Kevin Chiang
Romil Shah
Leo Dai
Aryan Barik
Krutik Doshi
Salman Alsabah

Rio Data

Introduction:

In recent years, social media platforms such as Twitter have become a rich source of information for various research fields, including social sciences, public health, and urban studies. One specific area of interest is the use of Twitter data to understand the socio-spatial dynamics of informal settlements, such as favelas in Brazil. In this project, we have been provided with a dataset of tweet information and tasked with creating a map that locates people who tweeted in favelas using Python. Our main objective is to read in the data, clean it, and develop a predictive model that identifies Twitter users who live in favelas. To accomplish this task, we will leverage various data science techniques and tools, including data preprocessing, exploratory data analysis, machine learning algorithms, and visualization. Finally, we will present our findings in a report using Jupyter Notebook and export it as a PDF. This project has the potential to shed light on the spatial patterns of Twitter usage in favelas and contribute to the broader understanding of the socio-economic dynamics of these informal settlements.

Methodology:

The methodology for this project involves several steps. Firstly, we set the file path and read the JSON file into a DataFrame using a JSON parser. We then print the first 5 rows of the DataFrame to inspect the data. To ensure that our analysis is focused on relevant columns, we remove any irrelevant columns from the DataFrame. We also drop any rows with missing or null values.

Next, we extract the two numbers from the 'coordinates' column and create new columns in the DataFrame. If necessary, we drop the original 'coordinates' column. We then extract the features for clustering and instantiate the KMeans object. We fit the data to the KMeans model and add the cluster labels to the DataFrame.

To create the map, we first center it at the mean of the data points. We then define the colors for the clusters and group the data points by their cluster labels. We iterate over each group and add the data points to the map with a different color. Finally, we save the map as an HTML file, download it, and click on the file to open the map. With this map we can compare the clustered points to the actual district regions on the map to see where potential favelas are. Any cluster that doesn't quite add up to the actual district on the map can be thus labeled as a favela. We can also use clues such as the size of each cluster and how dense or spread apart to help us determine whether it is a favela or not. Finally after finding the favela locations we can

use an algorithm to determine whether tweets are located near a favela or not to determine if the tweets come from favelas. Overall, this methodology leverages various Python libraries and techniques, including data preprocessing, machine learning, and visualization. It enables us to create a map that locates people who tweeted in favelas and contributes to our understanding of the socio-spatial dynamics of informal settlements.

Results summary:

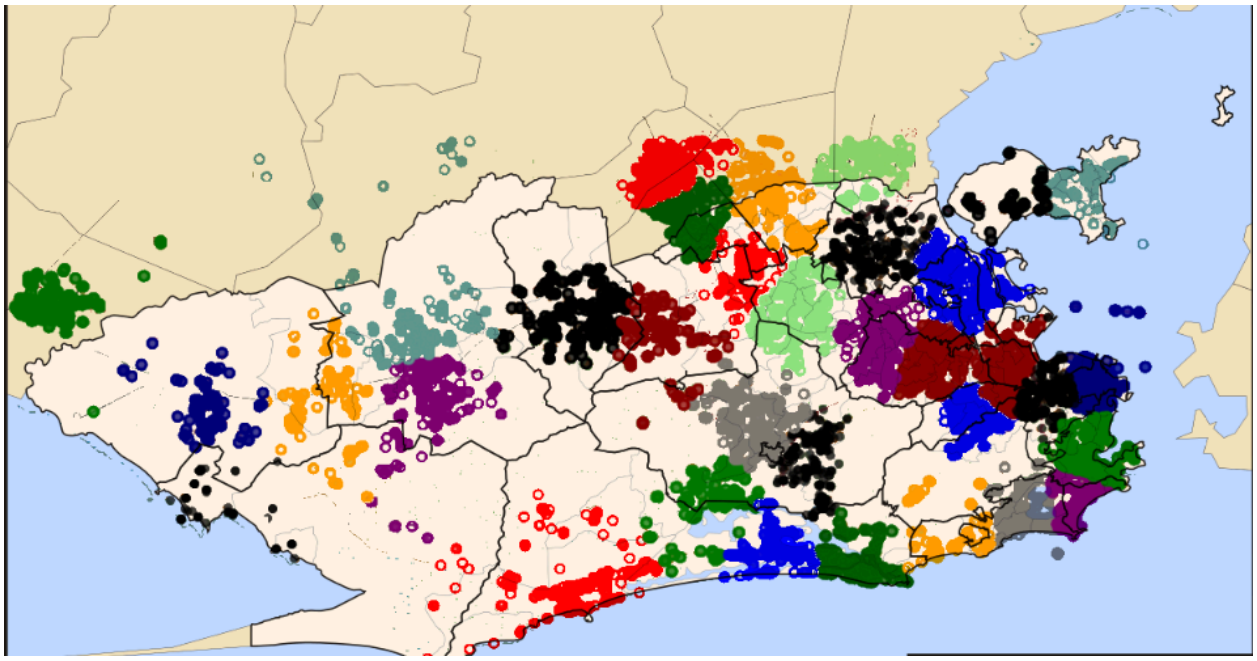

Figure 1: K-means clustering performed on provided Brazil data

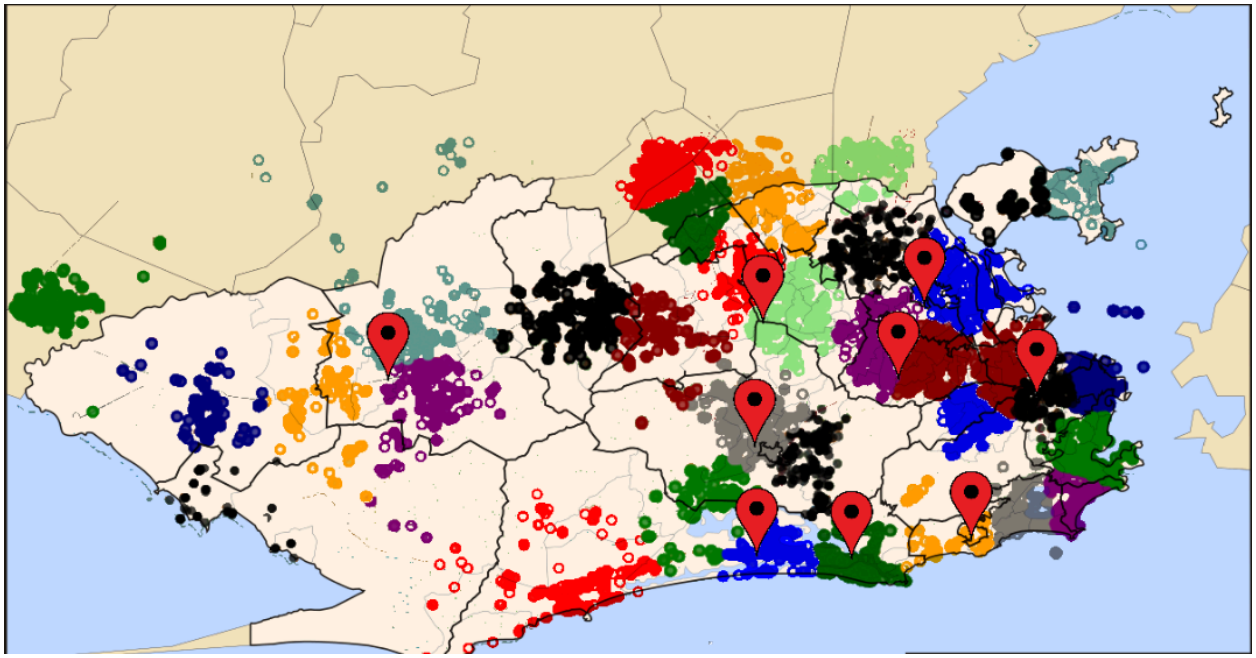

Figure 2: Pins of favela locations placed on the clustering map from Figure 1. Each pin represents an estimated favela location.

```
def isTweetFromFavela(lat, long):
  favelaCords = [[-22.9937, -43.2569],[-23.0044, -43.3297],[-22.9056, -435870],[-22.9409, -43.3758],
                 [-22.8773, -43.3727],[-22.8690, -43.2803],[-22.9098, -43.2947],[-22.9167, -43.2185]]

  for fLong, fLat in favelaCords:
    deltaLat = abs(lat - fLat)
    deltaLong = abs(long - fLong)
    if deltaLong <= 0.2 and deltaLat <= 0.2:
      return True
  return False

print(isTweetFromFavela(-22, 43))
```

Figure 3: Our defined function for determining whether a tweet is from a person living in a favela or not

For the model, we made a simple function that takes in the tweet data and returns whether it is from a favela or not based on location, as seen in Figure 3. From Figure 1, we can see how the clustering algorithm separated all the geographical points from where the tweets originated from. On Figure 2, we can see how certain clusters can be denoted as favelas, as seen from where the pins of true favelas overlap on those clusters.

Conclusion:

Finally, this project successfully applied data science techniques and tools to analyze Twitter data and create a map that locates people who tweeted in Brazilian favelas. Data preparation,

exploratory data analysis, machine learning techniques, and visualization were all part of the methodology. The generated map shows prospective favela sites and helps us to assess whether or not tweets are near a slum. This study has the potential to contribute to a better understanding of the socioeconomic dynamics of informal settlements and can be expanded to investigate additional research issues. Overall using the riodata really helped reinforce our knowledge of libraries such as pandas, sklearn, and numpy and it also helped us learn about new libraries like folium for the maps. It helped enable us to research python libraries that would help our data report and also enabled us to freely work on cleaning the data and organizing the data to be useful to us.

Recommendation:

If Brazilian authorities would like to address the various sociocultural and economic issues that can be caused by favelas, we would recommend them to utilize our geographical clustering model in order for them to find potential residential areas of concern. Our clusters generated from our model were able to accurately find the true locations of favelas.