**Effects of mental health on academic performance**

**2) Domain:** This study focuses on the domains ofhealth and education.

**3) Question:** The study between mental health and academic performance has the primary purpose of answering: ‘To what extent does the increase of mental health problems, whether it be student themselves with the mental illness or the society surrounding them, impact on a student’s ability to perform well in their academics?’ In answering this question, it is vital to explore the different sub-questions derived such as the effect of grade and how they are affected by mental illness so that more accurate deductions can be made.

**4) Datasets:**

<https://www.data.vic.gov.au/data/dataset/vcams-percentage-of-students-achieving-national-benchmark-in-literacy>

<https://www.data.vic.gov.au/data/dataset/vcams-percentage-of-students-achieving-national-benchmark-in-numeracy>

These 2 data sets show the proportion of students per grade (grade 3,5,7, and 9) who have achieved the national benchmark in literacy and numeracy respectively in the NAPLAN test within each LGA (local government area) of Victoria, of which there were more or less 79 LGAs. They had this data for each year from 2008-2005, hence storing 2530 rows of raw data. Both subject’s score were continuous data that had a typical range between 90-100. Both datasets also include summary data that is irrelevant to the study

<https://www.data.vic.gov.au/data/dataset/2015-local-government-area-profiles>

This data set provides a comprehensive range of data about each area in Victoria (approximately 400 attributes per LGA). However, the primary focus of this study will be based on the attribute ‘amount of registered mental health clients in each area’ which contains discrete values and has a typical range between 6-30. Again, each row is divided by 79 LGAs and has an irrelevant summary data.

**5) Pre-processing**

In initiating the data mining process, the datasets were converted from .xlsx to .csv using zamzar.com to be able to be read using python. Python was then used in all of the data cleaning process using its pandas library. The following was done in its cleaning process

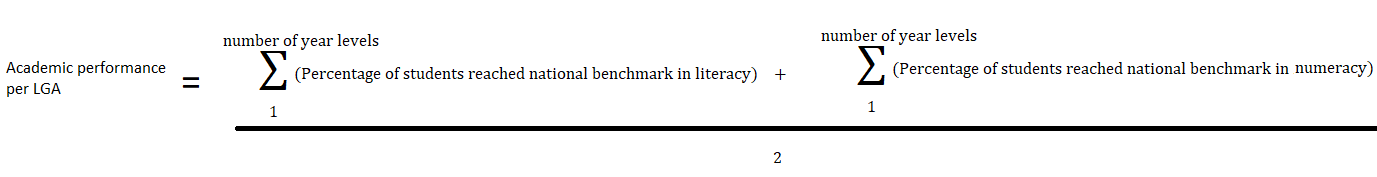
* Rows that provided irrelevant summary of the datasets were removed by examining where the raw data started and ended and then extracting only those data within that range. This was done to remove unnecessary information that could possibly affect integration and analysis of the data.
* Missing data were removed by detecting whether the relevant features (scores and LGA) were empty or labelled to be empty (labels such as ‘NDP’ and ‘Unknown LGA in Victoria’) and then removing them. The purpose of this was to uphold consistency and so that abnormal scores would not be generated, for example, if there were many and ‘Unknown LGA in Victoria’, then the total score generated from ‘Unknown LGA in Victoria’ would be abnormal.
* The literacy and numeracy scores had their ‘%’ sign removed and numbers be converted into integers so that the scores would have numerical value and for future mathematical processing and analysis.
* All LGAs on the three datasets were converted to lowercase so that there would be consistency in the case sensitivity which allows for the ease of future integration.
* Outlier detection was conducted by determining whether each academic performance per LGA was 3 interquartile ranges above the third quartile or 3 interquartile ranges below the first quartile. However, outliers were not removed as they provided interesting results and appeared to not have altered the results to a very drastic extent. Hence, outliers were only flagged and printed to identify the abnormal data.

**Challenges and limitations**

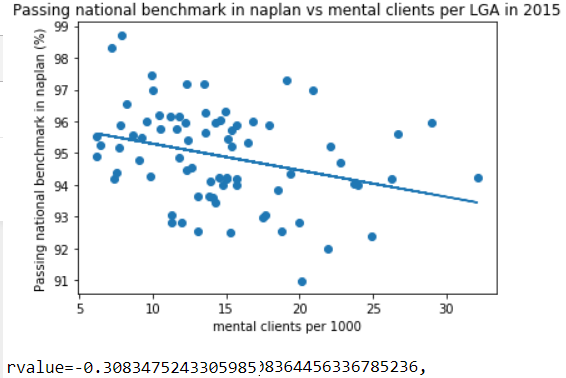
* The labels used to identify missing data (for example ‘Unknown LGA in Victoria’) was identified by manually examining the dataset. Hence, if there were different forms of identifiers for missing data (for example ‘unknown lga’), then there is a possibility that they have not been removed.
* Due to the nature of the study where the mean score of each LGA was the focus as opposed to individual scores, outlier detection was done during the integration phase after the ‘academic performance per LGA’ was determined. However, this implies that there could be outliers that largely affect the mean score of the LGAs.

**6) Integration**

Integration of the datasets required that the LGAs for each to be the same. This was done by first finding the intersection of LGAs between each year levels for the literacy and numeracy set. Then the intersection between the LGAs of the three datasets were generated and if any data point was not in the common LGA, it would not be processed to maintain consistency in the comparisons between LGAs. It was taken into account that this method had the possibility of removing too much data points but the value of keeping data consistent and accurate outweighed this limitation.

In this study, academic performance of each LGA is measured by finding the mean of the students who achieved the national benchmark in literacy and numeracy within each grade, that is

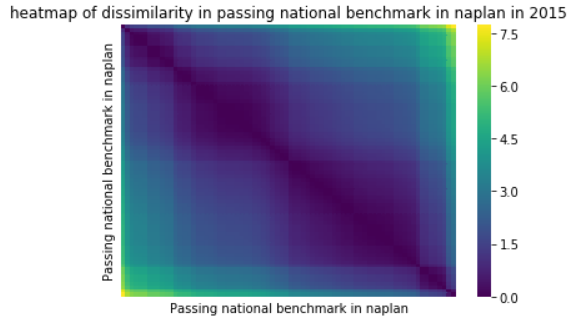
This was done in python by looping through the literacy and numeracy data sets and appending the scores into a list in a position where they are grouped up by year levels. Then the mean of each year levels was computed by adding up the groups of year levels (in both the literacy and numeracy dataset, there were 4 year levels per group) and then dividing by the number of year levels. Afterwards, to find the academic performance per LGA, the literacy’s and numeracy’s mean was added and then divided by 2. Once this was done, outliers could be detected and the advantages and disadvantages of doing this at in the integration stage was highlighted earlier.

The number of mental clients per 1000 for each LGA was then found using a for loop and appending the scores for each LGA. Now all the necessary components are ready to put into a data frame and hence the following data lists were put into the data frame: ‘LGA’, ‘Passing national benchmark in naplan’, ‘mental clients per 1000’, ‘is outlier’, ‘literature scores’, ‘numeracy scores’, ‘year 3 literature scores’, ‘year 5 literature scores’, ‘year 7 literature scores’, ‘year 9 literature scores’.

**7) Results**

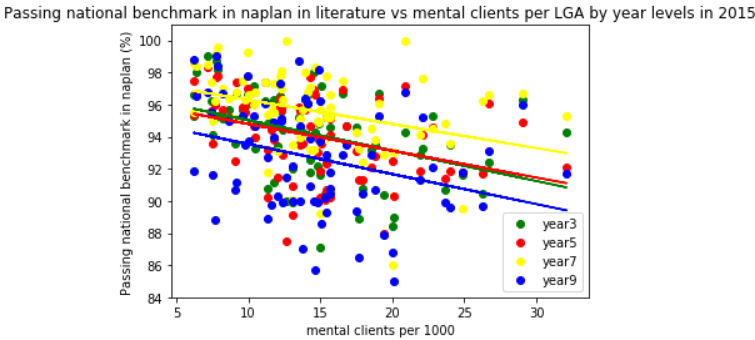
Using the data frame, the mental clients per 1000 and Passing national benchmark in the NAPLAN was visualised using a scatter plot for each LGA. Using the best fit function in python, a line of best fit was added to the visualisation to show the relationship between the two variables. Analysis of the data from the line regress function shows the Pearson correlation (rvalue). This information is summarised in figure 1. Additionally, outliers had their values printed though none was detected. The following information was deduced from this data:

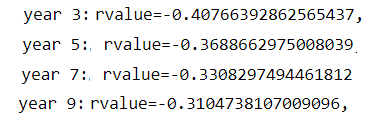
**Figure 1:** Scatter plot of academic success and mental clients

* As the number of mental clients within an LGA increased, the less likely students are to pass the national academic benchmark
* Although the Pearson correlation is not high, it is not insignificant data and suggests that mental health is indeed a small factor in passing the national benchmark.
* The lack of outliers suggests that the line of best fit is indeed representative of the overall data

For the purpose of identifying the key factors influencing the relationship between mental clients and passing rate, a heatmap was used to identify if there were any noticeable clusters within the data, as is shown in figure 2. This was done in python by copying the passing national benchmark rate feature, normalising the scores, sorting the data using the VAT algorithm then plotting the graph. The graph shows 4 distinct clusters on the diagonal. Using this data, it is reasonable to then investigate the effects of year levels on academic performance as there were 4 categories of year levels used.

**Figure 2**: shows the clustering of data using a heatmap



Likewise, an investigation on the effects of year levels was conducted. Figure 3.1 shows that year 7’s were most likely able to pass the national benchmark in literacy, year 3 an 5’s were moderately likely able to pass and year 9’s were least likely to pass within the year 2015. This graph also shows how mental illness effects all different types of year levels to roughly the same extent as indicated by the similar gradient in the line of best fit of each year level. However the more interesting data lies in figure 3.2 that shows the increase in year levels wears off the effects of mental health on a student’s ability to perform on the NAPLAN. Additionally, if we use both of the pieces of information from figure 3.1 and 3.2, we can deduce that the relationship between academic success and mental health is not affected by the passing rate. This can be inferred as the passing rate in ascending order is:

**Figure 3.2**: analysis of Pearson correlation between academic success and mental health

**Figure 3.1**: scatter plot of academic success and mental clients by year levels

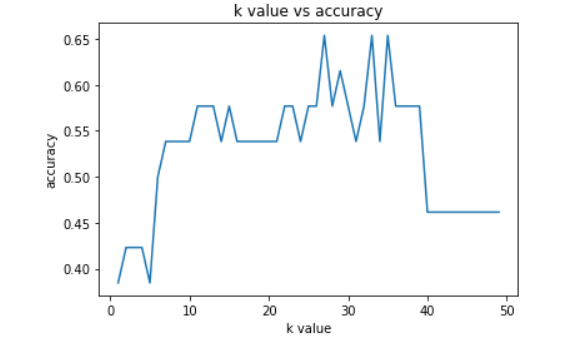
Year 7 – year 3/5 – year 9

While the strength of the Pearson correlation is in the ascending order of:

Year 3 – year 5 – year 7 – year 9

However, these deductions are merely based on the current data within the year 2015 and hence cannot be generalised without looking into the data and trends in other years. Furthermore, it was assumed that the literacy and numeracy datasets would produce the same results.

To test whether it is possible to predict the national benchmark passing rate of an LGA based on amount of mental clients, each data point was categorised by being flagged to either ‘is smart’ or ‘is not smart’, where ‘is smart’ is defined as a high number of students in the LGA that could pass the national benchmark (and is flagged as ‘1’) and ‘is not smart’ is the opposite (flagged as ‘0’). Strictly speaking, ‘is smart’ is defined by: ‘passing national benchmark percentage in an LGA’ > 95%, else, the data is flagged as ‘is not smart’.

After integrating the ‘is smart’ label into the data frame, the k nearest neighbour’s approach was used to predict the data. This was done on python by setting ‘mental clients per 1000’ to be the predictor of ‘is smart’ and then making 66% of the data to be the training set and then the rest to be the testing set. Normalisation of the data was then conducted. Since k nearest neighbours requires careful selection of k, a for loop was used to test the value of k from 1 to 50 and examine its effects on the accuracy of the prediction. The graph of this outcome is shown in figure 4.

**Figure 4**: a line graph depicting the effect on accuracy as the k value is changed

Using the knowledge that the k value should not be too high but also should try to maximise accuracy, a k value of 13 was chosen. Figure 5 shows the results generated at k=13.



**Figure 5:** accuracy of prediction to whether an LGA should be flagged ‘is smart’ at k = 13

Although the accuracy is not high, it is over 50% by a non-trivial margin given the size of the data set. The outcome indicates that if we know the amount of mental clients within an LGA, we are ‘more likely’ able to predict if the LGA will achieve an over 95% rate of passing the national benchmark. However, again, more data beyond the data within the year 2015 would be needed to gain a more precise ‘accuracy rate’ in prediction as the training and testing set is relatively small in comparison as to what is needed to make a concrete conclusion. Furthermore, the selection of the k value was based on human background knowledge and sight of the data visualisation and hence can be unreliable and inconsistent.

**8) Value**

These data sets have been processed, that is, by removing unhelpful data, in order to establish a focus on relevant details concerning the question at hand. In contrast, if the data was not processed, information that will not provide insight to the topic would only confuse and disturb the analysis of the data. The data sets have also been integrated so that the variables from different data sets are able to be compared with one another, hence with raw data, comparison between the large datasets would have been virtually impossible. Both analysis and visualisation allow people to see trends and hence for relationships between the variables to be seen. While with using raw data, it is extremely difficult to figure out whether there is a relationship between variables as the lack of consistency in the data hinders people from seeing patterns in the data. Collectively, these techniques have the ultimate effect of making the information that we extract out of the data more concise and clear, which in turn allows for precise conclusions to be made.

**9) Challenges and reflections**

The biggest challenges arose in the data pre-processing and integration stage where the large sets of data were just arbitrary and it was difficult to make sense of the data, hence difficult to know what to do in the next step. It was only until when the dataset was visualised where it could finally deliver meaning. However, even in the data visualisation process, it was difficult to know which type of visualisation was best suited for the type of information I had as factors such as ‘what type of information do I have’, ‘what type of information does the visualisation require’ and ‘what type of information does the visualisation show’ had to all be taken into account. This is exemplified when I tried doing a principal component analysis with my dataset where I assumed that I could use the graph because my academic performance score could be broken up into grades and subject. However, the principal component analysis required data to be divided by labels as opposed to being divided by features and as a result, I used a scatter plot and labelled each points by its category instead (completed work of this is **figure 3.1**).

**10) Question resolution**

The results attained shows that there is a good possibility that there is a relationship between mental health and academic success as the results show that there is indeed a correlation between the two variables. The approximate extent of the impact of mental health is also measured in this study as the Pearson correlation was consistently between the -30% to – 40% in both the total national pass rate and in all the pass rate by grade. Thus, Victorian schools and universities who are planning to implement or develop their support programs in helping disadvantaged students may use this information in determining which students might need help and to what extent. Hence both education facilities and students would be the beneficiary of this study.

**11) Code**

The line of best fit and its other operations in the main were sourced from the web while the k nearest neighbour and heatmap codes were taken from the University of Melbourne’s workshops and was adjusted to meet my data’s requirements. The rest were written from scratch using only python.

The major libraries used included matplotlib.pyplot which was used to create a large range of visualisations from the data. Additionally, the numpy library was pivotal in many of the mathematical functions and operations that was used in making the data. The pandas library necessary in many formatting, analysis and integration of the dataset.

**12) bibliography**

* Aziz Alto (20/1/17), ‘code for line of best fit of a scatter plot in python’, stack overflow,<https://stackoverflow.com/questions/22239691/code-for-line-of-best-fit-of-a-scatter-plot-in-python>
* Victorian Child and Adolescent Monitoring System (26/10/16), ‘VCAMS percentage of students achieving national benchmark in literacy’, <https://www.data.vic.gov.au/data/dataset/vcams-percentage-of-students-achieving-national-benchmark-in-literacy>
* Victorian Child and Adolescent Monitoring System (26/10/16, ‘VCAMS percentage of students achieving national benchmark in numeracy’, <https://www.data.vic.gov.au/data/dataset/vcams-percentage-of-students-achieving-national-benchmark-in-numeracy>
* DHHS (9/3/18), ‘2015 Local Government Area Profiles’, <https://www.data.vic.gov.au/data/dataset/2015-local-government-area-profiles>