Instagram User Analytics Report

Project Description

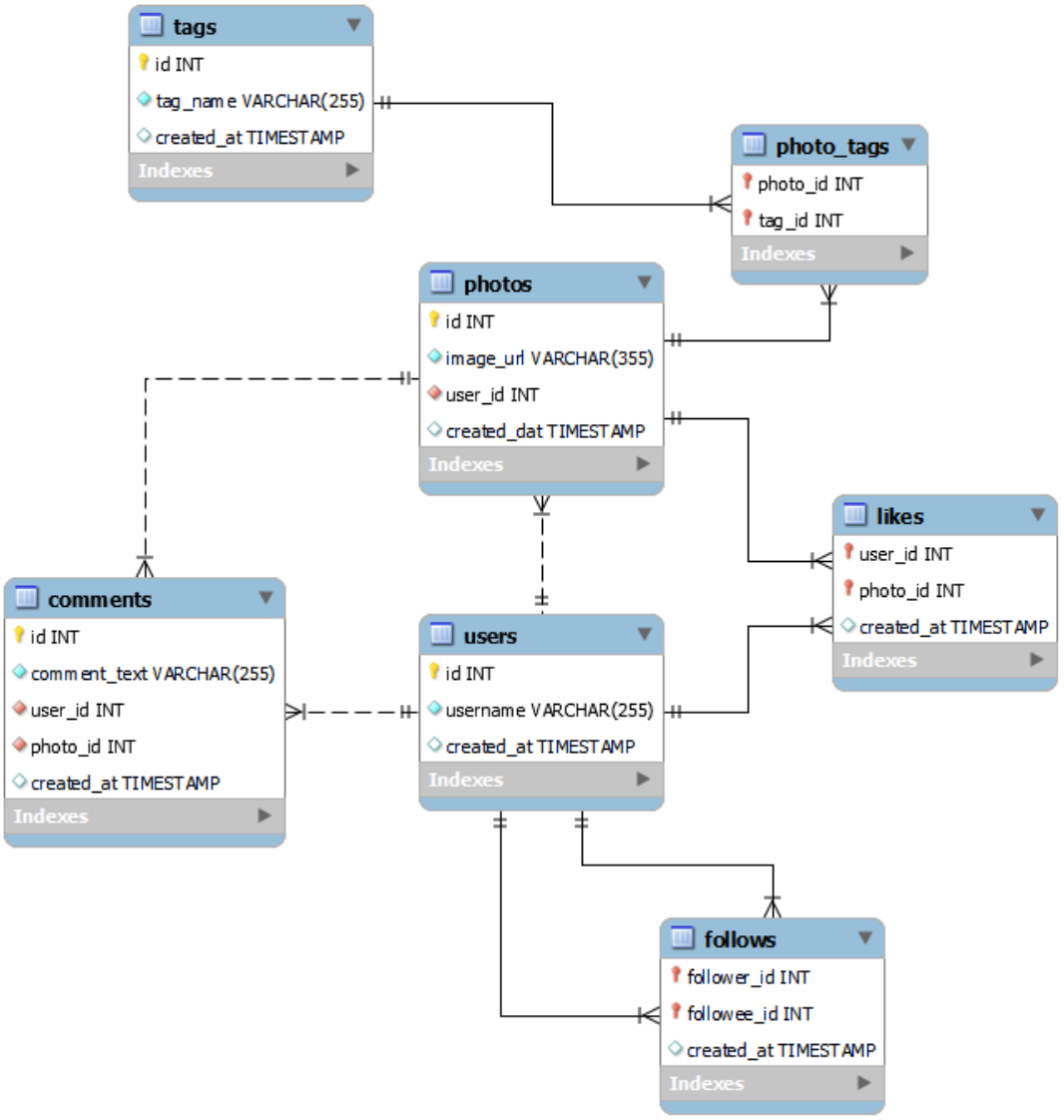
This project involved performing a detailed SQL-based analysis of Instagram's user data to uncover meaningful insights around engagement, growth patterns, fake accounts detection and content popularity trends. The goal was to gather data to enable key business decisions around marketing, new features, improving user experience and monitoring platform health.

The analysis could provide multiple teams at Instagram data-backed guidance around campaign management, new capability development and user experience improvements. Key use cases include:

- Marketing: Optimize spend based on user activity patterns
- Product: Prioritize tech investments into platform capabilities
- Operations: Improve moderation to ensure community culture

Database Schema

The MySQL database analyzed contained 7 tables to store relevant Instagram social data like users, photos, comments, follows and tags. The key entities and relationships are visualized below:



Entity Relationship Diagram

Approach

The analysis began with a comprehensive visualization of the dataset, emphasizing the identification of relationships and understanding of foreign keys. This involved exploring the raw data format to grasp the structure and relationships between different entities such as users, photos, comments, follows, and tags. The visualization phase laid the groundwork for formulating efficient SQL queries.

The subsequent steps involved developing query logic to extract meaningful insights. This included crafting queries to calculate user signups and posts per user, identifying inactive users, analyzing hashtag usage, and detecting potential fake accounts.

Complex joins across tables were a critical aspect of the analysis. Joining tables such as users to photos and likes to comments allowed for a holistic view of user engagement and

interaction patterns on the platform. By leveraging SQL relationships, the analysis was able to provide a nuanced understanding of user behaviour and content dynamics.

Tech Stack

- MySQL database for its free and open-source nature and its ubiquity allowing me to possess a highly transferable skill in my arsenal.
- Beekeeper Studio v4.0.3 Community Edition (SQL Query Editor) for its free and open-source nature, as well as superior and simplistic UI/UX as compared to MySQL Workbench.

Key Insights

Fake Account Detection

The analysis flagged accounts that exhibited abnormal behaviour of liking every photo in the database. This approach is grounded in the observation that it is not typically feasible for a genuine user to like every single photo on a platform with diverse content. Through this strategy, **13 Potential Fake Accounts** were detected. Uncovering potential bot accounts helps maintain the integrity of user engagement metrics and ensures a genuine and trustworthy platform.

```
# Potential bot/fake accounts
select
  user_id,
  u.username,
  count(*) num_likes
from
  likes l
  join users u on l.user_id = u.id
group by
  user_id
having
  count(*) = (
    select
      count(id)
    from
      photos
  );
```

user_id	username	num_likes
5	Aniya_Hackett	257
14	Jaclyn81	257
21	Rocio33	257
24	Maxwell.Halvorson	257
36	Ollie_Ledner37	257
41	Mckenna17	257
54	Duane60	257
57	Julien_Schmidt	257
66	Mike.Auer39	257
71	Nia_Haag	257
75	Leslie67	257
76	Janelle.Nikolaus81	257
91	Bethany20	257

User Engagement

For this analysis, a post by a user is defined as any interaction involving the user on the platform. This includes actions such as comments, likes, or photo uploads. By considering these varied forms of engagement, the analysis aims to capture the breadth and depth of user interactions on Instagram.

```
# Average posts per user and average photos per user
with
  comments_per_user as (
    select
      user_id,
      count(id) num_comments
    from
      comments
    group by
      user_id
  ),
  photos_per_user as (
    select
      user_id,
      count(id) num_photos
    from
      photos
```

```
    group by
      user_id
  ),
  likes_per_user as (
    select
      user_id,
      count(*) num_likes
    from
      likes
    group by
      user_id
  )
select
  avg(
    ifnull(num_comments, 0) + ifnull(num_photos, 0) + ifnull(num_
  ) avg_num_posts,
  avg(ifnull(num_photos, 0)) avg_num_photos
from
  users u
  left join comments_per_user cpu on u.id = cpu.user_id
  left join photos_per_user ppu on u.id = ppu.user_id
  left join likes_per_user lpu on u.id = lpu.user_id
where
  (
    num_comments < 257
    or num_comments is null
  )
  and (
    num_comments is not null
    or num_photos is not null
    or num_likes is not null
  );
```

Although the query for determining only the user engagement could have been much simpler, this complex query provides a more detailed insight into the data. It also facilitates filtering out potential fake accounts (first condition for the `and` operator in the `where` clause) and inactive users with no record of posting either a comment, a photo or a like on the platform (second condition for the `and` operator in the `where` clause).

1. *User Engagement for All Users:*

   Average Number of Posts per User = 165.3

Average Number of Photos per User = 2.6

This metric considers all users on the platform, providing a comprehensive view of overall user engagement. It reflects the average number of posts and photos created by users, contributing to the vibrancy of the platform. It serves as a baseline to understand the general trend of content creation and user participation.

2. *User Engagement for Users without Fake Accounts:*

   Average Number of Posts per User = 113.2

   Average Number of Photos per User = 3

   This metric focuses on genuine user engagement by excluding potential fake accounts. By filtering out accounts flagged as potential bots, the adjusted average provides a more accurate representation of authentic user activity. This metric is valuable for marketing and content strategies, as it reflects the behaviour of real users who contribute to the platform.

3. *User Engagement for Users without Fake Accounts and Inactive Users:*

   Average Number of Posts per User = 133

   Average Number of Photos per User = 3.5

   This metric takes a further refined approach by not only excluding potential fake accounts but also filtering out inactive users. This optimized average offers insights into user engagement from a pool of genuine and actively participating users. This metric is particularly useful for assessing the health and dynamism of the user community.

4. *User with the most number of likes on a single photo:*

   The user with the most number of likes on a single photo was identified. This metric serves the specific purpose of determining a contest winner and recognizing exceptional user engagement. This user is identified as:

```
# User with most likes on a single photo
with
  likes_per_photo as (
    select
      photo_id,
      count(user_id) no_of_likes
    from
      likes
    group by
      photo_id
  )
```

```
select
  p.user_id,
  u.username,
  p.id photo_id,
  p.image_url,
  lpp.no_of_likes
from
  photos p
  join likes_per_photo lpp on p.id = lpp.photo_id
  join users u on u.id = p.user_id
order by
  no_of_likes desc
limit
  1;
```

user_id	username	photo_id	image_url	no_of_likes
52	Zack_Kemmer93	145	https://jarret.name	48

The user with the most likes on a single photo stands out as a notable contributor to the platform, garnering significant attention and engagement from the Instagram community. This metric can be particularly valuable in the context of contests or promotional campaigns where recognizing and rewarding high-engagement users is a strategic initiative.

5. *Days with the most user sign-ups:*

   Knowing the days with the highest user signups allows Instagram to optimize marketing efforts and user acquisition strategies on specific days, maximizing growth potential.

```
# Day of week with most user sign-ups
select
  dayname(created_at) day_of_week,
  count(id) no_of_users
from
  users
group by
  dayname(created_at)
order by
  no_of_users desc;
```

day_of_week	no_of_users
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12

For the given database, **Thursday and Sunday** saw the most user sign-ups. Based on this, an optimum strategy can be to start running ad campaigns on Fridays, Saturdays and Sundays to align with the most number of recently joined users adding strong CTAs asking new users to engage right away. Additionally, a midweek campaign can be included with increasing intensity from Tuesdays through Thursdays to build momentum throughout the week culminating with the Sunday sign-up peak.

6. *Oldest Users and Inactive Users:*

   The analysis identified the 5 oldest users and users who have not uploaded a single photo since sign-up. Recognizing the longevity of certain users and identifying inactive users provides insights into user behaviour over time and highlights potential opportunities for re-engagement strategies. These users are identified as:

```
# 5 oldest users
select
  *
from
  users
order by
  created_at
limit
  5;
```

id	username	created_at
80	Darby_Herzog	2016-05-06 00:14:21
67	Emilio_Bernier52	2016-05-06 13:04:30
63	Elenor88	2016-05-08 01:30:41
95	Nicole71	2016-05-09 17:30:22
38	Jordyn.Jacobson2	2016-05-14 07:56:26

```
# Inactive users
select
  u.id,
  u.username
from
  users u
  left join photos p on u.id = p.user_id
where
  p.id is null;
```

id	username
5	Aniya_Hackett
7	Kasandra_Homenick
14	Jaclyn81
21	Rocio33
24	Maxwell.Halvorson
25	Tierra.Trantow
34	Pearl7
36	Ollie_Ledner37
41	Mckenna17
45	David.Osinski47
49	Morgan.Kassulke
53	Linnea59
54	Duane60
57	Julien_Schmidt
66	Mike.Auer39
68	Franco_Keebler64
71	Nia_Haag
74	Hulda.Macejkovic
75	Leslie67
76	Janelle.Nikolaus81
80	Darby_Herzog
81	Esther.Zulauf61
83	Bartholome.Bernhard
89	Jessyca_West

90	Esmeralda.Mraz57
91	Bethany20

Recognizing the 5 oldest users allows Instagram to celebrate and acknowledge the loyalty of long-term users, fostering a positive community spirit. Identifying inactive users presents an opportunity for Instagram to implement targeted re-engagement campaigns, encouraging users to become active contributors once again.

Content Analysis

Top 5 most used hashtags: Identifying popular hashtags enables Instagram to guide brands and content creators in maximizing reach and engagement. Instagram can collaborate with partner brands to create sponsored content using these popular hashtags, increasing visibility and user interaction. The top 5 hashtags present in the database are identified as follows:

```
# 5 most popular hashtags
select
  pt.tag_id,
  t.tag_name,
  count(pt.photo_id) tag_use_count
from
  photo_tags pt
  join tags t on pt.tag_id = t.id
group by
  pt.tag_id
order by
  tag_use_count desc
limit
  5;
```

tag_id	tag_name	tag_use_count
21	smile	59
20	beach	42
17	party	39
13	fun	38
18	concert	24

Results

The comprehensive SQL analysis delivered tangible value by uncovering usage trends and engagement velocity. Quantifying this value involves assessing the impact on decision-making and business outcomes.

Achievements:

- Successfully extracted and analyzed meaningful insights from a complex SQL database
- Uncovered trends around user engagement, growth patterns, and content popularity which can drive product decisions
- Identified fake accounts and inactive users to improve the health of the platform
- Displayed proficiency in using SQL for data analysis on business metrics

Value Provided:

- Quantified average user engagement levels to monitor community involvement over time
- Revealed best days for launching marketing campaigns based on sign-up data patterns
- Discovered most popular hashtags to help influencers reach wider audiences
- Pinpointed users with very high engagement as brand ambassadors
- Allowed data-driven decisions on new features to build and user experience improvements

Skills Gained:

- Building an SQL database from scratch and importing datasets
- Writing efficient SQL queries across multiple tables to extract insights
- Identifying relationships between metrics to derive actionable conclusions
- Unlocking dashboard-ready metrics through challenging SQL queries
- Structuring analytics reports tailored to stakeholder needs