Services vs. Safety: are crime and venue types related in San Francisco?

Jani Peurakoski, 30th May 2020

1. Introduction

1.1. Background

Neighborhoods have their reputations, and these may or may not reflect the actual reality perceived by the inhabitants. An example of this is Soho in London, UK, which used to be an active center of night life but has since the 1990's become increasingly gentrified, while still having a somewhat colorful reputation. Are the venues in the neighborhoods (as night clubs in the case of Soho) related to registered incidents of crime? San Francisco is an ideal location to study this issue, because of it various neighborhoods and its excellent datasets made available by the city government.

1.2. Problem

I investigated if reported incidents of crimes in San Francisco are linked to the various venues that are located in that incident neighborhood: as an example, if there are more bars or nightclubs in one neighborhood, will there be more reported incidents?

1.3. Interest

The information about neighborhoods is relevant information for both landlords, sellers of real estate and families with young children, that are comparing different neighborhoods when relocating.

2. Data acquisition, cleaning, methodology

2.1. Data Sources

I used the Foursquare location data of San Fransisco for the venue information in combination with the San Fransico Police Department (SFPD) data (https://data.sfgov.org/Public-Safety/Map-of-Police-Department-Incident-Reports-2018-to-/jq29-s5wp). Both of these datasets should be up-to-date and are interlinked by latitude and longitude coordinates; both datasets have also a 'neighborhood'-field. As the SFPD database is **very big**, I limited the dataset to incidents reported between 1st Jan 2020 and 25th May 2020; this period is already by all means sufficient with 46,382 incidents. The Foursquare data was accessed via their API and the SFPD data was available in CSV-format from their home page mentioned earlier.

2.2. Data cleaning

The SFPD data was well structured, but included incidents without coordinate data or neighborhood name, causing the dropping of ca. 2000 rows until the dataset included 44.040

incidents. Data formats were very good and did not need specific formatting. I deleted several columns, which I considered unnecessary including the following:

1. Incident Datetime: incident date and time combined.
2. Report Datetime: incident report date and time combined
3. Incident Number: reporting number of the incident
4. Row ID, Incident ID, CAD Number: SFPD internal codes for the incidents

While the Foursquare data was well structured, it proved to be limited by a maximum of 100 venues per neighborhood.

As an interesting detail, San Francisco has had a long-time dispute about where one neighborhood starts and the other one ends. For this reason, the city government recently published the coordinates of the neighborhoods (https://data.sfgov.org/Geographic-Locations-and-Boundaries/Analysis-Neighborhoods/p5b7-5n3h). I used the maximum amount of neighborhoods, 41 in all, as presented in the SFPD data.

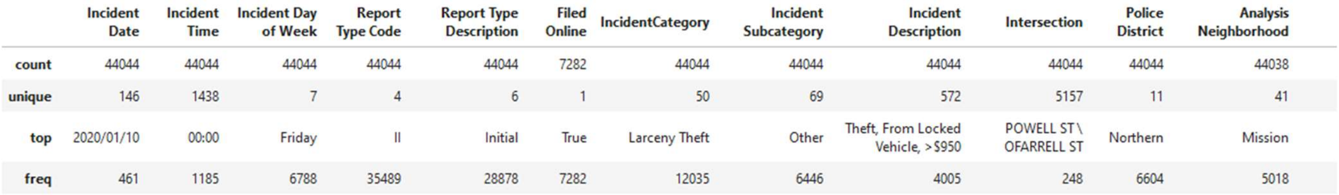
2.3. Methodology

I grouped the SFPD data and thereafter clustered it per neighborhood on Folium maps and subsequently investigated if there are some neighborhoods with higher amount of crime incidents, and if so, if there more of some specific type of venues in those neighborhoods compared to other neighborhoods that might be able to explain the higher frequency of reported crime incidents. I used one-hot coding.

SFPD records all incidents with incident-specific latitude and longitude coordinates, while naming the neighborhood of the incident. In order to get a realistic view on the true coordinate center of the neighborhood, I decided to calculate the mean of the neighborhood-specific latitude and longitude coordinates based on the reported incidents. For the purposes of this study this should associate the Foursquare data better with the location of the incidents.

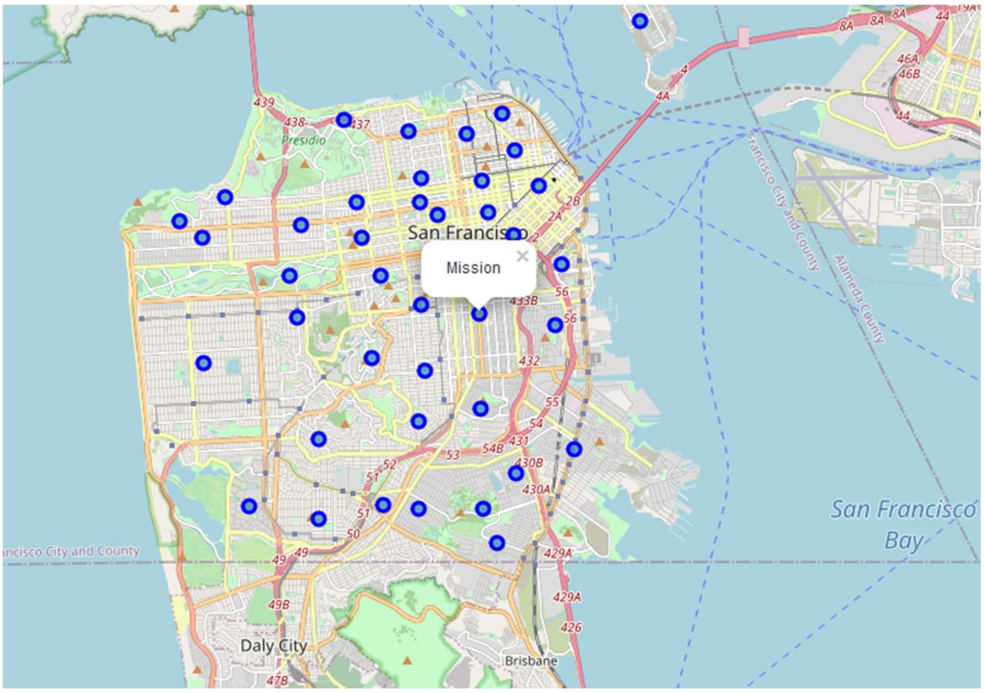
3. Exploratory Data Analysis

The first look at the description of the SFPD dataset gave a clue of a concentration of reported incidents to have taken place at the neighborhood called Mission.

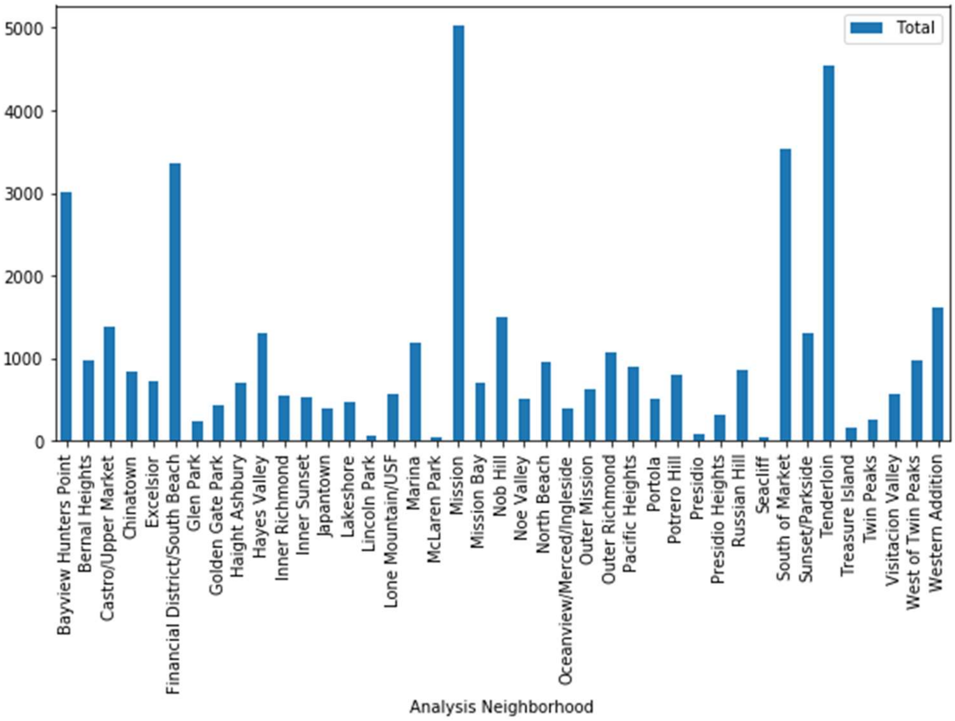


	Incident Date	Incident Time	Incident Day of Week	Report Type Code	Report Type Description	Filed Online	IncidentCategory	Incident Subcategory	Incident Description	Intersection	Police District	Analysis Neighborhood
count	44044	44044	44044	44044	44044	7282	44044	44044	44044	44044	44044	44038
unique	146	1438	7	4	6	1	50	69	572	5157	11	41
top	2020/01/10	00:00	Friday	II	Initial	True	Larceny Theft	Other	Theft, From Locked Vehicle, >$950	POWELL ST \ OFARRELL ST	Northern	Mission
freq	461	1185	6788	35489	28878	7282	12035	6446	4005	248	6604	5018

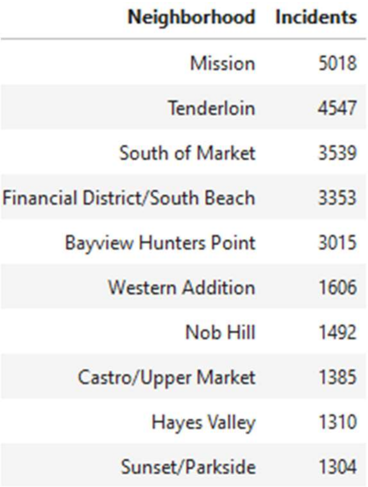
Below, the 41 neighborhoods of San Francisco, with Mission marked:



The first impression of Mission being the incident leader was further confirmed by comparing the neighborhood total incidents:

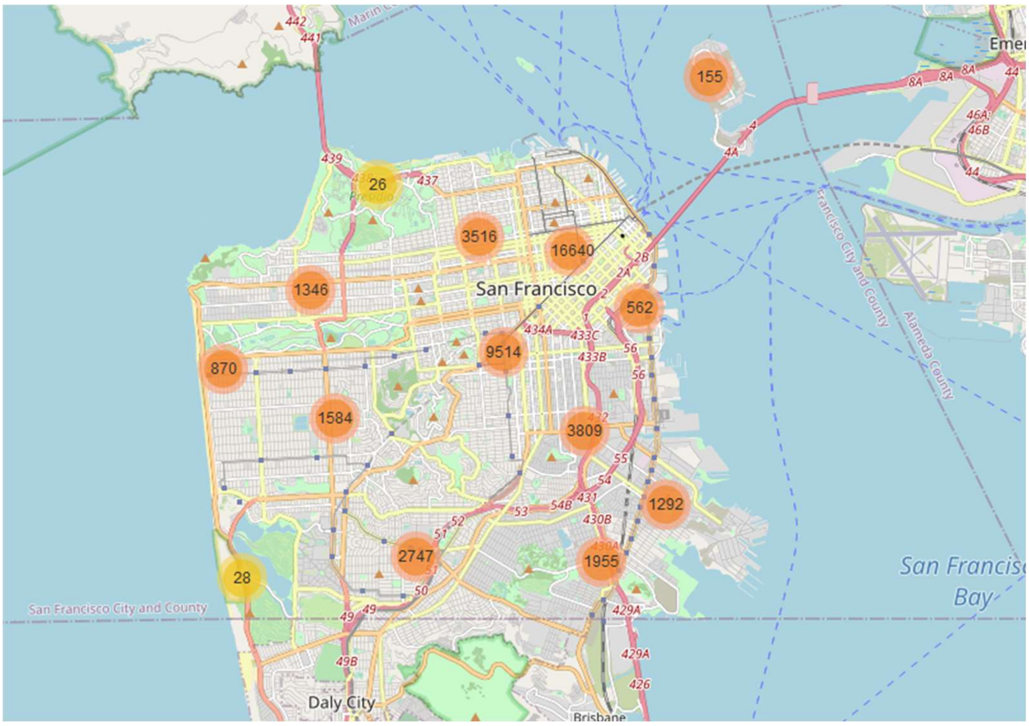


Interestingly, the slots 2 and 3 behind Mission are taken by the neighboring neighborhoods, Tenderloin and South of Market. Jointly these three neighborhoods cover 29.75% of San Francisco reported police incidents, indicating that Northeastern San Francisco is more prone to crime than other parts of the city. Below are the top 10 incident neighborhoods. Please recall, that the total reported incident amount for the period is 44,044.

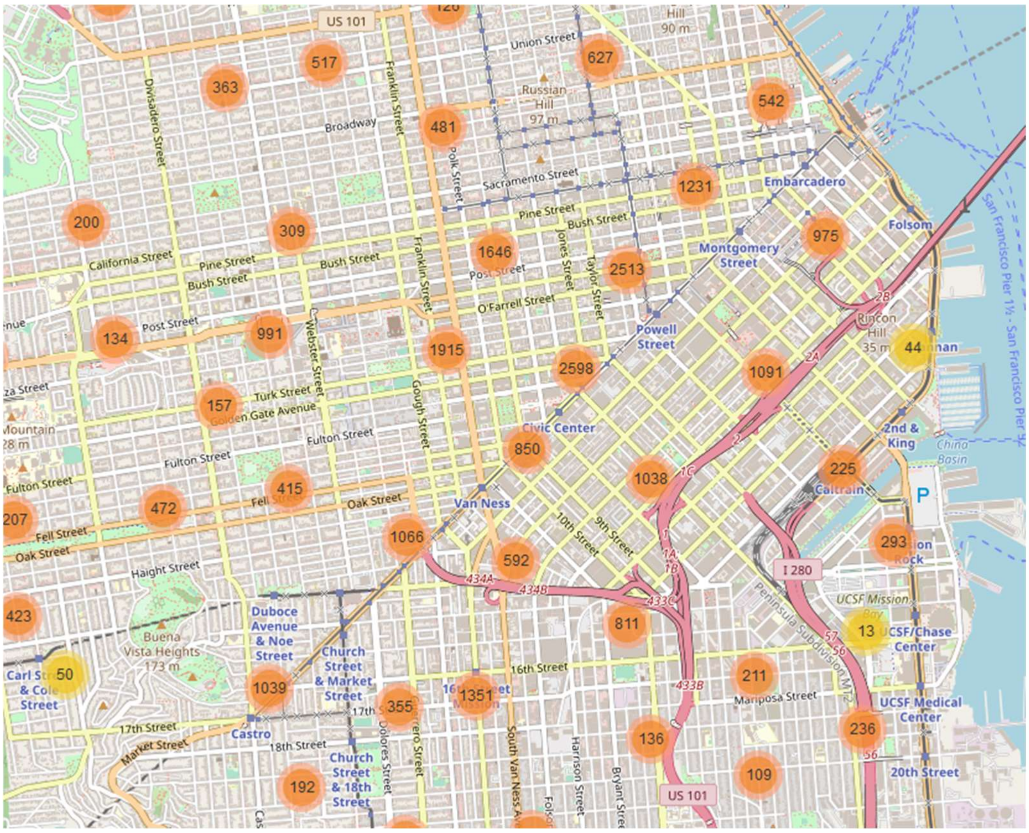


Neighborhood	Incidents
Mission	5018
Tenderloin	4547
South of Market	3539
Financial District/South Beach	3353
Bayview Hunters Point	3015
Western Addition	1606
Nob Hill	1492
Castro/Upper Market	1385
Hayes Valley	1310
Sunset/Parkside	1304

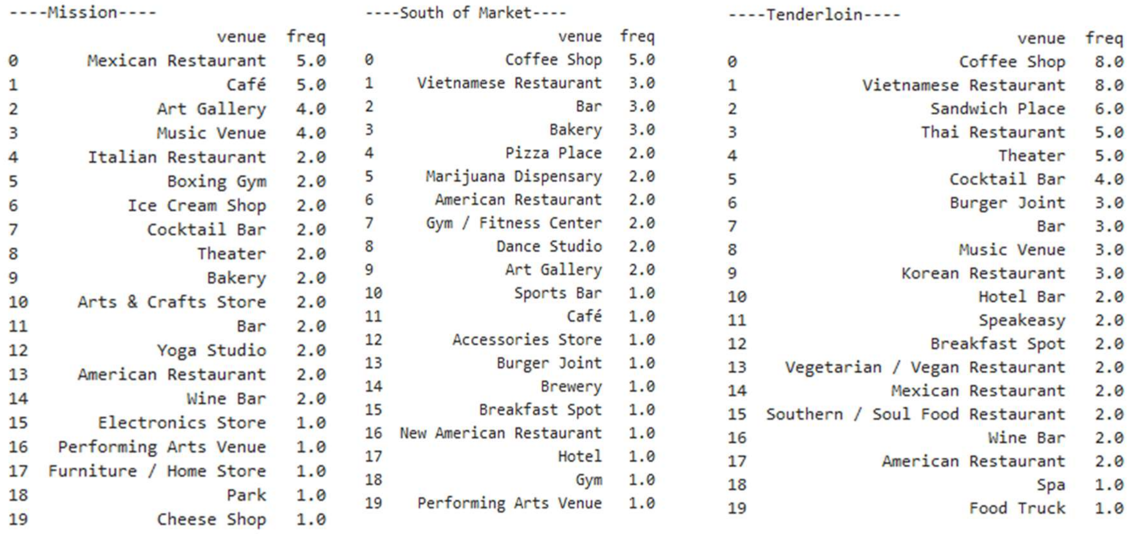
Clustering of the incidents (below) show a higher concentration in north-eastern part of the city, which was to be expected:



And in more detail:



Looking at the Foursquare data, it returned 2053 different venues in 307 categories. On neighborhood level, it was surprising to see that Mission has a smaller number of venues available, 86, compared to Tenderloin's 100 (which is the venue amount cap as well at Foursquare with a free license) or, for example, that of Chinatown's 99 or Marina's 100. Jointly, Mission, Tenderloin and South of Market include 242 venues, or 11.8% of all venues registered in Foursquare for San Francisco[1]. The top 20 most frequent venues for the top 3 incident neighborhoods are below:



```
----Mission----
                    venue  freq
0      Mexican Restaurant   5.0
1                    Café   5.0
2             Art Gallery   4.0
3             Music Venue   4.0
4      Italian Restaurant   2.0
5              Boxing Gym   2.0
6         Ice Cream Shop    2.0
7            Cocktail Bar   2.0
8                 Theater   2.0
9                  Bakery   2.0
10     Arts & Crafts Store  2.0
11                    Bar   2.0
12            Yoga Studio   2.0
13    American Restaurant   2.0
14               Wine Bar   2.0
15      Electronics Store   1.0
16    Performing Arts Venue 1.0
17  Furniture / Home Store  1.0
18                   Park   1.0
19            Cheese Shop   1.0
```

```
----South of Market----
                    venue  freq
0             Coffee Shop   5.0
1    Vietnamese Restaurant  3.0
2                     Bar   3.0
3                  Bakery   3.0
4             Pizza Place   2.0
5     Marijuana Dispensary  2.0
6      American Restaurant  2.0
7      Gym / Fitness Center 2.0
8            Dance Studio   2.0
9             Art Gallery   2.0
10             Sports Bar   1.0
11                   Café   1.0
12      Accessories Store   1.0
13           Burger Joint   1.0
14                Brewery   1.0
15         Breakfast Spot   1.0
16 New American Restaurant  1.0
17                  Hotel   1.0
18                    Gym   1.0
19   Performing Arts Venue  1.0
```

```
----Tenderloin----
                            venue  freq
0                     Coffee Shop   8.0
1           Vietnamese Restaurant   8.0
2                  Sandwich Place   6.0
3                 Thai Restaurant   5.0
4                         Theater   5.0
5                    Cocktail Bar   4.0
6                    Burger Joint   3.0
7                             Bar   3.0
8                     Music Venue   3.0
9               Korean Restaurant   3.0
10                      Hotel Bar   2.0
11                      Speakeasy   2.0
12                 Breakfast Spot   2.0
13   Vegetarian / Vegan Restaurant  2.0
14             Mexican Restaurant   2.0
15  Southern / Soul Food Restaurant 2.0
16                       Wine Bar   2.0
17            American Restaurant   2.0
18                            Spa   1.0
19                     Food Truck   1.0
```

[1] Note that the 100 venue cap is very likely to have an impact on the total amount of venues received from Foursquare. The true amount of venues in San Francisco and her neighborhoods remains therefore unknown.

Obviously, bars (including wine bars and cocktail bars) and nightclubs are outnumbered by coffee shops and various restaurants.

4. Conclusions

It is evident that reported crime incidents in San Francisco are clustered around the Northeastern part of the city. Mission is the clear leader in single incidents, and together with Tenderloin and South of Market cover almost 30% of all incidents. The first conclusion is, therefore that crime is more regional (or pan-neighborhood) than related to a single neighborhood.

It was a revelation to see that the venues in Mission, Tenderloin and South of Market were not something that I associate with criminality. As speculation:

1. This might have to do with the nature of Foursquare: people who use Foursquare intend to use it in their social media, possibly leading to a skewed dataset; users do not report all the venues they visit, or rate them. Do we only look at socially approved venues when looking at Foursquare?
2. Did the venue limit of 100 pre neighborhood have an impact on the result? Probably yes, but as Mission had only 86 venues, and still lead the statistics, it might not play a big role.

With the information at hand, I draw the following conclusion: Number of reported criminal incidents is not related to the venue types or number of venues in the neighborhood.

5. Suggestions for further studies

On May 29th 2020 San Francisco city published a new dataset that includes all registered business locations in the city (https://data.sfgov.org/Economy-and-Community/Registered-Business-Locations-San-Francisco/g8m3-pdis). This would probably be a better dataset to use when drawing conclusions about the link between venues and crime. Also, a closer look at the socioeconomical status of neighborhoods should be used when making comparison- a good outset would be the following report: (https://default.sfplanning.org/publications_reports/SF_NGBD_SocioEconomic_Profiles/2010-2014_ACS_Profile_Neighborhoods_v3AH.pdf). Another approach would be to follow the price development of real estate in the neighborhoods, by combining historical incident data with historical prices.