Analysis of Dietary Data - Exploring the Impact of Age and Gender on Calorie Consumption in a Diverse Population

Mahitha Vontimitta, Veronica Angelina, Lavanya Ranganatham

Indiana University Purdue University Indianapolis, IN 46202, USA

Abstract:

This study examined how age and gender relate to calorie intake in a diverse population using data from the USDA FoodData Central. Various statistical methods like data visualization, correlation analysis, non-parametric tests, linear and logistic regression were employed. The key findings showed significant differences in calorie consumption across age groups, with younger people tending to have higher calorie needs. Gender also played a role, with males generally consuming more calories than females. Additionally, there were potential interactions between age and gender influencing dietary patterns in a complex way across different demographic groups. The analysis highlighted the intricate relationships between these demographic factors and calorie intake.

Purpose:

The purpose of this project was to understand how age and gender affect calorie consumption in different population groups. The goal was to use this information to develop better nutrition guidelines and interventions tailored to specific age and gender demographics, in order to promote healthier eating habits and prevent diet-related diseases.

Introduction:

This project aims to see how age and being male or female affect how much food people eat. It wants to find out how these things work together to change how much people eat. This understanding is important for making plans to help people eat better and stay healthy. The main question is about how age and gender affect how many calories people consume and how they interact to make these differences. Using information from the USDA FoodData Central, which includes details about people and what they eat, the study looks at things like age, gender, and what nutrients people get from their food. The study used different methods to analyze the data, like looking at graphs and doing math tests to see if there were any patterns. The study found that there were big differences in how many calories people ate based on their age, and there were also differences between men and women. It also showed that age and gender together had a big impact on how much people ate. Younger people usually need more calories, and men usually eat more than women. In short, this research shows how age and gender affect what people eat, and it suggests that plans to help people eat healthier should think about these things. It also suggests that future studies could look at other things that might affect what people eat, and follow how eating habits change over time to see how it affects health.Data and Methods

Research Question:

To what extent do age and gender influence variations in calorie consumption among individuals in a diverse population, and how do potential interactions between these factors contribute to these variations?

Data:

The main data used came from the USDA FoodData Central and included two sets: one with details about people (DEMO.CSV) and another with information about their diet (DSQTOT.CSV). The information covered various factors such as age, gender, how many calories they consumed, and the amounts of protein, carbohydrates, fat, fiber, and sugar in their diet.

Methodologies:

The project employed various methodologies to analyze the data comprehensively:

Data Collection and Analysis:

Data were gathered from the USDA FoodData Central, comprising two datasets: DEMO.CSV for demographic information and DSQTOT.CSV for dietary data.

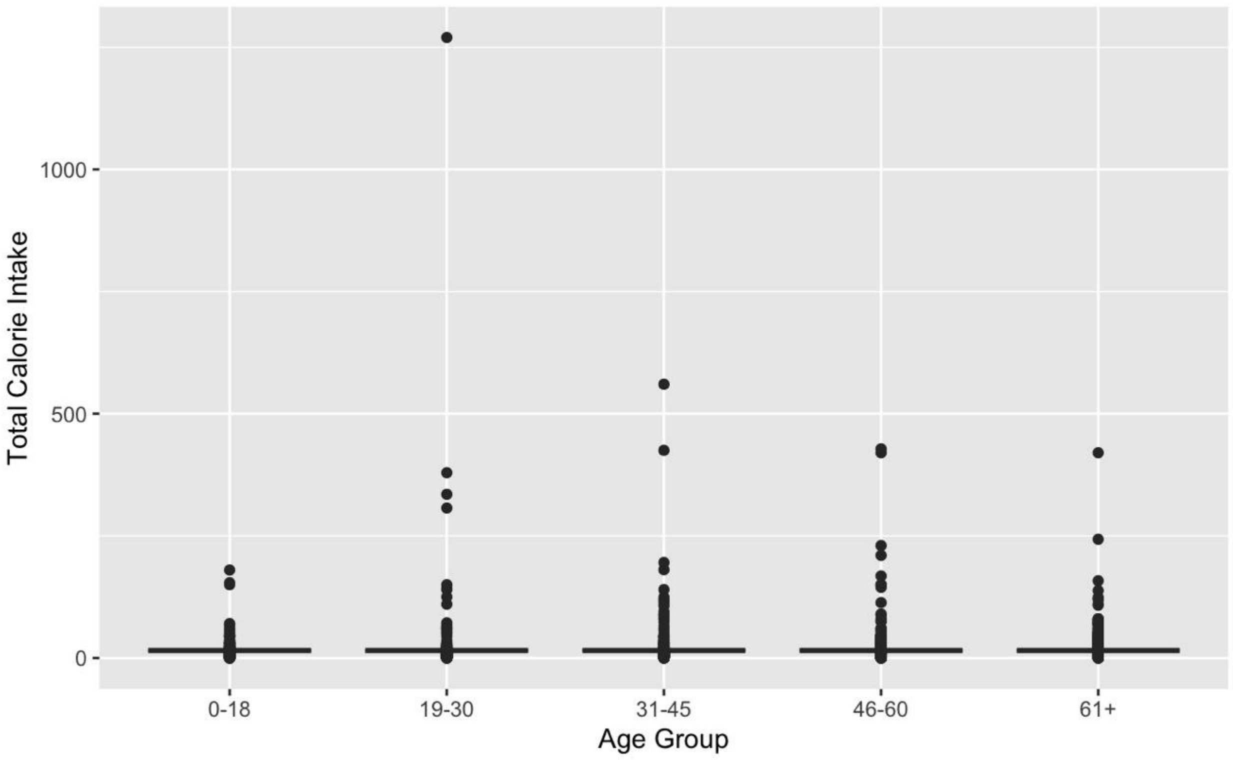
Variables like age, gender, and nutritional intake were considered to understand dietary habits and nutrient composition.

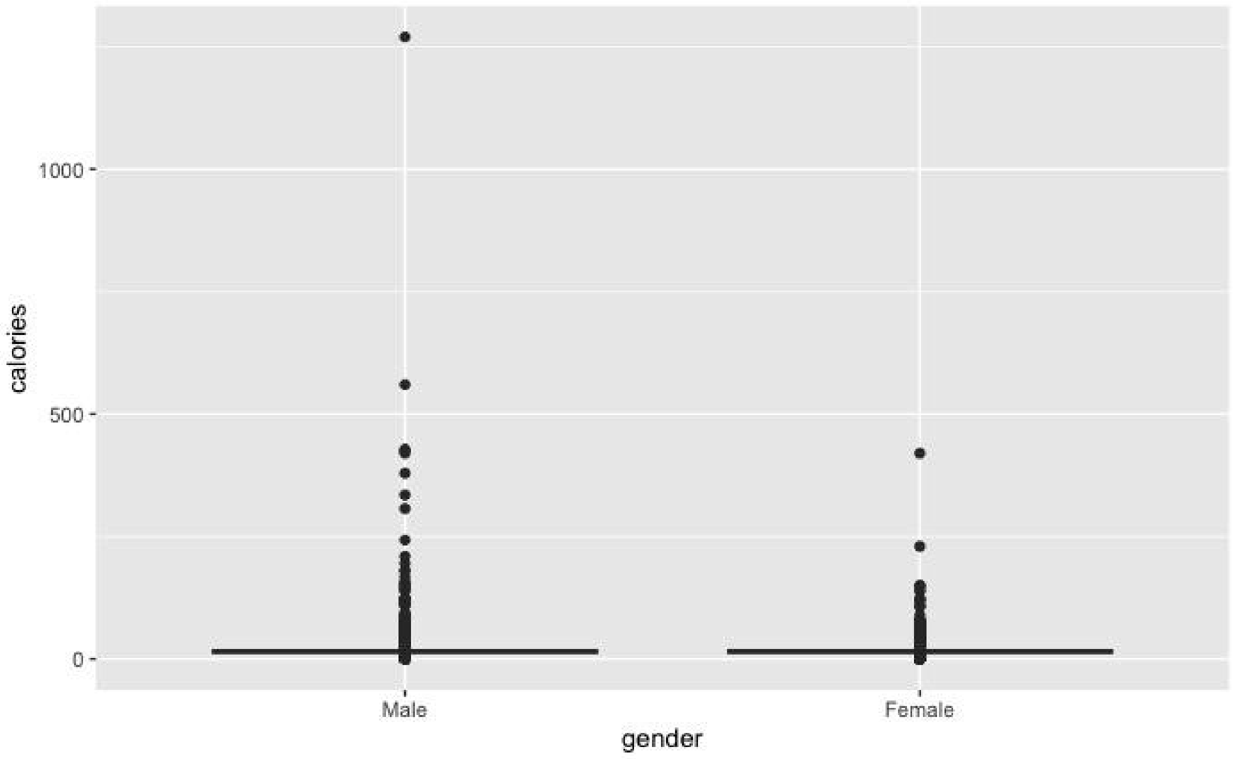
Data Cleaning and Merging:

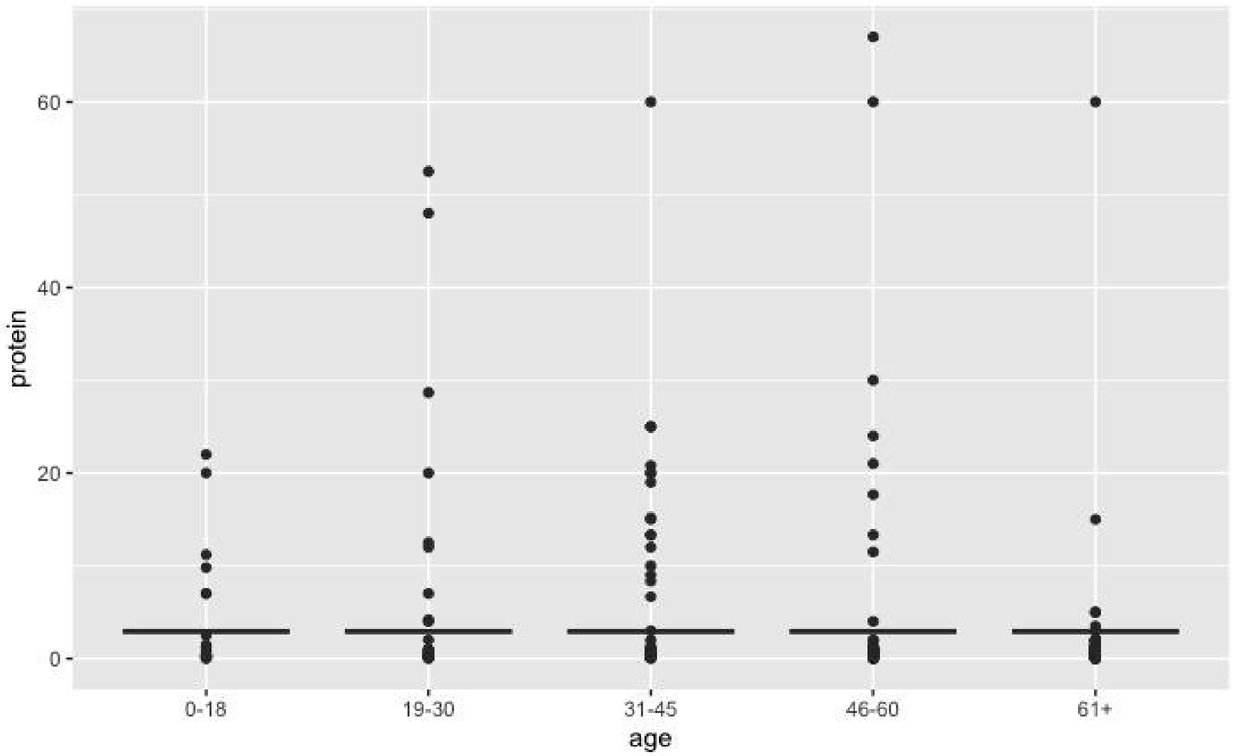
The datasets were merged based on a unique identifier, likely to combine demographic and dietary information for each participant.

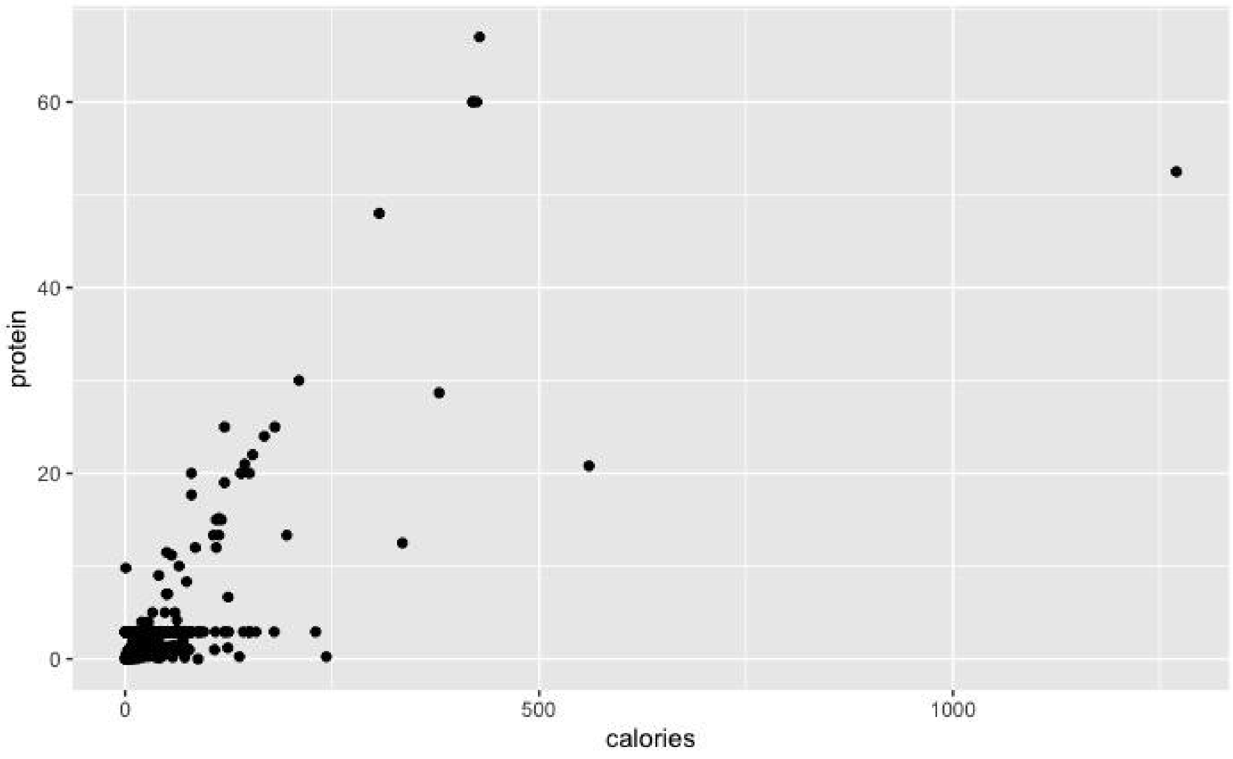
Data Visualization:

Box plots, density plots, and histograms were utilized to visualize the distribution of calorie intake across different demographic groups, helping identify patterns and outliers.


Total Calorie Intake
1000
500
0
0-18
19-30
31-45
46-60
61+
Age Group

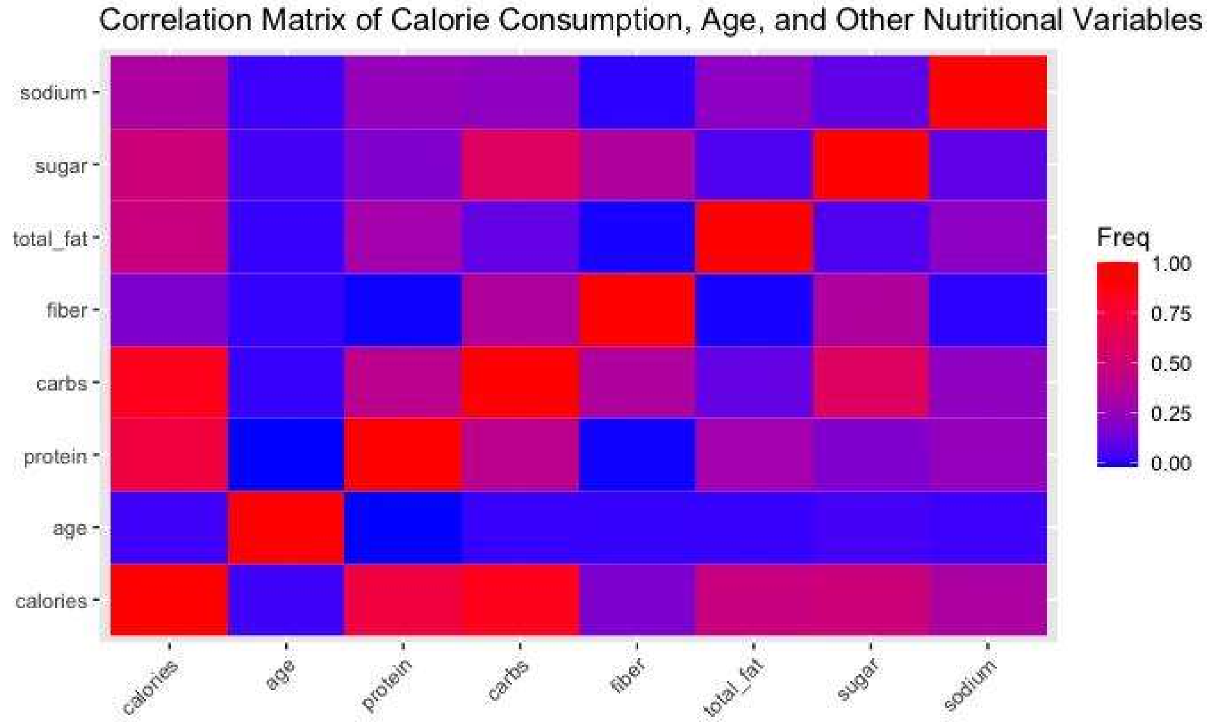

calories
1000
500
0
Male
Female
gender


protein
60
40
20
0
0-18
19-30
31-45
46-60
61+
age


protein
60
40
20
0
0
500
1000
calories

Correlation Analysis:

A correlation matrix was constructed by Spearsman method to examine relationships between variables, providing insights into associations between calorie intake, age, and nutritional factors.



Statistical Analysis:

1. Kruskal-Wallis Test:

This non-parametric test assessed differences in calorie consumption among age groups, considering the non-normal distribution of data.

2. Dunn's Test:

Post-hoc analysis compared differences between age groups to identify significant disparities in calorie intake.

3. Linear Regression:

This model evaluated the impact of age and gender on calorie consumption, examining potential interactions between these factors.

4. Logistic Regression:

This model predicted the likelihood of high calorie intake based on demographic factors, such as age and gender.

Overall, these methodologies facilitated a thorough exploration of how age, gender, and calorie consumption are interconnected, aiding in the identification of patterns and insights within the dataset.

Key Findings:

The Kruskal-Wallis test detected notable variations in calorie consumption across different age brackets, suggesting significant differences in dietary habits among age groups. Dunn's test further pinpointed specific pairwise distinctions among these age groups, offering detailed insights into which groups exhibited significantly different calorie intakes. Additionally, the linear regression analysis underscored the substantial impacts of both age and gender on calorie intake, revealing interactions between these factors that influence dietary patterns. Furthermore, the logistic regression analysis provided an understanding of the likelihood of individuals having a high calorie intake based on their demographic characteristics. In summary, younger age groups generally displayed higher calorie requirements, while males tended to consume more calories than females, highlighting the nuanced dynamics of calorie consumption within the studied population.

Applications and Future Directions:

Health organizations can create individualized dietary guidelines suited to various age groups and genders. Tailoring nutritional programs to meet the specific dietary requirements of diverse population segments is essential. Further research could investigate other factors affecting eating habits, such as socioeconomic status and cultural influences. Long-term studies can monitor shifts in dietary patterns over time and their implications for health outcomes.

Conclusions:

Age and gender significantly impact calorie intake in the studied population, shaping dietary habits. The relationship among age, gender, and calorie consumption is complex, with potential interactions influencing eating patterns. Tailored interventions and public health policies addressing age and gender disparities are crucial for promoting healthier eating behaviors and preventing diet-related illnesses. Customized strategies considering these differences are essential, as individuals have diverse nutritional needs based on their life stages and genders. Health authorities can develop personalized dietary guidelines and programs specific to different age and gender groups to enhance health outcomes. Additionally, exploring additional factors affecting dietary behaviors and conducting longitudinal studies can provide valuable insights for refining strategies to promote healthy eating and prevent diet-related diseases.

REFERENCES:

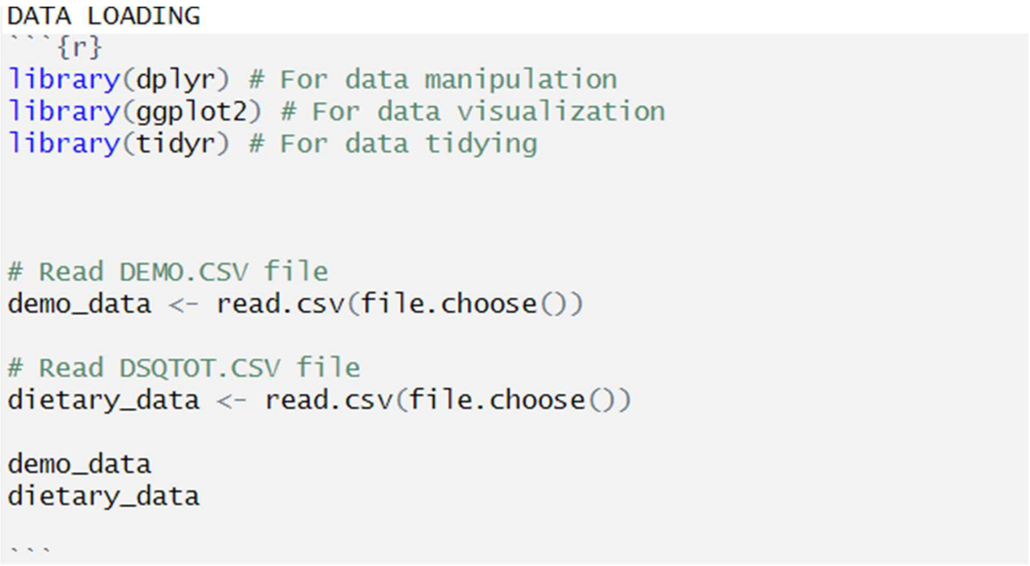
FoodData Central. (n.d.). https://fdc.nal.usda.gov/

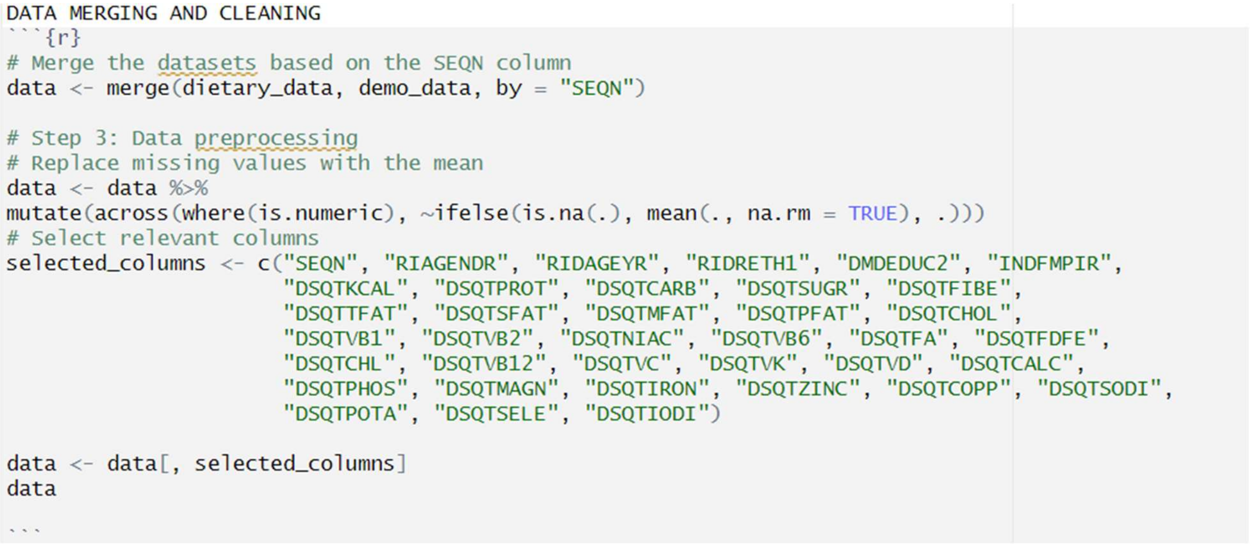
Mikkelsen, B. E., Beck, A. M., & Lassen, A. D. (2006). Do recommendations for institutional food service result in better food service? A study of compliance in Danish hospitals and nursing homes from 1995 to 2002–2003. *European Journal of Clinical Nutrition*, *61*(1), 129–134. https://doi.org/10.1038/sj.ejcn.1602488

Vaughan, L. E. (2006). Embedding Education into Diabetes Practice. *Journal of Human Nutrition and Dietetics*, *19*(3), 240–241. https://doi.org/10.1111/j.1365-277x.2006.00688.x

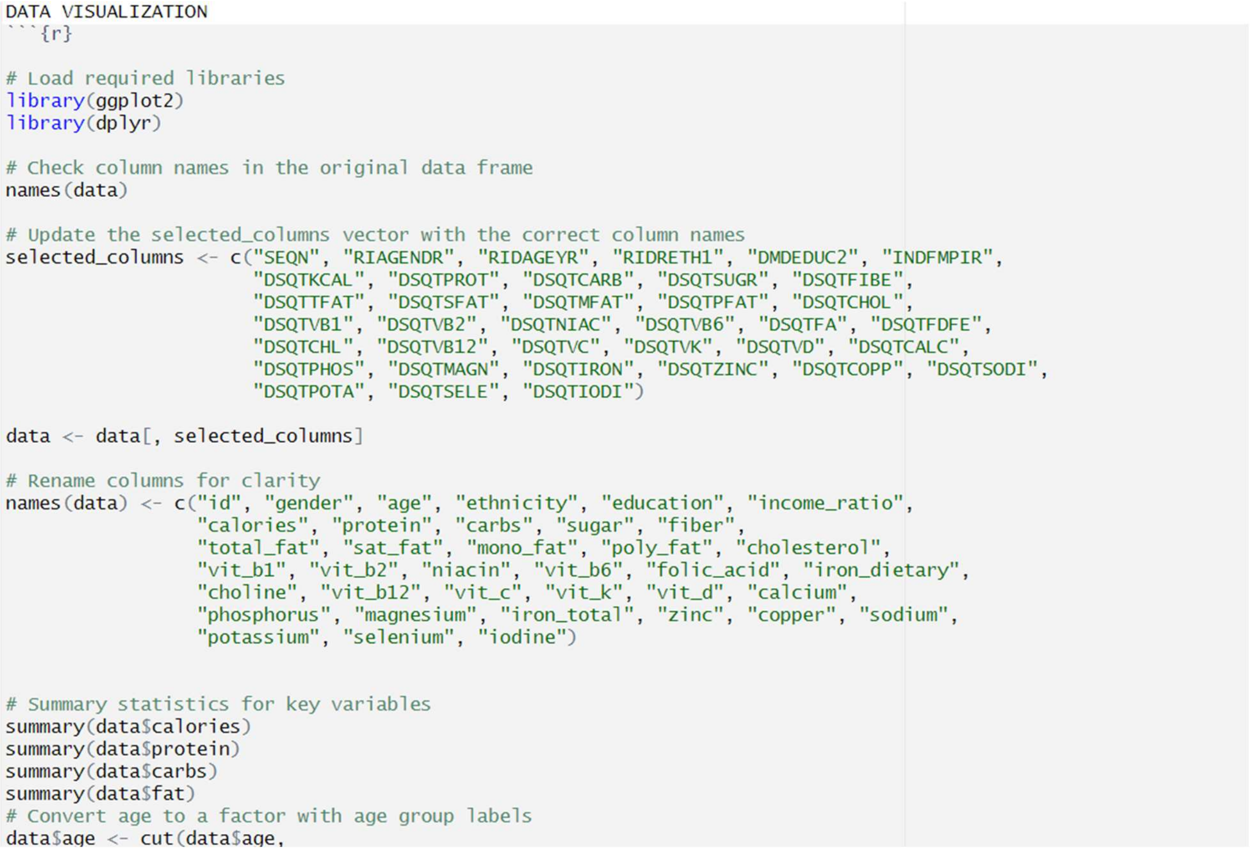
A. Appendix:

A1. Data Loading

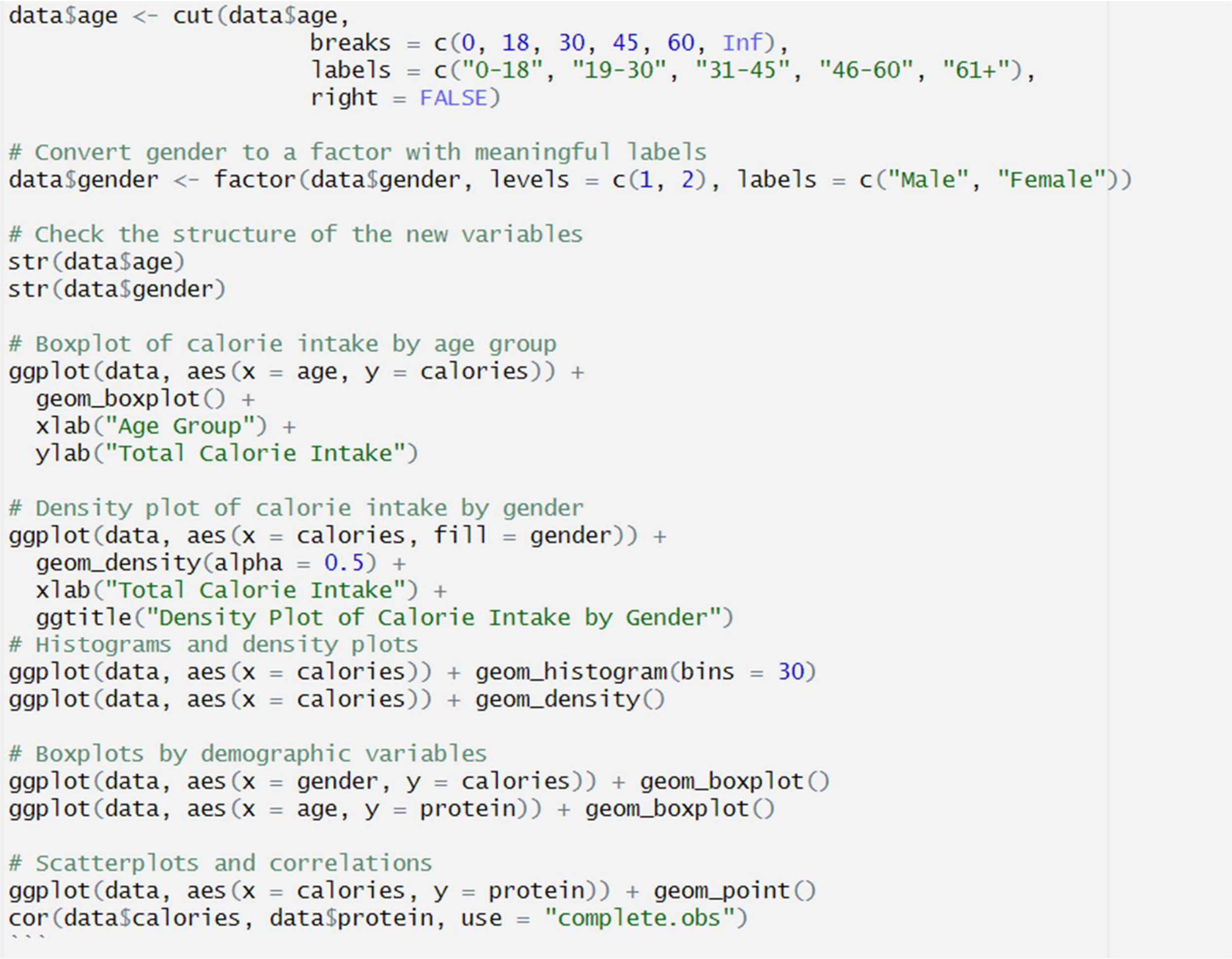


~~~
DATA LOADING
```{r}
library(dplyr) # For data manipulation
library(ggplot2) # For data visualization
library(tidyr) # For data tidying



# Read DEMO.CSV file
demo_data <- read.csv(file.choose())

# Read DSQTOT.CSV file
dietary_data <- read.csv(file.choose())

demo_data
dietary_data

~~~
```

A2. Data Merging and Cleaning



```
DATA MERGING AND CLEANING
```{r}
# Merge the datasets based on the SEQN column
data <- merge(dietary_data, demo_data, by = "SEQN")

# Step 3: Data preprocessing
# Replace missing values with the mean
data <- data %>%
mutate(across(where(is.numeric), ~ifelse(is.na(.), mean(., na.rm = TRUE), .)))
# Select relevant columns
selected_columns <- c("SEQN", "RIAGENDR", "RIDAGEYR", "RIDRETH1", "DMDEDUC2", "INDFMPIR",
                      "DSQTKCAL", "DSQTPROT", "DSQTCARB", "DSQTSUGR", "DSQTFIBE",
                      "DSQTTFAT", "DSQTSFAT", "DSQTMFAT", "DSQTPFAT", "DSQTCHOL",
                      "DSQTVB1", "DSQTVB2", "DSQTNIAC", "DSQTVB6", "DSQTFA", "DSQTFDFE",
                      "DSQTCHL", "DSQTVB12", "DSQTVC", "DSQTVK", "DSQTVD", "DSQTCALC",
                      "DSQTPHOS", "DSQTMAGN", "DSQTIRON", "DSQTZINC", "DSQTCOPP", "DSQTSODI",
                      "DSQTPOTA", "DSQTSELE", "DSQTIODI")

data <- data[, selected_columns]
data

```
```
```

A3. Data Visualization

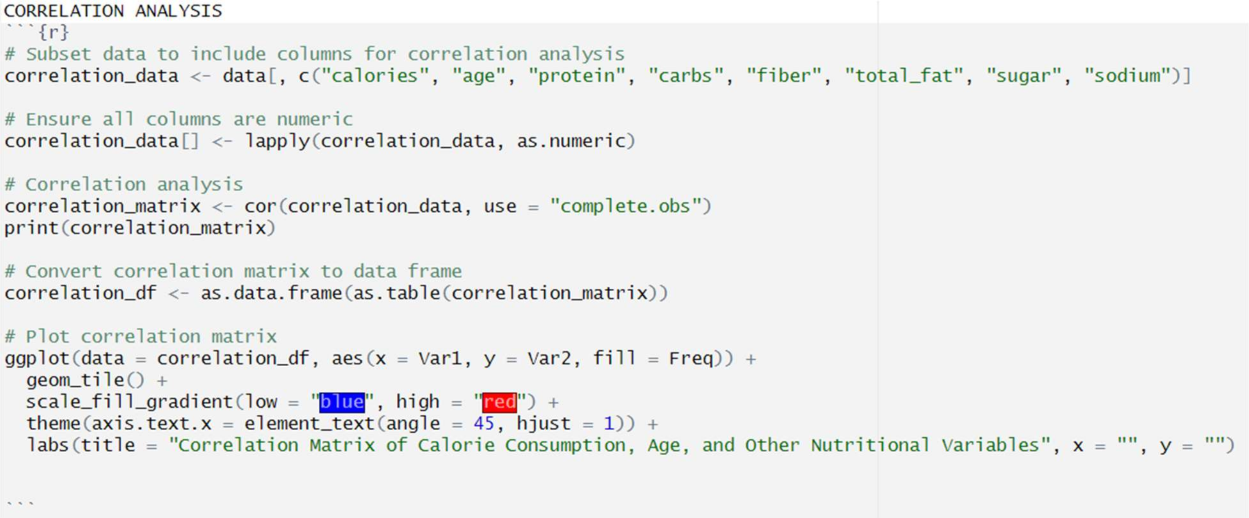


~~~
DATA VISUALIZATION
```{r}

# Load required libraries
library(ggplot2)
library(dplyr)

# Check column names in the original data frame
names(data)

# Update the selected_columns vector with the correct column names
selected_columns <- c("SEQN", "RIAGENDR", "RIDAGEYR", "RIDRETH1", "DMDEDUC2", "INDFMPIR",
                      "DSQTKCAL", "DSQTPROT", "DSQTCARB", "DSQTSUGR", "DSQTFIBE",
                      "DSQTTFAT", "DSQTSFAT", "DSQTMFAT", "DSQTPFAT", "DSQTCHOL",
                      "DSQTVB1", "DSQTVB2", "DSQTNIAC", "DSQTVB6", "DSQTFA", "DSQTFDFE",
                      "DSQTCHL", "DSQTVB12", "DSQTVC", "DSQTVK", "DSQTVD", "DSQTCALC",
                      "DSQTPHOS", "DSQTMAGN", "DSQTIRON", "DSQTZINC", "DSQTCOPP", "DSQTSODI",
                      "DSQTPOTA", "DSQTSELE", "DSQTIODI")

data <- data[, selected_columns]

# Rename columns for clarity
names(data) <- c("id", "gender", "age", "ethnicity", "education", "income_ratio",
                 "calories", "protein", "carbs", "sugar", "fiber",
                 "total_fat", "sat_fat", "mono_fat", "poly_fat", "cholesterol",
                 "vit_b1", "vit_b2", "niacin", "vit_b6", "folic_acid", "iron_dietary",
                 "choline", "vit_b12", "vit_c", "vit_k", "vit_d", "calcium",
                 "phosphorus", "magnesium", "iron_total", "zinc", "copper", "sodium",
                 "potassium", "selenium", "iodine")


# Summary statistics for key variables
summary(data$calories)
summary(data$protein)
summary(data$carbs)
summary(data$fat)
# Convert age to a factor with age group labels
data$age <- cut(data$age,
~~~



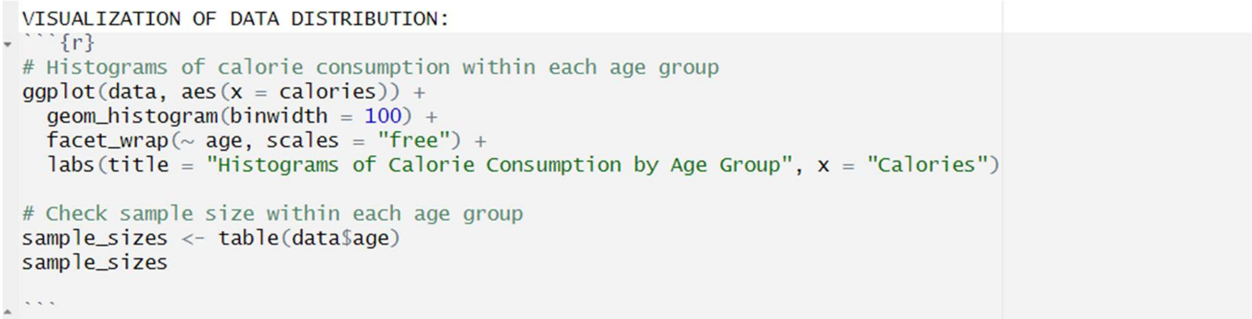
~~~
data$age <- cut(data$age,
                      breaks = c(0, 18, 30, 45, 60, Inf),
                      labels = c("0-18", "19-30", "31-45", "46-60", "61+"),
                      right = FALSE)

# Convert gender to a factor with meaningful labels
data$gender <- factor(data$gender, levels = c(1, 2), labels = c("Male", "Female"))

# Check the structure of the new variables
str(data$age)
str(data$gender)

# Boxplot of calorie intake by age group
ggplot(data, aes(x = age, y = calories)) +
  geom_boxplot() +
  xlab("Age Group") +
  ylab("Total Calorie Intake")

# Density plot of calorie intake by gender
ggplot(data, aes(x = calories, fill = gender)) +
  geom_density(alpha = 0.5) +
  xlab("Total Calorie Intake") +
  ggtitle("Density Plot of Calorie Intake by Gender")
# Histograms and density plots
ggplot(data, aes(x = calories)) + geom_histogram(bins = 30)
ggplot(data, aes(x = calories)) + geom_density()

# Boxplots by demographic variables
ggplot(data, aes(x = gender, y = calories)) + geom_boxplot()
ggplot(data, aes(x = age, y = protein)) + geom_boxplot()

# Scatterplots and correlations
ggplot(data, aes(x = calories, y = protein)) + geom_point()
cor(data$calories, data$protein, use = "complete.obs")
```
~~~

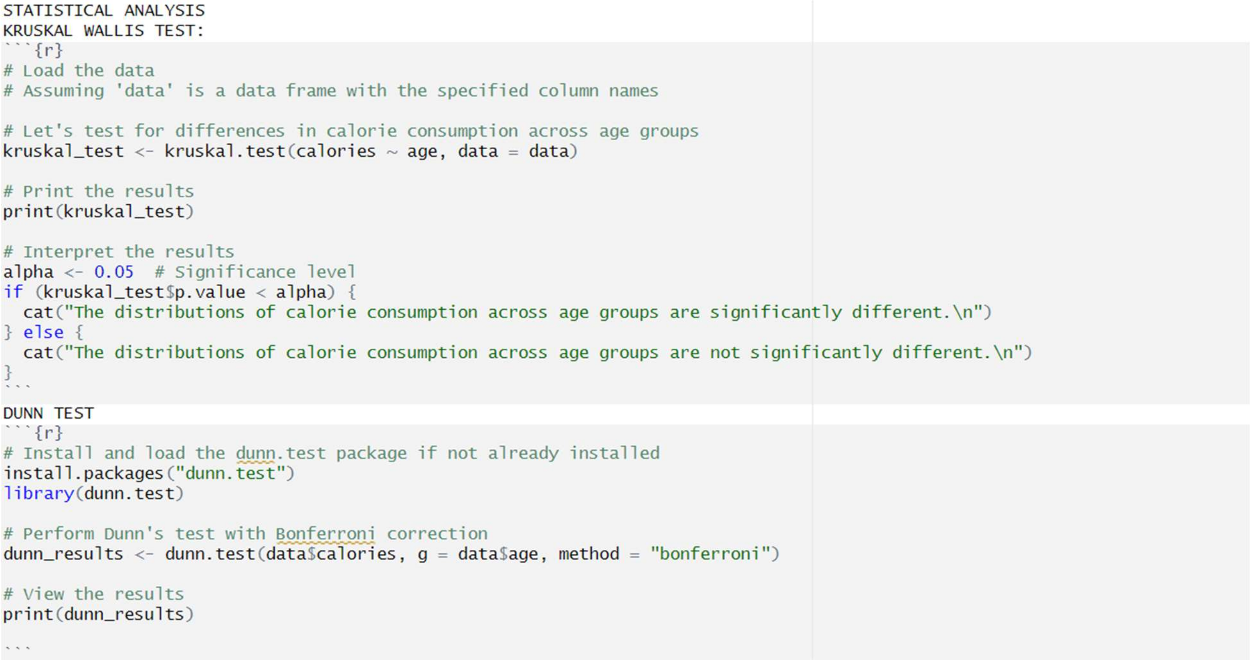
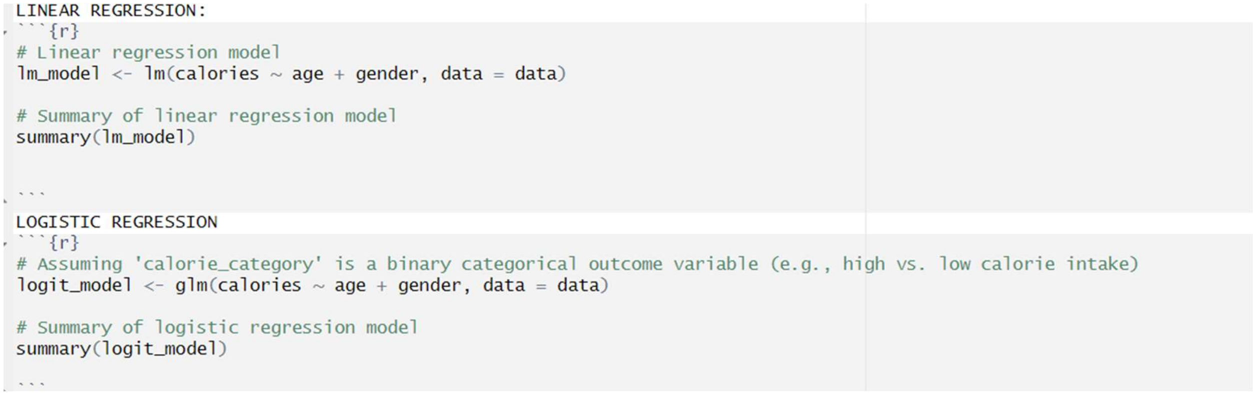
A4. Correlation Analysis



~~~
CORRELATION ANALYSIS
```{r}
# Subset data to include columns for correlation analysis
correlation_data <- data[, c("calories", "age", "protein", "carbs", "fiber", "total_fat", "sugar", "sodium")]

# Ensure all columns are numeric
correlation_data[] <- lapply(correlation_data, as.numeric)

# Correlation analysis
correlation_matrix <- cor(correlation_data, use = "complete.obs")
print(correlation_matrix)

# Convert correlation matrix to data frame
correlation_df <- as.data.frame(as.table(correlation_matrix))

# Plot correlation matrix
ggplot(data = correlation_df, aes(x = Var1, y = Var2, fill = Freq)) +
  geom_tile() +
  scale_fill_gradient(low = "blue", high = "red") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Correlation Matrix of Calorie Consumption, Age, and Other Nutritional Variables", x = "", y = "")


```
~~~

A5. Visualization of Data Distribution



~~~
VISUALIZATION OF DATA DISTRIBUTION:
```{r}
# Histograms of calorie consumption within each age group
ggplot(data, aes(x = calories)) +
  geom_histogram(binwidth = 100) +
  facet_wrap(~ age, scales = "free") +
  labs(title = "Histograms of Calorie Consumption by Age Group", x = "Calories")

# Check sample size within each age group
sample_sizes <- table(data$age)
sample_sizes

~~~
```

## A6. Stastical Analysis (Kruskal test, Dunn Test, Linear Regression & Logistic Regression)



```
STATISTICAL ANALYSIS
KRUSKAL WALLIS TEST:
```{r}
# Load the data
# Assuming 'data' is a data frame with the specified column names

# Let's test for differences in calorie consumption across age groups
kruskal_test <- kruskal.test(calories ~ age, data = data)

# Print the results
print(kruskal_test)

# Interpret the results
alpha <- 0.05  # Significance level
if (kruskal_test$p.value < alpha) {
  cat("The distributions of calorie consumption across age groups are significantly different.\n")
} else {
  cat("The distributions of calorie consumption across age groups are not significantly different.\n")
}
```
DUNN TEST
```{r}
# Install and load the dunn.test package if not already installed
install.packages("dunn.test")
library(dunn.test)

# Perform Dunn's test with Bonferroni correction
dunn_results <- dunn.test(data$calories, g = data$age, method = "bonferroni")

# View the results
print(dunn_results)

```
```



```
LINEAR REGRESSION:
```{r}
# Linear regression model
lm_model <- lm(calories ~ age + gender, data = data)

# Summary of linear regression model
summary(lm_model)


```
LOGISTIC REGRESSION
```{r}
# Assuming 'calorie_category' is a binary categorical outcome variable (e.g., high vs. low calorie intake)
logit_model <- glm(calories ~ age + gender, data = data)

# Summary of logistic regression model
summary(logit_model)

```
```
```