Venues vs Crimes in San Francisco

The Battle of Neighborhoods - Coursera Capstone Project

Final Report

Author: Angelo Di Marco

1. Introduction

1.1. Background

San Francisco is one of the main cities of United States of America, located in the so-called San Francisco Bay in the state of California [1]. It is the cultural, commercial and financial center of Northern California and in 2019 it reached 881,549 inhabitants, distributed on a surface of "just" 121.4 square kilometers. The population of San Francisco is strongly multiethnic and heterogeneous. As a result this city is the meeting point of a large variety of cultures which contribute to its peculiar identity in different ways. In particular, this diversity accounts for the presence of many different kind of venues including, e.g., restaurants with cuisines from every part of the world. These venues are spread over the different neighborhoods in which the city can be divided and the category they belong to can depend on their location.

Besides all the interesting, beautiful and "delicious" venues that San Francisco can offer to its inhabitants and especially to tourists from all over the world, this city is also, unfortunately, the scene of many criminal events. Every year, indeed, several thousands of criminal incidents are reported by the Police such as, for instance, murders, rapes, robberies, assaults, and gangs fights. Every neighborhood of the city is affected by these unpleasant happenings.

1.2. Business Problem

In this context, understanding how kind of venues and type of incidents/crimes are spatially distributed in San Francisco and finding out how they are correlated may have a relevant impact in the economy and safety of this American city. This might be a key point for every entrepreneur who wants to start an economical activity in the city and for the Police Department in reaching a solution for the criminality problem.

In this Data Science project, we used machine learning techniques to determine (1) how the different venues and crimes are located in San Francisco neighborhoods and (2) what is the correlation between them. The ultimate questions that we addressed are: Which location is more convenient for a new business activity? Did a particular incident/crime happen in a certain location because there are certain types of venues around there? Given a crime with no "certain" location, is it possible to predict where it occurred based on the venues spatial distribution?

1.3. Data Description

To answer the questions listed above we used several different datasets. First of all we considered all the venues in San Francisco available at the locations platform called Foursquare [2]. By means of API calls to foursquare.com [3], we got access to datasets containing name, category, latitude and longitude of the venues within a certain radius from a specified location. We chose the latter

to be the “center” of one of the 37 neighborhoods in which the city can be divided [4]. Each neighborhood is listed together with the coordinates of its “center” in a dataset created by the author of this report with the help of Google Maps [5]. Finally the “Police Department Incident Reports” provided us the data about the criminal events reported in San Francisco from 2018 to present [6]. Specifically, this dataset contains information like the type of incident (category, subcategory and description), the name of the neighborhood where the crime happened as well as the exact point on the map (latitude and longitude).

The analysis of the above mentioned datasets by means of machine learning techniques allowed us to achieve the goals of this project. All the details of our study of the venues and of the incidents/crimes of San Francisco can be found in the two Jupyter Notebooks [7] and [8], respectively. In the following chapters, we will present and discuss the main results.

2. Exploring San Francisco Neighborhoods

The city of San Francisco can be divided into neighborhoods. Several divisions exist and they are used for different purposes. In this project, we will consider two of them: the "Planning Neighborhood Groups Map” [4] and the “Analysis Neighborhoods” [9].

2.1. "Planning Neighborhood Groups Map”

The so-called "Planning Neighborhood Groups Map", available at this link [4], consists of 37 neighborhoods namely Seacliff, Outer Richmond, Golden Gate Park, Outer Sunset, Parkside, Lakeshore, Presidio, Inner Richmond, Inner Sunset, West of Twin Peaks, Oceanview, Presidio Heights, Haight-Ashbury, Twin-Peaks, Outer Mission, Crocker Amazon, Marina, Pacific Heights, Western Addition, Castro-Upper Market, Noe Valley, Diamond Heights, Glen Park, Russian Hill, Nob Hill, Downtown-Tenderloin, Mission, Bernal Heights, Excelsior, Visitacion Valley, North Beach,

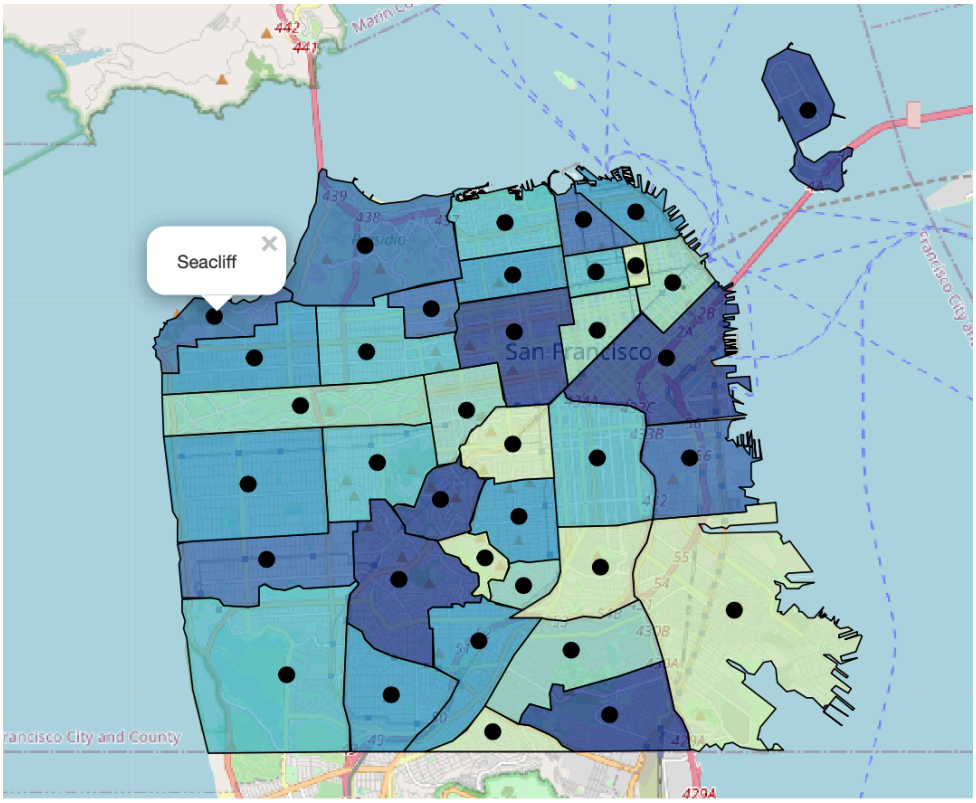

Figure 1 - Choropleth map of San Francisco showing the 37 neighborhoods of the "Planning Neighborhood Groups Map” division.

Chinatown, Financial District, South of Market, Potrero Hill, Bayview, Treasure Island. In Fig. 1, we show these neighborhoods on a Choropleth map. To obtain it we first imported this list of neighborhoods of San Francisco with the coordinates of their "centers" from the CSV file *SF_neighborhoods_latlon_ADM.csv* created by the author of this project. We used *pandas* library to convert it into a dataframe. Then, we loaded the GeoJson file containing the "borders" of the "Planning Neighborhood Groups Map" neighborhoods that can be downloaded at the link. After few adjustments to the names of the neighborhoods of the original GeoJson file, we got the new file *SF_planning_neighborhood_groups_map_ADM.geojson* and we imported it and used to create the Choropleth map in Fig.1. In the Jupyter Notebook [7], a click on the black spot on each neighborhood shows the corresponding name. Notice that the different colors are used only for a better visualization of the neighborhoods.

From Fig.1 we see that this division considers neighborhoods which have mostly a comparable size with each other and only few of them are maybe "too small". In addition, most of them have a "regular", "squared" shape. So, they offer a more even and regular division of the city. For these characteristics, in this project, we will focus on the "Planning Neighborhood Groups Map". We think that, among all the possible neighborhoods divisions, they are more suitable for the analysis of their most common venues obtained via Foursquare API calls.

2.2. "Analysis Neighborhoods”

Let us now have a look at the "Analysis Neighborhoods" division of San Francisco. It consists of 41 neighborhoods (see the link [9]) and are used, e.g., by the Police Department when reporting incidents/crimes occurring in the city. We show them in the Choropleth map of Fig.2. To create it we used the GeoJson file *SF_analysis_neighborhoods.geojson* containing the "borders" of the "Analysis Neighborhoods" available for download at the link (see the Jupyter Notebook [8] for further details). Also in this case, the different colors are used only for a better visualization of the neighborhoods.

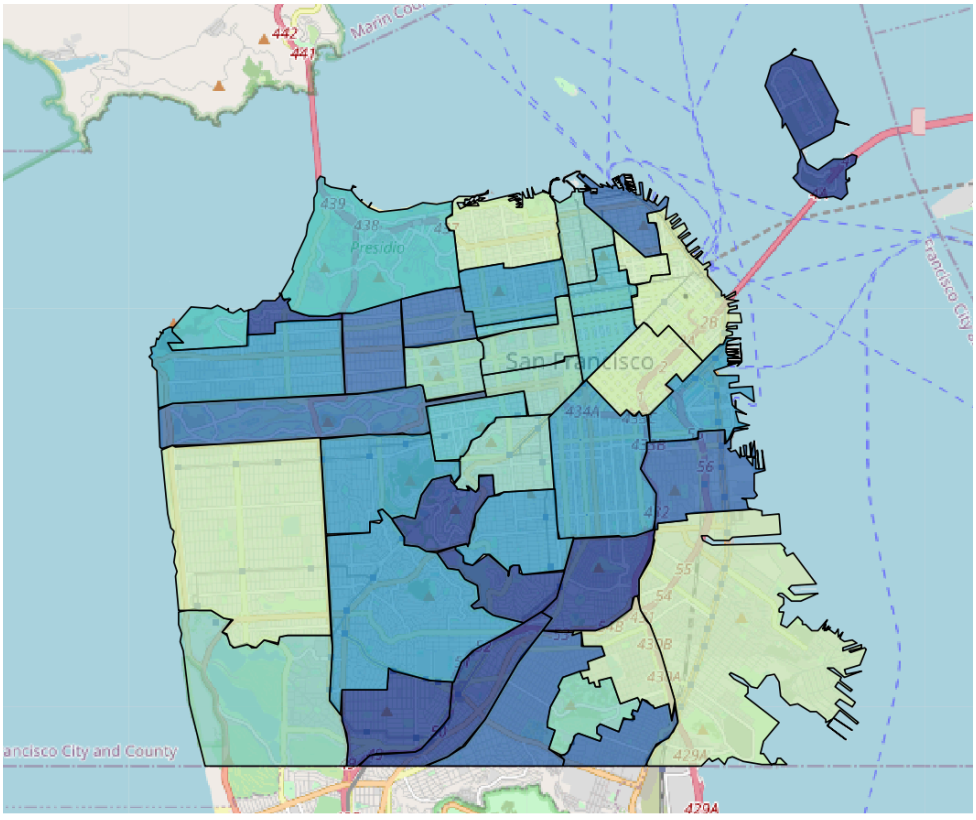

Figure 2 - Choropleth map of San Francisco showing the 41 neighborhoods of the "Analysis Neighborhoods” division.

We see that this type of division consists of both "large" and "very small" neighborhoods with shapes more irregular compared to the "Planning Neighborhood Groups Map" of Fig.1. As said, in the following, to have a more even and regular division of the city, we will focus on the neighborhoods of the "Planning Neighborhood Groups Map". The passage from one type of division to the other is discussed more in details in the Jupyter Notebook [8].

3. Venues in San Francisco: k-means Clustering

We start by examining the venues in San Francisco returned by Foursquare API calls and we use them to group the neighborhoods in clusters.

3.1. Foursquare API Calls

3.1.1. Neighborhoods as Circles

The maps of Fig.1 and Fig.2 gives a clear visual understanding of the borders and the areas of the different neighborhoods in San Francisco belonging to the "Planning Neighborhood Groups Map" and "Analysis Neighborhoods" divisions, respectively. However, to get the venues by means of the Foursquare API calls we will need the coordinates of the "center" of each neighborhood and the radius of the circle around this "center", rather than their shape. In other words, we have to represent the neighborhoods as circles. But how to choose center and radius of these circles in order to obtain the best result from Foursquare? As a first step, we identify San Francisco neighborhoods by means of their "centers" shown in the map below (see Fig.3) which follow the division given by the "Planning Neighborhood Groups Map". In the Jupyter Notebook [7], a popup with the name of the corresponding neighborhood appears by clicking on the blue markers.

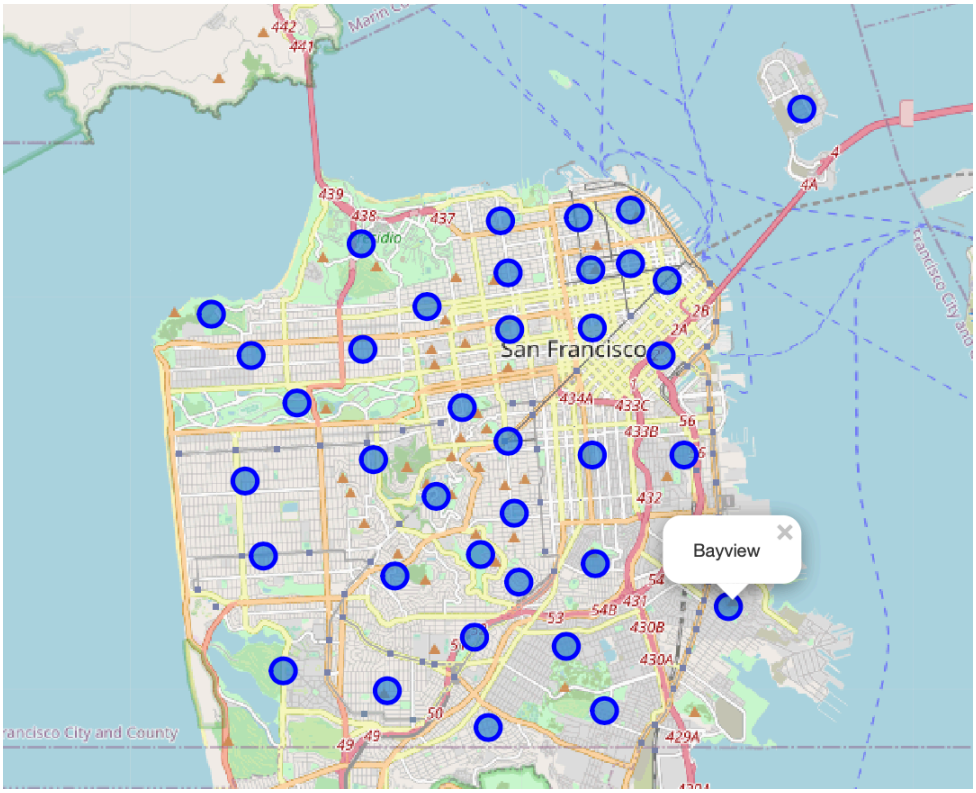

Figure 3 - Map of San Francisco. The blue circle markers are located at the "center" of each neighborhood of the "Planning Neighborhood Groups Map" division.

Second, the other important variable in the Foursquare API calls is the value (in meters) of the radius of the circle whose center is just the "center" of the neighborhood, in our case. To understand how this radius should be chosen in order to get an accurate clustering, we created several maps of San Francisco showing a circle around the location of each neighborhood with a specific radius in meters. We started by using 500 meters and we saw that a “small” portion of the neighborhoods was covered. Then we increased the value till 700 meters. In this case, the accessible area of the city is larger and most of the circles do not overlap. The main overlapping occurs in the city center namely the north-east part of San Francisco because there the “centers” are closer to each other (see Fig.3).

Finally, we increased further the value of the radius up to 1 kilometer, corresponding to 10 minutes walking distance from the "center" of a neighborhood till its circular border. The resulting circles are shown in Fig.4. In this case, the covered area of the city is much larger than the cases with 500 m and 700 m. Almost the entire surface of San Francisco can be explored in search for venues in this way. This holds especially for the “biggest” neighborhoods, as it can be noticed by comparing Fig.4 and Fig.1. On the other hand, there is also a larger overlap in many parts of the city, especially the north-east area. In this regard, the neighborhoods in the city center could be considered almost like a unique neighborhood when the radius is 1 km.

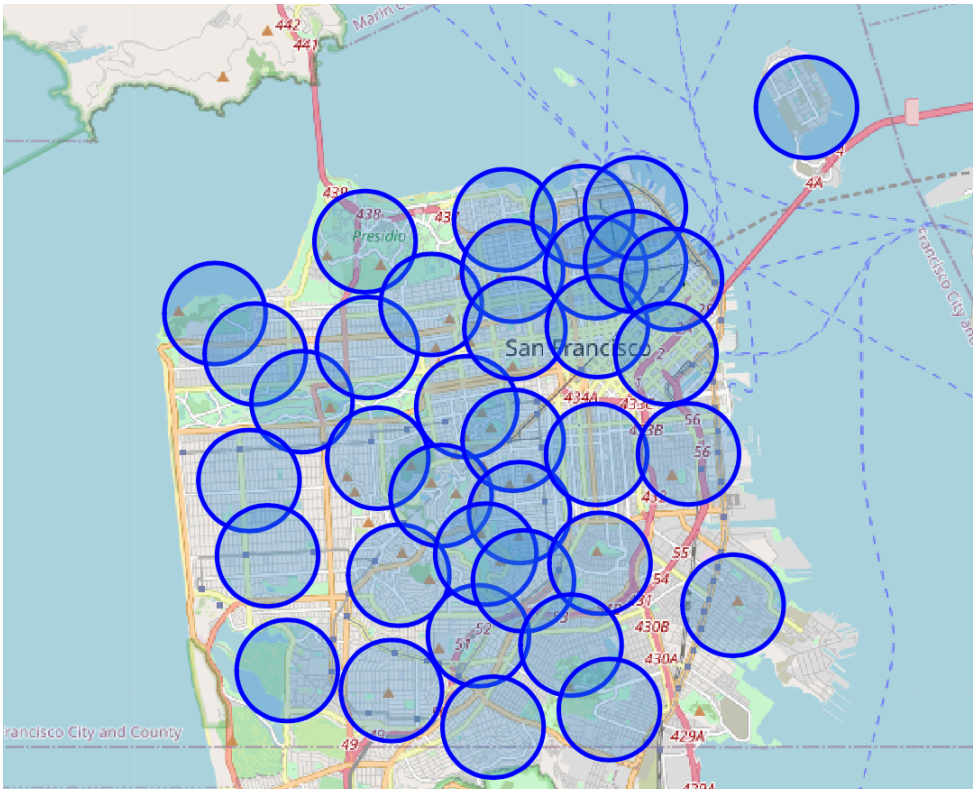

Figure 4 - Map of San Francisco. The blue circle markers are located at the “center” of each neighborhood of the "Planning Neighborhood Groups Map” division and their radius is 1 kilometer.

On the basis of these findings, we decided to use a radius of 1 km in making the Foursquare API calls. In this way, we will have access to a larger area of San Francisco and this radius guarantees that almost all the circles overlap a "bit" with the closest ones. We want this condition to be satisfied in order to take into account the fact that the venues at the border between two or more neighborhoods "belong" actually also to the closest neighborhoods around them, not only to the selected neighborhood.

Furthermore, using a radius of 1 km improves the venues search in such a way that the Foursquare API calls return more than 50 venues for most of the neighborhoods, making the comparison between them more even. On the contrary, when the radius is 500 m or 700 m, the

found venues are less than 10-20 for several neighborhoods and more than 50 venues are returned for several other neighborhoods.

Also, for a radius less than 1 km, we found that among the clusters obtained using the *k-means* method, there can be one cluster which contains most of the neighborhoods. In our analysis, this "issue" disappear by using 1 km.

3.1.2. Venues Collection

Once we decided center and radius of the circles representing the neighborhoods, we use them to build the URL to make Foursquare API calls with the "explore" option. To this end, in such URL, one needs to specify also the Foursquare CLIENT_ID and the CLIENT_SECRET. These parameters are provided by a developer account at Foursquare [3]. See the Jupyter Notebook [7] for more details.

In order to access the venues in each neighborhood, we sent then the GET request to Foursquare and we converted the result to a json variable. All the information about the venues is stored in the *items* key of this variable. Finally, we cleaned the json variable and structured it into a *pandas* dataframe to explore the venues in the neighborhoods. Figure 5 shows the first 5 rows of this dataframe.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Seacliff	37.784973	-122.501423	Legion of Honor	37.784571	-122.500670	Art Museum
1	Seacliff	37.784973	-122.501423	Lincoln Park	37.785436	-122.502022	Park
2	Seacliff	37.784973	-122.501423	Mile Rock Beach	37.787268	-122.506294	Beach
3	Seacliff	37.784973	-122.501423	Lands End Labyrinth	37.787934	-122.505813	Public Art
4	Seacliff	37.784973	-122.501423	Eagles Point	37.786714	-122.494810	Scenic Lookout

Figure 5 - First 5 rows of the dataframe containing the venues returned by the Foursquare API calls.

We got in total 2971 venues divided in 326 unique categories. These venues are not evenly distributed among the neighborhoods. In this regard, we noticed that the maximum number of venues returned by the API calls per neighborhood is 100 even if we set it to be 150. This should be related to the type of developer account used. An upgrade of it would return a higher number and it might improve the clustering of the neighborhoods presented in the following sections.

3.1.3. Venues Dataframe Pre-Processing

Before applying the clustering method to group the neighborhoods in terms of their most common venues, we performed some operations on the dataframe of Fig.5. First of all, we focused only on the attributes "Neighborhood", "Venue" and "Venue Category". We applied then the so-called "one hot encoding" to each venue obtained via Foursquare. This procedure consists on assigning 1 to the category of the considered venue and 0 to all the other possible categories. We created the corresponding "one hot" dataframe. Next, we grouped the rows of such dataframe by neighborhood and we calculated the mean of the frequency of occurrence of each category. The resulting dataframe, *neighborhoods_SF_grouped,* is the one that we used for the clustering of the neighborhoods, as explained in the next section.

3.2. Clustering Venues

At this point we proceed by grouping the neighborhoods of San Francisco into clusters based on the most common venues. Clustering is an unsupervised technique in machine learning used to assign each entry of a dataset to a cluster or, in other words, to attach a label to each element of

an unlabelled set. Specifically, the aim of the clustering strategy is that similar data points will end up into the same cluster. In our case, similar neighborhoods will belong to the same group.

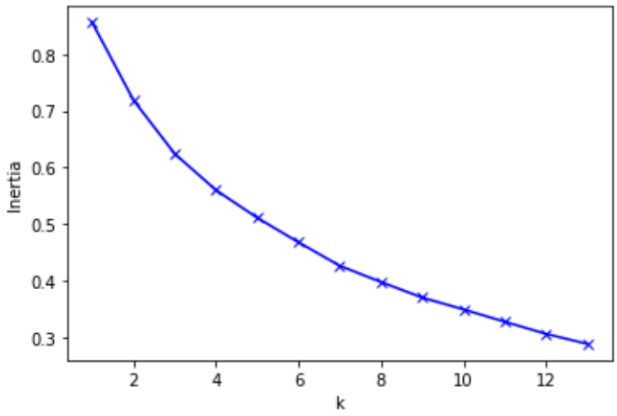

Figure 6 - Plot of the Inertia as a function of the number of clusters *k*.

Among the different clustering methods, in this project, we chose to use the *k-means* clustering technique which is one of the most popular in machine learning. The key feature of the *k-means* clustering is that the number of clusters into which the data points should be grouped into has to be decided before the clustering is executed. How to choose this number? One possibility is to use the elbow method to decide. In this regard, we consider the so-called *Inertia* namely the sum of squared distances of the data points from their cluster's center for increasing numbers of clusters and see if we can find a clear cluster number where the decrease in distortion starts to

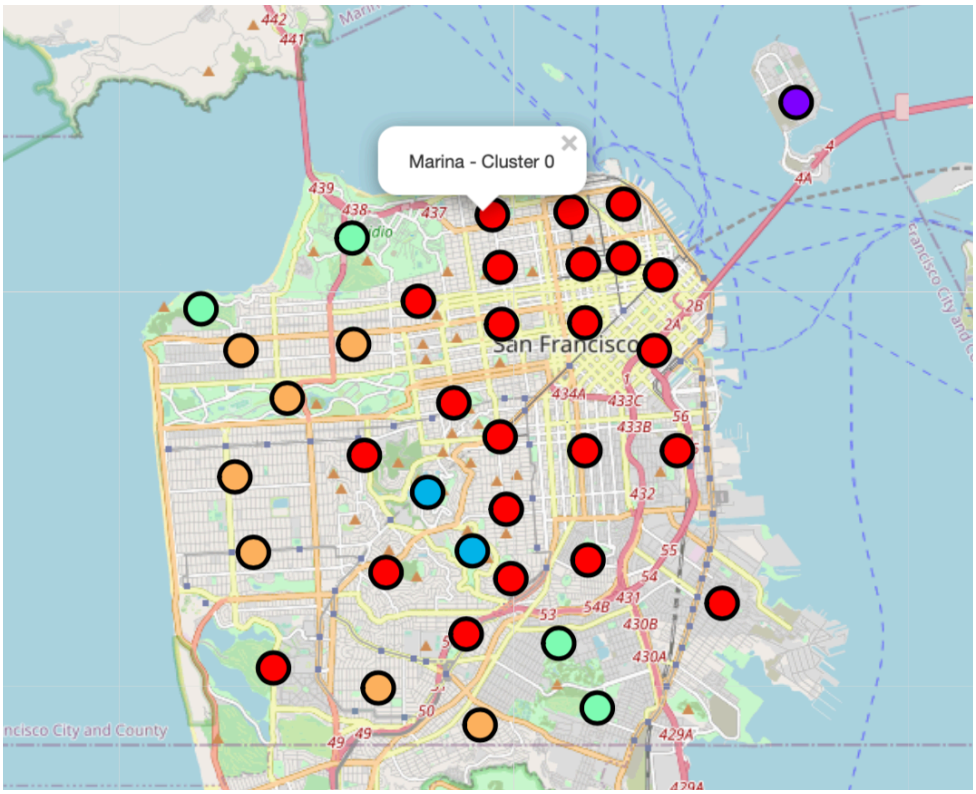

Figure 7 - Map of San Francisco showing the neighborhoods grouped in clusters. The circle markers are located at the “center” of each neighborhood. Each cluster has a different color: red (Cluster 0); orange (Cluster 1); light green (Cluster 2); blue (Cluster 3); violet (Cluster 4).

level off. We plot the Inertia in Fig.6 as a function of the number of clusters *k*. We see that the optimal number of clusters is 3, 4 or 5. After trying with different values, even larger than 5, we decided to choose 5. The resulting 5 clusters are shown in Fig.7 with different colors. In the Jupyter Notebook [7], a click on the circle markers on the map of Fig.7 shows a popup with the name of the neighborhood and the cluster number. The latter goes from 0 to 4.

From Fig.7, we notice that Cluster 0 (red), has more than half of the neighborhoods namely 23 out of 37. On the other hand, the smallest cluster, Cluster 4 (violet), contains only the neighborhood "Treasure Island", an outlier. To analyze more in details the content of each cluster we considered a dataframe per each cluster containing its neighborhoods together with the top 10 most common venues. An example is given in Fig.8, for Cluster 1. All the other dataframes can be found in the Jupyter Notebook [7]. A carful qualitative look at the venues listed in these dataframes suggests that Cluster 0 (red), the largest, is rather miscellaneous, the dominant venue categories seem "Coffee shop" and "Café". On the other hand, Cluster 1 (orange) contains mainly those neighborhoods where it is likely to find a Chinese Restaurant. Cluster 2 (light green) and Cluster 3 (blue) are similar and characterized by trails and parks. Finally, Cluster 4 (violet) namely "Treasure Island" has venues mostly in the "Food Truck" category as well as venues related to sport and music events.

	Neighborhood	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
1	Outer Richmond	1	Chinese Restaurant	Café	Bakery	Vietnamese Restaurant	Playground	Korean Restaurant	Sushi Restaurant	Deli / Bodega	Japanese Restaurant	Burrito Place
2	Golden Gate Park	1	Park	Vietnamese Restaurant	Bubble Tea Shop	Chinese Restaurant	Bakery	Deli / Bodega	Dumpling Restaurant	Garden	Coffee Shop	Playground
3	Outer Sunset	1	Chinese Restaurant	Liquor Store	Coffee Shop	Japanese Restaurant	Light Rail Station	Gym / Fitness Center	Café	Cosmetics Shop	Pet Store	Dessert Shop
4	Parkside	1	Chinese Restaurant	Dumpling Restaurant	Park	Sandwich Place	Pharmacy	Supermarket	Thai Restaurant	Bubble Tea Shop	Baseball Field	Bar
7	Inner Richmond	1	Bakery	Chinese Restaurant	Japanese Restaurant	Garden	Café	Korean Restaurant	Thai Restaurant	Burmese Restaurant	Vietnamese Restaurant	Asian Restaurant
10	Oceanview	1	Yoga Studio	Chinese Restaurant	Asian Restaurant	Pizza Place	Japanese Restaurant	Coffee Shop	Grocery Store	Poke Place	Convenience Store	Light Rail Station
15	Crocker Amazon	1	Pizza Place	Liquor Store	Mexican Restaurant	Baseball Field	Chinese Restaurant	Playground	Vietnamese Restaurant	Latin American Restaurant	Pool Hall	Soccer Field

Figure 8 - Dataframe for Cluster 1 containing its 7 neighborhoods and the top 10 most common venues.

In order to have a better understanding of the most common venues in each cluster in a more quantitative way, we examined also the total number of venues grouped per category in each cluster. Figure 9 shows the corresponding bar plot for Cluster 1. By combining the information from the dataframes and bar plots like the ones in Figs.8 and 9 respectively, we labelled the 5 clusters as follows:

- Cluster 0 - Coffee places (miscellaneous)
- Cluster 1 - Chinese restaurants and Asian cuisine
- Cluster 2 - Trails, parks, and café
- Cluster 3 - Trails and parks
- Cluster 4 - Food trucks, sport and music events places

We notice first of all that Cluster 2 and Cluster 3 could be considered as one cluster whose label would be "Trails and Parks". Further analysis is required to understand the reason why the *k-means* method differentiate them, but this is out of the scope of this project. On the other hand, since Cluster 0 is large and rather miscellaneous (even if it contains mainly coffee places), we decided to cluster it separately. In this way, we can achieve a better understanding of the venues that it contains. This sub-clustering is described in the next section.

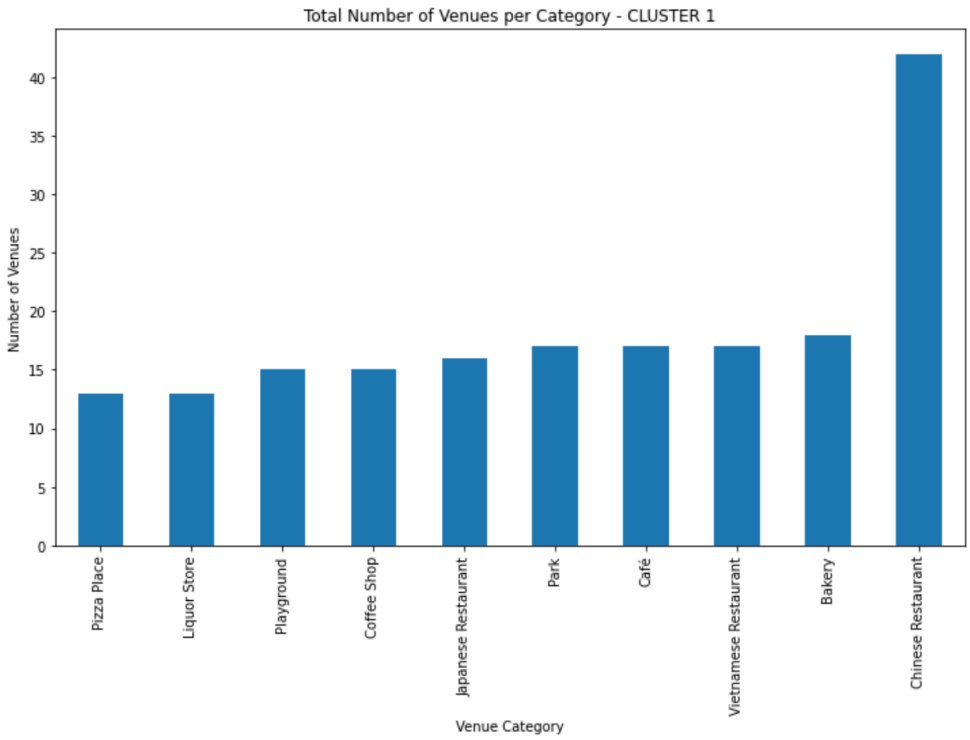

Figure 9 - Bar plot of the total number of venues per venue category for Cluster 1.

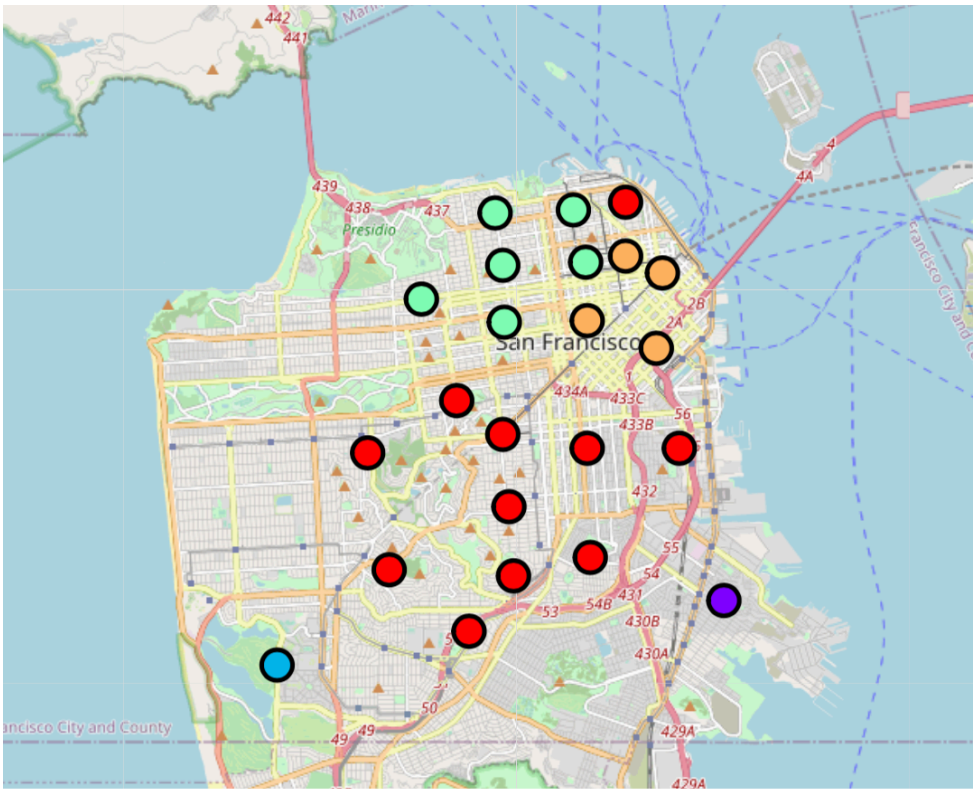

Figure 10 - Map of San Francisco showing the neighborhoods of Cluster 0 grouped in sub-clusters. The circle markers are located at the “center” of each neighborhood. Each sub-cluster has a different color: red (Sub-Cluster 0); orange (Sub-Cluster 1); light green (Sub-Cluster 2); blue (Sub-Cluster 3); violet (Sub-Cluster 4).

3.3. Sub-Clustering the “Large” Cluster

As we found and discussed in the previous section, Cluster 0 contains most of the neighborhoods of San Francisco. This "problem" is not solved by simply increasing the number of clusters. We decided then to further apply the *k-means* clustering method only to this "large" cluster. To obtain the sub-clusters we proceeded as described in the previous section. By means of the elbow method, we found that the optimal number of sub-clusters is 3, 4 or 5. After trying with different values, even larger than 5, we decided to use 5. The resulting sub-clusters are showed in the map of San Francisco in Fig.10. See the Jupyter Notebook [7] for further details. We see that the division is not even, one of the sub-clusters has half of the neighborhoods of the large cluster (Cluster 0). In addition, two small sub-clusters with just 1 element are created. These two outliers correspond to the neighborhoods "Lakeshore" and "Bayview". Notice also that the neighborhoods of the city center (north-east part of San Francisco) are grouped into two sub-clusters.

To determine the discriminating venue categories that distinguish each cluster, we considered first the top 10 most common venues. From this first analysis, we found that Sub-Cluster 0 is the largest and it is rather miscellaneous. The dominant categories seem to be "Café", "Park", and "Coffee shop". On the other hand, coffee shops are the dominant venues in the Sub-Cluster 1. Sub-Cluster 2 is also miscellaneous with many Italian restaurants. Finally, Sub-Cluster 3 and Sub-Cluster 4 have only the neighborhoods "Lakeshore" and "Bayview", respectively, as said. Both these sub-clusters are characterized by venues with mixed categories. We see that they do not have that much in common with the other clusters. In other words, it seems fine that they are two distinct sub-clusters. A more quantitative analysis by means of the total number of venues grouped per category in each cluster allowed us to label the 5 sub-clusters as follows:

- Sub-Cluster 0 - Coffee places (Italian and Mexican cuisine, parks)
- Sub-Cluster 1 - Coffee places (fitness, hotels)
- Sub-Cluster 2 - Italian cuisine (coffee places, parks)
- Sub-Cluster 3 - Miscellaneous (“Lakeshore”)
- Sub-Cluster 4 - Miscellaneous (“Bayview”)

Here Sub-Cluster 0 has more elements than the others and its neighborhoods contains mainly coffee shops but also other kind of venues are commonly reachable. A better understanding could be achieved by sub-clustering Sub-Cluster 0, but this is out of the scope of this project. Special attention is required by Sub-Cluster 3 and Sub-Cluster 4 which have only the neighborhoods "Lakeshore" and "Bayview", respectively. In fact, they contain miscellaneous venues and are clearly distinct from the other sub-clusters. On the other hand, Sub-Cluster 1 an 2 have well defined venue categories, “Coffee places” and “Italian cuisine”, respectively.

3.4. Merging Clusters and Sub-Clusters - Conclusions

In this chapter, we analyzed the neighborhoods of San Francisco in terms of the most common venues. We used the Foursquare API calls to get a list of the venues of each neighborhood, represented as a circle with radius of 1 km (see Fig.4). We then applied the *k-means* clustering method to group the neighborhoods into clusters and sub-clusters depending on the venues categories. Our findings are summarized in the map of Fig.11. We started by executing the clustering on all the 37 neighborhoods of the “Planning Neighborhood Groups Map” (see Fig.1). We found 5 clusters and we labelled them by considering the most common venues as follows:

- Cluster 0 - Coffee places (miscellaneous)
- Cluster 1 - Chinese restaurants and Asian cuisine
- Cluster 2 - Trails, parks, and café
- Cluster 3 - Trails and parks
- Cluster 4 - Food trucks, sport and music events places

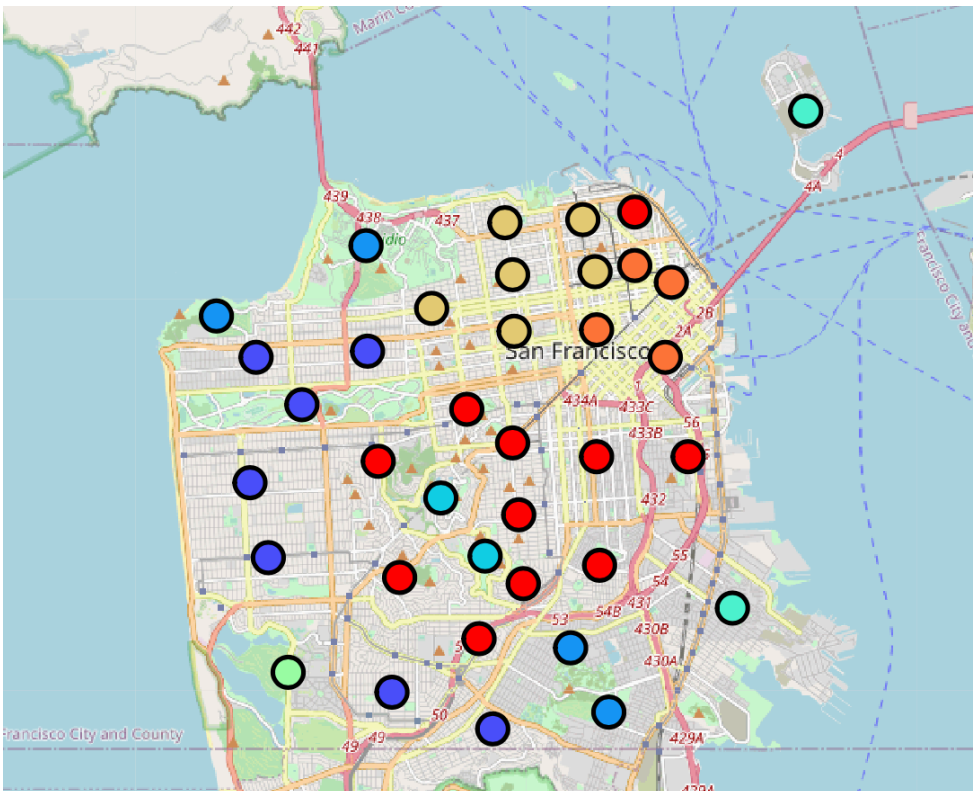

Figure 11 - Map of San Francisco showing the neighborhoods grouped in clusters. The circle markers are located at the “center” of each neighborhood. Each cluster and each sub-cluster has a different color: red (Sub-Cluster 0); orange (Sub-Cluster 1); yellow (Sub-Cluster 2); light green (Sub-Cluster 3); light blue-green (Sub-Cluster 4); violet (Cluster 1); blue (Cluster 2); light blue (Cluster 3); light blue-green (Cluster 4).

Here, Cluster 1 and Cluster 4 have a well defined label, whereas Cluster 2 and Cluster 3 could be actually one unique cluster. In addition, we found that Cluster 0 is over-represented and it has venues categories which are rather miscellaneous, mainly belonging to the "Coffee Shop" category. We decided then to cluster it separately and we obtained the following sub-clusters:

- Sub-Cluster 0 - Coffee places (Italian and Mexican cuisine, parks)
- Sub-Cluster 1 - Coffee places (fitness, hotels)
- Sub-Cluster 2 - Italian cuisine (coffee places, parks)
- Sub-Cluster 3 - Miscellaneous (“Lakeshore”)
- Sub-Cluster 4 - Miscellaneous (“Bayview”)

Here, Sub-Cluster 0 has half of the neighborhoods of the sub-clustered ones and it could be further clustered separately, but this is out of the scope of this project. On the other hand, Sub-Cluster 1 defines the neighborhoods in the city center where the presence of coffee places is dominant and where fitness centers and hotels are also common. Sub-Cluster 3 and Sub-Cluster 4 have a special variety of venues. They contain just 1 element namely the neighborhoods "Lakeshore" and "Bayview", respectively. The fact that these two big neighborhoods stand out from the others might be related to their peculiar geographical characteristics. In fact, "Lakeview" is identified by Lake Merced and the park around it whereas "Bayview" is represented mostly by the presence of the Hunters Point Naval Shipyard.

4. San Francisco Incidents/Crimes: k-means Clustering

After having clustered the neighborhoods of San Francisco using the most common venues, in this chapter, we will analyze and cluster the neighborhoods in terms of the incidents/crimes occurring in the city as reported by the Police Department.

4.1. Police Department Report Dataset

4.1.1. Importing, Exploring and Pre-Processing

Every year the Police Department of San Francisco reports thousands of incidents/crimes occurring in the city. All these events are registered in datasets that can be dowloaded for free at the link [10]. For this project, we focused on the incidents which took place from 2018 till present. We imported the corresponding dataset available at the link [6] and we stored it in the dataframe *incidents_SF_full*. This report contains 364275 rows (the incidents) and 36 attributes. Among the latter, the most relevant are: “Incident Datetime” and “Incident ID” which uniquely identify every incident; “Incident Category”, ”Incident Subcategory” and ”Incident Description” which categorize and describe each incident; “Analysis Neighborhood”, “Latitude” and “Longitude” which give the location of the (criminal) happening. In this regard, we see that, besides the geographical coordinates, the incidents/crimes are also located in terms of the neighborhoods where they occurred which correspond to the ones belonging to the "Analysis Neighborhoods" division described in Chapter 2 [9]. Figure 12 shows a slice of the first 5 rows of the *incidents_SF_full* dataframe.

	Incident Category	Incident Subcategory	Incident Description	Analysis Neighborhood	Latitude	Longitude
0	Offences Against The Family And Children	Other	Domestic Violence (secondary only)	Sunset/Parkside	37.762569	-122.499627
1	Non-Criminal	Other	Mental Health Detention	South of Market	37.780535	-122.408161
2	Missing Person	Missing Person	Found Person	Bayview Hunters Point	37.721600	-122.390745
3	Offences Against The Family And Children	Family Offenses	Elder Adult or Dependent Abuse (not Embezzleme...	Chinatown	37.794860	-122.404876
4	Assault	Simple Assault	Battery	Marina	37.797716	-122.430559

Figure 12 - First 5 rows of a slice of the *incidents_SF_full* dataframe containing the incidents/crimes reported by the Police Department of San Francisco.

Since we are interested in the type of crimes occurring in a certain neighborhood in order to cluster San Francisco neighborhoods, from the *incidents_SF_full* dataframe, we selected the attributes “Incident Category” and "Analysis Neighborhoods” and we created a new dataframe, *incidents_SF_relevant*. By counting and listing the unique values of both "Incident Category" and "Analysis Neighborhood” we found that there are several entries with a NaN value namely missing data. Thus, we cleaned the dataset by removing all the rows containing NaN values. As a result, we found that the number of unique neighborhoods and incident categories in the dataframe are 41, as expected, and 51, respectively. In addition, the number of registered events stored in *incidents_SF_relevant* decreased from 364275 to 344986.

4.1.2. From "Analysis Neighborhoods" to "Planning Neighborhood Groups Map"

As we discussed above, the report of the Police of San Francisco used in this project contains also the names of the neighborhoods where the incidents occurred. These neighborhoods are 41 and they belong to the so-called "Analysis Neighborhoods” (AN) list which is one of the ways the city can be divided (see Chapter 2). However, as said, we are interested in the "Planning Neighborhood Groups Map” (PN) division where the neighborhoods are 37 and the San Francisco area is more "evenly" divided. In the Jupyter Notebook [8], we described how we passed from the first division, AN, to the second one, PN. In doing this operation, we mainly compared the maps

with the two type of neighborhoods divisions showed in Fig.1 and Fig.2. Most of the AN are approximately the same as the PN and we just renamed them to match the PN ones. On the other hand, 6 neighborhoods of the PN are the result of merging of 2 or 3 of the AN. The merging operation is just a neighborhoods “renaming”. In this way, we got 34 neighborhoods instead of 37. This is because the neighborhoods "Parkside", "Diamond Heights" and "Crocker Amazon" are missing. The reason is that their data are included into the neighborhoods "Outer Sunset", "Glen Park" and "Excelsior", respectively, of the AN division. We included them also by copying the rows for "Outer Sunset", "Glen Park" and "Excelsior" and renaming them into "Parkside", "Diamond Heights" and "Crocker Amazon".

4.2. Incidents/Crimes Neighborhoods Distribution

As a first step in analyzing the incidents reported by the Police Department of San Francisco, we had a closer look at how they are distributed among the neighborhoods. By grouping and counting the incidents per neighborhood we got the bar plot in Fig.13 which shows the total number of incidents occurred from 2018 till present in each neighborhood. We see that the

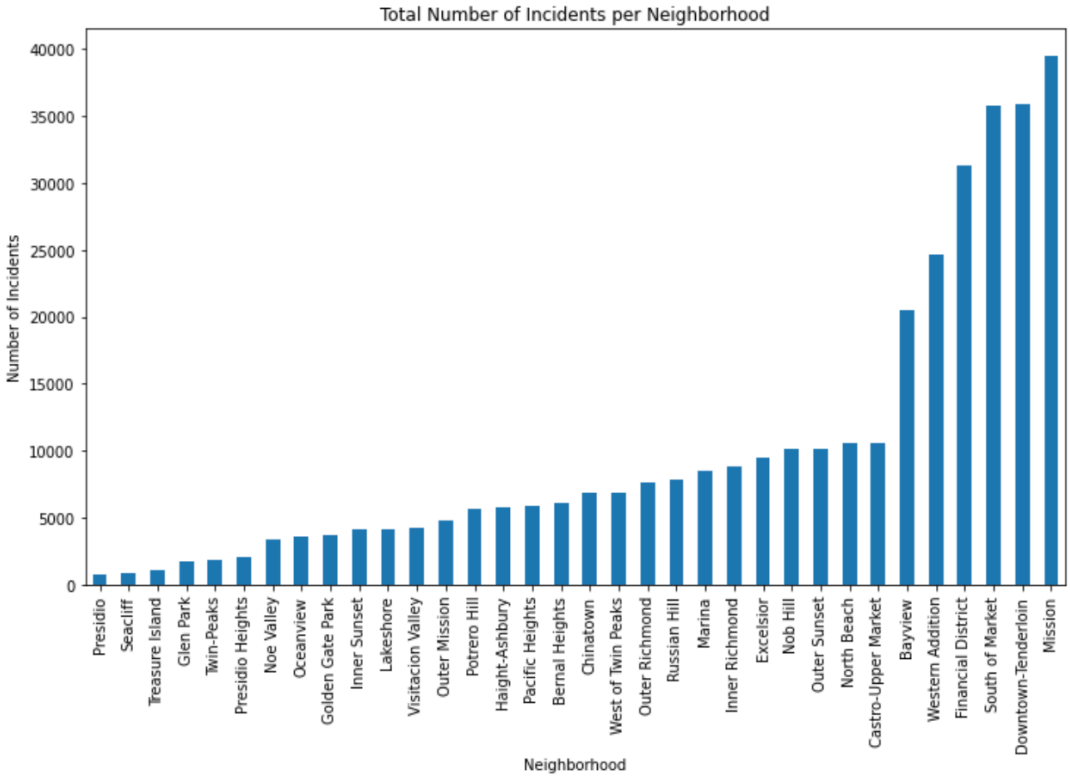

Figure 13 - Bar plot of the total number of incidents/crimes per neighborhood in San Francisco.

number of incidents is particularly large in 6 neighborhoods: (1) Mission, (2) Downtown-Tenderloin, (3) South of Market, (4) Financial District, (5) Western Addition, and (6) Bayview. The Choropleth map of Fig.14 gives a more concrete visualization of the crimes distribution in the territory of San Francisco. Excluding Bayview, the other 5 more affected neighborhoods are the ones of the city center namely the red and orange areas in the map of Fig.14.

4.3. Exploring Incidents/Crimes Categories

The spatial distribution of the incidents/crimes discussed in the previous section gives an idea of the areas of San Francisco which are mostly affected by the events reported by the Police Department. However, it is important to have a better understanding of the kind of happenings

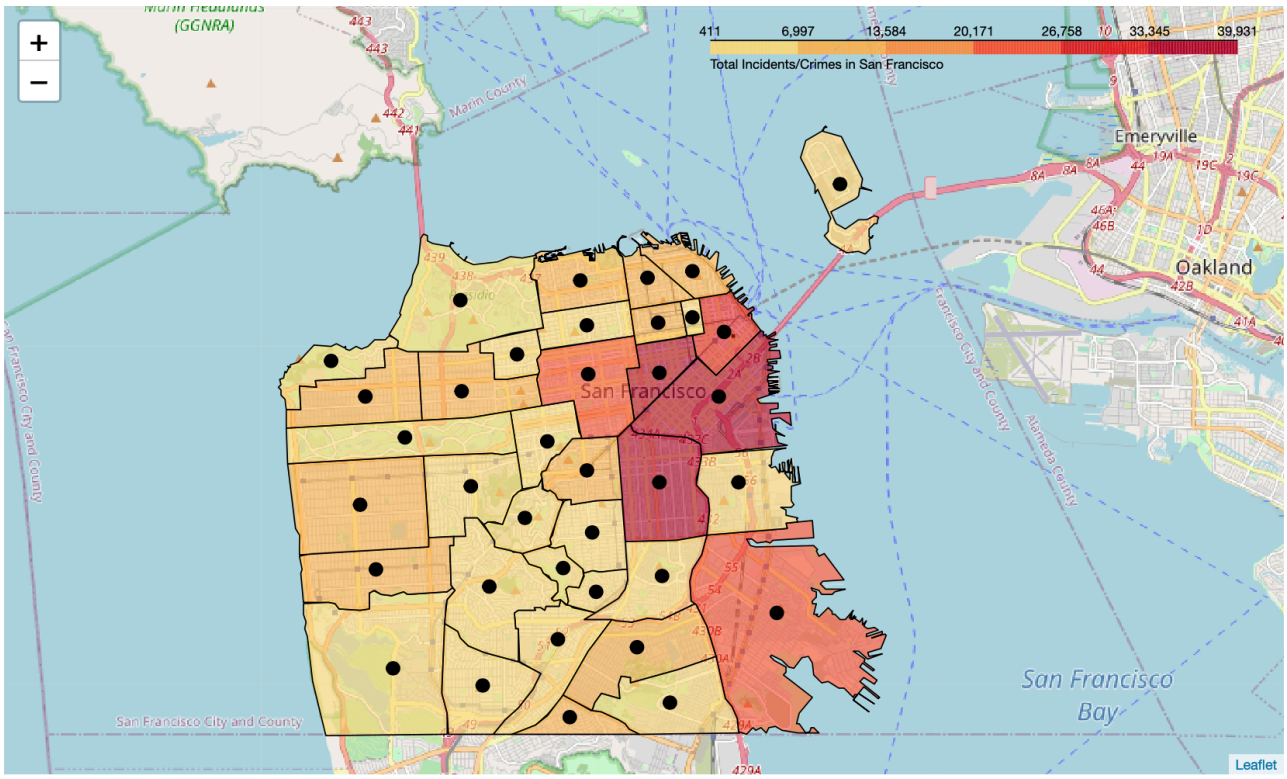

Figure 14 - Choropleth map of San Francisco showing the 37 neighborhoods of the "Planning Neighborhood Groups Map" division. Darker colors refer to a higher number of incidents/crimes in that neighborhood.

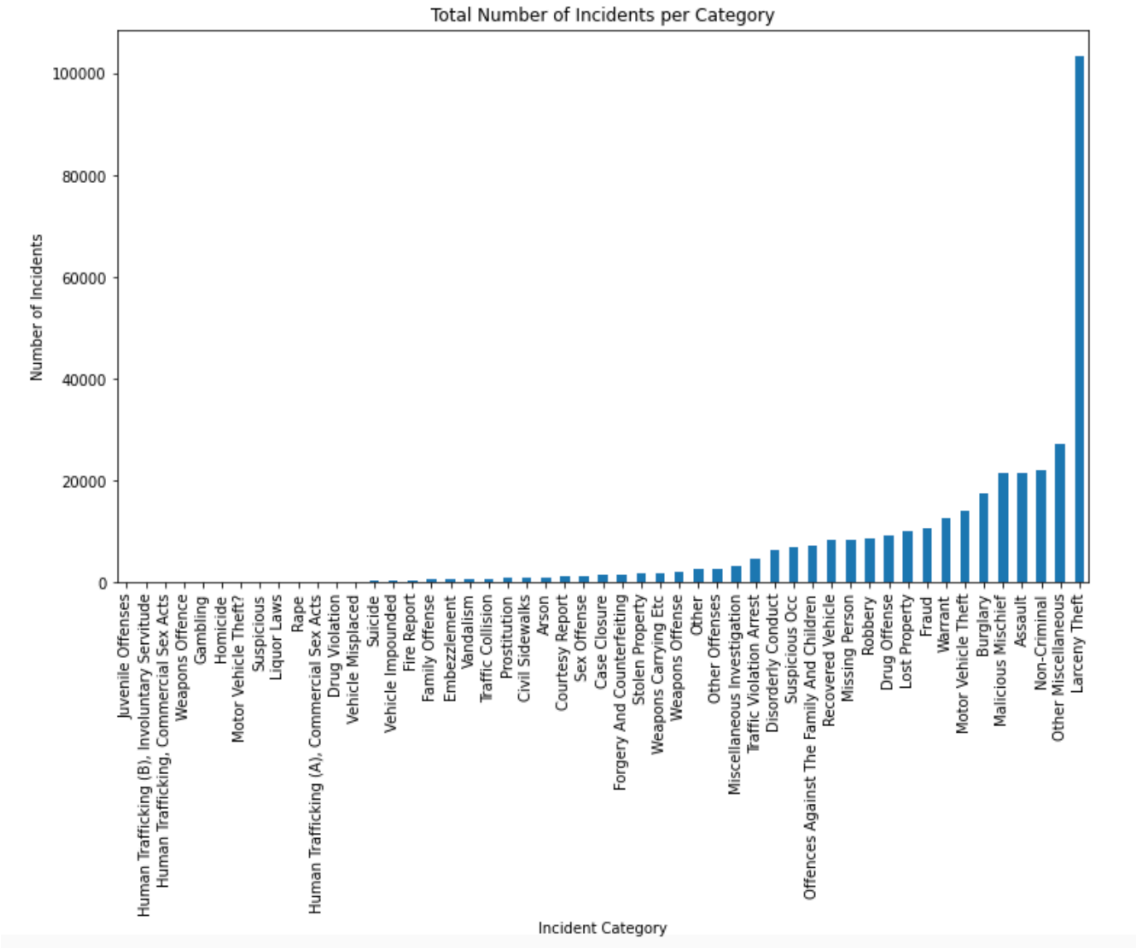

Figure 15 - Bar plot of the total number of incidents/crimes per category in San Francisco.

registered. To this end we plotted the total number of incidents per category, see Fig.15. Remarkably, it is evident that "Larceny Theft" is the most frequent type of crime in San Francisco, well above all the others. A more careful analysis shows that this is the most common incident category in every neighborhood (see the Jupyter Notebook [8]). In particular, "Larceny Theft" represents half of the total incidents/crimes in “Seacliff”, “Presidio”, and “Russian Hill”.

To check also the relevance of the other categories, we excluded the contribution of "Larceny Theft” and we obtained the plot of Fig.16. In this case, the most frequent type of incidents are: (1) “Other Miscellaneous”, (2) “Non-Criminal”, (3) “Assault”, (4) “Malicious Mischief”, (5) “Burglary”, and (6) “Motor Vehicle Theft”. Beside the first 3 categories, we notice that positions 4, 5 and 6 concern those kind of crimes related to the damage or stealing of properties.

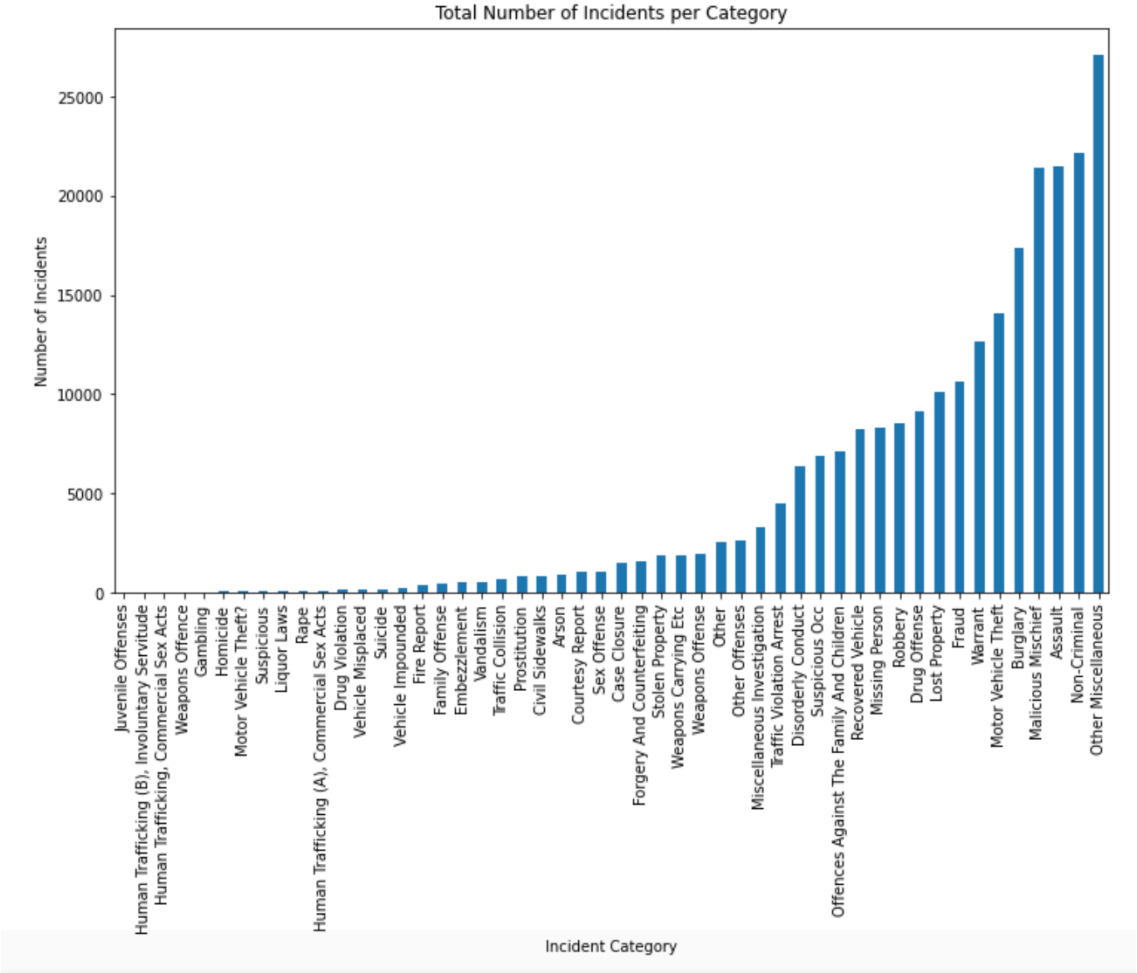

Figure 16 - Bar plot of the total number of incidents/crimes per category in San Francisco excluding "Larceny Theft” category.

4.4. Total Incidents/Crimes VS Time

Before dealing with the main part of this chapter namely the clustering of the neighborhoods of San Francisco in terms of the most common incidents/crimes, we finally analyzed how the total number of incidents/crimes behaves as a function of time. On this regard, from the main dataframe *incidents_SF_full*, we selected the attributes "Incident Date" and "Incident Category” and we counted the events for each date. The number of days reported from the first of January 2018 are 925. In Fig.17, we plotted the total number of incidents/crimes per day as a function of time. As pointed out by the linear regression included in Fig.17, there has been a decreasing trend of the number of incidents/crimes from 2018 till present, especially since February 2020.

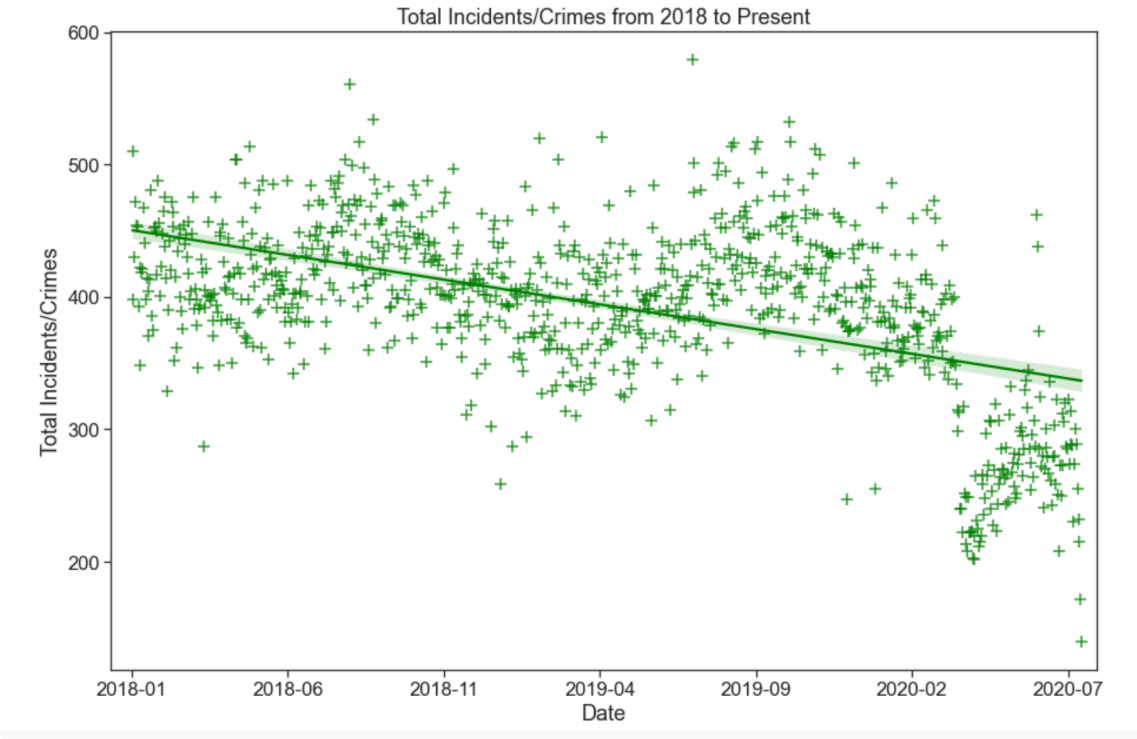

Figure 17 - Linear regression plot of the total number of incidents/crimes per day as a function of time from 2018-01-01 till present.

4.5. Clustering Incidents/Crimes

Thanks to the previous analysis of the incidents/crimes in San Francisco, we found that the 364275 events reported by the Police from 2018 till present are mainly located in the neighborhoods in the city center and lie mostly in the category called "Larceny Theft". When the latter is excluded, the most represented categories become (1) "Other Miscellaneous", (2) "Non-Criminal", (3) "Assault", (4) "Malicious Mischief", (5) "Burglary", and (6) "Motor Vehicle Theft".

To have a deeper understanding of how neighborhoods and incidents categories are linked to each other, we applied the clustering unsupervised technique to create groups of similar neighborhoods based on the type of "criminal" events occurred in San Francisco. As for the case of the venues treated in the previous chapter, we used the *k-means* clustering method. Since "Larceny Theft" category is the most common incident/crime category in each neighborhood (see

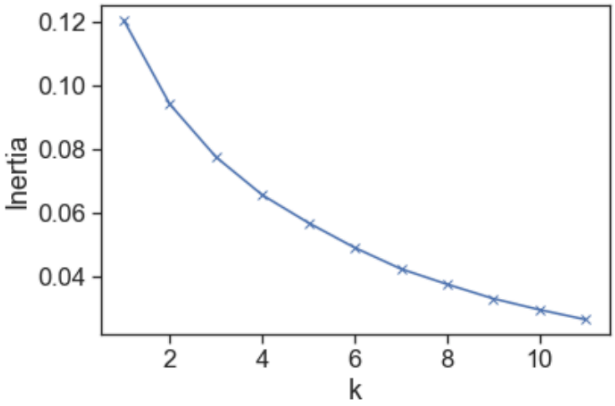

Figure 18 - Plot of the Inertia as a function of the number of clusters *k*.

Fig.15), we removed its contribution from the dataframe used for the clustering thus avoiding its strong bias on the obtained clusters (see the Jupyter Notebook[8] for details).

Without "Larceny Theft", we calculated the *Inertia* and plotted as a function of the number of clusters *k* to decide the optimal number of clusters using the elbow method. In this case, this is 3, 4 or 5, see Fig.18. After trying with different values, even larger than 5, we chose 7 and the resulting clusters are shown in the map of Fig.19 with different colors. In the Jupyter Notebook [8],

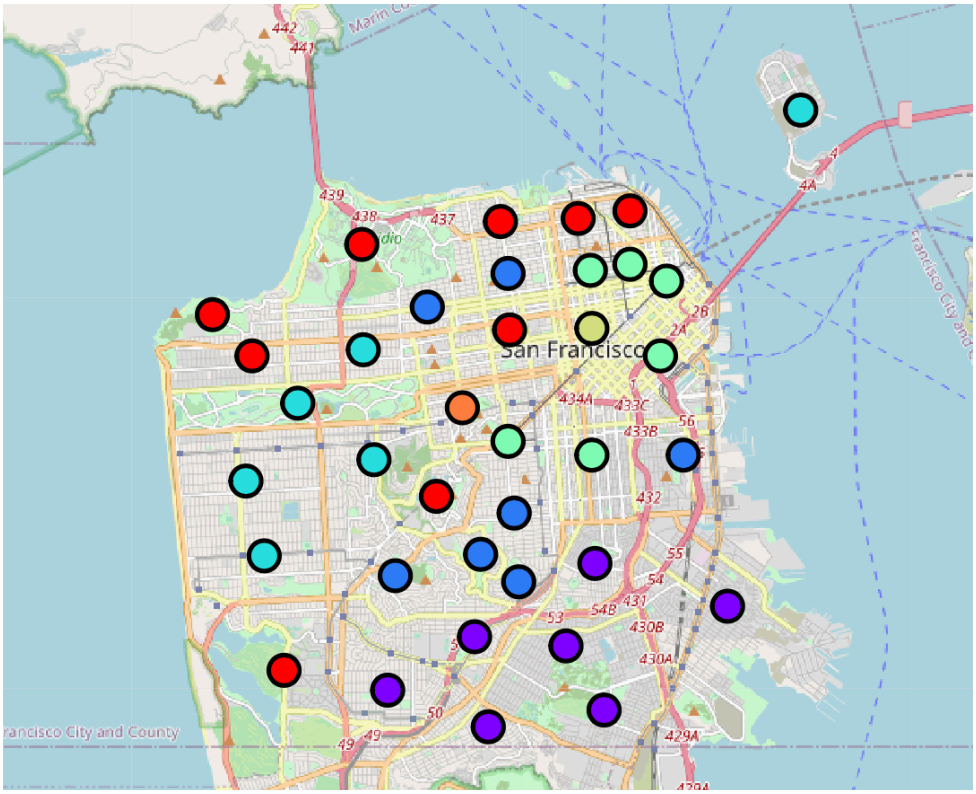

Figure 19 - Map of San Francisco showing the neighborhoods grouped in clusters. The circle markers are located at the “center” of each neighborhood. Each cluster has a different color: red (Cluster 0); violet (Cluster 1); blue (Cluster 2); light blue (Cluster 3); light green (Cluster 4); yellow (Cluster 5); orange (Cluster 6).

a click on the circle markers on the map of Fig.19 shows a popup with the name of the neighborhood and the cluster number. The latter goes from 0 to 6. From Fig.19 we see that, of the 7 clusters, Cluster 0 (red) is the largest one with 9 elements. The smallest clusters are 2, Cluster 5 (yellow) and Cluster 6 (orange) and contain only 1 neighborhood, namely "Downtown-Tenderloin" and “Haight-Ashbury”, respectively.

As in the previous chapter, in order to understand better the content of each cluster we created a dataframe per each cluster containing its neighborhoods together with the top 10 most common incidents/crimes. These dataframes can be found in the Jupyter Notebook [8]. The one for Cluster 3 is given in Fig.20, as an example. Qualitatively, from these dataframes we observed that in Cluster 0 (red), the largest, the dominant categories seem to be "Non-criminal", "Malicious Mischief" and "Other Miscellaneous". On the other hand, Cluster 1 (violet) and Cluster 4 (light green) contain mainly those neighborhoods where is likely to have "Other Miscellaneous" incidents. In addition, "Burglary" is clearly the main incident category in Cluster 2 (blue) whereas Cluster 3 (light blue) seems to be a "Non-Criminal" cluster (except for "Treasure Island" neighborhood). Finally, Cluster 5 (yellow) and Cluster 6 (orange), the 2 outliers, are characterized by "Drug Offense" crimes and by incidents mostly in the "Civil Sidewalks" category, respectively.

In the same way as for the clustering of the venues, to have a more quantitative point of view in the process of deciding the labels to assign to the obtained clusters, we considered also the total

	Neighborhood	Cluster Labels	1st Most Common Incident	2nd Most Common Incident	3rd Most Common Incident	4th Most Common Incident	5th Most Common Incident	6th Most Common Incident	7th Most Common Incident	8th Most Common Incident	9th Most Common Incident	10th Most Common Incident
2	Golden Gate Park	3	Non-Criminal	Other Miscellaneous	Malicious Mischief	Missing Person	Assault	Motor Vehicle Theft	Burglary	Recovered Vehicle	Lost Property	Warrant
3	Outer Sunset	3	Non-Criminal	Fraud	Malicious Mischief	Missing Person	Motor Vehicle Theft	Other Miscellaneous	Burglary	Assault	Offences Against The Family And Children	Lost Property
6	Inner Richmond	3	Non-Criminal	Other Miscellaneous	Malicious Mischief	Burglary	Missing Person	Motor Vehicle Theft	Assault	Fraud	Lost Property	Suspicious Occ
7	Inner Sunset	3	Malicious Mischief	Non-Criminal	Burglary	Missing Person	Motor Vehicle Theft	Other Miscellaneous	Fraud	Assault	Lost Property	Recovered Vehicle
33	Treasure Island	3	Assault	Burglary	Malicious Mischief	Non-Criminal	Missing Person	Other Miscellaneous	Motor Vehicle Theft	Suspicious Occ	Disorderly Conduct	Lost Property
36	Parkside	3	Non-Criminal	Fraud	Malicious Mischief	Missing Person	Motor Vehicle Theft	Other Miscellaneous	Burglary	Assault	Offences Against The Family And Children	Lost Property

Figure 20 - Dataframe for Cluster 3 containing its 6 neighborhoods and the top 10 most common incidents/ crimes.

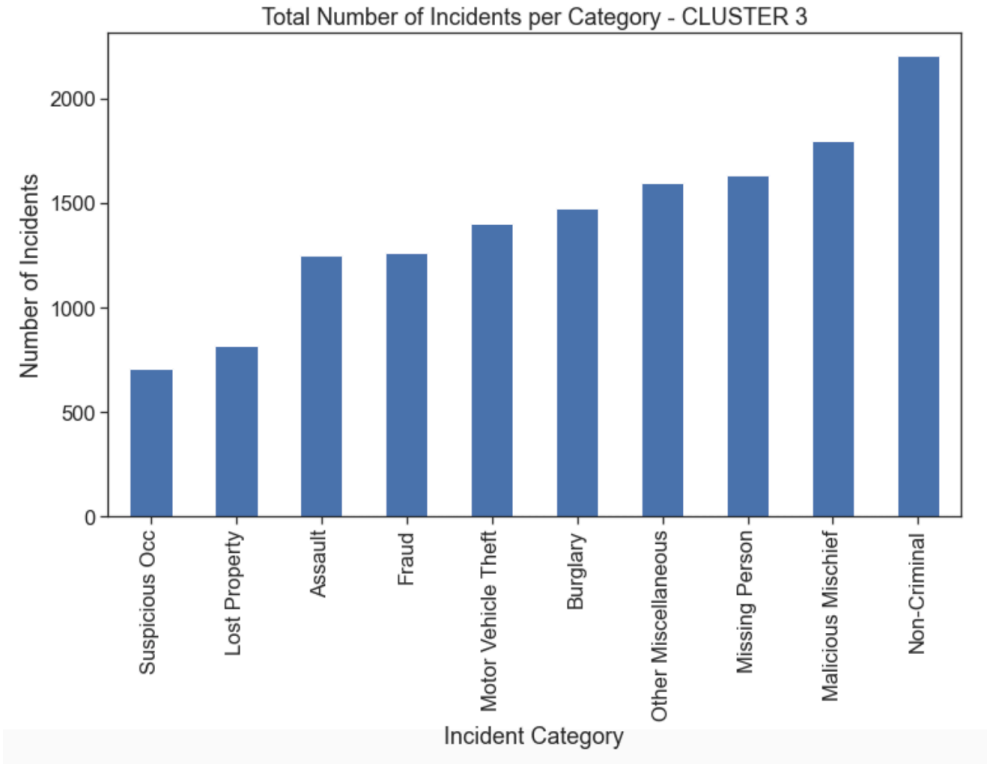

Figure 21 - Bar plot of the total number of incidents/crimes per incident category for Cluster 3.

number of incidents/crimes grouped per category in each cluster and we plotted it for the top 10 categories as shown, for instance, for Cluster 3 in Fig.21. The analysis of these kind of plots shows that among the most represented categories we find "Other Miscellaneous", "Non-Criminal" and "Malicious Mischief". However, we noticed that the sum of the elements in the categories "Malicious Mischief", "Burglary", "Robbery", "Motor Vehicle Theft" and "Lost Property", which are related to peoples properties, is well above all the other categories in each cluster. In other words, the category that one would call "Property at Risk" is the most represented everywhere in San Francisco (together with "Larceny Theft" that we removed from our analysis). Keeping that in mind and combining the information from the dataframes and bar plots like the ones in Fig.20 and 21, respectively, we labeled the 7 clusters as follows:

- Cluster 0 - Malicious Mischief - Miscellaneous - Non-Criminal
- Cluster 1 - Miscellaneous - Assault
- Cluster 2 - Burglary

- Cluster 3 - Non-Criminal - Missing Person
- Cluster 4 - Miscellaneous - Assault
- Cluster 5 - Drug Offense - Miscellaneous
- Cluster 6 - Civil Sidewalks - Non-Criminal

Here, beside Cluster 0, Cluster 1, and Cluster 4 which are rather miscellaneous, the other clusters have a more defined identity. In Cluster 2, we see that "Burglary" type of incidents are common while Cluster 3 is mostly "Non-Criminal" but crimes in the "Missing Person" category are frequent. Finally, Cluster 5 and Cluster 6 deserve a special attention. They are the smallest with just one neighborhood but they stand out from the other clusters for their peculiar type of crimes belonging to the categories "Drug Offense" and "Civil Sidewalks", respectively, as discussed previously. Why are these two neighborhoods so special from the incidents/crimes point of view? By checking the "Incident Description" attribute of the main dataframe, we found that in the "Drug Offense" category there are all those incidents related to drugs like possession and/or sale of Cocaine and Heroin. On the other hand, it is not really clear the meaning of "Civil Sidewalks". As a result of a careful search on internet, we found that this category refers to those incidents related to non-authorized occupation of sidewalks. Apparently, this involves mainly homeless people living on the sidewalks of "Haight-Ashbury" neighborhood. There has been a discussion about this kind of incident as reported in the article "The Sidewalks of San Francisco" [11].

By examining further the "Incident Description" attribute, it is also clear that the incidents reported in the categories "Non-Criminal" and "Other Miscellaneous" are both rather miscellaneous. A more careful analysis of these two categories is needed for an improved neighborhoods clustering, but this is out of the scope of this project.

4.6. Conclusions

In this chapter, we analyzed the neighborhoods of San Francisco in terms of the most common incidents/crimes reported by the Police Department. We used the data available at the link [6] to group the neighborhoods in clusters. To this end, we applied the *k-means* clustering method and we grouped the neighborhoods in 7 clusters depending on the incidents/crimes categories. We labelled them by considering the most common incidents as described at the end of the previous section. In our analysis, we removed the "Larceny Theft" category because it is the dominant one in every neighborhood. In addition, we took into account in a qualitative way the fact that the categories "Malicious Mischief", "Burglary", "Robbery", "Motor Vehicle Theft" and "Lost Property" give a large contribution to the total number of crimes in each neighborhood.

5. Combining Venues and Incidents Clusters

5.1. The Best Business Spot in San Francisco

The analysis of the neighborhoods in San Francisco carried out in the previous chapters gives a precise and clear qualitative understanding of how certain kind of venues and incidents determine the identity of each neighborhood. In particular, by examining and combining the two divisions in clusters showed in Figs.11 and 19, it is possible, for instance, to figure out which is the most convenient location where one could start a new business activity on the basis of both the already existing venues and the incidents/crimes that happened in that area.

As an example, let us focus on the case of opening a coffee shop or café. Where is the best business spot in San Francisco for this kind of venue? To answer this question, we start by considering the clusters obtained in terms of the most common venues of Fig.11. It turns out that Cluster 1 and Cluster 4 would be more suitable because in the corresponding neighborhoods the main venues are Chinese restaurants and food trucks, respectively, and coffee shops and café are less frequent. On the other hand, a new coffee shop opened in the other clusters would have to face a higher competition with similar venues. In particular, it would be rather inconvenient to start

this kind of business in the Sub-Cluster 1, the city center, where this kind of venue category is dominant.

On the basis of these considerations, let us now include in the discussion the clusters showed in Fig.19 which group the neighborhoods of San Francisco according to the most common incidents/crimes. By comparing the two maps of Figs.11 and 19, we see that Cluster 1 (venues) is split into Cluster 3, 1, and 0 (incidents) that we labelled as “Non-Criminal - Missing Person”, “Miscellaneous - Assault”, and “Malicious Mischief - Miscellaneous - Non-Criminal”, respectively. Differently, Cluster 4 (venues), which has only “Treasure Island” as neighborhood, belongs to Cluster 3 (incidents), characterized mainly by non-criminal events. However, a more carful analysis of the crimes happening on Treasure Island shows that there the main type of reported incidents are “Assault” and “Burglary” (see the Jupyter Notebook [8]). So, we exclude Cluster 4 (venues) from our list of possible best spot for the opening of a coffee shop or café since it does not seem a safe area. The only possibility would be then Cluster 1 (venues). Among the neighborhoods contained in the latter, we find that the optimal choice for our new business would be those ones belonging to Cluster 3 (incidents) because they are “Non-Criminal”.

In conclusion, our analysis suggests that the best spot in San Francisco where an entrepreneur can invest in starting an economical activity in the category “Coffee shop” or “Café” which is safer from criminal events is in Cluster 3 and more specifically in the neighborhoods Inner Richmond, Golden Gate Park, Outer Sunset, and Parkside. A similar analysis could be carried out for other kind of venues, different from “Coffee shop” or “Café”.

5.2. From Venues To Crimes

The careful comparison between the clusters showed in Fig.11 with the ones in Fig.19 allows also to give an answer to the following question: Does it exist a correlation between a given type of venue and a certain kind of incident/crime? In general, it is difficult to determine a clear connection between venues and crimes because it can be a complex problem. Despite that we show in the following that it is possible to correlate venues with crimes in a qualitative and straightforward way.

Let us first examine the venues clusters of Fig.11 which have a “better defined” or dominant venue category namely the clusters 1, 2, and 3 and sub-clusters 1 and 2. As discussed in chapter 3, the main venues in these groups of neighborhoods are Chinese restaurants, trails and parks, coffee places and restaurants with Italian cuisine, respectively. Notice that here we considered Cluster 2 and 3 as one cluster. Comparing these results with the incidents/crimes clusters of Fig.19, we can state that: non-criminal events are typically occurring in neighborhoods with Chinese restaurants; the presence of trails and parks is related to happenings in the “Malicious Mischief” category; the latter together with “Burglary” is the main incidents/crimes type in areas with Italian restaurants; the neighborhoods where coffee places are the most frequent venues are characterized by the occurrence of miscellaneous incidents, not necessarily criminal.

Secondly we deal with the outliers of the venues clusters which we found to be Cluster 4 and Sub-Cluster 3 and 4 containing the neighborhoods Treasure Island, Lakeshore, and Bayview, respectively. Although in all of them one can find venues belonging to a mix of categories, these neighborhoods can be identified by some specific type of venues which make them special. Let us examine them more in details and correlate to the corresponding incidents clusters. The first one, Treasure Island, is an island where it is common to find food trucks and where sport and music events can be hosted, as found by clustering in terms of venues. In this kind of neighborhood, we learn that assaults and burglary are common incidents/crimes, probably due to its isolated nature. The second one, Lakeshore, is characterized by the Lake Merced and the big park around it. In these case, the typical incidents/crimes are in the “Malicious Mischief” category, in agreement with what we discussed already above for Cluster 2 and 3 (venues) which contain neighborhoods with parks and trails. In addition, motor vehicle thefts seem to be also frequent in Lakeshore. Finally, from the third one, Bayview, we find that the presence of coffee places and Mexican restaurants but especially of the Hunters Point Naval Shipyard can be linked to the happenings of assaults and motor vehicle thefts.

As a final task, we apply the strategy discussed so far to the neighborhoods of Sub-Cluster 0. This is the largest venues cluster with 11 elements and, as shown in chapter 3, it is rather heterogeneous from the venues point of view. The most represented category is “Coffee Shop”. However one can easily find Italian and Mexican restaurants as well as parks in its neighborhoods. By comparing Sub-Cluster 0 with the incidents/crimes clusters discussed in chapter 4, we find out that the presence of these types of venues can be related to burglary events and happenings reported by the Police in the categories “Miscellaneous” and “Assault”. This result is in agreement with what we learnt from the correlation between the venues clusters 1, 2, and 3 and venues sub-clusters 1 and 2 with the corresponding incidents clusters and from the analysis of the outlier Bayview.

6. Conclusions and Future Developments

In this Data Science project, we dealt with the clustering of the neighborhoods of San Francisco in terms of the most common (i) venues and (ii) incidents/crimes. To this end, we applied the *k-means* clustering technique (i) to the datasets created from the venues obtained via Foursquare API calls and (ii) to the list of incidents reported by the Police Department of San Francisco from 2018 till present. The resulting clusters, showed in Figs.11 and 19, allow to gain valuable insights about the neighborhoods of this American city. As discussed in chapter 5, our analysis can be used, for instance, to figure out which is the best spot in the city where one could start a new business activity like the opening of a coffee shop. Also, we discussed how venues and incidents/crimes can be correlated to each other by comparing Figs.11 and 19. We found, for example, that one can determine the type of incidents/crimes happening in a certain cluster or sub-group of neighborhoods from the kind of most common venues. Our analysis, e.g., shows that typically the presence of Chinese restaurants in a certain neighborhood means that the happenings reported by the Police will be mostly in the “Non-criminal” category for that neighborhood.

The aim of this project was to provide an initial, qualitative but yet precise understanding of the nature or identity of the neighborhoods of San Francisco based on the most frequent venues and incidents/crimes categories. An improvement of the results discussed in the previous chapters as well as the discovery of new insights would be possible by means of a deeper analysis. Among the most relevant future developments we point out, first of all, the use of a neighborhoods division of San Francisco different from the “Planning Neighborhood Groups Map” such that it guarantees to choose an optimal value of center and radius of each circle used for the Foursquare API calls. In this regard, the dimensions of these circles could be adapted to the size of each neighborhood to avoid that the ones representing the smallest neighborhoods overlap too much and that the biggest ones are not covered enough, thereby improving the clustering in terms of the venues. In addition, a better or different grouping of San Francisco areas could be possible by using different clustering techniques than *k-means* such as hierarchical, fuzzy, and density-based clustering algorithms. Furthermore, new results could be achieved by grouping or removing those categories that are dominant and/or similar like “Coffee shop” and “Café” for the venues and “Malicious Mischief”, “Burglary”, “Motor Vehicle Theft”, “Robbery”, and “Lost Property” for the incidents/crimes. In particular, those events reported by the Police Department as “Non-Criminal” and “Other Miscellaneous” deserve a more careful investigation. All these further developments of this project would lead to more accurate results especially in finding more reliable qualitative correlations between type of venues and kind of incidents/crimes. In this regard, the use of a classification approach such as the decision tree algorithm would allow indeed to develop a quantitative and predictive model. The latter could be used by the Police, for instance, to determine the area or neighborhoods where an incident/crime with uncertain location might be occurred on the basis of the connection of the category of this event with a certain type of venue.

References

[1] San Francisco - Wikipedia - https://en.wikipedia.org/wiki/San_Francisco

[2] Foursquare - https://foursquare.com/city-guide

[3] Foursquare Developers - https://foursquare.com/developers/

[4] Planning Neighborhood Groups Map - https://data.sfgov.org/Geographic-Locations-and-Boundaries/Planning-Neighborhood-Groups-Map/iacs-ws63

[5] Google Maps - maps.google.com

[6] Police Department Incident Reports - https://data.sfgov.org/Public-Safety/Police-Department-Incident-Reports-2018-to-Present/wg3w-h783

[7] Venues Jupyter Notebook - https://github.com/angelodm/Coursera_Capstone/blob/master/San-Francisco_venues%26crimes/jupyter_notebooks/clustering_neighborhoods_San_Francisco_VENUES_.ipynb

[8] Crimes Jupyter Notebook - https://github.com/angelodm/Coursera_Capstone/blob/master/San-Francisco_venues%26crimes/jupyter_notebooks/clustering_neighborhoods_San_Francisco_CRIMES_.ipynb

[9] Analysis Neighborhoods - https://data.sfgov.org/Geographic-Locations-and-Boundaries/Analysis-Neighborhoods/p5b7-5n3h

[10] Data About San Francisco - https://datasf.org

[11] The Sidewalks of San Francisco - https://www.city-journal.org/html/sidewalks-san-francisco-13321.html