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

**Phase 2A: Concept Formulation and Preliminary Investigation**

**Domain**

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

**Question**

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

* Whether, and if so how, increased access to primary and secondary education is beneficial to surrounding communities?
* How does the school size / number of students in a school change its impact on surrounding communities?
* How do changes to primary and secondary education in LGAs effect crime rates, community well-being and economic indicators? How does this effect local communities?

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.

The chosen question provides innovative information in that it concerns an aggregation of various factors contributing to communities’ quality of life, rather than a single correlation to education. This explores the different impacts that education has on local societies, and how they can be broken down into their individual relationship.

**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>

**Processing, Integration, Analysis and Visualization**

The school and population data can be used in conjunction with each other to give a variety of values that can be integrated with the LGA profile dataset. Examples of this data includes the ratio of schools in an LGA to the population of that LGA, and the different kinds of schools that are documented (Government, Independent and Catholic). Visualizing the correlations we derive from this integration (e.g. using a scatter plot with a Pearson’s correlation coefficient), provides greater insight into the strength of the relationship between the two 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.

In this preliminary investigation I will use bar charts and scatter plots with Pearson correlations for a basic visualization of some of the correlations I intend to identify, however the remainder of the project is not limited to these forms of visualization. Heat maps, histograms and parallel coordinates are forms of visualizations that could prove to be useful in the final report.

**Summary of Initial Investigations**

**Preprocessing & Integration**

The necessary processing was carried out by either using the Python library, pandas, or through manual adjustments in Excel. The three CSV files were initially read into DataFrame objects.

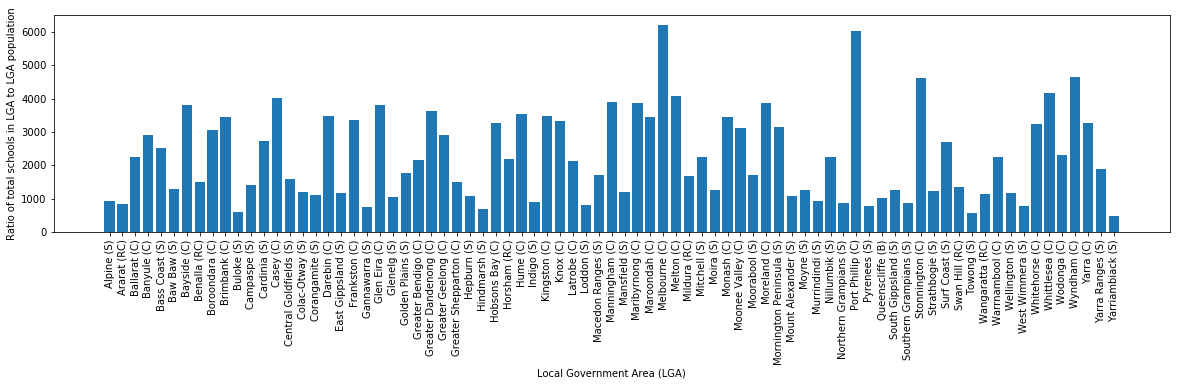
The **LGA profiles** dataset required little cleaning, since the AURIN user interface allows for the selection of specific attributes from the 402 that were made available. I therefore restricted my choices to attributes that were relevant to my question and avoided attributes with missing data, as having to remove LGAs for accuracy would skew my analysis greatly, given the relatively small number of LGAs (79). Four of the attributes chosen were 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 as percentage values. The data for drug use and possession offences, total criminal offences and family violence incidents were kept as a per 1000 people format, as percentage values would not be applicable in this context. Additionally, these three attributes form the crime aspect of my analysis, and as such would not be used in conjunction with the data stored as percentages of LGA populations. Some column names were edited to be shorter and more readable, and the homelessness attribute was changed to reflect percentage data instead of per 1000 data.

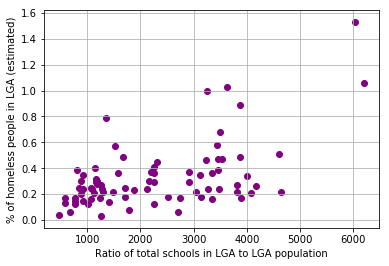
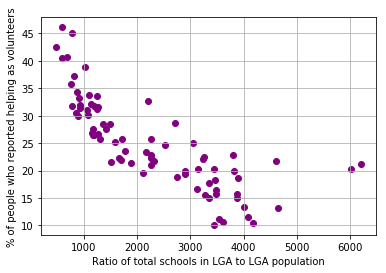
Both the **ERP by LGA** and **School Locations** datasets were similar in the way they were initially processed, because, unlike the **LGA profiles** data, the raw datasets had unnecessary data, which was to be ignored in the integration stage. Furthermore, there were minor discrepancies in the listed 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 to accommodate this issue for easier integration.

The data extracted from the **LGA profiles** were separated into three governing domains, namely crime, community and economy. Under crime, “Total offences”, “Drug usage and possession offences” and “Family violence incidents” were stored. For community, “People with low English proficiency”, “People born overseas”, “People who help as volunteers”, “People who believe other people can be trusted” and “People who believe multiculturalism makes life better” were stored. Finally, the economy domain contained “Households with mortgage stress”, “People with income below AUD $400 per week”, “Unemployment rate” and “Homeless rate”. 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 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.

**Initial Visualizations**

The bar chart below illustrates the ratio of schools in each LGA to the population of that LGA. It effectively shows how many people there are in an LGA for each school, and the variation in ratio between each LGA. This ratio is used below for the scatter plots, and the ratio indicates the number of people per school, meaning a **higher ratio is indicative of fewer schools over the population**.

 Pearson Correlation: 0.549287 Pearson Correlation: - 0.788744

**Left**: The % of homelessness

increases as ratio increases.

Less schools = more homeless.

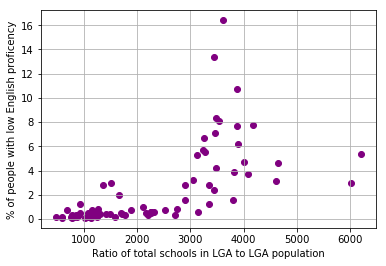
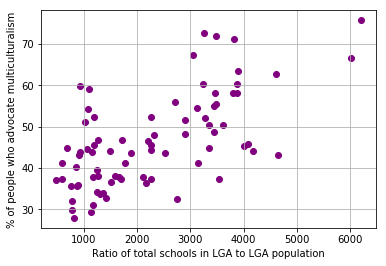
**Right:** The % of volunteering

decreases as ratio increases.

More schools = more reported

volunteers.

Pearson Correlation: 0.643057 Pearson Correlation: 0.643019

**Left:** The % of people with low

English ability increases as the

number of schools decreases.

**Right:** The % of people who

advocate for multiculturalism

increases as number of schools

decreases.

**Why the remainder of the project is feasible?**

As can be seen by the simplified explanation of the scatter plots above (shortened for brevity), the Pearson correlations all range from moderate to strong. Given that these are only four of the attributes chosen, the established correlations are promising for the remainder of the project. I said that an increasing **ratio** means that there are less schools over the LGA population, however this could also be indicative of larger schools enrolling more students, opening another possible interesting avenue of analysis, the size of schools. Overall, the relationships above suggests that answering my proposed questions is entirely feasible, as the previously mentioned domains of community, crime and economy seem to able to be adequately explored in the final report, and can therefore be brought together to determine the cumulative effects on local communities.