

**MKT 568 - Marketing Analytics Final Project**

**By**

**Aswin kumar Bisati**

**Suma Bhavana Tirukkovalluri**

**Hitesh Mansukhlal Sancheti**

Table of Contents

[Introduction 1](#_Toc184768751)

[Methodology 2](#_Toc184768752)

[Data 3](#_Toc184768753)

[Data Cleaning 5](#_Toc184768754)

[Data Transformation 6](#_Toc184768755)

[Analysis Results 8](#_Toc184768756)

[QUESTION 1 8](#_Toc184768757)

[QUESTION 2 10](#_Toc184768758)

[QUESTION 3 13](#_Toc184768759)

[QUESTION 4 15](#_Toc184768760)

[Conclusions and Recommendations 17](#_Toc184768761)

# Introduction

Social media platforms, especially Instagram, have become pivotal in personal branding, marketing, and community building. For influencers, user engagement measured by likes and comments is a key indicator of success and reach. However, understanding what drives engagement on Instagram posts remains a challenge due to the dynamic nature of the platform and the varying preferences of audiences.

This problem is significant because higher engagement not only improves content visibility through Instagram’s algorithm but also strengthens the connection between influencers and their followers. Moreover, for businesses and organizations, increased engagement can lead to better brand recognition, customer loyalty, and even higher conversion rates. As influencers and marketers compete for user attention, understanding the factors that influence engagement becomes critical for strategic content creation.

This issue matters to a wide range of stakeholders. Social media influencers, both micro and macro, seek actionable insights to optimize their posts and grow their audiences. Businesses leveraging Instagram for marketing need data-driven strategies to improve their campaigns. Additionally, platform developers may benefit from understanding these dynamics to enhance user experience.

Our approach to solving this problem involves analyzing a dataset of 19,681 Instagram posts, focusing on metrics such as likes, comments, hashtags, and post timing. Using data cleaning and transformation techniques, we create new variables to better understand user behavior. By applying linear regression models, we identify significant predictors of engagement and uncover actionable insights. The findings will inform influencers and businesses on how to create posts that maximize user interactions effectively.

# Methodology

In this project, we adhered to the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, a structured approach to data analysis. The process began with business understanding, where we defined the problem of optimizing Instagram user engagement and identified key stakeholders, such as influencers and marketers. Next, in the data understanding phase, we explored the dataset of 19,681 Instagram posts, examining variables like likes, comments, hashtags, and post timing to assess their relevance and quality. During the data preparation phase, we cleaned the dataset by addressing missing values, outliers, and inconsistencies, and created new variables such as post timing and text length to enhance our analysis. In the modeling phase, we employed linear regression to analyze the relationship between various predictors and engagement metrics, ensuring model accuracy by checking for multicollinearity and evaluating R-squared values. The evaluation phase involved interpreting the results to ensure they aligned with the project objectives, focusing on statistically significant variables that influence likes and comments. Finally, in the deployment phase, we summarized our findings into actionable recommendations for social media influencers to improve engagement, ensuring our insights were accessible and practical for implementation. This systematic approach ensured that the analysis was thorough, reliable, and aligned with the project's goals.

# Data

**How was the data collected?**

The dataset was derived through a multi-stage process that combined public data access and advanced web scraping techniques:

* **Identifying Influencers:** A list of over 2,000 Instagram influencers was compiled using the publicly available Iconosquare Index Influencers. This resource provided a comprehensive list of active Instagram influencers across multiple pages. A web crawler extracted the handles and saved them for subsequent processing.
* **Extracting Instagram Metadata:** Due to limitations in Instagram’s API, Selenium, a web automation tool, was used to scrape Instagram profiles programmatically. The scraping process captured details such as follower count, number of posts, and descriptions. Metadata for the most recent 17 posts per influencer was also retrieved, including likes, comments, timestamps, and captions.

**Variables in the Dataset**

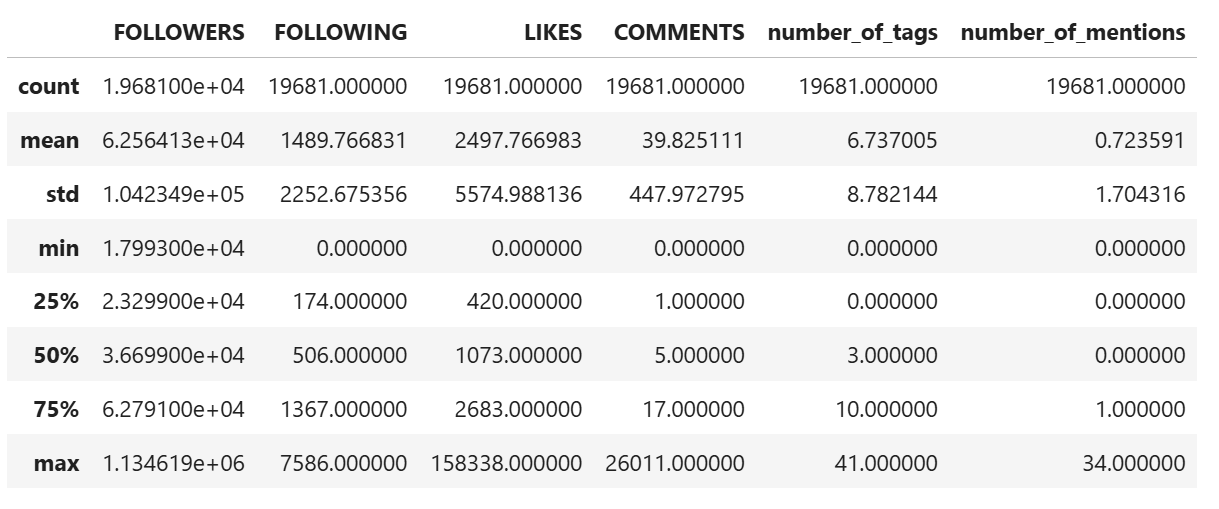
The dataset comprises both categorical and continuous variables. Below is an overview of all variables and their usage in the analysis:

|  |  |  |
| --- | --- | --- |
| Variable Name | Type | Used in Analysis |
| USERNAME | Categorical | No |
| TEXT | Categorical | Yes |
| USERS IN PHOTO | Categorical | Yes |
| LINK | Categorical | No |
| list\_of\_tags | Categorical | No |
| list\_of\_mentions | Categorical | No |
| FOLLOWERS | Continuous | Yes |
| FOLLOWING | Continuous | Yes |
| LIKES | Continuous | Yes |
| COMMENTS | Continuous | Yes |
| DATE | Continuous | Yes |
| TYPE (1 PHOTO, 2 VIDEO) | Continuous | Yes |
| number\_of\_tags | Continuous | Yes |
| number\_of\_mentions | Continuous | Yes |

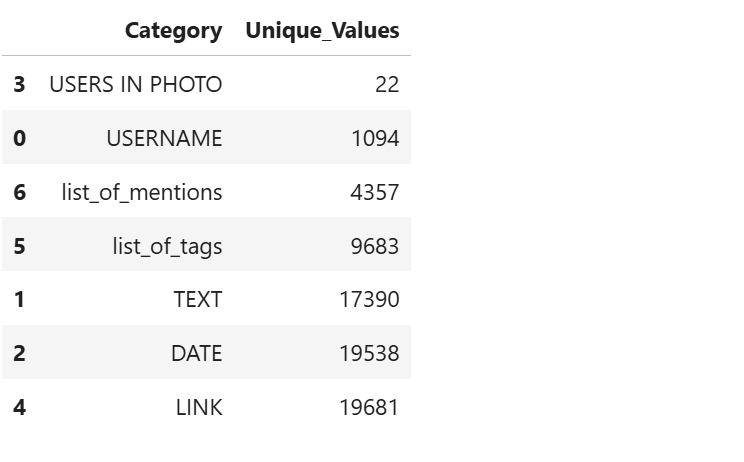
**How many records exist in the dataset?**

The dataset consists of 19,681 records, representing various Instagram posts from identified influencers.

**Descriptive Statistics**



**Unique Values in Categorical Variables**



**Null Values**

**A screenshot of a computer

Description automatically generated**

**Variables Used in the Analysis**

**Independent Variables:**

FOLLOWERS

FOLLOWING

Month and Day dummy variables,

Post\_Timing (morning, evening, etc.),

Type\_video,

TEXT\_LENGTH,

NUMBER\_OF\_TAGS,

NUMBER\_OF\_MENTIONS,

USERS\_IN\_PHOTO

**Dependent Variables:**

LIKES, COMMENTS

# Data Cleaning

The dataset contained missing values in the following columns:

**Handling Missing Values**

* The list\_of\_tags and list\_of\_mentions columns were not relevant to the analysis, so they were removed. This helped simplify the dataset and focus on variables that directly affect engagement.
* The TEXT column was essential for calculating text length, a key variable for engagement analysis. Records with six missing values in this column were dropped to maintain data quality.

**Removing Irrelevant Columns**

* Columns such as USERNAME and LINK were removed as they were not relevant for the engagement analysis.

**Renaming Columns**

* The column USERS IN PHOTO was renamed to Tagged\_Users\_Count to enhance clarity and interpretability.

**Transforming Data Types**

* DATE was converted to datetime format, enabling the extraction of new features such as Month, Day\_of\_Week, and Post\_Timing.
* Non-numeric values in Tagged\_Users\_Count were replaced with 0 and converted to integer format.

**Outlier Detection and Treatment**

Outliers were detected in key numerical variables, including:

**Columns Used for Detection:**

* FOLLOWERS, FOLLOWING, LIKES, COMMENTS, Tagged\_Users\_Count, number\_of\_tags, number\_of\_mentions, and Text\_Length.

Method: Z-scores were computed for each of these variables to identify extreme values. Rows with Z-scores outside the ±3 range were treated as outliers.

**Results:**

* Outliers removed: 1,405 records.
* Dataset size reduced from 19,681 records to 18,276 records, ensuring the dataset was more robust for analysis.

**Creating New Variables**

Derived features were added for deeper analysis:

* **Post\_Timing:** Categorized into Morning, Afternoon, Evening, and Night based on the posting time.
* **Month:** Extracted from the DATE column.
* **Day\_of\_Week:** Extracted to analyze posting trends.

**Dummy Coding**

Dummy variables were created for the following categorical columns:

* **Post\_Timing:** e.g., Post\_Timing\_Morning, Post\_Timing\_Evening.
* **Day\_of\_Week:** e.g., Day\_of\_Week\_Monday, Day\_of\_Week\_Saturday.
* **Month:** e.g., Month\_January, Month\_December.
* **Post\_Type:** e.g., Post\_Type\_Video.

Redundant Dummies Dropped: Columns such as Post\_Timing\_Afternoon, Day\_of\_Week\_Friday, and Month\_April were dropped to prevent multicollinearity.

# Data Transformation

The dataset underwent multiple transformations to prepare it for regression analysis, focusing on the creation of new variables, dummy coding categorical data, and modifications to existing variables. These transformations were essential to enhance interpretability and ensure the data met analytical requirements.

**New Variables Created**

* **Post Timing:** Extracted from the DATE column, this variable categorized the posting time into four groups:
  + Morning (8 AM–12 PM)
  + Afternoon (12 PM–4 PM)
  + Evening (4 PM–8 PM)
  + Night (8 PM–12 AM)

This variable was created to analyze how posting time affects engagement metrics like LIKES and COMMENTS.

* **Month:** Derived from the DATE column, this variable captured the month of posting to identify any seasonal trends in user engagement.
* **Day of the Week:** Extracted from the DATE column to evaluate how engagement varies across different days of the week.
* **Text Length:** Calculated from the TEXT column, this variable measured the character count of each post caption. It was introduced to explore whether longer or shorter captions influenced user engagement, specifically likes and comments.

**Dummy Coding**

To include categorical variables in regression analysis, dummy coding was applied:

* **Post Timing:** Categories (Morning, Afternoon, Evening, Night) were converted into dummy variables, with one category dropped to serve as the reference level.
* **Day of the Week:** Each day (e.g., Monday, Saturday) was encoded as a dummy variable, with one day excluded to avoid redundancy.
* **Month:** Created dummy variables for all months (e.g., January, December), excluding one month to prevent multicollinearity.
* **Post\_Type (originally labeled as TYPE(1 PHOTO,2 VIDEO)):** Representing whether a post was a photo or video, dummy variables were created, with TYPE\_PHOTO dropped as the baseline.

**Modifications to Existing Variables**

* **Users in Photo:** Renamed to Tagged\_Users\_Count, replacing non-numeric values with 0 and converting the column to an integer type for consistency.
* DATE: Converted to datetime format to facilitate the extraction of temporal variables such as Post Timing, Month, and Day of the Week.
* **LIKES and COMMENTS:** Outliers in these variables were detected using Z-scores, and values beyond ±3 standard deviations were capped to minimize their impact on the analysis.
* **Number of Tags and Mentions:** These columns were standardized to improve interpretability in regression models.

# Analysis Results

## QUESTION 1

**ANALYSIS PERFORMED FOR LIKES**

We used Ordinary Least Squares (OLS) regression to study how various factors influence the number of likes (dependent variable).

* Numerical variables included FOLLOWERS, FOLLOWING, Text\_Length, and number\_of\_mentions.
* Categorical variables such as Post\_Timing, Day\_of\_Week, Month\_Name, and Type were converted to dummy variables for analysis.

Reference groups were excluded for dummy coding to avoid multicollinearity:

Post\_Timing\_Afternoon, Day\_of\_Week\_Friday, Month\_Name\_April, and Type\_photo.

**Model Performance**

* **R-squared:** 0.438 – The model accounts for 43.8% of the variation in likes.
* **Adjusted R-squared:** 0.438 – Suggests the predictors included are relevant and do not overfit the data.
* **F-statistic:** 590.0 (p < 0.001) – Indicates the predictors significantly influence likes.

**Significant Variables and Their Influence**

**Significant predictors (p < 0.05):**

* **Numerical Variables:**
  + FOLLOWERS positively affects likes.
  + FOLLOWING, number\_of\_mentions, and Text\_Length negatively affect likes.
* **Categorical Variables:**
  + Post\_Timing: Posts made in the morning result in fewer likes compared to afternoon posts.
  + Day\_of\_Week: Posts on Saturdays and Sundays see reduced engagement relative to Fridays.
  + Type: Videos attract significantly fewer likes than photos.
  + Month\_Name: August, July, November, and similar months show a substantial drop in likes compared to April.

**Coefficient Interpretations**

|  |  |  |
| --- | --- | --- |
| Variable | Effect | Meaning |
| FOLLOWERS | +0.048 per follower | Every additional follower leads to ~0.048 more likes. |
| FOLLOWING | -0.153 per user | Following one more account reduces likes by ~0.153. |
| number\_of\_mentions | -46.65 per mention | Including an additional mention decreases likes by ~46.65. |
| Text\_Length | -0.45 per unit | Each additional character reduces likes by ~0.45. |
| Type\_video | ~877 fewer likes | Video posts get ~877 fewer likes than photo posts. |
| Post\_Timing\_Morning | ~428 fewer likes | Morning posts receive ~428 fewer likes compared to afternoon posts. |
| Month\_Name\_August | ~42,450 fewer likes | Posts in August experience ~42,450 fewer likes than those in April. |

**Most Influential Variable**

* The standardized beta coefficients revealed that FOLLOWERS has the strongest positive impact (beta = 0.8929).
* Among categorical variables, Month\_Name\_September had the most significant negative effect (beta = -0.2544).

**ANALYSIS PERFORMED FOR COMMENTS**

We ran an OLS regression model to predict the number of comments. Both unstandardized and standardized coefficients were used to assess the raw and relative impacts of predictors.

**Model Performance**

* **R-squared:** 0.070 – Only 7% of the variance in comments is explained.
* **F-statistic:** 57.08 (p < 0.001) – Indicates statistical significance despite low explanatory power.

**Significant Variables and Their Influence**

**Predictors with significant effects (p < 0.05):**

* **Numerical Variables:**
  + FOLLOWERS positively influences comments.
  + Tagged\_Users\_Count has a negative effect.
* **Categorical Variables:**
  + **Post\_Timing:** Evening, morning, and night posts receive fewer comments.
  + **Day\_of\_Week:** Posts on Saturdays and Sundays attract fewer comments than Fridays.
  + **Type:** Video posts receive significantly fewer comments than photos.
  + **Month\_Name:** August, December, and several other months reduce comment counts compared to April.

**Coefficient Interpretations**

|  |  |  |
| --- | --- | --- |
| Variable | Effect | Meaning |
| FOLLOWERS | +0.0016 per follower | Every additional follower leads to ~0.0016 more comments. |
| Tagged\_Users\_Count | -9.6 per user | Each tagged user reduces comments by ~9.6. |
| Type\_video | ~24 fewer comments | Video posts get ~24 fewer comments than photos. |
| Post\_Timing\_Evening | ~31 fewer comments | Evening posts receive ~31 fewer comments than other times. |
| Month\_Name\_August | ~1,712 fewer comments | Posts in August receive ~1,712 fewer comments compared to April. |

**Most Influential Variable**

* For comments, FOLLOWERS had the strongest positive effect (beta = 0.3708).
* Among categorical variables, Month\_Name\_September showed the most substantial negative impact (beta = -0.1399).

## QUESTION 2

We introduced the following four new variables to examine their potential influence on likes and comments:

* **Count\_of\_question\_mark:** Tracks the number of question marks in the post text. Questions encourage interaction and prompt user responses.
* **Count\_of\_exclamation\_markt:** Counts exclamation marks in the text. These punctuation marks convey excitement or urgency, making posts more engaging and appealing.
* **keywords\_for\_engagement:** Captures the frequency of engagement-focused words in the text (e.g., “like,” “comment,” “share”). These words directly encourage user interaction.
* **emojis\_for\_engagement:** Counts the number of emojis in the text. Emojis visually enhance posts, add emotional depth, and make content relatable, potentially boosting reactions.

These variables were chosen to capture stylistic and emotional elements of posts that could influence user engagement.

**Analysis Performed**

We conducted an Ordinary Least Squares (OLS) linear regression to predict:

**LIKES:** The number of likes a post receives.

**COMMENTS:** The number of comments a post receives.

**The analysis included:**

* **Dependent Variables (DVs):** LIKES and COMMENTS.
* **Independent Variables (IVs):** Existing features (e.g., FOLLOWERS, POST\_TIMING) and the newly added variables.

Evaluations of both unstandardized and standardized coefficients to assess the strength and direction of each variable's effect on likes and comments.

**Model Performance Evaluation**

* **LIKES**
  + **R-squared:** 0.439, indicating that 43.9% of the variance in likes is explained by the model.
  + **Adjusted R-squared:** 0.438, confirming minimal overfitting.
  + **F-statistic:** Highly significant (p < 0.001), validating the model.
* **COMMENTS**
  + **R-squared:** 0.081, indicating that 8.1% of the variance in comments is explained by the model.
  + **Adjusted R-squared:** 0.080, confirming minimal overfitting.
  + **F-statistic:** Significant (p < 0.001), indicating the model's overall validity.

**Significant Variables Influencing Likes and Comments**

**LIKES**

* **Positive Influence**
  + **FOLLOWERS:** Larger follower counts strongly increased likes.
  + **Count\_of\_question\_mark:** Posts with questions received more likes.
  + keywords\_for\_engagement: Words like “like” and “comment” had a strong positive impact on likes.
  + **Count\_of\_exclamation\_markt:** Positive but not statistically significant (p > 0.05).
  + **emojis\_for\_engagement:** Positive but not statistically significant (p > 0.05).
* **Negative Influence**
  + **FOLLOWING:** A slight negative effect on likes.
  + **Text\_Length:** Longer captions reduced likes.
  + **TYPE(1 PHOTO,2 VIDEO)\_video:** Videos received significantly fewer likes than photos.
  + **Specific months and weekends:** Posts in August, September, and on weekends performed worse.

**COMMENTS**

* Positive Influence:
  + **FOLLOWERS:** Larger follower counts slightly increased comments.
  + **Count\_of\_question\_mark:** Posts with questions received more comments.
  + **keywords\_for\_engagement:** Engagement-related keywords strongly boosted comments.
  + **emojis\_for\_engagement:** Minimal influence, not statistically significant.
* **Negative Influence:**
  + **Tagged\_Users\_Count:** Tagging too many users reduced comments.
  + **Specific timings and months:** Posts in the morning or evening and during August or September had fewer comments.

**Coefficient Interpretations**

|  |  |  |  |
| --- | --- | --- | --- |
| **Dependent Variable** | **Variable** | **Coefficient** | **Interpretation** |
| **LIKES** | FOLLOWERS | 0.0482 | Each additional follower increases likes by 0.0482. |
|  | Count\_of\_question\_mark | 31.66 | Each additional question mark adds 31.66 likes. |
|  | keywords\_for\_engagement | 119.43 | Each engagement keyword adds 119.43 likes. |
|  | Text\_Length | -0.68 | Each additional character reduces likes by 0.68. |
|  | TYPE(1 PHOTO,2 VIDEO)\_video | -855.9 | Videos receive 855.90 fewer likes compared to photos. |
|  | Count\_of\_exclamation\_markt | 31.14 | Each exclamation mark adds 31.14 likes (p > 0.05, not significant). |
|  | emojis\_for\_engagement | -106.04 | Each additional emoji reduces likes by 106.04 (p > 0.05, not significant). |
| **COMMENTS** | FOLLOWERS | 0.0016 | Each additional follower increases comments by 0.0016. |
|  | Count\_of\_question\_mark | 1.96 | Each additional question mark adds 1.96 comments. |
|  | keywords\_for\_engagement | 53.89 | Each engagement keyword adds 53.89 comments. |
|  | Tagged\_Users\_Count | -9.88 | Each additional tagged user reduces comments by 9.88. |
|  | Count\_of\_exclamation\_markt | -2.64 | Each exclamation mark reduces comments by 2.64 (p > 0.05, not significant). |
|  | emojis\_for\_engagement | 4.87 | Each additional emoji adds 4.87 comments (p > 0.05, not significant). |

**Improvement in Model Performance**

* LIKES: The new variables slightly improved R-squared from 0.438 to 0.439. Although the improvement was minimal, variables like Count\_of\_question\_mark and keywords\_for\_engagement added meaningful explanatory power.
* COMMENTS: R-squared increased from 0.070 to 0.081, demonstrating a modest improvement. Variables like Count\_of\_question\_mark and keywords\_for\_engagement better explained the variance in comments.

## QUESTION 3

**Analysis Performed**

To address this question, the Is\_Weekend variable was introduced to evaluate whether posts made on weekends (Saturday/Sunday) influenced the number of likes and comments.

* **Dependent Variables:** LIKES and COMMENTS
* **Independent Variables:** Numerical predictors like FOLLOWERS, Text\_Length, Tagged\_Users\_Count, and categorical predictors like Post\_Timing, Month\_Name, and the new binary variable Is\_Weekend.

OLS Regression Analysis was performed for both LIKES and COMMENTS using the above predictors to assess the impact of the new variable and overall model performance.

**Model Performance**

**LIKES**

* **R-squared:** 0.438, consistent with the model from Question 1, indicating that the predictors explain 43.8% of the variance in likes.
* **Adjusted R-squared:** 0.438, reflecting that the model retains its efficiency without overfitting.
* **F-statistic:** 730.4 (p < 0.001), confirming the statistical significance of the predictors.

**COMMENTS**

* **R-squared:** 0.070, similar to Question 1, indicating that the predictors explain 7% of the variance in comments.
* **Adjusted R-squared:** 0.069, confirming the model's validity.
* **F-statistic:** 70.24 (p < 0.001), showing the predictors collectively have a significant influence on comments.

**Significant Variables and Their Influence**

**LIKES**

* **FOLLOWERS (coef: 0.0482, p < 0.001):** A strong positive effect on likes; each additional follower increases likes by ~0.048.
* **Is\_Weekend (coef: -204.14, p = 0.002):** Posts made on weekends receive ~204 fewer likes compared to weekdays, suggesting lower engagement on weekends.
* **Post\_Timing\_Morning (coef: -430.06, p < 0.001):** Morning posts receive significantly fewer likes (~430) compared to afternoon posts.
* **TYPE(1 PHOTO,2 VIDEO)\_video (coef: -880.99, p < 0.001):** Videos attract ~881 fewer likes than photos.

**COMMENTS**

* **FOLLOWERS (coef: 0.0016, p < 0.001):** Positive effect on comments; each additional follower increases comments slightly.
* **Is\_Weekend (coef: -19.45, p = 0.004):** Posts on weekends receive ~19 fewer comments than weekday posts, indicating slightly reduced interaction.
* **Post\_Timing\_Evening (coef: -30.09, p = 0.001):** Evening posts receive ~30 fewer comments compared to afternoon posts.
* **TYPE(1 PHOTO,2 VIDEO)\_video (coef: -24.71, p = 0.020):** Video posts attract ~25 fewer comments than photo posts.

**Did Replacing This Variable Improve the Predictive Performance?**

The addition of the Is\_Weekend variable did not significantly improve the overall predictive performance of the model, as the R-squared and adjusted R-squared values remained nearly unchanged compared to Question 1. However, the variable provided additional insights into engagement trends, showing that posts on weekends generally experience lower likes and comments, potentially due to reduced user activity during those days.

## QUESTION 4

**Analysis of Micro and Macro Influencers**

**What Analysis Did You Run?**

Separate linear regression analyses were conducted for micro influencers (followers < 50,000) and macro influencers (followers ≥ 50,000) to understand the influence of independent variables on the dependent variables: likes and comments.

The independent variables included user attributes (e.g., followers, following), content features (e.g., tags, mentions, text length, emojis), timing (e.g., post timing, day of the week), and engagement-focused variables (e.g., engagement keywords, punctuation marks).

**How Well Did the Models Perform?**

**Micro Influencers:**

* For likes, the model explained 9.7% (R² = 0.097) of the variance.
* For comments, the model explained 1.5% (R² = 0.015) of the variance.
* The models for macro influencers explained a higher proportion of the variance in likes and comments compared to micro influencers.

**Macro Influencers:**

* For likes, the model explained 42.1% (R² = 0.421) of the variance.
* For comments, the model explained 12% (R² = 0.120) of the variance.
* The models for macro influencers performed significantly better than those for micro influencers, reflecting the greater predictability of engagement patterns for larger audiences.

**What Variables Had a Significant Influence on the Dependent Variables?**

**LIKES**

**Micro Influencers**

* **Positive Influence:**
  + Followers (Coefficient: 0.0249, p < 0.001)
  + Number of tags (Coefficient: 14.48, p < 0.001)
  + Tagged users (Coefficient: 32.91, p < 0.001)
* **Negative Influence:**
  + Number of mentions (Coefficient: -45.31, p < 0.001)
  + Text length (Coefficient: -0.63, p < 0.001)
  + Type (video) (Coefficient: -512.82, p < 0.001)
  + Post timing (morning) (Coefficient: -150.72, p = 0.001)

**Macro Influencers**

* **Positive Influence:**
  + Followers (Coefficient: 0.0543, p < 0.001)
  + Question marks (Coefficient: 78.85, p = 0.028)
  + Engagement keywords (Coefficient: 617.48, p < 0.001)
* **Negative Influence:**
  + Number of tags (Coefficient: -81.09, p < 0.001)
  + Text length (Coefficient: -1.63, p = 0.028)
  + Emojis (Coefficient: -215.31, p = 0.421)
  + Type (video) (Coefficient: -1445.83, p < 0.001)
  + Post timing (morning) (Coefficient: -937.72, p = 0.002)

**COMMENTS**

**Micro Influencers:**

* **Positive Influence:**
  + Engagement keywords (Coefficient: 6.96, p < 0.001)
  + Text length (Coefficient: 0.0227, p < 0.001)
  + Exclamation marks (Coefficient: 2.27, p < 0.001)
* **Negative Influence:**
  + Number of mentions (Coefficient: -0.29, p = 0.505)
  + Hashtags (Coefficient: -0.41, p < 0.001)
  + Type (video) (Coefficient: -1.70, p = 0.503)

**Macro Influencers:**

* **Positive Influence:**
  + Followers (Coefficient: 0.0020, p < 0.001)
  + Question marks (Coefficient: 8.23, p = 0.032)
  + Engagement keywords (Coefficient: 195.43, p < 0.001)
* **Negative Influence:**
  + Hashtags (Coefficient: -8.58, p < 0.001)
  + Text length (Coefficient: -0.20, p = 0.014)
  + Post timing (night) (Coefficient: -88.73, p < 0.001)
  + Type (video) (Coefficient: -43.51, p = 0.153)

**Did the Coefficients for Input Variables Differ Between the Two Models?**

Yes, the coefficients (impact strength) varied between micro and macro influencers:

**Hashtags:**

Increased likes for micro influencers (Coefficient: 14.48) but decreased likes for macro influencers (Coefficient: -81.09).

**Emojis:**

Negatively influenced likes for both groups but had a stronger negative effect for macro influencers (Coefficient: -105.46 for micro vs. -215.31 for macro).

**Engagement keywords:**

Positively influenced comments for both groups but had a much stronger effect for macro influencers (Coefficient: 195.43 for macro vs. 6.96 for micro).

**Text length:**

Increased comments for micro influencers (Coefficient: 0.0227) but decreased comments for macro influencers (Coefficient: -0.20).

**Recommendations for increasing engagement**:

* **For Micro Influencers:**
  + Use more hashtags and tags, as these positively influence likes.
  + Keep text concise for likes but slightly longer for more comments.
  + Utilize exclamation marks to boost comments.
* **For Macro Influencers:**
  + Avoid excessive hashtags, as they reduce likes and comments.
  + Leverage engagement keywords to drive both likes and comments.
  + Question marks in captions can encourage comments.
  + Avoid posting videos frequently, as they negatively impact likes.

# Conclusions and Recommendations

**Summary of the Project**

In this project, we analyzed 19,681 Instagram posts to identify the key factors that drive user engagement, measured by likes and comments. Using the CRISP-DM methodology, we cleaned the dataset, created new variables, and applied regression models to uncover significant predictors of engagement. Our analysis included examining variables like follower count, post timing, hashtags, and text length, while also comparing the engagement dynamics for micro and macro influencers.

**Key Findings (Data Analytics Perspective)**

**Positive Factors**:

* **Follower Count:** Strongly increases both likes and comments, making it the most critical predictor.
* **Engagement Keywords:** Boosts likes and comments significantly, especially for macro influencers.
* **Question Marks:** Promotes comments by encouraging user interaction.
* **Hashtags and Tags:** Positively influence likes for micro influencers.

**Negative Factors:**

* **Text Length:** Longer captions reduce likes.
* **Video Posts:** Attract significantly fewer likes and comments compared to photos.
* **Post Timing:** Posts made in the evening or on weekends generally perform worse.

**Key Findings (General Audience Perspective)**

We found that posts with concise captions, engaging questions, and specific keywords designed to prompt interaction tend to get more likes and comments. Posts made on weekdays, especially in the afternoon, perform better than those posted on weekends or during the evening. Photos generally receive higher engagement compared to videos. For influencers with smaller audiences, using hashtags and tagging other users can help increase likes.

**Recommendations for Social Media Influencers**

* **Post Timing:** Post on weekdays, preferably in the afternoon, for maximum visibility and engagement.
* **Caption Strategy:** Use concise captions with engaging questions and interaction-driven keywords.
* **Content Type:** Focus on photo posts rather than videos to attract more likes and comments.
* **Micro Influencers:** Leverage hashtags and tags to boost reach and engagement.
* **Macro Influencers:** Avoid excessive hashtags, but strategically use engagement keywords to maintain relatability and drive interactions.

By implementing these strategies, social media influencers can enhance their post performance and foster stronger connections with their followers.