Crime in low socio-economic suburbs

Domain and Question

For this project, I have chosen to look at the community. The report will be seeking to answer the following question:

Does the socio-economic status of a suburb affect its crime rate?

Whilst this will be the main question I seek to answer, I will explore the following sub-questions to add further depth to my investigation:

- Is there a specific attribute of socio-economic status that could be targeted to help lower crime?
- Is there a pattern crime averse home owners can use to pick low crime neighbourhoods?

Datasets

To aid my analysis, I have chosen two data sets:

Criminal Incidents Visualisation- year ending December 2017 (csv) This dataset contains the criminal incidents examined by the principal offence, region, suburb and type of incident. The dataset was produced by Crime Statistics Agency and can be downloaded as a xslx file here: https://www.crimestatistics.vic.gov.au/crime-statistics/latest-crime-data/download-data-6.

Table 1.1 *Schema of the reduced socio-economic status dataset (key attributes)* $Dimension: 13692 \times 6$

Suburb	*Rank(within Victoria)*	*State*	*Usual Population*	*Minimum resource score*	*Minimum education and income score*
Categorical 14.67% noise (incorrect format)	Continuous 2.09% noise 0.057% outliers Values range between 516-1180	Categorical	Continuous 5.8% noise 5.17% outliers Values range between 11-5074	Continuous 2.109% noise 0.0% outliers 508-1188	Continuous 1.72% noise 0.0% outliers 648-1171

Socio-Economic Indexes for Australia 2016 (csv) This dataset contains Australian suburbs' socio-economic ranking/score. This dataset was produced by the Australian Bureau of Statistics. The dataset can be downloaded as a xls file from: http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/2033.0.55.0012016?OpenDocument

Year	*Suburbs*	*Incidents*
Continuous Value range between 2008-2017	Categorical	Continuous 4.77% noise 8.11% outliers Values range between 1-15575

Table 1.2 *Schema of the reduced crime dataset (key attributes)* $Dimension: 1339 \times 3$

Pre-processing

The initial processing involved getting both datasets into a uniform size and format such that the two could be integrated. For simplicity, I will refer to the Criminal Incidents Visualisation- year ending December 2017 data set as *crime* and the Socio-Economic Indexes for Australia 2016 dataset as *socio-economic status*.

As both files were initially not CSV files, they were exported as CSV files on Excel. Data columns that were not necessary for the analysis were also deleted in Excel resulting in the schemas in Table 1.1/1.2.

As the socio-economic dataset had multiple excel sheets, I had to download each sheet as a separate CSV file and integrate each feature into the socio-economic dataset. As there were some missing suburbs in some of the features, I chose to delete missing suburbs to ensure accuracy of data was retained. Once I created an integrated csv file with all the socio-economic status data, I commenced the data cleaning stage of pre-processing.

The datasets then had to be cleaned using python code. The *socio-economic statu*s dataset contained noisy data with values in an incorrect form. To ensure consistency, values with unnecessary punctuation were stripped, numeric values were also converted to floats and suburb names for the socio-economic dataset were converted into uppercase objects. A similar process was done for the crime dataset but additional processing was required. As the socio-economic status dataset only contained data from 2016, I restricted the crime dataset to only values from 2016. Also, the crime dataset was not grouped by suburbs but rather had a list of different offences per suburb. This meant I had to group the dataset by suburb and then accumulate the total number of offences per suburb. I next calculated the common set between both dataset's suburbs was calculated and each dataset was reduced to only contain the common set of suburbs. Thus, majority of the pre-processing involved stripping values and type casting, as well as using the method **case deletion** (missing data removed which could cause bias but because dataset is large, bias minimised). Upon completion of my pre-processing, I had data on a total of 1048 suburbs.

Integration

The data was now in a format that allowed it to be integrated, and the two datasets were merged and written to a csv file, namely *integrated.csv*. As the project analysis requires examining the suburbs as the independent variable to produce a meaningful result, I went through the suburbs in the new data frame, and calculated the latitude and longitude for all suburbs using the *geopy* library, so that I could obtain a geographic map later which examines the crime and socio-economic status correlation. All suburbs that could not generate a longitude and latitude were deleted, resulting a data frame of 1043 objects.

Table 1.3 *Schema of the integrated dataset (contains key attributes)*

Crime	*Long_lat*	*Population*	*Suburb*	*Socio-economic Ranking*	*Resource*	*Education and Income*
Contains the number of incidents per suburb	Contains the longitude and latitude of suburbs	Contains the usual resident population in each suburb	Contains the suburb name	Contains the socio-economic ranking of suburbs	Contains the minimum resource score for suburbs (smaller indicating greater disadvantage)	Contains minimum score for education and income. (smaller indicating greater disadvantage)

Results

First, the Pearson Correlation and Mutual information score was calculated for the integrated dataset to determine the feasibility. The final values of **-0.3501** revealed that a higher socio-economic suburb had fewer crimes occurring, as expected. The Pearson Correlation does not determine if there is causality, but determines the correlation, which can be classed as '*moderate'* (*Cohen, 1998*). The normalised mutual information (NMI) value for the two features was 0.728 demonstrates a strong dependence between the two features as it is closer to 1 (closer to 1 means perfect correlation). However, must consider that the discretised bins were user dependent and thus the choice of values to discretise data on will vary the outcome.

Next, there was an analysis of the outliers present in the data using matplotlib. Whilst Table 1.1 and 1.2 show the relative noise and outliers in the each features, the following boxplots visually shows where the outlier values lie: **boxplot detection**. The definite outlier values were noted so that they could be removed if needed. The limitation of this technique is that it assumes a normal distribution and thus, the outliers may not be accurate.

Figure 1.1 *Crime Incidents Boxplot*

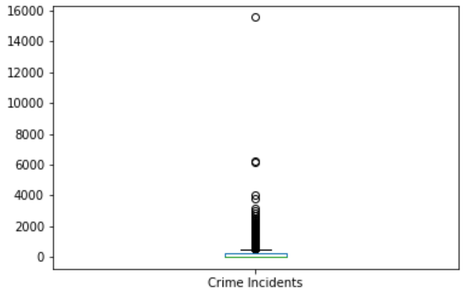
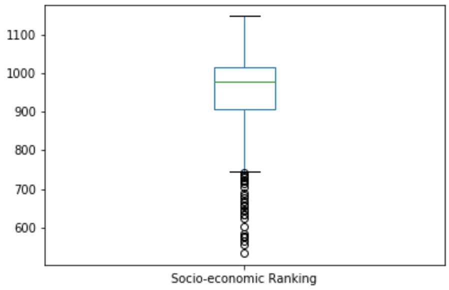
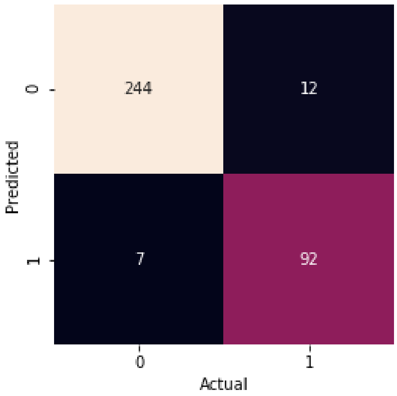


Figure 1.2 *Socio-economic Ranking boxplot*



Next, the K-NN algorithm was used to calculate the accuracy of the crime rate that could be observed in a specific suburb, using the attributes in the integrated dataset. The accuracy score of 0.9608 demonstrates that there is a relatively high accuracy when classifying an unknown suburb using 5 neighbours. Figure 1.3 shows the number of correct classifications, and whilst false positives and negatives exist, they have been minimised. However, the limitation is that the algorithm is **application dependent** and the value of k will affect the performance; k=5 may be too small and thus sensitive to noise.

Figure 1.3 *Confusion matrix*



The scatterplot and regression line graphs were produced in python, using the sklearn library along with matplotlib library.

Figure 1.4 *Socio-economic ranking vs Number of Crime Incidents*

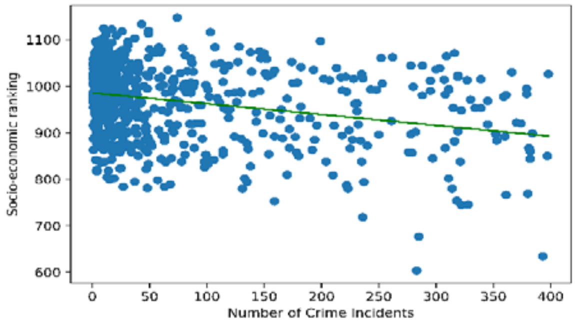


Figure 1.5 *Minimum resource score vs Number of Crime Incidents*

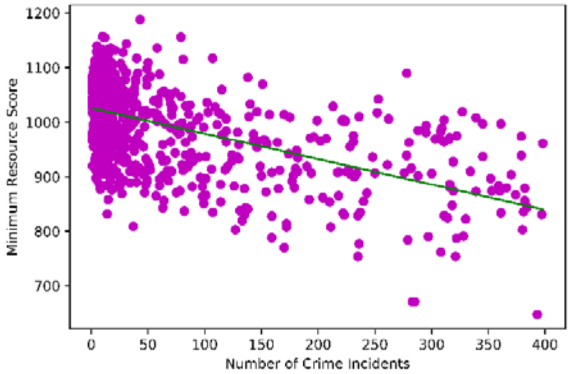


Figure 1.4 shows the correlation between the socio-economic rank and the number of crime incidents. Outliers were removed from this graph as the resulting graph with outliers greatly affected the regression line.Overall, most of the data points are clumped together near the regression line, however there is noticeably a large clump of data with few crime incidents but varying rankings. To analyse the effect of socio-economic state of a suburb and its crime in greater detail, we must analyse the factors that contribute to a suburbs 'socio-economic' status.

To examine further the potential factors that are affecting the socio-economic ranking, we examine Figure 1.5. Here, we can see a stronger linear correlation between the availability of resources and crime. Therefore, crime could potentially be tackled through providing more resources in areas with high crime rates.

Figure 1.6 *Minimum Education and Income Score vs Number of Crime Incidents*

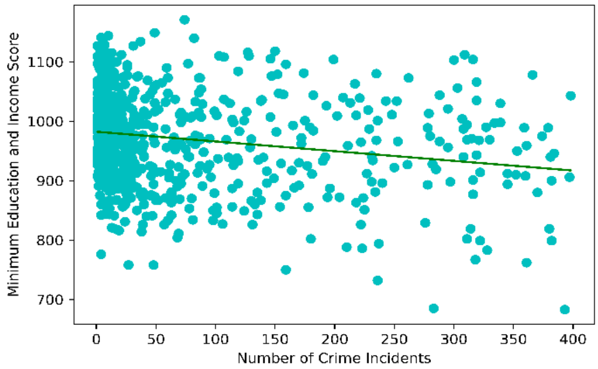


Figure 1.6 shows a weaker correlation between the minimum education and income of suburbs in Melbourne in relation to the number of crime incidents. Thus, this could demonstrate that approaches to lower crime through focusing on educating select suburbs may not necessarily be the most effective option.

Figure 1.7 *Parallel Coordinates*

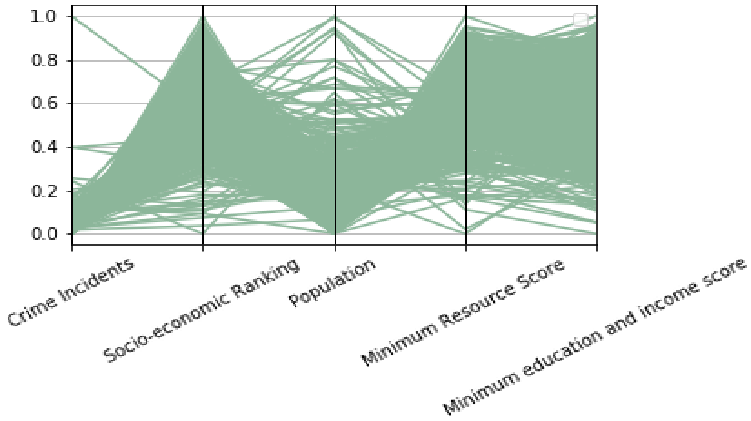


Figure 1.6 shows the distinct correlation and behaviour of each attribute, without outlier removal. Thus, we can observe a broad trend between the different attributes. The crime incidents and population are positively correlated, whilst the socio-economic ranking and minimum resource score are also positively correlated. The minimum education and income score does not reveal an obvious trend that could support the correlation between crime and socio-economic status.

The geographic heatmaps were produced to give a greater indication of whether there was a general pattern that could be observed from looking at each suburb individually. Figure 1.8 displays a random distribution of socio-economic rankings that do not fit any pattern when looking at it geographically. Figure 1.9 shows that crime tends to be higher in more metropolitan suburbs overall.

Figure 1.8 *Socio-Economic Ranking Geographic Heatmap*

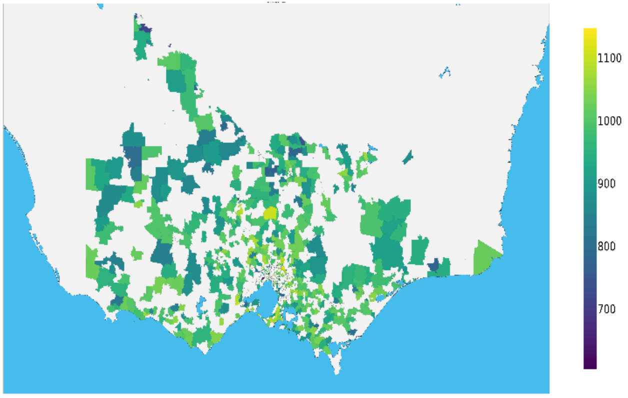
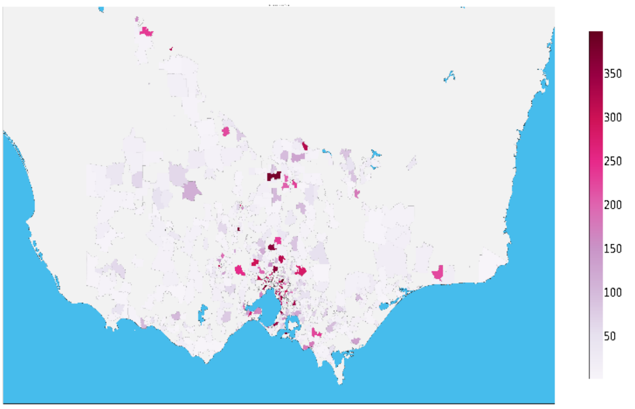


Figure 1.9 *Crime Incidents Geographic Heatmap*



The visualisation demonstrates that there is a correlation but there are limitations to our findings: **correlation** found but that does not mean **causality**, false positives (due to corrupt data or external factors that need to be considered and bias arising from deletion of suburbs (standard deviation altered).

Value

The final report that is produced through processing, integrating, analysing and visualising data offers a more significant analysis than using the raw data alone. The initial stage of processing and data cleaning, allowed data to be consistent and outliers detected (Figure 1.1,1.2) to minimise biased visualisations. Through determining what the outliers were, graphs could be produced to better represent the data without being affected by outlier data points.

Integration and visualisation allowed multiple features to viewed simultaneously, to assess potential dependencies. Such visual interpretation and aggregation of various data is not possible with raw data alone. Thus, the pre-processing, integration, analysis and visualisation allowed the data to be represented in an easier format, allowing various interpretations to be made.

Challenges and Reflections

- Question was too obvious initially so added sub-questions to give more depth and meaningfulness to the report. Also chose to look at specific factors which comprise of the *socio-economic status* of a suburb to provide policy makers with more insight
- Should have researched unfamiliar visualisations I wanted to complete prior to pre-processing; did not need to use geopy library to find longitude and latitude values as shapefile provided this information
- Dataset was large because analysis done by suburbs, resulting in many visualisation methods not being effective. For example, plotted parallel coordinates initially with legend as suburbs but too many suburbs to fit the legend.

Question Resolution

Upon analysing the data transformations and visualisations, I believe that there is a correlation between socio-economic status and crime, demonstrated through the various visualisations.

In terms of policy making, there was a significant correlation between the minimum resource available score and crime incidents, potentially validating the need for a policy which provides suburbs with high crime rates with more resources. This could potentially help reduce crime rates but further analysis would be required to support such a move. However, the minimum education and income score had very little correlation with the crime rate, suggesting that attempts to improve education in high crime suburbs may not produce a decrease in crime. Therefore, policy makers should order some data experts to carry out further investigations between the resource availability and crime rates in suburbs to determine if such a targeted approach to lowering crime could be effective.

For the general population, the analysis indicated that potential homebuyers may want to consider living in regional parts of Victoria as they tend to have lower crime (Figure 1.9). Those concerned about the availability of resources and more generally the socio-economic status of suburbs could move to suburbs with a lower population, as it tends to have more resources available; Figure 1.7 shows that suburbs with fewer persons have greater resource availability which could be attributed to assumption that there is fewer citizens to share with.

Code

All of the data visualisations and processing were done using python and additional libraries (pandas, numpy, sklearn, geopandas, geopy and matplotlib), but some of the code was adjusted from publically available code online. Where code is not original, it has been correctly attributed.

Bibliography

Cohen, J. [1998]. *Statistical power analysis for the behavioural sciences*.

Shapefile retrieved from:
http://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/1270.0.55.003July%202011