Analysis of San Francisco Neighborhoods to Open an Indian Restaurant

Siddharth Banerjee

1. INTRODUCTION

As per Wikipedia, California has the highest population of Indian-Americans in the United States of America, of which San Francisco has the highest number of Indians living in the city. Thus, it is an interesting prospect to open an Indian restaurant keeping in mind the high number of Indian population living there. In this analysis, I have attempted to analyze the best possible location to open an Indian restaurant in San Francisco. I have made use of the Foursquare API to explore neighborhoods in San Francisco City. I used the explore function based on a search criteria to search Indian restaurants in the city. Finally, I used the Folium library to visualize the neighborhoods in San Francisco City and their emerging clusters. This should help understand whether it would be a good idea to set up in a high or low concentrated place with Indian restaurants nearby other Indian restaurants keeping in mind competition from similar cuisine restaurant. Simultaneously, I have made use of San Francisco Crimes data obtained from the various Police Departments within the city and converted the addresses into their equivalent latitude and longitude values. The subsequent sections describe in detail my work as below.

2. DATA

This section describes how the data was obtained and the process of cleaning it up and converting it into Pandas data frames.

a. Data Collection:

The Foursquare API will be used to collect location information of Indian restaurants in San Francisco. This data will be obtained using my Client-Id and Client-Secret of my Foursquare account. The crimes data will be taken from the dataset that was used in the Coursera module “Generating Maps with Python”. The idea is to combine information extracted from these two sources and come up with the best possible solution to open the proposed Indian restaurant within San Francisco.

b. Data Cleaning:

The San Francisco crimes data has 150500 rows of data relating to crimes that were committed within the various police departments in the city. There are 13 features in this dataset that comprise of the incident number, category, description of the crime, day of week crime was committed, date, time, police district, resolution, address where the crime was committed, latitude, longitude, and the police department id. This dataset belongs to 2016 and details the various crimes reported under various police departments within the city.

3. METHODOLOGY

In this section, I will describe my thought process behind how to select the best location to open an Indian restaurant. Initially, I will explain the exploratory data process on the crimes data to understand the statistics of weekdays and the police districts and whether there are any relations of these variables with the crimes committed.

a. Feature Selection:

As one can see within the dataset, 150500 rows is a huge dataset with 13 features. Of this, one row has 'nan' value for its 'PdDistrict' and thus this row has been deleted. On probing further, I see that there are 107780 rows with 'Resolution' of the crime as None – this means that there might not have a crime committed at all for these 107780 rows. Thus, I decided to work with only those crimes that has a resolution of "Arrest, Booked". The benefit of this is twofold – first, this will increasing the computation power and second, this simplification will lead to use only those crimes data that were confirmed and pose a risk to the security of individuals within a police department. After cleaning this data, I was able to reduce the dataset to 39416 rows with just 3 features, selecting just the name of the police district, latitude, and longitudes of the rows.

b. Crime Data Variables

One of the variables within the crimes dataset is 'DayOfWeek' that denotes the day in the week the crime was committed. This is an important variable to understand the days of the week more crimes are committed because most patrons frequent restaurants on Fridays and Saturdays as it is considered the start of the weekend. So it becomes necessary to understand the insights from the crimes dataset and know the days of the week crimes were committed. To this regard, I grouped the rows based on the 'DayOfWeek' and the 'PdDistrict' to obtain the counts based on these conditions.

To visualize this information, I plotted box plots for these counts based on the 'DayOfWeek' and 'PdDistrict' in figures 1 and 2 respectively. The reason of plotting boxplots is because both the variables mentioned are categorical variables and the row counts for these variables is a continuous variable. Thus, boxplots are a better way to visualize the combination of categorical and continuous variables.

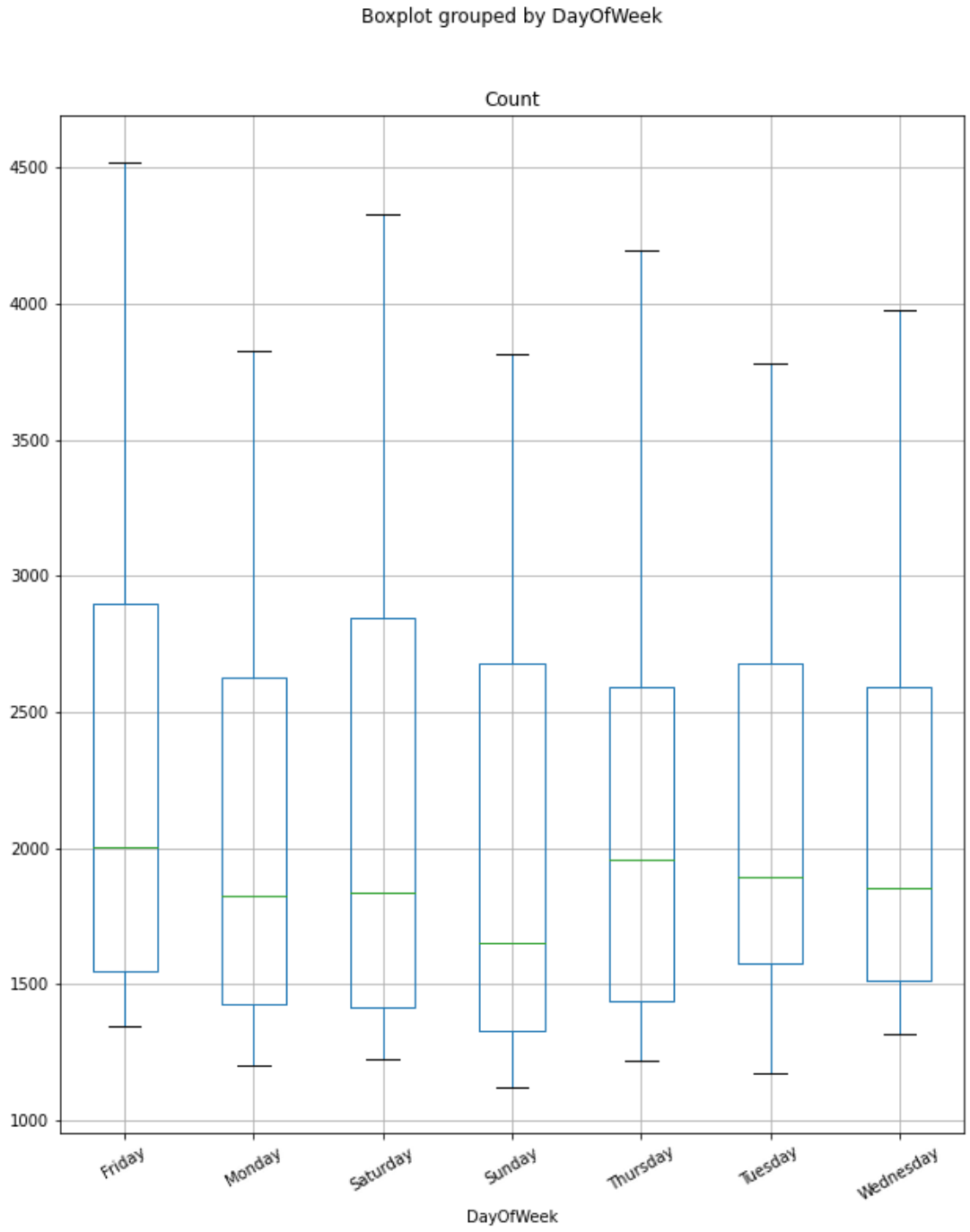

Figure 1. Boxplot of Crimes against ‘DayOfWeek’ variable

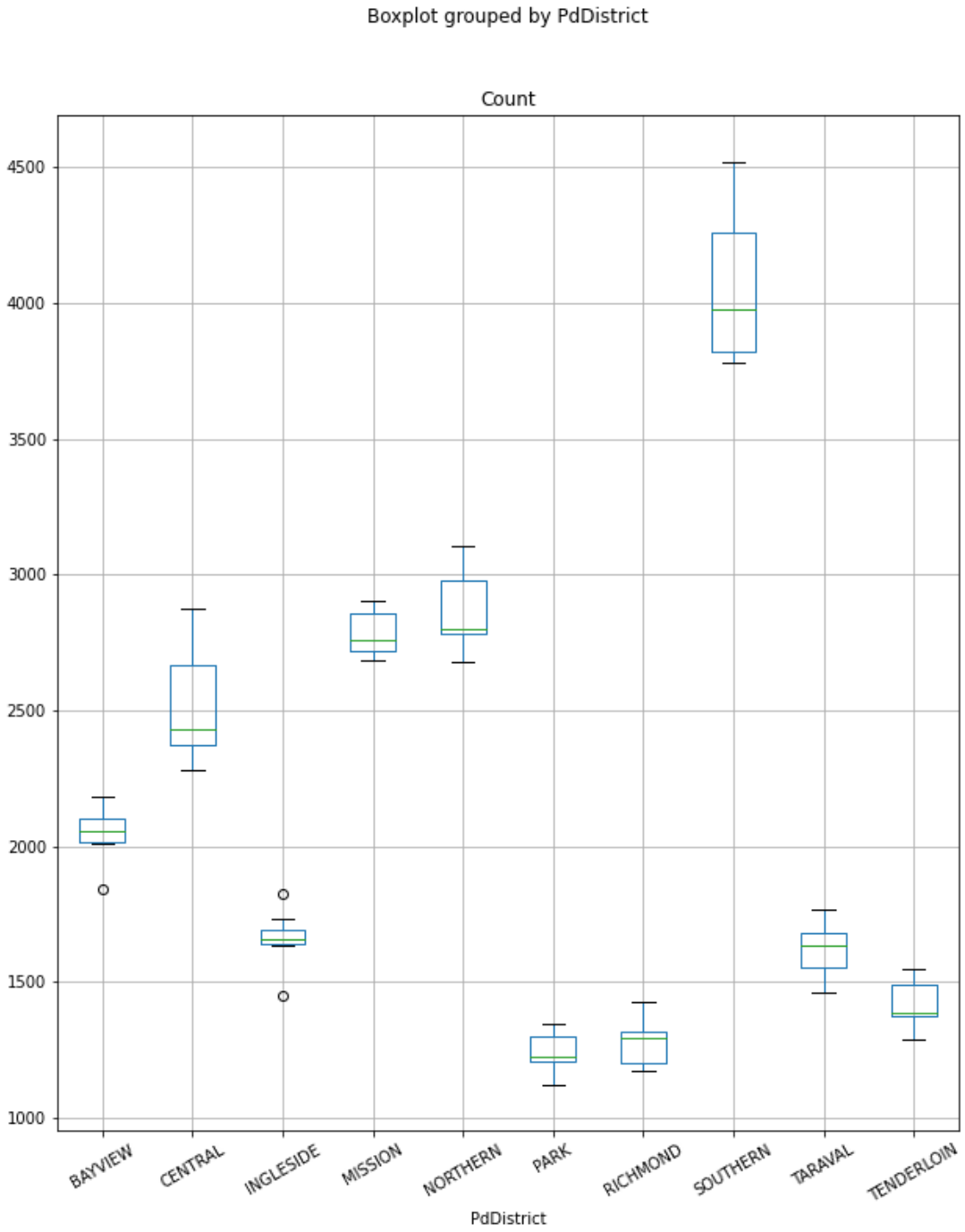

Figure 2. Boxplot of Crimes against 'PdDistrict' variable

c. San Francisco City data

I used the Geopy library to obtain the latitude and longitude values of San Francisco city. Using these location coordinates, I superimposed the map of the city setting the zoom level to 12. I added the markers to the various police districts and the locations of the crimes (taken from X, Y values in the crime dataset). These markers were plotted using Folium to obtain a visual represenatation of which police districts reported most crimes. Visualizing this data will help in understanding which parts of the San Francisco city recorded most crimes and will help in deciding which place to open a restaurant.

d. Foursquare API

The Foursquare API was used to obtain information about the various Indian restaurants in San Francisco city. By providing my client id and secret in the foursquare API, I was able to extract the Indian restaurants throughout the city using the relevant search criteria provided in my code. I was able to plot the locations of these Indian restaurants in the Folium map. This information will help understand existing popular locations for Indian restaurants with a greater concentration of restaurants nearby.

4. RESULTS

The initial exploratory analysis involved understanding the parts within San Francisco city with a greater number of crimes. This data was obtained from the SF crimes dataset. Figure 3 shows a line plot that shows the numbers of crimes under each police district. As can be seen, ‘Park’ police district has the least number of crimes whereas ‘Southern’ has the greatest number of crimes within the city. In addition, this statistic is also confirmed from the boxplot under police districts as shown in figure 2. So ‘Park’ police district could be a safe location with respect to crimes data to open the new Indian restaurant.

In addition, the boxplot shown in figure 1 does confirm that the most crimes are committed on Thursdays and Fridays, which are usually the days that are considered the start to the weekend. This may not be a great news for a prospective restaurateur like me, but for now I will opt to ignore this fact and continue with my search for the perfect location for the Indian restaurant.

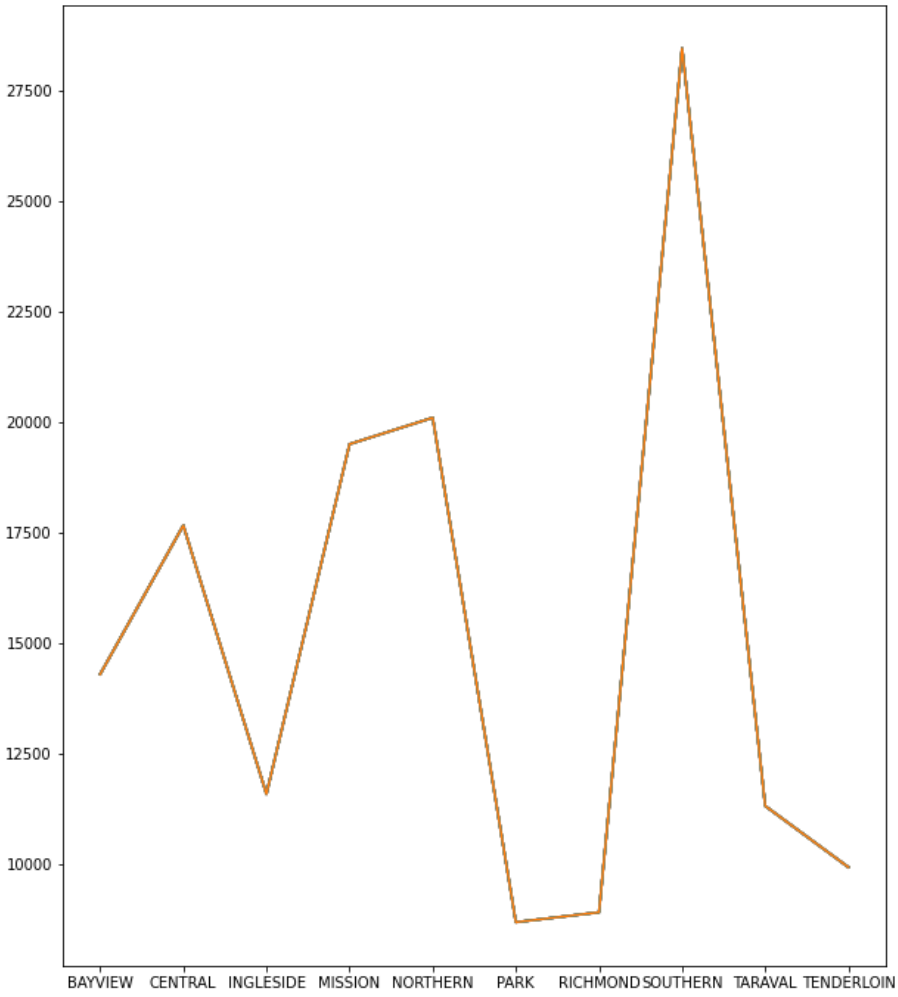

Figure 3. Crimes reported under police districts in SF City

From the initial 150500 rows, I chose to drop those rows for which 'PdDistrict' has a null value using the following command:

df_incident = df_incidents.dropna(subset = ['PdDistrict'], inplace = True)

Performing this command, I was able to remove one row with 'nan' value and the remaining 150499 rows were suitable for further analysis. Figure 4 shows the counts of all the features within the crimes dataset grouped with respect to the 'Resolution' feature. As can be seen, most

crimes were reported as ‘None’ and thus can yield little credibility to our analysis if considered. In addition, all the other resolution except for ‘Arrest, Booked’ are small values and thus I chose to ignore them as well. The only resolution considered in my analysis is ‘Arrest, Booked’ with 39416 rows.

```
In [8]:  1 df_incidents.groupby(['Resolution']).count()
```

Out[8]:

Resolution	IncidntNum	Category	Descript	DayOfWeek	Date	Time	PdDistrict	Address	X	Y	Location	PdId
ARREST, BOOKED	39416	39416	39416	39416	39416	39416	39416	39416	39416	39416	39416	39416
ARREST, CITED	144	144	144	144	144	144	144	144	144	144	144	144
CLEARED-CONTACT JUVENILE FOR MORE INFO	58	58	58	58	58	58	58	58	58	58	58	58
COMPLAINANT REFUSES TO PROSECUTE	2	2	2	2	2	2	2	2	2	2	2	2
EXCEPTIONAL CLEARANCE	371	371	371	371	371	371	371	371	371	371	371	371
JUVENILE BOOKED	1056	1056	1056	1056	1056	1056	1056	1056	1056	1056	1056	1056
JUVENILE CITED	3	3	3	3	3	3	3	3	3	3	3	3
JUVENILE DIVERTED	2	2	2	2	2	2	2	2	2	2	2	2
LOCATED	20	20	20	20	20	20	20	20	20	20	20	20
NONE	107780	107780	107780	107780	107780	107780	107779	107780	107780	107780	107780	107780
NOT PROSECUTED	22	22	22	22	22	22	22	22	22	22	22	22
PROSECUTED BY OUTSIDE AGENCY	1	1	1	1	1	1	1	1	1	1	1	1
PSYCHOPATHIC CASE	17	17	17	17	17	17	17	17	17	17	17	17
UNFOUNDED	1608	1608	1608	1608	1608	1608	1608	1608	1608	1608	1608	1608

Figure 4. Crimes data grouped by ‘Resolution’

Figure 5 below shows the plot of crimes in the various police districts plotted in Folium. The blue dots represent the location of the crimes for our data in consideration and the black dot represents the police districts. The latitude and longitude values for each police district was calculated by taking the mean of Y and X values respectively in the dataset using the following commands:

pd_loc = df_incidents.groupby('PdDistrict', as_index=False)['X'].mean()

pd_y = df_incidents.groupby('PdDistrict', as_index=False)['Y'].mean()

This resulted in two individual dataframes that were combined into one dataset for further calculations. ‘PdDistrict’ names have been assigned as labels to the folium map and the figure shows ‘Park’ police district since it has the lowest crimes for which arrests were made.

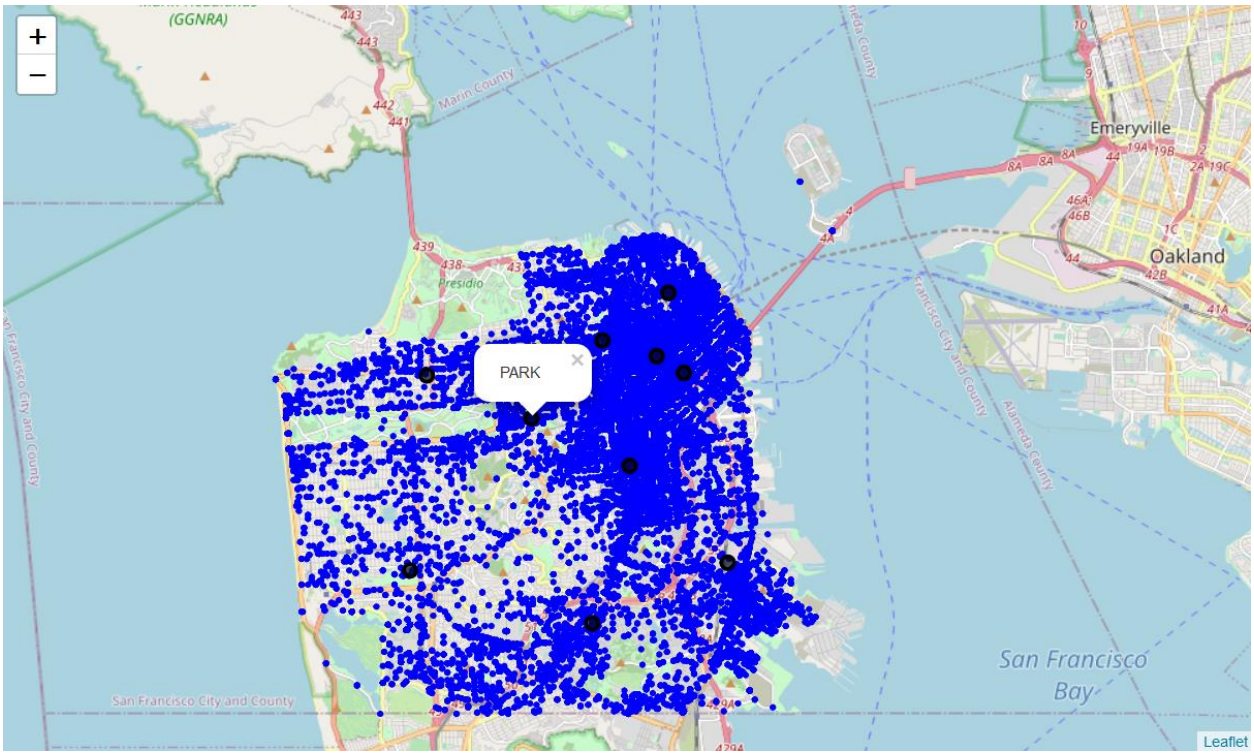

Figure 5. Crimes with respect to PoliceDistricts.

Next, I was able to get the locations of the Indian restaurants in San Francisco city using the Foursquare API with the code as given:

url=https://api.foursquare.com/v2/venues/search?&client_id={}&client_secret={}&v={}&query={}&ll={},{}&radius={}&limit={}'.format(…….) (please see code for complete information)

where …. signifies the values for all the required fields. This data is obtained in a .json format which I then transformed into Pandas dataframe. I further filtered this dataframe to obtain information on only the Indian restaurants within the city. Figure 6 shows these Indian restaurants that I was able to extract from this dataframe, these are indicated in red markers and the blue dot represents the lat, long of San Francisco city obtained earlier from the Geopy library.

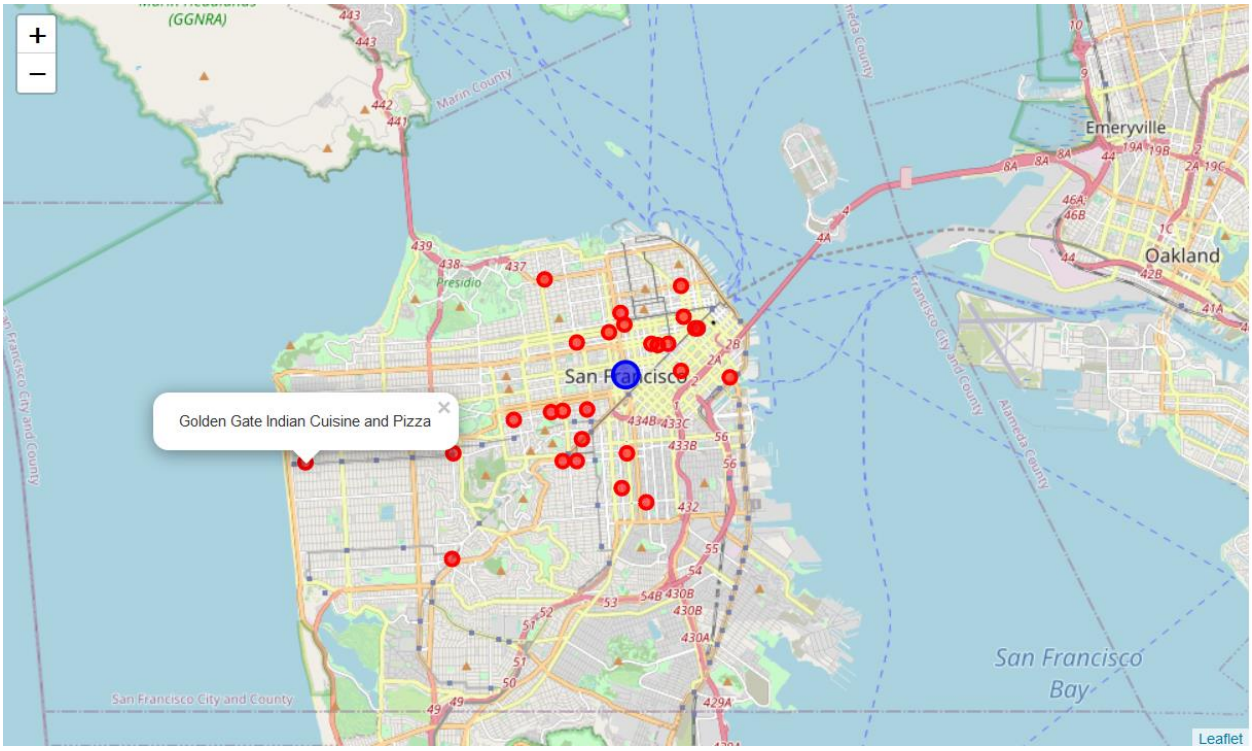

Figure 6. Indian restaurants in San Francisco city obtained from Foursquare API

Finally, I was able to plot all the markers as shown in figure 7 on the map as follows:

- The blue dots represent the points where the crimes were made
- The red dots represent the locations of Indian restaurants obtained from the Foursquare data
- The labels on the red dots indicate the names of the Indian restaurants

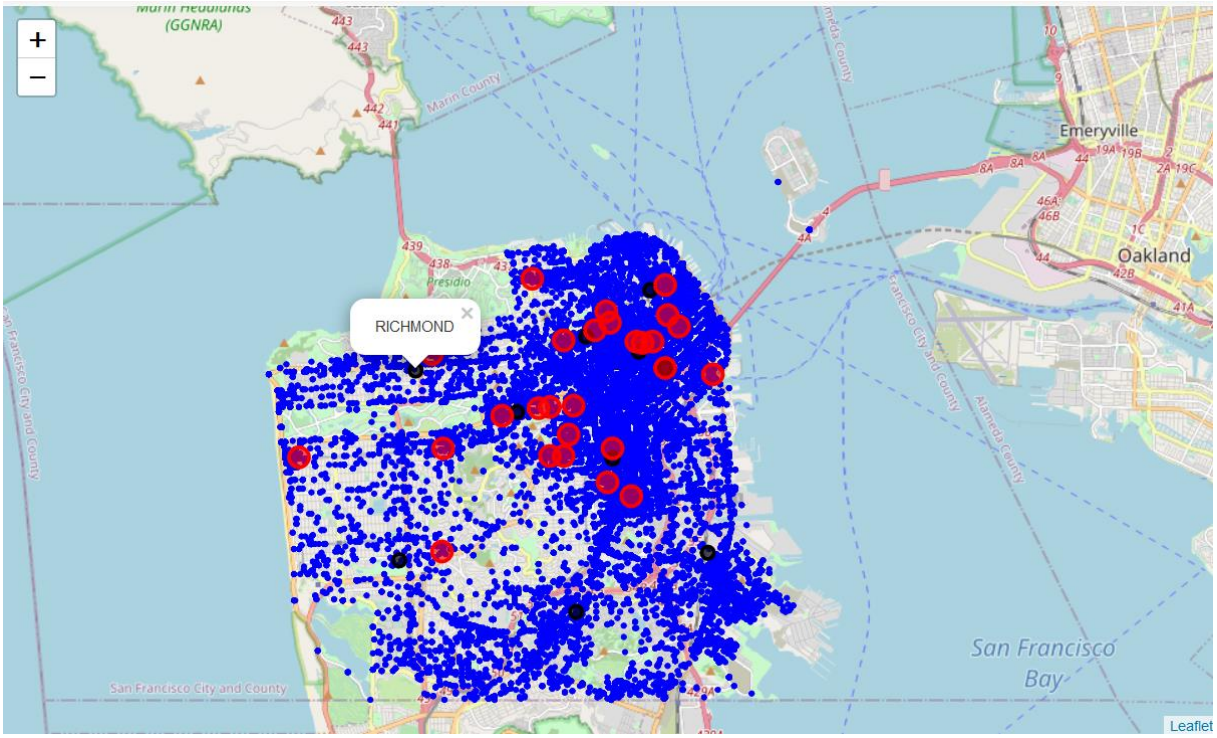

Figure 7. Overall mapping of data including crimes data and restaurants

5. DISCUSSION

Based on the data analysis and results visualization, I was able to visualize using Folium maps the crime rate per police division, and the number of restaurants that had a greater and lesser concentration of reported crimes. As one can see in figure 7, the Richmond police division has the fewest crimes reported per square unit of area. Additionally, there is just one Indian restaurant within this police division. Hence, the Richmond police district would be an ideal location to open the prospective Indian restaurant with respect to the crimes reported and arrests made within the city of San Francisco.

6. CONCLUSIONS

This analysis made use of a couple of data extraction, cleaning, and using it to obtain useful insights from these datasets using Python in a Jupyter notebook. The aspects analyzed are crime dataset obtained from Coursera's data visualization course and the Foursquare API to obtain locations of Indian restaurants within the San Francisco city. After performing a careful analysis and visualization of data, the Richmond police district area was found to be the best location to open an Indian restaurant for two prime reasons – lower crime rate per unit area and the presence of just one Indian restaurant in this police district. However, there could be multiple other factors one should look at before finalizing the location, such as, commercial rents in various areas of the city, availability of labor, etc. and these factors can be looked at in a future study.