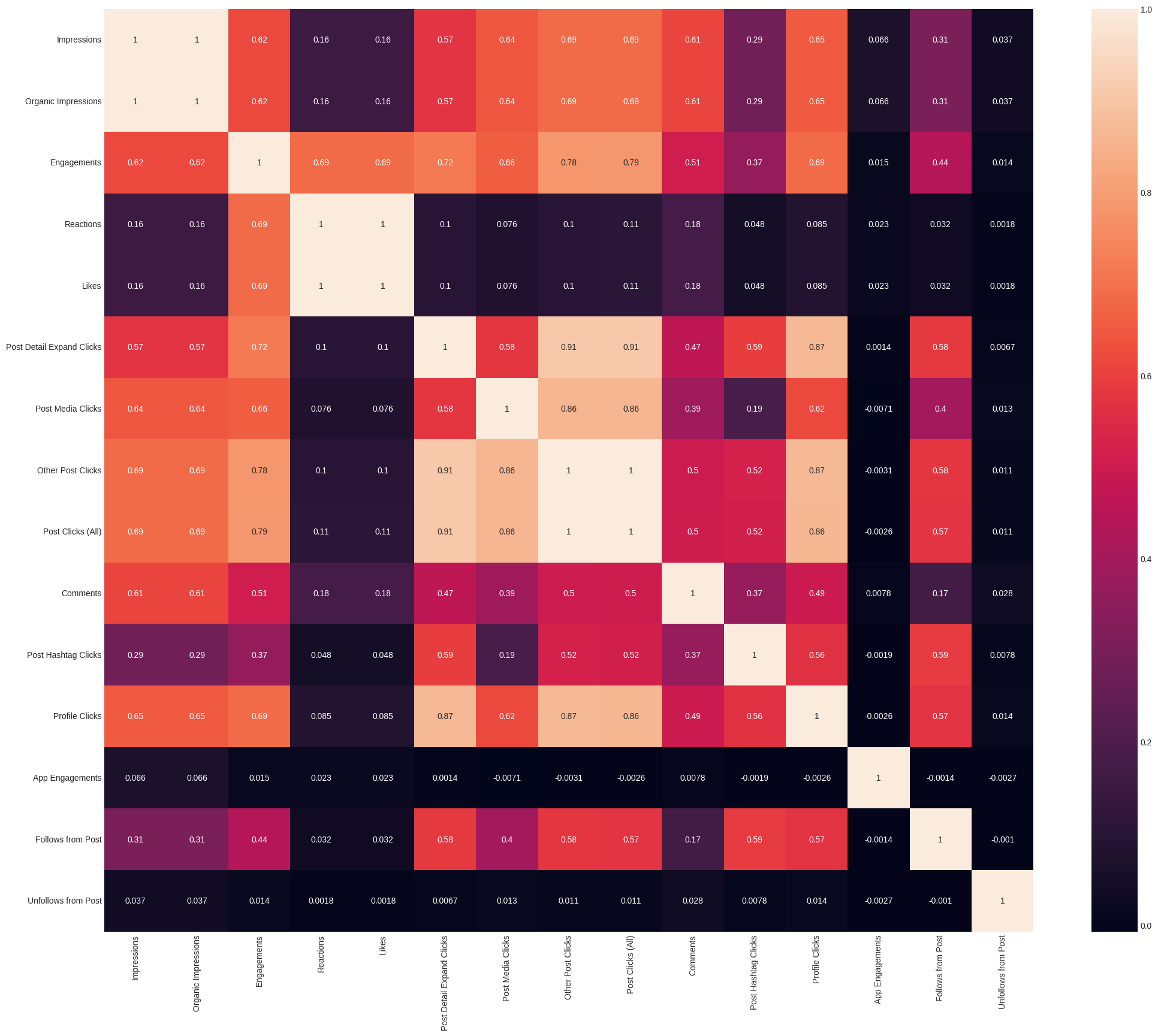


The same observable trends of the average values of Potential Reach over the months can also be seen over the weeks.

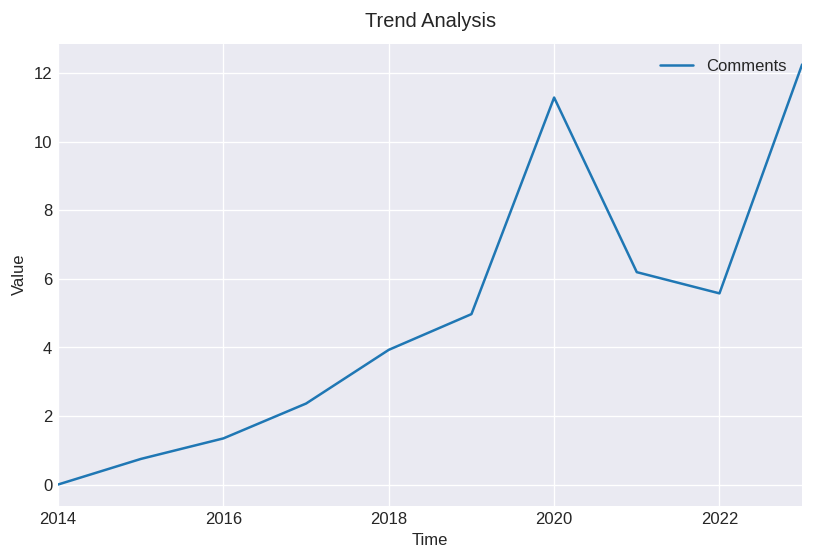




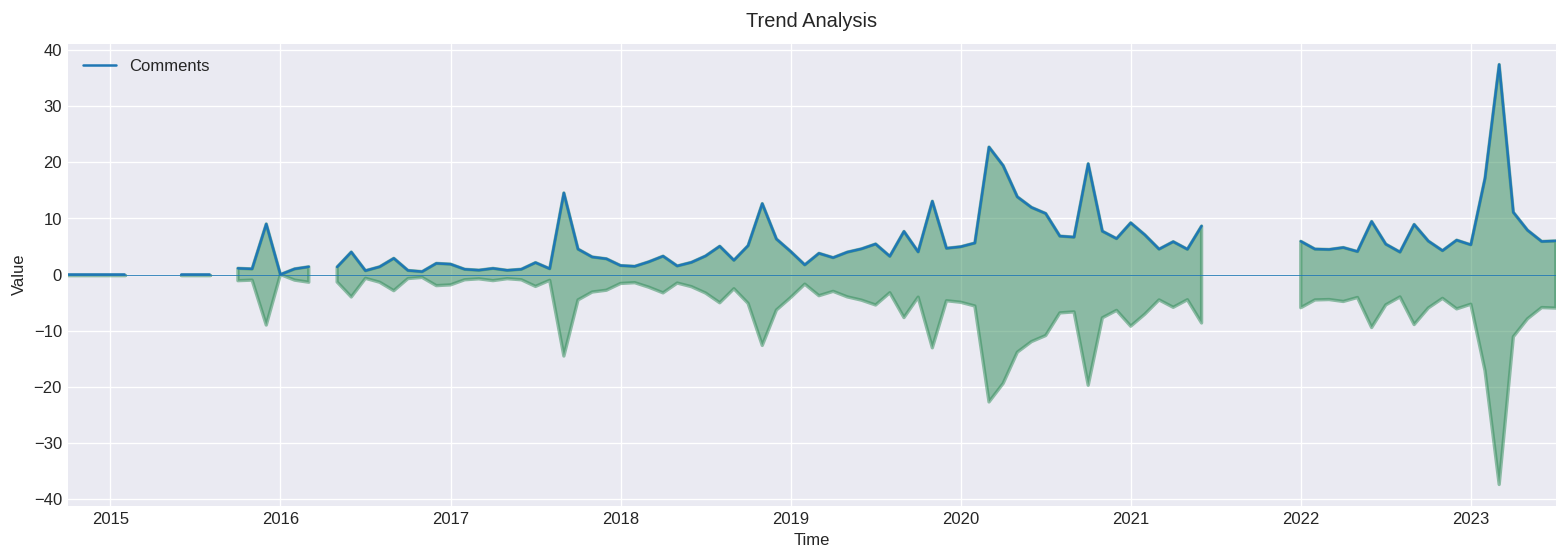
We can see the average values of our Post Media Clicks, Other Post Clicks, Post Clicks(All) across the weeks, and just as we see **Engagements and Other Metrics**: Engagements have a strong positive correlation with other post clicks (0.78) and post clicks (all) (0.79). This indicates that as engagements increase, the number of post clicks also tends to increase. This correlation has also been proven here.



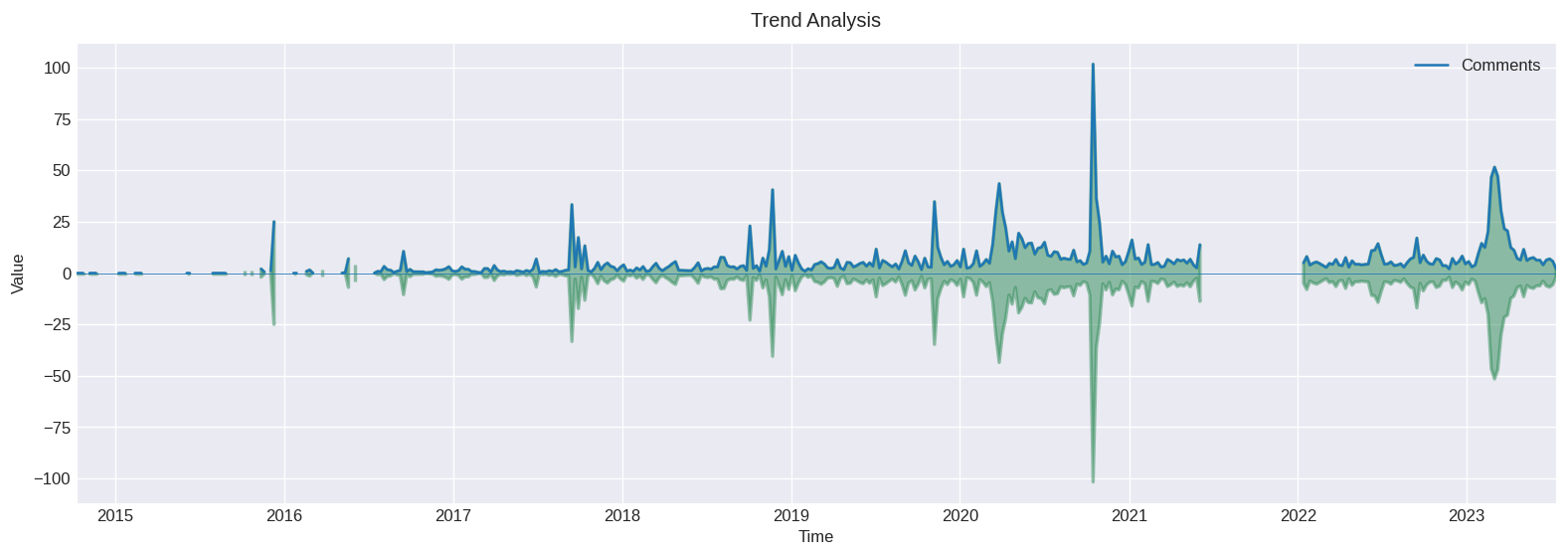
Comments has a strong correlation of about 0.61 with Impressions and Organic Impressions, 0.51 correlation with Engagements, Post Hashtag clicks has a strong correlation of about 0.59 with post detail expand click, follows from post, and Post Clicks(All), Profile Clicks has a 0.69 correlation with Impressions, Organic Impression and Engagements, 0.87 correlation with the click\_cols, Follows from Post has a correlation of 0.57 from Profile Clicks, and Post Hashtag Clicks and from Click Cols

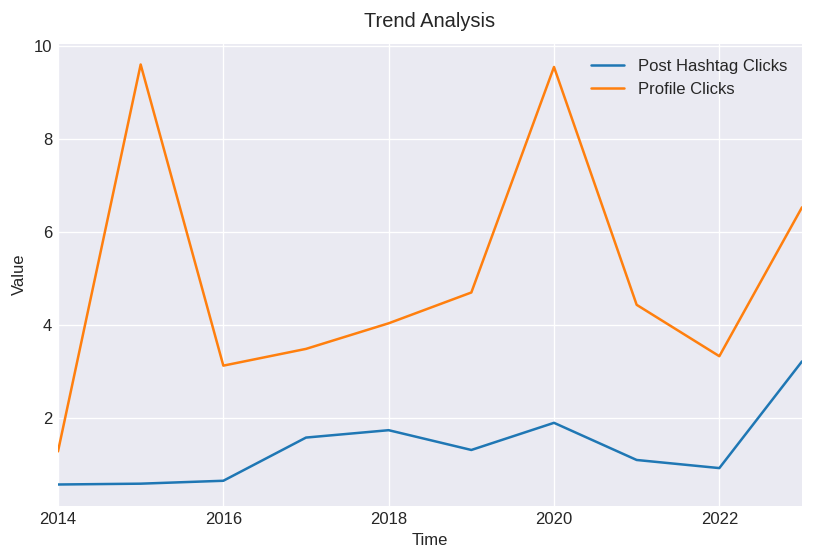
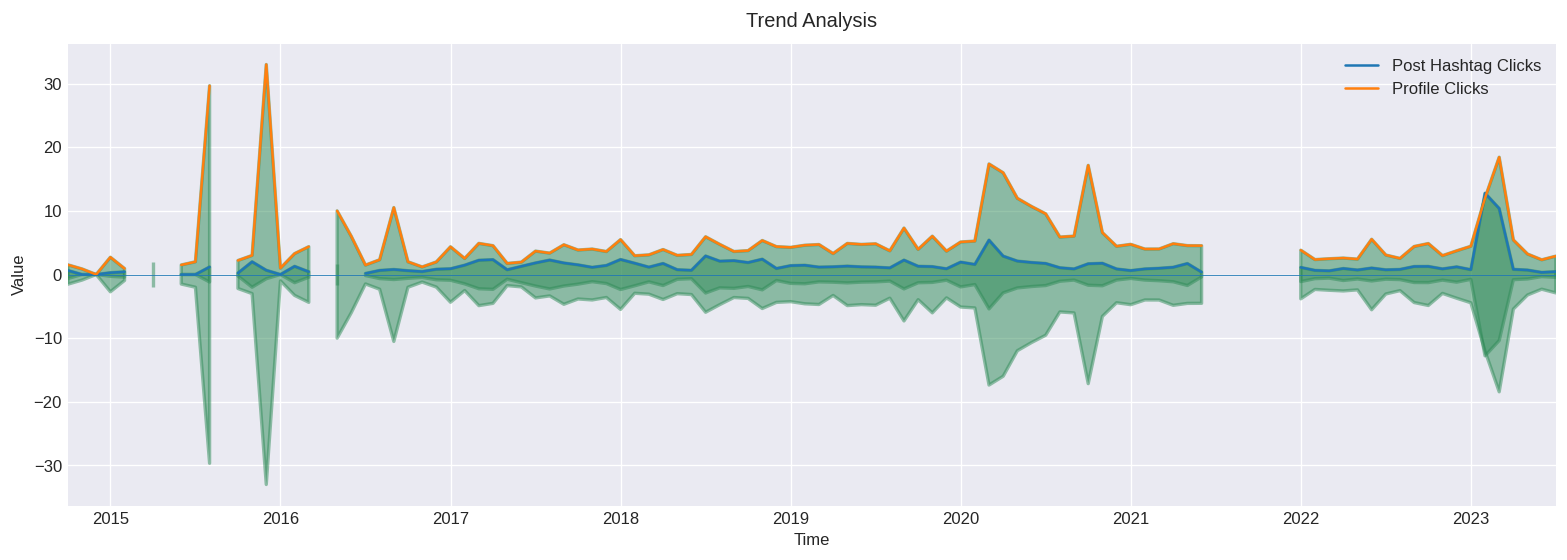


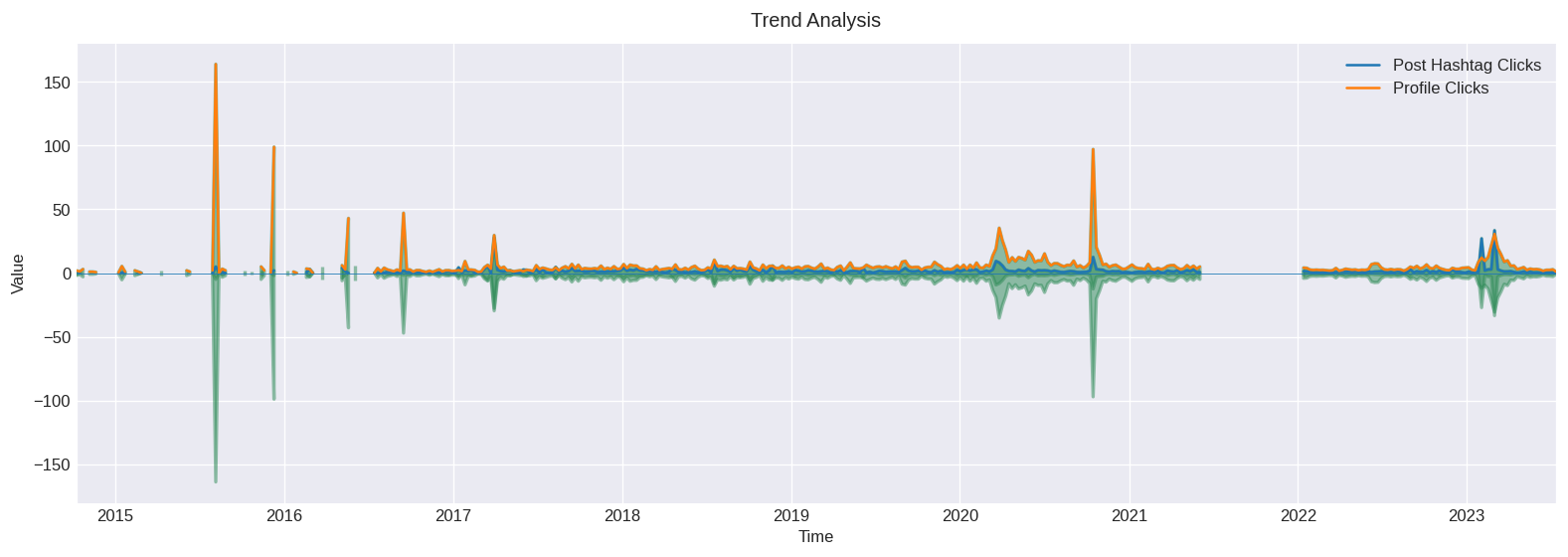
The Average Comments Increases over the Years, a sharp increase from 2019 till 2020 and then a sharp drop to 2021 and then 2022 before rising again



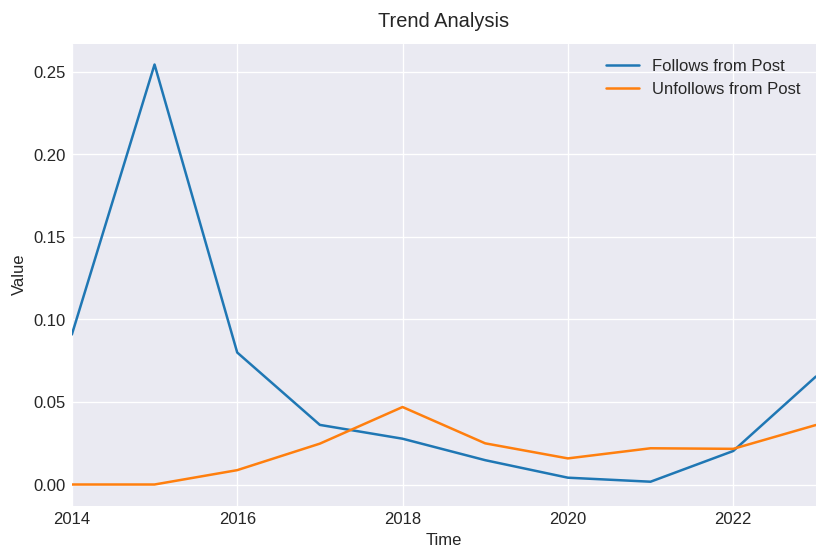
The Average Comments Increases over the Months though still rising and falling, we see sharp spikes at 2017-08, 2018-10, 2020-02 , 2020-06 and 2023-02.

  
Average Comments over the Weeks, we see spikes in 2017-10, 2018-11, 2020-02, 2020-08, and 2023-02, this is very similar to the trend of Engagements/Impression/Organic Impressions over the weeks, confirming their correlations

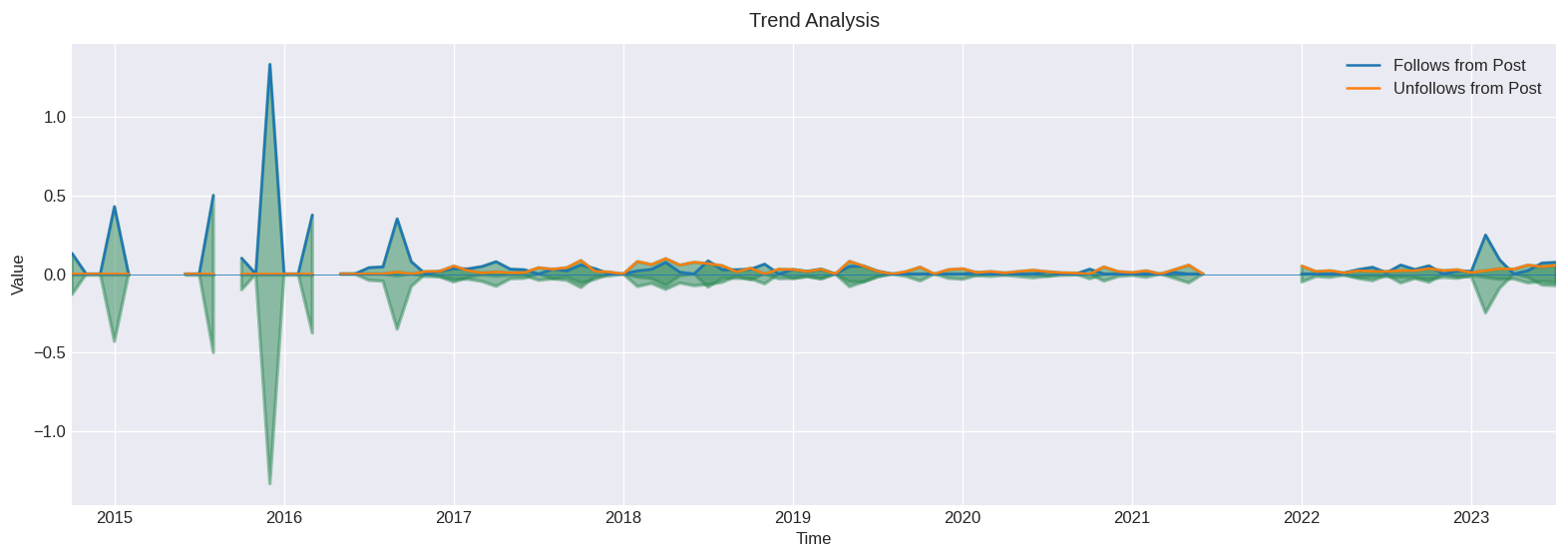
  




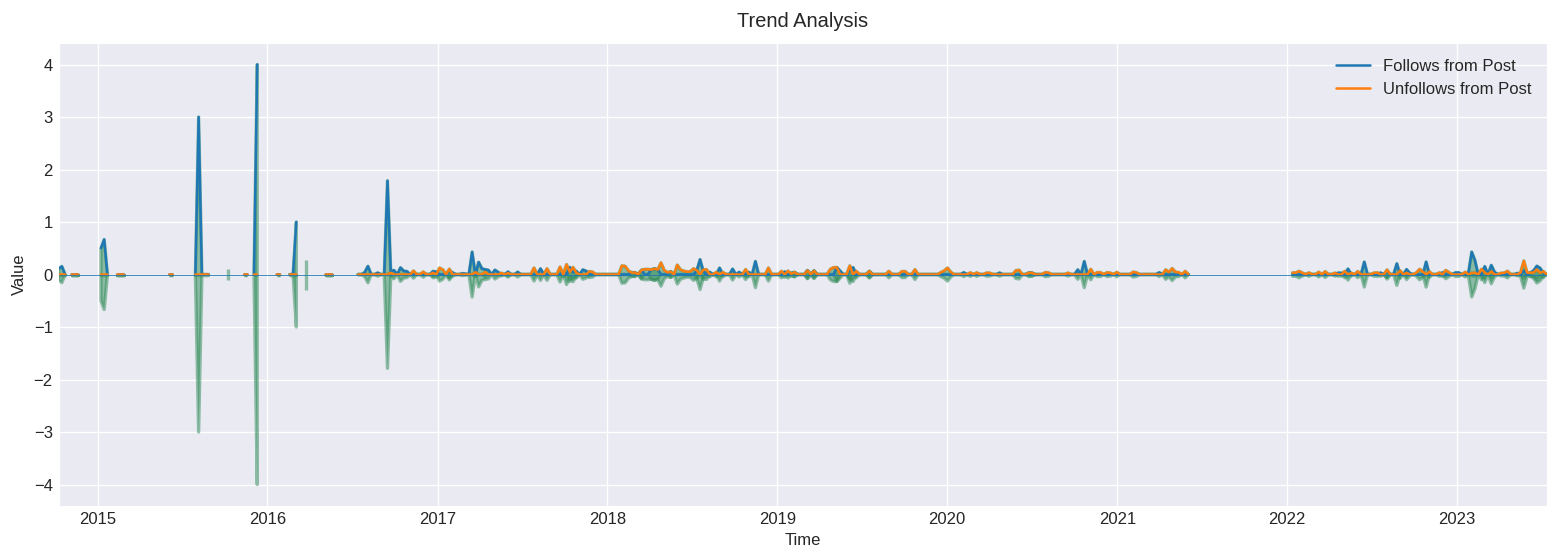
We can see that Post Hashtag Clicks, Profile Clicks trends across Weeks, Months and even Years, is very similar the trends of Engagements/Click Cols across Weeks, Months and even Years



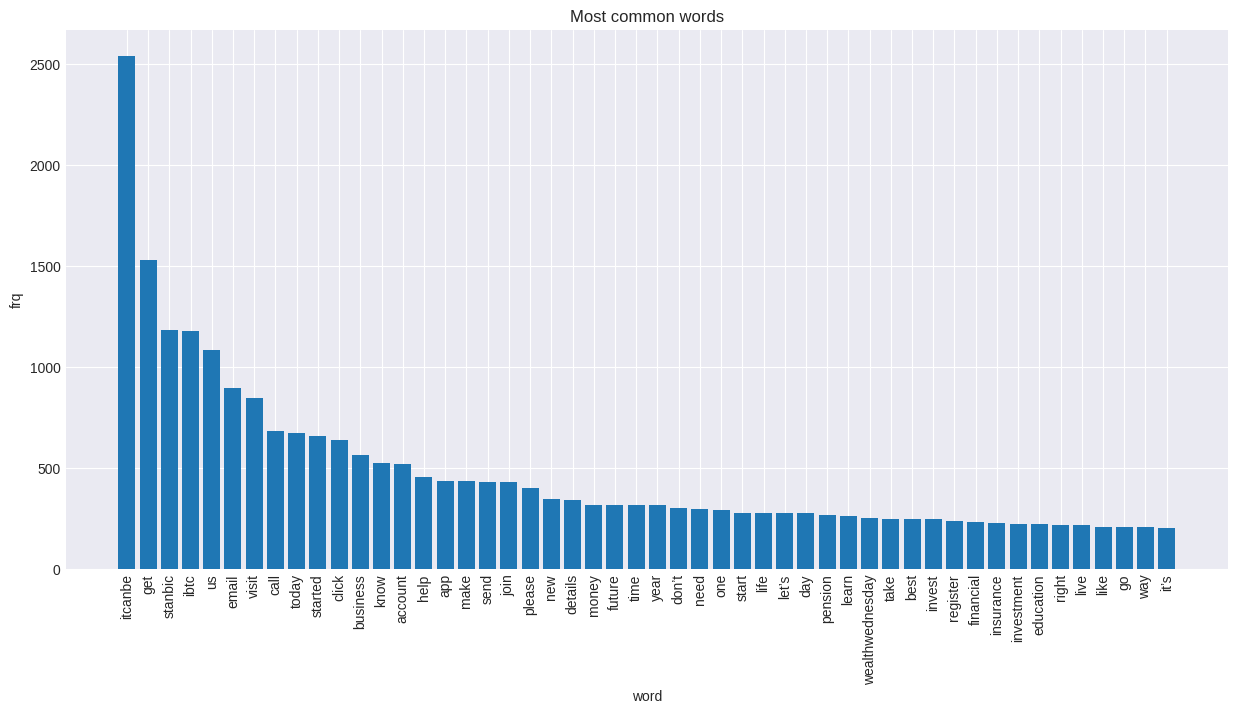
We haven’t had much Followers from our Post, Neither have we gotten much unfollowers across the years



2015-11 had a lot of followers

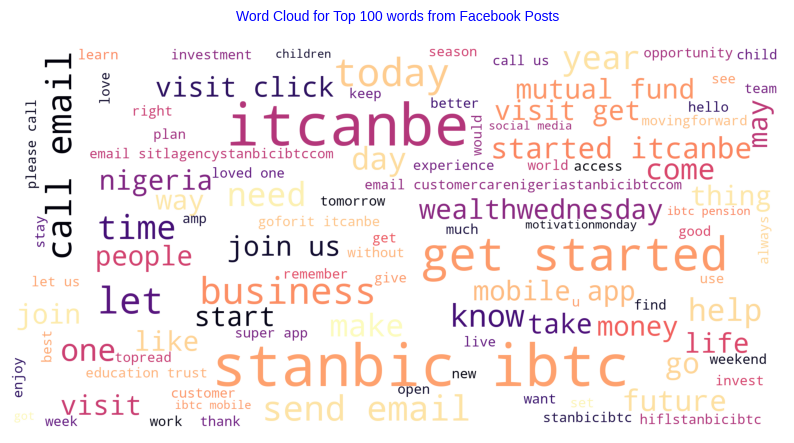


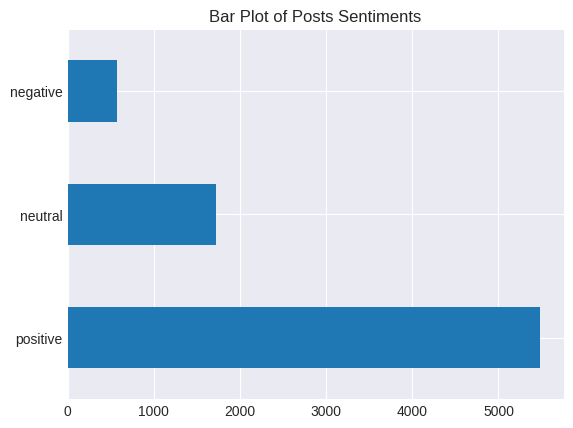
TOP 50 WORDS IN OUR TWITTER POSTS



WORDCLOUD FOR TOP 100

The most frequent words in our captions are itcanbe, stanbic, ibtc, visit, us, email, call, click, know, account and business, we are convincing people to join the company, invest in the company, create accounts, and make business decisions with us

  
Checking out our Word Cloud distribution we can see the most frequent words, such as customer care nigeria stanbic ibtc education trust wealth wednesday get started, we can see that most of our posts revolve around Creating Accounts, Investments, Personal Finance and Future Planning, We can also see the Sentiment Labels of our posts, A majority of our posts are Positive, or have a Positive outlook, we have about 5000+ Positive Posts, With < 2000 Posts being recognized as being Neutral, and < 1000 have a Negative Outlook.



Using Unsupervised Machine Learning Methods such as Topic Modeling:

Topic modeling is a way of abstract modeling to discover the abstract ‘topics’ that occur in the collections of documents. The idea is that we will perform unsupervised classification on different documents, which find some natural groups in topics. We can answer the following question using topic modeling.

What is the topic/main idea of the document?

Given a document, can we find another document with a similar topic?

How do topics field change over time?

Latent Dirichlet Allocation

Latent Dirichlet allocation is one of the most popular methods for performing topic modeling. Each document consists of various words and each topic can be associated with some words. The aim behind the LDA to find topics that the document belongs to, on the basis of words contains in it. It assumes that documents with similar topics will use a similar group of words. This enables the documents to map the probability distribution over latent topics and topics are probability distribution.

Let's break down these concepts:

1. **Topic Modelling using Latent Dirichlet Allocation (LDA)**: Imagine you have a huge pile of books but they don't have any labels or categories. You want to organize them, but reading each one would take forever. Topic Modelling is like a smart librarian who can skim through the books and sort them into different piles based on the main topics they cover. LDA is a popular method for doing this. It assumes that each book (or document) is a mix of various topics, and each topic is a collection of words. By looking at which words often appear together, LDA can figure out the possible topics and how much each document belongs to each topic.

2. **Named Entity Recognition (NER)**: This is like a smart highlighter. When you read a text, you can usually pick out names of people, places, organizations, dates, etc. NER does the same thing but automatically. It scans through the text and highlights these "named entities". This can be very useful in understanding who or what the text is about.

3. **Part-of-Speech Tagging (POS)**: This is like an English teacher analyzing a sentence. It involves labeling each word with its appropriate part of speech (like noun, verb, adjective, etc.). This helps to understand the grammatical structure of the sentence and can be used in many applications like text-to-speech, grammar checking, and more.

In the context of social media posts analysis:

- **Topic Modelling** can help us understand what are the main topics people are talking about.

- **NER** can tell us which specific entities (like people, places, or organizations) are being mentioned.

- **POS Tagging** can help us understand the context better by knowing the grammatical role of each word.

By combining these techniques, we can get a rich and structured understanding of unstructured social media data.

Using LDA to get the 8 main topics in our dataset

- Topic 0: **Promotions and Contests**: This topic seems to be about various offers and incentives that the bank provides to its customers, such as winning prizes, watching sessions, and using easewallet.

- Topic 1: **Nigeria and Africa**: This topic seems to be about the bank's presence and impact in Nigeria and Africa, as well as its involvement in the local economy, industry, and society.

- Topic 2: **Email and Investment:** This topic seems to be about the bank's email communication and investment services, such as sending details, visiting websites, and providing education and insurance.

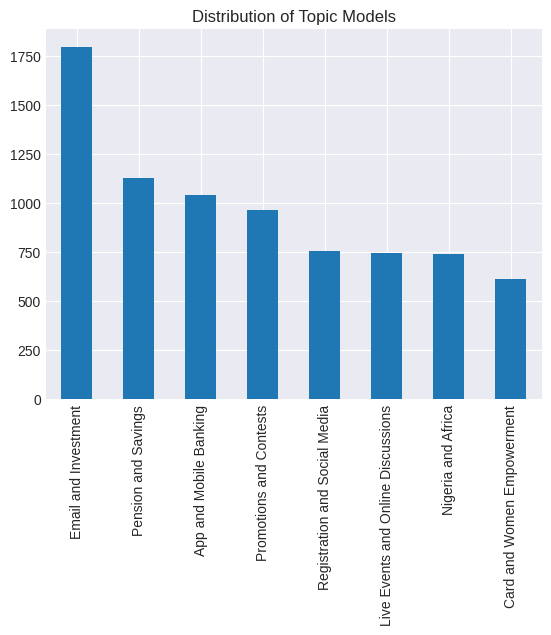
- Topic 3: **App and Mobile Banking:** This topic seems to be about the bank's app and mobile banking features, such as downloading, moving forward, and opening accounts.

- Topic 4: **Live Events and Online Discussions:** This topic seems to be about the bank's participation and hosting of live events and online discussions, such as smwlagos, motivationmonday, and business series.

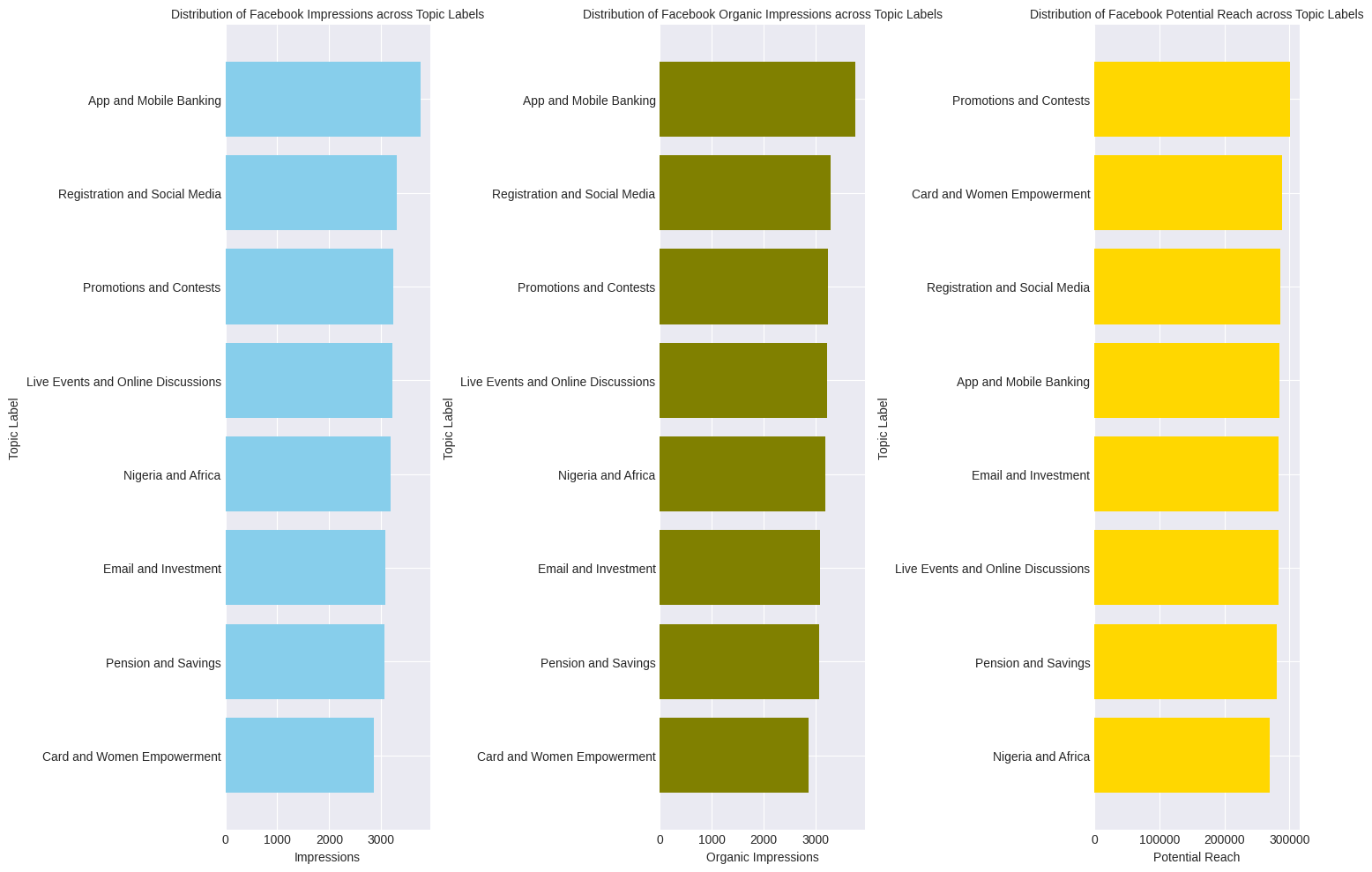
- Topic 5: **Pension and Savings:** This topic seems to be about the bank's pension and savings products, such as helping customers make their dreams come true, dialing for airtime, and visiting for more information.

- Topic 6: **Card and Women Empowerment:** This topic seems to be about the bank's card services and women empowerment initiatives, such as paying with cards, emailing for details, and supporting women's development.

- Topic 7: **Registration and Social Media:** This topic seems to be about the bank's registration process and social media presence, such as registering for events, saving for needs, and sharing top reads.



We track the average number of Impressions/Organic Impressions and Potential Reach across each of our topic labels and we observe the following



Based on the data, here are some insights that might be useful:

**App and Mobile Banking**: This topic has the highest average impressions, indicating that posts related to this topic are reaching the most people. This suggests that customers are highly engaged with content related to the bank’s app and mobile banking services.

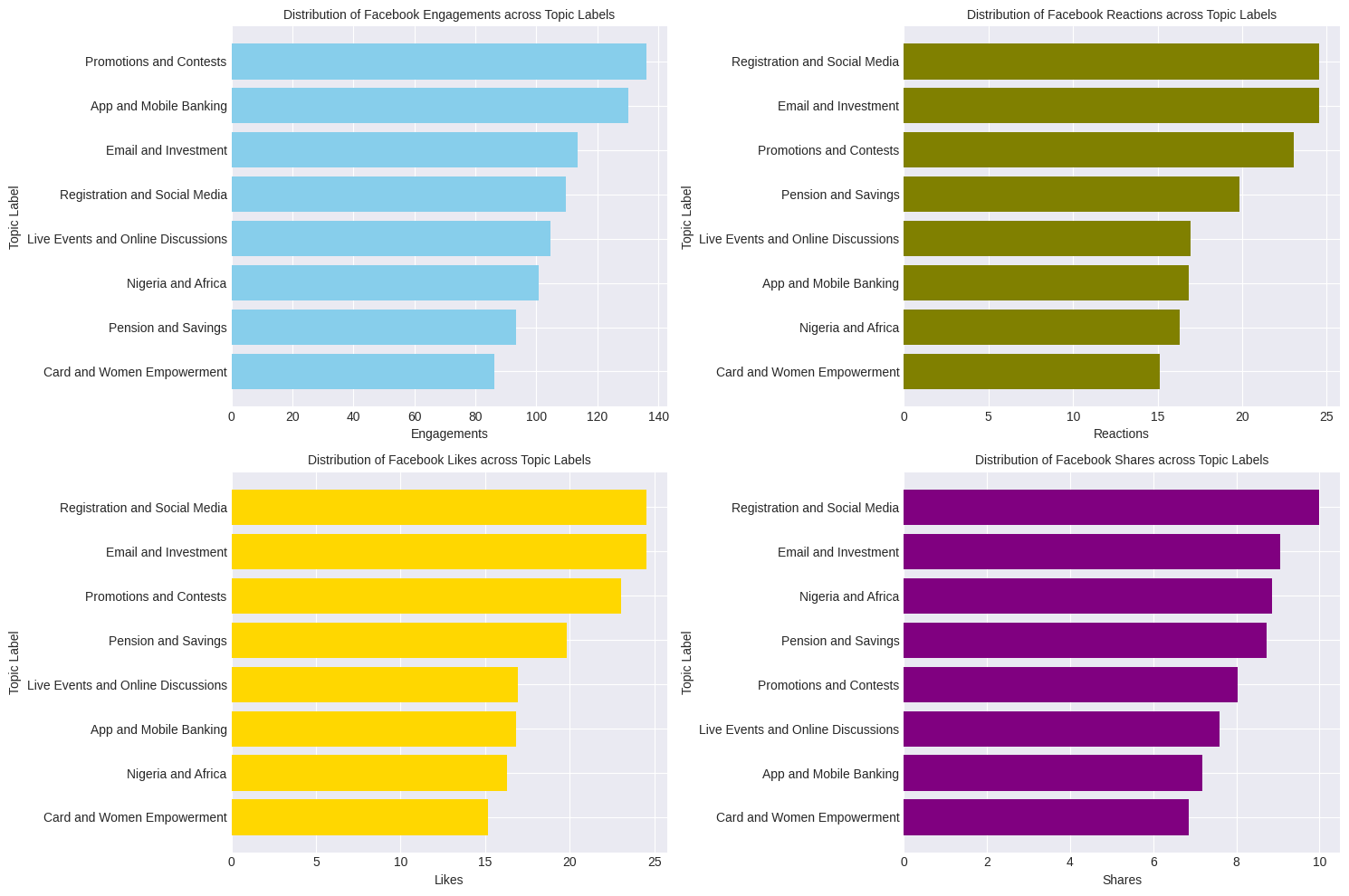
**Registration and Social Media**: Despite having slightly fewer impressions than “App and Mobile Banking”, this topic has a higher potential reach. This suggests that while these posts may not be viewed as often, they have the potential to reach a larger audience.

**Promotions and Contests**: This topic also has a high number of impressions and potential reach, indicating that customers are interested in promotional content and contests.

**Live Events and Online Discussions**: Posts on this topic have a slightly lower number of impressions but still maintain a high potential reach. This suggests that while these posts may not be viewed as often, they have the potential to engage a large audience.

**Nigeria and Africa**: This topic has fewer impressions but maintains a high potential reach, indicating that content related to the bank’s impact in Nigeria and Africa has the potential to engage a large audience.

**Email and Investment, Pension and Savings, Card and Women Empowerment**: These topics have the lowest average impressions but still maintain a high potential reach. This suggests that while these posts may not be viewed as often, they have the potential to engage a large audience.



Based on the data, here are some insights that might be useful:

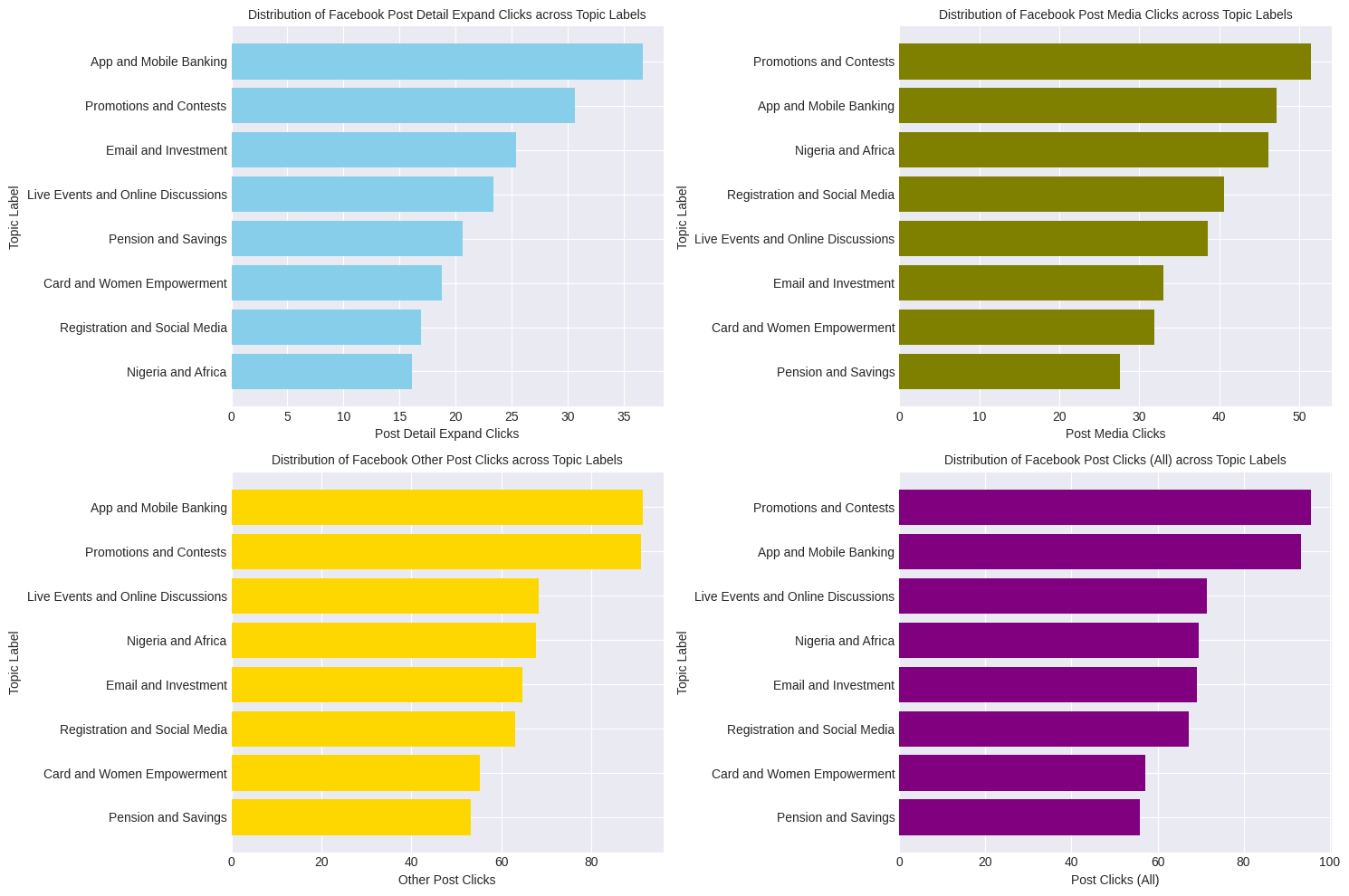
**Promotions and Contests**: This topic has the highest average engagements, reactions, likes, and shares. This suggests that customers are highly engaged with promotional content and contests.

**App and Mobile Banking**: This topic has the second highest average engagements, with a significant number of reactions, likes, and shares. This indicates that customers are actively interacting with content related to the bank’s app and mobile banking services.

**Email and Investment**: Despite having fewer engagements than “Promotions and Contests” and “App and Mobile Banking”, this topic has a higher average number of reactions, likes, and shares. This suggests that while these posts may not be viewed as often, they generate a high level of interaction when they are viewed.

**Registration and Social Media**: Similar to “Email and Investment”, this topic has fewer engagements but a higher average number of reactions, likes, and shares. This indicates that posts about registration processes and social media presence generate a high level of interaction.

Promotions and Contests lead in overall average engagements, including reactions, likes, and shares, indicating high engagement with promotional content and contests. Additionally, several other topics, such as 'App and Mobile Banking,' 'Email and Investment,' and 'Registration and Social Media,' also demonstrate above-average engagement levels. Notably, some topics may not have as high overall engagement but generate significant interaction when viewed, emphasising the importance of tailored content strategies. Finance-related topics consistently attract engagement. Interactive content, like contests, proves to be highly effective in driving interactions and expanding reach. Stakeholders should consider these insights to guide content strategies and explore growth opportunities



**Promotions and Contests and App and Mobile Banking Lead in Post Clicks:**

The "Promotions and Contests" and "App and Mobile Banking" topics have the highest average post clicks (all types combined), indicating that these topics are particularly effective in generating clicks and driving user engagement.

**Consistency in Click Metrics:**

Most topics demonstrate a consistent level of click metrics across different types, suggesting that the audience's interaction is well-distributed among these categories.

**Nigeria and Africa Topic Stands Out:**

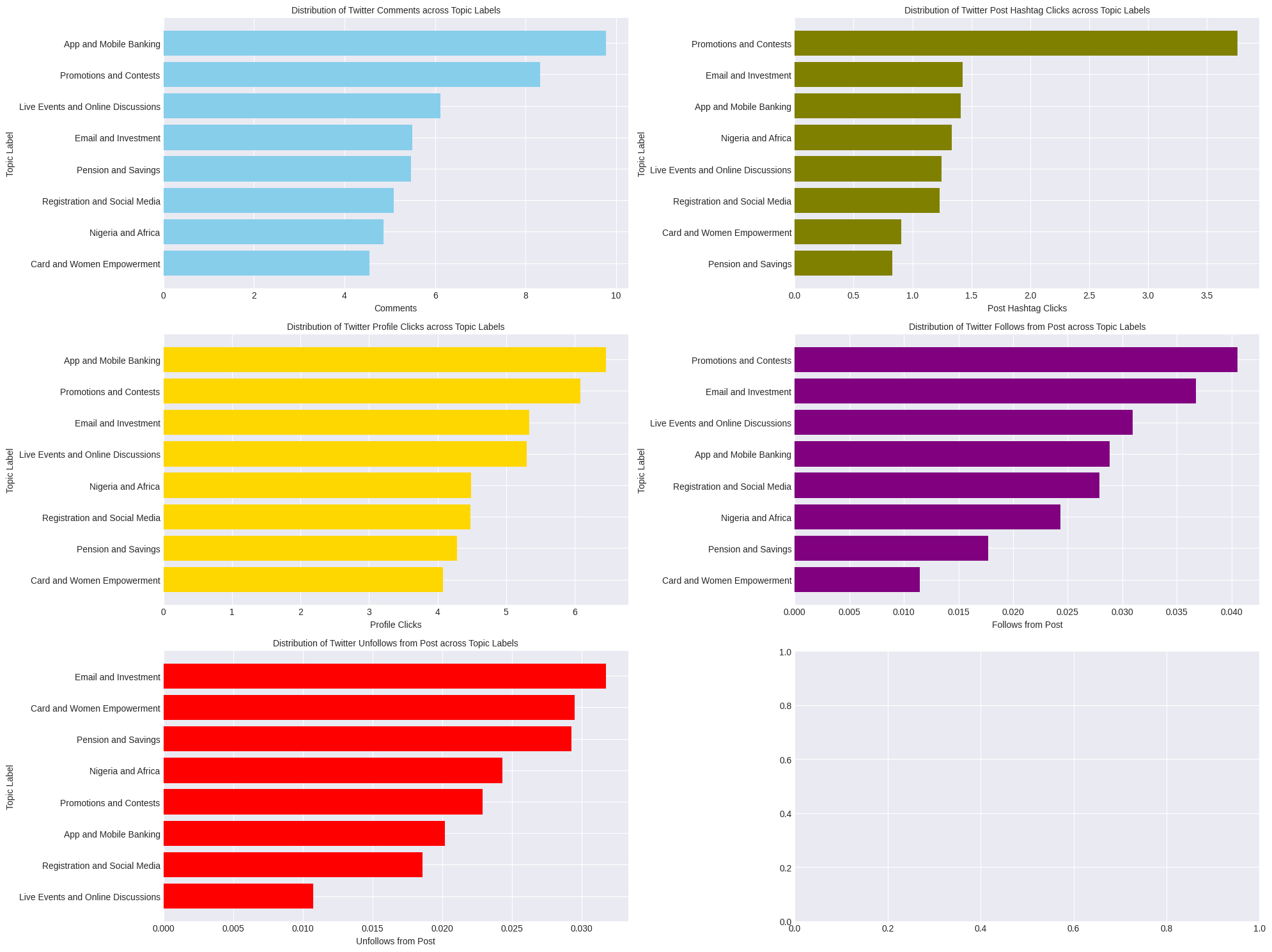
Despite having lower overall post clicks, the "Nigeria and Africa" topic has a relatively high number of "Post Media Clicks," indicating that users are engaging with multimedia content related to this topic.

**Email and Investment Topic's Click Performance:**

The "Email and Investment" topic, while not having the highest overall post clicks, shows a significant number of "Post Detail Expand Clicks," suggesting that users are actively expanding and exploring the details of content related to email and investment.

**Engagement in Live Events and Online Discussions:**

"Live Events and Online Discussions" generates a notable level of click activity, indicating that users are interested in engaging with content related to these topics, even though the overall engagement may not be the highest.



App and Mobile Banking has the highest avg count of comments and leads to the most Profile Clicks across all the topic labels, followed next by Promotions and Contents. Live Events and Discussion and Emails and Investments also generate a high amount of comments and profile clicks just below App and Mobile Banking and Promotions and Contents.

Promotions and Contents have the highest average values of Hashtag Clicks, and Follows from Post count, followed tightly by Email and Investment, and App and Banking.

Email and Investment, Card and Women Empowerment, Pensions and Saving are the leading topics for the Unfollow from Posts.

# Actionable Recommendations

Here are some possible actionable recommendations based on the insights from the data:

* **Promotions and Contests**: Continue to create and share promotional content and contests that appeal to the audience and generate high engagement. Consider offering incentives or rewards for sharing the content to increase the reach and visibility.
* **App and Mobile Banking**: Leverage the high engagement with app and mobile banking content to showcase the benefits and features of the bank’s app and mobile banking services. Encourage users to download the app, sign up for mobile banking, or use the app for transactions.
* **Email and Investment**: Increase the visibility and frequency of email and investment content to attract more clicks and views. Highlight the value proposition and differentiation of the bank’s email communication and investment services.
* **Registration and Social Media**: Optimize the registration process and social media presence to enhance user experience and satisfaction. Provide clear and easy instructions for registration, offer incentives or discounts for signing up, and create a consistent and engaging social media voice.
* **Live Events and Online Discussions**: Increase the promotion and awareness of live events and online discussions to attract more clicks and views. Use catchy headlines, compelling images, and relevant hashtags to generate interest and curiosity.
* **Nigeria and Africa**: Showcase the bank’s impact and contribution in Nigeria and Africa to build trust and loyalty among the audience. Share success stories, testimonials, or case studies that demonstrate the bank’s social responsibility and commitment.
* **Pension and Savings**: Increase the education and awareness of pension and savings products to attract more clicks and views. Provide useful tips, advice, or guidance on how to plan for retirement, save money, or achieve financial goals.
* **Card and Women Empowerment**: Increase the personalization and relevance of card and women empowerment content to boost engagement. Segment the audience based on their preferences, needs, or challenges, and tailor the content accordingly.

**Recommendations for High-Impression Topics (e.g., App and Mobile Banking, Registration and Social Media, Promotions and Contests):**

1. **Content Consistency:** Given the high impressions and potential reach, maintain a consistent content schedule for these topics to continue engaging the audience effectively.

**Recommendations for High-Engagement Topics (e.g., Promotions and Contests, App and Mobile Banking, Email and Investment):**

1. **Leverage Interactive Content:** As these topics generate the highest engagement, consider creating more interactive and participatory content, such as contests, polls, and surveys, to further boost engagement.

**Recommendations for High-Click Topics (e.g., Promotions and Contests, App and Mobile Banking, Nigeria and Africa):**

1. **Optimize Multimedia Content:** Since these topics lead in post clicks, focus on creating multimedia content like videos, infographics, and images to enhance user engagement and click-through rates.

**Recommendations for Topics with High Comments and Profile Clicks (e.g., App and Mobile Banking, Promotions and Contests, Live Events and Online Discussions):**

1. **Engagement Strategies:** Continue to encourage user interactions by responding promptly to comments and further promoting discussions. Encourage users to click on profiles for more information.

**Recommendations for Topics with High Hashtag Clicks and Follows from Post (e.g., Promotions and Contests, Email and Investment, App and Mobile Banking):**

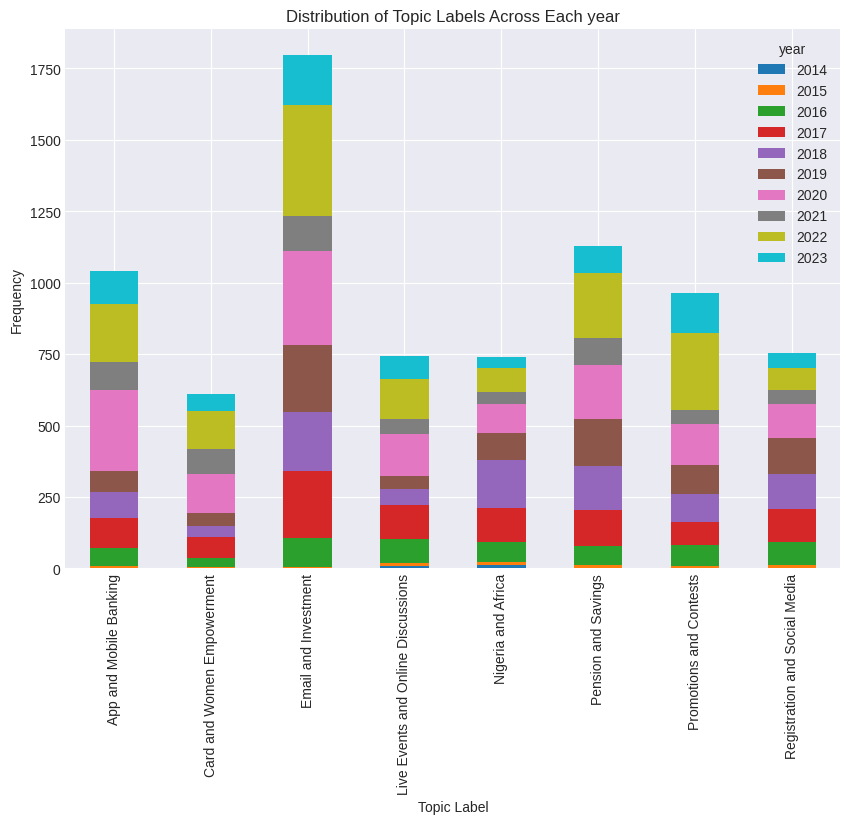
1. **Optimize Hashtag Usage:** Explore trending and relevant hashtags within these topics to increase discoverability and encourage users to follow your content. This can also expand your reach.

**Recommendations for Topics with Unfollows from Posts (e.g., Email and Investment, Card and Women Empowerment, Pensions and Savings):**

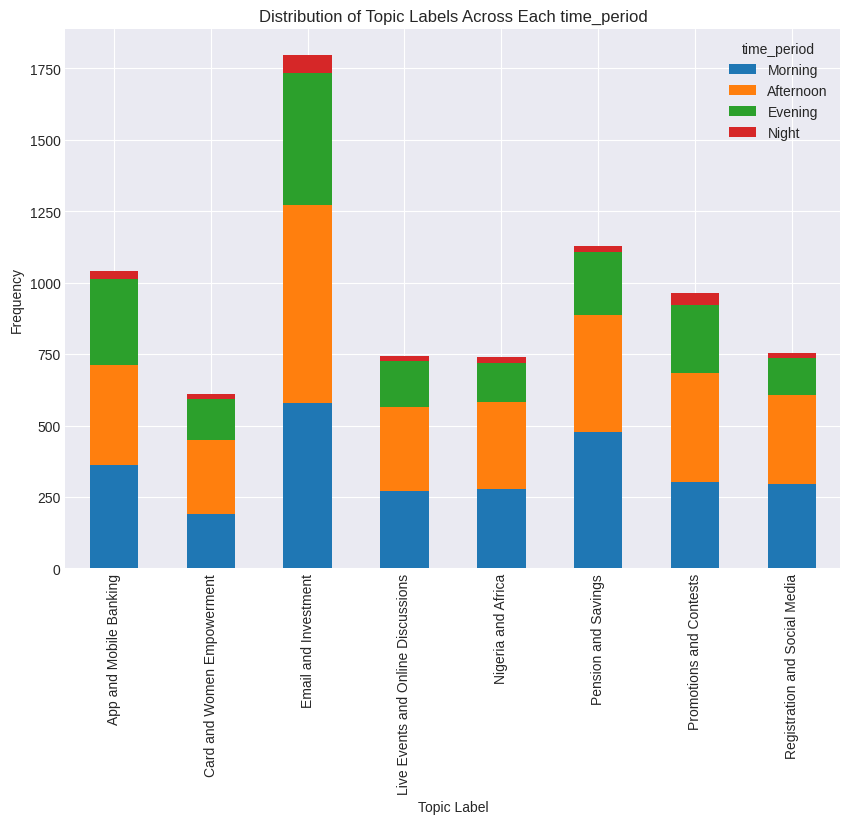
1. **Content Relevance:** Assess the content within these topics to ensure it aligns with the interests and expectations of your audience. Make adjustments to improve relevancy and reduce unfollows.

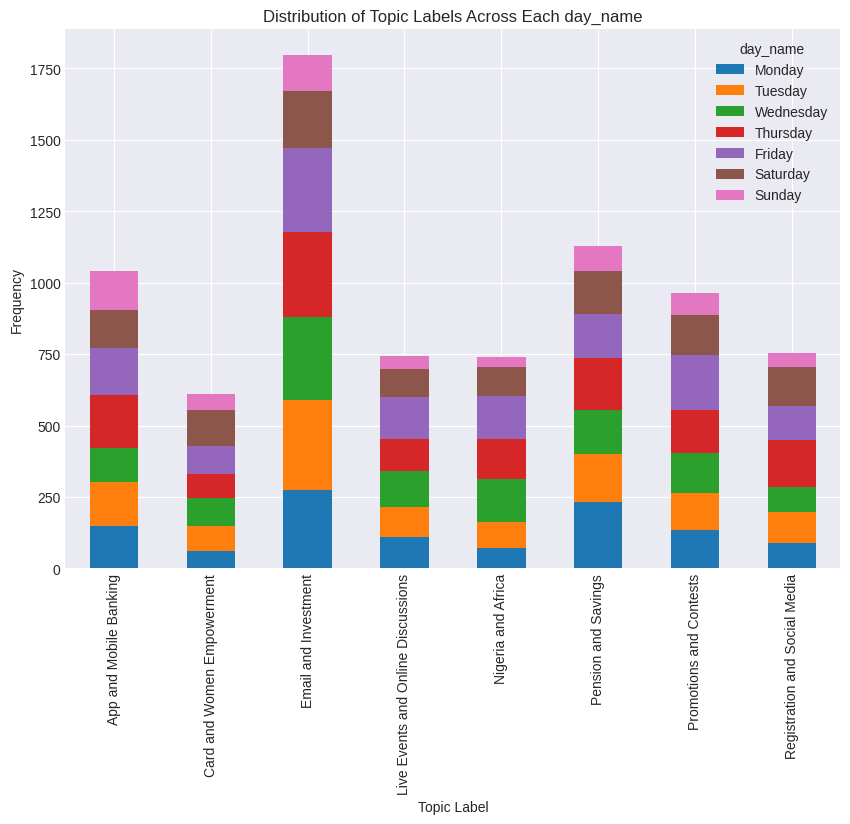
**Overall Content Strategy Recommendations:**

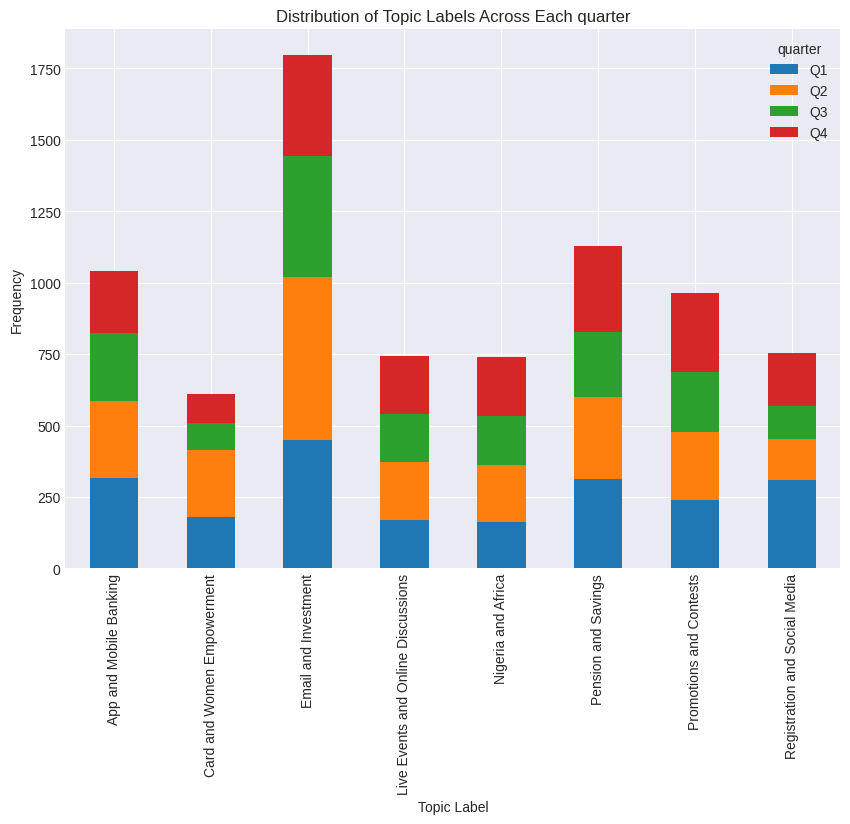
1. **Diversify Content Types:** Incorporate a variety of content types, including articles, videos, images, and infographics, to cater to different audience preferences.
2. **Audience Segmentation:** Segment the audience based on their interests and behavior, then tailor content accordingly to maximize engagement.
3. **Data-Driven Decisions:** Continuously monitor performance metrics and adjust content strategies based on the data to optimize results.
4. **A/B Testing:** Experiment with different content approaches and formats to identify what resonates most with the audience.
5. **Collaborate and Share Success Stories:** Share insights and success stories across teams to encourage collaboration and learning from the best-performing topics.

Below, we can see, the distribution of Topic labels across the years, We see the highest Topic labels as Email and Investments, pensions and savings, App and Mobile Banking  
  


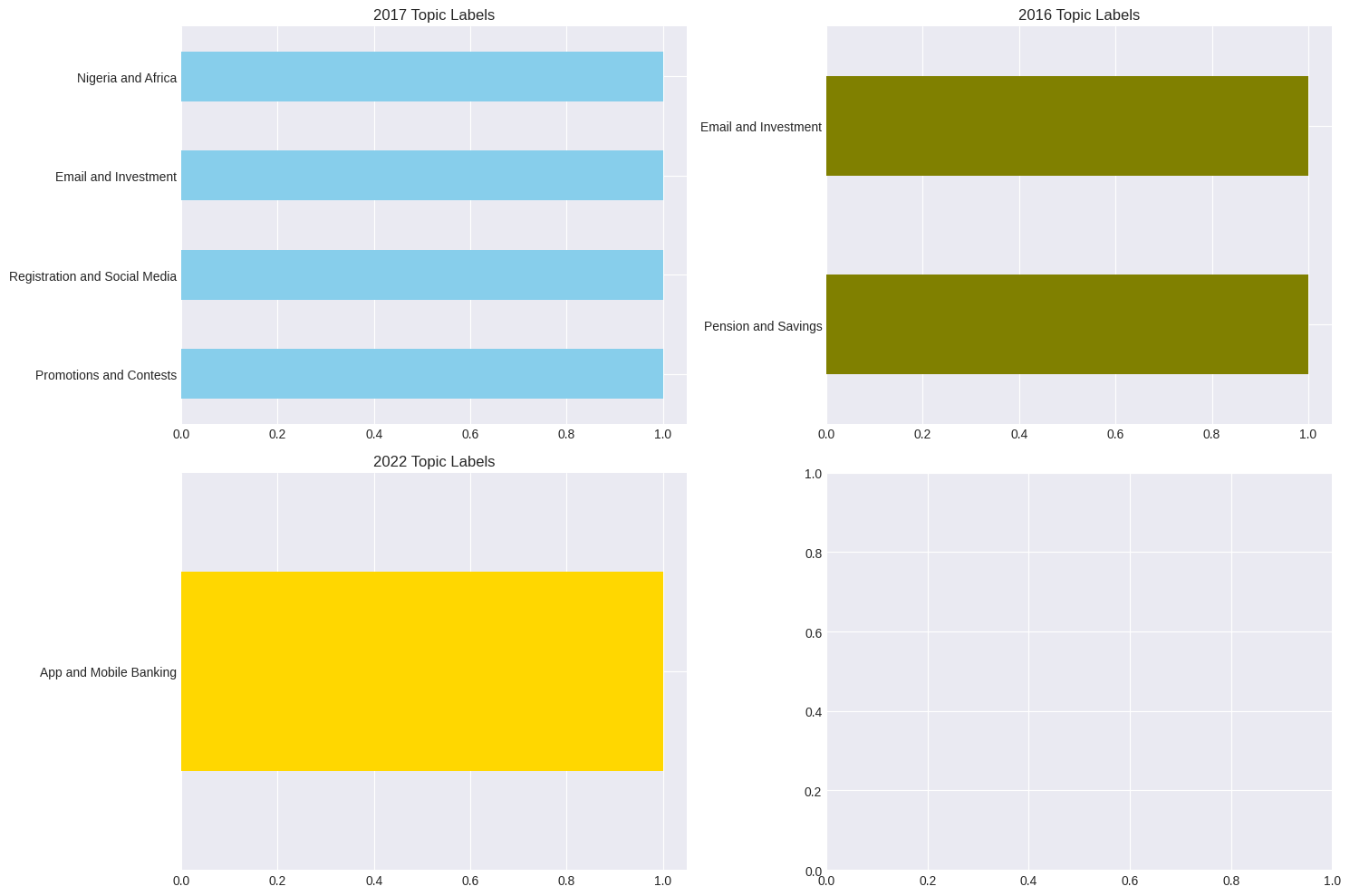
Below, we can also see the same across the Day Periods, We see the highest Topic labels as Email and Investments, pensions and savings, App and Mobile Banking







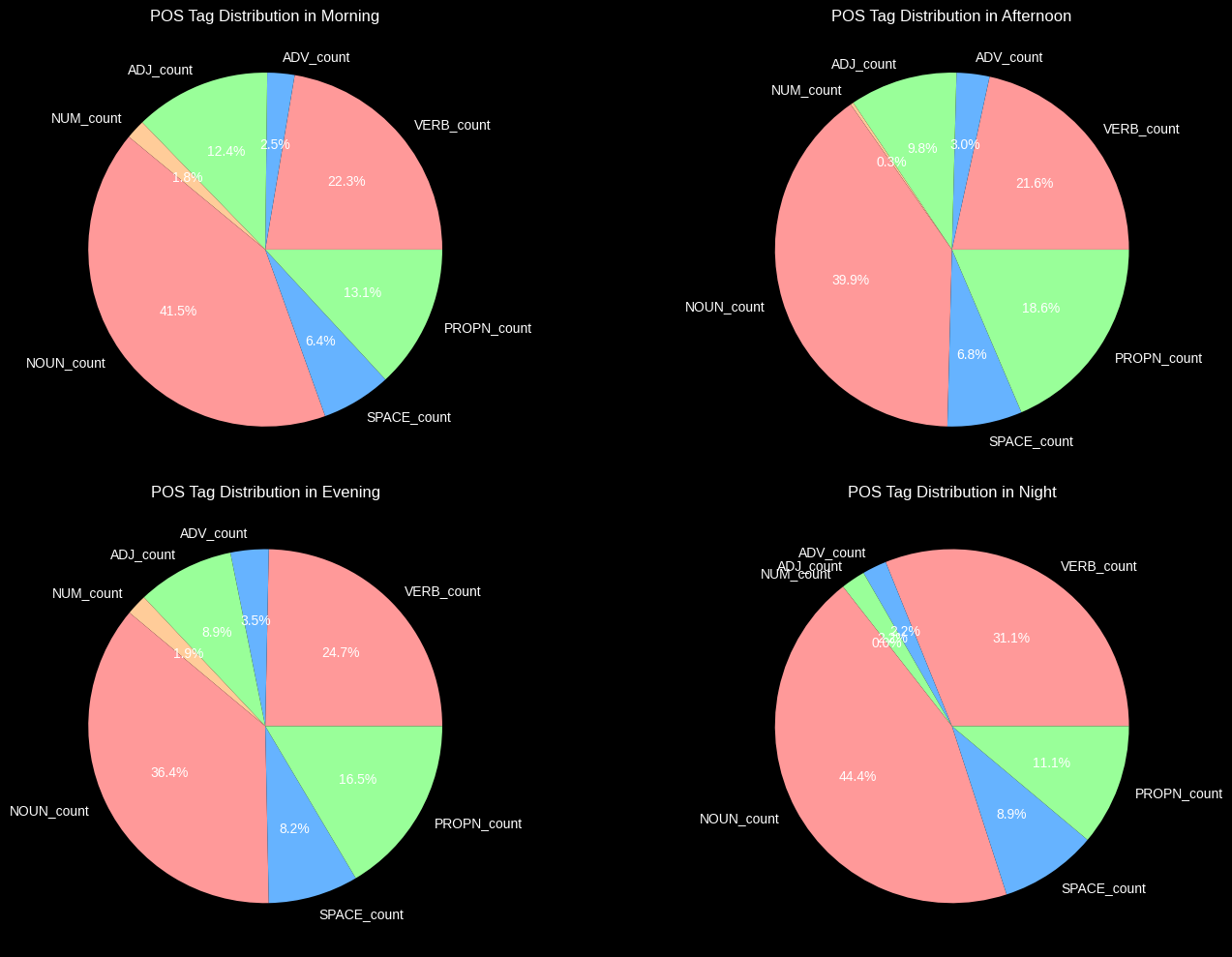
Distribution of Posts in the Top 1% Impression  
  
  
Our Insights are finally coming together as we can see and have a clearer explanation now, Posts that were in the top 1% Impression in the year 2017, we can see compared to the others 2017 had the highest number of topics being discussed. When our posts are distributed it helps to generate more insights.





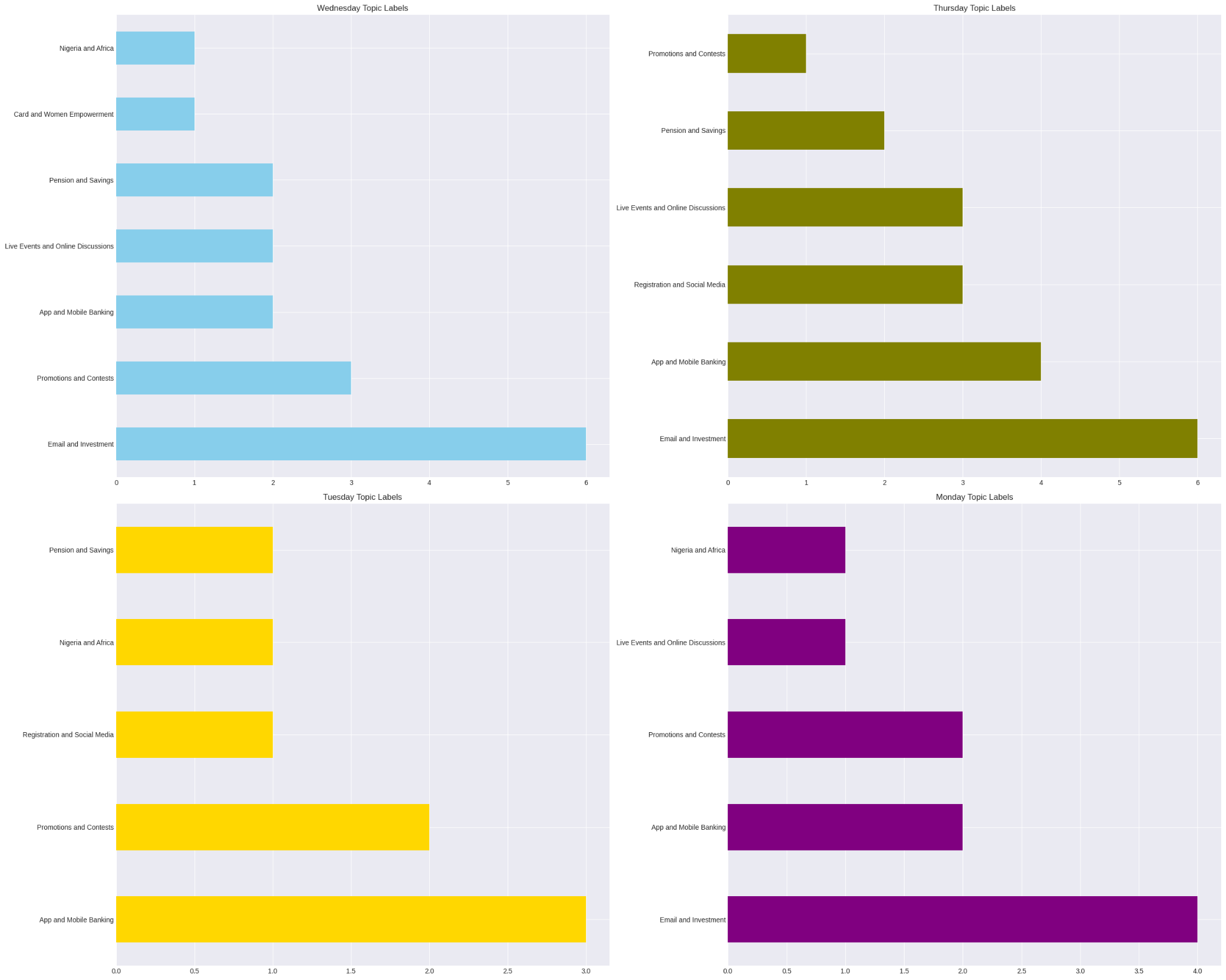
This also confirms our theory showing to us that the Morning and Afternoon Time Periods, Posts that were in the top 1% Impression during the Morning, Afternoon and Evening Time Periods, we can also see compared to the others Morning, Afternoon and Evening had the highest number of topics being discussed. When our posts are distributed it helps to generate more Impressions, so we had Important Dates, Organizations mentioned in our Posts which gathered more Impressions.





The Same theory can be applied to Top Posts in a Day with the Highest Impressions, we notice





# Conclusions

Twitter offers us good Impressions and Reach to our Users