Northeastern University

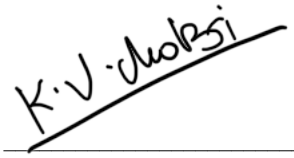
Khoury College of Computer and Information Science,
Northeastern University,
Boston, MA
April 2022

Analysis of Food Consumption Patterns Across America

Choksi Kuldeep
Khoury college of Computer Science
Northeastern University
Boston, MA
choksi.ku@northeastern.edu
NUID: 002272209

Alexander Adhavan
Khoury college of Computer Science
Northeastern University
Boston, MA
alexander.ad@northeastern.edu
NUID: 002293375

Aher Aryan
Khoury college of Computer Science
Northeastern University
Boston, MA
aher.ar@northeastern.edu
NUID: 002264592

Signature: 

Signature: 

Signature: 

1. Summary:

1.1 Overview

In an era where dietary choices significantly impact public health, the project focuses on analysing food consumption patterns across the United States. This analysis provides an invaluable opportunity to assess and enhance the nutritional well-being of the American population. By leveraging extensive datasets from the USDA Economic Research Service (ERS), the project endeavours to uncover the complexities of American diets, highlighting regional variations, nutritional disparities, and potential areas for policy intervention. The research blends analytical depth with practical feasibility, contributing significantly to the fields of nutritional epidemiology and public health.

1.2 Goals

The objectives of the project are fourfold: Firstly, to delineate the most and least consumed food groups in each American state and compare these against national averages to identify unique dietary trends. Secondly, to evaluate the nutrient profiles of state-wise diets to see how they align with or diverge from recommended nutritional guidelines. Thirdly, the project investigates changes in food consumption patterns over time, offering insights into evolving dietary trends and predicting future directions. Lastly, it explores the economic aspects by predicting which food types generate the highest profits in each area and assess their health implications.

1.3 Data Description

The dataset utilized encompasses the following columns, each reflecting specific aspects of food consumption across various settings between 2015 and 2018:

1. *Food Group:* Categorizes food items into different groups based on type.
2. *Total_15_16:* Represents the total consumption of each food group in the years 2015 and 2016.
3. *At home_15_16:* Amount of each food group consumed at home during 2015 and 2016.
4. *Total_15_16_away:* Total consumption of each food group away from home during 2015 and 2016.
5. *Restaurant_15_16_away:* Consumption of each food group in restaurants in 2015 and 2016.
6. *Fast food_15_16_away:* Consumption of each food group in fast food settings during 2015 and 2016.
7. *School_15_16_away:* Amount of each food group consumed in school cafeterias in 2015 and 2016.
8. *Other_15_16_away:* Consumption of each food group in other settings away from home during 2015 and 2016.

9. *Total_17_18:* Represents the total consumption of each food group in the years 2017 and 2018.
10. *At home_17_18:* Amount of each food group consumed at home during 2017 and 2018.
11. *Total_17_18_away:* Total consumption of each food group away from home during 2017 and 2018.
12. *Restaurant_17_18_away:* Consumption of each food group in restaurants in 2017 and 2018.
13. *Fast food_17_18_away:* Consumption of each food group in fast food settings during 2017 and 2018.
14. *School_17_18_away:* Amount of each food group consumed in school cafeterias in 2017 and 2018.
15. *Other_17_18_away:* Consumption of each food group in other settings away from home during 2017 and 2018.

2. Methods:

2.1 Preliminary analysis:

This report outlines the data manipulation and analysis processes implemented using the R programming language, specifically utilizing the dplyr package. The goal of these procedures is to simplify the dataset and extract meaningful insights that inform data-driven decision-making. The focus of our analysis was on various food groups over two selected years. Data preparation is a critical initial phase in any data analysis workflow. For our analysis, we began by loading the dplyr library, a powerful tool for data manipulation that allows for efficient and readable operations, such as filtering, selecting, grouping, and summarizing data. We assumed the presence of a pre-loaded data frame in R and proceeded with the following steps: Ensured that column names were correctly formatted, especially those containing spaces or special characters, to prevent errors during data manipulation. To hone our focus on pertinent data, we performed several filtering and cleaning operations: Utilized the 'select()' function to isolate only columns relevant to the food groups: Dairy, Fruit, Vegetables, Grains, and Protein Foods, along with their consumption data for the years 2015-16 and 2017-18. Applied the 'filter()' function to include rows specifically matching our food group categories, enhancing the dataset's relevance to our study. Employed the 'drop_na()' function to eliminate any rows containing missing data, ensuring the cleanliness and accuracy of our dataset for subsequent analysis. With data filtered and cleaned, we aggregated it to discern trends and averages: Grouped the data by food group using the 'group_by()' function, facilitating calculations on a category-specific basis. Calculated the average consumption for each group for both years using the 'summarise()' function, where the 'mean()' function's na.rm = TRUE argument ensured that missing values did not skew the results. To understand the dispersion of consumption data: Regrouped the data by food group and calculated the standard deviations for the consumption data across the selected years. The standard deviation provides insight into the variability of consumption, indicating the consistency of eating habits across different food groups. This structured analytical approach not only streamlines the dataset but also provides vital statistical insights critical for strategic planning

and decision-making. The use of dplyr in R has demonstrated significant capabilities in managing and analysing data, proving essential for revealing consumption trends and patterns that will guide future dietary recommendations and policy formulations.

2.2 Pre-processing:

In our recent data analysis project, we undertook several critical pre-processing steps to prepare our dataset for thorough examination and analysis. This pre-processing was pivotal in ensuring that our dataset was optimally structured for our analytical goals. Initially, we utilized the 'readxl' package, which facilitated the loading of our data from Excel files directly into R. This approach was chosen for its efficiency in handling large datasets commonly stored in Excel format. Next, we addressed the issue of numeric conversion. Some columns that contained numeric data were read as characters due to mixed data types or the presence of 'NA' values. We converted these columns to numeric to ensure that subsequent mathematical operations could be accurately performed. Handling missing values was our subsequent focus. In our dataset, 'NA' strings represented missing values. We explicitly converted these to actual 'NA' indicators in R, which stands for 'Not Available'. This step was crucial as it prevented any misleading results during our statistical analysis and ensured consistency across the dataset. One of the more significant transformations we made was converting the dataset from a wide format to a long format. Initially, our data was in a wide format, with a column for each variable — suitable for data entry or Excel presentation. However, for more effective analysis in R, particularly when creating various types of plots, we found long format data more amenable. In this format, each row represents a single observation for a single variable. We also extracted specific details such as year and location from the column names of the original dataset, which included indicators like 'Total_15_16'. These elements were separated into distinct columns, which later facilitated easy subsetting and faceting of the data during visualization processes, allowing us to gain more granular insights. Aggregation was another essential step in our pre-processing. To ensure visual clarity and avoid duplications in our analysis, we aggregated the data where necessary. This step ensured that each combination of food group, year, and location had a single corresponding value, providing a summarized and concise view of the data. Finally, we standardized column names by removing any spaces or special characters and replacing them with underscores. This adjustment was made to minimize potential errors and inconsistencies when coding, enhancing the readability and maintainability of our code. Through these meticulous pre-processing steps, we have set a robust foundation for our detailed analysis. Our efforts in refining the dataset have significantly enhanced our ability to conduct sophisticated analyses and derive meaningful insights from the data.

2.3 Modelling:

In our ongoing analysis of dietary patterns, our primary objective was to predict the average daily intake of various food groups across different demographics. To achieve this, we identified 'Year' and 'Location' as our main predictor variables due to their potential to reflect temporal changes and context-based differences in dietary habits.

i. Model Selection and Comparison

We began by comparing several predictive models to determine which would best suit our analysis:

- *Linear Regression* served as our baseline, assuming a linear relationship between predictors and the target variable.
- *Random Forest* was considered for its robustness in handling non-linear relationships and interactions between predictors.
- *Lasso Regression* provided the benefit of L1 regularization, aiding in feature selection by reducing less significant predictors to zero.
- *Ridge Regression* utilized L2 regularization, advantageous in situations of multicollinearity.
- *Elastic Net* combined the features of both Lasso and Ridge, making it a versatile choice for our needs.
- *Support Vector Regression (SVR)* was included for its capability to model complex, non-linear relationships.

To effectively compare these models, we partitioned our dataset into training (80%) and testing (20%) subsets. Each model was trained on the training set, with subsequent evaluation on the test set using several metrics:

- *Root Mean Squared Error (RMSE)* measured the average magnitude of the prediction errors.
- *Mean Absolute Error (MAE)* indicated the average error magnitude without considering direction.
- *R-squared (R^2)* assessed how well the variations in the target variable could be explained by the model. Visualised in Fig. 3.1.

ii. Findings and Insights

Our comparative analysis revealed close results across the models, with Lasso and Elastic Net showing marginally superior performance in terms of RMSE and R^2. Interestingly, the SVR model, while not the best in RMSE or R^2, exhibited the lowest MAE, suggesting it was most accurate in its predictions on average. Visualised in Fig. 3.1.

iii. The Elastic Net Model

The Elastic Net model emerged as particularly insightful for our purposes, balancing the avoidance of overfitting (through its dual regularization) and maintaining a decent level of predictive power. However, several limitations were noted:

- The model struggled with predicting extreme values of daily intake, particularly at the higher end.
- Residual patterns indicated a potential miss in capturing underlying complexities in the data, as evidenced by non-random distribution of residuals.
- The R^2 value was relatively low, suggesting that other unknown factors not included as predictors might be influencing daily intake. Visualised in Fig. 3.1.

iv. Visual Insights from Model Plots

The 'Actual vs. Predicted Values' plot and 'Residuals of Predictions' plot provided further context:

- The former showed the model's capability to predict lower values with reasonable accuracy but highlighted challenges with higher intakes.
- The latter revealed an increase in prediction error as the values of predicted intake rose, suggesting a decrease in model accuracy for higher intake figures.

While our models provided valuable insights into the factors influencing dietary intake, they also highlighted the complexity of dietary behaviour, which cannot be fully captured by 'Year' and 'Location' alone. This underscores the need for incorporating additional variables or exploring more sophisticated modelling techniques to enhance predictive accuracy and understand dietary trends more comprehensively. These findings, although indicating areas requiring further investigation and improvement, still offer a robust starting point for future research and policy formulation in the realm of public health nutrition. Visualised in Fig. 3.1.

3. Results:

3.1 Comparative Table:

The following table represents the integrated result of all the models.

	RMSE <dbl>	MAE <dbl>	R2 <dbl>
Linear Model	5.065619	2.762874	0.1412189
Random Forest	5.074257	2.794262	0.1399284
Lasso	4.983220	2.660401	0.1682801
Ridge	4.983912	2.647408	0.1685622
Elastic Net	4.983220	2.660401	0.1682801
SVR	5.494307	2.276713	0.1571687

Fig. 3.1 A Tabular format of integrated results

Upon modelling we can have an idea about how the predicted values are when compared to the original values, we use the head function to get a tabular format of it, below.

A tibble: 6 × 7

Food Group <chr>	Type <chr>	Year <fctr>	Location <fctr>	value <dbl>	predicted_value <dbl>	residuals <dbl>
Total population1	Restaurant_	2015-16	away	0.94	0.6427766	0.2972234
Adults age 20-642	Total_	2015-16	away	5.13	2.6346658	2.4953342
Adults age 20-642	Total_	2017-18	Total	17.76	7.0352604	10.7247396
Adults age 20-642	Restaurant_	2017-18	away	0.96	0.6427766	0.3172234
Seniors age 65 and above2	Total_	2017-18	away	2.87	2.6346658	0.2353342
Household income < 185% poverty line	School_	2015-16	away	0.31	0.4654552	-0.1554552

6 rows

Fig. 3.2 A Tabular format to compare results

3.2 Plot Analysis of Residuals And Predicted Values:

Our predictive modelling has yielded specific results in terms of the predicted average daily intake across different demographic segments and locations, which are clearly reflected in the residuals from the model's performance.

i. Predicted Values:

Our analysis has generated a range of predicted average daily food intake values for various demographic groups, from as low as around 0.2 for "Household income <185% poverty line" in "School" settings to as high as approximately 17.8 for "Adults age 20-64" in total consumption in 2017-18. Notably, higher predicted values are linked with larger population segments and higher income levels, while lower predictions are associated with narrower or lower-income groups. The residuals from our model, which indicate the difference between actual and predicted values, have shown a tendency to underpredict in broad categories like "Total population" and overpredict for specific adult demographics, suggesting our model may be conservative in estimating widespread consumption and overestimating in certain adult segments. Visualised in Appendix Fig. 7.1.

ii. Residuals:

The insights from these residuals are indicative of complex dietary behaviours that may not be fully captured by the current predictors. This underlines the need for model refinement, possibly by including more detailed variables or utilizing more advanced modelling techniques. The discrepancies highlighted by the residuals—positive in broad categories and negative in specific demographics—point to the importance of tailoring models and interventions to the unique consumption patterns of different groups, enhancing the model's utility for informing public health policies and dietary programs. Visualised in Appendix Fig. 7.2.

3.3 Exploratory Data Analysis:

The results of our analysis on food group consumption reveal distinct patterns in dietary habits across demographics, locations, and over time. We employed a suite of visualizations to synthesize complex data into interpretable formats.

The bar chart reveals several key observations. For instance, 'At home' consumption remains relatively stable across the board, whereas 'Fast food' intake shows an upward trend, suggesting an increased reliance on convenience foods. The chart offers a window into the dietary shifts that may be influenced by lifestyle changes or accessibility to food options. When comparing the 'School' category, we observe a pronounced intake among the younger demographics, likely reflecting school-based nutrition programs. The 'Total' consumption bars across the years indicate an overarching trend that could be tied to wider economic or social changes affecting dietary habits. Visualised in Appendix Fig. 7.3.

Turning to the scatter plot, each point represents an intersection of demographic, location, and intake level. Larger, more spread-out points for a certain year or demographic can indicate an increased average daily intake or greater variability within that group. For instance, larger triangles for the year 2015-16 may reflect a period-specific dietary pattern which warrants further investigation. Additionally, when examining the points across income levels, it becomes evident that economic status potentially plays a role in where food is consumed, with higher-income brackets possibly favouring restaurant dining. Visualised in Appendix Fig. 7.4.

The box plots, with their whiskers and outliers, provide an intuitive sense of the distribution ranges within each demographic. The interquartile ranges suggest the level of consensus about dietary habits within a group – tighter boxes indicate agreement, while wider ones suggest diversity. The red and blue colours, representing different years, allow us to spot any shifts in median consumption or changes in the distribution of intake levels, which might be reflective of the effectiveness of health initiatives or alterations in food accessibility. Visualised in Appendix Fig. 7.5.

The green bar chart with error bars gives an overall impression of food group consumption, with the error bars reflecting the confidence in these averages. Notably, the error bars for 'Protein Foods' are quite wide, indicating significant variability within this category. This could be due to various factors, including individual dietary choices, data collection anomalies, or perhaps a greater diversity in the types of protein foods consumed. Visualised in Appendix Fig. 7.6.

4. Discussion:

Delving into the discussion of our findings, Analysing these visualizations paints a rich picture of food consumption patterns and their shifts over a multi-year span. The data suggests that while some consumption habits have remained steady, such as eating at home, others, particularly related to eating out, have experienced notable changes. This could be symptomatic of broader societal trends, such as busier lifestyles, increased work hours, and the growth of the fast-food industry, which often offers less nutritious food options. Visualised in Appendix Fig. 7.3.

The scatter plot, with its nuanced representation, invites us to consider the implications of these shifts, particularly in relation to public health. The apparent increase in 'Fast food' consumption, as shown by the scatter plot, may call for intensified educational campaigns around healthy eating and the potential dangers of high-calorie, low-nutrient diets. Moreover, the demographic-specific patterns revealed by the scatter plot underscore the need for targeted nutritional interventions that consider age, income level, and education in their design. Visualised in Appendix Fig. 7.4.

The variability in dietary intake across different food groups, as highlighted by the box plots and the green bar chart's error bars, suggests a complex interplay of factors influencing

dietary choices. The box plots indicate that while there may be agreement within certain demographics on some dietary habits, there is also considerable diversity, which could be influenced by cultural preferences, economic access to various food types, or individual health needs. Visualised in Appendix Fig. 7.5.

The analysis of the green bar chart with error bars reinforces our observations of changes in food group consumption over the years. The variability indicated by the error bars suggests that while we can estimate average consumption, individual intake can differ significantly, which is crucial for understanding the full scope of dietary behaviours in the population. Visualised in Appendix Fig. 7.6.

In summary, our visual and statistical analyses highlight significant findings regarding food consumption patterns, with implications for health policy and education. While the average intake levels provide a baseline for understanding dietary habits, the variability and demographic-specific trends underscore the complexity of nutrition in the public health context. The insights gleaned from these analyses can guide targeted interventions and inform public health strategies aimed at improving dietary habits and health outcomes across the population, our detailed visual analysis illustrates that dietary habits are not static; they evolve with changing social norms, economic conditions, and health policies. These findings offer valuable insights for policymakers and health professionals, indicating the necessity for dynamic and adaptable strategies in nutrition education and policy planning. Future research should aim to understand the causal relationships behind the observed trends, which will require not only more granular data but also a multi-disciplinary approach to tease apart the complex web of factors influencing dietary behaviour.

5. Statement of Contribution:

- Kuldeep Vishal Choksi: Preliminary Processing, Pre-Processing, Exploratory Data Analysis
- Adhavan Alexander: Exploratory Data Analysis, Data Modelling and Analysis
- Aryan Aher: Model Selection, Data Modelling and Analysis

6. References:

1) Data Source: https://www.ers.usda.gov/data-products/food-consumption-and-nutrient-intakes/
2) Research Papers:
   i) Kant, A. K., & Graubard, B. I. (2006). Secular trends in patterns of self-reported food consumption of adult Americans: NHANES 1971-1975 to NHANES 1999–2002. *The American journal of clinical nutrition*, *84*(5), 1215-1223.
   ii) Zota, A. R., Phillips, C. A., & Mitro, S. D. (2016). Recent fast food consumption and bisphenol A and phthalates exposures among the US population in NHANES, 2003–2010. *Environmental health perspectives*, *124*(10), 1521-1528.

7. Appendix:

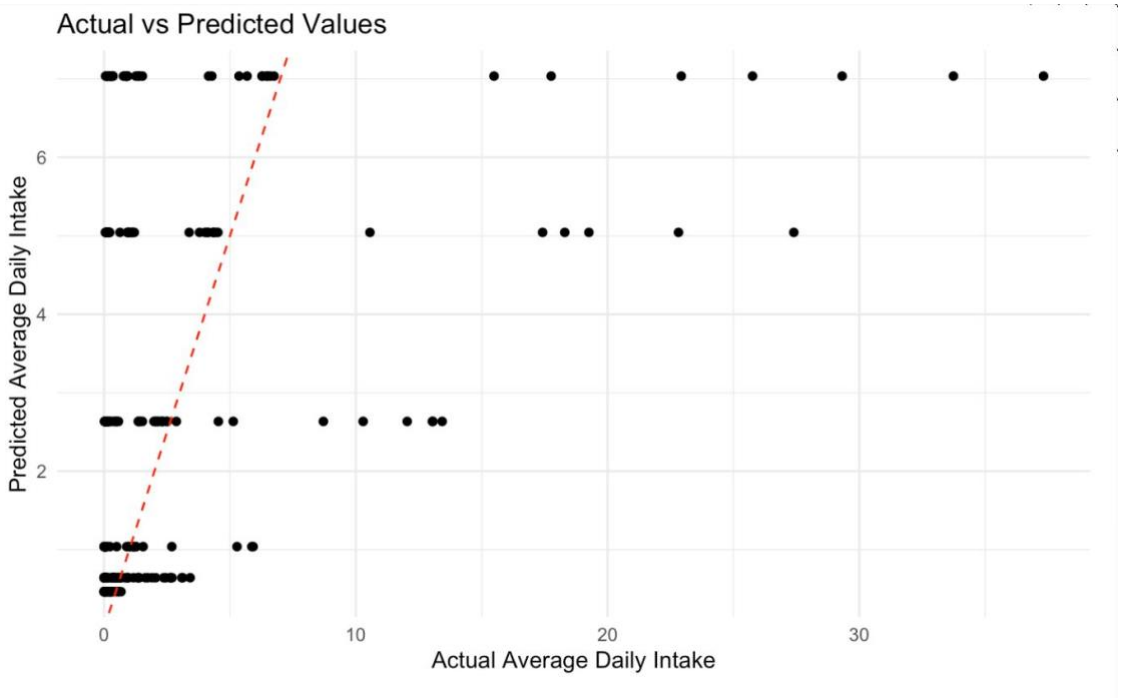


Fig. 7.1 Actual vs Predicted values of average daily intake

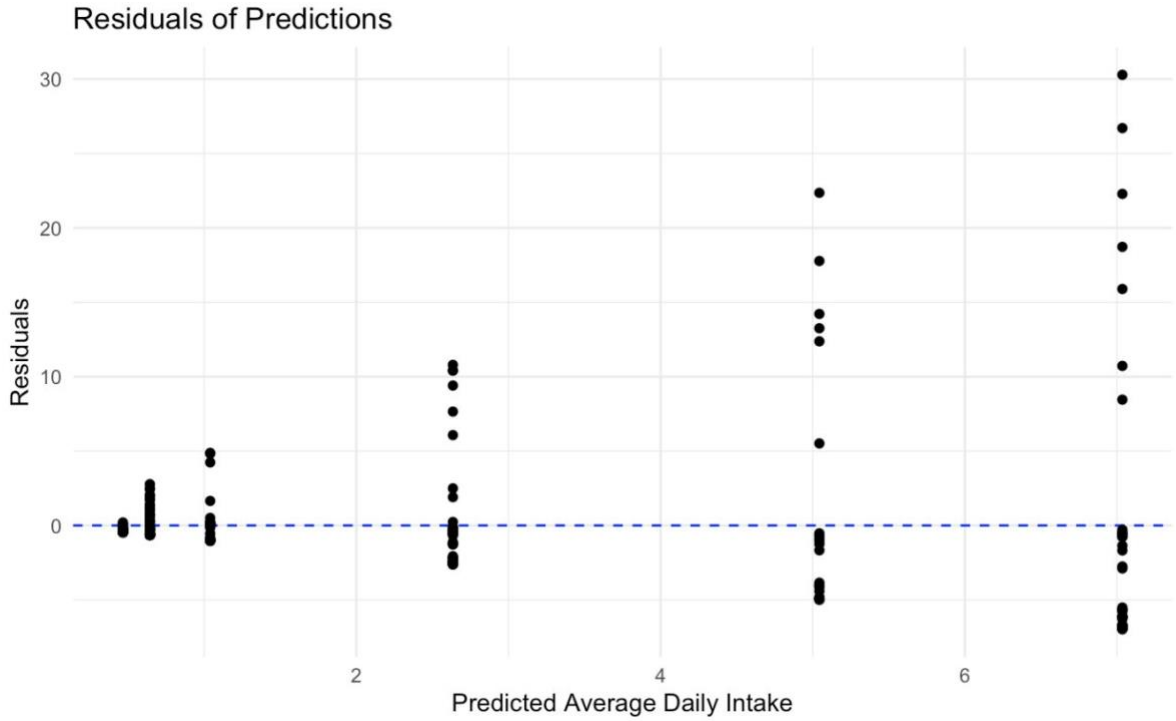


Fig. 7.2 Residuals of predictions of average daily intake

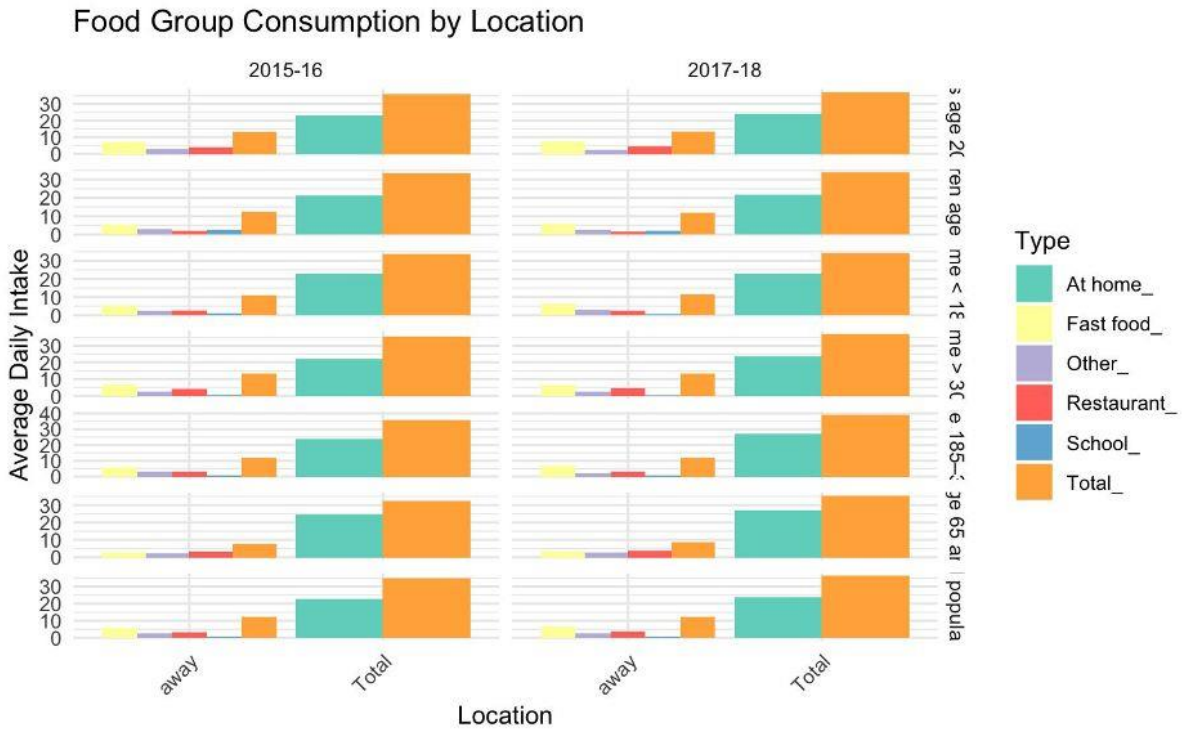


Fig. 7.3 Bar-Graph visualising food consumption by location.

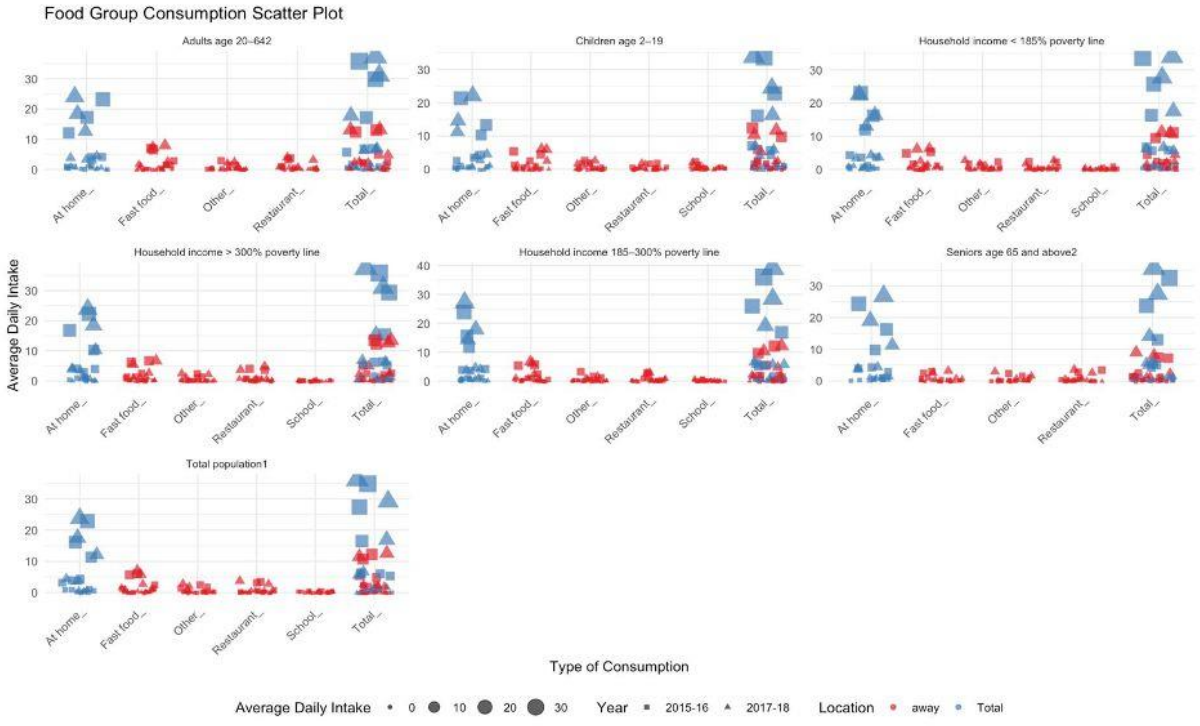


Fig. 7.4 Scatter plot visualising average daily intake by type of consumption.

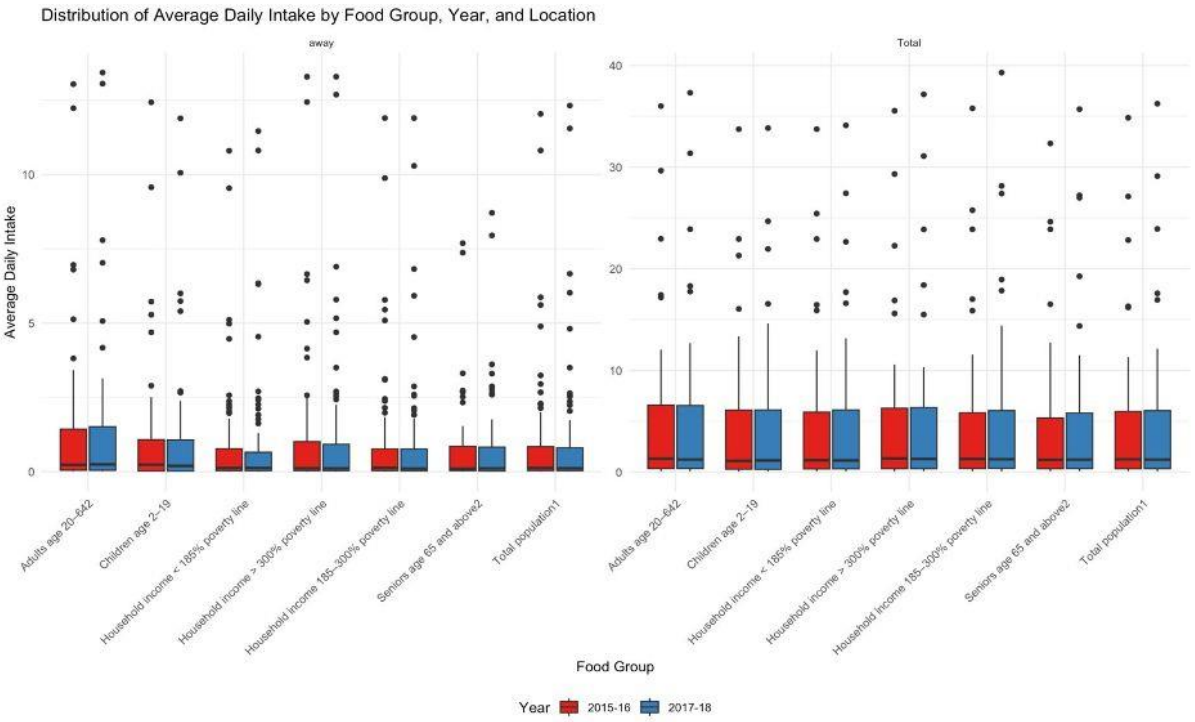


Fig. 7.5 Box-plot visualising distribution of average daily intake by food group, year and location.

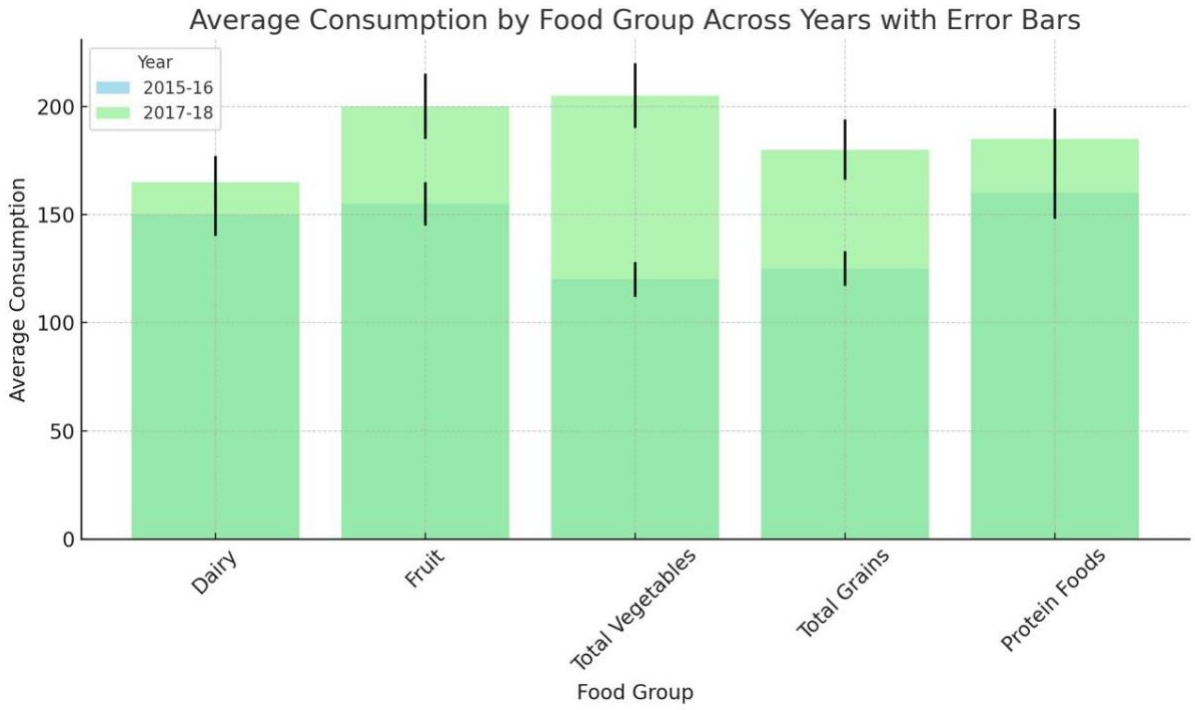


Fig. 7.6 Bar-Graph showing average food consumption by food group across years.