



**34\_** **Histogram of Tweets**

write a pyspark code to obtain a histogram of tweets posted per user in 2022. Output the tweet count per user as the bucket and the number of Twitter users who fall into that bucket.In other words, group the users by the number of tweets they posted in 2022 and count the number of users in each group.

**Difficult Level :** EASY

**DataFrame:**

**# Define the schema for the tweets DataFrame**

**schema = StructType([**

**StructField("tweet\_id", IntegerType(), True),**

**StructField("user\_id", IntegerType(), True),**

**StructField("msg", StringType(), True),**

**StructField("tweet\_date", StringType(), True)**

**])**

**# Create the tweets DataFrame**

**data = [**

**(214252, 111, 'Am considering taking Tesla private at $420. Funding secured.', '2021-12-30 00:00:00'),**

**(739252, 111, 'Despite the constant negative press covfefe', '2022-01-01 00:00:00'),**

**(846402, 111, 'Following @NickSinghTech on Twitter changed my life!', '2022-02-14 00:00:00'),**

**(241425, 254, 'If the salary is so competitive why won’t you tell me what it is?', '2022-03-01 00:00:00'),**

**(231574, 148, 'I no longer have a manager. I can\'t be managed', '2022-03-23 00:00:00')**

**]**

|  |
| --- |



**INPUT**

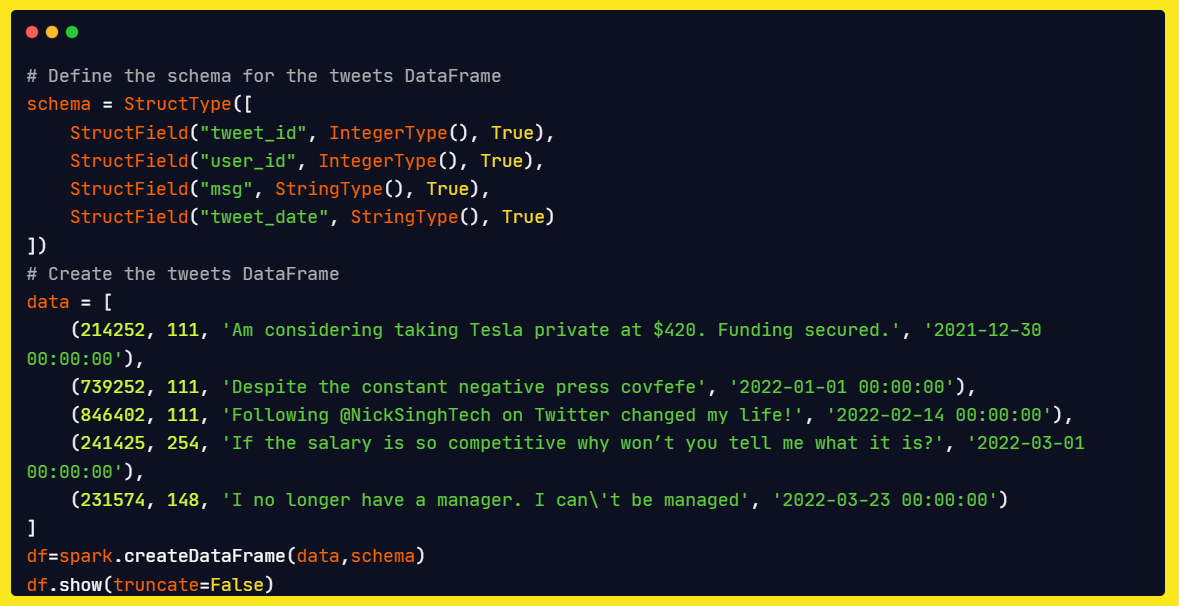
| **INPUT** | | | |
| --- | --- | --- | --- |
| **TWEET\_ID** | **USER\_ID** | **MSG** | **TWEET\_DATE** |
| 214252 | 111 | Am considering taking Tesla private at $420. Funding secured. | 2021-12-30 0:00:00 |
| 739252 | 111 | Despite the constant negative press covfefe | 2022-01-01 0:00:00 |
| 846402 | 111 | Following @NickSinghTech on Twitter changed my life! | 2022-02-14 0:00:00 |
| 241425 | 254 | If the salary is so competitive why won’t you tell me what it is? | 2022-03-01 0:00:00 |
| 231574 | 148 | I no longer have a manager. I can't be managed | 2022-03-23 0:00:00 |

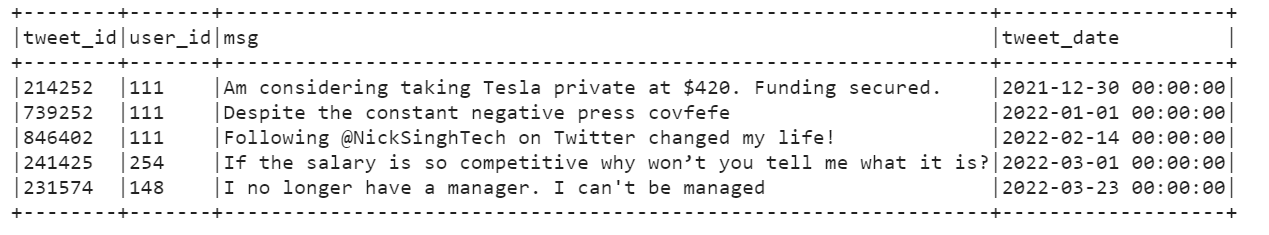
**OUTPUT**

| **OUTPUT** | |
| --- | --- |
| **BUCKET** | **USER\_NUM** |
| 1 | 2 |
| 2 | 1 |

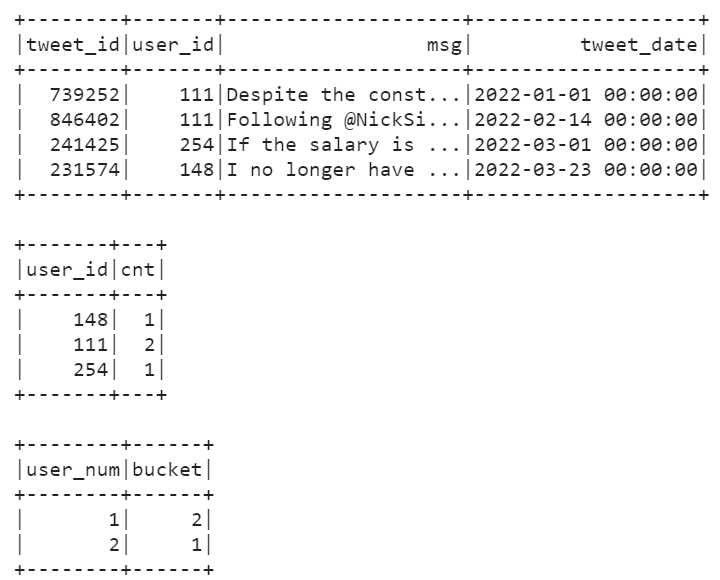


****

****

****

****

****

