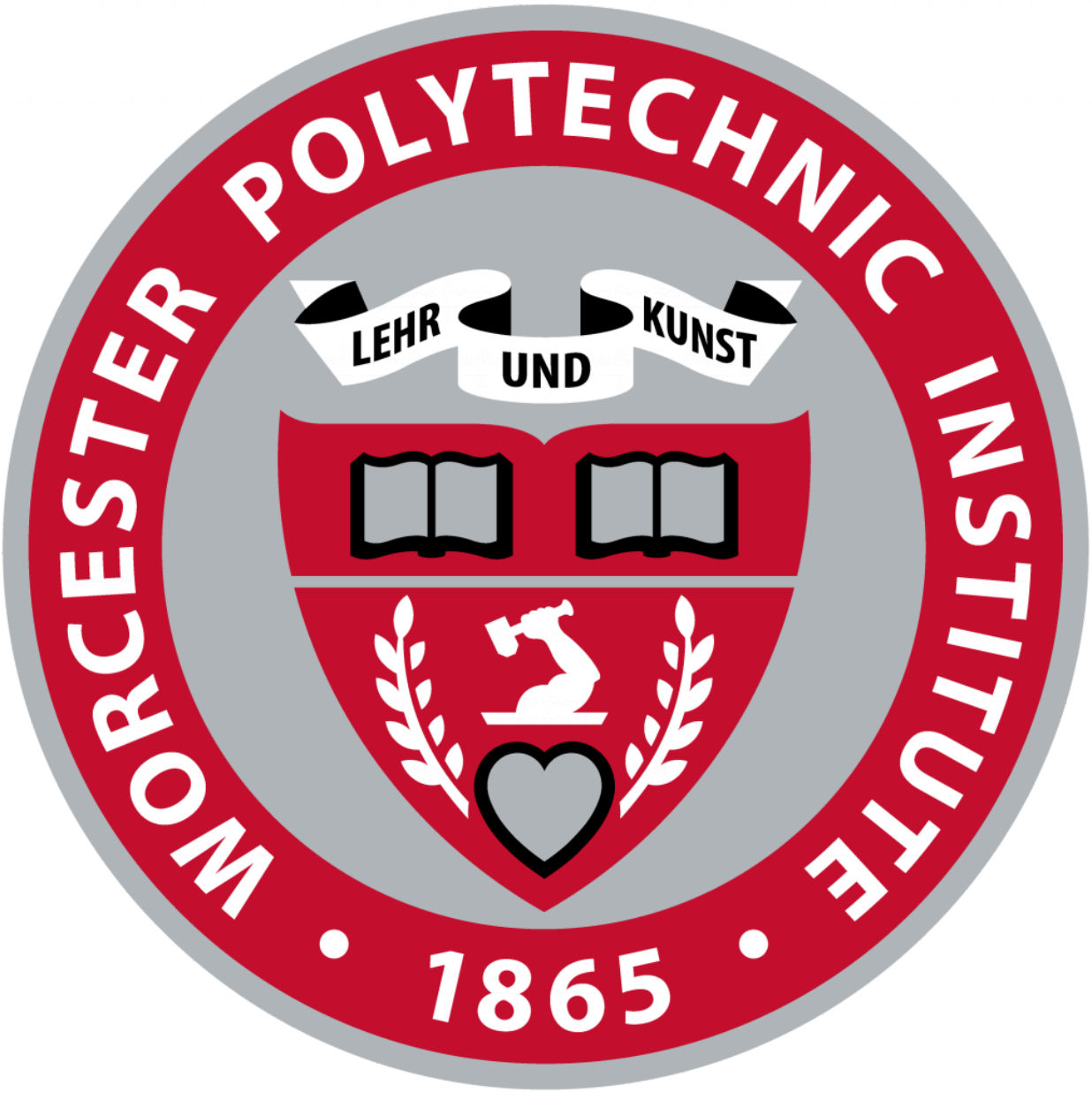

MKT 568 - Marketing Analytics Final Project

By

Ishan Deshpande
Brennan Hilger
Vishal Patidar

Table of Contents

Introduction

Problem Statement

Social media influencers often struggle to find ways to increase their user engagement, whether it be likes or comments, on their Instagram posts. While Instagram provides analytics specific to each account, these insights are limited to the data from the account of said influencer, making it hard to identify trends that apply more broadly. Without a deeper understanding of what drives engagement, influencers can find it challenging to create content that consistently resonates with their audience and expands their reach. With no innovative insights of what attracts their audience, how can up and coming influencers expect their audience to hunger for more content?

Why It's Important

Whether they be an influencer, a personal business, or any slight variation of an organization, the "Engagement" data is critical for any account using Instagram to connect with an audience. It helps gauge audience interest, improves visibility through Instagram's algorithm, and boosts all sorts of outreach. For businesses, high engagement often translates into better customer relationships, more website traffic, and higher sales. For influencers and organizations, Engagement data is the key to building influence, driving campaigns, and growing their communities. Without continuous engagement, the process of growing your audience, as well as income, can make achieving these goals much more difficult.

Who Cares?

The question of raising engagement to the average Instagram user may be exciting, but really cannot be beneficial to any account who isn't looking to monetize their posts. The topic of "Engagement" matters to influencers, marketers, businesses, and even government organizations who are constantly looking for different ways to grow their platform. Nonprofits can also bear fruits from bringing in various audiences through instagram; the more awareness for your cause, the more effectful your movement can become. Whether these accounts take part in promoting products, sharing informative content, or raising awareness for a charitable idea, all forms of business on Instagram benefit from understanding how to better engage their audience.

Our Plan

To solve this, we will analyze the large dataset of Instagram posts provided to us in order to figure out what drives likes and comments.

- We will start by exploring the data to understand trends and which factors seem to influence engagement.
- We then will look at specific features like post type (photo, video), captions, hashtags, posting time, and content themes.
- After this, we will use tools like statistical analysis and machine learning (linear regression) to identify patterns and predict what works best.
- Finally, we aim to create practical recommendations, like when to post, what type of content to focus on, and how to use captions and hashtags effectively.

Methodology

To analyze the dataset and identify strategies for maximizing Instagram engagement, our team followed the CRISP-DM methodology, as outlined below:

1. **Business Understanding**: Together, we defined the objective of increasing likes and comments by identifying key factors that drive user engagement on Instagram.
2. **Data Understanding**: The team collaboratively explored the dataset by examining trends, distributions, and important variables such as post type, hashtags, and posting times.
3. **Data Preparation**: We worked as a group to clean and process the data, addressing missing values, standardizing formats, and engineering meaningful features for analysis.
4. **Modeling**: Using a collaborative approach, we applied statistical techniques and machine learning algorithms to identify and predict the factors most strongly influencing engagement.
5. **Evaluation**: The team reviewed the model's performance to ensure the results were accurate and aligned with the project's goals.
6. **Deployment**: Finally, we combined our insights to create actionable recommendations, enabling influencers and businesses to design content that maximizes user engagement.

Data

How was the data collected?

The data collection process involved two main stages:

Gathering a List of Influencers: A list of over 2,000 Instagram influencers were sourced from the Iconosquare Index Influencers, a publicly available resource. As the list spanned more than 70 pages, a web crawler was used to extract the influencer handles and compile them into a usable format.

Scraping Instagram Profiles: Due to the limitations of Instagram's API, which allows only 60 requests per hour, Selenium, a web automation framework, was employed to programmatically crawl Instagram pages. The scraper collected metadata from influencer profiles, such as follower and following counts, number of posts, and profile descriptions. It also accessed the 17 most recent posts from each user, gathering granular details like image files (in JPG format), likes, comments, timestamps, and captions.

Variables

Variable Name	Type	Used in the Analysis
USERNAME	Categorical	No
FOLLOWERS	Continuous	Yes
FOLLOWING	Continuous	Yes

LIKES	Continuous	Yes
COMMENTS	Continuous	Yes
TEXT	Categorical	Yes
DATE	Continuous	Yes
TYPE (1 PHOTO, 2 VIDEO)	Continuous	Yes
USERS IN PHOTO	Categorical	Yes
LINK	Categorical	No
list_of_tags	Categorical	No
number_of_tags	Continuous	Yes
list_of_mentions	Categorical	No
number_of_mentions	Continuous	Yes

How many records exist in the dataset?

There are a total of 19681 records in the dataset.

Descriptive Statistics

Numerical variables - 'FOLLOWERS', 'FOLLOWING', 'LIKES', 'COMMENTS', 'number_of_tags', 'number_of_mentions'

	FOLLOWERS	FOLLOWING	LIKES	COMMENTS	number_of_tags	number_of_mentions
count	1.968100e+04	19681.000000	19681.000000	19681.000000	19681.000000	19681.000000
mean	6.256413e+04	1489.766831	2497.766983	39.825111	6.737005	0.723591
std	1.042349e+05	2252.675356	5574.988136	447.972795	8.782144	1.704316
min	1.799300e+04	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2.329900e+04	174.000000	420.000000	1.000000	0.000000	0.000000
50%	3.669900e+04	506.000000	1073.000000	5.000000	3.000000	0.000000
75%	6.279100e+04	1367.000000	2683.000000	17.000000	10.000000	1.000000
max	1.134619e+06	7586.000000	158338.000000	26011.000000	41.000000	34.000000

Categorical Variables

Categorical Variables - 'USERNAME', 'TEXT', 'DATE', 'USERS IN PHOTO', 'LINK', 'list_of_tags', 'list_of_mentions'

Variable	Unique number of values
USERS IN PHOTO	22
USERNAME	1094
list_of_mentions	4357
list_of_tags	9683
TEXT	17390
DATE	19538
LINK	19681

Null Values

Variable	Total Null Values
USERNAME	0
FOLLOWERS	0
FOLLOWING	0
LIKES	0
COMMENTS	0
TEXT	0
DATE	0
TYPE (1 PHOTO, 2 VIDEO)	0
USERS IN PHOTO	0
LINK	0
list_of_tags	5819
number_of_tags	0

list_of_mentions	12935
number_of_mentions	0

Variables used in the analysis

Independent Variables

Followers, Following, Month_Name_August, Month_Name_December, Month_Name_February, Month_Name_January, Month_Name_July, Month_Name_March, Month_Name_May, Month_Name_November, Month_Name_October, Month_Name_September, Day_of_Week_Monday, Day_of_Week_Saturday, Day_of_Week_Sunday, Day_of_Week_Thursday, Day_of_Week_Tuesday, Day_of_Week_Wednesday, Post_Timing_Evening, Post_Timing_Morning, Post_Timing_Night, Type_video, Text_Length, Number_of_tags, Number of Mentions, Number of Users in the Post

Dependent Variables

LIKES, COMMENTS

Data Cleaning

Missing Values:
Dropped 6 null records in the TEXT column.

Dropped Irrelevant Columns:
Removed list_of_tags and list_of_mentions columns as they were not needed for analysis.

Column Renaming and Transformation:
Renamed USERS IN PHOTO to Number_of_Users_in_Post.
Replaced non-numeric values with 0 and changed the data type to INT.

Dummy Encoding:
Applied dummy encoding to categorical variables (TYPE, POST_TIMING, DAY_OF_WEEK, MONTH_NAME), Dropped redundant dummy columns: Post_Timing_Afternoon, Day_Of_Week_Friday, Month_Name_April, TYPE_PHOTO.

Outlier Detection and Removal:
Selected variables: FOLLOWERS, FOLLOWING, Number_of_Users_in_Post, number_of_tags, number_of_mentions, Text_Length, LIKES, COMMENTS.
Calculated Z-scores and filtered out rows with Z-scores outside the ±3 range.
Reduced records from 19,675 to 18,276.

Data Transformation

New Variables Created:

MONTH: Extracted to display the month name for better understanding of data trends.
Day of Week: Created to represent the day of the week, helping identify any weekly patterns.
Post Timing: Categorized into Afternoon, Morning, Evening, and Night to capture post timing for analysis.
Text Length: Created to count the length of the text in each post, which could correlate with engagement.

Dummy Encoding:
Dummy encoding was applied to the following categorical variables:
TYPE (1 PHOTO, 2 VIDEO), POST_TIMING, DAY_OF_WEEK, MONTH_NAME
Dummy coding was used to convert categorical variables into a numeric format in order to complete our regression analysis.

Dropped Redundant Dummy Variables:
Dropped the following dummy columns to avoid redundancy:
Post_Timing_Afternoon, Day_Of_Week_Friday, Month_Name_April, and TYPE_PHOTO.

Changed Data Types:
Changed dummy-coded variables to integer data type to ensure they could be used in regression analysis.

Analysis Results

Question 1

LIKES

1. What analysis did you run for this question?
We ran an Ordinary Least Squares (OLS) regression analysis to model the relationship between LIKES (dependent variable) and several predictors:

Numerical predictors: FOLLOWERS, FOLLOWING, Text_Length, number_of_mentions, etc.
Categorical predictors: These were one-hot encoded into dummy variables for Type, Post_Timing, Day_of_Week, and Month_Name

To avoid multicollinearity, the following categories were dropped as reference groups:
Post_Timing_Afternoon (for Post_Timing),Day_of_Week_Friday (for Day_of_Week), Month_Name_April (for Month_Name), Type_photo (for Type).

2. How well did the model perform?

R-squared: 0.438

The model explains 43.8% of the variance in likes.

Adjusted R-squared: 0.438
The adjusted value confirms that the predictors included are meaningful and not inflating the fit unnecessarily.

F-statistic: 590.0 ($p < 0.001$)
The predictors collectively have a statistically significant impact on likes.

3. What variables had a significant influence on the dependent variable?

Based on the unstandardized coefficients and p-values ($p < 0.05$), the following variables significantly influenced LIKES:

Numerical Variables:
FOLLOWERS (positive effect), FOLLOWING, number_of_mentions, Text_Length (negative effects).

Categorical Variables:
Post_Timing: Post_Timing_Morning (negative relative to Afternoon).
Day_of_Week: Day_of_Week_Saturday and Day_of_Week_Sunday (negative relative to Friday).
Type: Type_video (negative relative to TYPE_PHOTO).
Month_Name: Strong negative effects for months like Month_Name_August, Month_Name_July, Month_Name_November, etc., compared to April.

4. Interpret the coefficients of significant variables:

Below are the interpretations for statistically significant predictors from the unstandardized model:

Variable	**Coefficient**	**Interpretation**
FOLLOWERS	0.0482	For every additional follower, the number of likes increases by ~0.048, holding all else constant.
FOLLOWING	-0.1530	For every additional user followed, the number of likes decreases by ~0.153.
number_of_mentions	-46.6480	Each additional mention reduces likes by ~46.65, holding other factors constant.
Text_Length	-0.4546	Each additional unit increase in

		text length decreases likes by ~0.45.
Type_video	-876.8623	Video posts receive ~877 fewer likes compared to photo posts (reference category).
Post_Timing_Morning	-427.7979	Posts made in the morning receive ~428 fewer likes compared to posts made in the afternoon.
Day_of_Week_Saturday	-240.3670	Posts on Saturdays receive ~240 fewer likes compared to posts on Fridays.
Day_of_Week_Sunday	-216.3180	Posts on Sundays receive ~216 fewer likes compared to posts on Fridays.
Month_Name_August	-42450.0000	Posts in August receive ~42,450 fewer likes compared to posts in April.
Month_Name_July	-40380.0000	Posts in July receive ~40,380 fewer likes compared to posts in April.
Month_Name_September	-38270.0000	Posts in September receive ~38,270 fewer likes compared to posts in April.
Month_Name_November	-38860.0000	Posts in November receive ~38,860 fewer likes compared to posts in April.

5. Which independent variable has the greatest impact on predicting the dependent variable?

Numerical Variables: FOLLOWERS (0.8929): Standardized beta shows that an increase in followers has the most positive impact on likes, **FOLLOWING (-0.0619):** Negative but relatively small effect compared to followers.

Categorical Variables: Month_Name_September (-0.2544): This has the largest negative standardized beta, indicating a strong reduction in likes in September compared to April.

COMMENTS

1. Analysis Run: We ran Ordinary Least Squares (OLS) Regression models to analyze the influence of various independent variables on COMMENTS. Two sets of models were created:

One with unstandardized coefficients, Another with standardized coefficients.
The unstandardized coefficients provide the raw effect of each variable, while the standardized coefficients allow for comparison of the relative impact of each variable on COMMENTS.

2. Model Performance

R-squared: 0.070, indicating that only 7% of the variation in COMMENTS is explained by the independent variables, **F-statistic:** 57.08, with a p-value of 5.40e-286, suggesting that the model is statistically significant but has relatively low explanatory power.

3. Significant Influences on COMMENTS

FOLLOWERS: Positive influence ($p < 0.001$), indicating that as the number of followers increases, COMMENTS also increase, **Number_of_Users_in_Post:** Negative influence ($p < 0.001$), meaning that as the number of users in the post increases, COMMENTS decrease, **Type_video:** Negative influence ($p = 0.023$), meaning that posts with video content tend to have fewer COMMENTS compared to other types of posts, **Post_Timing_Evening:** Negative influence ($p < 0.001$), indicating that posts made in the evening receive fewer COMMENTS, **Post_Timing_Morning:** Negative influence ($p = 0.044$), indicating that posts made in the morning receive fewer COMMENTS, **Post_Timing_Night:** Negative influence ($p < 0.001$), suggesting that posts made at night also receive fewer COMMENTS, **Day_of_Week_Saturday:** Negative influence ($p = 0.005$), posts made on Saturday receive fewer COMMENTS, **Day_of_Week_Sunday:** Negative influence ($p = 0.023$), posts made on Sunday receive fewer COMMENTS, **Month_Name_August:** Negative influence ($p < 0.001$), posts made in August receive fewer COMMENTS, **Month_Name_December:** Negative influence ($p = 0.023$), posts made in December receive fewer COMMENTS, **Month_Name_February:** Negative influence ($p < 0.001$), posts made in February receive fewer COMMENTS, **Month_Name_January:** Negative influence ($p < 0.001$), posts made in January receive fewer COMMENTS, **Month_Name_July:** Negative influence ($p < 0.001$), posts made in July receive fewer COMMENTS, **Month_Name_May:** Negative influence ($p = 0.010$), posts made in May receive fewer COMMENTS, **Month_Name_November:** Negative influence ($p < 0.001$), posts made in November receive fewer COMMENTS, **Month_Name_October:** Negative influence ($p < 0.001$), posts made in October receive fewer COMMENTS, **Month_Name_September:** Negative influence ($p < 0.001$), posts made in September receive fewer COMMENTS.

4. Interpretation of Coefficients (Unstandardized)

FOLLOWERS (0.0016): For each additional follower, the number of COMMENTS increases by 0.0016, **Number_of_Users_in_Post (-9.5987):** For each additional user in the post, the number of COMMENTS decreases by 9.6, **Type_video (-24.0824):** Posts with video content tend to have 24.08 fewer COMMENTS than other types of posts, **Post_Timing_Evening (-30.8510):** Posts made in the evening receive 30.85 fewer COMMENTS than posts made at other times, **Post_Timing_Morning (-21.8486):** Posts made in the morning receive 21.85 fewer COMMENTS than posts made at other times, **Post_Timing_Night (-35.7369):** Posts made at night receive 35.74 fewer COMMENTS than posts made at other times, **Day_of_Week_Saturday (-31.8558):** Posts made on Saturday receive 31.86 fewer COMMENTS than posts made on other days, **Day_of_Week_Sunday (-24.9646):** Posts made on Sunday receive 24.96 fewer COMMENTS than posts made on other days, **Month_Name_August (-1712.3496):** Posts made in August receive 1712.35 fewer COMMENTS compared to other months, **Month_Name_December (-568.8505):** Posts made in December receive 568.85 fewer COMMENTS than posts made in other months, **Month_Name_February (-770.8800):** Posts made in February receive 770.88 fewer COMMENTS than posts made in other months, **Month_Name_January (-1337.1623):** Posts made in January receive 1337.16 fewer COMMENTS compared to other months, **Month_Name_July (-1719.8850):** Posts made in July receive 1719.89 fewer COMMENTS than posts made in other months, **Month_Name_May (-36.9816):** Posts made in May receive 36.98 fewer COMMENTS compared to other months, **Month_Name_November (-1682.7615):** Posts made in November receive 1682.76 fewer COMMENTS compared to other months, **Month_Name_October (-1679.7924):** Posts made in October receive 1679.79 fewer COMMENTS compared to other months, **Month_Name_September (-1693.4865):** Posts made in September receive 1693.49 fewer COMMENTS compared to other months.

5. Greatest Impact on Predicting COMMENTS

The independent variable with the greatest impact on predicting COMMENTS is FOLLOWERS, with an unstandardized coefficient of 0.0016, indicating that each additional follower has a positive, but relatively small, effect on the number of COMMENTS.

6. Standardized Beta Coefficients

FOLLOWERS: Most influential variable; a 1 SD increase leads to a 0.37 SD increase in COMMENTS,
Notable Variables: September (-0.1399), August (-0.1123), and October (-0.1035) negatively impact COMMENTS.
Conclusion: FOLLOWERS has the strongest effect on COMMENTS.

Question 2

New Variables created

question_mark_count: Counts question marks in the TEXT. Posts with questions encourage interaction, leading to more likes and comments.

Exclamation_mark_count: Counts exclamation marks in the TEXT. Posts with exclamation encourage interaction, leading to more likes and comments.

engagement_keywords_count: Counts engagement-focused words in TEXT. Prompts like "like" or "comment" directly drive user interaction.

emoji_count: Counts emojis in TEXT. Emojis make posts more engaging and visually appealing, encouraging reactions.

LIKES

1. What analysis did you run for this question?

A linear regression analysis was performed to predict the number of LIKES using various independent variables, such as user and post attributes (e.g., FOLLOWERS, FOLLOWING, post timing, type, etc.) and new variables like question_mark_count, engagement_keywords_count, and others. Both unstandardized and standardized coefficients were evaluated to assess the strength and direction of influence.

2. How well did the model perform?

R-squared: 0.439, indicating that the model explains 43.9% of the variance in the number of likes.
Adjusted R-squared: 0.438, confirming the model's robustness with minimal overfitting.
F-statistic: Highly significant ($p < 0.001$), showing the overall model is statistically significant.

3. What variables had a significant influence on the dependent variable (LIKES)?

Variables with $p < 0.05$ were deemed significant:

Positive Influence:

FOLLOWERS: Strong positive impact, question_mark_count: Posts with questions increased likes, engagement_keywords_count: Words like "like" or "comment" positively impacted likes.

Negative Influence:

FOLLOWING: Minor negative impact, number_of_mentions, Text_Length, and Type_video: Reduced likes, Post_Timing_Morning, Day_of_Week_Saturday/Sunday, and specific Month_Names (e.g., August, September).

4. Interpretation of Significant Coefficients

For unstandardized coefficients:

FOLLOWERS: +0.0482 likes per additional follower, FOLLOWING: -0.1533 likes per additional followed account, Type_video: Videos get 855.9 fewer likes than photos, Month_Name_August: Posts in August get 42,360 fewer likes than in April, question_mark_count: +31.66 likes per question mark, engagement_keywords_count: +119.43 likes per engagement keyword.

For standardized coefficients:

FOLLOWERS: Largest positive influence (coef = 0.8919), indicating a substantial impact relative to other variables, question_mark_count: Small but positive effect (coef = 0.019). Type_video: Strong negative influence (coef = -0.0453).

Key Insights

Audience size (FOLLOWERS) is the strongest predictor of likes.
Posts with engagement-driven language or questions increase likes.
Video posts and long Text_Length negatively impact likes.
Posting in specific months (e.g., August, September) significantly reduces likes, possibly due to seasonal trends.
Avoid posting during weekends and mornings to maximize engagement.
To determine if the new variables improved the performance of the linear regression model, we need to compare key performance metrics between the original and updated models:

5. Key Comparison Metrics:

R-squared and Adjusted R-squared:

Original Model:

R-squared = 0.438, Adjusted R-squared = 0.438

New Model:
R-squared = 0.439, Adjusted R-squared = 0.438

The addition of new variables led to a very slight increase in R-squared (0.001) but did not improve Adjusted R-squared, indicating minimal improvement in model fit.

Significance of New Variables:

Among the newly added variables: question_mark_count and engagement_keywords_count are statistically significant ($p < 0.05$).
emoji_count and exclamation_mark_count are not statistically significant ($p > 0.05$).

Model Complexity:

Original Model: 26 predictors, New Model: 30 predictors.

Adding new predictors increases model complexity without a notable improvement in performance metrics.

Conclusion:

The addition of new variables had minimal impact on the overall performance of the model. While two of the newly added variables (question_mark_count and engagement_keywords_count) are statistically significant and contribute positively to predicting LIKES, the improvements in R-squared are negligible. This suggests that while the new variables might offer some explanatory power, their contribution is not substantial enough to significantly improve the model's performance.

COMMENTS

1. Analysis Performed:

For this question, an Ordinary Least Squares (OLS) regression was run with COMMENTS as the dependent variable. The independent variables included a mix of numeric features (e.g., FOLLOWERS, number_of_tags), categorical variables (e.g., Post_Timing, Day_of_Week), and newly added variables (question_mark_count, exclamation_mark_count, etc.). The goal was to assess the impact of these variables on the number of comments.

2. Model Performance:

The model explains only 8.1% of the variance in COMMENTS ($R^2 = 0.081$), indicating limited explanatory power. Adjusted R^2 (0.080) suggests minimal overfitting. Despite this, the model is statistically significant ($F = 57.88$, $p < 0.0001$), showing at least one predictor has a meaningful impact.

3. Variables with Significant Influence:

The following variables had p-values < 0.05, indicating statistically significant effects on the number of comments:

FOLLOWERS

Number_of_Users_in_Post, number_of_tags, Post_Timing_Evening, Post_Timing_Morning, Post_Timing_Night, Day_of_Week_Saturday, Day_of_Week_Sunday, Month_Name_August, Month_Name_December, Month_Name_February, Month_Name_January, Month_Name_July, Month_Name_November, Month_Name_October, Month_Name_September, question_mark_count, engagement_keywords_count

4. Interpretation of Significant Coefficients:

FOLLOWERS (0.0016): For every additional follower, the number of comments increases by 0.0016. While small, this is consistent with the idea that larger audiences generate more engagement, **Number_of_Users_in_Post (-9.8808)**:Each additional tagged user decreases the number of comments by approximately 9.88. This may reflect lower engagement when posts

appear less personal, **number_of_tags (-2.5712):** Adding one more hashtag reduces comments by about 2.57. Overuse of hashtags could appear spammy or reduce audience interest.

Post_Timing: Evening: Comments decrease by 31.39 compared to the reference group (likely "Afternoon"), **Morning:** Comments decrease by 23.13, **Night:** Comments decrease by 31.59. Posts in the afternoon appear to generate the most engagement.

Day_of_Week: Saturday: Comments decrease by 30.65 compared to the reference day, Sunday: Comments decrease by 23.17, Weekends see lower engagement compared to weekdays, **Months:** Months like August (-1703.85), September (-1688.91), and November (-1677.67) see large decreases in comments compared to the reference month, Seasonal factors heavily influence engagement, **question_mark_count (1.9578):** Each additional question mark increases comments by 1.96. Posts that ask questions encourage more audience interaction, **engagement_keywords_count (53.8903):** For every additional engagement keyword, comments increase by 53.89. This suggests that using targeted, engaging language has a strong positive effect on comment volume.

5. Key Comparison Metrics

R-squared and Adjusted R-squared:
Original Model: R-squared = 0.070, Adjusted R-squared = 0.069
New Model: R-squared = 0.081, Adjusted R-squared = 0.080
The addition of new variables led to a slight increase in R-squared (0.011) and Adjusted R-squared (0.011), indicating a modest improvement in model fit.

Significance of New Variables:

Among the newly added variables:
Statistically Significant ($p < 0.05$): question_mark_count, engagement_keywords_count
Not Statistically Significant ($p > 0.05$): emoji_count, exclamation_mark_count
Model Complexity: Original Model: 26 predictors, New Model: 30 predictors.
Adding new predictors increases model complexity but offers only a marginal improvement in performance metrics.

Conclusion:

The addition of new variables had minimal impact on the overall performance of the model. While question_mark_count and engagement_keywords_count are statistically significant and contribute positively to predicting COMMENTS, the improvements in R-squared and Adjusted R-squared are minor. This suggests that the new variables provide some explanatory power but are insufficient to meaningfully enhance the model's performance.

Question 3

1. Analysis Conducted

For LIKES: A multiple linear regression analysis was performed to evaluate the influence of 21 independent variables, including FOLLOWERS, FOLLOWING, Number_of_Users_in_Post, number_of_tags, number_of_mentions, Text_Length, Type_video, and time/post-related categorical variables, on the dependent variable LIKES.

For COMMENTS: A separate multiple linear regression analysis was run with the same independent variables to assess their effect on the dependent variable COMMENTS.

2. Model Performance

LIKES: R-squared: 0.438, Adjusted R-squared: 0.438, The model explains 43.8% of the variance in LIKES, suggesting a moderately strong fit. The F-statistic (730.4, $p < 0.001$) indicates the model is statistically significant overall.

COMMENTS: R-squared: 0.070, Adjusted R-squared: 0.069, The model explains only 7.0% of the variance in COMMENTS, indicating a weak fit. Despite this, the F-statistic (70.24, $p < 0.001$) shows the model is statistically significant.

3. Variables with Significant Influence

LIKES: The following variables significantly influenced LIKES ($p < 0.05$):

Positive Influence: FOLLOWERS (positive coefficient of 0.0482): Larger follower count leads to higher likes.

Negative Influence: FOLLOWING: Users with higher followings receive fewer likes, number_of_mentions and **Text_Length:** More mentions and longer text reduce likes, **Type_video:** Video posts receive significantly fewer likes than non-video posts, **Post_Timing_Morning and Is_Weekend:** Morning posts and weekend posts reduce likes, Several Month_Name variables (e.g., August, January, September): Posts in these months receive fewer likes.

COMMENTS:

The following variables significantly influenced COMMENTS ($p < 0.05$):

Positive Influence: FOLLOWERS (positive coefficient of 0.0016): Higher follower count results in more comments,

Negative Influence: Number_of_Users_in_Post: Fewer comments as the number of users tagged in the post increases, Type_video: Video posts receive fewer comments, Post_Timing_Evening, Post_Timing_Night, and Is_Weekend: These times/days are associated with fewer comments, Several Month_Name variables (e.g., August, January, September): Posts in these months receive fewer comments.

Summary

LIKES: The model has a decent explanatory power (R-squared = 0.438). FOLLOWERS had the strongest positive effect, while factors like Type_video, morning posts, and posts during certain months significantly decreased likes.

COMMENTS: The model's explanatory power is weak (R-squared = 0.070). Only FOLLOWERS had a positive influence, while variables like Number_of_Users_in_Post, Type_video, and specific timings/months negatively impacted comments.

Overall, the analysis shows that both LIKES and COMMENTS are influenced by follower count and various content/post-related factors, though the model for COMMENTS is less effective at explaining variability.

4. Did the variables you added to the model improve the performance of the linear regression model compared to the model you ran in Question 1?

Replacing the Day of Posting variables with the Is_Weekend variable had no impact on the model's performance.
The R squared value remained 0.438 for Likes in both scenarios.
The R squared value remained 0.070 for Comments in both scenarios.
Similarly, the adjusted R squared values stayed constant for both models.

Question 4

Comparison Table for LIKES (Micro and Macro Influencers)

Independent Variable	Micro Influencers: Coefficient (LIKES)	Micro Influencers: P-Value (LIKES)	Macro Influencers: Coefficient (LIKES)	Macro Influencers: P-Value (LIKES)
FOLLOWERS	0.0249	0.000	0.0543	0.000
FOLLOWING	-0.1122	0.000	-0.3833	0.000
Number_of_Users_in_Post	32.9064	0.000	50.6888	0.169
number_of_tags	14.4805	0.000	-81.0876	0.000
number_of_mentions	-45.3111	0.000	-73.2309	0.211
Text_Length	-0.6339	0.000	-1.6286	0.028
Type_video	-512.8156	0.000	-1445.8289	0.000

Post_Timing_Evening	3.7207	0.916	-341.9780	0.143
Post_Timing_Morning	-150.7177	0.001	-937.7213	0.002
Post_Timing_Night	-40.2358	0.222	-282.1315	0.189
Day_of_Week_Monday	41.4211	0.404	133.0422	0.690
Day_of_Week_Saturday	-34.7871	0.445	-586.6675	0.054
Day_of_Week_Sunday	-56.7248	0.206	-392.9263	0.186
Day_of_Week_Thursday	-40.0004	0.373	-22.8012	0.939
Day_of_Week_Wednesday	-10.6292	0.830	-29.6051	0.929
Day_of_Week_Tuesday	-18.6578	0.702	-297.0529	0.359
Month_Name_August	-2.191e-09	0.540	-4.781e+04	0.000
Month_Name_December	-1251.2716	0.217	-5.28e+04	0.000
Month_Name_February	-302.3779	0.504	-3.676e+04	0.000
Month_Name_January	-921.5362	0.363	-4.227e+04	0.000
Month_Name_July	-3.874e-10	0.542	-4.624e+04	0.000
Month_Name_March	189.5511	0.767	-1.992e-10	0.139
Month_Name_May	-294.7210	0.000	-1191.6251	0.003
Month_Name_No	-7.224e-11	0.542	-4.539e+04	0.000

vember				
Month_Name_October	0	nan	-4.333e+04	0.000
Month_Name_September	0	nan	-4.449e+04	0.000
question_mark_count	3.8795	0.286	78.8535	0.028
engagement_keywords_count	-11.3501	0.415	617.4792	0.000
emoji_count	-105.4646	0.002	-215.3063	0.421
exclamation_mark_count	125.0505	0.000	-122.5699	0.013

Comparison Table for COMMENTS (Micro and Macro Influencers)

Independent Variable	Micro Influencers: Coefficient (LIKES)	Micro Influencers: P-Value (LIKES)	Macro Influencers: Coefficient (LIKES)	Macro Influencers: P-Value (LIKES)
FOLLOWERS	0.0002	0.061	0.0020	0.000
FOLLOWING	-0.0013	0.000	-0.0196	0.001
Number_of_Users_in_Post	-0.0596	0.908	-7.8870	0.045
number_of_tags	-0.4118	0.000	-8.5831	0.000
number_of_mentions	-0.2935	0.505	-6.2782	0.315
Text_Length	0.0227	0.000	-0.1954	0.014
Type_video	-1.7008	0.503	-43.5138	0.153
Post_Timing_Evening	-0.1125	0.957	-84.2781	0.001
Post_Timing_Mor	-2.3826	0.352	-48.9538	0.122

ning				
Post_Timing_Night	-0.6587	0.732	-88.7268	0.000
Day_of_Week_Monday	7.5420	0.009	25.6234	0.470
Day_of_Week_Saturday	-1.9017	0.474	-86.9747	0.007
Day_of_Week_Sunday	-2.7484	0.294	-47.8028	0.131
Day_of_Week_Thursday	-2.7268	0.299	-39.0155	0.216
Day_of_Week_Wedneday	-1.4176	0.623	-43.3414	0.218
Day_of_Week_Tuesday	-0.6545	0.818	-36.2737	0.294
Month_Name_August	-2.197e-11	0.916	-2088.6590	0.000
Month_Name_December	-20.1699	0.733	-2024.1136	0.005
Month_Name_February	-4.4284	0.867	-1695.5706	0.000
Month_Name_January	-22.4290	0.705	-1865.5110	0.000
Month_Name_July	-3.88e-12	0.917	-2219.0823	0.000
Month_Name_March	-14.2691	0.703	-1.22e-11	0.395
Month_Name_May	-11.6615	0.001	-107.8221	0.012
Month_Name_November	-7.231e-13	0.917	-2114.4220	0.000
Month_Name_October	0	nan	-2074.9639	0.000

Month_Name_September	0	nan	-2113.2977	0.000
question_mark_count	0.1218	0.567	8.2256	0.032
engagement_keywords_count	6.9581	0.000	195.4264	0.000
emoji_count	-4.5091	0.026	13.4588	0.637
exclamation_mark_count	2.2737	0.000	-8.7956	0.095

Comparison of P-Values for Likes

Independent variables significantly influencing likes for micro but not for macro-influencers:

Number_of_Users_in_Post
Micro Influencers: P-Value = 0.000 (Significant)
Macro Influencers: P-Value = 0.169 (Not Significant)

number_of_mentions
Micro Influencers: P-Value = 0.000 (Significant)
Macro Influencers: P-Value = 0.211 (Not Significant)

Post_Timing_Evening
Micro Influencers: P-Value = 0.916 (Not Significant)
Macro Influencers: P-Value = 0.143 (Not Significant but closer to significance).

Independent variables significantly influencing likes for macro but not for micro-influencers:

Day_of_Week_Saturday
Micro Influencers: P-Value = 0.445 (Not Significant)
Macro Influencers: P-Value = 0.054 (Marginally Significant)

question_mark_count
Micro Influencers: P-Value = 0.286 (Not Significant)
Macro Influencers: P-Value = 0.028 (Significant)

Text_Length
Micro Influencers: P-Value = 0.000 (Significant)
Macro Influencers: P-Value = 0.028 (Significant but weaker).

Comparison of Coefficients for Likes

Independent variables with different coefficients:

number_of_tags
Micro Influencers: Coefficient = 14.4805 (Positive Influence)
Macro Influencers: Coefficient = -81.0876 (Negative Influence)
This indicates that the use of tags increases likes for micro-influencers but decreases likes for macro-influencers.

emoji_count
Micro Influencers: Coefficient = -105.4646 (Negative Influence)
Macro Influencers: Coefficient = -215.3063 (Stronger Negative Influence).
Emojis reduce likes for both groups, but the effect is stronger for macro-influencers.

Post_Timing_Morning
Micro Influencers: Coefficient = -150.7177
Macro Influencers: Coefficient = -937.7213
Morning posts reduce likes significantly more for macro-influencers.

Type_video
Micro Influencers: Coefficient = -512.8156
Macro Influencers: Coefficient = -1445.8289
Videos lead to fewer likes for both groups, but the impact is much larger for macro-influencers.

Comparison of P-Values for Comments

Independent variables significantly influencing comments for micro but not for macro-influencers:

Number_of_Users_in_Post
Micro Influencers: P-Value = 0.908 (Not Significant)
Macro Influencers: P-Value = 0.045 (Significant)

Day_of_Week_Monday
Micro Influencers: P-Value = 0.009 (Significant)
Macro Influencers: P-Value = 0.470 (Not Significant)

Independent variables significantly influencing comments for macro but not for micro-influencers:

Post_Timing_Night
Micro Influencers: P-Value = 0.732 (Not Significant)
Macro Influencers: P-Value = 0.000 (Highly Significant)

question_mark_count
Micro Influencers: P-Value = 0.567 (Not Significant)
Macro Influencers: P-Value = 0.032 (Significant).

Comparison of Coefficients for Comments

Independent variables with different coefficients:

number_of_tags

Micro Influencers: Coefficient = -0.4118 (Negative Influence)
Macro Influencers: Coefficient = -8.5831 (Stronger Negative Influence).

Text_Length

Micro Influencers: Coefficient = 0.0227 (Positive Influence)
Macro Influencers: Coefficient = -0.1954 (Negative Influence).
Text length has opposite effects: it increases comments for micro-influencers but decreases them for macro-influencers.

Post_Timing_Evening
Micro Influencers: Coefficient = -0.1125
Macro Influencers: Coefficient = -84.2781
Evening posts have almost no effect on micro-influencers but strongly reduce comments for macro-influencers.

engagement_keywords_count
Micro Influencers: Coefficient = 6.9581 (Positive Influence)
Macro Influencers: Coefficient = 195.4264 (Stronger Positive Influence).
Engagement keywords help both groups, but the effect is significantly larger for macro-influencers.

exclamation_mark_count
Micro Influencers: Coefficient = 2.2737 (Positive Influence)
Macro Influencers: Coefficient = -8.7956 (Negative Influence).
Exclamation marks improve comments for micro-influencers but reduce them for macro-influencers.

Summary of Observations

Significant Independent Variables:

Number_of_Users_in_Post:
Refers to the number of individuals tagged or featured in a post. This variable significantly influences the number of likes for micro-influencers but not for macro-influencers.

Micro-influencers' audiences may find collaborative or group posts more engaging and authentic, while macro-influencers' larger and more diverse audiences may not prioritize this aspect.

question_mark_count:
Indicates the number of question marks used in a post's caption. This variable significantly affects the number of likes and comments for macro-influencers but not for micro-influencers. Macro-influencer audiences may respond more to interactive or engaging content styles that encourage participation or curiosity.

Opposing Effects:

number_of_tags:

Refers to the count of hashtags used in a post. This positively affects likes for micro-influencers, as hashtags may enhance discoverability or signal collaboration, which resonates with niche audiences. However, it negatively impacts likes for macro-influencers, where excessive hashtags might come across as spammy or less authentic.

exclamation_mark_count:
Represents the number of exclamation marks in a caption. For micro-influencers, this positively impacts comments, reflecting an enthusiastic and engaging tone that connects with their audience. For macro-influencers, it negatively impacts comments, where overuse of exclamation marks might appear exaggerated or less professional to their broader audience.

Strength of Influence:

engagement_keywords_count: Refers to the use of words or phrases encouraging interaction (e.g., "like," "share," "comment"). This variable has a stronger positive influence on likes and comments for macro-influencers, suggesting that their larger audience base is more responsive to direct engagement calls.

emoji_count: Indicates the number of emojis used in a caption. Macro-influencers experience stronger positive impacts from this variable, as emojis may make content visually appealing to a diverse audience. On the other hand, micro-influencers see a less pronounced effect, possibly due to their audiences' focus on textual authenticity or relatability.

Conclusion and Recommendations

1. Summary of the Work Conducted

This project aimed to identify factors that drive Instagram engagement, such as likes and comments, for influencers by analyzing a dataset of 19,681 Instagram posts. The analysis involved regression models to evaluate the impact of variables such as follower count, hashtags, post timing, and content type. Processes such as data cleaning, data transformation, and feature engineering were also conducted to ensure reliable insights. Two types of influencers—micro and macro—were compared to understand audience behaviors across groups.

2. Key Findings (Data Analyst Terms) (Overall)

Likes: Significant positive predictors include FOLLOWERS, question_mark_count, and engagement_keywords_count, Type_video, Text_Length, and specific months (e.g., August, September) negatively influence likes, FOLLOWERS had the strongest impact, with each additional follower increasing likes significantly.

Comments: Positive predictors include FOLLOWERS, question_mark_count, and engagement_keywords_count, Negative predictors include number_of_tags, Type_video, and specific post timings like Post_Timing_Night, While FOLLOWERS was a significant positive predictor, number_of_tags and number_of_mentions had notable negative effects on comments.

3. Key Insights (General Audience Terms)

Interactive Captions: Questions and engagement keywords boost likes and comments, especially for macro-influencers.
Content Type: Photos outperform videos in engagement.
Timing: Afternoon posts perform best; morning and night posts lag.
Seasonal Trends: Engagement drops in August and September.
Macro Strategy: Macro-influencers gain more from direct calls to action.

4. Recommendations

Recommendations for Micro Influencers

For Likes:

Tag users and add relevant hashtags to attract attention.
Write concise captions; avoid lengthy text.
Limit mentions to keep posts focused.
Use exclamation marks to boost excitement.
Prefer photos over videos.
Post later in the day for better performance.
Post on weekdays, especially Mondays.
Avoid weekends, particularly Saturdays and Sundays.

For Comments:

Use longer captions to spark discussion.
Add engagement keywords and call-to-actions.
Avoid overusing tags and emojis to maintain authenticity.
Focus on content strategy in months like May to counter lower engagement.
Post in the afternoon and on Mondays, Avoid weekends.

Recommendations for Macro Influencers

For Likes:

Use keywords in captions to engage audiences.
Focus on photo content instead of videos.
Avoid excessive tagging or mentioning users.
Post in the morning or evening to align with audience activity.
Plan campaigns for high-engagement months.
Focus on weekdays; avoid weekends, especially Saturdays.

For Comments:

Add questions or action prompts in captions.
Minimize tags and mentions to boost authenticity.
Post during the day or evening for better engagement.
Focus on high-engagement months and avoid low-performing periods.
Focus on weekdays; they outperform weekends.