Investigating the Impact of Low-Calorie Diets on other Nutrients Intake

Group number: 47

Thuy Linh (Kylie) Le 1373049
Tung Lam Le 1374374
Ngoc Minh Vu 1375708
Ngoc Minh Pham 1312628

COMP20008 - Group Assignment 2

May 19, 2022

Table of Contents

1/ Aim

Pepsi's diet cola, Nestle's fat-free cocoa, or Kellogg's whole-grain breakfast cereal are only a few examples of companies introducing reduced-calorie products in response to consumers' demand. Reducing calories, a measure of energy in food and beverages, has gained considerable attention and can be a step towards a healthier lifestyle. However, low-calorie diets may not address overall nutritional needs and should be exercised with advice from professionals.

This project aims to investigate potential relationships between the amount of calories and the amount of other components in food. It also aims to identify nutrient components that are most affected by the reduction of calorie intake. This may give insights into the appropriateness of low-calorie diets for people with certain health conditions.

This project targets people who are looking to lose weight or maintain a healthy weight, where reducing calorie intake is often a key strategy. Target audiences also include community health workers who wish to promote balanced diets and food manufacturers who wish to tailor their low-calorie products to better address other nutritional needs of customers.

2/ Background

2.1/ Dataset

The dataset utilized in this project is the nutrient file from Australian Food Composition Database - Release 2 provided by Food Standards Australia New Zealand (Foodstandards.gov.au, 2016). It includes information of 1616 foods and beverages per 100 grams portion. Each food has a unique public food key as a string of letters and numbers, followed by a 5-digit numeric series representing the classification. The first 2 numbers of the 'Classification' form the Food group ID, which will be further discussed in part 6.3. This is followed by a string of Food Name, and, lastly, numeric values of up to 290 nutrients and food components contained. This dataset is initially formatted as an Excel spreadsheet with some columns containing large proportions of missing data.

2.2/ Target variable and Explanatory variables

'Energy with dietary fibre, equated (kJ)' will be referred to as the target variable throughout this report. It is calculated as follows:

Total energy = 17(Total protein) + 37(Total fat) + 17(Carbohydrate) + 37(Total dietary fibre)

(where total energy is measured in kilojoules per serve and other variables are measured in grams per serve) (www.mydailyintake.net, n.d.)

From this definition, the relationships between energy and protein, fat, carbohydrate, and fibre are already clear. Hence, variables relating to these nutrients will not be examined in this project. 283 other nutrients and food components are used as potential explanatory variables.

3/ Data Pre-processing

The dataset has some drawbacks: (1) a large number of missing values, and (2) different features' measurements and ranges. Before addressing these problems, regular expression is used to clean newline characters contained in features' names.

3.1/ Feature removal

As discussed in part 2.2, these features are removed: 'Energy, without dietary fibre, equated (kJ)', 'Protein (g)', 'Fat, total (g)', 'Total dietary fibre (g)', 'Available carbohydrate, without sugar alcohols (g)', 'Available carbohydrate, with sugar alcohols (g)'. 'Public Food Key', 'Food Name', and 'Classification' are also removed as this project's focus is on the relations between nutrient components.

All features having more than two-thirds of entries missing are removed as imputation techniques require a certain amount of data presented in order to work reasonably well. It is notable that lists of features that break the above threshold are constructed separately for train and test data. The lists are then merged, and features in the merged list are removed for both sets. This avoids the effect of "unlucky" splits where one set contains the majority of missing data, causing inaccurate separate imputation.

3.2/ Imputation and Scaling

Foods with similar amounts of other nutrients are likely to contain similar amounts of the nutrient with missing values. Hence, the chosen technique is K-nearest neighbors (KNN) imputation. Since KNN is distance-based, different scales for different features may lead to bias, so normalization is important.

MinMaxScaler is used since (1) the data has a significant amount of extreme values and large ranges, and (2) the upper and lower boundaries can be known from domain knowledge.

4/ Feature Selection

The pre-processed dataset still contains more than 70 features. Redundancy induces extra computational effort or overfitting, so reducing the dimensionality is needed.

4.1/ Zero-variance removal and Power transform

Variables with zero variance are removed since (1) no relationship between these variables and the target can be deduced, and (2) binning techniques only work well on non-zero variance variables.

Pearson correlation requires data to be normally distributed. An alternative method is Normalized Mutual Information (NMI). Continuous data needs to be discretized into bins using equal-width or equal-frequency binning. For heavily skewed data, it is impossible for each bin to have roughly the same number of objects. So for all methods mentioned, Power Transformer needs to be applied feature-wise to make data more Gaussian-like.

4.2/ Correlation

NMI with equal-width binning, NMI with equal-frequency binning, and Pearson correlation will be applied independently to find the most suitable method.

Before and after each method, correlation matrices are generated to visualize the effect of collinearity reduction (figure 1, 2) (For brevity, only figures for NMI with equal-frequency binning are given here. For other 2 methods, see appendix 1-4).

Predictors highly correlated with other predictors or slightly correlated with the target are deemed insignificant for the models and removed. Correlation greater than 0.5 or smaller than 0.1 is considered highly or slightly correlated, respectively.

Scatter graph matrices for 3 sets of remaining predictors are compared (figure 3 & appendix 5, 6). As least collinearity is shown in figure 3, NMI with equal-frequency binning may be the most suitable method.

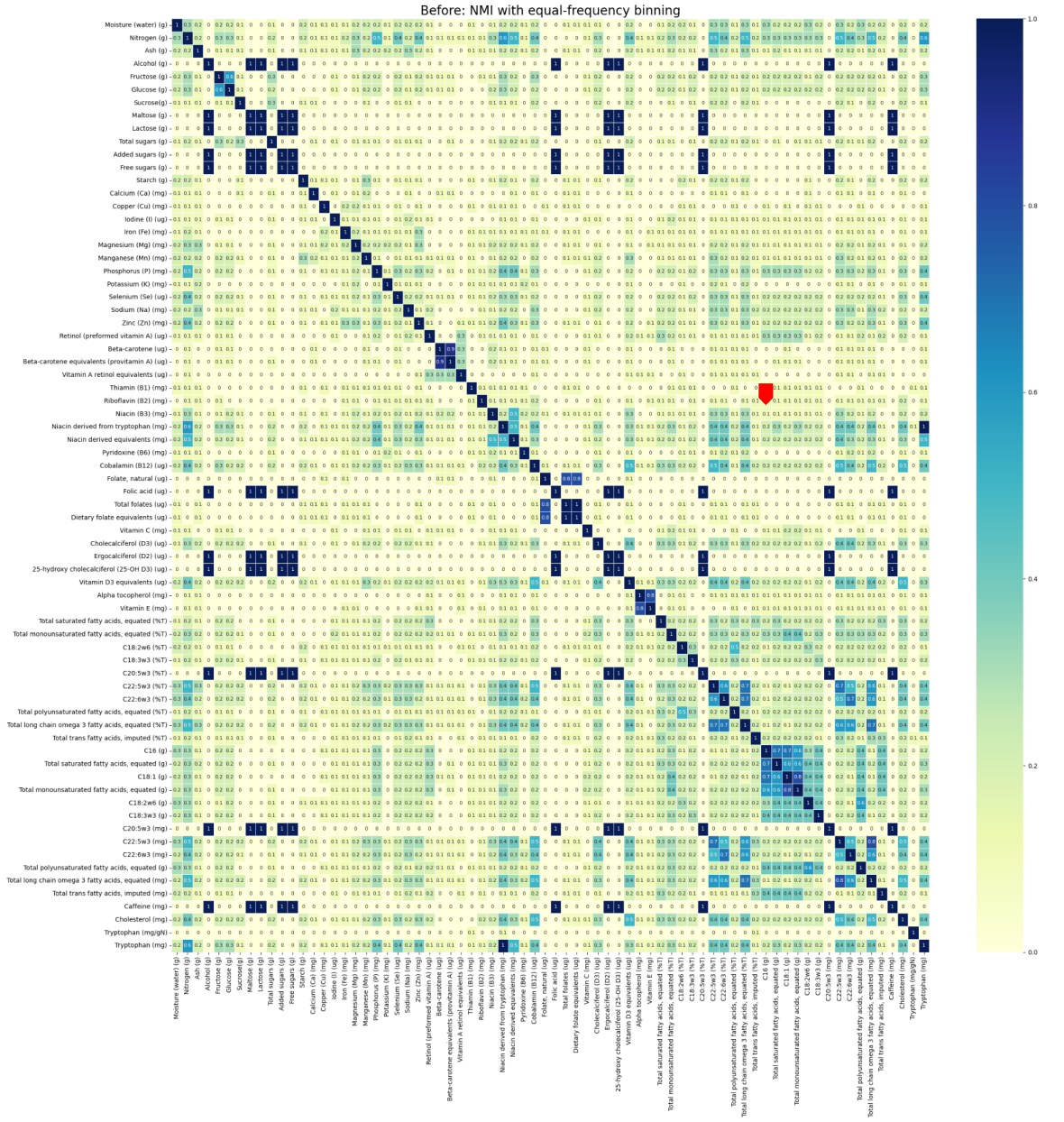


Figure 1: Before NMI with equal-frequency binning

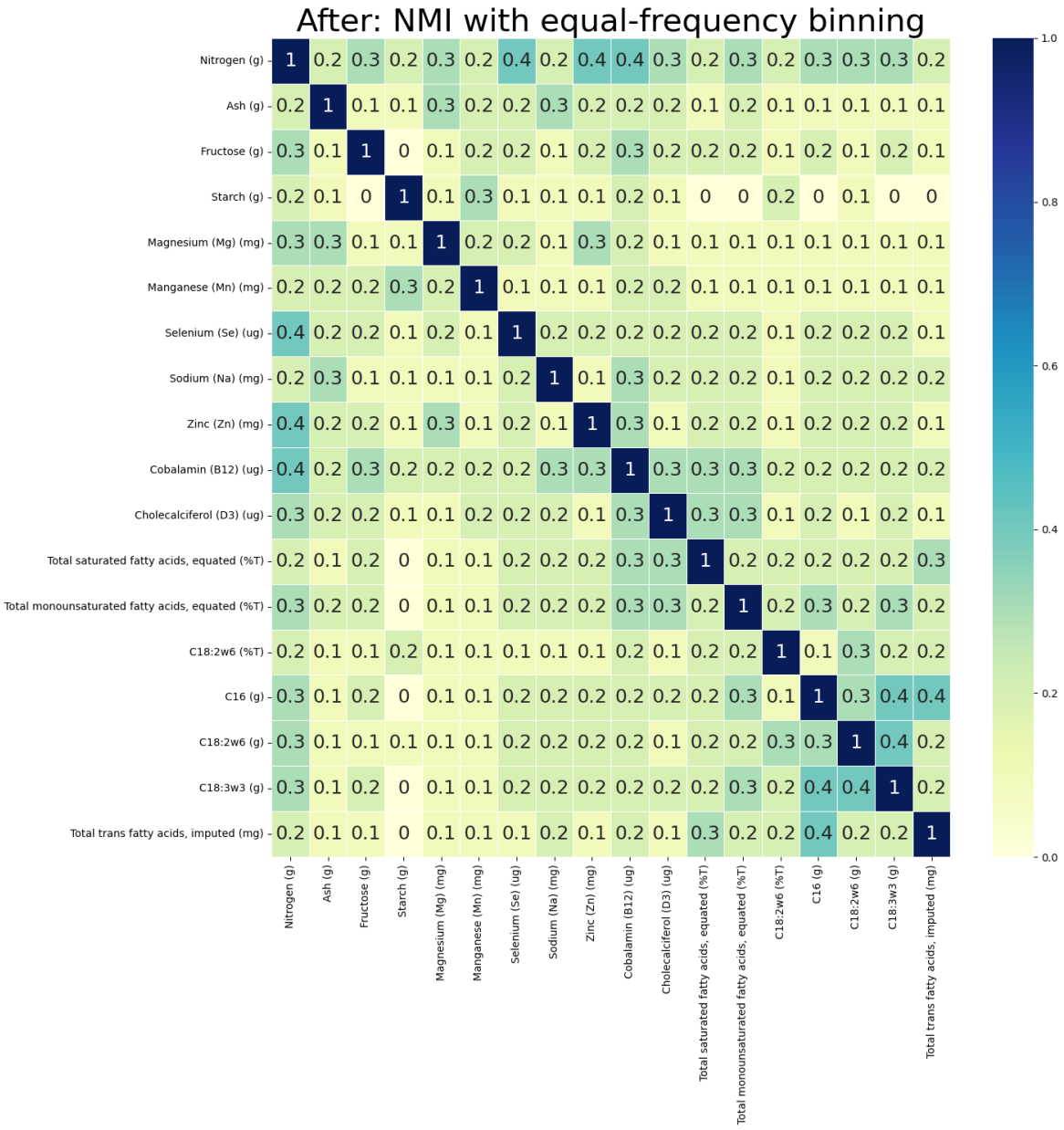


Figure 2: After NMI with equal-frequency binning

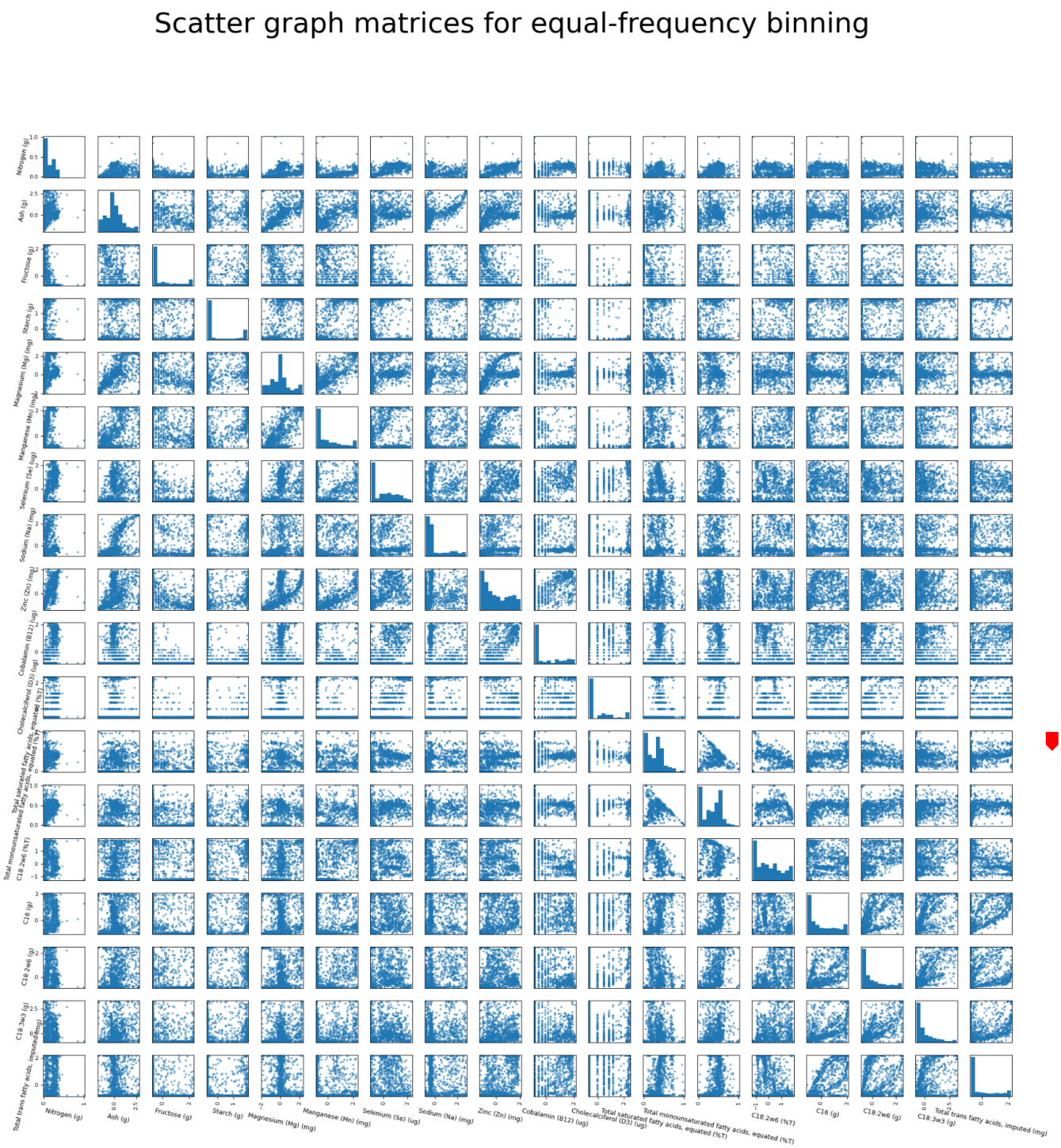


Figure 3: Scatter graph matrices for equal-frequency binning

5/ Regression models

5.1/ Preliminary analysis

Scatter plots and residual plots between each selected feature and the target feature are generated to determine whether linear relationships exist (figure 4, 5):

(1) Residual plots for 'Total saturated fatty acids, equated (%T)', 'Total monounsaturated fatty acids, equated (%T)', 'C18:2w6 (%T)' are somewhat symmetrically distributed and do not follow a clear pattern. 'Total saturated fatty acids, equated (%T)' and 'Total monounsaturated fatty acids, equated (%T)' plots show a tendency to cluster towards the middle of the plot. They satisfy the linear relationship with the target.

(2) Residuals for 'Nitrogen (g)', 'C16 (g)' tend to converge from left to right, which does not satisfy the condition for “perfect” linear regression fit. There may be a weak linear relationship.

(3) Residual plots for other predictors are either not evenly distributed vertically, have an outlier, or have a clear shape. They will not be considered for linear models.

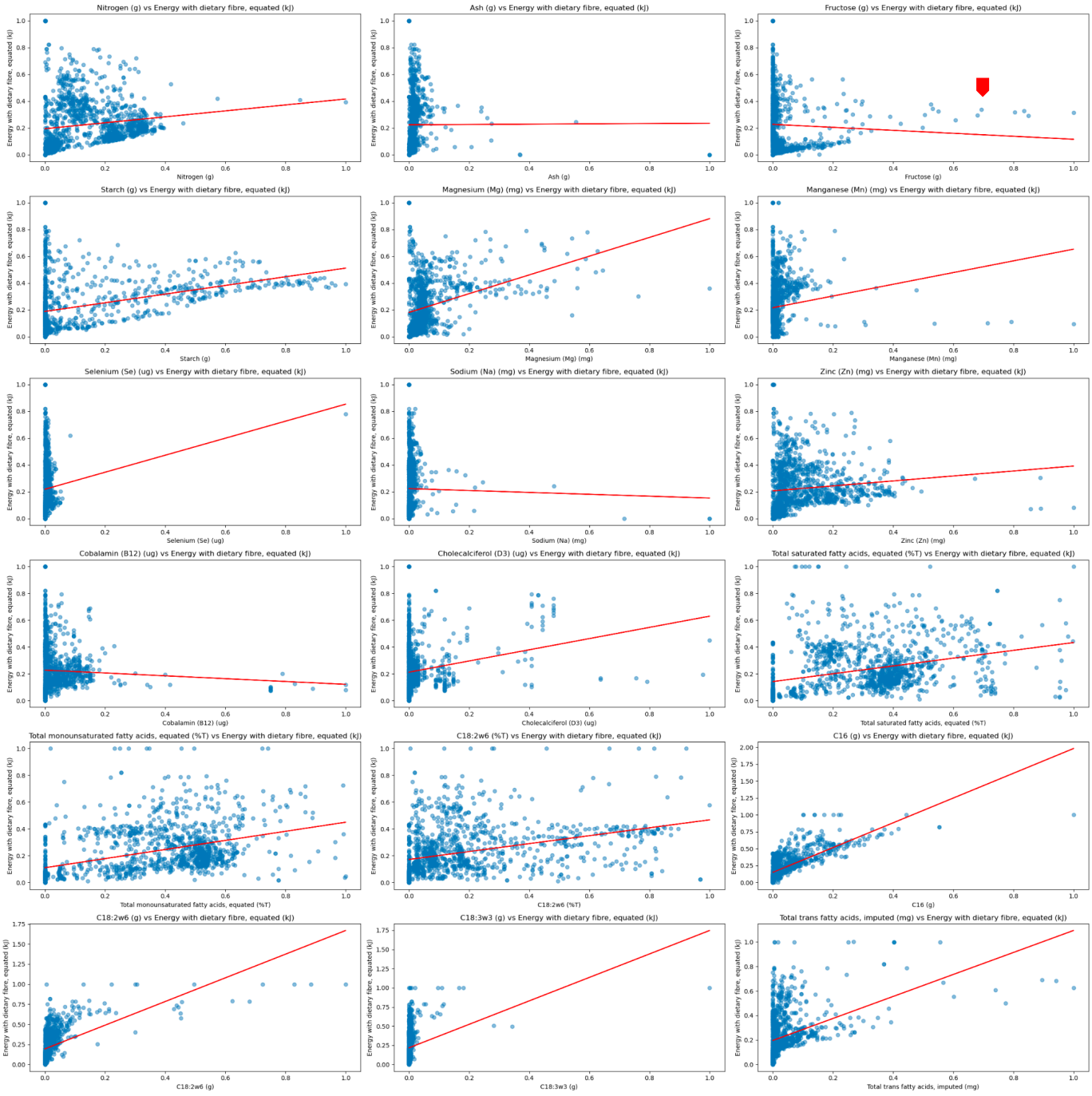


Figure 4: Scatter plots between selected features and the target feature

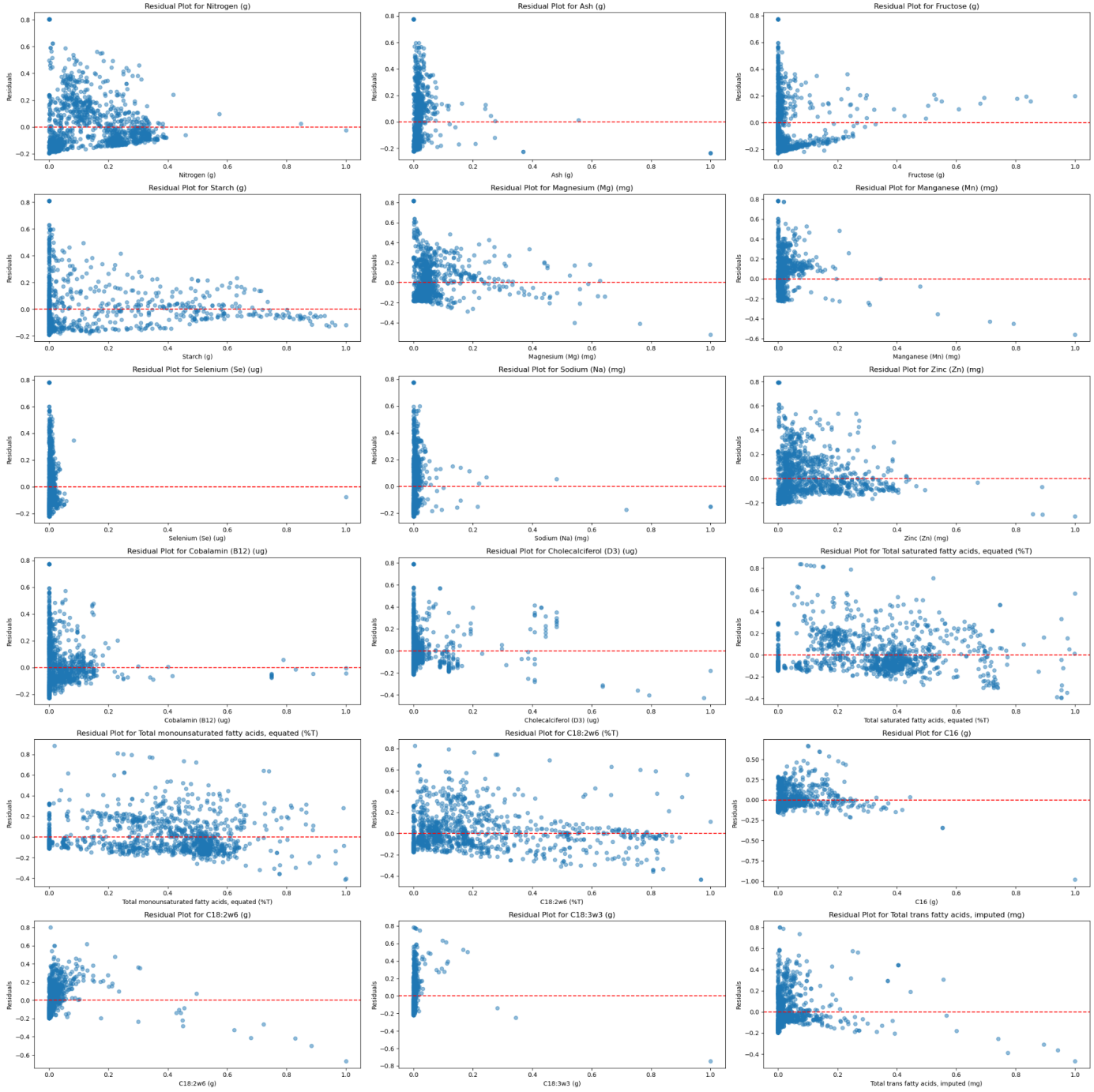


Figure 5: Residual plots for selected features

5.2/ Linear model 1

This model examine the relationship between 5 predictors ('Total saturated fatty acids, equated (%T)', 'Total monounsaturated fatty acids, equated (%T)', 'C18:2w6 (%T)', 'Nitrogen (g)', 'C16 (g)') and the response variable.

10-fold Cross Validation (CV) is used to:

(1) give the average interception and coefficients of the linear regression model, allowing for more reliable interpretation of the relation between predictors and the target variable

(2) evaluate the choice of above predictors using 3 metrics: R-squared on train set, R-squared on test set, and Mean Squared Error (MSE) on test set, noting that MSE is calculated on scaled data (table 1).

Obtained linear regression expression is as follows:

Energy = 0.0615 - 0.0166(Nitrogen) - 0.0067(Saturated fatty acids) + 0.1379 (Monounsaturated fatty acids) + 0.2945(Linoleic acid (C18:2w6)) + 1.7044(Palmitic acid (C16))

5.3/ Linear model 2

Since model 1 produces negative average R-squared on test set (table 1) a second model should be built on only predictors that show clearer linear relationship with target. The predictors chosen are: 'Total saturated fatty acids, equated (%T)', 'Total monounsaturated fatty acids, equated (%T)', and 'C18:2w6 (%T)'.

10-fold CV is performed as described in part 5.2, and evaluation scores are in table 1.

Obtained linear regression expression is as follows:

Energy = 0.05016 + 0.1613(Saturated fatty acids) + 0.2657(Monounsaturated fatty acids) + 0.2531 (Linoleic acid (C18:2w6))

5.4/ Regression tree

The chosen features contain outliers (figure 4), to which linear regression is sensitive. Regression tree should be utilized as it is not largely influenced by outliers. 10-fold CV on a train set is used to determine the tree depth minimizing MSE. Figure 6 implies that 5 may be an optimal depth where MSE is small and complexity level is acceptable. A regression tree is then visualized (figure 7). 10-fold CV is performed again to evaluate the performance of the tree regression model with max depth 5 using MSE metric (table 1).

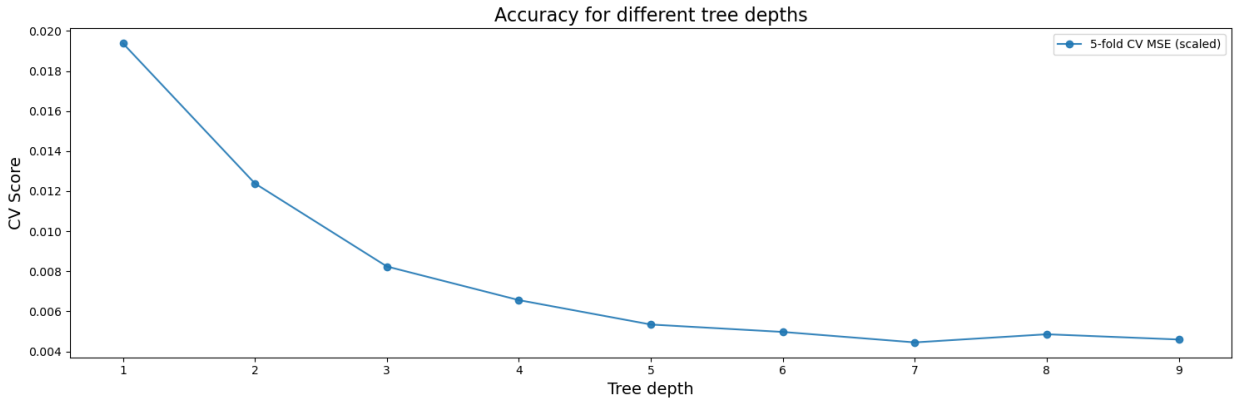


Figure 6: Average MSE values for different tree depths

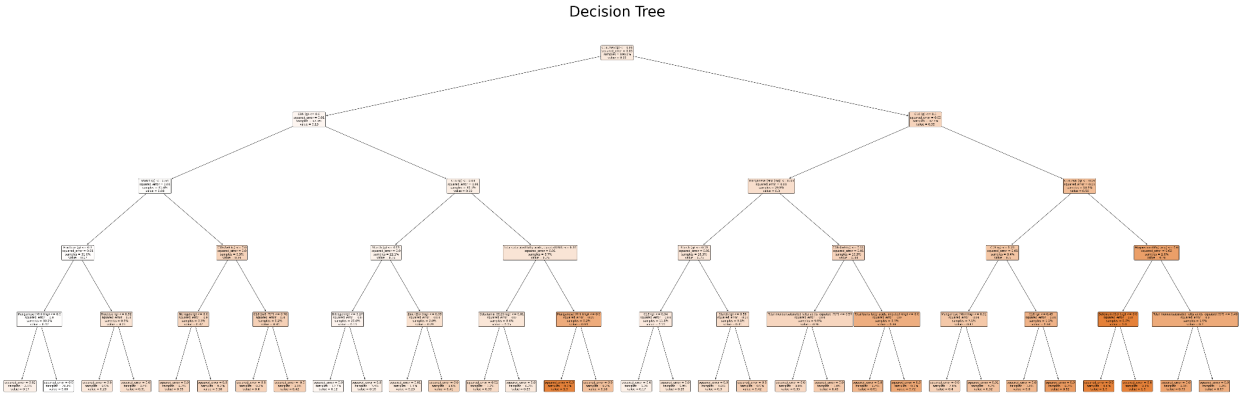


Figure 7: Regression tree

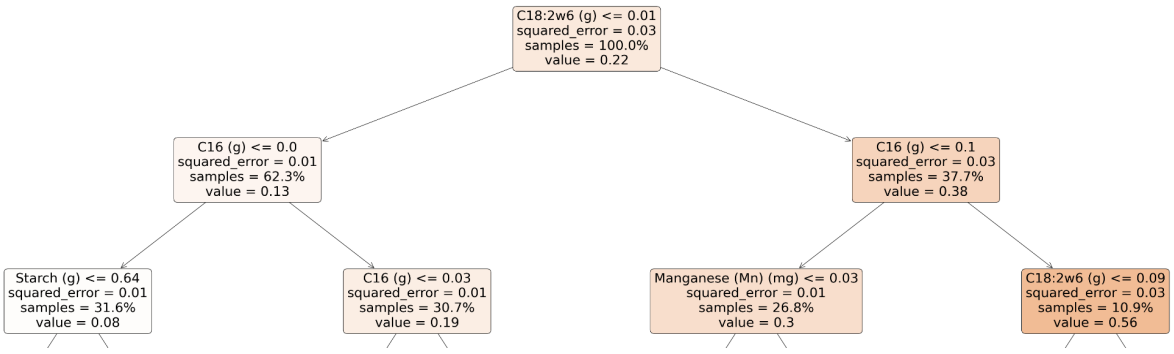


Figure 7 (broken up)

6/ Discussion

6.1/ Evaluation

	10-fold CV Train R-squared	10-fold CV Test R-squared	10-fold CV MSE (scaled)
Linear model 1	0.686	-0.089	0.041
Linear model 2	0.308	0.261	0.029
Regression Tree	n/a	n/a	0.023

Table 1

The residuals for Linear model 2 are more randomly distributed than those of Linear model 1 (figure 8, 9). The average R-squared on train data of Linear model 1 is significantly higher than that of Linear model 2. However, the average R-squared on test sets of Linear model 1 is not only smaller but also negative. So Linear model 1 does not generalize well, performing good on training data but worse than the mean of the target on test data. This overfitting problem may be due to noisy data. Linear model 2, only has a small discrepancy in average R-squared on train data and test data, and the average MSE of Linear model 2 is lower. Hence, residual analysis and evaluation scores conclude that Linear model 2 has a better performance than Linear model 1.

The Regression tree has the lowest average MSE, which indicates it may be the most suitable among 3 models, potentially due to the fact that tree regression is not as sensitive to outliers and over simplicity as linear regression.

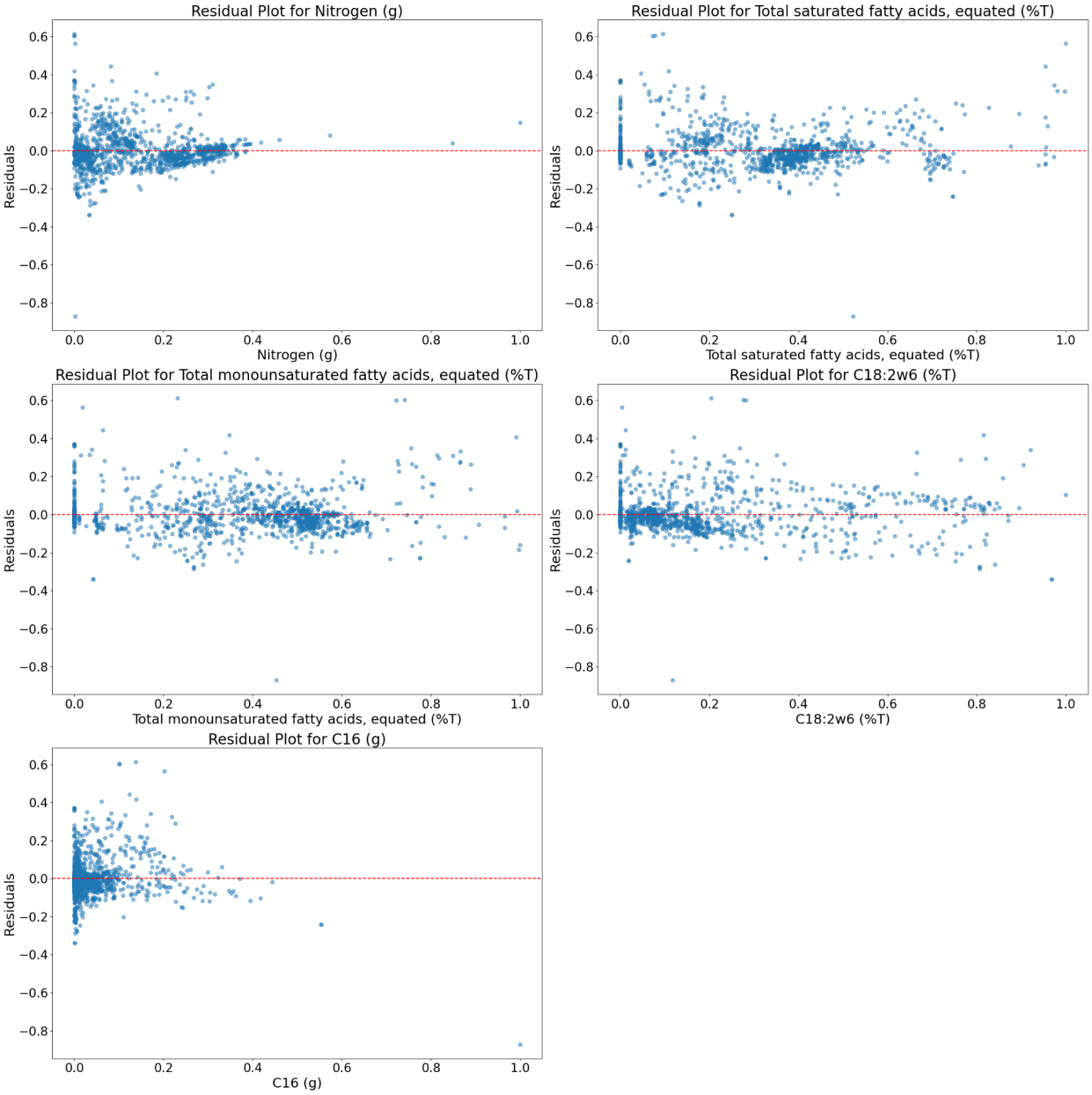


Figure 8: Residual plots for Linear model 1

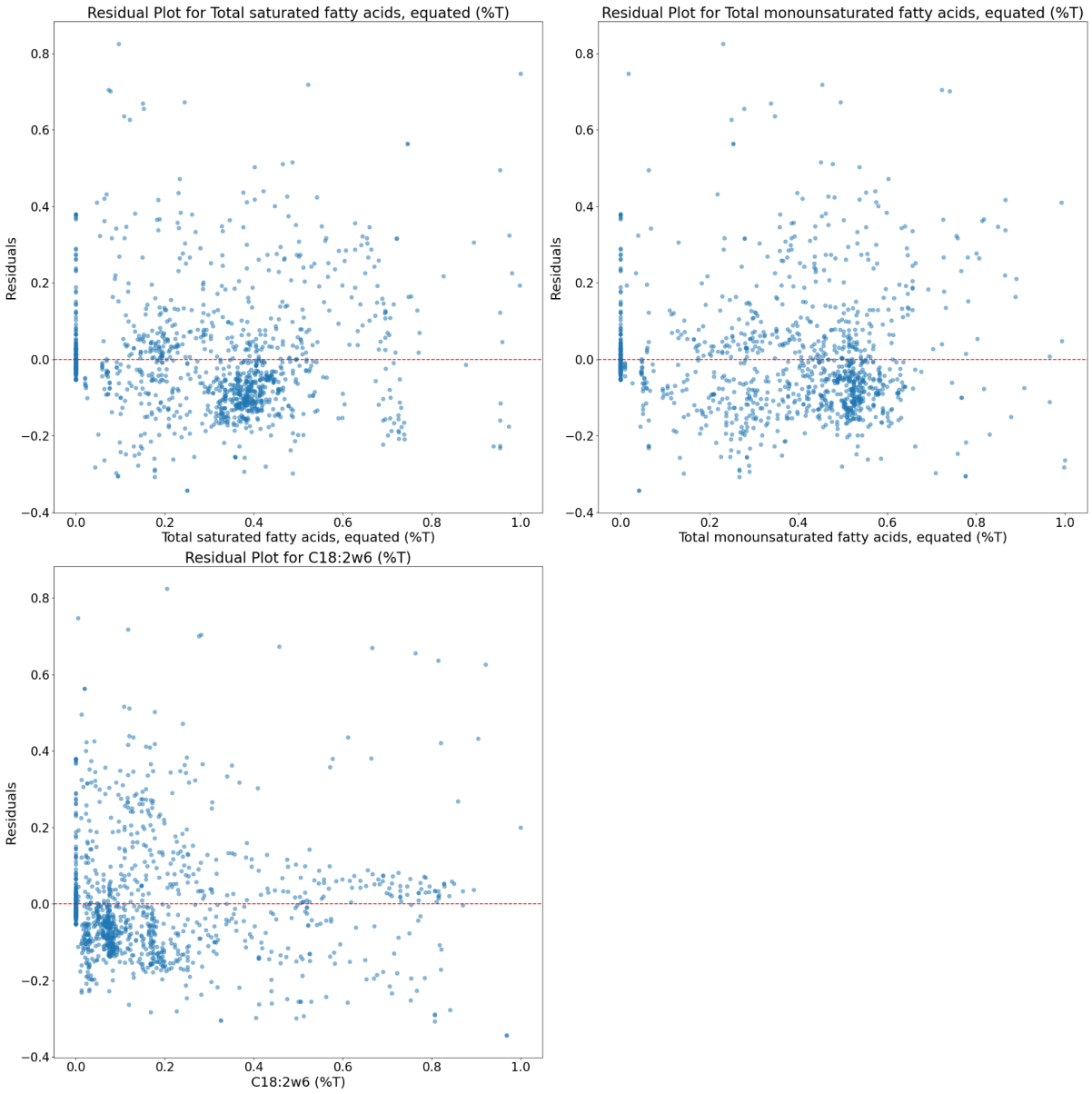


Figure 9: Residual plots for Linear model 2

6.2/ Conclusion

Positive coefficients in linear model 2 indicate that 'Total saturated fatty acids, equated (%T)', 'Total monounsaturated fatty acids, equated (%T)', and 'C18:2w6 (%T)' have positive relations with the target. Key variables for splitting in the Regression tree are 'C18:2w6 (%T)', 'C16 (g)', 'Manganese (Mn) (mg)', and 'Starch (g)' (figure 7). The scatter plots between each key variable and the target variable reveal positive relationships (figure 4). Hence, the decrease in energy may coincide with decreases in saturated fatty acids, monounsaturated fatty acids, linoleic acid (C18:2w6), palmitic acid (C16), manganese, and starch.

The decreases in saturated fatty acids and palmitic acid are desirable. However, monounsaturated fatty acids, linoleic acid, manganese, and starch should be preserved. Special attention should be paid to these beneficial nutrients while constructing low-calorie diets. Along with people with increased energy needs, this report suggests that people with inflammatory conditions or bone health concerns may also avoid low-calorie diets as linoleic acid and manganese intakes become important. From these findings, food manufacturers can focus on nutritional supplement products targeting customers who are following low-calorie meal plans. Community health workers can raise special attention to the mentioned nutrients, which can make low-calorie diets more balanced.

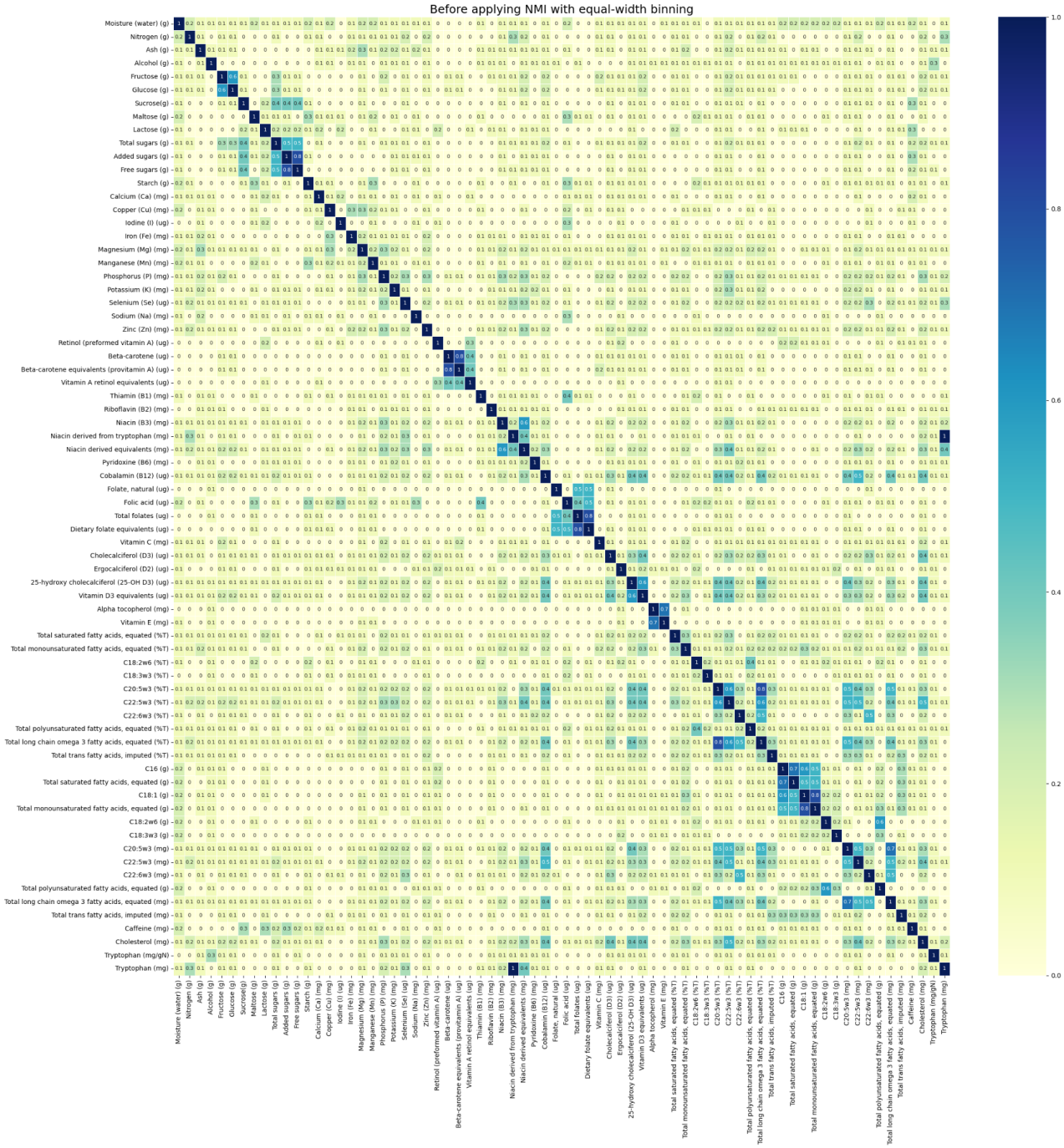
6.3/ Limitation and Improvement

KNN Imputation assumes that foods with similar amounts of other nutrients will have similar amounts of the nutrient we are trying to impute. Imputation can give more accurate results if the assumption is limited to foods of the same group. However, one limitation of the dataset is that some food groups only have a small number of foods presented (for example, special dietary foods (ID 12) with 2 items). Within such groups, it is much more likely for a feature to contain all NaN values, making imputation infeasible.

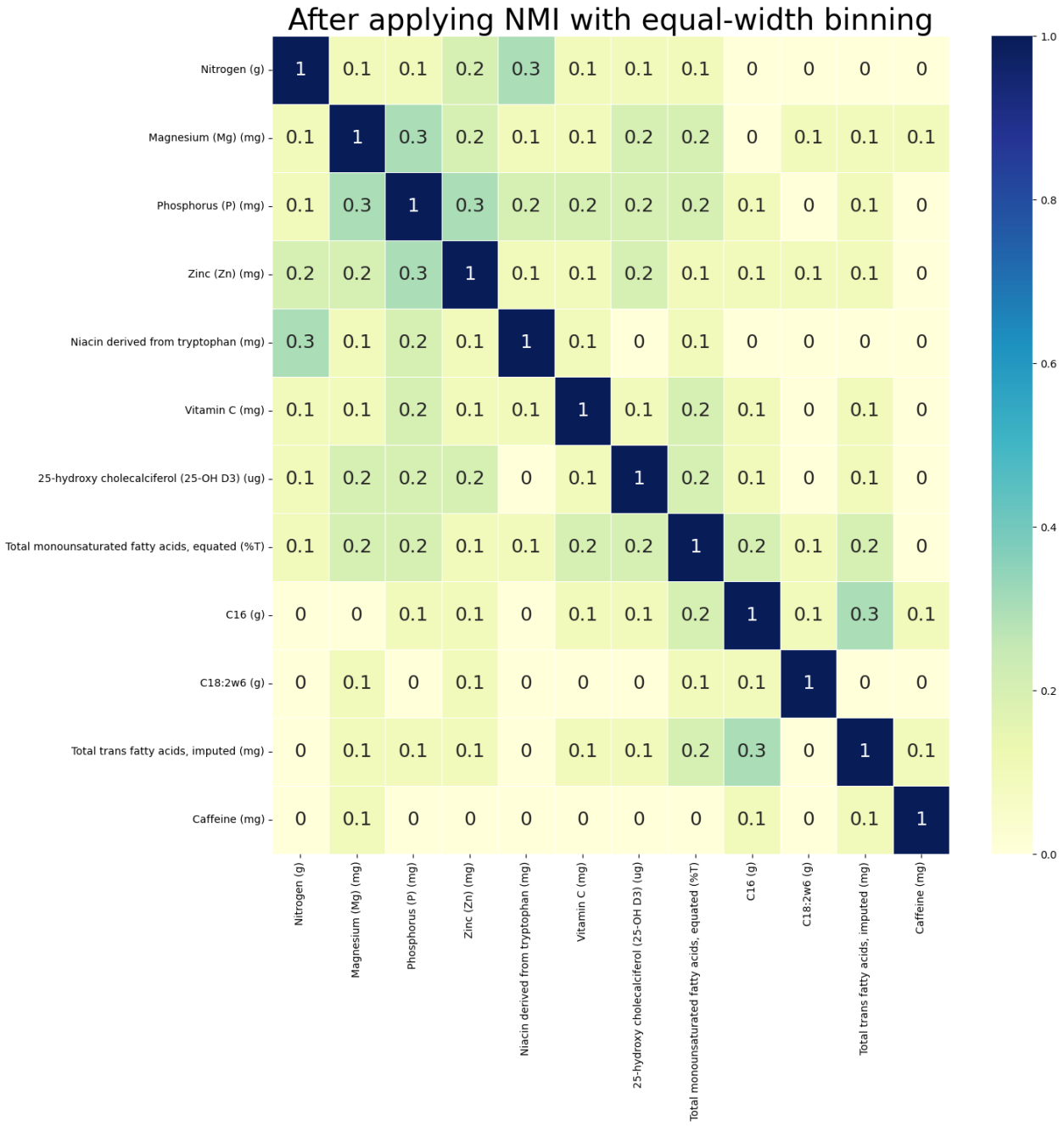
Many features have extreme values and are heavily right-skewed. After partly addressing this problem by Power Transform, plots on the diagonal of figure 3 still show skewed distribution, which hinders the performance of linear regression models. To improve, Interquartile Range (IQR) can be used to determine the boundaries, then removal or imputation can be performed to handle outliers.

The average R-squared on test sets of Linear model 2 is within the range 0.25-0.5 (table 1), implying that a small amount of variance is explained. Moreover, Regression Tree is believed to have better performance than both linear models. These points suggest non-linear relationships between explanatory variables and the target. Non-linear regression should be deployed to give more accurate models.

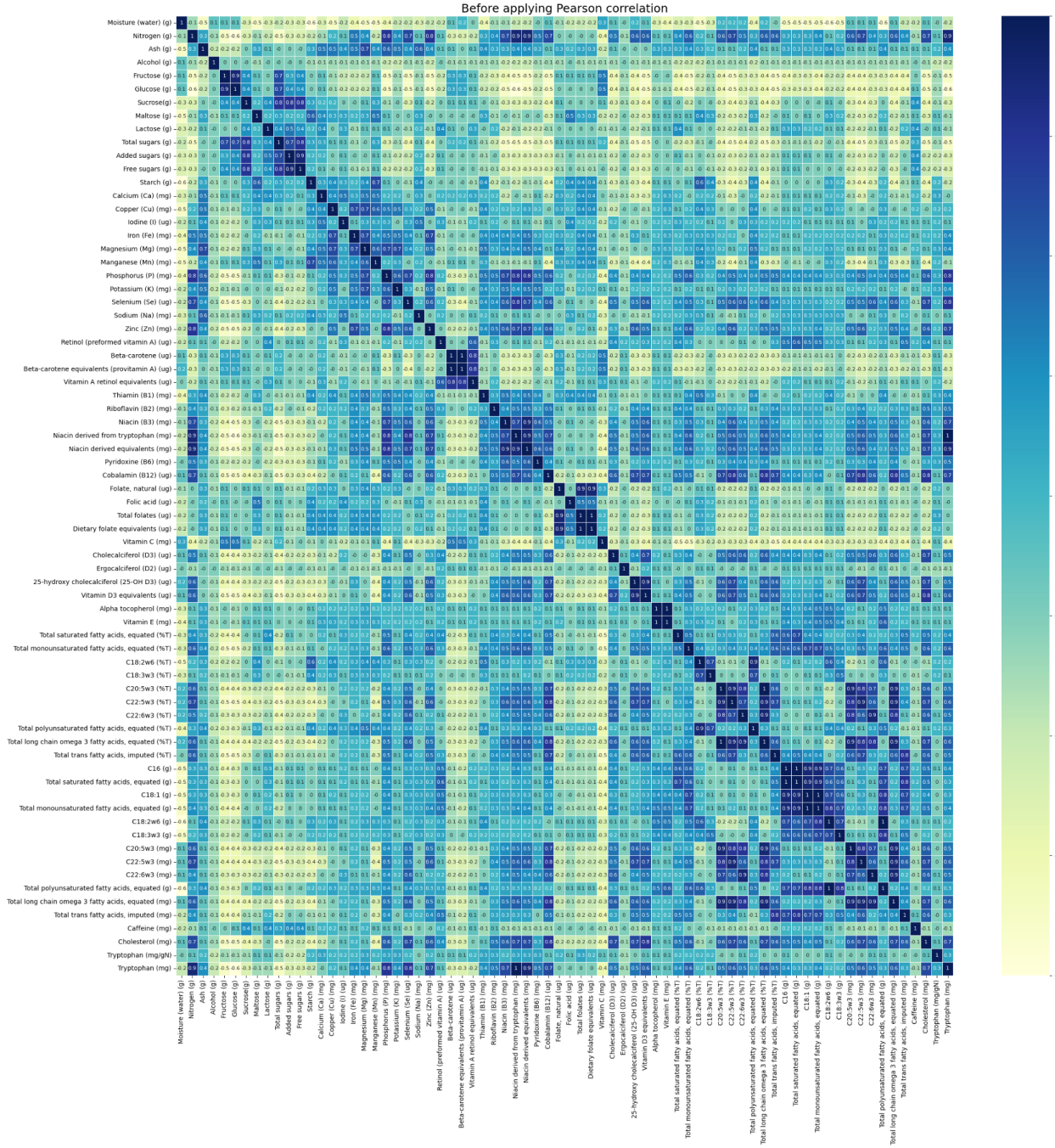
7/ Appendix



Appendix 1: Before NMI with equal-width binning



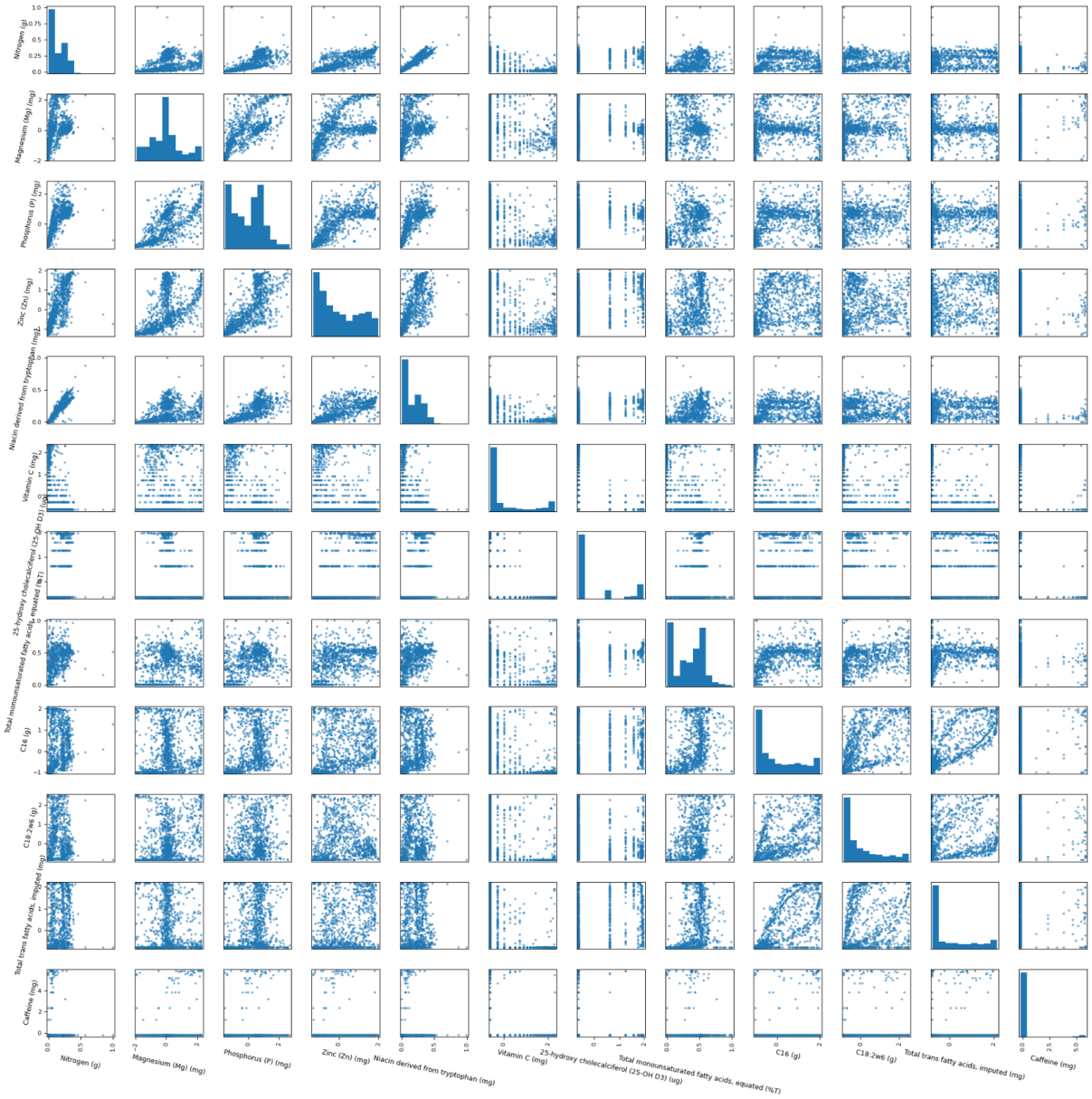
Appendix 2: After NMI with equal-width binning



Appendix 3: Before Pearson correlation

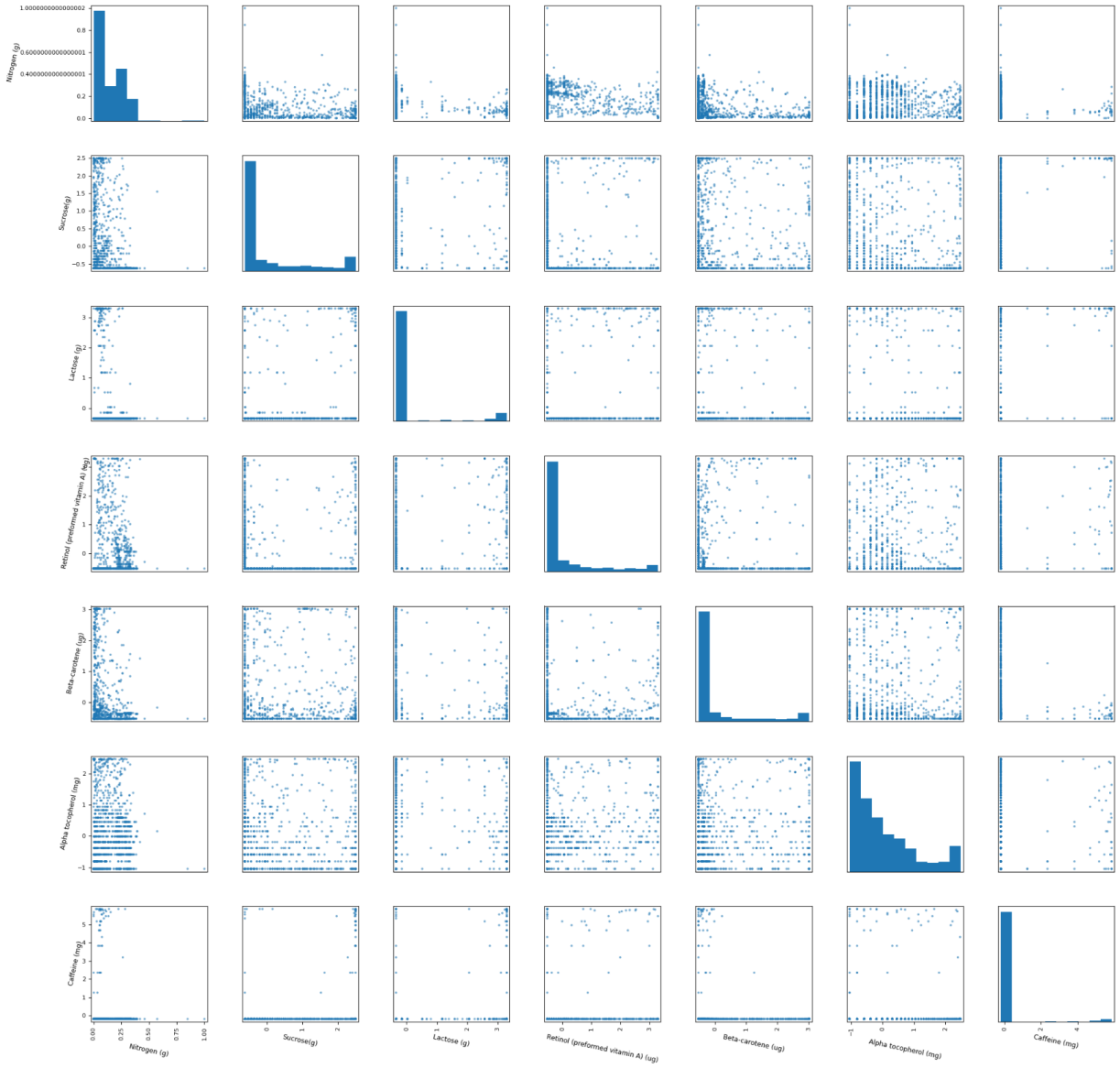
Appendix 4: After Pearson correlation

Scatter graph matrices for equal-width binning



Appendix 5: Scatter graph matrices for equal-width binning

Scatter graph matrices for Pearson correlation



Appendix 6: Scatter graph matrices for Pearson correlation

Reference list

Foodstandards.gov.au. (2016). *Download Excel files (Australian Food Composition Database - Release 1)*. [online] Available at: https://www.foodstandards.gov.au/science/monitoringnutrients/afcd/Pages/downloadableexcelfiles.aspx.

www.foodstandards.gov.au. (n.d.). *Classification of foods and dietary supplements*. [online] Available at: https://www.foodstandards.gov.au/science/monitoringnutrients/ausnut/classificationofsupps/Pages/default.aspx.

www.mydailyintake.net. (n.d.). *Daily Intake Guide: Healthy eating, made easy. Front-of-pack labelling for food and drink in Australia. - Calculating Food Energy*. [online] Available at: http://www.mydailyintake.net/calculating-energy/.