PRINCIPLES OF BIG DATA PROJECT

TRAVEL

TEAM MEMBERS : Shreyaa Sridhar
Vinay Maturi
Hari

SOURCE CODE

https://drive.google.com/drive/folders/1cj1nqNs5ntlsiEZfSP12TR1WTYtwikCS?usp=sharing

YOUTUBE LINK

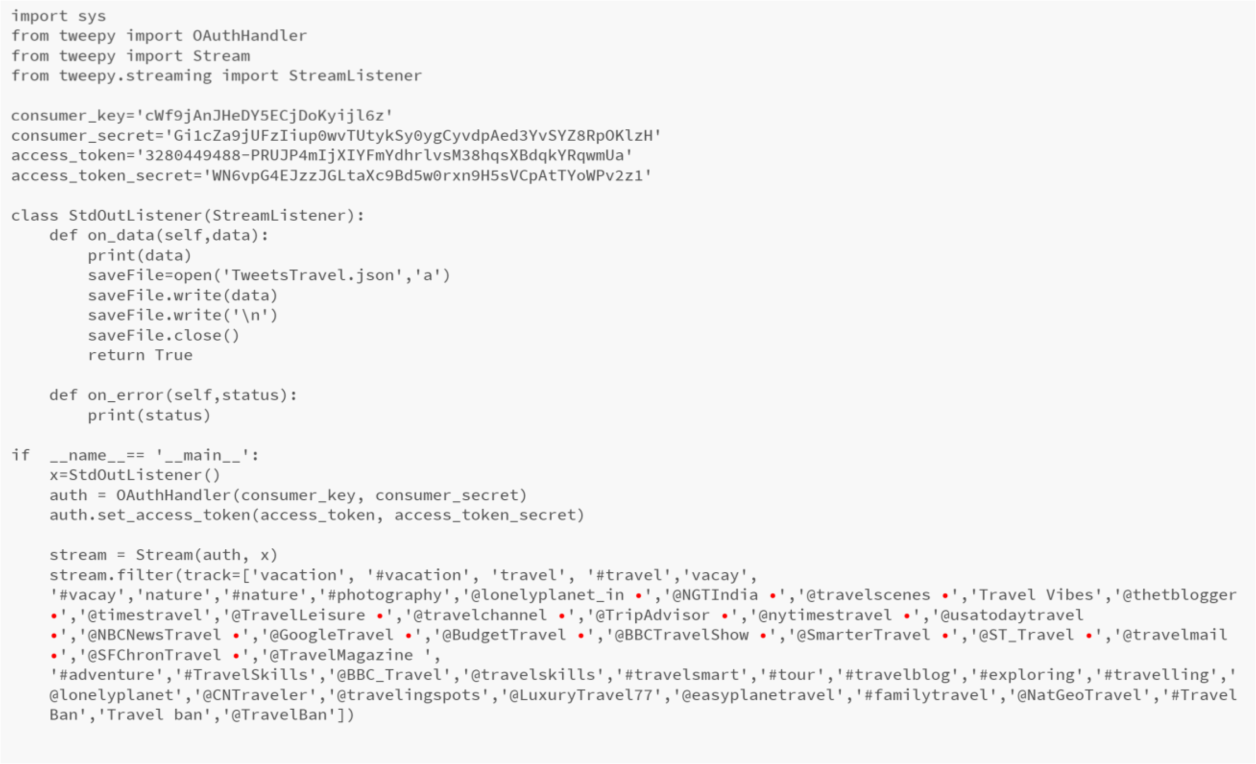
https://youtu.be/-ZggYl9NsL8

TECHNOLOGIES USED

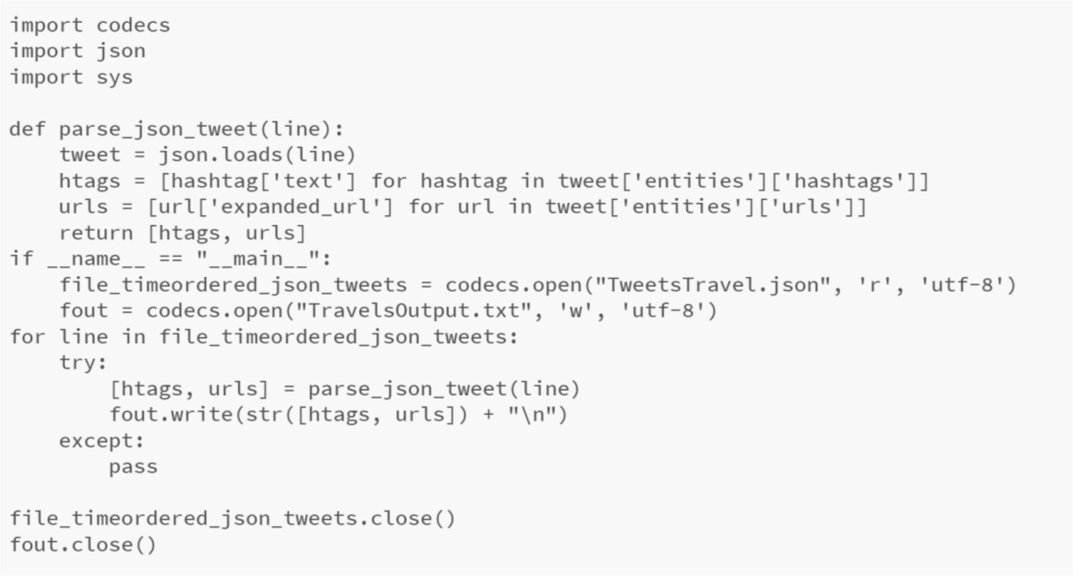
- Pycharm
- Cloudera
- IntelliJ
- Scala
- Spark
- Apache Hadoop
- Tableau

PHASE 1

- Collected tweets and saved as a JSON file.
- From the JSON file , extracted the hashtags and URLs and saved as text file.
- Word count on the hashtags and URLs using Apache Hadoop and Apache Spark.



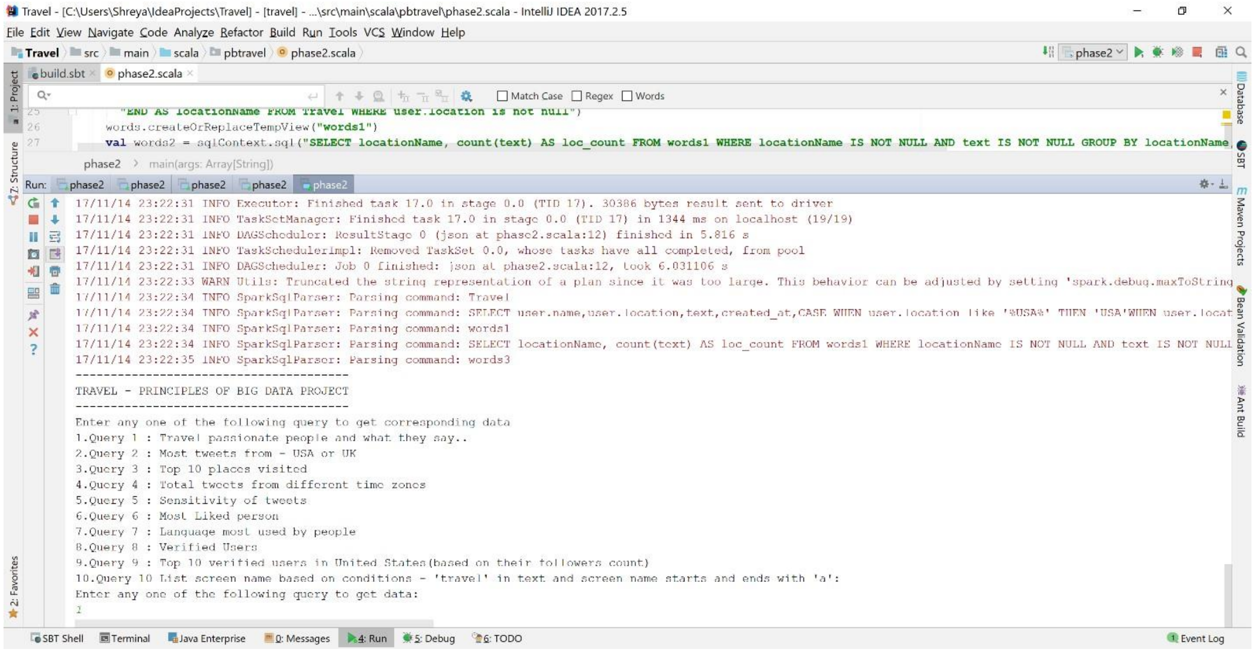
```
import sys
from tweepy import OAuthHandler
from tweepy import Stream
from tweepy.streaming import StreamListener

consumer_key='cWf9jAnJHeDY5ECjDoKyijl6z'
consumer_secret='Gi1cZa9jUFzIiup0wvTUtykSy0ygCyvdpAed3YvSYZ8RpOKlzH'
access_token='3280449488-PRUJP4mIjXIYFmYdhrlvsM38hqsXBdqkYRqwmUa'
access_token_secret='WN6vpG4EJzzJGLtaXc9Bd5w0rxn9H5sVCpAtTYoWPv2z1'

class StdOutListener(StreamListener):
    def on_data(self,data):
        print(data)
        saveFile=open('TweetsTravel.json','a')
        saveFile.write(data)
        saveFile.write('\n')
        saveFile.close()
        return True

    def on_error(self,status):
        print(status)

if  __name__== '__main__':
    x=StdOutListener()
    auth = OAuthHandler(consumer_key, consumer_secret)
    auth.set_access_token(access_token, access_token_secret)

    stream = Stream(auth, x)
    stream.filter(track=['vacation', '#vacation', 'travel', '#travel','vacay',
    '#vacay','nature','#nature','#photography','@lonelyplanet_in •','@NGTIndia •','@travelscenes •','Travel Vibes','@thetblogger
    •','@timestravel','@TravelLeisure •','@travelchannel •','@TripAdvisor •','@nytimestravel •','@usatodaytravel
    •','@NBCNewsTravel •','@GoogleTravel •','@BudgetTravel •','@BBCTravelShow •','@SmarterTravel •','@ST_Travel •','@travelmail
    •','@SFChronTravel •','@TravelMagazine ',
    '#adventure','#TravelSkills','@BBC_Travel','@travelskills','#travelsmart','#tour','#travelblog','#exploring','#travelling','
    @lonelyplanet','@CNTraveler','@travelingspots','@LuxuryTravel77','@easyplanetravel','#familytravel','@NatGeoTravel','#Travel
    Ban','Travel ban','@TravelBan'])
```



```
import codecs
import json
import sys

def parse_json_tweet(line):
    tweet = json.loads(line)
    htags = [hashtag['text'] for hashtag in tweet['entities']['hashtags']]
    urls = [url['expanded_url'] for url in tweet['entities']['urls']]
    return [htags, urls]
if __name__ == "__main__":
    file_timeordered_json_tweets = codecs.open("TweetsTravel.json", 'r', 'utf-8')
    fout = codecs.open("TravelsOutput.txt", 'w', 'utf-8')
for line in file_timeordered_json_tweets:
    try:
        [htags, urls] = parse_json_tweet(line)
        fout.write(str([htags, urls]) + "\n")
    except:
        pass

file_timeordered_json_tweets.close()
fout.close()
```

PHASE 2

ANALYTICAL QUERIES

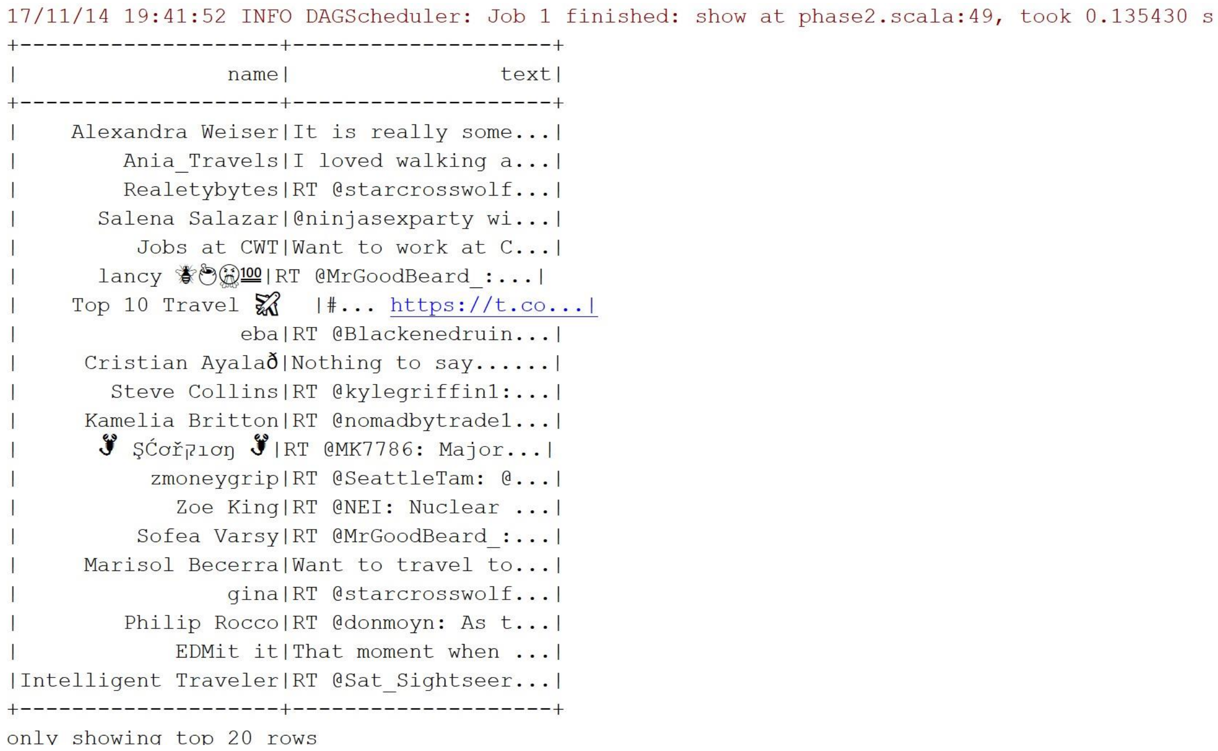

Travel - [C:\Users\Shreya\IdeaProjects\Travel] - [travel] - ...\src\main\scala\pbtravel\phase2.scala - IntelliJ IDEA 2017.2.5
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Travel src main scala pbtravel phase2.scala
phase2
build.sbt phase2.scala
□ Match Case □ Regex □ Words
phase2 main(args: Array[String])
Run: phase2 phase2 phase2 phase2 phase2
SBT Shell Terminal Java Enterprise 0: Messages 4: Run 5: Debug 6: TODO
Event Log
1: Project
7: Structure
2: Favorites
Database
SBT
Maven Projects
Bean Validation
Ant Build

```
        "END AS locationName FROM Travel WHERE user.location is not null")
      words.createOrReplaceTempView("words1")
      val words2 = sqlContext.sql("SELECT locationName, count(text) AS loc_count FROM words1 WHERE locationName IS NOT NULL AND text IS NOT NULL GROUP BY locationName,
```

```
17/11/14 23:22:31 INFO Executor: Finished task 17.0 in stage 0.0 (TID 17). 30386 bytes result sent to driver
17/11/14 23:22:31 INFO TaskSetManager: Finished task 17.0 in stage 0.0 (TID 17) in 1344 ms on localhost (19/19)
17/11/14 23:22:31 INFO DAGScheduler: ResultStage 0 (json at phase2.scala:12) finished in 5.816 s
17/11/14 23:22:31 INFO TaskSchedulerImpl: Removed TaskSet 0.0, whose tasks have all completed, from pool
17/11/14 23:22:31 INFO DAGScheduler: Job 0 finished: json at phase2.scala:12, took 6.031106 s
17/11/14 23:22:33 WARN Utils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.debug.maxToString
17/11/14 23:22:34 INFO SparkSqlParser: Parsing command: Travel
17/11/14 23:22:34 INFO SparkSqlParser: Parsing command: SELECT user.name,user.location,text,created_at,CASE WHEN user.location like '%USA%' THEN 'USA'WHEN user.locat
17/11/14 23:22:34 INFO SparkSqlParser: Parsing command: words1
17/11/14 23:22:34 INFO SparkSqlParser: Parsing command: SELECT locationName, count(text) AS loc_count FROM words1 WHERE locationName IS NOT NULL AND text IS NOT NULL
17/11/14 23:22:35 INFO SparkSqlParser: Parsing command: words3
----------------------------------------
TRAVEL - PRINCIPLES OF BIG DATA PROJECT
----------------------------------------
Enter any one of the following query to get corresponding data
1.Query 1 : Travel passionate people and what they say..
2.Query 2 : Most tweets from - USA or UK
3.Query 3 : Top 10 places visited
4.Query 4 : Total tweets from different time zones
5.Query 5 : Sensitivity of tweets
6.Query 6 : Most Liked person
7.Query 7 : Language most used by people
8.Query 8 : Verified Users
9.Query 9 : Top 10 verified users in United States(based on their followers count)
10.Query 10 List screen name based on conditions - 'travel' in text and screen name starts and ends with 'a':
Enter any one of the following query to get data:
2
```

1. **Travel passionate people and what they say..**

SELECT user.name, text FROM Travel WHERE text LIKE '%travel%'



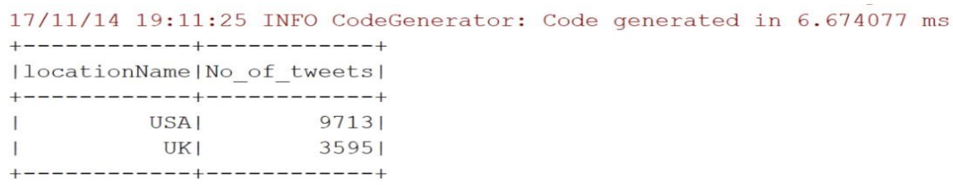
```
17/11/14 19:41:52 INFO DAGScheduler: Job 1 finished: show at phase2.scala:49, took 0.135430 s
+--------------------+--------------------+
|                name|                text|
+--------------------+--------------------+
|    Alexandra Weiser|It is really some...|
|        Ania_Travels|I loved walking a...|
|        Realetybytes|RT @starcrosswolf...|
|      Salena Salazar|@ninjasexparty wi...|
|         Jobs at CWT|Want to work at C...|
|     lancy 🐝🍯😤💯|RT @MrGoodBeard_:...|
|   Top 10 Travel ✈   |#... https://t.co...|
|                 eba|RT @Blackenedruin...|
|     Cristian Ayalað|Nothing to say......|
|       Steve Collins|RT @kylegriffin1:...|
|     Kamelia Britton|RT @nomadbytrade1...|
|    🦂 ŞĆσřקıσŋ 🦂|RT @MK7786: Major...|
|          zmoneygrip|RT @SeattleTam: @...|
|            Zoe King|RT @NEI: Nuclear ...|
|         Sofea Varsy|RT @MrGoodBeard_:...|
|     Marisol Becerra|Want to travel to...|
|                gina|RT @starcrosswolf...|
|        Philip Rocco|RT @donmoyn: As t...|
|            EDMit it|That moment when ...|
|Intelligent Traveler|RT @Sat_Sightseer...|
+--------------------+--------------------+
only showing top 20 rows
```

2. Most tweets from USA or UK

```
SELECT user.name,user.location,text,created_at," +
  "CASE WHEN user.location like '%USA%' THEN 'USA'" +
  "WHEN user.location like '%States%' THEN 'USA'" +
  "WHEN user.location like '%Kingdom%' THEN 'UK'" +
  "WHEN user.location like '%Scotland%' THEN 'UK'" +
  "WHEN user.location like '%England%' THEN 'UK'" +
  "WHEN user.location like '%Ireland%' THEN 'UK'" +
  "WHEN user.location like '%Wales%' THEN 'UK'" +
  "WHEN user.location like '%Europe%' THEN 'UK'" +
  "END AS locationName FROM Travel WHERE user.location is not null
```

```
SELECT locationName, count(text) AS loc_count FROM words1 WHERE
locationName IS NOT NULL AND text IS NOT NULL GROUP BY locationName,text
ORDER BY COUNT(text) DESC
```

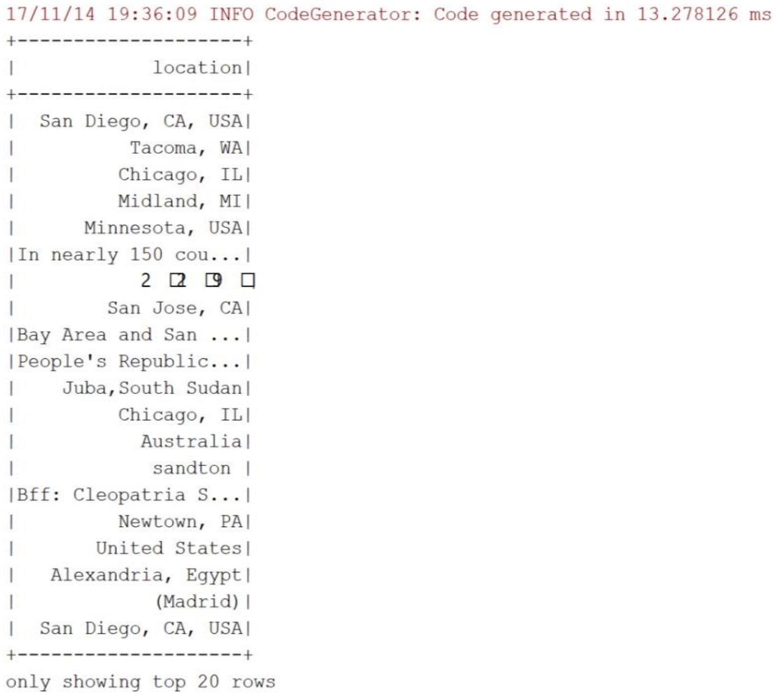
```
SELECT locationName, SUM(loc_count) AS No_of_tweets FROM words3 WHERE
locationName IS NOT
NULL AND loc_count IS NOT NULL GROUP BY locationName ORDER BY
SUM(loc_count) DESC
```



```
17/11/14 19:11:25 INFO CodeGenerator: Code generated in 6.674077 ms
+------------+------------+
|locationName|No_of_tweets|
+------------+------------+
|         USA|        9713|
|          UK|        3595|
+------------+------------+
```

3. **Top 10 places visited**

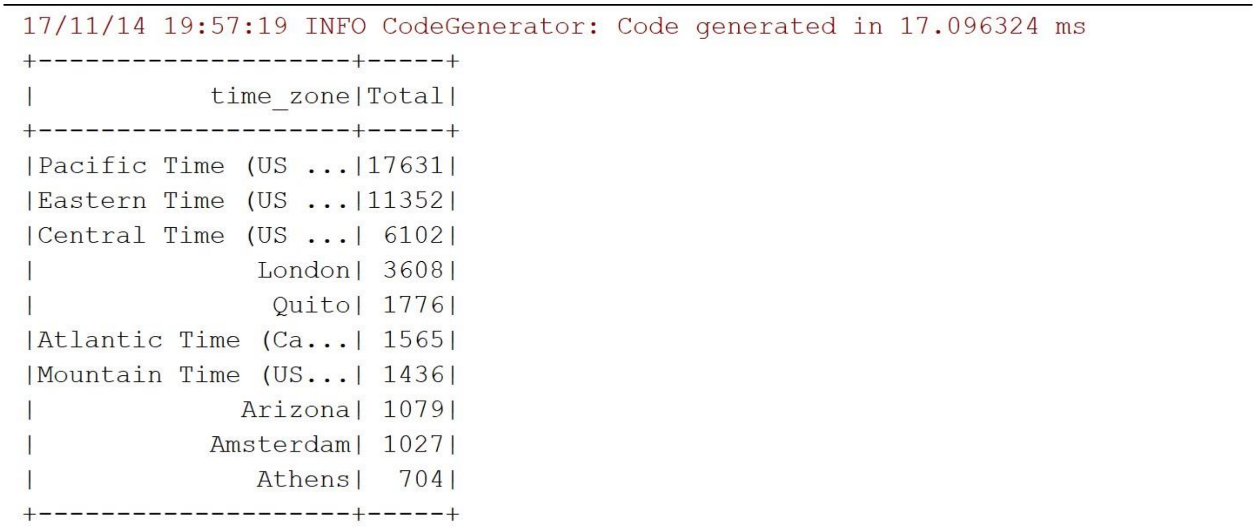
SELECT DISTINCT user.location FROM Travel WHERE user.location IS NOT NULL LIMIT 20



```
17/11/14 19:36:09 INFO CodeGenerator: Code generated in 13.278126 ms
+--------------------+
|            location|
+--------------------+
|  San Diego, CA, USA|
|          Tacoma, WA|
|         Chicago, IL|
|         Midland, MI|
|      Minnesota, USA|
|In nearly 150 cou...|
|            2 □2 □9 □
|        San Jose, CA|
|Bay Area and San ...|
|People's Republic...|
|    Juba,South Sudan|
|         Chicago, IL|
|           Australia|
|            sandton |
|Bff: Cleopatria S...|
|         Newtown, PA|
|       United States|
|   Alexandria, Egypt|
|            (Madrid)|
|  San Diego, CA, USA|
+--------------------+
only showing top 20 rows
```

4. **Total tweets from different time zones**

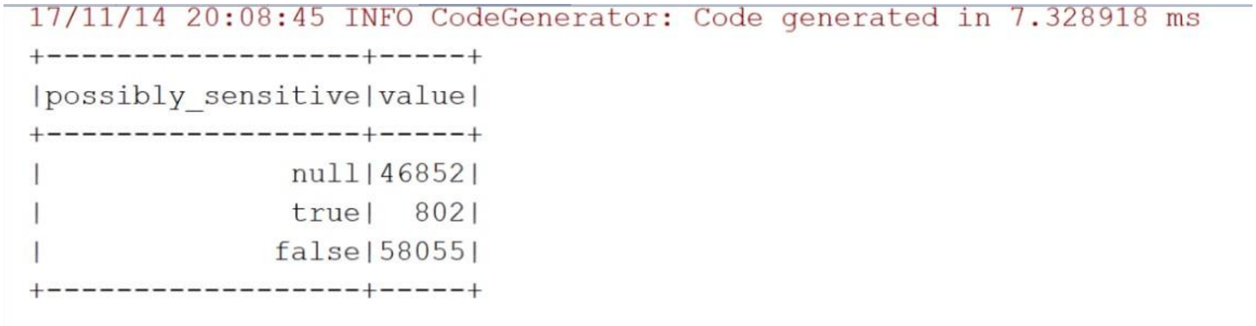
SELECT user.time_zone,count(text) AS Total FROM Travel WHERE user.time_zone IS NOT NULL GROUP BY user.time_zone ORDER BY Total DESC LIMIT 10



```
17/11/14 19:57:19 INFO CodeGenerator: Code generated in 17.096324 ms
+--------------------+-----+
|           time_zone|Total|
+--------------------+-----+
|Pacific Time (US ...|17631|
|Eastern Time (US ...|11352|
|Central Time (US ...| 6102|
|              London| 3608|
|               Quito| 1776|
|Atlantic Time (Ca...| 1565|
|Mountain Time (US...| 1436|
|             Arizona| 1079|
|           Amsterdam| 1027|
|              Athens|  704|
+--------------------+-----+
```

5. Sensitivity of tweets

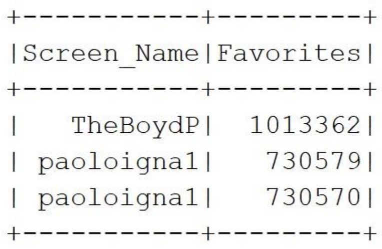
SELECT possibly_sensitive,count(*) AS value FROM Travel GROUP BY possibly_sensitive



```
17/11/14 20:08:45 INFO CodeGenerator: Code generated in 7.328918 ms
+------------------+-----+
|possibly_sensitive|value|
+------------------+-----+
|              null|46852|
|              true|  802|
|             false|58055|
+------------------+-----+
```

6. Most Liked person

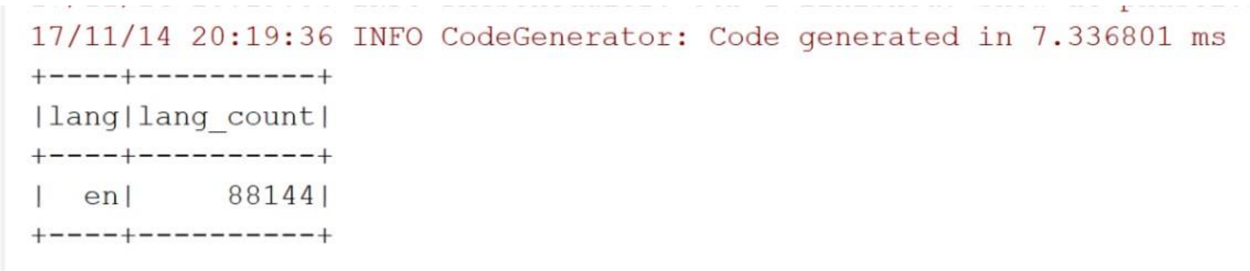
SELECT user.screen_name AS Screen_Name,user.favourites_count AS Favorites FROM Travel ORDER BY Favorites DESC LIMIT 3



```
+-----------+---------+
|Screen_Name|Favorites|
+-----------+---------+
|   TheBoydP|  1013362|
| paoloigna1|   730579|
| paoloigna1|   730570|
+-----------+---------+
```

7. Language most used by people

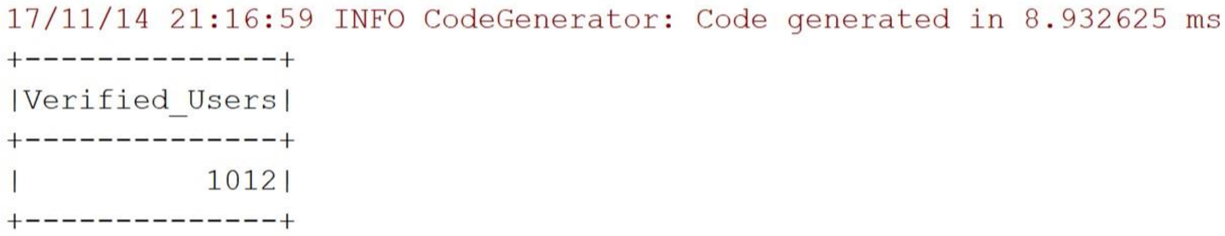
SELECT user.lang,count(*) AS lang_count FROM Travel WHERE user.lang IS NOT NULL GROUP BY user.lang ORDER BY lang_count DESC LIMIT 1



```
17/11/14 20:19:36 INFO CodeGenerator: Code generated in 7.336801 ms
+----+----------+
|lang|lang_count|
+----+----------+
|  en|     88144|
+----+----------+
```

8. Verified Users

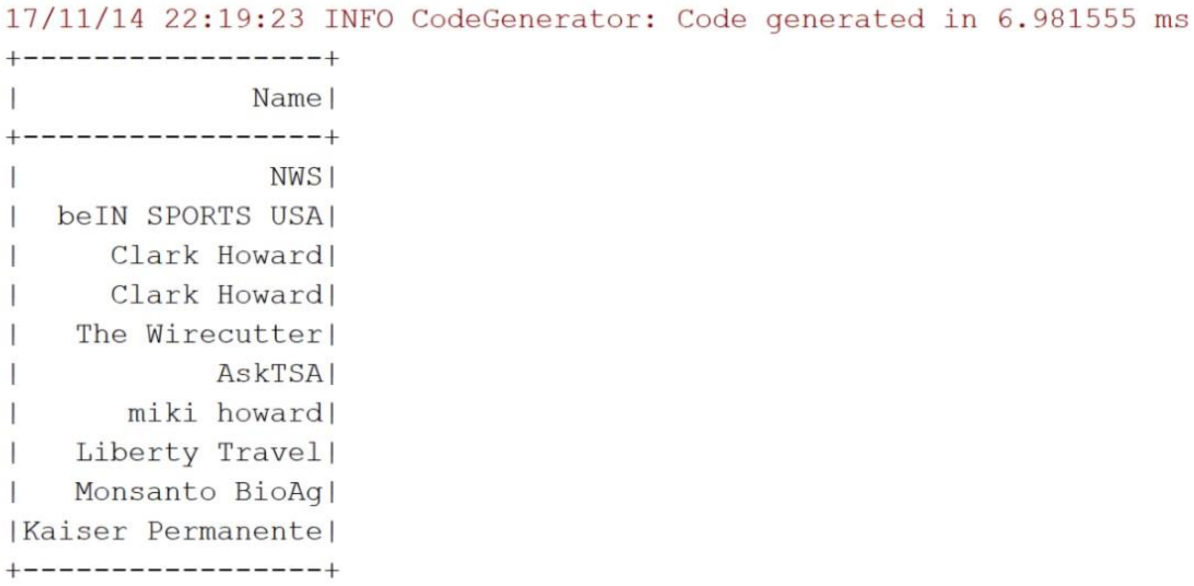
SELECT count(user.verified) AS Verified_Users FROM Travel WHERE user.verified=true



```
17/11/14 21:16:59 INFO CodeGenerator: Code generated in 8.932625 ms
+--------------+
|Verified_Users|
+--------------+
|          1012|
+--------------+
```

9. Top 10 verified users in United States(based on their followers count)

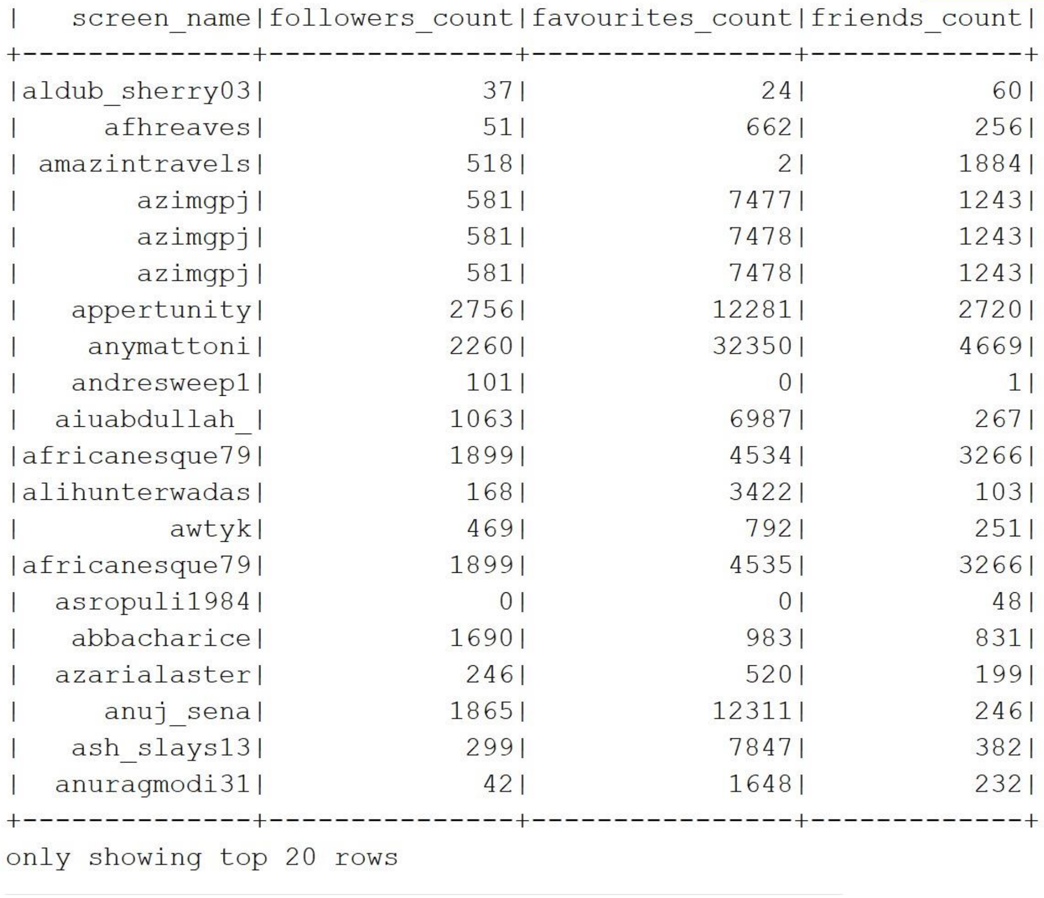
SELECT user.name AS Name FROM Travel WHERE user.verified=true AND user.location='United States' ORDER BY user.followers_count DESC LIMIT 10



```
17/11/14 22:19:23 INFO CodeGenerator: Code generated in 6.981555 ms
+-----------------+
|             Name|
+-----------------+
|              NWS|
|  beIN SPORTS USA|
|     Clark Howard|
|     Clark Howard|
|    The Wirecutter|
|           AskTSA|
|      miki howard|
|   Liberty Travel|
|   Monsanto BioAg|
|Kaiser Permanente|
+-----------------+
```

10. **List the User’s Screen Name, followers count , friends count and favourites count when it satisfies two conditions : ‘Travel’ in the text and screen name starts with ‘a’.**

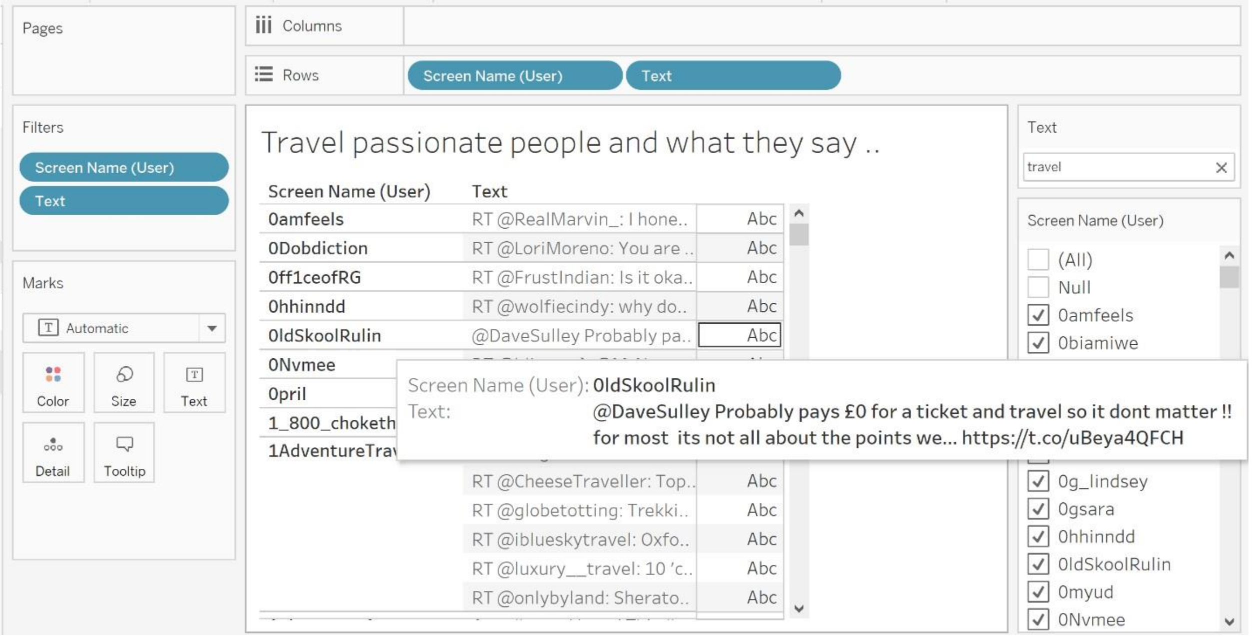
SELECT user.screen_name, user.followers_count,user.favourites_count, user.friends_count FROM Travel WHERE user.screen_name LIKE 'a%' AND text LIKE '%travel%'



```
|   screen_name|followers_count|favourites_count|friends_count|
+--------------+---------------+----------------+-------------+
|aldub_sherry03|             37|              24|           60|
|     afhreaves|             51|             662|          256|
| amazintravels|            518|               2|         1884|
|       azimgpj|            581|            7477|         1243|
|       azimgpj|            581|            7478|         1243|
|       azimgpj|            581|            7478|         1243|
|   appertunity|           2756|           12281|         2720|
|    anymattoni|           2260|           32350|         4669|
|   andresweep1|            101|               0|            1|
|  aiuabdullah_|           1063|            6987|          267|
|africanesque79|           1899|            4534|         3266|
|alihunterwadas|            168|            3422|          103|
|         awtyk|            469|             792|          251|
|africanesque79|           1899|            4535|         3266|
|  asropuli1984|              0|               0|           48|
|   abbacharice|           1690|             983|          831|
|  azarialaster|            246|             520|          199|
|     anuj_sena|           1865|           12311|          246|
|   ash_slays13|            299|            7847|          382|
|  anuragmodi31|             42|            1648|          232|
+--------------+---------------+----------------+-------------+
only showing top 20 rows
```

VISUALIZATION

QUERY 1 :



Pages

Columns

Rows: Screen Name (User) | Text

Filters: Screen Name (User), Text

Marks: Automatic

Color | Size | Text | Detail | Tooltip

Travel passionate people and what they say ..

Screen Name (User)	Text	
0amfeels	RT @RealMarvin_: I hone..	Abc
0Dobdiction	RT @LoriMoreno: You are ..	Abc
0ff1ceofRG	RT @FrustIndian: Is it oka..	Abc
0hhinndd	RT @wolfiecindy: why do..	Abc
0ldSkoolRulin	@DaveSulley Probably pa..	Abc
0Nvmee		
0pril		
1_800_choketh		
1AdventureTrav		
	RT @CheeseTraveller: Top..	Abc
	RT @globetotting: Trekki..	Abc
	RT @iblueskytravel: Oxfo..	Abc
	RT @luxury__travel: 10 'c..	Abc
	RT @onlybyland: Sherato..	Abc

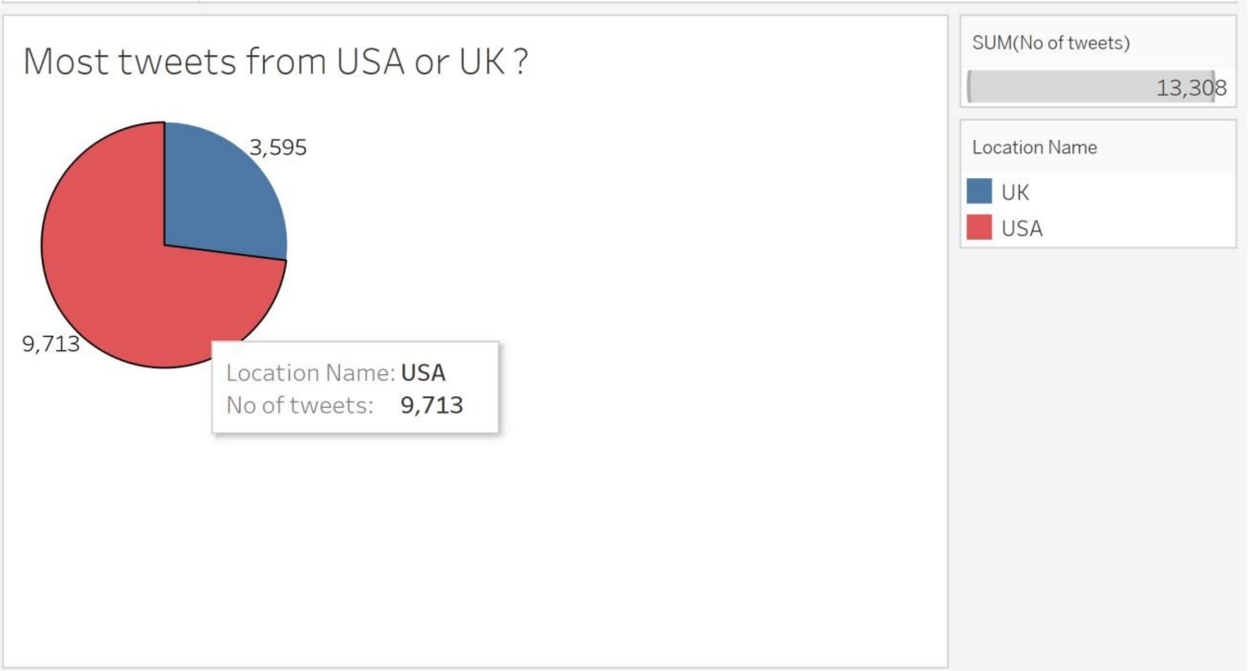
Screen Name (User): 0ldSkoolRulin
Text: @DaveSulley Probably pays £0 for a ticket and travel so it dont matter !!
for most its not all about the points we... https://t.co/uBeya4QFCH

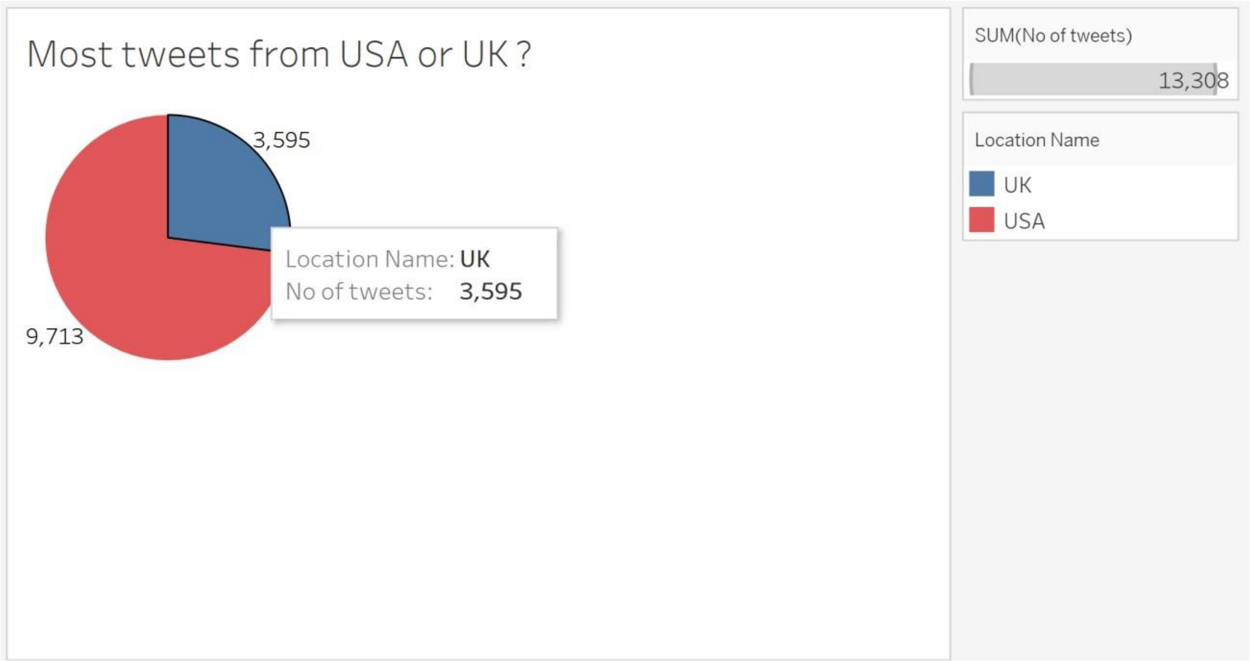
Text: travel

Screen Name (User)

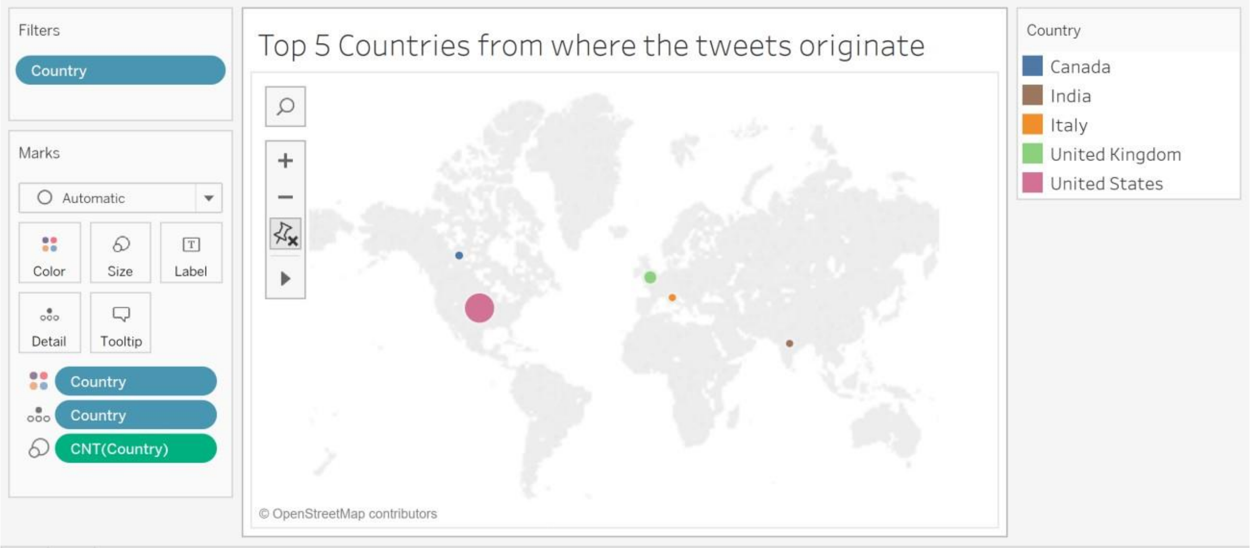
- □ (All)
- □ Null
- ☑ 0amfeels
- ☑ 0biamiwe
- ☑ 0g_lindsey
- ☑ 0gsara
- ☑ 0hhinndd
- ☑ 0ldSkoolRulin
- ☑ 0myud
- ☑ 0Nvmee

QUERY 2 :

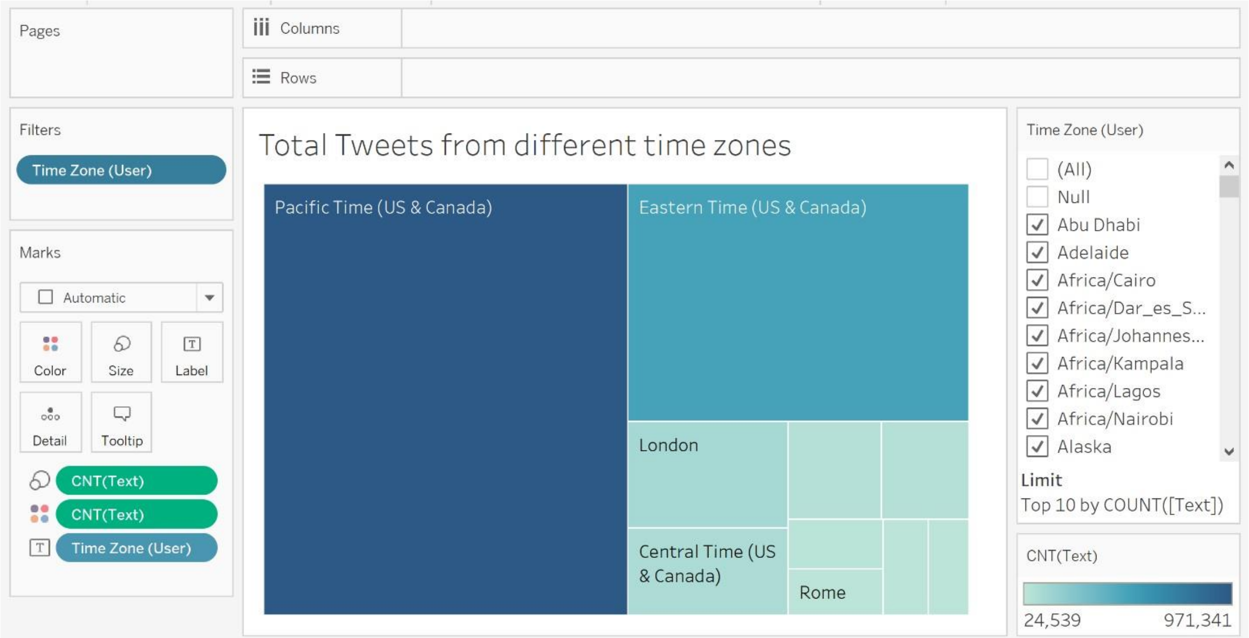



Most tweets from USA or UK ?
3,595
9,713
Location Name: UK
No of tweets: 3,595
SUM(No of tweets)
13,308
Location Name
🟦 UK
🟥 USA

QUERY 3 :


Filters
Country
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Country
Country
CNT(Country)
Top 5 Countries from where the tweets originate
© OpenStreetMap contributors
Country
Canada
India
Italy
United Kingdom
United States

QUERY 4 :

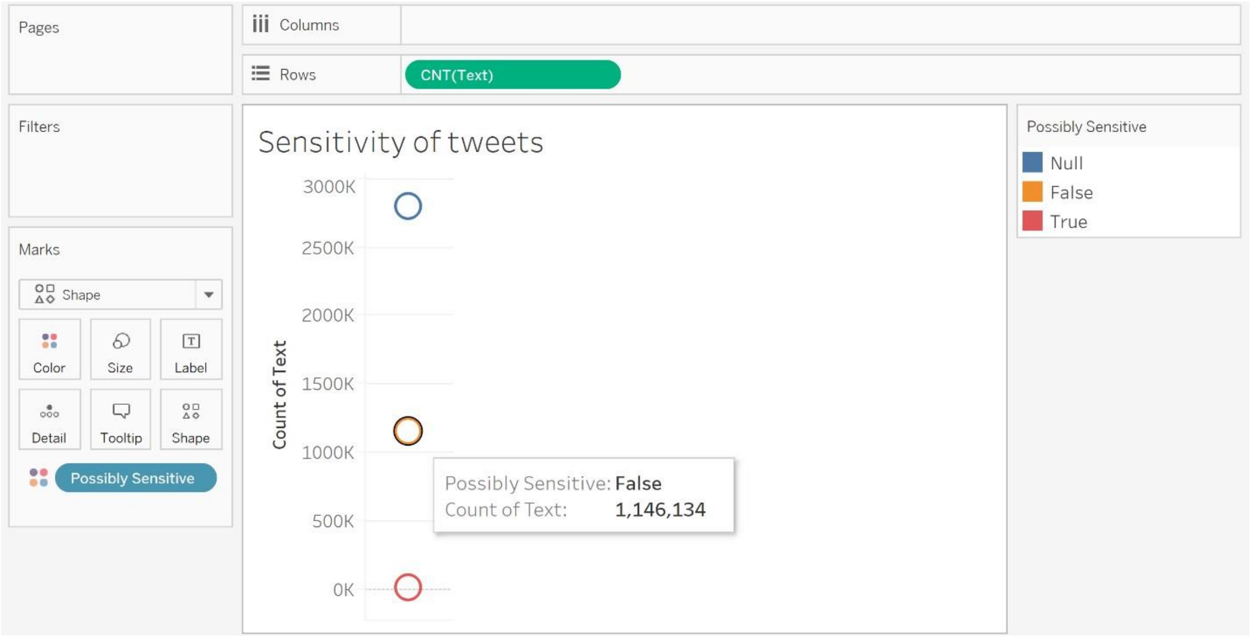


Time Zone (User)

- □ (All)
- □ Null
- ☑ Abu Dhabi
- ☑ Adelaide
- ☑ Africa/Cairo
- ☑ Africa/Dar_es_S...
- ☑ Africa/Johannes...
- ☑ Africa/Kampala
- ☑ Africa/Lagos
- ☑ Africa/Nairobi
- ☑ Alaska

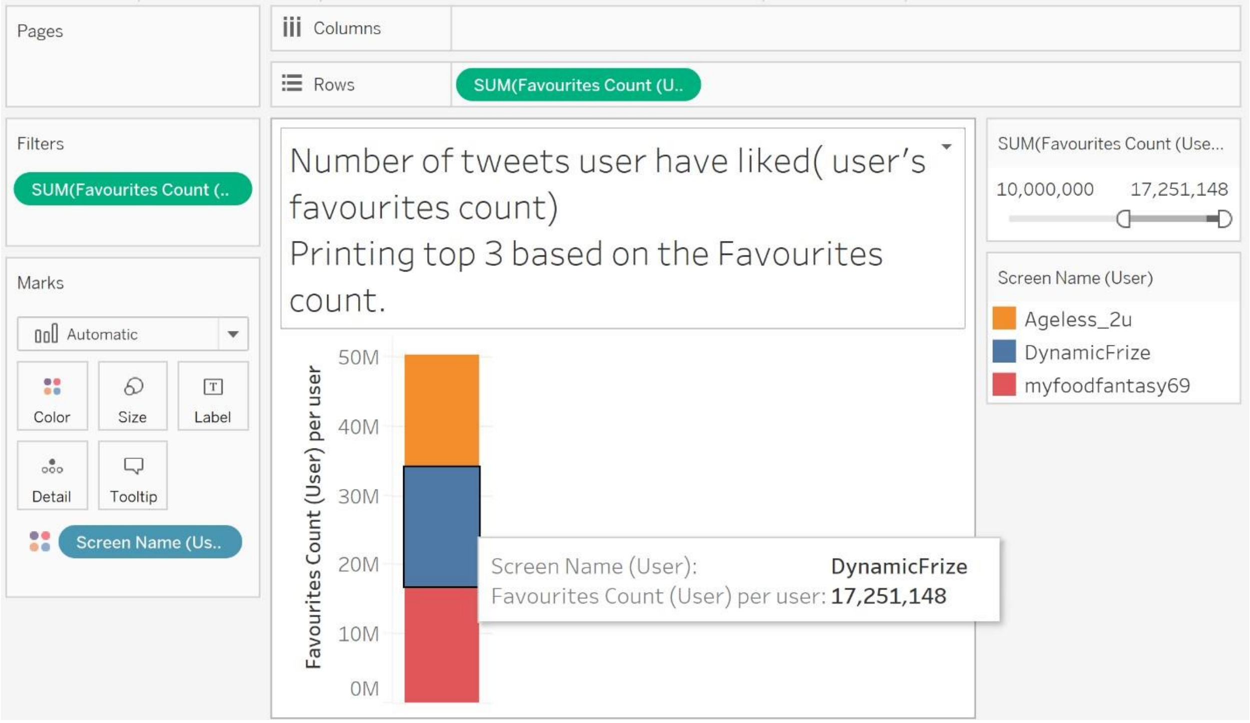
Limit
Top 10 by COUNT([Text])

QUERY 5 :



Possibly Sensitive: False
Count of Text: 1,146,134

QUERY 6 :



Rows: SUM(Favourites Count (U..

Filters: SUM(Favourites Count (..

Marks: Screen Name (Us..

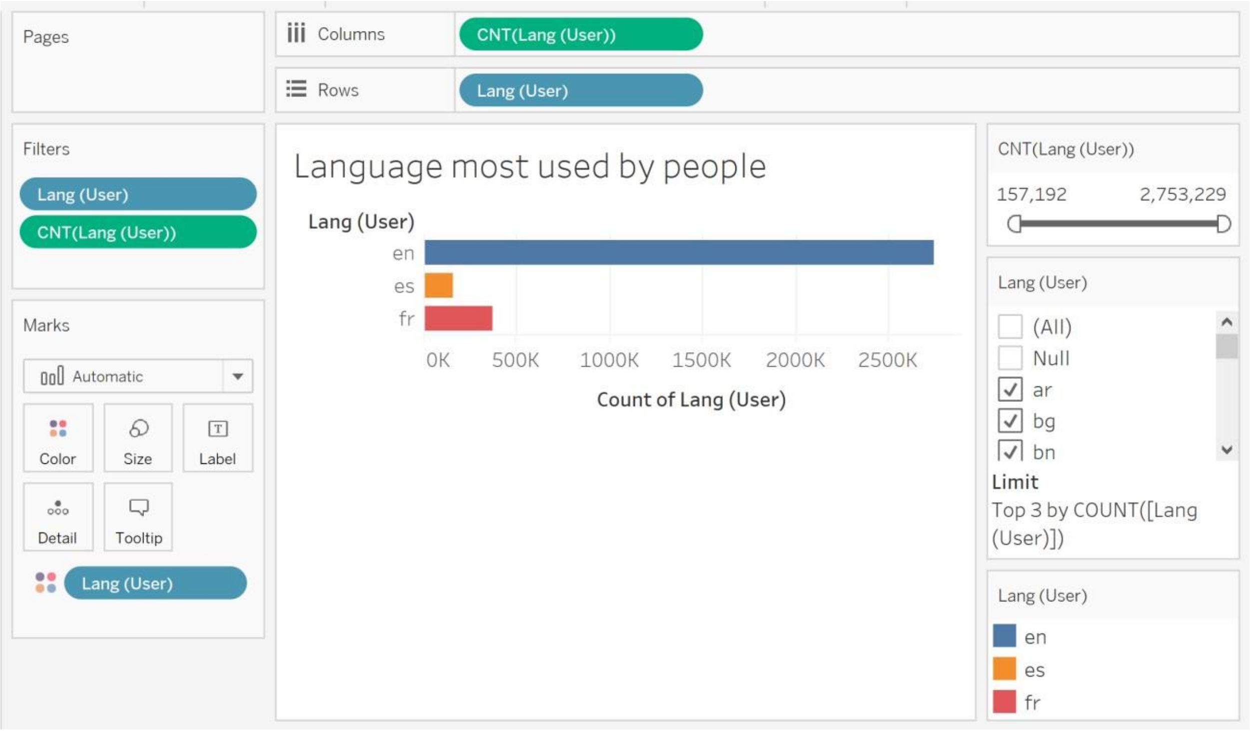
Number of tweets user have liked(user's favourites count)
Printing top 3 based on the Favourites count.

Screen Name (User): DynamicFrize
Favourites Count (User) per user: 17,251,148

SUM(Favourites Count (Use...
10,000,000 17,251,148

Screen Name (User)
🟧 Ageless_2u
🟦 DynamicFrize
🟥 myfoodfantasy69

QUERY 7:



Columns: CNT(Lang (User))

Rows: Lang (User)

Filters: Lang (User), CNT(Lang (User))

Marks: Lang (User)

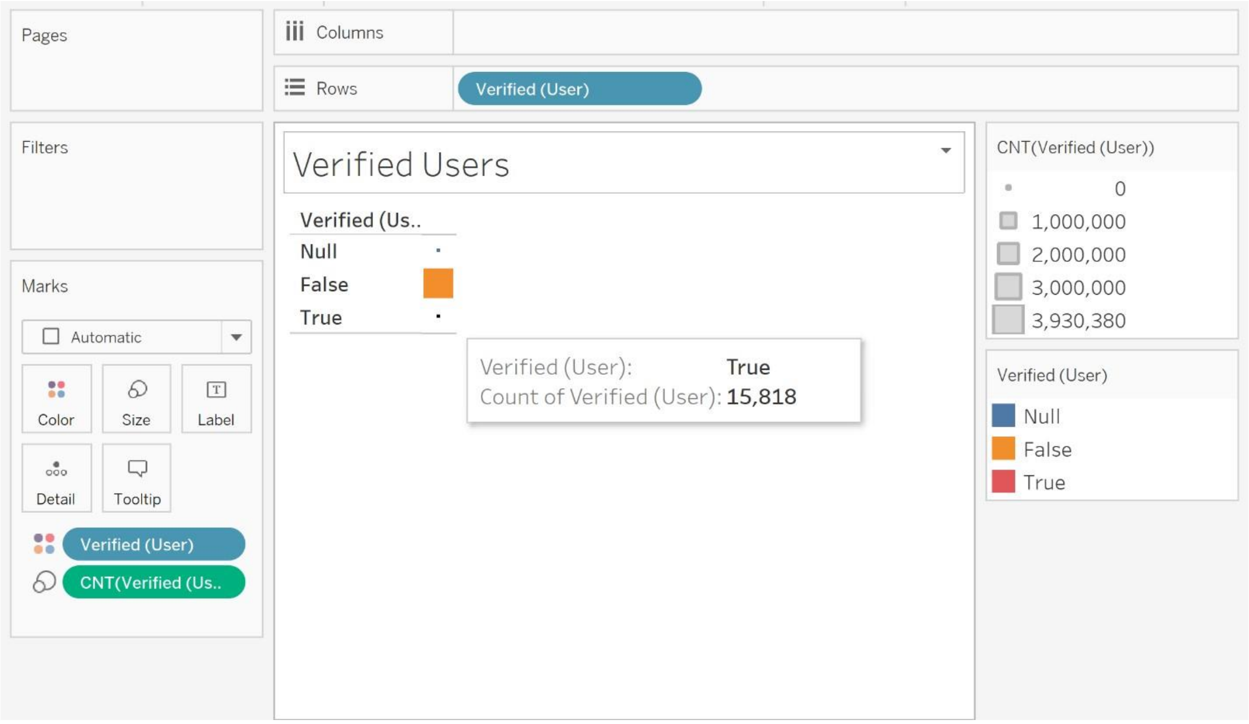
Language most used by people

CNT(Lang (User))
157,192 2,753,229

Lang (User)
□ (All)
□ Null
☑ ar
☑ bg
☑ bn
Limit
Top 3 by COUNT([Lang (User)])

Lang (User)
🟦 en
🟧 es
🟥 fr

QUERY 8 :



Rows: Verified (User)

Marks: Verified (User), CNT(Verified (Us..

Verified Users

Verified (Us..	
Null	·
False	🟧
True	·

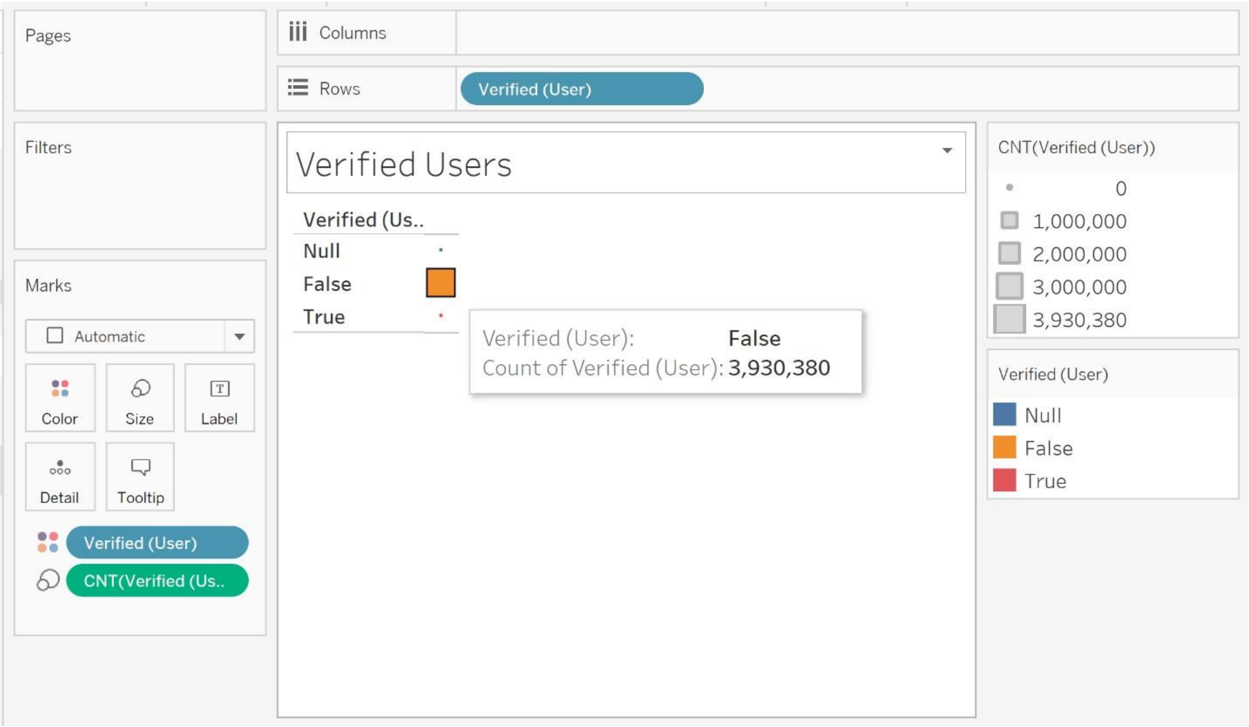
Verified (User): True
Count of Verified (User): 15,818

CNT(Verified (User))

- 0
- 1,000,000
- 2,000,000
- 3,000,000
- 3,930,380

Verified (User)

- 🟦 Null
- 🟧 False
- 🟥 True



Rows: Verified (User)

Marks: Verified (User), CNT(Verified (Us..

Verified Users

Verified (Us..	
Null	·
False	🟧
True	·

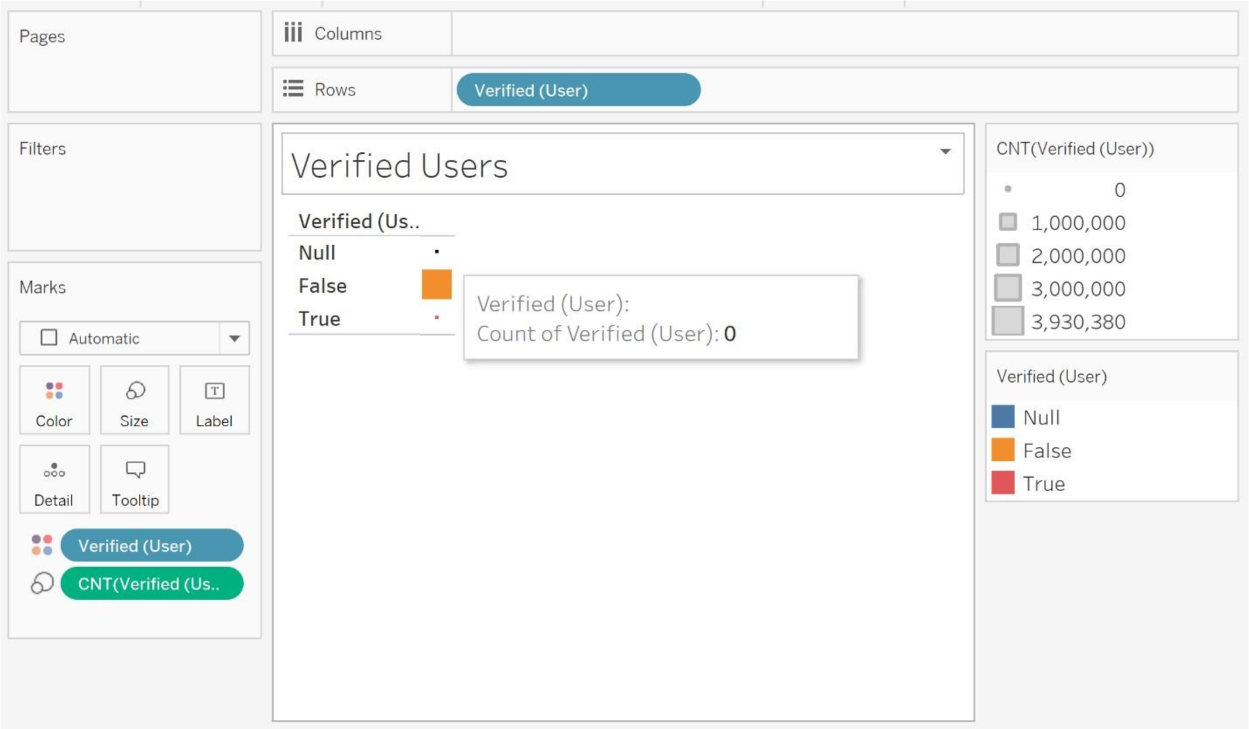
Verified (User): False
Count of Verified (User): 3,930,380

CNT(Verified (User))

- 0
- 1,000,000
- 2,000,000
- 3,000,000
- 3,930,380

Verified (User)

- 🟦 Null
- 🟧 False
- 🟥 True



Rows: Verified (User)

Marks: Automatic

Color: Verified (User)

Size: CNT(Verified (Us..

Verified Users

Verified (Us..	
Null	
False	🟧
True	

Verified (User):
Count of Verified (User): 0

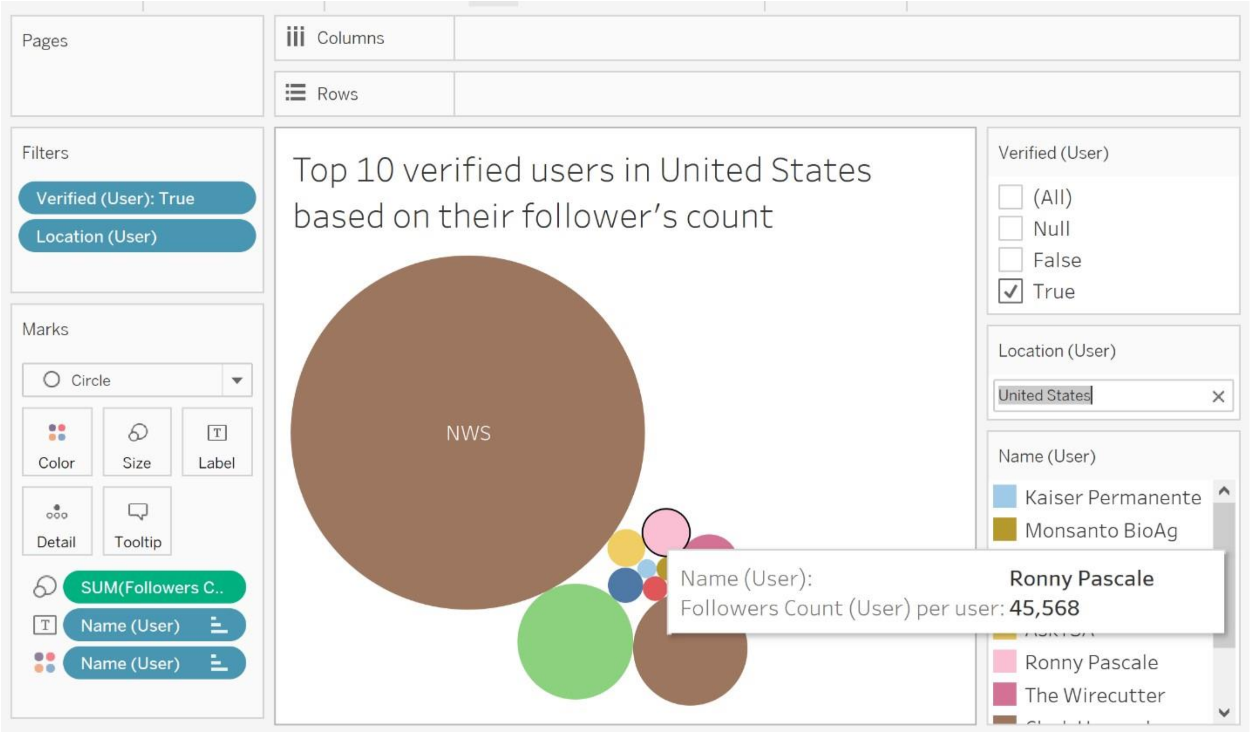
CNT(Verified (User))

- 0
- 1,000,000
- 2,000,000
- 3,000,000
- 3,930,380

Verified (User)

- 🟦 Null
- 🟧 False
- 🟥 True

QUERY 9 :



Filters: Verified (User): True; Location (User)

Marks: Circle

Size: SUM(Followers C..

Label: Name (User)

Color: Name (User)

Top 10 verified users in United States based on their follower's count

Name (User): Ronny Pascale
Followers Count (User) per user: 45,568

Verified (User)

- ☐ (All)
- ☐ Null
- ☐ False
- ☑ True

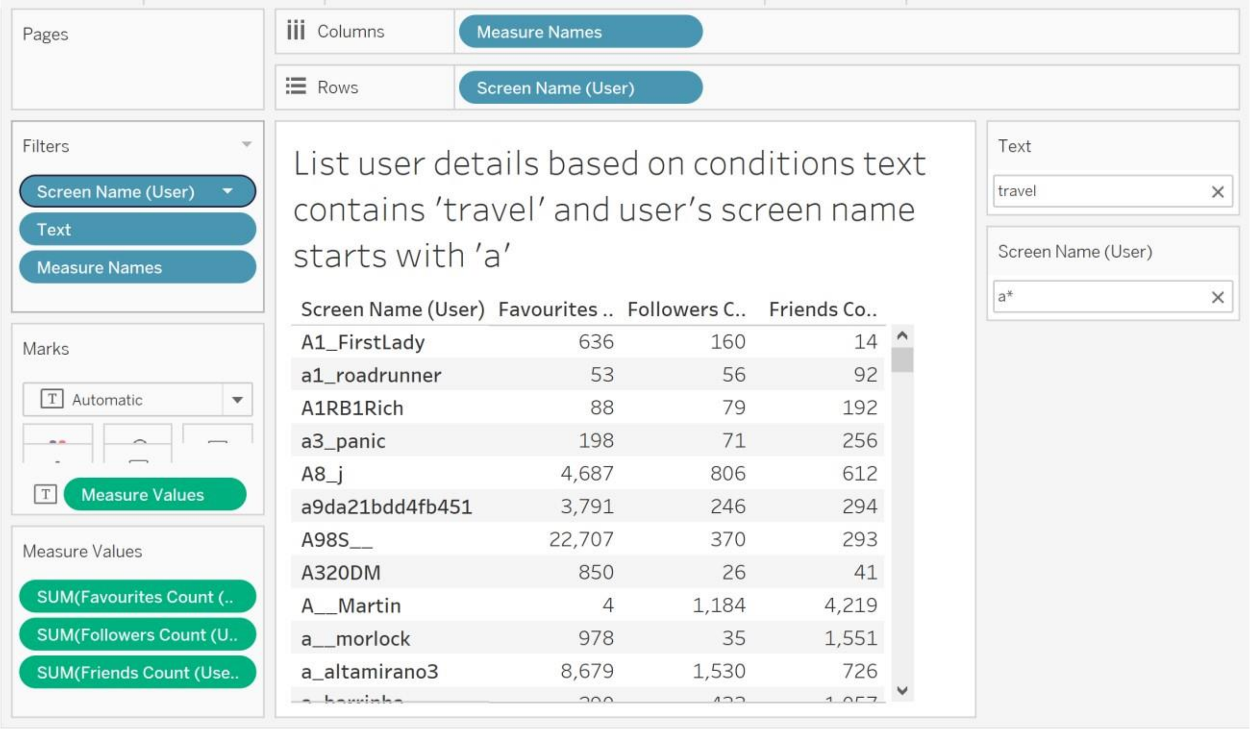
Location (User)

United States

Name (User)

- Kaiser Permanente
- Monsanto BioAg
- Ronny Pascale
- The Wirecutter

QUERY 10 :



Pages

Columns: Measure Names

Rows: Screen Name (User)

Filters: Screen Name (User), Text, Measure Names

Marks: Automatic

Measure Values

Measure Values: SUM(Favourites Count (.., SUM(Followers Count (U.., SUM(Friends Count (Use..

Text: travel

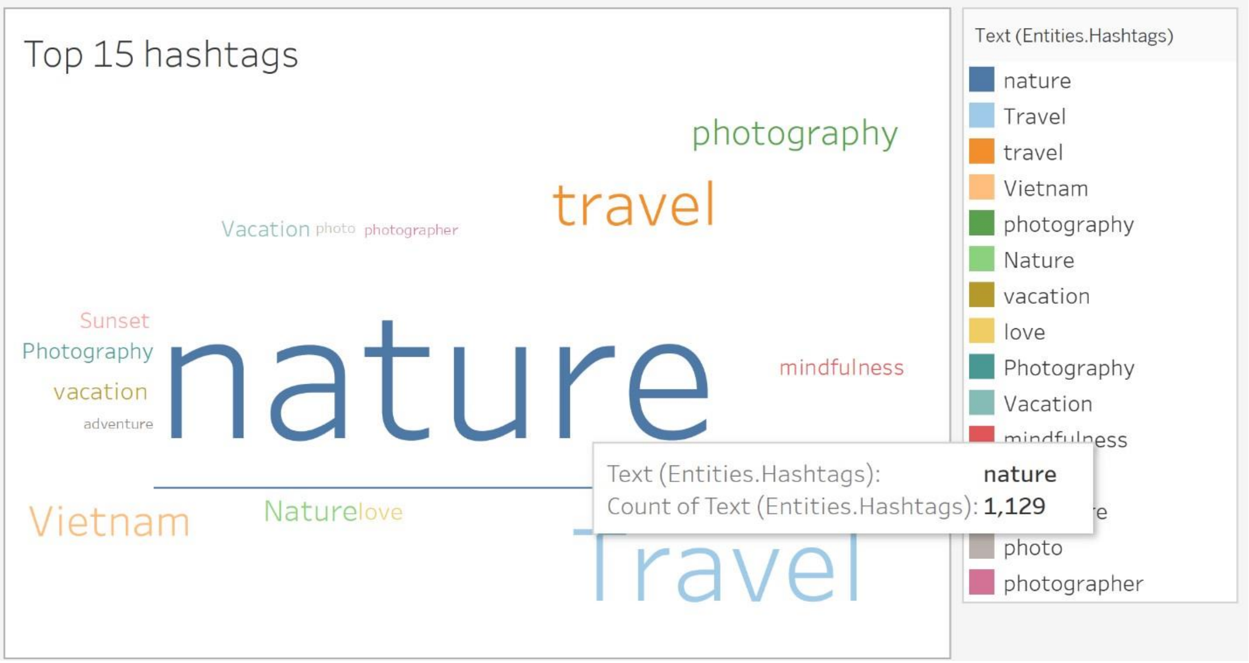
Screen Name (User): a*

List user details based on conditions text contains 'travel' and user's screen name starts with 'a'

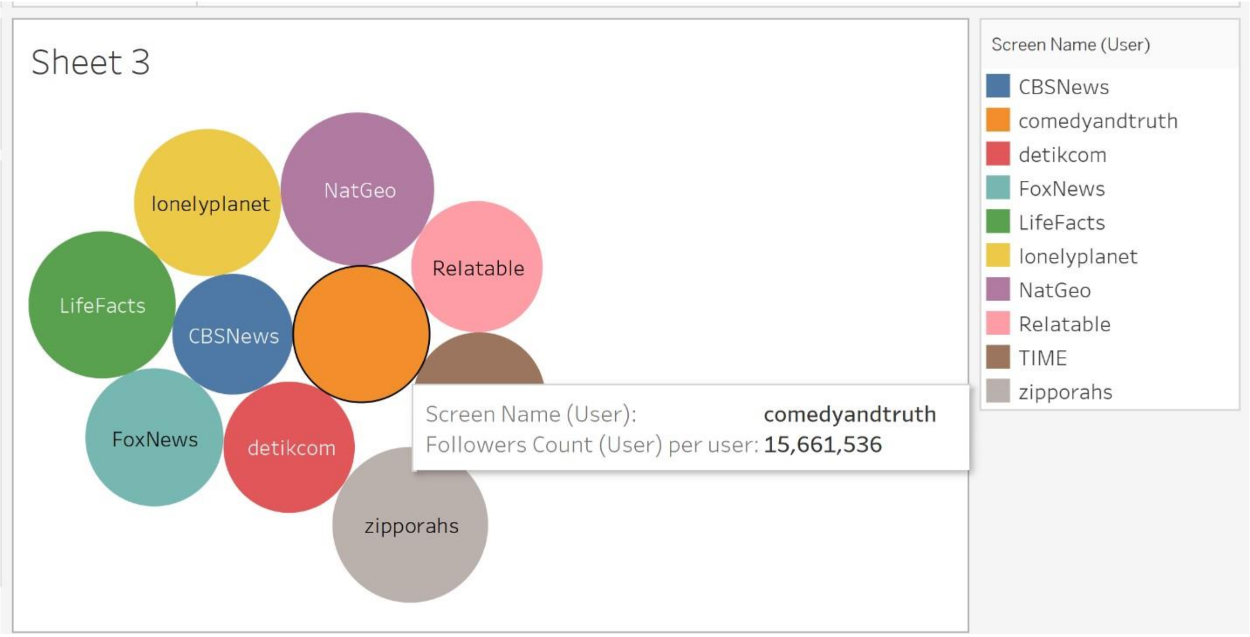
Screen Name (User)	Favourites ..	Followers C..	Friends Co..
A1_FirstLady	636	160	14
a1_roadrunner	53	56	92
A1RB1Rich	88	79	192
a3_panic	198	71	256
A8_j	4,687	806	612
a9da21bdd4fb451	3,791	246	294
A98S__	22,707	370	293
A320DM	850	26	41
A__Martin	4	1,184	4,219
a__morlock	978	35	1,551
a_altamirano3	8,679	1,530	726

EXTRA VISUALIZATIONS :

Top 15 Hashtags



Top 10 Users using their Follower's count



Screen Name (User): comedyandtruth
Followers Count (User) per user: 15,661,536