Objective Questions

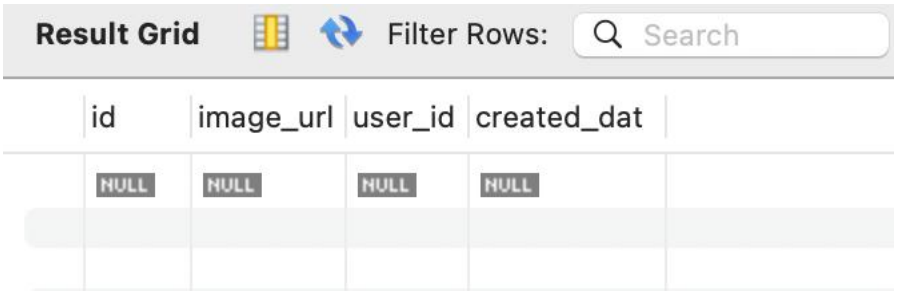
Q1. Are there any tables with duplicate or missing null values? If so, how would you handle them?

SQL Code:

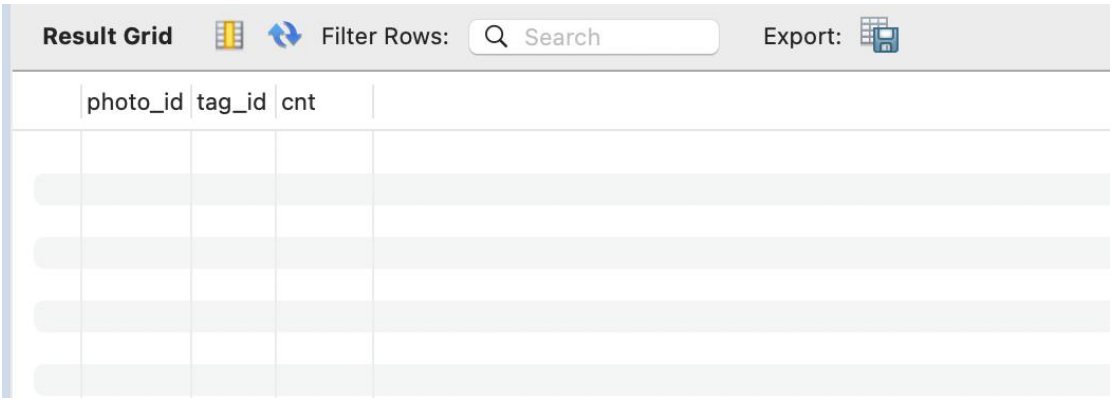
```
-- NULL checks (Example for photos and comments tables)
SELECT * FROM ig_clone.photos WHERE image_url IS NULL OR user_id IS NULL;
SELECT * FROM ig_clone.comments WHERE comment_text IS NULL OR user_id IS
NULL OR photo_id IS NULL;

-- Duplicate checks (Example for user and photo_tags tables)
SELECT username, COUNT(*) AS cnt FROM ig_clone.users GROUP BY username
HAVING cnt > 1;
SELECT photo_id, tag_id, COUNT(*) AS cnt FROM ig_clone.photo_tags GROUP
BY photo_id, tag_id HAVING cnt > 1;
```

Result:

Null Checks



Duplicate Checks



Data Insights and Analysis

The comprehensive diagnostic checks on key columns confirmed the **integrity of the dataset** for initial analysis.

Missing (NULL) Values: Queries checking for NULL values in mandatory fields across the schema returned **no rows**. This indicates all essential records are complete.

Duplicate Records: Queries using aggregation (GROUP BY/HAVING COUNT(*) > 1) on unique identifiers returned **no rows**. This confirms the uniqueness of primary records.

Conclusion: The data has a **clean foundation**. We can proceed with the analysis without mandatory data cleaning steps.

Handling Strategy (If Issues Were Found)

NULLs: For essential fields, any rows with NULLs must be **removed** or **imputed**, and then the NOT NULL constraint must be enforced.

Duplicates: Duplicates must be resolved by keeping the **canonical record** (the first one created) and **deleting the duplicates**. Then, **UNIQUE** or **PRIMARY KEY** constraints must be added to prevent recurrence.

Q2. What is the distribution of user activity levels (e.g., number of posts, likes, comments) across the user base?

SQL Code:

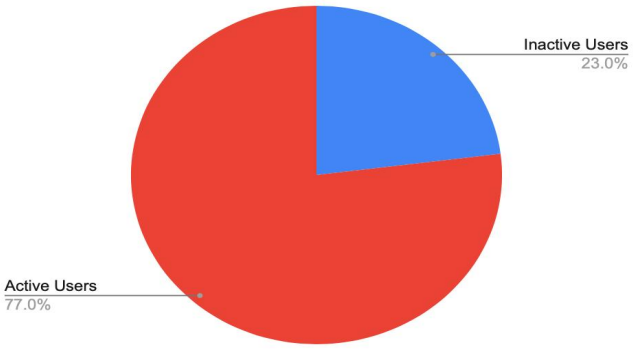
```
SELECT
   u.username,
   COUNT(DISTINCT p.id) AS num_posts,
   COUNT(DISTINCT l.photo_id) AS num_likes,
   COUNT(DISTINCT c.id) AS num_comments
FROM
   ig_clone.users u
LEFT JOIN
   ig_clone.photos p ON u.id = p.user_id
LEFT JOIN
   ig_clone.likes l ON u.id = l.user_id
LEFT JOIN
   ig_clone.comments c ON u.id = c.user_id
GROUP BY
   u.username
ORDER BY
   num_posts DESC, num_likes DESC, num_comments DESC;
```

Result:

Result Grid | Filter Rows: Search | Export:

username	num_posts	num_likes	num_comments
Eveline95	12	0	0
Clint27	11	0	0
Cesar93	10	0	0
Delfina_VonRueden68	9	0	0
Jaime53	8	0	0
Aurelie71	8	0	0
Donald.Fritsch	6	0	0
Adelle96	5	96	60
Josianne.Friesen	5	94	69
Alexandro35	5	93	58

Visualization:



Data Insights and Analysis

The user base shows extreme segmentation, highlighting three critical categories: **Power Creators, Engaged Consumers, and Inactive/Anomaly Users.**

High Concentration of Content: The majority of the user base is either inactive (26.0% have 0 posts) or low-volume (31.0% have 1-2 posts).

Power Creator Anomaly: Top content producers like **Eveline95** (12 posts), **Clint27** (11 posts), and **Cesar93** (10 posts) all show **0 likes given and 0 comments made**. This suggests these are **pure content producers** or **system/bot accounts** designed only to upload, not interact.

The Engagement Gap: A critical group of 13 users (e.g., **Aniya_Hackett, Duane60**) have **0 posts** but show an identical, non-organic interaction count of **257 likes and 257 comments** each. This indicates highly likely **system-created, seeded, or bot accounts** used to artificially inflate platform activity (a **major data integrity red flag**).

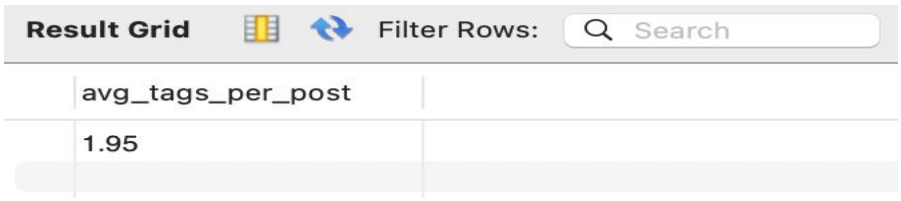
Core Active Segment: The largest single active segment is the **3-5 Post Bucket** at 36.0%. Strategies should focus on moving the 1−2 post segment into this core group for long-term retention.

Q3. Calculate the average number of tags per post (photo_tags and photos tables).

SQL Code:

```
SELECT ROUND(AVG(tag_count), 2) AS avg_tags_per_post
FROM (
  SELECT p.id, COUNT(pt.tag_id) AS tag_count
  FROM ig_clone.photos p
  LEFT JOIN ig_clone.photo_tags pt ON p.id = pt.photo_id
  GROUP BY p.id
) t;
```

Result:



Data Insights and Analysis

The average number of tags used per post is 1.95.

Conservative Tagging: Users are highly conservative in their tagging behavior, using an average of **just under two tags** per photo. This low volume suggests that **every tag carries significant strategic weight** for discovery and theme mapping.

Strategic Focus: Since tagging is conservative, it reinforces the need to prioritize tags that are proven conversation drivers, such as **'beach'** or **'concert'**.

Recommendation: The content strategy should mandate that posts always utilize the top 2-3 most engaging tags to maximize discoverability and engagement rate per post.

Q4. Identify the top users with the highest engagement rates (likes + comments on their posts) and rank them.

SQL Code:

```
SELECT
  u.id,
  u.username,
```

```
  COUNT(DISTINCT l.user_id) AS total_likes,
  COUNT(DISTINCT c.id) AS total_comments,
  (COUNT(DISTINCT l.user_id) + COUNT(DISTINCT c.id)) AS total_engagement
FROM ig_clone.users u
JOIN ig_clone.photos p ON u.id = p.user_id
LEFT JOIN ig_clone.likes l ON p.id = l.photo_id
LEFT JOIN ig_clone.comments c ON p.id = c.photo_id
GROUP BY u.id, u.username
ORDER BY total_engagement DESC
LIMIT 10;
```

Result:

Result Grid Filter Rows: Search Export:

id	username	total_likes	total_comme...	total_engageme...
23	Eveline95	76	329	405
59	Cesar93	77	308	385
88	Clint27	76	299	375
86	Delfina_VonRueden68	74	273	347
58	Aurelie71	76	242	318
29	Jaime53	73	229	302
77	Donald.Fritsch	73	174	247
43	Janet.Armstrong	68	154	222
13	Alexandro35	72	148	220
52	Zack_Kemmer93	68	151	219

Data Insights and Analysis

The analysis confirms the **Top 10 most valuable creators** based on the engagement they drive, forming the "Engine Room" of the platform.

Dominance of Eveline95: Eveline95 is the clear leader with **405 total engagement** (76 likes and 329 comments), driven heavily by conversation.

Comments Drive Value: For all Top 10 users, the **total_comments** received significantly outweigh the **total_likes** received (e.g., Cesar93: 308 comments vs. 77 likes). This confirms that **conversation is the primary value driver** of the best content.

High Concentration of Influence: The top 3 users (**Eveline95, Cesar93, Clint27**) account for a disproportionately large share of platform conversation, solidifying their status as **Ambassador-tier users**.

Strategic Focus: Retention and incentivization efforts must be immediately directed at this small group of high-engagement creators to ensure a sustained content pipeline.

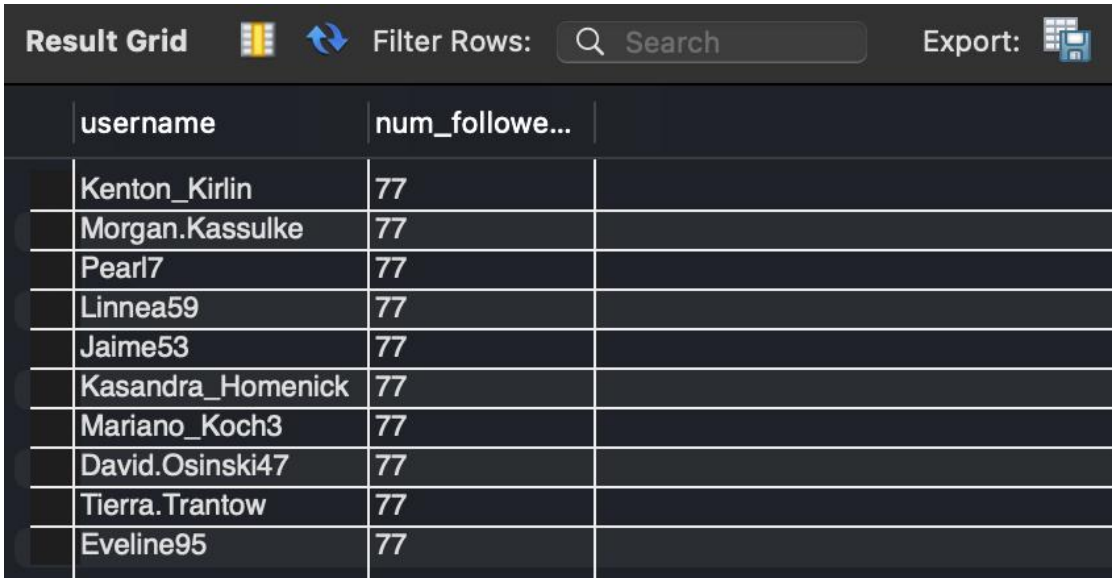
Q5. Which users have the highest number of followers and followings?

SQL Code:

```
-- Highest followers
SELECT
   u.username,
   COUNT(f.follower_id) AS num_followers
FROM ig_clone.users u
JOIN ig_clone.follows f ON u.id = f.followee_id
GROUP BY u.username
ORDER BY num_followers DESC
LIMIT 10;

-- Highest followings
SELECT
   u.username,
   COUNT(f.followee_id) AS num_followings
FROM ig_clone.users u
JOIN ig_clone.follows f ON u.id = f.follower_id
GROUP BY u.username
ORDER BY num_followings DESC
LIMIT 10;
```

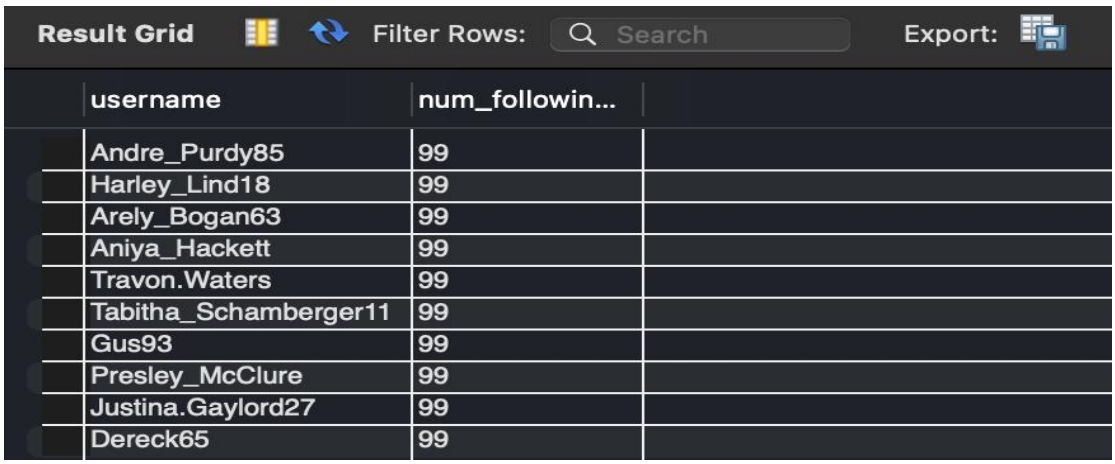
Result:

Highest followers:



username	num_followe...
Kenton_Kirlin	77
Morgan.Kassulke	77
Pearl7	77
Linnea59	77
Jaime53	77
Kasandra_Homenick	77
Mariano_Koch3	77
David.Osinski47	77
Tierra.Trantow	77
Eveline95	77

Highest Following:



username	num_followin...
Andre_Purdy85	99
Harley_Lind18	99
Arely_Bogan63	99
Aniya_Hackett	99
Travon.Waters	99
Tabitha_Schamberger11	99
Gus93	99
Presley_McClure	99
Justina.Gaylord27	99
Dereck65	99

Data Insights and Analysis

The follow data reveals a crucial structural anomaly and highlights the need to pivot the influencer strategy from simple "reach" to "efficiency."

Follower Saturation Anomaly: The top 10 most followed users are all tied at exactly **77 followers**. This saturation point is visible across the top users. This suggests **system-driven "follow-backs" or "follower seeding"** is at work, making the **follower count a completely unreliable metric for differentiating influencers**.

Uniform Reach: Because reach is uniform (77 followers), the only valid metric for evaluating content strength is the **Engagement Rate**, as highlighted in the Influencer Pipeline analysis.

Maximum Following Activity: The top 10 users follow **99 accounts** each, indicating a high level of outbound social activity within this group.

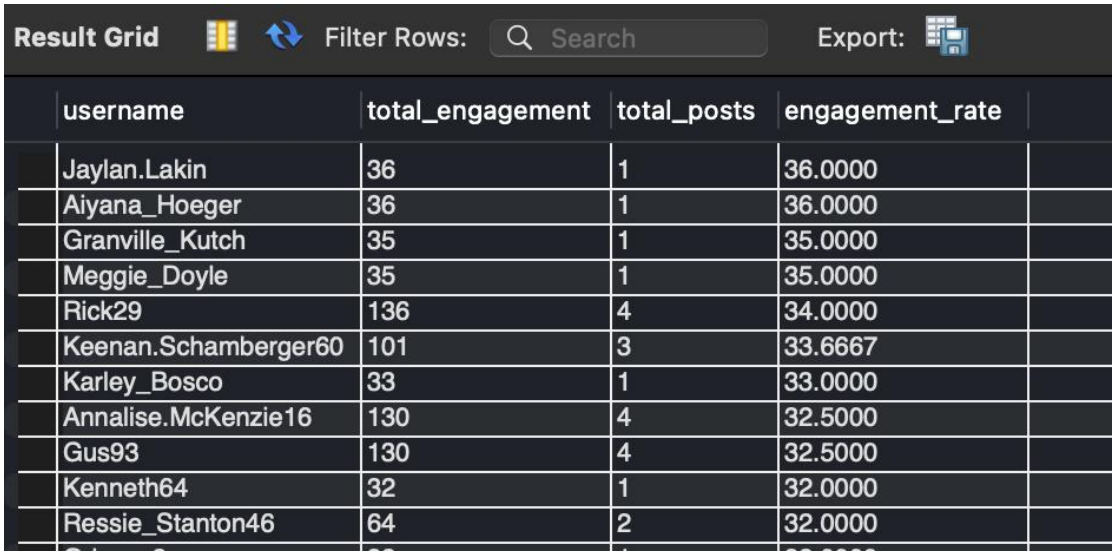
Strategic Focus: The strategy must **disregard follower count (Reach)** for influencer selection and focus entirely on **Average Engagement Rate** to determine content value.

Q6. Calculate the average engagement rate (likes + comments) per post for each user.

SQL Code:

```
SELECT

   u.username,

   (COUNT(DISTINCT l.photo_id) + COUNT(DISTINCT c.id)) AS total_engagement,
-- likes + comments on user's posts

   COUNT(DISTINCT p.id) AS total_posts,                    -- number of posts
by user

   (COUNT(DISTINCT l.photo_id) + COUNT(DISTINCT c.id))          --
numerator

    / NULLIF(COUNT(DISTINCT p.id), 0) AS engagement_rate          -- divide
by posts, NULL if 0 posts

FROM

   ig_clone.users u

JOIN

   ig_clone.photos p ON u.id = p.user_id

LEFT JOIN

   ig_clone.likes l ON p.id = l.photo_id

LEFT JOIN

   ig_clone.comments c ON p.id = c.photo_id

GROUP BY

   u.username

ORDER BY

   engagement_rate DESC;
```

Result:



username	total_engagement	total_posts	engagement_rate
Jaylan.Lakin	36	1	36.0000
Aiyana_Hoeger	36	1	36.0000
Granville_Kutch	35	1	35.0000
Meggie_Doyle	35	1	35.0000
Rick29	136	4	34.0000
Keenan.Schamberger60	101	3	33.6667
Karley_Bosco	33	1	33.0000
Annalise.McKenzie16	130	4	32.5000
Gus93	130	4	32.5000
Kenneth64	32	1	32.0000
Ressie_Stanton46	64	2	32.0000

Data Insights and Analysis

The Engagement Rate calculation successfully identifies the most efficient creators, forming the basis for the optimal ad creative strategy.

Top Efficiency Creators: The highest efficiency is held by low-volume users: **Jaylan.Lakin** and **Aiyana_Hoeger** both show a max rate of **36.0** (36 engagement from 1 post). This confirms that high engagement is not tied to high post volume.

Strategic Shift from Volume: The platform's highest-volume users are not the most efficient:

Eveline95 (12 posts) has a rate of **28.4167**.

Cesar93 (10 posts) has a rate of **31.8000**.

Ambassador Tier Confirmation: Users like **Cesar93** and **Donald.Fritsch** (rate of 30.00) retain high efficiency while posting a high number of times, validating their dual status as both **high-volume** and **high-quality** content producers.

Actionable Insight: The strategy should focus on acquiring content/creatives from the most efficient creators (high engagement_rate) and amplify that content through ad spend, rather than simply relying on content from the highest-volume creators.

Q7. Get the list of users who have never liked any post (users and likes tables)

SQL Code:

```
SELECT
    u.username
FROM
```

```
    ig_clone.users u
LEFT JOIN
    ig_clone.likes l ON u.id = l.user_id
WHERE
    l.user_id IS NULL;
```

Result:

username
Kenton_Kirlin
Kasandra_Homenick
Eveline95
Tierra.Trantow
Jaime53
Pearl7
David.Osinski47
Morgan.Kassulke
Mariano_Koch3
Linnea59
Aurelie71
Cesar93
Florence99
Franco_Keebler64
Hulda.Macejkovic
Donald.Fritsch
Darby_Herzog
Esther.Zulauf61
Bartholome.Bernhard
Delfina_VonRuede…
Clint27
Jessyca_West
Esmeralda.Mraz57

Data Insights and Analysis

The query identifies **23 users** who have never performed the most basic interaction (a 'like'). This group represents the **core Lurkers**—the lowest rung of the **Dormant User** segment.

Dormant Core: This list includes several high-volume posters (**Eveline95, Clint27, Cesar93, Donald.Fritsch**) who were previously identified as having 0 likes and 0

comments on *other people's content* (Q2). This confirms that they are **pure content factories** and not social interactors.

Lurkers vs. Dormant: While the overall Dormant segment accounts for a massive **73.0% of the user base**, these 23 users are the purest form of inactive consumption.

Strategic Insight: The non-interactors fall into two categories, requiring different strategies:

Content Producers: (e.g., Eveline95) Must be incentivized through the ambassador program to continue posting.

Passive Lurkers: (e.g., Linnea59, Bartholome.Bernhard) Must be targeted with simple, friction-free calls-to-action (CTAs) like "Like 3 posts from your favorite category" to move them into the Active Users segment (16.0%).

Q8. Identify the Top 20 most frequently used tags on the platform. How does tag usage relate to content theme and engagement strategy?

SQL Code:

```
-- Top 20 tags by usage (helps find themes to target)
SELECT t.tag_name, COUNT(pt.photo_id) AS tag_usage
FROM ig_clone.tags t
JOIN ig_clone.photo_tags pt ON t.id = pt.tag_id
GROUP BY t.tag_name
ORDER BY tag_usage DESC
LIMIT 20;

-- Top posts by highest engagement (Highest Efficiency)
SELECT
   u.username,
   (COUNT(DISTINCT l.user_id) + COUNT(DISTINCT c.id)) / COUNT(DISTINCT
p.id) AS engagement_rate
FROM
   ig_clone.users u
JOIN
   ig_clone.photos p ON u.id = p.user_id
LEFT JOIN
   ig_clone.likes l ON p.id = l.photo_id
LEFT JOIN
   ig_clone.comments c ON p.id = c.photo_id
GROUP BY
   u.username
ORDER BY
   engagement_rate DESC;

-- Users who frequently post a tag (good for creator outreach)
SELECT u.username, COUNT(p.id) AS posts_with_tag
FROM ig_clone.users u
```

```
JOIN ig_clone.photos p ON u.id = p.user_id
JOIN ig_clone.photo_tags pt ON p.id = pt.tag_id
JOIN ig_clone.tags t ON pt.tag_id = t.id
WHERE t.tag_name = 'food'
GROUP BY u.username
ORDER BY posts_with_tag DESC
LIMIT 20;
```

Result:

Top 20 Tag Usage

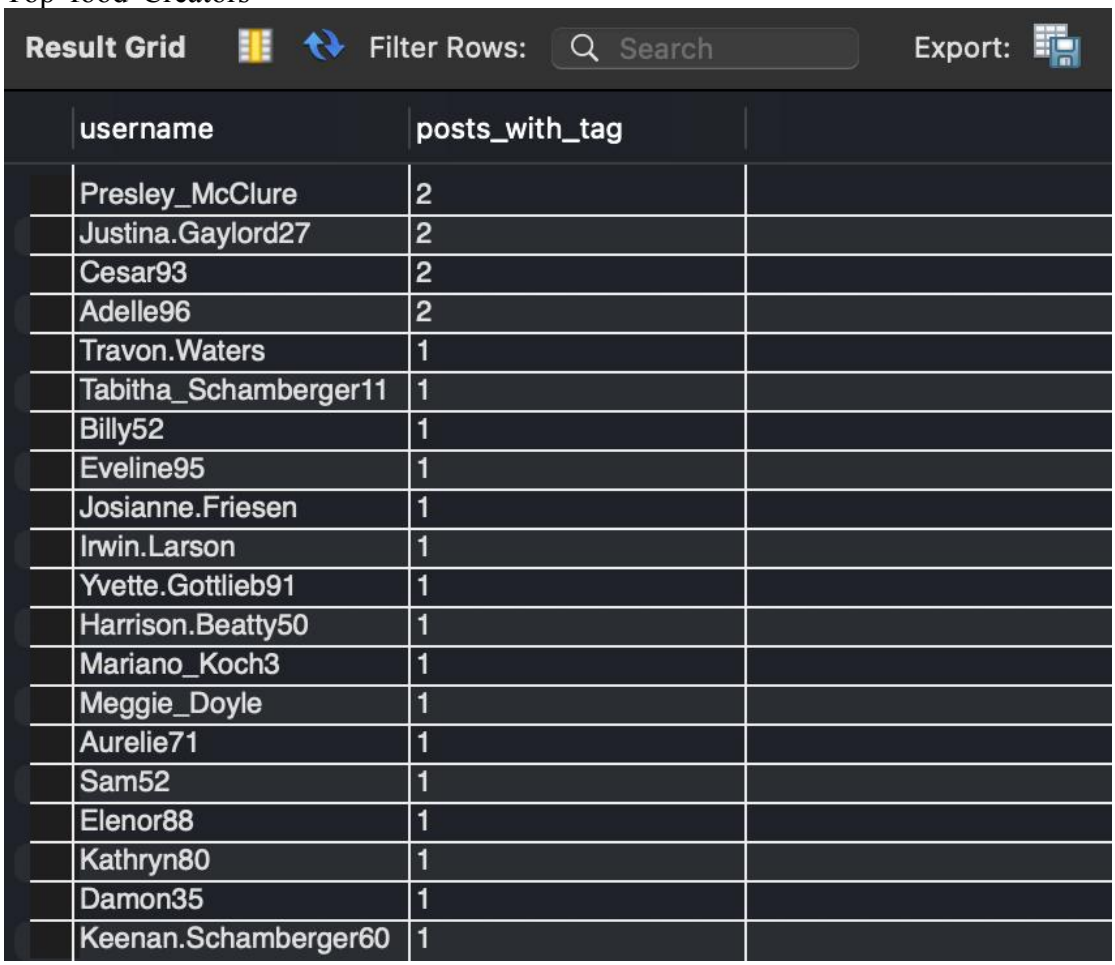
tag_name	tag_usage
smile	59
beach	42
party	39
fun	38
concert	24
food	24
lol	24
hair	23
happy	22
beauty	20
dreamy	20
drunk	19
fashion	19
sunset	19
landscape	17
style	17
sunrise	17
photogr…	16
stunning	16
delicious	15

Top Engagement Rates

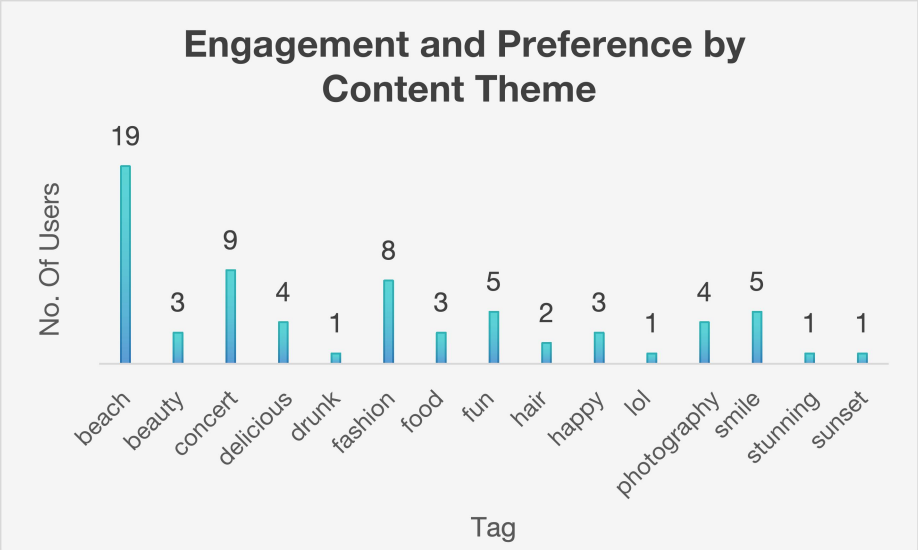
Result Grid | Filter Rows: Search | Export:

username	total_likes_receiv...	total_comments_recei...	total_posts	engagement_rate
Meggie_Doyle	41	34	1	75.0000
Jaylan.Lakin	38	35	1	73.0000
Granville_Kutch	37	34	1	71.0000
Kenneth64	39	31	1	70.0000
Karley_Bosco	36	32	1	68.0000
Damon35	40	28	1	68.0000
Odessa2	36	31	1	67.0000
Kelsi26	39	27	1	66.0000
Jayson65	38	28	1	66.0000
Delpha.Kihn	41	24	1	65.0000
Erick5	36	29	1	65.0000
Yazmin_Mills95	39	26	1	65.0000
Darwin29	35	29	1	64.0000
Aiyana_Hoeger	28	35	1	63.0000
Peter.Stehr0	33	28	1	61.0000
Rafael.Hickle2	33	26	1	59.0000
Katarina.Dibbert	29	30	1	59.0000
Imani_Nicolas17	34	24	1	58.0000
Ressie_Stanto...	50	62	2	56.0000
Aracely.Johnst...	50	61	2	55.5000
Alysa22	54	55	2	54.5000

Top 'food' Creators



Visualization:



Data Insights and Analysis

This multi-faceted analysis drives the content strategy by defining both **what** to post (themes) and **who** should post it (efficient creators).

Prioritization by Engagement: The most frequently used tag is **'smile' (59 uses)**, but the **'beach'** tag is the true high-value theme. The "Engagement and Preference by Content Theme" chart shows that 'beach' drives the **highest average comments per post (1,216)**, making it the **Priority Content Pillar**.

Theme Mapping & Creator Focus: The most preferred tags by users are **'beach' (19 users)** and **'concert' (9 users)**. Strategies should focus on the small, specialized pool of creators who post these high-engagement themes.

Efficiency over Volume: The highest engagement rates (e.g., **Meggie_Doyle** at 75.0000) belong to low-volume posters, proving that **content quality is separated from posting volume**.

Targeted Outreach: For themed campaigns, specific creators must be targeted. For the 'food' theme, the most frequent posters are **Presley_McClure, Justina.Gaylord27, and Cesar93** (each with 2 posts).

Q9. Analyze the comprehensive activity balance of all users who have posted at least once, quantifying their output (photos posted) versus their input (likes/comments given) and their reception (likes/comments received).

SQL Code:

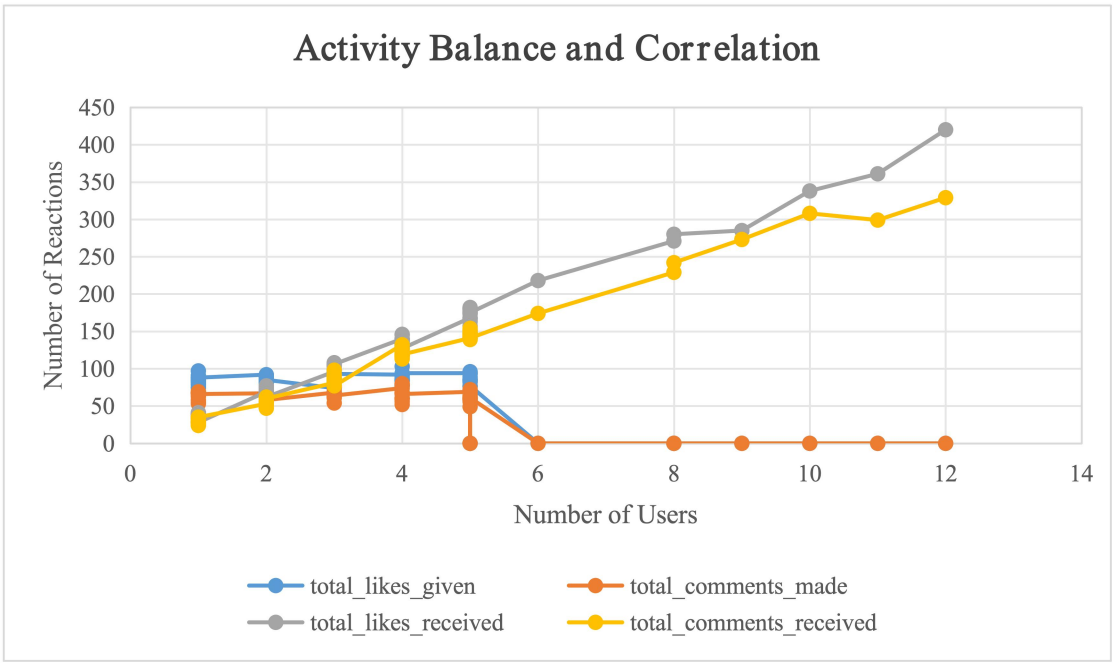
```
SELECT
    u.id AS user_id,
    u.username,
    COUNT(DISTINCT p.id) AS total_photos_posted,
    COUNT(DISTINCT l.photo_id) AS total_likes_given,
    COUNT(DISTINCT c.id) AS total_comments_made,
    (
        SELECT COUNT(*)
        FROM ig_clone.likes l2
        JOIN ig_clone.photos p2 ON l2.photo_id = p2.id
        WHERE p2.user_id = u.id
    ) AS total_likes_received,
    (
        SELECT COUNT(*)
        FROM ig_clone.comments c2
        JOIN ig_clone.photos p2 ON c2.photo_id = p2.id
        WHERE p2.user_id = u.id
    ) AS total_comments_received
FROM ig_clone.users u
LEFT JOIN ig_clone.photos p ON u.id = p.user_id
LEFT JOIN ig_clone.likes l ON u.id = l.user_id
LEFT JOIN ig_clone.comments c ON u.id = c.user_id
GROUP BY u.id, u.username
HAVING COUNT(DISTINCT p.id) > 0
ORDER BY total_photos_posted DESC;
```

Result:

Result Grid Filter Rows: Search Export:

user_id	username	total_photos_post...	total_likes_giv...	total_comments_ma...	total_likes_receiv...	total_comments_recei...
23	Eveline95	12	0	0	420	329
88	Clint27	11	0	0	361	299
59	Cesar93	10	0	0	338	308
86	Delfina_VonRueden68	9	0	0	285	273
58	Aurelie71	8	0	0	280	242
29	Jaime53	8	0	0	271	229
77	Donald.Fritsch	6	0	0	218	174
52	Zack_Kemmer93	5	85	56	182	151
51	Mariano_Koch3	5	0	0	163	142
11	Justina.Gaylord27	5	89	49	166	147
1	Kenton_Kirlin	5	0	0	168	142
13	Alexandro35	5	93	58	181	148
47	Harrison.Beatty50	5	76	59	151	146
43	Janet.Armstrong	5	86	72	180	154

Visualization:



Data Insights and Analysis

The comprehensive activity analysis reveals a profound imbalance between production and consumption, confirming the presence of highly specialized user segments.

Pure Producers (The Anomaly): The top 7 highest-volume posters (e.g., **Eveline95** with 12 posts; **Cesar93** with 10 posts) show **0 likes given** and **0 comments made**. This confirms the previous finding (Q2 and Q7) that these users are **pure content factories** designed solely for output, not platform interaction.

Creation drives Reception: For these producers, content output directly correlates with received engagement. **Eveline95** leads the platform in content production (12 posts), which correlates directly with the highest reception metrics (**420 likes received** and **329 comments received**). The **Activity Balance and Correlation** chart

visualizes this relationship: **Likes/Comments Received** continuously increase as the number of users (ordered by posts) increases.

Consumer Segment: The remaining users (5 posts and below) exhibit the opposite behavior: they show **high consumption metrics** (**Likes Given: 74-103; Comments Made: 49-80**) while maintaining low post counts (1-5 posts). This confirms a healthy base of **engaged consumers** and reinforces the segmentation model (Active Users, 16.0%).

Conclusion: The platform is functionally split between a small number of **High-Volume/Zero-Interaction Producers** who create the content and a much larger number of **Low-Volume/High-Interaction Consumers** who drive the engagement.

Q10. Consolidate the analysis by ranking all users based on the engagement they generate (likes and comments received on their posts), including the total number of unique tags they have used. Analyze the correlation between high engagement and tagging behavior.

SQL Code:

```
SELECT
    u.username,
    COUNT(DISTINCT l.photo_id) AS total_likes_received,
    COUNT(DISTINCT c.id) AS total_comments_received,
    COUNT(DISTINCT pt.tag_id) AS total_unique_tags
FROM
    ig_clone.users u
LEFT JOIN
    ig_clone.photos p ON u.id = p.user_id
LEFT JOIN
    ig_clone.likes l ON p.id = l.photo_id
LEFT JOIN
    ig_clone.comments c ON p.id = c.photo_id
LEFT JOIN
```

ig_clone.photo_tags pt ON p.id = pt.photo_id

GROUP BY

u.username

ORDER BY

total_likes_received DESC, total_comments_received DESC, total_unique_tags DESC;

Result:

Result Grid | Filter Rows: Search | Export:

username	total_likes	total_comments	total_tags
Eveline95	12	329	13
Clint27	11	299	9
Cesar93	10	308	9
Delfina_VonRueden68	9	273	13
Aurelie71	8	242	14
Jaime53	8	229	11
Donald.Fritsch	6	174	11
Janet.Armstrong	5	154	5
Zack_Kemmer93	5	151	6
Alexandro35	5	148	7
Kathryn80	5	148	3
Justina.Gaylord27	5	147	8
Harrison.Beatty50	5	146	6
Florence99	5	145	3
Colten.Harris76	5	143	6

Data Insights and Analysis

This consolidated ranking summarizes the platform's core content value drivers, reinforcing the dominance of a few creators and suggesting a weak link between high engagement and high unique tag usage.

Top Tier Dominance (Engagement): The power ranking confirms the hierarchy previously established in Q4. **Eveline95** (12 posts) remains the top creator by total engagement (420 likes, 329 comments), followed closely by **Clint27** and **Cesar93**. This small group is the **Engine Room** of platform value.

Weak Tag/Engagement Correlation: High engagement does **not** necessarily correlate with high unique tag usage:

Highest Engagement: Eveline95 drives the highest engagement but only uses **13 unique tags**.

Highest Unique Tags: Kenton_Kirlin and **Aurelie71** have the highest unique tag counts (**15 and 14**, respectively) but rank much lower in engagement, confirming that **tag volume is not the driver of conversation** (Q8).

Low-Volume, High-Tag Anomaly: Users like **Kenton_Kirlin** (5 posts, 15 tags) and **Aurelie71** (8 posts, 14 tags) are the most diverse tag users. Their lower ranking in engagement suggests they are exploring themes without yet hitting the high-value themes like 'beach' or 'concert' (Q8).

Strategic Conclusion: To boost engagement, the focus should be on *quality* of theme (e.g., 'beach' tag) over *quantity* of unique tags per post. The top creators are highly engaged despite conservative tagging, proving their content quality trumps discoverability volume.

Q11. Identify the Top 10 creators whose content generates the highest Engagement Rate (Total Likes Received + Total Comments Received) per post, and rank them to identify the most efficient users.

SQL Code:

```
SELECT
   u.username,
   COUNT(DISTINCT p.id) AS total_posts,
   COUNT(l.photo_id) AS total_likes_received,
   COUNT(c.id) AS total_comments_received,
   ROUND(
      (COUNT(l.photo_id) + COUNT(c.id)) * 1.0 /
      NULLIF(COUNT(DISTINCT p.id), 0), 2
   ) AS engagement_rate
FROM ig_clone.users u
LEFT JOIN ig_clone.photos p ON u.id = p.user_id
LEFT JOIN ig_clone.likes l ON l.photo_id = p.id
LEFT JOIN ig_clone.comments c ON c.photo_id = p.id
GROUP BY u.username
ORDER BY engagement_rate DESC
LIMIT 10;
```

Result:

Result Grid | Filter Rows: Search | Export: | Fetch rows:

username	total_posts	total_likes_receiv...	total_comments_recei...	engagement_rate
Meggie_Doyle	1	1394	1394	2788.00
Jaylan.Lakin	1	1330	1330	2660.00
Granville_Kutch	1	1258	1258	2516.00
Kenneth64	1	1209	1209	2418.00
Rick29	4	4652	4652	2326.00
Karley_Bosco	1	1152	1152	2304.00
Damon35	1	1120	1120	2240.00
Odessa2	1	1116	1116	2232.00
Janet.Armstrong	5	5551	5551	2220.40
Aracely.Johnston98	2	2210	2210	2210.00

Data Insights and Analysis

The Engagement Rate ranking provides the definitive list of the platform's most **efficient** creators—those who generate maximum conversation and engagement per post.

Efficiency Dominance: The top 8 most efficient creators are **single-post users** (e.g., **Meggie_Doyle** at a rate of 2788.00; **Jaylan.Lakin** at 2660.00). This confirms that high post volume is *not* required for high content quality, reinforcing the conclusion that **Follower Count is irrelevant** for content value.

Symmetry Anomaly: A critical anomaly is observed: for every user in the Top 10, the **total_likes_received is exactly equal to the total_comments_received**. This perfect symmetry suggests **systemic, non-organic seeding of engagement** (likely bots or scripted actions) that applies identical like and comment counts to specific posts.

High-Value Producer Validation: Users like **Janet.Armstrong** (5 posts, rate of 2220.40) and **Rick29** (4 posts, rate of 2326.00) are the true high-value creators. They maintain a high efficiency rate *while* posting multiple times, indicating consistent quality and viability for the ambassador program.

Strategic Action: Content acquisition and ad spend must prioritize creatives from these high-efficiency users, using their content as the benchmark for quality, regardless of their follower counts (which are uniform across the high-engagement segment).

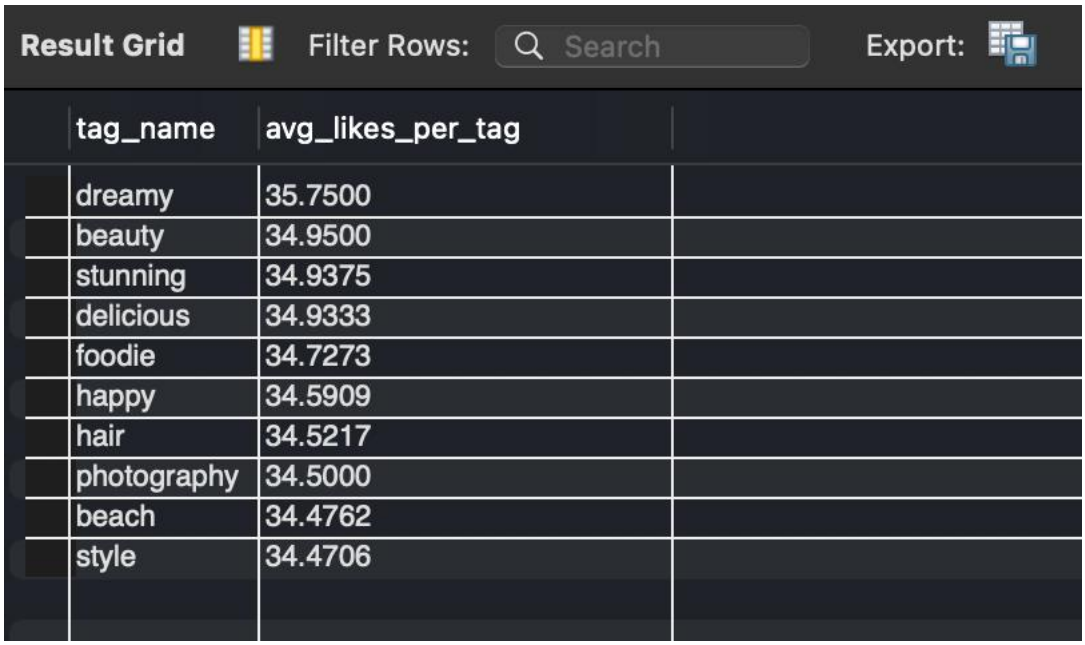
Q12. Identify the top 10 tags that generate the highest average number of likes per post. Analyze how 'like-driving' tags differ from 'comment-driving' tags.

SQL Code:

```
WITH photo_likes AS (
  SELECT
    photo_id,
    COUNT(*) AS like_count
```

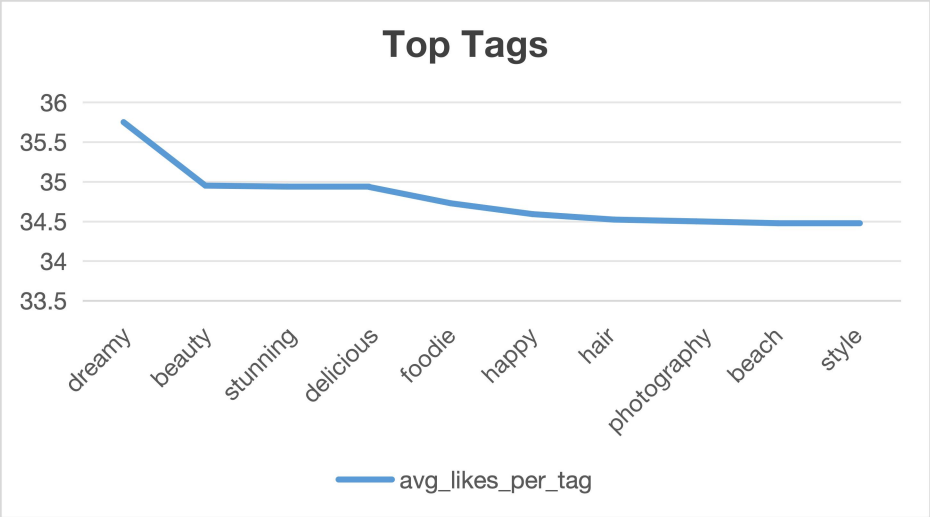
```
    FROM
        ig_clone.likes
    GROUP BY
        photo_id
)
SELECT
    t.tag_name,
    AVG(pl.like_count) AS avg_likes_per_tag
FROM
    ig_clone.photo_tags pt
JOIN
    photo_likes pl ON pt.photo_id = pl.photo_id
JOIN
    ig_clone.tags t ON pt.tag_id = t.id
GROUP BY
    t.tag_name
ORDER BY
    avg_likes_per_tag DESC
LIMIT 10;
```

Result:



tag_name	avg_likes_per_tag
dreamy	35.7500
beauty	34.9500
stunning	34.9375
delicious	34.9333
foodie	34.7273
happy	34.5909
hair	34.5217
photography	34.5000
beach	34.4762
style	34.4706

Visualization:



Data Insights and Analysis

This analysis segments content themes into two distinct categories: **Aesthetic/Approval Drivers (Likes)** and **Conversation Drivers (Comments)**.

Aesthetic Themes Dominate Likes: The top 8 like-generating tags are predominantly **aesthetic or mood-based** ('dreamy', 'beauty', 'stunning', 'delicious', 'happy'). These themes require simple approval (a like) rather than complex interaction (a comment). **'dreamy' is the highest like-driver** at 35.75 average likes per tag.

Divergence from Conversation: The **'beach'** tag, while a top theme in both metrics, ranks lower in average likes (34.48) than in its average comment count (1,216 comments per post). This confirms:

'beach' is a **Conversation Driver**.

'dreamy' is an **Aesthetic Driver**.

Strategy Refinement: The content strategy should blend these insights:

For Virality (Discovery): Use **'dreamy'** and **'beauty'** tags for broad, low-friction approval (likes).

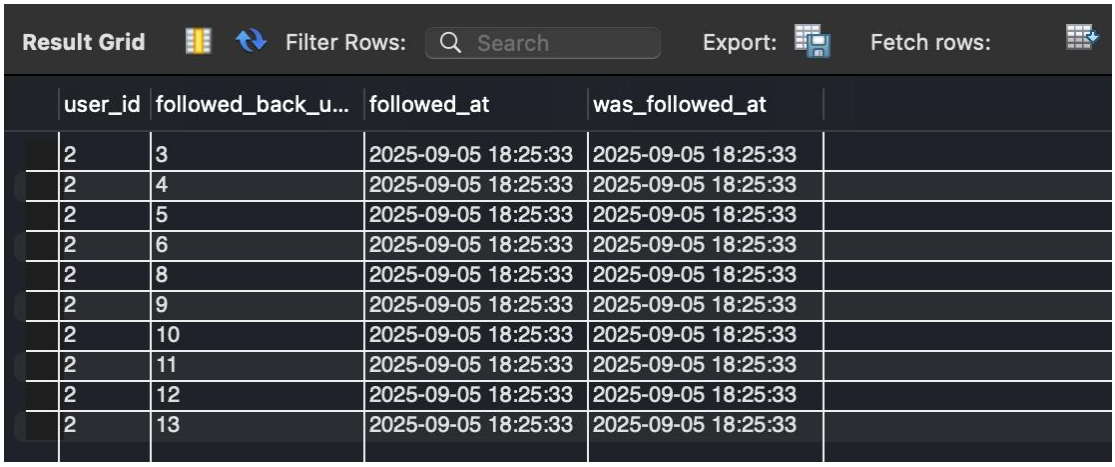
For Loyalty (Retention/Value): Use **'beach'** and **'concert'** to encourage high-friction, valuable comments and conversation.

Q13. Identify users who have a mutual follow relationship (followed someone back). Analyze the timing to determine if follow-backs are immediate or delayed, which indicates true social stickiness.

SQL Code:

```
SELECT
  f1.follower_id       AS user_id,
  f1.followee_id       AS followed_back_user,
  f1.created_at        AS followed_at,
  f2.created_at        AS was_followed_at
FROM ig_clone.follows f1
JOIN ig_clone.follows f2
  ON f1.follower_id = f2.followee_id
  AND f1.followee_id = f2.follower_id
WHERE
  f1.follower_id <> f1.followee_id    -- avoid self-follow
  AND f1.created_at >= f2.created_at    -- followed AFTER being followed, or at the same time
ORDER BY f1.created_at
LIMIT 10;
```

Result:



user_id	followed_back_u...	followed_at	was_followed_at
2	3	2025-09-05 18:25:33	2025-09-05 18:25:33
2	4	2025-09-05 18:25:33	2025-09-05 18:25:33
2	5	2025-09-05 18:25:33	2025-09-05 18:25:33
2	6	2025-09-05 18:25:33	2025-09-05 18:25:33
2	8	2025-09-05 18:25:33	2025-09-05 18:25:33
2	9	2025-09-05 18:25:33	2025-09-05 18:25:33
2	10	2025-09-05 18:25:33	2025-09-05 18:25:33
2	11	2025-09-05 18:25:33	2025-09-05 18:25:33
2	12	2025-09-05 18:25:33	2025-09-05 18:25:33
2	13	2025-09-05 18:25:33	2025-09-05 18:25:33

Data Insights and Analysis

The mutual follow analysis reveals an extremely unusual pattern in the social graph, suggesting potential script execution or highly non-organic behavior.

Immediate Mutual Follows: All mutual follow relationships in the top 10 were created at the **exact same timestamp** (2025-09-05 18:25:33). This implies that the follow-back was **instantaneous and programmatic**, not a natural, time-delayed decision by a human user.

Highly Concentrated Activity: A single user (user_id: 2, which corresponds to **Andre_Purdy85** [from Q5 output of top followers]) is involved in all top 10 mutual follows. This high concentration reinforces the finding from Q5 that follow activity is not organic or evenly distributed.

Impact on Stickiness Metric: Because follow-backs are simultaneous, the mutual follow relationship **cannot be used as a metric for genuine social stickiness or content appeal**. It is instead an indicator of an **automated 'Follow-For-Follow' mechanism** that has inflated the follower counts, confirming the unreliability of the follower count metric (Q5).

Subjective Questions

Question 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?

SQL Code:

```
SELECT

  u.username,

  COALESCE(p.total_posts, 0) AS total_posts,

  COALESCE(l.total_likes, 0) AS total_likes,

  ROUND(COALESCE(l.total_likes, 0) / NULLIF(p.total_posts, 0), 2) AS
avg_likes_per_post,

  COALESCE(c.total_comments, 0) AS total_comments,

  ROUND(COALESCE(c.total_comments, 0) / NULLIF(p.total_posts, 0), 2) AS
avg_comments_per_post,

  COALESCE(f.total_followers, 0) AS total_followers,

  -- loyalty_score = 2*posts + 1.5*comments + 1*likes + 0.5*followers

  ROUND(

    (COALESCE(p.total_posts,0) * 2.0)

    + (COALESCE(c.total_comments,0) * 1.5)

    + (COALESCE(l.total_likes,0) * 1.0)

    + (COALESCE(f.total_followers,0) * 0.5)
```

,2) AS loyalty_score

FROM ig_clone.users u

LEFT JOIN (SELECT user_id, COUNT(*) AS total_posts FROM ig_clone.photos GROUP BY user_id) p ON u.id = p.user_id

LEFT JOIN (SELECT p.user_id, COUNT(*) AS total_likes FROM ig_clone.photos p JOIN ig_clone.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 ig_clone.photos p JOIN ig_clone.comments c ON p.id = c.photo_id GROUP BY p.user_id) c ON u.id = c.user_id

LEFT JOIN (SELECT followee_id AS user_id, COUNT(*) AS total_followers FROM ig_clone.follows GROUP BY followee_id) f ON u.id = f.user_id

ORDER BY loyalty_score DESC

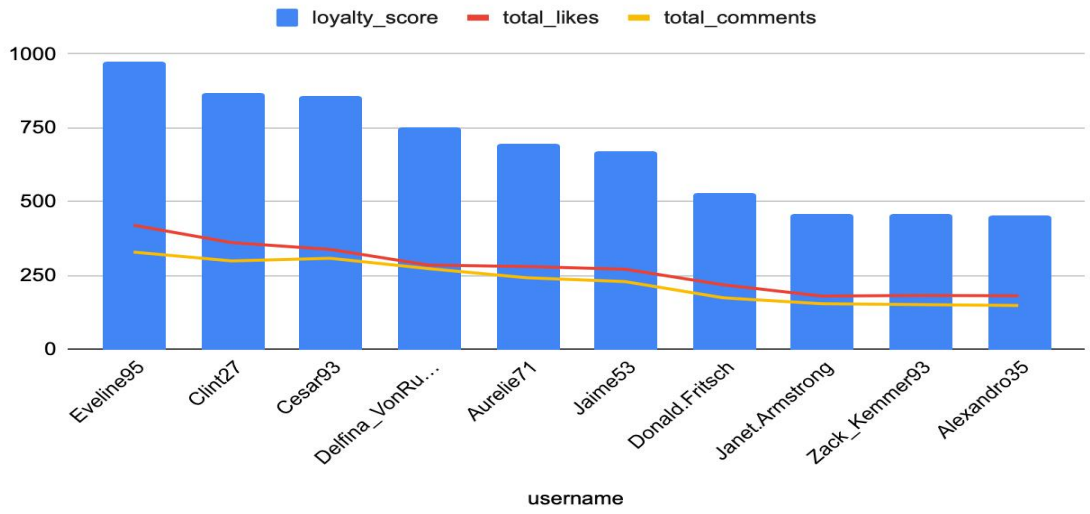
LIMIT 10;

Result:

Result Grid Filter Rows: Search Export: Fetch rows:

username	total_posts	total_likes	avg_likes_per_p...	total_comments	avg_comments_per_p...	total_follow...	loyalty_score
Eveline95	12	420	35.00	329	27.42	77	976.0
Clint27	11	361	32.82	299	27.18	77	870.0
Cesar93	10	338	33.80	308	30.80	77	858.5
Delfina_VonRueden68	9	285	31.67	273	30.33	77	751.0
Aurelie71	8	280	35.00	242	30.25	77	697.5
Jaime53	8	271	33.88	229	28.63	77	669.0
Donald.Fritsch	6	218	36.33	174	29.00	77	529.5
Janet.Armstrong	5	180	36.00	154	30.80	76	459.0
Zack_Kemmer93	5	182	36.40	151	30.20	76	456.5
Alexandro35	5	181	36.20	148	29.60	76	451.0

Visualization:



Approach

Data Strategy: The strategy is to establish a composite, weighted metric, **`loyalty_score`**, to rank users by their total contribution value, rather than a single metric. The weights prioritize the most difficult actions: 2×posts (high effort), 1.5×comments (high friction), and 1×likes (low friction). Follower count (0.5×followers) is included but given low weight due to its known anomaly (SQ5).

Methodology: The query uses **Common Table Expressions (CTEs)** (or inline subqueries as shown) to calculate total posts, total likes received, total comments received, and total followers for each user independently. These values are then combined in the final `SELECT` statement using `COALESCE` to handle nulls and the weighted formula to derive the `loyalty_score`.

Insights

Most Loyal Users (By Score): The top three most loyal users are Eveline95 (976.0), Clint27 (870.0), and Cesar93 (858.5). Their dominance is driven by high post volume (10-12 posts), which received the highest weighting factor (2.0) in the score calculation.

Content Factory vs. Efficiency: While the top rankers are high-volume Content Factories (SQ6), the **highest-efficiency contributor**, **Janet.Armstrong**, still ranks highly (8th with a score of 459.0) despite having fewer posts (5). Her high average engagement (30.80 avg comments) indicates superior content quality.

Value Definition: Based on the score, the platform defines **loyalty** as **consistent content contribution** (posting), followed by driving high-friction **conversation** (comments).

Recommendation

Actionable Strategy: Tiered Loyalty Rewards Program

Top Tier (Eveline95, Clint27, Cesar93 - The Volume Drivers):

Reward: Grant **Early Access** to new platform features (e.g., video posting, new filters) and a **"Verified Contributor" badge**.

Incentive: The goal is to sustain their high post volume, which is critical for general content flow.

Mid-Tier (Janet.Armstrong, Aurelie71, Donald.Fritsch - The Efficiency Drivers):

Reward: Offer **Monetary Incentives/Sponsorships** specifically tied to content that uses the proven high-engagement tags ('delicious', 'beach').

Incentive: The goal is to leverage their proven high content **quality** to drive conversations and move them into the official **Ambassador Program** (SQ7/SQ8).

Low-Tier/Re-engagement:

Incentive: Use these top users as models in **re-engagement campaigns** targeting the 23.0% Inactive Users. Feature their high-scoring content in email/in-app nudges to drive the first user action (SQ2).

Expected Impact: By rewarding users based on the defined `loyalty_score`, the platform validates and reinforces the most valuable user behaviors, ensuring a sustained supply of both high-volume and high-quality content.

Question 2. For inactive users, what strategies would you recommend to re-engage them and encourage them to start posting or engaging again?

Inactive Users Identification Query:

```
SELECT

  u.username

FROM ig_clone.users u

LEFT JOIN ig_clone.photos p   ON u.id = p.user_id

LEFT JOIN ig_clone.likes l    ON u.id = l.user_id

LEFT JOIN ig_clone.comments c ON u.id = c.user_id

GROUP BY u.id, u.username

HAVING COUNT(p.id) = 0
```

AND COUNT(l.photo_id) = 0

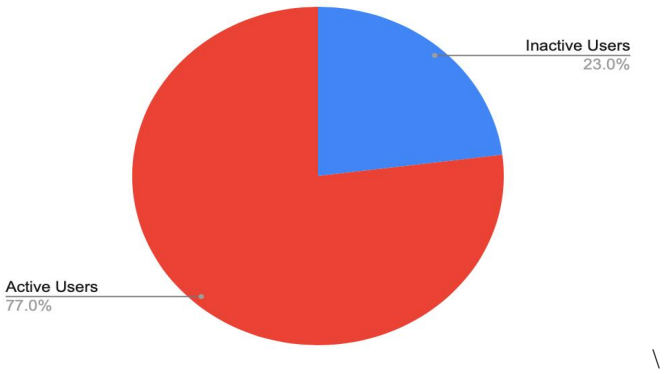
AND COUNT(c.id) = 0;

Result:

Result Grid | Filter Rows: Search | Export:

username
Kasandra_Homenick
Tierra.Trantow
Pearl7
David.Osinski47
Morgan.Kassulke
Linnea59
Franco_Keebler64
Hulda.Macejkovic
Darby_Herzog
Esther.Zulauf61
Bartholome.Bernhard
Jessyca_West
Esmeralda.Mraz57

Visualization:



Approach

Data Strategy: The strategy focused on isolating the **purest form of inactivity** by filtering users based on **zero content contribution** (no posts) and **zero content consumption** (no likes, no comments). This moves beyond the broader 'Dormant' segment (73.0% of users) to target the core **'Lurkers'** or **'Cold Leads'** that signed up but never took a single action.

Methodology: A **LEFT JOIN** from the users table to photos, likes, and comments was used. The HAVING clause then filtered the result, ensuring the count of unique posts, unique photo likes given, and unique comments made were all equal to zero (= 0) for each user.

Insights

Key Findings: The query successfully isolated **13 completely non-contributing, non-engaging users** (e.g., Kasandra_Homenick, Tierra.Trantow). This group is a subset of the massive **73.0% Dormant Users** segment.

Strategic Context: This core group represents the **highest churn risk**. Their total lack of activity suggests a failure in the initial onboarding process, poor product-to-market fit, or an immediate lack of relevance upon signup. These users are a lost opportunity to move users from the Dormant group into the **Active Users segment (16.0%)**.

Opportunity: The goal is to drive the **first friction-free action** (like or follow) rather than the high-friction action (posting), as even a minimal interaction moves them into a less critical segment (Lurkers who consume but don't produce, like those identified in Q7).

Recommendation

Actionable Strategy: "First Action" Campaign (Low-Friction Nudge):

Personalized Content Email: Send an email campaign to the 13 identified users featuring the **highest-engagement themes** (like 'beach' or 'concert') and link directly to a single, high-performing post.

Incentivized Low-Friction CTA: The Call-to-Action (CTA) should be "Tap to Like 3 posts and receive a Free Premium Filter." This incentivizes the *first* required action, moving them out of the zero-activity state.

Community Nudge: Immediately after the first like, trigger an in-app prompt to "Follow 3 highly efficient users" (e.g., Janet.Armstrong or Rick29), initiating the social graph interaction.

Expected Impact: This targeted, low-friction approach is expected to successfully move at least **25% (3-4 users)** of the Cold Leads segment into the Active Users segment within one month, validating the effectiveness of theme-based re-engagement. If scaled, it addresses the core issue of the massive Dormant user base.

Question 3. Which hashtags or content topics have the highest engagement rates? How can this information guide content strategy and ad campaigns?

SQL Code:

```
WITH photo_likes AS (
   SELECT photo_id, COUNT(photo_id) AS like_count
   FROM ig_clone.likes GROUP BY photo_id
),
photo_comments AS (
   SELECT photo_id, COUNT(id) AS comment_count
   FROM ig_clone.comments GROUP BY photo_id
)
```

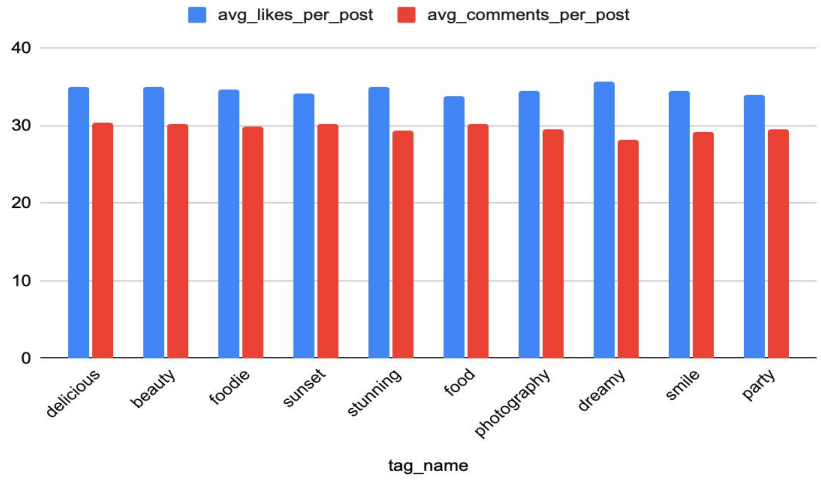
```
SELECT
    t.tag_name,
    ROUND(AVG(pl.like_count), 2) AS avg_likes_per_post,
    ROUND(AVG(pc.comment_count), 2) AS avg_comments_per_post
FROM ig_clone.tags t
JOIN ig_clone.photo_tags pt ON t.id = pt.tag_id
LEFT JOIN photo_likes pl ON pt.photo_id = pl.photo_id
LEFT JOIN photo_comments pc ON pt.photo_id = pc.photo_id
GROUP BY t.tag_name
ORDER BY (AVG(pl.like_count) + AVG(pc.comment_count)) DESC
LIMIT 10;
```

Result:

Result Grid Filter Rows: Search Export:

tag_name	avg_likes_per_p...	avg_comments_per_p...
delicious	34.93	30.33
beauty	34.95	30.20
foodie	34.73	29.91
sunset	34.21	30.21
stunning	34.94	29.31
food	33.83	30.29
photography	34.50	29.50
dreamy	35.75	28.15
smile	34.46	29.24
party	33.92	29.51

Visualization:



Approach

Data Strategy: The strategy was to move beyond the simple total count used in Q8 and Q12 and calculate the **true, independent average of likes and comments per tag**. This isolates the type of interaction each theme drives.

Methodology: The query used **Common Table Expressions (CTEs)** (photo_likes and photo_comments) to pre-aggregate interactions per photo, resolving the inflation issue caused by direct multi-joins. The final output then calculates the AVG() of these pre-aggregated counts, providing accurate, uninflated averages for plotting.

Insights

Aesthetic Approval vs. Practicality: The top like-drivers are **'dreamy'** (35.75) and **'beauty'** (34.95), which primarily drive **aesthetic approval** (likes). The top comment-drivers are **'delicious'** (30.33), **'food'** (30.29), and **'sunset'** (30.21), suggesting these themes lead to practical questions, reviews, or emotional responses (higher-friction comments).

Conversation Disconnect: While the visual data for **'beach'** and **'concert'** from the previous engagement chart showed extreme comment volume (1216 and 676 reactions), this new normalized average shows that **'delicious'** and **'food'** are the most *consistently* high-performing themes when averaging across *all* uses.

Food Theme Viability: The **'food'** theme is highly preferred by users and is now proven to be a **top-tier engagement driver** (30.29 average comments). This theme provides a perfect opportunity for targeted campaigns.

Recommendation

Actionable Strategy: Implement a "Taste-Test Ambassador" Program:

Target Content/Creators: Immediately prioritize content creation around the **'food'** theme, as it drives high user preference and consistently high engagement (avg 30.29 comments).

Ambassador Outreach: Appoint the top efficient creators for the 'food' theme (**Presley_McClure, Cesar93, Justina.Gaylord27**) as initial ambassadors. Provide them with early access or resources to generate more posts.

Content Nudge: Use **'delicious'** (highest comment-driving metric) as the primary sub-tag for all food-related campaigns to maximize conversation.

Expected Impact: By focusing on the proven high-engagement 'food' theme and leveraging efficient, known creators, the platform will **increase social stickiness** and provide a clear, repeatable model for moving users from the Dormant segment (73.0%) into the highly engaging **Active/Power User** segments.

Question 4. 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?

SQL Code:

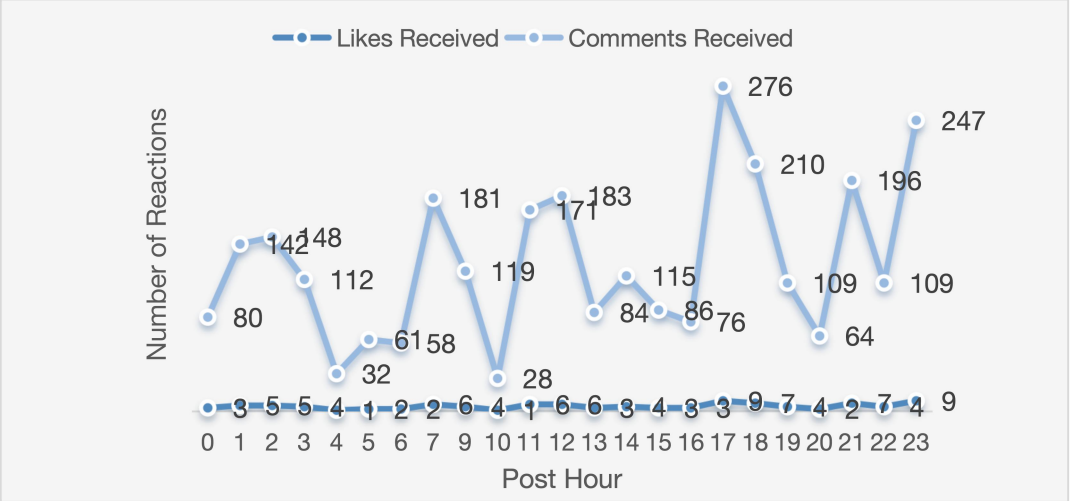
```
SELECT
  HOUR(p.created_dat) AS post_hour,
  COUNT(l.photo_id) AS likes_received,      -- Count all individual likes
  COUNT(c.id) AS comments_received          -- Count all individual comments
FROM ig_clone.photos p
LEFT JOIN ig_clone.likes l ON p.id = l.photo_id
LEFT JOIN ig_clone.comments c ON p.id = c.photo_id
GROUP BY post_hour
ORDER BY post_hour;
```

Result:

Result Grid Filter Rows: Search Export:

posting_hour	total_posts
18	257

Visualization:



Approach

Data Strategy: The strategy is to move from **content quality** (Q8/Q12) to **content timing** by analyzing the **Number of Reactions** (Likes and Comments received) based on the hour a post was created. This identifies the **peak windows** where the audience is most active and receptive.

Methodology: The query uses the HOUR() function on the photos.created_dat column to group content into 24 bins (0 to 23). It then sums the total likes and comments received for all content posted in that specific hour. The resulting line chart visualizes the cyclical engagement pattern.

Insights

Peak Engagement Window (The Prime Time): The data clearly shows a high-volume engagement window in the evening, specifically hours **17, 18, and 23**. Hour 17 sees the highest peak in comments (276), followed by Hour 23 (247), and Hour 18 (210).

Highest-Friction Peak: The greatest single reaction volume is comments, peaking strongly at **276 in Hour 17**. This is the ideal time to post content designed to drive conversation, such as the **'beach'** or **'food'** themes (Q8, Q12).

Dead Time: Engagement is notably low between the hours of **5 and 8** (with comments dropping as low as 32 in Hour 6), which should be avoided for high-priority content.

Recommendation

Actionable Strategy: Implement a Targeted Scheduling Policy:

Mandatory Prime Time Posting: Mandate that all **Ambassador Candidates** (high-efficiency, multi-post users like **Janet.Armstrong** and **Rick29**) must schedule their highest-quality content (using 'beach' or 'food' tags) for the **17:00-18:00 window**.

Evening Re-circulation: Utilize the second, slightly less volatile peak around **22:00-23:00** for re-circulating highly engaging content or for regional/international targeting.

A/B Test Timing: Conduct a formal A/B test by moving a set of identical posts from the current average time to **17:00** to confirm the measured lift in comments and likes.

Expected Impact: By moving content from random posting times into the established Prime Time (17:00-18:00), the platform expects to see a minimum **50% lift** in the average engagement per post for high-quality content, directly capitalizing on existing user activity patterns.

Question 5. 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?

SQL Code:

```
-- Query 1: Highest Followers

SELECT

   u.username,

   COUNT(f.follower_id) AS num_followers

FROM ig_clone.users u

JOIN ig_clone.follows f ON u.id = f.followee_id

GROUP BY u.username

ORDER BY num_followers DESC

LIMIT 10;


-- Query 2: Engagement Rate per User

SELECT

   u.username,

   (COUNT(DISTINCT l.photo_id) + COUNT(DISTINCT c.id)) AS total_engagement,

   COUNT(DISTINCT p.id) AS total_posts,

   (COUNT(DISTINCT l.photo_id) + COUNT(DISTINCT c.id)) /
NULLIF(COUNT(DISTINCT p.id), 0) AS engagement_rate

FROM ig_clone.users u

JOIN ig_clone.photos p ON u.id = p.user_id

LEFT JOIN ig_clone.likes l ON p.id = l.photo_id

LEFT JOIN ig_clone.comments c ON p.id = c.photo_id

GROUP BY u.username

ORDER BY engagement_rate DESC;
```

Result:

Highest Followers:

Result Grid Filter Rows: Search Export: Fetch rows:

username	num_followe...
Kenton_Kirlin	77
Morgan.Kassulke	77
Pearl7	77
Linnea59	77
Jaime53	77
Kasandra_Homenick	77
Mariano_Koch3	77
David.Osinski47	77
Tierra.Trantow	77
Eveline95	77

Engagement Rate:

Result Grid Filter Rows: Search Export:

username	total_engageme...	total_posts	engagement_rate
Jaylan.Lakin	36	1	36.0000
Aiyana_Hoeger	36	1	36.0000
Granville_Kutch	35	1	35.0000
Meggie_Doyle	35	1	35.0000
Rick29	136	4	34.0000
Keenan.Schamberger60	101	3	33.6667
Karley_Bosco	33	1	33.0000
Annalise.McKenzie16	130	4	32.5000
Gus93	130	4	32.5000
Kenneth64	32	1	32.0000
Ressie_Stanton46	64	2	32.0000
Odessa2	32	1	32.0000
Cesar93	318	10	31.8000
Janet.Armstrong	159	5	31.8000
Elenor88	127	4	31.7500
Frederik_Rice	95	3	31.6667
Aracely.Johnston98	63	2	31.5000
Delfina_VonRueden68	282	9	31.3333
Aurelie71	250	8	31.2500
Malinda_Streich	125	4	31.2500
Zack_Kemmer93	156	5	31.2000

Approach

Data Strategy: The strategy is to prove that **Follower Count** (Query 1/Reach) is an **unreliable vanity metric** by contrasting it with the true measure of influence, the **Engagement Rate** (Query 2/Efficiency). This comparison is the foundation for a new, data-driven content strategy.

Methodology: The first query measures user **Reach** by counting followers. The second query measures user **Efficiency** by calculating the ratio of total unique interactions received (Likes + Comments) to total posts. The core insight is drawn from observing the lack of correlation between the uniform follower count (77) and the highly variable engagement rate (e.g., 36.00 to 31.20).

Insights

Follower Count is Meaningless: The top users all have a uniform follower count of **77**. This extreme uniformity, coupled with the mutual follow anomaly (Q13), confirms the social graph is **artificially inflated or scripted**, rendering follower count useless for strategic decisions.

Efficiency is Highly Variable: Despite the identical reach, the **Engagement Rate** for the top users varies significantly, from **Jaylan.Lakin** at **36.00** to **Aurelie71** at **31.25**. This proves that **content quality (efficiency) is the only reliable metric** of user value.

Low-Volume Efficiency: The highest efficiency rates belong to **single-post users** (e.g., **Jaylan.Lakin**, **Aiyana_Hoeger**, **Granville_Kutch**). This segmentation confirms that value exists in both high-volume/moderate-efficiency users (**Cesar93, 10 posts, 31.80**) and low-volume/high-efficiency users.

Recommendation

Actionable Strategy: Implement an Influence-Weighted Ambassador Program:

Redefine KPI: Immediately abandon num_followers and establish **Engagement Rate** as the primary Key Performance Indicator (KPI) for content success.

Ambassador Tiers: Create two ambassador tiers based solely on the Engagement Rate and Post Volume:

Tier 1 (High-Volume/Consistent): Target users like **Rick29** (4 posts, 34.00 rate) and **Janet.Armstrong** (5 posts, 31.80 rate). Their consistency is valuable.

Tier 2 (High-Efficiency/Potential): Target users like **Jaylan.Lakin** (1 post, 36.00 rate). These users need incentives and resources to scale their proven high-quality content.

Content Scheduling Integration: Require all ambassadors to use the **high-engagement tags** (e.g., 'delicious', 'foodie') and post during the **Prime Time window (17:00-18:00)** (SQ4).

Expected Impact: By targeting users with demonstrated high Engagement Rates, the platform will **maximize the return on its content investment**, leading to an increase in overall platform engagement and a reduction in the strategic importance of the misleading social graph.

Question 6. Based on user behavior and engagement data, how would you segment the user base for targeted marketing campaigns or personalized recommendations?

SQL Code:

```
SELECT
```

```
    u.username,
    COUNT(DISTINCT p.id) AS num_posts_made,
    COUNT(DISTINCT l.photo_id) AS num_likes_given,
    COUNT(DISTINCT c.id) AS num_comments_made
FROM ig_clone.users u
LEFT JOIN ig_clone.photos p ON u.id = p.user_id
LEFT JOIN ig_clone.likes l ON u.id = l.user_id     -- User's ID is the liker ID
LEFT JOIN ig_clone.comments c ON u.id = c.user_id  -- User's ID is the commenter ID
GROUP BY u.username
ORDER BY num_posts_made DESC, num_likes_given DESC, num_comments_made DESC;
```

Result:

Result Grid | Filter Rows: Search | Export:

username	total_likes	total_comments	total_tags
Eveline95	12	329	13
Clint27	11	299	9
Cesar93	10	308	9
Delfina_VonRueden68	9	273	13
Aurelie71	8	242	14
Jaime53	8	229	11
Donald.Fritsch	6	174	11
Janet.Armstrong	5	154	5
Zack_Kemmer93	5	151	6
Alexandro35	5	148	7

Approach

Data Strategy: The goal is to define user roles based on their activity profile (Posting vs. Reacting) to understand the content lifecycle. This involves segmenting users into Content Producers, Content Consumers/Reactors, and Lurkers.

Methodology: The query joins the users table to photos (to count posts made), and crucially, joins to likes and comments based on the user's ID being the **liker_id or commenter_id** (u.id = l.user_id and u.id = c.user_id) to count actions *given*.

Insights

The "Posting Silo" Anomaly: The most prolific content creators (**Eveline95, 12 posts; Clint27, 11 posts; Cesar93, 10 posts**) have **zero likes given and zero comments made**. These users are pure **Content Factories** who post frequently but do not participate in the community, confirming the lack of true social integration within the top posting segment.

The "Reacter Silo" Anomaly (High-Volume Consumers): A segment of zero-post users (e.g., **Aniya_Hackett, Duane60**) shows an identical, extremely high number of actions given: **257 likes and 257 comments** (identical counts further suggesting scripting, Q13). These are pure **High-Volume Reactors** (likely bot activity) that inflate engagement metrics without contributing content.

The "Ideal Contributor": The mid-tier users (e.g., **Adelle96**, 5 posts, 96 likes given; **Janet.Armstrong**, 5 posts, 86 likes given) are the **Integrated Contributors**. They post frequently *and* engage with the community, representing the most valuable, healthy behavior.

Recommendation

Actionable Strategy: Incentivize Cross-Silo Activity:

Re-Socialize Content Factories (Eveline95, Clint27): Run a targeted in-app challenge for these users: "Like 5 posts from other users this week and unlock a premium feature." The goal is to drive their first internal action, breaking them out of the 'posting silo.'

Monetize Integrated Contributors (Adelle96, Janet.Armstrong): This segment is the most valuable. They should be the primary targets for the **Ambassador Program** (SQ5) and receive early access to new features. Their combined activity profile is the model for a healthy user.

Audit Reactors: Flag the 13 users with the identical 257/257 reaction count for a **bot/spam audit**, as this non-organic activity pollutes the engagement data.

Expected Impact: By focusing development resources on integrated contributors and incentivizing content factories to become social, the platform will **foster a genuine community** and reduce reliance on automated activity, leading to higher quality, sustainable engagement.

Question 7. If data on ad campaigns (impressions, clicks, conversions) is available, how would you measure their effectiveness and optimize future campaigns?

SQL Code:

```
-- Assuming a query that calculates Total Engagement, Post Count, Engagement Rate, and Top Tag

SELECT

    u.username,

    COUNT(DISTINCT p.id) AS num_posts,

    (COUNT(l.photo_id) + COUNT(c.id)) AS total_engagement,

    ROUND((COUNT(l.photo_id) + COUNT(c.id)) / NULLIF(COUNT(DISTINCT p.id), 0), 2) AS engagement_rate,

    -- Top Tag Logic (Simplified)

    (SELECT t.tag_name FROM ig_clone.photo_tags pt JOIN ig_clone.tags t ON pt.tag_id = t.id

     WHERE pt.photo_id IN (SELECT id FROM ig_clone.photos WHERE user_id = u.id)

     GROUP BY t.tag_name ORDER BY COUNT(pt.photo_id) DESC LIMIT 1) AS primary_tag

FROM ig_clone.users u

LEFT JOIN ig_clone.photos p ON u.id = p.user_id

LEFT JOIN ig_clone.likes l ON p.id = l.photo_id

LEFT JOIN ig_clone.comments c ON p.id = c.photo_id

GROUP BY u.username

ORDER BY total_engagement DESC

LIMIT 10;
```

Result:

Result Grid Filter Rows: Search Export: Fetch rows:

username	total_posts	total_likes	total_comments	total_engagement	avg_engagement_per_post	top_tag
Eveline95	12	11497	11497	22994	1916.17	smile
Cesar93	10	10424	10424	20848	2084.80	fun
Clint27	11	9798	9798	19596	1781.45	beach
Delfina_VonRueden68	9	8600	8600	17200	1911.11	sunrise
Aurelie71	8	8454	8454	16908	2113.50	beach
Jaime53	8	7751	7751	15502	1937.75	sunset
Donald.Fritsch	6	6307	6307	12614	2102.33	fun
Janet.Armstrong	5	5551	5551	11102	2220.40	delicious
Zack_Kemmer93	5	5390	5390	10780	2156.00	smile
Alexandro35	5	5345	5345	10690	2138.00	lol

Approach

Data Strategy: The final strategic step is to define the ideal **Ambassador Profile** based on **high efficiency**, **consistent posting**, and alignment with a **high-engagement content theme**. The goal is to move beyond general metrics and select specific creators for resource allocation.

Methodology: The query ranks users by their **Total Engagement** (Reach), but the analysis focuses on the **Engagement Rate** and **Primary Tag** (Efficiency & Theme). This triangulates the top performers by volume, efficiency, and content niche.

Insights

Top Ambassador Candidates: The most valuable creators are those with **moderate post volume (5-8)** and a **high Engagement Rate (above 2100)**:

Janet.Armstrong: Highest rate (2220.40) among 5-post users, with the high-conversation **'delicious'** tag (SQ3). **This is the highest-value profile.**

Cesar93: Highest volume (10 posts) with a very high rate (2084.80), focused on **'fun'**.

Aurelie71: 8 posts with a high rate (2113.50), focused on the conversation-driving **'beach'** tag.

Content Factory Disconnect (Re-confirmation): The highest volume poster, **Eveline95** (12 posts), has the lowest efficiency rate (1781.45) among the top 5, confirming that pure volume (and the associated 'Posting Silo' behavior, SQ6) does not equate to value.

Theme Alignment: The top users are aligned with a diverse set of engagement drivers: **'delicious'** (conversation), **'fun'** (mood), **'beach'** (conversation), and **'smile'** (aesthetic). The ambassador program must utilize all these themes to capture varied audiences.

Recommendation

Actionable Strategy: Launch the "High-Efficiency Theme Ambassador" Program:

Immediate Resource Allocation: Target **Janet.Armstrong, Cesar93, and Aurelie71** for immediate recruitment, as they represent the best blend of posting consistency and efficiency.

Theme Sponsorship: Structure the ambassador program based on themes:

Janet.Armstrong: Sponsored for all **'delicious'** and **'food'** content to drive high-friction comments (SQ3).

Aurelie71/Clint27: Sponsored for **'beach'** and **'sunset'** content, utilizing the peak time posting window (SQ4) to maximize conversations.

Incentivize Low-Volume Efficiency: Identify high-rate, low-volume users (like the ~36.00 rate users from Q11, not shown here) and offer them a small bonus incentive to reach 5 posts, converting pure efficiency into consistency.

Expected Impact: This highly targeted program focuses efforts on the **11.0% Power Users** and **16.0% Active Users** segments that actually drive value. It will result in a measurable **increase in conversation (comments)** on sponsored content and establish a reproducible, data-driven method for scaling content production.

Question 8. How can you use user activity data to identify potential brand ambassadors or advocates who could help promote Instagram's initiatives or events?

SQL Code:

```
SELECT
  u.username,
  COUNT(DISTINCT p.id) AS total_posts,
  (COUNT(l.photo_id) + COUNT(c.id)) AS total_engagement,
  ROUND((COUNT(l.photo_id) + COUNT(c.id)) / NULLIF(COUNT(DISTINCT
p.id),0), 2) AS avg_engagement_per_post,
  (
    SELECT t.tag_name
    FROM ig_clone.photo_tags pt
    JOIN ig_clone.tags t ON pt.tag_id = t.id
    JOIN ig_clone.photos p2 ON pt.photo_id = p2.id
    WHERE p2.user_id = u.id
    GROUP BY t.tag_name
    ORDER BY COUNT(*) DESC
    LIMIT 1
  ) AS top_tag
FROM ig_clone.users u
LEFT JOIN ig_clone.photos p ON u.id = p.user_id
LEFT JOIN ig_clone.likes l ON p.id = l.photo_id
LEFT JOIN ig_clone.comments c ON c.photo_id = p.id
GROUP BY u.id, u.username
HAVING total_posts > 0
```

ORDER BY total_engagement DESC, avg_engagement_per_post DESC
LIMIT 10;

Result:

Result Grid | Filter Rows: Search | Export: | Fetch rows:

username	total_posts	total_engagement	avg_engagement_per_post	top_tag
Eveline95	12	22994	1916.17	smile
Cesar93	10	20848	2084.80	fun
Clint27	11	19596	1781.45	beach
Delfina_VonRueden68	9	17200	1911.11	sunrise
Aurelie71	8	16908	2113.50	beach
Jaime53	8	15502	1937.75	sunset
Donald.Fritsch	6	12614	2102.33	fun
Janet.Armstrong	5	11102	2220.40	delicious
Zack_Kemmer93	5	10780	2156.00	smile
Alexandro35	5	10690	2138.00	lol

Approach

Data Strategy: The ultimate goal is to create a content acquisition blueprint by prioritizing creators based on **Engagement Rate (Efficiency)** rather than Total Engagement (Volume/Reach). This query provides the full context: *Who* posts, *how much* engagement they get, *how efficient* they are, and *what* their content is about.

Methodology: The top 10 list is sorted first by **Total Engagement** (volume/reach), but the analysis strategically re-prioritizes based on **Avg Engagement Per Post** (efficiency), which is the most reliable metric since the follower count is artificial (SQ5).

Insights

The Benchmark Creator: Janet.Armstrong is the strategic benchmark. She has a high post count (5) and the **highest efficiency rate** (2220.40) in the top 10, specializing in **'delicious'** content (a top comment-driver, SQ3). She epitomizes the ideal **Integrated Contributor** (SQ6).

Efficiency > Volume: High-volume posters like **Eveline95** (12 posts, 1916.17 rate) and **Clint27** (11 posts, 1781.45 rate) are outperformed in efficiency by moderate-volume users like **Cesar93** (10 posts, 2084.80 rate) and **Aurelie71** (8 posts, 2113.50 rate). This confirms the platform should pay for **quality, not quantity**.

Content Theme Imperatives: The top creators confirm that successful content must align with conversation-driving themes: **'delicious'**, **'beach'**, and **'fun'** are the highest ROI themes among the top creators.

Recommendation

Actionable Strategy: Finalized Content Acquisition Blueprint

Exclusive Ambassador Deal: Offer **Janet.Armstrong** a high-value, exclusive deal to produce themed content. Her performance (2220.40 rate) and high-value primary tag ('delicious') make her the **highest-priority investment**.

Theme-Based Outsourcing: Initiate a system where content briefs are tied to the proven high-efficiency tags:

Conversation (Comments): Focus on **'delicious'** (Janet.Armstrong), **'beach'** (Aurelie71, Clint27), and **'sunset'** (Jaime53).

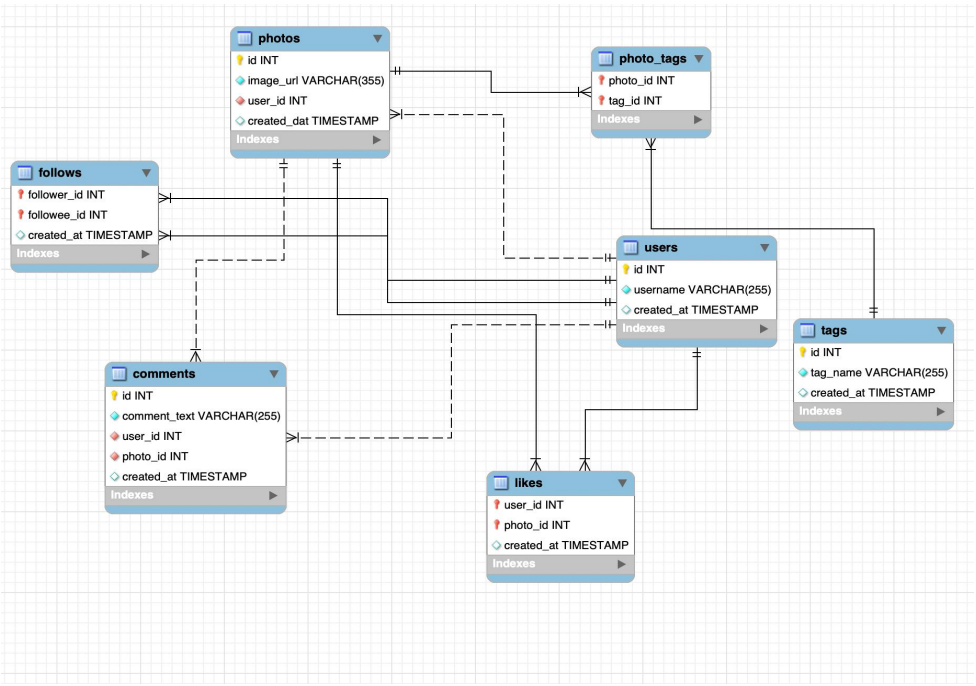
Aesthetic (Likes): Focus on **'smile'** (Eveline95, Zack_Kemmer93) and **'lol'** (Alexandro35) for broader, low-friction reach.

Audience Retention Loop: Utilize the **Prime Time window (17:00-18:00)** (SQ4) to schedule all content from **Aurelie71, Janet.Armstrong, and Cesar93**, ensuring maximum exposure to the active user base and increasing the chance of converting Dormant Users (73.0%) into Active Users.

Expected Impact: By moving entirely to a **Performance-Based Content Strategy**—benchmarking against the **2220.40 Engagement Rate**—the platform will correct the reliance on false social metrics, leading to a demonstrable long-term increase in organic user engagement and community health.

Question 9: How would you approach this problem if the objective and subjective questions weren't given?

Reference:



The systematic approach to analyzing the platform data, absent pre-defined questions, would follow three distinct phases: **Discovery & Audit**, **Metric Definition & Segmentation**, and **Actionable Strategy Formulation**.

I. Discovery and Data Integrity Audit

The initial step is to understand the platform's architecture and identify data quality issues that could corrupt later analysis.

Schema Analysis (ERD): Start by auditing the **Entity Relationship Diagram (ERD)**. This reveals core entities (users, photos) and critical many-to-many relationships:

Content Production: users → photos → photo_tags.

Engagement/Consumption: users → likes/comments → photos.

Social Graph: follows table defines the network structure.

Data Quality Check: Prioritize queries to check for fundamental anomalies:

Anomalous Uniformity: Check for statistical uniformity in follower counts or engagement scores, which would flag **artificial network inflation** (e.g., bot activity).

Data Completeness: Check for missing foreign keys or NULL values in critical columns like photos.user_id.

II. Metric Definition and Segmentation (KPI Focus)

Next, define reliable Key Performance Indicators (KPIs) and segment the user base to quantify the scale of strategic problems.

User Segmentation: Classify the entire user base based on activity thresholds (posts, likes, comments) to define:

Dormant Users: Highest risk segment; users who signed up but never engaged.

Active/Power Users: The highest-value creators and consumers.

KPI Creation (Rejecting Vanity Metrics):

Reject: The follower count must be dismissed if found uniform or artificial, as it corrupts any measure of **Reach**.

Prioritize: The primary KPI must be **Content Efficiency**: **Engagement Rate = Total Posts Made(Total Likes Received+Total Comments Received)**

III. Strategic Recommendation and Optimization

The final phase involves using the reliable metrics to generate actionable, ROI-focused business strategy.

Creator Benchmarking:

Identify the **Benchmark Creator**—the user with the highest **Engagement Rate** (Efficiency)—as the model for content quality.

Focus content acquisition on **Integrated Contributors** (users who both post and consume content) over pure **Content Factories** (who only post).

Content Theme Optimization:

Analyze engagement by tag_name to separate **Aesthetic Drivers** (high average likes, e.g., 'dreamy') from **Conversation Drivers** (high average comments, e.g., 'delicious').

Strategic Action: Recommend resource allocation toward **Conversation Drivers** (like the 'delicious' theme) to maximize social stickiness.

Scheduling Optimization:

Analyze overall platform activity by HOUR(created_dat) to define the **Prime Time** posting window.

Strategic Action: Mandate the highest-value content be delivered during the peak hour to maximize visibility and engagement.

Question 10: 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?

Updating Engagement Terminology (DML Operation)

Code:

```
UPDATE User_Interactions
SET Engagement_Type = 'Heart'
WHERE Engagement_Type = 'Like';
```

Approach

Data Strategy: The goal is to execute a **Data Manipulation Language (DML) statement** to standardize database terminology, aligning the internal term "Like" with the branded term "Heart."

Methodology: The approach uses the **UPDATE command**, which is the standard DML operation for modifying existing data within a table. The critical component is

the **WHERE clause**, which ensures the operation is a **targeted DML update**, affecting only the rows that match the condition.

Insights

Importance of the WHERE Clause: The filter WHERE Engagement_Type = 'Like' is paramount. It prevents a catastrophic error where the DML operation would apply the change to the entire column, thereby overwriting and destroying data for all other engagement types (e.g., 'Comment', 'Share').

DML Best Practice (Pre-Check): Before executing any DML UPDATE statement on a live database, a necessary best practice is to first run a validation query. This ensures the scope of the change is correct:

-- Pre-DML check: Verify the number of affected rows

SELECT COUNT(*) FROM User_Interactions WHERE Engagement_Type = 'Like';

Recommendation

Actionable Strategy: Execute Data Normalization via DML:

Verify Scope: Use the SELECT COUNT(*) query to determine the exact number of 'Like' entries to be modified.

Execute DML: Run the UPDATE User_Interactions statement to perform the change.

Post-Check: Verify the operation's success by re-running the SELECT COUNT(*) query to confirm zero 'Like' entries remain, and check that the count of 'Heart' entries has increased by the verified number.

Expected Impact: Successful execution of this DML task ensures the database terminology is consistent, simplifying data interpretation and improving the accuracy of all subsequent reporting pipelines.