COMP20008 Assignment 2

Data Science Project Report

Steven Nguyen, Lina Zhu, Sen Turner

1 Introduction

The prevention and treatment of depression and mental health issues is increasing its significance in contemporary society. This raises this question: how do different childhood developmental factors impact an individual's susceptibility to depression? By asking this question we aim to identify and isolate external childhood developmental factors that may impact on the health and wellbeing of our communities. Our focus on childhood and youth development also allows us to understand the liveability of Victorian communities, as we will potentially be able to identify areas that need improved living conditions for young people.

To answer our question we focused on multiple datasets; the Victorian Public Health Survey, VCAMS (Victorian Child and Adolescent Monitoring System), AEDC (Australian Early Development Census) and the Education State. These allow us to examine a variety of factors influencing school aged individuals. Each of these datasets was recorded by LGA (Local Government Area) or DHS (Department of Health Services Area).

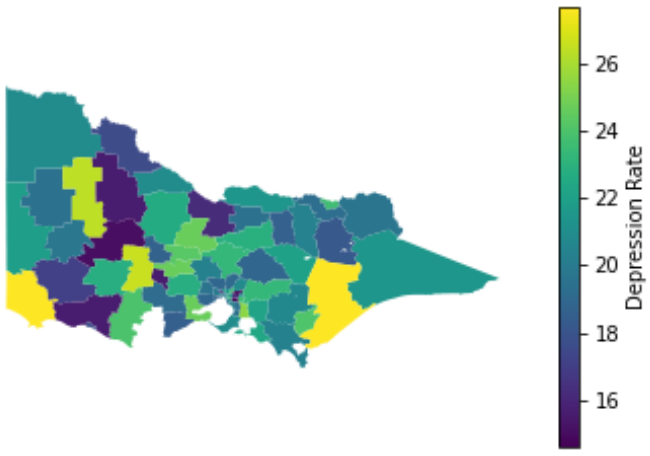


Figure 1: Depression Rates in Victoria by LGA

The Victorian Public Health Survey collects information about the health and wellbeing of adult Victorians, making it our depression rate data source. VCAMS tracks and measures a variety of indicators for young people's health, wellbeing, safety and learning. The AEDC dataset focuses on the development of children across Australia, recording the proportion of children that are developmentally at risk or vulnerable. The Education State dataset is focused on the Victorian school system, including data surrounding the number of enrollments, teachers and expulsions.

2 Wrangling and Analysis Methods

Because we use a variety of data sets, there were many data files we needed to process and collect into one table. Thus, automating workflows in Python was key. In the initial wrangling steps we first examined files to study the structure of our data sets before attempting to read them in via Python and Pandas.

Year	DHS AREA	Indicator_Calc	RSE
2013	Mallee	19.5%	13.7%
2013	Loddon Area	18.5%	14.1%
2013	Hume Moreland Area	17.8%	11.5%
2013	North Eastern Melbourne Area	16.7%	10.3%
2013	Ovens Murray Area	16.9%	13.7%
2013	Goulburn Area	19.0%	13.0%
2013	Outer Eastern Melbourne Area	21.6%	9.7%

Figure 2: Example structure of a VCAM spreadsheet.

One example issue we found was that in the VCAMs data sets, we found that while most files had a regular tabular format, a few files had differing column names or extra columns that made it harder to use the same function.

Year	LGA_KEY	LGA_DESC	INDICATOR_NUM_FULL	Numerator	Denominator	Indicator_Calc
2009_10	20110	Alpine (S)	20.1	12	2,549	4.7
2009_10	20260	Ararat (RC)	20.1	29	2,435	11.9
2009_10	20570	Ballarat (C)	20.1	158	21,730	7.3
2009_10	20660	Banyule (C)	20.1	119	25,696	4.6
2009_10	20740	Bass Coast (S)	20.1	81	5,796	14.0
2009_10	20830	Baw Baw (S)	20.1	57	10,021	5.7
2009_10	20910	Bayside (C)	20.1	32	21,795	1.5
2009_10	21010	Benalla (RC)	20.1	66	3,107	21.2
2009_10	21110	Boroondara (C)	20.1	63	35,534	1.8
2009_10	21180	Brimbank (C)	20.1	212	43,111	4.9

Figure 3: Example of a nonconforming VCAM spreadsheet.

Some other issues included -

- `.pdf` format data.
- Inconsistent area names. ('Greater Melbourne' vs 'Greater Melbourne *Area*')
- Non-numeric or missing data which had to be imputed.
- Teacher data being improperly split for Catholic schools (requiring collation of secondary and primary school data)

We used a combination of manual cleaning via Excel where it was trivial to do so or not worth automating, and Python for more automated cleaning, particularly for the VCAMs datasets of which there were many.

Data sets were linked together by merging features over common keys – either LGAs or DHS Areas together with respective depression rates in young people for each region. It was not of interest to aggregate LGA areas into DHS areas to combine the data sets as that would result in less data points in our plots which could potentially make us lose some information.

Data was then analysed by searching for correlations between depression rate and our chosen features in order to see whether or not youth experiences dictate future depression rates in adulthood. We also wanted to see whether or not childhood features were correlated with each other to see if there were any underlying systemic issues. To achieve this we made a

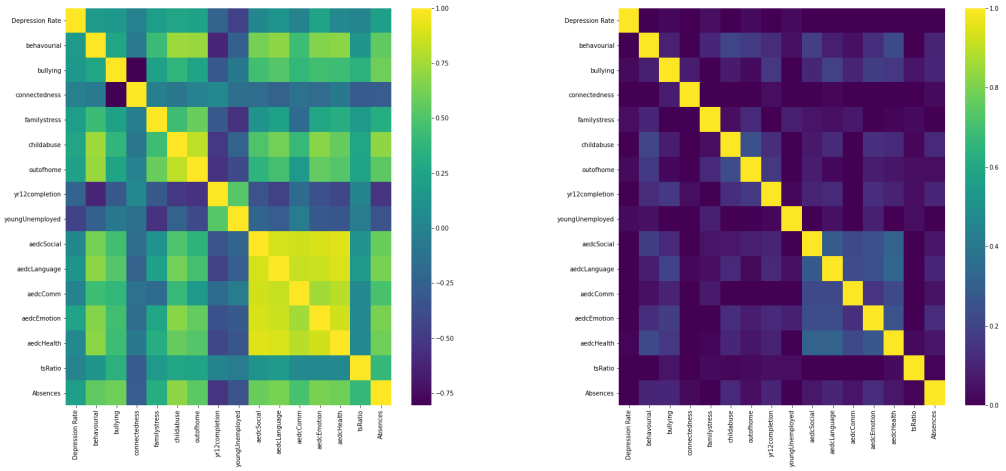
pairs plot between all the relevant features initially to see if there were any correlations at all in order to guide further study. We then calculated pairwise correlation coefficients and mutual information scores, and used a variety of visualisation tools on our data.

It was also of interest to see which LGA areas were 'most liveable' overall for youth. This was done by normalising each column in our tables by first assuming a normal distribution, then calculating a Z-score. This was done over simple min-max normalisation in order to reduce the effect of any possible outliers and to ensure a relative scale for each column. For each LGA we took the average of these as a rough index for the youth liveability in that area, based on the given features. The higher the score, the better the region performs *relative* to other LGAs in Victoria.

3 Results

3.1 Correlation

Through our Pearson Correlation Coefficient heatmap we found overall low Pearson correlation values between rate of depression and individual childhood factors, suggesting minimal linear correlation. This was also demonstrated by our mutual information map. However, the heat maps highlight some clear linear correlations between childhood factors, particularly the AEDC indicators as well as bullying rate and connectedness to school[1].



(a) Correlation Coefficient Heatmap

(b) Mutual Information Heatmap

Figure 4: Pairwise correlation matrix heatmap plots.

The low correlation between depression and childhood factors was largely to be expected, as depression is caused by a multitude of factors and can't be expected to be linearly mapped by any individual variable. Thus, even small correlations are of interest. Rate of absences had the largest mutual information score of 0.1187, while rate of children at risk of development emotionally had the largest Pearson value, the relationships between these factors and depression are visualised below.

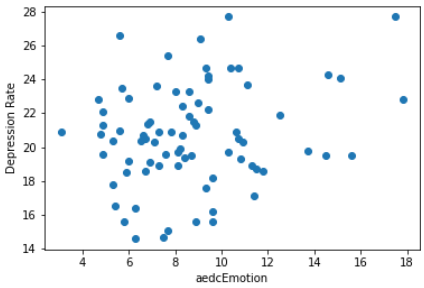


Figure 5: Plot of depression rate against proportion of children at risk (AEDC)

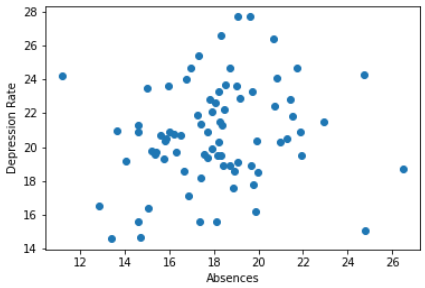


Figure 6: Plot of depression rate against abesences per 1000 students.

Visually there is a slight positive relationship between absences and depression. However, it

[1]Specific definitions for VCAMs factors can be found in the readme.

is difficult to identify whether this is a result of early onset depression influencing absences or absences increasing susceptibility to depression. Overall, it was difficult to say from these visualisations whether there are any worthwhile correlations between depression rates and any individual factors.

Through our parallel coordinate plots we were able to visually identify some strong linear correlations within factors - for example bullying, family stress and connectedness, as seen below. This potentially suggests that high levels of family stress or children lacking connection leads to increased bullying in schools. Low, medium and high in this plot refers to low, medium and high rates of depression in each LGA (which are represented by a single line).

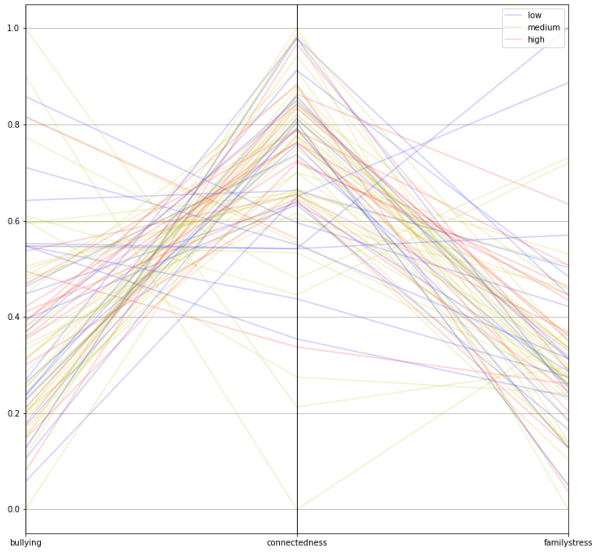


Figure 7: A section of a parallel plot between all the LGA features.

According to the pair plot of dhsData, the overall correlation is relatively low which is insignificant in concluding a clear causation relation between any of the factors. In general, positive correlation can be seen between factors that both contribute to positive impact (mental health access vs physical activities) or negative impact (psychological (stress) vs cyberbullying) on individuals. However the more interesting correlation between cyberbullying and depression rate shows a weak negative correlation which conflicts with expectation.

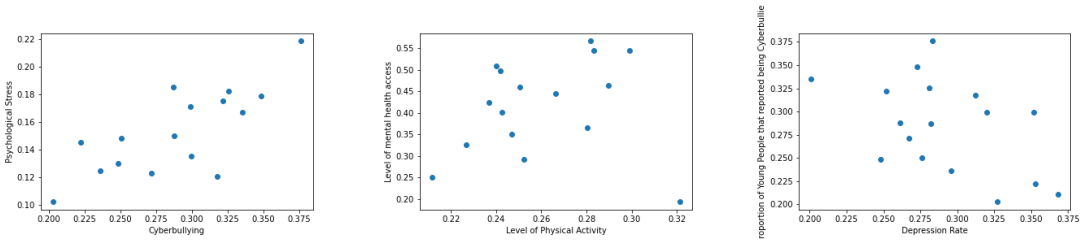
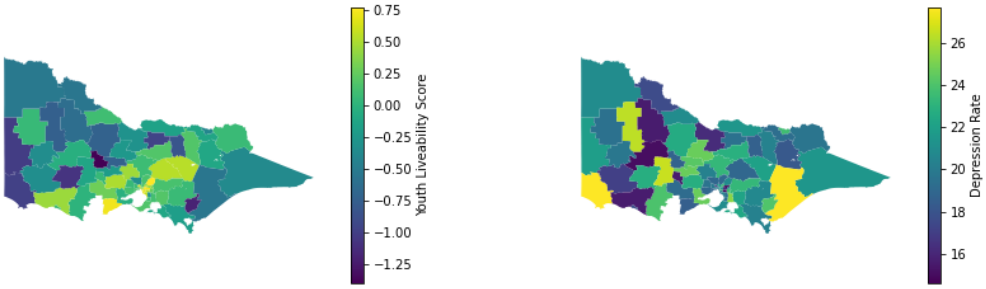


Figure 8: Comparison of DHS Data plots.

3.2 Youth Liveability Index

With our normalised 'youth livability score' and depression rates we produced separate choropleths. The visual congestion of high scoring areas in the centre of the 'youth liveability' map suggests that areas closer to Melbourne tend to be better for youth while more rural areas are worse.



(a) Youth Liveability Index Map

(b) Depression Rate Map

Figure 9: Comparison of depression rate against overall score for each region.

To make this clearer, we used K-means clustering with $k = 3$ in order to classify areas by the score.

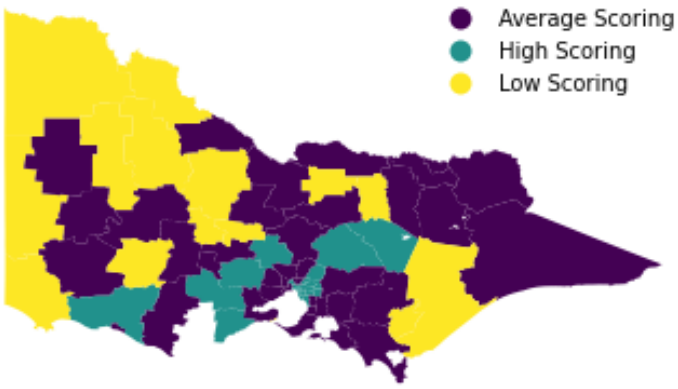
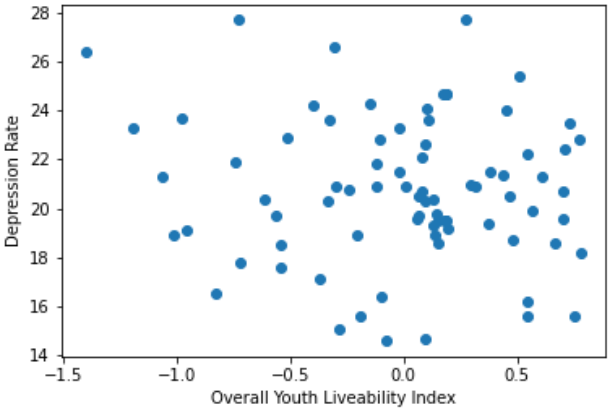


Figure 10: Map clustered by liveability score using K-means.

As we can see, there is a wide spread of areas that are performing about average, but low scoring areas tend to be concentrated in rural North Western Victoria.

Finally, we created a scatterplot comparing liveability index and depression rate. This graph demonstrates a slight negative correlation, albeit very wide, suggesting that increasing youth liveability may have a slight impact on decreasing depression later in life. However, the lack of strong correlation suggests there are more causal factors to identify than just childhood liveability.



4 Conclusion

Overall the data is not enough to conclude that there is any significant relationship between depression rate in adults and childhood factors. While there are some slight correlations that warrant further investigation there is nothing significant enough to make outright conclusions. Depression rates as seen on the map vary significantly between areas, and there are likely a lot of causes depending on the individual. Moreover, since we are simply doing an observational study of populations in an area, there is no way to account for any confounding factors - obvious ones being socioeconomic differences between rural and urban populations.

This analysis could potentially be improved by combining the DHS and LGA data through aggregation, meaning we don't need to analyse these factors separately, or simply by analysing more features. Additionally, ideally the liveability index would have weighting introduced, as some factors are more important when considering liveability. An increased focus on this overall welfare/liveability may also improve our conclusions. It is somewhat unlikely that any individual variable significantly impacts depression, but overall child welfare is more likely to show correlation. Similarly, it would potentially benefit the analysis to place additional focus on individuals rather than sole focus on populations in specific regions. This would allow us to look at individuals without the bias across a population in a local government area.

Based on the results we gathered as well as the limitations of our design, we cannot make strong conclusions as to whether depression rates and childhood factors are indeed highly correlated, despite most of the selected features influencing young individuals in a clearly negative way. However, the supplementary information gathered on liveability in regions for young, school-aged individuals is still valuable - showing a clear difference between rural and urban areas. Moreover we did find that quite a few features were highly correlated, strongly suggesting that one issue leads to other issues in a child's life. With this information we can pinpoint ways in order to improve young people's lives, particularly in struggling areas.

Links to Data Sets

https://www.education.vic.gov.au/about/research/Pages/vcamsindicator.aspx
https://www.aedc.gov.au/resources/detail/public-table-by-local-government-area-(lga)-2009-2018
https://www.education.vic.gov.au/Documents/school/teachers/profdev/careers/TSDR-2018-final-report.docx
https://www2.health.vic.gov.au/public-health/population-health-systems/health-status-of-victorians/survey -data-and-reports/victorian-population-health-survey/victorian-population-health-survey-2011-12
https://www.bettersafercare.vic.gov.au/reports-and-publications/vphs2018