Data Exploration Project - FIT5147

ABSTRACT

This report examines crime incidents in Victoria, Australia, over a ten-year period and investigates potential relationships between crime rates and socio-economic indicators. Spatial visualizations identify areas with high crime rates, while correlation analyses reveal weak negative associations with socio-economic factors. The findings underscore the complexity of crime dynamics and emphasize the need for further research. Despite data limitations, this analysis provides valuable insights for policymakers and researchers working to address crime-related challenges in Victoria.

Hoang Do

Table of Contents

Introduction

Melbourne, hailed as the world's third most liveable city, recently eclipsed Sydney, which sits at the fourth position on the global liveability scale (Glynn, 2023). However, beneath the glamorous veneer of this vibrant metropolis lies a disconcerting reality. Melburnians find themselves grappling with an escalating criminal phenomenon, particularly involving the city's youth. Incidents are not only on the rise but have taken on a more perilous and ubiquitous nature. As reported by Davis (2023), the year leading up to March witnessed an alarming surge in youthful criminal activity, with a staggering 44.6 percent increase in cases involving offenders aged between 10 and 14. Notably, this age group accounted for a staggering 6,418 incidents, while the number of burglaries committed by this demographic soared by an astonishing 86.7 percent. This worrisome trend poses a heightened concern for international students, who are identified as being particularly vulnerable to such criminal activities. Consequently, as an international student myself, this project aims to comprehensively analyse crime rates and incidents across various categories, including geographical regions, local government areas, and postcodes. The overarching goal is to raise awareness among newly arrived international students about potential risks associated with different suburbs before they decide to rent accommodation. Furthermore, this research will delve into the potential influence of socio-economic factors on crime rates. Nonetheless, it is crucial to acknowledge the constraints imposed by data limitations, such as the availability of data collected only over a five-year period or data that has not been updated for the year 2023. Consequently, this analysis will provide a snapshot of the existing relationship rather than a longitudinal assessment. In essence, by performing a comprehensive and in-depth analysis, this report seeks to address two critical questions:

* Are there some areas in Victoria more dangerous than the other?
* Are there any socio-economic indicators associated with the highly dangerous areas?

Data Wrangling

Criminal Incidents Data

The first and main data for this project will be criminal incidents data fetched from Crime Statistics Agency.

The Crime Data comprises five distinct sheets, each offering a unique perspective on criminal activities spanning from April 2013 to March 2023. These sheets meticulously document criminal incidents and rates per 100,000 population across various geographic and categorical dimensions. Sheet 1 meticulously catalogues criminal incidents and associated rates within police regions and local government areas. Meanwhile, Sheet 2 delves deeper into the data, scrutinizing the statistics based on principal offenses, local government areas, and police service areas. Sheet 3 shifts its focus to granular details, reporting criminal incidents based on principal offenses, local government areas, and even postcodes or suburb/town classifications. This comprehensive dataset serves as a valuable resource for conducting in-depth analyses and generating insights into crime patterns and trends over the past decade.

Criminal incidents and rate per 100,000 population by police region and local government area - April 2013 to March 2023

For Sheet 1, the data wrangling process involves addressing specific issues to prepare the dataset for subsequent analysis. Firstly, we identify missing values in the "Rate per 100,000 population" variable, which occur in two distinct locations: Justice Institutions and Immigration Facilities (includes, prisons, youth justice facilities and immigration detention centres) and Unincorporated Vic (small islands (administered by the state) and ski resorts (administered by management boards) which are unincorporated. These locations have populations of less than 100,000, rendering the calculation of rates impractical. Consequently, we opt to remove observations pertaining to these two areas, as they are not pertinent to our analysis of the general resident and study groups. This is accomplished using the na.omit() function.

Secondly, we identify that the "Year Ending" variable consistently holds the value "March" offering no additional information beyond signalling the data collection and publication timing, which is uniform across all records. Therefore, it is deemed redundant for our analysis and subsequently removed.

Lastly, it comes to our attention that total figures for each police region within each year are recorded as individual observations within the dataset. These records, which denote aggregate statistics, are not aligned with the granularity of our analysis and, therefore, are excluded to maintain the integrity and relevance of the dataset.

Criminal incidents and rate per 100,000 population by principal offence, local government area and police service area - April 2013 to March 2023

For sheet 2, the data wrangling process encompasses several essential steps to streamline the dataset for our specific analytical goals. First, we eliminate the "Year Ending" variable for reasons akin to those previously mentioned—its uniform value of "March" throughout the dataset adds no meaningful information and is therefore removed.

Furthermore, recognizing that our analysis primarily centres on criminal rates and offenses within specific local government areas, we opt to exclude the "Police Service Area" and "PSA Rate per 100,000 population" variables, as our focus does not extend to examining these aspects.

Lastly, it is worth noting that the dataset exhibits no missing values, as confirmed by utilizing the sum(is.na()) function. If the function returns 0, it means that there is no missing value. This cleanliness ensures that our dataset is robust and free from data gaps, setting the stage for subsequent in-depth analysis.

Criminal incidents by principal offence, local government area and postcode or suburb/town - April 2013 to March 2023

For sheets 3, the data wrangling process is relatively straightforward, with minimal adjustments required to prepare the datasets for analysis.
In sheet 3, the "Year Ending" variable is removed for reasons similar to those previously discussed. No further data wrangling is deemed necessary, as the dataset does not exhibit missing values, as confirmed by the sum(is.na()) function.

Satisfaction with your standard of living ranked from 0-100 (very dissatisfied to very satisfied)

This data is fetched from University of Canberra and Health Research Institute.
In the process of preparing a complex dataset for analysis, we embark on an extensive journey of data wrangling and clarification. This dataset, comprising diverse variables and facets, demands meticulous attention to detail to extract meaningful insights.

Filtering for Victoria
Our first imperative is to focus exclusively on the state of Victoria, Australia. To achieve this, we employ the filter() function, isolating the relevant subset of data for in-depth analysis.

Acknowledging Data Limitations
It is essential to acknowledge the dataset's inherent limitations. A notable constraint is the relatively modest sample size, with only 3,273 participants partaking in the Regional Wellbeing Survey. Additionally, approximately 90% of these participants reside in regional areas, excluding major metropolitan zones such as Melbourne. This demographic reality is crucial to consider, given that our primary focus group for this analysis, international students, primarily inhabits urban areas.

Age Distribution Matters
Another essential aspect to address is the age distribution within the dataset. A striking majority of individuals—4,920 out of 10,648—are aged 65 and above. Out of these, 4,416 reside in regional areas. Given the unique perspectives of this age group regarding the satisfaction of standard living, it is imperative to recognize that their viewpoints may significantly differ from those of the broader population.

Simplified Data Representation
To simplify our analysis, we opt to utilize average scores, ranging from 0 to 100, representing the spectrum from very dissatisfied to very satisfied. These scores are further rounded to the nearest integer using the round() function.

Removing Unnecessary Rows
Our data refinement process extends to removing extraneous rows that do not align with our analytical objectives. The first three rows, specifically pertaining to whole Victoria, regional areas, and urban regions, are deemed irrelevant and are promptly eliminated using the slice() function from the dplyr package.

Handling Grouped LGAs
Intricacies also arise regarding Local Government Areas (LGAs), some of which are grouped due to geographical proximity and the limited number of participants. To address this, we systematically disentangle these grouped LGAs, treating them as individual entities while maintaining the assumption that the average scores remain consistent.

Cleaned Parentheses and Characters
We meticulously cleanse the dataset by eliminating characters inside parentheses, which denote region types like RDA or grouped LGAs. This step enhances the dataset's clarity and readability, achieved through the gsub("\\([^\\)]+\\)", "") function.

Addressing Duplicate Values
The dataset presents instances of duplicate values for regions with identical average scores but identified in distinct types of regions. These duplications are meticulously removed using the distinct() function.

Separation of Grouped LGAs
Further data wrangling ensues as we substitute "and" with commas and proceed to employ the separate_rows() function to segregate and organize the data. This step facilitates a deeper analysis of regional nuances by calculating the mean average score for each region.

Trimming Trailing Spaces
Post-separation, it is discerned that certain observations harbor superfluous spaces at their conclusion. These extraneous spaces are methodically excised utilizing the trimws() function from the tidyr package.

Absence of Missing Values
In our pursuit of data cleanliness and robustness, we methodically inspect for missing values using the sum(is.na()) function. Deliberate scrutiny confirms that the dataset remains devoid of any missing values.

Accounting for Missing Data
It is essential to recognize that certain regions (LGAs) may exhibit limited participant representation, leading to missing values when integrated with crime data. This is deemed an inherent limitation of the dataset, which we acknowledge as we progress toward comprehensive analysis.

Finally, this data will be merged into other criminal incidents data above for further analysis using the dplyr package's left_join() function.

Index of relative socio-economic disadvantage ranked within state
This data is fetched from Australia Bureau Statistics website.
In the process of preparing socio-economic data for analysis, we embark on a series of meticulous data wrangling steps to ensure that the dataset is conducive to meaningful insights. The dataset, in its raw form, contains various variables, but for our analytical purposes, we will selectively retain only those that are pertinent to our analysis. Specifically, we will retain variables such as LGA (Local Government Area), Score, ranking within State or Territory, and Rank, while discarding any extraneous variables that do not contribute to our objectives. To further streamline our dataset, we narrow our focus to LGAs within the state of Victoria. This geographical restriction aligns with our analytical goals and ensures that our analysis remains regionally relevant. In the original Excel file, column labels are often merged, leading to variable names that may lack clarity. To enhance the dataset's readability and interpretability, we opt to rename the variable names, providing them with more descriptive and distinct labels. One observation, "Unincorporated Vic," is identified as an outlier and is subsequently removed from the dataset. This action is taken to maintain dataset integrity and align with the specific focus of our analysis. Afterwards, a meticulous data cleaning process ensues,

involving the removal of parentheses and any accompanying characters that may obscure the dataset's clarity. The versatile gsub() function is employed to execute this task. Moreover, we observe that the "Score" and "Rank" variables are initially categorized as character class data. To facilitate numerical analyses, we convert these variables into the numeric class using the as.numeric() function. Finally, data integrity remains a paramount concern throughout this process. We conduct a thorough check for missing values within the dataset, employing the vis_dat() function for a comprehensive assessment. Our examination concludes that the dataset is devoid of any missing values, ensuring the robustness of our subsequent analyses.

It is essential to understand that the "Score" variable in this dataset represents socio-economic disadvantage. In this context, a higher score indicates a more pronounced socio-economic disadvantage, and conversely, the highest-ranked LGAs will exhibit the lowest scores. This nuanced interpretation underscores the importance of considering this variable within its specific socio-economic context.

Finally, this data will be merged into other criminal incidents data above for further analysis using the dplyr package's left_join() function.

Index of education and occupation ranked within state

This data is fetched from Australia Bureau Statistics website.
The dataset's provenance from the reputable Australian Bureau of Statistics (ABS), housed within a single Excel file, warrants a standardized data wrangling approach akin to the procedure employed for the Index of Relative Socio-Economic Disadvantage dataset. Subsequently, an exhaustive review for missing values, facilitated through the vis_dat() tool, confirms the dataset's integrity by revealing its absence of missing values.

Finally, this data will be merged into other criminal incidents data above for further analysis using the dplyr package's left_join() function.

It is essential to underscore that within this dataset, lower scores correlate with relative socio-economic advantage, signifying that areas with lower scores enjoy a higher socio-economic status. Consequently, Local Government Areas (LGAs) achieving higher ranks exhibit lower scores, a reflection of their superior socio-economic standing. This nuanced interpretation underscores the pivotal role of the score variable, thereby enhancing its significance in subsequent analyses.

Data Checking

Criminal Incidents Data

In the process of data consistency checking, we employed the 'unique()' function to meticulously scrutinize all character variables within the five sheets, ensuring the absence of misspellings and verifying that the years fell within the predefined dataset range of April 2013 to March 2023. This examination yielded a commendable outcome, with all variables across the five sheets exhibiting consistency.

Turning our attention to the box plot (*figure 1*), a notable discovery emerged during our exploration of the Police Region data. The presence of multiple outliers within the North West Metro and Eastern regions sparked further investigation. Subsequently, when

analysing the Local Government Area (LGA) data *(figure 2)*, three specific LGAs, namely Melbourne, Latrobe, and Yarra, stood out with significantly elevated means compared to their counterparts. Notably, these three LGAs are situated within either the North West Metro or Eastern regions, providing a rationale for the observed outliers.

These findings collectively highlight Melbourne, Latrobe, and Yarra as the top three areas with the highest incidence of criminal activity in Victoria, as measured by the number of criminal incidents per 100,000 population. Furthermore, our examination of the pie chart *(figure 3)* representing the distribution of criminal incidents across various offence categories reveals that a majority, amounting to 60.7%, can be attributed to property and deception offences, which can be further categorized into subtypes such as arson, burglary, and theft, among others.

Satisfaction with your standard of living ranked from 0-100 (very dissatisfied to very satisfied)

With comprehensive data cleaning undertaken during the data wrangling phase, extensive data checking is deemed unnecessary for this dataset. Visualizations generated via the vis_dat() function (*figure 4)* affirm data consistency, the absence of missing values, adherence to naming conventions, and correct data types. Additionally, examination through boxplots reveals an overall high standard of living score, with a minimum of 75. Notably, outliers, characterized by scores of 86, emanate from participants residing in Macedon Ranges and Moorabool. These findings align with the summary statistics provided by the fivenum() function (75 77 78 80 86).

Index of relative socio-economic disadvantage ranked within state & education and occupation ranked within state

These two datasets have undergone thorough cleaning during the data wrangling phase, as supported by the visualizations generated using the vis_dat() function *(figure 6)*. Furthermore, the examination of boxplots *(figure 5)* for the Score variable reveals a consistent pattern with no apparent outliers. This observation is corroborated by the summary statistics of the Score variable for both datasets, reinforcing the data's integrity and reliability.
fivenum(SED_index$Score)
[1] 889.0 964.0 994.0 1032.5 1099.0

fivenum(EandO_index$Score)
[1] 889 956 979 1041 1160

Data Exploration

Question 1: Are there some areas in Victoria more dangerous than others?

To address this question, spatial analysis is conducted using an additional geometric dataset sourced from the Data Vic website. Initially, a data integration approach involves merging the comprehensive dataset of total criminal incidents with this newly acquired dataset, leveraging the `left_join()` function, while linking the data by the Local

Government Area (LGA). Within the realm of data exploration, the analysis focuses on simplification by computing the average incident rates per 100,000 population, aggregated by Local Government Area, spanning a decade. However, for the data visualization component of the project, users will have the flexibility to apply year-based filters to enable in-depth examination.

The spatial visualization *(figure 7)* outcomes corroborate the findings from the boxplot *(figure 2)* analysis. Notably, the three most hazardous areas, exhibiting the highest average incidents per 100,000 population, are identified as Melbourne, Latrobe, and Yarra, vividly depicted as the three brightest-blue tiles in the spatial visualization.

Question 2: Are there any socio-economic indicators associated with highly dangerous areas?

Satisfaction with your standard of living

Correlation Analysis

We calculate Pearson correlation coefficients between crime rates (Rate per 100,000 population) and average score for satisfaction with standard of living by using cor.test() function. This function will also determine whether the correlation is statistically significant.

```
> correlation_test

        Pearson's product-moment correlation

data:  ssl_crime_tot$`Rate per 100,000 population` and ssl_crime_tot$Average_Score
t = -2.1568, df = 51, p-value = 0.03575
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.51886631 -0.02042279
sample estimates:
       cor
-0.2891206
```

The result indicates there is a negative relationship between criminal incidents and satisfaction with living standard index.

Scatterplots

Limited data availability and some missing values constrain the scope of the correlation analysis between crime rates and socio-economic indicators. The available data, restricted to the year 2021 and lacking scores for certain LGAs, hinders a comprehensive examination. Nevertheless, despite these limitations, a preliminary analysis through scatterplots *(figure 8)* hints at a potential negative relationship between the two variables.

Linear Regression

```
Call:
lm(formula = `Rate per 100,000 population` ~ Average_Score, data = ssl_crime_tot)

Residuals:
    Min      1Q  Median      3Q     Max
-3855.2 -1686.7     5.1   978.9 10386.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   27517.0    10152.7   2.710  0.00913 **
Average_Score   -278.4      129.1  -2.157  0.03575 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2481 on 51 degrees of freedom
Multiple R-squared:  0.08359,   Adjusted R-squared:  0.06562
F-statistic: 4.652 on 1 and 51 DF,  p-value: 0.03575
```

Similarly, the linear model reveals a negative relationship between the examined variables. Nevertheless, it is crucial to emphasize that data limitations and a relatively low R-squared value cast doubts on the model's reliability.

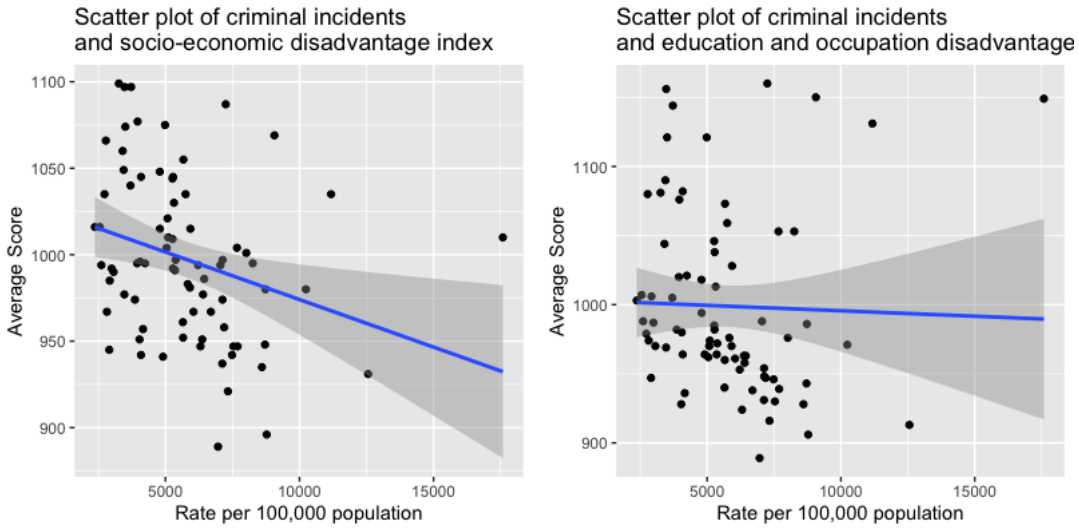
Index of relative socio-economic disadvantage ranked within state & education and occupation ranked within state

Correlation Analysis

	Rate per 100,000 population.x	Score_SED	Score_EandO
Rate per 100,000 population.x	1.00000000	-0.2909682	-0.02995949
Score_SED	-0.29096820	1.0000000	0.87082650
Score_EandO	-0.02995949	0.8708265	1.00000000

The correlation matrix reveals that there is a negative relationship between the average scores of socio-economic disadvantage and education and occupation with criminal incidents. However, these relationships are relatively weak, with correlation coefficients of -0.29 for socio-economic disadvantage and -0.03 for education and occupation.

Scatterplots



The scatterplot analysis demonstrates a negative relationship between criminal incidents and the socio-economic disadvantage index. Conversely, the scatterplot depicting the relationship between criminal incidents and the education and occupation disadvantage index shows an almost straight line, indicating a relatively weak negative relationship between these variables.

Linear Regression

```
Call:
lm(formula = `Rate per 100,000 population.x` ~ Score_SED + Score_EandO,
    data = full_crime_tot)

Residuals:
    Min      1Q  Median      3Q     Max
-4055.0 -1794.8   237.5  1017.4  7284.6

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28440.100   5371.596   5.295 1.15e-06 ***
Score_SED     -58.006     10.465  -5.543 4.24e-07 ***
Score_EandO    35.201      7.529   4.675 1.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2146 on 75 degrees of freedom
Multiple R-squared:  0.2912,    Adjusted R-squared:  0.2723
F-statistic: 15.41 on 2 and 75 DF,  p-value: 2.478e-06
```

The coefficient analysis reveals that for every one-point increase in the index of education and occupation disadvantage, there is an average increase of 35 criminal incidents per 100,000 population. Surprisingly, the socio-economic disadvantage index shows the opposite effect, with each point higher in the index associated with an average decrease of 58 criminal cases per 100,000 population.

Conclusion

In conclusion, this comprehensive analysis of crime incidents and socio-economic indicators in Victoria has shed light on various aspects of crime rates and their potential associations with socio-economic factors. Our investigation revealed that Melbourne, Latrobe, and Yarra were identified as the areas with the highest average crime incidents per 100,000 population, emphasizing their relatively higher risk. Furthermore, the distribution of crime incidents across offense categories highlighted that property and deception offenses accounted for the majority, with subcategories such as arson, burglary, and theft contributing significantly. Correlation analyses were conducted to explore the relationship between crime rates and socio-economic indicators. The findings suggested a weak negative association between crime rates and indices of socio-economic disadvantage and education and occupation disadvantage. It is essential to acknowledge the limitations of this analysis, including data availability, data quality, and the absence of temporal causality assessments. Nevertheless, this study contributes to a better understanding of crime patterns and their potential relationships with socio-economic factors in Victoria.

Reflection

The completion of this project presented a myriad of challenges, yet it was also an immensely rewarding experience. Throughout the process, I acquired a profound understanding of data acquisition, manipulation, and transformation, ultimately enabling me to derive meaningful insights and answers to the research questions posed. The project also gives me the opportunity to hone my data visualization and wrangling skills, further enhancing my proficiency in these critical areas of data analysis.

As elucidated in the report, the socio-economic datasets utilized in this study possess inherent limitations, including biases and insufficient observations, which could potentially introduce inaccuracies and render the answers to the second research question less reliable. This underscores the importance of exercising increased caution during the data acquisition phase, with thorough metadata examination becoming an integral part of the dataset selection process. In future projects, I am committed to applying these lessons to ensure a more robust and accurate analysis, ultimately contributing to the advancement of research in this domain.

Bibliography

Glynn, L. (2023, June 22). Melbourne named the world's third most liveable city. *TimeOut.* https://www.timeout.com/melbourne/news/melbourne-named-the-worlds-third-most-liveable-city-062223

Davis, M. (2023, September 8). 'We've had enough': Radio host Neil Mitchell calls for 'serious rethink' to combat Victoria's rising youth crime problem. *Sky News.* *https://www.skynews.com.au/australia-news/weve-had-enough-radio-host-neil-mitchell-calls-for-serious-rethink-to-combat-victorias-rising-youth-crime-problem/news-story/3fb30a11e2c8f8a8d1b1fe6ed7031433*

Appendix

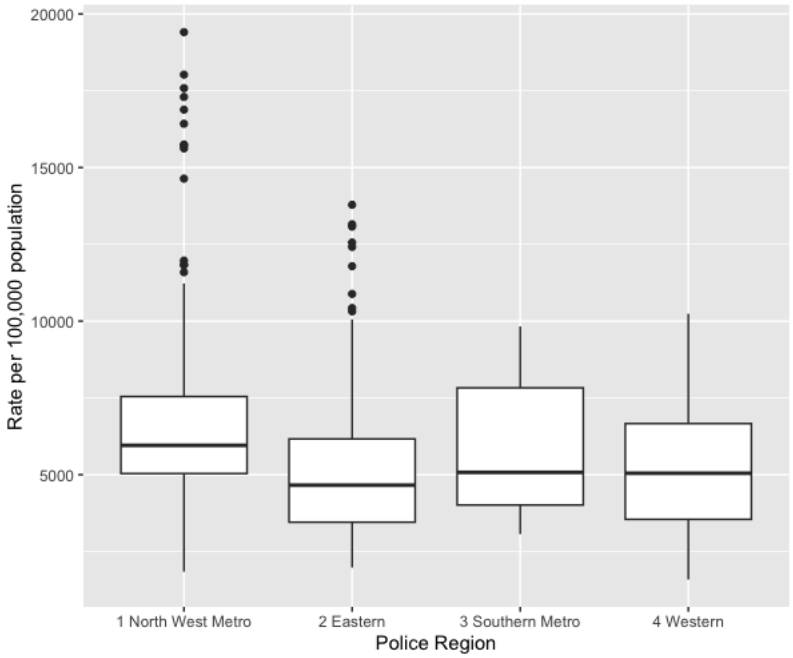


Figure 1: Box plot of rate per 100,000 population for each police region

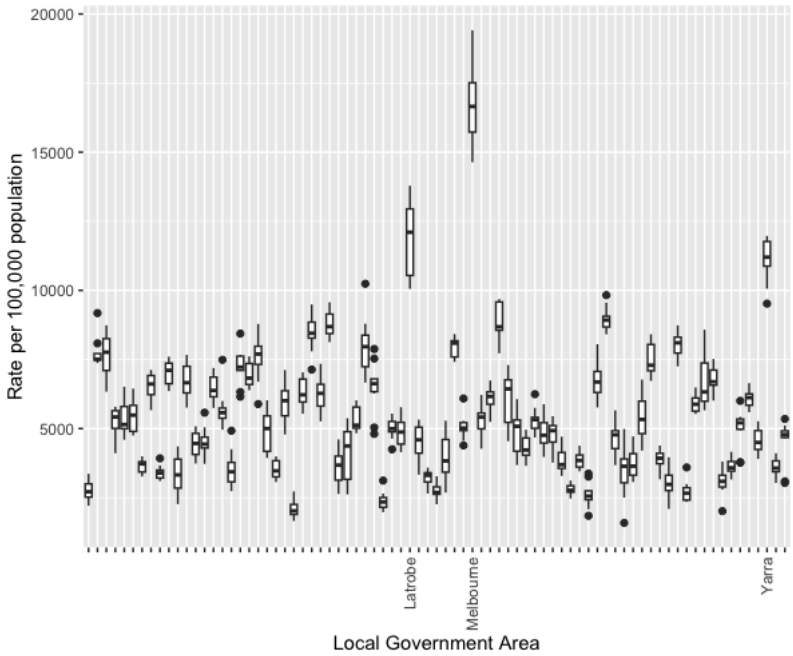


Figure 2: Box plot of rate per 100,000 population for each local government area

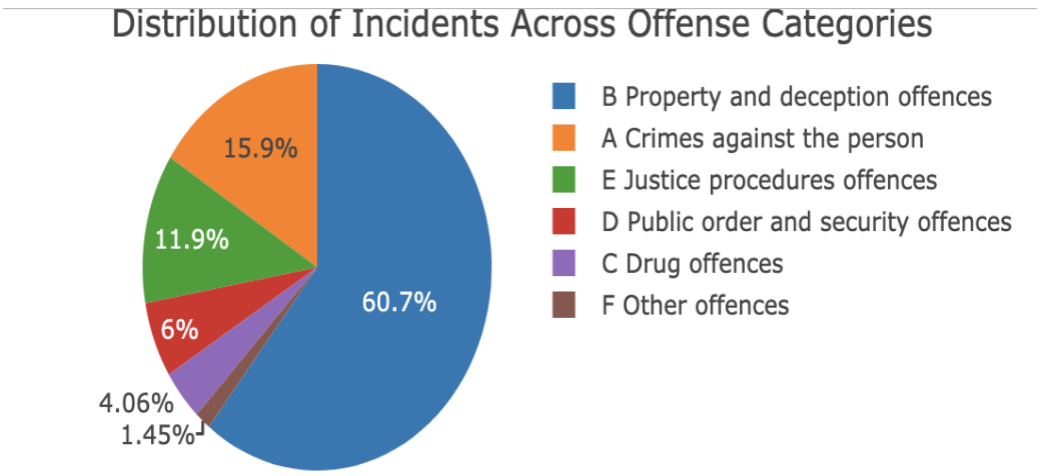


Figure 3: Pie chart for distribution of incidents across offense categories

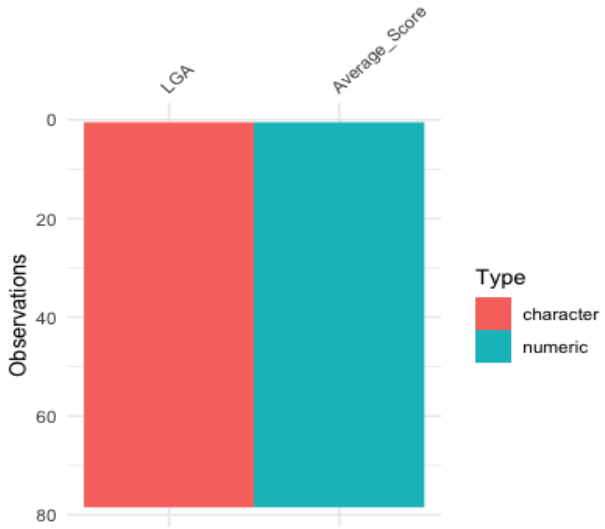


Figure 4: Variable class and missing values indication for each variable in Satisfaction with your standard of living dataset

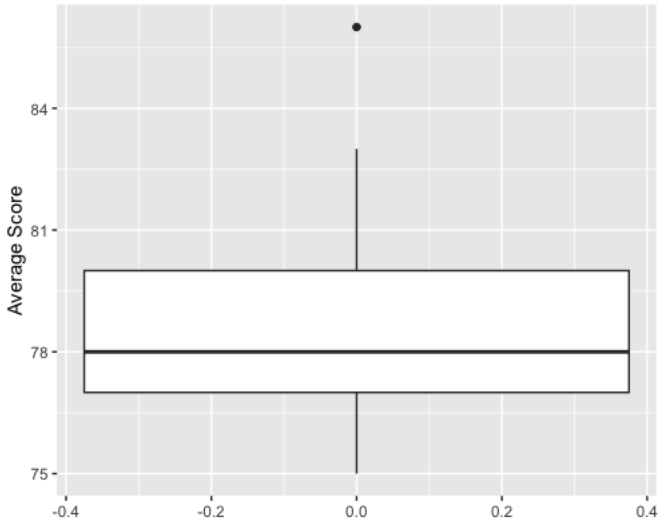


Figure 5: Box plot of average score of Satisfaction with your standard of living index

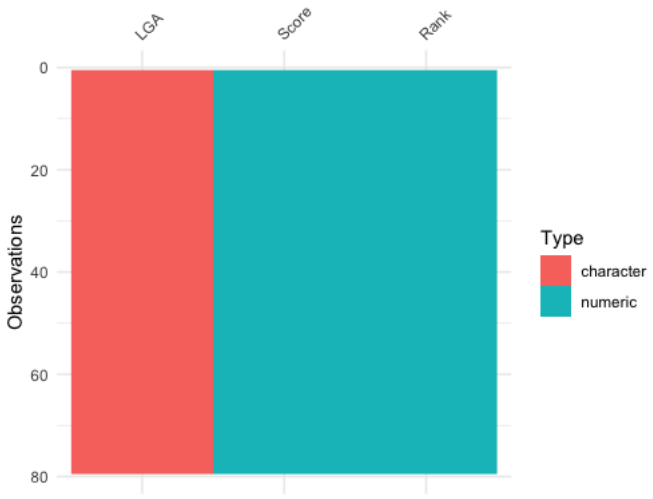


Figure 6: vis_dat() result for Index of relative socio-economic disadvantage ranked within state & education and occupation ranked within state data set

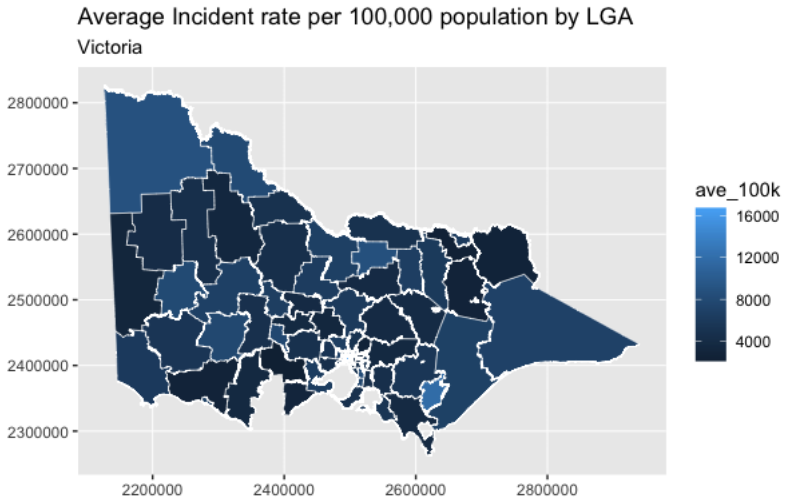


Figure 7: Spatial Visualisation for average criminal incident rate per 100,000 population by local government area

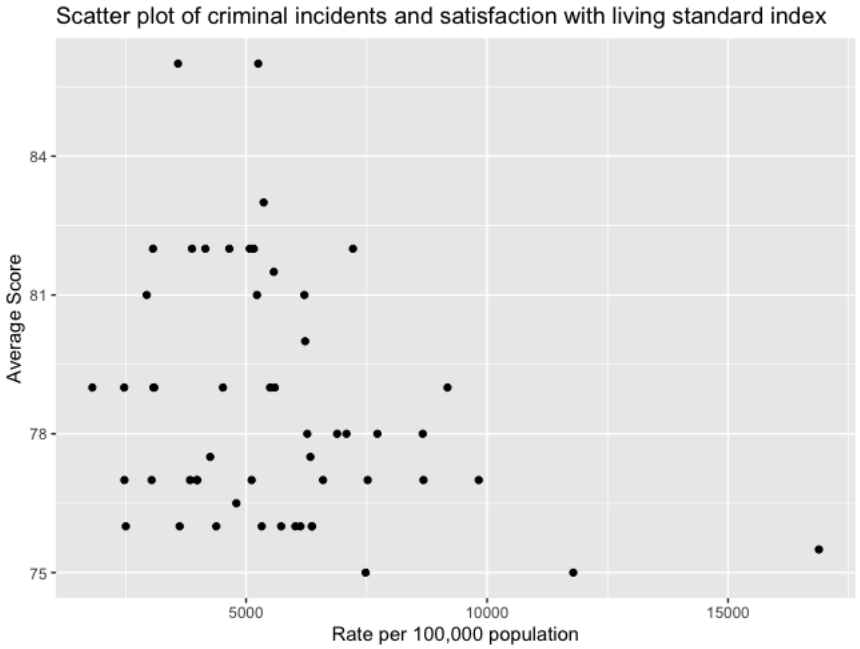


Figure 8: Scatterplot of criminal incidents per 100,000 population and satisfaction with living standard index