“Identifying correlations between homelessness and LGA populations”

Domain

The scope of this study is within the domain of Public amenities and communities.

Question

This study seeks to give an insight on the correlations between homelessness and populationin Victorian Local Government Areas(LGAs), in the span of a six-year period (2011-2016). The open data wrangled from this study would be of great benefit to the Victorian Government, by conveying the LGAs which need more updated support, by inspecting other reasons potentially causing homelessness. Not only will this study observe the correlations with homelessness, it will also discover the various non-intuitive reasons which lead to homelessness. Some other features that will be correlated against homelessness are: drugs, social housing, access to community services, low income, education, multiculturalism and assistance from family and friends. These complementary topics were chosen as they are possible reasons behind homelessness.

Innovative Information

This study will provide innovative information since the homelessness crisis is still a prevalent subject across Victoria. Moreover, no concrete solution by the government has been found to combat this ongoing issue, which is a serious problem for the people of Victoria. By wrangling various open datasets, the results obtained will identify the relationships between homelessness and populations of Victoria’s LGAs. Not only will this investigation provide insightful information about Victoria’s homeless populations, it will demonstrate a different angle on which areas need more attention from the Victorian Government.

Datasets

The following datasets are to be used:

- **Dataset 1:** Local Government Area(LGA) profiles data 2015 for VIC – csv
  This dataset contains most recent information on homeless populations per LGA, and key information such as LGA names and codes. Some of the relevant attributes taken from this dataset which concern homelessness are: LGA Code (2015), LGA Name (2015), social housing, education, accessibility to support services, support from close ones and ethnicity/backgrounds.

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

- **Dataset 2:** VIF 2015 **-** LGA Yearly Estimated Resident Population 2011 - 2031 for Victoria – csv
  Estimations of populations per LGA, specifically only looking at years 2011-2016. Similarly, to the above dataset, LGA names and codes are given, and population sizes are whole numbers.

  URI: https://data.aurin.org.au/dataset/vic-govt-delwp-vic-govt-delwp-vif2015-erp-1yr-2011-2031-lga2011

Processing, Integration, Analysis and Visualization

For the data to be meaningful to this investigation, unnecessary columns and redundant data will be removed from the dataset. Therefore, the datasets will be first loaded into panda’s data frames, then pre-processed to remove unnecessary data. This is beneficial over keeping the raw data unchanged,

as unnecessary data will be filtered out. For example, in dataset 1, some LGAs are missing. This would mean some LGAs would have be filtered out in the estimated population dataset, to allow both datasets to be consistent with respect to LGAs. Additionally, aspects such as missing data and outliers will have to be considered before moving onto the integration stage. These modifications of the data will allow the datasets to be more useful in answering the question.

After careful inspection and cleansing of the datasets, visualizations using python libraries such as matplotlib can be used to model the data with the goal of discovering more useful relationships between the datasets. This Is beneficial over the raw data as visual motives in the data can be highlighted through visualizations such as bar charts and scatterplots. An example would be to plot comparisons in populations between both datasets, and seeing how much of each LGAs population is homeless.

Summary of Initial Investigations

Before moving on to the integration of the data with this investigation, some preprocessing tasks were needed. This included removing empty NaN rows in the datasets, which was observed through python data frames from pandas. This verified that only valid, usable data remained behind. It was also observed that both datasets contained identical columns, LGA codes and LGA names. In dataset 2, these columns were reordered to have the same orientation as dataset 1, meaning that both these columns were made to be the first two columns in the data frames. This decision was made to make comparisons between both datasets trivial, and allowed consistency between them. Additionally, some of the column names were too long, and had to be shortened to allow indexing of columns to be a straightforward task.

The first task to testing if this plan is feasible was by comparing the homeless population from dataset 1 with the estimated populations in dataset 2. This comparison showed that the increase in population per LGA led to a higher homelessness count, however there are some issues with this intuitive observation. Firstly, the population counts are estimated, and don't convey the truthful population per LGA. This could lead to inaccurate results. Secondly, the homeless populations per LGA are based of 2011, and interestingly this is the most recent data concerning homelessness in LGA, as the dataset was released in November 2016. It is obvious that if the population increases each year, then applying the same homelessness data from 2011 would result in a linear correlation between rising homelessness and population. To resolve this issue, some other categories which indirectly involve homelessness were considered.

Top 5 most homeless LGAs:

Rank	Local Government Area	Homeless per 1000 Population
1	Port Philip (C)	15.3 (1.53 %)
2	Melbourne (C)	10.6 (1.06 %)
3	Greater Dandenong (C)	10.3 (1.03 %)
4	Yarra (C)	10.0 (1 %)
5	Maribyrnong (C)	8.9 (0.89 %)

The table on the previous page shows the 5 most homeless populated LGAs. This table was created to show some of the high homelessness populations in Victoria. It is possible to remove these LGAs from the dataset, because they are obvious outliers just from inspecting the integrated data.

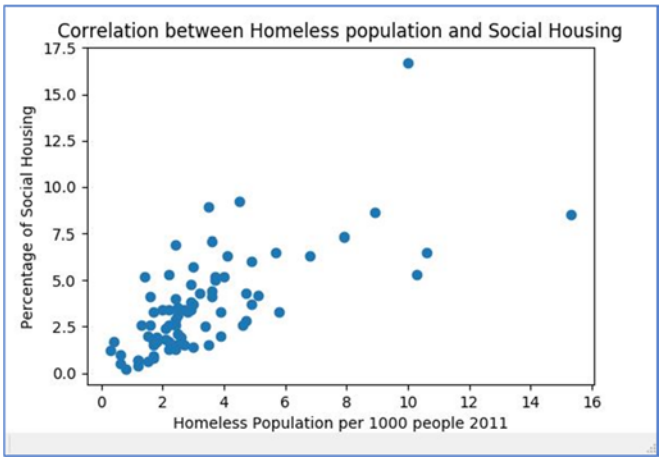

Figure 1: Scatter plot of correlation between homelessness and social housing in Victoria's Local Government Areas

Since the data was preprocessed and integrated, some visualizations were made to detect correlations between homelessness and different categories. When plotting the homelessness population (2011) against the social housing percentages (2014-2015) using plt. scatter () from matplotlib, It was evident that there was a moderately positive Pearson's r correlation of 0.6908. The higher the homeless population, the higher the percentage of social housing was found in each LGA. Outliers from Yarra (C) and Port Phillip (C), which was expected since they had high homelessness, therefore more social housing would be needed.

Another lopsided observation with scatterplots was found in comparing the homeless population with the percentage of people who could get help from friends/family/neighbors. This showed a negative Pearson's r correlation of -0.5781, whereby the higher the homeless population increased, the less support people had from friends/family/neighbors. This was expected, as places with less support in these areas would obviously lead to more homelessness. A more non-intuitive correlation was also found with people born overseas. This produced a positive correlation of 0.5643 in the scatterplot, which reinforced that people from ethnic backgrounds suffer from homelessness in LGAs. When plotting homelessness populations against percentage of people who could access community services and resources, this provided very little correlation, -0.0898 to be exact. This was not expected, as it was hypothesized that less access to community services would lead to more homelessness, but this doesn't seem to be the case. Both datasets are from the same years (2011), so it is fair to assume that this correlation is true. Education was another feature that needed to be examined, since it is a fair assumption that not having finished high school could have an impact on homelessness. It was seen that a -0.4890 correlation was found, and that the increase in homelessness per LGA could not have depended on if high school was finished.

Feasibility of project

After seeing many interesting results from the initial investigations, it is promising that more interesting results will come from this investigation. This project is feasible as it provides a new angle to the homelessness problem in Victoria, which is what Victoria needs to prevent this issue from becoming worse in the future. Instead of looking at the most common reasons for homelessness such as unemployment, family violence and poverty, it examines non-intuitive causes that could contribute to this rising epidemic. It is also feasible because looking at the different causes to homelessness can provide innovative information for the government.

Many interesting results have already been found, and perhaps the scope of the question will need to be modified. This experiment shows that more categories can be tested to yield more interesting results. The current successful results show many interesting correlations with homelessness in Victoria, and provides many different starting points to solving this problem.