An investigation into the correlation between crime rates and police station presence in San Francisco

Coursera Capstone Project – by Theresa Hughes

1. Introduction/Business Problem

This analysis will explore the relationship between crime rates and police station presence in San Francisco. The analysis will look at the number of criminal instances versus the number of police stations in each district of San Francisco.

There will be 2 aspects to the analysis. Firstly, I will investigate which of the districts of San Francisco is safest in terms of number of criminal incidents. Limitations of this investigation are that crime rates will not be scaled based on district size, e.g. crime rate per 100sqm, and no other factors will be considered.

Secondly, I will investigate if there is a correlation between the number of police stations and the number of incidents of crime using a regression plot. The expectation is that there will be a negative correlation between the two variables, i.e. more police stations leads to less crime.

The first part of the analysis will be of interest to people looking to become residents of San Francisco to give them a view of where the safest areas are. The second part of the analysis will be of use to the Police Department of San Francisco to understand whether establishing police stations in areas with higher crime rates might deter criminal activity and make it a safer area.

2. Data

The analysis will leverage data on crime in San Francisco as provided during the Coursera Python course. Data is downloaded in csv format from https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DV0101EN-SkillsNetwork/Data%20Files/Police_Department_Incidents_-_Previous_Year__2016_.csv. This dataset contains a description of the crime, time, date, geo-location, what the outcome of the report was, i.e. prosecution, etc. The analysis will extract from this dataset criminal instances such as theft, assault, etc. and focus only on the crimes which have a general impact, e.g. suicide is no immediate danger to the general public; it is an individual instance. This data will be grouped by district to count the number of crimes per district. We can then identify the district with the highest crime rate. This data will also be used in the regression analysis for part two.

The analysis will also be accessing data from the Foursquare API to find the location of all police stations in San Francisco. As the San Francisco crime dataset was based on 2016, the Foursquare API will be accessed with a version date of 2016 for consistency. Data will be reviewed and any entries which are deemed to not be valid police stations will be removed. This data contains the names and locations (coordinates) of all police stations which will allow us to plot them on a map. The number of police stations will then be counted per district in order to run the regression against the number of crimes for the second part of the analysis.

3. Methodology

This analysis is split into four parts: data setup, data cleansing, data visualisation, and statistical analysis. Each will be discussed in turn in the following sections. There are a number of assumptions being made in this analysis:

1. All police stations retrieved by the Foursquare API are actual, functioning police stations (unless known otherwise)
2. All police stations are opened 24 hours
3. There has been no material change in the levels of crime in each district since 2016.

3.1 Part One: Data Setup

This part of the project reads in all required libraries and packages needed throughout the project. We also read in the GeoJson file of the San Francisco Police Department districts to allow us to graph the data. This GeoJson file was downloaded from: https://data.sfgov.org/Public-Safety/Current-Police-Districts/wkhw-cjsf. We then create a map of San Francisco, using this GeoJson file to map the outlines of each PD district. We next import the crime dataset, dropping unneeded fields - Incident Number, Day of the week, time, location, Police Department ID. The last step in this setup part of the project is to call the Foursquare API to access data on police station locations in San Francisco. The API is called using the version from 2016 which is identified as the reporting year of the crime dataset.

3.2 Part Two: Data Cleansing/Wrangling

The data wrangling part of the analysis is quite important as this is where we ensure the data we have available is in the correct format in order to draw insightful conclusions for the project. We transform the raw police station data into a pandas dataframe listing all police stations in San Francisco. We then drop 2 entries which are deemed to not be valid police stations. One of these is a “community room” and the other is an “old” police station in the Mission district. A quick Google search showed that this station has not operated in many years and has been replaced by the new Mission station. After dropping these 2 observations, the final dataframe is ready.

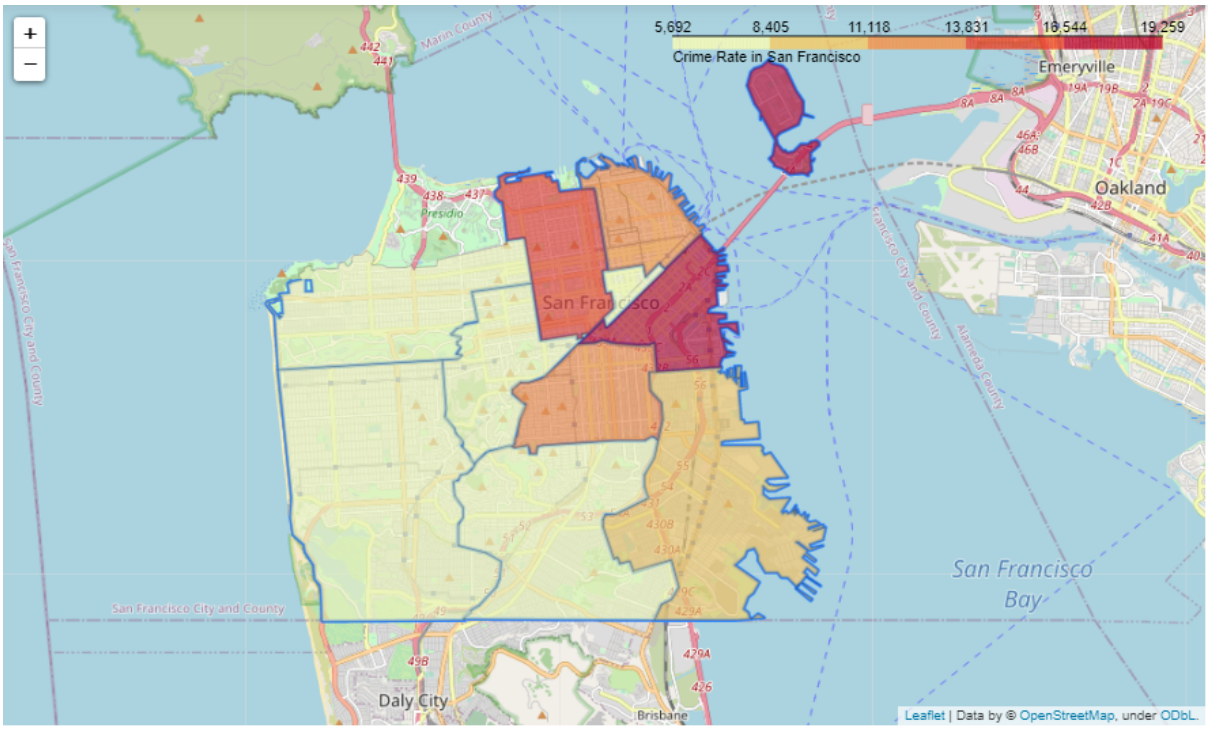
The crime dataset is cleansed next. We drop any incidents that are not publically impactful, e.g. suicide, warrants, runaway, gambling. We then combine a number of crime categories together for simplicity, e.g. theft, robbery and burglary, etc. are all combined under “Theft/Burglary”, and both forcible and non-forcible sex offenses are combined together. We then recheck the number of observations (crimes) in the dataset following the data wrangling, to ensure there are still sufficient number of instances to conduct a robust analysis.

Once data is cleansed, we group the number of crimes by PD district to ascertain which area has the highest crime rate. The data is copied into a dataframe and sorted in descending order - highest to lowest crime rates. Again, note that this analysis is limited in that we do not consider the area covered by each district and as such figures are not scaled to be crime rates per unit of area.

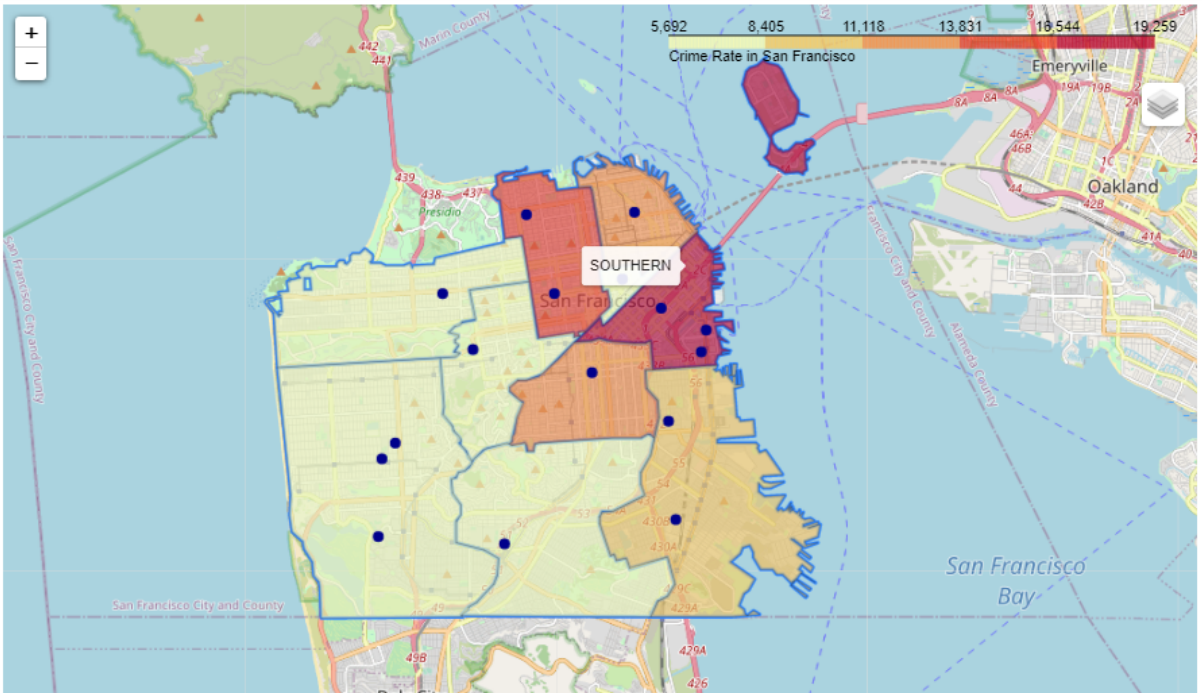
	PdDistrict	No. Incidents
0	SOUTHERN	19258
1	NORTHERN	14425
2	CENTRAL	12404
3	MISSION	12394
4	BAYVIEW	9347
5	INGLESIDE	7578
6	TARAVAL	7187
7	TENDERLOIN	6190
8	RICHMOND	6041
9	PARK	5692

3.3 Part Three: Data Visualisation

In this part of the project we visualise the data using choropleth maps. We use the district polygons from the GeoJson file in combination with the data frame created in the previous part of the analysis showing the number of crimes in each district to graphically depict the safety scale of the San Franciscan districts. We will also plot the police station locations to get a feel for where these are located. We can clearly see from the map which is the most dangerous district by consulting the colour scale (darkest red).



To verify that this matches up with the resulting data frame from the previous step, i.e. that the dark red district is indeed the Southern district, we will add district labels to the map.

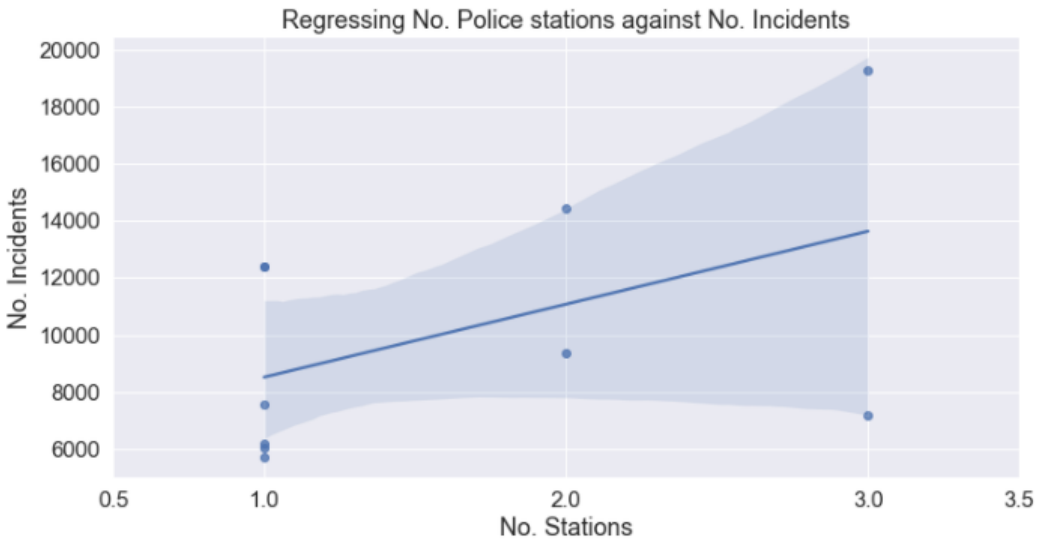


The above map visually confirms that the Southern district has the highest crime rate. We can also see the locations of all the police stations in San Francisco for reference.

3.4 Part Four: Statistical Analysis

In this final part of the project we will run a linear regression, modelling the number of police stations against the number of criminal offenses. In order to run the regression, we first need to add a PD District field to the police stations dataframe to be able to merge with the dataframe showing the number of crimes in each district. The PD District field on the original data from Foursquare (neighbourhood) is missing for all but one police station. We can see from the above map which district each of the stations is in but it would be quite tedious to manually enter the district for every station. Instead we will use the “within()” and “contains()” functions to assess in which district each station is. These functions work by checking whether a set of coordinates is within a particular polygon. Each district is a polygon shape with coordinates for all outer boundaries in the GeoJson file. We will loop through each district and check if any of the police stations are contained within. We are then able to add a district field to the police station dataset, count the number of stations in each district, merge with the crime rate dataframe, and run the regression.

A regression plot will be created which will visually depict the relationship between the number of police stations and the number of criminal instances. The R-Squared will also be calculated which is a numeric statistical measure of the predictiveness of our model, i.e. how well can we predict crime rates based on our knowledge of the number of police stations in a particular area.




```
R-squared:  23.25%
```

4. Results

The analysis allows us to make a number of conclusions on the safety of different areas in San Francisco. The results show us that the Southern district of San Francisco is the most dangerous in terms of the number of criminal occurrences, with 19,258 incidents occurring in 2016. Conversely, the safest area is the Park region with 5,692 incidents occurring (also in 2016). This means that the Southern district has over 3 times the crime rate of the Park region.

The results of the regression plot, and the resulting R-squared, show us that there is a very weak relationship between the number of police stations and crime rates. This means that we cannot rely on our model to accurately predict how the number of police stations in an area might impact on the levels of criminal activity in that same area. However, this weak relationship does not necessarily mean that there is no relationship between the two variables, but is driven by the very narrow range of the dependent variable (number of police stations) which ranges from 1 to 3.

Aside from the weak correlation, the plotted stations on the San Francisco map in Section 3.3 also suggest that establishing police stations does not necessarily deter crime, with the Southern region being the prime example - it has the largest crime rate yet also the most police stations (3) tied with the Taraval region which also has 3 stations and one of the lowest crime rates.

5. Discussion

The results in Section 4, can be used by the San Francisco Police Department to select locations for new police stations. While we saw from the regression that there is no clear link between the number of police stations and the number of criminal incidents in each district, it would be a logical conclusion that areas with higher crime rates could benefit from a greater police presence in the

form of more stations. Therefore the Southern district should be the obvious choice of where to establish a new station.

The results can also be helpful to those looking to move to San Francisco, or those wishing to buy property in the city. When safety is the only factor taken into consideration, and facilities such as schools, gyms, offices, etc. are disregarded, the Southern region would clearly be the least optimal choice of location to move to. This is based on this district having the highest crime rate in all of San Francisco. The Park area would be much more preferable to live in due to having the lowest crime rate. Looking at the map in Section 3.3, it is interesting to note also that western San Francisco has much less crime than East. This may be driven by other factors such as population density, other facilities, etc. but this analysis is not intended to analyse this aspect.

6. Conclusion

In this project, I have looked at the city of San Francisco in the context of criminal activity, with data based on the year 2016. The main conclusions to be drawn from the analysis are that the Southern region is the least safe area of the city, while Park is the safest area; there is no clear relationship between crime rates and police station presence in San Francisco, and that western San Francisco is generally safer, experiencing less crime than the Eastern districts. The results of the analysis would inform 2 recommendations: 1). Those looking to move to the city should look to live nearer the west coast if their primary concern is safety, 2). The San Francisco Police Department should establish more police stations in the Southern district to try and combat the high crime rates in the region.

This analysis could be further enhanced by investigating the relationship between distance from a police station and the number of crimes occurring. The analysis could also benefit from utilising more recent data. Exploring other facilities in or aspects of each of the districts would also be beneficial in addressing the question of which area would be most attractive to live in.