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

Coursera Capstone Project – by Theresa Hughes

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.

Data

The analysis will leverage data on crime in San Francisco as provided during the Coursera Python course. 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 ad locations (coordinates) of all police stations which will allow us to map them. 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.