Insights from Social Media Analytics

Nanke Williams
October 2023

Outline

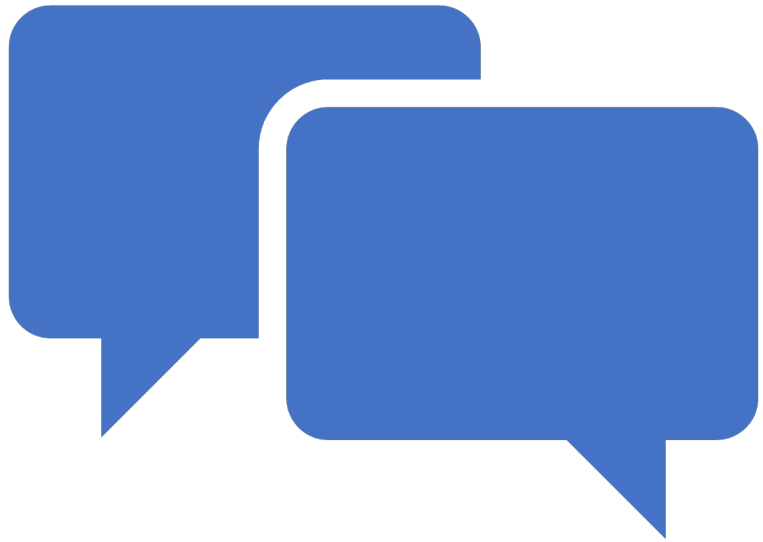
- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Data Collection
  - Data Wrangling
  - Exploratory Data Analysis with Visualization
  - Interactive Visual Analytics with Tableau
  - Statistical Analysis
- Summary of all results
  - Findings from Exploratory Data Analysis
  - Findings from Interactive Analytics
  - Findings from Statistical Analytics

Introduction

- Project background and context
    - Playhouse Communication is one of Nigeria's leading digital marketing agencies. They combine design and media planning with cutting-edge tech solutions to reimagine what marketing is all about.
    - The goal of the project is to decode a treasure trove of social media data for one of their high-profile clients and transform it into game-changing insights.
- Project Deliverables
    - We will determine responses to Key Questions derived during Exploratory Analysis.
    - We will use Statistical Analysis to test whether the hypothesis raised from charts are significant and these relationships truly exist.
    - Make recommendations based on insights generated from Exploratory and Statistical Analysis.



Section 1

Methodology

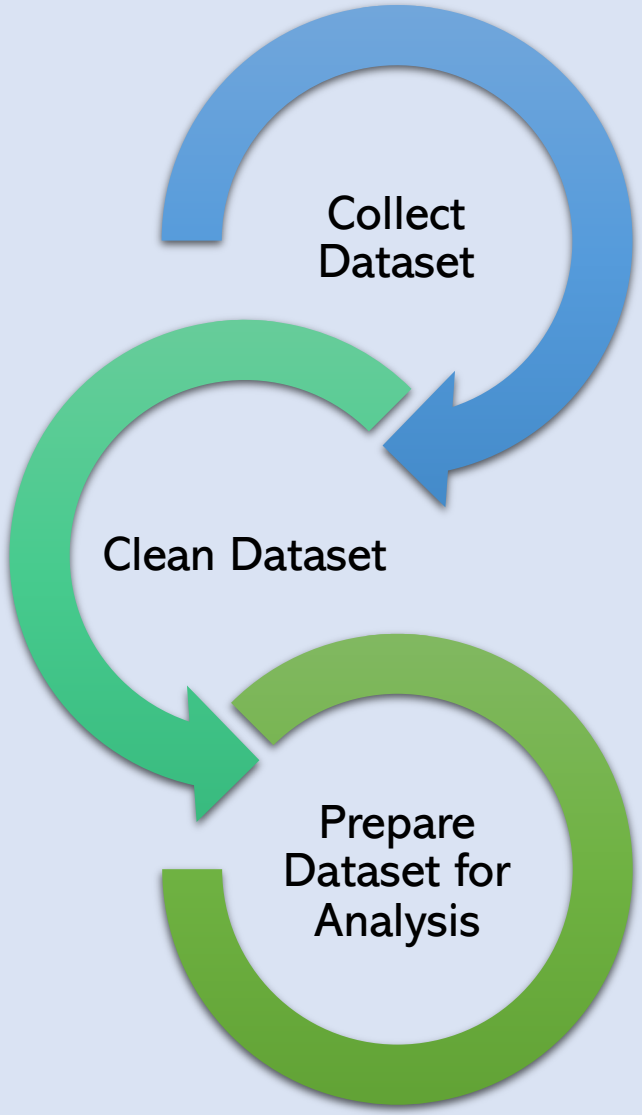
Methodology

- Executive Summary
- Data preparation methodology.
- Perform exploratory data analysis (EDA) with visualization
- Perform interactive visual analytics using Tableau
- Perform statistical analysis

Data Preparation

- The dataset was collected from four social media websites: Facebook, Instagram, Twitter & LinkedIn.
- Each social media dataset had 147 unique features which allowed for merging them into a single dataset.
- The dataset was cleaned and analyzed in a jupyter notebook environment using Python.

Please see the jupyter notebook for more information.



11 rows | 147 rows × 1 columns pd.DataFrame

	9680
Date	4/9/2022 4:00 pm
Post ID	1512807586038493187
Network	Twitter
Post Type	Tweet
Content Type	Photo
...	...
Card Impressions	NaN
Card Teaser Impressions	NaN
Card Teaser Clicks	NaN
Poll Votes	NaN
Tags	Stanbic IBTC DiSEP

Please see the jupyter notebook for more information.

Data Preparation – Sample Record

- Each record had a feature stating:
  - The social media network they originated from.
  - The date and time the Post was sent.
  - The type of content they were, the text of the Post.
  - The name of the sender of the post.
  - General and specific social media metrics collected by the network.

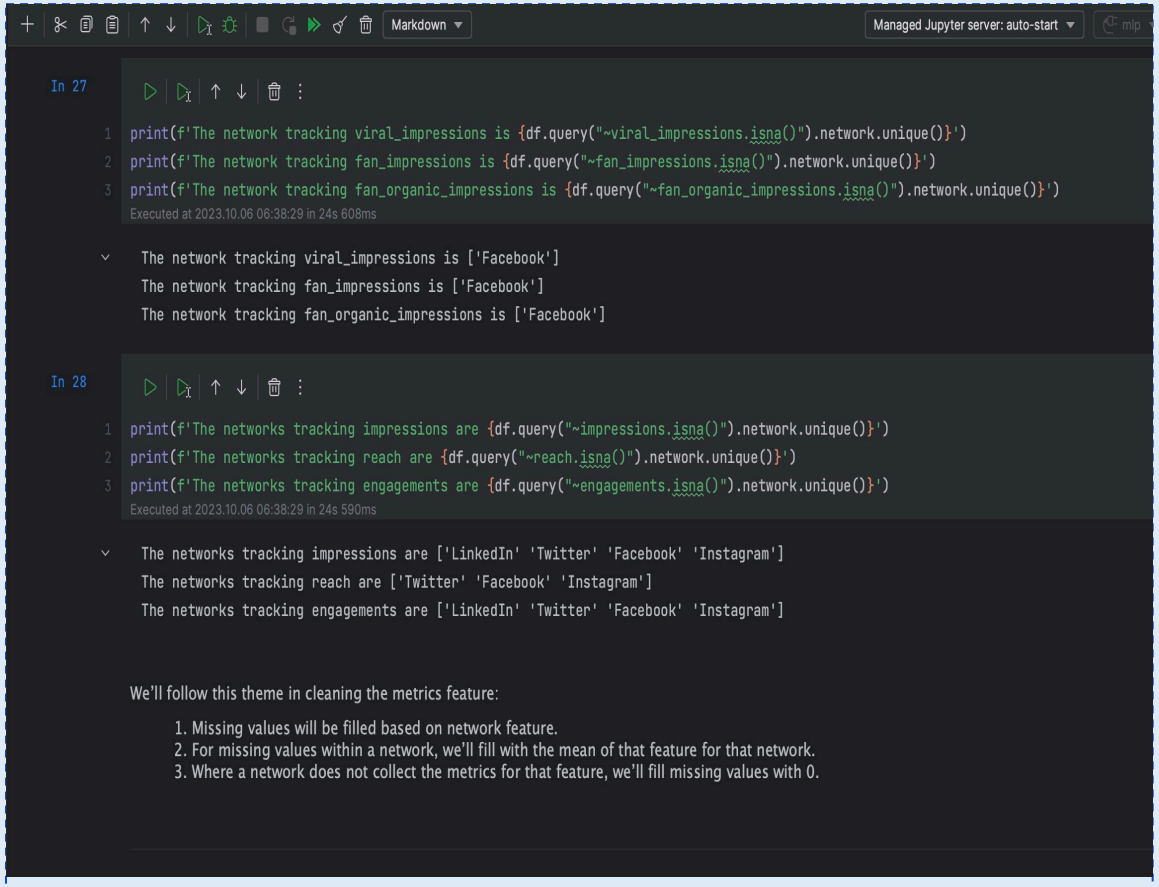
Data Preparation – Data Cleaning Strategy

- Empty rows & duplicate rows were removed.
- The date feature was cleaned and converted to datetime type.
- New columns for hashtags, post word counts, character counts, and sentiments were added.
- Each column was converted to its appropriate datatype
- Datatypes to be changed
- More than 30 columns were removed because they contained no data.
- The data in ‘Potential Reach’ which represented ‘Reach’ data from Twitter was moved to the ‘Reach’ column.
- Impute missing values for numeric and categorical features.
- Set Date as Index.

Please see the jupyter notebook for more information.

Data Preparation – Data Imputation Strategy

- There are certain metrics that are network-specific, i.e., tracked and collected by only certain networks.
- Our data imputation strategy for missing values is:
  - Where the metric is collected by a social media network, missing values will be imputed using the average of that metric filtered by the relevant social network.
  - Where the metric is not collected by a social network, missing values will be imputed using zero, representing that the metric is not collected by the social media network.
  - For categorical or text-based features, missing values were imputed using text stating that no value was provided for the record.



Please see the jupyter notebook for more information.

Data Preparation – Summary of the Dataset

```
********************************
Instagram    10000
Facebook      9803
Twitter       8529
LinkedIn      7760
Name: Network, dtype: int64
********************************
```

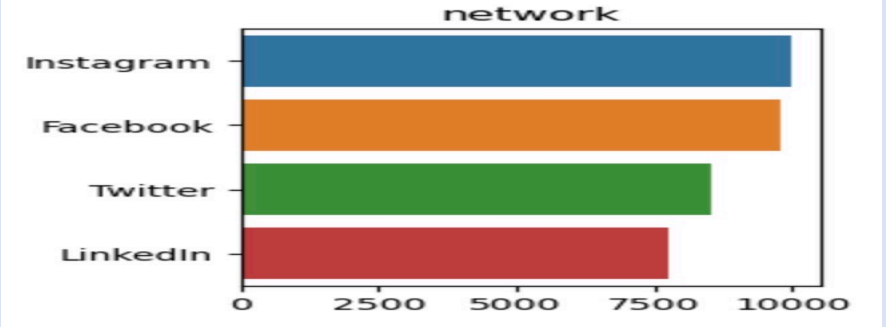
Please see the jupyter notebook for more information.

- The dataset has 36,092 rows.
- There are 3 categorical features, 4 text-based features and 105 numeric features.
- There are 4 distinct networks, 7 distinct content types and 12 distinct senders in the dataset.
- The dataset has date records spanning the 11-year period ranging from 2013 to 2023

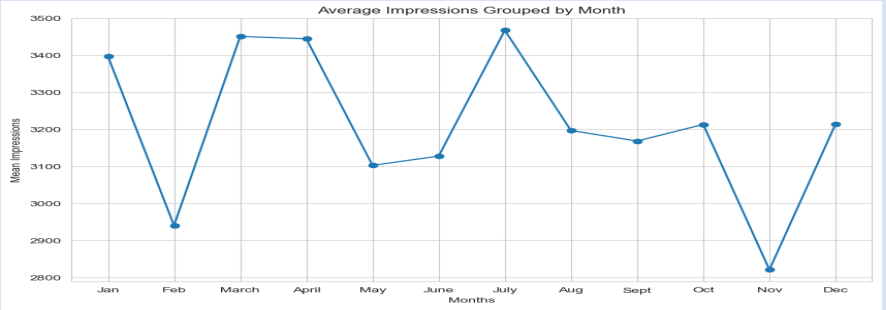
Exploratory Data Analysis with Visualization

- We visualized the relationship between a few variables in the dataset to determine:
  - If variables are correlated.
  - How categorical variables are distributed over numeric metrics.
  - What trends over time are evident in variables
- We visualized variable distributions to make assumptions for statistical analyses.
- We used word clouds to visualize hashtags to understand what posts are talking about.

- Example - Bar Chart Spread of Social Media Networks



- Example – Line Chart of Monthly Impressions



Please see the jupyter notebook for more information.

Exploratory Data Analysis with Visualization

- We conducted exploratory data analysis to determine:
  - What posts are talking about.
  - What kind of posts are popular.
  - What are the most engaging types of posts.
  - Which platform yields the highest engagement for the client.
  - What are the peak times for user engagement.
  - What are the noticeable trends over time (e.g., increasing likes, decreasing shares).
  - What are the sentiments of posts and how does it affect engagement.

Please see the jupyter notebook for more information.

Interactive Dashboard

- We built an interactive dashboard with Tableau
- We plotted pie charts and bar charts showing the spread of generally tracked metrics segmented by the categorical variables.
- We plotted line charts to track trends in engagement metrics.

Please see the Dashboard for more information.

Statistical Analysis

- We conducted statistical analysis to test the following hypotheses:
  - Whether the relationships between features are statistically significant.
  - Whether the strength and direction of the relationship between variables is statistically significant.
  - Whether there is a significant difference between 2 groups or variables.

Plot the data

Determine the hypothesis

Test the hypothesis

Make conclusions.

Please see the [jupyter notebook]() for more information.

Insight Generation

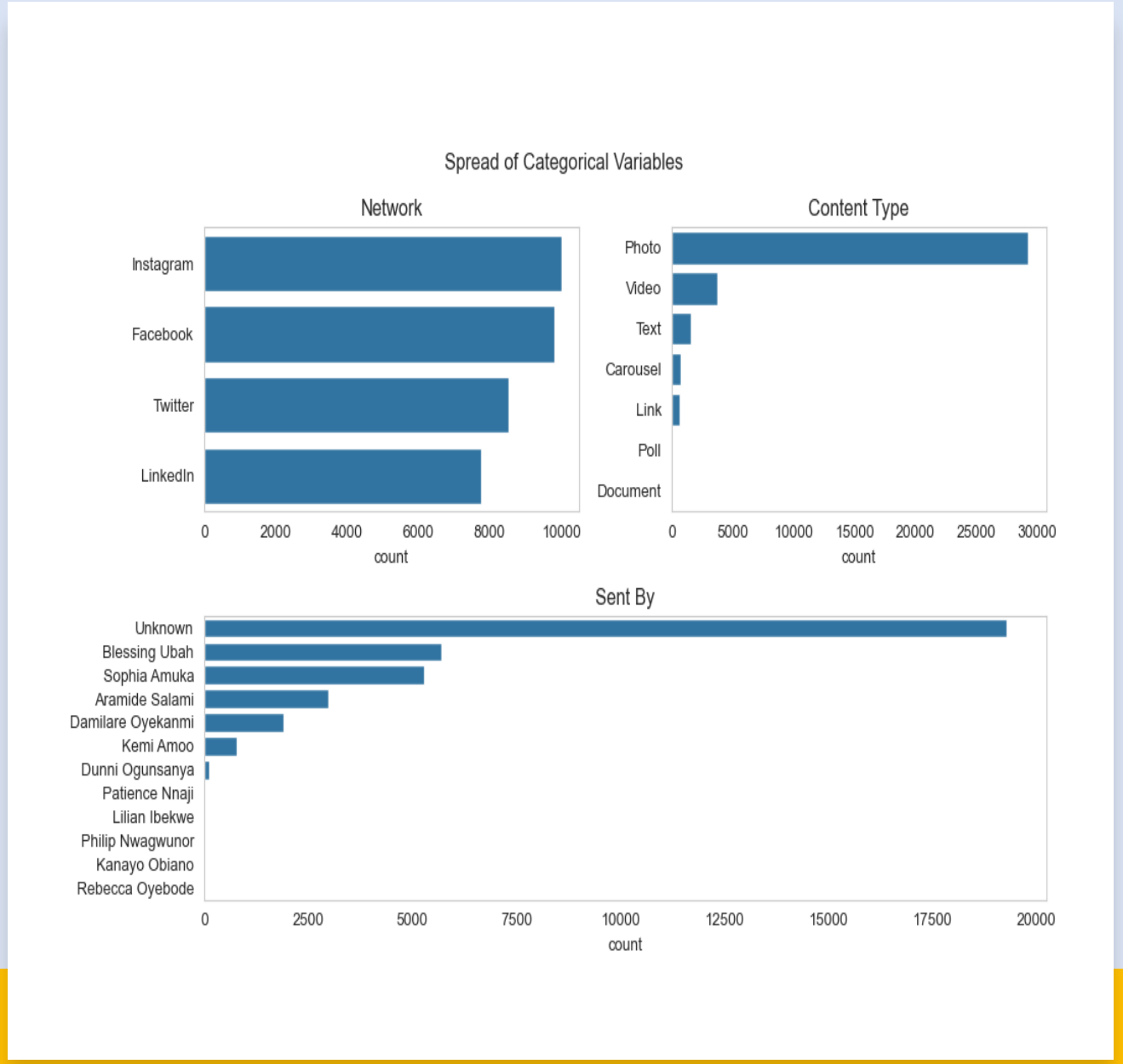
- Insights from Exploratory Analysis
- Insights from Statistical Analysis
- Insights summarized by Interactive Dashboard
- Insights from Predictive Analysis

Section 2

Insights Drawn From Exploratory Analysis

Categorical Features

- There are 3 categorical variables: Network, Content Type & Sent by.
- From Network, Instagram has the highest number of records in the dataset and LinkedIn has the least.
- From Content Type, Photo is the feature with the most number of records and Document is the feature with the least.
- Most posts are sent by an unknown sender, but from known senders, Blessing Ubah has the highest number of posts.



Please see the jupyter notebook for more information.

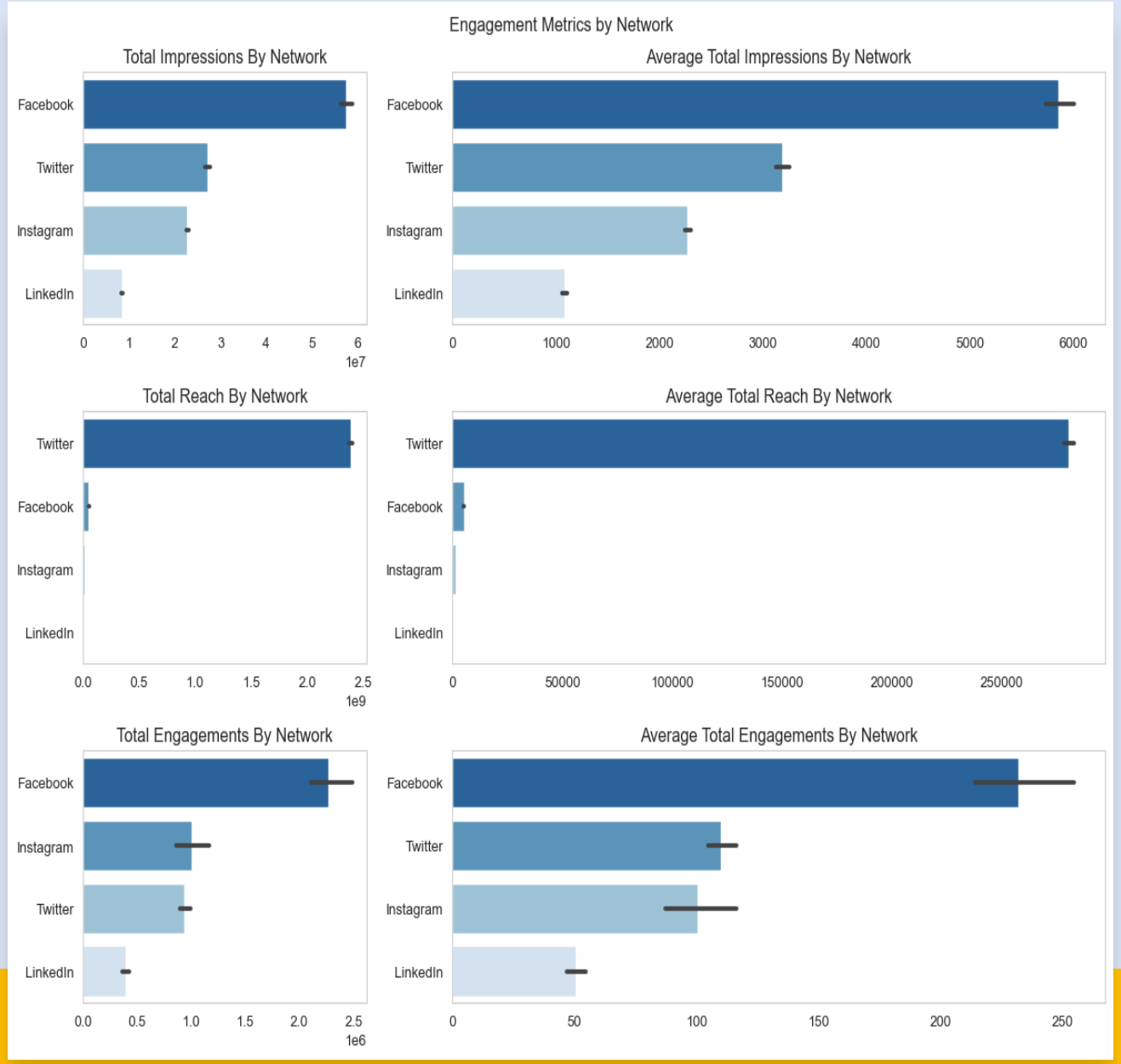
Categorical Features - Network

- The 3 main engagement metrics are Impressions, Reach & Engagements.
- Impressions is the total number of times a piece of content is displayed on a user's screen.
- Reach refers to the unique number of users who have seen a piece of content at least once during a specific reporting.
- Engagements encompass various types of interactions that users have with a piece of social media content.

Please see the jupyter notebook for more information.

Categorical Features - Network

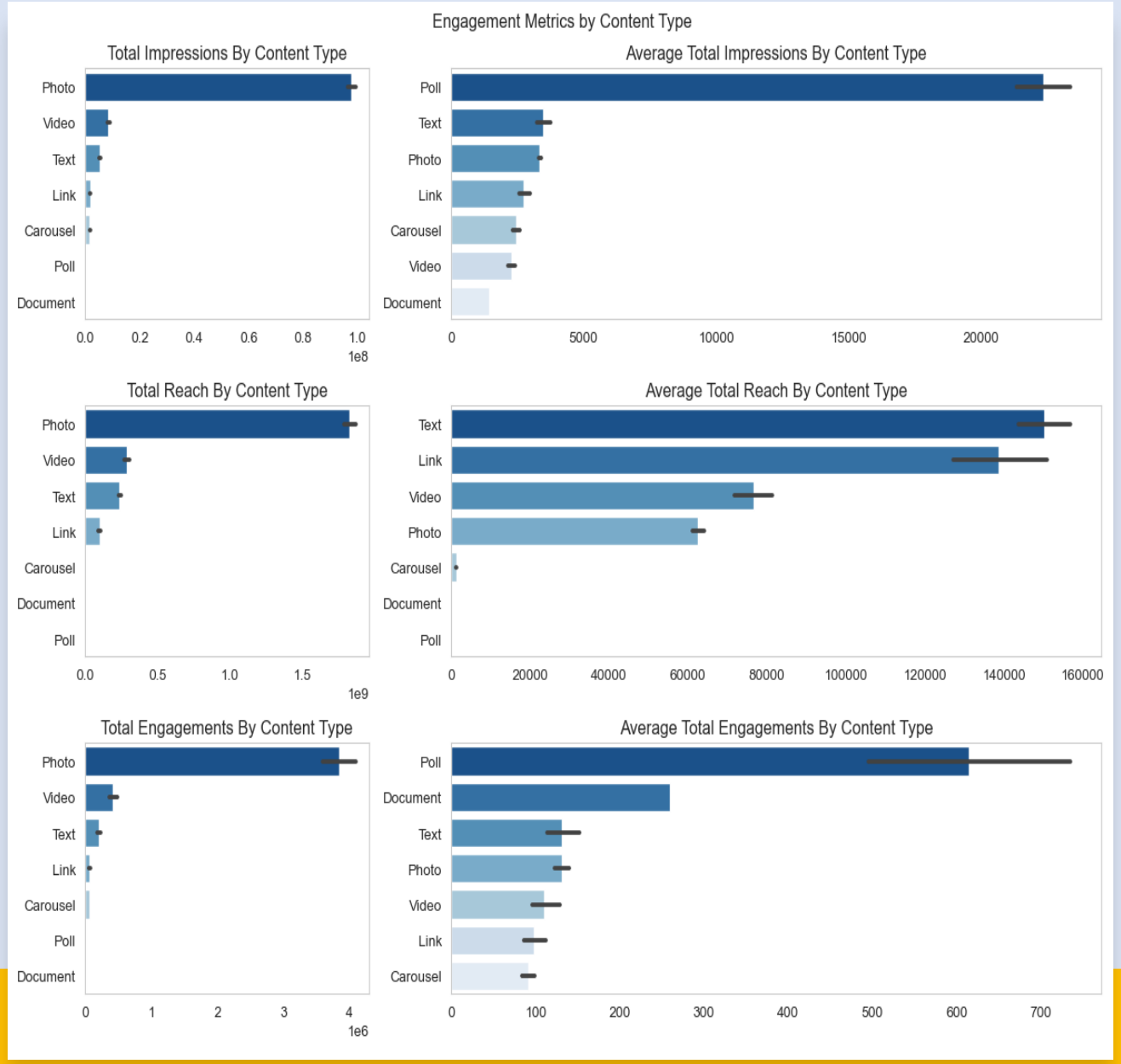
- Facebook is the network with the highest number of impressions, nearly twice the number of the second highest, Twitter
- Twitter has the highest Reach. This is likely because Twitter track Reach as Potential Reach, which is likely wider than actual reach.
- LinkedIn does not track Reach or any equivalent metric.
- Facebook has the highest number of engagements.



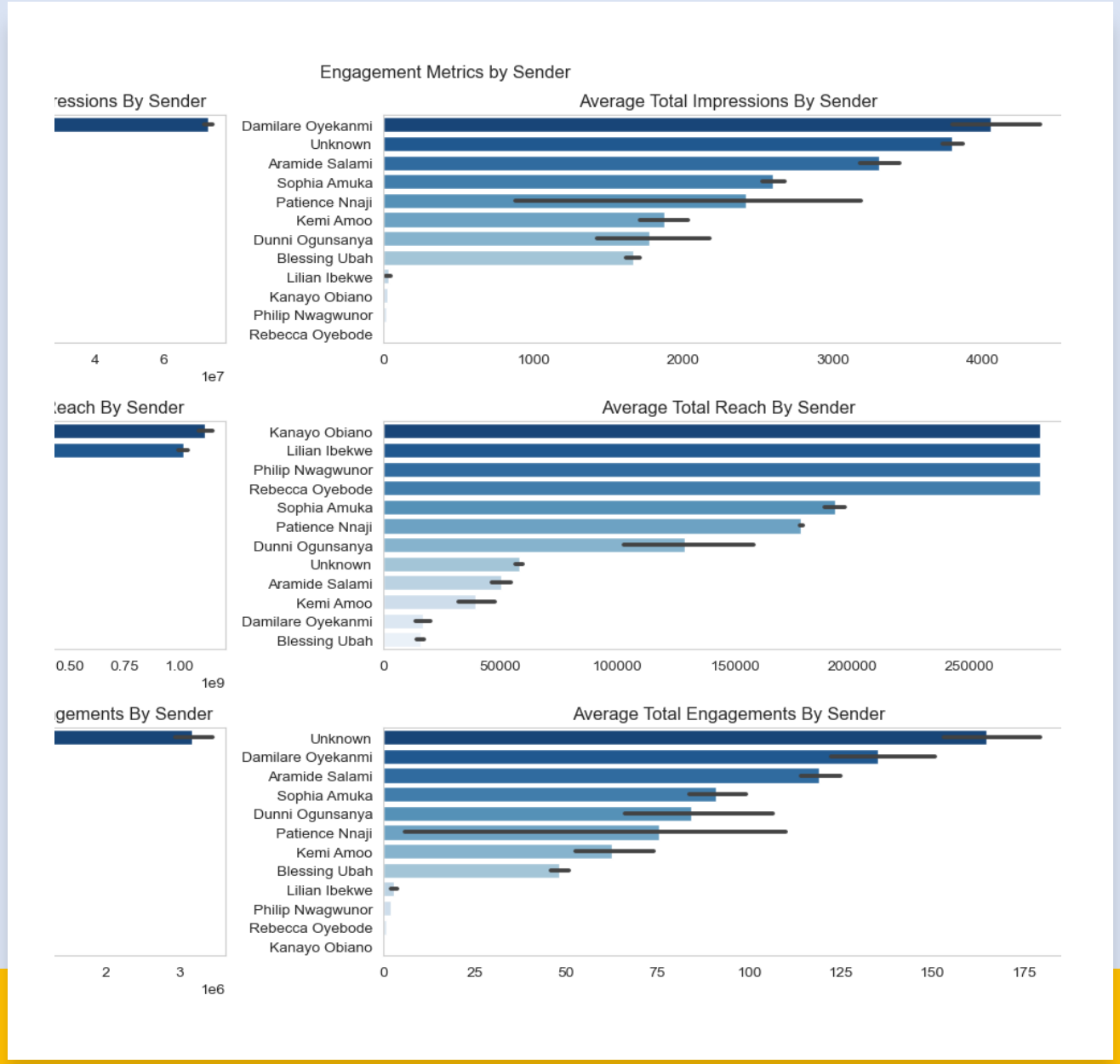
Please see the jupyter notebook for more information.

Categorical Features – Content Type

- Photo posts have the highest total engagement metrics.
- On average, Polls have the highest impressions and engagements. However, there are only 2 poll records in the dataset.
- On average, text posts have the highest reach.



Please see the jupyter notebook for more information.



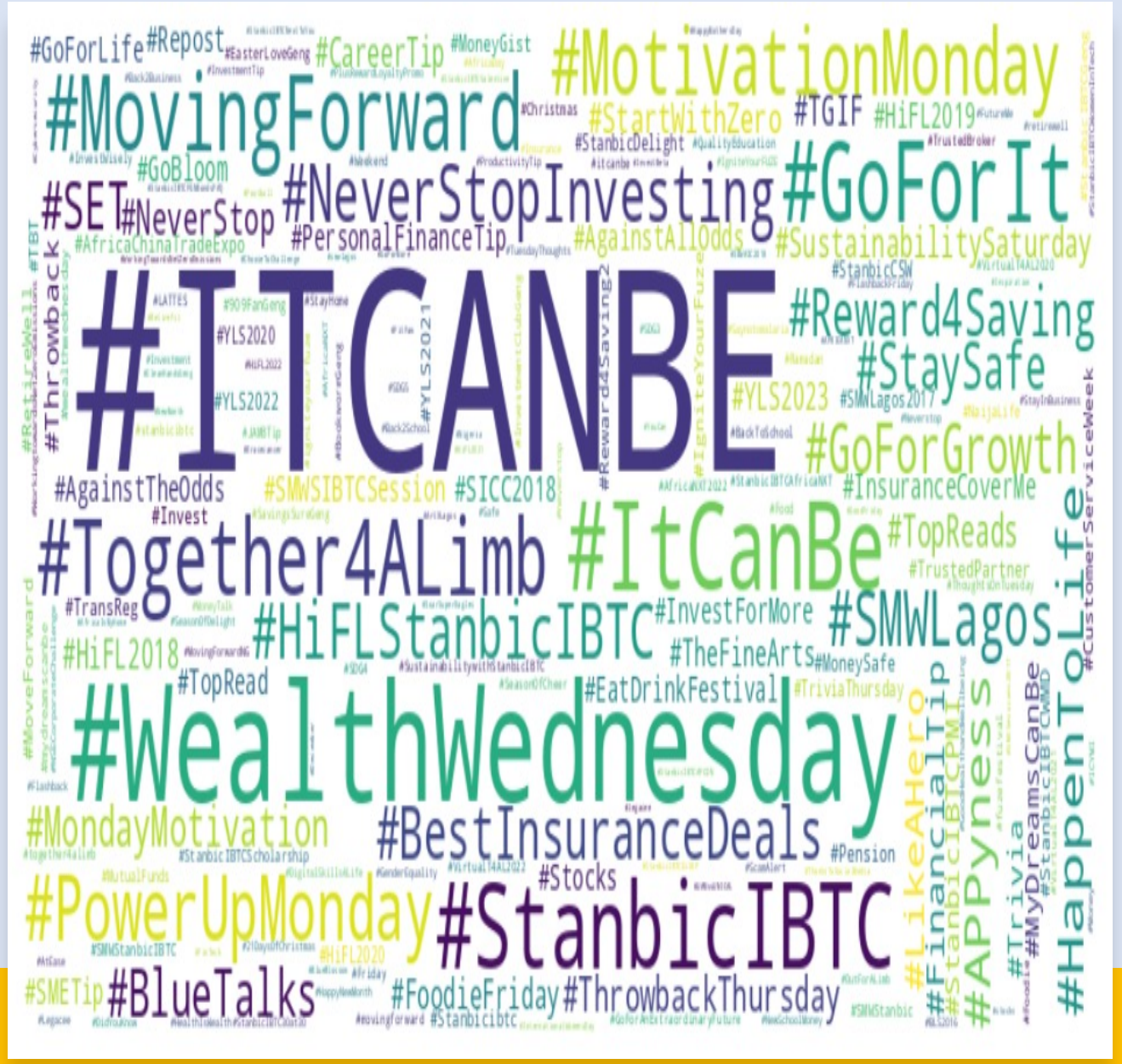
Categorical Features – Sender

- The highest total values for all metrics is held by the Unknown poster.
- On average, the highest impressions are from posts sent by Damilare Oyekanmi.
- The average highest reaches are from posts sent by Kanayo Obiano, Lilian Ibekwe, Philip Nwagwunor and Rebecca Oyebode.
- The posts with the highest engagements on average are from the unknown poster, with Damilare Oyekanmi coming in second.

Please see the jupyter notebook for more information.

Text-Based Features

- There are 4 text-based features: post, linked content, tags and hashtags.
- We used hashtags to determine what subjects the posts are talking about.
- The most frequent hashtag is #ITCANBE



Please see the jupyter notebook for more information.

Numerical Features – Social Media Metrics

- There are 9 types of numeric features / metrics collected in the dataset:
  - General metrics (collected by at least 3 of the 4 networks)
  - Impression metrics
  - Reach Metrics
  - Engagement Metrics
  - Reaction metrics
  - Activity metrics
  - Click metrics
  - User-based metrics
  - Video metrics

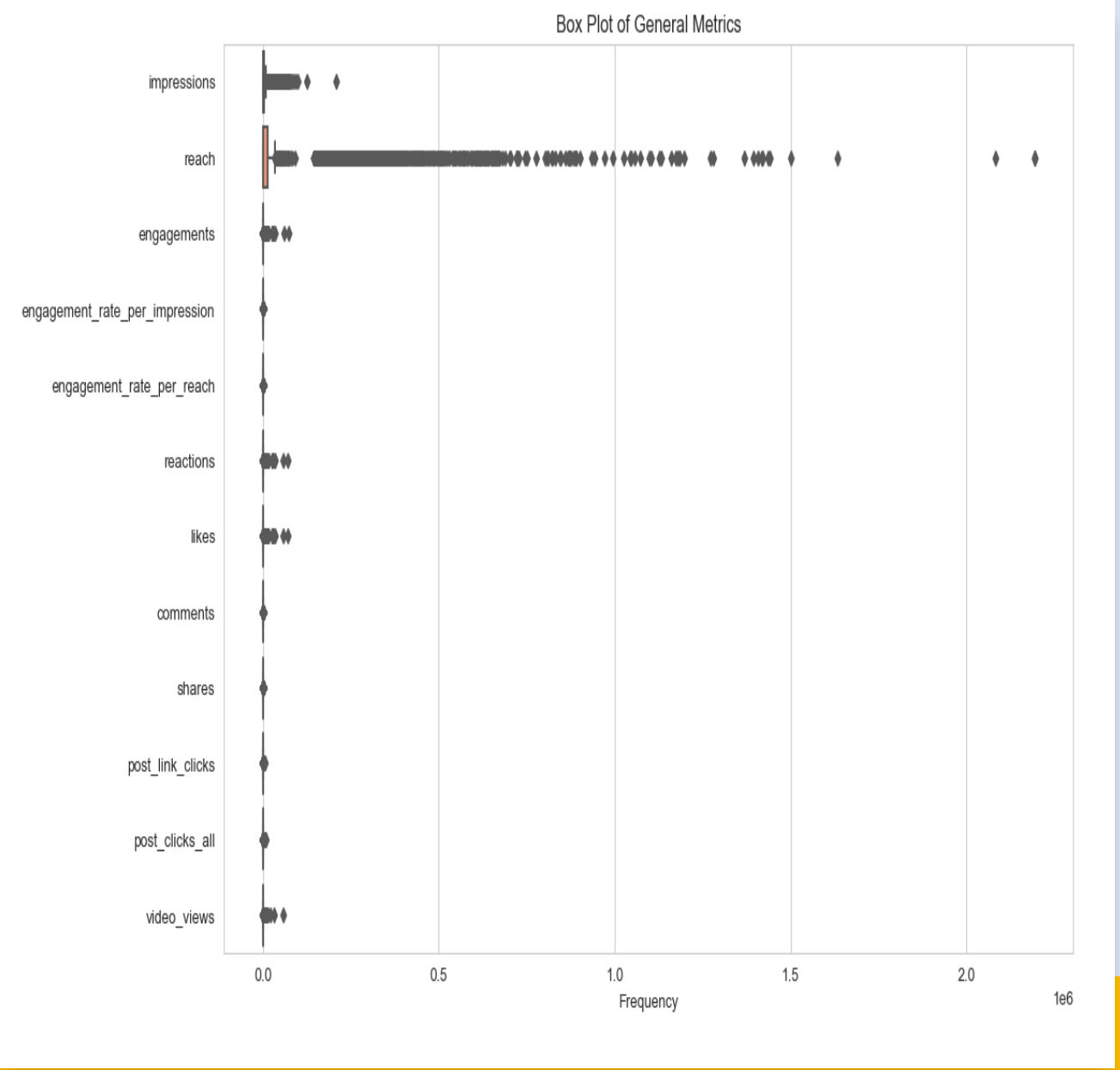
Please see the jupyter notebook for more information.

General Metrics

There are 12 general metrics:

- Impressions
- Reach
- Engagements
- Engagement rate per impression
- Engagement rate per reach
- Reactions
- Likes
- Comments
- Shares
- Link clicks
- All clicks
- Video views

This classification will be used for the statistical analysis section.



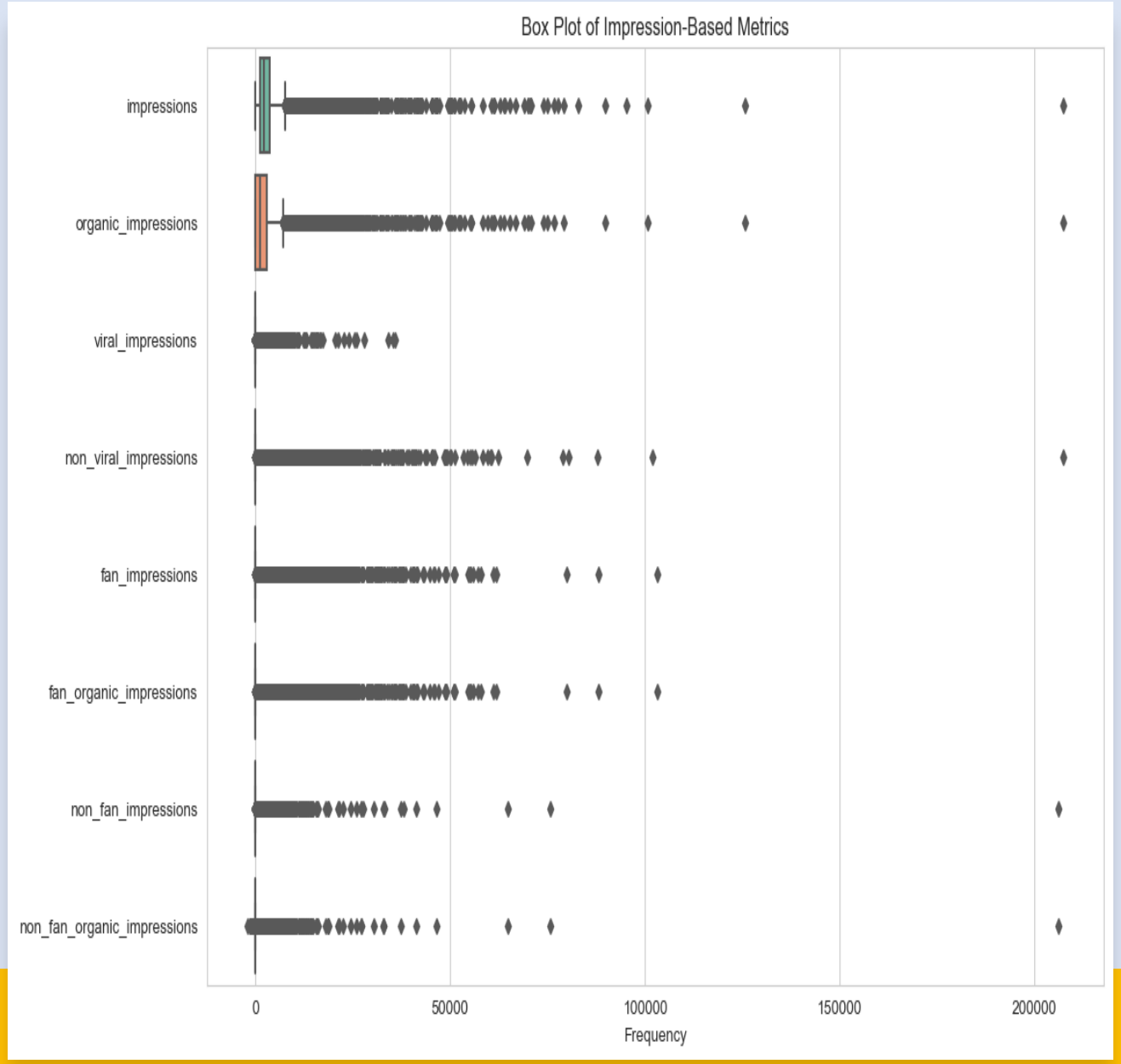
Please see the jupyter notebook for more information.

Impression Metrics

There are 8 impression metrics:

- Impressions
- Organic impressions
- Viral impressions
- Non-viral impressions
- Fan impressions
- Fan organic impressions
- Non-fan impressions
- Non-fan organic impressions

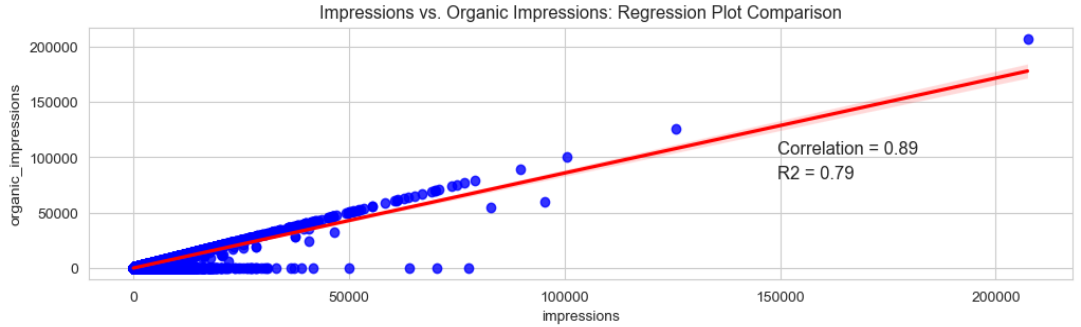
Beyond impressions and organic impressions, other impression metrics are only collected by Facebook.

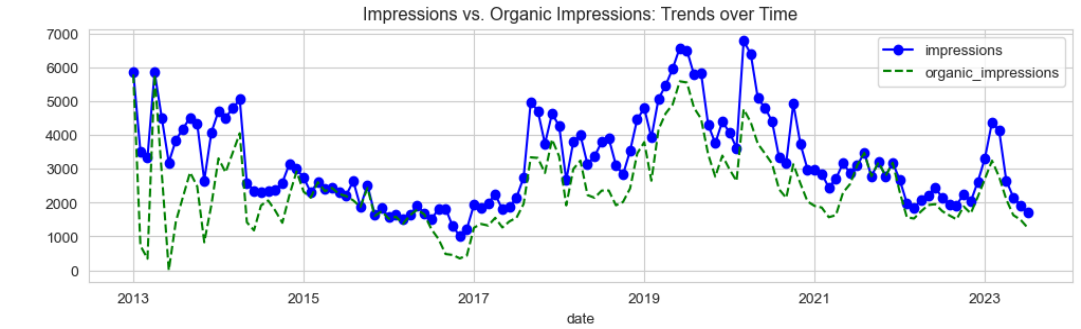


Please see the jupyter notebook for more information.

Impression Metrics – Impression vs. Organic Impressions

- Organic impressions are impressions generated without any paid promotion or advertising.
- Impressions and organic impressions are nearly identical.
- This implies that nearly all impressions are organic.
- The values on the zero axis represent impressions from Twitter, which does not collect organic impressions data.
- Trend data shows that organic impressions are consistently lower than actual impressions, reflecting missing data from Twitter. However, the trend is identical.

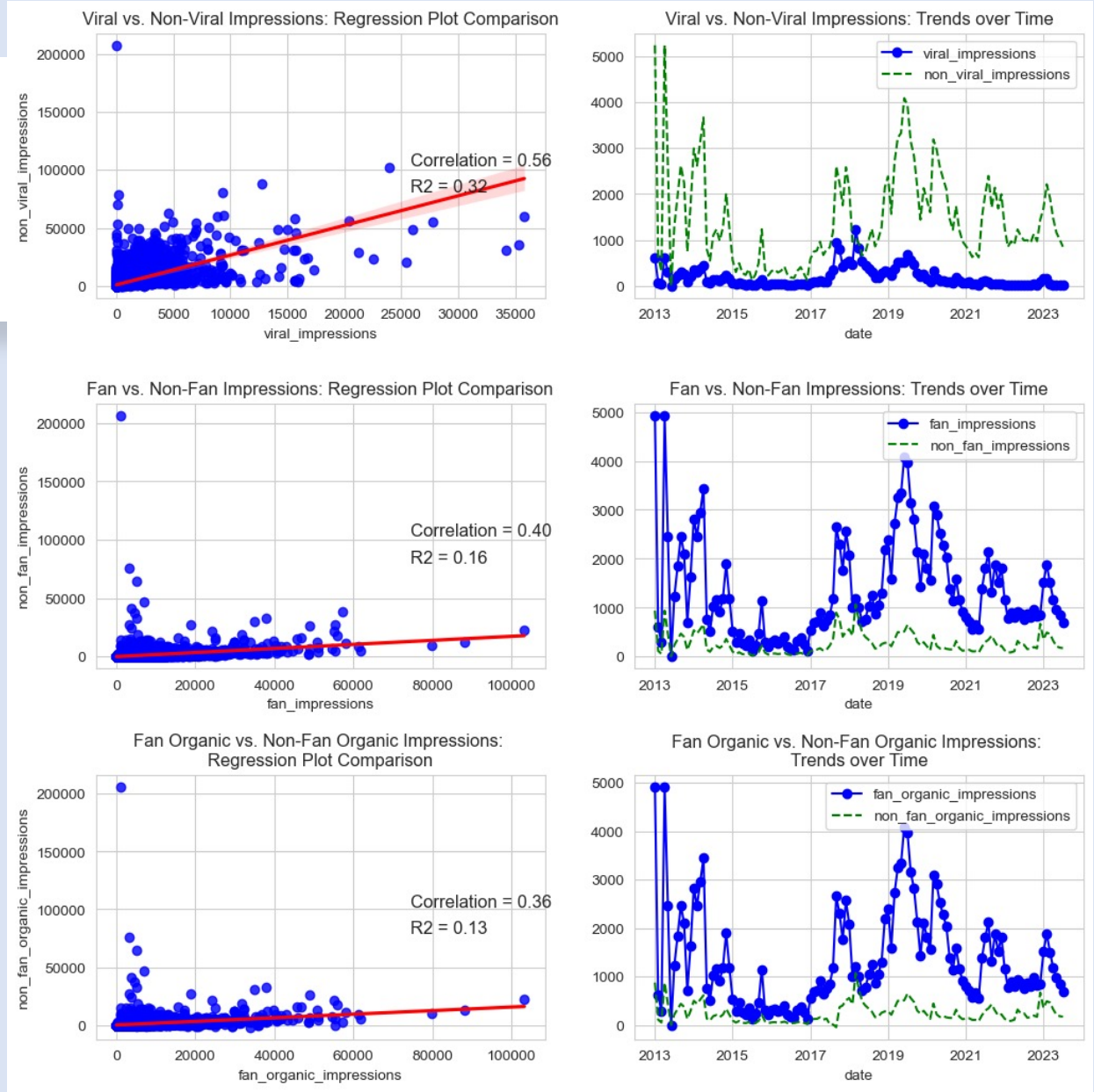




Please see the jupyter notebook for more information.

Impression Metrics – Other Impressions

- Fan impressions are from users who have liked or followed a Facebook Page (fans) and and non-fan impressions are from those who have not (non-fans).
- Viral impressions are from content that spread rapidly and widely across Facebook. Non-viral impressions are those from the limited audience of the poster's fan or followership.
- From the regression plots, there is some correlation between each metric and its opposite.
- The trend charts show that a significant amount of the poster's impressions are from its fans and followers.



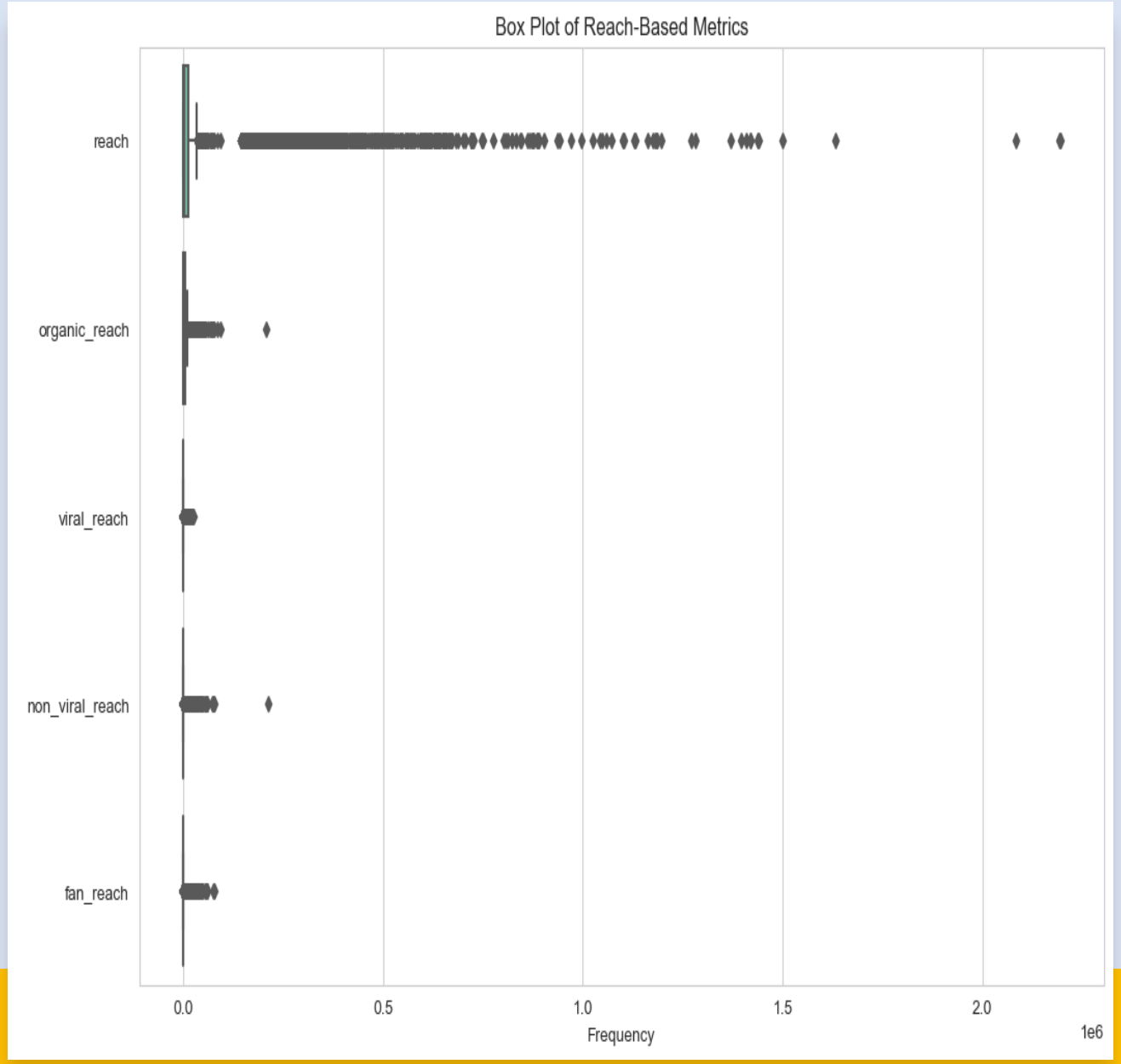
Please see the jupyter notebook for more information.

Reach Metrics

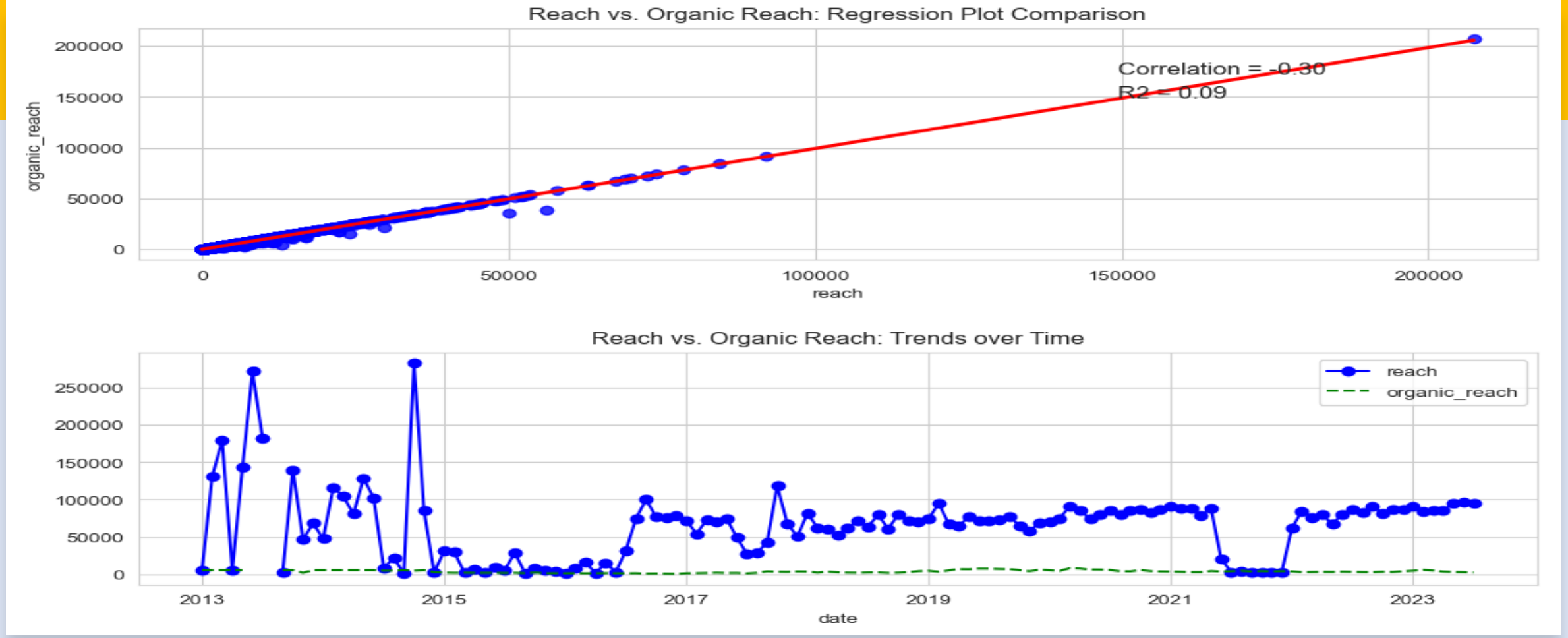
There are 5 reach metrics:

- Reach
- Organic reach
- Viral reach
- Non-viral reach
- Fan reach

Beyond reach and organic reach, other reach metrics are only collected by Facebook.



Please see the jupyter notebook for more information.



Reach Metrics – Reach vs. Organic Reach

- Organic reach refers to the number of unique users who have seen a piece of content through unpaid means.
- Reach and organic reach are nearly identical.
- This implies that nearly all reach is organic.
- There are no reach nor organic reach data attributable to LinkedIn.
- Reach has significantly larger values than organic reach due to the large values from Twitter's potential reach.
- Trend data shows that organic reach is consistently lower than actual reach, reflecting missing data from Twitter.

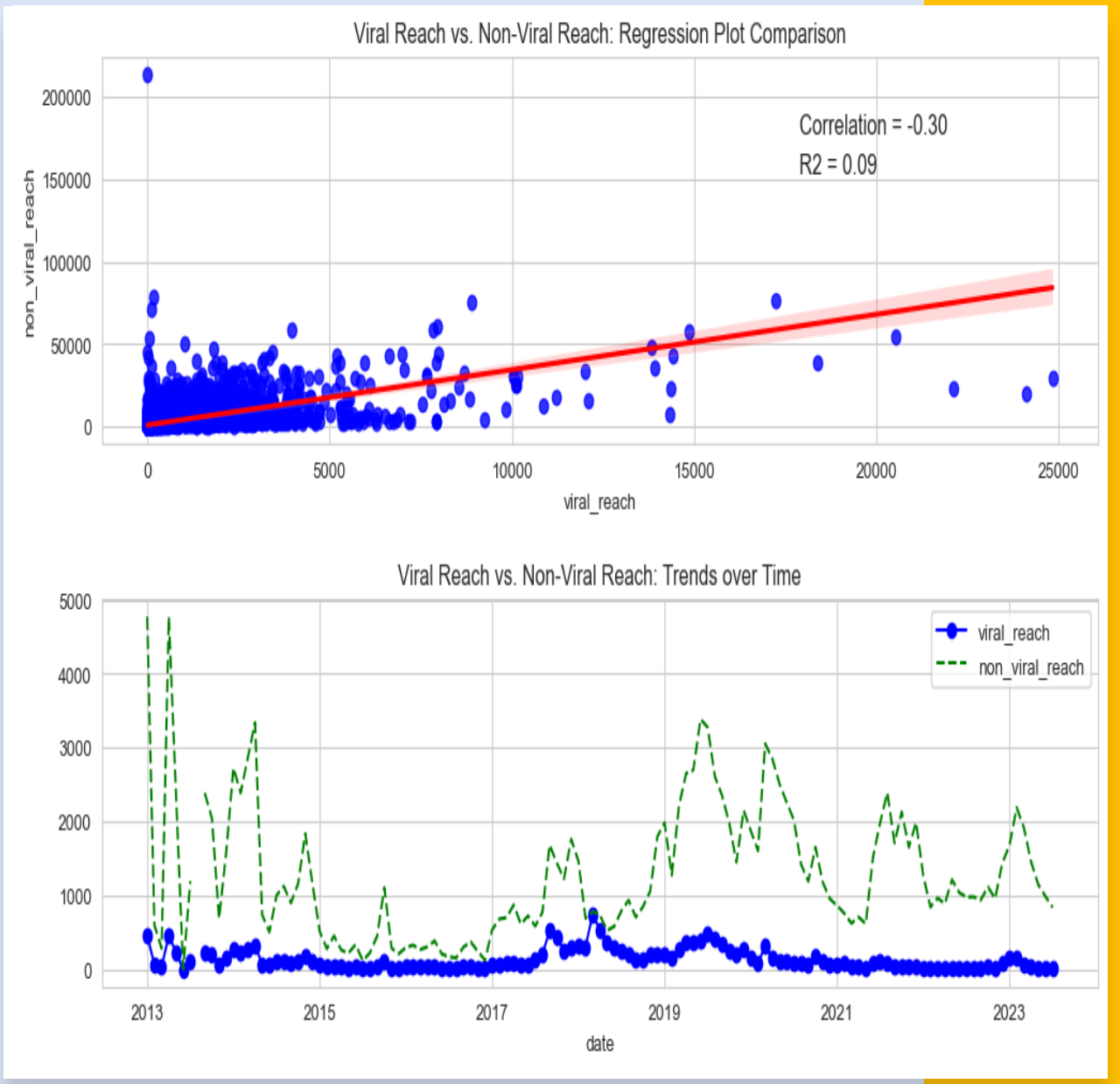
Please see the jupyter notebook for more information.

Reach Metrics – Other Reach

Viral reach indicating that a piece of content has gone viral and reached a wide audience, while non-viral reach pertains to content that has not achieved viral status.

Fan reach occurs when content is seen by users who have already liked or followed a specific Facebook Page (fans).

The trend charts show that a significant amount of the poster's reach are from its fans and followers.

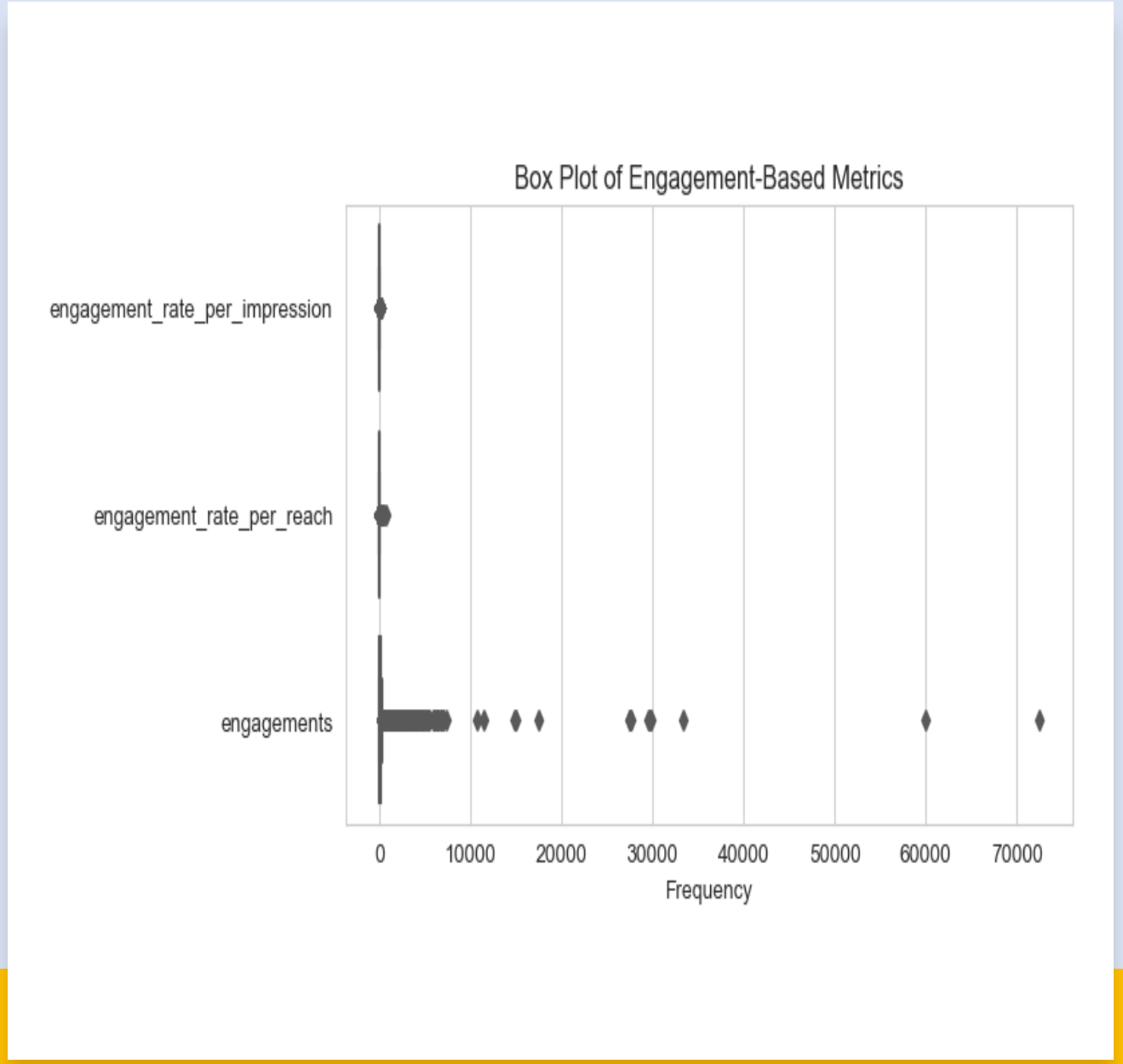


Please see the [jupyter notebook]for more information.

Engagement Metrics

There are 3 engagement metrics:

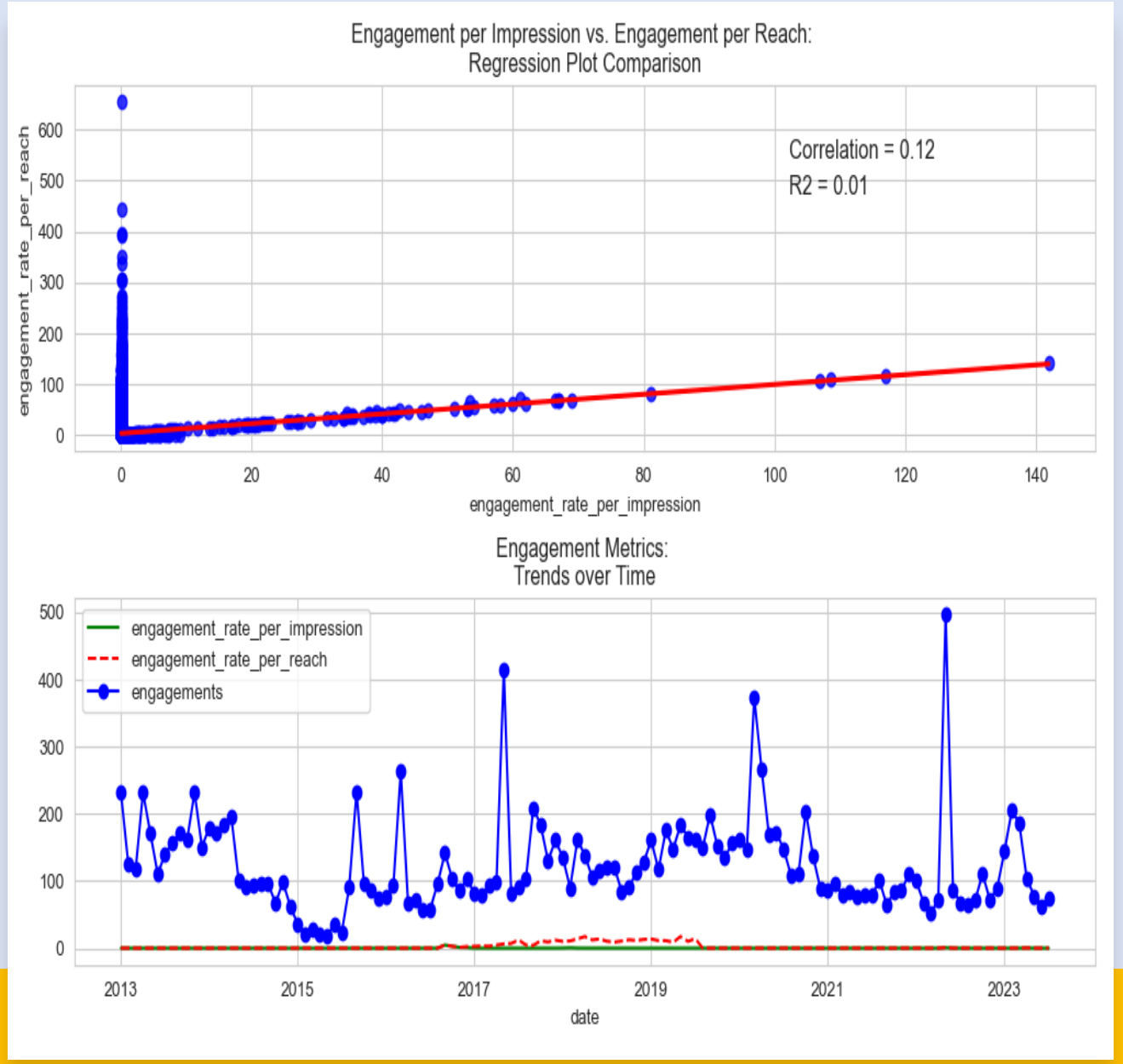
- Engagements
- Engagement per impression
- Engagement per reach



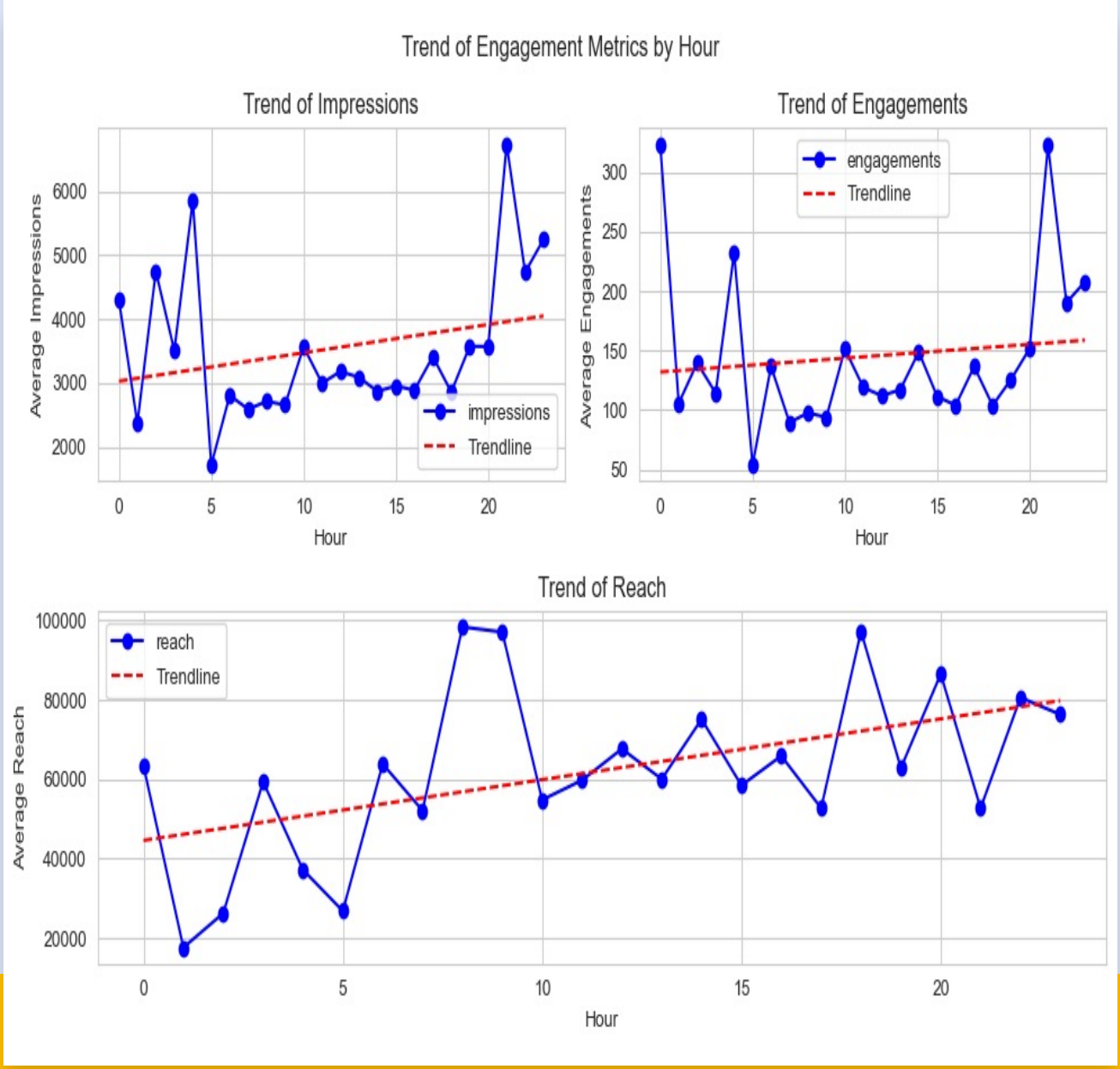
Please see the jupyter notebook for more information.

Engagement Metrics

- Engagement per impressions and engagement per reach reflect the ratio of engagement and impressions/ reach.



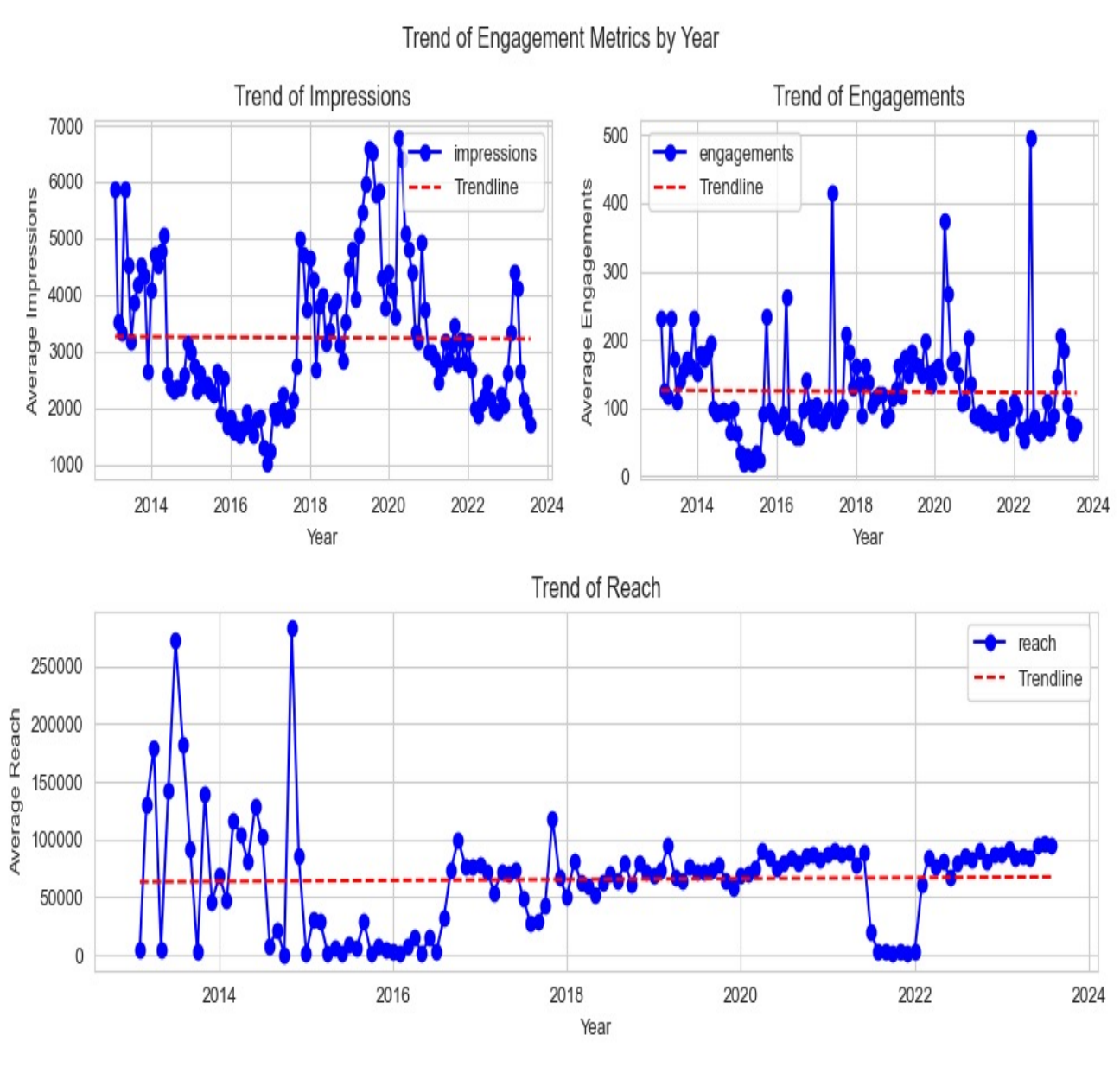
Please see the [jupyter notebook]() for more information.



Engagement Metrics by Hour

- When aggregated by Hour, engagements see an upward trend as the day progresses across all engagement metrics.
- 9pm is the Hour for peak impressions and engagements on average.
- 8am and 6pm are the hours of peak reach on average.

Please see the jupyter notebook for more information.



Engagement Metrics by Year

- When aggregated by Year, engagements are flat across all engagement metrics.
- 2020 is the year for peak impressions on average.
- 2022 is the year for peak engagements on average.
- 2015 is the year of peak reach on average.

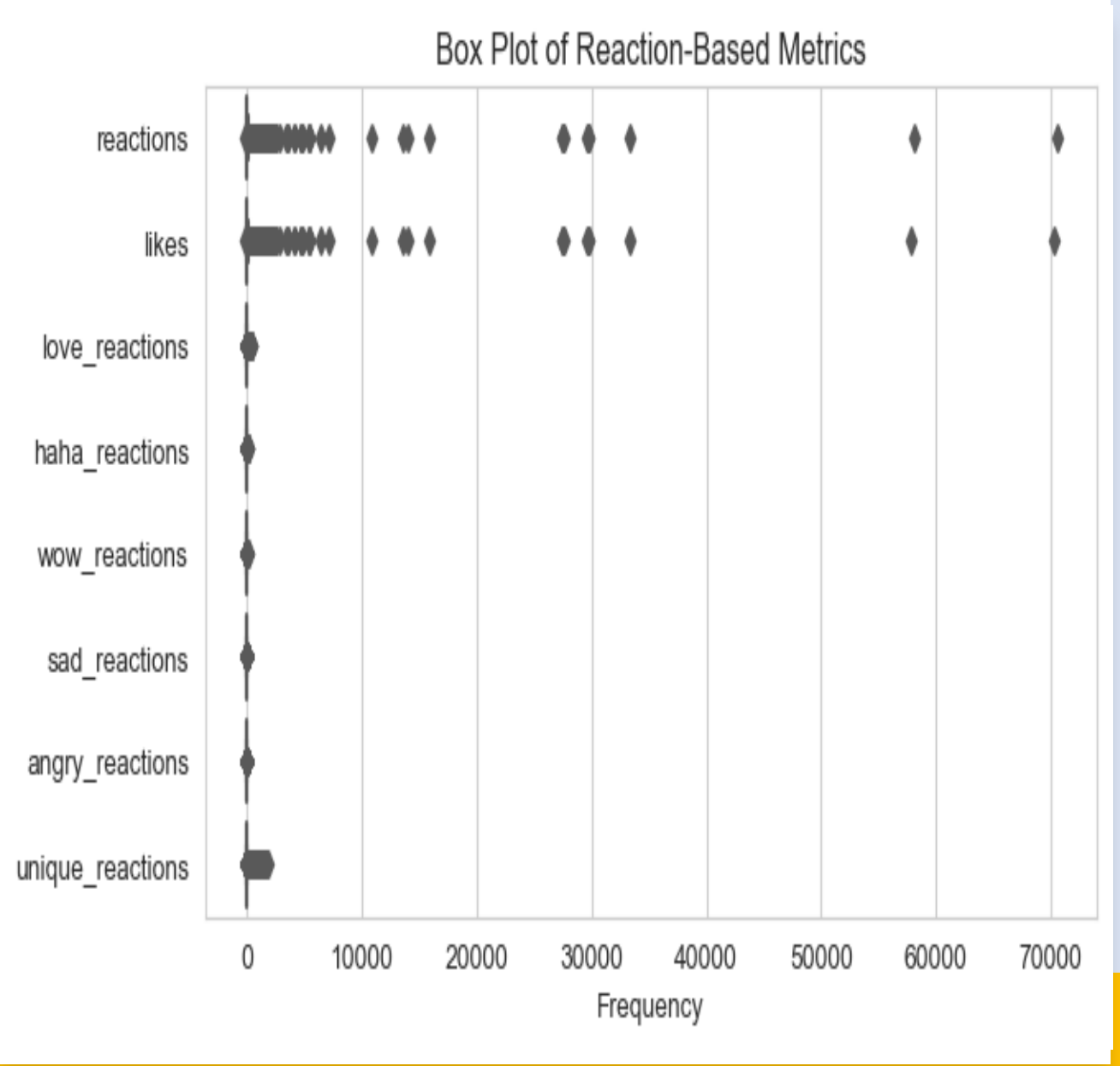
Please see the jupyter notebook for more information.

Reaction Metrics

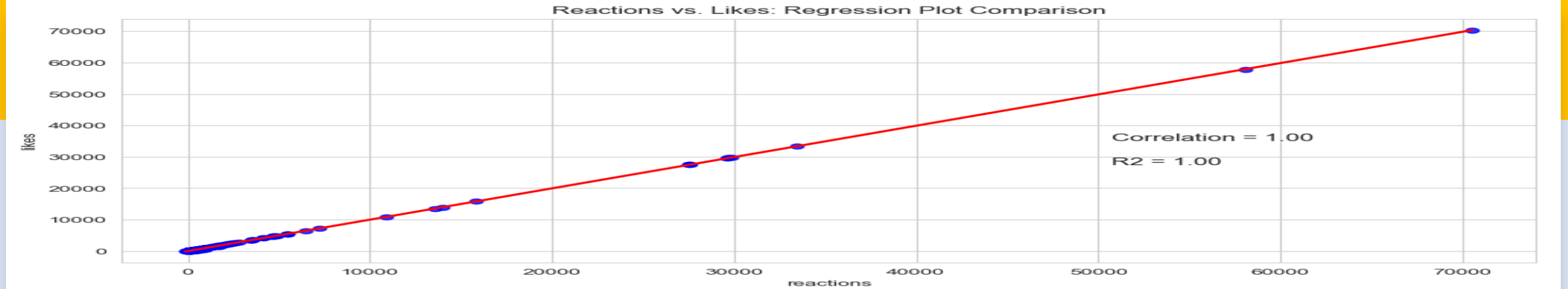
Reaction metrics provide insights into how users are responding to posted content. There are 8 reaction metrics:

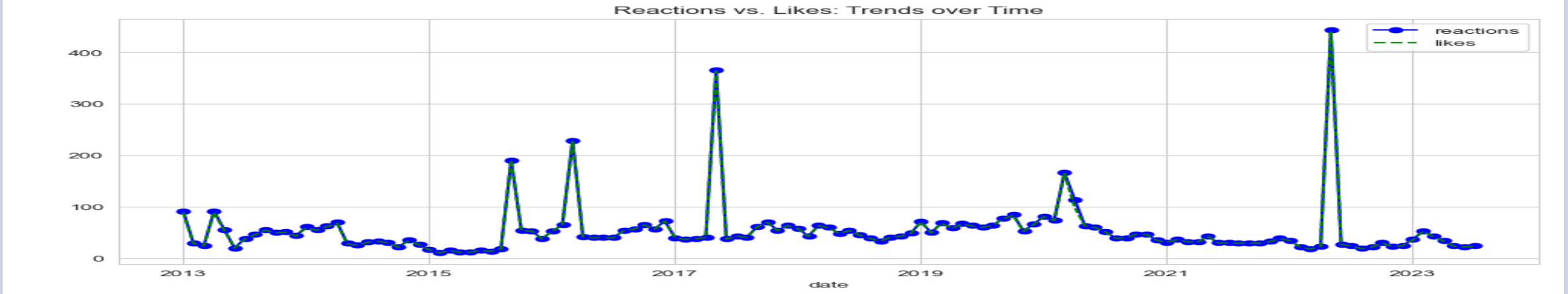
- Reactions
- Likes
- Love reactions
- Haha reactions
- Wow reactions
- Sad reactions
- Angry reactions
- Unique reactions

Beyond reactions and likes, other reaction metrics are only collected by Facebook.



Please see the jupyter notebook for more information.





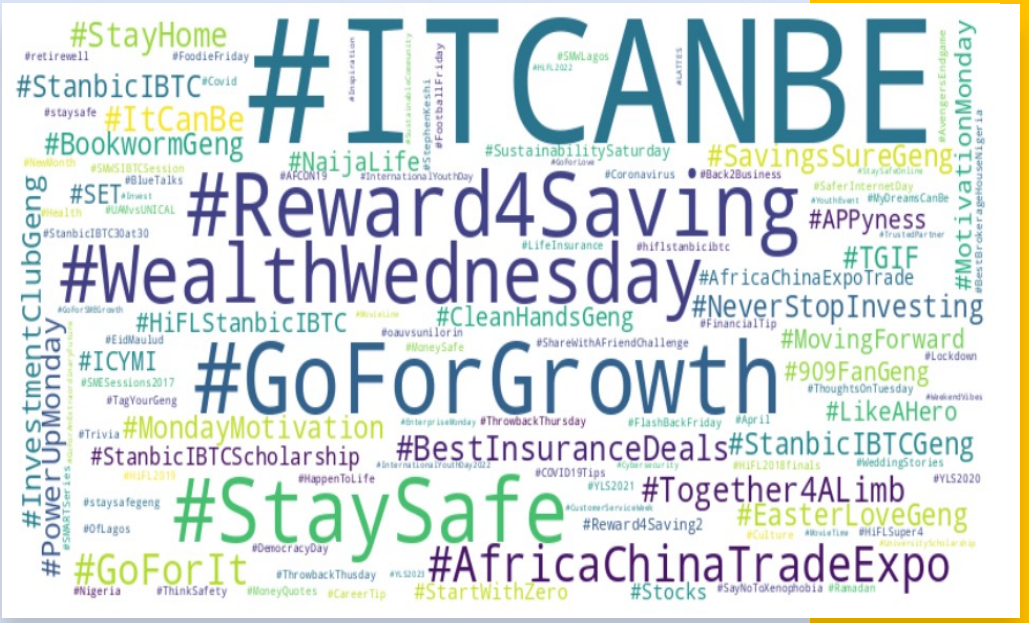
Reaction Metrics – Reactions vs. Likes

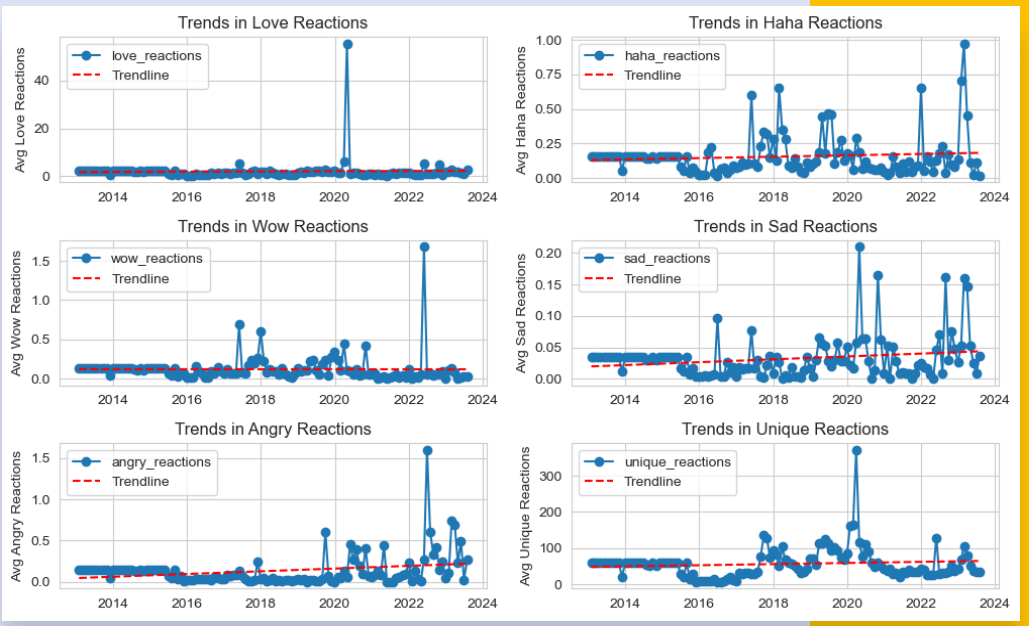
- Users use likes to express positive reaction to a post. Reactions allow users to express a wider range of engagement a range of emotions in response to content.
- In our dataset, reactions and likes are identical.

Please see the jupyter notebook for more information.

Reaction Metrics – Other Reactions

- Apart from love and wow reactions, other reaction metrics have seen an increase in values over time.
- Given the relationship between reactions and engagements, this may imply that the poster's followers have become more engaging.
- The highest trend increases are from sad and angry reactions. When compared to the word cloud of hashtags from posts with sad reactions greater than the trendline, the cause of this increase is unclear.





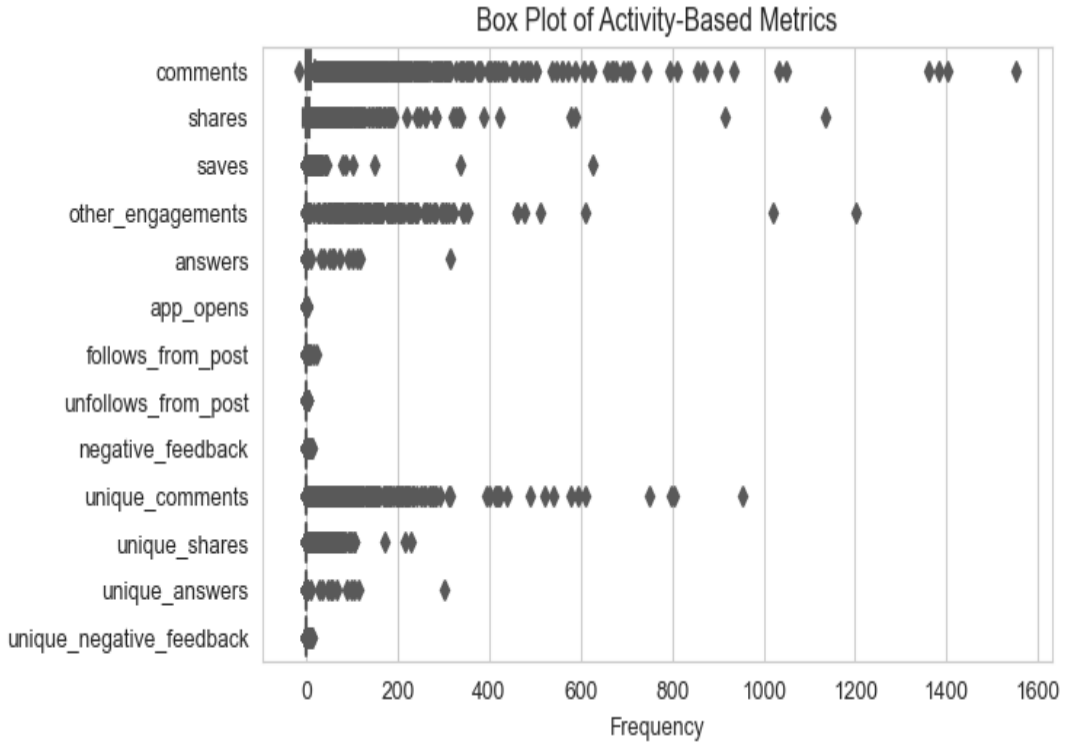
Please see the jupyter notebook for more information.

Activity Metrics

There are 13 activity metrics. Comments is the only metric common to all and shares is collected by all but Instagram.

Segmented by network, the activity metrics are:

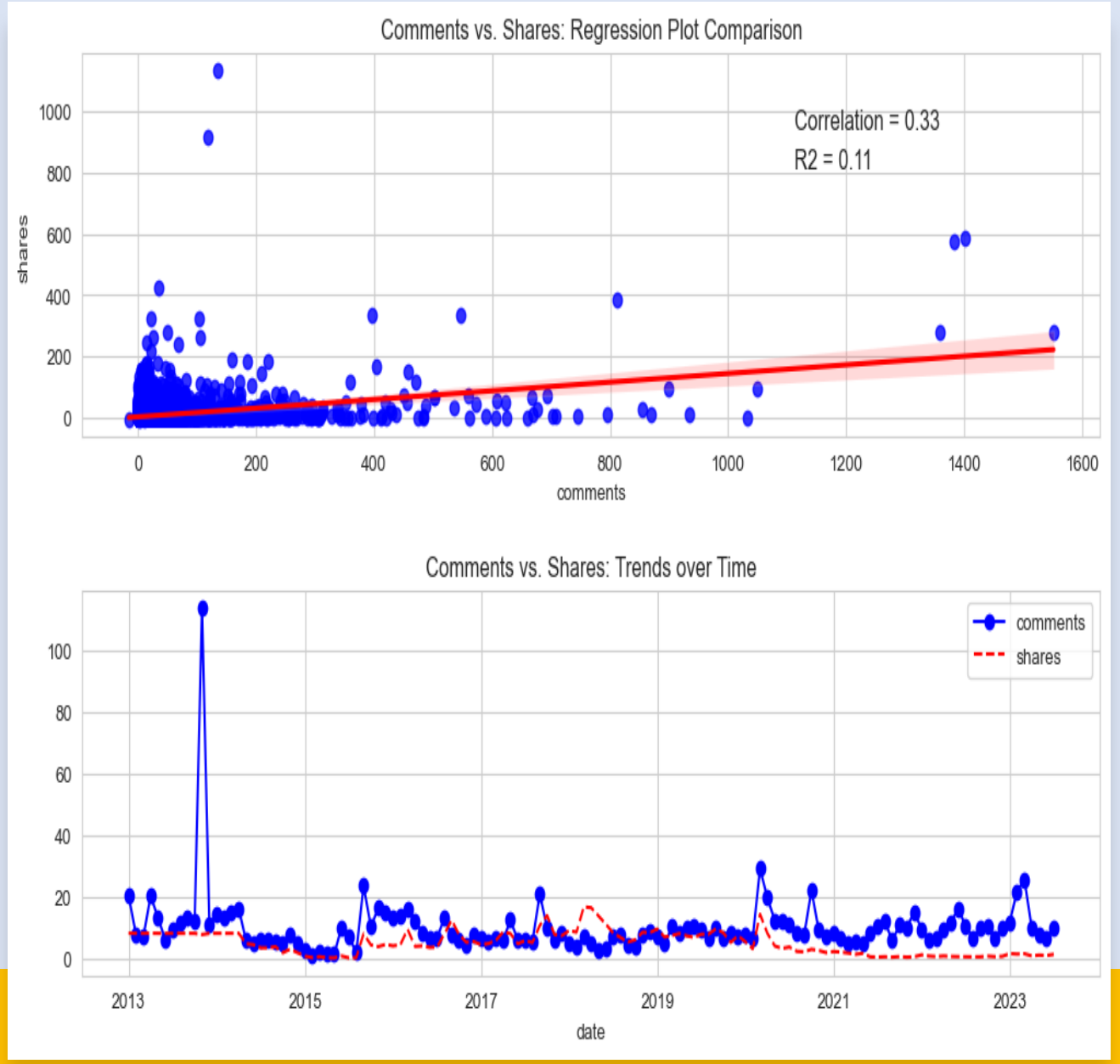
- Facebook – answers, negative feedback, unique comments, unique shares, unique answers, & unique negative feedback.
- Twitter – other engagements, app opens, follows from post, & unfollows from post.
- Instagram – saves.



Please see the jupyter notebook for more information.

Activity Metrics – Comments & Shares

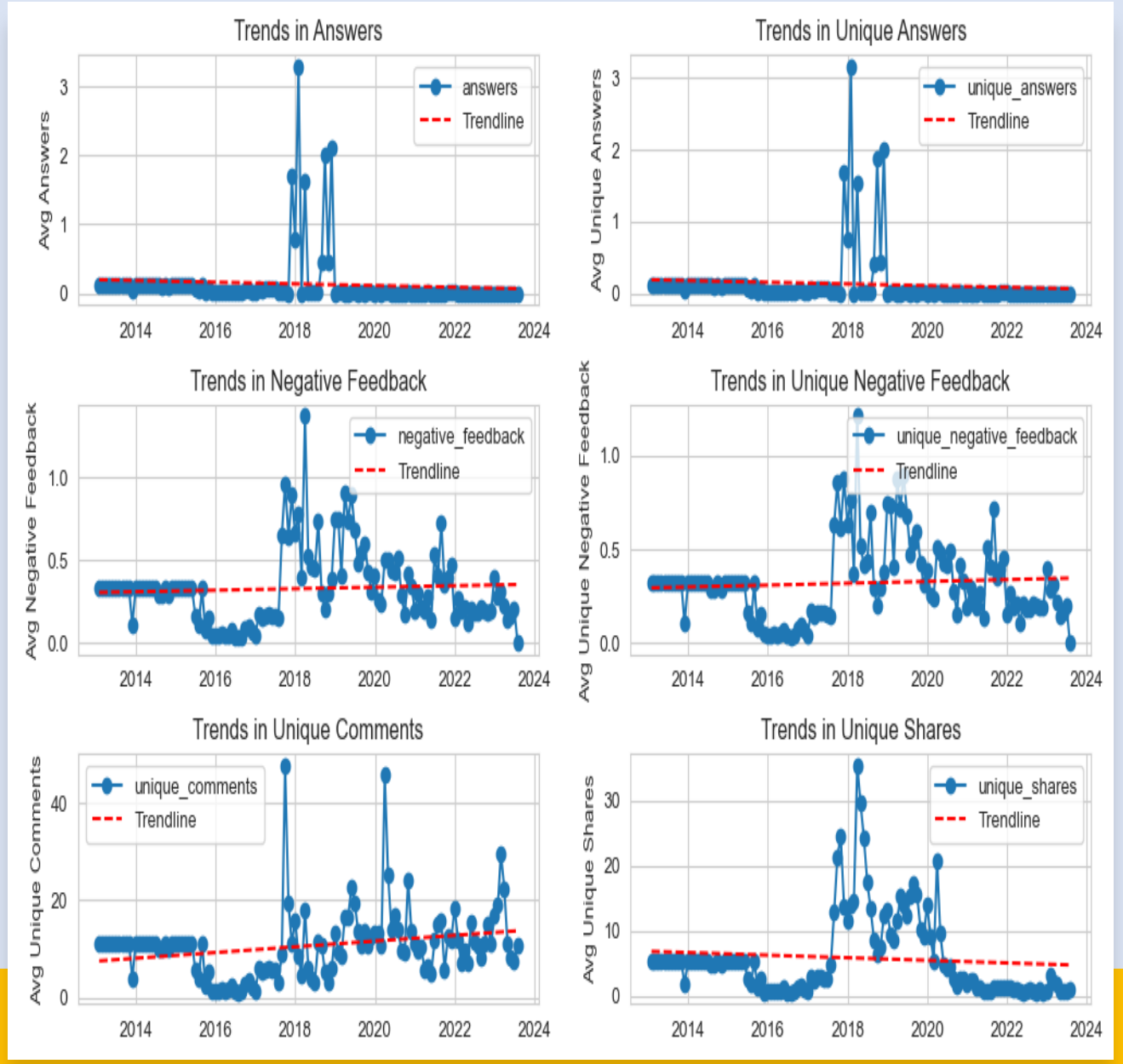
- Comments & shares are slightly correlated. The extension along the 0-axis may represent values from Instagram which does not collect shares data.
- The trend of comments and shares are similar, however posts on average have more comments than shares.



Please see the jupyter notebook for more information.

Activity Metrics – Facebook

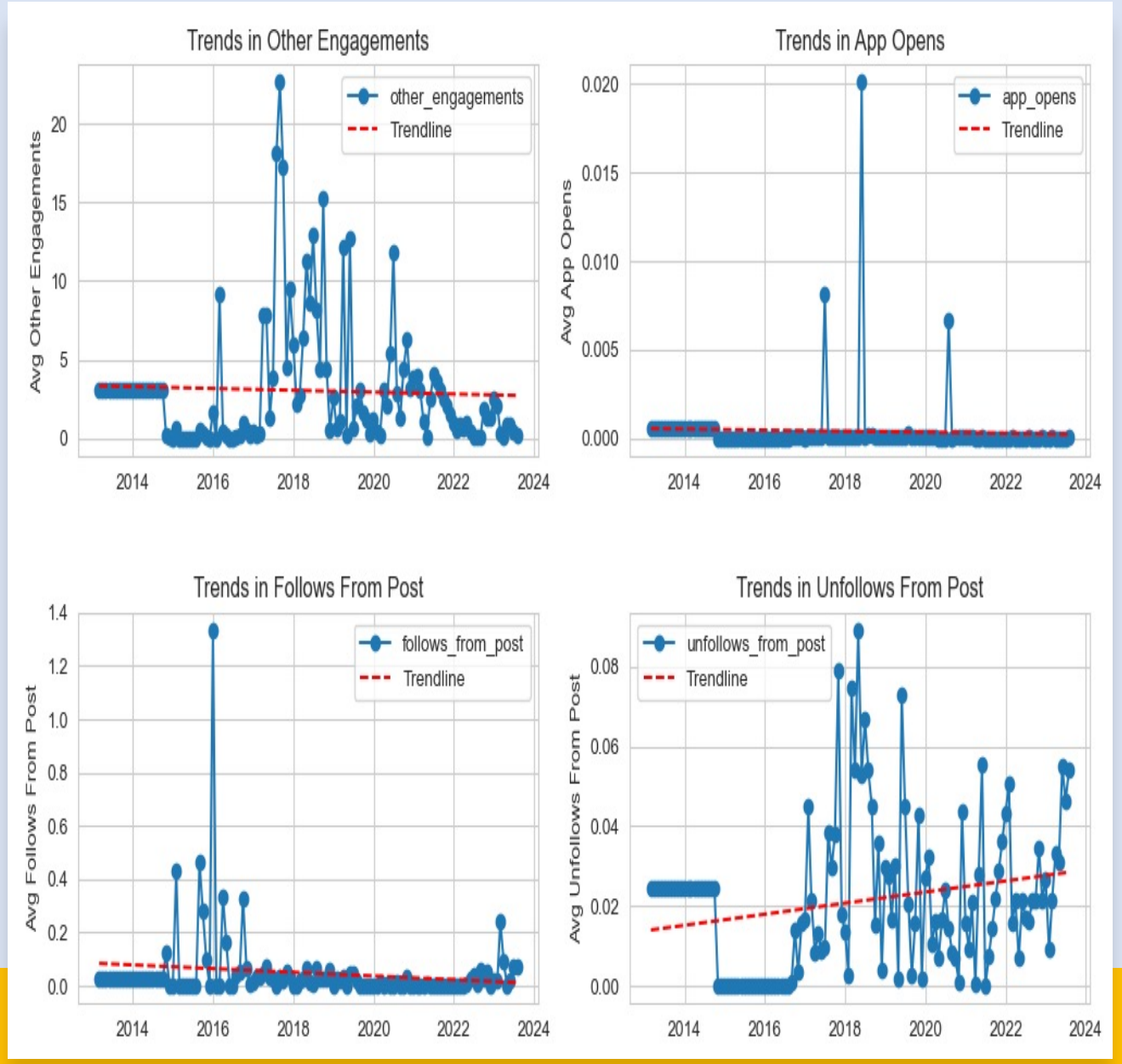
- Trend in answers is on a decline while trends in negative feedback is on the rise.
- Unique comments are also rising while unique shares are declining.
- While we can't determine the cause of this from hashtags, senders or content types, we can see from all charts that there is a significant increase in metric values from 2018 that starts to fall around 2020-2021.
- This may be the result of changing algorithms and user patterns.



Please see the jupyter notebook for more information.

Activity Metrics – Twitter

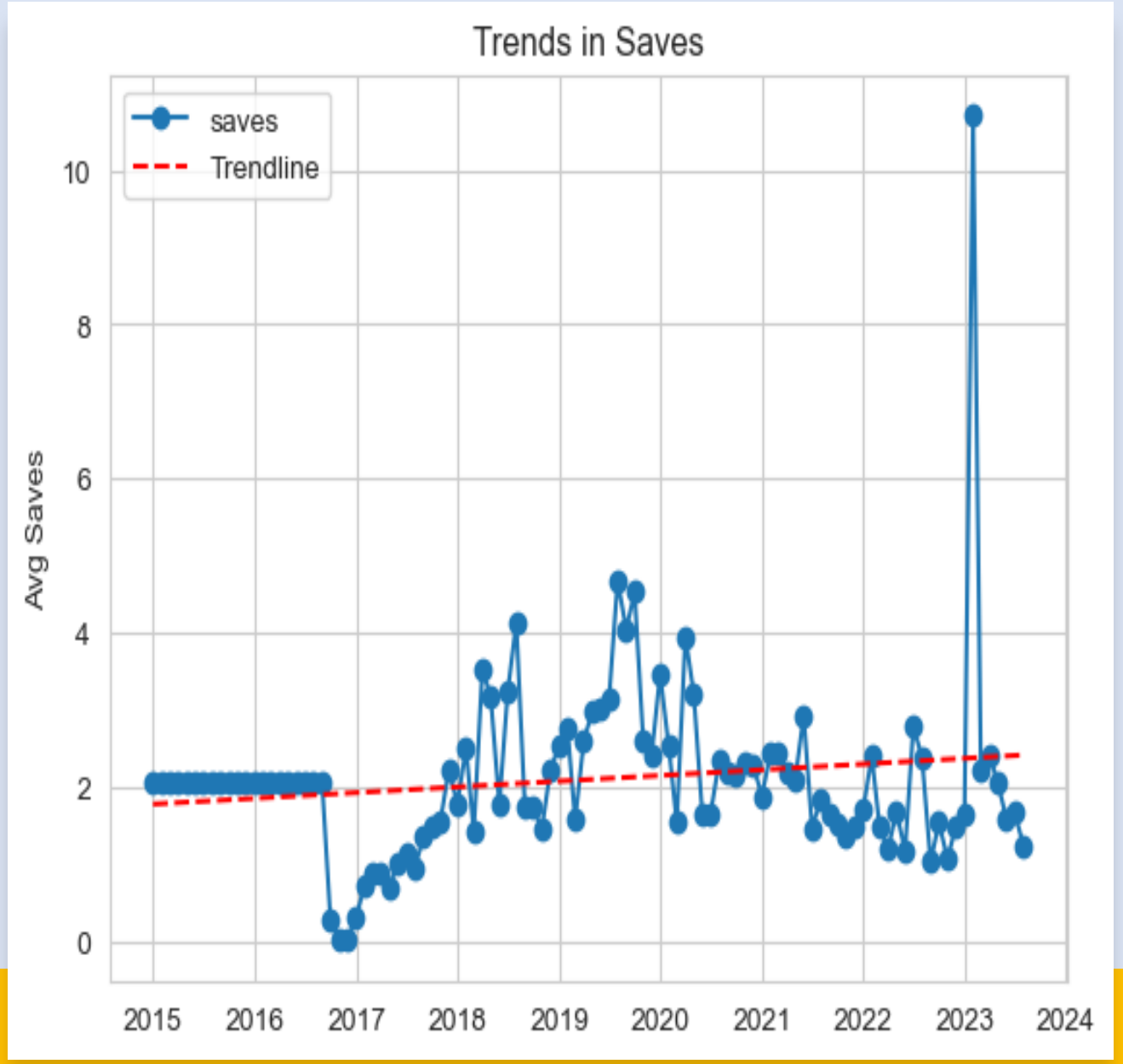
- Trends from all twitter activity metrics are on the decline except unfollows from post which is on the incline.
- This is likely due to changing user behaviour on the platform.
- While we can't determine the cause of this from hashtags, senders or content types, we can see from all charts that there is a significant increase in metric values from 2018 that starts to fall around 2020-2021.
- This may be the result of changing algorithms and user patterns.



Please see the [jupyter notebook]for more information.

Activity Metrics – Instagram

- Instagram saves are on the increase on the average despite a decline from the upward trend peak in 2019.



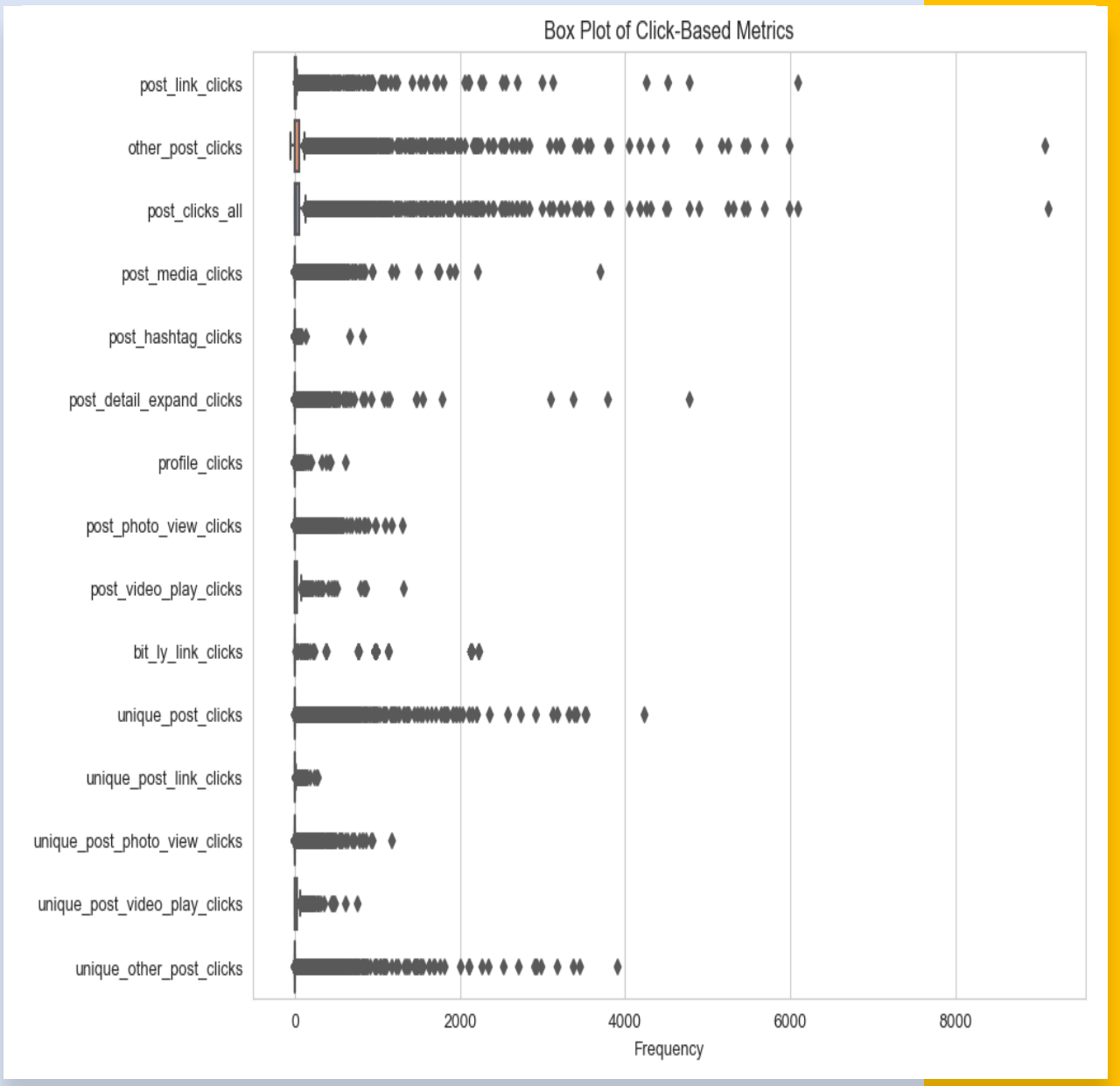
Please see the jupyter notebook for more information.

Click Metrics

There are 15 click metrics. Instagram does not collect any click metrics. Of the remaining network, link clicks and all clicks are tracked by all of them.

Segmented by network, the click metrics are:

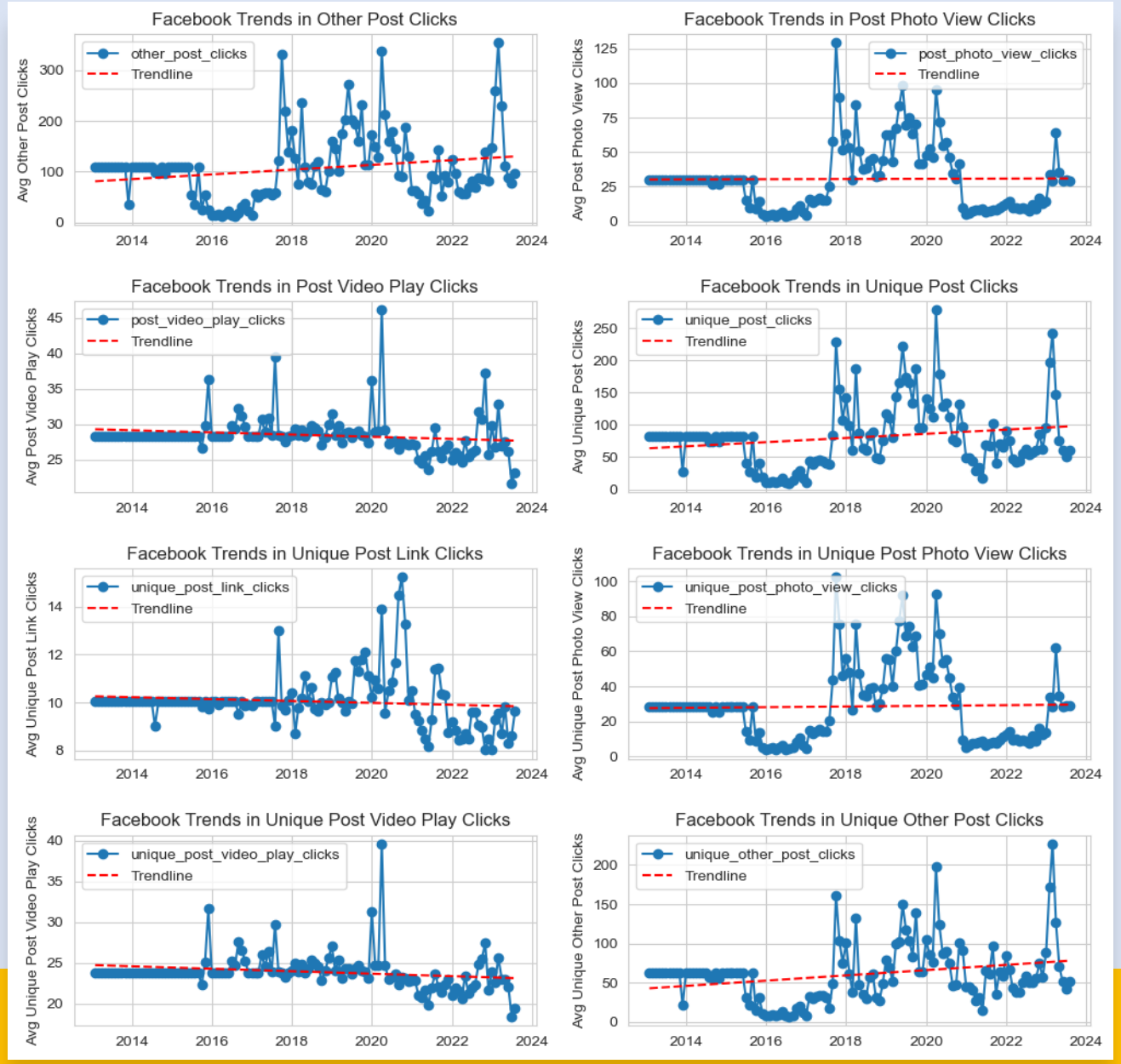
- Facebook – other post clicks, post photo view clicks, post video play clicks, unique post clicks, unique post link clicks, unique post photo view clicks, unique post video play clicks & unique other post clicks.
- Twitter – other post clicks, post media clicks, post hashtag clicks, post detail expand clicks, profile clicks & bit.ly link clicks.



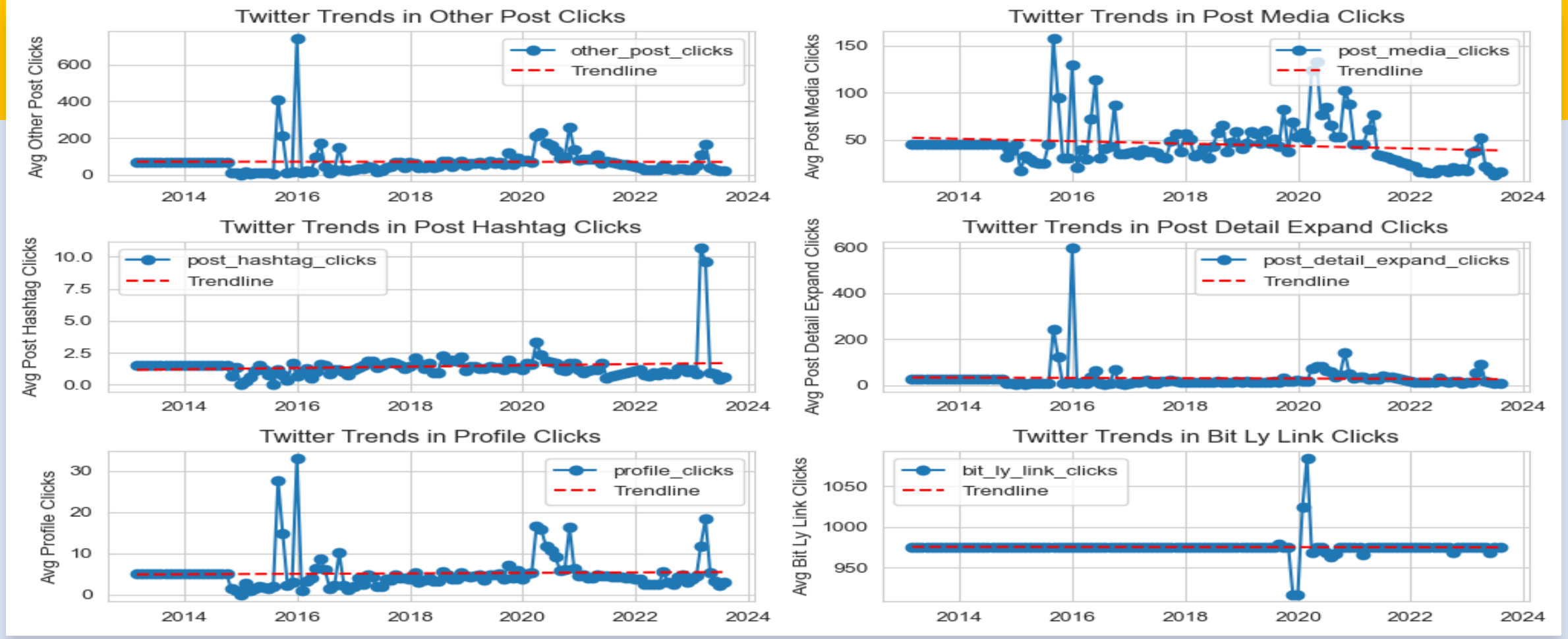
Please see the jupyter notebook for more information.

Click Metrics – Facebook

- The trend for clicks on posts (possibly to read more) is on the rise, but the trend for link and video post clicks are on the decline.
- This may be because the poster's Facebook fans either prefer text posts to video / link posts and/ or text posts are cheaper on user data than other posts.
- The poster must have a posting strategy for each social media network that includes how its user base engages with the platform.



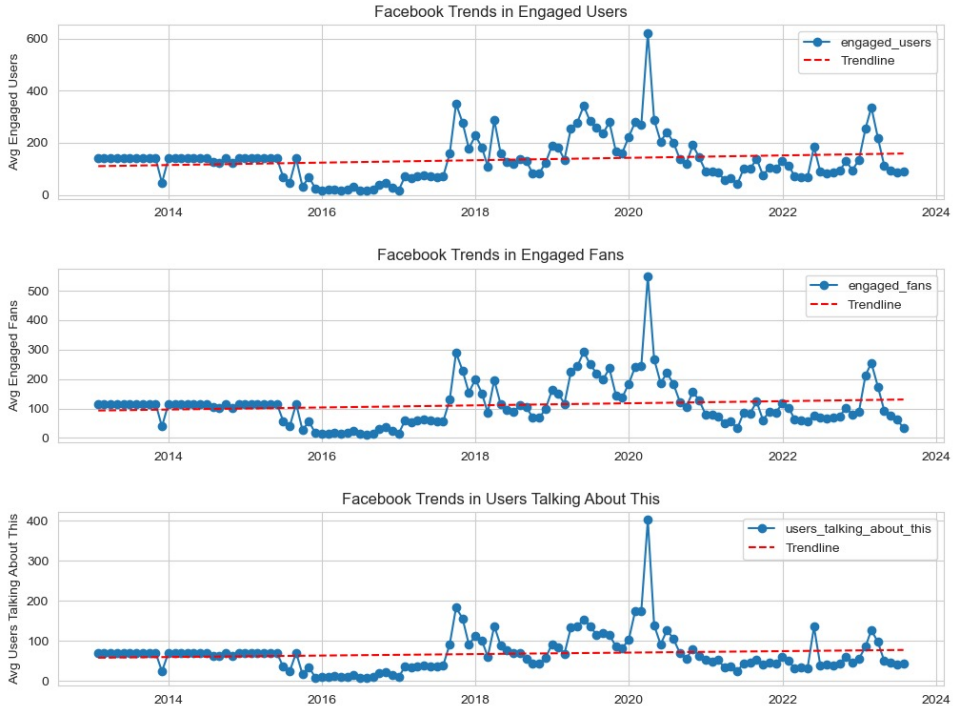
Please see the jupyter notebook for more information.

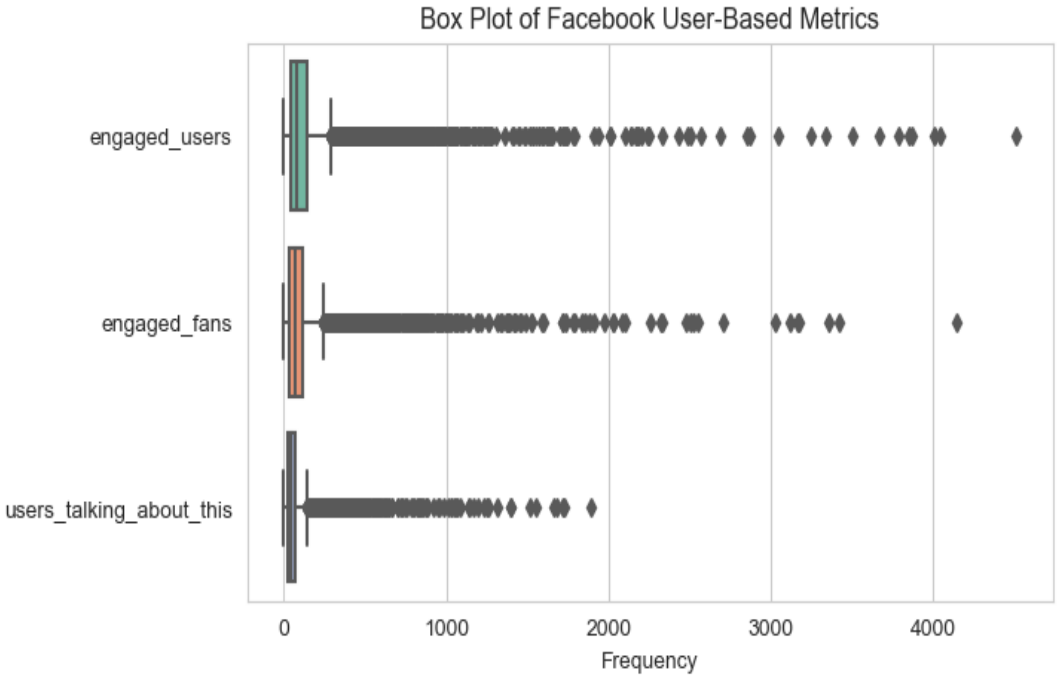


Click Metrics – Twitter

- Most twitter click trends are flat.
- Media clicks trend is on the decline, facing a sharp decline from 2021.
- Hashtag clicks is trending slightly upward, which is line with how users use Twitter.

Please see the jupyter notebook for more information.



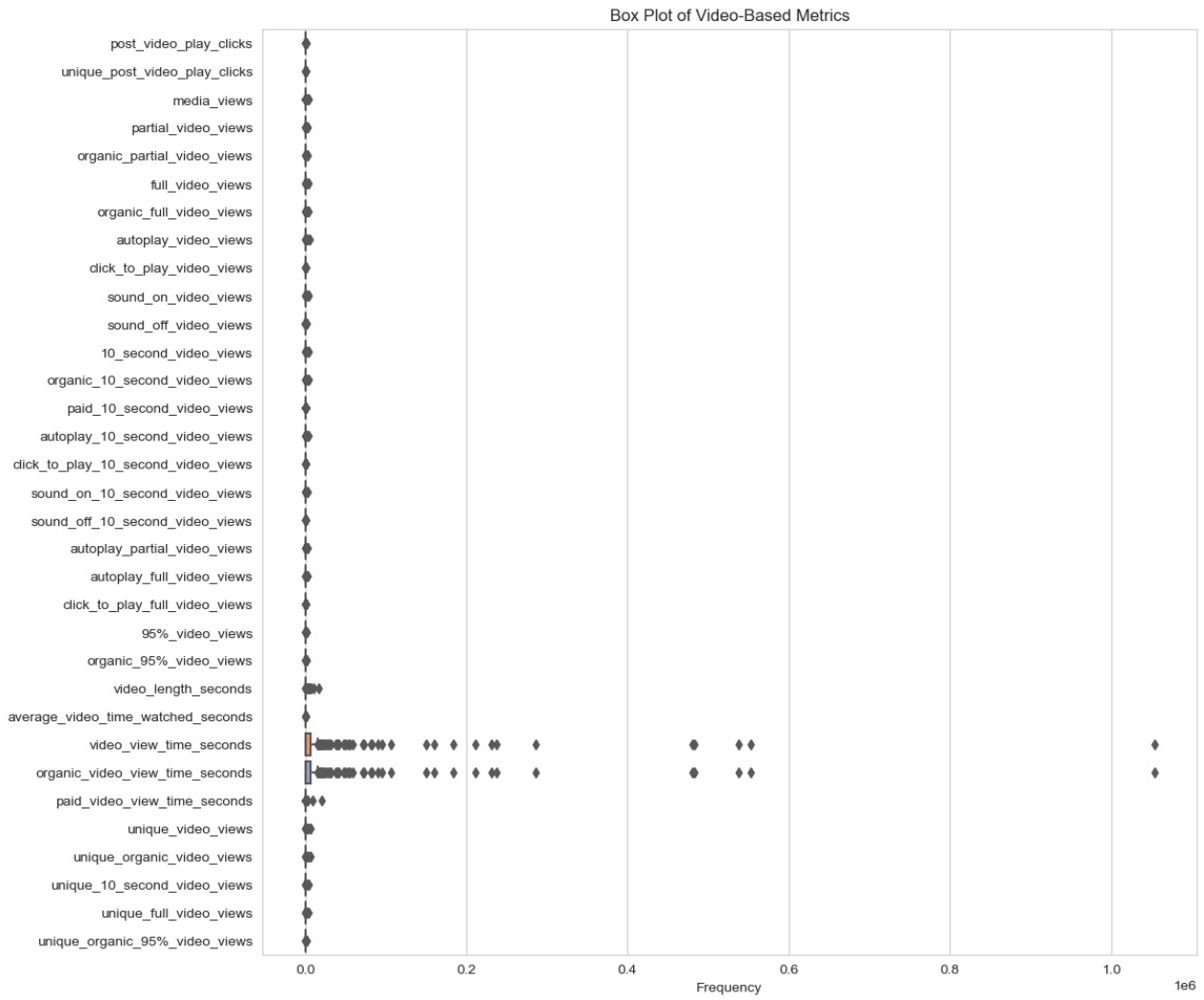


User-Based Metrics

- User based metrics are only tracked by Facebook.
- All these metrics are on an upward trajectory, implying that the poster's followers are increasingly engaging with the content on its page.
- Across all metrics, there is a significant spike in values between 2018 and 2020. This reflects a trend across many other metrics.

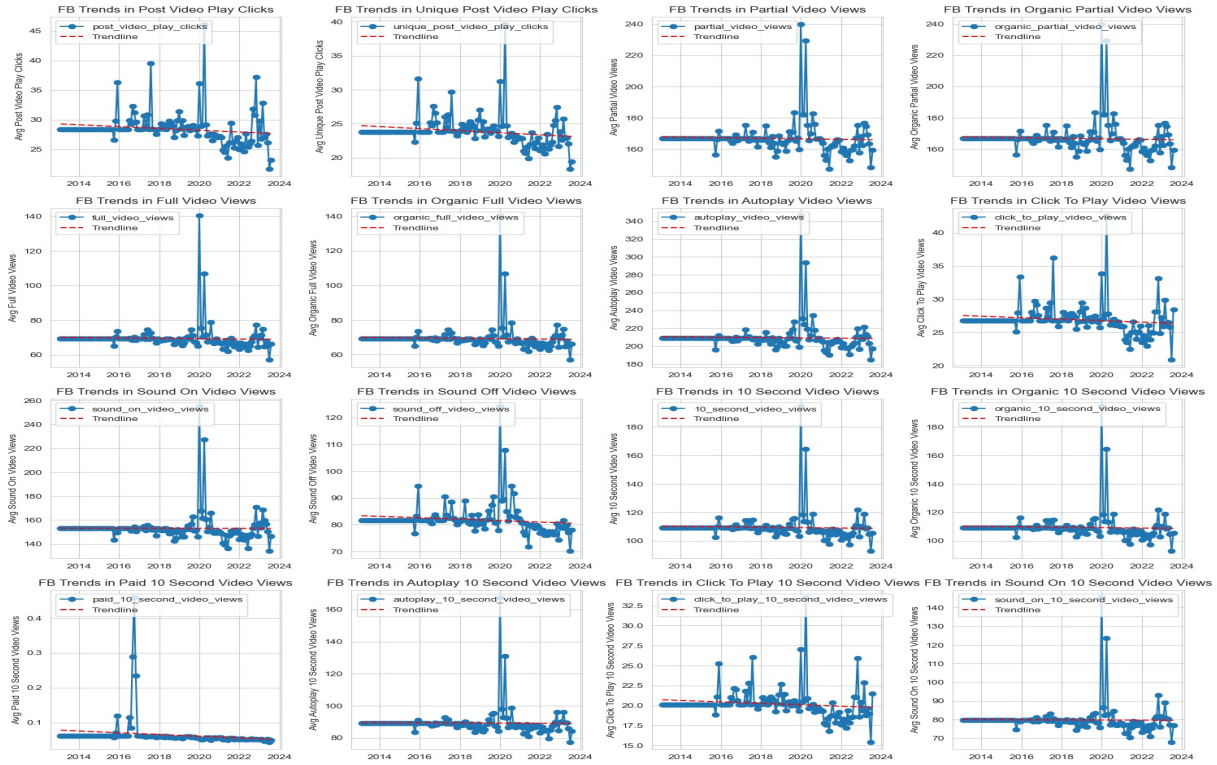
Video Metrics

- There are 32 video metrics all of which are tracked by Facebook except media views, which is only tracked by Twitter.



Video Metrics – Facebook

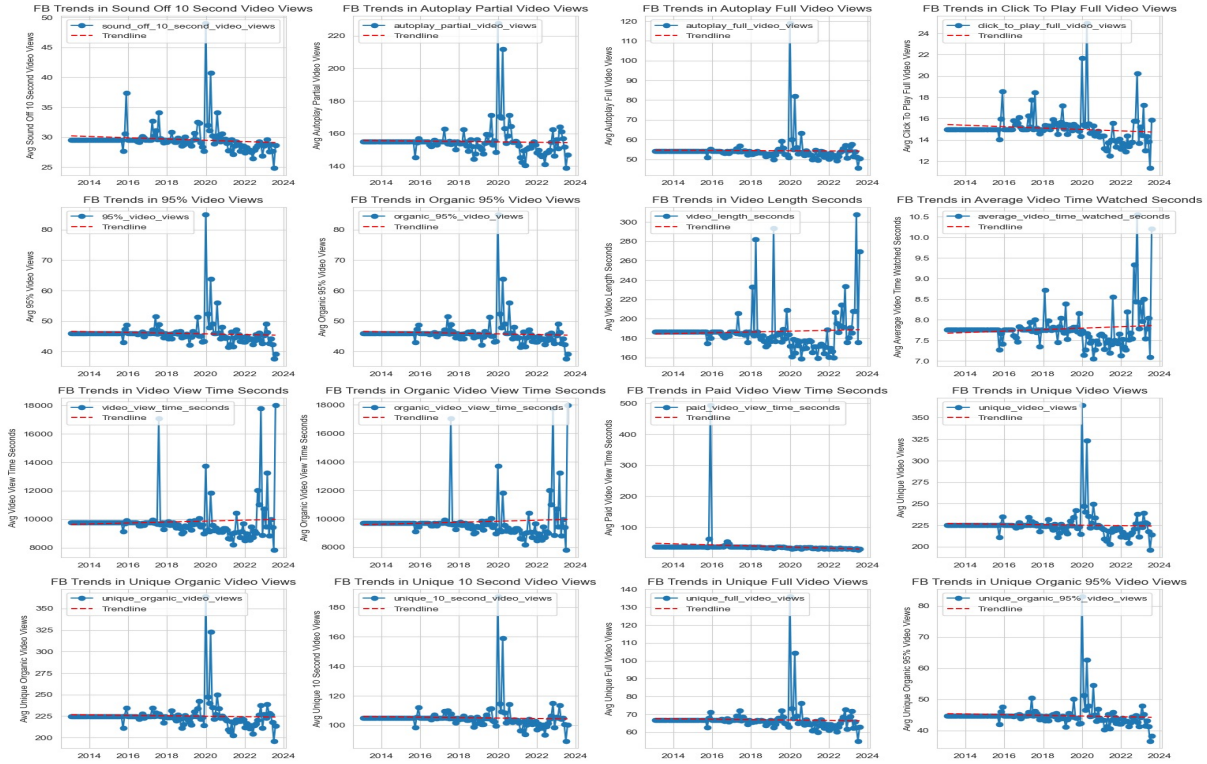
- All Facebook video metrics are either flat or trending downward.
- This confirms the trend we saw in clicks about video which has seen a sharp decline.



Please see the jupyter notebook for more information.

Video Metrics – Facebook 2

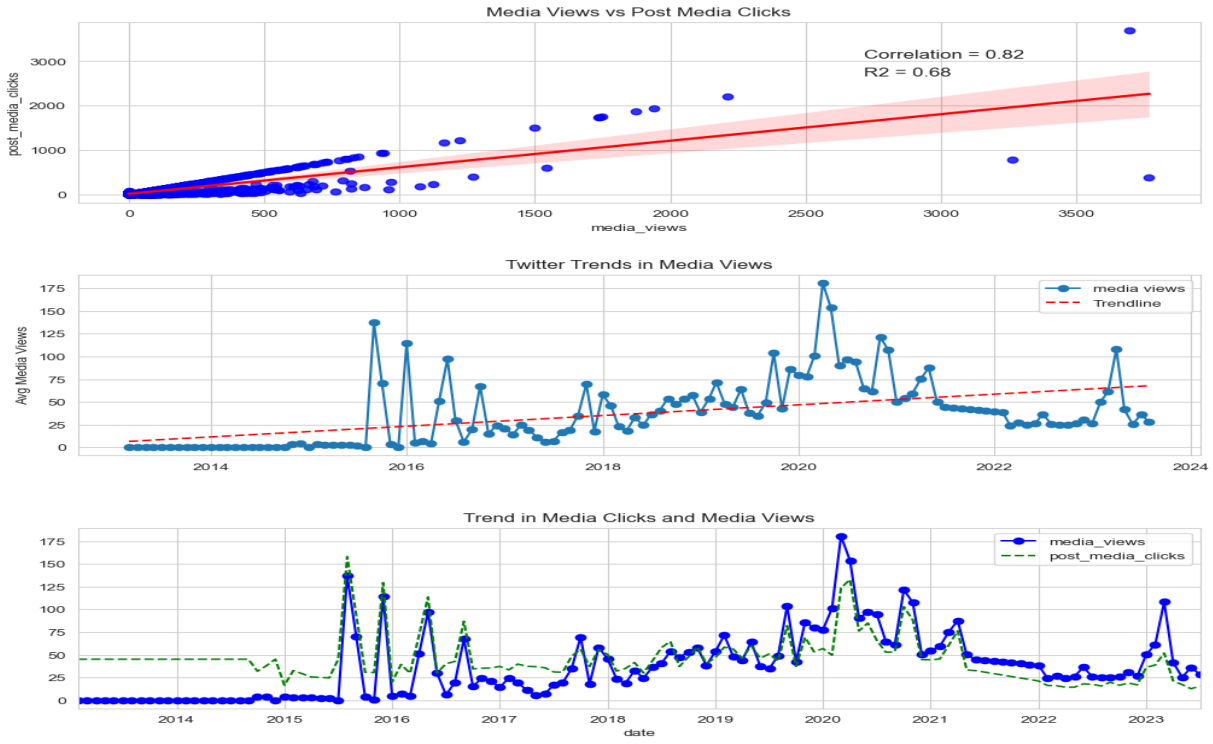
- All Facebook video metrics are either flat or trending downward.
- This confirms the trend we saw in clicks about video which has seen a sharp decline.



Please see the jupyter notebook for more information.

Video Metrics – Twitter

- Twitter media view trends are on the upward incline. This is contrasted with post media clicks which are on the decline.
- Media clicks and media views are positively correlated yet move in negative direction over time.
- The trendlines for both metrics are similar but there is a slight downtrend in media clicks trend.



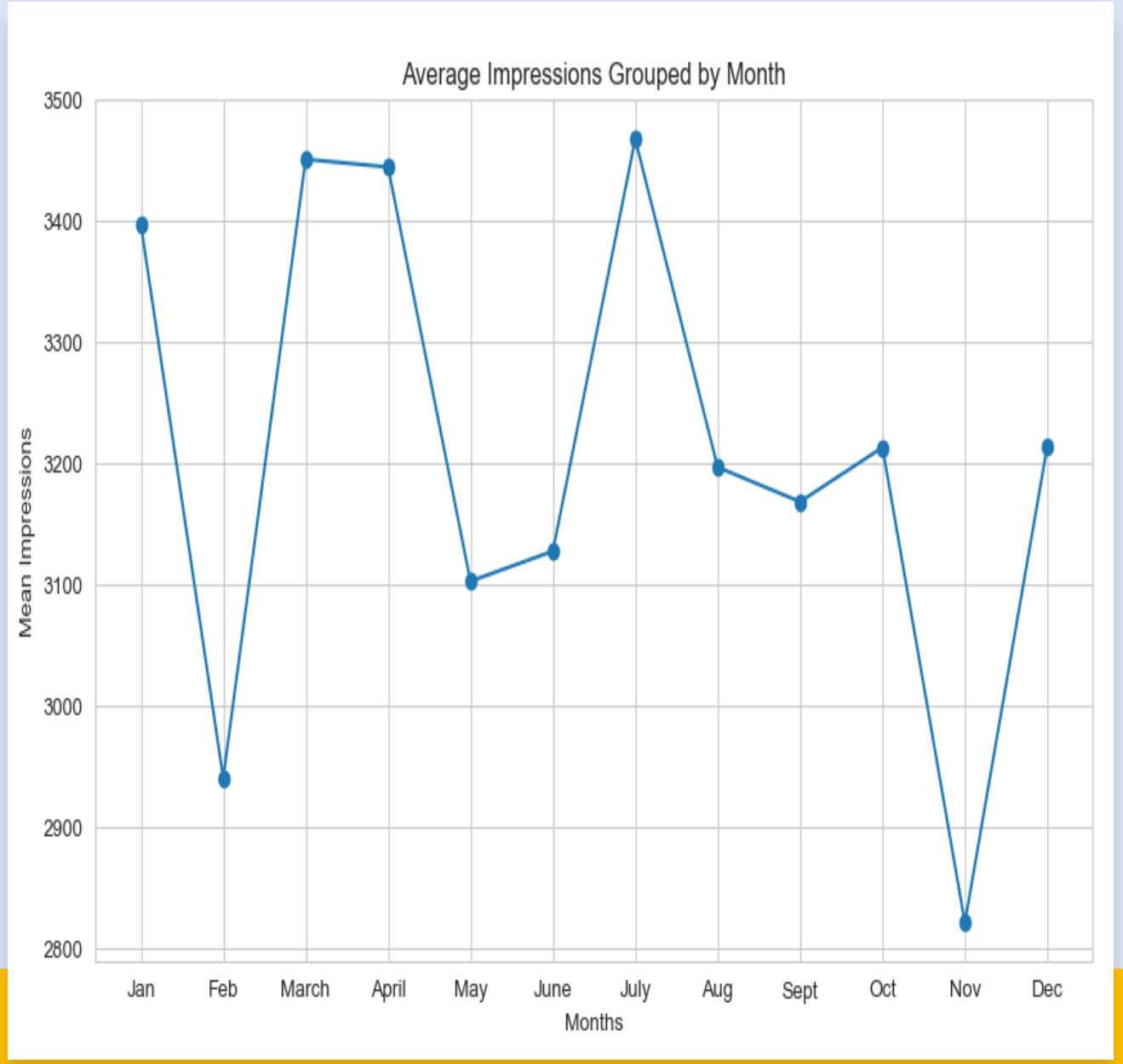
Please see the jupyter notebook for more information.

Section 3

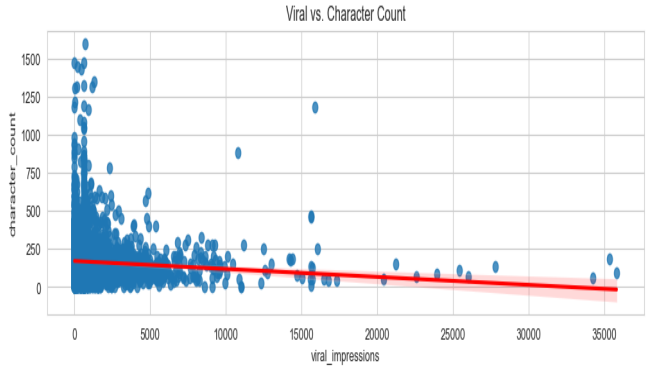
Key Questions Addressed in Exploratory Analysis

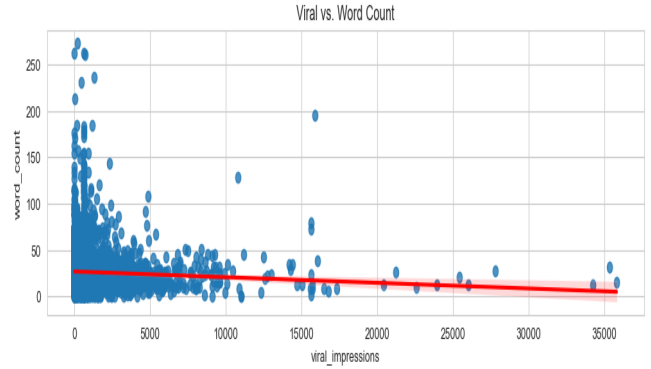
Are Impressions Seasonal?

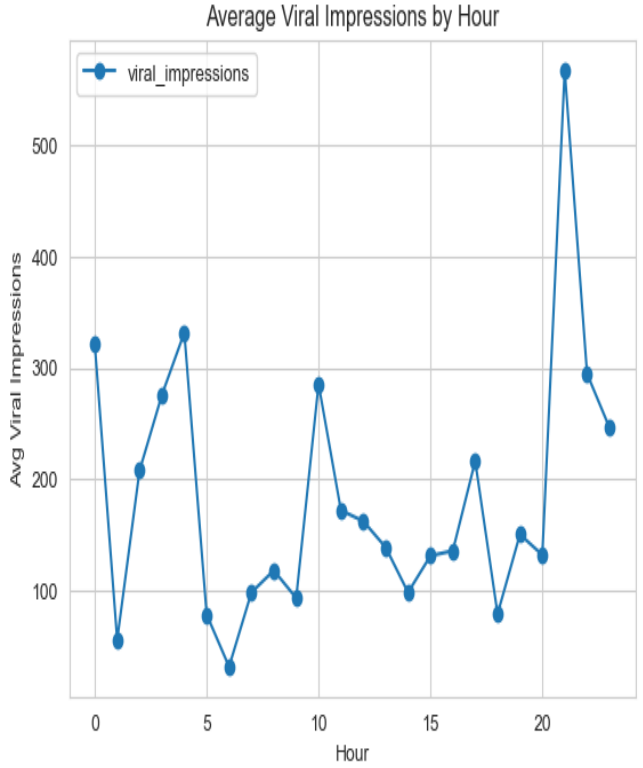
- When grouped by months, Impressions seem to be on an upward trend between November and April (the dry months in Nigeria's seasonal calendar), with the exception of a dip in February.
- The lowest months for impressions are the months between May and November (the wet months in Nigeria's seasonal calendar), with exception of a peak in July.
- This implies that impressions are seasonal.

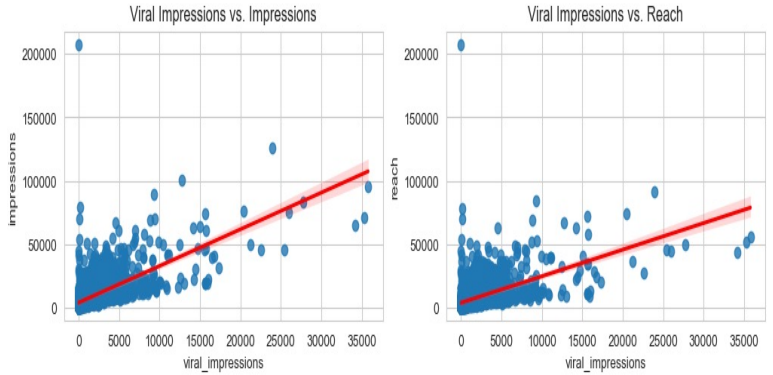


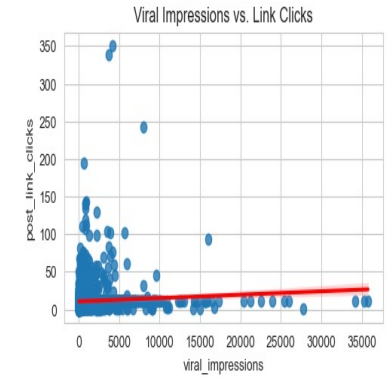
Please see the jupyter notebook for more information.

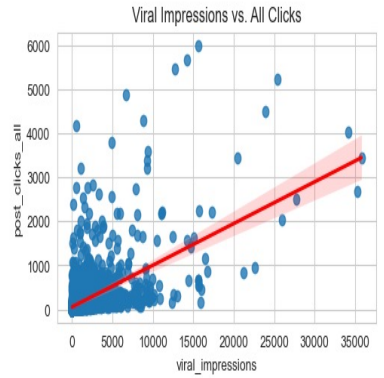








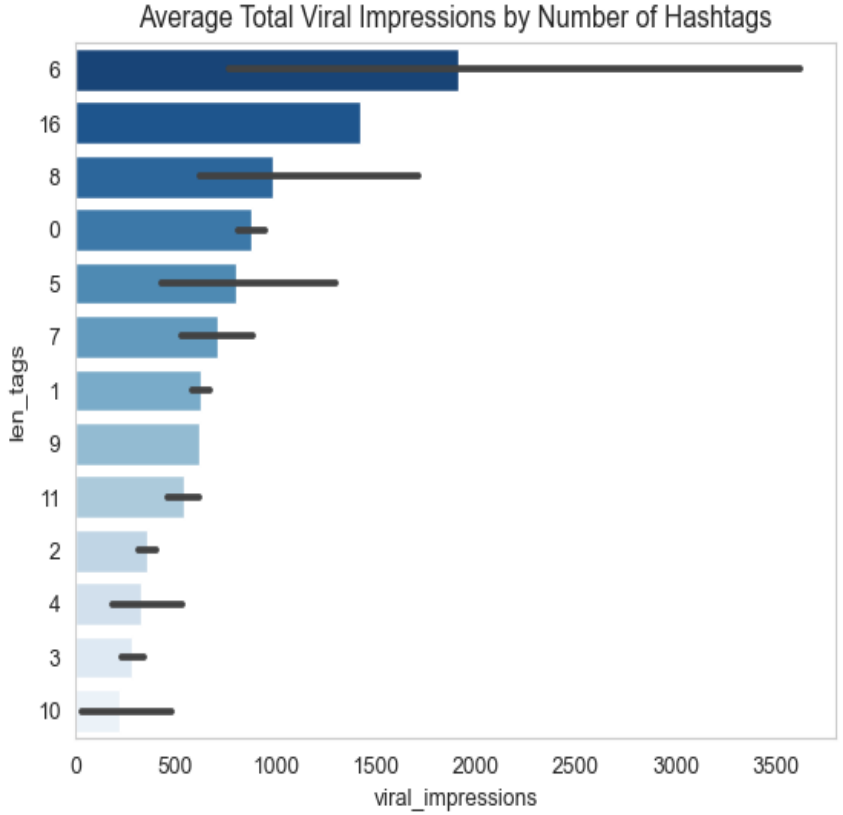


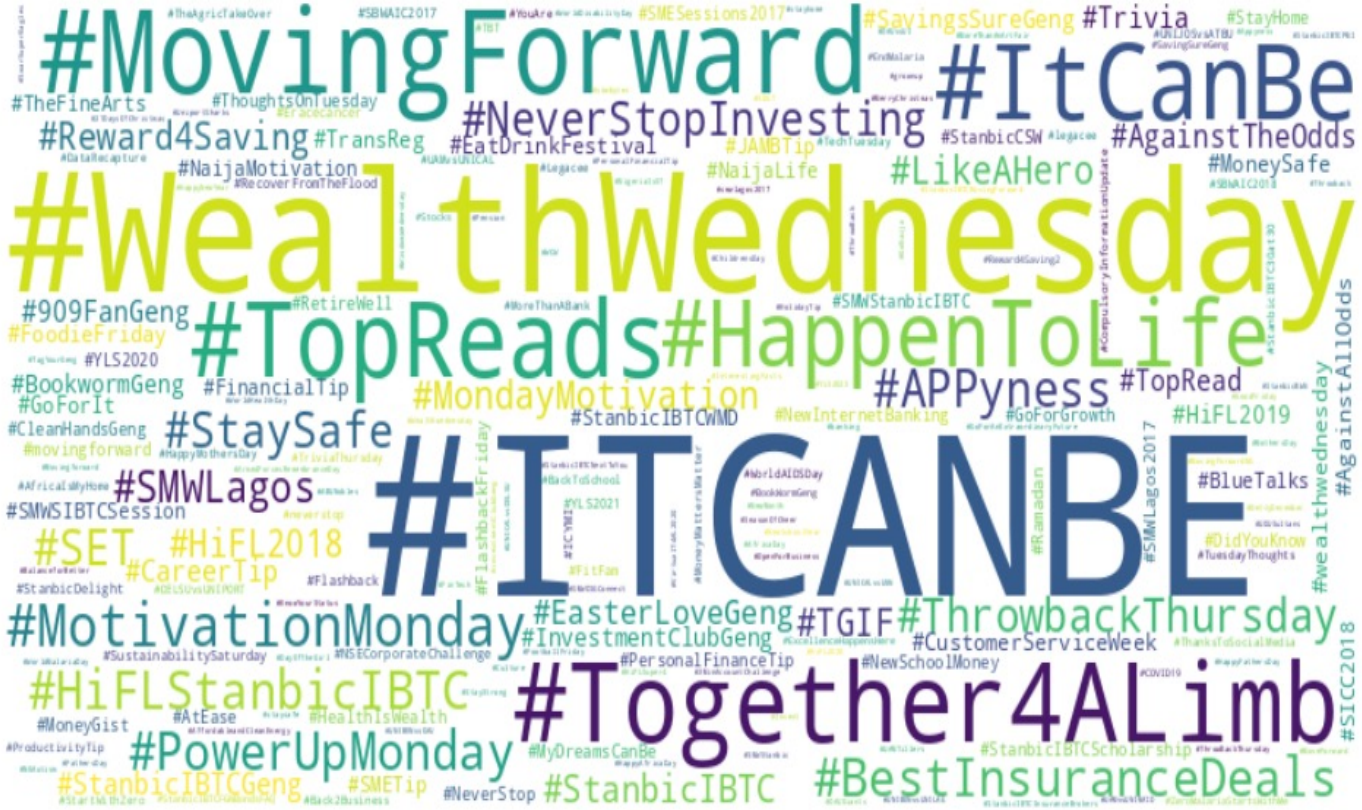


What trends favor virality?

- Virality is not strongly correlated with word or character counts.
- On average the most viral posts are those made around 9pm.
- Impressions, reach and all clicks are positively correlated with virality. However, this does not tell us which variable is dependent on the other.
- Expanding impressions, reach and posting interesting content during evening time generally seems to positively favour virality.

Please see the jupyter notebook for more information.





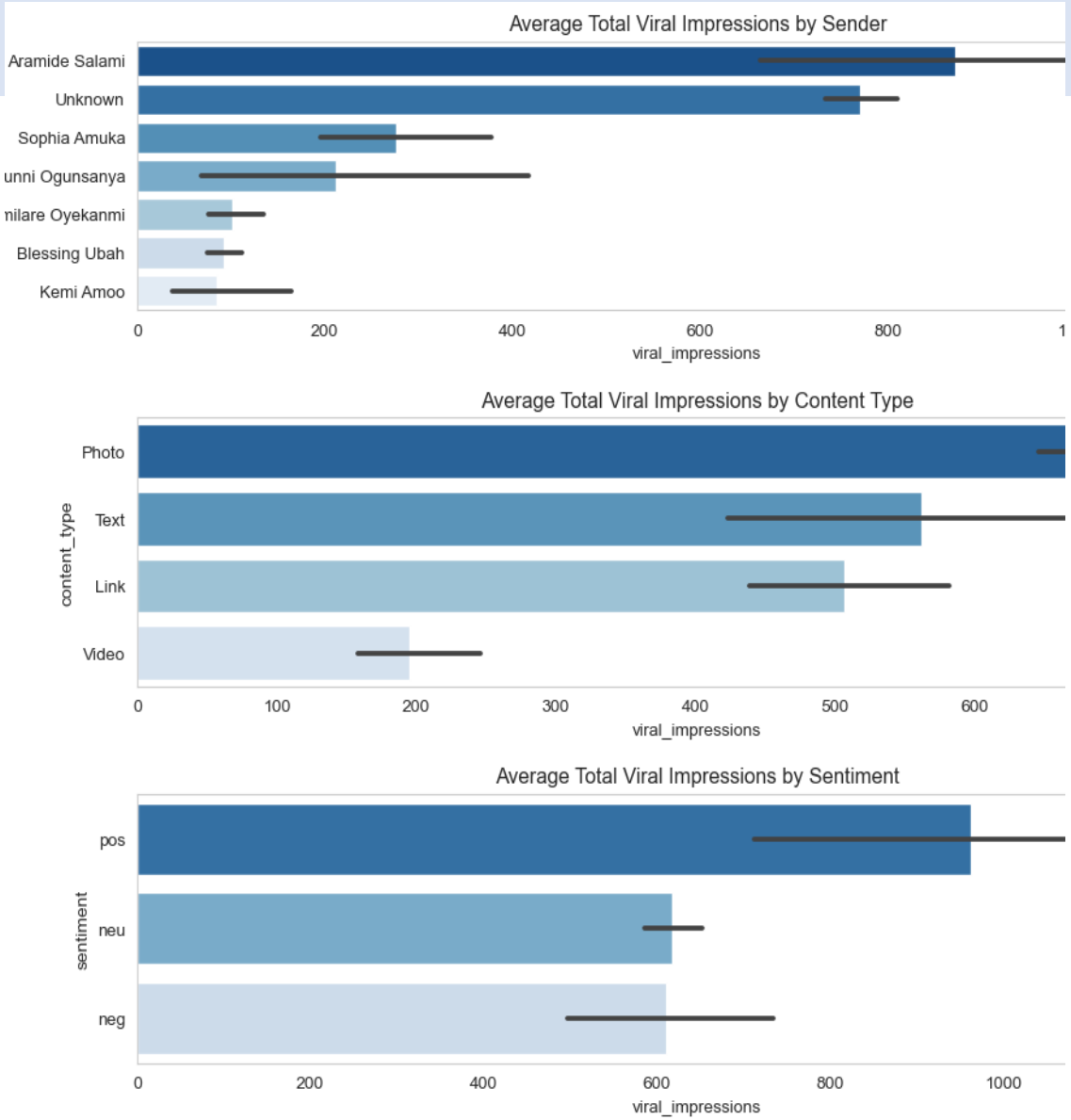
What Hashtags follow Above Average Viral Posts

- The word cloud of hashtags for posts with viral values in excess of the average looks no different than the general word cloud for hashtags of the whole dataset.
- This may imply that hashtags are not a clear indicator of virality.
- However, more viral posts have about 6 hashtags.

Please see the jupyter notebook for more information.

Which Categorical Variables Favour Virality?

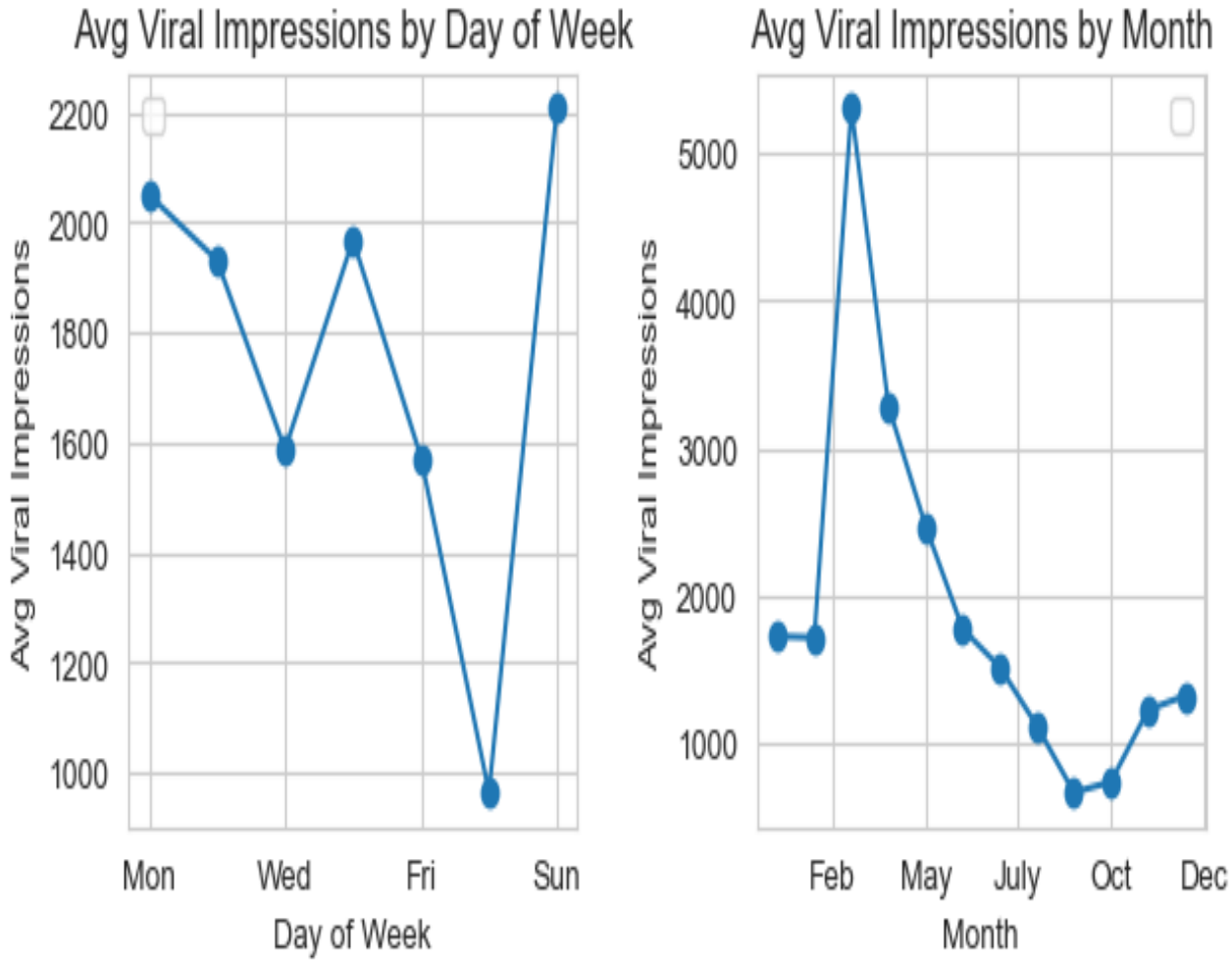
- On average, Aramide Salami is the sender whose posts have the highest amount of virality.
- Photo posts generally are more viral than others.
- Posts with a positive sentiment are more likely to go viral.



Please see the jupyter notebook for more information.

When do Posts Go Viral?

- On average, posts sent on Sundays are the most viral.
- The month of March is the hottest month for virality.



Please see the jupyter notebook for more information.

Section 4

Insights Drawn From Statistical Analysis

Statistical Analysis - Hypothesis Testing

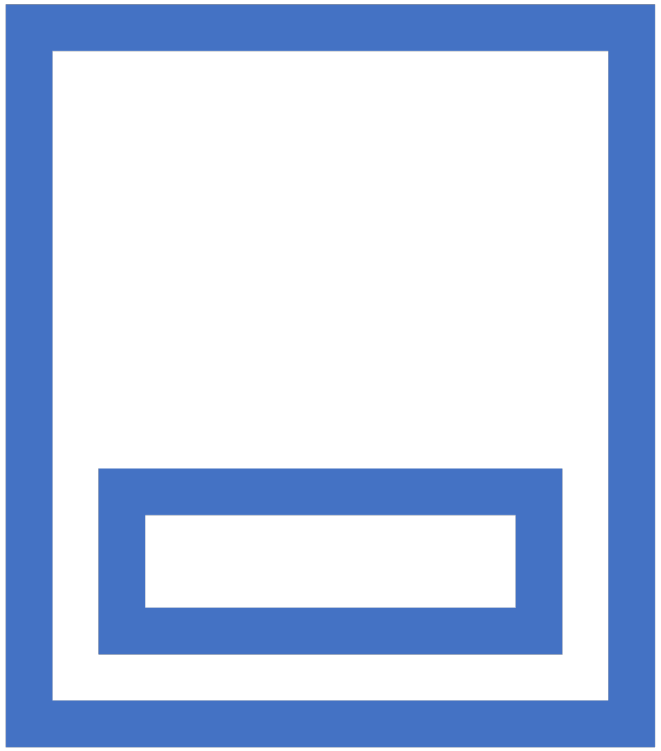
- Statistical tests are a set of mathematical techniques used to make inferences about populations or data based on a sample of that data.
- They are used to draw conclusions about the characteristics of a population or the relationship between variables.
- Our exploratory analysis has raised several hypotheses about the relationship between variables in our dataset.
- We'll use statistical tests to test the significance of those theses and make inferences from them, including ANOVA, Regression Analysis, Correlation & T/U-Tests.

Please see the jupyter notebook for more information.

Statistical Tests - ANOVA

- **Analysis of Variance (ANOVA)**, which tests whether there are any statistically significant differences among the means of the numeric variable across the categorical groups. The ANOVA test returns an F-statistic and a p-value.
- The F-statistic indicates how much the variance in the dependent variable can be attributed to the independent or categorical variable. Higher F-values suggest larger differences among group means relative to the variation within each group implying that there might be significant differences among group means.
- The p-value tells us the probability of observing the observed F-statistic by random chance. Smaller p-values (typically less than a chosen significance level, often 0.05) indicate stronger evidence that the F-statistic is not due to chance, i.e., it is statistically significant. This indicates that there are significant differences among the group means.

Please see the jupyter notebook for more information.



Statistical Tests – Correlation & T/U-Tests

- **Correlation analysis**, which measures the strength and direction of the linear relationship between two continuous numeric variables. The Pearson correlation coefficient (r) is commonly used for normally distributed data, while the Spearman rank correlation coefficient (ρ) is suitable for non-normally distributed data or when the relationship may be nonlinear.
- **T-Test or Mann-Whitney U Test**, which compares the means of two numeric variables between two groups to determine whether there is a significant difference between the two groups. T-test is for normally distributed data and Mann-Whitney U test is for non-normally distributed data. A higher absolute statistic value (whether t-statistic or U-statistic) is often associated with a smaller p-value, which indicates stronger evidence against the null hypothesis, i.e., we would accept the visuals from our plot.

Please see the jupyter notebook for more information.

Statistical Tests – Regression Analysis

Regression analysis, which models the relationship between two numeric variables by fitting a linear equation to the data. It helps us understand how one variable changes as the other changes. Linear regression coefficients helps us assess the strength and direction of the relationship.

The results to be interpreted from a regression analysis include coefficients, standard errors, T-statistics & p-values, r-squared, F-statistic, residuals, Durbin-Watson statistic, confidence intervals and assumption checks.

Coefficients, which represents the change in a dependent variable that is associated with a one-unit change in the corresponding independent variable, holding all other variables constant.
Standard errors, which measure the variability or uncertainty in the coefficient estimates. Smaller errors indicate more precise estimates.
T-statistics & P-values, which measures how many standard errors the coefficient is away from 0.
R-squared, which measures the proportion of variance in the dependent variable explained by the independent variable. A higher r-squared indicates that a larger portion of the variance in the dependent variable is explained by the independent variables.
F-statistic, which tests the overall significance of the model. A low p-value for the F-statistic suggests that the model is statistically significant.
Residuals, which are the difference between observed and predicted values. Ideally, residuals should be randomly distributed and exhibit no systemic patterns.
Durbin-Watson statistic, which checks for autocorrelation in the residuals. A value close to 2 suggests no significant autocorrelation.
Confidence Interval, which provide a range within which the true population parameter (the coefficient) is likely to fall. It's often more informative than just looking at point estimates (coefficients)..
Assumption checks, which check for the model's adherence to the assumptions of linearity, independence of errors, homoscedasticity (constant variance of residuals), and normality of residuals.

Please see the jupyter notebook for more information.

Is Content Type Related to any Engagement Metric?

- In the EDA section on Categorical Features – Content Type, we saw that on average, Poll was the highest content type by Impressions and Engagements, Text was the highest content type by Reach.
- When this relationship is tested for statistical significance, the relationship between engagements and content type is not statistically significant and the chart should be rejected for inference.
- Between impressions and reach, the relationship between reach and content type is more statistically significant.
- We can therefore infer that text posts are more likely to have wider reach Facebook provides the highest impression and engagements and Twitter provides the highest reach.

```
Impressions v. Content Type
F-Statistic: 44.904
P-Value: 4.685334090909321e-55
```

```
Engagements v. Content Type
F-Statistic: 1.201
P-Value: 0.3020611620705095
```

```
Reach v. Content Type
F-Statistic: 195.701
P-Value: 1.973390643751267e-246
```

Please see the jupyter notebook for more information.

Which Network yields the highest Engagement?

- In the EDA section on Categorical Features – Network, we saw that on average, Facebook provides the highest impression and engagements, and Twitter provides the highest reach.
- When these relationships are tested for statistical significance, all of them are statistically significant, inferring that network explains some variance (has some relationship) within each engagement variable.
- The smallest explained variance is with Engagements and the highest is with Reach.
- We can therefore infer that Facebook will likely yield the highest impressions and Twitter will yield the highest Reach.

```
Impressions v. Social Media Platform
F-Statistic: 2260.482
P-Value: 0.0
```

```
Engagements v. Social Media Platform
F-Statistic: 107.516
P-Value: 2.6683347837541804e-69
```

```
Reach v. Social Media Platform
F-Statistic: 64477.094
P-Value: 0.0
```

Please see the jupyter notebook for more information.

Which Sender posts the most engaging posts?

- In the EDA section on Categorical Features – Sender, we saw that on average, the highest impressions and engagements are from posts sent by Damilare Oyekanmi and the highest reach are from posts sent by Kanayo Obiano, Lilian Ibekwe, Philip Nwagwunor and Rebecca Oyebode.
- When these relationships are tested for statistical significance, all of them are statistically significant, inferring that sender explains some variance (has some relationship) within each engagement variable.
- The smallest explained variance is with Engagements and the highest is with Reach.
- We can therefore infer that posts by Damilare Oyekanmi will likely yield the highest impressions and posts from Kanayo Obiano, Lilian Ibekwe, Philip Nwagwunor and Rebecca Oyebode will yield the highest Reach.

```
Impressions v. Social Media Platform
F-Statistic: 2260.482
P-Value: 0.0
```

```
Engagements v. Sender
F-Statistic: 13.268
P-Value: 1.1386325334241122e-25
```

```
Reach v. Sender
F-Statistic: 712.846
P-Value: 0.0
```

Please see the jupyter notebook for more information.

What are the peak times for user engagement?

- In the EDA section on Engagement Metrics by Hour, we saw that on average when aggregated by Hour, engagements see an upward trend as the day progresses across all engagement metrics.
- When tested using Spearman's rank correlation, we see that hour is only slightly correlated with any engagement metric.
- Regression analysis also showed that there is no linear relationship that can be modeled between hour and any engagement metric.
- The combination of high u-statistic and low p-value infers that there is a significant and consistent difference in each metric when grouped by hour. This may infer that Hour is a predictor of each engagement metrics and 9pm is a great time for posting.

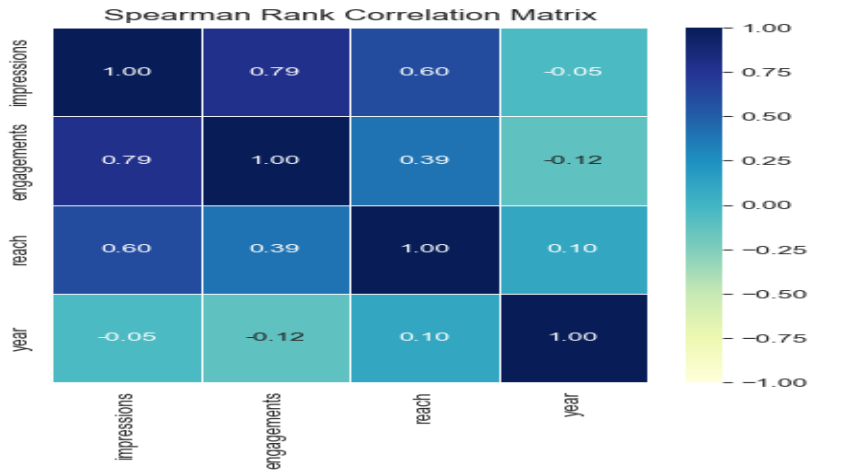
```
Impressions v. Hour
Spearman's rho (ρ): 0.068
P-value: 7.1104869322006 47e-38

Engagements v. Hour
Spearman's rho (ρ): 0.054
P-value: 1.3194053181149577e-24

Reach v. Hour
Spearman's rho (ρ): 0.030
P-value: 7.01850083588354 95e-09
```

```
Impressions v. Hour
U-statistic: 58905390.000
P-value: 0.0

Engagements v. Hour
U-statistic: 137914759.500
P-value: 0.0

Reach v. Hour
U-statistic: 422413876.500
P-value: 0.0
```

Please see the jupyter notebook for more information.

Is Year a predictor of user engagement?

- In the EDA section on Engagement Metrics by Year, we saw that on average when aggregated by Year, impressions and engagements see an upward trend as the year progresses and reach sees the reverse.
- When tested using Spearman's rank correlation, we see that there is a more significant correlation between year and all engagement metrics than compared with Hour.
- Regression analysis also showed that there is no linear relationship that can be modeled between year and any engagement metric.
- The combination of high u-statistic and low p-value infers that there is a significant and consistent difference in each metric when grouped by year. This may infer that Year is a predictor of each engagement metrics and each later will on average see higher engagement metrics.
- Year is likely a better time-period predictor for impression metrics than other time periods.

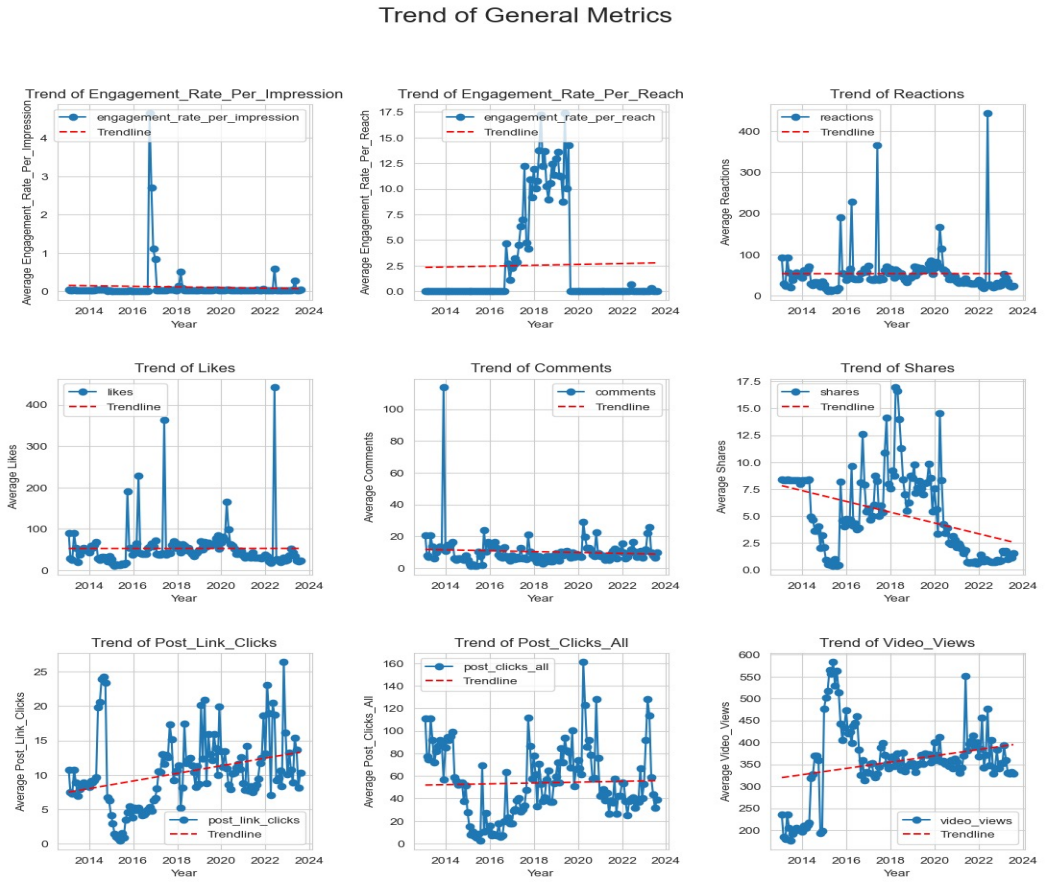



```
Impressions v. Year
U-statistic: 613749293.500
P-value: 3.755335723181876e-41


Engagements v. Year
U-statistic: 1297790323.500
P-value: 0.0


Reach v. Year
U-statistic: 633116653.500
P-value: 7.1179546569044 22e-11
```

Please see the jupyter notebook for more information.

Testing Trends in General Metrics

- For other general metrics beyond engagements, we see a significant upward trend in link clicks and video views and a significant decline in shares. Other metrics have flat trendlines or only slight inclines.



Please see the jupyter notebook for more information.

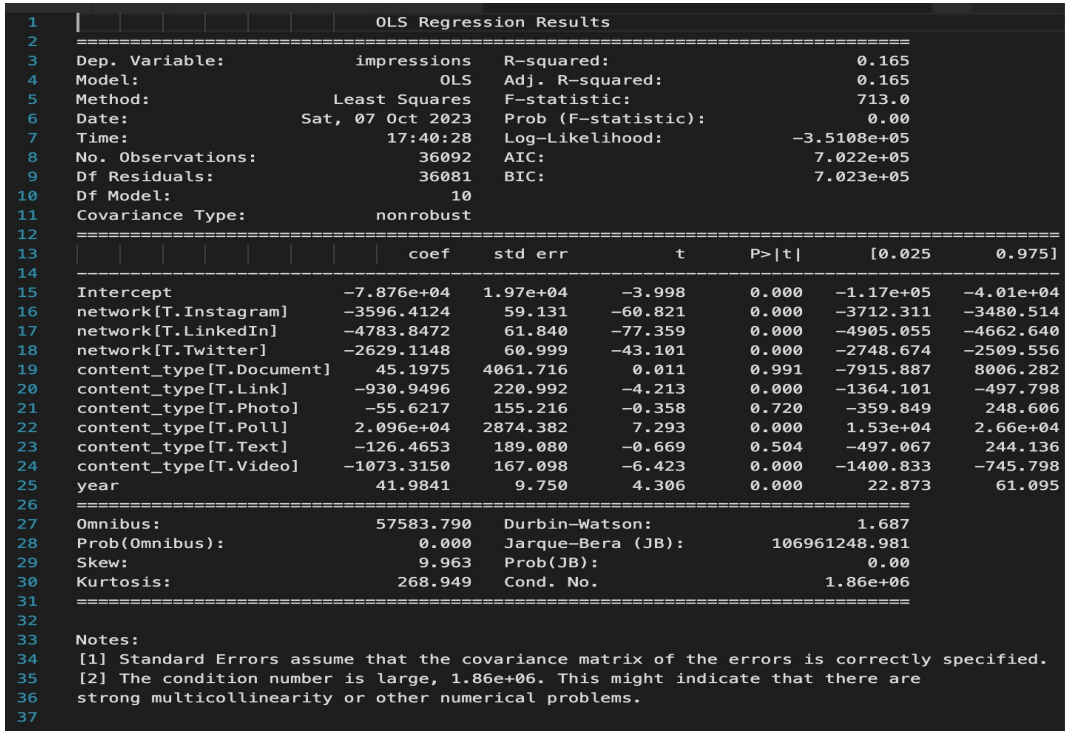
Testing Trends in General Metrics - Correlation

- Pearson correlation poorly assesses the relationship between variables because they are not normally distributed.
- When tested using spearman rank correlation, we find that year is more correlated with general metrics than hour.
- In particular, the highest correlations are between year and comments, likes, reactions and shares.
- Many metrics are correlated with each other and explaining the variance in certain variables requires models that combine many independent variables. In regression analysis, this is called Multiple Linear Regression.

Please see the jupyter notebook for more information.

Testing Trends in General Metrics – Regression Analysis - Impressions

- The model with the highest r squared for modeling the impressions feature is the model with parameters as network, content type and year. R-squared is low at 16.5%
- Parameter coefficients for document, photo and text content type are statistically insignificant, meaning that their coefficient values are no different from zero.
- According to this model, all features except Poll content type and year negatively impact the values in the impression feature.
- The model's F-statistic (713, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters is related to the impressions feature.
- The confidence intervals for most statistically significant features are left of zero on the number line, i.e., they are negative values.
- The standard errors on a few significant features are high implying that any predictions from the model will be far from the observed values.
- The Durbin-Watson statistic (1.687) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.



OLS Regression Results

Dep. Variable:	impressions	R-squared:	0.165
Model:	OLS	Adj. R-squared:	0.165
Method:	Least Squares	F-statistic:	713.0
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	17:40:28	Log-Likelihood:	-3.5108e+05
No. Observations:	36092	AIC:	7.022e+05
Df Residuals:	36081	BIC:	7.023e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.876e+04	1.97e+04	-3.998	0.000	-1.17e+05	-4.01e+04
network[T.Instagram]	-3596.4124	59.131	-60.821	0.000	-3712.311	-3480.514
network[T.LinkedIn]	-4783.8472	61.840	-77.359	0.000	-4905.055	-4662.640
network[T.Twitter]	-2629.1148	60.999	-43.101	0.000	-2748.674	-2509.556
content_type[T.Document]	45.1975	4061.716	0.011	0.991	-7915.887	8006.282
content_type[T.Link]	-930.9496	220.992	-4.213	0.000	-1364.101	-497.798
content_type[T.Photo]	-55.6217	155.216	-0.358	0.720	-359.849	248.606
content_type[T.Poll]	2.096e+04	2874.382	7.293	0.000	1.53e+04	2.66e+04
content_type[T.Text]	-126.4653	189.080	-0.669	0.504	-497.067	244.136
content_type[T.Video]	-1073.3150	167.098	-6.423	0.000	-1400.833	-745.798
year	41.9841	9.750	4.306	0.000	22.873	61.095

Omnibus:	57583.790	Durbin-Watson:	1.687
Prob(Omnibus):	0.000	Jarque-Bera (JB):	106961248.981
Skew:	9.963	Prob(JB):	0.00
Kurtosis:	268.949	Cond. No.	1.86e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.86e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Please see the jupyter notebook for more information.

Testing Trends in General Metrics – Regression Analysis - Reach

- The model with the highest r squared for modeling the reach feature is the model with parameters as network, content type and year. R squared is high at 85.3%
- Parameter coefficients for document, photo and poll content type are statistically insignificant, meaning that their coefficient values are no different from zero.
- All features except Twitter network, video content type and year negatively impact the values in the reach feature.
- The model's F-statistic (>2k, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters is related to the reach feature.
- The confidence intervals for most statistically significant features are left of zero on the number line, i.e., they are negative values.
- The standard errors are generally high implying that any predictions from the model will be far from the observed values.
- The Durbin-Watson statistic (1.8) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.
- However, this model confirms the charts we've plotting regarding the Reach variable.

Dep. Variable:	reach	R-squared:	0.853
Model:	OLS	Adj. R-squared:	0.853
Method:	Least Squares	F-statistic:	2.090e+04
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	13:02:18	Log-Likelihood:	-4.4129e+05
No. Observations:	36092	AIC:	8.826e+05
Df Residuals:	36081	BIC:	8.827e+05
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-9.198e+06	2.4e+05	-38.346	0.000	-9.67e+06	-8.73e+06
network[T.Instagram]	-3792.9583	720.005	-5.268	0.000	-5204.190	-2381.727
network[T.LinkedIn]	-7646.9700	752.988	-10.156	0.000	-9122.848	-6171.092
network[T.Twitter]	2.786e+05	742.748	375.151	0.000	2.77e+05	2.8e+05
content_type[T.Document]	-1.134e+04	4.95e+04	-0.229	0.819	-1.08e+05	8.56e+04
content_type[T.Link]	-1.411e+04	2690.902	-5.245	0.000	-1.94e+04	-8838.384
content_type[T.Photo]	2890.2890	1889.979	1.529	0.126	-814.127	6594.705
content_type[T.Poll]	-1.134e+04	3.5e+04	-0.324	0.746	-7.99e+04	5.73e+04
content_type[T.Text]	-2.516e+04	2302.319	-10.927	0.000	-2.97e+04	-2.06e+04
content_type[T.Video]	5184.0712	2034.665	2.548	0.011	1196.068	9172.075
year	4556.0084	118.723	38.375	0.000	4323.307	4788.709

Omnibus:	64051.678	Durbin-Watson:	1.800
Prob(Omnibus):	0.000	Jarque-Bera (JB):	160888250.250
Skew:	12.502	Prob(JB):	0.00
Kurtosis:	329.129	Cond. No.	1.86e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.86e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Please see the jupyter notebook for more information.

Testing Trends in General Metrics – Regression Analysis – Engagement Rate per Reach

- The model with the highest r squared for modeling the engagement rate per reach feature is the model with parameters as network, content type, impressions, reach, engagements, reactions and likes. R squared is low at 22.2%
- This model has been filtered to remove parameters with insignificant coefficients. This raises concerns of whether the model is overfit to its parameters and if it can generalize to the fuller dataset.
- All features except Instagram network, impressions and reactions negatively impact the values in the engagement rate per reach feature.
- The model's F-statistic (616, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters is related to the reach feature.
- The confidence intervals for most statistically significant features are left of zero on the number line, i.e., they are negative values.
- The standard errors are generally low implying that any predictions from the model will be close to the observed values.
- The Durbin-Watson statistic (1.895) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.

OLS Regression Results

Dep. Variable:	engagement_rate_per_reach	R-squared:	0.222
Model:	OLS	Adj. R-squared:	0.221
Method:	Least Squares	F-statistic:	616.0
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	17:40:36	Log-Likelihood:	-86656.
No. Observations:	19464	AIC:	1.733e+05
Df Residuals:	19454	BIC:	1.734e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.9107	0.843	5.828	0.000	3.259	6.562
network[T.Instagram]	4.2013	0.338	12.430	0.000	3.539	4.864
content_type[T.Link]	-4.3113	1.710	-2.521	0.012	-7.663	-0.960
content_type[T.Photo]	-3.2538	0.796	-4.088	0.000	-4.814	-1.694
content_type[T.Video]	-4.2951	0.913	-4.706	0.000	-6.084	-2.506
impressions	0.0080	0.000	62.210	0.000	0.008	0.008
reach	-0.0091	0.000	-64.553	0.000	-0.009	-0.009
engagements	-0.0080	0.001	-7.184	0.000	-0.010	-0.006
reactions	0.0462	0.014	3.367	0.001	0.019	0.073
likes	-0.0372	0.014	-2.749	0.006	-0.064	-0.011

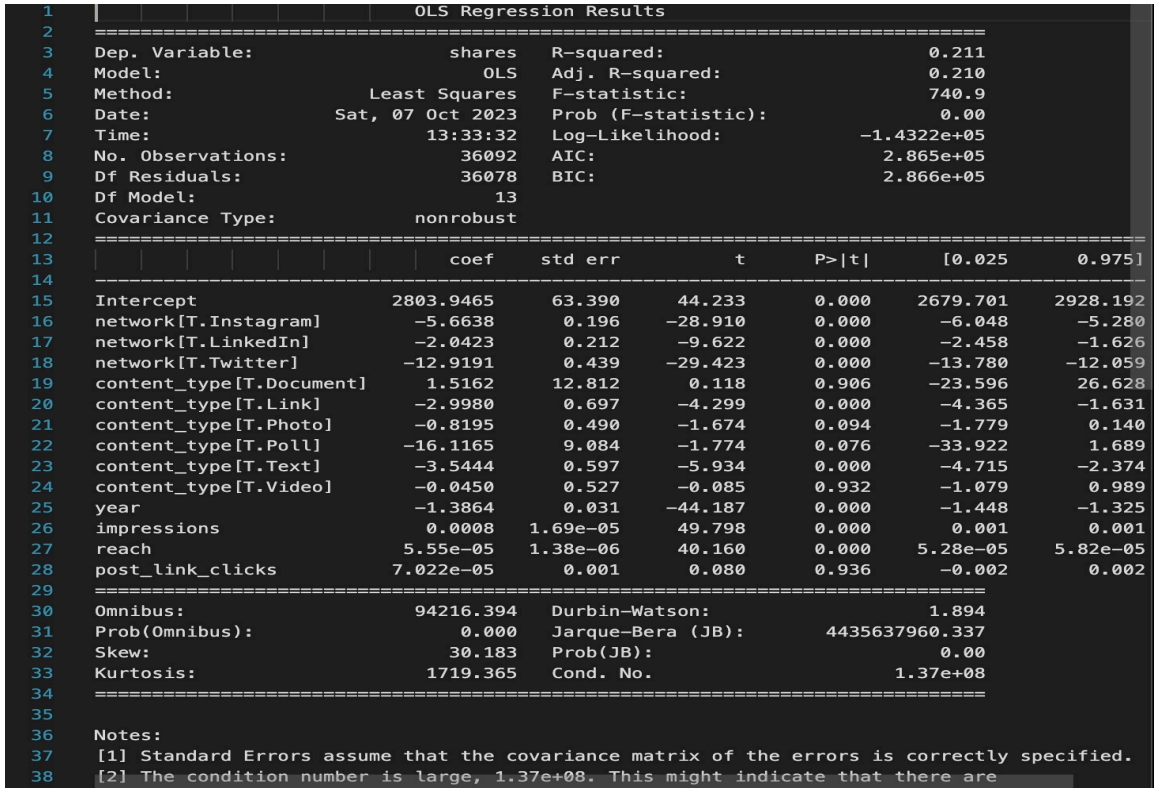
Omnibus:	21837.189	Durbin-Watson:	1.895
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4322460.936
Skew:	5.523	Prob(JB):	0.00
Kurtosis:	75.165	Cond. No.	1.12e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified
[2] The condition number is large, 1.12e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Please see the jupyter notebook for more information.

Testing Trends in General Metrics – Regression Analysis – Shares

- The model with the highest r squared for modeling the engagement rate per reach feature is the model with parameters as network, content type, year, impressions, reach and link clicks. R squared is low at 21.1%
- All features except impressions negatively impact the values in the engagement rate per reach feature.
- The model's F-statistic (740, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters is related to the reach feature.
- The confidence intervals for most statistically significant features are left of zero on the number line, i.e., they are negative values.
- The standard errors are generally low implying that any predictions from the model will be close to the observed values.
- The Durbin-Watson statistic (1.894) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.



OLS Regression Results

Dep. Variable:	shares	R-squared:	0.211
Model:	OLS	Adj. R-squared:	0.210
Method:	Least Squares	F-statistic:	740.9
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	13:33:32	Log-Likelihood:	−1.4322e+05
No. Observations:	36092	AIC:	2.865e+05
Df Residuals:	36078	BIC:	2.866e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2803.9465	63.390	44.233	0.000	2679.701	2928.192
network[T.Instagram]	−5.6638	0.196	−28.910	0.000	−6.048	−5.280
network[T.LinkedIn]	−2.0423	0.212	−9.622	0.000	−2.458	−1.626
network[T.Twitter]	−12.9191	0.439	−29.423	0.000	−13.780	−12.059
content_type[T.Document]	1.5162	12.812	0.118	0.906	−23.596	26.628
content_type[T.Link]	−2.9980	0.697	−4.299	0.000	−4.365	−1.631
content_type[T.Photo]	−0.8195	0.490	−1.674	0.094	−1.779	0.140
content_type[T.Poll]	−16.1165	9.084	−1.774	0.076	−33.922	1.689
content_type[T.Text]	−3.5444	0.597	−5.934	0.000	−4.715	−2.374
content_type[T.Video]	−0.0450	0.527	−0.085	0.932	−1.079	0.989
year	−1.3864	0.031	−44.187	0.000	−1.448	−1.325
impressions	0.0008	1.69e−05	49.798	0.000	0.001	0.001
reach	5.55e−05	1.38e−06	40.160	0.000	5.28e−05	5.82e−05
post_link_clicks	7.022e−05	0.001	0.080	0.936	−0.002	0.002

Omnibus:	94216.394	Durbin-Watson:	1.894
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4435637960.337
Skew:	30.183	Prob(JB):	0.00
Kurtosis:	1719.365	Cond. No.	1.37e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.37e+08. This might indicate that there are

Please see the jupyter notebook for more information.

Testing Trends in General Metrics – Regression Analysis – Link Clicks

- The model with the highest r squared for modeling the link clicks feature is the model with parameters as network and content type. R squared is very high at 96.4%
- This model has been filtered to remove parameters with insignificant coefficients. This raises concerns of whether the model is overfit to its parameters and if it can generalize to the fuller dataset.
- The model's F-statistic (9958, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters is related to the reach feature.
- The standard errors are very low implying that any predictions from the model will be close to the observed values.
- The Durbin-Watson statistic (2.0) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.
- However, the sample size is about 2% of the fuller dataset.

OLS Regression Results

Dep. Variable:	post_link_clicks	R-squared:	0.964
Model:	OLS	Adj. R-squared:	0.964
Method:	Least Squares	F-statistic:	9958.
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	17:41:06	Log-Likelihood:	-2418.5
No. Observations:	745	AIC:	4843.
Df Residuals:	742	BIC:	4857.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.999e-16	0.229	-8.74e-16	1.000	-0.449	0.449
network[T.LinkedIn]	282.6667	2.548	110.937	0.000	277.665	287.669
content_type[T.Document]	-31.6667	4.406	-7.187	0.000	-40.316	-23.017
content_type[T.Poll]	314.3333	3.598	87.369	0.000	307.270	321.396

Omnibus:	393.767	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4238126.510
Skew:	-0.000	Prob(JB):	0.00
Kurtosis:	372.500	Cond. No.	4.64e+17

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.46e-33. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.

Please see the jupyter notebook for more information.

Testing Trends in General Metrics – Regression Analysis – Link Clicks 2

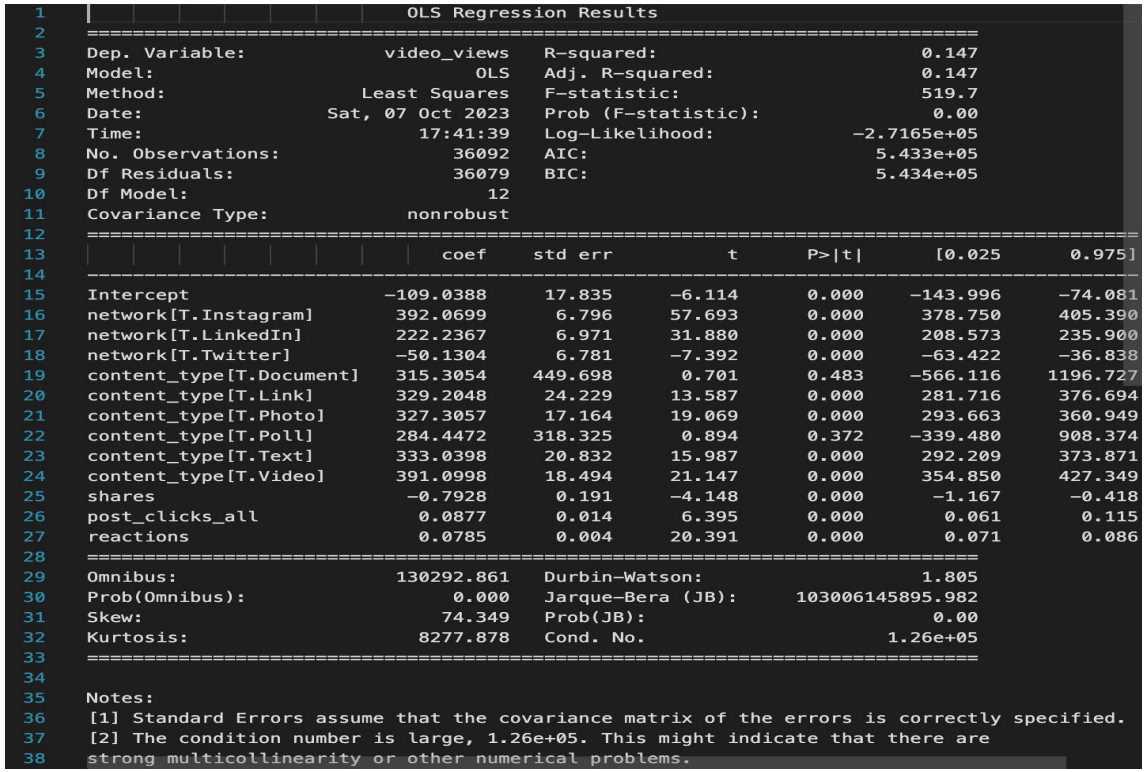
- When we generalize the model to the fuller dataset, the r—squared drops to 2%. This supports the theory that the previous model was overfit.
- However, standard error remains low, f-statistic remains significant, and Durbin-Watson statistic remains near perfect.
- This implies that the independent features explain some of the variance in the links click feature however, the relationship is likely non-linear and using a line equation to model the dataset is inadequate

```
                            OLS Regression Results
==============================================================================
Dep. Variable:       post_link_clicks   R-squared:                       0.026
Model:                           OLS   Adj. R-squared:                  0.025
Method:                Least Squares   F-statistic:                     105.1
Date:               Sat, 07 Oct 2023   Prob (F-statistic):          3.65e-195
Time:                       13:42:08   Log-Likelihood:            -2.0790e+05
No. Observations:              36092   AIC:                         4.158e+05
Df Residuals:                  36082   BIC:                         4.159e+05
Df Model:                          9
Covariance Type:           nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       10.7005      3.033      3.528      0.000       4.756      16.645
network[T.Instagram]           -10.7005      1.117     -9.579      0.000     -12.890      -8.511
network[T.LinkedIn]             21.3444      1.168     18.279      0.000      19.056      23.633
network[T.Twitter]              -3.5871      1.154     -3.107      0.002      -5.850      -1.324
content_type[T.Document]       218.9552     76.874      2.848      0.004      68.281     369.630
content_type[T.Link]             1.7940      4.142      0.433      0.665      -6.323       9.911
content_type[T.Photo]            0.5448      2.934      0.186      0.853      -5.206       6.296
content_type[T.Poll]           564.9552     54.401     10.385      0.000     458.328     671.582
content_type[T.Text]            -1.0896      3.560     -0.306      0.760      -8.067       5.888
content_type[T.Video]           -4.4892      3.161     -1.420      0.156     -10.686       1.707
==============================================================================
Omnibus:                   107070.927   Durbin-Watson:                   1.991
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       8745889946.235
Skew:                          42.651   Prob(JB):                         0.00
Kurtosis:                    2413.074   Cond. No.                         262.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Please see the jupyter notebook for more information.

Testing Trends in General Metrics – Regression Analysis – Video Views

- The model with the highest r squared for modeling the video views feature is the model with parameters as network, content type and year. R squared is low at 14.7%
- All features except Twitter network and shares positively impact the values in the engagement rate per reach feature.
- The model's F-statistic (519, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters is related to the reach feature.
- The confidence intervals for most statistically significant features are right of zero on the number line, i.e., they are positive values.
- The standard errors are generally low implying that any predictions from the model will be close to the observed values.
- The Durbin-Watson statistic (1.805) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.



OLS Regression Results

Dep. Variable:	video_views	R-squared:	0.147
Model:	OLS	Adj. R-squared:	0.147
Method:	Least Squares	F-statistic:	519.7
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	17:41:39	Log-Likelihood:	−2.7165e+05
No. Observations:	36092	AIC:	5.433e+05
Df Residuals:	36079	BIC:	5.434e+05
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	−109.0388	17.835	−6.114	0.000	−143.996	−74.081
network[T.Instagram]	392.0699	6.796	57.693	0.000	378.750	405.390
network[T.LinkedIn]	222.2367	6.971	31.880	0.000	208.573	235.900
network[T.Twitter]	−50.1304	6.781	−7.392	0.000	−63.422	−36.838
content_type[T.Document]	315.3054	449.698	0.701	0.483	−566.116	1196.727
content_type[T.Link]	329.2048	24.229	13.587	0.000	281.716	376.694
content_type[T.Photo]	327.3057	17.164	19.069	0.000	293.663	360.949
content_type[T.Poll]	284.4472	318.325	0.894	0.372	−339.480	908.374
content_type[T.Text]	333.0398	20.832	15.987	0.000	292.209	373.871
content_type[T.Video]	391.0998	18.494	21.147	0.000	354.850	427.349
shares	−0.7928	0.191	−4.148	0.000	−1.167	−0.418
post_clicks_all	0.0877	0.014	6.395	0.000	0.061	0.115
reactions	0.0785	0.004	20.391	0.000	0.071	0.086

Omnibus:	130292.861	Durbin-Watson:	1.805
Prob(Omnibus):	0.000	Jarque-Bera (JB):	103006145895.982
Skew:	74.349	Prob(JB):	0.00
Kurtosis:	8277.878	Cond. No.	1.26e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.26e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Please see the jupyter notebook for more information.

Testing Trends in General Metrics – Regression Analysis - Results

- We can confidently accept that features like year, network and content type are related to Reach.
- For other models with low r squared, we can only determine which feature are closely related to the dependent variable by plotting residual plots of each independent variable and the residuals.
- Other general metrics like engagements, engagements per impressions, reactions, likes, comments, and all clicks have poor linear models.
- This may imply that the features cannot be linearly modeled, and other types of non-linear modeling techniques should be considered.

Please see the jupyter notebook for more information.

Testing Trends in General Metrics – U - Test

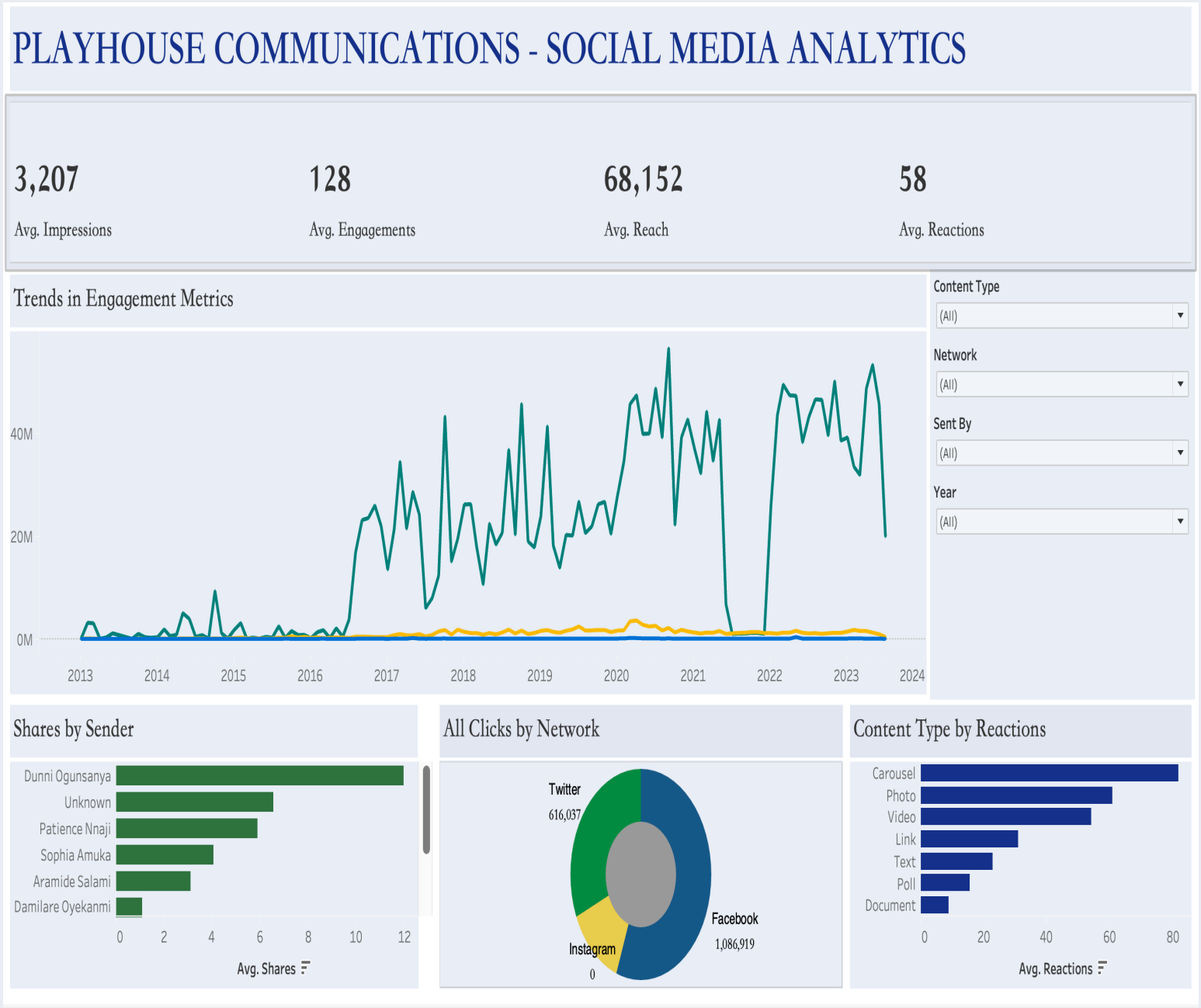
- We've already seen that the features are not normally distributed. Therefore, we'll favor the Mann-Whitney U-Test to test the significance of group means.
- From our analysis, there is there is a statistically significant difference between year and each general metric indicating that the relationship might be non-linear.
- This confirms our conclusion from the regression analysis.

```
engagement_rate_per_impression : U-Test with Year:
U-statistic: 1302632464.000, P-value: 0.0

engagement_rate_per_reach : U-Test with Year:
U-statistic: 1302632464.000, P-value: 0.0

reactions : U-Test with Year:
U-statistic: 1301297060.000, P-value: 0.0

likes : U-Test with Year:
U-statistic: 1301297060.000, P-value: 0.0

comments : U-Test with Year:
U-statistic: 1302632464.000, P-value: 0.0

shares : U-Test with Year:
U-statistic: 1302632464.000, P-value: 0.0

post_link_clicks : U-Test with Year:
U-statistic: 1302127176.000, P-value: 0.0

post_clicks_all : U-Test with Year:
U-statistic: 1300394760.000, P-value: 0.0

video_views : U-Test with Year:
U-statistic: 1299071289.500, P-value: 0.0
```

Please see the jupyter notebook for more information.

Section 5

Build a Dashboard with Tableau



Please see the Dashboard for more information.

Summary Dashboard

- We prepared an interactive dashboard in Tableau expressing the insights from the exploratory analysis.
- Visit Tableau for more information.

Section 6

Predictive Modelling

Predictive Modelling

- Our initial statistical analysis was unable to successful model the relationship between many features in the general metrics list and the correlated features.
- Our initial assumption was that these variables are not linearly related.
- We tested this assumption using predictive analytics and deploying linear and non-linear models to the entire general metrics dataset.

Please see the jupyter notebook for more information.

Predictive Modelling – Linear Models

- We modeled the data using Linear Regression and Linear Support Vector Machine Models.
- The independent features in the model where the remaining general metrics and the categorical features.

Please see the jupyter notebook for more information.

Predictive Modelling – Linear Models

- We modeled the data using Linear Regression and Linear Support Vector Machine Models.
- The independent features in the model where the remaining general metrics and the categorical features.
- Engagements, Reactions, Reach, Comments and All Clicks returned near perfect models with generally low errors.
- We used tree models to determine which features were most important to the model results.

```
Impressions
Linear Regression RMSE: 3833.82
Linear Regression R-squared: 0.11
Linear SVR RMSE: 7484.23
Linear SVR R-squared: -2.39


Engagements
Linear Regression RMSE: 8.99
Linear Regression R-squared: 1.00
Linear SVR RMSE: 14.67
Linear SVR R-squared: 1.00


Reach
Linear Regression RMSE: 41332.47
Linear Regression R-squared: 0.92
Linear SVR RMSE: 160793.78
Linear SVR R-squared: -0.23


Reactions
Linear Regression RMSE: 9.00
Linear Regression R-squared: 1.00
Linear SVR RMSE: 8.75
Linear SVR R-squared: 1.00


Comments
Linear Regression RMSE: 12.07
Linear Regression R-squared: 0.85
Linear SVR RMSE: 8.77
Linear SVR R-squared: 0.92
```

```
Shares
Linear Regression RMSE: 7.90
Linear Regression R-squared: -4.55
Linear SVR RMSE: 8.65
Linear SVR R-squared: -5.66


Post_Link_Clicks
Linear Regression RMSE: 115.11
Linear Regression R-squared: 0.24
Linear SVR RMSE: 139.01
Linear SVR R-squared: -0.11


Post_Clicks_All
Linear Regression RMSE: 9.59
Linear Regression R-squared: 1.00
Linear SVR RMSE: 9.09
Linear SVR R-squared: 1.00


Video_Views
Linear Regression RMSE: 801.12
Linear Regression R-squared: 0.06
Linear SVR RMSE: 1016.91
Linear SVR R-squared: -0.52
```

Please see the jupyter notebook for more information.

Predictive Modelling – Tree Models

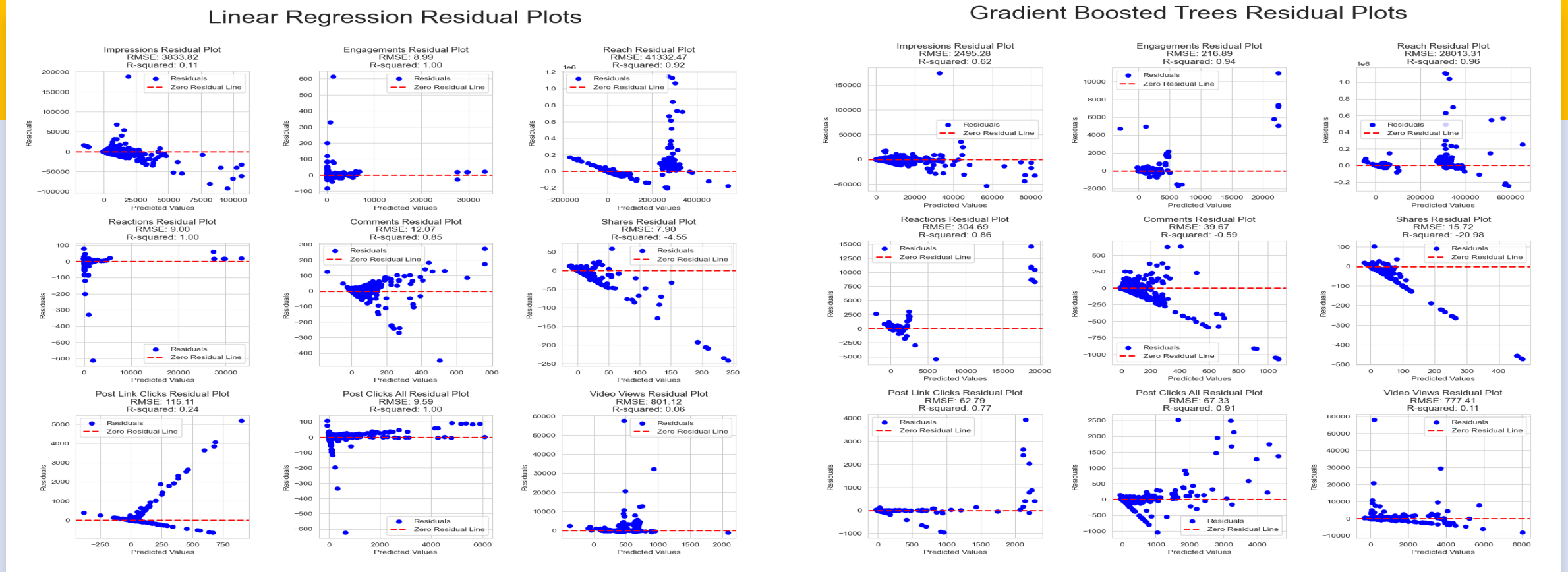
- We modeled the data using Decision Tree, Random Forest, Gradient Boosted Trees and XGBoost. The Gradient Boosted Trees performed best overall.
- All the tree models out-performed the linear models for variables with poor linear models.
- Tree model errors where on average higher than linear model errors.
- The tree models in python's Sci-kit learn library include a method that lists the most important independent variables.

```
Impressions
Decision Tree RMSE: 3045.00
Decision Tree R-squared: 0.44
Random Forest RMSE: 2669.82
Random Forest R-squared: 0.57
Gradient Boosting RMSE: 2485.39
Gradient Boosting R-squared: 0.63
XGBoost RMSE: 2640.38
XGBoost R-squared: 0.58


Engagements
Decision Tree RMSE: 346.76
Decision Tree R-squared: 0.84
Random Forest RMSE: 135.05
Random Forest R-squared: 0.98
Gradient Boosting RMSE: 215.72
Gradient Boosting R-squared: 0.94
XGBoost RMSE: 426.02
XGBoost R-squared: 0.75


Reach
Decision Tree RMSE: 29973.97
Decision Tree R-squared: 0.96
Random Forest RMSE: 28726.32
Random Forest R-squared: 0.96
Gradient Boosting RMSE: 28361.92
Gradient Boosting R-squared: 0.96
XGBoost RMSE: 28813.26
XGBoost R-squared: 0.96
```

```
Reactions
Decision Tree RMSE: 319.80
Decision Tree R-squared: 0.85
Random Forest RMSE: 172.82
Random Forest R-squared: 0.95
Gradient Boosting RMSE: 304.69
Gradient Boosting R-squared: 0.86
XGBoost RMSE: 379.87
XGBoost R-squared: 0.78


Comments
Decision Tree RMSE: 42.61
Decision Tree R-squared: -0.83
Random Forest RMSE: 35.32
Random Forest R-squared: -0.26
Gradient Boosting RMSE: 39.96
Gradient Boosting R-squared: -0.61
XGBoost RMSE: 30.33
XGBoost R-squared: 0.07


Shares
Decision Tree RMSE: 19.93
Decision Tree R-squared: -34.33
Random Forest RMSE: 17.12
Random Forest R-squared: -25.07
Gradient Boosting RMSE: 15.65
Gradient Boosting R-squared: -20.79
XGBoost RMSE: 12.88
XGBoost R-squared: -13.75
```

```
Post_Link_Clicks
Decision Tree RMSE: 62.84
Decision Tree R-squared: 0.77
Random Forest RMSE: 71.06
Random Forest R-squared: 0.71
Gradient Boosting RMSE: 63.41
Gradient Boosting R-squared: 0.77
XGBoost RMSE: 62.96
XGBoost R-squared: 0.77


Post_Clicks_All
Decision Tree RMSE: 108.35
Decision Tree R-squared: 0.77
Random Forest RMSE: 70.70
Random Forest R-squared: 0.90
Gradient Boosting RMSE: 67.38
Gradient Boosting R-squared: 0.91
XGBoost RMSE: 75.95
XGBoost R-squared: 0.89


Video_Views
Decision Tree RMSE: 785.34
Decision Tree R-squared: 0.09
Random Forest RMSE: 770.07
Random Forest R-squared: 0.13
Gradient Boosting RMSE: 779.00
Gradient Boosting R-squared: 0.11
XGBoost RMSE: 747.39
XGBoost R-squared: 0.18
```

Please see the jupyter notebook for more information.



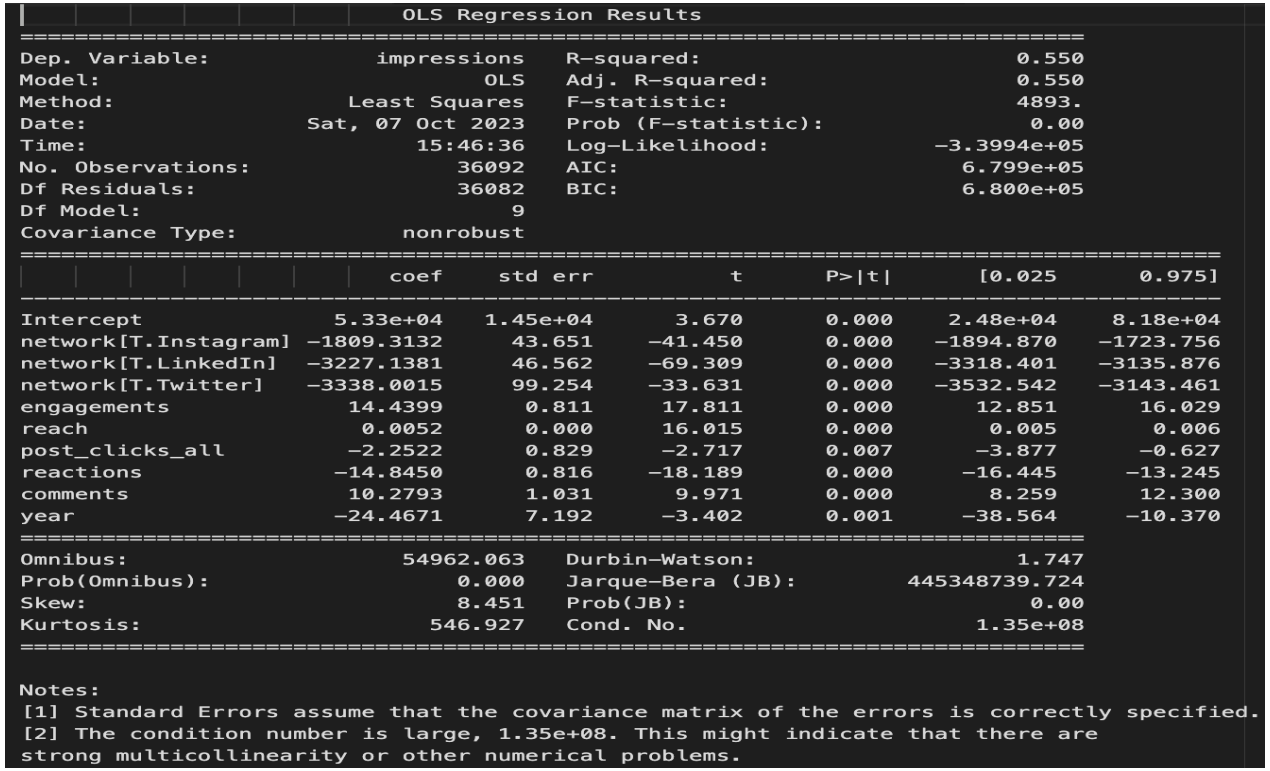
Important Features & Residual Plots

- The Residual Plots from the Linear Regression and Gradient Boosted Trees Models show heteroskedasticity and existence of outliers in the model parameters.
- This may imply that parameter estimates are inefficient and confidence intervals are unreliable.
- However, when we conduct regression analysis with the features that were important to the gradient boosted trees model, our regression metrics significantly improve across all general metrics.

Please see the jupyter notebook for more information.

Predictive Models - Regression Analysis - Impressions

- Using the important features from our predictive models, r-squared for Impressions increase by nearly 40 points to 55%
- There are no statistically insignificant parameters in this model.
- The model's F-statistic (4893, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters is related to the Impressions feature.
- The standard errors are significantly smaller in this model implying that any predictions from the model will be close to the observed values.
- The Durbin-Watson statistic (1.747) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.



OLS Regression Results

Dep. Variable:	impressions	R-squared:	0.550
Model:	OLS	Adj. R-squared:	0.550
Method:	Least Squares	F-statistic:	4893.
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	15:46:36	Log-Likelihood:	-3.3994e+05
No. Observations:	36092	AIC:	6.799e+05
Df Residuals:	36082	BIC:	6.800e+05
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.33e+04	1.45e+04	3.670	0.000	2.48e+04	8.18e+04
network[T.Instagram]	-1809.3132	43.651	-41.450	0.000	-1894.870	-1723.756
network[T.LinkedIn]	-3227.1381	46.562	-69.309	0.000	-3318.401	-3135.876
network[T.Twitter]	-3338.0015	99.254	-33.631	0.000	-3532.542	-3143.461
engagements	14.4399	0.811	17.811	0.000	12.851	16.029
reach	0.0052	0.000	16.015	0.000	0.005	0.006
post_clicks_all	-2.2522	0.829	-2.717	0.007	-3.877	-0.627
reactions	-14.8450	0.816	-18.189	0.000	-16.445	-13.245
comments	10.2793	1.031	9.971	0.000	8.259	12.300
year	-24.4671	7.192	-3.402	0.001	-38.564	-10.370

Omnibus:	54962.063	Durbin-Watson:	1.747
Prob(Omnibus):	0.000	Jarque-Bera (JB):	445348739.724
Skew:	8.451	Prob(JB):	0.00
Kurtosis:	546.927	Cond. No.	1.35e+08

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.35e+08. This might indicate that there are strong multicollinearity or other numerical problems.

Please see the jupyter notebook for more information.

Predictive Models - Regression Analysis – Engagements

- Using features relevant to the Gradient Boosted Trees prediction model, the r-squared for the Engagements feature moved 100 points from 0% to 100%.
- All parameter coefficients are statistically significant.
- The model's F-statistic (20m, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters (likely all) is related to the Engagements feature.
- The range of values for the confidence intervals is very small indicating that the model provides a precise estimate of the parameters, it fits the data well and the estimated parameter is highly reliable.
- The standard errors are very low implying that any predictions from the model will be nearly identical to the observed values.
- The Durbin-Watson statistic (2.002) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.

OLS Regression Results

Dep. Variable:	engagements	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	2.068e+07
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	15:46:55	Log-Likelihood:	-1.4867e+05
No. Observations:	36092	AIC:	2.973e+05
Df Residuals:	36087	BIC:	2.974e+05
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.0189	0.084	12.108	0.000	0.854	1.184
post_clicks_all	0.9985	0.000	2013.784	0.000	0.997	0.999
reactions	0.9996	0.000	7514.437	0.000	0.999	1.000
comments	1.0046	0.003	358.999	0.000	0.999	1.010
shares	1.0562	0.006	170.997	0.000	1.044	1.068

Omnibus:	102690.433	Durbin-Watson:	2.002
Prob(Omnibus):	0.000	Jarque-Bera (JB):	7059036550.505
Skew:	38.006	Prob(JB):	0.00
Kurtosis:	2168.237	Cond. No.	711.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Please see the jupyter notebook for more information.

Predictive Models - Regression Analysis - Reach

- This model only increases the r-squared by 1 point to 86.2%
- Parameter coefficients for video views are statistically insignificant, meaning that their coefficient values are no different from zero.
- The model's F-statistic (>2k, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters is related to the Reach feature. This is unchanged from the correlation analysis.
- The standard errors are significantly lower than the correlation-based regression analysis.
- The Durbin-Watson statistic (1.8) implies that the model's errors are independent and there is no strong case for overfitting or underfitting. This is unchanged from the correlation analysis.

OLS Regression Results

Dep. Variable:	reach	R-squared:	0.862
Model:	OLS	Adj. R-squared:	0.862
Method:	Least Squares	F-statistic:	2.057e+04
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	15:47:48	Log-Likelihood:	-4.4006e+05
No. Observations:	36092	AIC:	8.802e+05
Df Residuals:	36080	BIC:	8.803e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.211e+07	2.29e+05	-52.889	0.000	-1.26e+07	-1.17e+07
network[T.Instagram]	9011.7611	753.549	11.959	0.000	7534.782	1.05e+04
network[T.LinkedIn]	2325.7796	807.701	2.880	0.004	742.661	3908.899
network[T.Twitter]	2.778e+05	734.037	378.393	0.000	2.76e+05	2.79e+05
year	5989.1229	113.324	52.850	0.000	5767.005	6211.241
shares	912.1054	20.930	43.580	0.000	871.083	953.128
impressions	1.2351	0.084	14.680	0.000	1.070	1.400
day	147.6381	28.239	5.228	0.000	92.289	202.987
comments	-117.4595	9.350	-12.563	0.000	-135.786	-99.133
post_clicks_all	10.2339	1.909	5.360	0.000	6.492	13.976
reactions	-3.6059	0.434	-8.314	0.000	-4.456	-2.756
video_views	0.6943	0.557	1.246	0.213	-0.398	1.787

Omnibus:	59533.088	Durbin-Watson:	1.779
Prob(Omnibus):	0.000	Jarque-Bera (JB):	108239784.150
Skew:	10.737	Prob(JB):	0.00
Kurtosis:	270.423	Cond. No.	5.12e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.12e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Please see the jupyter notebook for more information.

Predictive Models - Regression Analysis – Reactions

- Using features relevant to the Gradient Boosted Trees prediction model, the r-squared for the Reactions feature moved from 0% to 94%.
- All parameter coefficients are statistically significant.
- The model's F-statistic (18k, p-value 0) is significant implying that the model is statistically significant and at least one (likely all) of the model parameters is related to the Reactions feature.
- The range of values for the confidence intervals is very small indicating that the model provides a precise estimate of the parameters, it fits the data well and the estimated parameter is highly reliable.
- The standard errors are very low implying that any predictions from the model will be nearly identical to the observed values.
- The Durbin-Watson statistic (1.901) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.

OLS Regression Results

Dep. Variable:	reactions	R-squared:	0.940
Model:	OLS	Adj. R-squared:	0.940
Method:	Least Squares	F-statistic:	1.880e+05
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	15:48:20	Log-Likelihood:	-2.3476e+05
No. Observations:	36092	AIC:	4.695e+05
Df Residuals:	36088	BIC:	4.696e+05
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-23.2141	0.907	-25.608	0.000	-24.991	-21.437
engagements	0.9827	0.001	681.358	0.000	0.980	0.986
comments	-3.0093	0.029	-103.822	0.000	-3.066	-2.952
shares	-3.3866	0.066	-50.993	0.000	-3.517	-3.256

Omnibus:	59551.186	Durbin-Watson:	1.901
Prob(Omnibus):	0.000	Jarque-Bera (JB):	280998179.845
Skew:	-10.332	Prob(JB):	0.00
Kurtosis:	434.773	Cond. No.	772.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Please see the jupyter notebook for more information.

Predictive Models - Regression Analysis – Comments

- Using features relevant to the Gradient Boosted Trees prediction model, the r-squared for the Comments feature moved from 0% to 41%.
- The model's F-statistic (837, p-value 0) is significant (and higher than the correlation-based model) implying that the model is statistically significant and at least one of the model parameters is related to the reach feature.
- The standard errors for significant features are generally low implying that any predictions from the model will be close to the observed values.
- The Durbin-Watson statistic (1.8) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.

OLS Regression Results

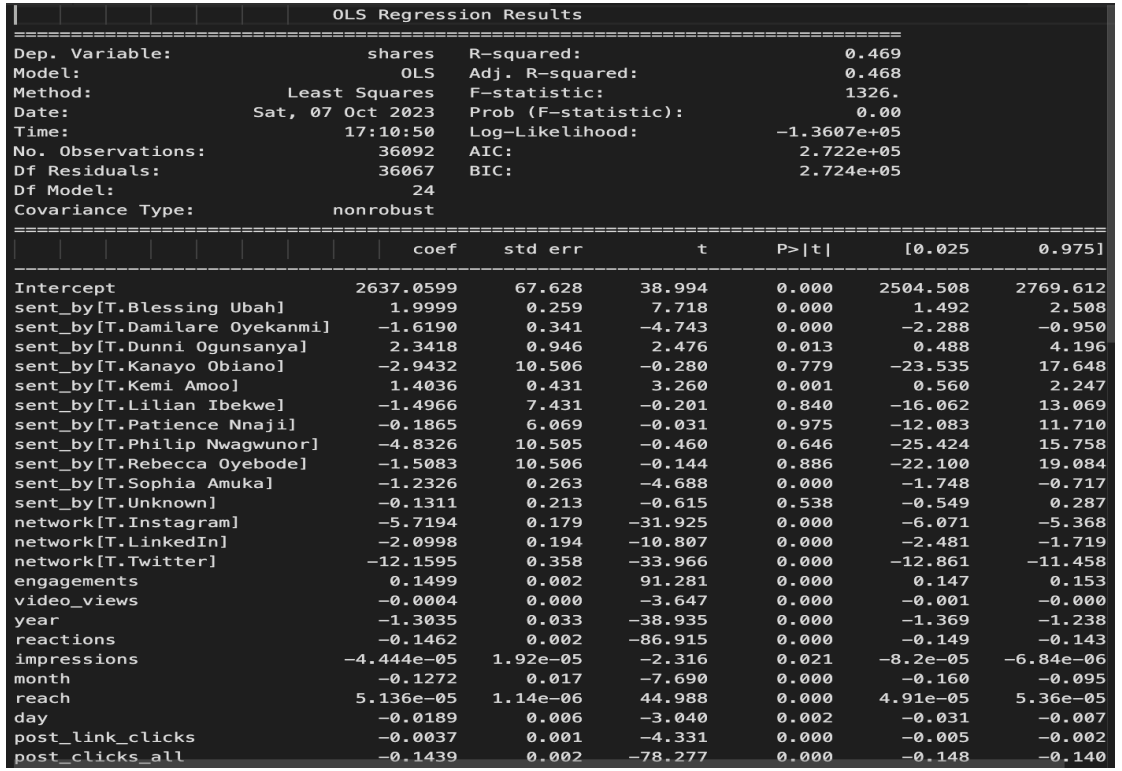
Dep. Variable:	comments	R-squared:	0.411
Model:	OLS	Adj. R-squared:	0.410
Method:	Least Squares	F-statistic:	837.8
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	17:11:48	Log-Likelihood:	-1.6880e+05
No. Observations:	36092	AIC:	3.377e+05
Df Residuals:	36061	BIC:	3.379e+05
Df Model:	30		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-919.8162	177.934	-5.169	0.000	-1268.573	-571.060
network[T.Instagram]	1.5924	0.465	3.422	0.001	0.680	2.505
network[T.LinkedIn]	-3.9258	0.483	-8.123	0.000	-4.873	-2.979
network[T.Twitter]	1.6354	0.914	1.788	0.074	-0.157	3.428
sent_by[T.Blessing Ubah]	3.3460	0.644	5.197	0.000	2.084	4.608
sent_by[T.Damilare Oyekanmi]	1.5647	0.847	1.847	0.065	-0.095	3.225
sent_by[T.Dunni Ogunsanya]	5.2054	2.343	2.221	0.026	0.613	9.798
sent_by[T.Kanayo Obiano]	0.7534	26.026	0.029	0.977	-50.258	51.765
sent_by[T.Kemi Amoo]	2.4994	1.067	2.343	0.019	0.409	4.590
sent_by[T.Lilian Ibekwe]	1.5612	18.412	0.085	0.932	-34.527	37.649
sent_by[T.Patience Nnaji]	-3.7238	15.041	-0.248	0.804	-33.205	25.757
sent_by[T.Philip Nwagwunor]	3.2622	26.023	0.125	0.900	-47.744	54.268
sent_by[T.Rebecca Oyebode]	0.4926	26.024	0.019	0.985	-50.515	51.500
sent_by[T.Sophia Amuka]	4.2374	0.655	6.474	0.000	2.954	5.520
sent_by[T.Unknown]	0.4921	0.534	0.921	0.357	-0.555	1.539
content_type[T.Document]	2.2394	26.032	0.086	0.931	-48.784	53.263
content_type[T.Link]	4.0233	1.436	2.802	0.005	1.208	6.838
content_type[T.Photo]	2.1013	1.019	2.061	0.039	0.103	4.099
content_type[T.Poll]	-29.1884	18.467	-1.581	0.114	-65.385	7.008
content_type[T.Text]	8.0165	1.235	6.491	0.000	5.596	10.437
content_type[T.Video]	3.0977	1.087	2.850	0.004	0.967	5.228
engagements	0.0160	0.000	71.774	0.000	0.016	0.016
reach	-3.571e-05	2.91e-06	-12.271	0.000	-4.14e-05	-3e-05
shares	0.1489	0.012	12.535	0.000	0.126	0.172
impressions	0.0016	4.7e-05	34.467	0.000	0.002	0.002

Please see the jupyter notebook for more information.

Predictive Models - Regression Analysis – Shares

- Using features relevant to the Gradient Boosted Trees prediction model, the r-squared for the Shares feature moved from 21% to 47%.
- The model's F-statistic (1326, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters is related to the reach feature.
- The standard errors are generally low implying that any predictions from the model will be close to the observed values.
- The Durbin-Watson statistic (1.848) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.



OLS Regression Results

Dep. Variable:	shares	R-squared:	0.469
Model:	OLS	Adj. R-squared:	0.468
Method:	Least Squares	F-statistic:	1326.
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	17:10:50	Log-Likelihood:	-1.3607e+05
No. Observations:	36092	AIC:	2.722e+05
Df Residuals:	36067	BIC:	2.724e+05
Df Model:	24		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	2637.0599	67.628	38.994	0.000	2504.508	2769.612
sent_by[T.Blessing Ubah]	1.9999	0.259	7.718	0.000	1.492	2.508
sent_by[T.Damilare Oyekanmi]	-1.6190	0.341	-4.743	0.000	-2.288	-0.950
sent_by[T.Dunni Ogunsanya]	2.3418	0.946	2.476	0.013	0.488	4.196
sent_by[T.Kanayo Obiano]	-2.9432	10.506	-0.280	0.779	-23.535	17.648
sent_by[T.Kemi Amoo]	1.4036	0.431	3.260	0.001	0.560	2.247
sent_by[T.Lilian Ibekwe]	-1.4966	7.431	-0.201	0.840	-16.062	13.069
sent_by[T.Patience Nnaji]	-0.1865	6.069	-0.031	0.975	-12.083	11.710
sent_by[T.Philip Nwagwunor]	-4.8326	10.505	-0.460	0.646	-25.424	15.758
sent_by[T.Rebecca Oyebode]	-1.5083	10.506	-0.144	0.886	-22.100	19.084
sent_by[T.Sophia Amuka]	-1.2326	0.263	-4.688	0.000	-1.748	-0.717
sent_by[T.Unknown]	-0.1311	0.213	-0.615	0.538	-0.549	0.287
network[T.Instagram]	-5.7194	0.179	-31.925	0.000	-6.071	-5.368
network[T.LinkedIn]	-2.0998	0.194	-10.807	0.000	-2.481	-1.719
network[T.Twitter]	-12.1595	0.358	-33.966	0.000	-12.861	-11.458
engagements	0.1499	0.002	91.281	0.000	0.147	0.153
video_views	-0.0004	0.000	-3.647	0.000	-0.001	-0.000
year	-1.3035	0.033	-38.935	0.000	-1.369	-1.238
reactions	-0.1462	0.002	-86.915	0.000	-0.149	-0.143
impressions	-4.444e-05	1.92e-05	-2.316	0.021	-8.2e-05	-6.84e-06
month	-0.1272	0.017	-7.690	0.000	-0.160	-0.095
reach	5.136e-05	1.14e-06	44.988	0.000	4.91e-05	5.36e-05
day	-0.0189	0.006	-3.040	0.002	-0.031	-0.007
post_link_clicks	-0.0037	0.001	-4.331	0.000	-0.005	-0.002
post_clicks_all	-0.1439	0.002	-78.277	0.000	-0.148	-0.140

Please see the jupyter notebook for more information.

Predictive Models - Regression Analysis – Link Clicks

- Using features relevant to the Gradient Boosted Trees prediction model, the r-squared for the Link Clicks feature moved from 2% to 27%.
- The model's F-statistic (2197, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters is related to the Link Clicks feature.
- The standard errors are significantly lower in this model implying nearly accurate predictions.
- The Durbin-Watson statistic (1.996) implies that the model's errors are independent and there is no strong case for overfitting or underfitting.

OLS Regression Results

Dep. Variable:	post_link_clicks	R-squared:	0.268
Model:	OLS	Adj. R-squared:	0.267
Method:	Least Squares	F-statistic:	2197.
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	16:07:17	Log-Likelihood:	-2.0274e+05
No. Observations:	36092	AIC:	4.055e+05
Df Residuals:	36085	BIC:	4.056e+05
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	15.7410	0.854	18.422	0.000	14.066	17.416
network[T.Instagram]	-1.9386	0.996	-1.947	0.052	-3.891	0.013
network[T.LinkedIn]	14.0578	1.103	12.748	0.000	11.896	16.219
network[T.Twitter]	5.9453	2.199	2.703	0.007	1.635	10.256
post_clicks_all	0.2760	0.003	107.863	0.000	0.271	0.281
reach	-5.481e-05	6.96e-06	-7.871	0.000	-6.85e-05	-4.12e-05
impressions	-0.0060	0.000	-52.983	0.000	-0.006	-0.006

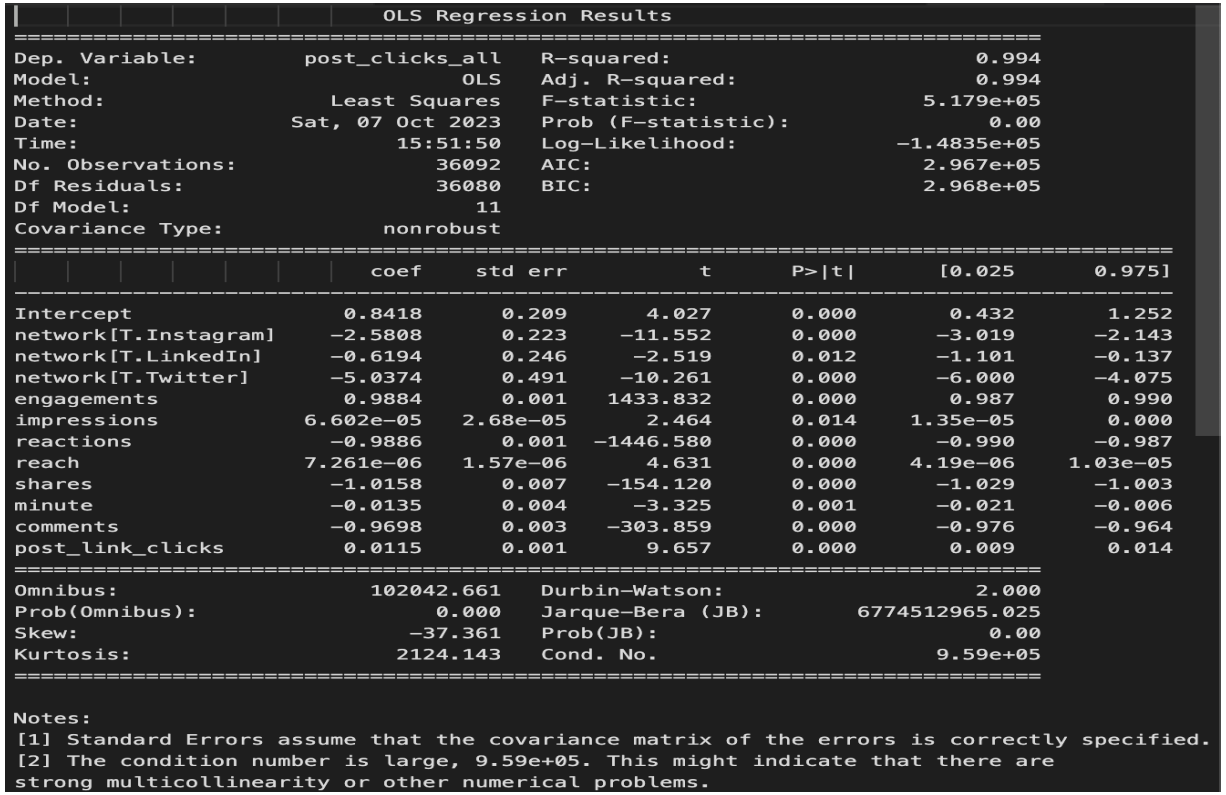
Omnibus:	86787.350	Durbin-Watson:	1.996
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2649823048.539
Skew:	24.470	Prob(JB):	0.00
Kurtosis:	1329.518	Cond. No.	9.49e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.49e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Please see the jupyter notebook for more information.

Predictive Models - Regression Analysis – All Clicks

- Using features relevant to the Gradient Boosted Trees prediction model, the r-squared for the Engagements feature moved from 0% to 99.4%.
- The model's F-statistic (51k, p-value 0) is significant implying that the model is statistically significant and at least one of the model parameters (likely all) is related to the reach feature.
- The standard errors are very low implying that any predictions from the model are nearly identical to observed values.
- The Durbin-Watson statistic (2.0) implies that the model's errors ae independent and there is no strong case for overfitting or underfitting.



OLS Regression Results

Dep. Variable:	post_clicks_all	R-squared:	0.994
Model:	OLS	Adj. R-squared:	0.994
Method:	Least Squares	F-statistic:	5.179e+05
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	15:51:50	Log-Likelihood:	-1.4835e+05
No. Observations:	36092	AIC:	2.967e+05
Df Residuals:	36080	BIC:	2.968e+05
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.8418	0.209	4.027	0.000	0.432	1.252
network[T.Instagram]	-2.5808	0.223	-11.552	0.000	-3.019	-2.143
network[T.LinkedIn]	-0.6194	0.246	-2.519	0.012	-1.101	-0.137
network[T.Twitter]	-5.0374	0.491	-10.261	0.000	-6.000	-4.075
engagements	0.9884	0.001	1433.832	0.000	0.987	0.990
impressions	6.602e-05	2.68e-05	2.464	0.014	1.35e-05	0.000
reactions	-0.9886	0.001	-1446.580	0.000	-0.990	-0.987
reach	7.261e-06	1.57e-06	4.631	0.000	4.19e-06	1.03e-05
shares	-1.0158	0.007	-154.120	0.000	-1.029	-1.003
minute	-0.0135	0.004	-3.325	0.001	-0.021	-0.006
comments	-0.9698	0.003	-303.859	0.000	-0.976	-0.964
post_link_clicks	0.0115	0.001	9.657	0.000	0.009	0.014

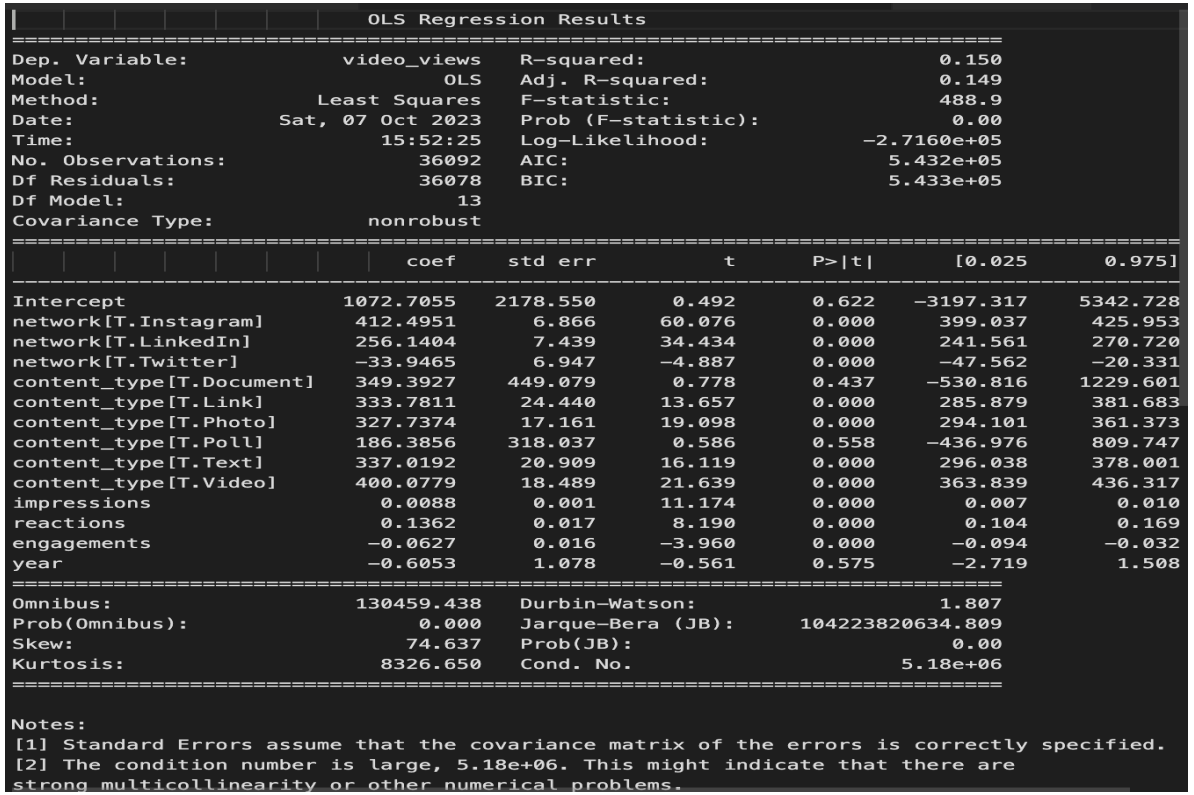
Omnibus:	102042.661	Durbin-Watson:	2.000
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6774512965.025
Skew:	-37.361	Prob(JB):	0.00
Kurtosis:	2124.143	Cond. No.	9.59e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.59e+05. This might indicate that there are strong multicollinearity or other numerical problems.

Please see the jupyter notebook for more information.

Predictive Models - Regression Analysis – Video Views

- This model is identical to the correlation model, indicating that Video Views may not be linearly related to the other features in the dataset.
- However, when we consider that the tree models were unable to provide any better predictions than this, it is likely the parameters or features related to Video Views are not being used in our models.
- This means we have been unable to explain a significant amount of the variance in this feature.



OLS Regression Results

Dep. Variable:	video_views	R-squared:	0.150
Model:	OLS	Adj. R-squared:	0.149
Method:	Least Squares	F-statistic:	488.9
Date:	Sat, 07 Oct 2023	Prob (F-statistic):	0.00
Time:	15:52:25	Log-Likelihood:	-2.7160e+05
No. Observations:	36092	AIC:	5.432e+05
Df Residuals:	36078	BIC:	5.433e+05
Df Model:	13		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1072.7055	2178.550	0.492	0.622	-3197.317	5342.728
network[T.Instagram]	412.4951	6.866	60.076	0.000	399.037	425.953
network[T.LinkedIn]	256.1404	7.439	34.434	0.000	241.561	270.720
network[T.Twitter]	-33.9465	6.947	-4.887	0.000	-47.562	-20.331
content_type[T.Document]	349.3927	449.079	0.778	0.437	-530.816	1229.601
content_type[T.Link]	333.7811	24.440	13.657	0.000	285.879	381.683
content_type[T.Photo]	327.7374	17.161	19.098	0.000	294.101	361.373
content_type[T.Poll]	186.3856	318.037	0.586	0.558	-436.976	809.747
content_type[T.Text]	337.0192	20.909	16.119	0.000	296.038	378.001
content_type[T.Video]	400.0779	18.489	21.639	0.000	363.839	436.317
impressions	0.0088	0.001	11.174	0.000	0.007	0.010
reactions	0.1362	0.017	8.190	0.000	0.104	0.169
engagements	-0.0627	0.016	-3.960	0.000	-0.094	-0.032
year	-0.6053	1.078	-0.561	0.575	-2.719	1.508

Omnibus:	130459.438	Durbin-Watson:	1.807
Prob(Omnibus):	0.000	Jarque-Bera (JB):	104223820634.809
Skew:	74.637	Prob(JB):	0.00
Kurtosis:	8326.650	Cond. No.	5.18e+06

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.18e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Please see the jupyter notebook for more information.

Conclusions

- Most of the metrics in the general metrics list are linearly related with each other, the categorical features and date.
- This means that increasing or reducing the values of a specific metric is achievable by increasing or decreasing the values of its related independent variables.
- We used predictive modeling to determine the best features that increase model accuracy.
- Impressions is better modeled using tree models than a linear model. However, linear models are able to explain about 55% of the variance in impressions.
- We are unable to successfully model the Video Views features to significantly increase accuracy.

Please see the jupyter notebook for more information.

Thank you!