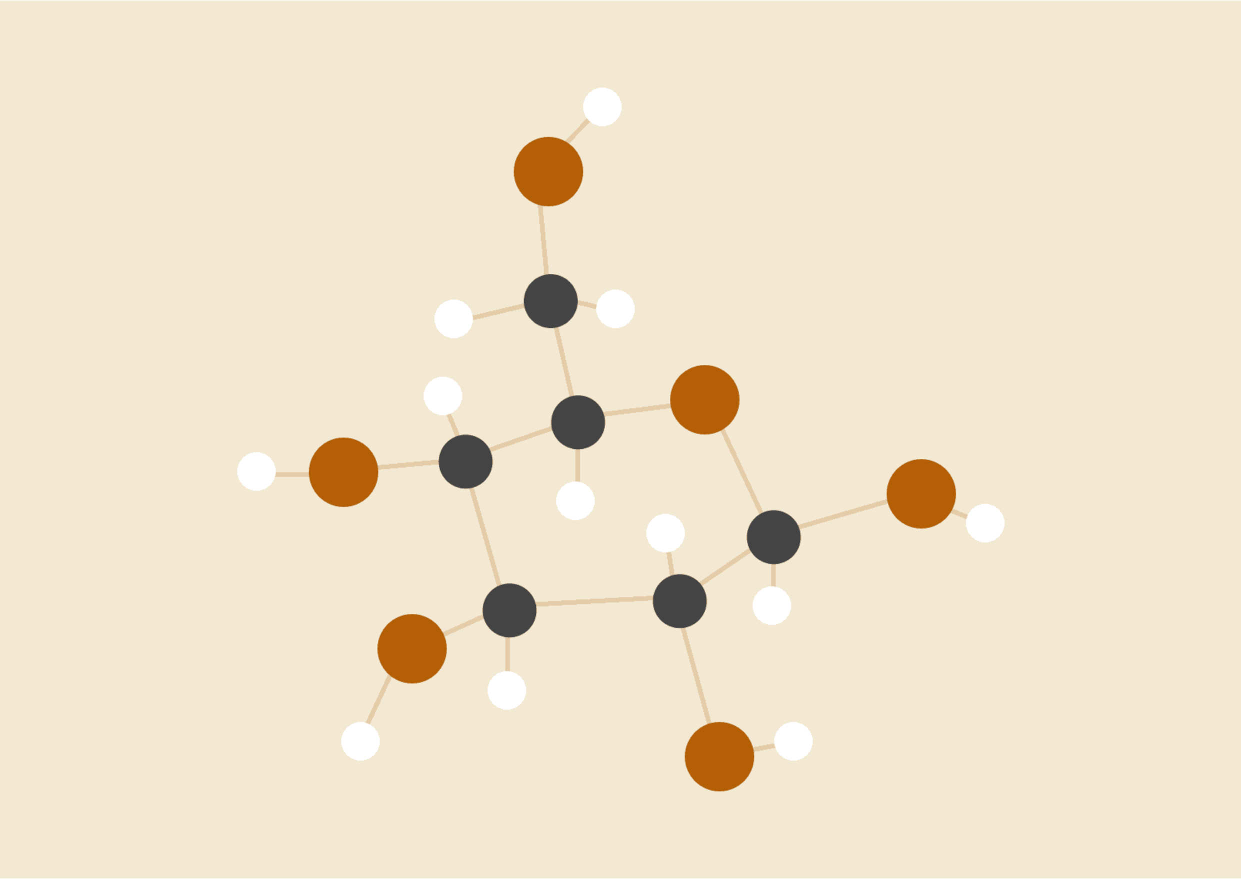
END OF INTERNSHIP REPORT

PROTEIN PREDICTION



Hung Long Nguyen – 1157436

RESEARCH QUESTION

Can we predict the protein content of a food item based on its other nutrient values? Which nutrients have the greatest impact on protein content?

TARGET AUDIENCE

Food scientists or nutritionists are interested in predicting the protein content of food products, or individuals or companies involved in the food industry may be interested in quality control or labelling requirements related to protein content.

Proteins are essential nutrients that play a vital role in the human body. They are the building blocks of many tissues, including muscles, skin, hair, and nails, and are involved in numerous biological processes such as cell growth and repair, immune function, and hormone production.

Therefore, this report will investigate a model that can predict the protein content of food products through other nutrition, which is crucial for quality control and labelling requirements. With increasing health awareness and changing consumer preferences, the demand for high-quality protein-rich food products has been on the rise.

This report will explore various techniques and methods for predicting the protein content of food products, and the study results will allow people to pay more attention to other factors when buying any new food brands in order to build a good eating habit for society.

DATASET

The data source [AFCD22] for this project is the nutrient file from Australian Food Composition Database - Release 2 which is provided in Excel form. It contains the nutrient data available for each food, with the nutrient data provided in two ways:

Per 100 g – all foods and all beverages are reported per 100 g edible portion.

Per 100 mL – beverages and other liquid foods only, reported per 100 mL edible portion.

The initial dataset includes entries with a value of zero, indicating that the associated nutrient has undergone testing and has been determined to have no value. Conversely, the absence of entries signifies that the nutrients in those specific food items have not been assessed, resulting in the presence of missing values.

Quick Description:

Field	Description of field
Key	Food identification code used to identify each food.
Nutrient ID	Nutrient identification code (Shorthand way of presenting the nutrient e.g., Protein is 'PROT').
Description	Full nutrient name e.g. 'Protein'.
Scale	Units in which the nutrient is presented e.g., grams.
Value	Value of the nutrient reported.
Category	Nutrient category the nutrient belongs to e.g., Calcium belongs to the 'MNS' category.

Cited from Australian Food Composition Database - Release 2 - January 2022

DATA WRANGLING AND FEATURE SELECTION

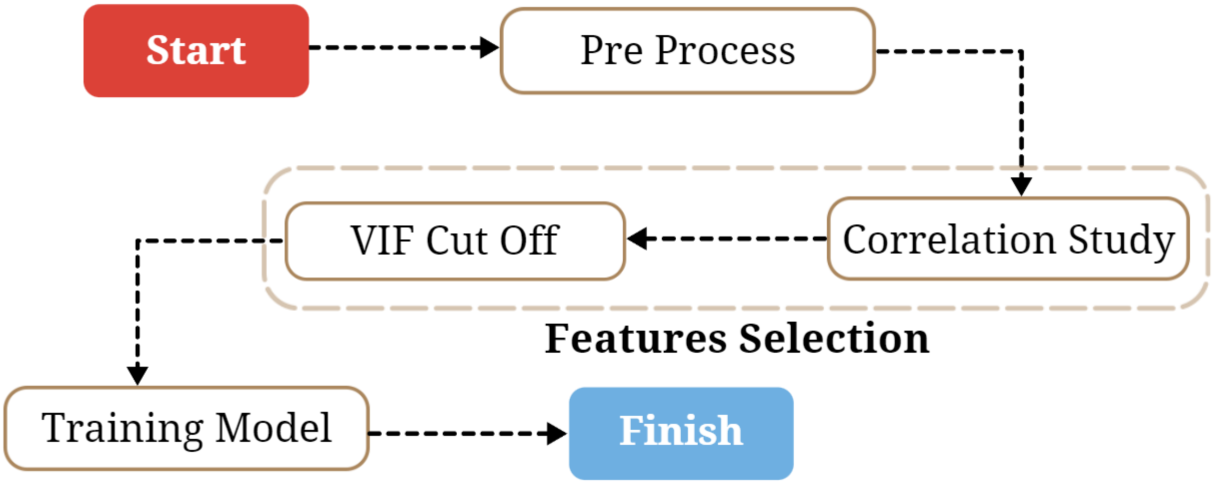


Fig 1: Modelling steps

1. Data Pre-processing:

The source data contains a high number of missing values, a large amount of data that is irrelevant to the model, which could significantly affect the performance of the predicting model. Thus, it is inevitable that the data must undergo a pre-processing and wrangling process.

To be able to use our data, we must first clean and then aggregate our data into the desired values. Fortunately, the source data was rather clean and devoid of human errors.

The original dataset contains entries with zero, indicating that the corresponding nutrient has been tested and found to have no value. On the other hand, missing entries indicate that the nutrients in those particular foods have not been tested, leading to the presence of missing values. Therefore, it is necessary to address the issue of missing values in the dataset.

Using simpler methods like filling missing values with zeros or using means/medians can negatively impact the model's performance and result in inaccurate estimates and unreliable predictions. These methods introduce biases, distort patterns, and oversimplify relationships between variables. They disregard the potential variability and interdependencies among variables, compromising the model's ability to capture the true underlying structure of the data. To address missing data without compromising performance, advanced techniques like Random Forest, Non-Negative Matrix Factorization, Jaccard similarity, or K-nearest-neighbors are recommended. These methods consider interdependencies and provide a more accurate and reliable approach to handling missing data, ensuring trustworthy results. However, they can be excessively complicated within the scope of this project.

Therefore, we opted to remove features with missing values as a method to address this concern. This decision was made after careful consideration, as it represents an optimal approach for handling missing values without compromising the integrity of the regression model or unnecessarily escalating the complexity of the process.

2. Feature selection

Prior to building the model, it is crucial to perform feature selection as it enables the reduction of feature dimensions. This is essential since having redundant features can have significant consequences, such as adding unnecessary computational effort, resulting in prolonged training times. Moreover, it can lead to overfitting, which impairs the model's generalisability to new samples and can cause poor readability and interoperability of the model.

2.1. Correlation cutoff:

There are more than 120 nutrients present in the data, it is obvious that we do not need all these nutrients, and some are irrelevant to our model.

Therefore, in the first step of feature selection, we will be cutting off nutrients that are not correlated with protein. Correlations that are lower than 0.3 are considered to be insignificant or uncorrelated and will be removed suggested by researcher Jacob Cohen. After filtering, we obtain the remaining nutrients presented in Table 1.

Table 1:

Available carbohydrate, with sugar alcohols	-0.311739
Available carbohydrate, without sugar alcohols	-0.398878
Total sugars	-0.307079
Vitamin D3 equivalents	0.324242
C22:5w3	0.340197
Cholesterol	0.407516
25-hydroxy cholecalciferol (25-OH D3)	0.424564
Niacin (B3)	0.441776
Zinc (Zn)	0.633874
Niacin derived equivalents	0.656785
Phosphorus	0.675865
Niacin derived from tryptophan	0.913233
Tryptophan	0.913233
Nitrogen	0.998240

2.2. Correlation matrix:

Upon closer examination of the filtered features, we have observed significant correlations among certain features, leading to concerns regarding multicollinearity.

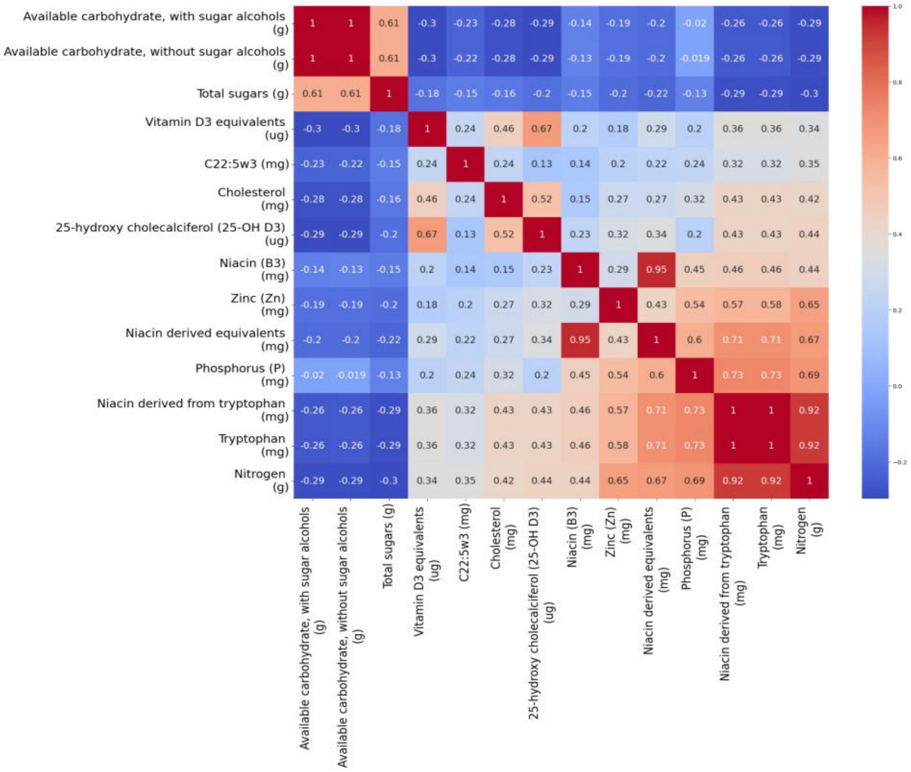


Fig 2: Correlation heat map

Filtered features with high multicollinearity pose a challenge in multiple regression models as the inputs are interdependent, making it difficult to estimate the impact of the independent variables on the dependent variable. The model may produce regression coefficients that are not statistically significant, as the underlying effect is being accounted for multiple times. This makes it challenging to determine which independent variable is affecting the dependent variable.

2.3. Variance inflation factor:

To ensure the model is properly specified and functioning correctly, we have implemented the Variance Inflation Factor (VIF) method, which measures how an independent variable's variance is influenced by its correlation with other independent variables. When significant multicollinearity exists, the VIF will be large for the variables involved. Combining or eliminating collinear variables can resolve the issue.

The VIF formula is:

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where: R_i^2 = unadjusted coefficient of determination for regressing the i^{th} independent variable on the remaining ones.

Available carbohydrate, with sugar alcohols	3193.69694554369
Available carbohydrate, without sugar alcohols	3185.9549535588467
Total sugars	2.0006075444663494
Vitamin D3 equivalents	2.3630150396266254
C22:5w3	1.366654659369165
Cholesterol	2.086135826978604
25-hydroxy cholecalciferol (25-OH D3)	2.8173226950062906
Niacin (B3)	493.414102049965
Zinc (Zn)	3.023319880878934
Niacin derived equivalents	971.0610695168245
Phosphorus	6.241521127443208
Niacin derived from tryptophan	283190.90575510653
Tryptophan	283210.7659990532
Nitrogen	16.06390513960722

In general terms,

VIF equal to 1 = variables are not correlated

VIF equal to 5 = variables are moderately correlated

VIF greater than 10 = variables are highly correlated

In this model, a VIF cutoff value of 10 has been chosen as the threshold. The application of this criterion which removes features with a VIF value greater than 10, has resulted in the following outcomes.

Total sugars, Vitamin D3 equivalents, C22:5w3, Cholesterol, 25-hydroxy cholecalciferol (25- OH D3), Zinc (Zn), Phosphorus (P).

MODELLING ANALYSIS

1. Analysis Methods:

The objective of this study was to establish a predictive model for protein by examining the correlation between protein and various nutrient factors. Since all our data is given numerically, clustering or classification techniques would be an inappropriate oversimplification. The regression models' coefficients and intercepts were employed to identify the nutrients that have a significant correlation with protein.

2. Preliminary Analysis:

Looking at Figure 3-9 below, we see that the relationships of the chosen features and protein appear to be somewhat linear indicating that linear regression may be an appropriate model to fit and that transformations of the variables are likely, not necessary.

Figure 3: Total sugars (g) vs Protein (g)

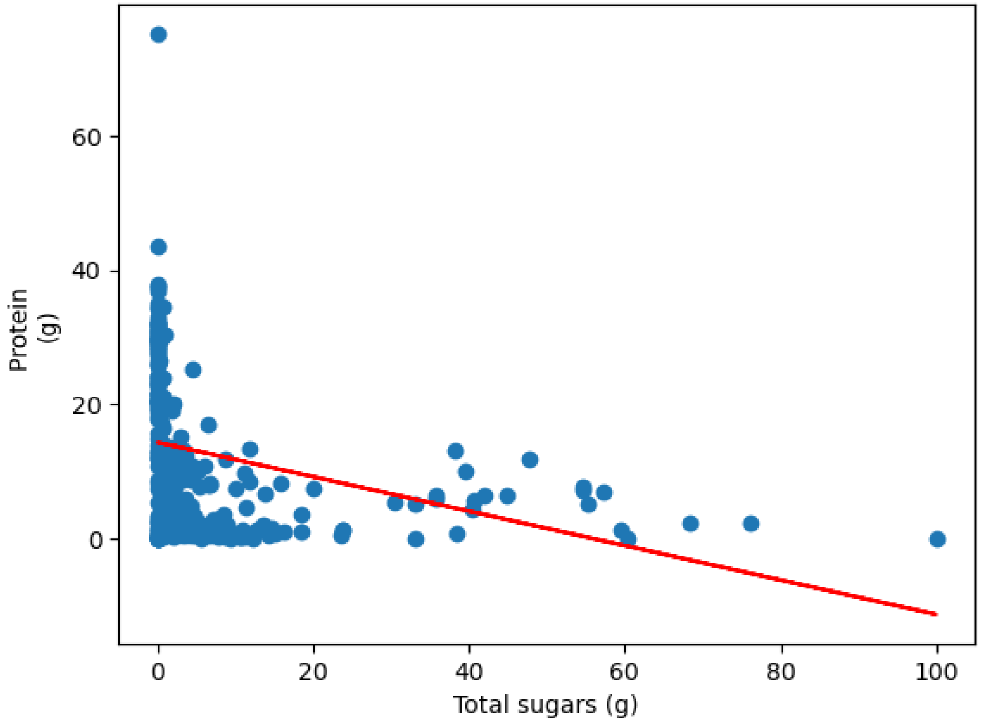


Figure 4: Vitamin D3 equivalents (ug) vs Protein (g)

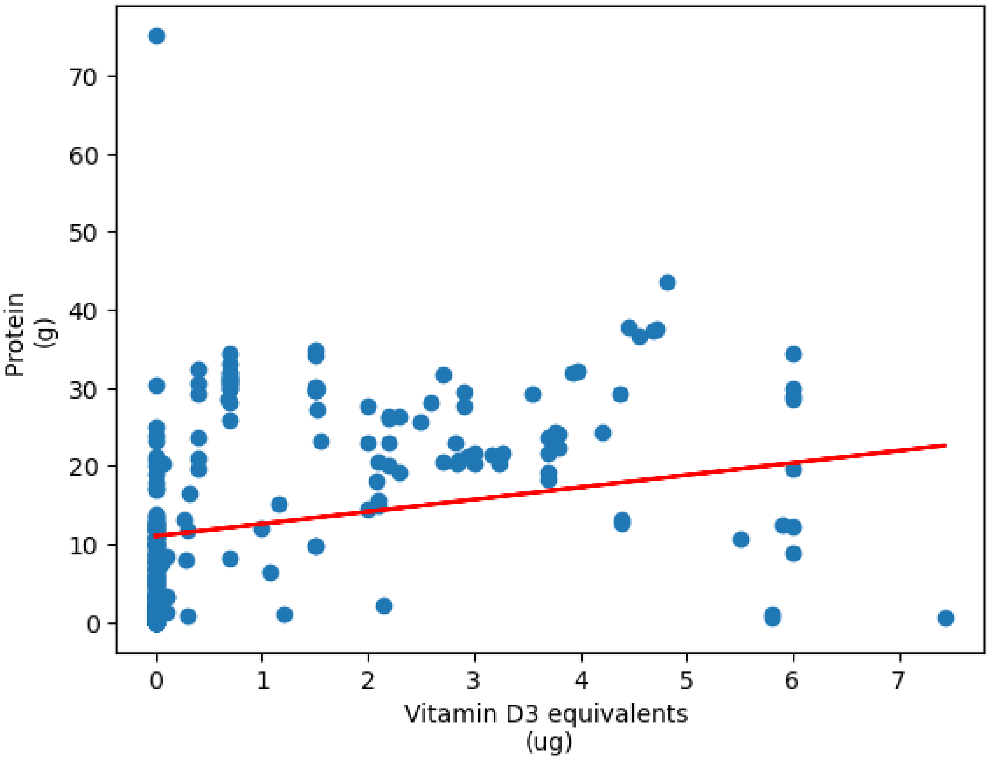


Figure 5: C22:5w3 (mg) vs Protein (g)

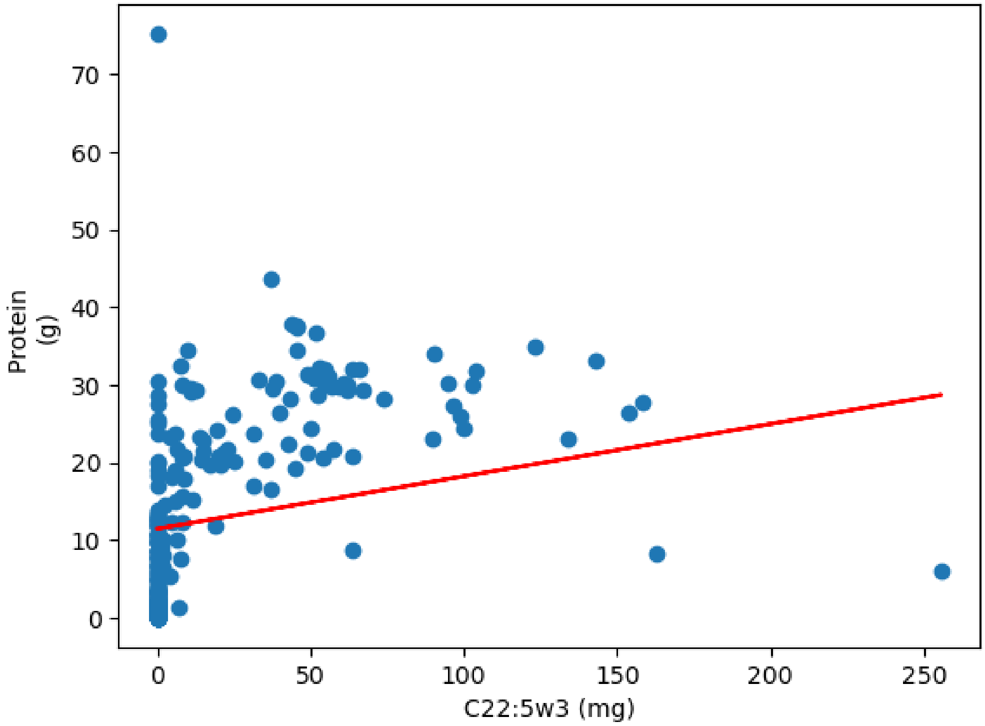


Figure 6: Cholesterol (mg) vs Protein (g)

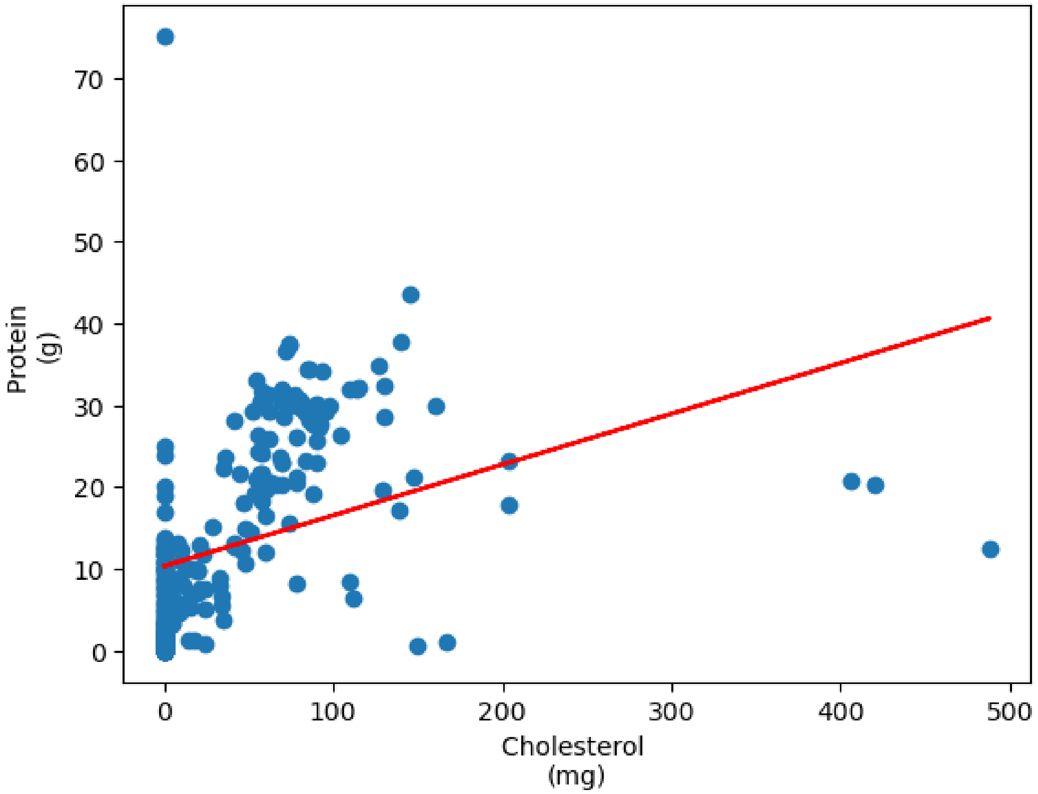


Figure 7: 25-hydroxy cholecalciferol (25-OH D3) (ug) vs Protein (g)

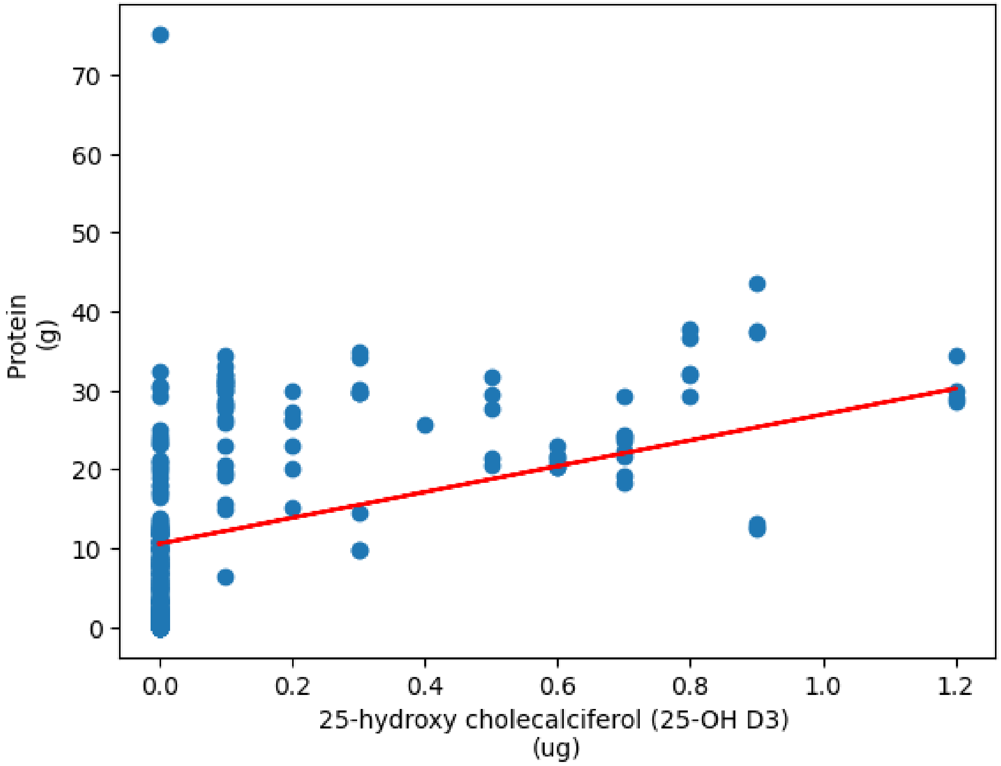


Figure 8: Zinc (Zn) (mg) vs Protein (g)

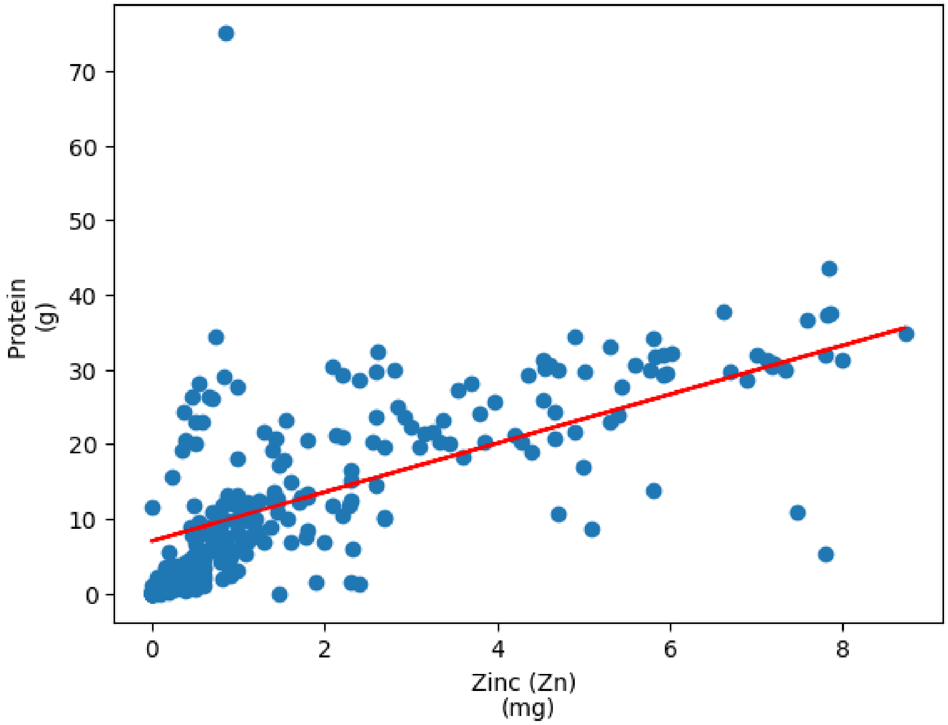
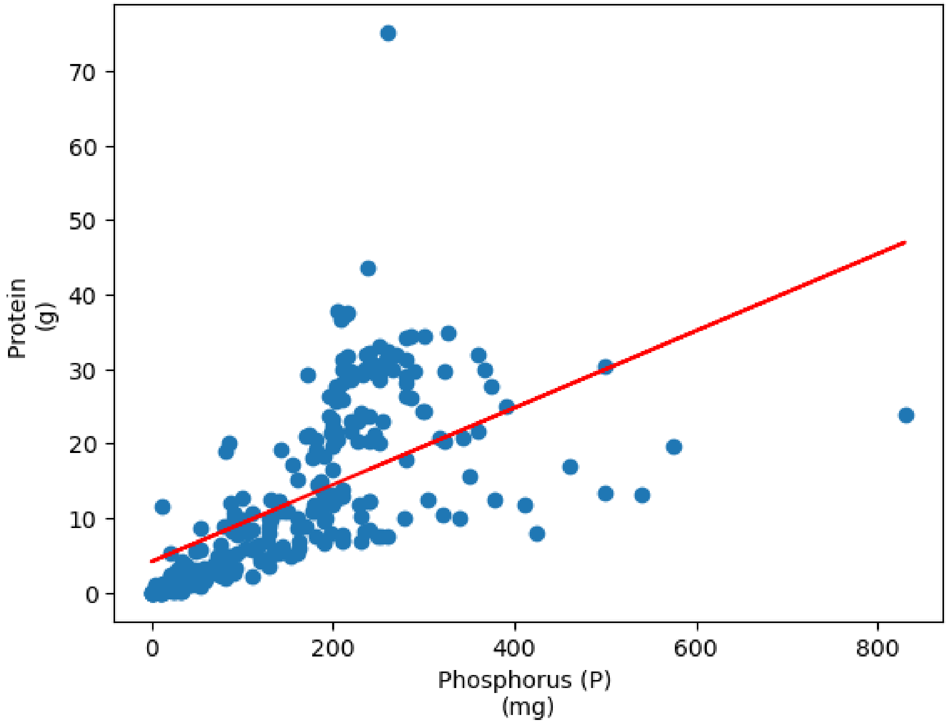
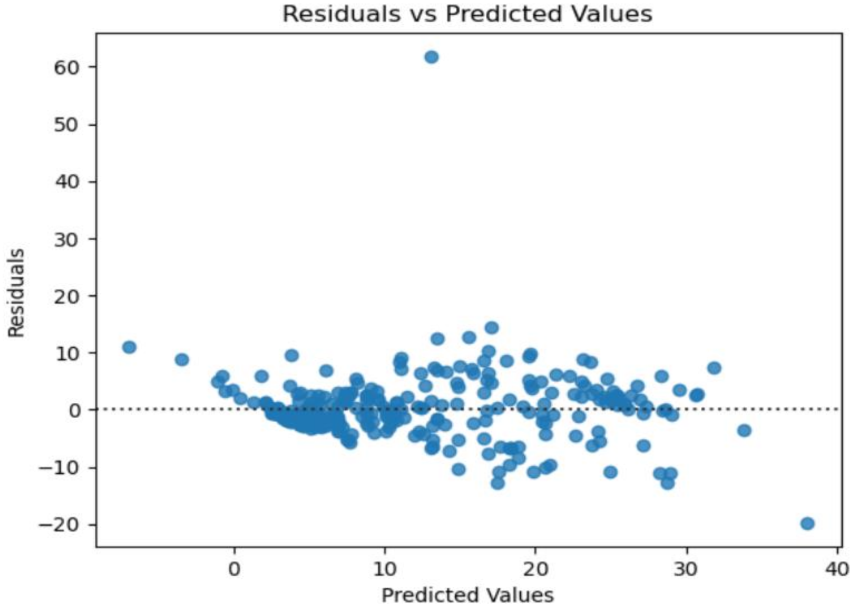


Figure 9: Phosphorus (P) (mg) vs Protein (g)



In order to ensure that our linear regression model is an appropriate fit for the data, we conducted a preliminary analysis by generating a scatter plot for the residuals and predicted values. By examining the plot, we were able to gain insight into the variability of the residuals and how they change across the range of predicted values.



Upon close inspection of the plot, we observed that the residuals appeared to be fairly constant with few outliers and exhibited equal variability across the entire range of predicted values. This consistent pattern in the residuals is an indication of homoscedasticity, which suggests that our linear regression model is a good fit for the data.

Furthermore, the fact that there is no discernible bias or trend in the residuals indicates that the model has effectively captured the majority of the variation in the data. As a result, we can be confident in the reliability and accuracy of our linear regression model, and we can move forward with our analysis knowing that the model has effectively accounted for the underlying patterns in the data.

3. Modelling:

3.1 Train test split:

Before further analysis, the dataset was partitioned into training and testing subsets with an 80/20 split. The training subset was used for initial exploration and model development, while the testing subset was utilised to evaluate the model's performance and generalisability. In addition, k-fold cross-validation was employed to minimise the impact of chance fluctuations and improve the reliability of the results. This model splits the data into 10 folds, then trains and tests the model on the various folds and averages the results. This leads to our model not being overfitted to a particular split.

3.2 Linear Regression Function:

This model is then allowed to learn on the training set, outputting the regression interceptions and coefficients, so that we can obtain a functional expression for this ternary linear regression model:

$$\begin{aligned}\boldsymbol{Protein} = {} & \mathbf{3.470} - \mathbf{0.104} \times (\boldsymbol{Total\ sugars}) - \mathbf{0.038} \times (\boldsymbol{Vitamin\ D3}) \\ & + \mathbf{0.029} \times (\boldsymbol{C22{:}5w3}) + \mathbf{0.009} \times (\boldsymbol{Cholesterol}) \\ & + \mathbf{7.720} \times (\mathbf{25} - \boldsymbol{hydroxy\ cholesterol}) + \mathbf{1.486} \times (\boldsymbol{Zinc}) \\ & + \mathbf{0.032} \times (\boldsymbol{Phosphorus})\end{aligned}$$

DISCUSSION

1. Conclusions:

The linear regression model built can explain the objective problem of this report [CHAP7]. People who want to choose high-quality food products based on protein content should focus on 25 - hydroxy cholecalciferol, zinc phosphorus or even cholesterol and C22:5w3 (aka Fat Acid). Among those nutrients, phosphorus stands out as the highest coefficient, which is quite reasonable because scientists have already proven that "...All proteins contain nitrogen and sulfur atoms and may also contain phosphorus atoms and traces of other elements…" (Chemistry Libre texts). Therefore, phosphorus is a good source of nutrients that can predict the protein content and people can rely on them to choose and pick the healthiest product.

Model	Training R^2	Testing R^2	10-fold CV RMSE
Model 1	0.66	0.75	5.80

The interrelationships among these metrics can aid us in comprehending the model's performance. The Training R^2 score of 0.66 shows that the model has some ability to predict the target variable on the training data but there is still room for improvement. The Testing R^2 score is better at 0.75, indicating that the model is capable of generalising well to unseen data. Furthermore, the 10-fold cross-validation Root Mean Square Error (RMSE) value of 5.80 suggests that the model exhibits relatively low bias or variance and performs well in general.

2. Limitations and Improvements:

2.1 Limitations:

- The performance metrics in the table may not be representative of the model's performance on unseen data.
- The model's performance may be affected by the quality and the quantity of the training data if it is not diverse enough or does not include enough examples.
- The choice of hyper-parameters can have a significant impact on the model's performance.

2.2 Improvements:

- Collect more data: Collecting more diverse and representative data can help improve the performance of the model.
- Experiment with different hyper-parameters: Adjusting the hyper-parameters of the model can optimise the performance. Moreover, try out different values using grid research or randomised search to find the best combination.
- Use more advantage techniques: Exploring more advanced techniques such as assembling, stacking or transfer learning to improve the performance and reduce bias or variance.

REFERENCES

[AFCD22] Cited from Australian Food Composition Database - Release 2 - January 2022

[CHAP7] Cited from chapter 7: Amino Acids, Proteins, and Enzymes - Chemistry libretexts

[PFPM] Nordic Statistical Meeting (2022), Predicting a food product’s missing nutritional values using machine learning and matching algorithms from natural language processing

[FCT] National Library of Medicine (2020), Evaluating missing value imputation methods for food composition databases

[JC] Jacob Cohen, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences

[VIF] Investopedia(Feb 2023), Variance Inflation Factor (VIF)