**MACHARIA ANASTASIA**

**TWITTER DATA ANALYSIS CHALLENGE**

**Step 1: Hypothesis Selection**

Null hypothesis (h0) = Valentines Celebration is part of the African Culture

Alternative hypothesis (**h1**) = Valentines Celebration is not part of the African Culture

**Step 2: Hashtag Selection, Period, Dataset Size**

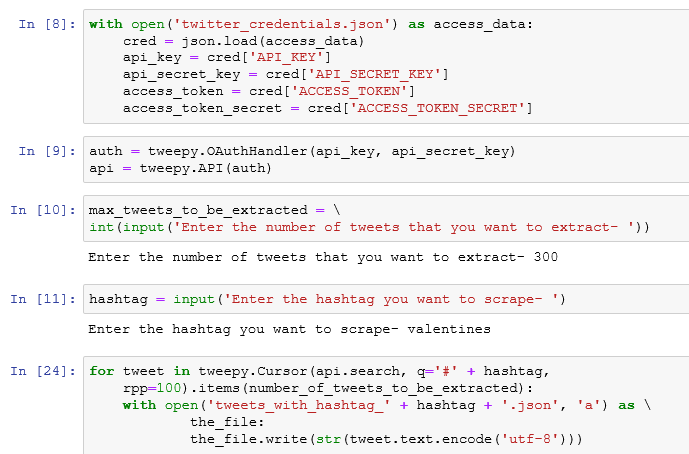
Hashtag: #valentines

Period: 1st February 2020 to 15th Feb

Size of data: 300 entries

**Step 3: Data Scrapping from Twitter**

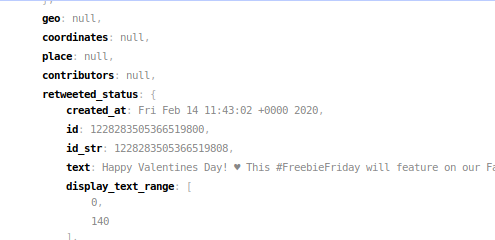
The Tweepy library was used to scrap data from the Twitter API using the Twitter developer’s platform credentials.



**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.





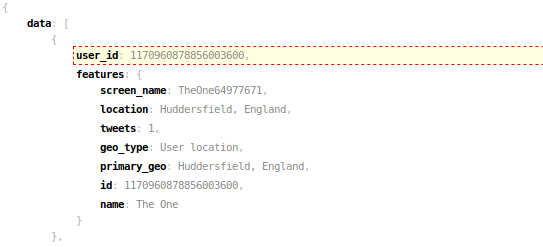
**Step 4: Data Analysis**

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.





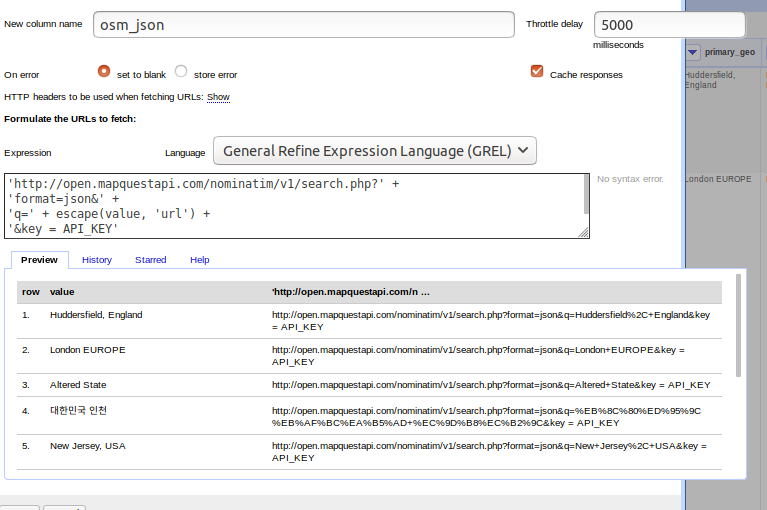
The resultant json file contains a dictionary with the following features:



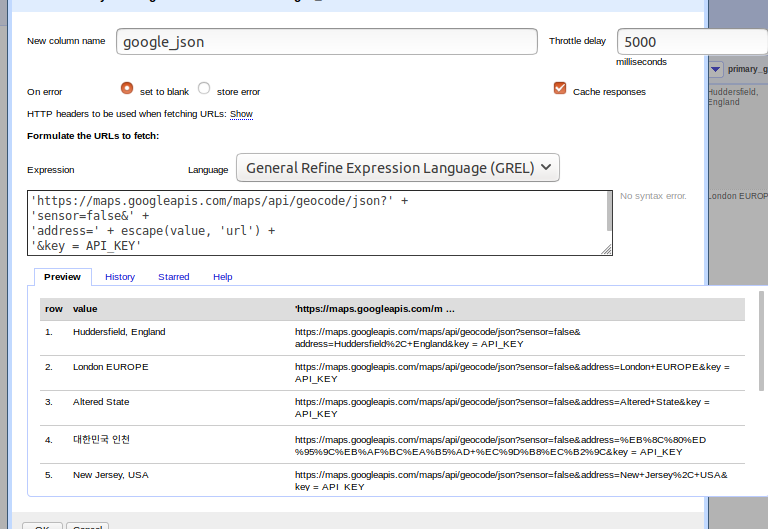
**Step 5: Geocoding the Location details**

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**



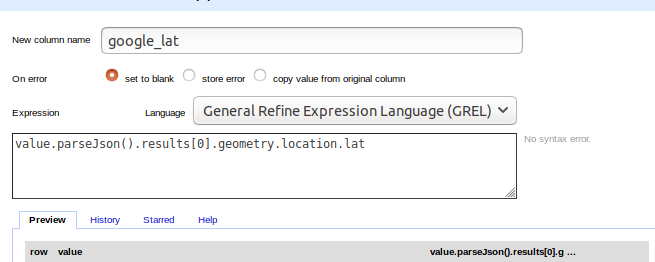
**Using OpenRefine to geocode using Google API**



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

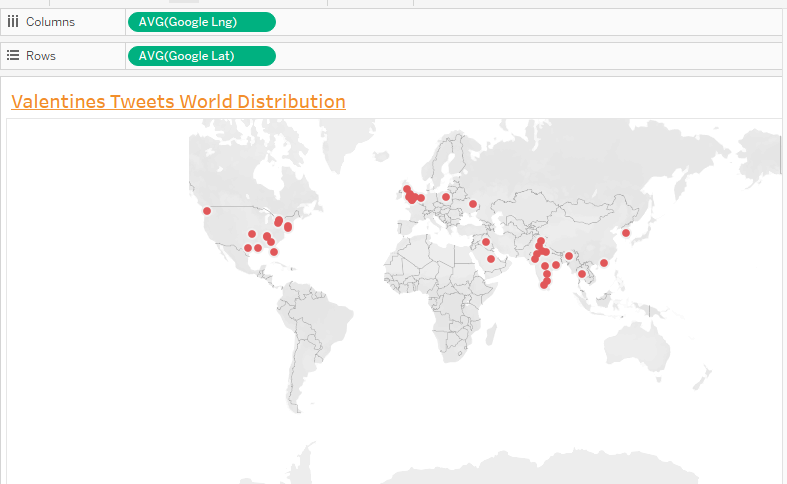
**Extracting latitude values as a column using .parseJson( ) function**

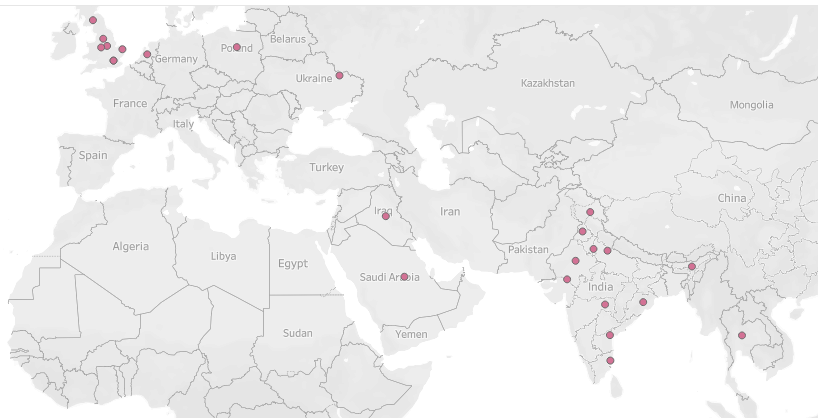


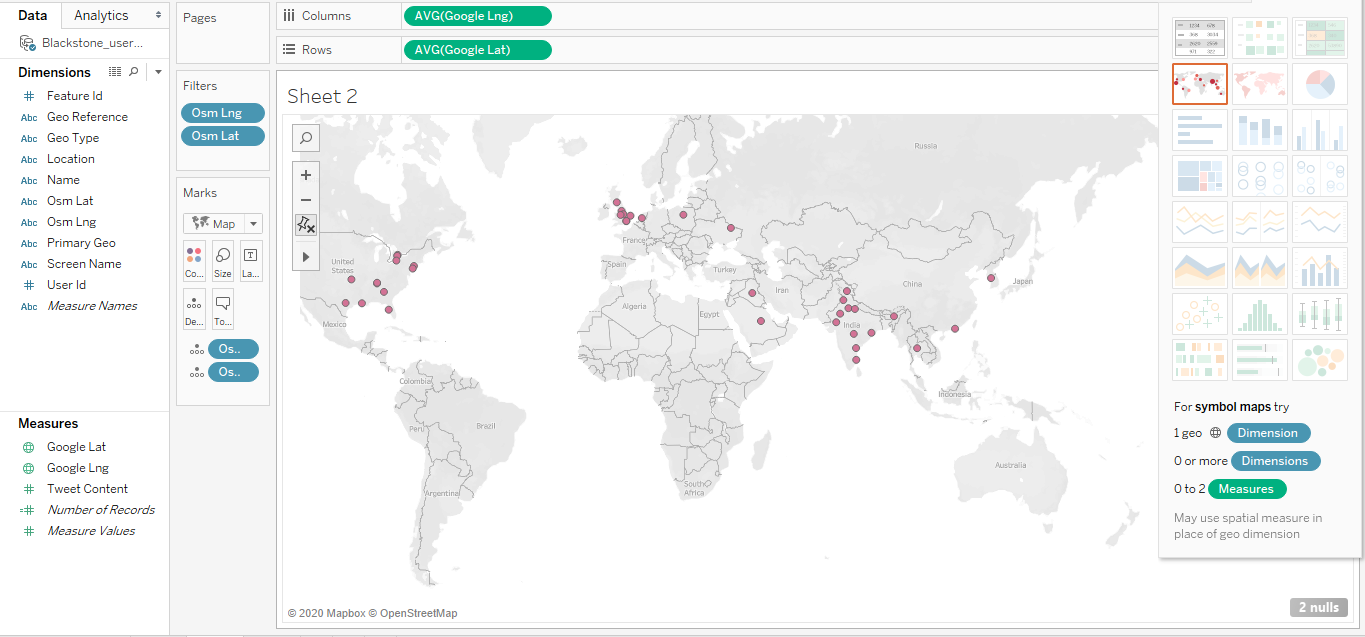


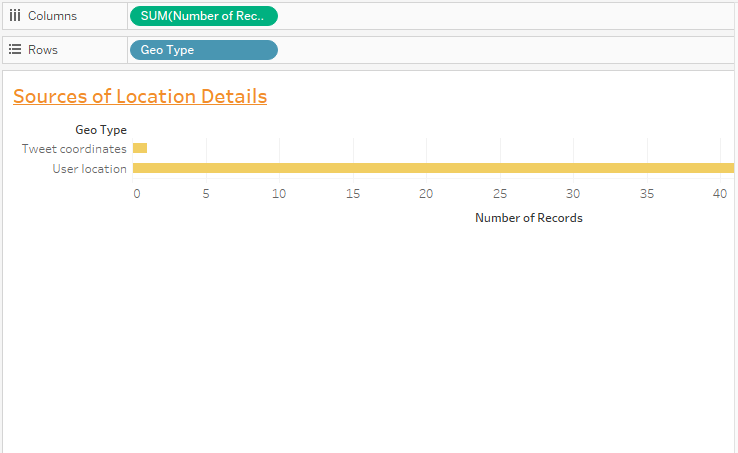
**Step 6: Data Visualization**

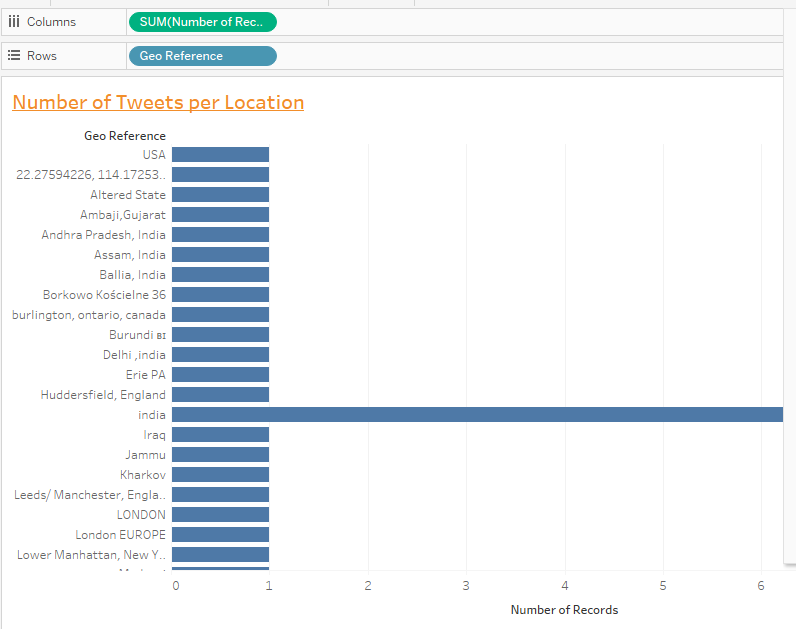
Tableau public software was used to visualize the data to help in testing the hypothesis and for better interpretation of the results. The geodata from the user tweets were plotted on a Symbol Map with the geo coordinates plotted.





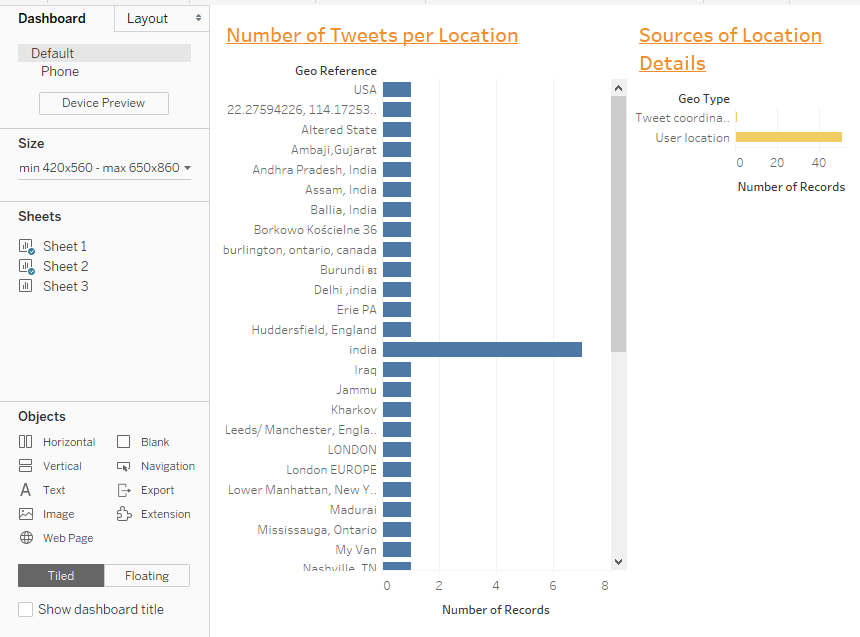






**Step 7: Hypothesis Testing and Conclusion**

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