COMP20008 Data Science Project Proposal

Steven Nguyen, Sen Turner, Lina Zhu

Mental health is an increasingly concerning issue due to rapidly growing rates of mental illness, with one in five Australians aged 16-85 experiencing a mental illness in any given year (Department of Health, 2009).

Currently, there is a great emphasis placed onto support groups, effective treatment and fast diagnosis. However, it is imperative that we look into mitigating the factors that place an individual in a vulnerable state in the first place. In our research we propose to look at how childhood education (6-18) and developmental factors affect the rate of depression in adulthood. We believe that this question is worth tackling as it has direct implications on the health and wellbeing of Victorians, as poor mental health decreases overall wellbeing. Moreover poor mental health leads to further issues especially at an early onset, and can be costly on both the community and the individual. We hope that the findings will be useful to educators, parents and the Victorian Department of Education and Early Childhood to adjust their approach to education and childhood development to best account for the future wellbeing of our young Victorians.

One data set which we will primarily look at is the Victorian Public Health Survey (2018) (VPHS), which contains a multitude of information surrounding health issues such as alcohol use, obesity and depression, split by factors such as age, region and socioeconomic status. This is in the form of an Excel (macro enabled) spreadsheet. The other data set is another Excel spreadsheet from the Australian Early Development Census (AEDC), which contains indexes of how children have developed physically, mentally, socially etc. by the time they reach school. Finally we will examine statistics from the Victorian Department of Education (VDET), focusing on factors we think may severely affect quality of education such as teacher demand, expulsion rates and class sizes. The data sets provided by the VDET come in inconsistent formats, including Excel spreadsheets and Word document tables.

We aim to link these data sets together by creating general aggregates over location, gender, and ethnicity, which will hopefully allow us to create differences and draw comparisons. For example, each data set listed above looks at measurements by area, and so we can combine common regions such as 'North-West Melbourne'. We expect to have some difficulties as some data sets will have different granularities or define different areas with different names.

For Excel data sets, we can convert the data using `pandas` and produce dataframes. However for certain VDET data sets, because they are in Word documents, we will first manually isolate tables of interest before using `pandas`. In doing the outlined process we hope to produce more informative tables that will have both educational factors and depression rates that can then be used for further analysis.

Once the data has been combined in tables we will emphasise finding correlation between our factors of interest and depression rates, meaning a primary focus on visualisation as our analysis methodology. This can be achieved through creating graphs with `matplotlib` like scatter plots to visualise the data. Scatter plots are ideal in that they can visually demonstrate correlation between education and depression that would not be clearly communicated in the raw data. This will hopefully lead to isolating areas that should be improved for the betterment of Victorian's mental health.

With the VDET data, effectiveness of schooling can also be looked at, which will provide more in-depth information behind the quality of schooling in Victoria, which of course can then be linked to the research question. For example we can look at parallel coordinate plots between factors such as expulsion rate and teacher-student ratios to see if there are any deeper links. By looking at certain

issues in schooling we can possibly identify systemic issues and other correlations which will give further direction towards the factors in education that influence mental health and aid our stakeholders (educators).

Along with challenges brought up throughout the proposal, potential challenges and risks might include the accurate identification of causations from correlated factors, we need to be careful in labelling something a causal factor due only to a correlation. More specifically, we will likely have to be careful of mutual causation and look at our analysis more critically. Additionally, the small range of available dataset might lead to incomplete conclusion of the question about the social phenomenon relating to the public's mental health status.

Links to Data Sets

https://www.education.vic.gov.au/about/department/Pages/factsandfigures.aspx
https://www.aedc.gov.au/data/downloads
https://www.bettersafercare.vic.gov.au/sites/default/files/2020-12/VPHS%202018%20Dashboard.xlsm