

MACHARIA ANASTASIA

TWITTER DATA ANALYSIS CHALLENGE

Step 1: Hypothesis Selection

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

Alternative hypothesis (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.

```
id: 1170960878856003600,  
id_str: 1170960878856003584,  
name: The One,  
screen_name: TheOne64977671,  
location: Huddersfield, England,  
url: null,  
description: Just Following Town,  
translator_type: none,  
protected: false,  
verified: false,  
followers_count: 2,  
friends_count: 26,  
listed_count: 0,  
  
geo: null,  
coordinates: null,  
place: null,  
contributors: null,  
retweeted_status: {  
  created_at: Fri Feb 14 11:43:02 +0000 2020,  
  id: 1228283505366519800,  
  id_str: 1228283505366519800,  
  text: Happy Valentines Day! ♥ This #FreebieFriday will feature on our F:  
  display_text_range: [  
    0,  
    140  
  ]  
}
```

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.

Dictionary to store new data

```
user_data = {
    "user_id": tweet['user']['id'],
    "features": {
        "name": tweet['user']['name'],
        "id": tweet['user']['id'],
        "screen_name": tweet['user']['screen_name'],
        "tweets": 1,
        "location": tweet['user']['location'],
    }
}

user_data1 = list(user_data)

if tweet['coordinates']:
    user_data["features"]["location"] = tweet['coordinates'][tweet['coordinates'][1]][1] +
    ", " + tweet['coordinates'][tweet['coordinates'][1]][0]
    user_data["features"]["geo_type"] = "Tweet coordinates"
elif tweet['place']:
    user_data["features"]["primary_geo"] = tweet['place']['full_name'] + ", "
    + tweet['place']['country']
    user_data["features"]["geo_type"] = "Tweet place"

else:
    user_data[1][4] = user_data[[1][4] + user_data[1][4]
    user_data["features"]["geo_type"] = "User location"

if user_data["features"]["primary_geo"]:
    users_with_geodata['data'].append(user_data)
    geo_tweets += 1
```

```
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

New column name: Throttle delay: milliseconds

On error: ☒ set to blank ☐ store error ☒ Cache responses

HTTP headers to be used when fetching URLs: [Show](#)

Formulate the URLs to fetch:

Expression: Language: No syntax error.

Preview

row	value	'http://open.mapquestapi.com/n ...
1.	Huddersfield, England	http://open.mapquestapi.com/nominatim/v1/search.php?format=json&q=Huddersfield%2C+England&key = API_KEY
2.	London EUROPE	http://open.mapquestapi.com/nominatim/v1/search.php?format=json&q=London+EUROPE&key = API_KEY
3.	Altered State	http://open.mapquestapi.com/nominatim/v1/search.php?format=json&q=Altered+State&key = API_KEY
4.	대한민국 인천	http://open.mapquestapi.com/nominatim/v1/search.php?format=json&q=%EB%8C%80%ED%95%9C%EB%AF%BC%EA%B5%AD+%EC%9D%B8%EC%B2%9C&key = API_KEY
5.	New Jersey, USA	http://open.mapquestapi.com/nominatim/v1/search.php?format=json&q=New+Jersey%2C+USA&key = API_KEY

Using OpenRefine to geocode using Google API

New column name: Throttle delay: milliseconds

On error: ☒ set to blank ☐ store error ☒ Cache responses

HTTP headers to be used when fetching URLs: [Show](#)

Formulate the URLs to fetch:

Expression: Language: No syntax error.

Preview History Starred Help

row	value	'https://maps.googleapis.com/m ...
1.	Huddersfield, England	https://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=Huddersfield%2C+England&key = API_KEY
2.	London EUROPE	https://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=London+EUROPE&key = API_KEY
3.	Altered State	https://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=Altered+State&key = API_KEY
4.	대한민국 인천	https://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=%EB%8C%80%ED%95%9C%EB%AF%BC%EA%B5%AD+%EC%9D%B8%EC%B2%9C&key = API_KEY
5.	New Jersey, USA	https://maps.googleapis.com/maps/api/geocode/json?sensor=false&address=New+Jersey%2C+USA&key = API_KEY

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

New column name

osm_lat

On error

☒ set to blank

☐ store error

☐ copy value from original column

Expression

Language

General Refine Expression Language (GREL) ▾

value.parseJson()[0].lat

No syntax error.

Preview

History

Starred

Help

row	value	value.parseJson()[0].lat
1.	[{"place_id":"140129","licence":"Data © OpenStreetMap contributors, ODbL 1.0. https://www.openstreetmap.org/copyright","osm_type":"node","osm_id":"26703087","boundingbox":["53.6066645","53.6866645",-1.8222482",-1.7422482"],"lat":"53.6466645","lon":"-1.7822482","display_name":"Huddersfield, Yorkshire and the Humber, England, HD1 2AN, UK","class":"place","type":"town","importance":0.35197523239197,"icon":"http://Vip-10-98-171-248.mq-us-east-1.ec2.aolcloud.net/nominatim/images/mapicons/Vpoi_place_town.p.20.png"},{"place_id":"105416","licence":"Data © OpenStreetMap contributors, ODbL 1.0. https://www.openstreetmap.org/copyright","osm_type":"node","osm_id":"3123212","boundingbox":["53.643511","53.653511",-1.7898033,-1.7798033"],"lat":"53.648511","lon":"-1.7848033","display_name":"Huddersfield, St George's Square, Springwood, Highfields, Kirklees, Yorkshire and the Humber, England, HD1 1JB, UK","class":"railway","type":"station","importance":0.30751342970414,"icon":"http://Vip-10-98-171-248.mq-us-east-1.ec2.aolcloud.net/nominatim/images/mapicons/Vtransport_train_station2.p.20.png"}]	53.6466645

New column name

google_lat

On error

☒ set to blank

☐ store error

☐ copy value from original column

Expression

Language

General Refine Expression Language (GREL) ▾

value.parseJson().results[0].geometry.location.lat

No syntax error.

Preview

History

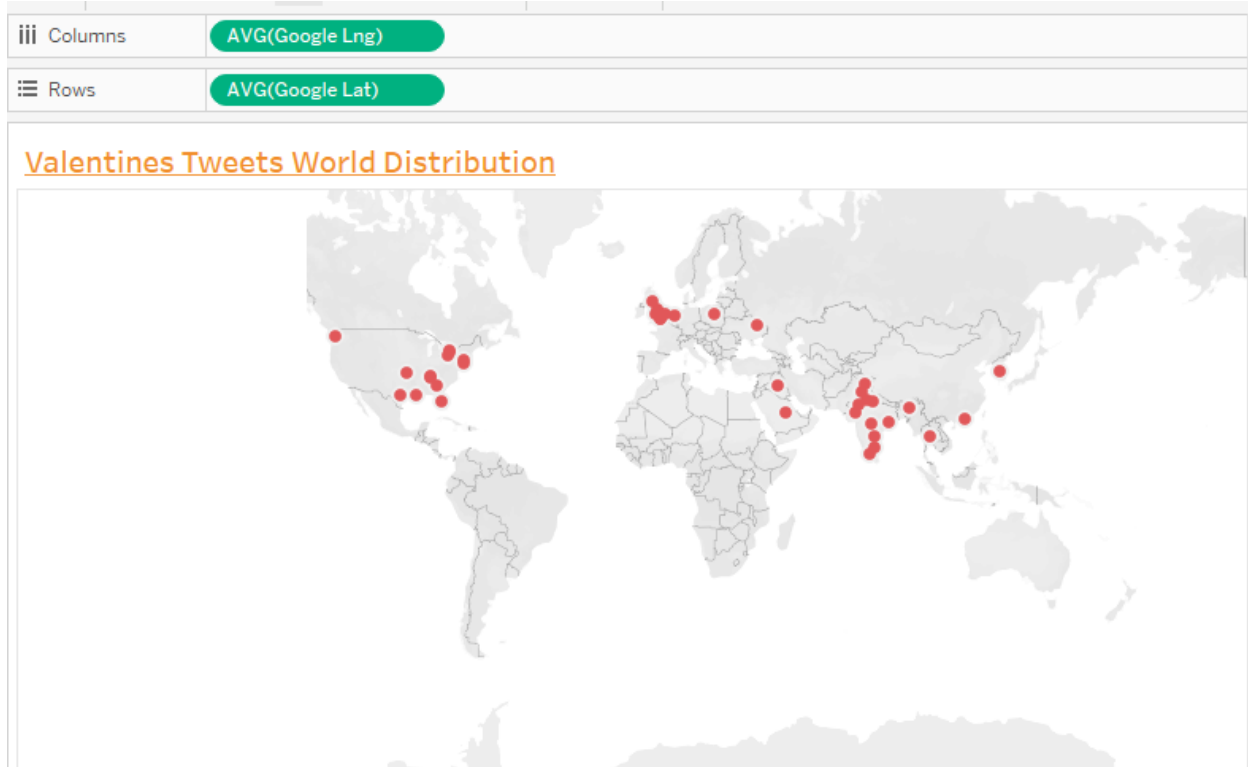
Starred

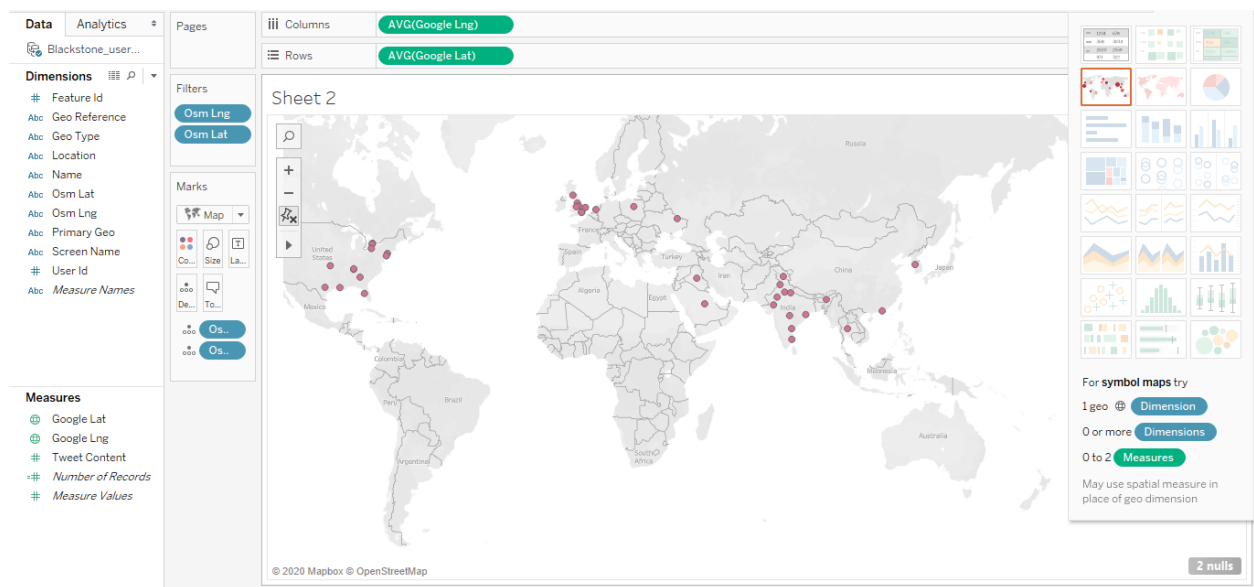
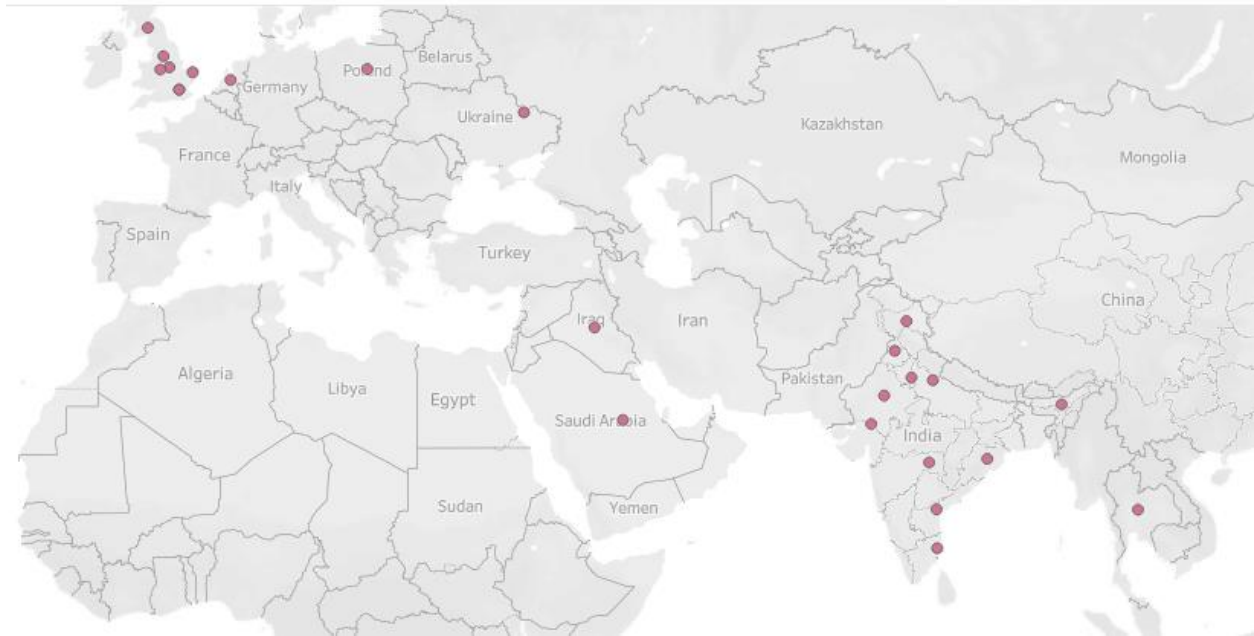
Help

row	value	value.parseJson().results[0].g ...
-----	-------	------------------------------------

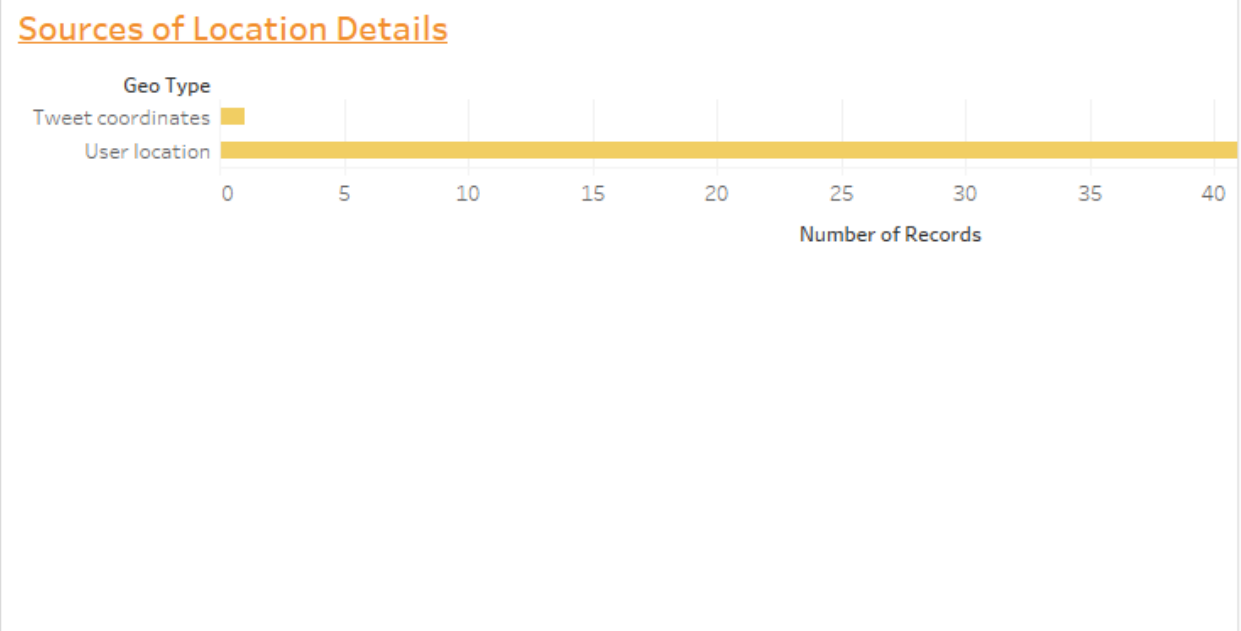
Step 6: Data Visualization

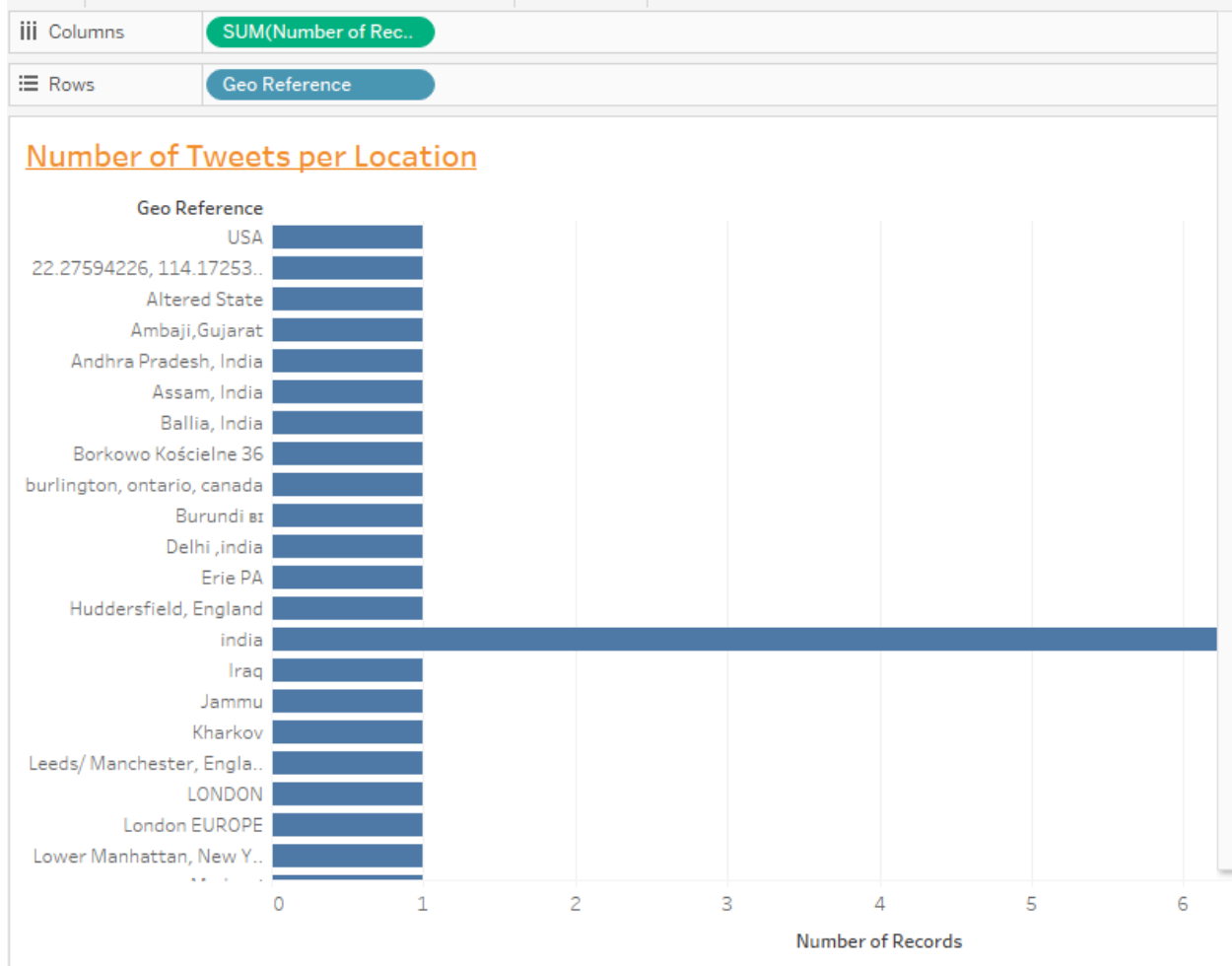
Tableau public software was used to visualize the data to help in testing the hypothesis and for better interpretation of the results. The geodata from the user tweets were plotted on a Symbol Map with the geo coordinates plotted.





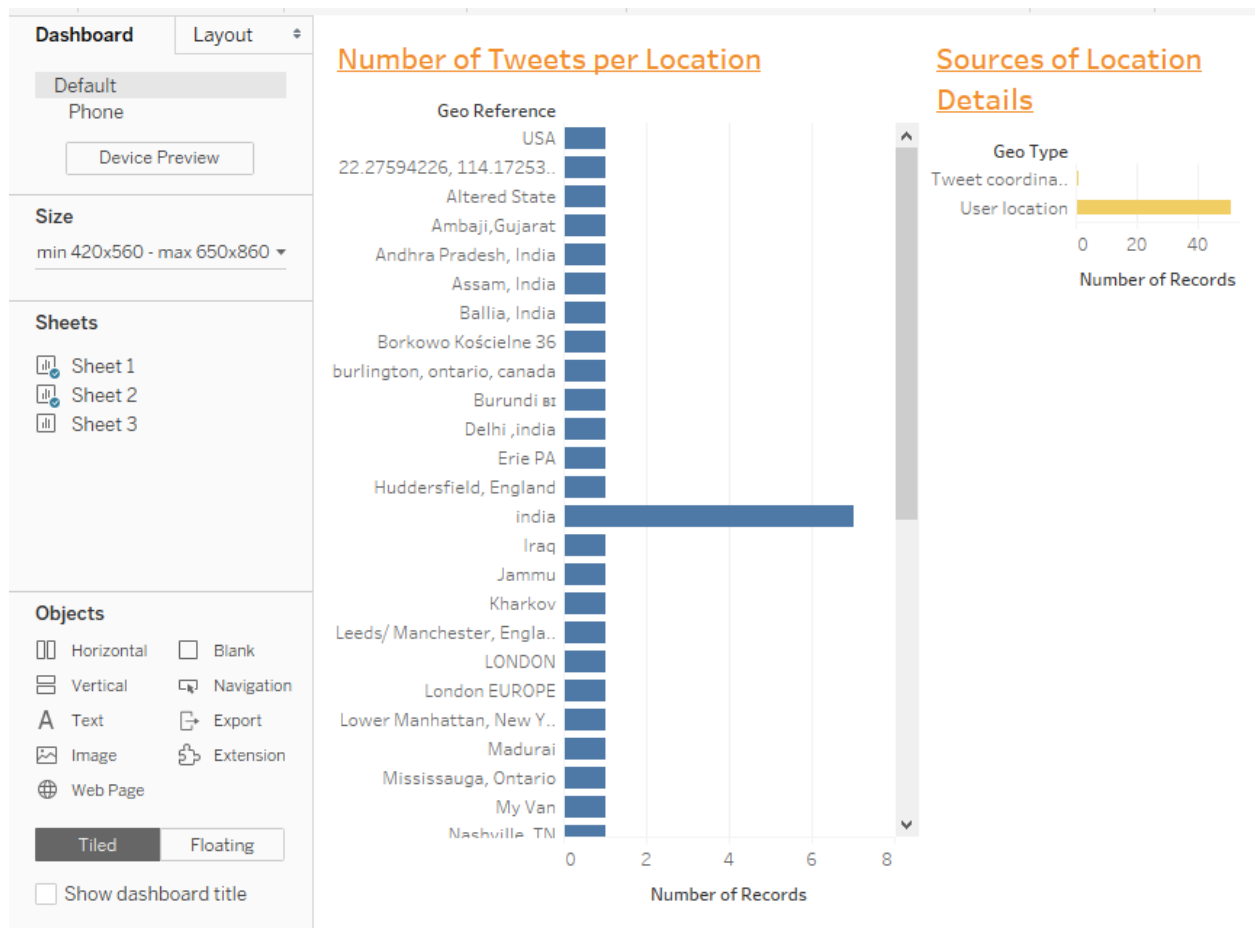
Columns	SUM(Number of Rec..
Rows	Geo Type





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