Data Analysis: Obesity Prevalence

Code

1 Introduction

Obesity is a significant public health concern, influenced by various socio-economic and lifestyle factors. This analysis uses data from the 2008-2012 Scottish Health Surveys to explore the relationship between demographic variables, lifestyle habits, and obesity prevalence in Scotland. The dataset includes information on age, gender, employment, dietary habits, and obesity status.

The objective is to examine trends in obesity prevalence over the survey years and identify differences in obesity rates based on socio-economic and lifestyle factors. The findings will inform public health strategies aimed at addressing obesity.

Section 2 provides an exploratory analysis of the dataset, including trends in obesity across different groups. Section 3 presents the results of a regression analysis, including model diagnostics to assess the model's fit and assumptions. Concluding remarks are in Section 4.

2 Exploratory data analysis

2.1 Distribution of each variable

Figure 1 illustrates the distribution of key categorical variables in the dataset, including Age Group, Employment Status, Vegetable and Fruit Consumption, and Sex. Age Group distribution indicates that middle-aged individuals (35-64 years) constitute the largest proportion, while younger (16-24 years) and older (75+ years) groups have lower representation, with females slightly outnumbering males across most age groups. Employment Status shows that females have a higher proportion in the employed category but also dominate in "looking after home/family" and "permanently unable to work" categories, reflecting gender differences in social roles. Vegetable and Fruit consumption data reveal that most individuals report regular intake, with females exhibiting a higher proportion of consumption compared to males, potentially indicating gender-related differences in health behaviors. Sex distribution suggests a slightly higher number of females than males in the dataset, though the overall distribution remains relatively balanced.

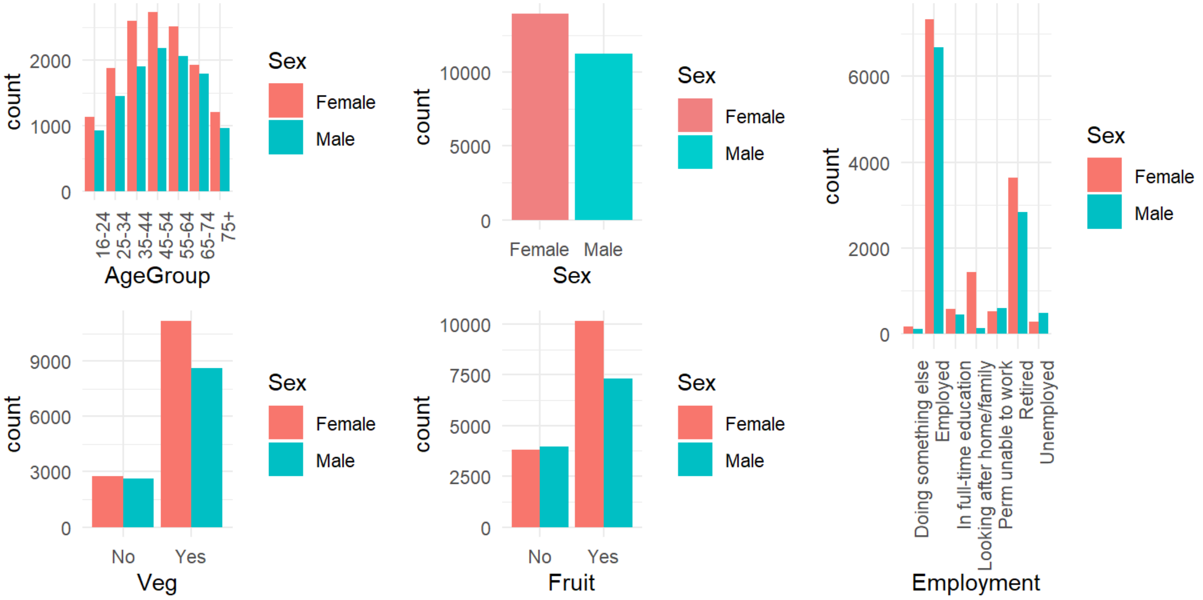

Figure 1: Distribution of Age Group, Employment Status, Vegetable Consumption, Fruit Consumption, and Sex by Gender in the Dataset.

2.2 Obesity changes and trend over time (2008-2012)

Figure 2 shows that the yearly obesity rates in Scotland from 2008 to 2012 show periodic fluctuations but remain relatively high. The rate dropped to its lowest point in 2009 (28.79%), then rose sharply in 2010 to its five-year peak

(30.47%), before gradually stabilizing in 2011-2012. This fluctuation in 2010 may reflect an increase in health risk factors, while the slowdown after 2011 could be linked to health interventions and improved awareness of physiological health management. Overall, obesity rates fluctuated within a range of 29% to 30.5% during this period.

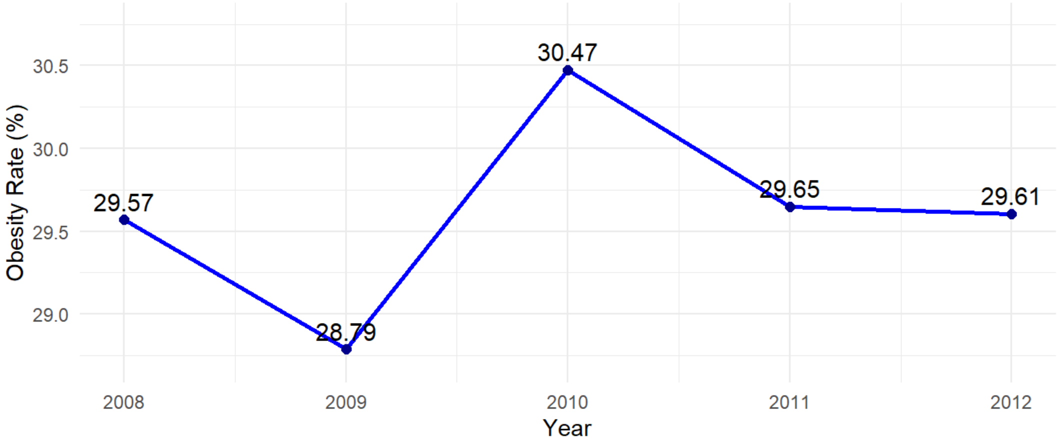

Figure 2: Yearly Obesity Rate Trend in Scotland (2008-2012)

2.3 Group-wise Obesity Rate Analysis

In this section, we explore the obesity rates by various grouping variables (Age Group, Employment, Vegetable Intake, and Fruit Intake) stratified by Sex. The obesity rates are counted and generates Figure 3.

Obesity rates peak at ages 55-64 before declining, with gender differences more pronounced in younger groups, where females have higher rates, but this trend reverses slightly in middle age due to metabolic and lifestyle changes. Employment status strongly influences obesity, with the highest prevalence among those permanently unable to work and the lowest among students. Caregiving roles and job-seeking categories also reveal gender disparities, reinforcing the link between socioeconomic conditions and obesity.

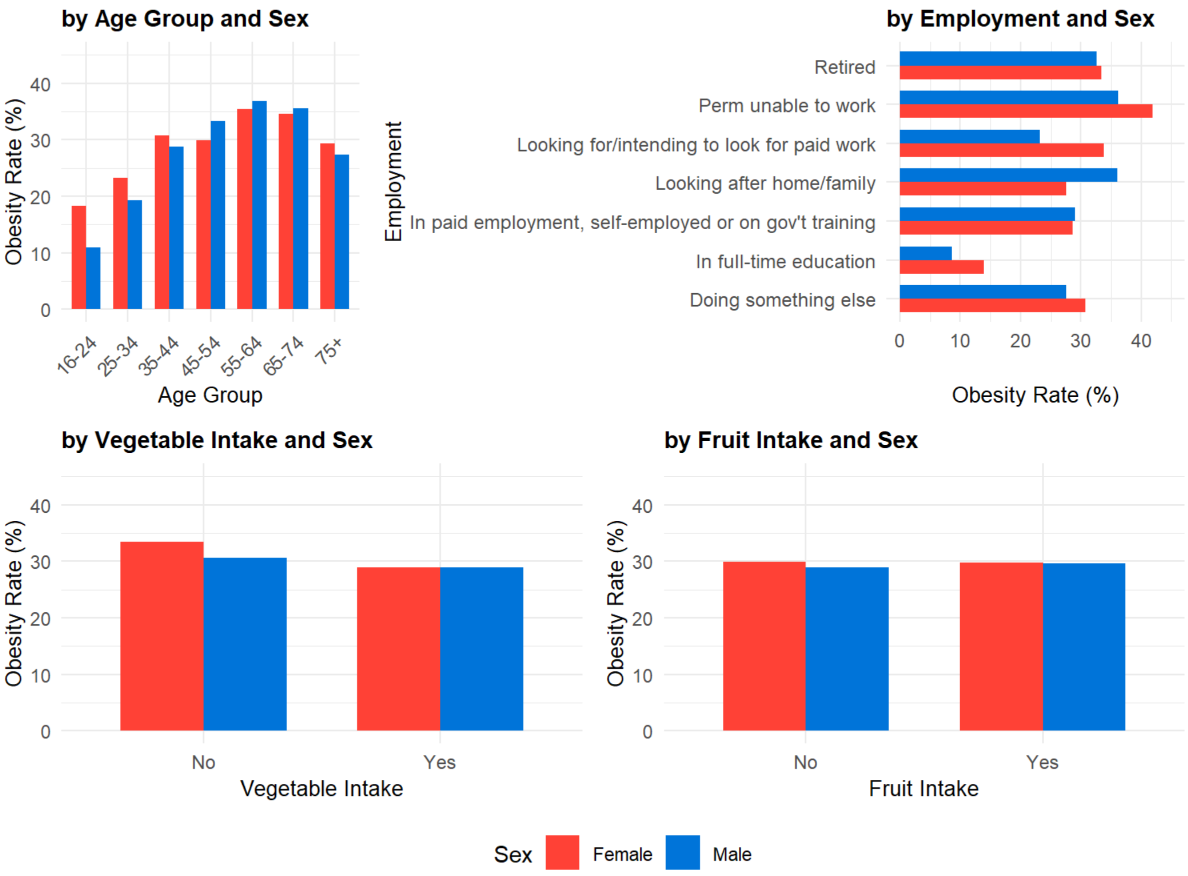

Figure 3: Obesity Rates by Demographic and Lifestyle Factors.

Dietary habits play a role, as higher vegetable intake is linked to lower obesity rates, particularly in females, suggesting a protective effect. However, fruit consumption shows little association with obesity, indicating potential differences in dietary impact. Age remains the most consistent predictor, while employment status shows the greatest variation, particularly for students and those unable to work. Gender differences fluctuate across factors, highlighting the complexity of obesity risk.

2.4 Variable Relationships

2.4.1 Correlation Analysis

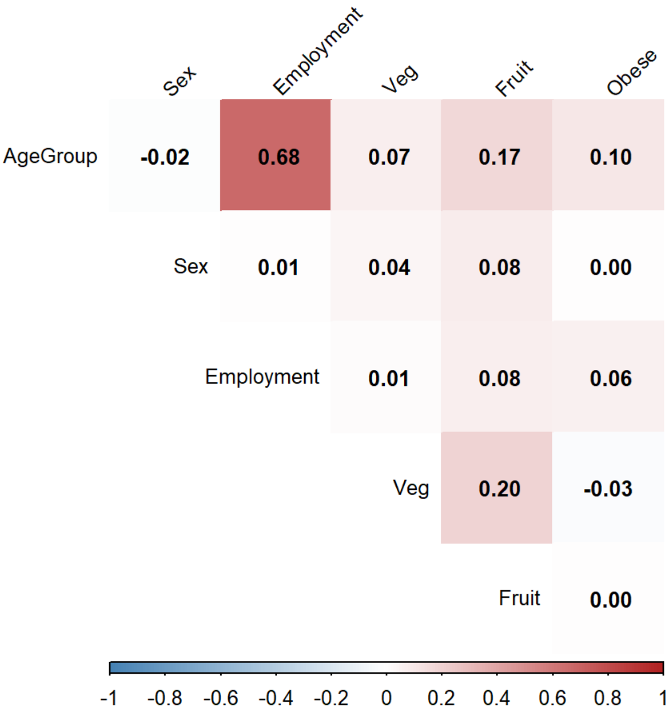

Figure 4: Correlation Heatmap.

In Figure 4, age shows the strongest correlation with obesity (r = 0.10), followed by employment status (r = 0.06), while vegetable intake is negatively correlated (r = -0.03). Sex and fruit intake exhibit near-zero correlations, aligning with chi-square results, reinforcing their weaker influence.

2.4.2 Potential Influencing factors

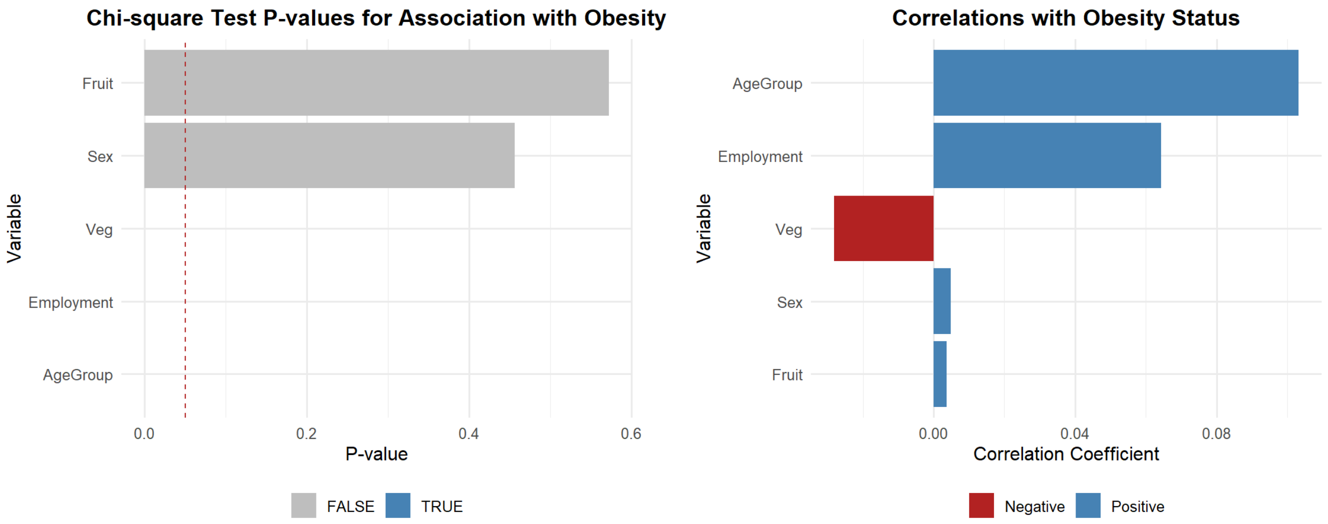

Figure 5: Potential Factor Tests.

Based on the Figure 5, the following factors appear to have the strongest association with obesity status:

```
Significant factors (chi-square): AgeGroup, Employment, Veg
```

```
Top correlating factors: AgeGroup, Employment
```

Age emerges as the most critical factor in obesity risk, followed by employment status and vegetable intake, while sex and fruit intake show weaker associations. The interaction between age and employment suggests socioeconomic factors play a key role in obesity prevalence.

3 Formal Data Analysis

3.1 Variable Selection & Regression Modelling

In this analysis, the primary objective is to understand the relationship between various socio-economic and lifestyle factors and the likelihood of an individual being obese. We will employ a logistic regression model where the response variable is Obese (Yes/No), and the predictors include Year, AgeGroup, Sex, Employment, Veg, and Fruit. We start by fitting the full logistic regression model containing all explanatory variables, which can be written as:

$$y_i = \beta_0 + \beta_{\text{Year}} \cdot \text{Year} + \beta_{\text{AgeGroup}} \cdot \text{AgeGroup} + \beta_{\text{Sex}} \cdot \text{Sex} + \beta_{\text{Employment}} \cdot \text{Employment} + \beta_{\text{Fruit}} \cdot \text{Fruit} + \beta_{\text{Veg}} \cdot \text{Veg}$$

- y_i denotes the log-odds of obesity for individual i,
- β_0 is the intercept, representing the baseline log-odds of obesity,
- β_{Year}, β_{AgeGroup}, β_{Sex}, $\beta_{\text{Employment}}$,$\beta_{\text{Fruit}}$ and β_{Veg} indicates the change in log-odds of obesity for each unit change in the respective variable.

Table 1: AIC comparison between full and stepwise logistic regression models with retained variables.

Model	AIC	Retained_Variables
Full Model	30080.98	All
Stepwise Model	30078.14	AgeGroup, Sex, Employment, Veg

Stepwise regression with backward selection will be used to determine whether the full model can be reduced based on the Akaike information criterion (AIC). Hence, the model which results in the lowest AIC will result in the final model fitted to the data. Based on Table 1, we can conclude the stepwise Logistic Regression model, with an AIC of 30078.14 is a better fit than the full model, with AIC of 30080.98. The Stepwise function has retained variables of AgeGroup, Sex, Employment and Veg, so the only removed variables are fruit and year.

Table 2: Stepwise logistic regression model coefficients by variable.

Variable Category	Regression Coefficient (β)	Standard Error	Z-Value	P-Value
Age Group	0.719	0.082	8.820	1.952e-36
Employment Status	−0.072	0.150	−0.392	1.243e-04
Intercept	−1.372	0.151	−9.116	7.829e-20
Sex	−0.044	0.029	−1.532	1.256e-01
Veg	−0.199	0.034	−5.909	3.434e-09

Table 2 presents the regression coefficients, standard errors, z-values, and p-values for each variable category in the stepwise logistic regression model, assessing the statistical significance and impact of different variable on the odds of obesity. We obtain the Stepwise logistic regression model looks like this:

$$y_i = -1.372 + 0.719 \cdot \text{AgeGroup} - 0.044 \cdot \text{Sex} - 0.072 \cdot \text{Employment} - 0.199 \cdot \text{Veg}$$

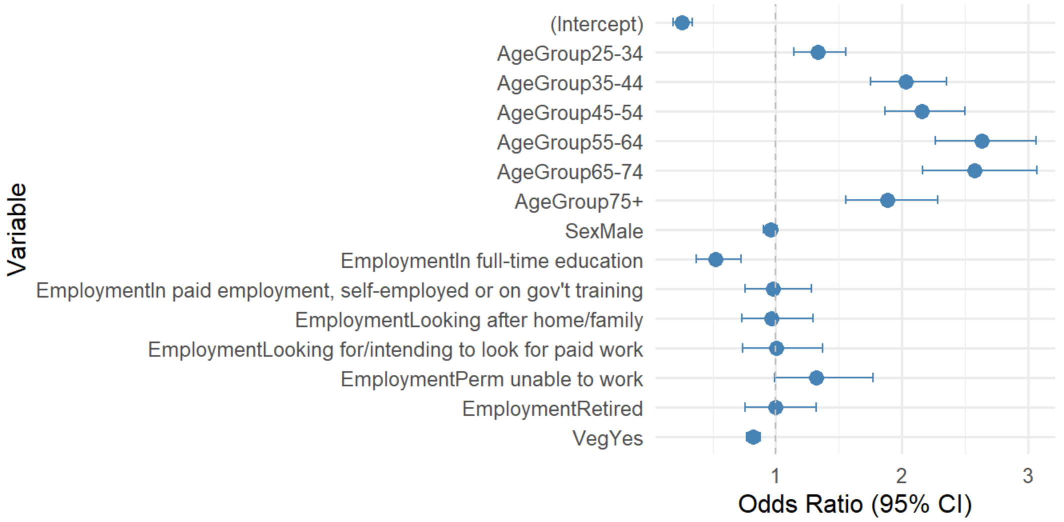

Figure 6: Forest plot of Odds Ratios from the Stepwise Logistic Regression model with 95% confidence intervals.

Figure 6 illustrates the impact of various factors on obesity prevalence. The results show that the older the age, the higher the risk of obesity, especially in the 55-74 age group, where the risk is highest. Unemployed or inactive individuals have a higher likelihood of obesity, while full-time students have the lowest obesity risk. Additionally, vegetarians have a lower risk of obesity, suggesting that dietary habits play an important role in obesity development. Gender does not have a significant effect on obesity, with males having an odds ratio close to 1. Overall, age, employment status, and dietary habits are the main factors influencing obesity, while gender has a weaker effect.

3.2 Model Diagnostics & Evaluation

3.2.1 Check Collinearity (VIF Check)

Table 3 presents the Variance Inflation Factors (VIF) to assess multicollinearity among the predictor variables. The VIF values for all variables are below 5, indicating no severe multicollinearity issues. However, AgeGroup (VIF = 3.68) and Employment (VIF = 3.82) show relatively higher values compared to other predictors, suggesting a moderate correlation with other variables. The tolerance values further confirm that multicollinearity is not a significant concern in this model.

Table 3: Variance Inflation Factor (VIF) for stepwise logistic regression Model Predictors.

Term	VIF	VIF_CI_low	VIF_CI_high	SE_factor	Tolerance	Tolerance_CI_low	Tolerance_CI_high
AgeGroup	3.68	3.602947	3.759847	1.918398	0.2717207	0.2659683	0.2775506
Sex	1.05	1.038757	1.066248	1.025023	0.9517725	0.9378685	0.9626888
Employment	3.82	3.740094	3.903928	1.954692	0.2617238	0.2561522	0.2673730
Veg	1.02	1.007090	1.035363	1.007886	0.9844127	0.9658448	0.9929601

3.2.2 Hosmer-Lemeshow Test

Table 4: Hosmer-Lemeshow Test for stepwise logistic regression Model Fit.

Test Statistic	Value	Degrees of Freedom	P-value
Chi-squared	12.4680	8.0000	0.1320

Table 4 assesses the goodness-of-fit for logistic regression models. The chi-squared statistic is 12.468 with 8 degrees of freedom, and the p-value (0.1320) is much greater than 0.05. This indicates insufficient evidence to reject the null hypothesis, suggesting that the model provides a reasonable fit to the data.

3.2.3 ROC Curve & AUC

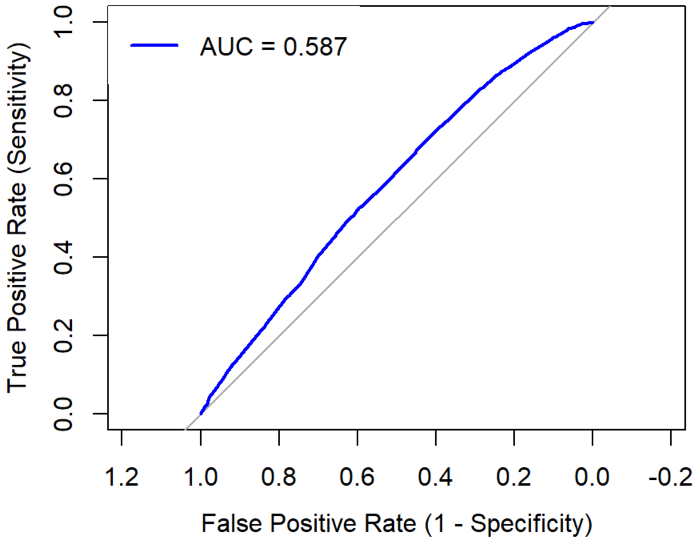

Figure 7: ROC Curve for stepwise logistic regression Model with AUC

Figure 7 evaluates the discriminative ability of the stepwise logistic regression model. The area under the curve (AUC) is 0.587, which is only slightly better than random chance (AUC = 0.5). This suggests that the model has weak predictive performance, indicating that further refinement, such as feature selection or alternative modeling approaches, may be necessary to improve classification accuracy.

3.2.4 Confusion Matrix & Accuracy

Table 5 evaluates the classification performance of the stepwise logistic regression model. The overall accuracy is 67.33%, indicating a moderate predictive ability. However, the model exhibits a high false negative rate, misclassifying 6,262 obese individuals as not obese. This suggests that while the model performs relatively well in identifying non-obese individuals, its ability to correctly classify obesity cases is limited. The imbalance in misclassification may impact its practical utility in applications requiring accurate obesity detection.

Table 5: Confusion Matrix and Accuracy for stepwise logistic regression Model.

Predicted Class	Actual Class	Count
Not Obese	Not Obese	15,777
Obese	Not Obese	1,978
Not Obese	Obese	6,262
Obese	Obese	1,207
Model Accuracy: 0.6733		

4 Conclusions

This study examined the prevalence of obesity in Scotland from 2008 to 2012, analyzing its association with socio-economic and lifestyle factors. The exploratory analysis revealed that obesity rates fluctuated over time but remained consistently high, with variations across demographic and behavioral groups. The stepwise logistic regression model identified age, employment status, and vegetable intake as significant predictors of obesity, while gender had a weaker influence. Model diagnostics confirmed a reasonable fit, but the predictive performance was limited, with a high false negative rate and an AUC of 0.587. These findings highlight the complexity of obesity determinants and suggest that targeted public health interventions should focus on high-risk groups, particularly older adults and those with socio-economic disadvantages, while promoting healthier dietary habits.

The collaborative code and data files are available in the GitHub repository: https://github.com/FSUniAccount/Group11.