MACHARIA ANASTASIA

TWITTER DATA ANALYSIS CHALLENGE

Step 1: Hypothesis Selection

Null hypothesis (h_0) = Valentines Celebration is part of the African Culture

Alternative hypothesis ($\mathbf{h_1}$) = Valentines Celebration is not part of the African Culture

Step 2: Hashtag Selection, Period, Dataset Size

Hashtag: #valentines

Period: 1st February 2020 to 15th Feb

Size of data: 300 entries

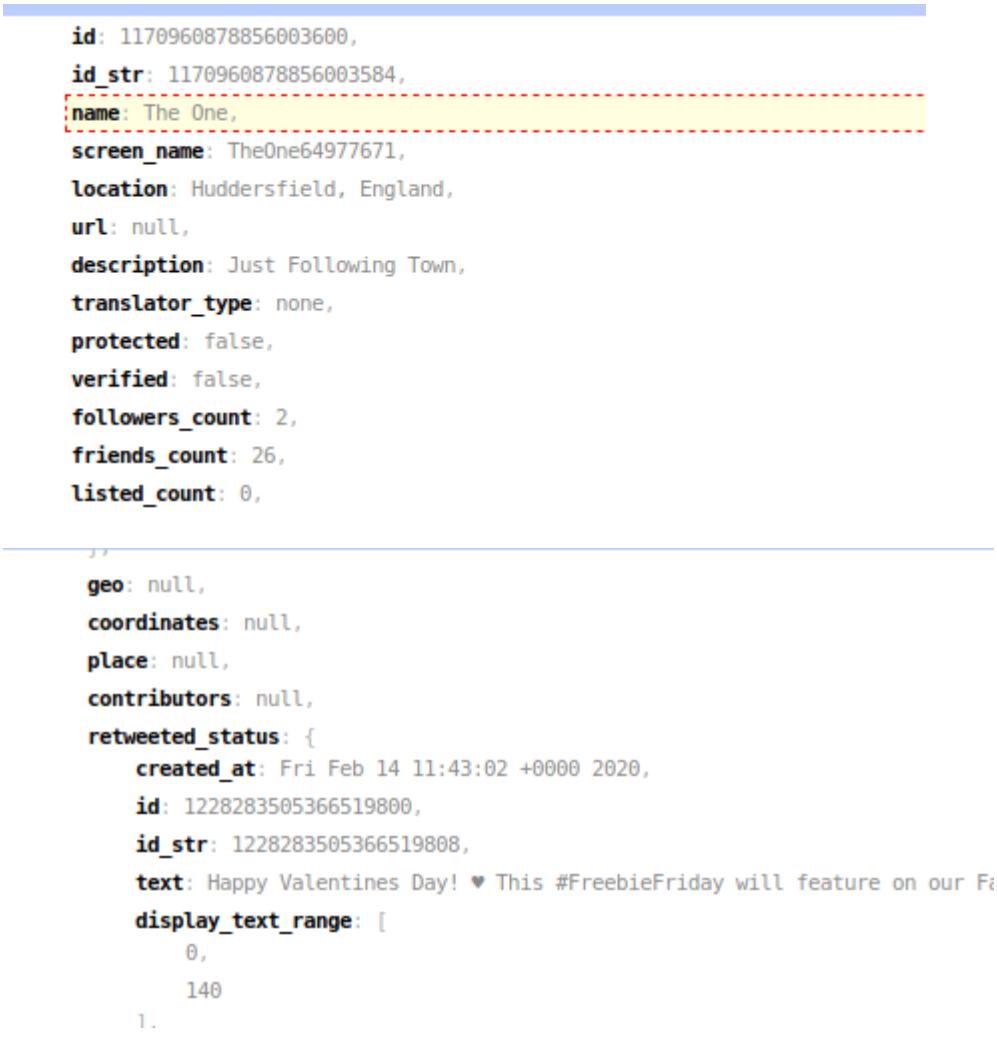
Step 3: Data Scrapping from Twitter

The Tweepy library was used to scrap data from the Twitter API using the Twitter developer's platform credentials.

```
In [8]: with open('twitter_credentials.json') as access_data:
            cred = json.load(access_data)
            api_key = cred['API_KEY']
            api_secret_key = cred['API_SECRET_KEY']
            access_token = cred['ACCESS_TOKEN']
            access_token_secret = cred['ACCESS_TOKEN_SECRET']

In [9]: auth = tweepy.OAuthHandler(api_key, api_secret_key)
        api = tweepy.API(auth)

In [10]: max_tweets_to_be_extracted = \
         int(input('Enter the number of tweets that you want to extract- '))

         Enter the number of tweets that you want to extract- 300

In [11]: hashtag = input('Enter the hashtag you want to scrape- ')

         Enter the hashtag you want to scrape- valentines

In [24]: for tweet in tweepy.Cursor(api.search, q='#' + hashtag,
             rpp=100).items(number_of_tweets_to_be_extracted):
             with open('tweets_with_hashtag_' + hashtag + '.json', 'a') as \
                     the_file:
                     the_file.write(str(tweet.text.encode('utf-8')))
```

Raw Data Collection

valentines_tweets. json contains the raw data with 300 randomly picked entries with tweets containing valentines hashtag. The screenshots below shows some features from the tweet object.



Step 4: Data Analysis

From the raw data analysis is performed to extract data that will be meaningful to the study. The location feature of the tweet object was used to get the location details of the twitter user. If the location was not specified, the tweet coordinates were extracted to be used as the location details. Other meaningful features like username, user_id , screen_name, primary_geo and type of tweet were picked.


```
with open('valentines_tweets_loc_details_extract.json', 'w') as fout:
fout.write(json.dumps(users_with_geodata, indent=4))
```

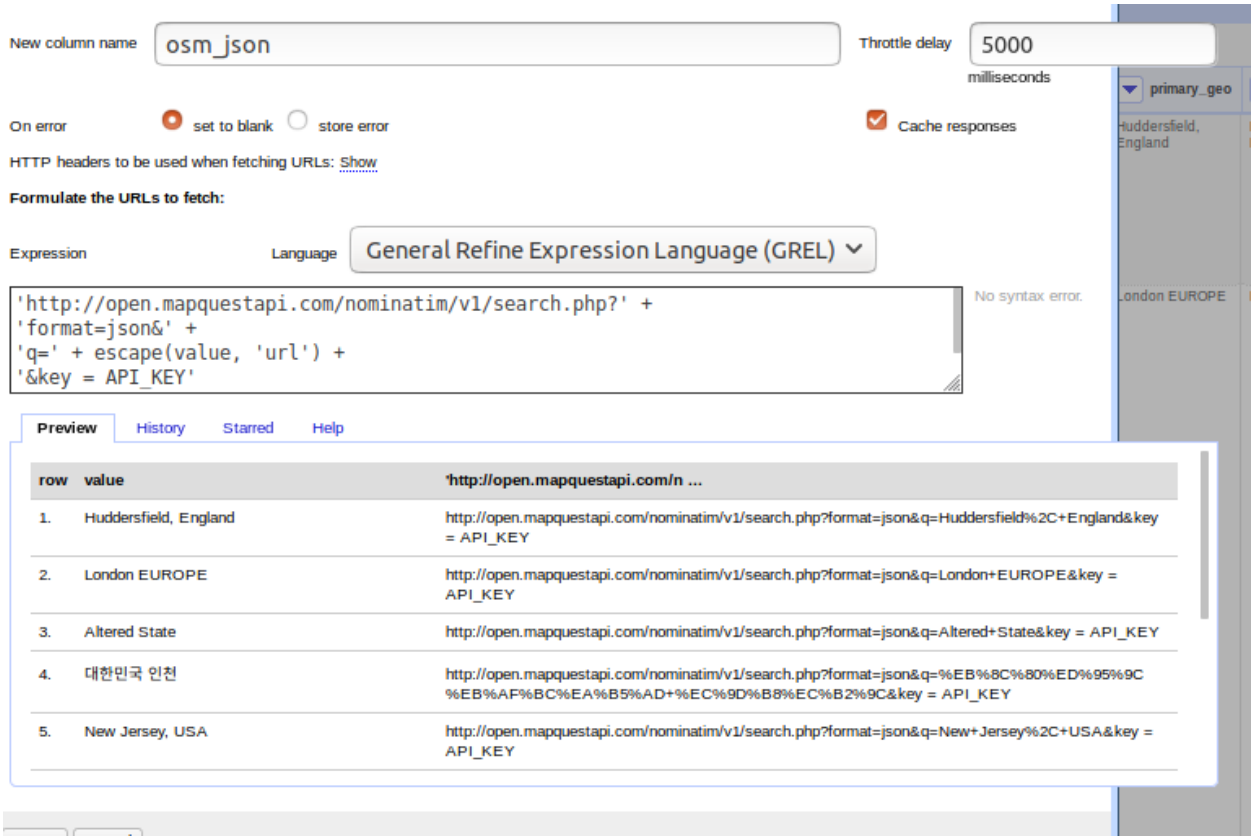
The resultant json file contains a dictionary with the following features:


```
{
    data: [
        {
            user_id: 1170960878856003600,
            features: {
                screen_name: TheOne64977671,
                location: Huddersfield, England,
                tweets: 1,
                geo_type: User location,
                primary_geo: Huddersfield, England,
                id: 1170960878856003600,
                name: The One
            }
        },
```

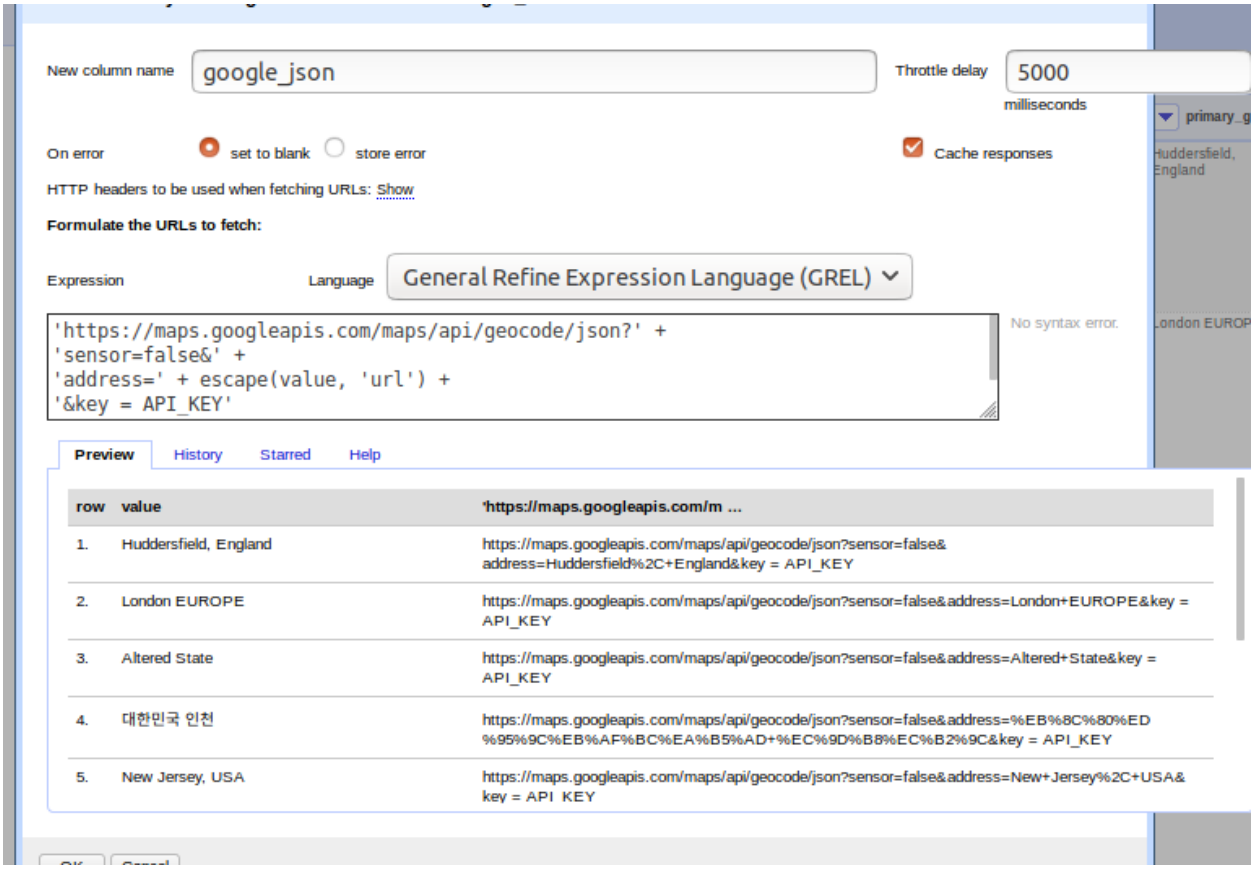
Step 5: Geocoding the Location details

The Google Geocoding API and OpenStreetMap API were used to get the location addresses in order to get the latitude and longitude details. Both APIs were used for purposes of comparing accuracy and determine which API was more reliable.

Using OpenRefine to geocode using OpenStreetMap API

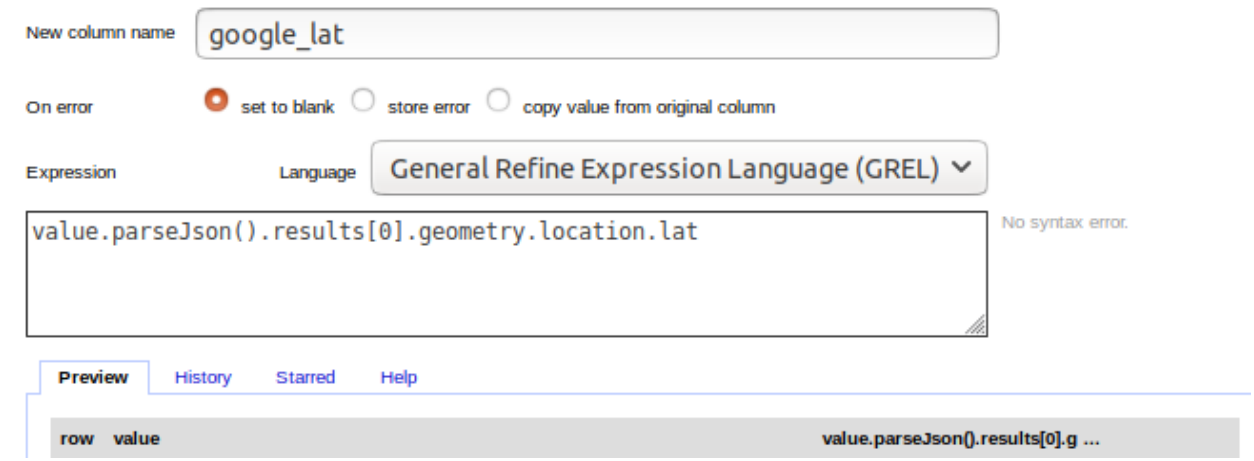


Using OpenRefine to geocode using Google API



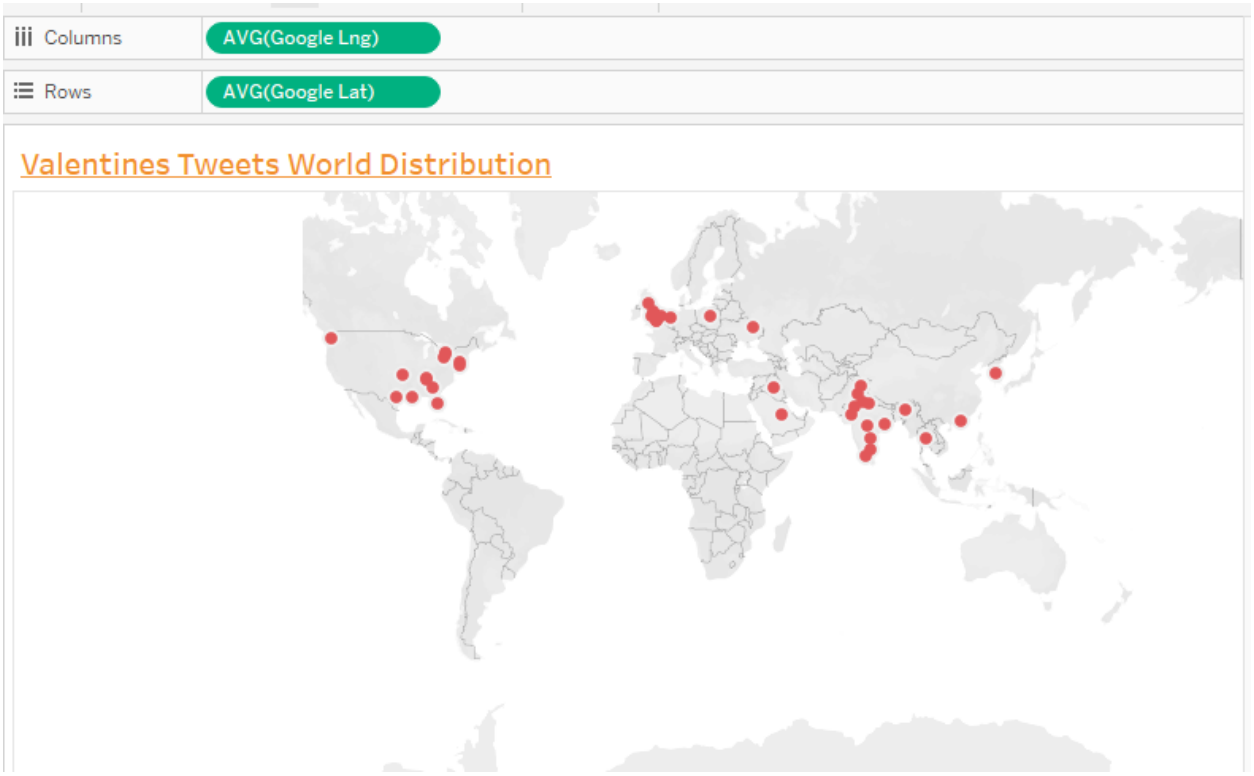
From the geocoded json object returned by the APIs, latitude and longitude details were extracted to individual columns using OpenRefine javascript function.

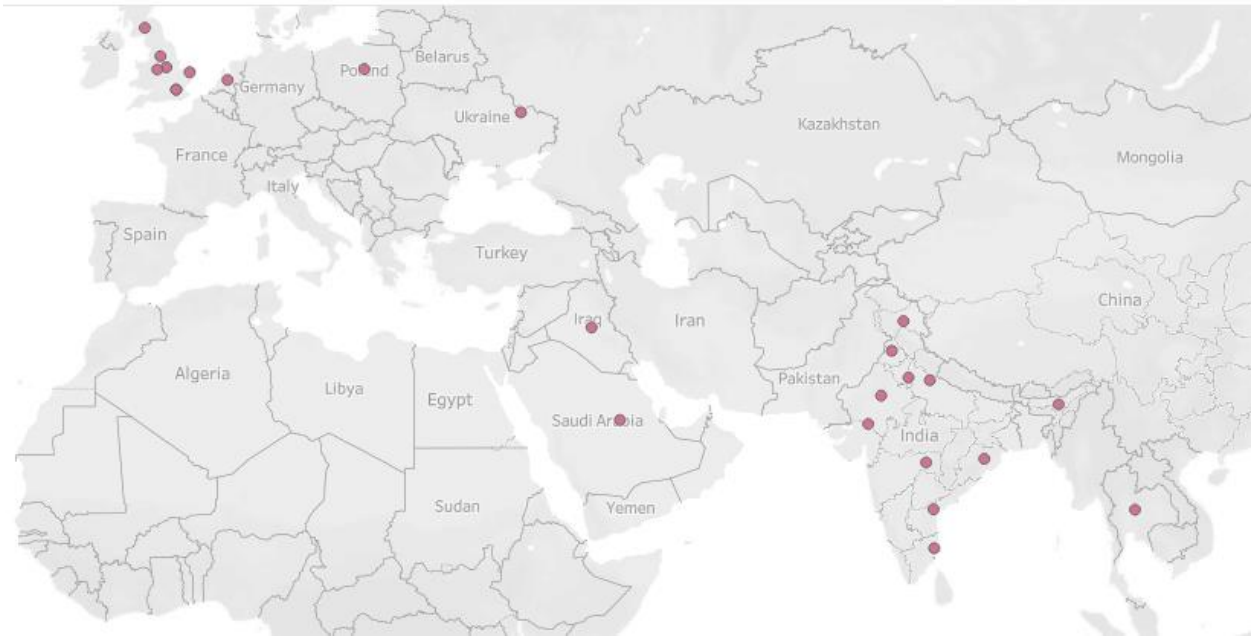
Extracting latitude values as a column using .parseJson( ) function

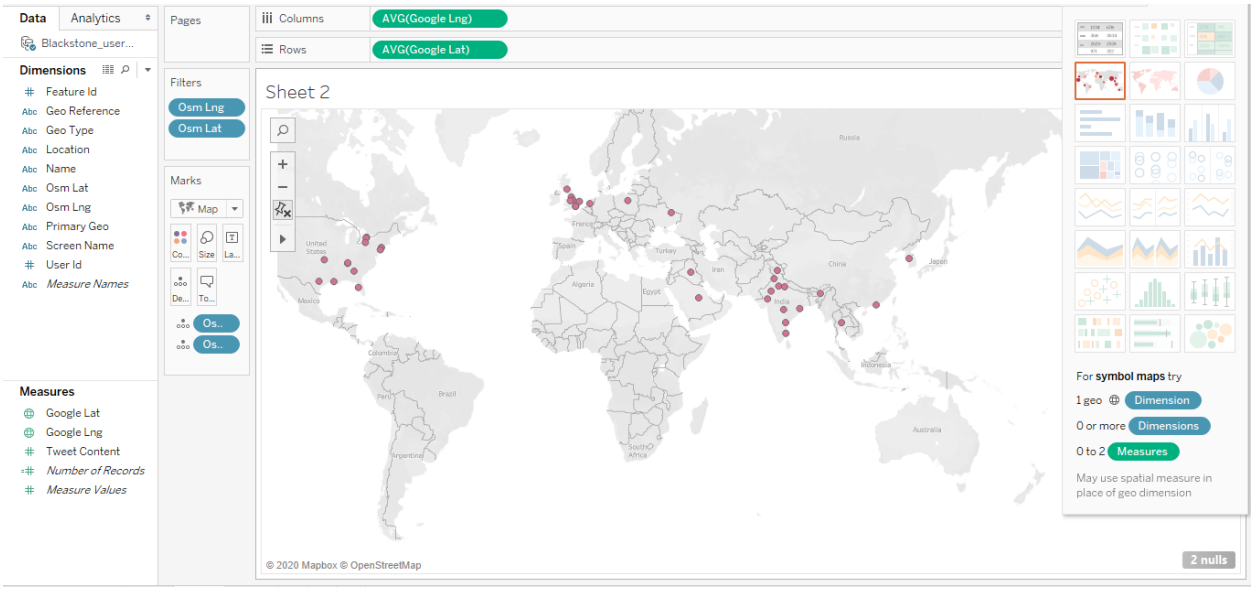


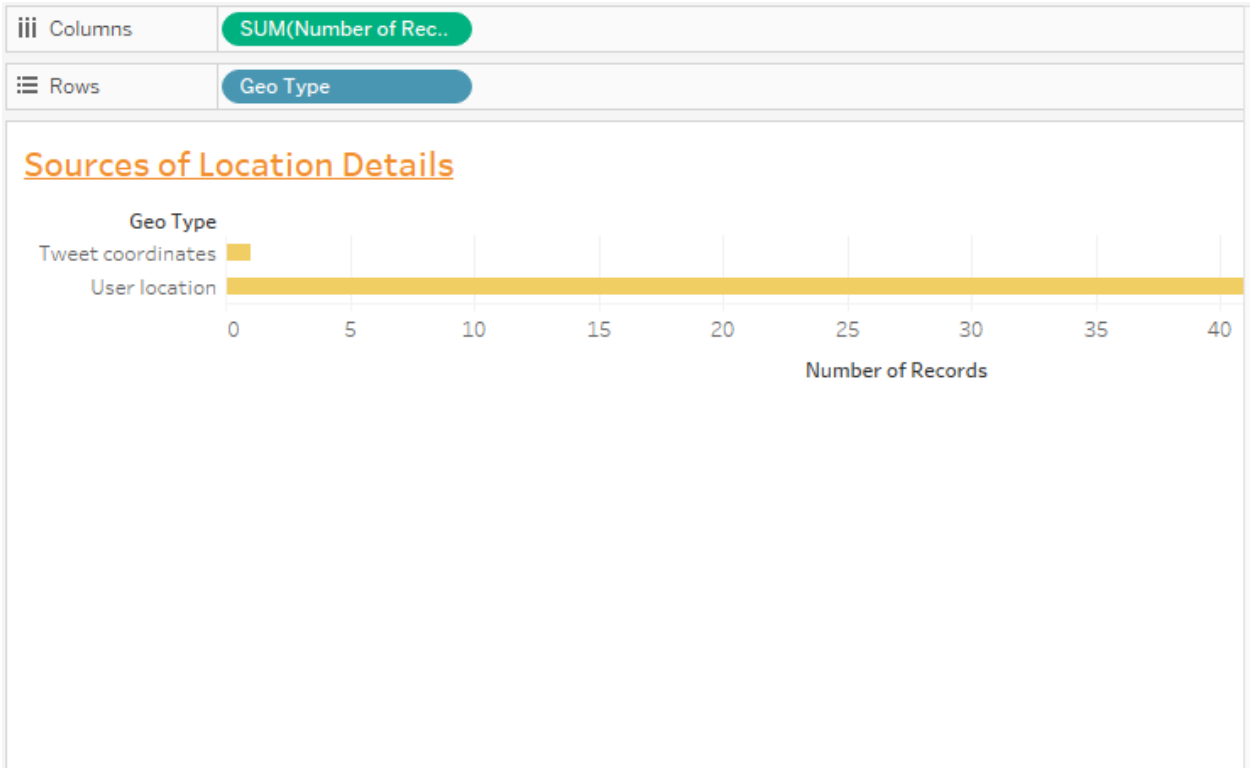
Step 6: Data Visualization

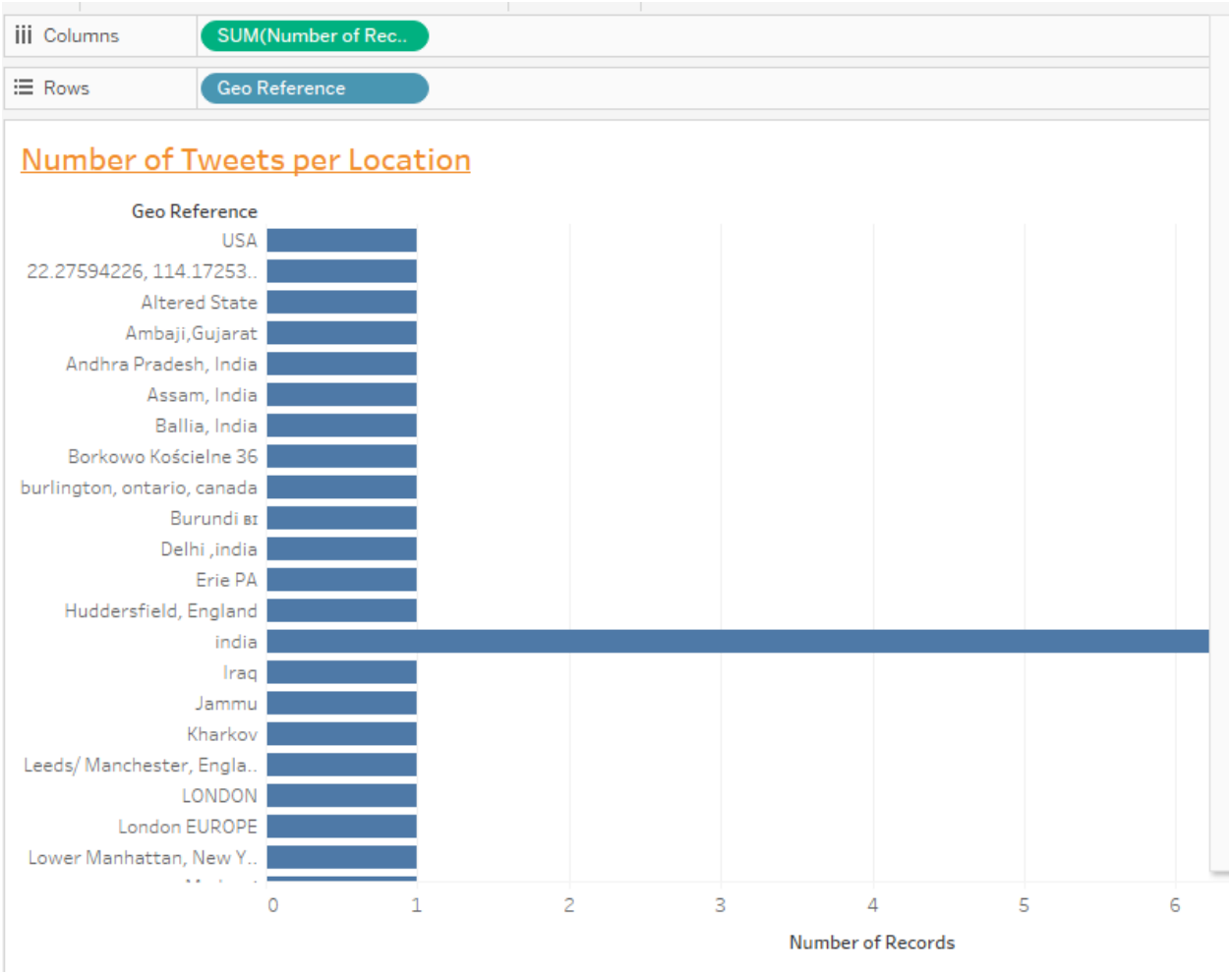
Tableau public software was used to visualize the data to help in testing the hypothesis and for better interpretation of the results. The geodata from the user tweets were plotted on a Symbol Map with the geo coordinates plotted.




Belarus
Poland
Germany
Ukraine
Kazakhstan
France
Mongolia
Italy
Spain
Turkey
China
Iraq
Iran
Algeria
Libya
Egypt
Pakistan
Saudi Arabia
India
Sudan
Yemen

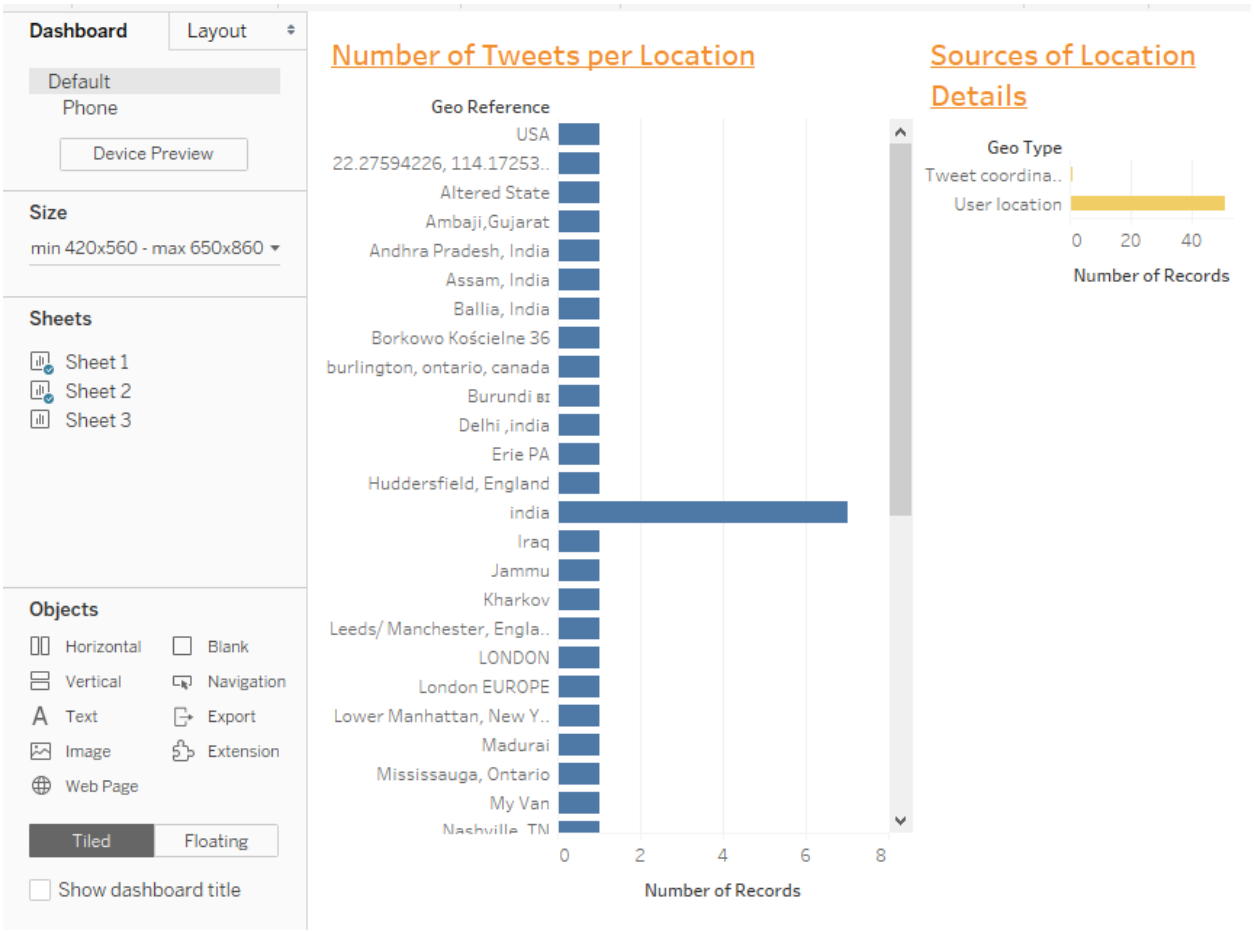

Data
Analytics
Blackstone_user...
Dimensions
Feature Id
Geo Reference
Geo Type
Location
Name
Osm Lat
Osm Lng
Primary Geo
Screen Name
User Id
Measure Names
Measures
Google Lat
Google Lng
Tweet Content
Number of Records
Measure Values
Pages
Filters
Osm Lng
Osm Lat
Marks
Map
Columns
AVG(Google Lng)
Rows
AVG(Google Lat)
Sheet 2
For symbol maps try
1 geo Dimension
0 or more Dimensions
0 to 2 Measures
May use spatial measure in place of geo dimension
© 2020 Mapbox © OpenStreetMap
2 nulls


Columns
SUM(Number of Rec..
Rows
Geo Type
Sources of Location Details
Geo Type
Tweet coordinates
User location
0
5
10
15
20
25
30
35
40
Number of Records



Step 7: Hypothesis Testing and Conclusion

Out of the 57 samples of the tweet objects with geo data, none of them were from Africa and there we can reject the null hypothesis and adapt the alternative hypothesis concluding that Valentines celebrations are not part of the African culture.



Dashboard Screenshot