Bhargavi Poyekar
CH33454

Assignment 3

DATA ANALYSIS AS A SUPPLEMENT TO VISUALIZATION

Analysis Algorithm Used: Clustering

Analytic Goals:

Our project aims to analyze the social network across geographic locations. Hence, I performed an analysis of the following:

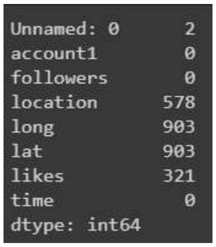
1. Clustering profiles based on the location.
2. Clustering posts based on the number of likes, and followers.
3. Analyze the location posts with the most likes.

Preprocessing:

Since we created our own dataset, we had to perform web scraping to collect it. The raw data collected comprised a lot of noise and non-essential columns. I cleaned and preprocessed the data by removing the null values, and unnecessary columns and transforming the data. Final dataset consists of the attributes:

Dataset Description:
Account1- User profile name- data type: string
Followers- No. of followers- data type: original-string, transformed-int
Location- Location name of post- data type: string
Long- Longitude coordinate- data type: float
Lat- Latitude Coordinate- data type: float
Likes- Likes on the post- data type: original-string, transformed-int
Time- Time of the post- data type: Time

Null Values: Missing Values were '_', so I replaced '_' with Nan values.



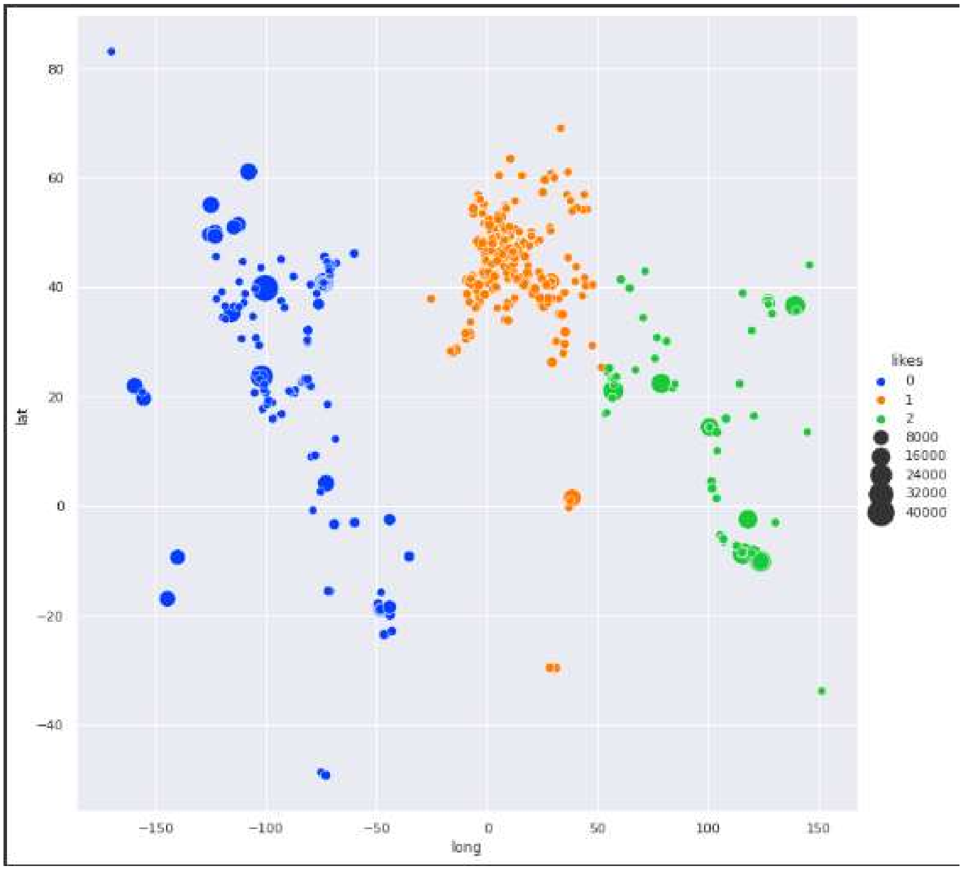
```
Unnamed: 0      2
account1        0
followers       0
location      578
long          903
lat           903
likes         321
time            0
dtype: int64
```

Dataset consisted of more than 2400 points, so I removed the null values, as the remaining points were enough for me to analyze the dataset and it didn't make any sense to replace missing values with average or mode values here.

Data Transformation: The likes and followers columns were present as strings and likes were in the format 'no. of likes + likes'. So I split the values, removed the trailing 'likes' and converted them to int.

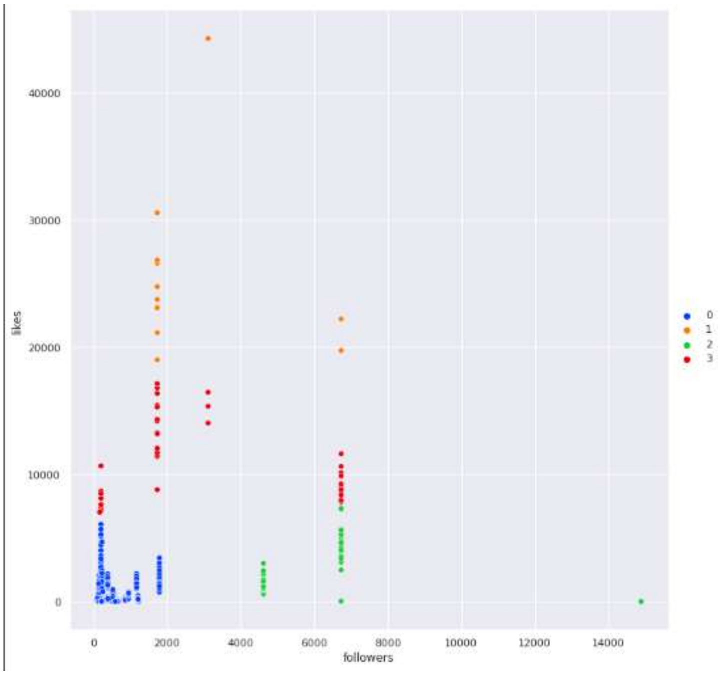
Clustering:

1. **Clustering posts based on the location**: I used latitude and longitude coordinates for getting the exact location of the post and performed K-means clustering on it. I used the elbow method to decide the number of clusters.



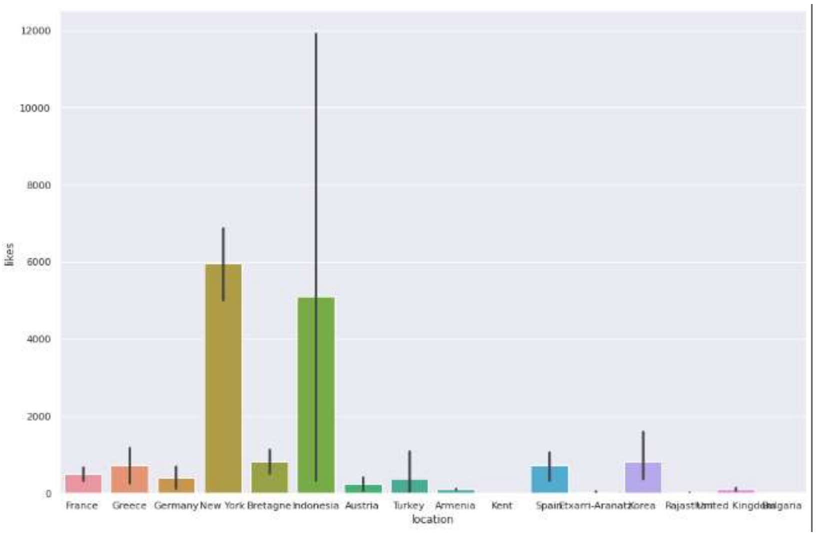
The marks used here are points and the visual channels used are size and color. The color represents the clusters and hence a categorical attribute. The size represents the number of likes for the posts. We can also see here that the posts belonging to cluster 1(blue) have the most number of likes and posts belonging to cluster 2(orange) has the least number of likes.

2. **Clustering posts based on the number of likes and followers.**



There are 4 clusters formed based on the elbow method. I hoped to find some kind of trend or correlation between the number of followers and likes, but I think the dataset is not enough for this analysis.

3. **Analyze the location posts with the most likes.**



The above plot shows location vs the number of likes, and from this, I could find that posts from New York and Indonesia have more likes.

Difficulties: Since the data is scraped from the web, I had to do a lot of preprocessing on it, and many values were missing, so I had to compromise on the size of the dataset.

Limitations: Right now the dataset consists of less number of users and the location is not extracted properly. Perfection of the dataset will improve our project.

Conclusion: This analysis can be useful for making our system more interactive and informative.

References:

1. https://seaborn.pydata.org/generated/seaborn.scatterplot.html
2. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
3. https://towardsdatascience.com/seaborn-relplot-in-python-visualising-relationships-in-data-ee39138d53aa
4. https://stackoverflow.com/questions/13413590/how-to-drop-rows-of-pandas-dataframe-whose-value-in-a-certain-column-is-nan