RELATIONSHIP BETWEEN NUMBER OF POLICE INCIDENTS AND VENUES CATEGORY IN THE NEIGHBORHOODS OF SAN FRANCISCO

Fernando Luiz Trazzi Junior

February 05, 2021

1. INTRODUCTION

1.1. Background

The city of San Francisco is one of the most populous in the United States. Its surface covers an area of 121 km² at the north end of the San Francisco Peninsula, being formed by 21 neighborhoods and internationally recognized for its numerous hills. The city is the financial, cultural and transportation center of the San Francisco Bay area, being one of the richest cities in the country.

Throughout the Data Visualization course, we used some information from the city of San Francisco to generate a map and present police statistics on each of the city's 10 police districts in 2016. In addition, in the introduction to this Capstone Project, we learned to search for information about different places, such as stores, restaurants, parks, etc., of a neighborhood through the Foursquare API. From this data, we learn to generate clusters, each grouping similar neighborhoods.

Based on these two types of information, the problem described below is proposed.

1.2. Problem

Is it possible to associate the number of police incidents to neighborhoods with similar venues categories?

For example, a neighborhood that contains more bars, restaurants and stores, could have a greater number of police incidents than a neighborhood that is more residential, with more parks and leisure areas.

1.3. Interests

People who are choosing a place to live in San Francisco could make a tradeoff between neighborhoods and choose a neighborhood that has many locations such as shops, restaurants and bars, in other words, having more facilities nearby, but with more police occurrences, or a neighborhood more residential and with fewer police occurrences.

It is worth mentioning that this study can also be extended to other cities that have these public data at hand.

2. DATA

2.1. Data Sources

Data from a few different sources were used, described below:

- Dataset of police events in the city of San Francisco, divided among the various police districts, which occurred during the year 2016. This dataset was used in the Data Visualization course and can be downloaded at the following link: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DV0101EN-SkillsNetwork/Data%20Files/Police_Department_Incidents_-_Previous_Year__2016_.csv. From this dataset, a table was generated with the number of occurrences in each of the 10 police districts in the city.
- A geojson file from the city of San Francisco, obtained from the following link: https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DV0101EN-SkillsNetwork/labs/FinalModule_Coursera/data/san-francisco.geojson. In this file, are the outline of each of the city's police districts.
- Table with the neighborhoods and their zipcodes, obtained from the link: http://www.healthysf.org/bdi/outcomes/zipmap.htm. Using the pandas library, the website was scrapped, and the table was obtained. Using the uszipcode library it was possible to obtain the latitude and longitude of each neighborhood.
- To get the most common venues of given neighborhood of San Francisco was used the Foursquare API.

2.2. Data Cleaning and Feature Selection

The police incident dataset has more than 150000 lines and contains a lot of important information. The head of this data set is shown in the table below.

	IncidntNum	Category	Descript	DayOfWeek	Date	Time	PdDistrict	Resolution	Address	X	Y	Location	PdId
0	120058272	WEAPON LAWS	POSS OF PROHIBITED WEAPON	Friday	01/29/2016 12:00:00 AM	11:00	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421	(37.775420706711, -122.403404791479)	12005827212120
1	120058272	WEAPON LAWS	FIREARM, LOADED, IN VEHICLE, POSSESSION OR USE	Friday	01/29/2016 12:00:00 AM	11:00	SOUTHERN	ARREST, BOOKED	800 Block of BRYANT ST	-122.403405	37.775421	(37.775420706711, -122.403404791479)	12005827212168
2	141059263	WARRANTS	WARRANT ARREST	Monday	04/25/2016 12:00:00 AM	14:59	BAYVIEW	ARREST, BOOKED	KEITH ST / SHAFTER AV	-122.388856	37.729981	(37.7299809672996, -122.388856204292)	14105926363010
3	160013662	NON-CRIMINAL	LOST PROPERTY	Tuesday	01/05/2016 12:00:00 AM	23:50	TENDERLOIN	NONE	JONES ST / OFARRELL ST	-122.412971	37.785788	(37.7857883766888, -122.412970537591)	16001366271000
4	160002740	NON-CRIMINAL	LOST PROPERTY	Friday	01/01/2016 12:00:00 AM	00:30	MISSION	NONE	16TH ST / MISSION ST	-122.419672	37.765050	(37.7650501214668, -122.419671780296)	16000274071000

Table 1. Head of police incident dataset

However, the most important information in this project is the number of occurrences in each police district. After a transformation of the dataframe, this data was obtained, which is shown in the graph below.

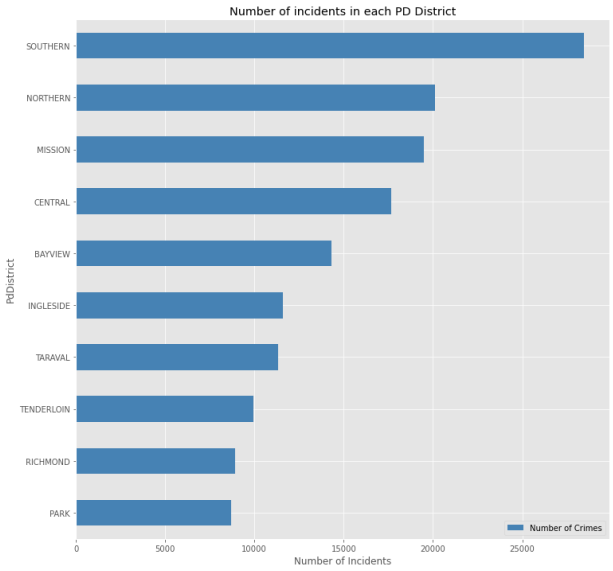

Figure 1. Number of police incidents in each PD District

As previously described, the San Francisco neighborhood dataset was obtained from a website, by selecting the table that contains the zip code for each neighborhood. Then, a library was used to obtain the latitude and longitude of each neighborhood. The result is shown in the table below:

	Zip Code	Neighborhood	Latitude	Longitude
1	94102	Hayes Valley/Tenderloin/North of Market	37.780	-122.420
2	94103	South of Market	37.780	-122.410
3	94107	Potrero Hill	37.770	-122.390
4	94108	Chinatown	37.791	-122.409
5	94109	Polk/Russian Hill (Nob Hill)	37.790	-122.420
6	94110	Inner Mission/Bernal Heights	37.750	-122.420
7	94112	Ingelside-Excelsior/Crocker-Amazon	37.720	-122.440
8	94114	Castro/Noe Valley	37.760	-122.440
9	94115	Western Addition/Japantown	37.790	-122.440
10	94116	Parkside/Forest Hill	37.740	-122.480
11	94117	Haight-Ashbury	37.770	-122.440
12	94118	Inner Richmond	37.780	-122.460
13	94121	Outer Richmond	37.800	-122.700
14	94122	Sunset	37.760	-122.480
15	94123	Marina	37.800	-122.440
16	94124	Bayview-Hunters Point	37.730	-122.380
17	94127	St. Francis Wood/Miraloma/West Portal	37.730	-122.460
18	94131	Twin Peaks-Glen Park	37.750	-122.440
19	94132	Lake Merced	37.720	-122.480
20	94133	North Beach/Chinatown	37.800	-122.440
21	94134	Visitacion Valley/Sunnydale	37.720	-122.410

Table 2. Neighborhoods, latitude and longitude

It should be stressed that PD districts are different from neighborhoods. There are a total of 10 PD Districts and 21 neighborhoods, meaning more than one neighborhood can belong to a police district. To search for venues data in each neighborhood, the Foursquare API was used, which will be discussed in greater detail in the next chapter.

3. METHODOLOGY

3.1. Neighborhood Analysis

The Foursquare API searches for all venues within a given radius from a latitude and longitude, with a response limited to 100 venues. In this project, venues were sought within a radius of 1000 meters from the latitude and longitude of each neighborhood in San Francisco, resulting in the bar graph below. Note that for some neighborhoods the limit of venue number is not reached.

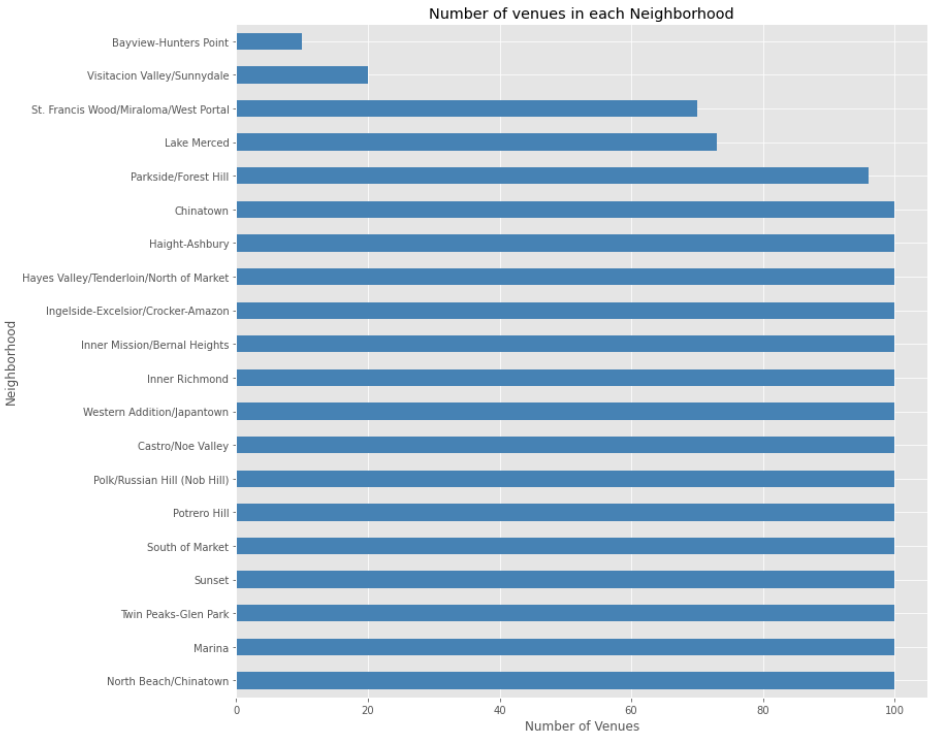

Figure 2. Number of venues in each Neighborhood

The next step was to create a dataframe using the one-hot-encoding technique related to the category of each venue found. Afterwards all venues were grouped by neighborhood and category, then the average of each category in each neighborhood was applied to normalize the analysis. Finally, a dataframe was created with the top 10 venue categories in each neighborhood, so it was possible to cluster the neighborhoods according to the venue categories.

3.2. Clustering

To cluster the neighborhoods, it was used the unsupervised machine learning K-means algorithm, one of the most common cluster methods. To not guessing the number of clusters, the elbow method was used to define the optimal number of clusters. First, we run K-means for many different amounts of clusters to calculate its inertia, or the within clusters sum of squares (WCSS). The method consists of plotting inertia as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

Using the dataframe with the top 10 venues categories by neighborhood, the following graph was generated

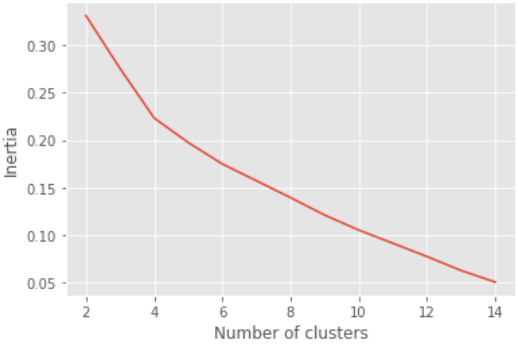

Figure 3. Elbow method.

As we can see, the inflection point, or elbow, happens when the number of clusters equals 4. This way, the neighborhoods was divided into 4 clusters according to their top 10 venue categories. So, it was run K-means to cluster the neighborhood into 4 clusters. K-means will then partition our neighborhoods into 4 groups, where the neighborhoods in each cluster are similar to each other in terms of the top 10 venues categories included in the dataset.

4. RESULTS

4.1 Examining Clusters

After running the K-means, the 4 neighborhood clusters were obtained in terms of the venue's categories.

Cluster 0 (red in the map below) consists of one neighborhood and the most common venue's categories are related to outside recreation.

Cluster 1 (purple) consists of 14 neighborhoods and the most common venue's categories are related to different types of bars, restaurants and general stores.

Cluster 2 (cyan) consists of 4 neighborhoods and this cluster is formed by different kinds of restaurants.

Cluster 3 (green) consists of 1 neighborhood and this cluster a mixed venue's categories, from park and gym to coffee shop and brewery.

4.2 Plotting the Map

To analyze the results, a map was plotted with the relationship of the all information obtained here in this project.

With the police incident dataset, the Folium library was used to plot a Choroplet map, thus highlighting the number of police incidents in each district.

On top of this map, colored circles were plotted indicating the cluster of each of the San Francisco neighborhoods.

The result of this is on the map below, where we can see the two pieces of information and gain insights into the relationship between the number of police incidents and the categories of venues in each neighborhood.

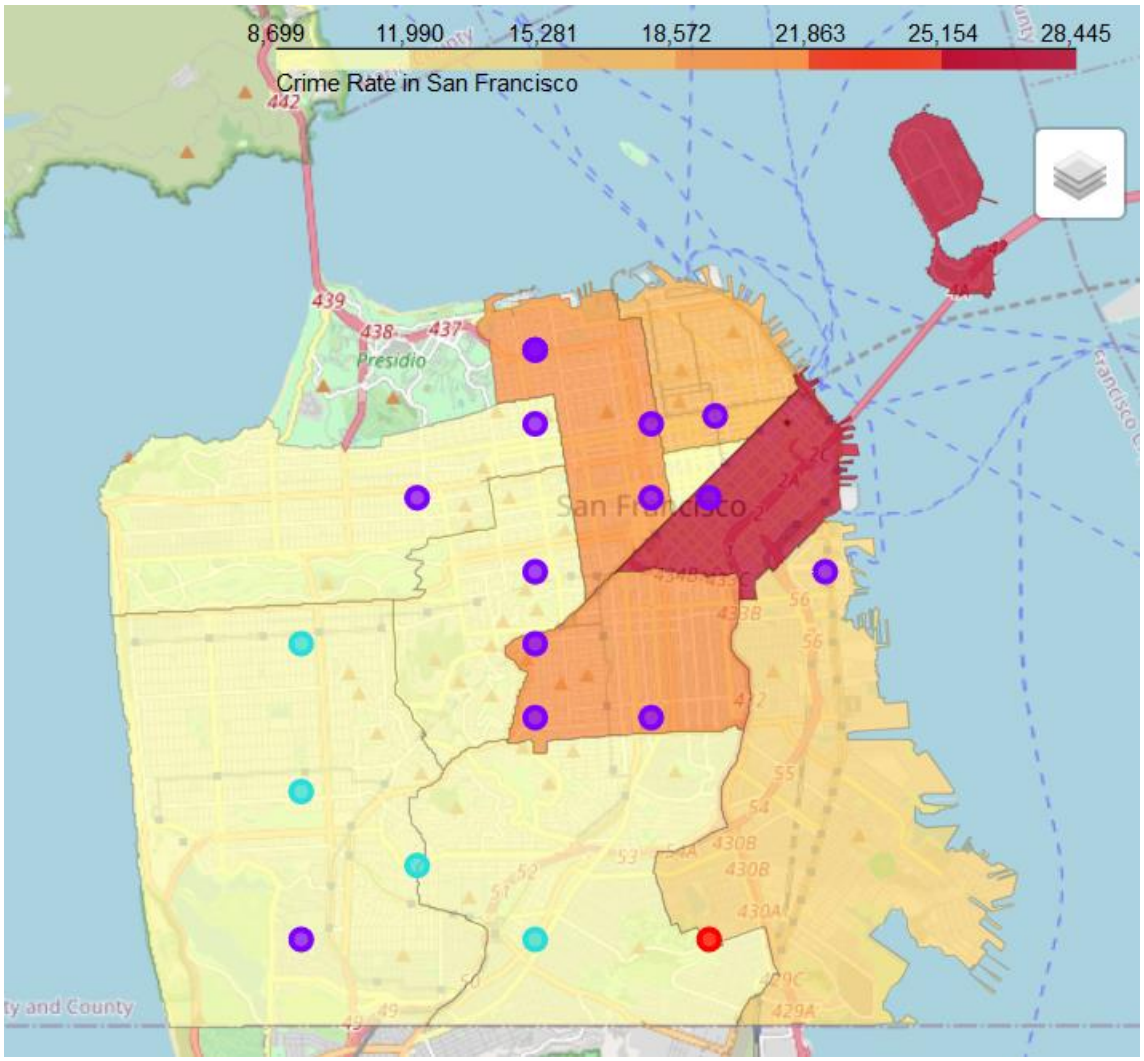

Figure 4. Map of San Francisco

5. DISCUSSION

As we can see on the map above, the police districts with the highest number of police incidents are located mainly in the neighborhoods of cluster 1 (purple), mostly composed of venue's categories of bars, restaurants and shops in general, that is, districts with greater number of people both day and night.

On the other hand, neighborhood clusters number 0 (red), 2 (cyan) and 3 (green), are located mainly in districts with fewer police incidents, perhaps because of the neighborhood's characteristic there is less movement of people both during both day and night.

Only from this map it is not possible to determine that there is a cause and effect between the type of neighborhood and the number of police incidents, however it indicates that there may be this type of relationship between the two variables.

6. CONCLUSION

There may be a relationship between the number of police incidents and the type of neighborhood in a neighborhood. However, this project is not intended to be a definitive analysis on this subject, it only shows a perception when these two datasets are compared on the same map.

Others can deepen this study in the city of San Francisco or even take this project to other cities where this data can be easily available.