The Effects of Primary and Secondary Education on Communities in Victorian Local Government Areas (LGAs)

Domain

The domain of this study primarily involves a combination of education and communities, with a minor focus on the economy and crime.

Question

This study aims to analyze education data across Victoria, and answer the **research questions**:

- *How does the number of schools in an LGA relative to its population, and the types of schools, change the impact made on surrounding communities?*
- *How do changes to education availability in LGAs effect local community well-being and economic/physical quality of life?*

To answer these questions, we must first define "local communities" in the context of this investigation as the population within each LGA. For the sake of maintaining focus, the quality of life indicators are restricted to economic and physical safety, and community well-being is measured through a combination of individual opinions and community actions per person in each LGA.

Datasets

- **Local Government Area (LGA) profiles data 2015 for VIC**
  The first dataset used in this study contains the 2015 profiles (released in November 2016) for each of the LGAs of Victoria. It illustrates the social conditions of each LGA, as well as factors that influence quality of life, such as employment, transport and social engagement. Collected from AURIN and attributed to the Government of Victoria – Department of Health and Human Services, the dataset is a CSV file and can be found at the following URI: https://data.aurin.org.au/dataset/vic-govt-dhhs-vic-govt-dhhs-lga-profiles-2015-lga2011
- **ERP by LGA (ASGS 2016), 2001 to 2016**
  The second dataset used contains the most recent Estimated Resident Population (ERP) by LGA, and is the official measure of the Australian population. Collected from the Australian Bureau of Statistics, this dataset is a CSV file, and can be found at the following URI: http://stat.data.abs.gov.au/Index.aspx?DataSetCode=ABS_ERP_LGA2016
- **School Locations – 2017**
  The final dataset details the information collected as part of the ongoing registration of schools in Victoria, which includes the name, school type, school sector, address, phone number, co-ordinates and local government area name. From the Victorian Government Data Directory, this dataset is a CSV file and can be found at the following URI: https://www.data.vic.gov.au/data/dataset/school-locations-2017

Pre-processing

The necessary processing was carried out by either using the Python library pandas, or through manual adjustments in Excel. In Excel, some column names were edited to be shorter and more readable, and the three CSV files were initially read into DataFrame objects using Pandas.

The **LGA profiles** dataset required little cleaning, since the AURIN user interface allows for the selection of specific attributes from the 402 that were made available. I therefore restricted my choices to attributes that were relevant to my question and avoided attributes with missing data, as having to remove LGAs for accuracy would skew my analysis, given the relatively small number of LGAs (79). Two of the attributes chosen was given as per 1000 people values, while the rest were given as percentages. The data for estimated homelessness per 1000 people was converted to a percentage of the LGA's population, as it will be used in visualization with other data given in a similar format. The data for total criminal offences was kept as a per 1000 people format, as percentage values would not be applicable in this context. Additionally, this attribute forms the crime aspect of my analysis, and as such would not be used in conjunction with the data stored as percentages of LGA populations. Finally, the homelessness attribute was changed to reflect percentage data instead of per 1000 data.

Both the **ERP by LGA** and **School Locations** datasets were similar in the way they were initially processed, because, unlike the **LGA profiles** data, the raw datasets had unnecessary data, which was to be ignored in the integration stage. Furthermore, there were minor discrepancies in the listed names of LGAs in all three datasets (such as "Latrobe (C) (Vic.)" instead of "Latrobe (C)" in the **ERP by LGA** dataset), which prompted adjustments to their names for easier future integration.

The data extracted from the **LGA profiles** were separated into three governing domains, namely crime, community and economy. Under crime, "Total offences per 1000" was stored. For community, "People with low English proficiency", "People who help as volunteers", "People who believe multiculturalism makes life better", "People who rated their community as active" and "People who attended a local community event" were stored. Finally, the economy domain contained "Homeless rate". These make up the **7 community indicators** I intend to use for a portion of the remaining analysis. The intention behind this separation was to analyse the relationships between school sizes and school access rates and each of the three domains, hopefully connecting these relationships in the final analysis to achieve a more complete look at the effect of primary and secondary education on local LGA communities.

Since the data on schools is given as a total number of school per LGA, and each LGA would have varying populations, having just the number of schools available to each LGA community would not be enough to adequately compare relationships between schools and their effects on communities.

$$\boxed{People\ per\ school\ in\ LGA = \frac{Estimated\ LGA\ Population}{Total\ Schools\ in\ LGA}}$$

Figure 1: *Calculation used for People per school in LGA ratio*

This ratio (**Figure 1**) effectively shows the number of people per school, meaning fewer schools increases the value, so a **higher ratio is indicative of fewer schools relative to the population.**

Integration

Since the LGA profiles dataset was already clean, I decided to use it as the base of my final integrated DataFrame. As all three datasets were read into DataFrames, the desired attributes of "total_schools", "GOV_schools", "INDEPENDENT_schools", "CATHOLIC_schools", "Primary", "Secondary" and "Primary/Secondary" from **School Locations**, as well as "Est_pop_2016" from **ERP by LGA**, could be added as new attributes to new clean LGA_profiles dataset. Since **School Locations** and **ERP by LGA** were not indexed by LGA names in the same way as **LGA profiles**, I used for loops to parse through each row of the two CSV files to extract to correct corresponding data that was then added to the clean LGA_profiles dataset. To connect the datasets, an inner join was made on the respective attributes that detailed "LGA Names", which were cleaned in the previous stage.

In the final integrated DataFrame, the **7 community indicators** illustrated above in the pre-processing portion were stored as attributes, to allow for efficient visualization. A possible limitation of my pre-processing methods was the variable years of data collection. Since the range was within five years, I decided that the differences compared to true values could be overlooked as the benefits of having complete datasets outweighs the negatives. Additionally, using an existing data structure as the base of the final clean structure could result in unwanted data remaining.

Results

➢ Distribution of schools and school types in Victoria. Schools to population relationships.

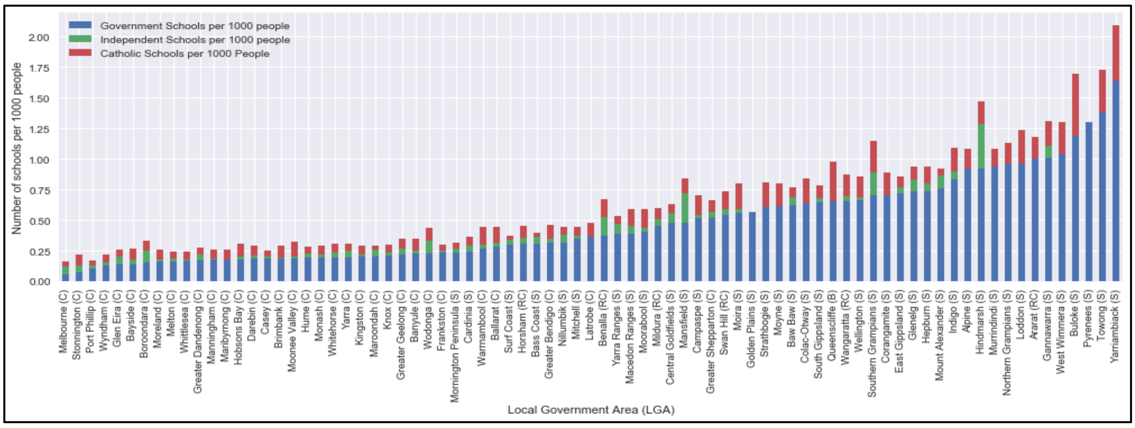

Figure 2: *Stacked bar chart showing distribution of school types per 1000 people*

The distribution of schools over LGAs (**Figure 2**) demonstrates the spread of schools as they relate to LGA populations. It is sorted by Government Schools per 1000 because they make up the majority, thereby showing how the distribution of other school types changes the ratio of government schools per 1000 people increases.

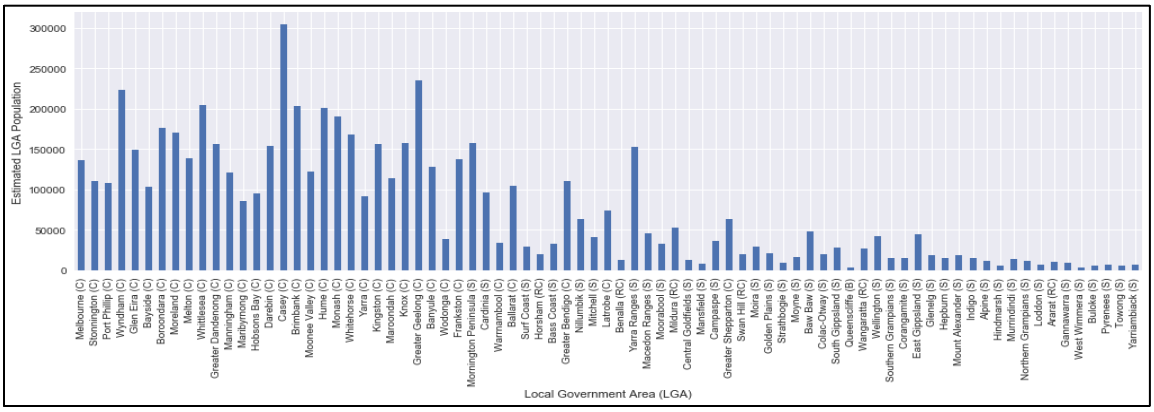

Figure 3: *Bar chart showing LGA populations, with LGAs sorted according to Figure 2*

Figure 3 shows an almost inverse trend to **Figure 2**, in that LGAs with lower total populations tend to have more schools per 1000 people. This suggests closer-knit communities in low population areas, which allows us to progress to the analysis of community indicators and how they relate to this.

➢ Scatter plots of "People per school in LGA" against the 7 community indicators.

The "People per school in LGA" values were calculated using the equation shown above (**Figure 1**).

Figures 4 and **5** concern the physical/economic safety quality of life indicators. The Pearson Correlations are moderate to strong, and suggest that, as the number of schools relative to population decreases, crime and homeless rates increase.

Pearson Correlation: 0.404212

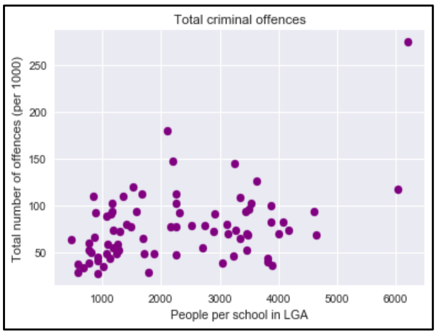

Figure 4: *Total offences*

Pearson Correlation: 0.549287

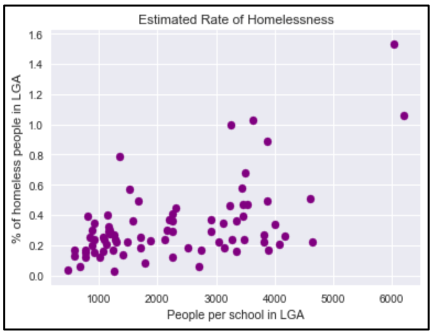

Figure 5: *Homeless rate*

Pearson Correlation: - 0.788744

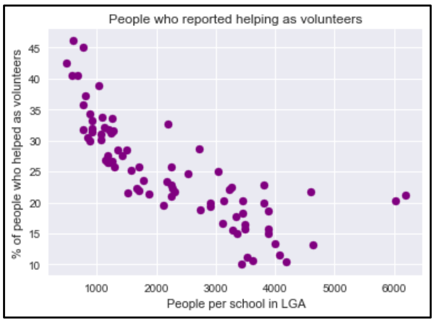

Figure 6: *Volunteering reports*

Pearson Correlation: - 0.79246

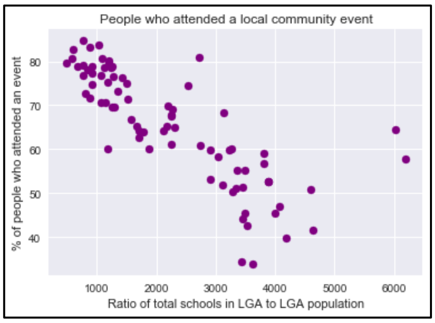

Figure 7: *Local event attendees*

Pearson Correlation: -0.811423

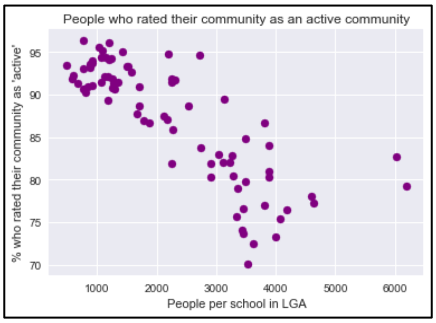

Figure 8: *People who rated communities as active*

Figure 6 has a strong negative relationship, and shows that the % of volunteering decreases as the number of schools relative to LGA population decreases (people per school in LGA increases). This means that a higher concentration of schools in an LGA correlates to more volunteers in the community. **Figure 7** and **8** both show similar strong negative linear correlations, and suggest that a decrease to the number of schools relative to an LGA results in decreases to the percentage of people that attend community events and people who rate their community as an active community. We can conclude from these findings that a higher number of schools relative to an LGA's population correlates to increased volunteering, attendance at community events and ratings of local communities as active.

Pearson Correlation: 0.643057

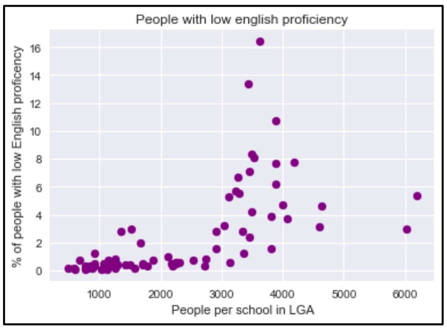

Figure 9: *Low English proficiency*

Pearson Correlation: 0.643019

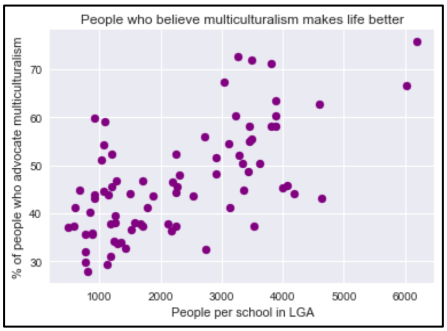

Figure 10: *Multiculturalism makes life better*

Figures 9 and **10** have very similar and strong correlation values, suggesting linear correlation. However, upon inspecting the scatter plots, there does not appear to be a direct relationship, unlike **Figure 7** and **8** above. This shows the importance of visualization for human analysis. It is suggested that, as the number of schools relative to LGA population decreases, the % of people with low English proficiency increases, and the % of people who advocate for multiculturalism increases.

- **<u>Heatmap to illustrate relationships between different school types and community indicators.</u>**

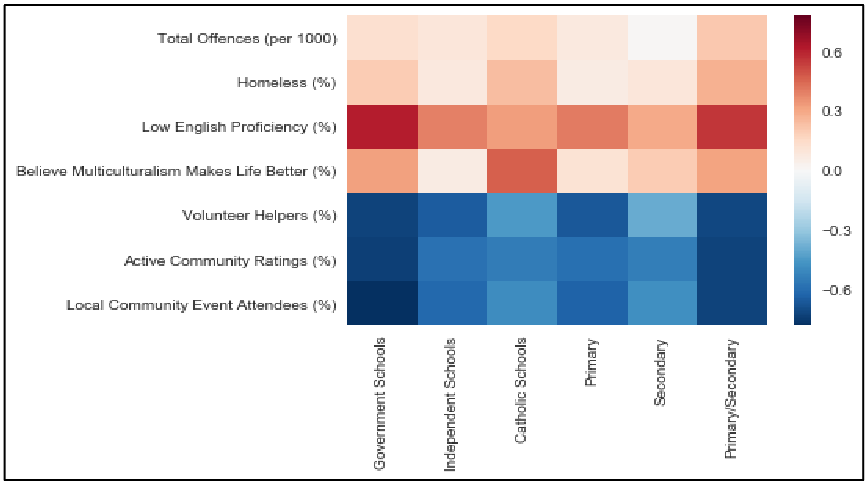

Figure 11: *Heatmap showing Pearson Correlations between each school type and community indicators*

Using the scatter plots above as reference, this heatmap (**Figure 11**) can be used to analyse the correlations between my selected community well-being indicators and different school types. Stronger Pearson Correlations indicate a closer similarity to the trend shown above for total schools in an LGA, for that specific school type. The interesting results will be explored in the **question resolution** section.

Value

Through the visualization of the correlations we derive from this integration (e.g. using a scatter plot with a Pearson's correlation coefficient), we are provided with greater insight into the strength of the relationships between the data sets. Visualization prepares data for human analysis, as it allows us to spot patterns, trends and outliers that we would otherwise be unable to see.

The pre-processing and integration portion of this report was crucial because it produced valuable and relevant information with regards to the initial research question. The raw datasets had an excess of unrelated information and various formats which, through processing and integration, were consolidated into a single data structure that was used to illustrate the correlations between school to population ratios and school types, and their implications.

Challenges and Reflections

A significant challenge in this report was maintaining focus on answering the proposed research questions, since having an aggregation of various community-related attributes made it easy to broaden the scope of the investigation too much. Additionally, I found that I did not have enough space to include all my desired visualization methods, including boxplots to detect outliers and exploring clustering methods. I decided that I needed to use the available space for visualizations that were more integral to answering the investigation questions.

I attempted to use VAT visualization to find clusters across combinations of my selected community indicators, however I struggled to produce any meaningful clusters, so I therefore avoided clustering as a method of finding relationships. The decision to analyse seven different community indicators may have prevented analysis as detailed as I would have liked, which means I should have kept the desired length of the final report in mind when deciding the scope of the investigation.

Question Resolution

To answer the first research question, we consider the analysis of **Figure 2, 3** and **11**. The bar charts suggest that LGAs with lower populations have more close-knit communities, due to the higher ratio of schools to population in these areas. How the types of schools influence the effects of education is seen in **Figure 11**. Government and Primary/Secondary schools have higher correlations since they are the majority of schools in Victoria. Independent and Catholic schools can be seen to change the correlation, where Catholic schools have the largest departure from total school correlations, showing the impact of school type.

As for the second research question, the scatter plots of "People per school in LGA" against my 7 selected community indicators demonstrate various interactions between school to population ratio and their effect on local communities. As detailed above, the Pearson correlation values ranged from moderate to strong, and show correlations that suggest the beneficial effects of a higher number of schools in an LGA relative to the population of said LGA.

Whether these correlations are indicative of causation is unclear, however this is the benefit of choosing multiple community indicators, that the correlations found can be aggregated in how they are considered within the domain of local community well-being. This suggests that a relationship does exist.

The results of this study could be used by policy makers or authorities such as the Victorian Government to benefit residents in each LGA by achieving general or specific goals pertaining to LGA communities through increased spending education, or changes to the education system. This is because improvements to citizens' quality of life, and by extension how close-knit local communities are, will always be an area of investment that garners government attention.

Code

Approximately **220 lines of Python code** was written (in a Jupyter Notebook) for the pre-processing, integration and visualization stages of this project. I used the declaration of the VAT algorithm in Workshop 6, as well as adapting some of the boxplot code from Workshop 5 when experimenting with different visualizations which amassed to a total value of about **315 lines of code**, however these methods were not used in the final report.

The Python libraries used for the final results were **NumPy**, **Matplotlib**, **Pandas** and **Seaborn**.