"Social Benefits and Crime Rate: Identifying correlations"

Ang Mink Chen

Domain
The domain of the study is based on communities.

Question
"Can the rate of offences be reduced by giving helps and benefits through social programs to those in need?"

The aim of this study is to identify correlations between social benefits, specifically Department of Social Services (DSS) payments, and offences rate in Victoria. DDS payments are benefits or helps allocated by the government to specific groups of residents, for example, in forms of age pensions, youth allowance, low income card, etc. Finding the appropriate correlations could potentially help the government to better manage their funds and policies in hope of minimizing the offence rate of the local communities.

Datasets
The following datasets were used:

- Department of Social Services - DSS Payments by 2014 Local Government Area: Contains attributes of 25 different types of DSS payments and the number of recipients of each type of DSS payment in all Victoria LGA. The data in it are recorded quarterly between 2016-2017. It is a comprehensive dataset and is sufficiently detailed to provide an accurate analysis.
  URI: "https://data.gov.au/dataset/dss-payments-by-local-government-area"

- Crime Statistics Agency - LGA Number of offences in Victoria by offence type 2008-2017: Includes attributes of 35 different types of offences and the offence counts recorded of each type of offence in all Victoria LGA. The data are recorded from 2008-2017. As this dataset is provided by the CSA, it can be relied upon to be accurate and correct. URI:
  "https://data.aurin.org.au/dataset/vic-govt-csa-lga-vic-crime-stats-2008-2017-lga2011"

- Australian Bureau of Statistics - Estimated Resident Population by LGA (ASGS 2016), 2001 - 2016: Contains the total number of estimated resident population (ERP) of each LGA in Australia. It covers the data from year 2001-2016. Only the data in year 2016 are required for the normalization of data values in the datasets schema. URI: "http://stat.data.abs.gov.au/Index.aspx?DataSetCode=ABS_ERP_LGA2016"

These datasets were chosen as they are published from reliable sources, and have common LGA attributes, which simplifies the integration processes of the datasets.

Pre-processing

Pre-processing of the datasets was done by using **pandas** library in Python. The data recorded in the datasets were stored and processed as the **DataFrame** object in Python, which it is designed to work with large matrices of data. The following processes were done to prepare the datasets to ensure successful integration:

- **Removing unnecessary data.** Columns that do not contain any important information or value were removed from the datasets in hope of eliminating redundancy and unrelated data to allow for easier manipulation of the data. Moreover, the data in the datasets were filtered by selecting the data recorded only in year 2016 and in Victoria LGA.
- **Removing missing values**. Only the "Offence" dataset out of the three contains missing values. Majority of the missing values exist only within several columns of the dataset. It is more preferable to remove the columns that contain missing values as the removing rows with missing values would significantly reduce the reliability of the dataset. (Too many rows contain missing values, hence deleting them would render the datasets useless).
- **Handling outliers**. The outliers in the data are not removed from the dataset. The reason being that removing the outliers from the dataset would potentially alter the research findings, in which it increases or decreases the correlations of the datasets artificially, leading to false correlations being identified.

- **Normalization**. The data in both of the "DSS Payments" and the "Offence" datasets were normalized to a percentage. It was done by dividing the value over the estimated resident population accordingly for each LGA, then multiple them by 100. Hence, each value in the data represents a percentage of the estimated resident population.

- **Multiple records of data**. As the data in the "DSS Payment" dataset were recorded quarterly, there were four values recorded for each LGA. Hence, the average of the values was the calculated and used to replace the old values in the dataset. Although the accurate value of the data value can be retained through this approach, the new values lack a natural representation of the data as it contains non-integer number.

Integration

To connect the two datasets and find their general correlation, the total number of DSS payments recipients and the grand total offence counts were calculated and extracted to be integrated into a new dataset via their corresponding LGA code. A scatter plot was then generated along with its Pearson correlation coefficient (**Figure 1**).

Due to the multidimensionality of the data, the general trends and correlations exist between two datasets are very likely to have been affected by substantial amount of internal and external factors. Consequently, no informative results could be obtained.

The next step was to calculate the Pearson correlation for every pair of attributes from different datasets in hopes of finding meaningful results. Hence, 525 Pearson correlation coefficients were generated and plotted in a matrix to visualize the results (**Figure 2)**.

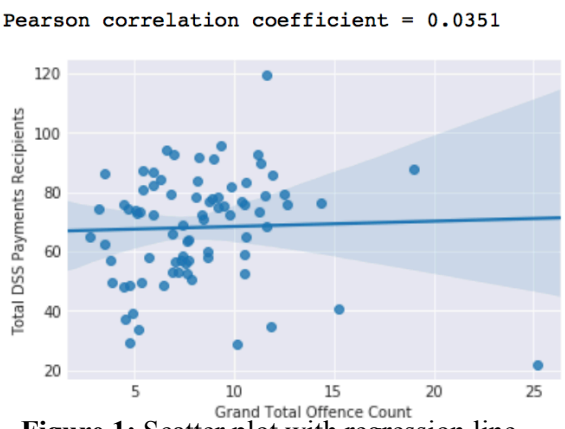


Figure 1: Scatter plot with regression line

This was done by making all types of offences columns of a new dataset, with each row represents a type of DSS payment, and each value represents its corresponding Pearson correlation coefficient. Lastly, **seaborn** Python library was used to generate the heatmap visualization. Several pairs of attributes with exceptionally interesting correlations were selected for further investigation.

Results

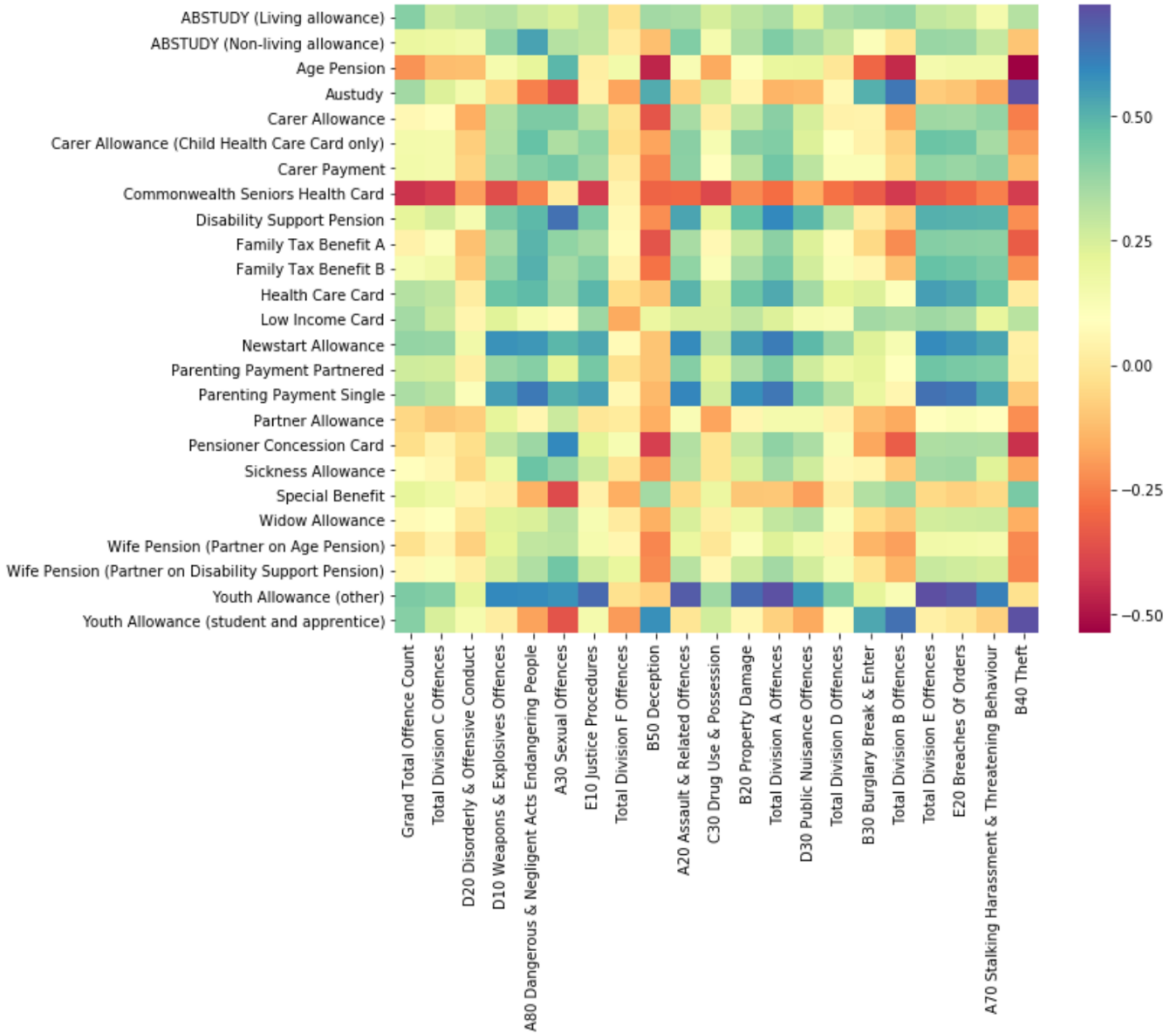
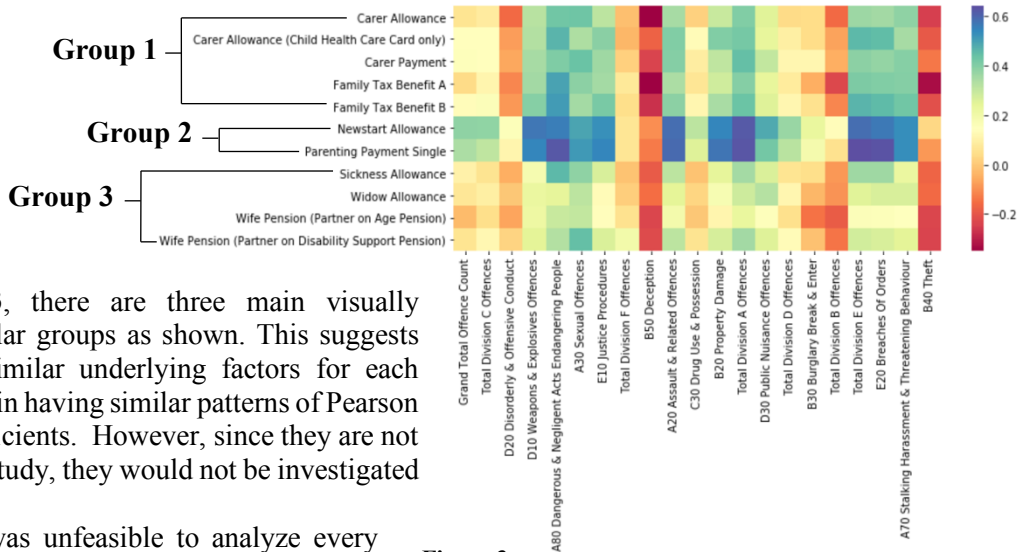


Figure 2: Heatmap of Pearson correlation coefficients

Upon examining the correlation heatmap, both strongly positive and negative correlations can be easily identified. A selection of interesting results is further discussed below:

- The strongest positively correlated pair of attributes exists between 'Youth Allowance (other)' and 'Total Division E Offences' ($r \approx 0.72$).
- The strongest negatively correlated pair of attributes exists between 'Age Pension' and 'B40 Theft' ($r \approx -0.54$).
- Areas with higher percentage of Commonwealth Seniors Health Care Card recipients tend to have lower percentage of overall offences count (with average $r \approx -0.29$).
- Areas with higher percentage of Youth Allowance (other) recipients tend to have a trend of higher percentage of overall offences count (with average $r \approx 0.43$).
- Offences such as 'B50 Deception' and 'B40 Theft' have more negative correlations relatively to the other offences, which suggests that their offence counts are more likely to be reduced through increasing the total number of DSS payments recipients.
- Similar visual patterns of Pearson correlation coefficients can be identified and are selected to be visualized as the followings (**Figure 3**):



From **Figure 3**, there are three main visually identifiable similar groups as shown. This suggests that there are similar underlying factors for each group that result in having similar patterns of Pearson correlation coefficients. However, since they are not the focus of the study, they would not be investigated any further.

Figure 3: Heatmap of selected Pearson correlation coefficients

Considering it was unfeasible to analyze every correlation in the heatmap, several pairs of attributes with interesting correlations were chosen arbitrarily to be examine further. Scatter plot was used then to visualize detailed correlations of the chosen pairs of attributes, along with a regression line drawn in the plot to better show its correlation.

```
Pearson correlation is  0.7227850212884126
```

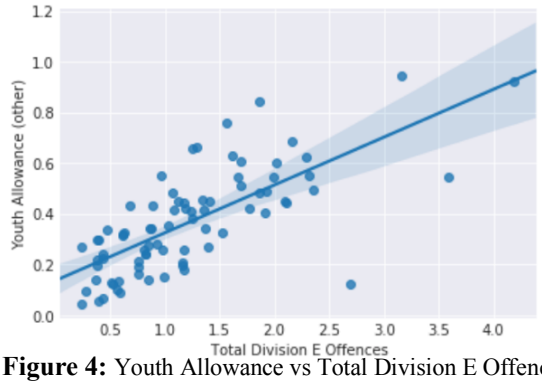


Figure 4: Youth Allowance vs Total Division E Offences

```
Pearson correlation is  0.6467334002518121
```

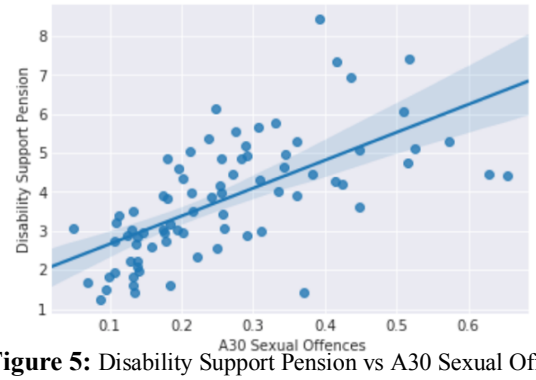


Figure 5: Disability Support Pension vs A30 Sexual Offences

```
Pearson correlation is  -0.5375589872312175
```

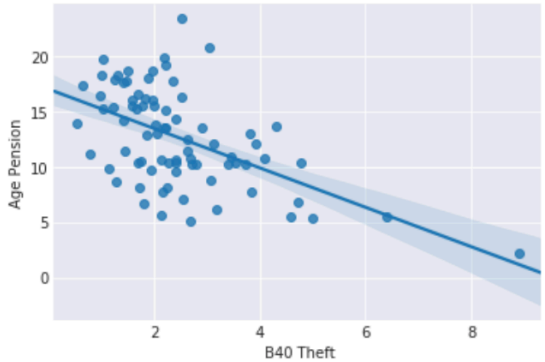


Figure 6: Age Pension vs B40 Theft

```
Pearson correlation is  -0.40575948335543244
```

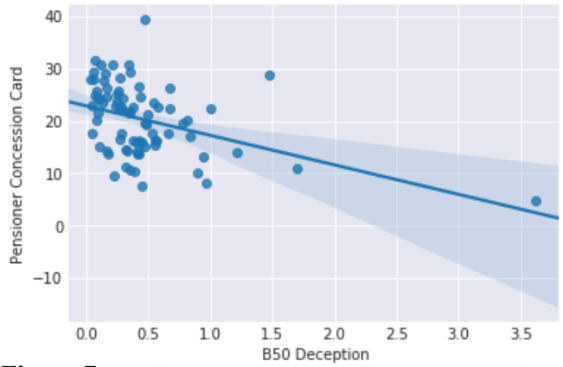


Figure 7: Pensioner Concession Card vs B50 Deception

It is worth mentioning that the results might be slightly skewed by the presence of the outliers in the datasets, however the real accuracy retained by not removing the outliers increases the likelihood of finding more interesting results as it does not artificially improve the final results. Some of the correlations found are not particularly unexpected, for example the one as shown in **Figure 7**, as it can be argued that one who receives a pensioner concession card (a form of DSS payments) is less likely to deceive to gain benefits from the other. Conversely, correlation such as the one shown in **figure 5** is certainly less likely to be anticipated, as there is no direct link or association between the two attributes.

Value

The correlations and results obtained from the wrangling processes of the data in this study would not able to be deduced directly from the raw values of the datasets as the data was incomprehensible in its original form. Due to the multidimensional nature of the datasets, it is very unlikely that any meaningful information and knowledges can be inferred from the raw datasets. By taking advantages of the processing power of computers, human-readable and informative visualisations of the data were processed and produced from over 130,000 data points in this study. Informative explanations and correlations were able to be deduced which help to answer the questions posed earlier in the study by analysing the plots and heatmaps produced.

Challenges and Reflections

While there are ton of datasets on the topic of this study existing on the internet ready to be downloaded and processed, it was very difficult to find a set of datasets that has matching attributes and properties that simplify the integration processes later in the research. Moreover, due to the multidimensionality of the datasets, it was difficult to visualise the data to produce interesting and meaningful correlations and interpretations. Multiple approaches were used and failed to generate the appropriate correlations.

Although the results produced from the study had shown unexpected and interesting correlations between certain variables and attributes, it is worth noting that more often than not, correlation does not imply causality. Furthermore, the relatively importance and effects of possible confounding factors that affected the correlations and results of the wrangled data directly and indirectly have not been analysed or identified in this study, which in turns it further undermines the significance of the outcomes. For instance, while the effect of number of DSS payments recipients can be determined and studied, the effect of the differences in individual amounts received by each DSS payment recipient have not been taken into the consideration of this study, hence it might affect the overall correlations of the datasets ultimately.

Question Resolution

The results produced from the study are not able to directly answer the question posed earlier in the study. The reason being that scope of the question posed is too generic, as both supportive and opposing results have been produced from the study, there is no one definitive answer. More accurate and informative results can be achieved with more detailed statistical analysis. Moreover, further investigation could identify unknown factors and refine the results to be more applicable.

However, the results of the study could be valuable to Victorian government and the DSS in managing the government's budget for social benefit program with hopes of minimizing the crime rate for the local communities. Specifically, based on the indications of correlations generated, it could provide insights to the government on the how types of DSS payments could potentially affect the crime rate of a certain LGA.

Code

All codes were written in Python programming language. The major Python libraries used are **pandas, matplotlib, numpy, datetime, scipy,** and **seaborn**. CSV dataset files were opened and read using **pandas** methods. Methods from **numpy, datetime,** and **pandas** were used to prepare, process, and integrate the data for detailed analysis. Correlations between each pair of attributes of datasets are calculated and generated using codes written independently with minor implementations of functionality from the major libraries. Scatter plots and heatmaps were then generated using plotting method of **seaborn** library.