COMP20008 Elements of Data Processing

Assignment 2

Data Science Project

Group Number: 12

Group Member: Jiayu Wang 1039580

Dongyu Li 1058304

Dafei Lin 1043058

Hanqing Tian 1031969

Word Count: 1993

Comprehensive research to investigate factors contributing to mental illness across different areas in Australia

Introduction

As of 2019, around 45% of the Australian population has experienced mental illness to a certain extent and this number is still rising (Australian Institute of Health and Welfare, 2019). Researchers in this area have taken their steps focusing on factors that could explain this rising trend (**Fig 1-1**). And according to the Global Burden of Disease Study, Australia had the top share of the population with mental or substance disorders in the world (**Fig 1-2**). This phenomenon poses a threat to Australians' life quality, still, there is not enough evidence to explain why. Therefore, our team collected the proportions of mental illness and related potential risks in different regions of Australia in recent years and divided them into factors related to individual health and factors not related to health. This project aims to provide more evidence in this area so that further measures can be better taken.

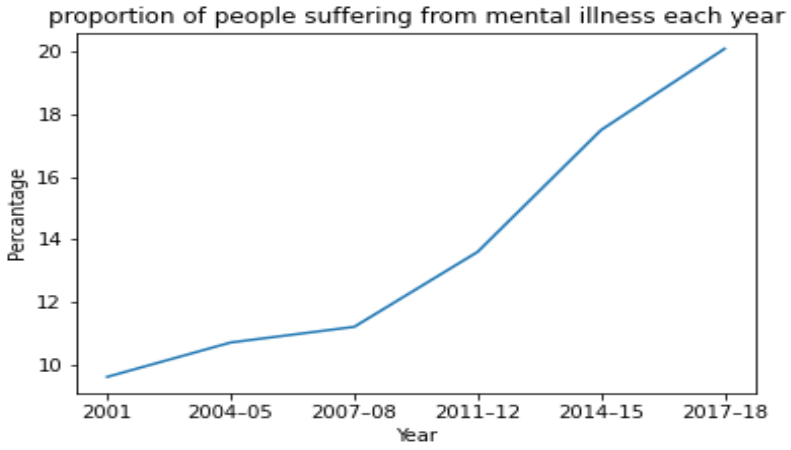


Fig 1-1. The trend of the proportion of people suffering from mental illness in Australia

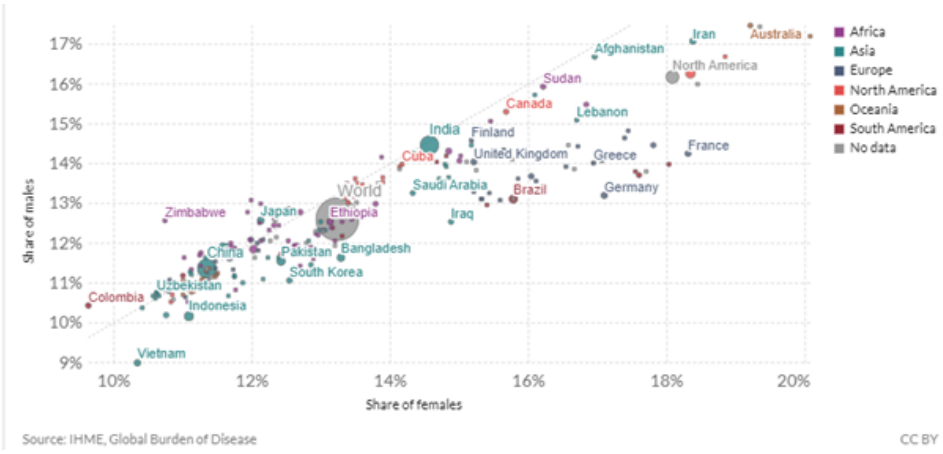


Fig 1-2. Share of population with mental or substance disorders, male vs. female, 2017

Datasets

The datasets used in this research are:

(1) Social Health Atlas of Australia by Local Government Areas (LGA)

This data set was published by PHIDU in 2021 which contains information on a range of population characteristics, including demography, socioeconomic status, health status and risk factors. It contains over 500 attributes for all LGAs across Australia.

(2) PHIDU - Admissions - Principal Diagnosis: Persons (LGA) 2016-2017

This dataset contains data regarding hospital admissions during 2016-2017 by principal diagnosis of mental health disease, nervous system disease and other diseases.

The two datasets share the same attribute, namely, the LGA codes for areas in Australia. Hence we used this to link the two datasets together.

Methodology and Results

Regression-based study

Pre-processing

We selected 15 potentially meaningful health-relevant variables (2017-2018) and 15 non-health-relevant variables (2016-2017). The response variables for health and non-health relevant feature variables are different, as the mental illness rate for 2017-2018 and hospital mental-illness admission rate for 2016-2017. This is because non-health feature data are only available for 2016-2017 (Australia does census every 5 years).

Feature selection

We calculated the correlation matrix (Pearson correlation) between variables and created heatmaps[1] (**Fig 2-1** and **Fig 2-2**). The reason for using Pearson correlation as the feature identifier is because both the independent variable and dependent variables are continuous variables and will be fitted into a regression model.

For health-relevant features, the dependent variable is %mental. We identified several correlated variables with correlation coefficients greater than 0.3.

For non-health related features we do not observe significant correlation between %mental. But the correlations between %people_in_the_social_housing and %persons_living_in_crowded_dwellings or %aboriginal_population_as_proportion_of_total_population are very high.

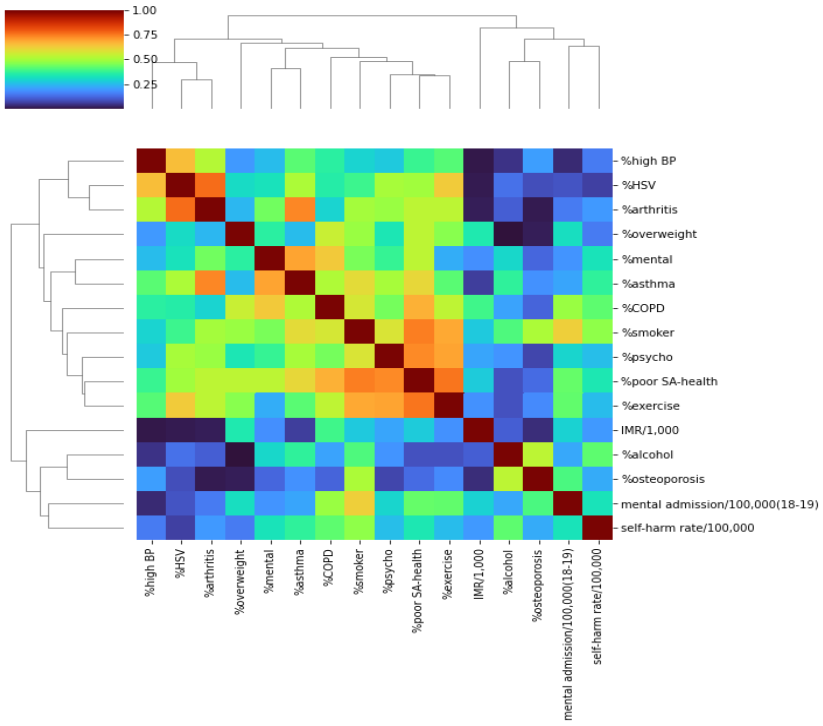


Fig 2-1. The heatmap of correlation coefficient among health-related features. (Note: see **appendix: abbreviation of variables**)

[1] see table 1 in appendix for correlation matrix

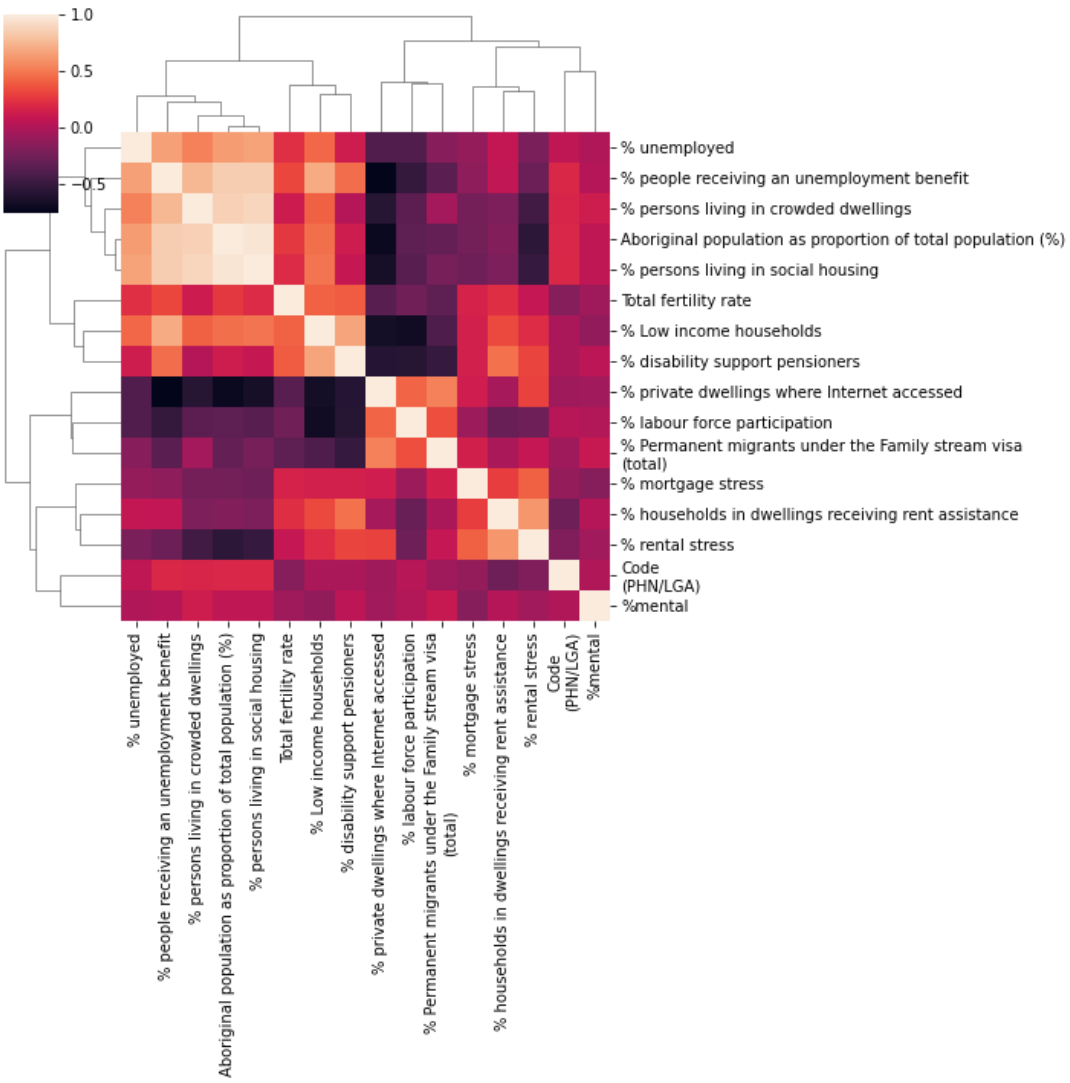


Fig 2-2. The heatmap of correlation coefficient among non-health-related features

Deciding suitable models to fit

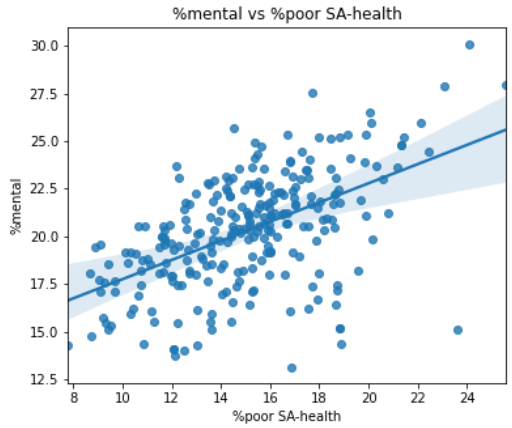
For health-related model, we first try out different regression models using unmodified variables and observe their fit (pseudo R^2). We tried 3 different assumptions of underlying distribution of our dependent variable namely Gaussian, Poisson and Gamma with corresponding pseudo R^2 that are 0.7896, 0.1385 and 0.6722 (**Supplementary Table 2**). We identified Gaussian seems to have the best fit and therefore build a OLS regression model.

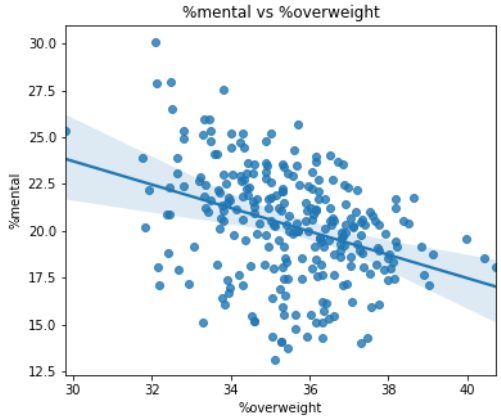
For non-health-related model, Lasso regression is used to test linear model fit and variable selection. 6 out of 14 variables are punished with the correlation coefficient of 0 (**Supplementary Fig 1**). The model yields a R^2 of 0.1215, this means even the remaining variables seem to have little impact on mental health.

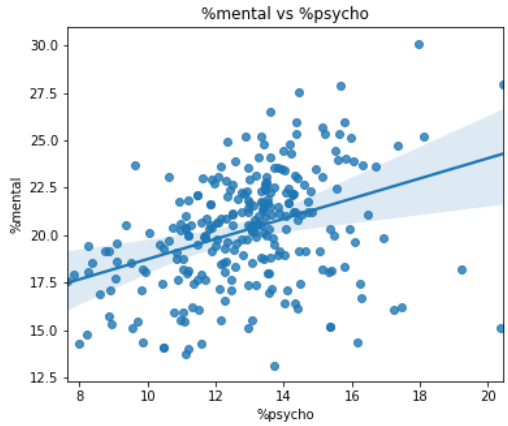
Plot each health-related feature against dependent variable %mental

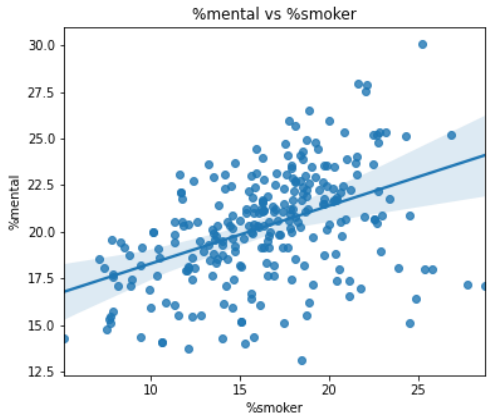
Results are shown below in **Fig 3-1**. From the graph we observe most variables have a linear relationship with %mental. However, we notice %alcohol and %smoker seems to have a non-linear(quadratic) relationship with %mental. We transform those variables by taking the natural logarithm as shown in **Fig 3-2**.

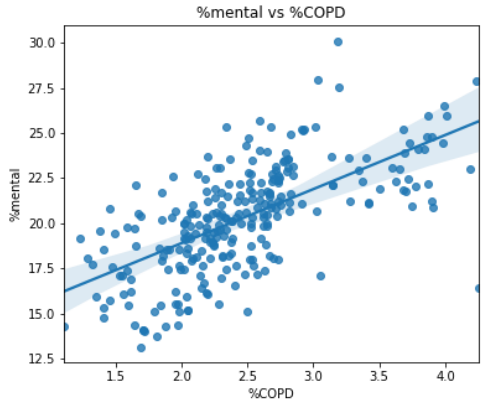
After log transformation, relationship between log_%alcohol and %mental still looks non-linear. We then construct a new variable %alcohol_squared(%alcohol raise to the power of 2) and use both %alcohol and %alcohol_squared later in the regression model. The intuition is such the coefficient on %alcohol will be positive and coefficient on %alcohol_squared to be negative to capture the initial positive but then negative relationship between %alcohol and %mental as %alcohol increase.

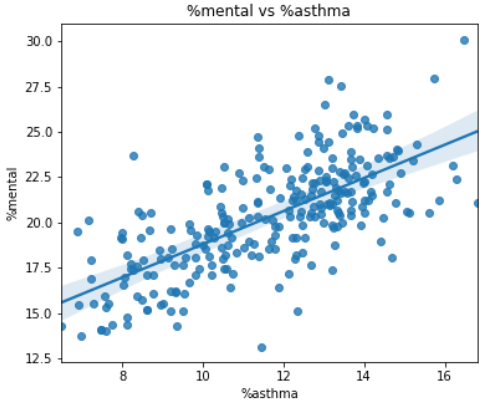












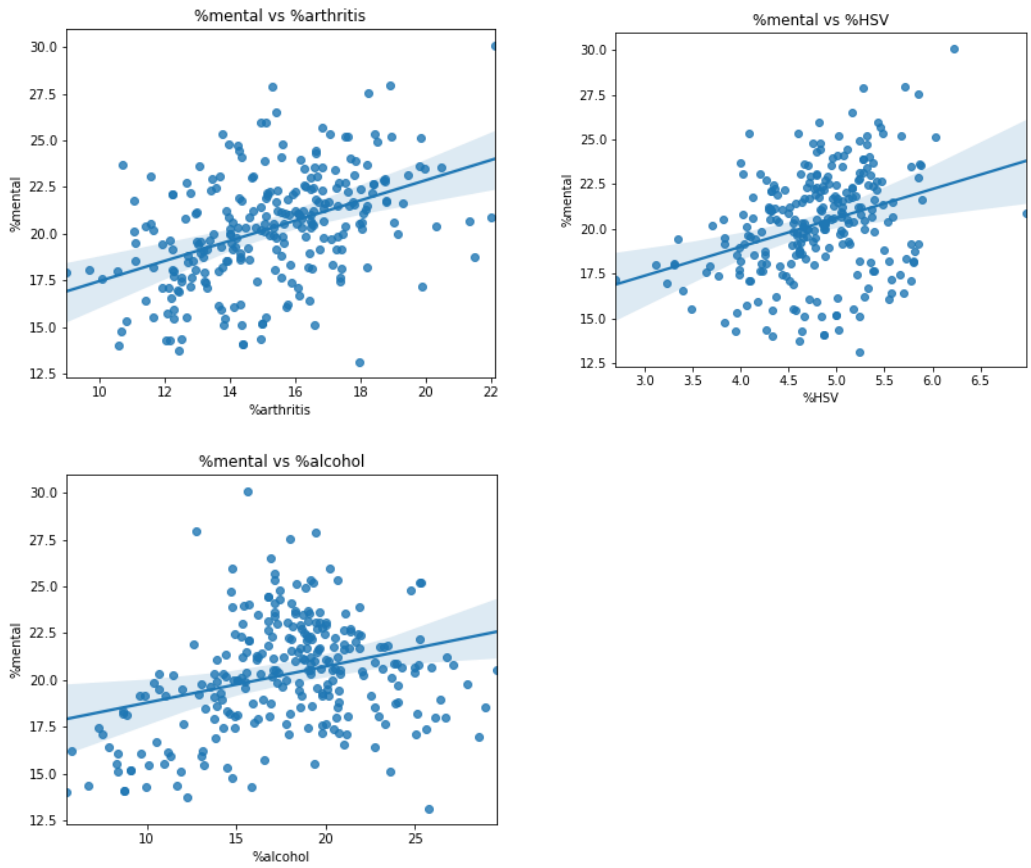


Fig 3-1. Linear regression plots

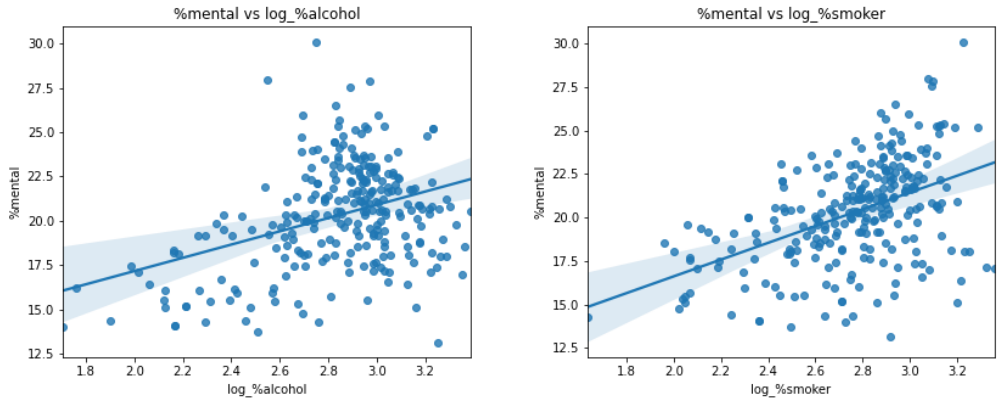
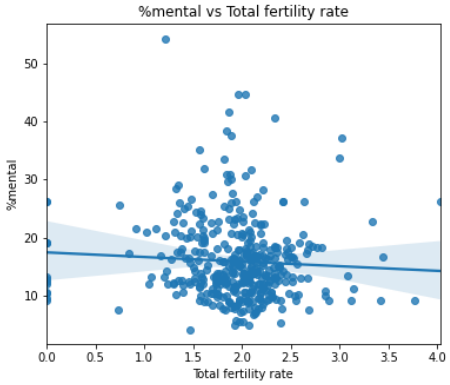
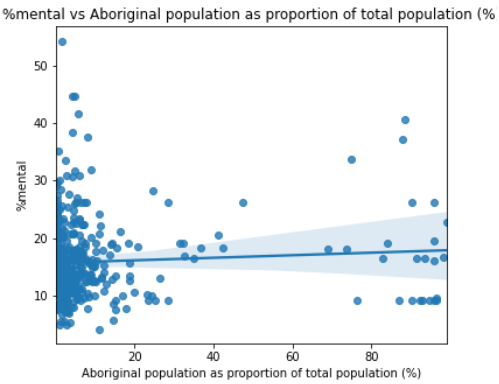


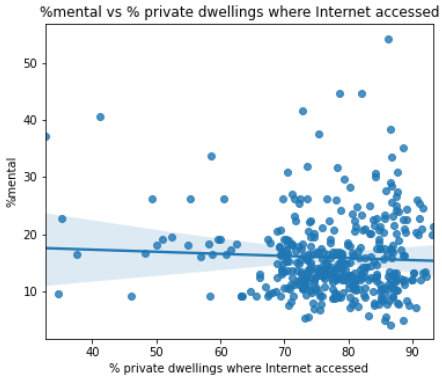
Fig 3-2. Linear regression between response variable and logarithmized feature varables

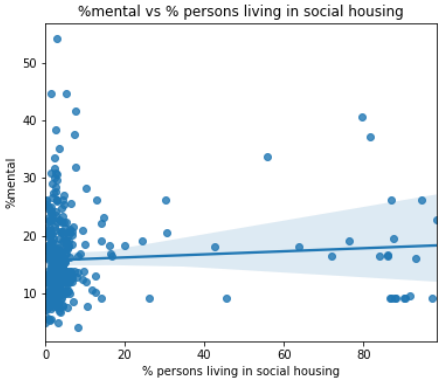
Plot each non-health related feature against dependent variable %mental

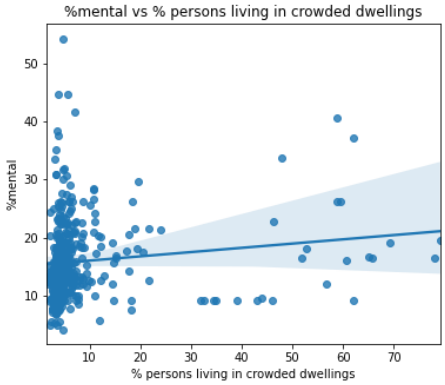
Fig 3-3 shows that most non-health related variables have no linear relationship with mental illness as we expected from the lasso regression. And it is easy to find that these data are distributed centrally.

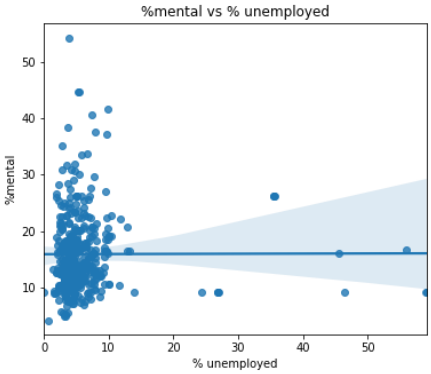

%mental vs Total fertility rate
%mental
Total fertility rate

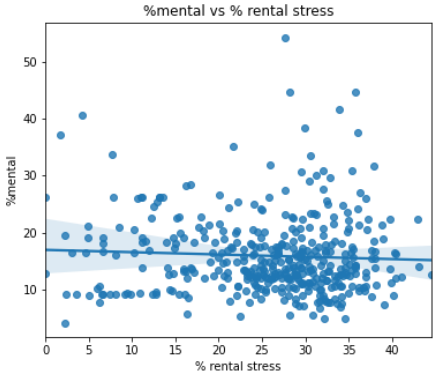

%mental vs Aboriginal population as proportion of total population (%
%mental
Aboriginal population as proportion of total population (%)

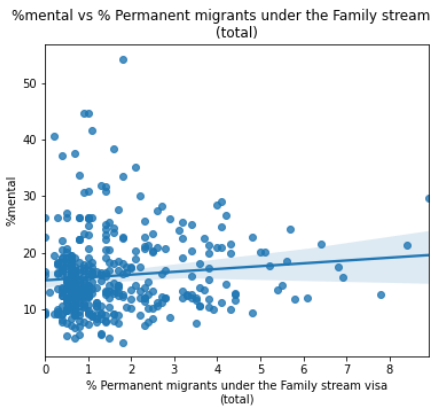

%mental vs % private dwellings where Internet accessed
%mental
% private dwellings where Internet accessed


%mental vs % persons living in social housing
%mental
% persons living in social housing


%mental vs % persons living in crowded dwellings
%mental
% persons living in crowded dwellings


%mental vs % unemployed
%mental
% unemployed


%mental vs % rental stress
%mental
% rental stress


%mental vs % Permanent migrants under the Family stream visa (total)
%mental
% Permanent migrants under the Family stream visa (total)

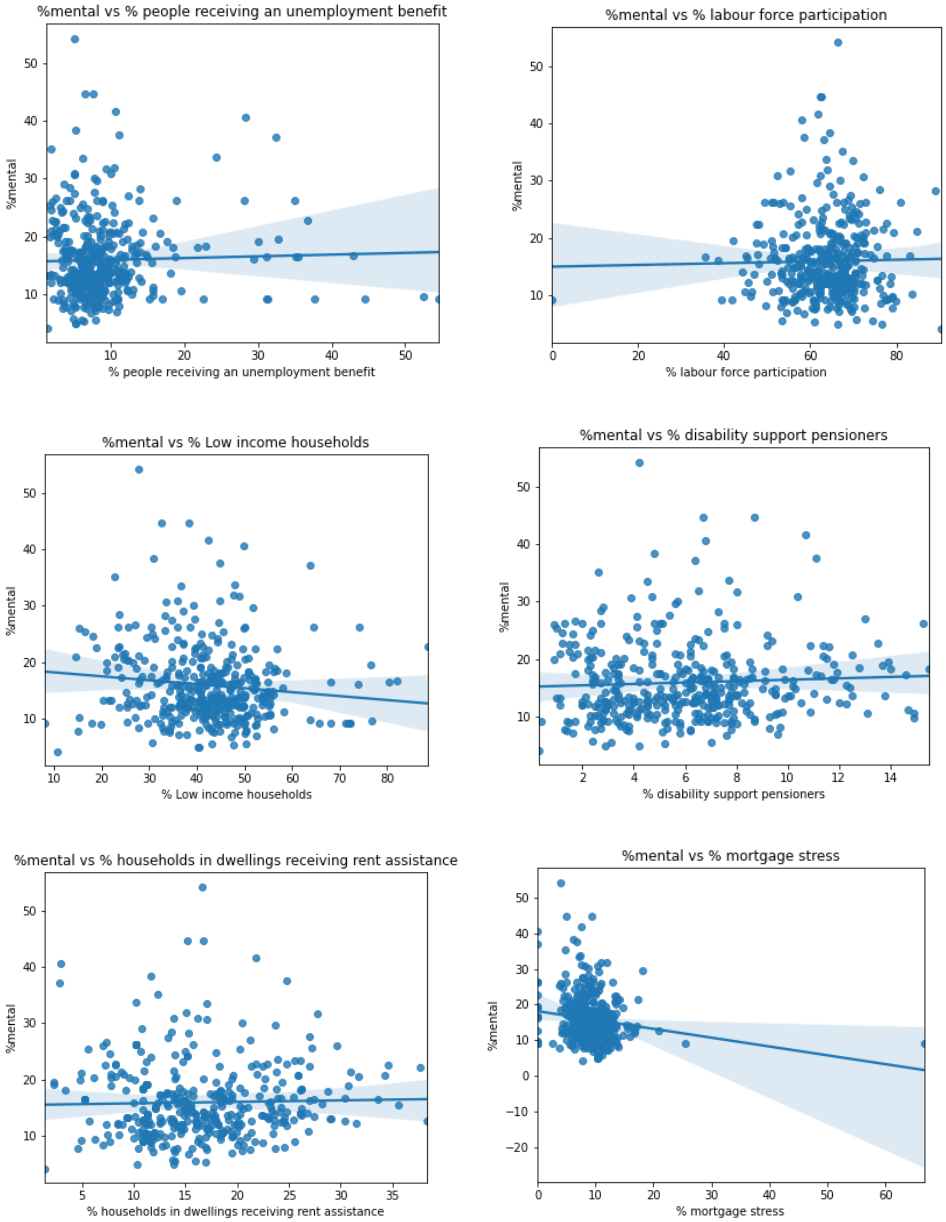


Fig 3-3. The linear regression plots for each non-health related features against %mental

Regression model result

The regression output utilises the full dataset with the purpose of finding significant variables that could explain %mental. We applied heteroscedasticity-consistent standard errors (White HC) to avoid potential bias in standard errors due to heteroskedasticity. Results are shown as **Table 1** and **Table 2**.

Table 1. Model coefficient estimates and significance tests of health-relevant features

OLS Regression Results						
Dep. Variable:	%mental	R-squared:	0.67			
Model:	OLS	Adj. R-squared:	0.657			
Method:	Least Squares	F-statistic:	58.08			
Date:	Sun, 10 Oct 2021	Prob (F-statistic):	6.62E-60			
Time:	0.153576389	Log-Likelihood:	-516.23			
No. Observations:	265	AIC:	1054			
Df Residuals:	254	BIC:	1094			
Df Model:	10					
Covariance Type:	HC0					
	coef	std err	z	P>|z|	[0.025	0.975]
const	5.576	4.576	1.218	0.223	-3.393	14.545
%poor SA-health	0.1962	0.078	2.51	0.012	0.043	0.349
%psycho	0.0686	0.098	0.702	0.483	-0.123	0.26
%overweight	-0.142	0.085	-1.673	0.094	-0.308	0.024
log_%smoker	-0.9725	0.941	-1.033	0.302	-2.818	0.873
%alcohol	1.1841	0.174	6.809	0	0.843	1.525
%alcohol_squared	-0.0313	0.005	-6.587	0	-0.041	-0.022
%HSV	0.3819	0.495	0.771	0.44	-0.588	1.352
%asthma	0.5292	0.114	4.661	0	0.307	0.752
%COPD	0.8236	0.308	2.676	0.007	0.22	1.427
%arthritis	-0.1258	0.143	-0.88	0.379	-0.406	0.154
Omnibus:	14.004	Durbin-Watson:	1.91			
Prob(Omnibus):	0.001	Jarque-Bera (JB):	27.282			
Skew:	-0.245	Prob(JB):	0.00000119			
Kurtosis:	4.493	Cond. No.	14900			

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)

Table 2. Model coefficient estimates and significance tests of non-health-relevant features

OLS Regression Results				
Dep. Variable:	%mental	R-squared:	0.19	
Model:	OLS	Adj. R-squared:	0.161	
Method:	Least Squares	F-statistic:	5.975	
Date:	Thu, 14 Oct 2021	Prob (F-statistic):	1.07E-10	
Time:	0.392708333	Log-Likelihood:	-1297.6	
No. Observations:	401	AIC:	2625	
Df Residuals:	386	BIC:	2685	
Df Model:	14			
Covariance Type:	HC0			
	coef	std err	z	P>|z|
const	24.6255	9.008	2.734	0.006
% labour force participation	0.036	0.046	0.787	0.431
% Internet accessed	-0.0856	0.085	-1.01	0.313
% persons living in crowded dwellings	0.1472	0.089	1.651	0.099
Total fertility rate	-0.4076	0.737	-0.553	0.58
% Permanent migrants	0.8441	0.357	2.365	0.018
%Aboriginal population)	-0.029	0.082	-0.352	0.725
%receiving rent assistance	0.0511	0.067	0.759	0.448
% persons living in social housing	0.1704	0.106	1.614	0.107
% mortgage stress	-0.1935	0.105	-1.834	0.067
% rental stress	0.1362	0.078	1.752	0.08
% Low income households	-0.3386	0.069	-4.906	0
% unemployment benefit	-0.3388	0.117	-2.895	0.004
% disability support pensioners	1.2226	0.226	5.403	0
% unemployed	0.0361	0.06	0.598	0.55
Omnibus:	117.255	Durbin-Watson:	1.676	
Prob(Omnibus):	0	Jarque-Bera (JB):	343.395	
Skew:	1.361	Prob(JB):	2.71E-75	
Kurtosis:	6.625	Cond. No.	3220	

Notes:
[1] Standard Errors are heteroscedasticity robust (HC0)
[2] The condition number is large, 3.22e+03. This might indicate that there are strong multicollinearity or other numerical problems.

For health-relevant features, we obtain a R^2 of 0.67 which means the model can explain 67% of variation in %mental. By examining p-values we noticed %poor SA_health, %alcohol, %alcohol_squared, %asthma, %COPD are all significant at 5% level.

For non-health related variables. The R^2 is 0.19, indicating that this model is not very good as we expected. However, we can observe that some variables are significant at 5% level, such as %disability_support_pensioners, %low_income_households, %permanent migrants and %unemployment benefits.

Predictive model building

Because we observe a good fit for health-related model, we decided to also build a predictive regression model using health-related variables. This predictive regression model is built using the 80/20 training and testing split and fitting a linear model.

The R^2 for the test data set is 0.58 which means the model can explain 58% of variation for difference in %mental (**Fig 4**). Also, we noticed the model suffers from heteroskedasticity as errors seem to be non-constant which confirms the choice of using heteroscedasticity-consistent standard errors in the previous analysis.

Fig 4. Model fit and Error plot

Machine-Learning-based study

A. Mutual Information (MI)-based dependence study between variables

Based on the previous results of regression-based study, we found the linear correlation between some features (several health-relevant features and most non-health relevant features) and response is not significant. Hence, we applied MI to capture the potential non-linear dependence between features and response.

Pre-processing

Split data into training and test datasets with test size = 0.3.

MI score between each feature and response on training set

Since all the selected features are numeric, we used mutual_info_regression() function in sklearn.feature_selection to calculate the MI score between each feature and response, separately for health and non-health relevant features and their corresponding response.

Results for health-relevant features are shown below as in **Table 3-1** and **Fig 5-1**. We can see features self-harm rate, %asthma, %COPD, %poor SA-health have more MI / higher dependence with response variable %mental (MI greater than 0.24).

Results for non-health-relevant features are shown below as in **Table 3-2** and **Fig 5-2**. We can see features % households_in_dwellings_receiving_rent_assistance has more MI / higher dependence with response variable (MI greater than 0.319).

Table 3-1. MI score of each health-relevant feature

self-harm rate/100,000	0.331431
%asthma	0.294440
%COPD	0.282454
%poor SA-heath	0.246361
%smoker	0.180931
%alcohol	0.147433
%HSV	0.138125
%psycho	0.117894
%high BP	0.097659
%osteoporosis	0.070467
%exercise	0.069454
%arthritis	0.066276
%overweight	0.051531
IMR/1,000	0.032661

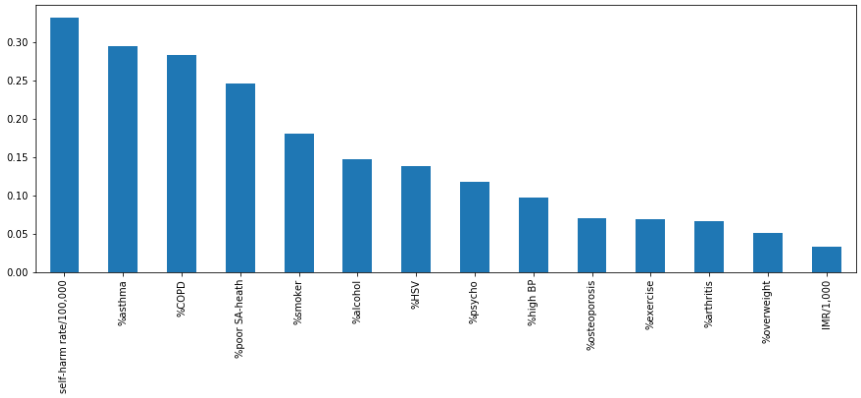


Fig 5-1. Bar plot of the MI scores of health-relevant features, calculating the MI between features and response.

Table 3-2. MI score of each non-health feature

% households in dwellings receiving rent assistance	0.319587
% mortgage stress	0.132458
% persons living in social housing	0.110370
Aboriginal population as proportion of total population (%)	0.094294
% persons living in crowded dwellings	0.072575
% Permanent migrants under the Family stream visa\n(total)	0.050758
% private dwellings where Internet accessed	0.044418
% rental stress	0.044137
% people receiving an unemployment benefit	0.038009
% Low income households	0.037435
Total fertility rate	0.017937
% unemployed	0.003410
% disability support pensioners	0.000000
% labour force participation	0.000000

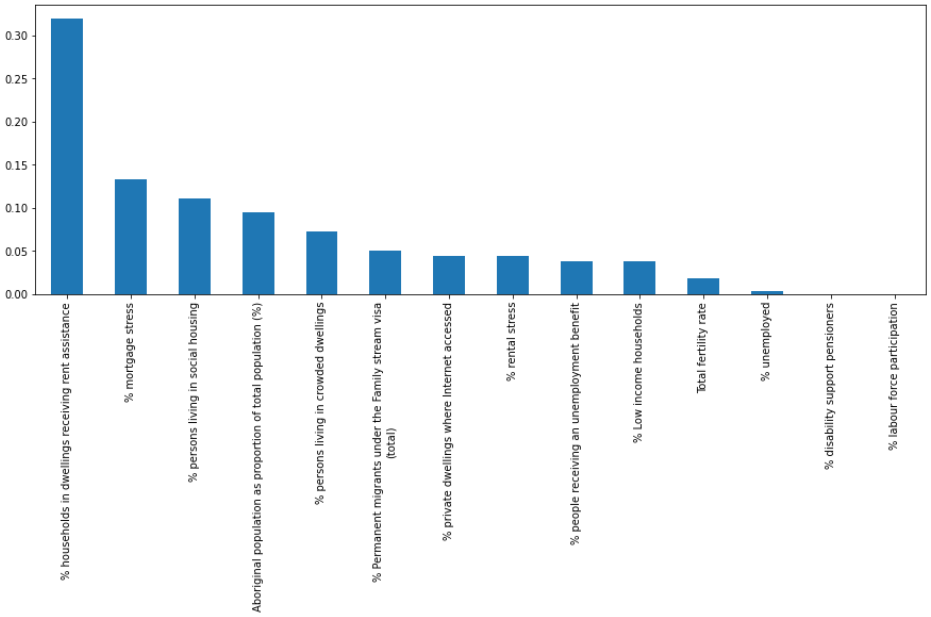


Fig 5-2. Bar plot of the MI scores of non-health-relevant features, calculating the MI between features and response.

MI score between features on training set

Similarly, we used mutual_info_regression() function to calculate the MI scores between features. This is because for the following machine learning models we do not want to use highly mutually dependent features, which can lead to the model put too much weight on some features.

The results are shown as heatmaps below (**Fig 6.a,b**) plotting from **Supplementary Table 3**. We can see within health-relevant features, %poor SA-heath has high MI with %smoker (MI = 0.697) and %COPD (MI = 0.595). Within non-healthy-relevant features, %disability_support_pensioners has high MI with %people_receiving_an_unemployment_benefit (MI = 0.7) and %Low_income_households(MI = 0.626). Here, high MI indicates high dependence.

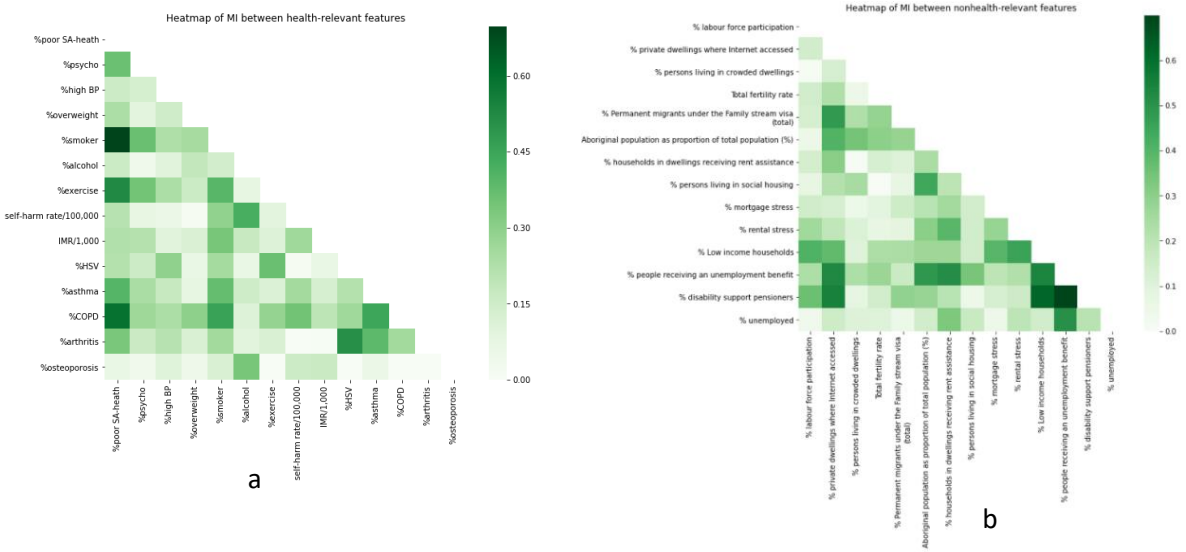


Fig 6. Heatmap for MI between features **a**. health-relevant **b**. non-health-relevant

B. Mutual Information (MI)-based decision tree model

Pre-processing

For both the training and test datasets, we binned the numeric response variables according to the distribution of the 2 responses for health and non-health features respectively (**Fig 7a,b**).

For health-relevant features, response_value <= 17.5 labelled 0; response_value > 17.5 and < 27, labelled 1; response_value > 17.5 labelled 2. For non-health-relevant features, response_value <= 10 labelled 0; response_value > 10 and < 30, labelled 1; response_value > 30 labelled 2.

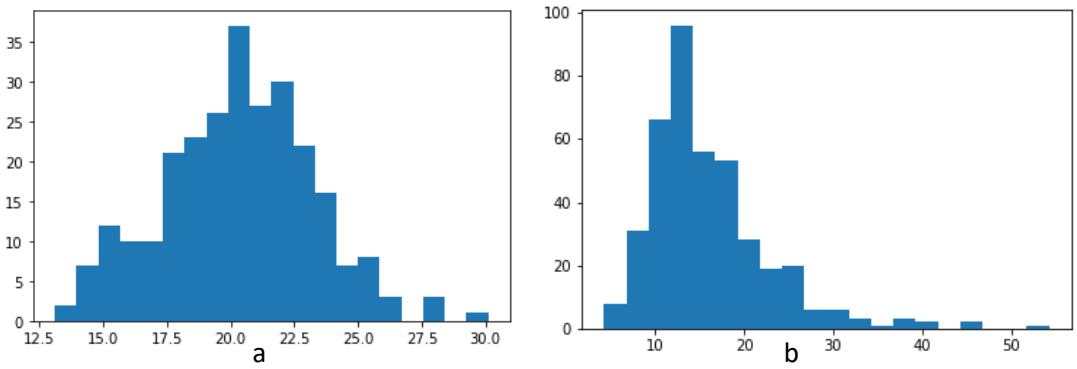


Fig 7. Histograms of two response variables **a**. Response variable %mental for health-relevant features **b**. Response variable for non-health-relevant features

Feature selection

Based on the MI scores between variables calculated in the last section, we chose 3 features respectively for health-relevant-feature model and non-health-relevant-feature model. The choosing standard is the features with top MI scores with response variable while with low MI scores with each other, since we want features to have high dependence with response while low dependence with each other.

Selected features are '%smoker', '%asthma', '%COPD' for health-relevant-feature model, and '%households_in_dwellings_receiving_rent_assistance', '%persons_living_in_social_housing', '% mortgage_stress'.

Decision tree model building, visualization and evaluation

Two decision tree models were built based on entropy using selected features as shown in **Fig 8-1** and **Fig 8-2.** The accuracy scores for the 2 models are around 0.90 for health-relevant-feature model and 0.77 for non-health-relevant-feature model. The max_depths of the two model are 3 and 2 respectively, chosen by the optimal accuracy scores via multiple trials setting value range from 1 to 8.

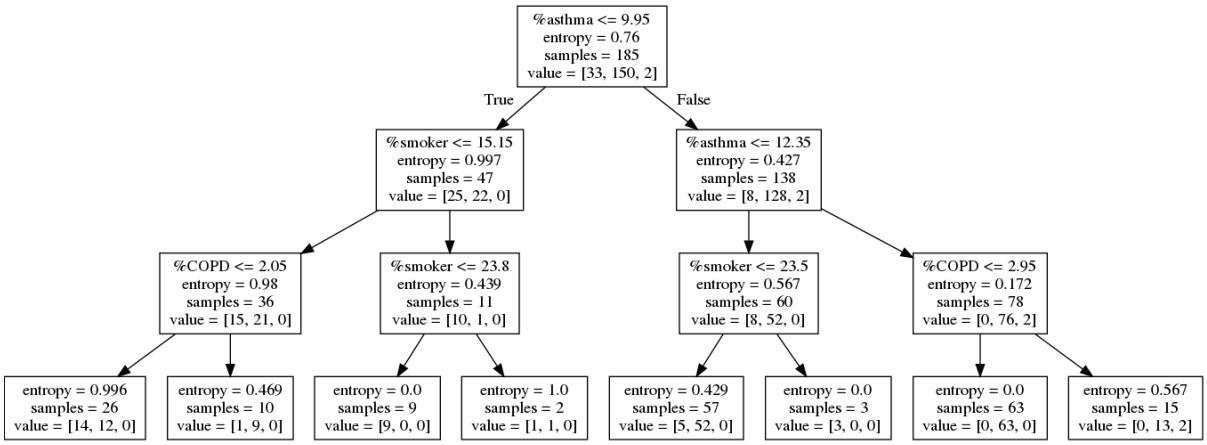


Fig 8-1. Visualization of health-relevant-feature decision tree model

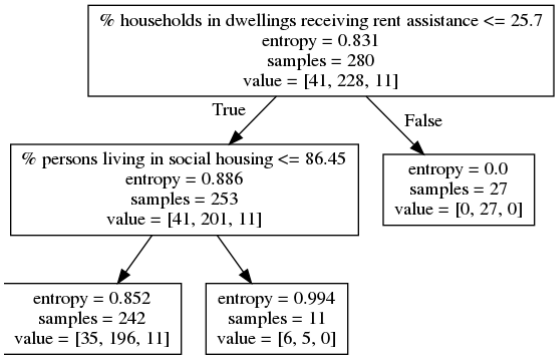


Fig 8-2. Visualization of non-health-relevant-feature decision tree model

C. PCA visualization

Pre-processing

Data normalization

Both health-relevant and non-health-relevant datasets were logarithmized for normalization. The contrast histograms of raw data and logarithmized data are shown below (**Fig 9-1** and **Fig 9-2**) to support the necessity.

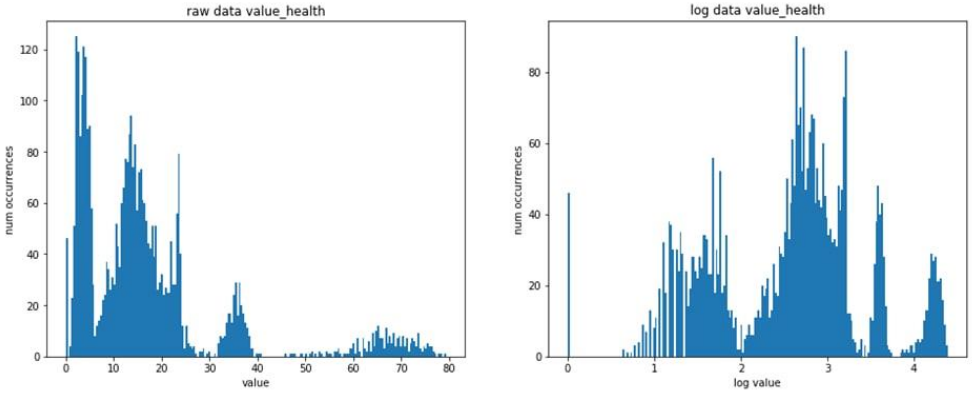


Fig 9-1. Raw vs logarithmized health-relevant feature variable data

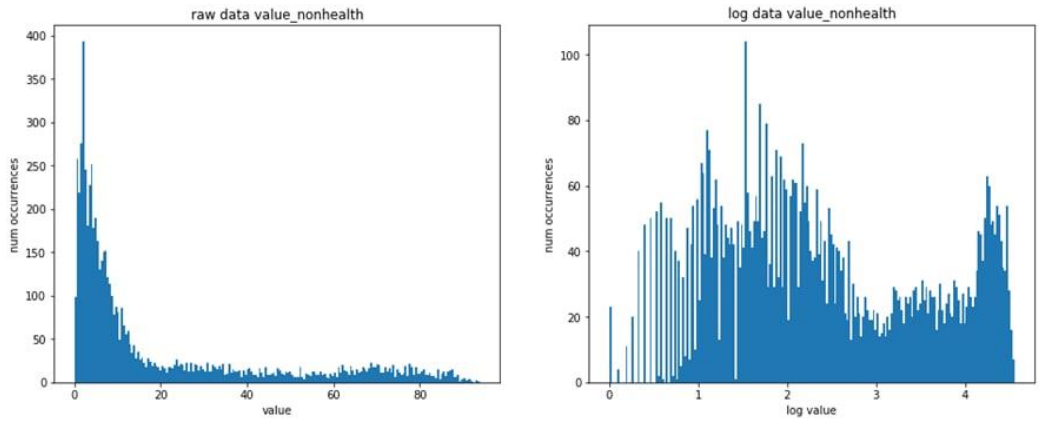


Fig 9-2. Raw data vs logarithmized non-health-relevant feature variable data

Binning

We binned the numeric response variable the same as we did in section: 'Mutual Information (MI)-based decision tree model- (1) pre-processing', where label 0 stands for low-rate, 1 for the medium rate, and 2 for the high rate of mental illness. We did the binning for better visualization (clustering of data) in PCA plots.

PCA analysis and visualization

To explain 90% of the variance, we first need to determine the number of components by the explained variance ratio. And the expected component number for health-relevant features is 4, for non-health-relevant features is 5.

Then we visualized all the components and observed the separation between each mental health illness level. And the importance of each PC is demonstrated by the explained variance.

From the overall PCA Figs of health-relevant features (**Fig 10-1**), the separation between groups is significant to some degree. And the most significant graph for clustering does not involve PC1 and PC2 (**Fig 10-2**), which reveals that main health-relevant features do not contribute to the mental-health rate increase. And for the non-health-relevant features (**Fig 10-3**), the PCA graph did not represent any significance for mental-health level separation.

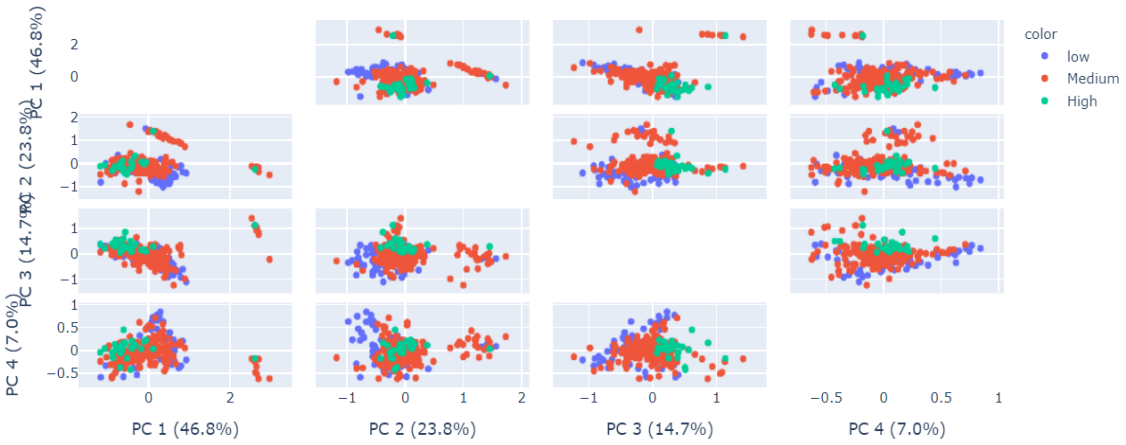


Fig 10-1. Visualization of PCA for health-relevant dataset

Fig 10-2. PC3 and PC4 of the health-relevant dataset

Fig 10-3. Visualization of PCA for non-health-relevant dataset

Significance of results

We first identified significant variables based on correlation and MI that could potentially explain variations in mental-illness rate across different areas. These variables/features are most of the 15 health-relevant features especially '%smoker', '%asthma', '%COPD', 3 of the non-health=relevant features '%households_in_dwellings_receiving_rent_assistance', '%persons_living_in_social_housing', '% mortgage_stress'. This would help the Australian government to identify potential causes of mental diseases and hopefully take steps to prevent the mental illness rate from further rising.

We built a predictive regression model that captures 58% of variation of mental illness rate by using identified health indicators which will further aid local governments to estimate their policy effects on improving mental health through improvements in other health indicators.

We built an MI-based decision tree model that can be used for prediction and classification. Specifically, if relevant features are obtained, we could predict the risk level of mental-illness rate in different areas and classify them into low-risk, medium-risk, high-risk categories. Areas labelled high-risk need more attention from the relevant organization and help, for example, increasing community psychological counselling services.

We also found, as **Fig 11** shows, it is also believed that people with mental conditions are more likely to commit self-harm behaviours such as suicide. This indicates more attention could be paid to people's mental conditions.

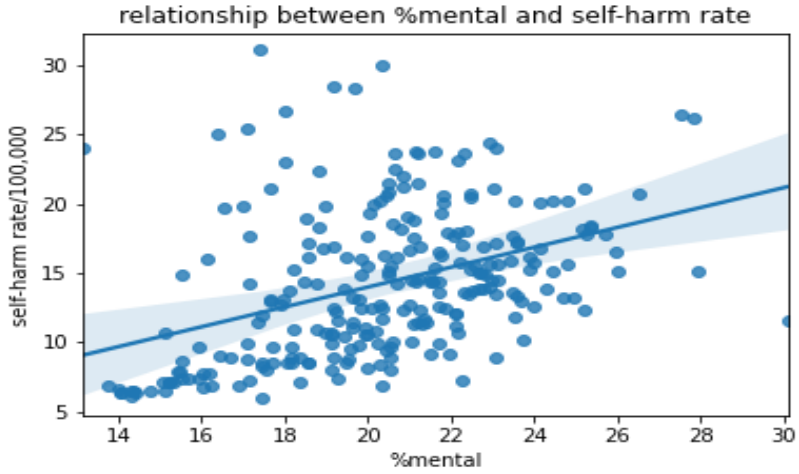


Fig 11. Relevance between regional mental health disorder rate and self-harm rate

Limitations and future improvement

One limitation of our report is that the analysis is divided into two sections due to the inconsistent period of health-relevant and non-health-relevant feature variables. As a result, the significance of health-relevant and non-health-relevant features cannot be put together for comparison or used in the same prediction/classification model.

For the regression-based study section, we observe from the heatmap and correlation matrix some fairly strong correlation between independent variables (e.g. 0.52 between %asthma and %COPD, 0.97 between %Aboriginal population and %people in the social housing) which may indicate that our model suffers from some level of multicollinearity which could bias the standard errors and significance tests of our regression model. We could improve this problem by identifying highly correlated variables in the model and do further variable transformation/modification to eliminate such multicollinearity.

Furthermore, for the non-health-relevant features, although there is no simple linear relationship, we can try to find other regression relationships or multivariate correlation analysis in the following analysis. And in the data preprocessing, the outliers deserve more attention and find out why these values are so extreme in some areas. Also, more authoritative information can be obtained through other literature or data reports to support this project.

For the Machine-Learning-based study section, due to the relatively sufficient sample size, cross-validation was not applied in the decision-tree building part, which can be improved in the future to increase the reliability of the model building and evaluation.

Appendix

Some abbreviations of variables

Estimated number of people with mental and behavioural problems ASR per 100 - **%mental**

Mental conditions admissions ASR per 100,000(1 year later for 2018-2019) - **mental admission/100,000(18-19)**

People with poor self assessed health level ASR per 100 - **%poor SA-heath**

Estimated number of people aged 18 years and over with high or very high psychological distress, based on the Kessler 10 Scale (K10) ASR per 100 - **%psycho**

Estimated number of people aged 18 years and over who had high blood pressure (modelled estimates) ASR per 100 - **%high BP**

Estimated number of people aged 18 years and over who were overweight (but not obese) (modelled estimates) ASR per 100 - **%overweight**

Estimated number of people aged 18 years and over who were current smokers (modelled estimates) ASR per 100 **- %smoker**

Estimated number of people aged 18 years and over who consumed more than two standard alcoholic drinks per day on average (modelled estimates) ASR per 100 - **%alcohol**

Estimated population, aged 18 years and over, who undertook low, very low or no exercise in the previous week (modelled estimates) ASR per 100 - **%exercise**

Deaths from suicide and self-inflicted injuries, 0 to 74 years Average annual ASR per 100,000 -

self-harm rate/100,000

Infant mortality Average annual IMR per 1,000 – **IMR/1,000**

Estimated number of people with heart, stroke and vascular disease ASR per 100 - **%HSV**

Estimated number of people with asthma ASR per 100 - **%asthma**

Estimated number of people with chronic obstructive pulmonary disease (COPD) ASR per 100 - **%COPD**

Estimated number of people with arthritis ASR per 100 - **%arthritis**

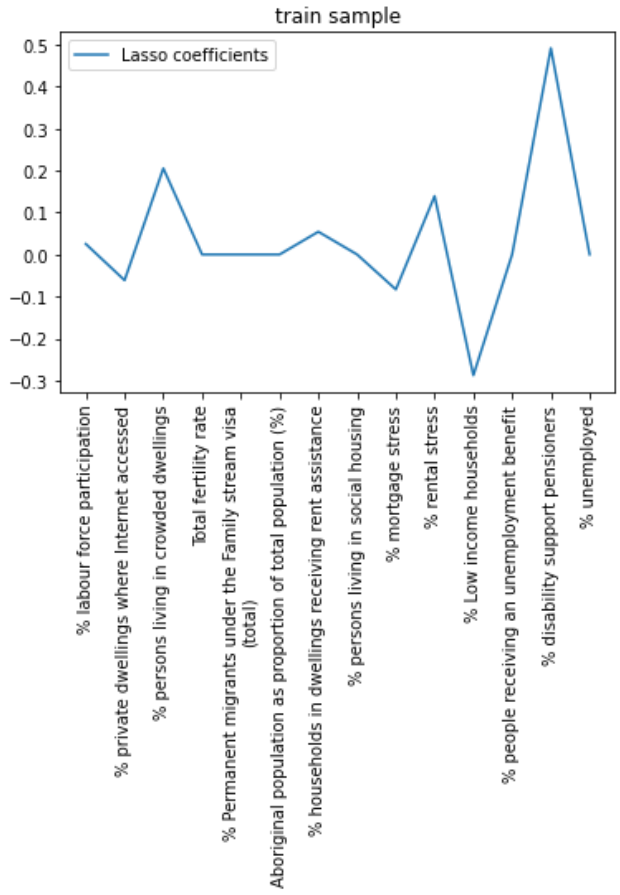
Estimated number of people with osteoporosis ASR per 100 - **%osteoporosis**

Supplementary Table.1

	%mental	mental ad	%poor SA-	%psycho	%high BP	%overwei	%smoker	%alcohol	%exercise	self-harm	IMR/1,000	%HSV	%asthma	%COPD	%arthritis	%osteopo
%mental	1.00	0.19	0.54	0.38	0.25	-0.36	0.45	0.30	0.23	0.33	0.18	0.33	0.70	0.64	0.44	-0.12
mental ad	0.19	1.00	0.43	0.30	-0.03	-0.32	0.63	0.22	0.43	0.33	0.29	0.09	0.22	0.49	0.15	-0.40
%poor SA-	0.54	0.43	1.00	0.74	0.39	-0.54	0.75	0.09	0.76	0.34	0.28	0.50	0.61	0.68	0.54	-0.13
%psycho	0.38	0.30	0.74	1.00	0.28	-0.34	0.59	-0.19	0.70	0.26	0.22	0.51	0.51	0.45	0.49	-0.07
%high BP	0.25	-0.03	0.39	0.28	1.00	-0.20	0.30	0.05	0.42	0.15	-0.01	0.66	0.42	0.37	0.53	0.21
%overwei	-0.36	-0.32	-0.54	-0.34	-0.20	1.00	-0.49	0.00	-0.47	-0.15	-0.35	-0.31	-0.25	-0.56	-0.24	-0.02
%smoker	0.45	0.63	0.75	0.59	0.30	-0.49	1.00	0.41	0.70	0.47	0.28	0.39	0.60	0.58	0.50	-0.51
%alcohol	0.30	0.22	0.09	-0.19	0.05	0.00	0.41	1.00	-0.09	0.42	-0.11	-0.13	0.37	0.21	0.11	-0.54
%exercise	0.23	0.43	0.76	0.70	0.42	-0.47	0.70	-0.09	1.00	0.25	0.19	0.64	0.42	0.54	0.54	-0.18
self-harm	0.33	0.33	0.34	0.26	0.15	-0.15	0.47	0.42	0.25	1.00	0.20	0.07	0.38	0.43	0.20	-0.23
IMR/1,000	0.18	0.29	0.28	0.22	-0.01	-0.35	0.28	-0.11	0.19	0.20	1.00	0.02	0.06	0.39	0.02	0.04
%HSV	0.33	0.09	0.50	0.51	0.66	-0.31	0.39	-0.13	0.64	0.07	0.02	1.00	0.52	0.36	0.77	0.08
%asthma	0.70	0.22	0.61	0.51	0.42	-0.25	0.60	0.37	0.42	0.38	0.06	0.52	1.00	0.52	0.74	-0.19
%COPD	0.64	0.49	0.68	0.45	0.37	-0.56	0.58	0.21	0.54	0.43	0.39	0.36	0.52	1.00	0.29	-0.12
%arthritis	0.44	0.15	0.54	0.49	0.53	-0.24	0.50	0.11	0.54	0.20	0.02	0.77	0.74	0.29	1.00	0.01
%osteopo	-0.12	-0.40	-0.13	-0.07	0.21	-0.02	-0.51	-0.54	-0.18	-0.23	0.04	0.08	-0.19	-0.12	0.01	1.00

Supplementary Table.2

Gamma distribution			
Dep. Variable:	%mental	No. Observations:	212
Model:	GLM	Df Residuals:	203
Model Family:	Gamma	Df Model:	8
Link Function:	inverse_power	Scale:	0.010663
Method:	IRLS	Log-Likelihood:	-454.66
Date:	Tue, 12 Oct 2021	Deviance:	2.2368
Time:	0.530162037	Pearson chi2:	2.16
No. Iterations:	6	Pseudo R-squ. (CS):	0.6722
Covariance Type:	nonrobust		
Poisson distribution			
Dep. Variable:	%mental	No. Observations:	212
Model:	GLM	Df Residuals:	203
Model Family:	Poisson	Df Model:	8
Link Function:	log	Scale:	1
Method:	IRLS	Log-Likelihood:	-545.21
Date:	Tue, 12 Oct 2021	Deviance:	62.904
Time:	0.531041667	Pearson chi2:	64.3
No. Iterations:	4	Pseudo R-squ. (CS):	0.1385
Covariance Type:	nonrobust		
Gaussian distribution			
Dep. Variable:	%mental	No. Observations:	212
Model:	GLM	Df Residuals:	203
Model Family:	Gaussian	Df Model:	8
Link Function:	identity	Scale:	3.5601
Method:	IRLS	Log-Likelihood:	-430.81
Date:	Tue, 12 Oct 2021	Deviance:	722.7
Time:	0.532025463	Pearson chi2:	723
No. Iterations:	3	Pseudo R-squ. (CS):	0.7896
Covariance Type:	nonrobust		



Supplementary Fig 1. Lasso coefficient plot

Supplementary Table 3: MI between health-relevant and non-health-relevant features

	%poor SA-heath	%psycho	%high BP	%overweight
%poor SA-heath	0	0.36448	0.160507	0.235526
%psycho	0	0.00000	0.133870	0.093479
%high BP	0	0.00000	0.000000	0.152210
%overweight	0	0.00000	0.000000	0.000000
%smoker	0	0.00000	0.000000	0.000000
%alcohol	0	0.00000	0.000000	0.000000
%exercise	0	0.00000	0.000000	0.000000
self-harm rate/100,000	0	0.00000	0.000000	0.000000
IMR/1,000	0	0.00000	0.000000	0.000000
%HSV	0	0.00000	0.000000	0.000000
%asthma	0	0.00000	0.000000	0.000000
%COPD	0	0.00000	0.000000	0.000000
%arthritis	0	0.00000	0.000000	0.000000
%osteoporosis	0	0.00000	0.000000	0.000000

	%smoker	%alcohol	%exercise	self-harm rate/100,000
%poor SA-heath	0.697248	0.161021	0.525720	0.210073
%psycho	0.365354	0.042698	0.349066	0.075430
%high BP	0.229547	0.099130	0.236054	0.062433
%overweight	0.248500	0.184513	0.162993	0.012394
%smoker	0.000000	0.142095	0.396567	0.292655
%alcohol	0.000000	0.000000	0.079221	0.422696
%exercise	0.000000	0.000000	0.000000	0.093543
self-harm rate/100,000	0.000000	0.000000	0.000000	0.000000
IMR/1,000	0.000000	0.000000	0.000000	0.000000
%HSV	0.000000	0.000000	0.000000	0.000000
%asthma	0.000000	0.000000	0.000000	0.000000
%COPD	0.000000	0.000000	0.000000	0.000000
%arthritis	0.000000	0.000000	0.000000	0.000000
%osteoporosis	0.000000	0.000000	0.000000	0.000000

	IMR/1,000	%HSV	%asthma	%COPD	%arthritis
%poor SA-heath	0.225556	0.219107	0.400874	0.595485	0.336721
%psycho	0.216176	0.160191	0.240023	0.264913	0.164931
%high BP	0.099473	0.295355	0.176887	0.236337	0.211573
%overweight	0.122232	0.070100	0.079705	0.299420	0.124872
%smoker	0.341051	0.251384	0.375174	0.458790	0.258286
%alcohol	0.173696	0.069016	0.152840	0.109090	0.095522
%exercise	0.109385	0.366191	0.114729	0.285694	0.124257
self-harm rate/100,000	0.262678	0.010945	0.253363	0.352963	0.000000
IMR/1,000	0.000000	0.069254	0.131601	0.198581	0.005317
%HSV	0.000000	0.000000	0.218492	0.262379	0.511801
%asthma	0.000000	0.000000	0.000000	0.448561	0.389420
%COPD	0.000000	0.000000	0.000000	0.000000	0.257025
%arthritis	0.000000	0.000000	0.000000	0.000000	0.000000
%osteoporosis	0.000000	0.000000	0.000000	0.000000	0.000000

	%osteoporosis
%poor SA-heath	0.068167
%psycho	0.036780
%high BP	0.104848
%overweight	0.044129
%smoker	0.124363
%alcohol	0.340748
%exercise	0.000000
self-harm rate/100,000	0.159111
IMR/1,000	0.168574
%HSV	0.000000
%asthma	0.058030
%COPD	0.000000
%arthritis	0.000000
%osteoporosis	0.000000

Data1. MI metrics between health-relevant features

	% labour force participation
% labour force participation	0.000000
% private dwellings where Internet accessed	0.146140
% persons living in crowded dwellings	0.004439
Total fertility rate	0.144601
% Permanent migrants under the Family stream vi...	0.132083
Aboriginal population as proportion of total po...	0.049776
% households in dwellings receiving rent assist...	0.138313
% persons living in social housing	0.074944
% mortgage stress	0.149017
% rental stress	0.261447
% Low income households	0.411899
% people receiving an unemployment benefit	0.229925
% disability support pensioners	0.365037
% unemployed	0.035755

	% private dwellings where Internet accessed
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.133624
Total fertility rate	0.229152
% Permanent migrants under the Family stream vi...	0.482000
Aboriginal population as proportion of total po...	0.407417
% households in dwellings receiving rent assist...	0.307482
% persons living in social housing	0.216919
% mortgage stress	0.133916
% rental stress	0.190096
% Low income households	0.382247
% people receiving an unemployment benefit	0.532511
% disability support pensioners	0.551193
% unemployed	0.152365

	% persons living in crowded dwellings
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.000000
Total fertility rate	0.050326
% Permanent migrants under the Family stream vi...	0.226888
Aboriginal population as proportion of total po...	0.347437
% households in dwellings receiving rent assist...	0.006361
% persons living in social housing	0.243584
% mortgage stress	0.054760
% rental stress	0.114145
% Low income households	0.105705
% people receiving an unemployment benefit	0.230459
% disability support pensioners	0.076691
% unemployed	0.111068

	Total fertility rate
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.000000
Total fertility rate	0.000000
% Permanent migrants under the Family stream vi...	0.289350
Aboriginal population as proportion of total po...	0.302552
% households in dwellings receiving rent assist...	0.129129
% persons living in social housing	0.000000
% mortgage stress	0.090972
% rental stress	0.074279
% Low income households	0.235366
% people receiving an unemployment benefit	0.274905
% disability support pensioners	0.145748
% unemployed	0.106507

	% Permanent migrants under the Family stream visa\n(total)
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.000000
Total fertility rate	0.000000
% Permanent migrants under the Family stream vi...	0.000000
Aboriginal population as proportion of total po...	0.282544
% households in dwellings receiving rent assist...	0.111150
% persons living in social housing	0.074290
% mortgage stress	0.156267
% rental stress	0.088923
% Low income households	0.238181
% people receiving an unemployment benefit	0.161738
% disability support pensioners	0.290125
% unemployed	0.054795

	Aboriginal population as proportion of total population (%)
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.000000
Total fertility rate	0.000000
% Permanent migrants under the Family stream vi...	0.000000
Aboriginal population as proportion of total po...	0.000000
% households in dwellings receiving rent assist...	0.240535
% persons living in social housing	0.443155
% mortgage stress	0.208442
% rental stress	0.306088
% Low income households	0.270563
% people receiving an unemployment benefit	0.490128
% disability support pensioners	0.278872
% unemployed	0.144715

	% households in dwellings receiving rent assistance
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.000000
Total fertility rate	0.000000
% Permanent migrants under the Family stream vi...	0.000000
Aboriginal population as proportion of total po...	0.000000
% households in dwellings receiving rent assist...	0.000000
% persons living in social housing	0.199943
% mortgage stress	0.255999
% rental stress	0.394186
% Low income households	0.270264
% people receiving an unemployment benefit	0.516963
% disability support pensioners	0.210184
% unemployed	0.330268

	% persons living in social housing
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.000000
Total fertility rate	0.000000
% Permanent migrants under the Family stream vi...	0.000000
Aboriginal population as proportion of total po...	0.000000
% households in dwellings receiving rent assist...	0.000000
% persons living in social housing	0.000000
% mortgage stress	0.142425
% rental stress	0.145604
% Low income households	0.144491
% people receiving an unemployment benefit	0.338979
% disability support pensioners	0.046699
% unemployed	0.167325

	% mortgage stress
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.000000
Total fertility rate	0.000000
% Permanent migrants under the Family stream vi...	0.000000
Aboriginal population as proportion of total po...	0.000000
% households in dwellings receiving rent assist...	0.000000
% persons living in social housing	0.000000
% mortgage stress	0.000000
% rental stress	0.280483
% Low income households	0.398444
% people receiving an unemployment benefit	0.199104
% disability support pensioners	0.129213
% unemployed	0.048220

	% rental stress
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.000000
Total fertility rate	0.000000
% Permanent migrants under the Family stream vi...	0.000000
Aboriginal population as proportion of total po...	0.000000
% households in dwellings receiving rent assist...	0.000000
% persons living in social housing	0.000000
% mortgage stress	0.000000
% rental stress	0.000000
% Low income households	0.458814
% people receiving an unemployment benefit	0.224353
% disability support pensioners	0.161052
% unemployed	0.194378

	% Low income households
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.000000
Total fertility rate	0.000000
% Permanent migrants under the Family stream vi...	0.000000
Aboriginal population as proportion of total po...	0.000000
% households in dwellings receiving rent assist...	0.000000
% persons living in social housing	0.000000
% mortgage stress	0.000000
% rental stress	0.000000
% Low income households	0.000000
% people receiving an unemployment benefit	0.543188
% disability support pensioners	0.625684
% unemployed	0.142659

	% people receiving an unemployment benefit
% labour force participation	0.000000
% private dwellings where Internet accessed	0.000000
% persons living in crowded dwellings	0.000000
Total fertility rate	0.000000
% Permanent migrants under the Family stream vi...	0.000000
Aboriginal population as proportion of total po...	0.000000
% households in dwellings receiving rent assist...	0.000000
% persons living in social housing	0.000000
% mortgage stress	0.000000
% rental stress	0.000000
% Low income households	0.000000
% people receiving an unemployment benefit	0.000000
% disability support pensioners	0.699773
% unemployed	0.509046

	% disability support pensioners	\
% labour force participation	0.00000	
% private dwellings where Internet accessed	0.00000	
% persons living in crowded dwellings	0.00000	
Total fertility rate	0.00000	
% Permanent migrants under the Family stream vi...	0.00000	
Aboriginal population as proportion of total po...	0.00000	
% households in dwellings receiving rent assist...	0.00000	
% persons living in social housing	0.00000	
% mortgage stress	0.00000	
% rental stress	0.00000	
% Low income households	0.00000	
% people receiving an unemployment benefit	0.00000	
% disability support pensioners	0.00000	
% unemployed	0.20656	

	% unemployed
% labour force participation	0.0
% private dwellings where Internet accessed	0.0
% persons living in crowded dwellings	0.0
Total fertility rate	0.0
% Permanent migrants under the Family stream vi...	0.0
Aboriginal population as proportion of total po...	0.0
% households in dwellings receiving rent assist...	0.0
% persons living in social housing	0.0
% mortgage stress	0.0
% rental stress	0.0
% Low income households	0.0
% people receiving an unemployment benefit	0.0
% disability support pensioners	0.0
% unemployed	0.0

Data2. MI metrics between non-health-relevant features