



Store, Analyze and Visualize “tweets” from Twitter

Team members:

Gulnoza Khakimova

Khushbu Kolhe

Lakshmi Chamala



Instructor: Dr. Praveen Rao

CS5540 Principles of Big Data Management

Phase 2

Design Steps:

In this phase we have executed queries on collected tweets (in phase 1) with following topics: "Bitcoin", "Forbes", "Winter Olympics" and displayed their visualization.

Libraries:

- Spark Core - org.apache.spark:spark-core_2.11:2.0.0.2
- Spark SQL - org.apache.spark:spark-sql_2.11:2.0.0.2

APIs:

Twitter public REST APIs - GET followers/ids.

Resource URL: <https://api.twitter.com/1.1/followers/ids.json> returns a collection of userIDs for every user following the specified user.

Technologies:

- Scala – to run Spark Programs.
- HTML5, CSS3 – to design user interface and front-end development.
- JavaScript – to do API calls and visualize.
- IntelliJ
- Spark
- Tableau - for visualization

Team members contribution:

- 1.Khushbu Kohle - Queries: 1, 2, 3 ,10
- 2.Gulnoza Khakimova - Queries: 4, 5, 6
- 3.Lakshmi Chamala - Queries: 7, 8, 9,

Tableau software was used to visualize output results.

Queries:

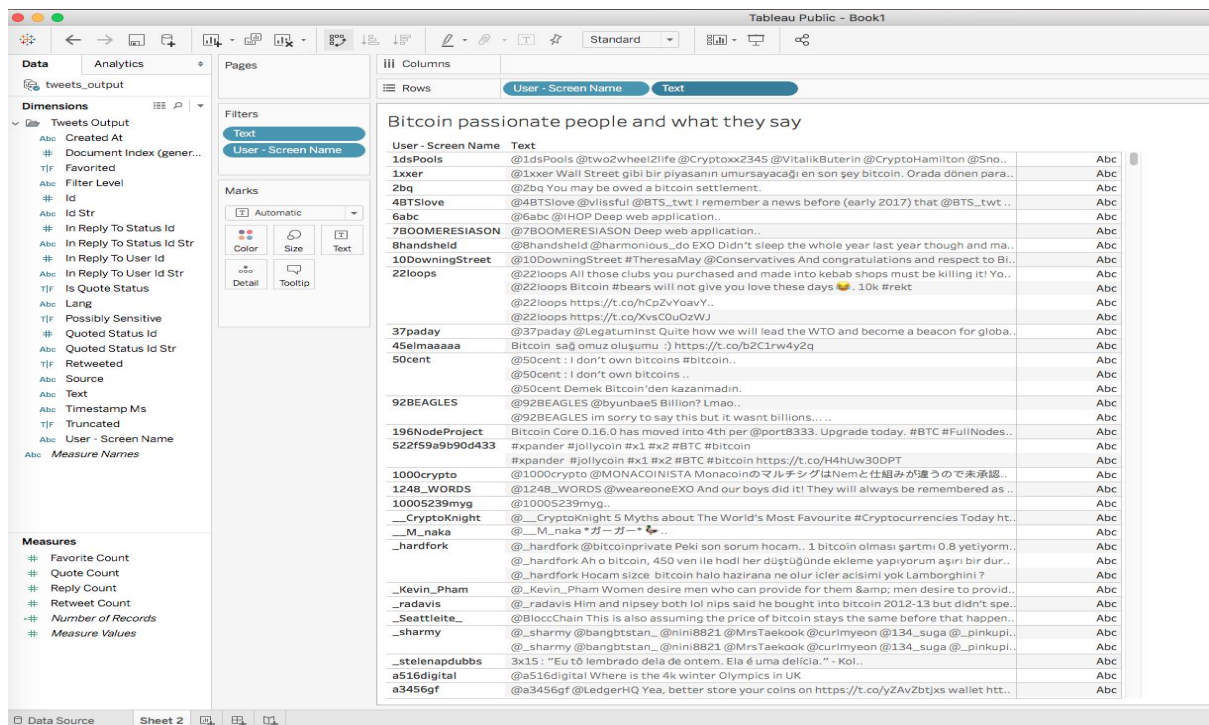
Query 1: Total tweets from different time zones

Command : SELECT user.time_zone,count(text) AS Total FROM Bitcoin WHERE user.time_zone IS NOT NULL GROUP BY user.time_zone ORDER BY Total DESC LIMIT 10

Output 1:

time_zone	Total
Pacific Time (US ...	23143
Eastern Time (US ...	2886
London	2288
Jakarta	1780
Bangkok	1431
Central Time (US ...	1355
Amsterdam	962
Seoul	951
Paris	922
Tokyo	911

Visualization 1:



Query 2: Bitcoin passionate people and what they say..

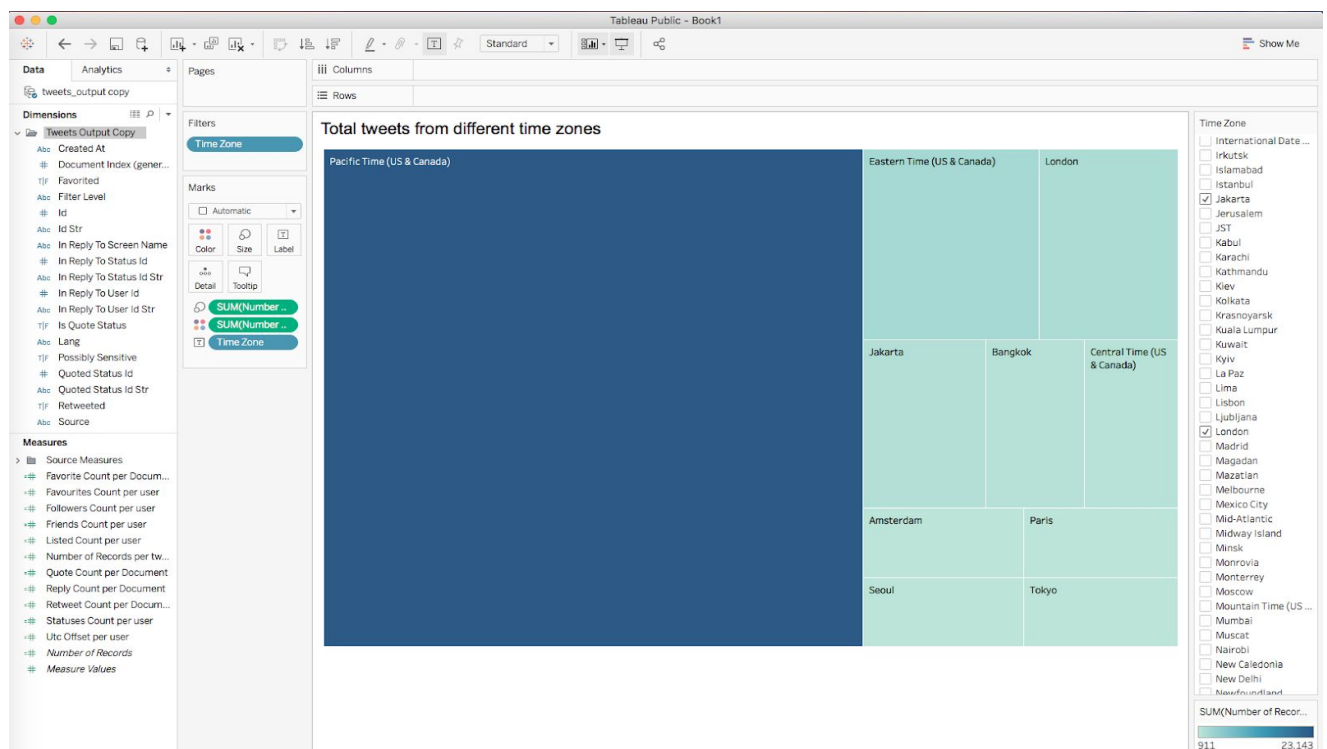
Command: SELECT user.name, text FROM Bitcoin WHERE text LIKE '%bitcoin%'

Output 2:

name	text
Ainsley Selwyn	#SpendMoneyWisely...
Colette Heron	#GetThatMoney How...
Cryptoo Currency	WizSec: Neither W...
infocryptos	Ethereum Price Te...
Luz Grove	#GoGetTheMoney Ho...
infocryptos	Bitcoin Buyers Ta...
BitBrokers Inc	Coinbase tells 13...
Samara Baker	#MoneyMature How ...
Ramil vale	RT @bitcoinus_io:...
Marsha Osborne	#MoneyTeam4DaWin ...
Уна Илья	RT @bitcoinus_io:...
ibrahim yildirim	RT @GizmenNalbant...
Ladonna Long	#EasyMoney How do...
Les Echos	Les banques veule...
Kirsty Desford	#selfieformoneysl...
Kevi	@Cointelegraph @m...
Corin	Much gratitude to...
Alicia Pole	#pokermoney Why b...
Yono	RT @coinspectator...
Willow Pinnock	#nodarkmoney How ...

only showing top 20 rows

Visualization 2:



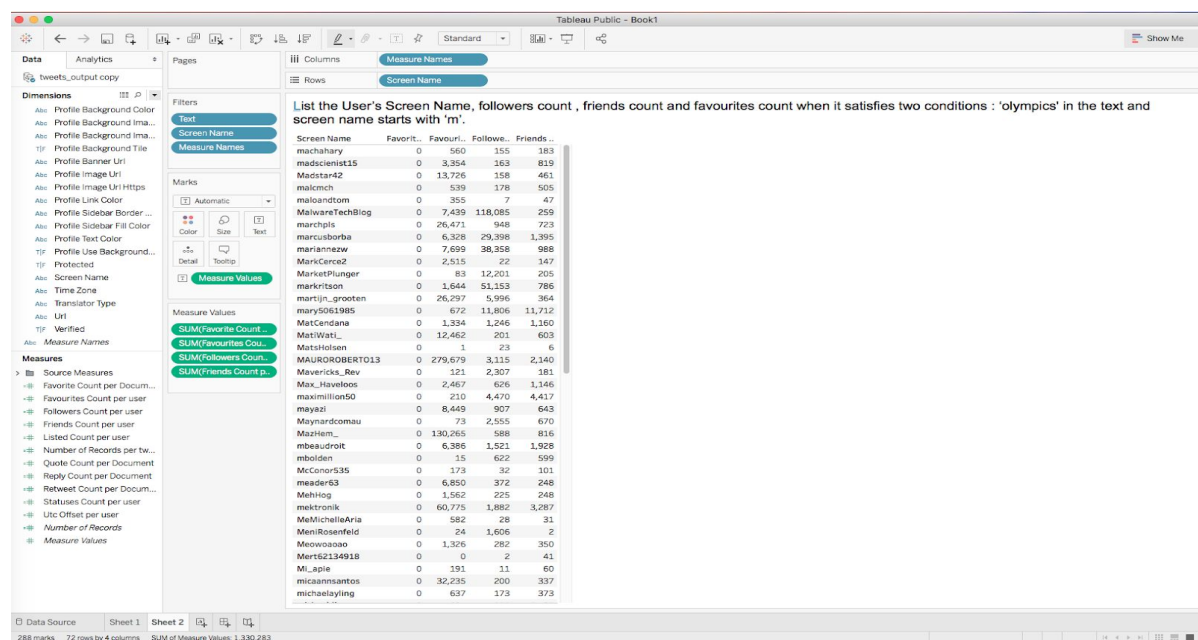
Query 3: List the User's Screen Name, followers count , friends count and favourites count when it satisfies two conditions : 'olympics' in the text and screen name starts with 'm'.

Command : SELECT user.screen_name, user.followers_count, user.favourites_count, user.friends_count FROM Bitcoin WHERE user.screen_name LIKE 'm%' AND text LIKE '%olympics%'

Output 3:

screen_name	followers_count	favourites_count	friends_count
michaela_16	497	19961	331
mareekitaz	573	21256	259
meonomous	432	12884	572
mikaajoy	176	1437	63
mochinbik	1603	71095	208
micaannsanatos	200	32235	337
myeoningdae	290	1534	432
minxgyu_k	51	872	70
melodyuh	248	33690	792
mizashuhada	2390	2264	368
marcwebber	6050	7919	4202
minthope0918	462	29628	1574
mrsohxx	129	13414	282
mathsaretrash	46	351	855
mona013000	243	13476	200
march9238	1247	9734	395
marshasofieya	27	189	211
mojgan_re	492	11802	466
maya_munoz_76_	208	6940	269

Visualization 3:



Query 4:

Count number of tweets based on location of the user:

"SELECT Count(id) AS Total, user.location AS Location FROM TweetTable WHERE user.location IS NOT NULL GROUP BY user.location ORDER BY COUNT(id) DESC"

Output 4:

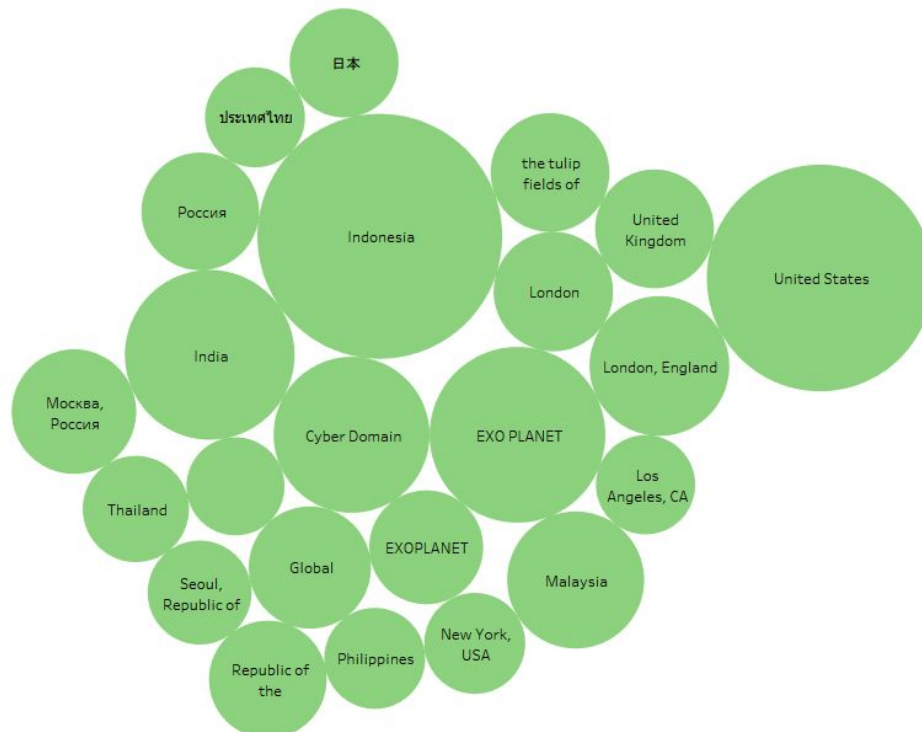
18/04/20 20:28:23 INFO DAGScheduler: Job 1 fi
18/04/20 20:28:23 INFO CodeGenerator: Code ge

Total	Location
1733	Indonesia
1602	United States
761	Cyber Domain
631	India
608	London, England
560	Malaysia
485	Москва, Россия
450	Global
441	the tulip fields ...
441	United Kingdom
429	Россия
426	Republic of the P...
415	London
386	EXO PLANET
368	日本
338	Seoul, Republic o...
321	New York, USA
306	ประเทศไทย
305	Los Angeles, CA
301	Ile-de-France, Fr...

only showing top 20 rows

18/04/20 20:28:23 INFO SparkSQLParser: Parse

Visualization 4:



Query 5:

Select users with most number of statuses who mentioned “Forbes” in their tweets:

```
"SELECT user.name AS Name, user.statuses_count AS Statuses FROM
TweetTable WHERE text LIKE '%forbes%' AND user.verified = True ORDER BY
user.statuses_count DESC"
```

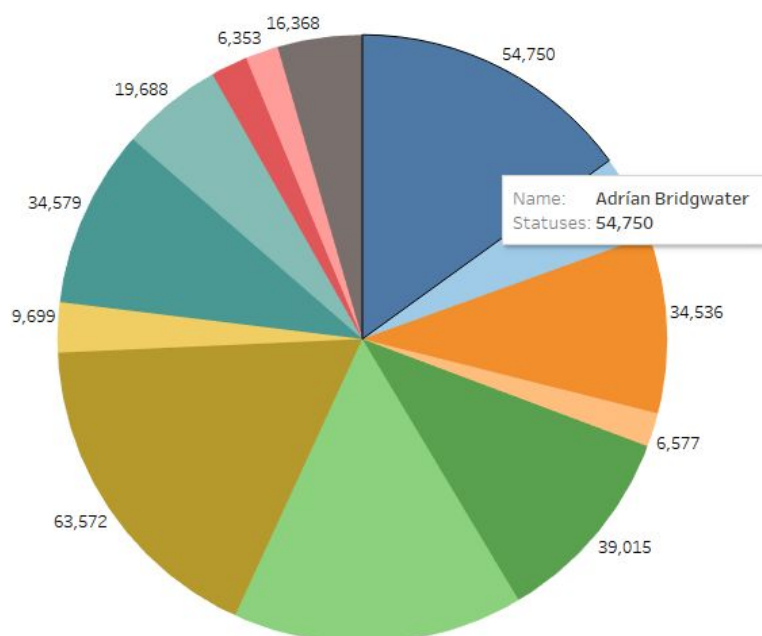
Output 5:

```
18/04/20 20:28:30 INFO CodeGenerator
18/04/20 20:28:30 INFO SparkSqlParse
```

Name	Statuses
Rebecca Enonchong	63572
Kyle Matthews	56273
Adrian Bridgwater	54750
SHEROES	34579
Commun.it	34536
stéphane koch	19688
Janet Novack	19508
Janet Novack	19507
TRACI BINGHAM	16368
Cheryl Richardson	16272
Sébastien AUDOUX	9699
Tech Data	7135
Dave in Osaka	6577
tedreed	6353

```
18/04/20 20:28:30 INFO FileSourceStr:
18/04/20 20:28:30 INFO FileSourceStr:
```

Visualization 5:

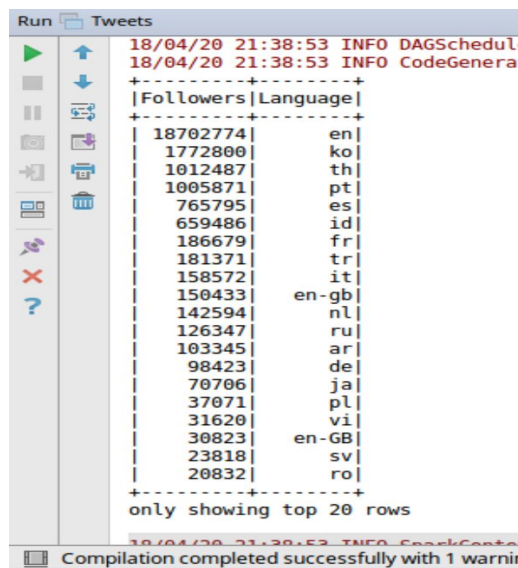


Query 6:

Select total number of follower count based user's language:

```
"SELECT SUM(user.followers_count) AS Followers, user.lang AS Language FROM
TweetTable WHERE retweeted_status.text LIKE '%Olympics%' GROUP BY
user.lang ORDER BY SUM(user.followers_count) DESC"
```

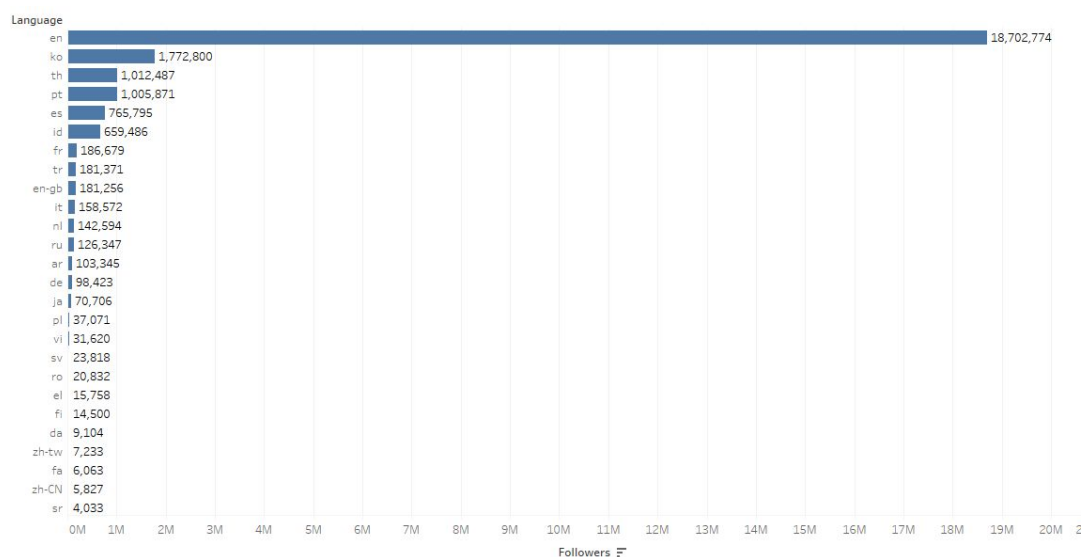
Output 6:



Followers	Language
18702774	en
1772800	ko
1012487	th
1005871	pt
765795	es
659486	id
186679	fr
181371	tr
158572	it
150433	en-gb
142594	nl
126347	ru
103345	ar
98423	de
70706	ja
37071	pl
31620	vi
30823	en-GB
23818	sv
20832	ro

only showing top 20 rows

Visualization 6:



Query 7:

Tweets based on No:of retweet status:

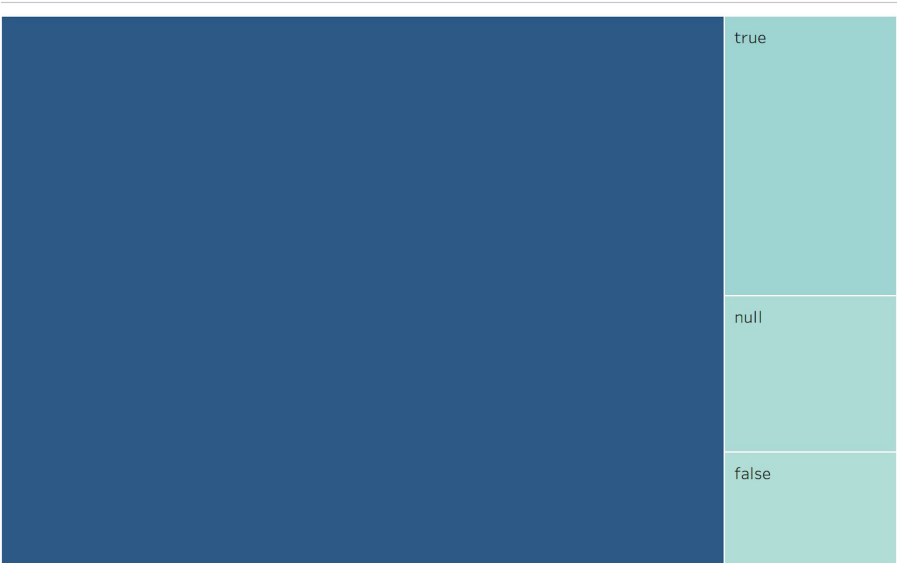
```
SELECT user.name,retweeted_status.truncated,retweet_count FROM TweetTable
```

Output 7:

새계정	false	0
.8/04/29 14:06:30 INFO TaskSchedulerImpl: Removed TaskSet 13.0, whose tasks have all completed, from pool		
loz	true	0
.8/04/29 14:06:30 INFO DAGScheduler: Job 7 finished: show at SparkTransformation.scala:64, took 0.352663 s		
Nico Braune	null	0
Ainsley Selwyn	null	0
구독계	true	0
chichi	true	0
beatrice	false	0
love yourself cloudy	true	0
Bety Eng	true	0
PORRNESIAN PARRAP...	true	0
.	true	0
WANCHU	true	0
Kamidzy	true	0
Nolympic Soldier	true	0
CryptoChloe	false	0
evi	true	0
Colette Heron	null	0
Cryptoo Currency	null	0
Polar_YT iHeart...	true	0
Cryptoo Currency	null	0

only showing top 20 rows

Visualization 7:



Query 8:

Tweets based on more no:of users in the particular month:

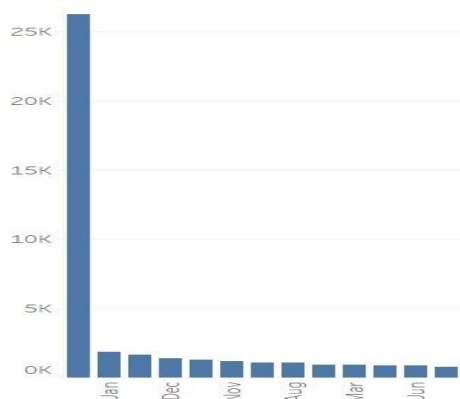
```
select user.name, SUBSTRING(user.created_at,5,3) AS  
month,SUBSTRING(user.created_at,12,8) AS time," + "Count(*) as cnt from  
TweetTable where SUBSTRING(user.created_at,5,3) like '%Dec%' group by user
```

Output 8:

name	month	time
patra	Feb	09:25:09
LiVe Store [hiatus]	Jan	14:52:18
Tatian	Jul	01:15:07
Mjanie Baron	Nov	08:53:59
babyjksinheaven	Jan	07:49:13
Александр Богатый	Sep	16:47:59
Roman	Sep	17:48:34
Work Hard, Play H...	Dec	09:30:58
Pavel	Feb	07:13:40
Putra97	Nov	16:05:21
angkasaku	Feb	14:15:23
🌱	Feb	15:24:22
b!uette ♡ 60 days...	Jan	04:25:59
Sahistya Dhanes	Oct	11:25:22
Bitcoin Retweet Bot	Jan	14:06:54
IT-virtual entity	May	23:52:40
cryptobeard	Oct	02:21:55
SC Howard #Boycot...	Mar	22:57:41
Selim	Jan	12:52:30
Shaherezade	Mar	00:38:55

only showing top 20 rows

Visualization 8:



Number of tweets in the particular month

Query 9:

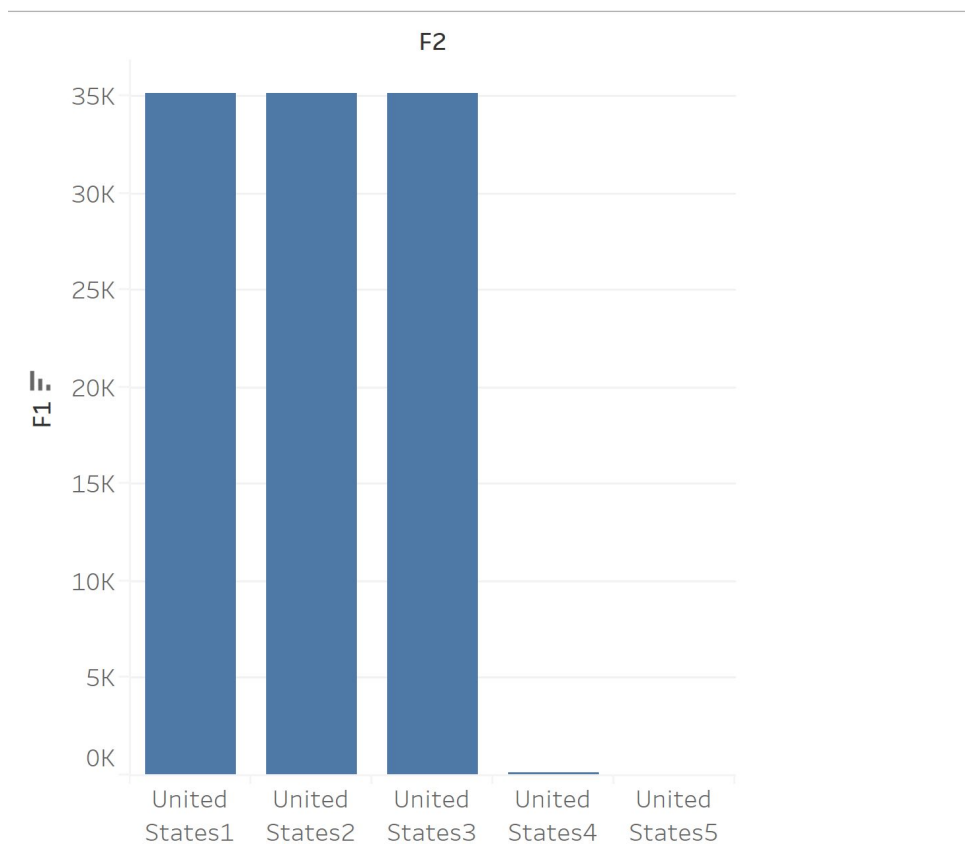
Tweets based on favourites count where location is US

select user.favourites_count,place.country from TweetTable where user.location='United States' and place.country is not null

output 9:

```
(35090,United States1)
(35090,United States2)
(35090,United States3)
(127,United States4)
(3,United States5)
```

visualization 9:



Query 10: Tweets based on verified users

Command: *SELECT count(user.verified) AS Verified_Users FROM Bitcoin WHERE user.verified=true*

output 10:

Verified_Users
643

visualization 10:

