**COMP20008 Phase 3 Report**

Armaan Dhaliwal-Mcleod | 837674

## Title

Contributing factors to Homelessness in Victoria’s Local Government Areas

## 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 following question:

**“Are socioeconomic and demographic factors causing homelessness in Victoria’s LGAs?”**

The open data wrangled from this this investigation would be of great benefit to the Victorian Government, by inspecting potential causes to homelessness in Victoria. Not only will this study look at correlations between homelessness and various impacting features, it will also attempt to discover non-intuitive causes for this issue. The candidate features that will be looked at are: homelessness, social housing, drug usage, affordable rental housing, unemployment, cultural diversity, education, social support and population.

## Datasets

The following datasets are to be used:

* **Dataset 1:** Local Government Area(LGA) profiles data 2015 VIC -csv

This dataset covers a wide range of information concerning LGAs, including topics such as diversity, population and social engagement.

**Schema:** Specific to this study, The following attributes were taken from this dataset: LGA Code(2015), LGA Name(2015), Homeless people (estimated) per 1000 population (2011), Percentage of Population born Overseas(2011), Number of Social housing dwellings (2014-2015), Drug usage and possession offences per 1000 population (2015), Percentage of rental housing that is affordable (2015), Unemployment rate (2015), Percentage of people who did not complete year 12 (2011), Percentage of people who are definitely able to get help from family, friends and neighbours (2011).

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

This dataset gives the yearly estimated populations per LGA over the span of years 2011-2031. The year 2015 was chosen to match the profile data above.

**Schema:** In relevance to this study, the following attributes were taken from this dataset: LGA Code (2015), LGA Name (2015) and Estimated Resident population 2015.

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

These datasets were chosen from a reliable source, AURIN, and both have common attributes lga\_code and lga\_name.

## 

**Figure 1:** Sample profile data

## 

**Figure 2**: Sample population data

## Preprocessing

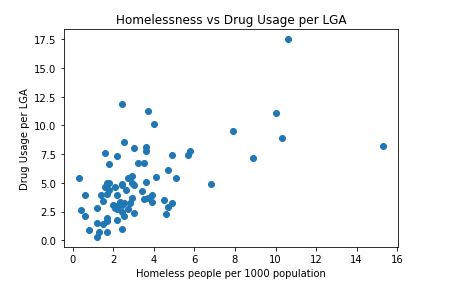
Many of the preprocessing tasks were accomplished with the *pandas* library. Data was read into *pandas Data Frames,* since this was the most trivial way to categorize and manipulate the data. Given that the datasets contained a substantial amount of data concerning each LGA in Victoria, it was essential to clean and only keep data which was required to answer the question. Since both datasets contained two matching columns, LGA code and LGA name, it was assumed that combining these datasets would be a trivial task. The following preprocessing tasks were done, in preparation for a successful integration:

* Whitespace was removed from column names in the data frames, for indexing to be more straightforward.
* Merging of the two datasets into one data frame was done, since both datasets contain the same primary keys LGA code and LGA name. Specifically, the tpop\_2015 column was taken from the second dataset, and moved into the first dataset.
* Checking of LGA consistency was examined between both datasets, and any LGA that was not present in both datasets was removed.
* The data columns were read in with extremely long column names, and were shortened to allow more concise, meaningful headers.
* Checking of missing values and unnecessary data was performed. This included percentages that were over one hundred, NaN entries and zero values. Since both datasets were very clean to begin with, none of these issues were found.
* Ordering of columns were changed, with LGA code and LGA name to be the first two columns, as each row should be headed by these identifiers.
* Although it was evident that there would be outliers in the data, such as Melbourne, it was decided that they would be left unchanged. This choice was made because highly populated LGAs such as Melbourne are hotspots for homelessness, and leaving such data out of the investigation would lead to false correlations.

## Limitations of Preprocessing

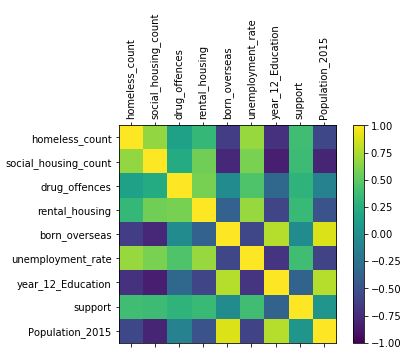
Some limitations were encountered in the preprocessing stage. Despite the profile dataset being 2015, some of the attributes recent profile data was in 2011, which could cause inconsistencies in the data. Despite this, the profile data was released on 2016, and uses the most updated information available. This could lead to slightly inaccurate results, but no other updated datasets were available. Another limitation was found in the homeless data for each LGA per 1000 population. The values were given in decimal, which indicates the average homeless count. This allowed this part of the data to be awkward, as it was expected that the data would be put into whole numbers, since it is only possible to have ‘whole’ persons. This could be a limiting feature if you wanted to model some data on how many people are homeless in each LGA. Moreover, some features such as People who finished grade 12 might not necessarily include the population of interest, the homeless. These various limitations may produce slightly inaccurate results, but there is not much you can do about this.

Integration

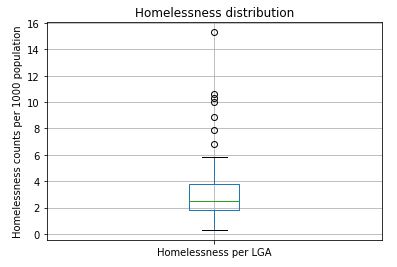
To allow correlations between homelessness and the various features to be easier, simply comparisons between the homeless count column and the rest of the features was performed. Due to the many features being included into the data, it was difficult to determine if any of socioeconomic features contained useful information. To decide if the investigation is feasible, some scatter plots of the of the features against homelessness per LGA were generated using *matplotlib*.

For example, just by looking at some of the results, it is evident that some correlations exist. For example, when looking at plot between homelessness and drug usage per LGA (**Figure 3**), it is evident that there is a positive linear correlation. The Pearson correlation evaluated to 0.5643, showing a strong correlation between the two variables. The outliers, Melbourne and Yarra are evidently observed in this visualization, but for the purposes of this investigation, they were not removed.

**Figure 3:** Scatter plot demonstration

The limitations of simply using scatter plots to determine correlations was not substantial to make classification model of this data would be needed to answer the question.

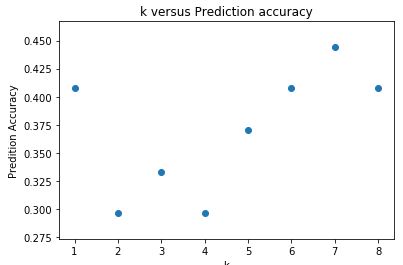
## Results

After analyzing the correlations from the scatter plots, it was immediately clear that there were potential correlations in the data. The correlations between homelessness and the various features per LGA are shown in the correlation matrix (**Figure 4**). The highest correlations with homelessness were with social housing (0.69), drug usage (0.58) and social support (-0.57). Social housing and drugs usage was expected, but social support provides interesting results. It seems that the more homeless areas are, the less social support they have. Negative Correlations were also found with Education (-0.48) and rental housing affordability (-0.37), showing that areas with people with less high school qualifications lead to more homelessness, and areas that are less affordable cause more homeless people. Not only were their high socioeconomic correlations, there were demographic correlations with people born overseas (0.56). This was an interesting result, as it was unexpected that areas with a more ethnic background have more homeless people. Other correlations with population (0.33) and unemployment rate (0.35) were also expected.

**Figure 4**: Correlation Matrix

The above shows moderate to strong correlations in the data, when comparing homelessness against the other features. Although these comparisons provide useful information, more visualizations and analysis are needed. To quantify the quality of these results, larger models such as K-NN and decision tree classifications are needed. Since the data was already imported into data frames, it was essential to firstly divide the data into class labels and features. Feature X, homeless Count is the class label, and the other data, Feature Y, are the remaining features. One issue was that the class label was not divided into low, medium and high categories. This was achieved by first creating a boxplot **(Figure 5**), that showed the distribution of homeless counts. It was decided that low, medium and high thresholds would be based off the whiskers of the boxplot, since this would give the best even distribution between these three categories. Values lower or equal to 2.1 would be classified as low, values bigger than 2.1 and smaller or equal to 3.5 would be classified as medium, and anything bigger than 3.5 would be high. These thresholds were chosen as then provide a close distribution of values. Specifically, out of the 79 LGA values, 25 were mapped to low, 29 were mapped medium, and 25 were mapped to high. If the distributions were lopsided, results would be consequently skewed.

**Figure 5:** distribution of homeless data

Since the class label was finalized, 66% of instances would be training data, and the rest 33% would be testing data. Since the data was split into testing and training data, the data needed to be normalized. After the data was normalized, cross validation was used to determine the best value of k (**Figure 6**). It was found that 7 as the value of k, produced the best overall accuracy of 0.44 when comparing the prediction to the actual class label. Overall, all these results do show and predict that homelessness is an issue in Victoria’s LGAs, but some problems still occur. Outliers such as Melbourne and Yarra could skew results, and it was concluded that perhaps removing them from the data could show more accurate results. These strange results could also be from the variation of years used in the datasets, whereby some data is recent from 2011, and some data is from 2015, which could change how predictions are made. This study would be more accurate if the data had consistent years. In terms of the KNN analysis, it was clear that the model had the highest accuracy, when k was set to 7, which shows that the more neighbors were included, the more accurate the model became. This predicts that the homelessness problem in Victoria’s LGAs is from many causes, including the more less intuitive factors such as ethnicity.

**Figure 6**: Scatter plot of k versus prediction accuracy

## Value

This study provided a more innovative way to look at the causes of homelessness in Victoria. Without applying the preprocessing, integration and visualizations, it would be difficult to see other reasons why homelessness is getting worse. Given the way the mutli-dimensional data was given, it is not obvious to determine the correlations in the data, and any useful information would be skipped if not looked deeper into with software such as Python. By applying these methods, especially the visualizations, correlations between seriocomic/demographic factors and homelessness could be made easier to find for human interpreters.

## Challenges and Reflections

Some of the major issues I had were:

* Creating a clear research question. It was difficult to decide a research question that provided innovative data. I originally only compared homelessness with population, and realized that this would not provide any useful information.
* Finding data that had matching years. Many of the data I could find was scattered from 2011 to 2015, and this could skew results. It was assumed that the most recent data, even it was last updated in 2011, would be most useful for this study.
* The KNN model needs more work, as it does not provide a fool-proof prediction of homelessness and its associating factors. It does provide a base to work from this issue, but it is far from making any final predictions.
* A substantial amount of the data is estimated, which can provide inaccurate results.
* Despite this study showing correlations with socioeconomic/demographic factors and homelessness, it does not consider the fact that many homeless people could be constantly on the move from one LGA to another. Dealing with this issue would be out of scope of the study, as too many other factors are involved.

## Question Resolution

The study question was answered to the best of its ability, with the current resources available. Perhaps other models, such as K means clustering could provide more accurate results on the behavior of the features in this study. Further research and updated datasets would be needed to provide more useful information on this study. The results and data wrangled from study would be very useful to the Victorian Government, as it considers various factors than can potentially affect homelessness. Specifically, this can inform the Victorian Government about LGAs that need more help with factors affecting homelessness. This study demonstrates that many more non-intuitive factors contribute to homelessness, and we need to act now to resolve this issue.

## Code

The major libraries used this project were: *pandas, numpy, matplotlib, plotly* and *sklearn.* The code for the preprocessing and integration stages, in *prepintegration.py*, were written from scratch, with a lot of the code being inspired from phase 1. The only exception of code that was taken from public sites was the correlation matrix code. In terms of the K-NN modelling in *knnmodel.py*, this code was taken from Workshop 8, and referenced in the files.