COMP20008: ELEMENTS OF DATA PROCESSING

ASSIGNMENT: PHASE 4

STUDENT NAME: TANISHA WIJESINGHE

NUMBER OF PAGES : 8

RISING CRIME AND EDUCATION

2. Domain

The two domains of this project are education and community based. The focus of this report is to investigate how rising education levels impact criminal offence rates and the community as a whole.

3. Question

This report aims to answer the following question:
"Could increased spending on education be used as an effective tool to fight the current offence rates in Melbourne?"

Victoria has experienced a steady rise in offence rates from 2011 to the year ending 31st December 2015 (Crime Statistics Agency, 2015). For investigation purposes, this project will examine the offence rates from 2015 and compare them with data on Victorian schools in the same year. If this report finds a negative correlation between the number of Government schools and the offence rate in each Local Government Area, the Victorian Government could use this information to help reduce offence rates in Melbourne.

4. Datasets

The first of the two chosen datasets (named 'all_school.csv') listed all schools as of 07/03/2016 in Victoria by the Department of Education and Training. The data is updated annually and is stored in CSV format. Each school name is coupled with the Local Government Area it belongs to and its education sector (Government, Catholic, Independent). It further comes with information such as *Address_Town, Address_State,* and *Address_Postcode.*

Link: https://www.data.vic.gov.au/data/dataset/school-locations

The second dataset (named 'crimebylocationdatatable.csv') contained a detailed description of the offence rates for Victoria for the period of 2011-2015 as reported and recorded by Victoria Police on the LEAP database. It is also available in CSV format. For each year, offences were categorised by Local Government Area and then divided into six (6) main offence divisions (*Crimes against the person, Property and deception offences, drug offences, Public order and security offences, Justice procedures offences, other offences)*. Each division was further divided into subdivisions, which allowed for more thorough analysis.

Link:http://www.crimestatistics.vic.gov.au/home/crime+statistics/year+ending+31+december+2015/download+data

Keeping pre-processing in mind, these datasets were chosen after carefully assessing their quality and suitability for the project. Accuracy and credibility of the datasets were ensured as the datasets were both sourced from reputable websites. The datasets were then evaluated on completeness, concluding that they contained no missing, incomplete or disguised values. Available time periods were also taken into consideration, ensuring that both datasets contained information relevant to the year 2015 and were updated regularly. The fact that both datasets were in CSV format and their first rows contained the column names facilitated interpretation. Finally, to enable smooth integration, we chose datasets that could be categorised by the LGAs (Local Government Areas) in Victoria.

5. Pre-processing

In preparation for data mining, we first extracted both datasets and then carefully screened them to ensure consistency. Initially, Data cleaning was carried out. Both CSV files were opened in Excel and cleared of any formatting used on them for beautification (colors, headers, number formatting etc.

were all removed). Column names were edited to ensure no spaces between words ('_'s were substituted). Next, columns were renamed to ensure consistency of data (eg: '*LGA name*' in the education CSV was renamed '*Local_Government_Area'* to be consistent with the crime data set).

Data reduction was undertaken to reduce noise and extract features that contained only useful information. Irrelevant data columns were deleted using Excel. This resulted in easier visualisation and faster analysis of the data. From '*all-schools-list.csv*': *Entity_Type, School_No, School_Status, Address_Line_1, Address_Line_2, Address_Town, Address_State, Address_Postcode, Postal_Address_Line_1, Postal_Address_Line_2, Postal_Town, Postal_State, Postal_Postcode, Full_Phone_No, LGA_ID*. From '*crimebylocationdatatable.csv*': *Police Region, Police Service Area*. All instances that fell under the LGA '*unincorporated vic*', were deleted in '*all-schools-list.csv*' so to ensure that both datasets had the same unique Local Government Area names.

Data transformation was then carried out using python functions to ensure smooth integration.
Since integration was to be done on the 'local_government_area' column, it was necessary that the LGA data in both data sets were uniform. This was done using a CSV reader and then a CSV writer. The column data was edited using functions that stripped unnecessary characters and converted the rest to lowercase. Finally, the same input files were overwritten with the modified data.

Visualisations were created in order to analyse the data and detect patterns within it. All Graphs were plotted using a combination of Python functions, the Pandas data frame, Numpy and Matplotlib. The data presented in the tables below were also found using python functions. Data transformation was first done with the help of Boolean indexing in Pandas to filter the relevant data needed for mapping each graph. Then depending on the type of graph chosen, the list values were passed into Matplotlib in the correct order and formatted according to taste. The main issue encountered was that it was difficult to compare information between different LGAs and years as population growth and size masked the findings. Therefore, graphs were mapped again, taking population growth into consideration. This was done using the relevant estimated resident population for the year, as provided by the Australian Bureau of Statistics (ERP was included in 'crimebylocationdatatable,csv'). This provided a clear and unbiased analysis of the real changes in offence rates and concentration of schools.

$$offence\ rate = \frac{(offence\ count)}{ERP} * 100{,}000$$

6. Integration

Integration was completed without complications as thorough preprocessing had been carried out in preparation for it. After opening both datasets using the Pandas data frame, a list of the unique LGA's was created to use as an index on both datasets. While iterating through the list of LGAs, the current LGA and Boolean indexing was used to filter the data files and the relevant x and y coordinates were found.

In detail, when plotting the number of schools vs. the total number of crime in each LGA for the year 2015, a 'for loop' would iterate through the list of LGAs, and for each LGA, two functions would be called. The first would use Boolean indexing on 'all_school.csv' and find the schools relevant to the current LGA. Then using the value_counts() method, the number of schools per LGA would be returned and appended to a list of x coordinates. Next, the second function would then be called and using similar logic but using the .sum() method, to return the total amount of crime in that LGA for the year 2015. This value will be appended to a list of y coordinates. At the end of the iterations, the two lists of x and y coordinates were used for visualisation purposes.

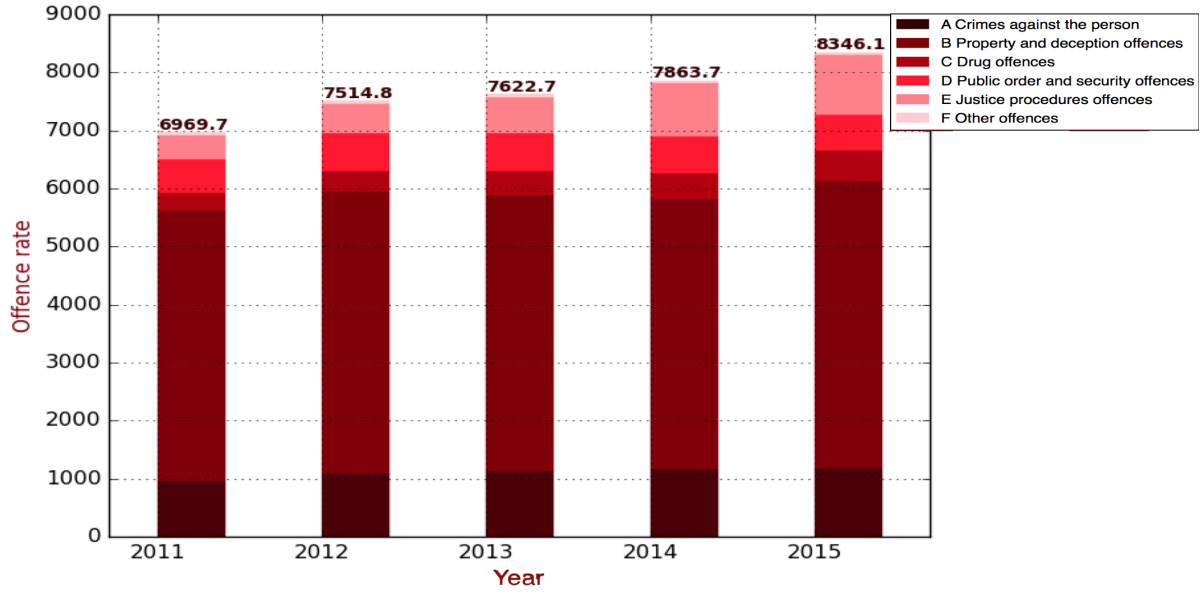
In order to plot the k-means distribution, the x and y coordinates were written to an output CSV file which was later read using a CSV reader in the Notebook containing the K-means code. The code was sourced from the elements of data processing week 5 tutorial.

The limitation of this method was that no new dataset was created by merging the two existing datasets. However, given the complexity and subdivisions in our crime dataset, this allowed for more flexibility and the possibility to delve into more detail at each step.

7. Results

7.1Movement in the rate of offences for Victoria

Year	2011	2012	2013	2014	2015
Offence count	385969	423275	437,493	459,370	496,685
Offence rate per 100,000 people	6969.7	7514.8	7622.7	7863.7	8346.1
% Increase per 100,000 people	-	7.8%	1.4%	3.2%	6.1%
Outcome	RISE	RISE	RISE	RISE	RISE



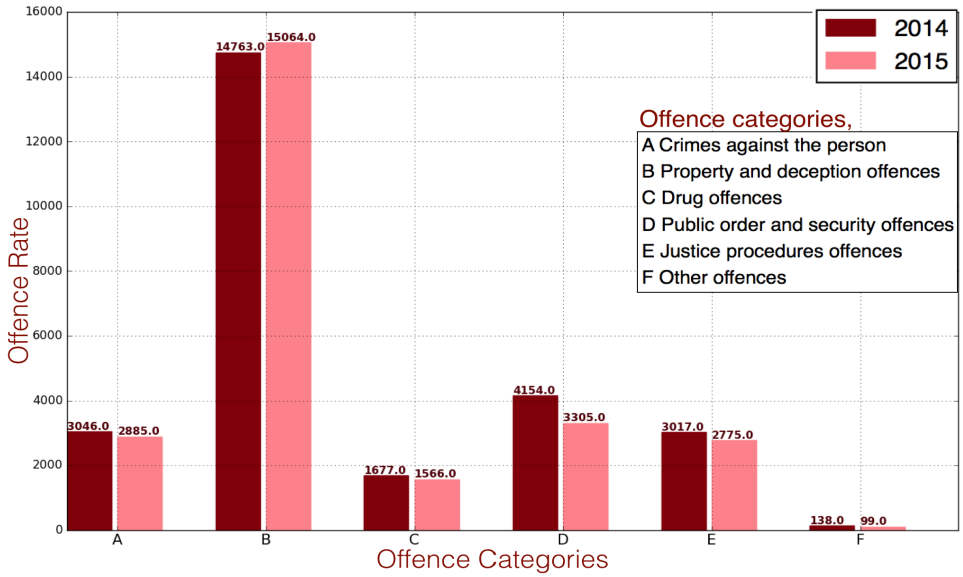
Graph 1: Victoria's total offences for the years 2011-2015 by offence categories.

Graph 1 shows that during the period 2011-2015, Victoria has experienced a rise in its offence rate per 100,000 people. The table provides information on the offence counts in Victoria and the corresponding percentage increase per 100,000 people. Over this time period, offences have risen at an average of 4.6% annually. Given the increase in the offence rate, I believe that this project should draw the Government's attention.

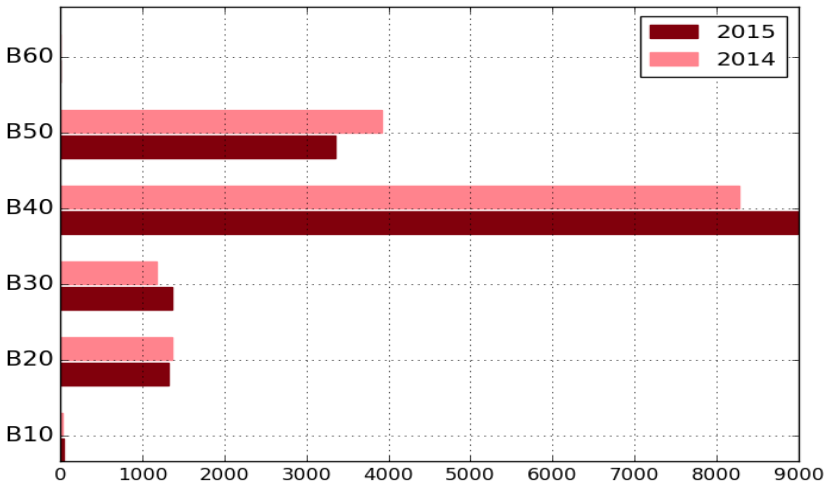
7.2. Movement in the rate of offences for Melbourne

For the year ending December 31st, 2015, the total offence count for Melbourne was 33,334, representing a 1.8% rise from 2014's offence count of 32,747. However after taking population growth into account, Melbourne experienced an overall fall of 4.1% in the offence rates per 100,000 people. This is demonstrated below in the table and Graph 2.

Offence Category	Offense count 2014	Offence count 2015	%Change	%Change per 100,000 people	outcome
A	3723	3743	0.5%	-5.3%	fall
B	18041	19543	8.3%	2%	RISE
C	2050	2031	-0.9%	-6.6%	fall
D	5077	4288	-15.5%	-20.4%	fall
E	3687	3600	-2.4%	-8%	fall
F	169	129	-23.7%	-28.3%	fall



Graph 2: Melbourne's offences for the years 2014 and 2015 by offence category



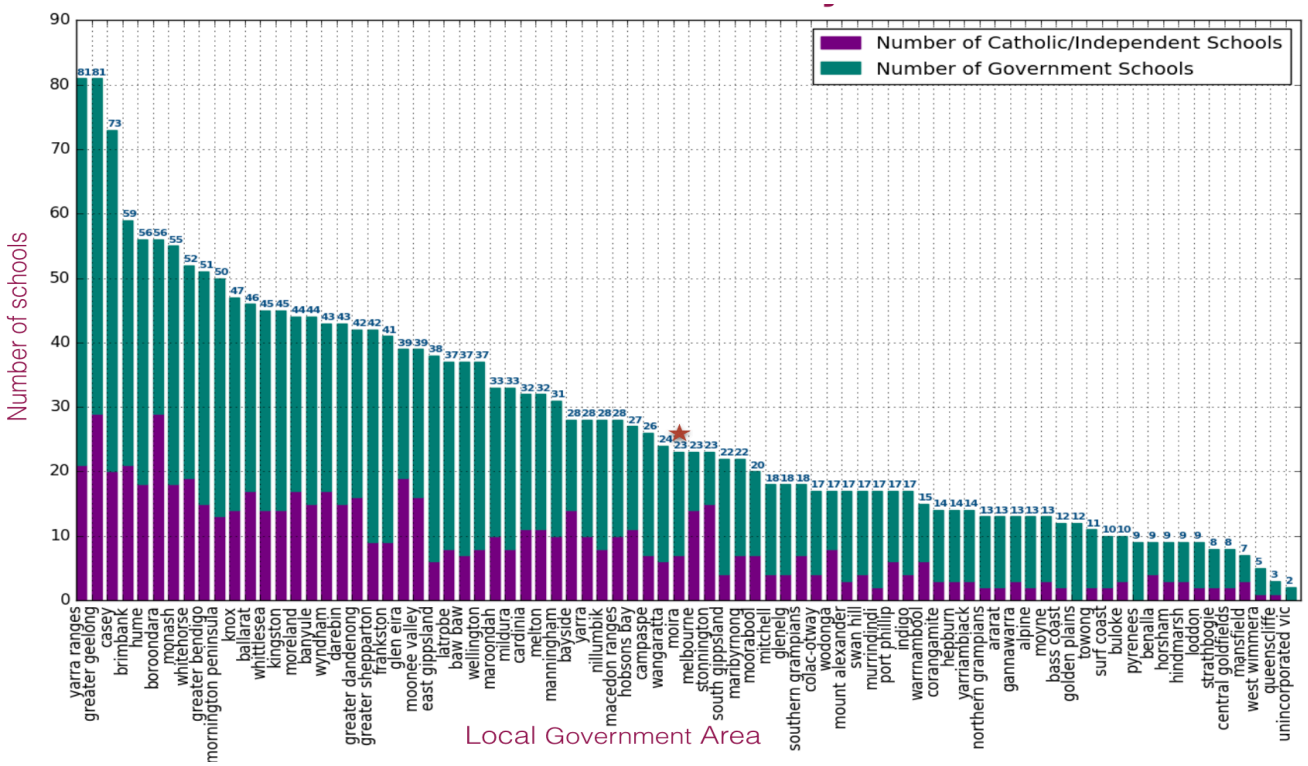
Graph 3: Melbourne's 'Property and Deception offences' further broken down into its offence subcategory for 2014 and 2015

Graph 2 shows that from the year 2014 to 2015, Melbourne experienced a 2% rise in 'property and deception offences'. Graph 3 breaks down the subdivisions of 'B: property and deception offences' for further analysis. It indicates that this rise in property and deception was driven by the rise in subdivisions **B10 (Arson), B30 (Burglary/Break and enter) and B40 (Theft).**

It is important to note that while Melbourne has experienced an overall fall in its offence rate, it is still experiencing a certain level of crime. Concluding that there is still scope for improving the offence rate in Melbourne

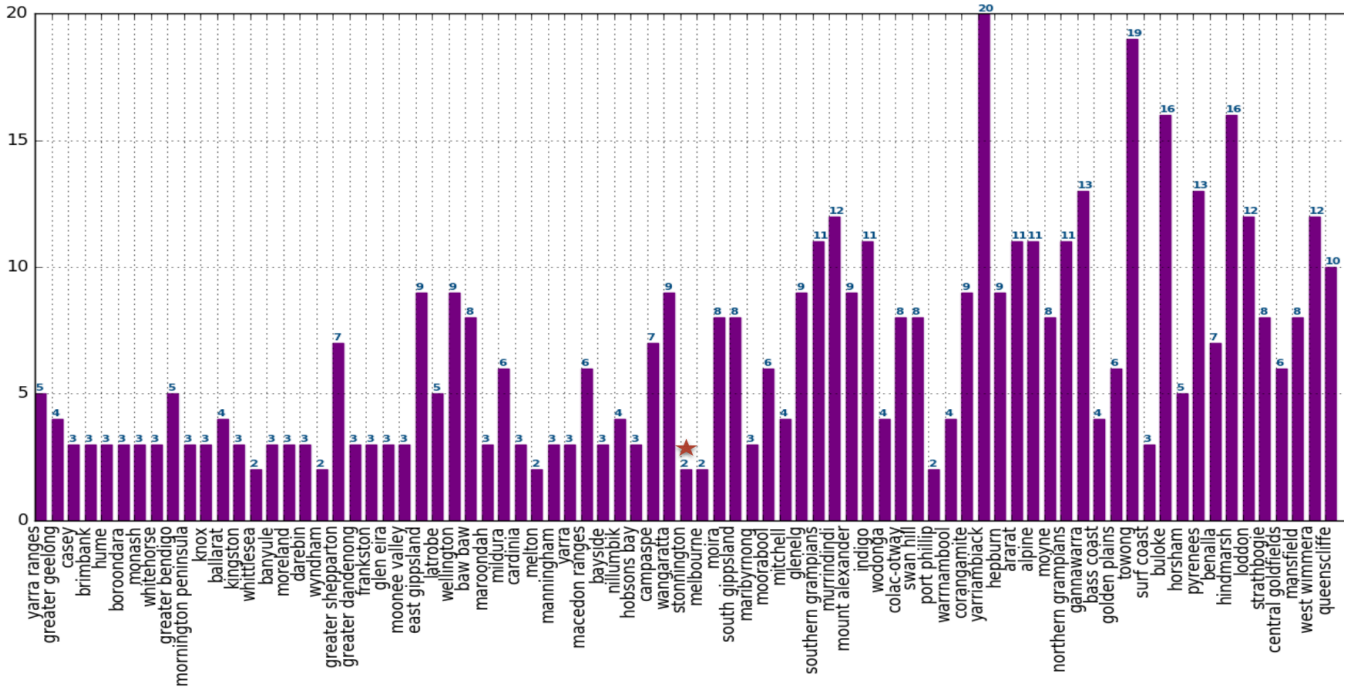
7.3. Distribution of schools in Victoria

LGA	Government schools	Other schools	Total Number of Schools	Schools per 10,000 people
Yarra Ranges	60	21	81	4
Melbourne	**9**	**14**	**23**	**2**
Moira	16	7	23	8.8
Queenscliffe	2	1	3	9.8



Graph 4: A distribution of all LGA's with the corresponding number of schools in each LGA.

Graph 4 shows that Melbourne is located midway in the distribution with a total of 23 schools (Melbourne is indicated by the orange star). In the table above, a comparison of Melbourne with Moira, which has the same number of schools, is shown. It is revealed that Melbourne only has nine (9) Government schools while Moira has sixteen (16). Furthermore, graph 5, which illustrates the concentration of schools per 10,000 people, shows that Melbourne (indicated by the orange star) has one of the lowest ratios of just 2 schools per 100,000 people (while Moira has 8.8).

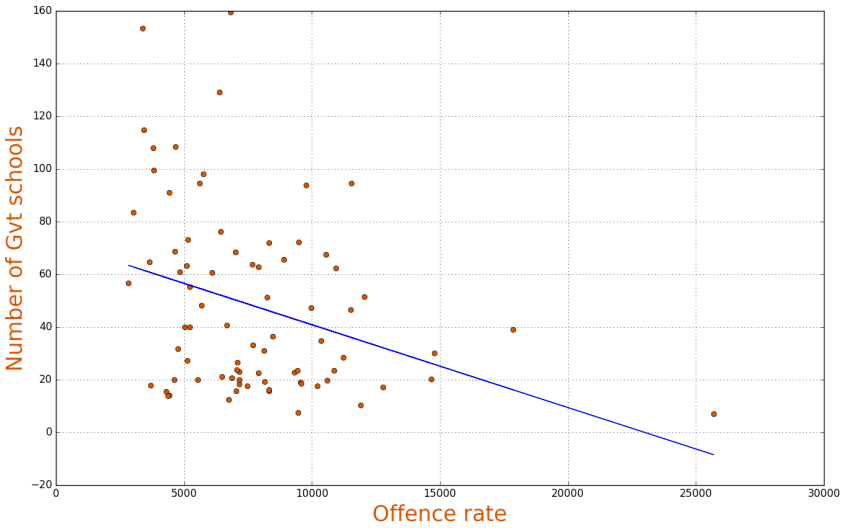


Graph 5: The number of schools per 10,000 people by local Government Area

In conclusion, there is scope in Melbourne for further investment in educational institutions by the Government. However the limitation of this analysis is that there is insufficient data to compare the size of each school, arguably making our conclusion incomplete.

7.4. Scatter plot and line of best fit

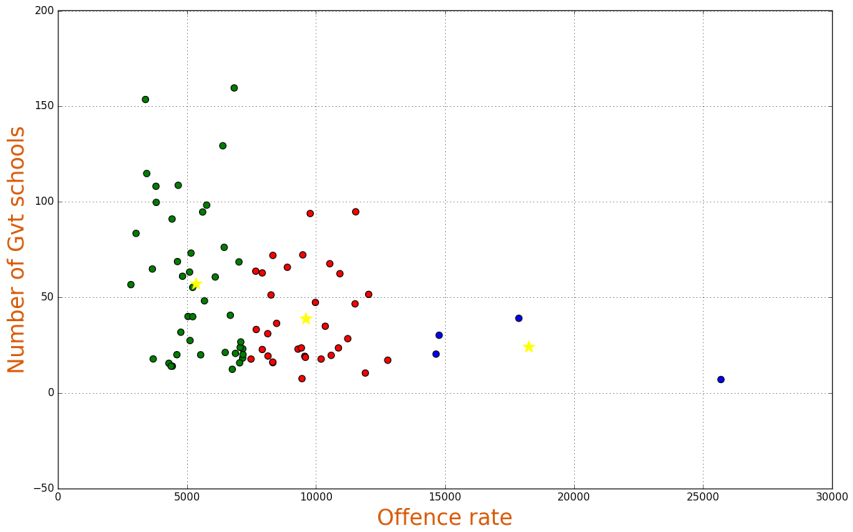
To discover if there is a correlation between the number of Government schools and the offence rate, a scatter plot was generated and a line of best fit was mapped (as seen in Graph 6). Each dot corresponded to one of the LGAs. Population growth was taken into consideration to ensure meaningful results (per 100,000 people). From Graph 6, it can be concluded that there is in fact a negative relationship between the two variables. Therefore, for higher offence rates, the Local Government Area has a lower number of Government schools.



Graph 6: Correlation between Government schools and Offence rates

The Pearson's Correlation Coefficient was calculated using python code to determine the strength of this relationship. It came to be **-0.33,** which translates to a moderate relationship (Cohen, 1988). However, causality is not proven in this analysis. Therefore, while there is sufficient information to prove that there is a correlation, this does not mean that the rising number of government schools is the cause in the drop of the offence rate experienced by the LGA.

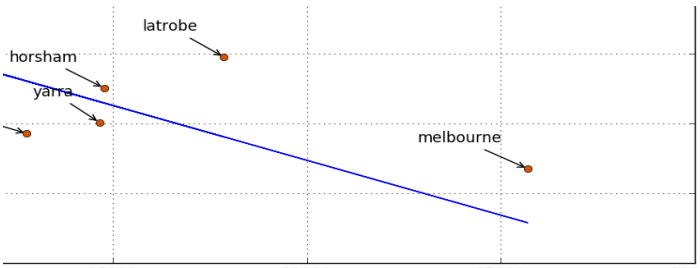
7.5. K-means clustering and outlier detection



Graph 7: Correlation between Government schools and Offence rates

To assist the government in determining which Local Government Areas they should invest more funds in, the K-means algorithm was used to split the merged data into 3 clusters. Graph 7 indicates a blue cluster, which has a relatively high offence rate and a low number of Government schools in comparison to the other clusters (The cluster has the lowest 'center' represented by the yellow star, which is the mean of all the points in the cluster).

Graph 8 helps distinguish which areas are represented by the blue points in Graph 7. As shown below, the blue points represent the Local Government Areas Yarra, Horsham, Latrobe and Melbourne. Melbourne seems to be worse off according to Graph 8.



Graph 8 – Naming of LGAs in the blue cluster

While analysing Graph 7, certain data points can be identified to be within a somewhat proximate distance from the cluster centeroids, while some points are a long distance away. These distant points could be considered outliers. It is interesting that Melbourne is the most likely outlier in the blue cluster as it lies an abnormal distance away from the rest of the data points. Elimination of this point would be unwise though as it is the Local Government Area that is being critically analysed in this project. It is not, however, too surprising that Melbourne constitutes an outlier given the fact that it is the CBD, it would have a relatively high wealth accumulation and a high offence rate.

In conclusion, there is sufficient evidence to prove that there is a correlation between the number of Government schools in a Local Government Area and its corresponding offence rate. Further evidence was found to indicate scope for improvement in Melbourne, as it appears to be the area with the highest offence rate and the lowest number of Government schools in the whole of Victoria.

8. Value

While the raw data chosen holds information relevant to this project, on its own it has no real value. The task of choosing the most adequate datasets for meaningful statistical tests, that would also allow to be merged in the process, involved very detailed consideration. Later, pre-processing this data was the most important and time-consuming step of this project and added the most value to the raw data. Thorough pre-processing gave deep insight into the nature of the data sets and allowed for more suited/interesting questions to be asked. Each step prepared and transformed the data to ensure that later analysis did not produce false or misleading information. A combination of computer and human inspection was employed.

Visualisation of graphs rather than presenting plain numbers allowed for the absorbance of information in a new, more constructive way. These graphical illustrations made it easier to find patterns and relationships and gain knowledge. This led to more insightful conclusions. Trends seen in graphs made it possible to make predictions and be proactive in dealing with the rise in the offence rate. For example, Graph 1, which shows how the offence rate in Victoria changes from the years 2011 – 2015, indicates an upward trend in the offence rate, which implies that future years are likely going to experience a similar rise. Finally, successfully integrating the datasets enabled us to find relationships between variables that had not been found before. Through integration, it was possible to find an influencing factor that could be contributing towards the upward trend in offences.

9. Challenges and reflections

Finding suitable datasets for this project was the most challenging step. Many days were taken to finalise two datasets that held enough relevant information to answer the question this project posed. It was also important to ensure that these datasets could be merged to form one dataset later in the project. This meant they needed to be categorised in the same fashion. The initial stages of preprocessing went smoothly. Learning how to implement the high level libraries was somewhat

challenging. Ensuring that the data was in the right format before parsing it into Matplotlib for visualisations and then creating appealing graphs, was time consuming but straightforward.

The main challenge faced was when comparing data across years and LGAs. It was difficult to gauge the changes accurately as the growth in population across years and the differences in population size was the main reason for the changes in value. Therefore all calculations needed to be revised to allow for accurate comparisons. Merging of the data was also challenging but after a few trials and errors, it was successfully done.

10. Question resolution

There are three main factors that have been analysed in this project that will contribute to answering the initial question. Firstly, Melbourne is currently experiencing a certain level of crime and a rise in the offence rate in the category 'property and deception offences'. Secondly, Melbourne does not have a large concentration of schools to cater for its population size. Also, among the few schools, only 9 are public schools. Lastly, that there is a correlation between the number of Government schools in each LGA and the offence rate in that area. Therefore in relation to the question this paper was trying to answer, we conclude that government spending on education could be used as an effective tool to reduce Melbourne's offence rate.

These results should interest the Government, as they should now consider taking further steps to invest in education. A few things they could investigate are facilitating and improving the current teaching methods in public schools. Future students would benefit from a higher standard of education. This would enlighten the students to make well-informed decisions, leading to students who would have otherwise dropped out of high school, now preferring to further their education. It should also interest future students that these schools would now entail a higher percentage of graduates enrolling into TAFE or Universities. This would allow them to get higher paid jobs in the future. The Government could look into constructing new schooling facilities and expanding the logistics of the existing schools. This will also benefit future students, as it will increase the number of available/cheap schooling options. The work force should be interested that the construction demands from infrastructure expansion as well as the increased staffing needs would result in the creation of new jobs. This could help reduce the unemployment level in Victoria. The Government should ensure the availability of extracurricular activities after school hours to students whose parents are employed full time and are only able to pick them up after work. This, whilst giving the parents a peace of mind, would also ensure that the students are in a safe and peaceful environment

All this, combined with the eventual fall in offence rate would improve the average citizen's standard of living and promote economic growth in Melbourne. In conclusion, the results of this project are of importance to the Government, future and current students, the unemployed workforce and the community as a whole.

11. Code

Around 430 lines of code were written before integration. It is important to note that code was written in separate files for each of the graphs shown below, resulting in some repetition. While it is not the most efficient approach, given the time constraint, this approach was chosen. To merge the data sets and then find a correlation within the new data, around 112 lines of code were written. The code, which carried out the k-means clustering, was sourced from the elements of data processing week 5 tutorial. A rough total of 540 lines of code were written from scratch. All code was in Python and every graph in this project was coded using the major Python libraries, CSV, Pandas, Numpy and Matplotlib. The details of each step taken was written above.

12 .Bibiliography

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillside, NJ: Lawrence Erlbaum Associates.
- Crime Statistics Agency. (March, 2015). http://www.crimestatistics.vic.gov.au/home/crime+statistics/year+ending+31+december+2015/recorded+offences