Suggestions for Influencers: Increasing Likes & Comments on Instagram

Marketing Analytics Final Project Report

Olivia Cava

Shubham Chaudhary

Sai Krishna Chaitanya Srikonda

MKT 568

Prof. Reshadi

Fall 2024

Table of Contents

Executive Abstract

Instagram influencers struggle to grow their platforms due to a lack of actionable insights on how to increase likes and comments on their posts.

To address this problem, we analyzed a dataset of over 18,000 Instagram posts using a structured data mining approach based on the CRISP-DM methodology. The dataset included engagement metrics such as likes and comments, content attributes like hashtags, captions, and post types, and timing variables. Our process involved cleaning and transforming the data, creating new variables, and running regression analyses to identify factors influencing engagement. Key questions guiding our analysis were: *When should influencers post?* and *What should they post?*

Our findings revealed that timing and content significantly impact engagement. Posts during evenings and nights generated higher likes, while caption features such as exclamation marks positively influenced likes and comments. On the other hand, videos received lower engagement compared to photos. Hashtags, while helpful in moderation, were negatively correlated with comments when overused. For micro-influencers, including multiple people in photos and using emotional punctuation boosted engagement, while macro-influencers benefited more from strategic hashtag use and thought-provoking captions. Despite the relatively strong model for predicting likes ($R^2 = 39.2\%$), predicting comments proved more challenging, with an R^2 of only 3.2%, suggesting unexplored factors may drive comment engagement.

In conclusion, these insights offer actionable recommendations for influencers to optimize their posting strategies, enabling them to grow their reach and foster deeper audience connections.

Introduction

Instagram influencers face significant challenges in understanding what actions can help them increase engagement, specifically likes and comments. Despite their desire to grow their platforms, the absence of clear, actionable insights leaves influencers unsure about optimizing their content and posting strategies.

Engagement is crucial for influencers, as it directly impacts their visibility, follower retention, and monetization opportunities. From Instagram's perspective, higher engagement rates contribute to app activity, advertising revenues, and user retention. Additionally, users benefit from highly engaging influencer content, as it fosters inspiration, entertainment, and a sense of community.

This issue concerns several stakeholders: influencers who want to build their brands, Instagram as a business seeking sustained app usage and revenue growth, and everyday users who rely on influencers for entertainment, advice, and social connection. Solving this problem has widespread benefits, from improving individual influencers' livelihoods to enriching the overall Instagram user experience.

Our plan focuses on identifying optimal strategies for Instagram influencers by analyzing their historical data. Specifically, we aim to answer the questions:

1. **“When should I post?”** by evaluating time-based patterns of engagement.
2. **“What should I post?”** by examining content characteristics such as hashtags, post type (photo or video), and caption features.

To achieve this, we conducted a data-driven study using the CRISP-DM framework. Key steps included data cleaning, feature engineering, and regression analysis to uncover actionable insights.

Methodology

Our team followed the CRISP data mining methodology, a standardized 6-step process widely used in data science projects (Hotz, 2018). We began with Business Understanding, using the Final Project Instructions Document as our guide. In Data Understanding, we assessed the dataset's format and quality, followed by Data Preparation where we cleaned and transformed the data. During the Modeling phase, we applied various data mining techniques learned in class. We then evaluated our results and completed the Deployment phase by creating a report, presentation, and Instagram-targeted infographic. This structured approach ensured an organized and effective study (Hotz, 2018).

Data

Corentin Dugué and team at Towards Data Science collected Instagram influencer data using the Iconosquare Index Influencers directory, which provided access to over 2,000 verified influencer profiles (Dugué, 2017). Due to Instagram's API limitations of 60 requests per hour, the team developed a custom web scraper to collect profile metadata and information about users' 17 most recent posts. To focus specifically on influencers rather than celebrities, they filtered out posts exceeding 200,000 average likes and 1,000,000 followers (Dugué, 2017). The original dataset contained approximately 30,000 records, though Prof. Reshadi later acquired and cleaned a sample for student use.

The dataset includes both categorical and continuous variables. LINK serves as the unique identifier for each post, while USERNAME identifies the posting influencer. The continuous variables include FOLLOWERS, FOLLOWING, LIKES, and COMMENTS, which track engagement metrics. Categorical variables include TEXT (post captions), DATE (posting time), TYPE (photo or video), List_of_tags (hashtags used), and List_of_mentions (users mentioned). Corresponding continuous variables number_of_tags and number_of_mentions track the quantity of hashtags and mentions. USERS IN PHOTO counts individuals present in images.

Additional variables were created during analysis to extract more insights. These include Month_of_post, day_of_post, and timing_of_post (categorized as morning, afternoon, evening, and night), all derived from the DATE variable. Length_of_post was created to measure caption length. The dataset contained 19,681 records before cleaning and transformation.

Numerical Variable Descriptive Statistics

In the Python analysis, date and post type (video/photo) were detected as numerical variables due to dummy coding, but were excluded from the metrics table as they are categorical in nature. The DATE variable is included to show the collection period's range, though it lacks a standard deviation value due to its datetime format..

Numerical Variables	**Minimum**	**Maximum**	**Average**	**Standard Deviation**
FOLLOWERS	1.799300e+04	1.134619e+06	6.256413e+04	1.042349e+05
FOLLOWING	0	7586.000000	1489.766831	2252.675356
LIKES	0	158338.000000	2497.766983	5574.988136
COMMENTS	0	26011.000000	39.825111	447.972795
DATE	2016-03-27 08:21:35	2017-05-02 17:07:30	2017-04-21 05:07:52.777196288	NaN
USERS IN PHOTO	0	20.0	1.072261	2.399112

number_of_tags	0	41.000000	6.737005	8.782144
number_of_mentions	0	34.000000	0.723591	1.704316
length_of_post	1	2191.000000	151.703864	180.574815

Categorical Variable Unique Values and Counts

For each categorical variable included in this dataset, and prior to any dummy coding or feature engineering, there are a set of unique values. Below is a table that provides this information. It is important to note that this table was created after removing outliers in the data cleaning steps. At this point, the dataset included a total of 18505 rows.

Categorical Variables	Unique Values and Counts
USERNAME	1,071 unique influencers
TEXT	16,290 unique captions, includes empty captions
LINK	18,505 unique links, no duplicates
list_of_tags	8,976 unique tag lists, includes posts without tags
list_of_mentions	3,827 unique mention lists, includes posts without mentions
month_of_post	Heavily skewed: April (17,235), May (1,248), February (12), March (5), January (3), December (2)
day_of_post	Evenly distributed: Sunday (2,957), Monday (2,791), Tuesday (2,192), Wednesday (2,025), Thursday (2,952), Friday (2,855), Saturday (2,733)
timing_of_post	Night (6,924), Evening (4,699), Afternoon (4,662), Morning (2,220). More night posts, fewer morning posts.
TYPE(1 PHOTO, 2 VIDEO)	Photos (16,727), Videos (1,778). Predominantly photos.

Null Values Present in Dataset

Prior to any data cleaning or transformation, there were a few variables with null values that had to be treated later on. The following table describes all variables with null values, where variables without any null values are not included in the table. It is important to note these numbers are gathered before any data cleaning steps, and the dataset at this point had 19681 entries.

Variables	Number of Null Values
TEXT	6
USERS IN PHOTO	8527
list_of_tags	5819
list_of_mentions	12935

Independent and Dependent Variables

Our analysis examines how different factors affect Instagram post engagement (likes and comments). We used num_likes and num_comments as dependent variables, while independent variables included username, num_followers, num_following, post_caption, month_of_post, day_of_post, timing_of_post, post_type, num_users_in_photo, list_of_tags, num_hashtags, list_of_mentions, and num_mentions. Some variables were later dummy coded (for example, month_of_post was split into separate month columns) and new variables were added to the independent variables list during analysis.

Independent Variables	Dependent Variables
FOLLOWERS, FOLLOWING, DATE, USERS IN PHOTO, number_of_tags, number_of_mentions, length_of_post, list_of_tags, list_of_mentions, month_of_post, day_of_post, timing_of_post, TYPE(1 PHOTO, 2 VIDEO)	LIKES, COMMENTS

Data Cleaning

Our team's main data cleaning challenge was deciding between multiple valid approaches for handling data issues. For instance, with missing values in continuous variables, we could choose between replacing nulls with zero, mean, mode, or median values. We learned to base these decisions on context while maintaining transparency about our choices to better understand their impact on results. Real-world data cleaning proved more complex than classroom examples, requiring additional steps beyond the project guidelines, which we detail in this section and the Data Transformation section. Following the CRISP methodology, our data cleaning process was iterative rather than linear, often requiring us to revisit previous steps as new challenges emerged. For clarity, we've organized our process into three main phases: exploration, general cleaning, and feature engineering/data transformation. The final phase included additional preparation specific to linear regression analysis.

During initial data exploration, we analyzed an Instagram dataset containing 19,681 rows and 14 columns. The columns included user metrics (username, follower count, following count), post engagement metrics (likes, comments), content details (caption, post date, media type, users tagged, hashtags, mentions), and post URLs. While rich in information, the data required cleaning before analysis.

Before dropping any columns, we reviewed data types and identified that 'USERS IN PHOTO' was incorrectly stored as an object type instead of integer, with 8,527 rows containing '-' values. We handled these missing values by first converting them to null values, then to a continuous variable, and finally filling nulls with the column mean rather than zeros to avoid skewing the already zero-heavy distribution. While we planned to eventually remove the link column as it served only as a unique identifier, we retained it during initial inspection. We standardized variable naming conventions by converting inconsistent formats (all caps vs. lowercase with underscores) to a single cohesive style and improved variable names for clarity and future use.

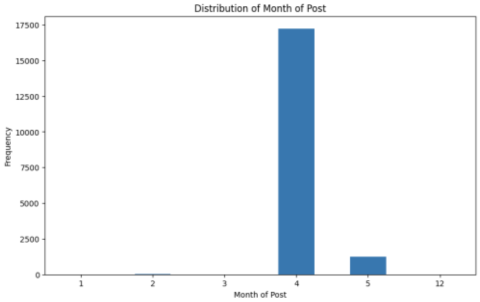
We filled missing values in post_caption, list_of_tags, and list_of_mentions with empty strings rather than placeholder text to avoid skewing length-based calculations. After reviewing descriptive statistics for all numerical variables, we confirmed their ranges were reasonable and required no additional processing. We standardized spelling in categorical variables and verified data integrity by checking for both complete and partial duplicates using the link_to_post column. Finding no duplicates, we then dropped the link_to_post column as it wasn't needed for analysis.

In depth cleaning

During in-depth cleaning, we converted the post_type column (1=photo, 2=video) into dummy variables is_photo and is_video, then dropped is_photo. We identified outliers by calculating z-scores for continuous variables, using a threshold of 3. All continuous variables except num_following contained outliers, which we removed after analyzing their ranges, means, and frequencies. This reduced our dataset from 19,681 to 18,505 rows while maintaining 14 columns.

Values	**num_followers**	**num_likes**	**num_comments**	**num_users_in_photo**	**num_hashtags**	**num_mentions**
Outlier Values Range	(406776, 1134619)	(19235, 158338)	(1402, 26011)	(7, 20)	(34, 41)	(6, 34)
Mean Value of Column	62564.12524770083	2497.7669833849905	39.8251105126772	1.040953203597378	6.737005233473909	0.723591280930847
Number of Outliers	337	239	70	356	15	382

We standardized non-dummy variables for regression analysis, verifying means near 0 and standard deviations near 1. Analysis of the correlation matrix and VIF scores guided our feature engineering. We combined Evening and Night posting times due to their strong negative correlation (-0.45). Given the highly skewed monthly distribution (17,235 posts in April versus 2-1,248 in other months), we replaced individual month dummies with a single non_april_month indicator. We noted, though, that making recommendations based on month may not be fair, due to the unbalanced nature of this variable.



After feature engineering, we validated our decisions through updated correlation matrix and VIF analysis. The highest VIF values were Evening_or_Night (2.48) and Monday (1.72), both well under the critical threshold of 5. The non_april_month variable showed appropriate VIF (1.45) and correlations. While length_of_post and num_hashtags maintained a correlation of 0.61 (VIFs: 1.65, 1.62), they appeared to capture distinct features. Engagement metrics (num_followers, num_following, num_users_in_photo) had low VIF values (<1.1), indicating independent contributions.

Data Transformation

We used several key data transformation techniques to enhance the clarity, consistency, and usability of the dataset for analysis. We renamed several columns to improve readability and make them more descriptive. For instance, 'USERNAME' was renamed to 'username', 'FOLLOWERS' was renamed to 'num_followers', and 'LIKES' was renamed to 'num_likes'. We did this to ensure that the column names were consistent and accurately reflected the data they contained, making it easier to understand and work with the dataset.

We identified and corrected erroneous values in the 'num_users_in_photo' column. Specifically, we replaced entries marked as '-' with None to represent missing or invalid data. We also changed the data type of this column to a nullable integer type (Int64), allowing it to handle missing values while preserving its numerical integrity for analysis. We also did systematic handling of missing values across various columns. We filled missing values in the 'num_users_in_photo' column with the mean value of the column to ensure consistency in the dataset. For categorical columns like 'post_caption', 'list_of_tags', and 'list_of_mentions', we replaced missing values with empty strings to maintain the structure and avoid errors during data processing. These transformations were essential for preparing the dataset for further analysis, ensuring it was clean, consistent, and structured in a way that would allow for meaningful insights.

Analysis Results

Question 1

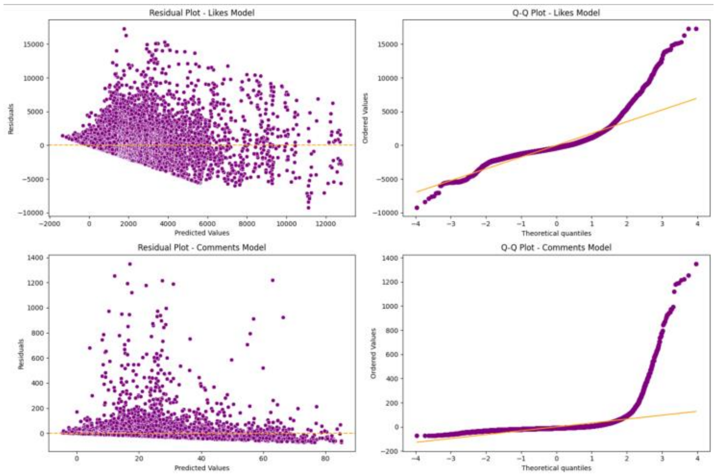
Model Performance Summary Table:

Metric	Likes Model	Comments Model
R-squared	39.2%	3.2%
F-statistic	744.4	38.67
RMSE	1927.81	59.33

5 Most Influential Variables Table:

Rank	Likes Model	Effect	Comments Model	Effect
1	Number of followers	+0.576	Number of followers	+0.139
2	Number of following	-0.149	Length of post	+0.077
3	Video posts (vs photos)	-0.084	Number of following	-0.063
4	Non-April month	-0.038	Number of hashtags	-0.054
5	Morning posts	-0.030	Number of users in photo	-0.025

Residuals & Q-Q Plot for Likes and Comments Model



Looking at how well we can predict Instagram engagement, we found some interesting patterns. Our model was decent at predicting likes (explaining about 39% of what drives likes) but wasn't great at predicting comments (only explaining about 3% of what drives comments).

The number of followers had the strongest impact on likes - when an account's follower count was one standard deviation above average, it increased likes by 0.58 standard deviations. Following count had the second strongest effect, with a one standard deviation increase leading to a 0.15 standard deviation decrease in likes. Posts with videos (versus photos) had the third strongest effect, decreasing likes by 0.08 standard deviations. Posts made outside of April decreased likes by 0.04 standard deviations, and morning posts decreased likes by 0.03 standard deviations. It is important to remember, though, that there was a heavily uneven distribution of months present in the dataset. So, actual recommendations regarding months should not be used.

For comments, the patterns were similar but the effects were smaller. A one standard deviation increase in followers was associated with a 0.14 standard deviation increase in comments. Posts that were one standard deviation longer than average saw a 0.08 standard deviation increase in comments. Accounts with following counts one standard deviation above average experienced a 0.06 standard deviation decrease in comments. Using more hashtags (one standard deviation above average) led to a 0.05 standard deviation decrease in comments. What's really interesting is how much more predictable likes are compared to comments. While these factors give us a good idea about what drives likes, comments seem to be influenced by other things we haven't captured in our data, maybe things like the post's content or how engaging the caption is.

All in all, morning posts tend to get fewer likes, so influencers might want to experiment with posting at other times. While April showed better engagement overall, we are cautious about reading too much into that since we do not have even data across all months. The characteristics of influencer accounts, especially your follower count, and content choices such as post type, length, and timing, matter more for predicting likes than comments.

Question 2

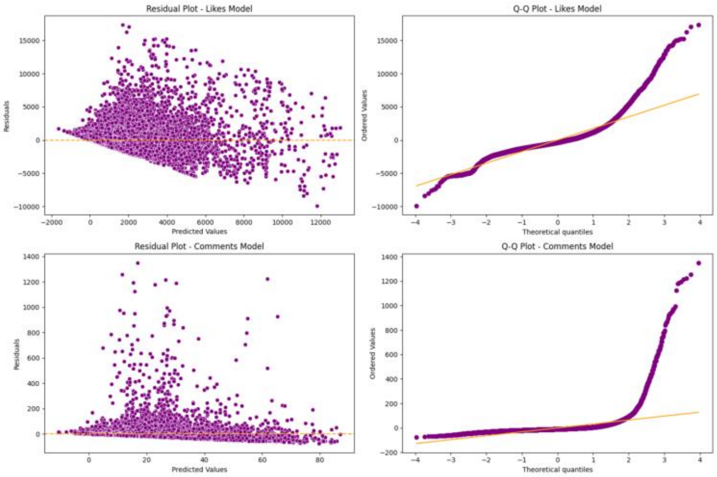
New Variables Created Table (with our reasoning behind each one):

Variable	Description	Rationale
count_of_question	Number of question marks in caption	Questions invite engagement and responses
count_of_exclamation	Number of exclamation points in caption	Indicates enthusiasm and emotional content
popular_tag_count	Count of top 25 hashtags used	Popular hashtags increase visibility

Model Performance Before and After New Variables Created:

Metric	Original Likes Model	New Likes Model	Original Comments Model	New Comments Model
R-squared	39.2%	39.8%	3.2%	3.4%
F-statistic	744.4	644.0	38.67	34.38
RMSE	1927.81	1917.56	59.33	59.28

Residuals & Q-Q Plot for Likes and Comments Model



Based on the analysis, our team created three new variables to study social media engagement: count of question marks in captions, count of exclamation points, and the usage of top 25 popular hashtags. These variables were chosen because questions naturally invite user interaction, exclamation points indicate enthusiasm and emotional content, and popular hashtags potentially increase post visibility.

The regression analysis revealed interesting patterns in how these variables affect engagement. For likes, exclamation points showed a significant positive effect, with each standard deviation increase in exclamation points corresponding to a 0.073 standard deviation increase in likes. Popular hashtags also positively influenced likes, with a 0.039 standard deviation increase in likes for each standard deviation increase in popular tag usage. Question marks, however, did not significantly impact likes. For comments, the effects were different - a one standard deviation increase in question marks was associated with a 0.030 standard deviation increase in comments, while exclamation points led to a 0.019 standard deviation increase in comments. Interestingly, popular hashtags showed a negative effect on comments, with each standard deviation increase in popular tag usage associated with a 0.025 standard deviation decrease in comments.

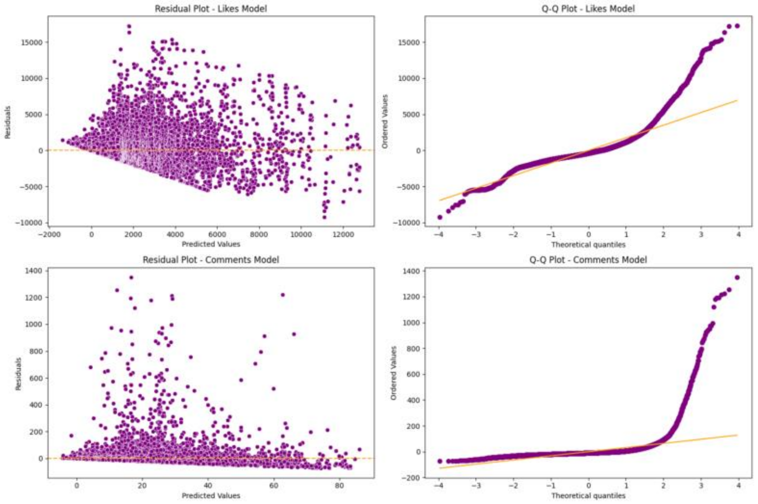
The overall model performance showed modest improvements after adding these new variables. The likes model's R-squared increased from 39.2% to 39.8%, with the RMSE improving from 1927.81 to 1917.56. The comments model saw similar small gains, with R-squared increasing from 3.2% to 3.4% and RMSE improving slightly from 59.33 to 59.28. While these improvements are relatively small, they suggest our new variables contribute meaningfully to understanding engagement patterns. The residual plots and Q-Q plots revealed some heteroscedasticity and heavy tails, indicating the models have difficulty predicting extreme engagement values accurately. Despite these limitations, the statistical significance of most new variables suggests they provide valuable insight into the factors driving social media engagement.

Question 3

Model Performance Before and After “is_weekend” Was Created:

Metric	Original Likes Model	is_weekend Likes Model	Original Comments Model	is_weekend Comments Model
R-squared	39.2%	39.2%	3.2%	3.2%
F-statistic	744.4	1082.0	38.67	55.89
RMSE	1927.81	1928.27	59.33	59.34

Residuals & Q-Q Plot for Likes and Comments Model



Our team conducted a regression analysis to examine the impact of weekend posting versus weekday posting on social media engagement. We replaced the individual day-of-week variables with a single binary "is_weekend" variable to determine if this simplification would improve model performance. The regression results showed that the "is_weekend" variable was not statistically significant for either likes (p = 0.248) or comments (p = 0.982). The standardized coefficient for likes was -0.007, suggesting that weekend posts had a very slight negative effect on likes compared to weekday posts, though this difference was not statistically significant. For comments, the near-zero standardized coefficient of 0.0002 indicated virtually no difference between weekend and weekday posting.

Comparing model performance, the simplified weekend model did not improve predictive power. For the likes model, while the F-statistic increased from 744.4 to 1082.0, the R-squared remained unchanged at 39.2%, and the RMSE slightly worsened from 1927.81 to 1928.27. Similarly, for the comments model, though the F-statistic increased from 38.67 to 55.89, the R-squared held steady at 3.2%, and the RMSE marginally increased from 59.33 to 59.34. Other variables showed more significant influence on engagement. For likes, follower count (standardized beta = 0.576, p < 0.001), following count (standardized beta = -0.150, p <

0.001), and video posts (standardized beta = -0.085, p < 0.001) had strong effects. For comments, follower count (standardized beta = 0.139, p < 0.001), following count (standardized beta = -0.063, p < 0.001), and post length (standardized beta = 0.077, p < 0.001) were the most influential factors. These results suggest that the timing of posts between weekdays and weekends is less important for engagement than other factors like account characteristics and content type.

Question 4

Performance of Micro vs Macro Influencer Models:

Metric	Micro Likes Model	Micro Comments Model	Macro Likes Model	Macro Comments Model
R-squared	12.1%	1.2%	32.8%	3.6%
F-statistic	125.8	11.36	199.0	15.33
RMSE	1322.27	48.30	2741.48	77.69
Sample Size	12,776	12,776	5,729	5,729

Micro Influencer vs Macro Influencer Likes Coefficients and p-values:

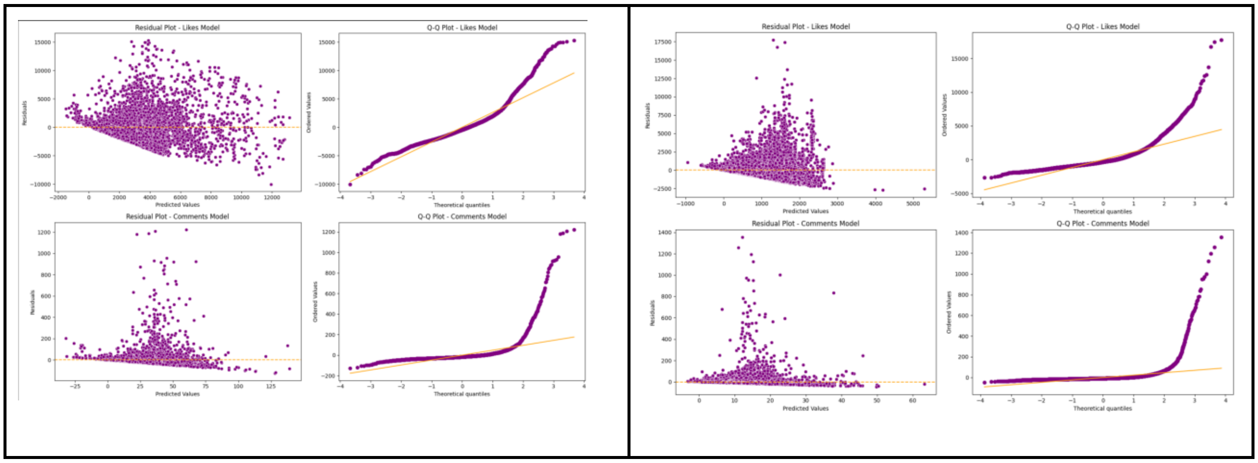
Independent Variable	Micro Influencer Standardized Beta	Micro Influencers p-value	Macro Influencers Standardized Beta	Macro Influencers p-value
num_followers	+0.160	<0.001	+0.506	<0.001
num_following	-0.184	<0.001	-0.169	<0.001
num_users_in_photo	+0.039	<0.001	-0.017	0.146
num_hashtags	-0.004	0.767	-0.091	<0.001
num_mentions	-0.028	0.002	-0.013	0.258
post_is_video	-0.093	<0.001	-0.100	<0.001
length_of_post	-0.067	<0.001	-0.029	0.061
non_april_month	-0.046	<0.001	-0.032	0.004
is_morning	-0.036	<0.001	-0.051	<0.001

Independent Variable	Micro Influencer Standardized Beta	Micro Influencers p-value	Macro Influencers Standardized Beta	Macro Influencers p-value
num_followers	+0.160	<0.001	+0.506	<0.001
count_of_question	+0.007	0.407	+0.004	0.708
count_of_exclamation	+0.116	<0.001	+0.063	<0.001
popular_tag_count	+0.156	<0.001	-0.024	0.099
evening_or_night	-0.010	0.302	+0.023	0.066
is_weekend	-0.008	0.363	-0.010	0.357

Micro Influencer vs Macro Influencer Comments Coefficients and p-values:

Independent Variable	Micro - Comments Coef.	Micro - Comments p-value	Macro - Comments Coef.	Macro - Comments p-value
num_followers	+0.038	<0.001	+0.097	<0.001
num_following	-0.064	<0.001	-0.064	<0.001
num_users_in_photo	+0.013	0.153	-0.029	0.045
num_hashtags	+0.019	0.136	-0.132	<0.001
num_mentions	-0.007	0.460	-0.022	0.120
post_is_video	-0.006	0.478	+0.009	0.495
length_of_post	+0.057	<0.001	+0.098	<0.001
non_april_month	-0.018	0.041	-0.014	0.294
is_morning	-0.014	0.181	-0.015	0.314
count_of_question	+0.010	0.256	+0.076	<0.001
count_of_exclamation	+0.034	<0.001	-0.001	0.941
popular_tag_count	-0.036	0.001	+0.018	0.319
evening_or_night	-0.020	0.051	+0.036	0.015
is_weekend	+0.001	0.877	-0.000	0.982

Macro Influencers Residuals & Q-Q Plot for Likes and Comments Models	Micro Influencers Residuals & Q-Q Plot for Likes and Comments Models



Our team conducted a comparative regression analysis examining how different factors affect engagement for micro versus macro influencers. Since the variables are standardized, the coefficients represent the change in the dependent variable in terms of standard deviations for a one standard deviation increase in the independent variable. This analysis revealed several distinct patterns in how various factors influence engagement differently based on influencer size.

For likes, the analysis showed several noteworthy differences between micro and macro influencers. Number of users in photos had a positive effect for micro-influencers (standardized beta = +0.039, p < 0.001) but was not significant for macro-influencers (standardized beta = -0.017, p = 0.146). Hashtag count showed an opposite pattern, significantly affecting macro-influencers (standardized beta = -0.091, p < 0.001) but not micro-influencers (standardized beta = -0.004, p = 0.767). Popular tag count demonstrated an interesting reversal: it had a positive influence for micro-influencers (standardized beta = +0.156, p < 0.001) versus a marginally negative effect for macro-influencers (standardized beta = -0.024, p = 0.099).

The comment analysis revealed equally interesting points. Exclamation points significantly increased comments for micro-influencers (standardized beta = +0.034, p < 0.001) but had no significant effect for macro-influencers (standardized beta = -0.001, p = 0.941). Question count showed the opposite pattern, significantly affecting macro-influencer comments (standardized beta = +0.076, p < 0.001) but not micro-influencers (standardized beta = +0.010, p = 0.256). Post length also showed different effects, with longer posts negatively impacting micro-influencer likes (standardized beta = -0.067, p < 0.001) but positively affecting macro-influencer comments (standardized beta = +0.098, p < 0.001).

These findings suggest that engagement strategies should be tailored based on influencer size. Micro-influencers benefit more from including multiple people in photos, using popular hashtags, and incorporating emotional punctuation, while macro-influencers see better results from strategic hashtag usage and incorporating questions in their captions. The residual plots and Q-Q plots show similar patterns of heteroscedasticity for both groups, indicating that the models' prediction accuracy varies with engagement levels. This suggests that while these factors are significant, other unmeasured variables likely play important roles in determining viral success.

Conclusions and Recommendations

In this study, we analyzed Instagram influencer data containing over 18,500 posts from 1,071 unique influencers to understand the factors that drive post engagement through likes and comments. Using multiple linear regression models, we examined how various factors such as follower count, posting time, content type, and caption characteristics affect engagement levels. While our initial combined model explained approximately 39% of the variance in likes and 3% of the variance in comments, we chose to focus our recommendations on separate models for micro and macro influencers, despite their lower explanatory power. This decision was made because the differentiated models better capture the unique dynamics and engagement patterns specific to each influencer category, providing more targeted and actionable insights for influencers based on their audience size. Through feature engineering and iterative analysis, we explored the impact of new variables such as question marks, exclamation points, and popular hashtags, which provided additional insights into engagement patterns.

Our analysis revealed several key factors that significantly influence Instagram engagement. From a technical perspective, follower count was the strongest predictor of engagement, with a one standard deviation increase in followers associated with a 0.576 standard deviation increase in likes and a 0.139 standard deviation increase in comments. The impact of content strategies varied notably between micro-influencers and macro-influencers. In general audience terms, we found that micro-influencers saw better engagement when they included more people in their photos and used popular hashtags, while their posts with longer captions received fewer likes. For macro-influencers, the patterns were different, where using questions in captions and writing longer captions increased comments, but using too many hashtags decreased likes. Across both groups, we found that morning posts generally received less engagement than posts at other times of day, and photo posts typically performed better than videos for generating likes.

Based on our findings, we recommend different strategies for micro and macro influencers to optimize their Instagram engagement. For micro-influencers, we recommend including more people in photos to increase likes, using exclamation points to drive comments, and keeping captions concise since longer captions were associated with decreased likes. For macro-influencers, we suggest incorporating questions in captions to boost comments, writing longer, more detailed captions to encourage discussion, and being selective with hashtag usage as excessive hashtags were linked to decreased engagement. For all influencers, we found that photo posts generally performed better than videos for generating likes, and morning posts tended to receive less engagement than posts at other times of day. While these strategies can help optimize engagement, it's important to note that social media success involves many factors beyond what our models could capture. Future research could explore more advanced techniques like clustering analysis and neural networks to uncover additional patterns in influencer engagement.

References

Hotz, N. (2018, September 10). *What is CRISP DM?* Data Science PM. https://www.datascience-pm.com/crisp-dm-2/

Dugué, C. (2017, May 14). *Predicting the number of likes on Instagram*. Medium. https://towardsdatascience.com/predict-the-number-of-likes-on-instagram-a7ec5c020203