Maleakhi Agung Wijaya - 784091

EFFECTS OF VOLUNTEERING ACTIVITIES ON CRIME RATE

2. Domain

The domain for this research will be about community, specifically crime and volunteering activities.

3. Question

Crime rate has been increasing for the past few years and it has become a major issue in Victoria (Percy, 2016). This report will tackle this issue and try to answer the following question: **"Could increase in volunteering participation reduce offence/ crime rates?"**.

For analysis purposes, this report will investigate correlation between variables, specifically volunteering participation and offence rates. If we find a negative correlation, this information can further be explored and used by Department of Justice and Regulation Victoria as potential solution to reduce offence rates in Victoria.

4. Datasets

The first of the chosen dataset was taken from AURIN (named "community-data.csv") sourced from Department of Health and Human Services which was released on November 2016. This dataset contains information about percentage of people participation in community (volunteer_ rating_participation, get_help_by_neighbour_percentage, active_community_percentage, attend_community_event_percentage, valued_by_society_percentage) that was categorised based on LGA (lga_code, lga_name).

Link: https://data.aurin.org.au/dataset/vic-govt-dhhs-vic-govt-dhhs-lga-profiles-2015-lga2011

The second dataset was also taken from AURIN (named "offence_data.csv") source from Crime Statistics Agency and which was released on 2017. This dataset contains information about offences in 2012 – 2016 that includes crime type A - F (*Crimes against person, Property and deception offences, drug offences, Public order and security offences, Justice procedures offences, other offences)*, total number of offences (offence_total), and population number (lga_erp) that was categorised based on LGA (lga_code, lga_name).

Link: https://data.aurin.org.au/dataset/vic-govt-csa-lga-vic-crime-statistics-2012-2016-lga2011

5. Pre-Processing

In pre-processing step, our aims are to make both datasets consistent, normalised, and clean. Pre-processing starts by removing unnecessary columns using AURIN filter tools on both datasets. Then, CSV files was read to DataFrame, column names were changed (i.e. lga_code11 -> lga_code) for consistency and smooth integration using inner join.

Afterward, offence_data was transformed so that it displays valuable information about offence rate (normalised_offence_data). Note that before this transformation, offence count (a - f_total, offence_total) is not normalised and therefore biased by number of population in an LGA. To solve the issue, every row was iterated and changed so that the number of offence is changed to a normalised version using formula from Crime Statistics Agency to calculate offence rate / 100,000 people (number of populations in each LGA is given as LGA_ERP).

$$offence\ rate = \frac{number\ of\ crimes}{number\ of\ populations} * 100{,}000$$

From normalised_offence_data, an average of offence rate from 2012 – 2016 was calculated because community_data is taken from 2011 census, which have different time periods. Despite different time periods, government of Victoria – Department of Health and Human Services confirmed the validity of community_data for 2012 – 2016 periods when it was released in 2016.

Lastly, outlier detection was performed on average offence data (average_df) and community_data to eliminate noisy values. Outliers detection was done using a function (outlier_detection) that return lga_name for a data that are outside outer fence. The detection can also be visualised using the box and whisker plot which was shown on Figure 1 and 2.

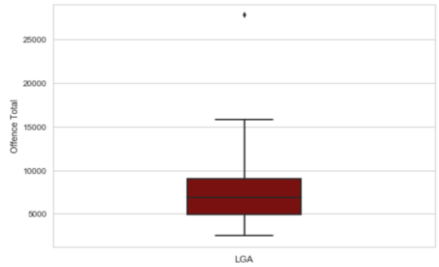


Figure 1.

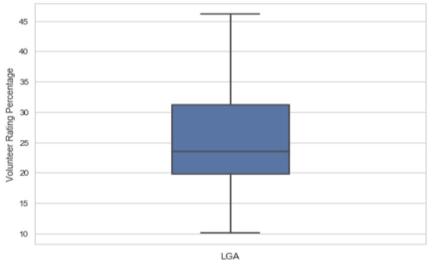


Figure 2.

Based on Figure 1 boxplot, Melbourne was detected as outliers with 27799 offence rate. However, it will be included in later analysis as using domain knowledge Melbourne still lies within normal range. Reason behind high crime rate in Melbourne is due to "… the result of the massive influx of commuters and revellers who visited the city each day" (Fontana, 2017). For the volunteer rating percentage (Figure 2), no outlier was detected, and so all data is clean and will be included in later analysis.

6. Integration

Integration was done to merge offence_data and community_data using its index (lga_code). This was done by df.merge() function using inner join mode. Columns after merging is lga_code, a_total, b_total, c_total, d_total, e_total, f_total, offence_total, lga_name, get_help_by_neighbour _percentage, active_community_percentage, attend_community_event_percentage, valued_by_society_percentage, and volunteer_rating_percentage.

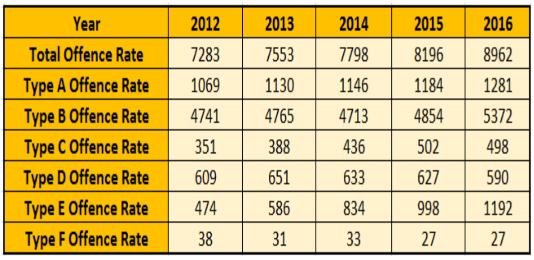
To perform decision tree analysis, another column which consists of discrete values regarding crime level (discrete_crime) will be appended to merged dataframe. First, discrete labels were chosen by inspecting the data and classify low, medium, high crime rate according to offence total on every LGA (offence_total < 5000 = low, 5000 < offence_total < 10,000 = medium, offence_total > 10,000 = high). Then, data was iterated using for loop and discrete value was assigned for every LGA.

After all required columns have been added, duplicated columns such as lga_name was deleted and merged dataframe was saved to CSV file for reusability. Overall, the process went smoothly due to consistency in column name and data cleaning which was done on pre-processing phase.

7. Results

7.1 Offence Rate in Victoria

To depicts the movement of offence rate throughout 2012 – 2016, table (Figure 3) containing information about specific types of crimes (A - F) as well as total crime rates was displayed. In addition, bar plot was also given to help visualising the information.



Year	2012	2013	2014	2015	2016
Total Offence Rate	7283	7553	7798	8196	8962
Type A Offence Rate	1069	1130	1146	1184	1281
Type B Offence Rate	4741	4765	4713	4854	5372
Type C Offence Rate	351	388	436	502	498
Type D Offence Rate	609	651	633	627	590
Type E Offence Rate	474	586	834	998	1192
Type F Offence Rate	38	31	33	27	27

Figure 3.

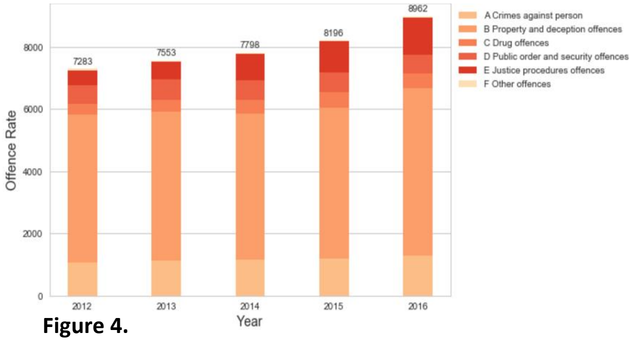
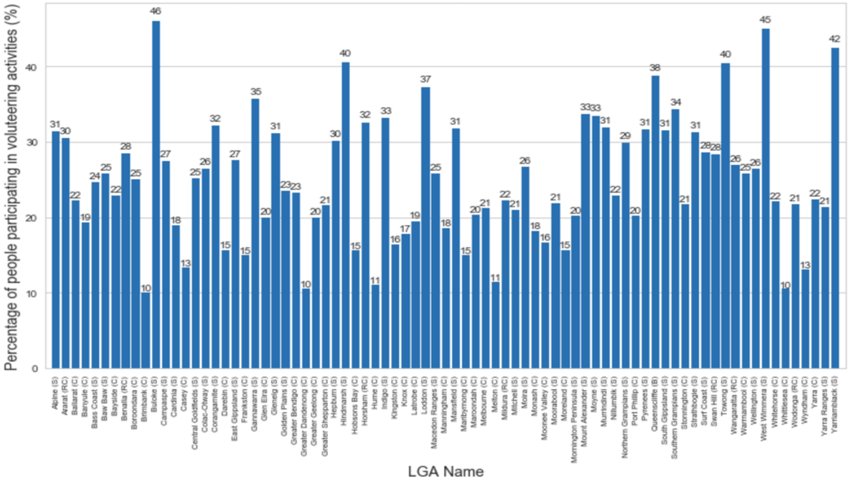


Figure 4.

Table (Figure 3) and bar plot (Figure 4) communicates that some segments (A, C, E) of offence types is increasing while other segments (B, D, F) is fluctuating during the periods. However, total offence rate still consistently increasing throughout the years by about 3-9% every year. This fact shows that overall crime rates has consistently increasing for the last 5 years and therefore a major issue that government and society of Victoria need to be aware of.

7.2 Volunteering Participation in Victoria



Bar plot (Figure 5) shows that participation percentage in every LGA is below 50% with highest participation is at 46% (Buloke). In addition, participation percentage in numerous LGA still falls below 20%. In conclusion, there is room for government and volunteering organisations to encourage participation and increase the number of volunteering events in every LGA.

7.3 Correlation Analysis

To identify if there is correlation between offence rates and number of volunteering participation, scatter plot and line of best fit was plotted (Figure 6). In addition, to strengthen the argument that volunteering participation might reduce crime rate, additional data related to community was added. Correlation between these data and crime rate was then computed using Pearson Correlation and the result is displayed on the heat map (Figure 7)

Observing Figure 6 and 7, moderately negative correlation with value of -0.32 can be found between volunteering participation and offence rate. Moreover, using others community data, there is also negative correlations between good and active community with offence rate (see Figure 7 for correlation value). Even though correlation does not imply causality, but by taking various community factors, there is a high possibility that increasing community participation will decrease crime rates. Hence, increasing volunteering participation which allows individuals to know and actively participate in their community, might decrease the crime rate experienced by the LGA.

Figure 6.

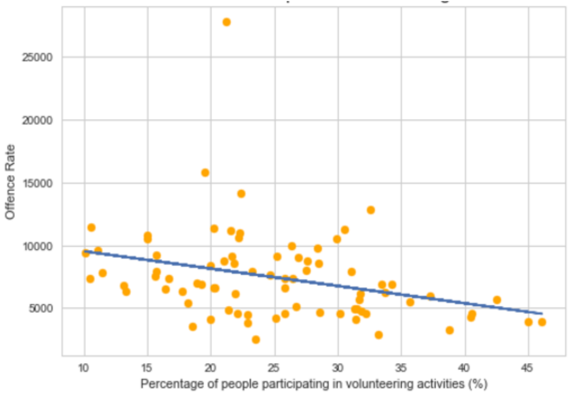
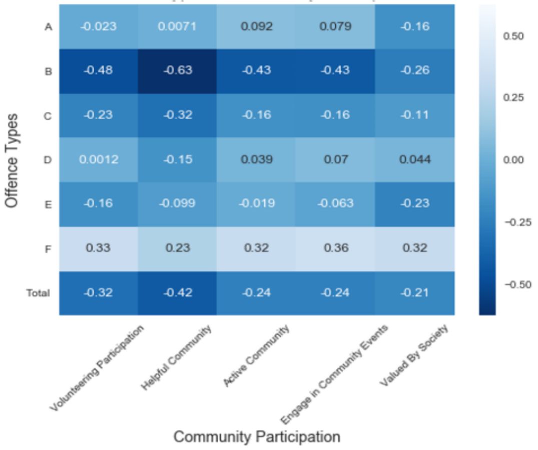
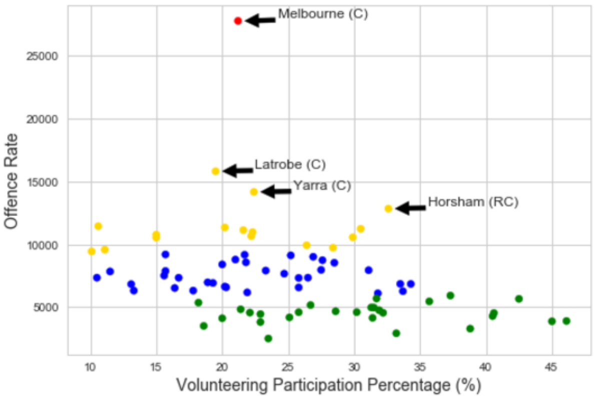


Figure 7.



7.4 K-Means Clustering



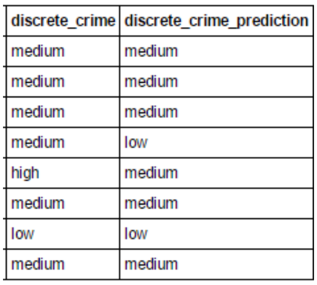
K-Means Clustering was performed to help government and volunteering organisations decide which area they should focus first. Four clusters are used and points were divided to most important and important, medium importance, and lower importance. Inspecting Figure 8, red and yellow coloured clusters is the highest offence rate with low volunteering participation, blue has medium offence rate and low volunteering participation, and green has low offence rate with decent volunteering participation percentage. With this clustering method, two information can be drawn. First, real evidence to prove that there is correlation between volunteering participation and crime rate that was depicted by Melbourne (high crime rate and below average volunteering participation). Second, information for government and volunteering organisation on place that they should invest first (Melbourne, Latrobe, Yarra, Horsham which was represented by red and yellow clusters).

7.5 Prediction Analysis

To further explore and maximise use of the data, prediction analysis such as decision tree and linear regression model was created. Decision tree model are used to determine the level of crime rate (low, medium, high) in an area given community data about the area (Figure 9). It was created using Python sklearn library with 75% accuracy. Using this model, government can predict the level of crime in an area given community data and focuses more on area with high crime rate prediction.

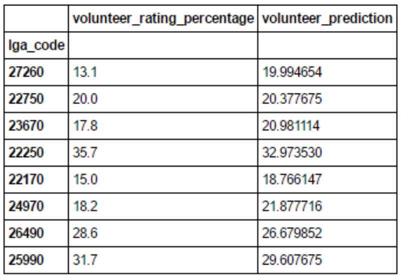
Apart from decision tree model, linear regression model was also used to predict volunteer rating participation given crime data about an area (Figure 10). It was also created using the same Python sklearn library with mean square error of 12.63. Using the linear regression model, government and volunteer organisations can predict the percentage of people participating in volunteering activities and decide whether to organise more events in the area.

Accuracy is the main issues and limitations for both prediction models. Both models have accuracy around 70% which do not have the best prediction capability. This limitation occurs due to lack of availability of the data and can be resolved by fitting the model with more data.



discrete_crime	discrete_crime_prediction
medium	medium
medium	medium
medium	medium
medium	low
high	medium
medium	medium
low	low
medium	medium

Figure 9.



	volunteer_rating_percentage	volunteer_prediction
lga_code		
27260	13.1	19.994654
22750	20.0	20.377675
23670	17.8	20.981114
22250	35.7	32.973530
22170	15.0	18.766147
24970	18.2	21.877716
26490	28.6	26.679852
25990	31.7	29.607675

Figure 10.

8. Values

Pre-processing, integration, analysis and visualising data add values compared to having raw data. Pre-processing add values as it helps to identify the nature of the data. This information is then used to construct research questions and deciding what analysis to perform. In addition, with proper pre-processing noisy data can be detected and removed thus leaving accurate and consistent data for further analysis. As for integration, it can be used to capture relationship between different datasets that hasn't been discovered previously. Moreover, with integrated data, analysis and visualisation can be done easily.

Analysis and visualisation add values by making it easier for human eyes to spot pattern and interpreting large information. As an example, bar chart can be used to display the increasing rate of crime, heat map to display Pearson Correlation performed on community data and crime rates, as well as scatter plot and best fit that are used to visualise correlation.

9. Challenges and Reflections

Finding datasets was the most challenging part of the project. Problem occurs as not many datasets are credible and support the research question. Furthermore, I need to find datasets that has the same primary key (LGA) for merging. The initial phase of pre-processing and integration went smoothly as it has been planned when choosing datasets. Choosing which analysis method to perform was also hard because I need to make sure that the analysis provides some insight and help to answer the question that are proposed by the project. As for the learning process, it was done smoothly by reading documentations and forums for matplotlib, sklearn, pandas, and seaborn.

10. Question Resolution

Victoria is currently experiencing an increase in crime rates for about 3-9% per year for the past 5 years. As for volunteering participation, it was below 50% on every LGA in Victoria. Therefore, crime is a major problem in Victoria and as there are room for improvement in volunteering participation, it will be interesting if we can reduce crime through volunteering. Result has showed that there is a moderately negative correlation between crime rates and volunteering participation. In addition, if we broaden the scopes, it was also shown that by comparing numerous community participation with crime, we can see that all of it is negative correlated with crime. Hence, there is high probability that volunteering participation can reduce crime rate since volunteering encourages individuals to connect and participate in their community. Concrete example of Melbourne was also observed as it is a location with highest crime rate that has relatively low volunteering participation. With this, we can say with confident that volunteering participation is an effective tool to reduce crime rate.

Because of this interesting finding, clustering analysis and prediction analysis was done to help government (Department of Justice and Regulation Victoria, Department of Human Services Victoria) and volunteering organisations which will be interested on the result. Clustering and predictions analysis were used to assist them to find location which they should focus more as there are high crime rate with low volunteering participation (i.e Melbourne).

11. Code

The code for this project were all written from scratch. Around 100 lines of code were written for pre-processing and integration phase, which includes cleaning, merging, and outlier analysis. As for visualisation and analysis around 300 lines of codes were written. These includes bar plot, heat map, correlation analysis, K Means Clustering, decision tree, and regression. As for the library, I used pandas and numpy for data structure, matplotlib and seaborn for plotting, sklearn for K Means Clustering, decision tree, and regression.

12. Bibliography

- http://www.theage.com.au/victoria/inner-melbourne-worst-for-crime-20110920-1kjm8.html
- http://www.abc.net.au/news/2016-09-29/victorian-crime-rate-spikes-as-opposition-warns-of-crime-tsunami/7890832
- https://www.crimestatistics.vic.gov.au/