Relation Between crime rate and the surrounding Businesses in a Neighbourhood

1. Introduction

This report is part of the IBM data science Coursera capstone project and will give an outline of the process I went through in the project by giving a detailed summary of the data, the analysis and methods and finally my results and conclusion.

The main question that we are trying to answer through this report is “Is there a connection between the amount of crime in a neighbourhood and the number or type of nearby businesses.” The type of people that would be interested in this analysis would be the local police and law makers. This will help them determine the cause of certain crimes, predict the effect that a new business might have on the local area and finally make decisions on where to locate certain resources based on the crime rate and the surrounding area. Additionally, this study can be of use to new business owners searching for the best place to open a business. They may be interested in the crime data as the safety of their property and customers is the top priority.

Throughout the project data was gathered and cleaned from two sources as will be outlined in the report. The data works together to produce this analysis and to round it off, k-means clustering is used to try and group the neighbourhoods based on certain characteristics.

2. Data

The first step of the project after we had the question in place was to try and find the right data to help with the analysis. The project required for the foursquare API to be used as a form of location data analysis. The foursquare data needs to be supplemented with geographical coordinates and possibly a second dataset to analyse. This section will outline the two sources of data used as well as give details on the foursquare API and how it was used.

2.1 Foursquare API

Foursquare is the leading location data provider in the space providing location data for some of the biggest players such as Uber. The foursquare API gives us the ability to search for certain types of venues in a given location based on the Latitude and Longitude coordinates we give it. Then foursquare returns data on all the different businesses in an area which is very useful for our use case.

2.2 Crime data

The main idea of this data science analysis revolves around crime data. So, at this stage it was time to choose the city we wanted to work with and analyse the crime data in its different neighbourhoods. The city chosen was San Francisco and this was inspired by an article found of someone who had also completed this capstone project (see reference). The San Francisco crime data was collected from a government website. This data of course was cleaned, and some exploratory analysis was performed to get some insights from the data.

3. Method

This section will outline the method used including how the data was gathered as well as the cleaning and analysis done on each dataset. First, we will start with the crime data which was loaded using the panda's library into a data frame. The data was in the form of a csv file and so was relatively easy to download from the San Francisco government website. Seen below is the dataset obtained directly from the website and in later sections we will see some of the cleaning done to the data.

Crime data

```
In [3]: crime = pd.read_csv('https://data.sfgov.org/api/views/wg3w-h783/rows.csv?accessType=DOWNLOAD')
        print(crime.shape)
        crime.head()
```

```
(418994, 36)
```

Out[3]:

	Incident Datetime	Incident Date	Incident Time	Incident Year	Incident Day of Week	Report Datetime	Row ID	Incident ID	Incident Number	CAD Number	...	SF Find Neighborhoods	Current Police Districts	Current Supervisor Districts	A N
0	2018/01/01 09:26:00 AM	2018/01/01	09:26	2018	Monday	2018/01/01 09:27:00 AM	61893007041	618930	171052174	173641140.0	...	88.0	2.0	9.0	1
1	2018/01/01 02:30:00 AM	2018/01/01	02:30	2018	Monday	2018/01/01 08:21:00 AM	61893105041	618931	180000768	180010668.0	...	90.0	9.0	1.0	7
2	2018/01/01 10:00:00 AM	2018/01/01	10:00	2018	Monday	2018/01/01 10:20:00 AM	61893275000	618932	180000605	180010893.0	...	20.0	4.0	10.0	3
3	2018/01/01 10:03:00 AM	2018/01/01	10:03	2018	Monday	2018/01/01 10:04:00 AM	61893565015	618935	180000887	180011579.0	...	NaN	9.0	1.0	2
4	2018/01/01 09:01:00 AM	2018/01/01	09:01	2018	Monday	2018/01/01 09:39:00 AM	61893607041	618936	171052958	180011403.0	...	106.0	6.0	3.0	6

5 rows × 36 columns

Figure 1 Initial Crime dataset

In figure 1 above we see the crime dataset, this initially consisted of 36 columns and some of which were of no use to us in this situation. Which is why it was necessary to drop most of the columns and only keep ones that will help us throughout the project. In terms of exploratory analysis, we needed to see which neighbourhoods had the most crime and by what margins. This was done by grouping the crime data based on the neighbourhood which gives the number of crimes in each area throughout a specific time frame.

Below we can see the final dataset we obtained after grouping and cleaning.

Out[5]:

	Neighborhood	Incident_type
18	Mission	44423
35	Tenderloin	41121
5	Financial District/South Beach	34398
33	South of Market	33867
0	Bayview Hunters Point	24424
40	Western Addition	13036
2	Castro/Upper Market	12292
22	North Beach	11922
34	Sunset/Parkside	11805
20	Nob Hill	11541

Figure 2 Cleaned Crime Dataset

Next, we gathered the venues data using the foursquare API. The first step in doing so was to create a developer account on the foursquare website and get the credentials that allow us to make requests to the website. Then we needed to gather the Latitude and Longitude coordinates of each neighbourhood which was done using the python geopy library. This gave us a data frame that had the top 10 neighbourhoods along with the corresponding coordinates which we can then use as a query to the foursquare API. We used this to retrieve all the venues within a 5-mile radius of the neighbourhoods.

Out[16]: (289, 7)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Mission	37.7599	-122.4148	Kahnfections	37.758873	-122.415152	Bakery
1	Mission	37.7599	-122.4148	Mission Cliffs	37.760654	-122.412474	Climbing Gym
2	Mission	37.7599	-122.4148	Moxie Yoga	37.758979	-122.414899	Yoga Studio
3	Mission	37.7599	-122.4148	Hit Fit SF	37.759795	-122.412651	Boxing Gym
4	Mission	37.7599	-122.4148	flour + water	37.759062	-122.412334	Italian Restaurant

Figure 3 Foursquare Data

4. Analysis

Using the figures above we were able to perform some exploratory analysis and produce a couple of insightful graphs on the dataset. Graph 1 includes a bar char of the top 10 neighbourhoods for crime in San Francisco and clearly shows a large disparity between the top 2 neighbourhoods especially and the rest.

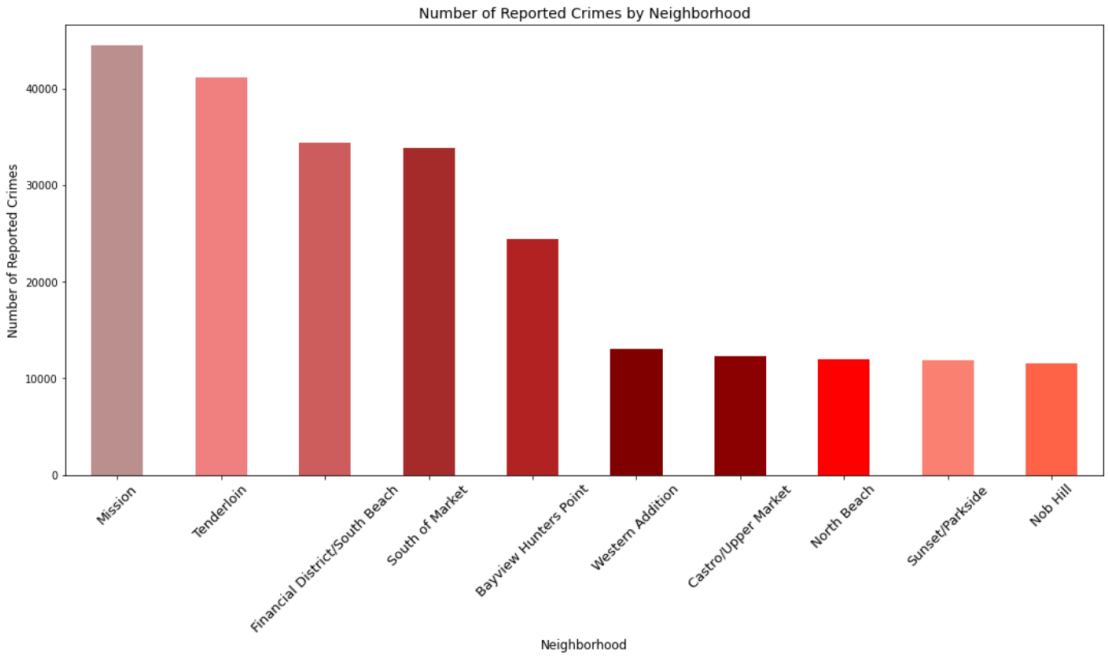

Figure 4 Top 10 Neighbourhoods with crime in San Francisco

The next graph was an outline of the most popular types of businesses in the area as returned by the foursquare API. This allows us to get a feel for the type of businesses that are most common in San Francisco and was clear that the food and beverage industry was the most popular by far, holding 4 of the top 5 positions and 4 of the top 4.

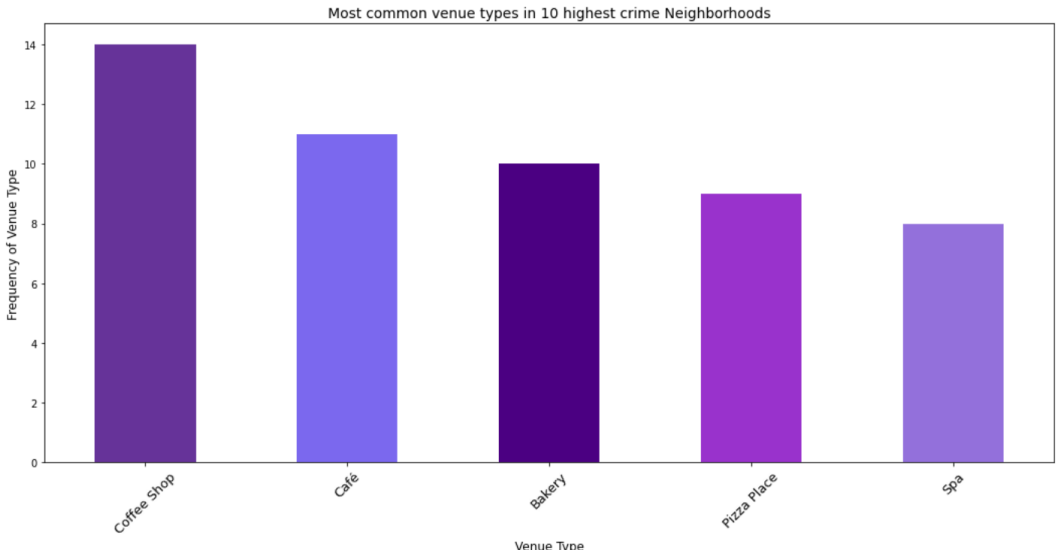

Figure 5 Most common business types in top 10 neighborhoods for crime

The above two graphs act as the exploratory analysis for our two datasets. The first outlines the top 10 neighbourhoods with crime in the San Francisco area which narrowed down all the neighbourhoods to the ones with the highest crime and so the reasons for these numbers can be analysed. The second graph shows us the most popular types of businesses in these crime ridden neighbourhoods, however the analysis was not very telling as most of the venues in the top 5 belong to the food and beverage industry.

5. Results

The refined and cleaned data sets were then merged together with all the important pieces of information such as the number of crimes and the rankings of all the most common neighbourhoods in the area to allow for clustering of our top 10 neighbourhoods. The aim was to try and spot any patterns for the neighbourhoods with the highest crime numbers.

This clustering analysis was then plotted using folium on a map of San Francisco with each point on the map colour coded according to the cluster its been placed in as well as the size based on the number of crimes.

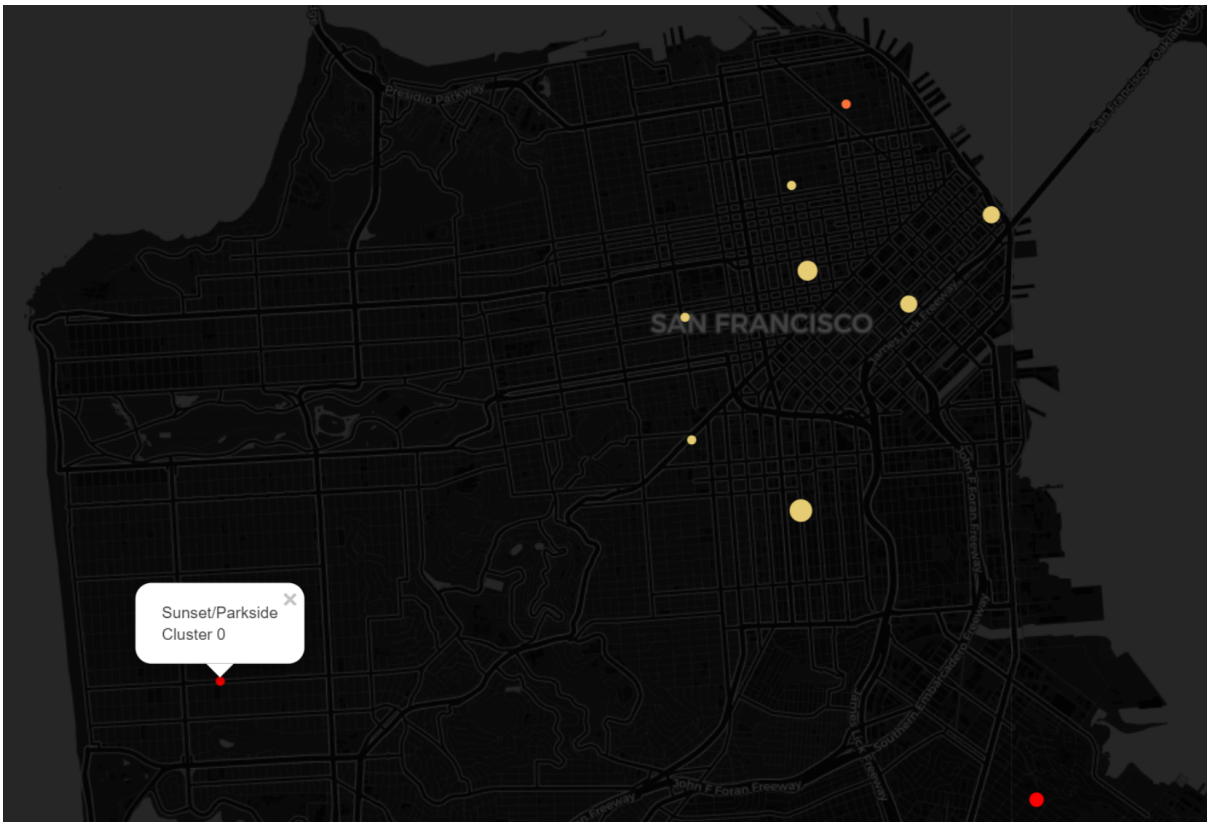

Figure 6 Final Map with clustered neighborhoods

6. Discussion and Conclusion

From the clusters obtained of the top 10 neighbourhoods with crime in San Francisco, we were able to make some deductions on the crime data as well as the data on the different venues in the neighbourhoods. Out of the top 10 neighbourhoods for crime from our dataset, 7 of them were placed

into cluster number 1. Of these 7 neighbourhoods 5 of which were in the top 5 neighbourhoods for crime incidents in San Francisco. On the other hand, cluster 0 and 2 contained 2 of the bottom 3 neighbourhoods when it comes to crime from the dataset. Further inspecting the aggregated data which contains the top 30 venues in each area of San Francisco we can gather some more incites on the data. First, for Financial district, Castro and south of market gyms/fitness centres were present in the top 5 venues in those area. These 3 neighbourhoods did rank highly on the crime ranking but were not towards the top. Interestingly, the two run away areas for the top neighbourhoods for crime had a similar distribution of venues in the top 5. This included cafes and bar with the highest rated health and fitness place ranking 23rd for mission. Perhaps this is a contributing factor to the crime in an area, the number of places there are to expend energy and exercise.

Other ways that these insights can be helpful is for anyone looking to start up a new business in these neighbourhoods. From the data, new owners would do well to avoid most of the neighbourhoods in cluster 1 as these are mostly made up of the top crime neighbourhoods in the area. Anyone looking to open a business weather that be a restaurant, school or bar would want their customers to feel safe especially when there are children involved. Additionally, the list for the top 30 most common businesses in each venue can help new investors identify any market gaps they may want to fill. Having a look at the less common businesses found in the neighbourhoods can give new owners a competitive advantage in an area with very little competition.

To conclude, we were able to use the various tools for an end-to-end data science project. These included the foursquare api to fetch data on the different venues as well as mining different datasets to obtain interesting data for our research.

7. References

[1] https://github.com/davisnix/Opening-a Bar/blob/master/Opening%20a%20Bar%20in%20San%20Francisco%20.ipynb