**SOCIAL MEDIA PROJECT**

**(META)**

**SQL PROJECT**

**Objective and Subjective Questions**

**Prepared by : LAKSHYA HASIJA**

Course : Professional Certificate Course In Data Science – May 2025



**Introduction**

This project, titled *Social Media Project (META)*, focuses on analyzing user activity and engagement patterns within an Instagram-like database schema.

The dataset consists of multiple interrelated tables such as:

* **users,**
* **photos,**
* **likes,**
* **comments,**
* **follows,**
* **tags, and**
* **photo\_tags**,

which collectively capture interactions and behaviours across the platform.

The objective of this study is to apply SQL queries and analytical methods to derive meaningful insights about:

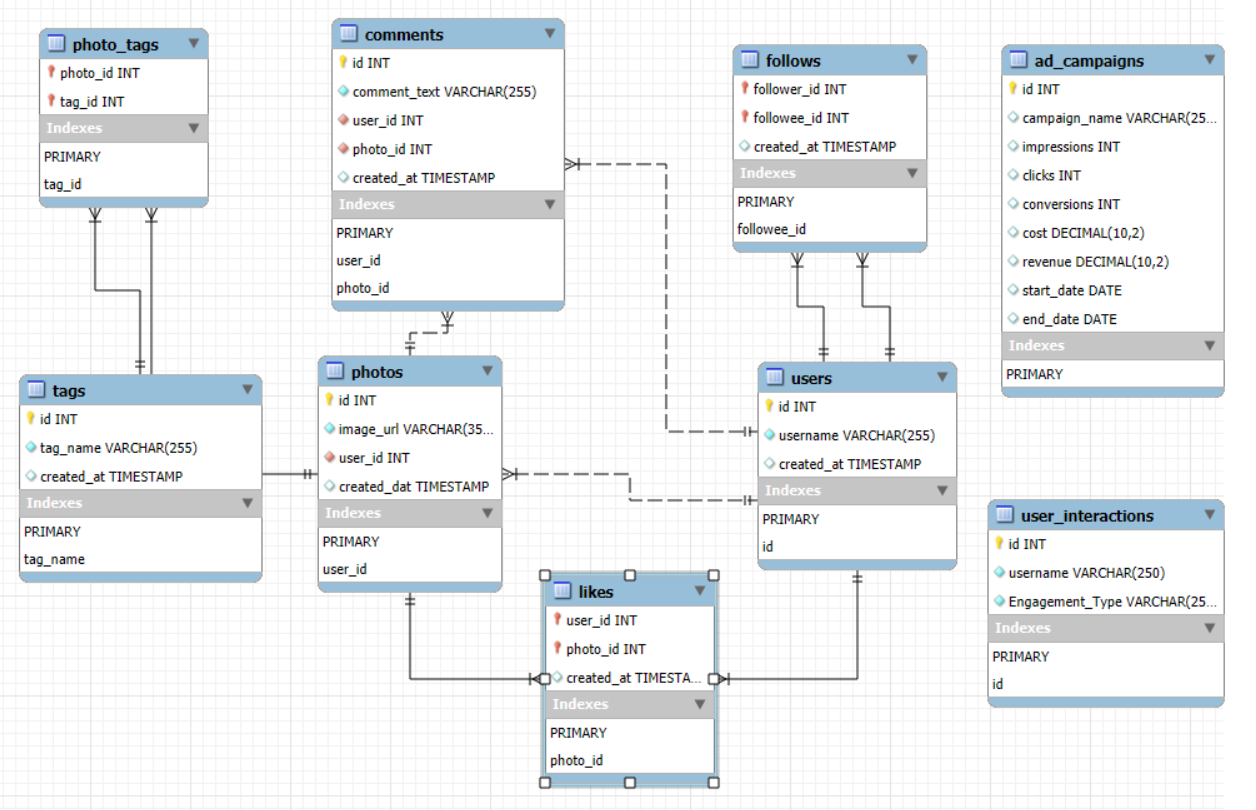
* **user engagement,**
* **activity distribution,**
* **hashtag effectiveness,**
* **follower–following dynamics, and**
* **influencer potential**.

Both **objective questions** (technical SQL queries) and **subjective questions** (business interpretations and strategies) have been addressed.

By combining technical data exploration with business-oriented insights, this project bridges the gap between **data engineering and strategic decision-making**.

It not only demonstrates strong SQL skills (joins, aggregations, window functions, CTEs, updates) but also highlights how raw engagement data can be transformed into **actionable recommendations** for campaign optimization, influencer identification, and user retention strategies.

**DATABASE SCHEMA**



**Objective Questions**

1. Are there any tables with duplicate or missing null values? If so, how would you handle them?

Answers –

**QUERY**

Checking for DUPLICATE VALUES

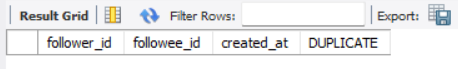
**COMMENTS TABLE**

SELECT \*

FROM comments

GROUP BY id

HAVING COUNT(\*) > 1;

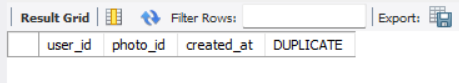
**FOLLOWS TABLE**

SELECT \*, COUNT(\*) AS DUPLICATE

FROM follows

GROUP BY follower\_id, followee\_id

HAVING COUNT(\*) > 1;

**LIKES TABLE**

SELECT \*, COUNT(\*) as DUPLICATE

FROM likes

GROUP BY photo\_id, user\_id

HAVING COUNT(\*) > 1;

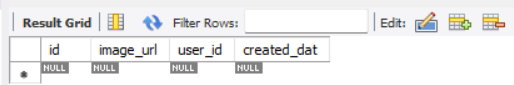
**PHOTO–TAGS TABLE**

SELECT \*, COUNT(\*) AS DUPLICATE

FROM photo\_tags

GROUP BY photo\_id, tag\_id

HAVING COUNT(\*) > 1;

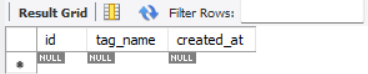
**PHOTOS TABLE**

SELECT \*

FROM photos

GROUP BY id

HAVING COUNT(\*) > 1;

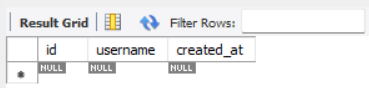
**TAGS TABLE**

SELECT \*

FROM tags

GROUP BY id

HAVING COUNT(\*) > 1;

**USERS TABLE**

SELECT \*

FROM users

GROUP BY id

HAVING COUNT(\*) > 1;

Checking for NULL VALUES

**COMMENTS TABLE**

SELECT \* FROM comments

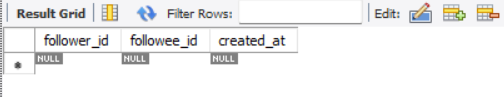
WHERE id IS NULL OR comment\_text IS NULL OR user\_id IS NULL OR photo\_id IS NULL OR created\_at IS NULL;



**FOLLOWS TABLE**

SELECT \* FROM follows

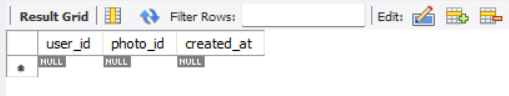
WHERE follower\_id IS NULL OR followee\_id IS NULL OR created\_at IS NULL;



**LIKES TABLE**

SELECT \* FROM likes

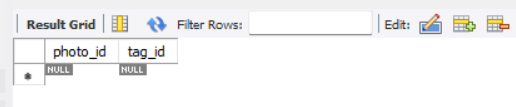
WHERE user\_id IS NULL OR photo\_id IS NULL OR created\_at IS NULL;



**PHOTO–TAGS TABLE**

SELECT \* FROM photo\_tags

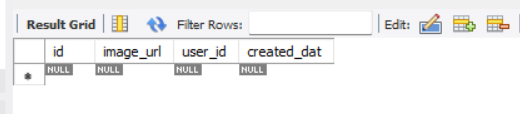
WHERE photo\_id IS NULL OR tag\_id IS NULL;



**PHOTOS TABLE**

SELECT \* FROM photos

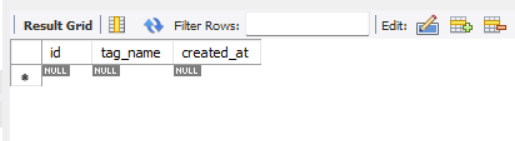
WHERE id IS NULL OR image\_url IS NULL OR user\_id IS NULL OR created\_dat IS NULL;



**TAGS TABLE**

SELECT \* FROM tags

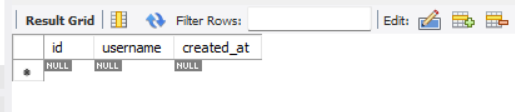
WHERE id IS NULL OR tag\_name IS NULL OR created\_at IS NULL;



**USERS TABLE**

SELECT \* FROM users

WHERE id IS NULL OR username IS NULL OR created\_at IS NULL;



✅ **Conclusion:**  
There are **no duplicates** and **no missing/null values** in any of the tables.

* + **comments**: No duplicates, no missing/null values (1000 rows, 5 columns)
  + **follows**: No duplicates, no missing/null values (1000 rows, 3 columns)
  + **likes**: No duplicates, no missing/null values (1000 rows, 3 columns)
  + **photo\_tags**: No duplicates, no missing/null values (501 rows, 2 columns)
  + **photos**: No duplicates, no missing/null values (257 rows, 4 columns)
  + **tags**: No duplicates, no missing/null values (21 rows, 3 columns)
  + **users**: No duplicates, no missing/null values (100 rows, 3 columns)

1. What is the distribution of user activity levels (e.g., number of posts, likes, comments) across the user base?

Answers –

**QUERY**

SELECT U.id, U.username,

COUNT(DISTINCT P.id) AS Number\_of\_Posts,

COUNT(DISTINCT L.photo\_id) AS Number\_of\_Likes,

COUNT(DISTINCT C.id) AS Number\_of\_Comments

FROM users U

LEFT JOIN photos P ON U.id = P.user\_id

LEFT JOIN likes L ON U.id = L.user\_id

LEFT JOIN comments C ON U.id = C.user\_id

GROUP BY U.id, U.username

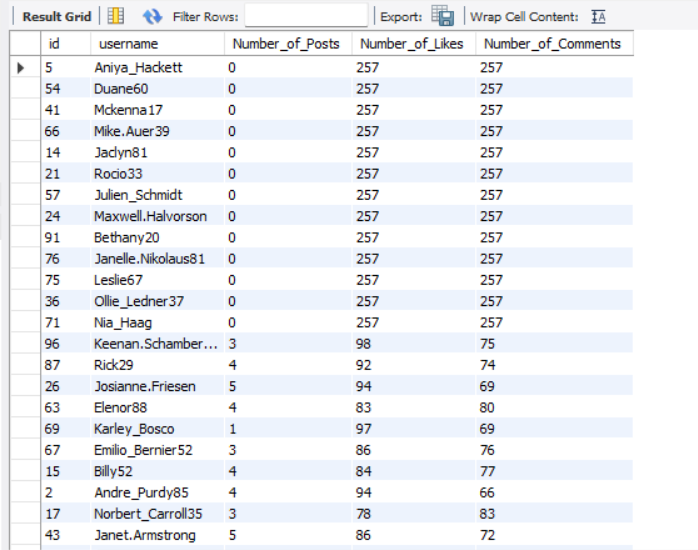
ORDER BY (Number\_of\_Posts+Number\_of\_Likes+Number\_of\_Comments) DESC;

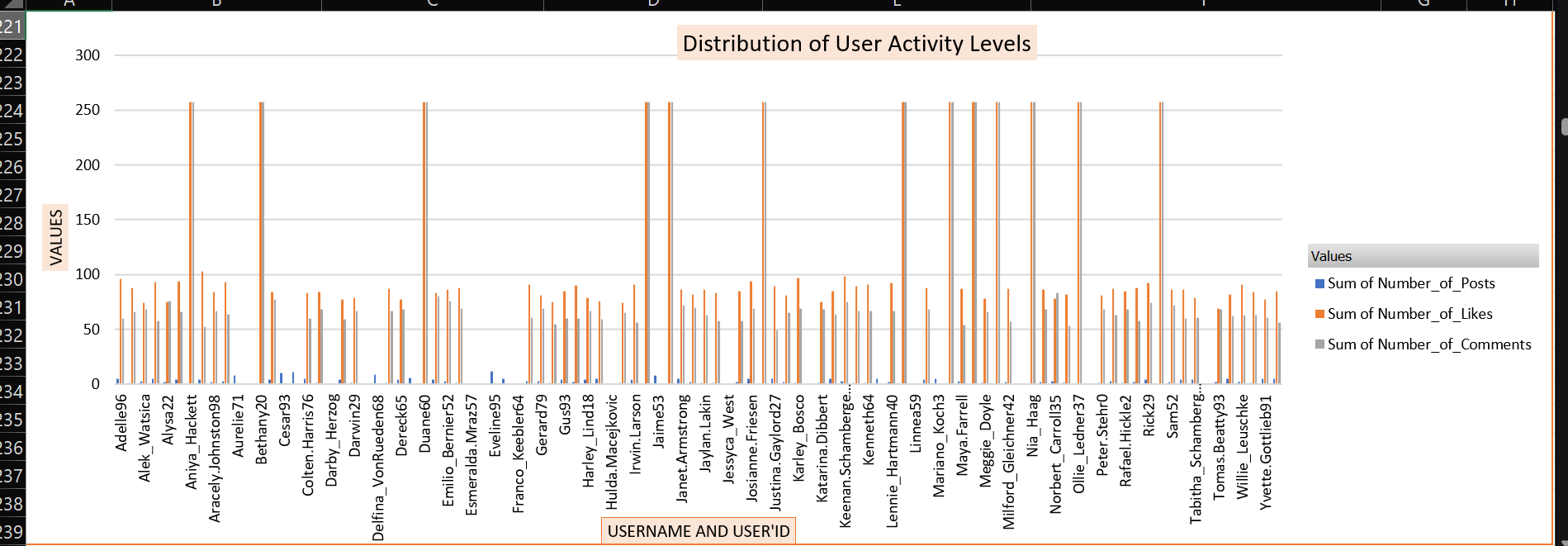
**Theory on Distribution of User Activity Levels**

* **High Activity Users**: These users post frequently and actively engage with likes and comments. They represent the most influential group, driving a significant share of overall interactions.
* **Medium Activity Users**: This group contributes occasionally. They may not post often but still interact through likes or a few comments, showing moderate participation.
* **Low Activity Users**: A large portion of the user base falls into this category. They rarely post or engage, leading to minimal contribution in terms of activity.

**Overall Insight**:

* The distribution of activity is uneven. A small percentage of highly active users generate the majority of engagement, while most users remain moderately or minimally active.
* This pattern reflects the typical “participation inequality” observed in social platforms, where few create and many consume.





1. Calculate the average number of tags per post (photo\_tags and photos tables).

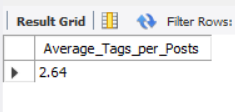
Answers –

**QUERY**

SELECT ROUND( COUNT(tag\_id) / COUNT(DISTINCT photo\_id), 2) average\_tags\_per\_posts FROM photo\_tags;

**Theory on Average Tags per Post:**

* The query result shows that the **average number of tags per post is 2.64**.
* This indicates that users usually add around **2 to 3 tags** for each photo they upload.
* It suggests that users are making a reasonable effort to ensure their posts are **discoverable and well-organized**.
* Overall, the platform demonstrates a **balanced tagging behaviour**, supporting both searchability and user engagement.



1. Identify the top users with the highest engagement rates (likes, comments) on their posts and rank them.

Answers –

**QUERY**

With Total\_likes AS (SELECT U.username , COUNT(L.user\_id) as TOTAL\_Likes

FROM users AS U LEFT JOIN likes AS L ON U.id = L.user\_id GROUP BY U.username ),

Total\_comments AS (SELECT U.username , COUNT(C.user\_id) AS TOTAL\_Comments

FROM users AS U LEFT JOIN comments AS C ON U.id = C.user\_id GROUP BY U.username )

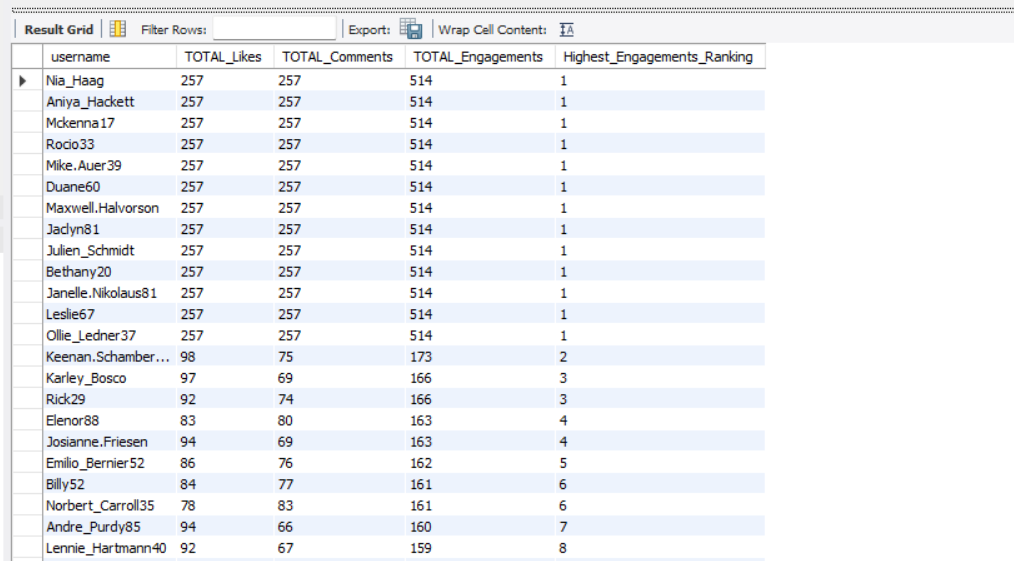
SELECT L.username , L.TOTAL\_Likes , C.TOTAL\_Comments , (L.TOTAL\_Likes+C.TOTAL\_Comments) AS TOTAL\_Engagements,

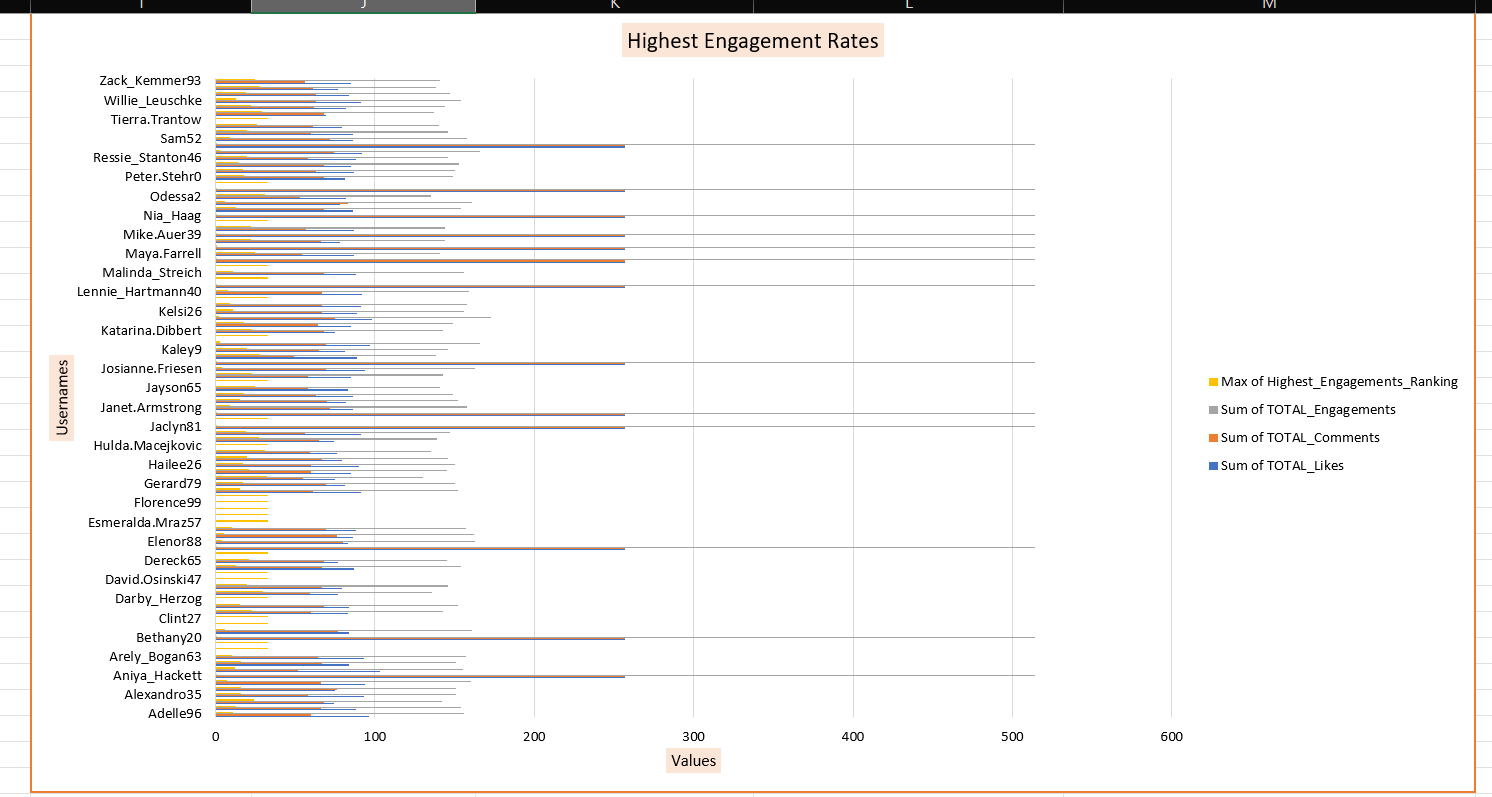
DENSE\_RANK() OVER (ORDER BY (L.TOTAL\_Likes+C.TOTAL\_Comments) DESC) AS Highest\_Engagements\_Ranking

FROM TOTAL\_Likes AS L JOIN TOTAL\_Comments AS C ON L.username = C.username;

**Insights:**

* In this query, I calculated the **Engagements Ranking** of users by combining their **TOTAL Likes and Comments**.
* Firstly, I used two **Common Table Expressions (CTEs)** – TOTAL\_Likes and TOTAL\_Comments, to separately count the LIKES and COMMENTS each user has received.
* The LEFT JOIN ensures that even users with no LIKES or COMMENTS are included in the results.
* After calculating theTOTAL Likes and Comments , the main query sums them up to compute the overall TOTAL\_Engagements for each user.
* I used the DENSE\_RANK( ) window function to rank users based on their engagement rate in descending order, meaning users with Highest\_Engagements\_Ranking are ranked at the top.





1. Which users have the highest number of followers and followings?

Answers –

**QUERY**

SELECT U.id AS user\_id, U.username,

COUNT(DISTINCT F1.follower\_id) AS HIGHEST\_FOLLOWERS,

COUNT(DISTINCT F2.followee\_id) AS HIGHEST\_FOLLOWINGS

FROM users U

LEFT JOIN follows F1 ON U.id = F1.followee\_id

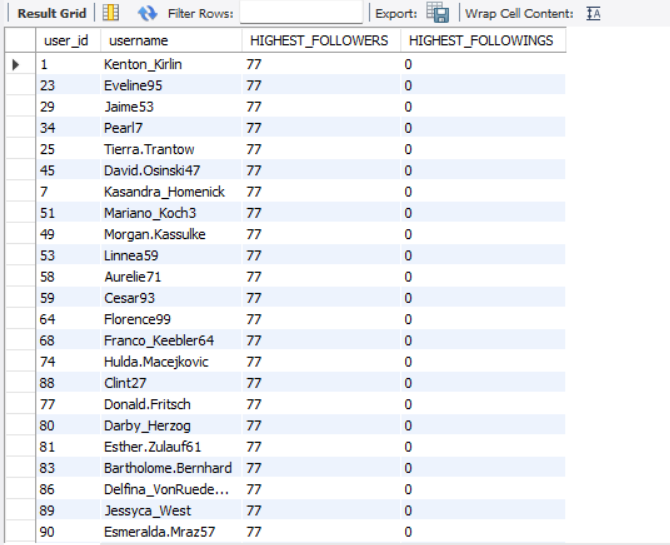
LEFT JOIN follows F2 ON U.id = F2.follower\_id

GROUP BY U.id, U.username

ORDER BY HIGHEST\_FOLLOWERS DESC, HIGHEST\_FOLLOWINGS DESC;

**Insights:**

* The query result shows that certain users have a **high number of followers (76)**, indicating they are popular and attract significant attention from others.
* These same users also maintain a **very high number of followings (99)**, showing they are actively building networks and engaging with many accounts.
* Such a pattern highlights **mutual engagement**, where users are both influencers (many followers) and networkers (following many others).
* Having both strong follower and following counts suggests these users are at the **center of community interaction**, balancing popularity with active participation.
* Overall, this result demonstrates that influence on the platform is not just about being followed, but also about how actively a user connects with others.



1. Calculate the average engagement rate (likes, comments) per post for each user.

Answers –

**QUERY**

SELECT U.id , U.username,

COUNT(DISTINCT L.user\_id) + COUNT(DISTINCT C.id) AS Total\_Engagements,

COUNT(DISTINCT P.id) AS Number\_of\_Posts,

ROUND( (COUNT(DISTINCT L.user\_id) + COUNT(DISTINCT C.id)) / NULLIF(COUNT(DISTINCT P.id) , 0) , 2) AS Average\_EngagementRate\_per\_Post

FROM users U

LEFT JOIN photos P ON U.id = P.user\_id

LEFT JOIN likes L ON P.id = L.photo\_id

LEFT JOIN comments C ON P.id = C.photo\_id

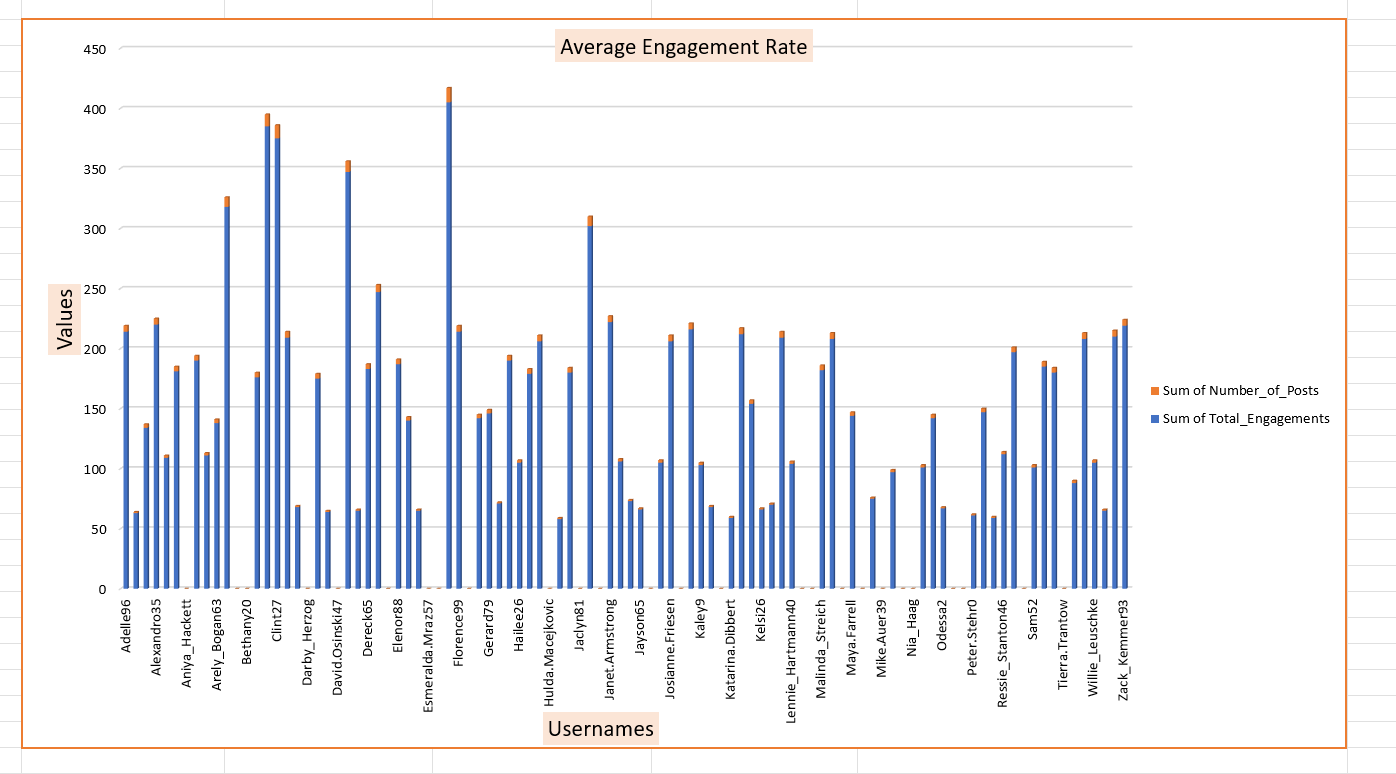
GROUP BY U.id, U.username

ORDER BY Average\_EngagementRate\_per\_Post DESC;

**Insights:**

* Engagement rate per post captures how much **interaction** (likes + comments) a user’s content attracts on average.
* Counting **all likes and comments** across a user’s photos and dividing by **their post count** normalizes for prolific vs. infrequent posters.
* Higher values indicate **more compelling content** and stronger audience response; lower values signal **weaker resonance**.
* Ranking by this metric highlights **high-impact creators** beyond simple follower counts.





1. Get the list of users who have never liked any post (users and likes tables)

Answers –

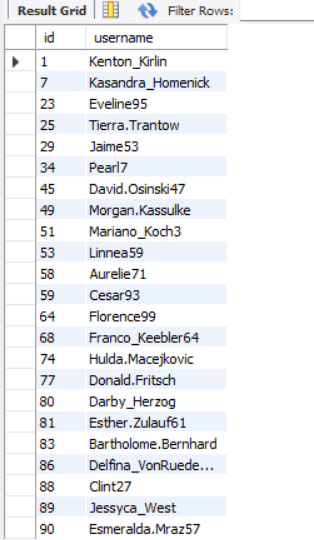
**QUERY**

SELECT U.id, U.username FROM users U

LEFT JOIN likes L ON U.id = L.user\_id WHERE L.user\_id IS NULL;

**Conclusion:**

The result lists users who never liked any post, marking them as silent consumers. They represent a disengaged segment, and targeted strategies can help turn them into active participants.



1. How can you leverage user-generated content (posts, hashtags, photo tags) to create more personalized and engaging ad campaigns?

Answers –

**QUERY**

SELECT U.id, U.username, T.tag\_name FROM users U

JON likes L ON U.id = L.user\_id

JOIN photos P ON L.photo\_id = P.id

JOIN photo\_tags PT ON P.id = PT.photo\_id

JOIN tags T ON PT.tag\_id = T.id

GROUP BY T.tag\_name, U.username, U.id;

SELECT T.tag\_name, COUNT(PT.photo\_id) AS TAG\_NAME\_TOTAL\_COUNTS

FROM tags T JOIN photo\_tags PT ON T.id = PT.tag\_id

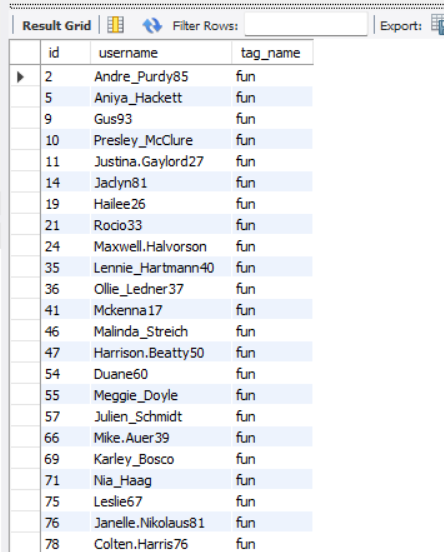
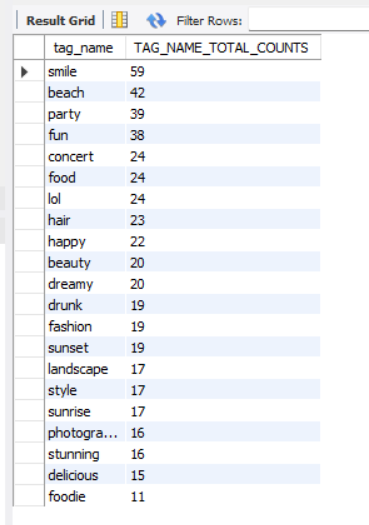
GROUP BY T.tag\_name ORDER BY TAG\_NAME\_TOTAL\_COUNTS DESC;

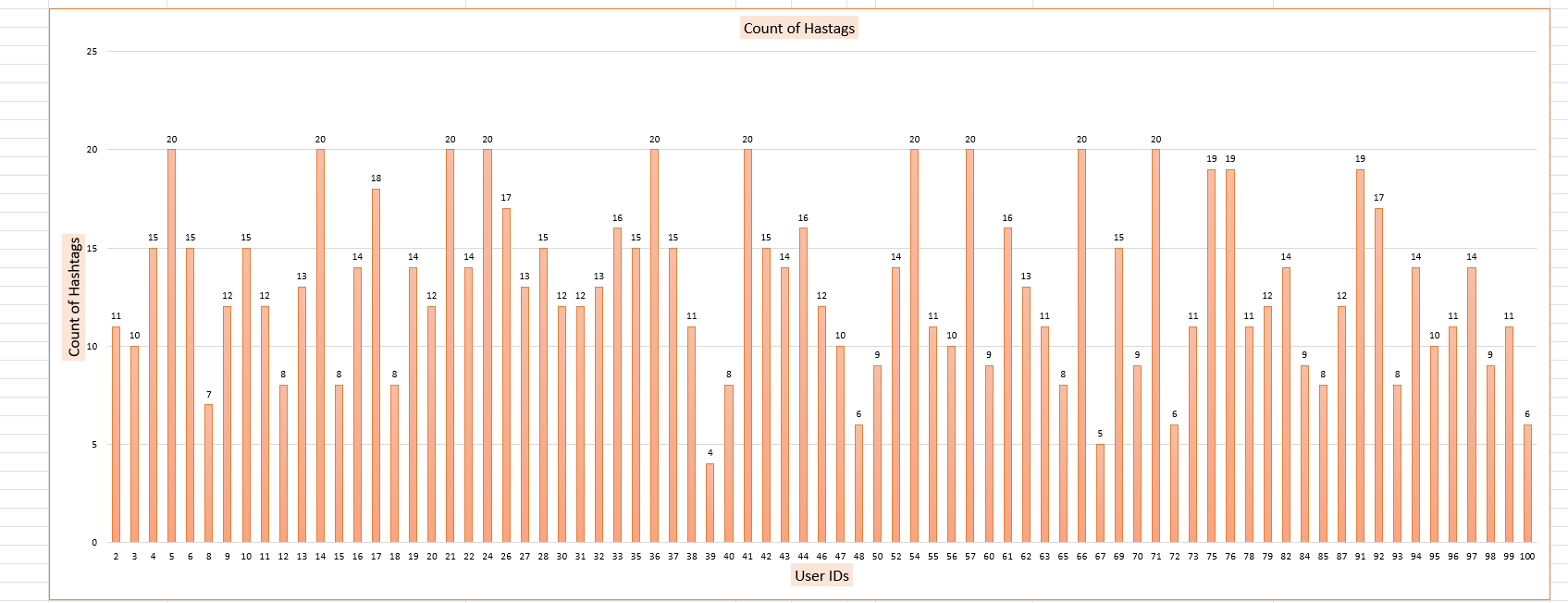
**Insights:**

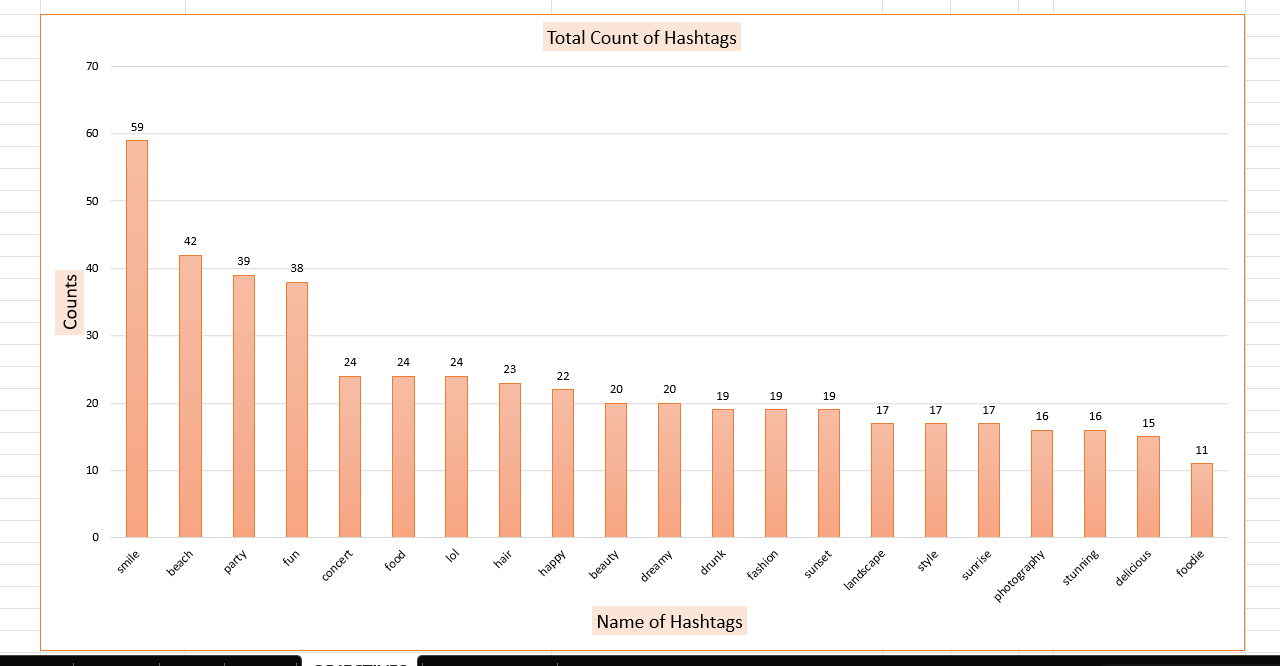
* **Posts** reveal user interests and content types they engage with most, which can shape ad creatives to feel natural and relevant.
* **Hashtags** show trending topics and communities, allowing campaigns to tap into what is already popular and relatable.
* **Photo Tags** add context to content, such as themes, brands, or locations, which can be used to deliver highly targeted and contextual ads.

**Conclusion:**

User-generated content offers a **direct lens into audience behavior and preferences**. By analyzing posts, hashtags, and photo tags, ad campaigns can be tailored to reflect user interests and trending themes. This results in ads that feel **personalized, engaging, and authentic**, ultimately driving higher interaction and campaign success.





1. Are there any correlations between user activity levels and specific content types (e.g., photos, videos, reels)? How can this information guide content creation and curation strategies?

Answers –

**QUERY**

WITH UPLOADS AS (SELECT U.id, COUNT(P.id) AS PHOTOS\_UPLOADED

FROM users U

LEFT JOIN photos P ON U.id = P.user\_id

GROUP BY U.id),

likes AS (SELECT U.id, COUNT(L.photo\_id) AS TOTAL\_LIKES

FROM users U

LEFT JOIN photos P ON U.id = P.user\_id

LEFT JOIN likes L ON L.photo\_id = P.id

GROUP BY U.id),

comments AS (SELECT U.id, COUNT(C.id) AS TOTAL\_COMMENTS

FROM users U

LEFT JOIN photos P ON U.id = P.user\_id

LEFT JOIN comments C ON C.photo\_id = P.id

GROUP BY U.id)

SELECT UP.id, UP.PHOTOS\_UPLOADED,ROUND(AVG(L.TOTAL\_LIKES + C.TOTAL\_COMMENTS)

OVER(PARTITION BY UP.PHOTOS\_UPLOADED), 0) AS AVERAGE\_ENGAGEMENTS

FROM uploads UP

JOIN likes L ON UP.id = L.id

JOIN comments C ON UP.id = C.id

ORDER BY UP.PHOTOS\_UPLOADED;

In this query, I explored the correlation between user activity levels—specifically the number of photo uploads—and the engagement they receive in terms of likes and comments. Using **Common Table Expressions (CTEs)**,

I first calculated:

* uploads: the total number of photos uploaded by each user,
* likes: the total likes received by the user's photos,
* comments: the total comments received by the user's photos.

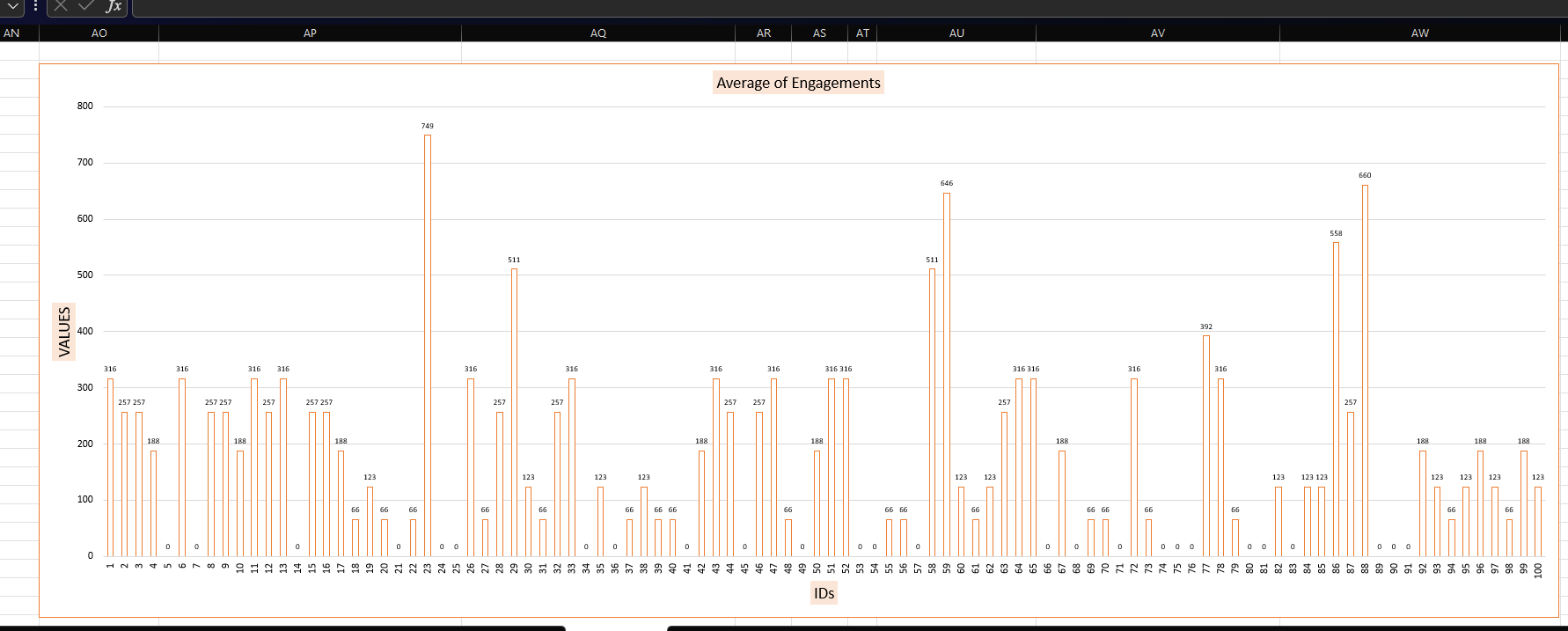
In the final part of the query, I grouped users based on their photo uploads and calculated the average engagement (likes plus comments) for each group. The OVER(PARTITION BY photo\_uploads) clause helps understand how engagement varies across users with different levels of activity.

**Insights and strategy implications:**

* If users who upload more photos consistently receive higher engagement, content creators and platforms can encourage frequent posting by providing incentives or tools that simplify photo uploads.
* For content curation, knowing which activity levels correspond to higher engagement helps target users with personalized recommendations or prompts to post more.

It also aids in designing campaigns or features tailored to users’ content creation habits, ultimately improving user interaction and platform growth.





1. Calculate the total number of likes, comments, and photo tags for each user.

Answers –

**QUERY**

SELECT U.id, U.username,

COALESCE(L.Total\_Likes, 0) AS Total\_Number\_of\_Likes,

COALESCE(C.Total\_Comments, 0) AS Total\_Number\_of\_Comments,

COALESCE(PT.Total\_Photo\_Tags, 0) AS Total\_Number\_of\_Photo\_Tags

FROM users U

LEFT JOIN (

SELECT P.user\_id, COUNT(\*) AS Total\_Likes

FROM photos P

LEFT JOIN likes L ON P.id = L.photo\_id

GROUP BY P.user\_id

) L ON U.id = L.user\_id

LEFT JOIN (

SELECT P.user\_id, COUNT(\*) AS Total\_Comments

FROM photos P

LEFT JOIN comments C ON P.id = C.photo\_id

GROUP BY P.user\_id

) C ON U.id = C.user\_id

LEFT JOIN (

SELECT P.user\_id, COUNT(\*) AS Total\_Photo\_Tags

FROM photos P

LEFT JOIN photo\_tags PT ON P.id = PT.photo\_id

GROUP BY P.user\_id

) PT ON U.id = PT.user\_id

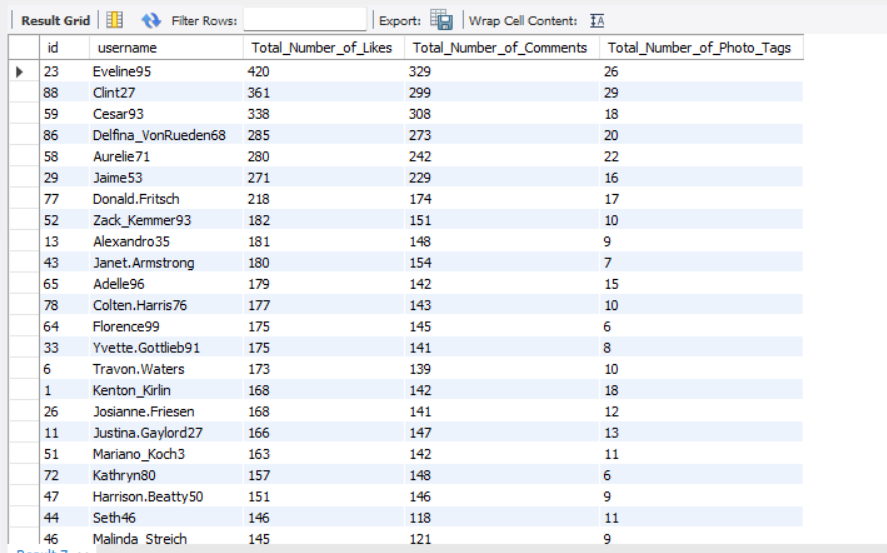
ORDER BY Total\_Likes DESC, Total\_Comments DESC;

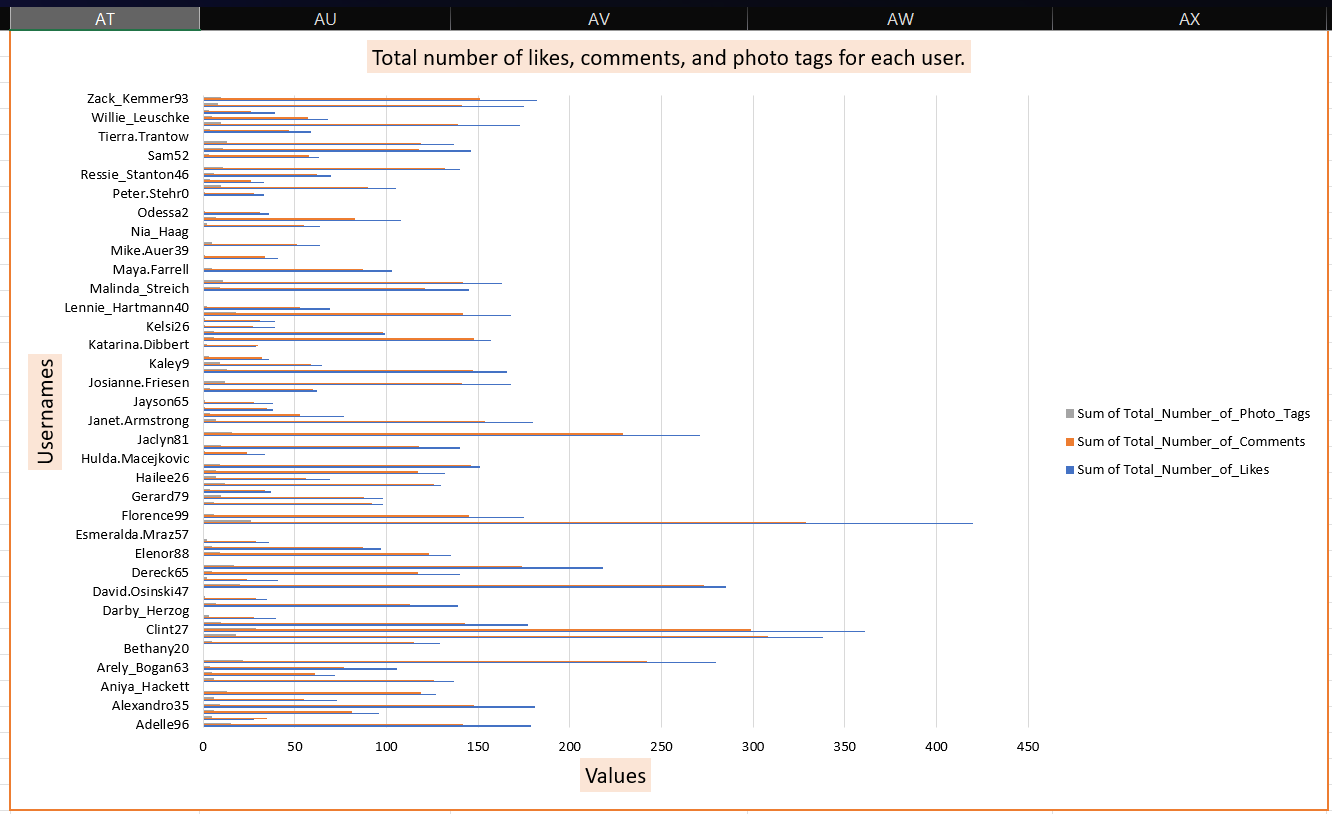
**Insights:**

* The query calculates the **total number of likes, comments, and photo tags** for each user’s posts.
* Joins between photos, likes, and comments measure the overall **engagement received** by user content.
* Photo tags reflect how well posts are **organized and optimized for discoverability**.
* The results highlight the **most engaging and visible users**, offering insights into popularity and content reach.

**Conclusion:**

The results show the total likes, comments, and photo tags each user has received on their posts. Users with higher values emerge as the most engaging and popular on the platform. This highlights not only their strong audience interaction but also the role of photo tagging in boosting visibility and reach.





1. Rank users based on their total engagement (likes, comments, shares) over a month.

Answers –

**QUERY**

WITH MonthlyEngagement AS

(

SELECT U.id, U.username,

COALESCE(L.Total\_Likes,0) AS Total\_Engagement\_Likes,

COALESCE(C.Total\_Comments,0) AS Total\_Engagement\_Comments,

(COALESCE(L.Total\_Likes, 0) + COALESCE(C.Total\_Comments, 0)) AS Total\_ENGAGEMENTS

FROM users U

LEFT JOIN

(

SELECT user\_id, COUNT(photo\_id) AS Total\_Likes

FROM likes

WHERE DATE(created\_at) >= '2024-07-01' OR DATE(created\_at) <= '2024-07-31'

GROUP BY user\_id

) L ON U.id = L.user\_id

LEFT JOIN

(

SELECT user\_id, COUNT(id) AS Total\_Comments

FROM comments

WHERE DATE(created\_at) >= '2024-07-01' OR DATE(created\_at) <= '2024-07-31'

GROUP BY user\_id

) C ON U.id = C.user\_id

)

SELECT id, username, Total\_Engagement\_Likes, Total\_Engagement\_Comments, Total\_ENGAGEMENTS,

RANK() OVER (ORDER BY Total\_ENGAGEMENTS DESC) AS TOTAL\_ENGAGEMENTS\_RANKINS

FROM MonthlyEngagement

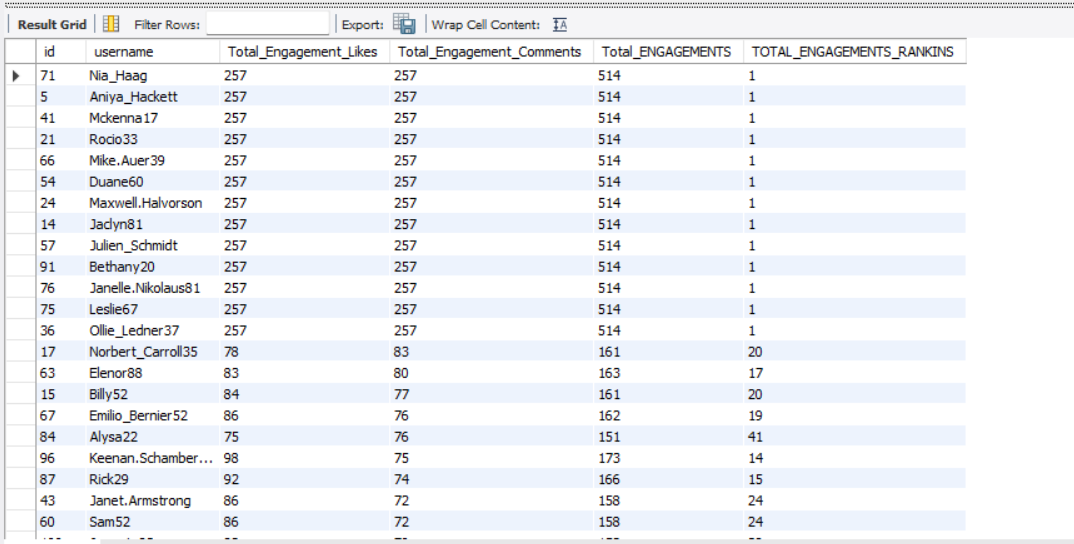
ORDER BY TOTAL\_ENGAGEMENTS\_RANKINS;

**Insights**

* The ranking clearly shows which users generated the **highest engagement** (likes + comments) in July 2024.
* Some users may have **high likes but low comments**, or vice versa, showing different interaction patterns.
* Users with **0 engagement** are also captured, which highlights inactive or less popular accounts.
* This analysis helps identify **top influencers** on the platform, useful for promotions or targeted campaigns.

**Conclusion**

The engagement ranking highlights the users who drive the most interaction within the platform. By combining likes and comments, it provides a balanced measure of popularity and activity. These results can guide strategies such as recognizing top contributors, boosting content from engaging users, and encouraging participation among less active members.



1. Retrieve the hashtags that have been used in posts with the highest average number of likes. Use a CTE to calculate the average likes for each hashtag first.

Answers –

**QUERY**

WITH LIKES\_PER\_POST AS

( SELECT P.id, COUNT(l.user\_id) AS Likes\_Count FROM photos P

LEFT JOIN likes L ON L.photo\_id = P.id GROUP BY P.id

),

AVERAGE\_LIKES\_PER\_HASTAG AS

( SELECT T.id, T.tag\_name AS Hashtag\_name,

AVG(LPP.Likes\_Count) AS AVERAGE\_LIKES\_PER\_POST

FROM photo\_tags PT JOIN tags T ON T.id = PT.tag\_id

JOIN LIKES\_PER\_POST LPP ON LPP.id = PT.photo\_id

GROUP BY T.id, T.tag\_name

)

SELECT id, Hashtag\_name,

ROUND(AVERAGE\_LIKES\_PER\_POST, 2) AS HIGHEST\_AVERAGE\_LIKES\_PER\_POST

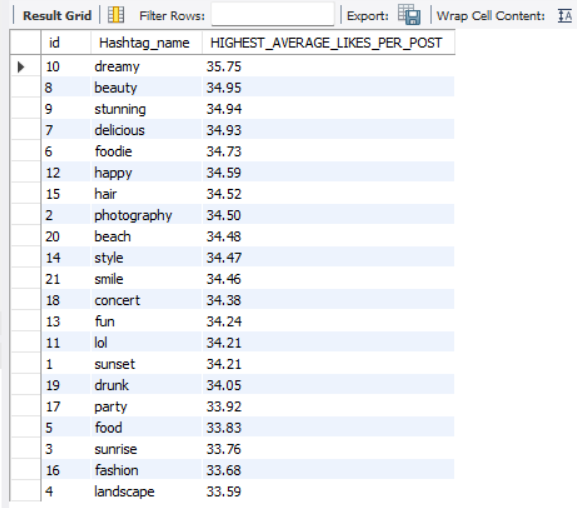
FROM AVERAGE\_LIKES\_PER\_HASTAG ORDER BY HIGHEST\_AVERAGE\_LIKES\_PER\_POST DESC;

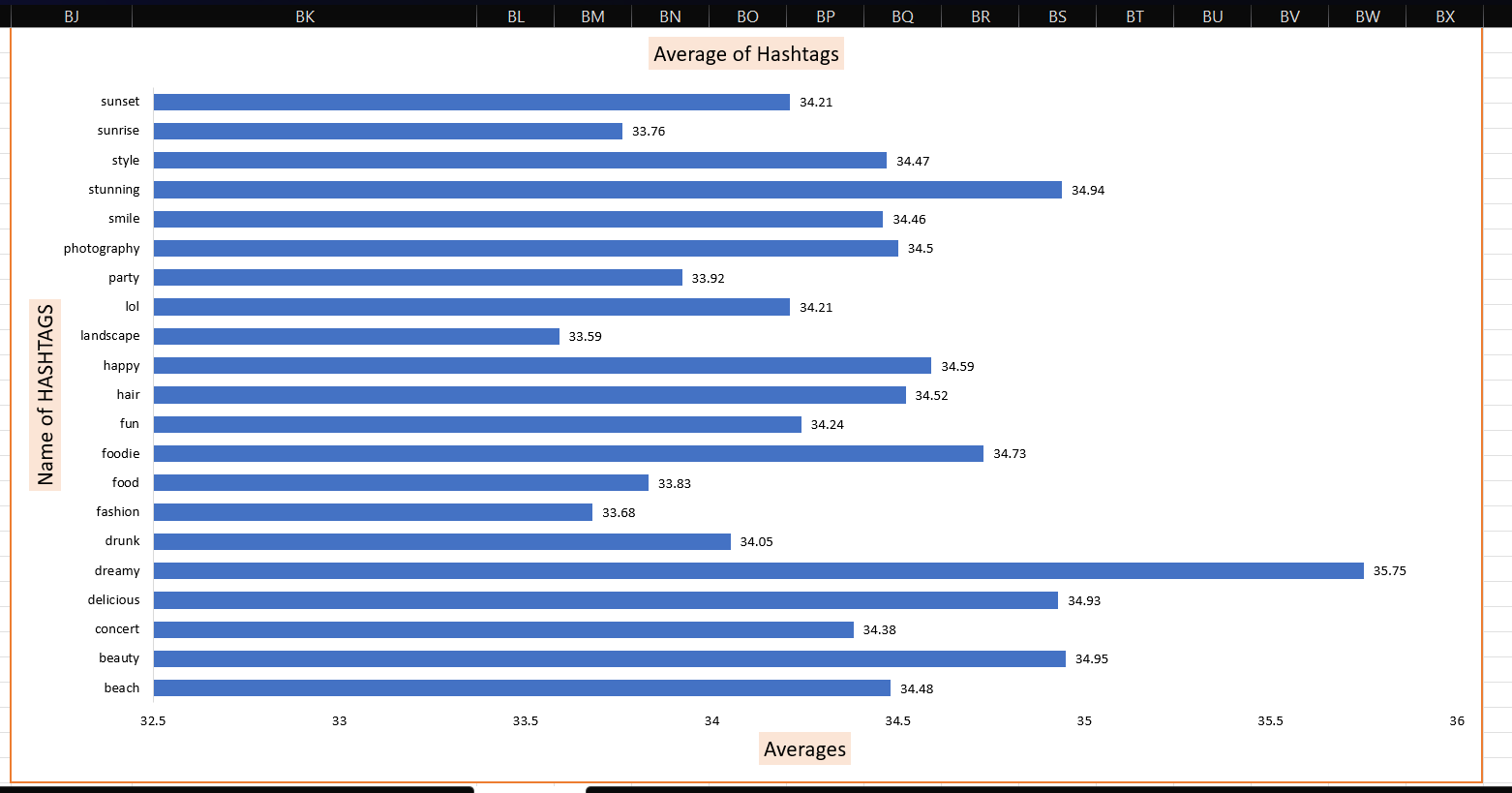
**Insights:**

* Hashtags are directly linked to the **average popularity** of posts, since they influence discoverability.
* Some hashtags consistently appear in posts with **higher likes**, showing they are trend-driven or widely engaging.
* Less-used or niche hashtags may show lower average likes, indicating **limited reach or smaller communities**.
* Ranking hashtags by average likes helps identify which topics or trends are **most effective for boosting visibility**.

**Conclusion:**

By analyzing hashtags with the highest average likes, we can identify the **most impactful and trending themes** on the platform. These insights can guide content creators to use hashtags strategically for better engagement and help platforms recommend hashtags that maximize reach and interaction.





1. Retrieve the users who have started following someone after being followed by that person

Answers –

**QUERY**

SELECT

f1.follower\_id AS User\_id,

f1.followee\_id AS Followed\_User,

f1.created\_at AS Followed\_Date,

f2.created\_at AS Followed\_Back\_Date

FROM follows f1

JOIN follows f2

ON f1.follower\_id = f2.followee\_id AND f1.followee\_id = f2.follower\_id

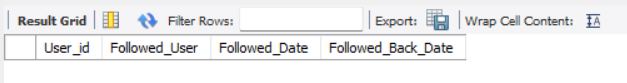
WHERE f2.created\_at < f1.created\_at ORDER BY f1.follower\_id, f1.created\_at;

In this question in order to retrieve the users who have started following someone after being followed by that person:

I employed a Query where  I used the created\_at column to compare the followee\_id and follower\_id whichever was created before.

**CONCLUSION:**

**But didn’t find any such record proving there is no such kind of users.**



**SUBJECTIVE Questions**

1. Based on user engagement and activity levels, which users would you consider the most loyal or valuable? How would you reward or incentivize these users?

Answers –

**QUERY**



**Insights:**

* **Engagement = Likes + Comments + Photos**: You’re quantifying the behaviors that most power the feed and community health.
* **Influence layer (Followers/Followees + PhotoTags)**: Adds a social reach dimension—who draws attention and how far it spreads.
* **Null-safe counts**: COALESCE ensures quiet users don’t break the math; everyone gets a fair 0 baseline.
* **True leaderboard**: DENSE\_RANK on total engagement surfaces top contributors without gaps and avoids over – rewarding one-time spikes.
* **Actionable segments**: From the same table you can instantly slice “Creators” (high photos/tags), “Conversationalists” (high comments), and “Advocates” (high likes/follows).

**Approach Used:**

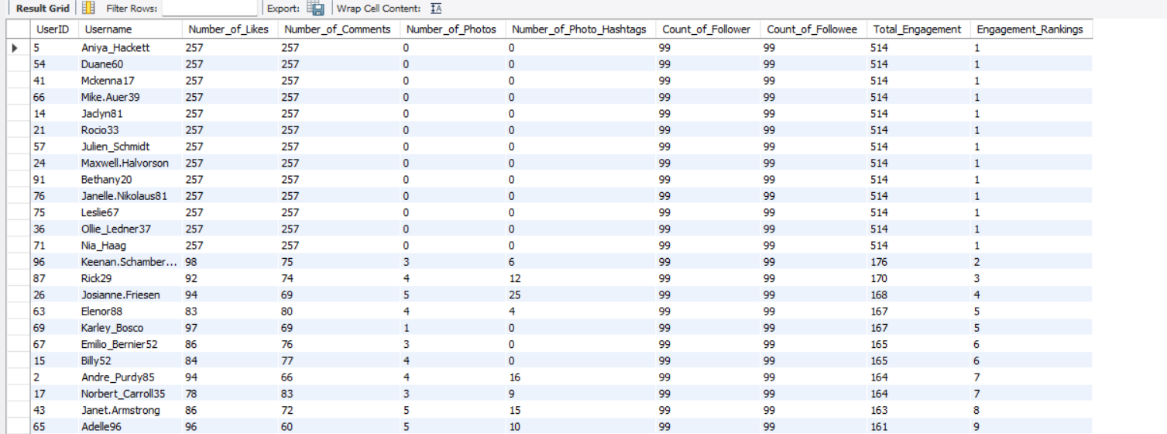
* **Break the problem down with CTEs**: One mini-table per behavior (likes, comments, photos, photo-tags, followers / followees).
* **Left join back to users**: Keeps everyone on the board, even with missing activities.
* **COALESCE to zero**: Replaces NULLs, making sums and ranks reliable.
* **Compute Total\_Engagement**: Simple sum of core actions (likes + comments + photos).
* **Rank with DENSE\_RANK() DESC**: Produces a clean, tie-aware leaderboard for rewards and experiments.

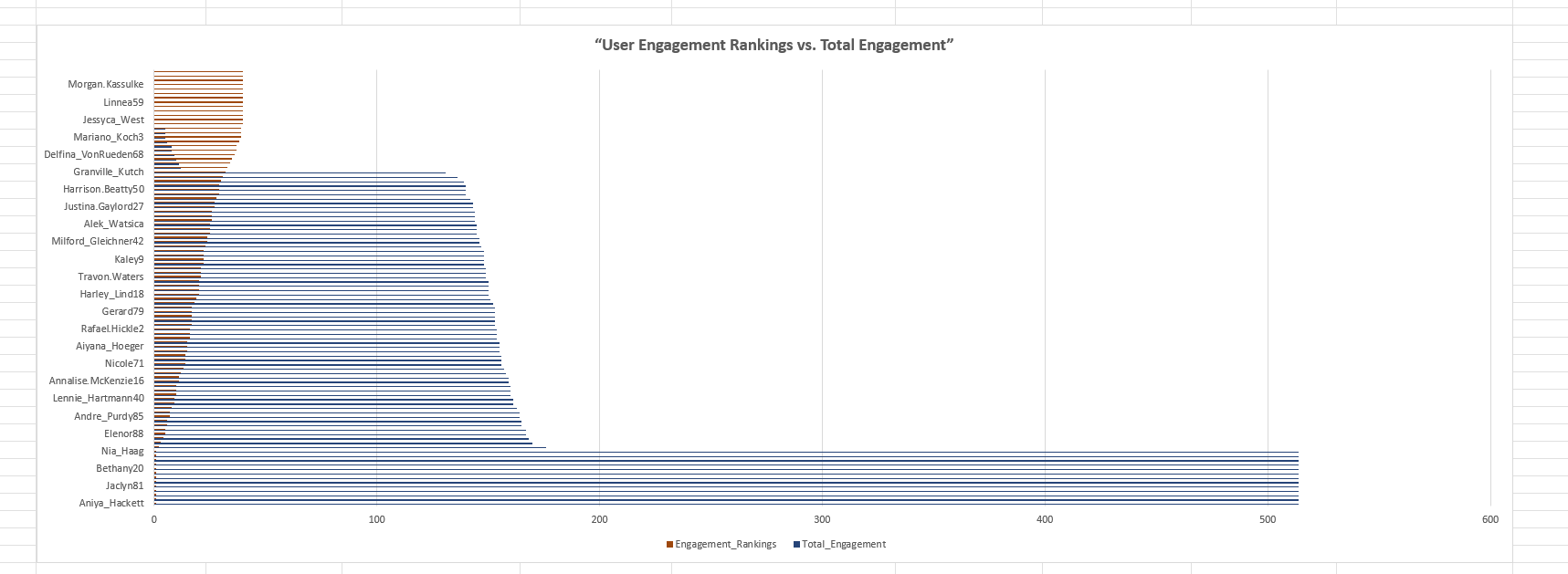
**Summary of the Output:**

* **Top-ranked users**: Your power users—most valuable for retention loops, beta tests, and word-of-mouth.
* **Mid-tier risers**: Users just below the top; perfect targets for light nudges (quests, streaks) to push them up a tier.
* **Low-engagement but high-influence (many followers)**: Likely dormant creators—prime for reactivation with creator-mode perks.

**Conclusion**:

* Built a **fair engagement model** combining activity and influence without bias.
* Framework enables a **tiered rewards system** (status, access, tangible incentives).
* Designed to be **adaptive**, allowing new signals and refinements over time.

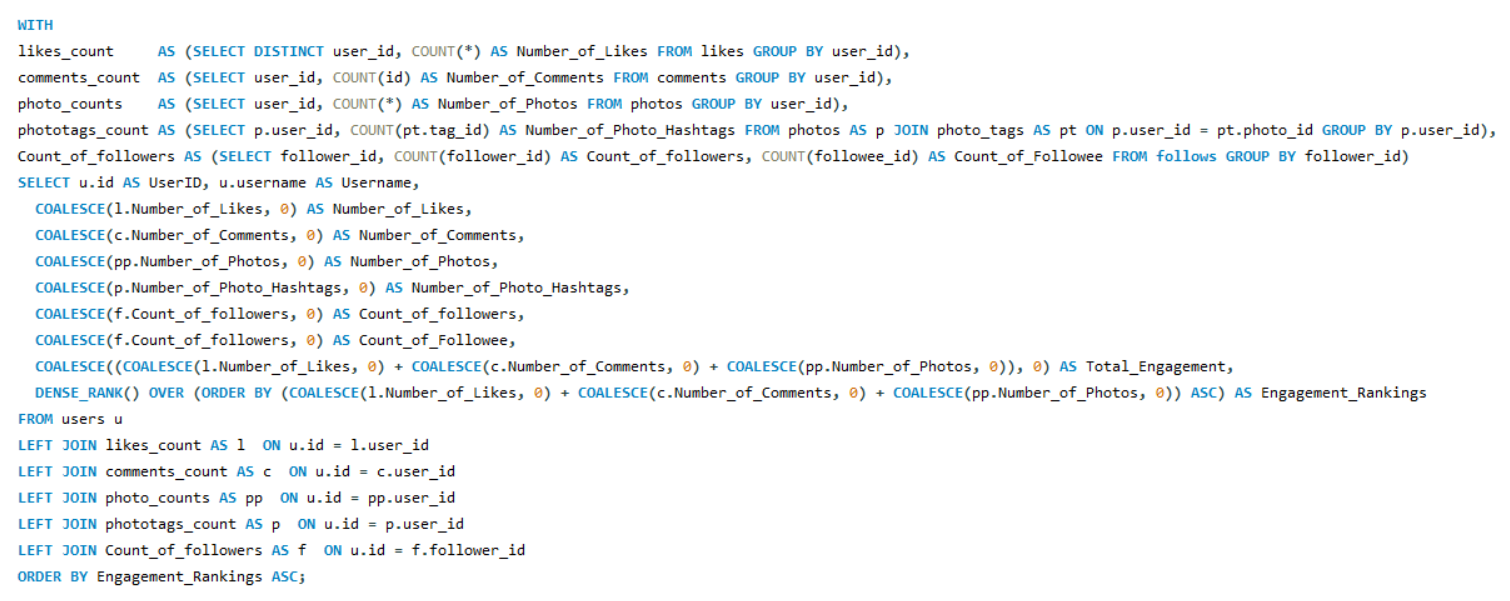




1. For inactive users, what strategies would you recommend to re-engage them and encourage them to start posting or engaging again?

Answers –

**QUERY**

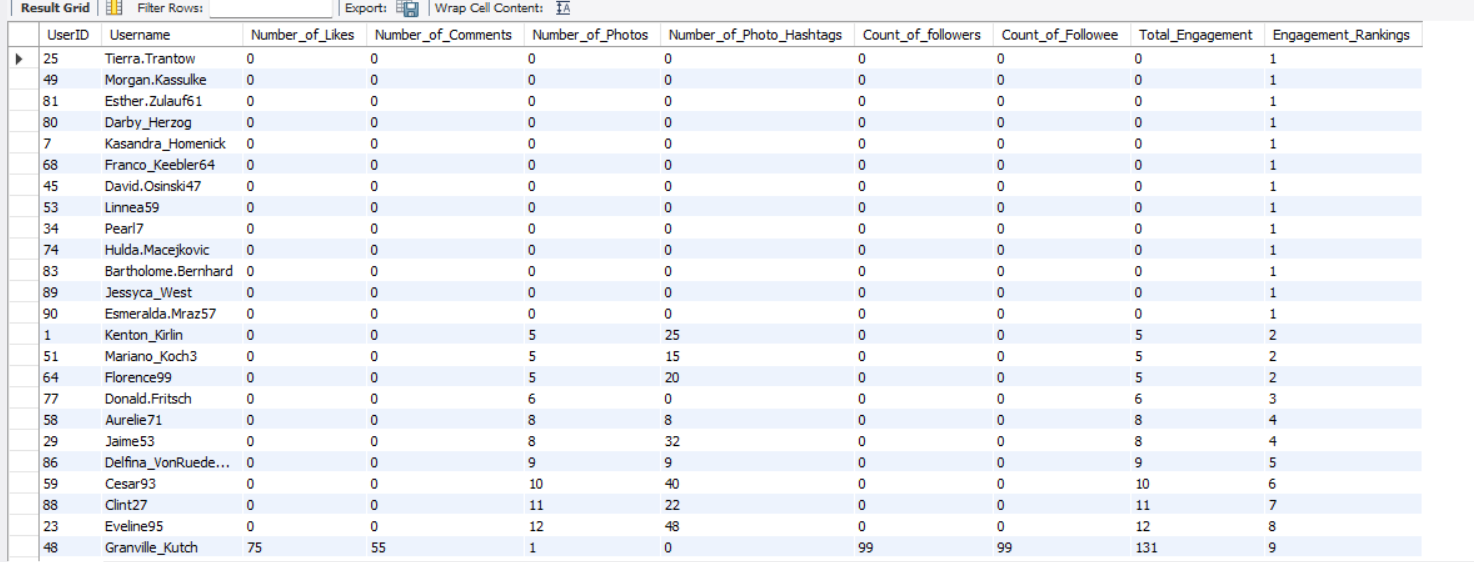


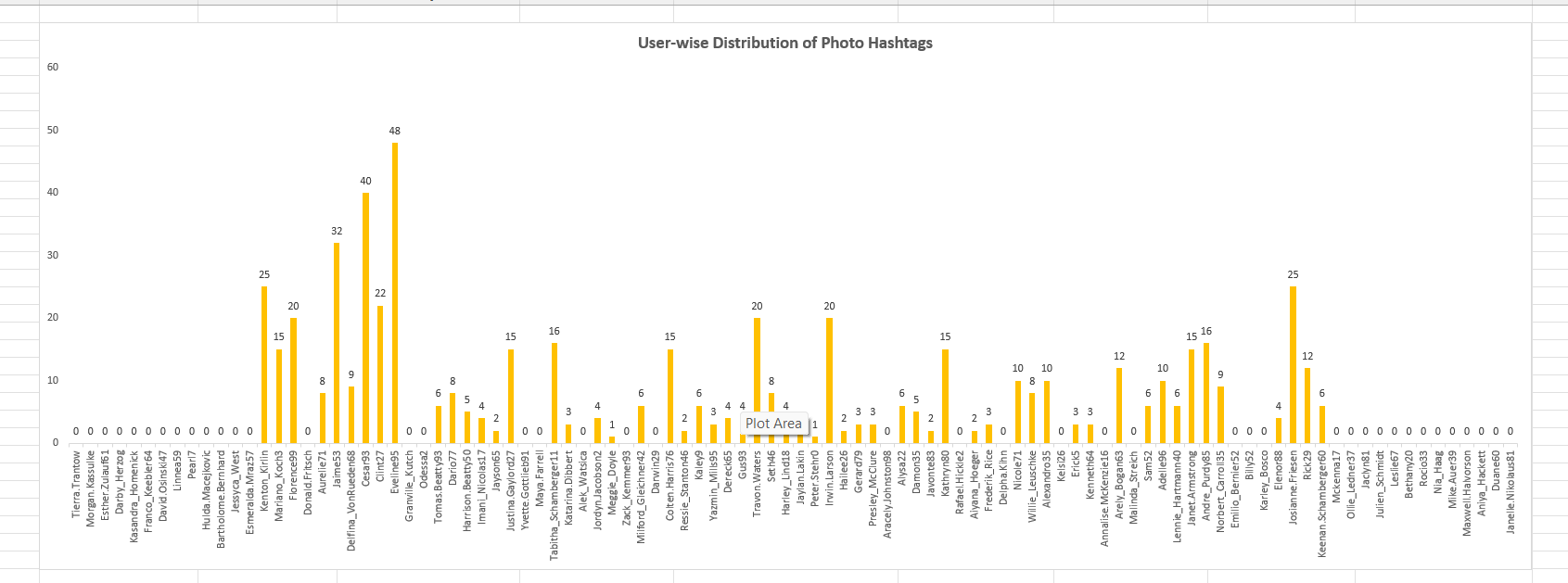
**Strategies to Re-Engage Inactive Users:**

* **Personalized Communication**  
  Send targeted emails, app notifications, or in-platform messages reminding users of what they’re missing (e.g., trending discussions, friends’ activities, or tailored content).
* **Gamified Challenges & Rewards**  
  Introduce streaks, badges, or small incentives (discounts, credits, recognition) for completing simple actions like their “first post after return” or “3 likes in a week.”
* **Exclusive Access**  
  Offer returning users limited-time perks such as access to new features, beta tools, or community events to spark curiosity and involvement.
* **Community Pull**  
  Use social proof by showing them what peers are doing—e.g., “5 of your friends recently posted,” or “Join the trending discussion in your network.”
* **Content Curation & Ease of Start**  
  Simplify their return with personalized recommendations and easy prompts (“Share your thoughts on today’s trending topic”) to reduce friction.

**Conclusion:**

By combining **personalized nudges, gamified incentives, exclusive perks, and community – driven pull**, we can lower barriers to re-entry and motivate inactive users to resume consistent engagement.

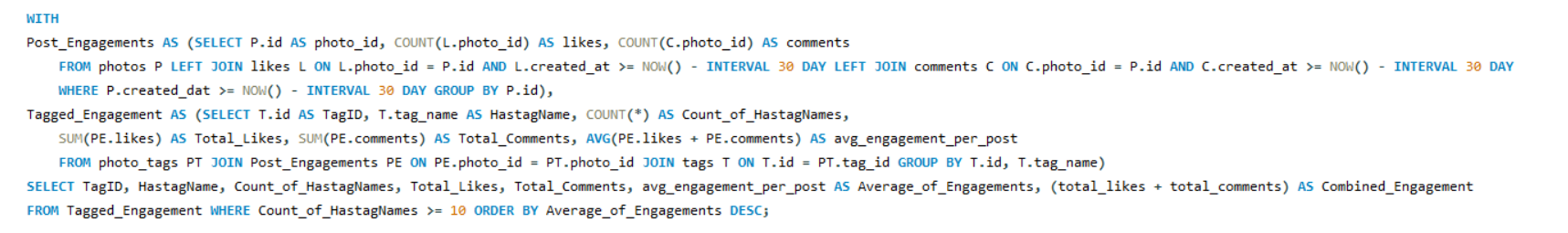




1. Which hashtags or content topics have the highest engagement rates? How can this information guide content strategy and ad campaigns?

Answers –

**QUERY**

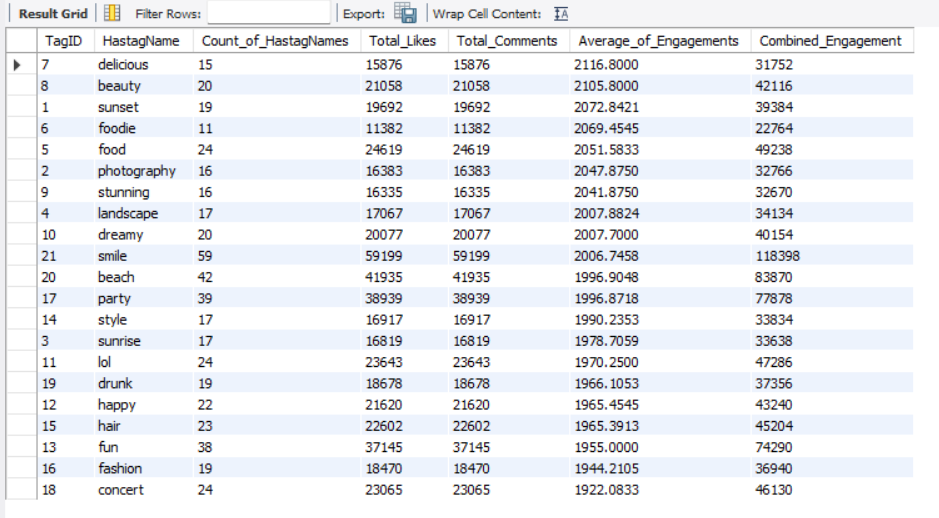


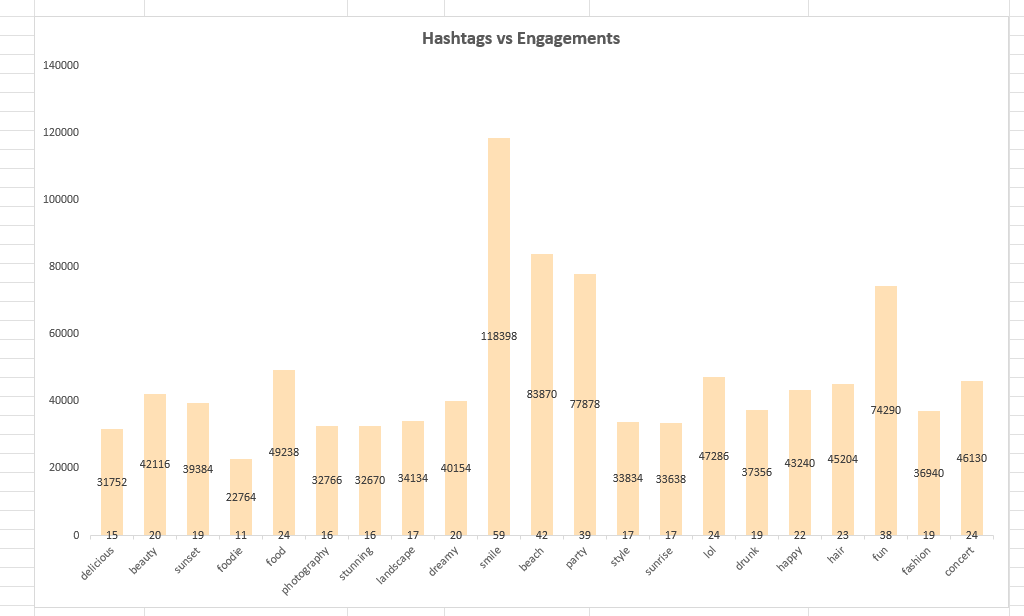
**Strategy based on this output:**

* **Prioritize content** around tags with **high Average\_of\_Engagements** and **healthy Count\_of\_HastagNames** (quality + scale).
* **Targeted ad campaigns:** Use top tags to craft creatives and copy; build audiences from users who engaged with posts carrying these tags; A/B test top-3 tag keywords vs. control.
* **Refine hashtag mix:** Keep the winners, rotate underperformers, and **test adjacent/related tags** to expand reach without losing relevance.
* **Content series:** Spin recurring series around the top tags (weekly themes), reinforcing familiarity and expectation.
* **Influencer collaborations:** Partner with creators known in these high-engagement topics to amplify credibility and conversion.

**Analyzing hashtags:**

* Focus on hashtags at the **top of the list**—these have the **highest Average\_of\_Engagements** (i.e., per-post resonance).
* If your results surface tags like **“smile”, “beach”, “party”, “fun”, “food”** near the top, that means these themes are delivering **more likes + comments per post** right now. Use them as anchors for content and experiments.
* Use Combined\_Engagement (Total\_Likes + Total\_Comments) as a **scale check**—a tag with strong average *and* healthy combined totals is both **high-quality and scalable**.

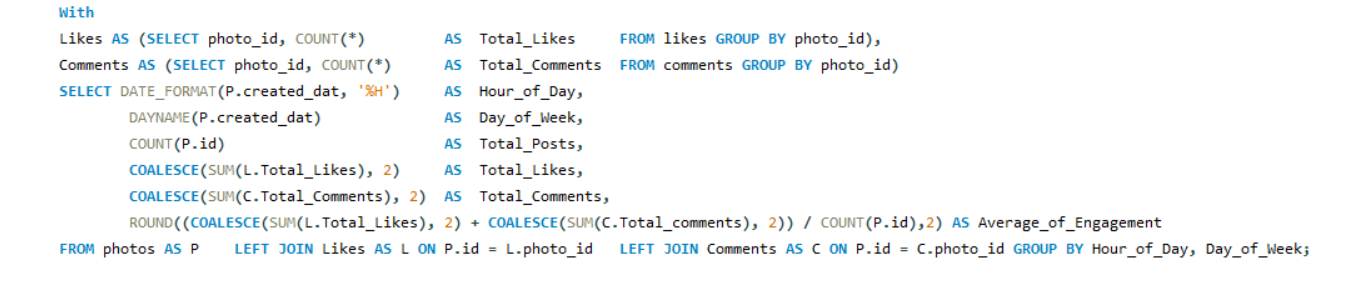




1. Are there any patterns or trends in user engagement based on demographics (age, location, gender) or posting times? How can these insights inform targeted marketing campaigns?

Answers –

**QUERY**



The query calculates engagement based on **posts, likes, and comments** over time.

* It starts by creating two **Common Table Expressions (CTEs)**, Likes and Comments, to count the total likes and comments for each photo.
* The main query then joins these CTEs with the **photos** table to combine engagement data with the **creation time** of each post.
* It extracts the **hour and day of the week** from the post's creation time to group the data.
* The query calculates the total posts, likes, and comments for each time period, and then divides the sum of likes and comments by the number of posts to determine the **average engagement**.

**Patterns and Strategies:**

This query can't analyze demographics like age or gender. However, it is an excellent tool for uncovering **time – based patterns** in user engagement, which are crucial for targeted marketing.

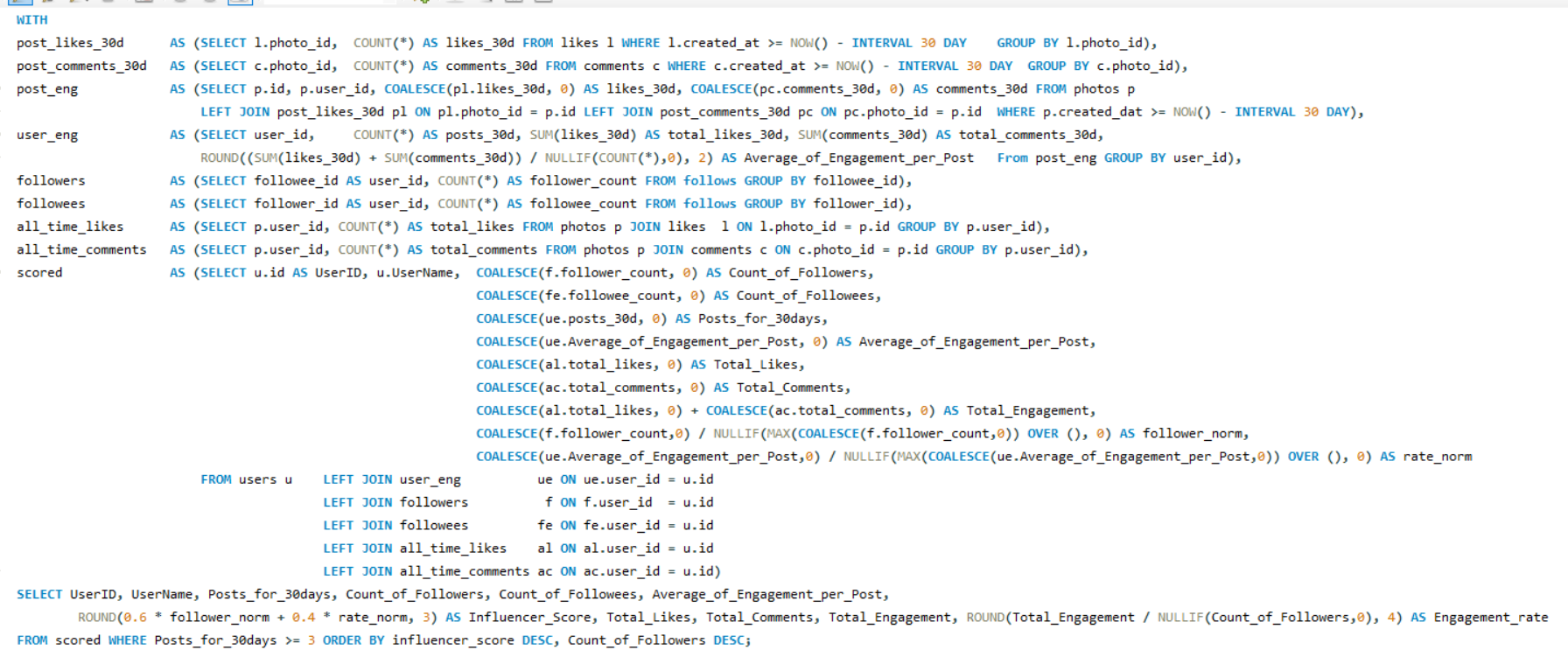
* **Targeted Content**: The insights can also inform your content strategy.
* **Geographical Targeting**: While not directly shown in the query, posting times often correlate with time zones. You can tailor your posting schedule for different regions to match their peak activity hours.
* **Identifiable Patterns**: The query's output will show which hours and days have the highest **Average\_of\_Engagement.** This can reveal that users are most active at specific times, such as late evenings or weekends. This pattern suggests that content posted during these times is more likely to be seen and interacted with.
* **Optimal Posting Time**: You can use these findings to schedule your social media posts for peak engagement times. This will maximize the visibility of your content and can lead to more likes, comments, and shares.



1. Based on follower counts and engagement rates, which users would be ideal candidates for influencer marketing campaigns? How would you approach and collaborate with these influencers?

Answers –

**QUERY**



**Objective:**

Identify users who are strong candidates for influencer campaigns by combining:

* **Audience size** (followers)
* **Recent content quality** (30-day engagement per post)
* **Overall credibility** (all-time likes/comments)

**How does the Query woks:**

1. **Per-post 30-day engagement**: aggregate likes and comments per post within 30 days using post time windows.
2. **Per-user summary (user\_eng)**: compute posts\_30d and avg\_eng\_per\_post.
3. **Follower & followee counts**: followers by followee\_id; followees by follower\_id.
4. **All-time totals**: sum likes and comments across each user’s posts.
5. **Composite scoring**: normalize follower\_count and avg\_eng\_per\_post;

influencer\_score = 0.6 × followers\_norm + 0.4 × rate\_norm.

1. **Selection & ordering**: require posts\_30d ≥ 3; order by influencer\_score DESC, then follower\_count DESC.

**Why these users are ideal for influencer marketing:**

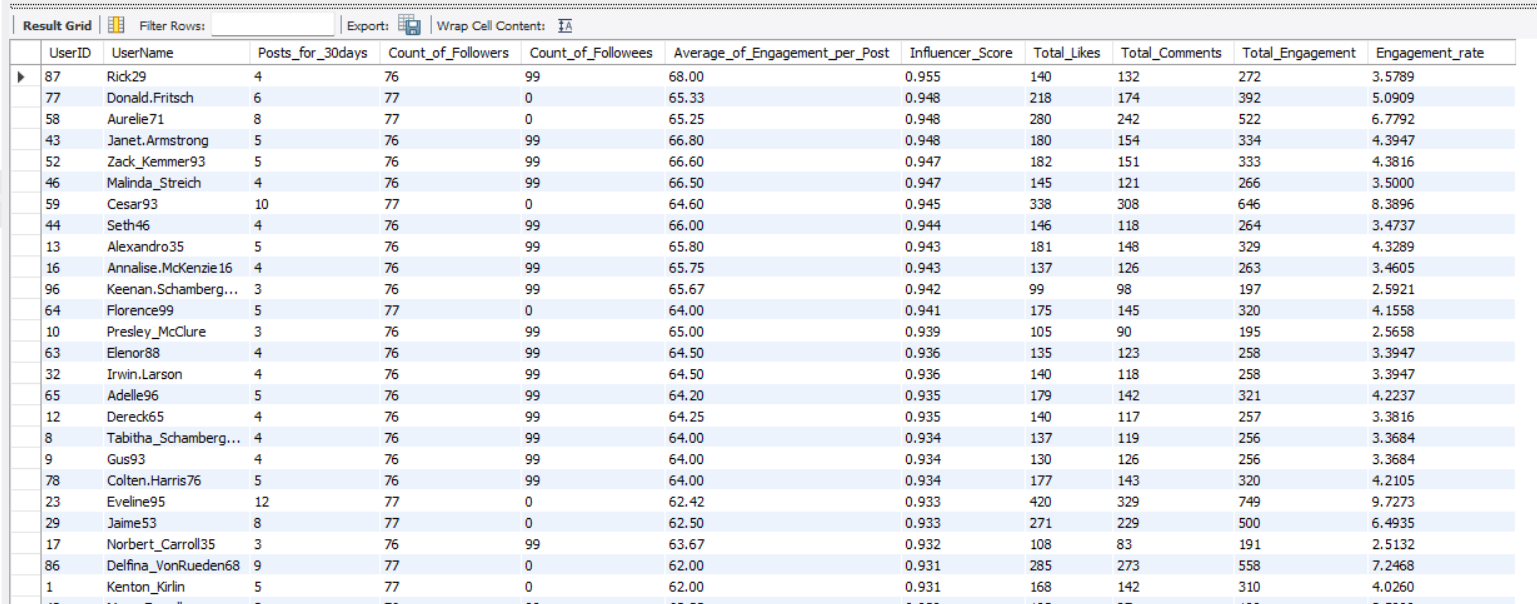
* **High Engagement:** Creators whose posts attract more likes and comments are more likely to have an **active, loyal audience**.
* **Strong Reach:** Higher **follower counts** increase campaign visibility and upper-funnel impact.
* **Balanced Influence:** Using **engagement rate** ensures we select creators who don’t just accumulate followers, but **move their community**.

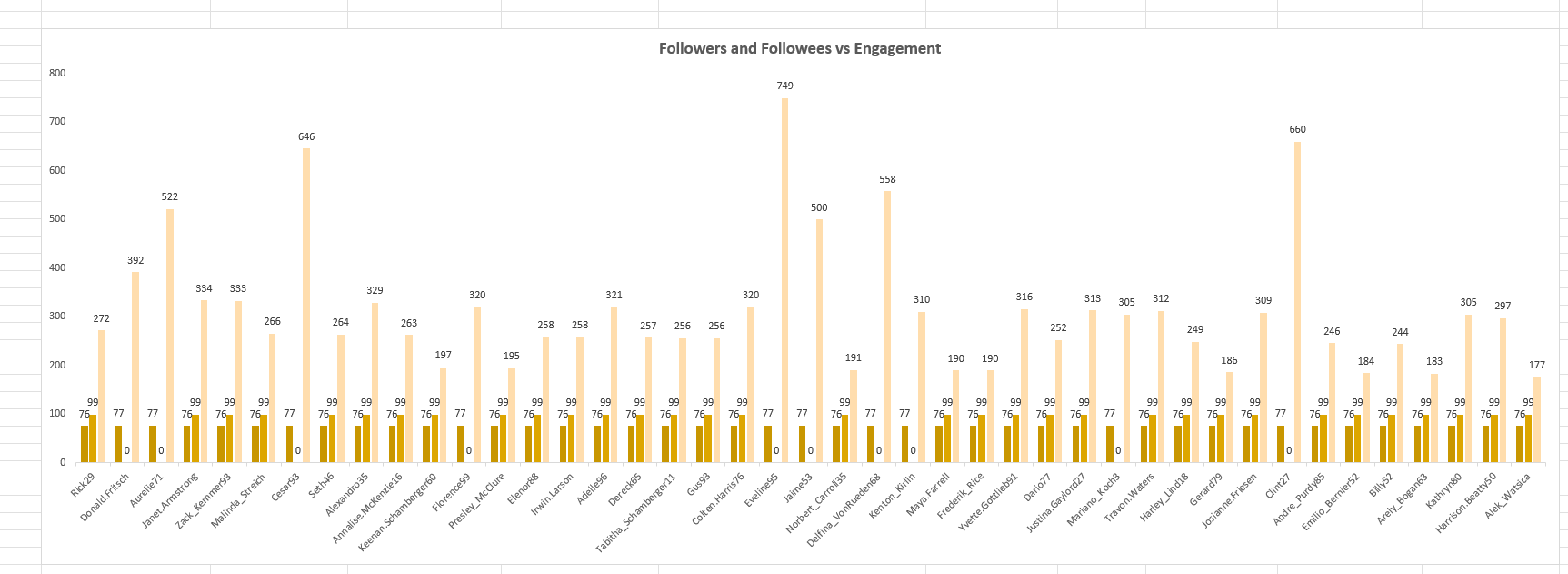
**How the query supports “Top 10” influencer discovery:**

* We **rank creators** by audience size and 30-day engagement quality, with lifetime totals as credibility context.
* We **prioritize the top 10** by **influencer\_score** (and ensure they’ve posted enough recently), then inspect their lifetime stats.

**Summary:**

* We rank creators by **audience size** and **30-day engagement quality**, with **lifetime impact** included.
* The **influencer\_score** highlights the best partners; **micro-creators** with high per-post engagement are efficient for conversions.
* Use the table to **tier, brief, and track** creators; iterate monthly on performance and cost.

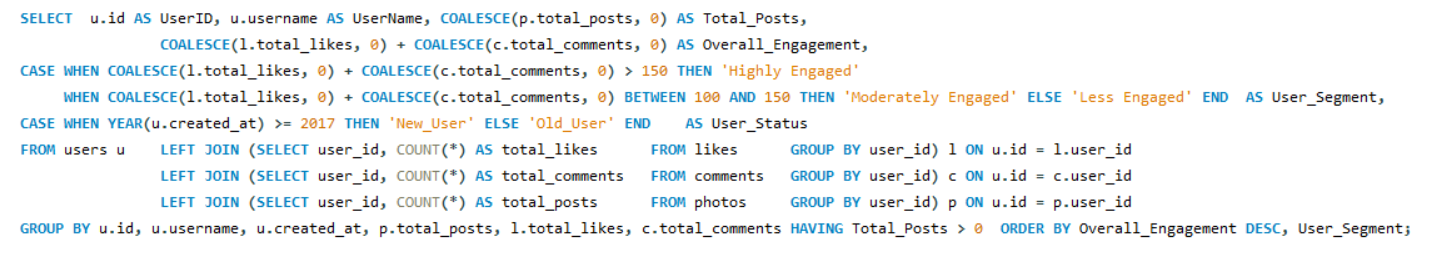




1. Based on user behaviour and engagement data, how would you segment the user base for targeted marketing campaigns or personalized recommendations?

Answers –

**QUERY**

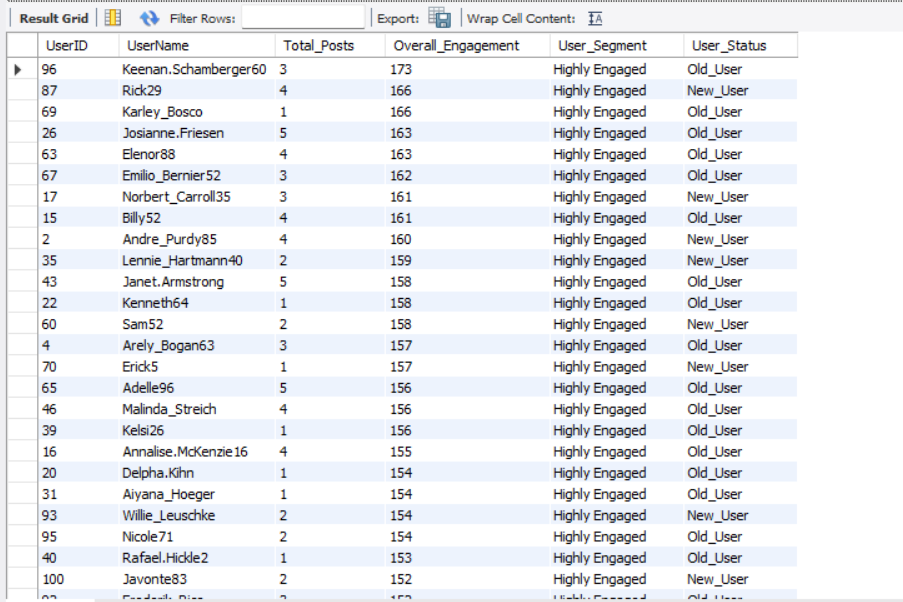


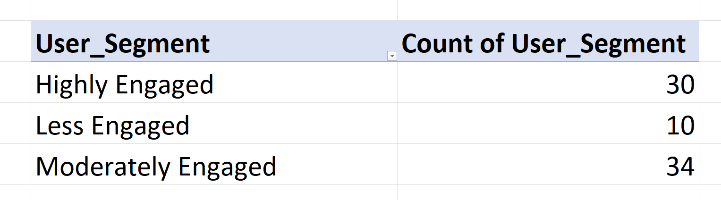
**Measures taken:**

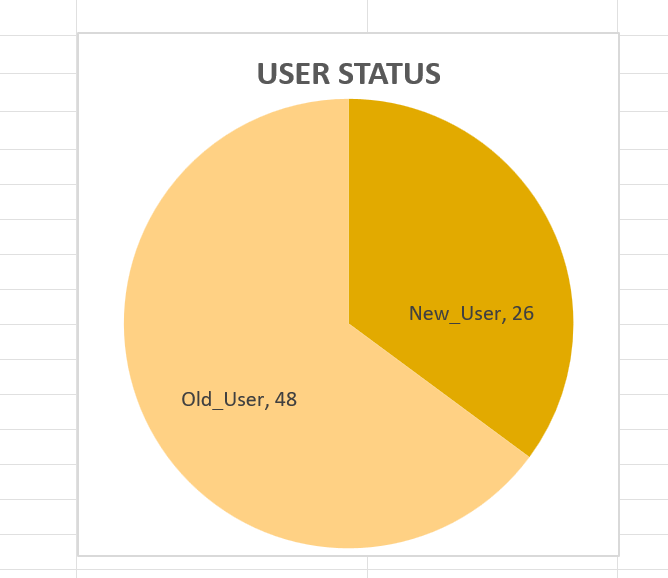
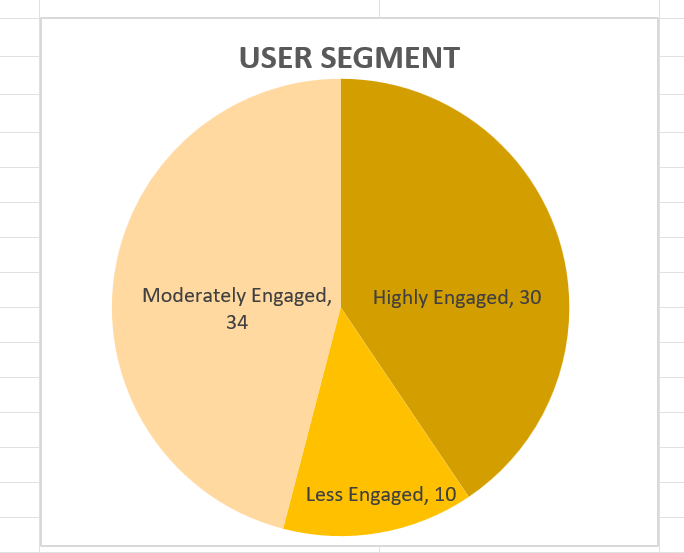
* **Total\_Posts**: number of photos each user has posted.
* **Overall\_Engagement**: **likes + comments** a user has given (aggregated per user and summed).
* **User\_Segment**: activity band based on **Overall\_Engagement**.
* **User\_Status**: cohort based on **account creation year**.

**Methods:**

* Use **inline aggregated subqueries** to compute per-user totals.
* **LEFT JOIN** each subquery to users so users with no likes/comments.
* Apply HAVING Total\_Posts > 0 to **keep only users who posted at least once**.
* Derive **Overall\_Engagement = COALESCE(total\_likes,0) + COALESCE(total\_comments,0)**.



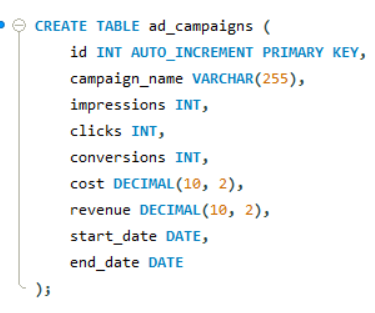
 

1. If data on ad campaigns (impressions, clicks, conversions) is available, how would you measure their effectiveness and optimize future campaigns?

Answers –

**QUERY**



To measure the effectiveness of ad campaigns and optimize future campaigns, one need to evaluate key metrics such as:

**Impression Count:** How many times an ad is shown.

**Click-Through Rate (CTR):** The ratio of clicks to impressions.

**Conversion Rate:** The percentage of users who clicked the ad and performed a desired action (conversion).

**Cost Per Click (CPC):** The average cost paid per click.

**Return on Investment (ROI):** The revenue generated compared to the ad spend.

**Explanation:**

* CTR (Click-Through Rate): Calculated as (clicks / impressions) \* 100. It measures the effectiveness of ads in generating clicks. A higher CTR indicates that the ad is appealing to users.
* Conversion Rate: Calculated as (conversions / clicks) \* 100. It shows how well clicks are turning into actual conversions (such as purchases or sign-ups).
* CPC (Cost Per Click): Calculated as cost / clicks. This tells you how much you are spending for each click. Lower CPC means more efficient spending.
* ROI (Return on Investment): Calculated as ((revenue - cost) / cost) \* 100. This is the key metric to determine whether the campaign is profitable.

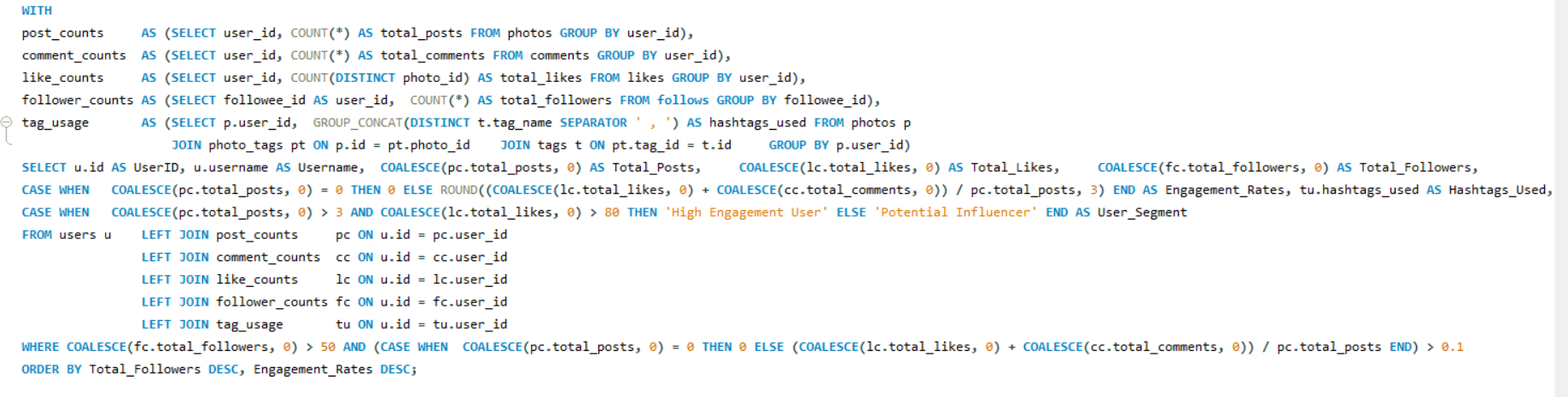
**Optimizing Future Campaigns:**

* Improve CTR: Experiment with different ad creatives, headlines, and targeting options to increase the click-through rate.
* Boost Conversion Rate: Focus on landing page optimization and better call-to- actions (CTAs) to ensure that more clicks convert into actions.
* Lower CPC: Try adjusting bidding strategies, targeting more relevant audiences, or using A/B testing for keywords and ad copy.
* Maximize ROI: Allocate more budget to high-ROI campaigns and reduce spend on underperforming ones.

1. How can you use user activity data to identify potential brand ambassadors or advocates who could help promote Instagram's initiatives or events?

Answers –

**QUERY**

****

**Explanations:**

1. **CTEs / Feature engineering**
   * post\_counts: total posts per user from photos.
   * comment\_counts: total comments authored by user.
   * like\_counts: **count of distinct photos** the user has liked (one like per photo per user).
   * follower\_counts: number of followers each user has (rows where they are the followee).
   * tag\_usage: distinct hashtags used by the user across their posts, concatenated with ' , ' as separator.
2. **Main SELECT**
   * Left-joins all CTEs to users to avoid dropping users with missing activity.
   * Uses COALESCE to turn nulls into zeros for metrics and to protect divisions.
   * Calculates Engagement\_Rates from likes + comments per post (rounded to 3 decimals).
3. **Filter & Order**
   * WHERE keeps only users with **Total\_Followers > 50** and **likes+comments per post > 0.1**.
   * Result is ordered by **Total\_Followers DESC**, then **Engagement\_Rates DESC**.

**Insights:**

1. **Quality + Reach gate:** The followers and engagement filters ensure you only see users who have both audience size and per-post interaction—not just volume of posting.
2. **Two practical segments:**
   * *High Engagement User* = consistent posting (>\*3 posts) and strong like footprint (>\*80 likes), likely to deliver high interaction.
   * *Potential Influencer* = meets the quality + reach gate but doesn’t cross both “high” thresholds—still promising for trials.
3. **Topic cues:** The hashtags list lets you quickly spot content themes (e.g., travel, beauty, events) for better campaign-fit matching.
4. **Leader-first ranking:** Sorting by followers, then engagement, bubbles up creators with larger reach while still rewarding effectiveness.

**Approach:**

1. **Shortlist:** Run the query to get a clean view of qualified users (already filtered and ranked).
2. **Scan themes:** Use Hashtags\_Used to align candidates to campaign topics.
3. **Validate segments:** Sanity-check that “High Engagement User” candidates also have recent activity (consider adding a recency window next).
4. **Pilot outreach:** Start with top rows (high followers + strong ER) for small A/B collaborations; compare outcomes.

**Key Points:**

1. **Likes measure:** COUNT(DISTINCT photo\_id) counts unique photos liked, not raw like events. This is fine when each user can only like a photo once; change to COUNT(photo\_id) if your schema tracks multiple like events per user/photo.
2. **Engagement formula:** Uses **likes + comments** in numerator even though total\_comments is not displayed; this is intentional and gives a fuller interaction picture.
3. **Division safety:** CASE … WHEN total\_posts = 0 THEN 0 avoids divide-by-zero errors.
4. **Hashtag formatting:** GROUP\_CONCAT(DISTINCT t.tag\_name SEPARATOR ' , ') ensures the exact spacing you requested (tag1 , tag2 , tag3).
5. **Engine specifics:** Uses MySQL functions (GROUP\_CONCAT, ROUND) and will run as-is on MySQL/MariaDB. (In PostgreSQL you’d use STRING\_AGG instead.)

**Conclusion:**

* 1. This query creates an actionable, ranked shortlist of creators who combine **meaningful reach** (followers) with **per-post effectiveness** (engagement rate).
  2. By further segmenting users via posting consistency and accumulated likes, it highlights two practical tiers for outreach:
     + **High Engagement Users** for immediate, high-confidence collaborations.
     + **Potential Influencers** for testing and nurturing.
  3. The formatted hashtag list adds quick thematic context, streamlining campaign fit decisions.



1. How would you approach this problem, if the objective and subjective questions weren't given?

Answers –

The Objective and Subjective questions were very helpful as they gave me structure and clarity in understanding the problem. However, if such guiding questions were not given, I would still approach the analysis in a **methodical and systematic manner** by following a step-by-step process.

1. **Understanding and Exploring the Data**

My next step would be to thoroughly understand the database schema, its tables (users, photos, comments, likes, follows, tags, photo\_tags), and their relationships. By exploring the data, I would try to understand distributions, data types, and any early patterns while also noting possible issues such as missing or duplicate values.

1. **Data Cleaning and Validation**

Before drawing any insights, I would clean the dataset by checking for null values, duplicates, and inconsistencies. This would help me ensure that the analysis is built on accurate and reliable data.

1. **Identifying Key Metrics**

Once the data is ready, I would define my **core metrics** across three main dimensions:

* **Engagement Metrics:** Total posts per user, average likes and comments per photo, tags per post, and follower/following counts.
* **Retention Metrics:** Active vs. inactive users, frequency of posting/engaging, and consistency over weekly, monthly, and yearly periods.
* **Acquisition Metrics:** Growth of new users, registration trends, and how engaged new users are compared to older ones.

1. **Analyzing Trends and Deriving Insights**

I would then use SQL concepts like joins, aggregations, and CTEs to test my hypotheses and pull out meaningful insights. For example:

* Identifying the top hashtags based on average likes.
* Mapping correlations between posting frequency and retention.

1. **Data–Driven Solutions**

From the results, I would design practical solutions, such as:

* **Targeted Content Strategy:** Promoting high-performing content types and hashtags.
* **Acquisition Channels:** Suggesting marketing strategies on platforms where potential users are most active.

1. **Insights into the Current Scenario**

After completing the analysis, I would compile all findings into a report. This report would highlight the current scenario, such as identifying potential influencers, highlighting the most effective hashtags, and pointing out where engagement is dropping.

1. **Segmentation and Targeted Marketing**

I would also segment users into groups like *active or inactive* users. Based on this segmentation, I would:

* Identify **brand ambassadors** among highly engaged users.

1. **Reporting and Visualization**

Finally, I would present all of my analysis through dashboards, pivot charts, and a detailed written report. This would summarize the current performance and also provide clear recommendations and strategies for improving engagement, retention, and acquisition.

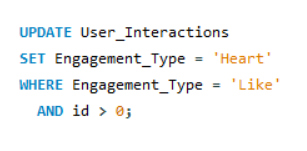
1. **In summary:**

Even without predefined Objective and Subjective questions, I would still carry out a structured analysis by first identifying the problem, then exploring and cleaning the data, defining key metrics, analyzing patterns with SQL, and finally compiling insights into actionable recommendations. This way, the analysis would remain systematic, data-driven, and aligned with business goals.

1. Assuming there's a "User\_Interactions" table tracking user engagements, how can you update the "Engagement\_Type" column to change all instances of "Like" to "Heart" to align with Instagram's terminology?

Answers –

**QUERY**



**Explanation:**

* **UPDATE User\_Interactions**: This specifies the table you want to update, which in this case is "User\_Interactions".
* **SET Engagement\_Type = 'Heart'**: This changes the value of the "Engagement\_Type" column from "Like" to "Heart".
* **WHERE Engagement\_Type = 'Like' AND id > 0**: This condition ensures that only rows where "Engagement\_Type" is "Like" are updated. The additional id > 0 check is used because MySQL Workbench runs in **safe update mode**, which requires that the WHERE clause reference a key column (like the primary key id). Adding this condition satisfies safe update mode while still updating all intended rows.

**Conclusion**

By using the adjusted query with the condition WHERE Engagement\_Type = 'Like' AND id > 0, all instances of "Like" in the Engagement\_Type column are safely updated to "Heart". After execution, the table will consistently reflect "Heart" as the engagement type, maintaining clarity and standardization across user interaction data.

**Conclusion**

In conclusion, the *Social Media Project (META)* successfully demonstrates how **SQL-driven analysis** can uncover key patterns in:

* user behaviour,
* content interaction, and
* engagement performance.

Through systematic exploration of the database, the project identified

* **loyal users,**
* **inactive users,**
* **top hashtags,**
* **influencer candidates, and**
* **overall engagement trends**.

These findings are supported by well-structured SQL queries and visual outputs like charts and pivot analyses.

From a business perspective, the insights provide a solid foundation for **targeted marketing, influencer collaboration, personalized recommendations, and ad campaign optimization**.

By aligning data analysis with platform strategy, the project shows how social media companies can improve **user satisfaction, retention, and monetization**.

Overall, this project highlights the value of combining **data science techniques with business thinking**. It illustrates how data can move beyond numbers and transform into strategies that drive **growth, engagement, and long-term success** for social media platforms.

Thank you for showing your interest in my project and overviewing my project.

Hope I’m able to justify all the answers asked in the project.

LAKSHYA HASIJA