POPULATION/DRUG AND EDUCATION

DOMAIN

The domains this project is going to cover are community and education.

QUESTION

This report aims to answer the following questions:

1. "Do the no. of schools in a Local Government Area have a positive correlation with the population of the Local Government Area?"
2. "Does the rate of Drug crimes have a correlation with the no. of schools in the Local Government Area?"

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 other relevant data such as the Local Area Government it belongs to, education sector, Address etc. The dataset was published by the **Department of Education and Training** which is a reputable source.

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 and the Incidents recorded related to drug crime. The format of this dataset is XLSX and the dataset was published by the **Crime Statistics Agency** is also a reputable source.

EXTRA DATASET

https://www.planning.vic.gov.au/land-use-and-population-research/victoria-in-future-2016/victoria-in-future-data-tables

PREPROCESSING

Preprocessing of the datasets was done using pandas library for Python. All the datasets were scrutinized to provide consistency throughout the entire process of data mining. The datasets selected were rather large and had a lot of extraneous columns along with some data which prevented the data from good integration. To get rid of all this and provide consistency to the data and to prepare it for good integration data reduction or cleaning was carried out. There were some outliers in the data, but it was decided to leave them alone as they provided gainful information.

a. Data tables - Criminal Incidents Visualization - year ending September 2017.xlsx- The following was done to this dataset: 'Table 03and04' was parsed into a dataframe and then the estimated residential population data of each local government area from 2012 to 2017 was sorted into another dataframe. Column names were edited to eliminate all white space between words and replace them with '_'. Unwanted columns like 'Police Region' and 'Incidents Recorded' were deleted. All 'Nan' values were changed to zero

b. school.csv: The following was done to this dataset:Column name 'LGA_Name' was changed to 'Local_Government_Area' to provide consistency with the other datasets. Converted LGA names to uppercase. Unwanted parts from LGA names were eliminated like (C), (RS) and (S)

c. VIF2016_LGAs_VIFSAs_ERP_5yr_age_sex_2011_2031.xlsx- The following was done to this dataset:Rows were skipped to select the desired data. Unwanted parts from LGA names were eliminated like (C), (RS) and (S). Values were rounded to the nearest integer.

INTEGRATION

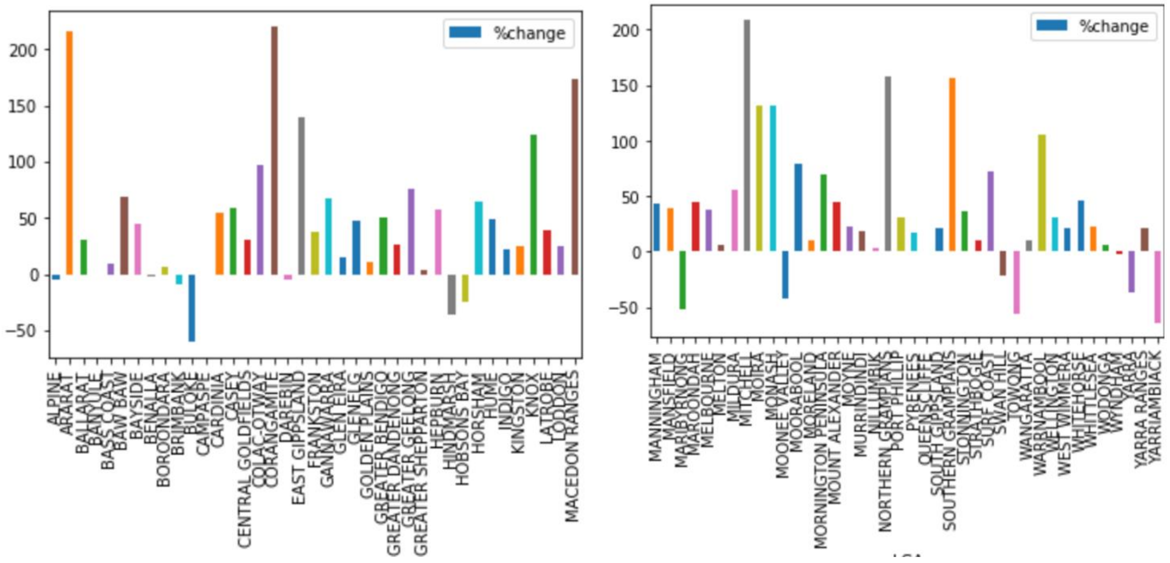
Thorough preprocessing helped in the smooth integration of data. To reduce number of columns in the dataframe the different type of drug offences were grouped together as one. All the local government areas which weren't present in all the datasets were removed with the help of looping and Boolean indexing. The local government areas were used as index in all the datasets to find correlation between the datasets. All the unwanted regions were removed like 'Unicorporated Victoria' to just focus on the 79 main local government areas.

Since we were covering the years 2012-2017 most of the columns in the dataframes were renamed to the appropriate year and then the data was stored under those columns according to their year and local government area**(index)**. To count the number of schools in a local government area the **.size()** along with **.groupby()** function was used. Plotting was done with the number of schools on the x-axis and the estimated residential population on the y-axis, a line of best fit was also plotted for a more comprehensible visualization. A scatter plot was done for all the different kinds of data and the correlation coefficient was also found for most of them which would help us gain valuable insight.

RESULTS

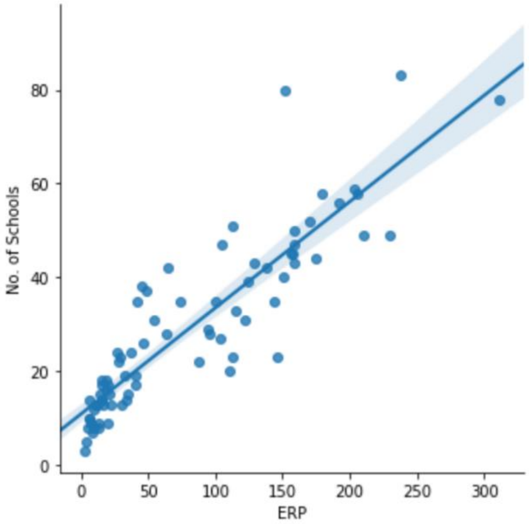
DRUG-CRIME RATE

Graph 1 shows the change in percentage of drug related crimes in the period 2012-2017, Victoria has experienced a 25.1% rise in the overall drug relate crime incidents. The graph provides information on the rate of change of crime(y-axis) in the corresponding local government area(x-axis). From the graph we can easily point out the increased crime rate in major LGAs.



Graph 1: Drug crime rate vs LGAs

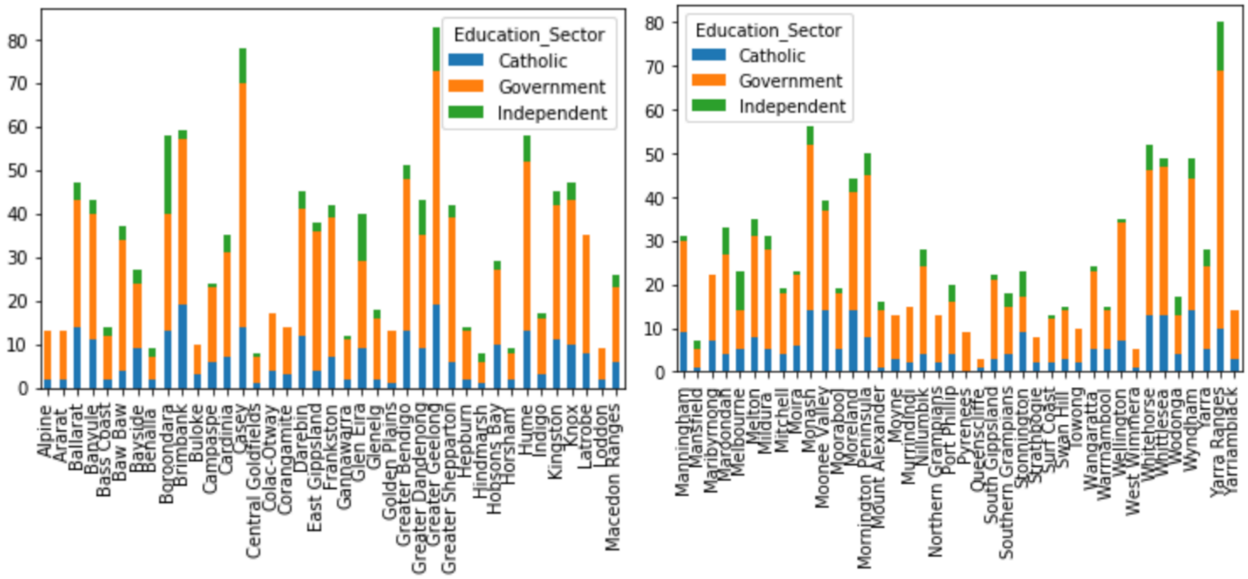
DISTRIBUTION OF SCHOOLS AND ESTIMATED RESIDENT POPULATION OF LOCAL GOVERNMENT AREAS



Graph 2: No. of schools in LGAs vs ERP

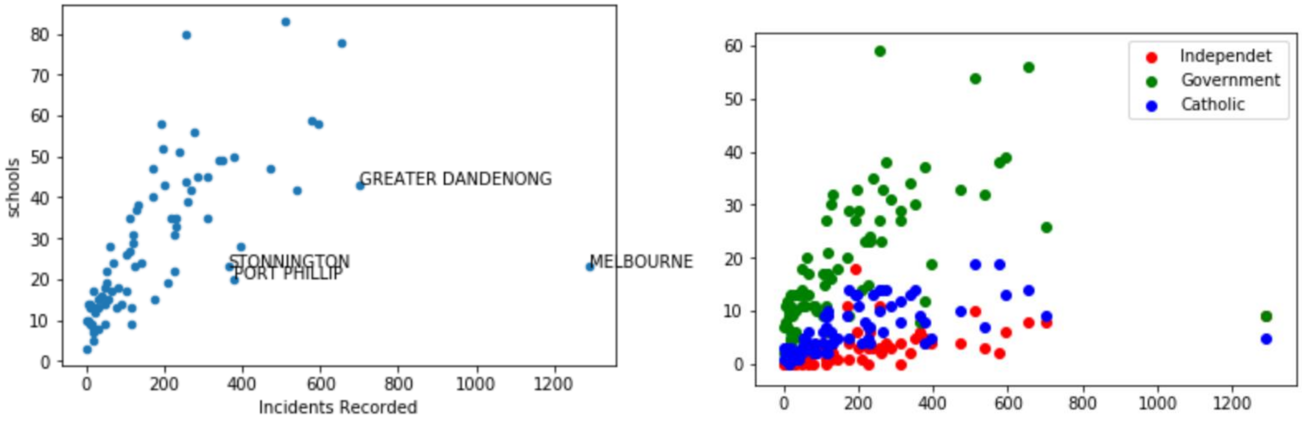
Graph 2 shows a relation between the no. of schools and the population in a Local Area Government. This clearly shows that higher the population, higher the no. of schools but this isn't true in all cases even if it shows a positive line of best fit. Some of the Local Area Governments with a similar population have a higher number of schools in their LGA and some areas with high population have lower number of schools. number of schools.

Graph 3 shows the distribution of different type of schools ('Catholic', 'Independent', 'Government'). This visualisation helps the Victorian Government to identify areas which need more investment in the education sector. Melbourne being the prime location in Victoria has less government schools compared to other less populated LGAs.



Graph 3: Distribution of schools

DRUG INCIDENTS VS SCHOOLS IN THE LOCAL GOVERNMENT AREAS



Graph 4: Number of schools vs Drug crime*(left)* and Types of schools vs Drug crime*(right)*

Graph 4 shows positive correlations between schools and drug related crimes. The left graph indicates that the higher the number of schools in the LGA, higher is the drug offence rate but there are some outliers which have been labelled. These outliers were not removed since they include major LGAs and high drug offence rates. The graph on the right plots different types of schools(y-axis) against offence rate(x-axis) and all the three kind of education sectors have a correlation between the range **0.52-0.59** which are good correlations. Therefore, while there is sufficient information to prove that there is a correlation, this does not mean that the number of schools is the cause of high drug offence rate but since drug offence is a major problem related to teenagers this does throw some light on the matter.

VALUE

This project had many benefits over the raw datasets since they were indecipherable on their own. The drug offence related data was spread across the entire dataset which made it difficult to progress further with our study and so was the same case with the school dataset. By taking advantage of computer-aided data processing, over many data points were condensed into human-readable and easily visualised results.

All this helped in finding out areas where the government can invest more in respect to education and also at the same time fight the growing drug related crimes. The data integration also helped in finding a good relation between number of schools and estimated residential population for the year 2017.

CHALLENGES AND REFLECTIONS

There were a few hurdles which came up in my study:

- I originally planned to just focus on population and number of schools in a LGA, however during the course of the project, my assessors in phase 2 raised an interesting questions about the difficulties in creating new knowledge. This made me modify my question and bring new data related to drug offences into question. I didn't have to use any new datasets which saved time.
- Some of the LGA names in the two datasets were different because of which we had some inconsistencies during the initial stages of the project, but which were later resolved.
- The biggest challenge was finding the right datasets, which took me personally 2 days.

QUESTION RESOLUTION

Victoria is experienced a steady rise in crime rate since the year 2012 and drug related crimes have had a major contribution to it. Victoria is also experiencing a steady rise in population and in the next 14 years it is projected to increase by a million or more.

The first question will help provide information to the Victorian Government which would help them to find areas where they need to invest more in education 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 report clearly shows LGAs like Melbourne need more government schools (Graph 3).

Drug related crimes among school goers is a big problem, the second question will help us answer if drug incidents are more common in LGAs with more schools? Getting a correlation between the range of 0.52-0.59 is not negligible, this should help the Victorian Police and Education department to make improvements by introducing new programs.

CODE

Around 200 lines of code were written, some of the investigation was left out in the report from the coding bit since we had a page limit of 5 pages. All the important aspects of the investigation have been covered in the report. All code was in Python and every graph in this project was coded using the major Python libraries like Pandas, Numpy, seaborn and Matplotlib.

BIBLIOGRAPHY

- Crime Statistics Agency.
- Data.vic.gov.au