ELEMENTS OF DATA PROCESSING: PHASE 2

ANEESH CHATTARAJ (826860)

POPULATION AND EDUCATION

Domain

The two domains of this project are education and community.

Question

The report aims to answer the following question:

"Do the no. of schools in a Local Government Area have a positive correlation with the population of the Local Government Area?"

This would be informative to the Victorian Government and also people starting new families who would want to live in suburbs or areas where they'll have enough options to send their kids to school in the future. This project will also investigate the growing population of the Local Government Areas in Phase 3 which would also be beneficial for the Victorian Government to plan accordingly.

Datasets

The datasets used for this project are listed as follows:

1. https://www.data.vic.gov.au/data/dataset/school-locations-2017
   This lists all the schools as of 2017 in Victoria by the Department of Education and Training. This data is updated every year. The format type is CSV which is good for data accessibility. This file consists of the school names and also other relevant data such as the Local Area Government it belongs to, education sector, Address and etc.

2. https://www.crimestatistics.vic.gov.au/crime-statistics/latest-crime-data/recorded-offences-5
   This data was mainly used for the Estimated Residential Population of the Local Government Areas which would be of use in the project.

Benefits of processing, integration, analysis and visualization over raw data

Raw data can be very difficult and mind-numbing when trying to understand it to draw conclusions about the provided data. When looking at all the raw data at once, it is difficult to interpret what it means. This is where data processing and integration come in, with the help of various programming languages like R, Python, SQL, etc. along with their various built in functions it becomes easy to play with raw data which would in turn help us understand it better. These can help us to group data, visualize it in the form of various graphs.

Processing data can also help get rid of unwanted attributes or data contained in it, which would make it easier for us to understand. Help us perform various functions make new data out of existing data by comparing other data, sort them in a particular order and etc. All this would not only ease our understandability of the provided data but reduce cost as well which is really important for very big firms. It also helps in tackling errors which might have been made during the creation of the raw data. All this could help firms and even governments work more efficiently.

Investigation

The two datasets were extracted, and preprocessing of the data was done using pandas library for Python. The datasets were reasonably large, so a data cleanup was done to get rid of any unwanted columns and rows. Along with this some of the extracted column names were edited to get rid of spaces between words like 'Local Area Government' and 'Education Sector'.

In the two datasets some of the data in the 'Local Area Government' column was transformed to provide consistency by eliminating unwanted parts of the column values. This was also done with panda and string functions. This provided smooth integration between the two datasets. All missing or 'Nan' values were either removed or changed to '0', to provide consistency and avoid errors in the code.

Integration: The Local Area Government name was used in the ***'groupby'*** function to group and count the number of schools in that Local Government Area. The 'ERP' was divided by 1000 to get estimated population per 1000 people so that we don't get high plot values in our results.

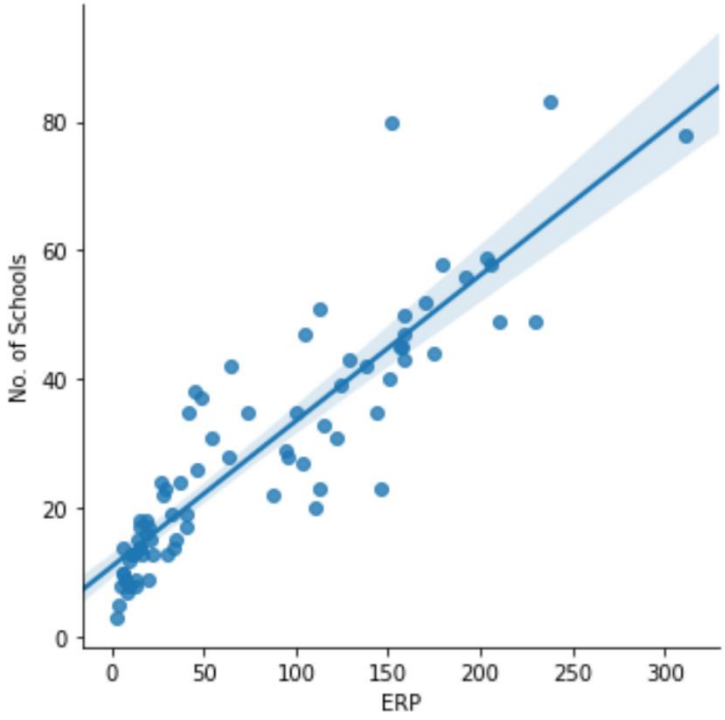
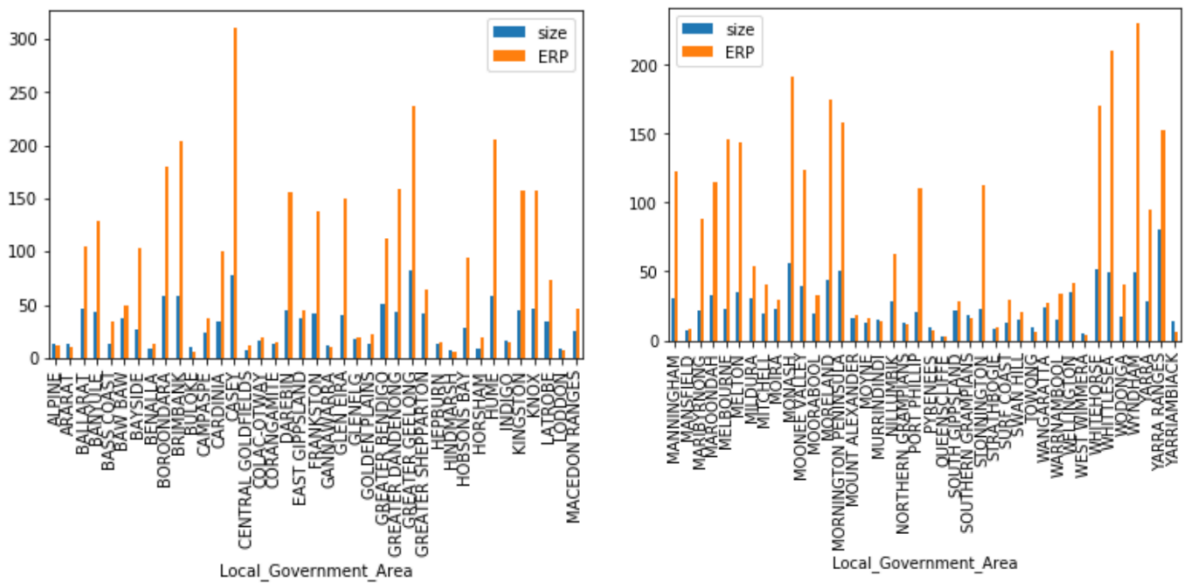


FIGURE 1

Figure 1 shows a relation between the no. of schools and the population in a Local Area Government. This clearly shows that higher the population, higher the no. of schools but this isn't true in all cases even if it shows a positive line of best fit. Some of the Local Area Governments with a similar population have a higher number of schools in their LGA and some areas with high population have lower number of schools. This could be due to various factors like the size of the LGA, average age of the people living in that Local Area Government, etc.



The figures above compare the no. of schools in a Local Government Area and the estimated residential population per 1000 people of 79 LGAs. It shows a trend between the two values. This visualization helps people who want to start early and live in Local Government Areas which has more schools.

Why the Project is feasible?

Many part of this project have still haven't been investigated like the estimated rise in population by the year 2031 in each Local Government Area which would help the Victorian Government to ready itself, distribution of types of schools in the Local Area Governments which could help people looking to start new families to settle early in the Local Government Area which has the most no. of preferred school type(government, catholic, private) and also point out where more new schools are necessary. Additional datasets will be used in future research which would help in getting a clearer visualization of the project. Hence, this project does have a future and is feasible to continue.