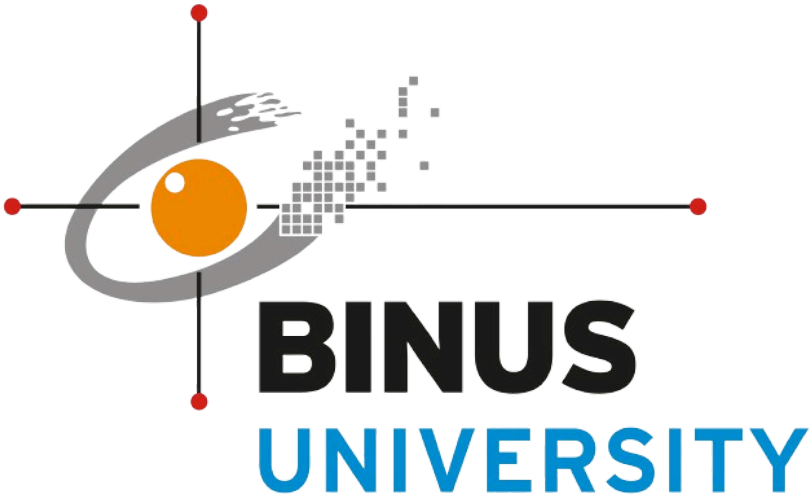


Basic Statistic Final Report

Lecturer: Raymond Bahana ST., M.Sc

Class: L3AC

Computer Science Program

School of Computing and Creative Arts

Bina Nusantara International University

Jakarta, 2024

Name: Gabriel Anderson
NIM: 2702256315

Name: Rafael Anderson
NIM: 2702255981

Name: Albertus Santoso
NIM: 2702334885

Table of Contents

I. Background

In the current digital world, the number of social media accounts are increasing at a rapid rate [1]. In 2024, the number of social media users increased from 4.72 billion to 5.02 billion users across all social media platforms from the previous year, which is an astounding increase of 320 million new users [2].

Instagram is one of the more common social media platforms, who has the 4th most active users which is over 2 billion users [3]. Instagram growth are high with a 50 million growth rate from 2023 to 2024[4].Instagram users spend around 31.4 hours daily with a daily users of around 500 million [5].The continuously increasing number of Instagram users has increased the competition between brands and influencers to capture as much engagement rate from the users. Instagram being the 4th most popular social media platform, it is a platform often used by influencers to promote products, where diverse sizes of businesses utilize this platform, since it is extremely helpful in increasing their brand awareness [6] .

Instagram has a steady increase in ad reach with a 12.2 percent increase year to year [7]. However, finding which instagram accounts that receive the overall most rate of engagement from the users is a challenge faced by most brands trying to find a suitable promoter for their brand. The engagement rate is calculated by getting the average of the total likes and total comments, which is then divided by the amount of followers, and then multiplied by 100. This means that the average engagement rate doesn't necessarily represent the number of users, as it is proportional to the number of followers.

It is crucial to find the amount of active users by making it proportional to the number of followers. As brands would prefer quality over quantity, brands would rather find Instagram accounts with audiences interested in the content of that account than an account with more likes but a lower engagement rate. Interested audiences will talk and share the product, providing additional exposure to the promoted product [8]. This is also important for companies or people trying to find a suitable influencer to promote their product or service. The size of an influencer is strongly correlated with the cost of asking them to promote a product or service. For example, the cost of sponsoring an influencer with a larger following will generally be more expensive than an influencer with a smaller following. Thus, by finding smaller influencers with a decent or larger engagement rate, it may give brands higher quality audiences and be charged less amount for the sponsor.

II. Hypothesis

Null Hypothesis: There is no significant relationship between average likes and the engagement rate prediction for influencer campaigns.

Alternative Hypothesis: There is a significant relationship between average likes and the engagement rate prediction for influencer campaigns.

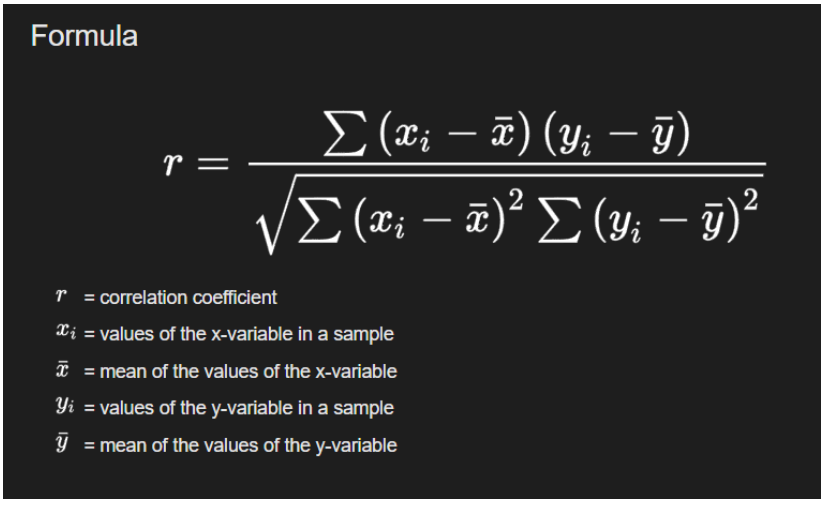
III. Analyze statistic concept

To test our hypothesis, we used several statistical tests to determine whether there is a significant relationship between average likes and average engagement rate for influencer campaigns. The tests we used are the **Pearson Correlation Coefficient**, **Kendall's Tau**, and **Spearman's Rank Correlation** tests. We used these tests to test our hypothesis by calculating the p value. The p value is used to either reject or accept our hypothesis.

Results:

Pearson Correlation: 0.2286398943688465, p-value: 2.1195395286744633e-12

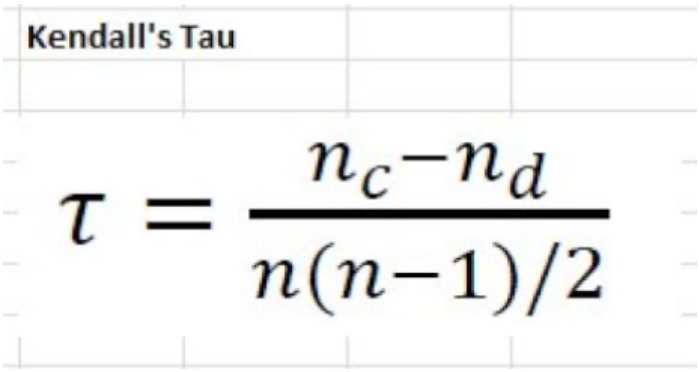
Formula:



Conclusion: The Pearson correlation is more than 0, and the p-value is smaller than 0.05. Reject the null hypothesis: There is a significant relationship between Average Likes and Engagement Rate.

Kendall's Tau: 0.1480610021852201, p-value: 1.6871536539164292e-11

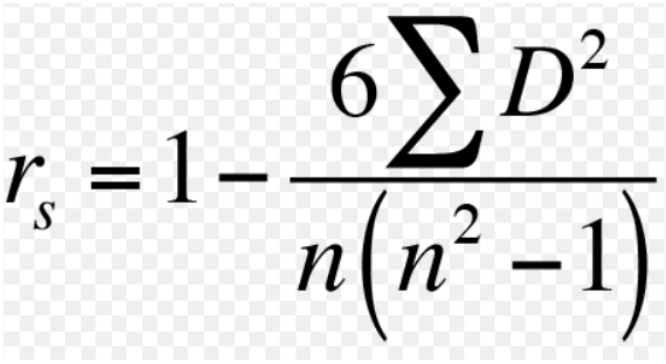
Formula:



Conclusion: Kendall's Tau is greater than 0, and the p-value is lower than 0.05. Reject the null hypothesis: There is a significant relationship between Average Likes and Engagement Rate.

Spearman Correlation: 0.20790619019460516, p-value: 1.839566304675373e-10

Formula:



Conclusion: The spearman's correlation is greater than 0, and the p-value is lower than 0.05. Reject the null hypothesis: There is a significant relationship between Average Likes and Engagement Rate.

From the results, we can conclude that there is a **significant** relationship between average likes and average engagement rate for influencer campaigns with the Pearson Correlation Coefficient having the best result (Highest correlation).

Code for statistical test:
https://github.com/gamakagami/FoDS-FinalProject/blob/main/statistical_test.ipynb

IV. Data Visualization

We used 8 csv files which are **‘mega_global_cleaned.csv’, ‘macro_global_cleaned.csv’, ‘micro_global_cleaned.csv’, ‘nano_global_cleaned.csv’, ‘mega_indo_cleaned.csv’, ‘macro_indo_cleaned.csv’, ‘micro_indo_cleaned.csv’, ‘nano_indo_cleaned.csv’,** and merged them. The final size of the data after cleaning is (922, 9), with 922 records and 9 attributes. Based on the data, it can be separated into several aspects. Which includes the origin of the account, which is either Indonesian or Global, and by the following size of the account, including mega (More than 1M), macro (100k - 1M), micro (10K - 100K), and nano (1K - 10K).

Dataset snippet:

	Followers	Average_Likes	Average_Comments	Average_Engagement_Rate	Engagement_Rate_(60_Days)	Posting_frequency_(last_60_days)	Posts_(image)	Posts_(videos)	Posts_(carousel)
2	47352198	458723.76	4422.62	0.9781	0.4896	132.0	9.0	34.0	89.0
3	39581466	118171.13	692.59	0.3003	0.0018	4.0	0.0	0.0	4.0
4	38776943	139324.51	763.73	0.3613	0.1303	100.0	42.0	44.0	14.0
5	36318320	152328.63	2451.69	0.4262	0.22	83.0	8.0	46.0	29.0
6	34486065	145525.51	948.45	0.4247	0.2491	98.0	5.0	56.0	37.0
7	1245315	164240.73	1246.38	13.2888	0.2431	1.0	0.0	0.0	1.0
8	1808490	70429.07	1255.41	3.9638	0.2944	5.0	1.0	4.0	0.0
9	80883913	3674629.51	16979.32	4.5641	5.2609	26.0	7.0	9.0	10.0
10	33412540	272957.06	1596.6	0.8217	0.692	47.0	1.0	11.0	35.0
11	2047559	185356.19	630.77	9.0834	4.2291	23.0	0.0	13.0	10.0
12	1811551	26212.84	145.84	1.455	0.7947	13.0	0.0	2.0	11.0
13	27521485	731834.07	3964.1	2.6735	4.405	7.0	0.0	2.0	5.0
14	1180207	154382.52	1755.34	13.2297	0.5307	1.0	0.0	0.0	1.0
15	7994281	201394.06	628.37	2.5271	1.5076	29.0	4.0	8.0	17.0
16	35137412	291341.05	2557.12	0.8364	0.4342	30.0	4.0	26.0	0.0
17	27421149	150723.02	3042.12	0.5608	0.1551	35.0	2.0	24.0	9.0
18	7146979	700613.67	5303.04	9.8771	5.8659	4.0	0.0	0.0	4.0
19	7900796	94032.52	642.67	1.1983	0.6597	48.0	16.0	16.0	16.0
20	8277231	266408.27	1143.38	3.2324	3.1181	2.0	0.0	0.0	2.0
21	8660319	55255.08	150.44	0.6398	0.0008	12.0	1.0	4.0	7.0

Attributes Table:

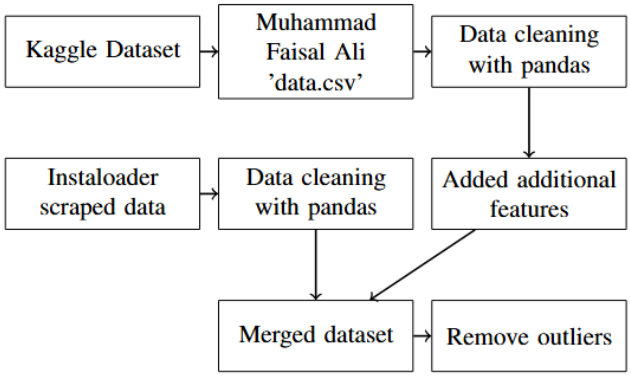
The description of all attributes and sample data are present in the following table:

Attribute	Description	Sample Data
Username	Username of the account	cristiano
Followers	Follower count of the account	47352198
Average Likes	The average likes of the account	458723.76
Average Comments	The average comments of the account	4422.62
Average Engagement Rate	Value calculated from the average likes, comments, and followers	0.9781
Engagement Rate (60 Days)	Same as the average engagement rate, but only containing data from posts posted the last 60 days	0.4896
Posting Frequency (60 Days)	Number of posts posted (60 Days)	132
Posts (Image)	Number of posts of the image type (60 Days)	9
Posts (Video)	Number of posts of the video type (60 Days)	34
Posts (Carousel)	Number of posts of the carousel type (60 Days)	89
Average Hashtags / Post (60 Days)	Average number of hashtags used for posts posted last 60 days	1.0

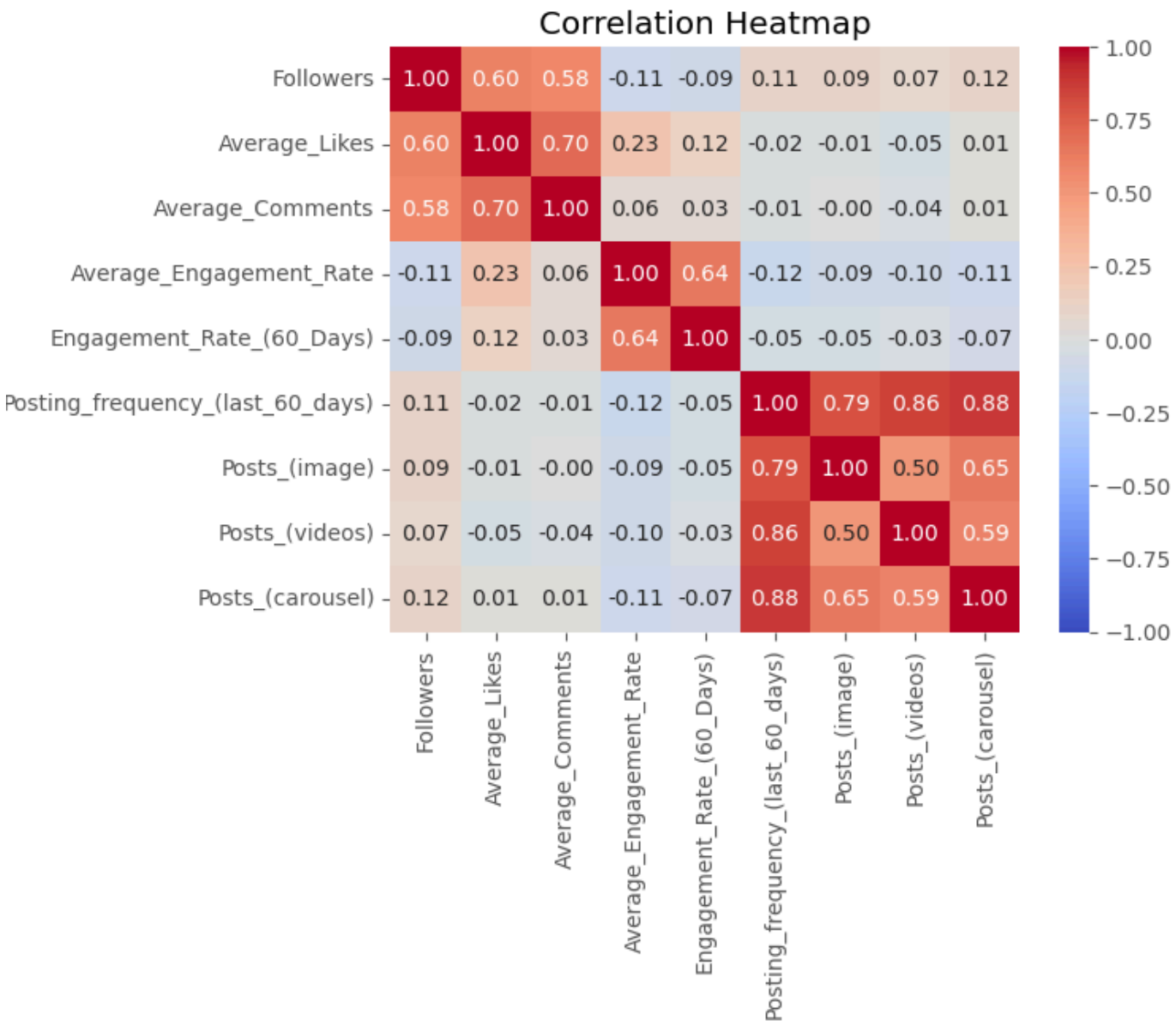
Kaggle and Instaloader scraping are the main sources of our datasets. The dataset from kaggle was manually scraped by the author from the socialbook website [9]. To process our data to be readable, we first need to format the CSV files. Since the accounts scraped and the data we got from Kaggle have their respective CSV files, the datasets need to be formatted individually. The data from Kaggle is cleaned by removing unneeded columns that are not used in the analysis. It is also added with additional attributes, to match the structure of the scraped data. For example, data from the last 60 days are added to the Kaggle data. Then, both csv files are merged, and the outliers are removed from the data by using the Z-score method.

The scraping codes can be accessed through Github:
https://github.com/gamakagami/FoDS-FinalProject/tree/main/data%20scraping

The full process can be seen in the following image:



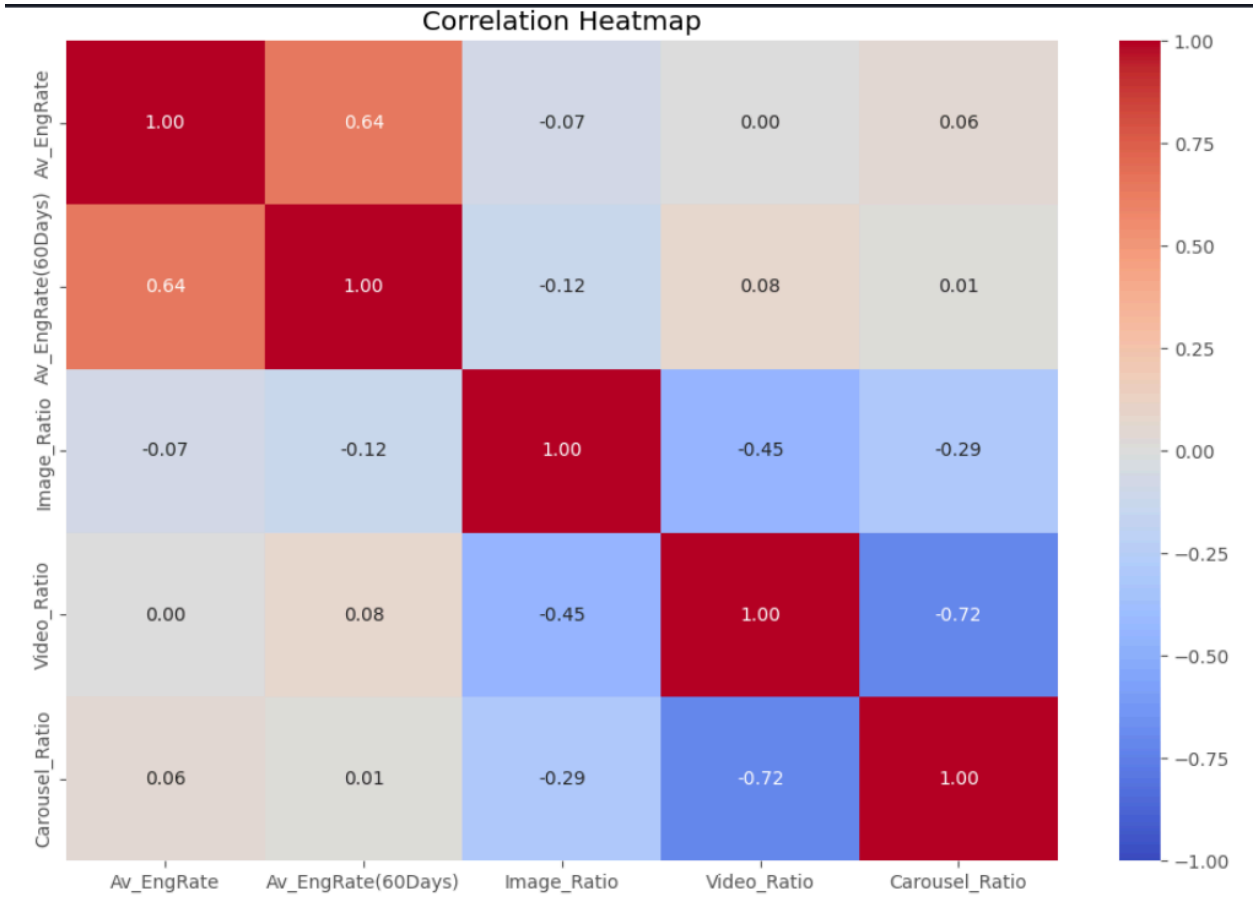
Kaggle dataset link: https://www.kaggle.com/datasets/syedjaferk/top-200-instagrammers-data-cleaned



Correlation matrix

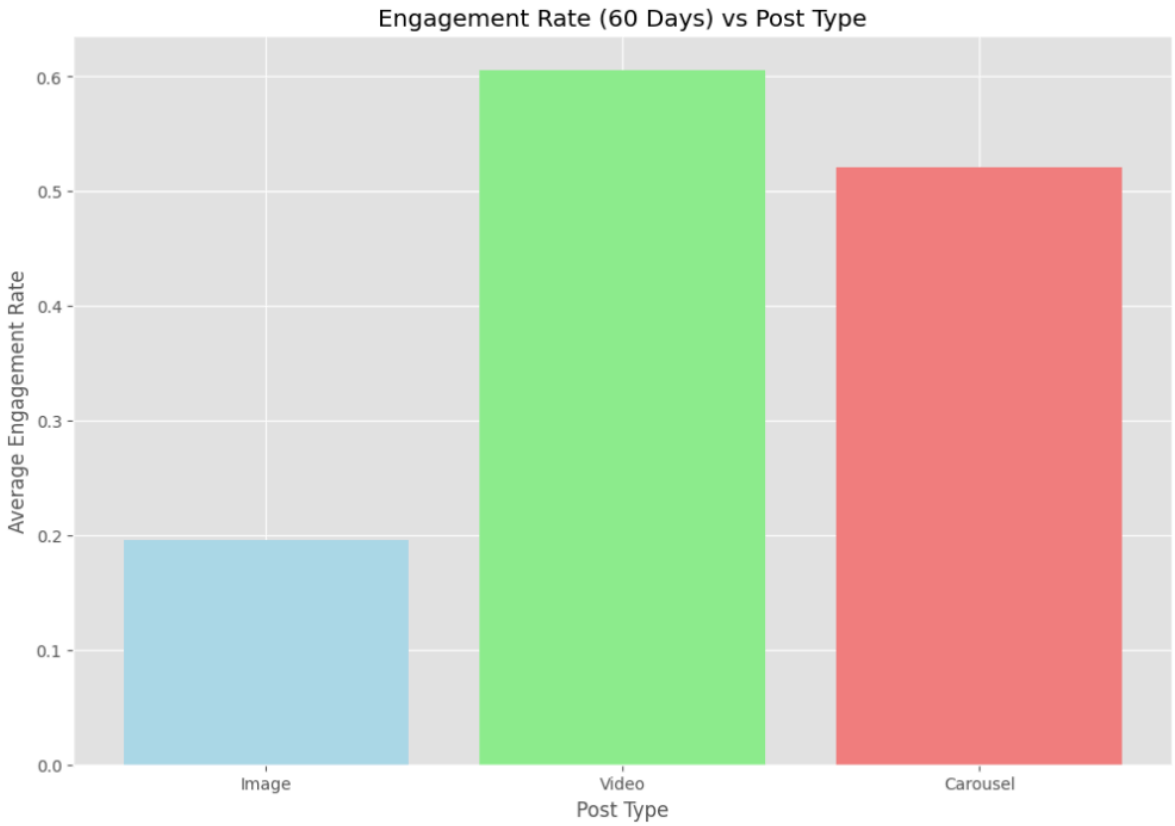
In the correlation matrix, the variable with the huge correlation with our target variable (Average Engagement Rate) is the average engagement rate of the last 60 days, with a correlation of 0.64. Followed by the average likes and average comments with a correlation of 0.23 and 0.06 respectively.

Key Insight: Average Engagement Rate (60 Days) has the highest correlation with the overall engagement rate.



In this matrix, we wanted to find whether the ratio of each post out (Post type count / Posting Frequency) of the posting frequency had an impact on the engagement rate, as the count of the posts itself may not be reliable during an analysis. Turns out, there is a subtle correlation between the video ratio and carousel ratio, with both engagement rates (Overall and 60 days).
Note: The data only contains accounts with posting frequency more than 0.

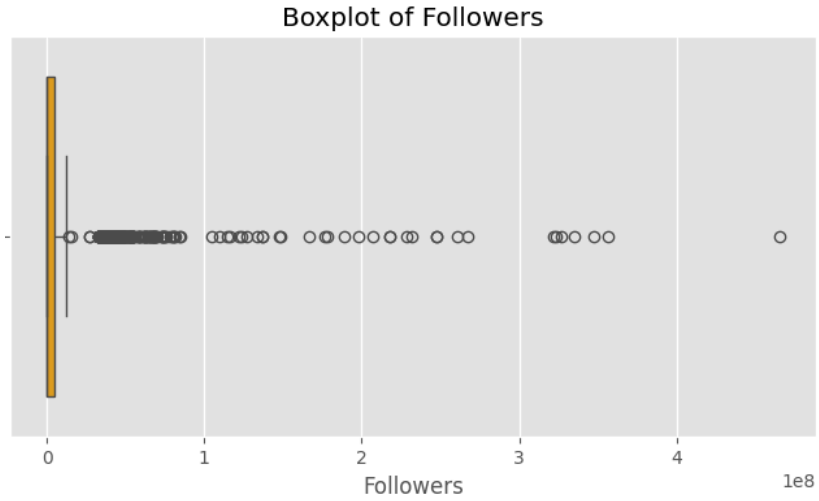
Key Insight: There is a subtle correlation between the video ratio and carousel ratio, with both engagement rates (Overall and 60 days).



On each record where that specific type of post is not equal to 0, it is divided by the posting frequency, and multiplied to the average engagement rate of the last 60 days. The records where the posting frequency is 0 are removed, to guarantee an accurate analysis. The process is done three times, each for a specific post type. As a result, 3 new data frames are formed. Then, the median of the average engagement rate is calculated (Less sensitive to outliers), and it is found that the video post type receives the most engagement rate.

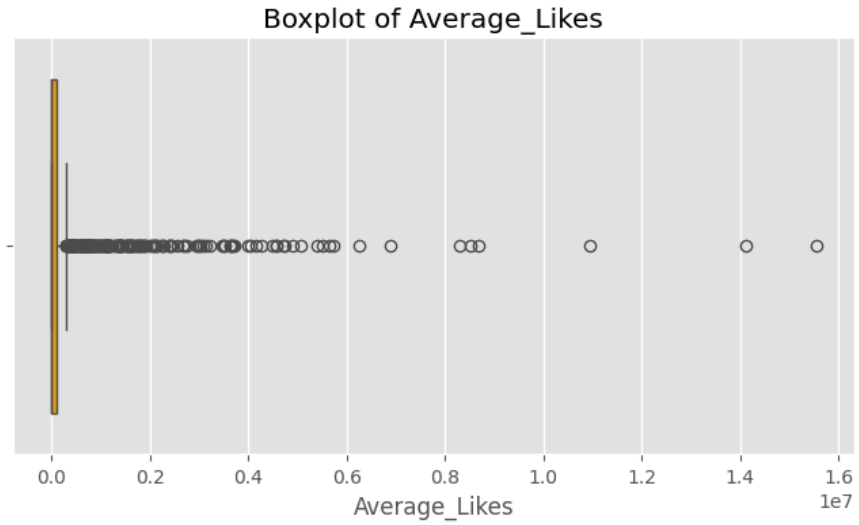
Key Insights: Video post type received the most engagement.

Box Plot



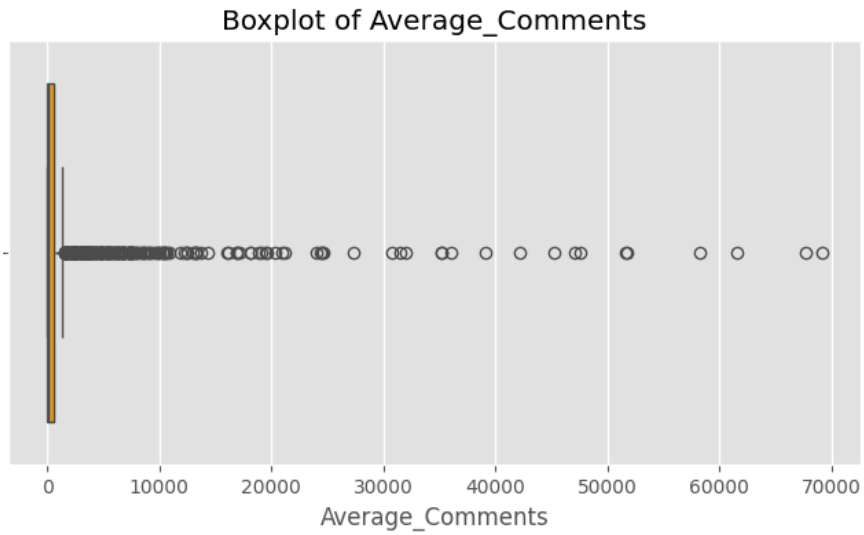
For the followers, the majority are outliers since we collect accounts from several follower range such as mega (> 1M Followers), macro (100K - 1M Followers), micro (10K - 100K Followers), and nano (1 - 10K Followers) accounts.

Key Insight: Accounts from a wide range of followers are used



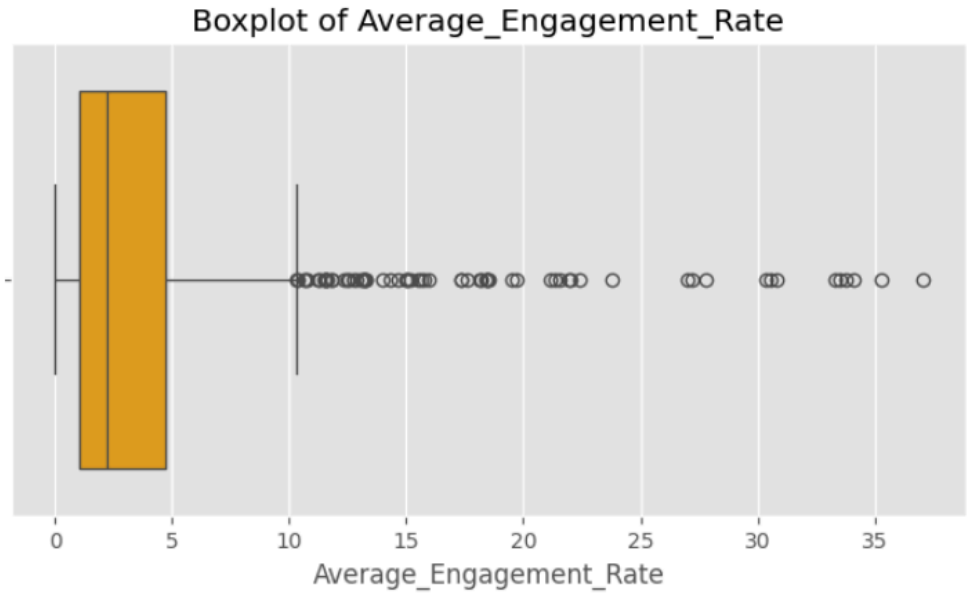
Since accounts from a wide range of followers are used, the average likes will be similar, since the average likes has high correlation with the followers count.

Key Insight: Accounts from a wide range of followers are used



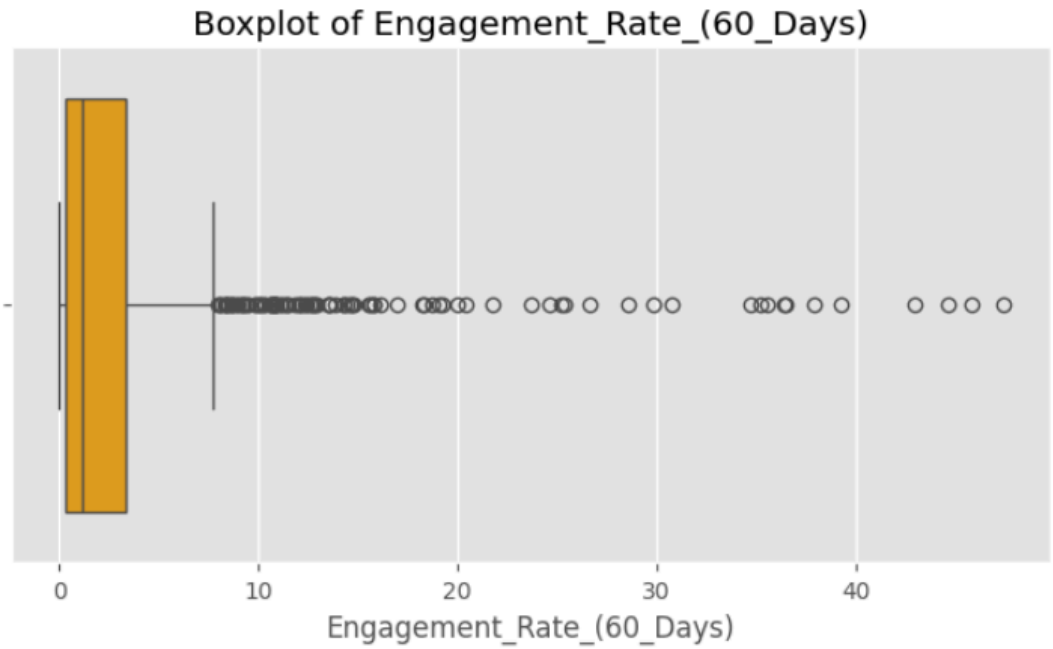
Similar to the average likes, the average comments are also proportional to the follower count. Making it very diverse since diverse accounts are used (In terms of followers).

Key Insight: Accounts from a wide range of followers are used



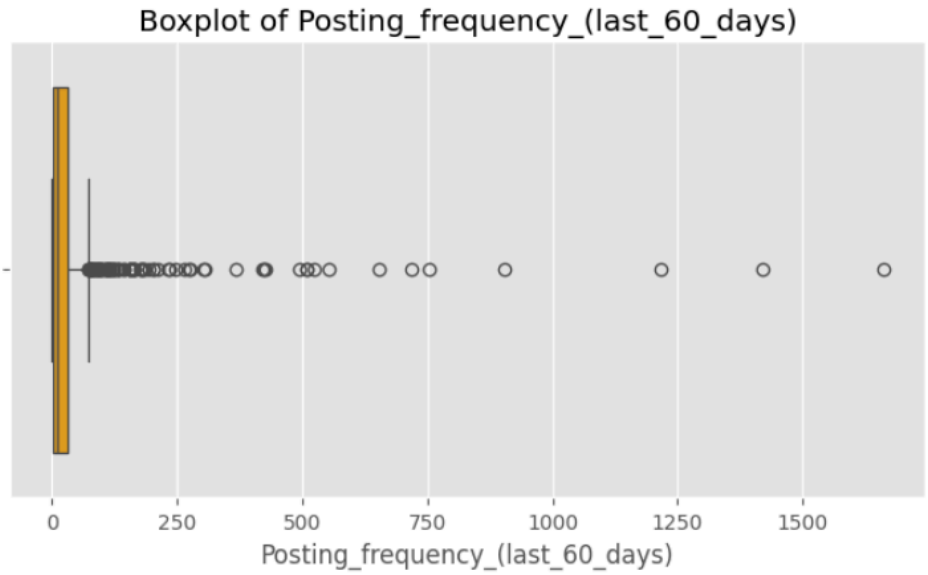
In this boxplot, it shows that the majority of the accounts have an engagement rate ranging from 1 to 5, where the median is approximately 2. Outliers start from 10 up to 37.

Key Insight: Median of average engagement rate is approximately 2.



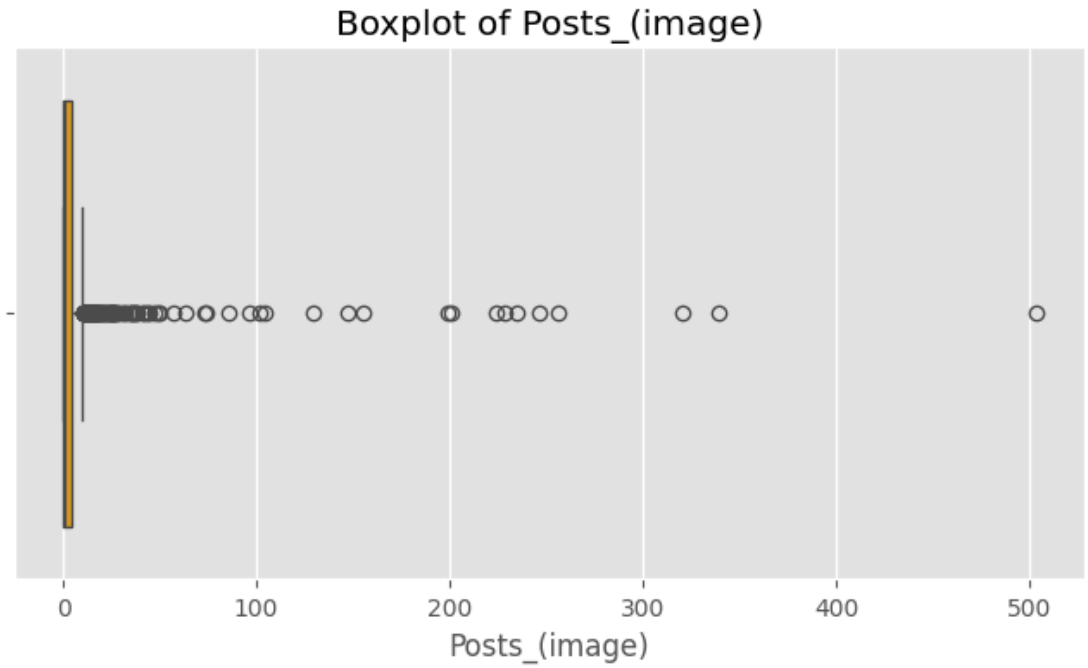
Based on the following boxplot, the lower quartile for the average engagement rate of the last 60 days is significantly lower than the overall average engagement rate. On the median, the value is also lower than the overall engagement rate as well. Additionally, the outliers have greater values, ranging up to 48.

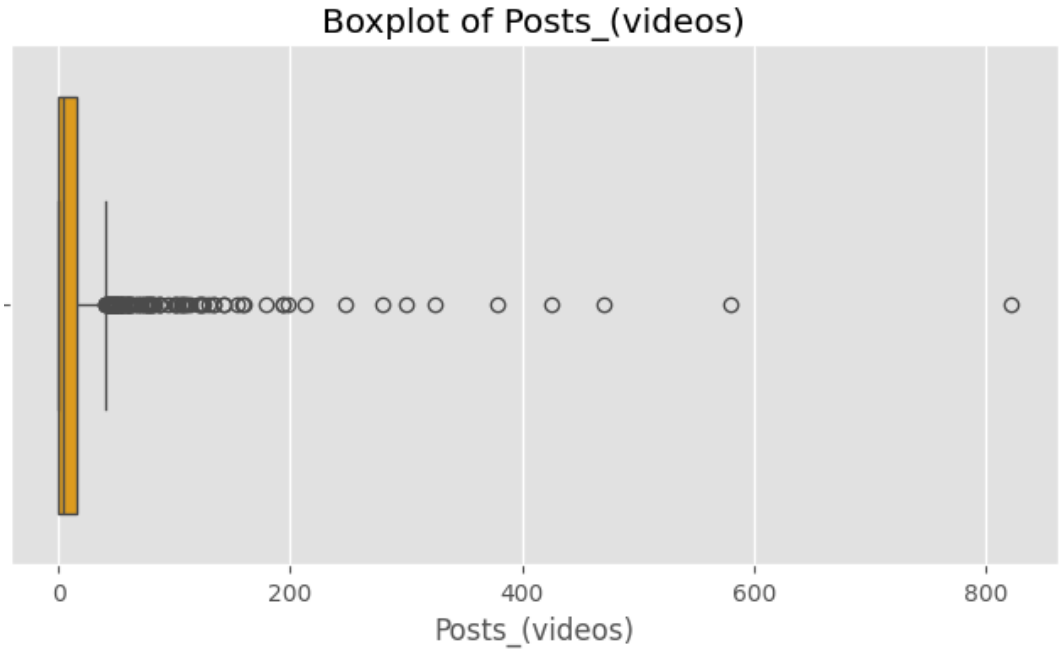
Key Insights: The median of the engagement rate of the last 60 days is lower than the overall engagement rate, containing more outliers with higher values as well.

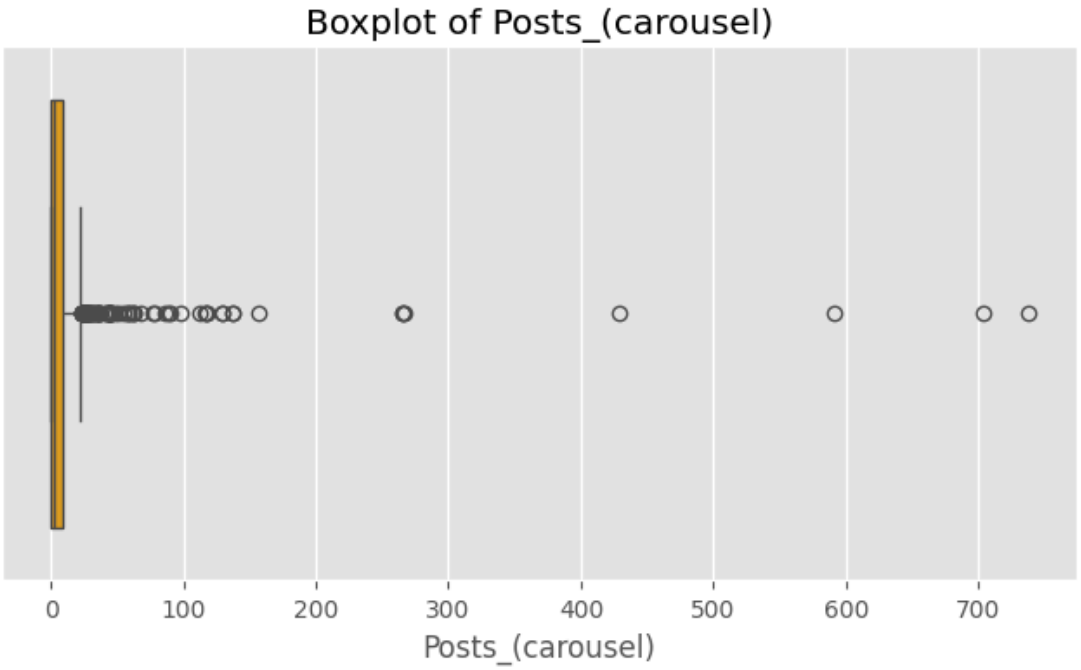


The posting frequency of the last 60 days contains more outliers than previous box plots, as it is not a linear data type. Most posts have a few posts for the last 60 days, ranging from 0 to 40 for the lower quartile and the upper quartile. It contains lots of outliers ranging up to 1600 posts.

Key Insights: Most accounts posted in a range between 0 to 40 in the last 60 days.



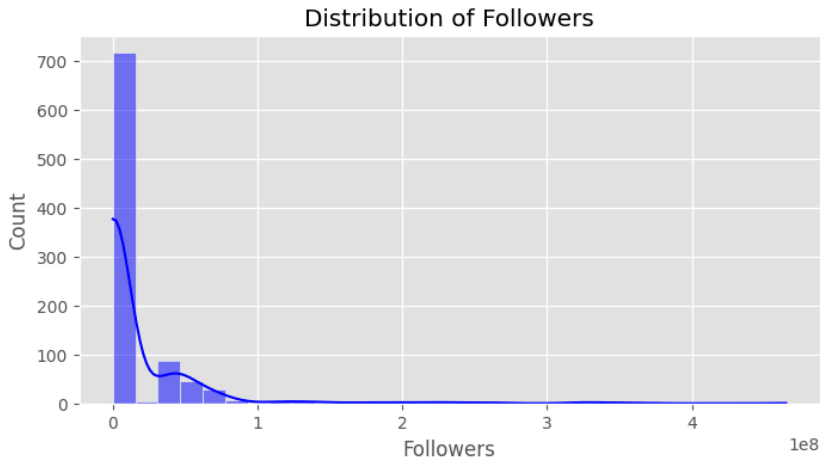




The boxplot for the type of images is similar to the boxplot of the posting frequency, since they rely on it. The majority of the boxplots have a few posts of their specific post type, where there are many outliers as well, since these post types count are not a linear data type

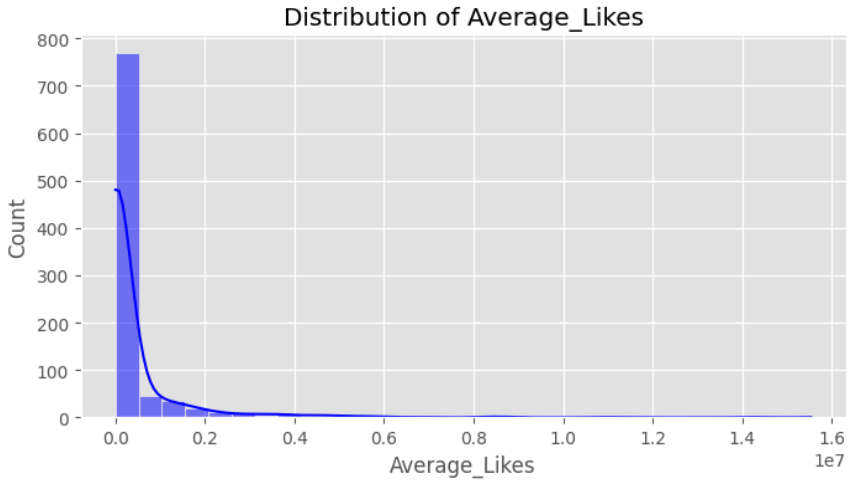
Key Insight: The majority of the boxplots have a few posts of their specific post type

Distribution Plot



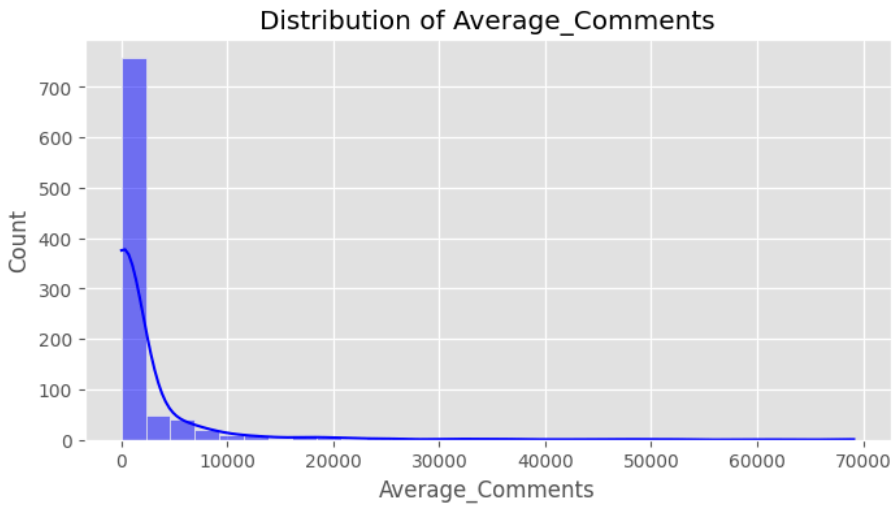
The distribution is highly skewed to the right, indicating that the majority of accounts have relatively low follower counts. A small number of accounts have extremely high follower counts, creating a long tail in the distribution.

Key Insight: The majority of accounts have relatively low follower counts, with a small number of accounts exhibiting extremely high follower counts, creating a long-tail distribution.



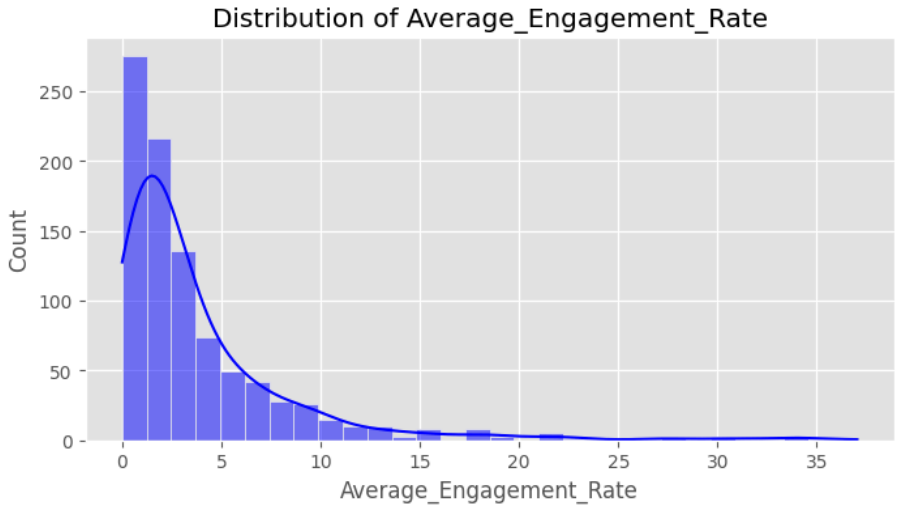
The distribution is highly skewed to the right, indicating that the majority of accounts have relatively low average likes. A small number of accounts receive extremely high average likes, creating a long tail in the distribution.

Key Insight: Most accounts receive a relatively low number of average likes, while a small subset achieves extremely high average likes, leading to a long-tail distribution.



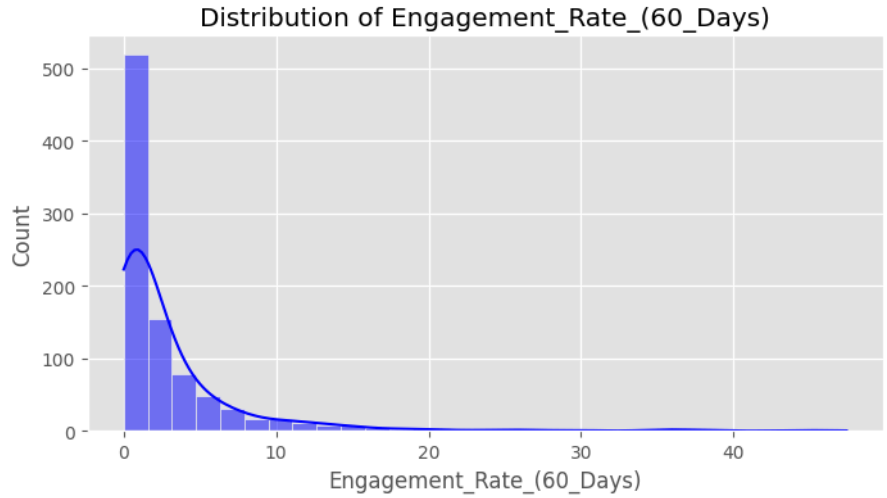
The distribution is highly skewed to the right, indicating that the majority of accounts have relatively low average comments. A small number of accounts receive extremely high average comments, creating a long tail in the distribution.

Key Insight: The majority of accounts receive relatively low average comments, with a small number of accounts having exceptionally high average comments, forming a long-tail distribution.



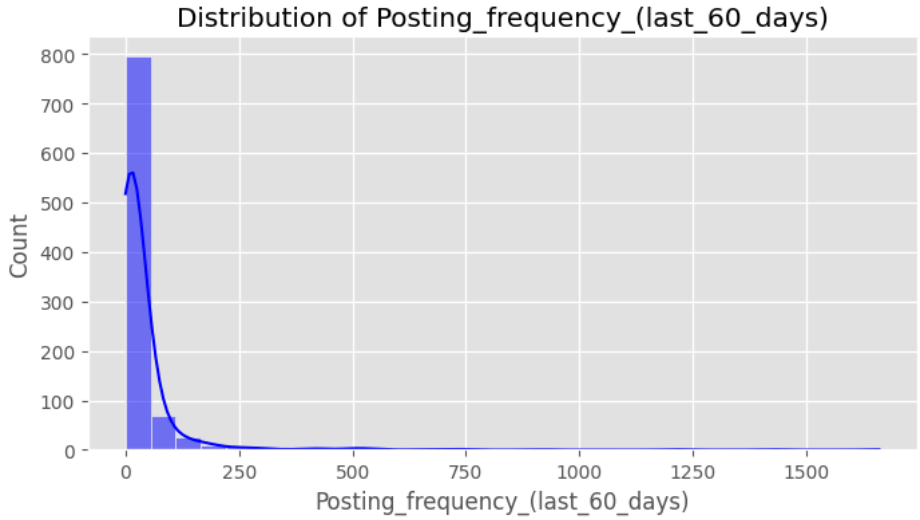
The distribution is highly skewed to the right, indicating that most accounts have a low engagement rate with most of them below 5 percent. A small number of accounts have an extremely high average engagement rate, creating a long tail in the distribution.

Key Insight: Most accounts have a low engagement rate, typically below 5 percent, with a few accounts achieving extremely high engagement rates, resulting in a long-tail distribution.



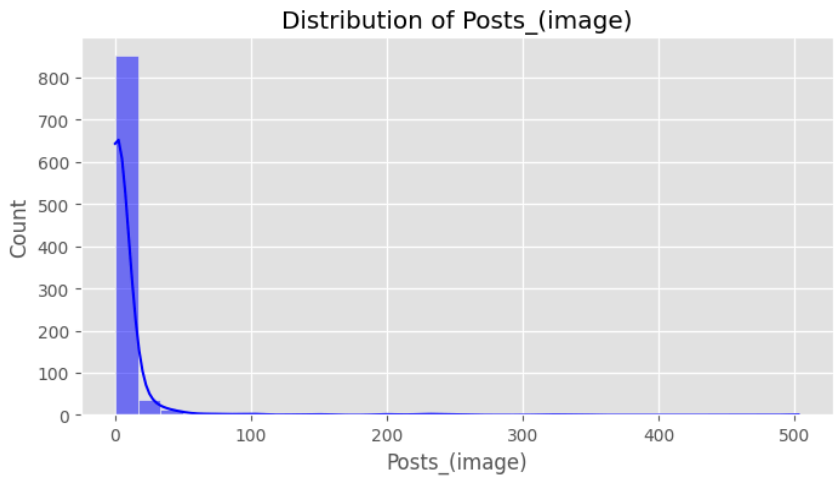
The distribution is highly skewed to the right, indicating that most accounts have a low engagement rate over the last 60 days with most of them below 10 percent. A small number of accounts have an extremely high engagement rate over the last 60 days, creating a long tail in the distribution.

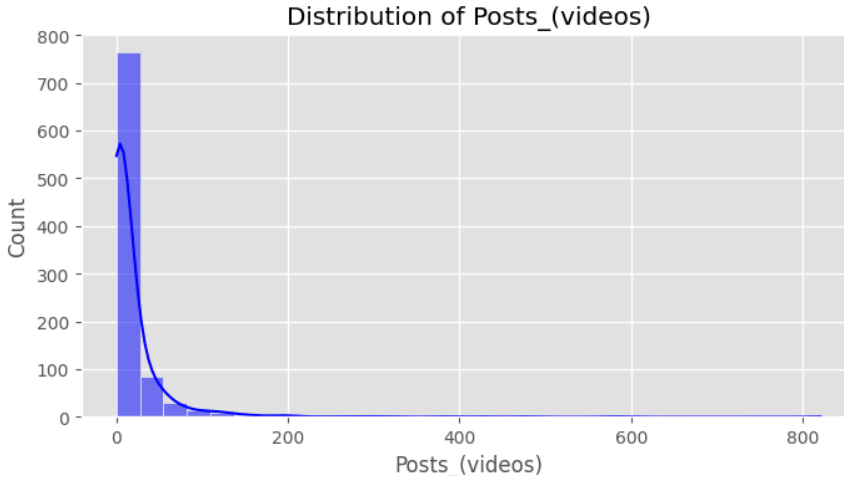
Key Insight: Most accounts exhibit a low engagement rate over the last 60 days, generally below 10 percent, with a minority showing significantly higher engagement rates, forming a long-tail distribution.

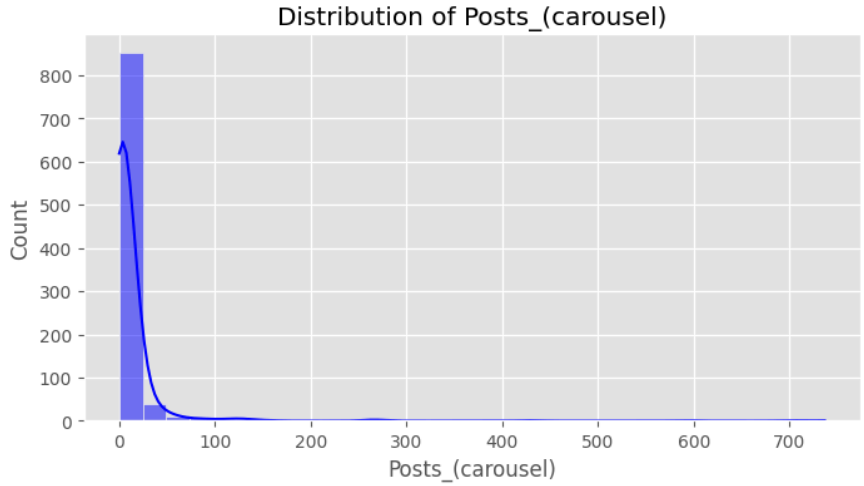


The distribution is highly skewed to the right, indicating that most accounts have a low posting frequency in the last 60 days. Majority of accounts have a posting frequency below 250 over the last 60 days with a minority of accounts having a large number of posting frequency over 1500, creating a long tail in the distribution.

Key Insight: The majority of accounts post infrequently over the last 60 days, with posting frequencies below 250, while a small number of accounts post more than 1500 times, creating a long-tail distribution.

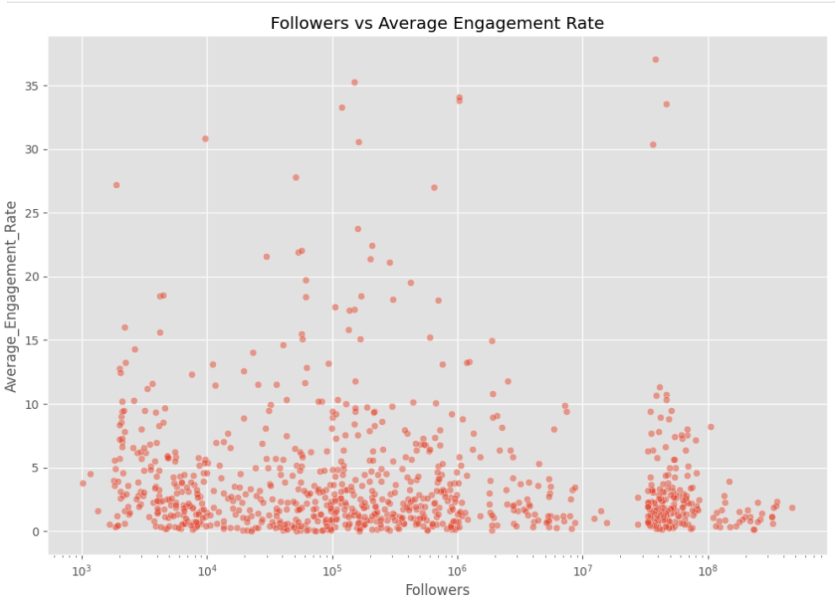

Distribution of Posts_(image)
Count
Posts_(image)
0
100
200
300
400
500
800
700
600
500
400
300
200
100
0


Distribution of Posts_(videos)
Count
Posts (videos)
0
200
400
600
800
800
700
600
500
400
300
200
100
0


Distribution of Posts_(carousel)
Count
Posts_(carousel)
0
100
200
300
400
500
600
700
800
700
600
500
400
300
200
100
0

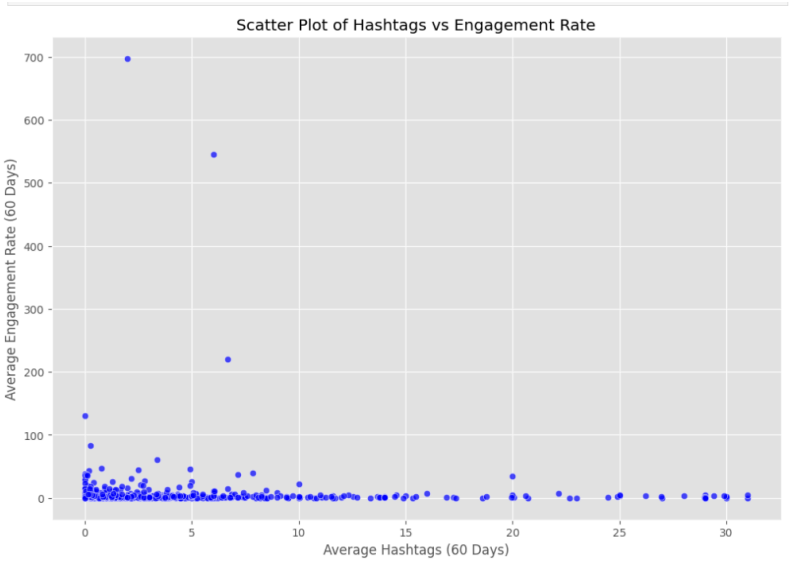
The distribution of images, videos, and carousels are highly skewed to the right as they rely on the posting frequency over the last 60 days. These distributions are mostly the same as the posting frequency with videos having the most frequency followed by carousels, and images.

Key Insight: The distributions of post types (images, videos, carousels) mirror the skewness of posting frequency, with videos being the most frequent post type, followed by carousels and images.



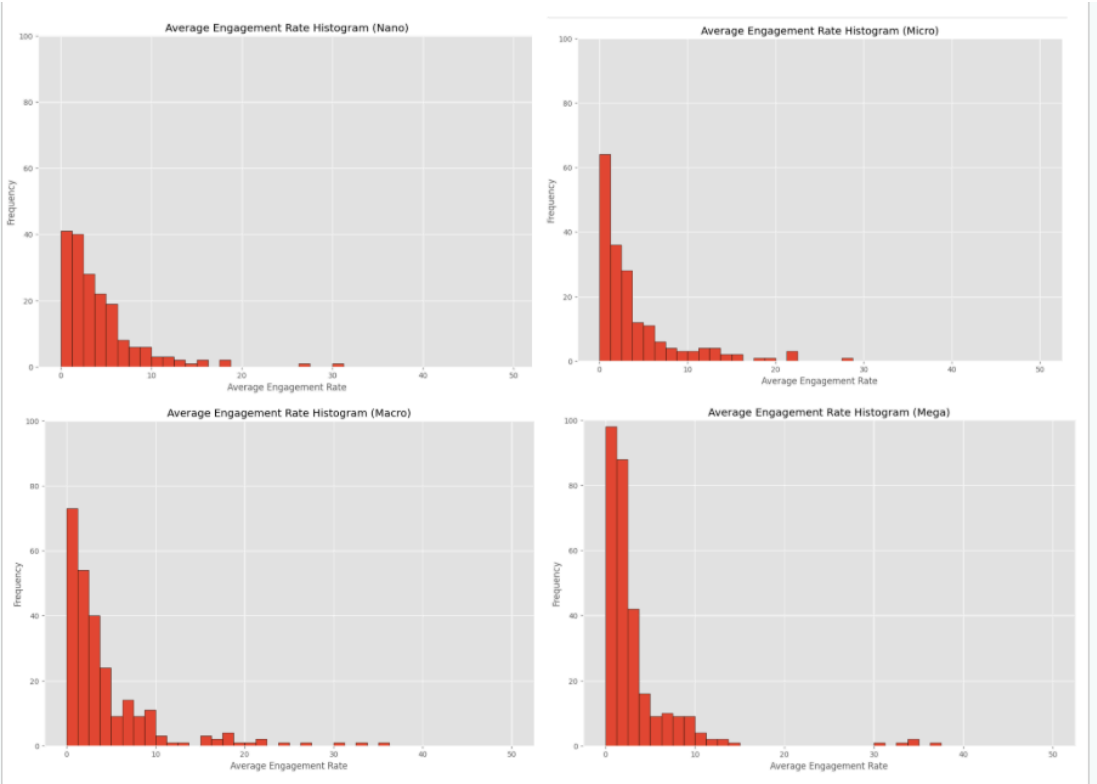
Based on the following scatterplot representing the average engagement rate and the followers, it shows that the more followers an account has, doesn't mean more engagement rate. Making average engagement rate not linear.

Key Insight: Average engagement rate is not linear.



Another scatter plot represents the relation between the average hashtags used with the average engagement rate of the 60 days. It shows that more hashtags used don't guarantee a better engagement rate, making both variables have no correlation with each other. The posts with smaller amount of average hashtags were even to have more engagement rate than those with bigger amount of hashtags used.

Key Insight: The average hashtags used have no correlation with the engagement rate.



Based on the following histograms representing the engagement rate for different account sizes, it can be seen that mega accounts are able to reach the highest engagement rate. On the other hand, the macro accounts fall short below mega in reaching the peak, but it is more consistent (Less Outliers). Additionally, the micro and nano accounts are more packed in the beginning, indicating low engagement rates.

Key Insights: On overall, nano accounts have the lowest average engagement rate, where micro have the lowest peak, and mega have the highest peak, followed by macro.

V. Conclusion

This study explores the relationship between features and the average engagement rates on Instagram to help brands in identifying suitable influencers for their marketing campaigns. We used a dataset consisting of 922 accounts spanning diverse influencer sizes (nano (1k - 10k), micro (10k - 100k), macro (100k - 1M), and mega (>1M), we examined key metrics, including average likes, comments, engagement rates (overall and last 60 days), and post type ratios.

Based on the result using the Pearson correlation coefficient test analysis, we can reject the null hypothesis as P-value < 0.05, indicating there is a significant relationship between average likes and engagement rate. Which means that our alternative hypothesis is supported.

For the main insights received from the data, it is shown that the average engagement rate is not linear, indicating larger accounts do not guarantee a higher engagement rate. Additionally, from the posts posted, videos receive the most engagement, followed by carousel, and images. It is also shown that the number of hashtags used per post have little to no impact towards the engagement rate. The main variables that have the highest correlation with the average engagement rate are average engagement rate of the 60 days, average likes, and average comments. For the average engagement rate of the 60 days (Which has the highest correlation with the engagement rate), has a subtle correlation to the ratio of post types posted.

Overall, this research highlights that effective influencer selection depends on analyzing engagement quality and audience relevance, rather than focusing solely on follower count. Brands should prioritize influencers with high engagement rates relative to their audience size, particularly those excelling in creating engaging video content. For influencers, these insights emphasize the need to maintain consistent, high-quality posting strategies to remain competitive in the rapidly evolving Instagram landscape.

By aligning these findings with marketing goals, stakeholders can optimize resource allocation, improve campaign effectiveness, and foster better connections between brands and target audiences.

VI. Presentation

PPT link: https://www.canva.com/design/DAGZbvmqOIg/8FXAP1hayWKL0m3Hgp9gBQ/edit

Code (Github): https://github.com/gamakagami/FoDS-FinalProject

Merged Dataset:
https://github.com/gamakagami/FoDS-FinalProject/blob/main/data/final/cleaned/merged_z5.csv

VII. References

[1] A. Radu, “Social media statistics you should know in 2024,” SocialBee. Available: https://socialbee.com/blog/social-media-statistics/ (accessed Dec 2024).

[2] S. Larson, “Social media users 2024 (global data & statistics),” PrioriData. Available: https://prioridata.com/data/social-media-usage/ (accessed Dec 2024).

[3] Backlinko, “Instagram statistics: Key demographic and user numbers,” BackLinko. Available: https://backlinko.com/instagram-users (accessed Dec 2024).

[4] N. Kumar, “How many people use instagram 2024 [new data],” DemandsAge. Available: https://www.demandsage.com/instagram-statistics/. (accessed Dec 2024).

[5] K. Kuligowski, “12 reasons to use Instagram for your business,” Business. https://www.business.com/articles/10-reasons-to-use-instagram-for-business/ (accessed Dec. 18, 2024).

[6] J. Zote, “Instagram statistics you need to know for 2024 [updated],” SproutSocial. Available: https://sproutsocial.com/insights/instagram-stats/ (accessed Dec 2024).

[7] I. Media, “Quality vs. quantity in social media,” InklingMedia. Available: https://inklingmedia.net/quality-vs-quantity-in-social-media/ (accessed Dec 2024).

[8] B. Schivinski, G. Christodoulides, and D. Dabrowski, “Measuring consumers’ engagement with brand-related social-media content: Development and validation of a scale that identifies levels of social-media engagement with brands,” Journal of Advertising Research, vol. 56, no. 1, pp. 1–18, 2016.

[9] Socialbook, “Top 200 instagrammers,” n.d. Available: https://socialbook.io/instagram-channel-rank/top-200-instagrammers.