

NAME: NARINDER SINGH
DATA ANALYTICS TRAINEE

TASK 2: INSTAGRAM USER ANALYTICS

ANALYSIS ON THE GIVEN POINTS

(A). MARKETING:

1. Rewarding the most loyal users
2. Reminding the inactive users to Start Posting
3. Declaring the contest winners
4. Hashtag Researching
5. Launching AD Campaign

(B). INVESTOR METRICS:

6. User Engagement
7. Bots and Fake Accounts

PROJECT DESCRIPTION:

This project focuses on using SQL and *MySQL Workbench* to understand *user interactions* as well as engagement data from Instagram. Working in the product team as a data analyst, I need to provide valuable information that could be used *for strategic decision making* on improving user experience, helping to build features and supporting relevant marketing efforts.

The approach relies on SQL queries to be used as instruments of exploring the database and seeking out essential information that can address particular sets of *business issues and objectives* designed in this research. The project includes a number of activities that aim to reveal *user behavior and engagement* options in order to make specific recommendations.

APPROACH:

Understanding the Database Schema:

To do so, I would first familiarize myself with the ways in which the given database of Instagram is constructed. Additionally, I will specify the main tables to be used and define their associates as well as *attributes for analysis* available.

Data Exploration:

- Investigate the information contained in each appropriate table to determine what details are present. For instance, *attributes of the users table* are id, username and created_at timestamp etc. However, the fields of the photo table (that contain information about photos posted by users) are id, image_url, user_id, created_at etcetera.
- Locate *primary keys, foreign keys* and all joins needed to pull out meaningful information.

Query Formulation:

- Developed *SQL queries* for the details and the relevant questions provided by the management.
- Utilizing queries related to aggregate, filter and selection statements to access the required data.

Task Execution:

- Perform activities included in the management team's instructions, to determine five eldest Instagram users.
- Using SQL queries to *extract the required information* and presenting them in an easy-to-understandable format.

Insight Generation:

- Analyzing the data gathered from the queries to derive valuable information that can be linked with *business goals*.
- Characterizing the results in terms of user activity, *engagement metrics* and areas that can be improved further.

Presentation of Insights:

- Efficiently conveying the identified findings of our analysis to stakeholders, i.e., product manager and others on the team.
- Giving practical recommendations with respect to the findings to inform the procedures of *decision making*.

Tech-Stack Used:

In this project, I applied **MySQL Workbench 8.0 CE** to perform the relevant SQL queries towards the accomplishment of the provided tasks such as identification of top five oldest users on Instagram from given database, determination and recommendation of top five most used hashtags on Instagram platform plus calculation of average number of posts per user on Instagram for evaluating *user engagement* among others.

I opted to use **MySQL Workbench 8.0 CE** as my preferred GUI because the application has been developed for use with *MySQL databases* and is fully compatible with our scripts' SQL dialect. Moreover, there are advanced options available in the tool such as syntax highlighting and *auto-completion* which assist in writing *efficient SQL queries* as well.

MARKETING:

1. LOYAL USER REWARD:

Top 5 oldest Instagram users- These are the people who use the platform for a very long period.

- **SELECT * FROM users:** This part of the query is selecting all columns (*) from user's table.
- **ORDER BY created_at** clause picks out the chosen rows by created_at column. Order by sorts in ascending order by default, so this will sort it from the oldest created_at timestamp to the newest.
- **LIMIT 5:** This condition restricts the set of results to the top 5 items on a sorted list.

Program/Query:

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

MARKETING:

LOYAL USER REWARD:

Top 5 oldest Instagram users- These are the people who use the platform for a very long period.

Result:

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

Therefore, the top five longest users of Instagram are Darby_Herzog'(started using on 2016-05-06 00:14:21), Emilio_Bernier52, Elenor88, Nicole 71, and Jordyn.Jacobson2.

MARKETING:

2. Inactive User

Engagement:

Program/Query:

```
SELECT username
FROM users
LEFT JOIN photos
ON users.id=photos.user_id
WHERE photos.id IS NULL;
```

Sending inactive users a reminder to schedule activity shots (users who have never posted one single photo on Instagram for now)

- The part of the query, (SELECT username FROM users), selects the column named username from table users.
- The clause LEFT JOIN photos ON users.id = photos.user_id joins rows between the two tables, users and photos by using the user_id. When a user has never posted a photo, the fields from the photos table will be NULL for that user.
- The clause WHERE photos.id IS NULL filters all users, who have posted a photo and leaves only those users, which never posted a photo. This is due to the fact that if a user has never posted a picture, in the photos table, the id field will be NULL for that particular user.

MARKETING:

Inactive User

Engagement
Result:

Sending inactive users a reminder to schedule activity shots (users who have never posted one single photo on Instagram for now)

username
Aniya_Hackett
Bartholome.Bernhard
Bethany20
Darby_Herzog
David.Osinski47
Duane60
Esmeralda.Mraz57
Esther.Zulauf61
Franco_Keebler64
Hulda.Macejkovic
Jadyn81
Janelle.Nikolaus81
Jessyca_West
Julien_Schmidt
Kasandra_Homenick
Leslie67
Linnea59
Maxwell.Halvorson
Mckenna17
Mike.Auer39
Morgan.Kassulke
Nia_Haag
Ollie_Ledner37
Pearl7
Rocio33
Tierra.Trantow

A total of 26 users do not have posted one photo on Instagram.

MARKETING:

3. Contest Winner Declaration

Program/Query:

```
SELECT
username, photos.id, photos.image_url, count(likes.user_id) As total
FROM photos
INNER JOIN likes
ON likes.photo_id=photos.id
INNER JOIN users
ON photos.user_id=users.id
GROUP BY photos.id
ORDER BY total DESC
LIMIT 1;
```

The team launched a competition and now the user who gets most likes on a single photo is supposed to win that contest; and they want to announce the winner. In this contest, determine the winner and present their details to the team.

- FROM photos is the primary table for which we are retrieving data.
- The join statement INNER JOIN likes ON likes.photo_id=photos.id links the photos table and the likes table in which photo_id in the likes table matches id of the photos' one. This will yield a new table in which each row represents the data about one photo and about a like of that photo.
- INNER JOIN users ON photos.user_id = users.id combines the above table with the users table where user_id of photos matches id of users. This provides us with a new table in which each row contains the details of a particular photo, such as like and users who uploaded it.
- GROUP BY photos.id groups the rows based on photos' id. This implies that for each photo we will have one record in our result set which contains details about that photo, the user who posted it and all the likes it received.
- count(likes.user_id) As total sums up the likes in every photo. Since we have grouped by photos.id, then this will get us the number of likes per photo.
- ORDER BY total DESC subsequently sorts the rows according to total likes, in a descending manner. This implies that the most liked pictures should come first.
- LIMIT 1 restricts the set of results to only the first row. We ordered in descending order by total likes, so this will be the photo with the highest number of likes.

MARKETING:

Contest Winner Declaration

The team launched a competition and now the user who gets most likes on a single photo is supposed to win that contest; and they want to announce the winner. In this contest, determine the winner and present their details to the team.

Result:

username	id	image_url	total
Zack_Kemmer93	145	https://jarret.name	48

The winner of the competition is user named Zack_Kemmer93, since his photo with photo_id 145 has been liked by substantial number of users i.e. 48 likes which is a higher value than any other entry in the contest.

MARKETING:

4. Hashtag Research:

Program/Query

```
SELECT
tags.tag_name,
COUNT(*) As total
FROM photo_tags
JOIN tags
ON photo_tags.tag_id = tags.id
GROUP BY tags.id
ORDER BY total DESC
LIMIT 5;
```

One of the partner brands is asking the hashtags to use in order for the post to be seen by many people on the platform.(Top 5 most popular #Hashtags on Instagram)

- SELECT tags.tag_name, COUNT(*) As total clause picks out the tag_name from the tags table and count how many times each such tag_id appears in the photo_tags table(which represents how many times that specific tag has been used).
- FROM photo_tags JOIN tags ON photo_tags.tag_id = tags.id is the portion of the query where photo_tags table is joined with tags table on condition that tag_id in the photo_tags table is paired with id in the tags table. This provides us with a transformed table such that each row describes the association between a photo and tag, along with corresponding tag.
- GROUP BY tags.id clause is used to group the selected rows by tags id. This, therefore, implies that each tag will have one row in our result set which has the tag_name and the number of times it has been used.
- ORDER BY total DESC clause sorts the rows by count of each tag in descending order. This implies that the most used tags will come first.
- LIMIT 5 restricts the output to only the first 5 lines. As we sorted by total count in a descending order, these will be the top five hashtags.

MARKETING:

Hashtag Research:

One of the partner brands is asking the hashtags to use in order for the post to be seen by many people on the platform.(Top 5 most popular #Hashtags on Instagram)

Output:

tag_name	total
smile	59
beach	42
party	39
fun	38
concert	24

Among the prevalent hashtags on the platform, the top 5 frequently utilized ones include Smile, beach, party, fun, and concert with Smile having been utilized 59 times.

MARKETING:

5. Ad Campaign Launch:

Program/Query

```
SELECT  
DAYNAME (created_at) As day, count(*) as total  
FROM users  
GROUP BY day  
ORDER BY total DESC  
LIMIT 1;
```

The team wants to find out the optimal day of the week to launch ads. Find the weekday with the highest number of registrations on Instagram. Offer considerations about ad campaign timing.

- FROM users: This is the primary table from which we are retrieving data.
- SELECT DAYNAME(created_at) As day, count(*) as total: This portion of the query is based on selecting day of the week from created_at time stamp (DAYNAME(created_at)) and count of users(count(*)) from users table.
- GROUP BY day: This clause aggregates the chosen rows according to the day of a week. This implies that we will have one row in our result set for each day of the week, which contains the day and total number of users registered on that particular day.
- ORDER BY total DESC: This clause sorts rows according to the number of users in descending order. This implies that the days with highest registration will come first.
- LIMIT 1: This particular clause confines the result set to a single first row only. Since we sorted by total count in descending order, this is the day with the highest number of registrations.

MARKETING:

Ad Campaign Launch:

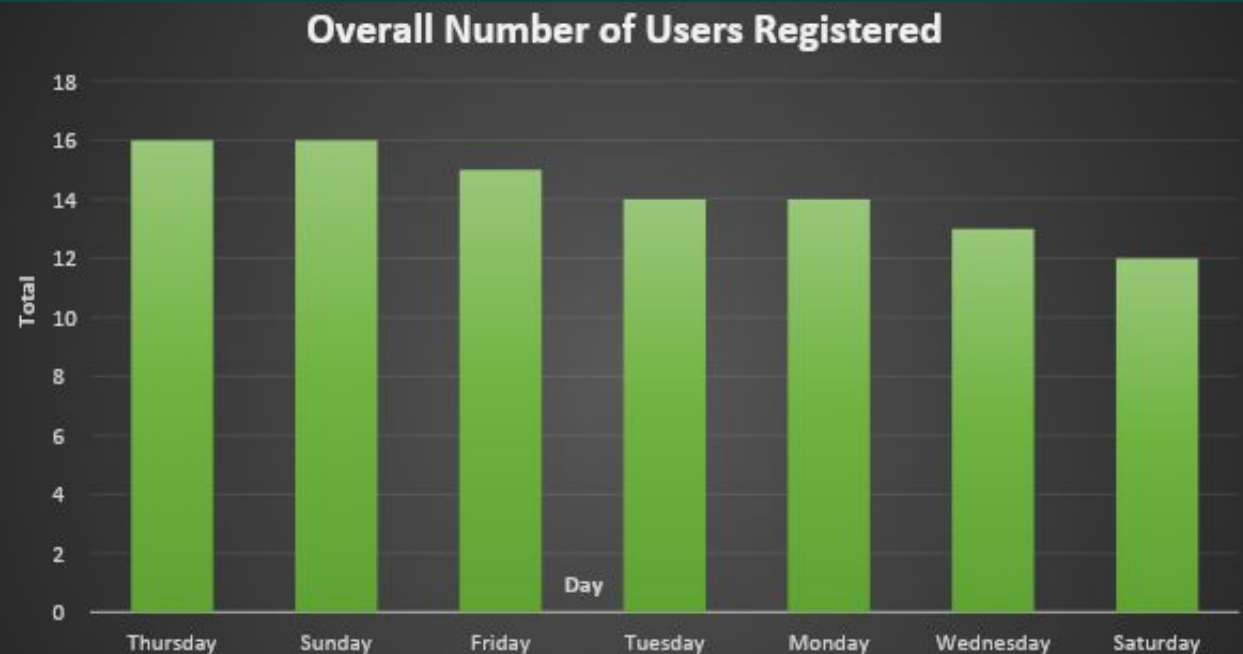
The team wants to find out the optimal day of the week to launch ads. Find the weekday with the highest number of registrations on Instagram. Offer considerations about ad campaign timing.

Most users registered on Thursday and Sunday with 16 registrations. Ad campaigns should therefore be started on these two days by the

Result:

day	total
Thursday	16

day	total
Thursday	16
Sunday	16
Friday	15
Tuesday	14
Monday	14
Wednesday	13
Saturday	12



B. INVESTOR METRICS

1. USER ENGAGEMENT:

Program/Query:

```
SELECT (SELECT Count(id)
FROM photos) / (SELECT Count(DISTINCT user_id)
FROM photos) AS Average_posts_per_User,

(SELECT Count(id)
FROM photos) / (SELECT Count(id)
FROM users) AS Ratio_of_Total_Posts_to_Total_Users;
```

Investors are interested to know whether users continue being active on Instagram or make less number of posts.

- Find the mean number of posts per user on Instagram.
- Also indicate the number of total photos on Instagram divided by total users.
- $(\text{SELECT Count(id) FROM photos}) / (\text{SELECT Count(DISTINCT user_id) FROM photos})$ AS Average_posts_per_User section of the query determines the average posts per user. It achieves this by dividing the total number of posts ($\text{Count(id) FROM PHOTOS}$) with the total distinct users who have made a post ($\text{Count(DISTINCT user_id) FROM PHOTOS}$).
- In this case, the $(\text{SELECT Count(id) FROM photos})$ selects the count of total number of photos from the photos table while $(\text{SELECT Count(DISTINCT user_id) FROM photos})$ select counts for users that are distinct who posted at least one photo.
- $(\text{SELECT Count(id) FROM photos}) / (\text{SELECT Count(id) FROM users})$ AS Ratio_of_Total_Posts_to_Total_Users is the part of the query that computes the proportion of all posts to total number of users by dividing the number of posts made ($\text{Count(id) FROM photos}$) with the total number of users ($\text{Count(id) FROM users}$).

B. INVESTOR METRICS

USER ENGAGEMENT:

- Investors are interested to know whether users continue being active on Instagram or make less number of posts.
- Find the mean number of posts per user on Instagram.
 - Also indicate the number of total photos on Instagram divided by total users.

Result:

Average_posts_per_User	Ratio_of_Total_Posts_to_Total_Users
6.9459	5.1400

total_photos_divide_total_users
5.1400

- The mean number of posts per user for Instagram is 6.9459.
- On dividing the aggregate count of photos on Instagram by the overall number of users, we get 5.1400

B. INVESTOR METRICS

2. Bots and Fake Accounts:

Program/Query:

```
select user_id, username, count(*) as total_likes_per_user
from users
inner join likes
on users.id = likes.user_id
group by likes.user_id
having total_likes_per_user = (select count(*) from photos);
```

The investors need to understand whether the platform is filled with fake and dummy accounts.

- Find out users who have shown likes to each and every photo on the site, because such activities are not common for a normal user.
- Initially, we choose the user_id column in the photos table.
- This is then followed by selecting the username column from users table
- Next, we choose the count function and select the count(*) to determine the total number of likes from the likes table
- Next, we inner join users and likes table based on users.id and likes.user_id, using the on function/clause.
- Then with the function group by, we group the required output table using likes.user_id .
- We then find the values from the count(*) on photos with similar values with the total_likes_per_user.

B. INVESTOR METRICS

Bots and Fake Accounts:

The investors need to understand whether the platform is filled with fake and dummy accounts.

- Find out users who have shown likes to each and every photo on the site, because such activities are not common for a normal user.

Result:

user_id	username	total_likes_per_user

username	num_likes

No user throughout the site has liked all photos. Accordingly, using the provided dataset, it is not feasible to have a single bot account.

Results and the Impacts of the Analysis:

- By running relevant SQL queries, we can derive **Business insights** for various teams such as product, marketing, and development by tracking how users interact with our digital products (software or mobile application) through the analysis of data stored on our database.
- However, more importantly, it is the general **business goals** and specific needs of different departments in an organization that is essential as running SQL queries to get insights is just one aspect of the same.
- Marketing team can set rewards for most loyal customers, send **promotional emails** to inactive users as well as to declare the winner during a contest; can use popular hashtags and most active day of the week for brand promotions or initiating the **advertising campaigns**.

- One of the very good growth success metric is the level of ***user engagement***.
- Company can eliminate the bots and fake accounts from the platform in order to ***improve user experience***.
- Along with the information derived from the given dataset, the best day to conduct ***ad campaigns*** can depend on other factors such as the audiences that are targeted, type of ads chosen and the ***user activity patterns*** on the Instagram, among others.