





Business problem and details of dataset

The task is to analyze Twitter usage patterns among different age groups to understand how younger and older generations engage with the platform. The goal is to uncover insights that can inform marketing strategies, user engagement plans, and product development.

The dataset includes:

- 1. Tweets: This table comprises user_id , tweet
- 2. Users: This table includes user_id , username , and user
- 3. Followers: This table contains user_id and Requirements: Link. timestamp, and tweet text. age. follower_id pairs

Platform used

Deepnote

Upload datasets

The three datasets were dragged and dropped into the file section of Deep note. Since the SQL can not directly interact with CSV files, therefore, tables were created and CSV files were fed into them.

Creating tables

3 tables - users, tweets and followers were created.

The tables were created using the below format:

```
CREATE TABLE table_name (
column1 datatype constraint,
column2 datatype constraint,
column3 datatype constraint,
....
);
```

Appropriate constraints were added while creating the tables to ensure the accuracy and reliability of data.

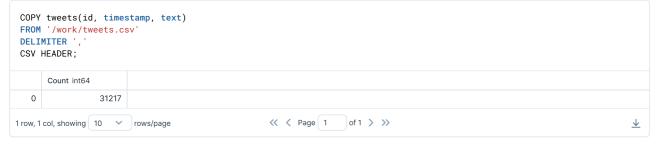


Inserting data to the tables

The format used on this platform of DeepNote is as follows:

COPY table_name(column1, column2......)
FROM '<path to csv file'
DELIMITER ','
CSV HEADER;







Describing tables and displaying first 5 rows

Tables were described using the keyword 'describe'. The output tells us about the column name, column datatype, if it allows null values or not, if it is primary key or foreign or none, any default value, and any extra value.

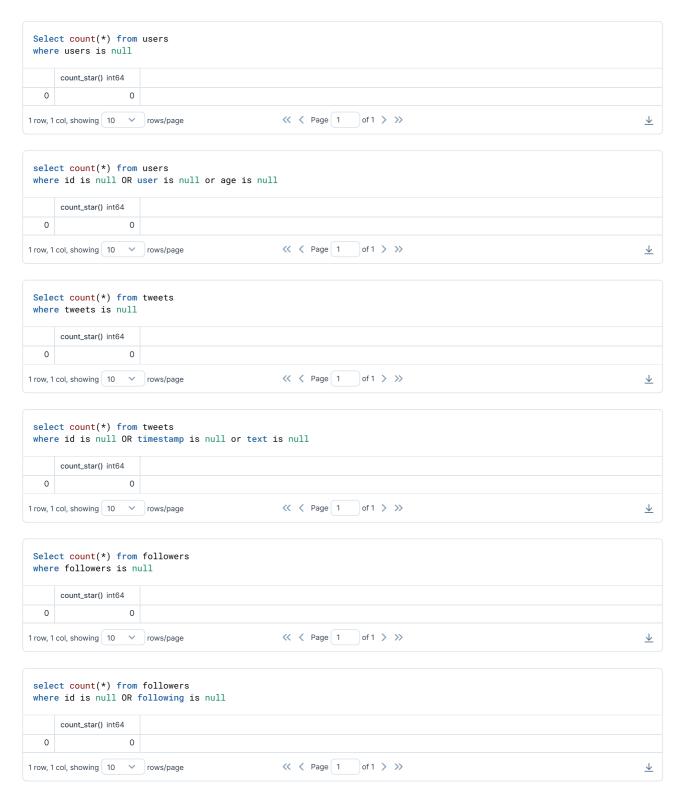
The first 5 rows were displayed using the the keyword 'limit' which only displays 5 rows.





Checking for null values

We aim to count the number of rows from tables where there are any null values. This can be done at the level of entire table or column wise. Both the ways are discussed.



We conclude that there are no null values in any of the tables.

Data analysis

Tweet count by age group

For the above query we consider the tables user and tweets. The users table has id, user and age. The tweets table has id, timestamp and text. The id column is the common between these two tables.

Therefore, to get the number of tweets age wise, we need to inner join (because only the common ids would be considered for counting) two mentioned tables on id column, count the text column and group by age. We further order the output in descending order using the 'order by' and 'desc'.

The format for inner join is:

SELECT column_name(s)
FROM table1
INNER JOIN table2
ON table1.column_name = table2.column_name;



Young people tweet more as compared to old people

Number of followers by age group

We consider the tables followers and users.

Followers has id and following and users has id, user, age as columns. The two tables can be inner joined on id column, grouped by age, and order by average of number of followers.

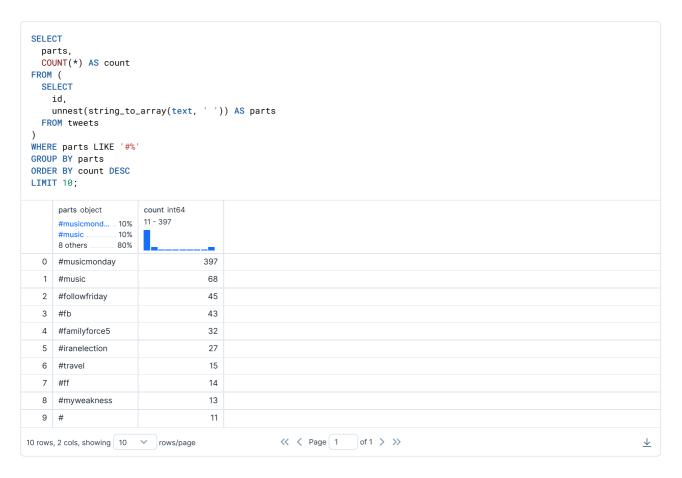
To see the number of followers, we counted the 'following' column from followers table as it would give the person who has been followed and by how many.



Top 10 hashtags

To achieve this, the text column in tweets data set was divided into words and words starting with # were counted and sorted in decending order.

The subquery SELECT unnest(string_to_array(t.text, '')) AS parts extracts individual words from the text field in tweets by splitting the text on spaces and unnesting the resulting array into separate rows.



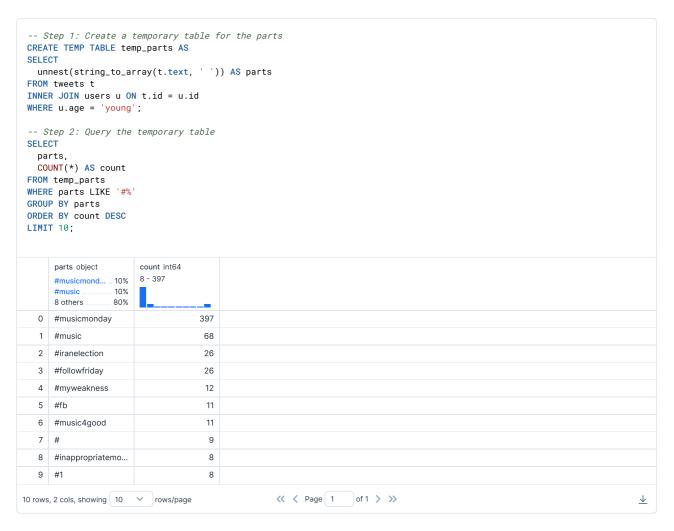
The top 10 hashtags used by younger users.

To count the hashtags, we used tweets table.

we need to extract hashtags from tweets.text. For this we had to read the text column from tweets table, search for words starting with '#' through regular expression matching.

To group it by age group, we inner joined with users table on id.

Make a temporary table which will bifurcate the words in the tweets.text column and count the words starting with #. This temporary table was queried for the count of top 10 hashtag used by young people



Whether young or older users generally follow users in their same age group.

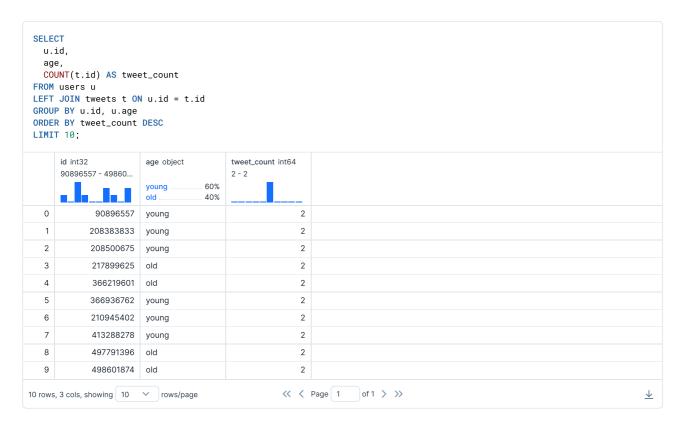
We used followers table to find of the id and following belong to same age group. If yes then it was considered as 1 and number of instances of 1s were averaged.



We observe that more older people follow users in their own age group.

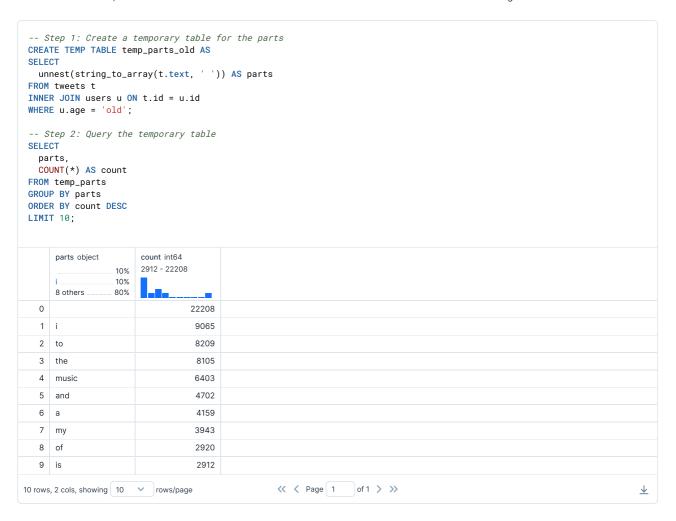
Most Active Users by Age Group

The aim is to find the user_id with maximum tweets. We used tweets and user tables.

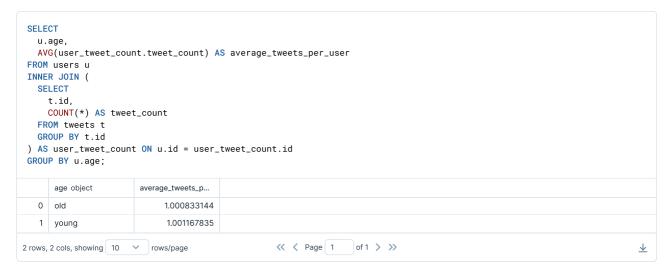


Most Common Words in Tweets by Older Users

From the tweets table, the text column was divided into words and words were counted and sorted in descending order.



Average Number of Tweets Per User by Age Group



We observe that there is not much difference in the number of tweets done by both old and young people.