Contents

Aim

The aim of this project is to investigate whether there is a relationship between health-related factors and the crime rate in Victoria, as well as to determine which health factors have the greatest impact on crime. The results may be beneficial to policymakers attempting to reduce crime rates. It will also be useful in informing health officials as to the greater social impact of their decisions. The project hence helps to expand the understanding of health and crime in Victoria as well as improving liveability in the state.

Datasets

Two datasets were utilised in this investigation: the Victorian Population Health Survey 2017, and the Crime Statistics Agency Data Tables - Criminal Incidents (2021).

Both datasets were initially in excel file format. The former dataset contained health-related data, organised by local government area (LGA) and gender for the year 2017. The other dataset contained data on the crime rate per 100,000 people, for the years 2011-2020 within each LGA. The crime rate was the response variable for our investigation.

The health factors used as explanatory variables in the investigation were as followed:

- Anxiety/Depression: the percentage of people diagnosed with anxiety and depression
- Alcohol: the percentage of people who do not abstain from alcohol
- Fruit Consumption: the percentage of people who complied with daily fruit guidelines
- Overweight: the percentage of people who were considered overweight
- Physical Activity: the percentage of people who complied with physical activity guidelines
- Smoking: the percentage of smokers

LGA was the common attribute between the datasets, hence was used to link the health and crime data. Furthermore, the only crime data from 2017 was used to complement the year of record for the health data.

Pre-processing and Wrangling

Before linking the datasets, they first needed to be cleaned. Firstly, the initial format of the datasets was not machine readable and so had to be manually reformatted. Next, as each health factor was categorised by gender, we combined the gendered data by taking the mean to get a single gender-neutral value. This change is appropriate because the corresponding gendered variables exhibited high degrees of correlation (see Figure 1) and hence this transformation reduces the collinearity and dimensionality of our data. Merging on a column only works if the key values are exact matches so regular expression was performed on the LGA column in the crime data to remove additional tags and spaces that were not present for the LGA values in the other dataset. The datasets were then merged.

Finally, the numbers in crime rate data were in string format so regex was again performed to remove commas before converting the data to a numeric type.

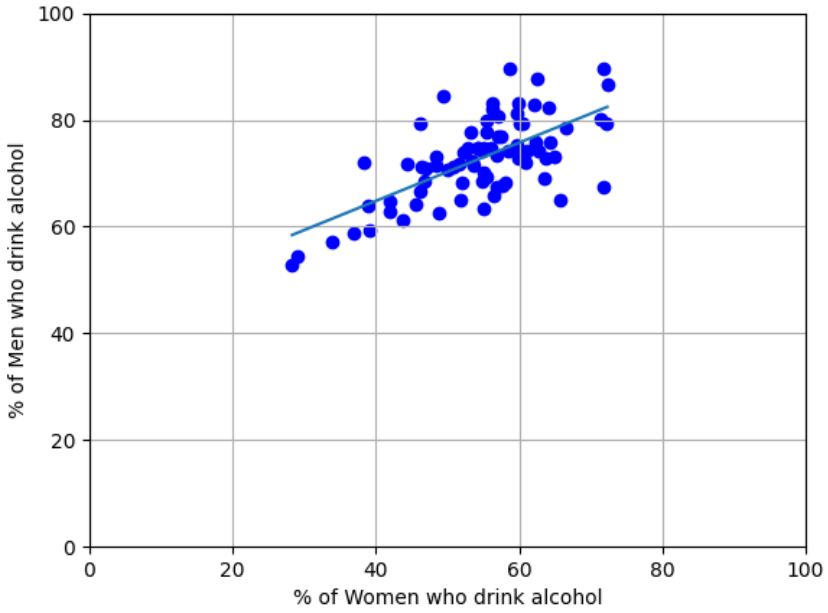


Figure 1: Scatter plot of 'Alcohol Women %' vs 'Alcohol Men %'

Analysis Methods

Methods

Considering the project aim and datasets chosen, some analytical techniques were more relevant than others. As this investigation deals with continuous data and inferring the relationship between variables, we chose to use techniques such as Pearson correlation, linear regression, and regression trees. The coefficients and goodness of fit of the regression models will help indicate which health factors are related to crime. We deliberately chose not to use clustering and classification techniques because they assign records into discrete classes/clusters which does not help us answer our question about determining which health factors are important determinants of crime (a continuous variable).

Training-test Split

Before further analysis, the data was split into training and test sets (80/20 split). The training set was used for preliminary analysis and model fitting, whereas the test set was used to assess the fit and generalisability of the models. Furthermore, k-fold cross-validation was also used to assess the fit, to reduce the impact of an 'unlucky' split.

Preliminary Analysis

To get a preliminary understanding of how each health factor is related to the crime rate, we created scatter plots between each predictor variable and the response (in the training set), and calculated the respective Pearson correlation.

On initial inspection of the plots (Figures 2-7), we see that the relationships appear to be somewhat linear indicating that linear regression may be an appropriate model to fit and that transformations of the variables is likely not necessary. Furthermore, this linear relationship suggests that Pearson correlation is a suitable metric for getting an initial understanding of how variables are related. However, we noticed that one point had a significantly higher crime rate than other points and appears to be highly influential on the line of best fit. This point corresponds to 'Melbourne' which, as the CBD, we expect to receive a greater traffic of people and hence a likely higher crime rate. We exclude this outlier point and redo the plots (Figures 2-7). The correlation coefficients changed drastically after removing this point (consider 'overweight' in Table 1 e.g.), indicating that removal was appropriate.

Looking at the correlation coefficients in Table 1, we observe that 'overweight' exhibits a trivial level of correlation with crime rate, while other variables show small levels of correlation. This gives us a cursory indication that some health factors may be related to crime rate but that this relationship is not strong. The variables 'alcohol', 'smoking' and 'physical activity' have the highest correlation coefficient indicating they may be more influential on crime rate. No causal relationship can be determined from these coefficients. The correlation also allows us to do some initial feature filtering: we can ignore 'overweight' in our models. The correlation between explanatory variables was also calculated (see Appendix) and it was found that some variables exhibited high degrees of collinearity ('Physical activity' and 'alcohol' e.g.). This will limit the interpretability of the multiple regression model so was kept in mind.

Table 1: Pearson correlation coefficients of health predictors with crime rate

Variable	Alcohol	Anxiety/ depression	Fruit consumption	Overweight	Physical Activity	Smokers
Correlation including outlier	-0.302	0.038	-0.012	-0.229	-0.179	0.041
Correlation excluding outlier	-0.283	0.177	-0.167	0.001	-0.215	0.220

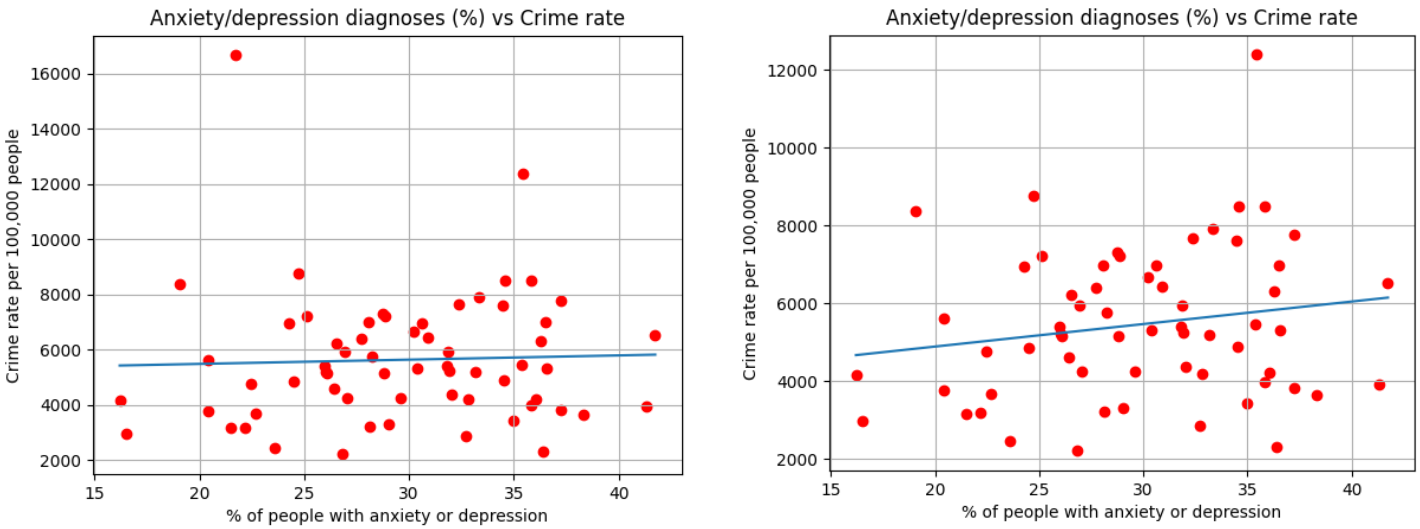


Figure 2: Anxiety/depression feature plotted against crime rate.

All LGA (left). Outliers removed (right).

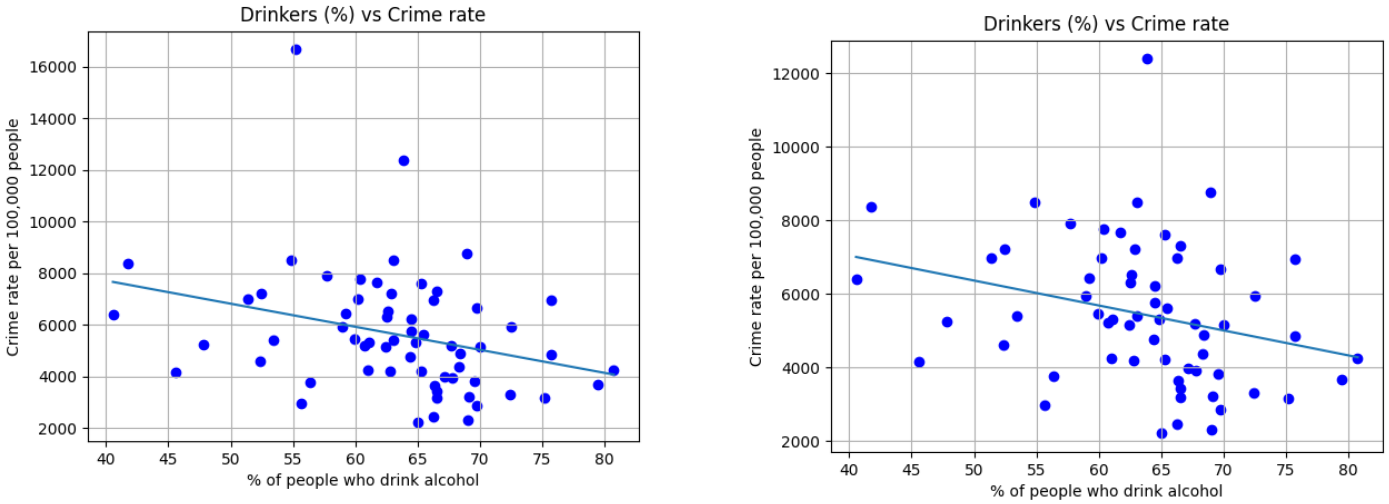


Figure 3: Alcohol consumption feature plotted against crime rate.

All LGA (left). Outliers removed (right).

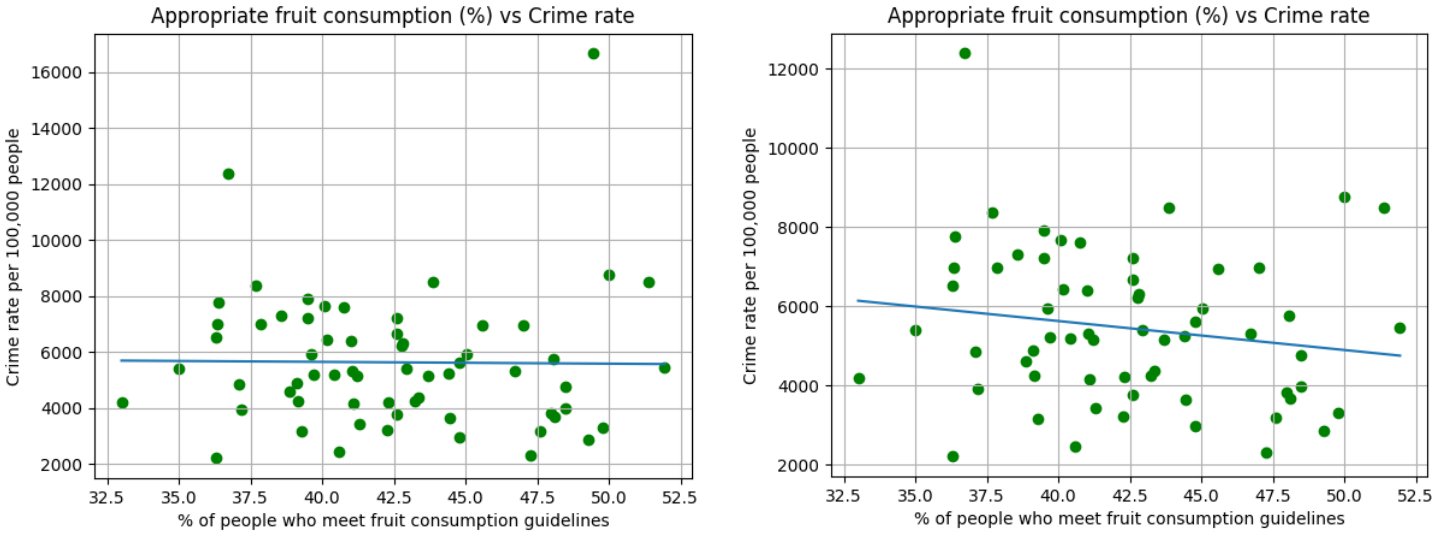


Figure 4: Fruit consumption feature plotted against crime rate.

All LGA (left). Outliers removed (right).

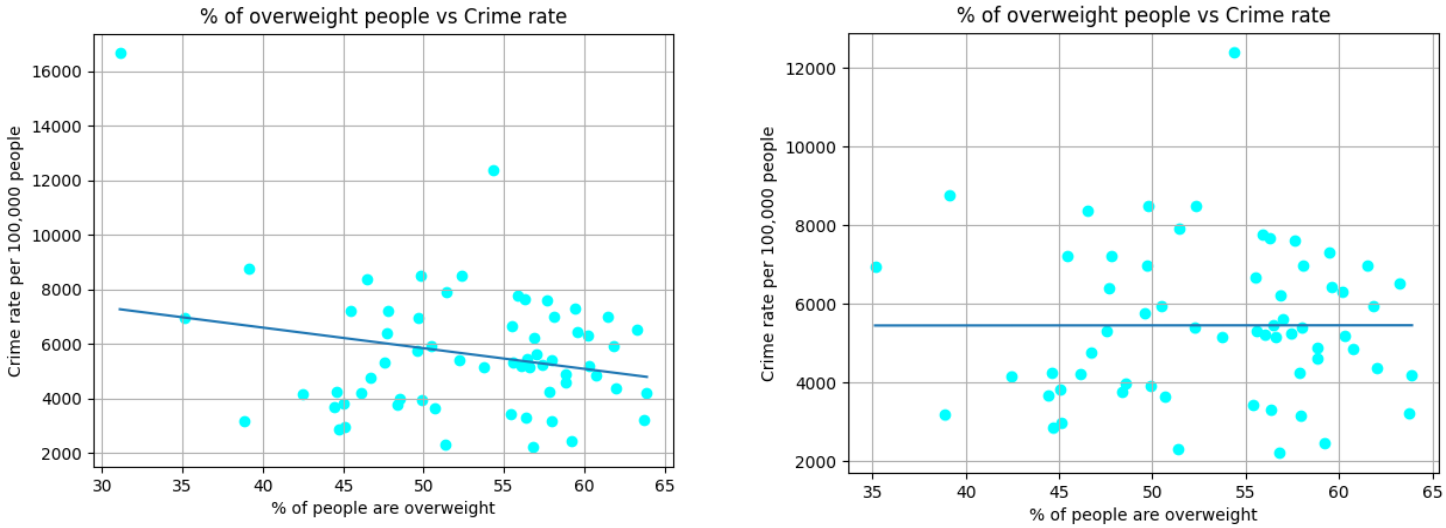


Figure 5: Overweight feature plotted against crime rate.

All LGA (left). Outliers removed (right).

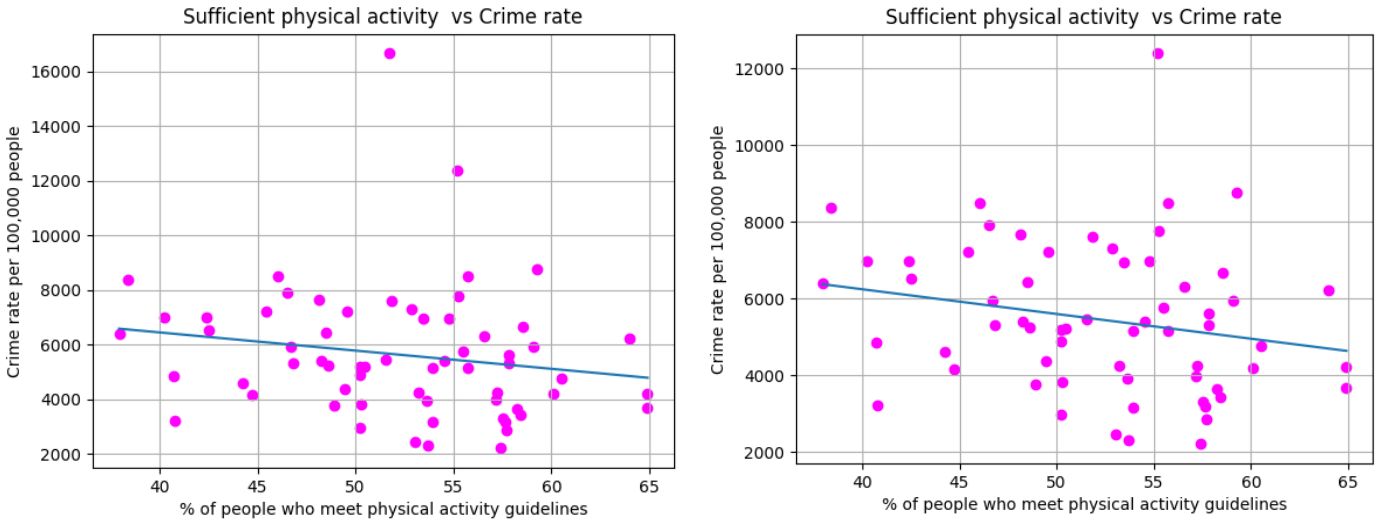


Figure 6: Physical activity feature plotted against crime rate.

All LGA (left). Outliers removed (right).

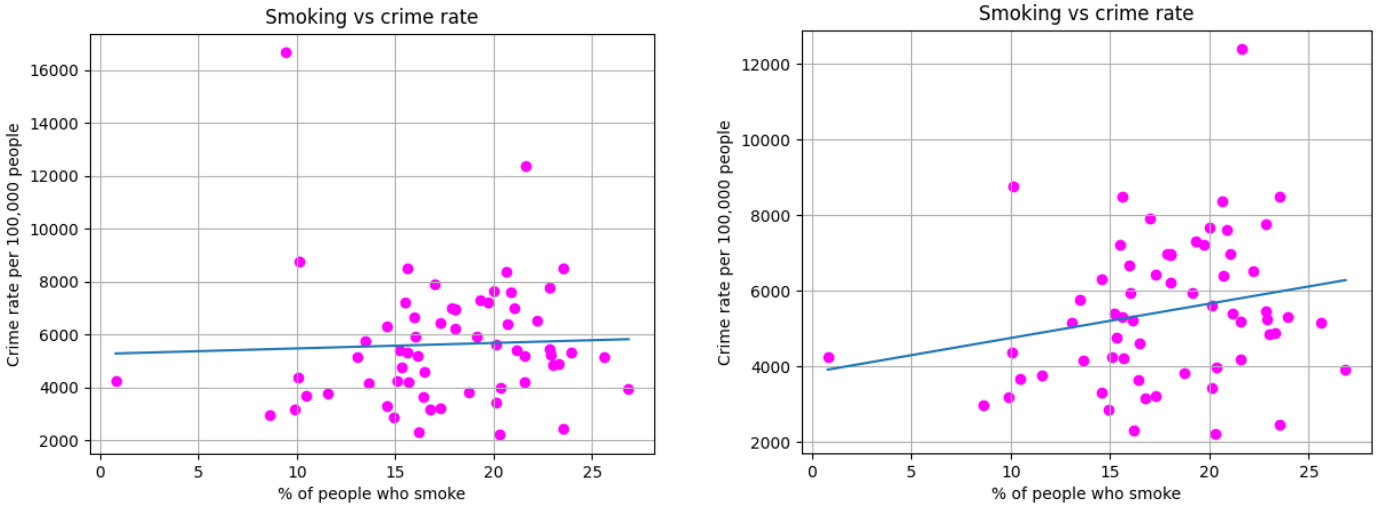


Figure 7: Smoking feature plotted against crime rate.

All LGA (left). Outliers removed (right).

Modelling

To get a better understanding of whether the health factors have a meaningful relationship with the crime rate we fit linear regression models. Interpretation of all models is left to the next section. We initially fit a regression model including all predictors except 'overweight' to our training set. We calculated its test and training R^2 as well as its test MSE (Table 2). Furthermore, we performed 10-fold cross validation and calculated the same metrics, averaged across the folds, to give more reliable results. To account for the collinearity present between the explanatory variables, we also fit a simple regression with 'alcohol' as the only predictor (variable with highest correlation coefficient). The residuals of the regression models appeared to be distributed randomly, indicating that the assumptions for linear regression were satisfied (see appendix). Other regression models were fit but for brevity are not discussed here.

Linear regression may be too simple and is adversely affected by collinearity, so we also fit a regression tree to possibly resolve these issues. Regression trees work similarly to classification trees, but leaf nodes instead give the *mean* value of objects in the node as a prediction. We use 10-fold cross-validation to determine the tree depth that minimises the MSE. The optimal depth was found to be two and the tree can be visualised in the appendix.

Discussion

In this section we interpret the models and discuss the significance of the results. The low correlation coefficients of each health factor suggests that some of them may have a weak relation with crime rate. The health factors 'alcohol' and 'smoking' had the greatest absolute correlation with crime (-0.283, and 0.220, respectively), indicating that an increase in drinkers or a decrease in smokers coincides with a lower crime rate. However, the fitting of our models revealed that the correlations we observed were likely spurious or insignificant, and not reflective of a meaningful relationship between health factors and crime rate. For both regression models, we find that the mean R^2 after doing 10-fold CV is negative which occurs when the model is a worse predictor of crime rate than simply using the average (see table 2). Looking at the regression tree (see appendix), we see again that 'alcohol' and 'smoking' are key variables for splitting and may be important but, based on MSE, the regression tree performs even worse than the linear regression models (see table 2). This reinforces the idea that the relationships observed between health factors and the crime rate were spurious and not statistically significant. This conclusion is valuable as it communicates to policyholders wishing to reduce crime that health factors are not a useful means of doing so. Other factors should instead be pursued. It also indicates that health-related polices are unlikely to have severe unintended consequences with regards to the crime rate.

Table 2: Goodness of fit of models

	Training R^2	Test R^2	10-fold CV R^2	10-fold CV MSE (scaled)
Model 1	0.14	0.04	-0.40	0.038
Model 2	0.08	0.00	-0.24	0.038
Regression tree	n/a	n/a	n/a	0.043

Evaluation

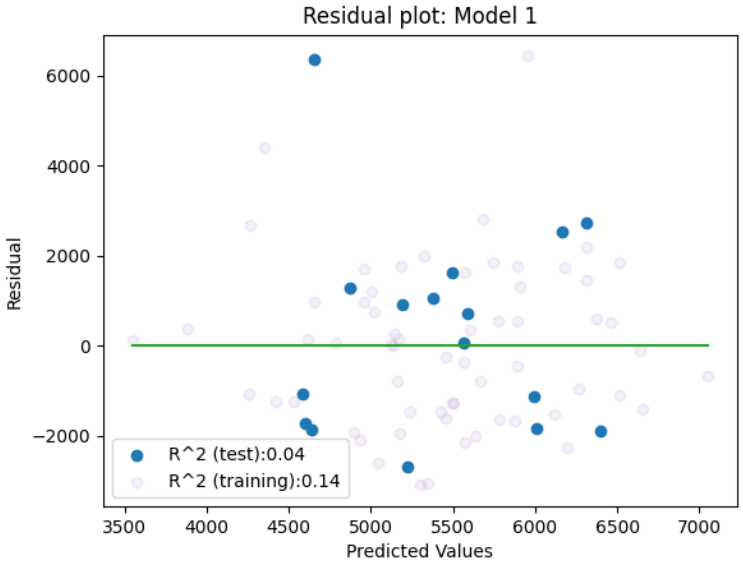
In this section we examine the limitations of our results and propose possible improvements for future research. Firstly, the results are limited as only a subset of possible health factors was considered; there may be important health factors that were not included in the analysis which future studies could incorporate. Secondly, there was significant correlation between some of the explanatory variables. This issue was inevitable given we looked only at health factors and limits the interpretability of regression models (not too important given our models fit poorly), however, future studies may wish to address this by creating a single health index variable. Finally, linear regression assumes that the effect of each variable is additive which may be inappropriate. Other models such as GLMs may be more suitable.

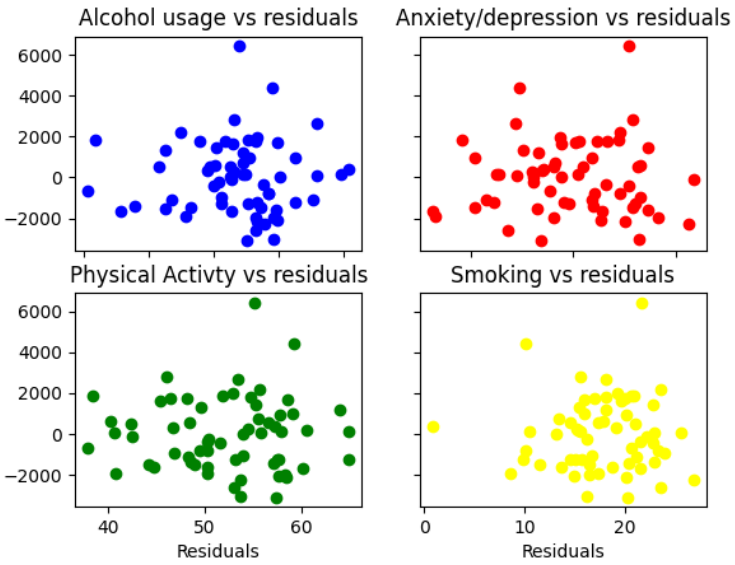
Appendix

Correlation matrix between explanatory variables

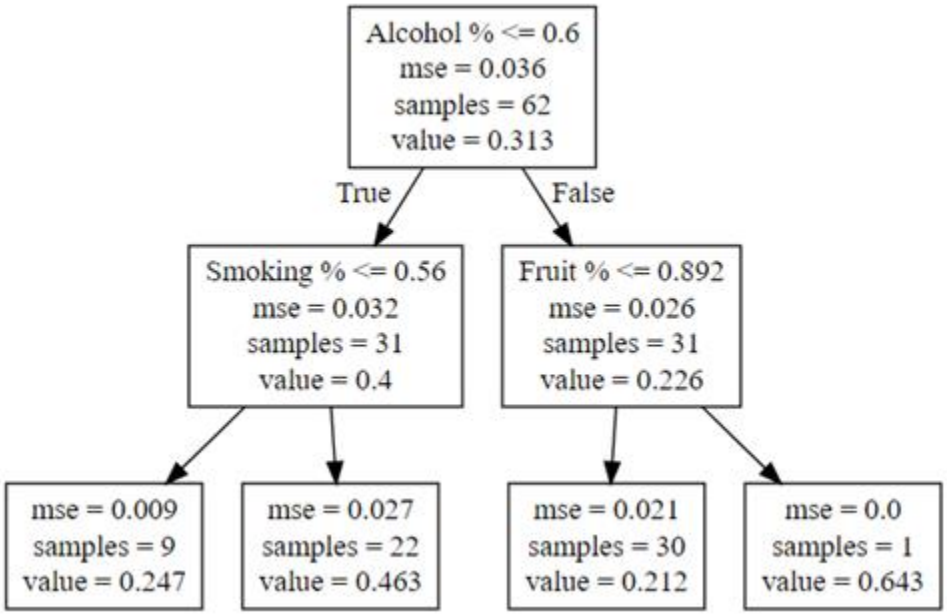
		Alcohol %	Anxiety/depression %	Fruit %	Overweight %	Physical Activity %	Smoking %
1	Alcohol %	1.0	0.13703132812170252	0.23219759340182913	0.010624744288490525	0.5472610323327002	-0.22377190140715927
2	Anxiety/depression %	0.13703132812170252	1.0	-0.08131740966826993	0.3396039928458715	0.04418197221055248	0.42032225735438533
3	Fruit %	0.23219759340182913	-0.08131740966826993	1.0	-0.4947607915392833	0.32774304716346586	-0.3676304268385496
4	Overweight %	0.010624744288490525	0.3396039928458715	-0.4947607915392833	1.0	-0.21395619177812933	0.4254979024770028
5	Physical Activity %	0.5472610323327002	0.04418197221055248	0.32774304716346586	-0.21395619177812933	1.0	-0.25724233687003184
6	Smoking %	-0.22377190140715927	0.42032225735438533	-0.3676304268385496	0.4254979024770028	-0.25724233687003184	1.0

Residual analysis plots





Regression tree



References

Crime Statistics Agency Data Tables - Criminal Incidents. (2021, March 19). Retrieved from DATA VIC: https://discover.data.vic.gov.au/dataset/crime-by-location-data-table

Victorian Population Health Survey 2017 - VHISS. (2020, May 11). Retrieved from DATA VIC: https://discover.data.vic.gov.au/dataset/victorian-population-health-survey-2017-vhiss