DB Report

The table provided was not normalized and having following columns created_at, text, tweet_id, in_reply_to_screen_name, in_reply_to_status_id, in_reply_to_user_id, retweet_count, tweet_source, retweet_of_tweet_id, hashtag1, hashtag2, hashtag3, hashtag4, hashtag5, hashtag6, user_id, user_name, user_screen_name, user_location, user_utc_offset, user_time_zone, user_followers_count, user_friends_count, user_lang, user_description, user_status_count, user_created_at

There are total five subject areas:

- 1. Tweet_User
- 2. Tweet_Reply
- 3. Tweet_Retweet
- 4. Tweet_source
- 5. Tweet_hashtag

For each subject area is required individual table as per normalization rule. To remove redundancy.

However, we can create a table for language but in this situation is not required.

After performing 3rd Normalized from tables which will populate looks like this:

```
Table Tweet_User
(
user_id,
user_name
user_screen_name
user_location
user_utc_offset,
user_time_zone
user_followers_count,
user_friends_count,
user_lang
user_description
user_status_count,
user_created_at
)
TABLE Tweet_Reply
(
     tweet_Reply_id,
     tweet_id ,
     in_reply_to_screen_name,
```

in_reply_to_status_id,

in_reply_to_user_id

)

```
Table Tweet_Retweet
```

TABLE Tweet_source

```
tweet_id ,
tweet_source,
text,
retweet_id,
tweet_Reply_id,
user_id ,
created_at
)
```

TABLE Tweet_hashtag

```
hashtagid, hashtag, tweet_id
```

)

Functional Dependencies:

```
user_id -> user_name , user_screen_name
user_lang , user_created_at -> user_id
user_time_zone -> u user_id

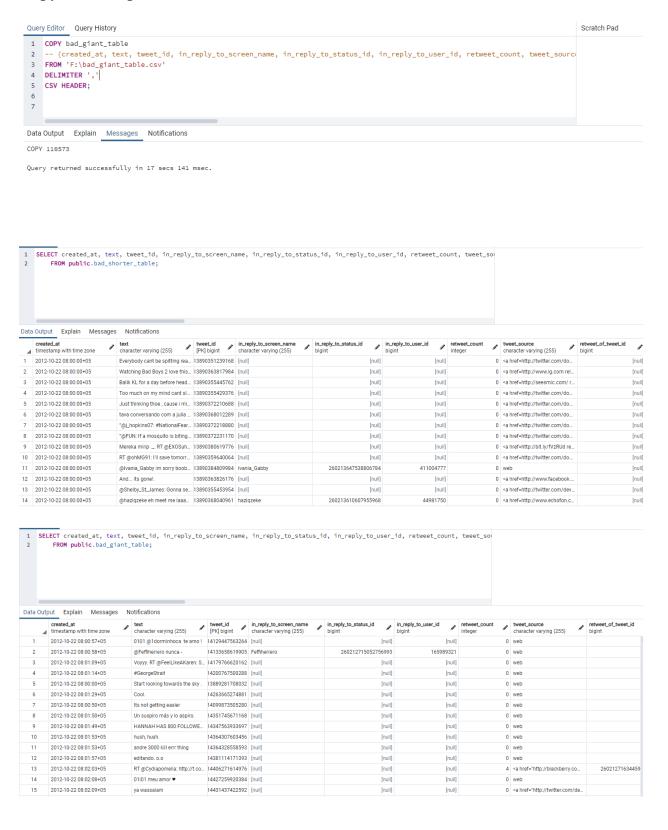
tweet_Reply_id -> tweet_id

retweet_id -> retweet_of_tweet_id

hashtagid-> hashtag
```

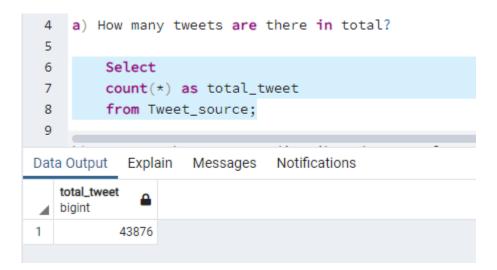
Note: Before loading data into final tables I've removed duplicated while populating data into new tables.

Copy the data provided:



SQL Queries:

1



18 th

```
10 b) How are these tweets distributed across languages? Write a query that shows,
11 for every language ( user_lang ) the number of tweets in that language.
12
    Select
user_lang,
count(*) as total_tweet_Per_Language
  from Tweet_user
13
14
15
group by 1;
18
Data Output Explain Messages Notifications
user_lang

∠ character varying (10)
                    1 fr
                                            221
2
                                            121
3 en
                                           69279
4 fi
                                             1
5 ru
                                            382
6
                                             2
7
   ja
                                           9596
8 sv
                                              2
9 fil
                                              9
10 eu
                                              1
11 de
                                             71
12 nl
                                             60
13 id
                                           1360
14 zh-cn
                                             21
15 zh-tw
                                             14
16 ar
                                           1328
17 no
                                             1
```

217

```
c) compute, for each language, the fraction of total tweets that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that have that language setting, as well as the fraction of the number of users that language setting, as well as the fraction of the number of users that language setting, as well as the fraction of the number of users that language setting, as well as the fraction of users that language setting, as well as the fraction of users that language setting, as well as the fraction of users that language setting, as well as the fraction of users that language setting, as well as the fraction of users that language setting, as well as the fraction
```

Data Output	Explain	Messages	Notifications
	long		connected

4	user_lang character varying (10)	user_id integer	fraction_user_lang numeric	fraction_user numeric
1	en	226803667	0.000022791503327559485824	0.000022791503327559485824
2	ja	170019464	0.000022791503327559485824	0.000022791503327559485824
3	pt	310658484	0.000022791503327559485824	0.000022791503327559485824
4	es	275738606	0.000022791503327559485824	0.000022791503327559485824
5	en	232351994	0.000022791503327559485824	0.000022791503327559485824
6	en	51323231	0.000022791503327559485824	0.000022791503327559485824
7	ja	152568208	0.000022791503327559485824	0.000022791503327559485824
8	en	26862229	0.000022791503327559485824	0.000022791503327559485824
9	en	166209840	0.000022791503327559485824	0.000022791503327559485824
10	en	58362155	0.000022791503327559485824	0.000022791503327559485824
11	en	361627962	0.000022791503327559485824	0.000022791503327559485824
12	en	239788600	0.000022791503327559485824	0.000022791503327559485824
13	en	29499656	0.000022791503327559485824	0.000022791503327559485824
14	en	228236772	0.000022791503327559485824	0.000022791503327559485824
15	en	244262497	0.000022791503327559485824	0.000022791503327559485824
16	en	106826585	0.000022791503327559485824	0.000022791503327559485824
17	en	59858229	0.000022791503327559485824	0.000022791503327559485824

```
39 a) What fraction of the tweets are retweets ?
40
41 Select
42
          t.tweet_id,
           (\textbf{count}( \ \texttt{rt.tweet\_id}) \ \star \ \textbf{1.0} / \ (\textbf{select count}(\star) \ \ \textbf{from} \ \ \texttt{Tweet\_source}) \ ) \ \textbf{as} \ \ \texttt{Fraction\_Tweet}
43
44
           from Tweet_source t
45
         inner join Tweet_Retweet rt
46
           on t.retweet_id = rt.retweet_id
47
          group by 1
48
49
50
```

Data Out	Data Output Explain Messages Notifications		
4	tweet_id [PK] bigint	fraction_tweet numeric	
1	260214557283336192	0.000022791503327559485824	
2	260219334582992896	0.000022791503327559485824	
3	260216721514852352	0.000022791503327559485824	
4	260219468796530688	0.000022791503327559485824	
5	260221456879857665	0.000022791503327559485824	
6	260217409393274880	0.000022791503327559485824	
7	260217124168036352	0.000022791503327559485824	
8	260220852929433600	0.000022791503327559485824	
9	260215245136621568	0.000022791503327559485824	
10	260219997274656768	0.000022791503327559485824	
11	260217317135360000	0.000022791503327559485824	
12	260214066545586176	0.000022791503327559485824	
13	260217702973571072	0.000022791503327559485824	
14	260214725038727168	0.000022791503327559485824	
15	260219540091326464	0.000022791503327559485824	
16	260214804734685188	0.000022791503327559485824	

```
b) Compute the average number of retweets per tweet.
51
52
53
   Select
54
        t.tweet_id,
55
        avg( rt.retweet_count) as Avg_Retweet
56
        from Tweet_source t
57
        inner join Tweet_Retweet rt
58
        on t.retweet_id = rt.retweet_id
59
        group by 1
60
61
```

Data Output Explain Messages Notifications

4	tweet_id [PK] bigint	avg_retweet numeric
1	260214557283336192	1.0000000000000000000000000000000000000
2	260216721514852352	1.0000000000000000000000000000000000000
3	260219468796530688	1.0000000000000000000000000000000000000
4	260221456879857665	0.0000000000000000000000000000000000000
5	260217409393274880	1.0000000000000000000000000000000000000
6	260217124168036352	2.0000000000000000
7	260220852929433600	9.0000000000000000
8	260215245136621568	1.0000000000000000000000000000000000000
9	260219997274656768	1.0000000000000000000000000000000000000
10	260217317135360000	1.0000000000000000000000000000000000000
11	260214066545586176	0.0000000000000000000000000000000000000
12	260217702973571072	1.0000000000000000000000000000000000000
13	260214725038727168	0.0000000000000000000000000000000000000
14	260219540091326464	2.0000000000000000
15	260214804734685188	10.0000000000000000
16	260214544700432384	3.0000000000000000

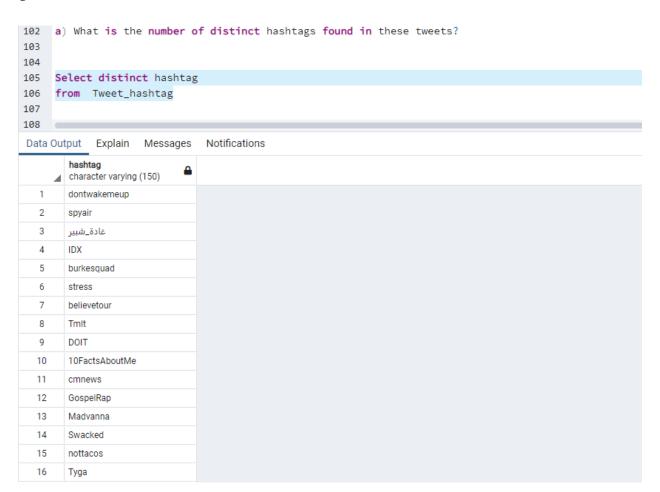
```
64 c) What fraction of the tweets are never retweeted?
66 Select
67
      t.tweet_id,
       (count( t.tweet_id) * 1.0/ (select count(*) from Tweet_source) ) as Fraction_Tweet
68
69
        from Tweet_source t
70
       left join Tweet_Retweet rt
        on t.retweet_id = rt.retweet_id
71
72
        where t.retweet_id is null
73
        group by 1
74
75
```

Data Ou	tput Explain	Messages Notificat	ions
4	tweet_id [PK] bigint	•	fraction_tweet numeric
1		260215006052876288	0.000022791503327559485824
2		260217652667092993	0.000022791503327559485824
3		260216780260274176	0.000022791503327559485824
4		260215683760140291	0.000022791503327559485824
5		260214263677845504	0.000022791503327559485824
6		260214435635933184	0.000022791503327559485824
7		260219183583866880	0.000022791503327559485824
8		260220056015884288	0.000022791503327559485824
9		260219734212091904	0.000022791503327559485824
10		260214653714587648	0.000022791503327559485824
11		260219686917128193	0.000022791503327559485824
12		260217921098362880	0.000022791503327559485824
13		260215064773160960	0.000022791503327559485824
14		260215127704494080	0.000022791503327559485824
15		260221301724168192	0.000022791503327559485824

```
76 d) What fraction of the tweets are retweeted fewer times than the average number of retweets (and what does this say about the distribution)?
 78 select TT.tweet_id,
79 TT.Fraction_Tweet,
80 TT.Avg_Retweet from
 81 (
 82 Select
      t.tweet_id,
  (count( rt.tweet_id) * 1.0/ (select count(*) from Tweet_source) ) as Fraction_Tweet,
  avg( rt.retweet_count) as Avg_Retweet
  from Tweet_source t
  inner join Tweet_Retweet rt
  on t.retweet_id = rt.retweet_id
  group by 1
) TT
 83
       t.tweet_id,
 84
 85
 86
 87
 88
 89
90
 90 ) TT
91 where TT.Fraction_Tweet < TT.Avg_Retweet
```

ata	Output	Explain	Messages	Notifications

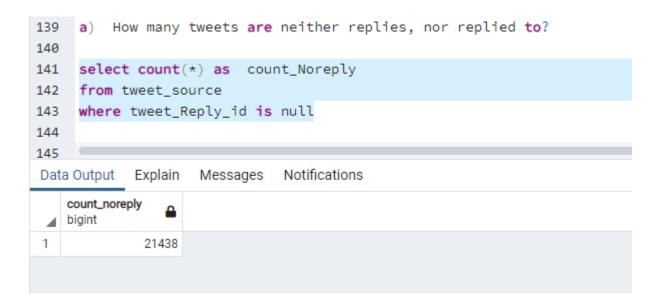
4	tweet_id [PK] bigint	fraction_tweet numeric	avg_retweet numeric
1	260213890351239169	0.000022791503327559485824	106.0000000000000000
2	260213890351239170	0.000022791503327559485824	1.0000000000000000000000000000000000000
3	260213890351255552	0.000022791503327559485824	5027.0000000000000000
4	260213890355453952	0.000022791503327559485824	112.00000000000000000
5	260213890359635968	0.000022791503327559485824	1.0000000000000000000000000000000000000
6	260213890368012291	0.000022791503327559485824	2.0000000000000000
7	260213890368016384	0.000022791503327559485824	7.0000000000000000
8	260213890368024576	0.000022791503327559485824	1.0000000000000000000000000000000000000
9	260213890368040962	0.000022791503327559485824	80.0000000000000000
10	260213890372222976	0.000022791503327559485824	30.000000000000000



```
110 b) What are the top ten most popular hashtags, by number of usages?
111
112
113 Select hashtag,
114 count( hashtag ) as Top_Hashtags
115 from Tweet_hashtag
116 group by 1
117 order by 2 desc limit 10
118
119
400 A Martin a manual adular. For each January the transfer and a martin broken de the
Data Output Explain Messages Notifications
    hashtag
                           top_hashtags
  bigint
    ReasonsIFailAtBeingAGirl
                                      467
 2
    RED
                                      205
    oomf
                                      188
   HonestyHour
                                      172
    TeamFollowBack
                                      139
 5
                                      130
    EresGuapaSi
 6
    10PeopleYouTrulyLove
                                      126
 7
 8
    TweetLikeAGirl
                                      98
    ImSingleBecause
                                      97
10 WeAllGotThatOneFriend
                                      96
```

```
120 c) Write a query giving, for each language, the top three most popular hashtags in that language.
122 Select hashtag,user_lang, count( hashtag )
123 from Tweet_hashtag
124 join tweet_source
on Tweet_hashtag.tweet_id = tweet_source.tweet_id
126 join tweet_user
127     on tweet_source.user_id = tweet_user.user_id
128 group by 1,2
129 order by 3 desc limit 3
Data Output Explain Messages Notifications
 hashtag character varying (150) a user_lang character varying (10) a count bigint
 1 ReasonsIFailAtBeingAGirl en
                                                     184
 2 RED
                                                     84
 3 MagicMike
                                                     78
```

11



```
147 b) If a user user1 replies to another user user2 , what is the probability that they have the same language setting?
148
149 Select user_lang,
150 COUNT(user_lang)*1.0/(SELECT count(tweet_id) FROM tweet_source) as Prob_UselangSetting
151 from tweet_source as ts
152 JOIN tweet_reply as tu
153  on ts.tweet_Reply_id= tu.tweet_Reply_id
154 join tweet_user as tul
on ts.user_id=tu1.user_id
156 group by 1
157
158
Data Output Explain Messages Notifications
                        prob_uselangsetting numeric
    user_lang
  1 ar
                                                0.00423921961892606436
 2 ca
                                             0.000022791503327559485824
 3 de
                                                0.00027349803993071383
 4 en
                                                0.32728598778375421643
 5
                                                0.08601513355820949950
    es
 6
    fil
                                             0.000045583006655118971647\\
    fr
                                                0.00109399215972285532
 8
    id
                                                0.00556112681192451454
 9
                                                0.00013674901996535691
   it
 10 ja
                                                0.06212963807092715836
 11 ko
                                                0.00394293007566779105
 12 msa
                                             0.000045583006655118971647
 13 nl
                                                0.00025070653660315434
 14 no
                                             0.000022791503327559485824
```

```
159 c) How does this compare to the probability that two arbitrary users have the same language setting?
160 Throughout, you may create views that support your queries.
161
162 WITH lang_setting AS(
163 Select ts.user_id,
164 COUNT(user_lang)*1.0/(SELECT count(tweet_id) FROM tweet_source) as Prob_UselangSetting
165 from tweet_source as ts
166 JOIN tweet_reply as tu
167 on ts.tweet_Reply_id= tu.tweet_Reply_id
168 join tweet_user as tul
169 on ts.user_id=tu1.user_id
170 AND tu1.user_lang = tu1.user_lang
171 group by 1
172
173 Select * from lang_setting ORDER BY RANDOM() LIMIT 2
174
Data Output Explain Messages Notifications
                   prob_uselangsetting
numeric
   user_id

    integer

 1
               548028773
                                         0.000022791503327559485824
 2
               746909557
                                          0.000022791503327559485824
```