Analysis of dietary data

Exploring Calorie Consumption and Nutritional Factors

Group 10:

Veronica Angelina Itta

Lavanya Ranganatham

Mahitha Vontimitta

INTRODUCTION

- Calorie intake is essential for maintaining health, with age and gender being significant factors influencing dietary habits.

- However, the extent to which these factors affect calorie consumption in diverse populations is not fully understood.

- The purpose of the study is to explore how age and gender influence calorie intake and whether there's an interaction between them. By examining these relationships in a demographically diverse population, we hope to gain insights that can inform targeted interventions and public health strategies to promote healthier eating habits.

RESEARCH QUESTION:

What is the extent to which age and gender influence the variability in calorie consumption among individuals in a demographically diverse population, and how do potential interactions between age and gender contribute to these differences?

AIM:

The aim of this project is to analyze nutritional and demographic data to uncover relationships, patterns and influences on dietary habits.

HYPOTHESIS

Null Hypothesis (H0):

There is no significant association between age and calorie intake among individuals in the population. Similarly, there is no significant difference in calorie intake between genders.

Alternative Hypothesis (H1):

There is a significant association between age and calorie intake among individuals in the population. Additionally, there is a significant difference in calorie intake between genders.

DATA COLLECTION AND ANALYSIS

- **Link to the Dataset:** https://fdc.nal.usda.gov/
- The primary dataset for this project has been sourced from the USDA FoodData Central (FDC), a rich repository of nutritional information for a wide array of food items.
- The analysis utilizes two datasets: DEMO.CSV, which contains demographic information, and DSQTOT.CSV, which contains dietary data.
- **Variables**:

The variables of this study are Age, Gender, Calorie Intake, Protein, Carbs, Fat, Fiber, and Sugar. These factors assess dietary habits, energy intake, and nutrient composition, offering insights into health patterns and informing strategies for healthier lifestyles and disease prevention.

METHODOLOGY

- DATA LOADING: The datasets were loaded using the read.csv() function in R.

```r
{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
```

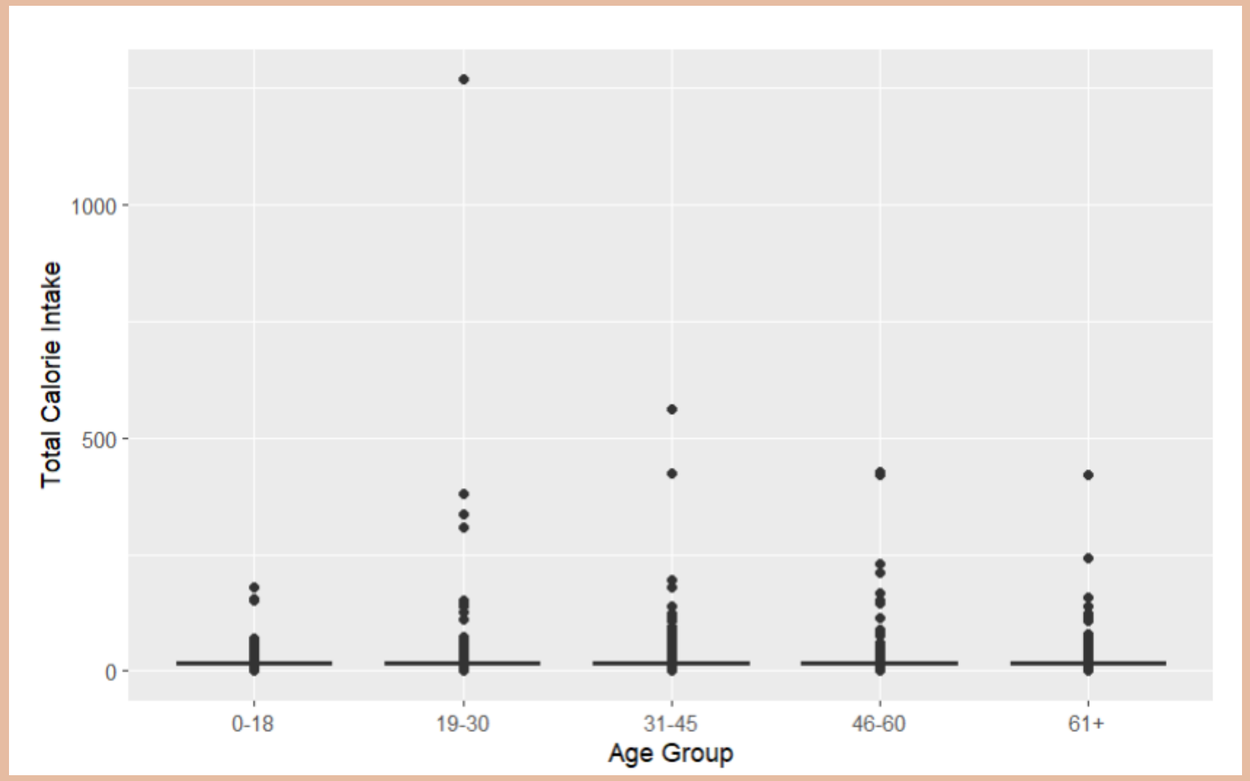
- DATA CLEANING AND MERGING: The datasets were merged based on the "SEQN" column, which serves as a unique identifier for each participant.

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

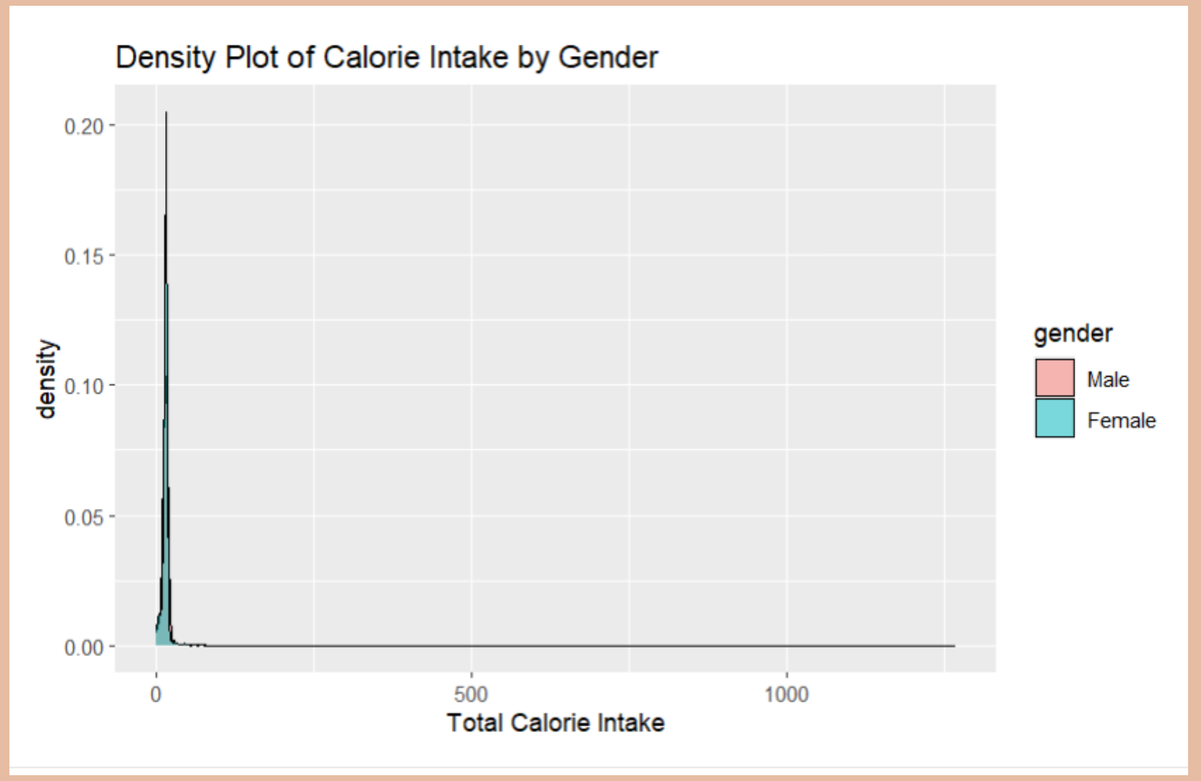
```r
# Select relevant columns
selected_columns <- c("SEQN", "RIAGENDR", "RIDAGEYR", "RIDRETH1", "DMDEDUC2", "INDFMPIR",
                      "DSQTKCAL", "DSQTPROT", "DSQTCARB", "DSQTSUGR", "DSQTFIBE",
                      "DSQTTFAT", "DSQTSFAT", "DSQTMFAT", "DSQTPFAT", "DSQTCHOL",
                      "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
```

DATA VISUALIZATION

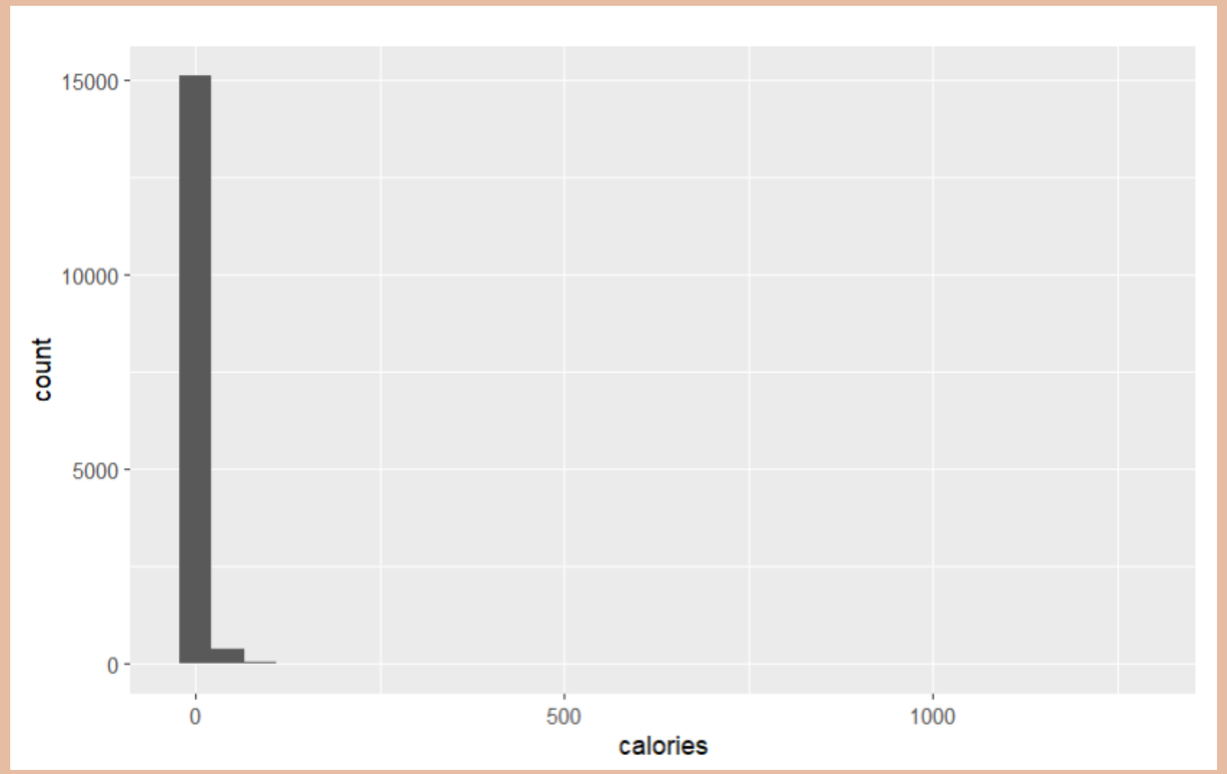


Box Plot of Total Calorie Intake by Age Group

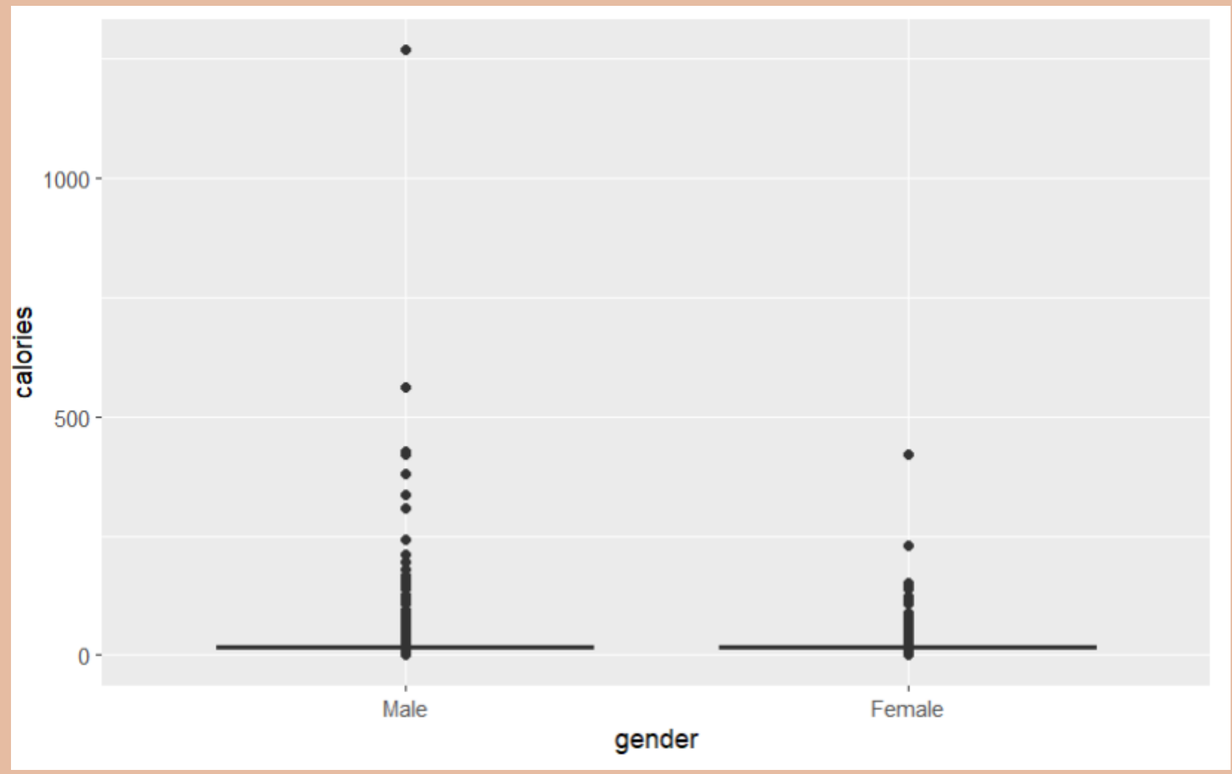


Density Plot of Calorie Intake by Gender

DATA VISUALIZATION

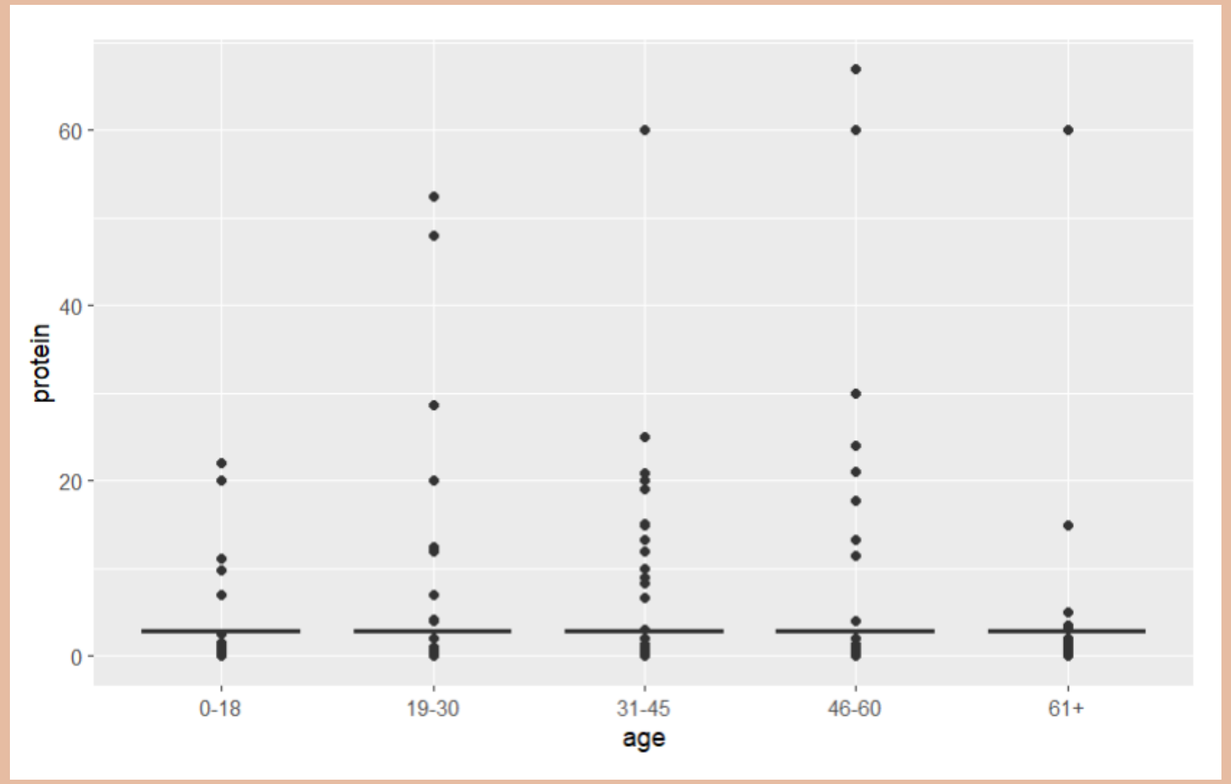


Histogram of Calorie Intaket

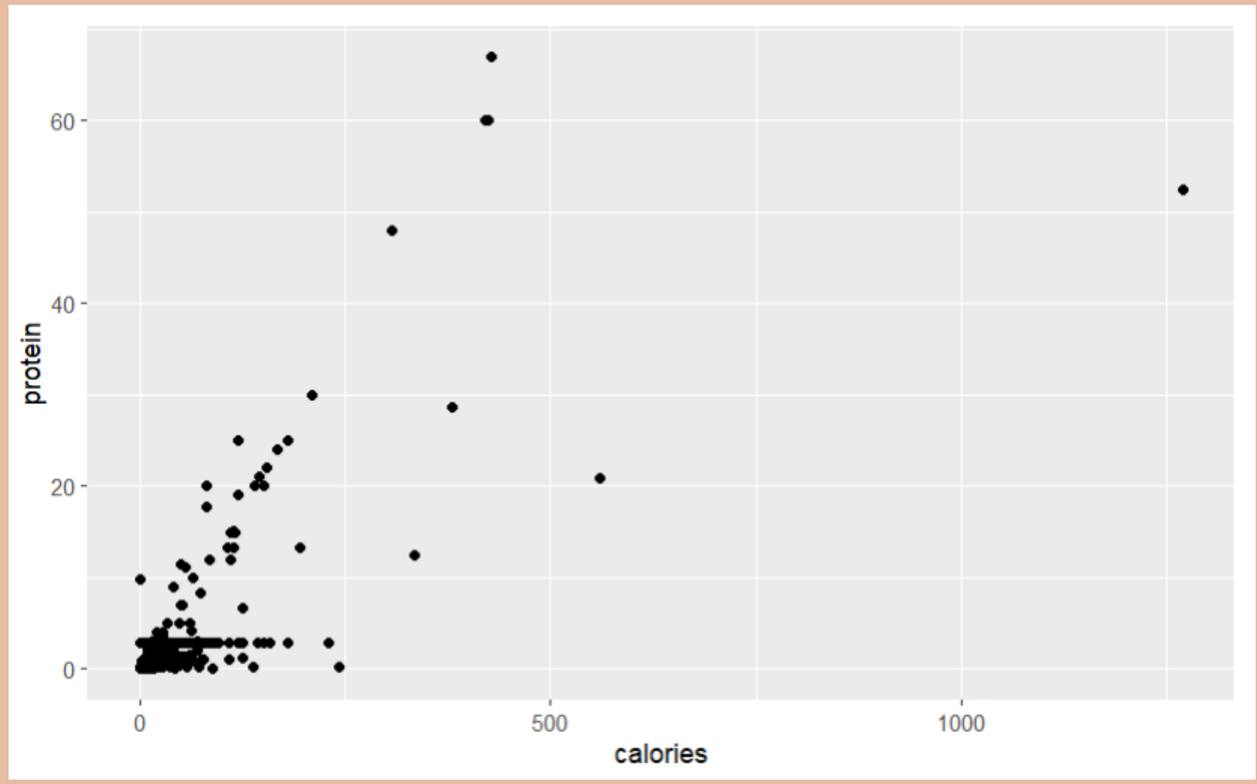


Add a little bit of body text

DATA VISUALIZATION

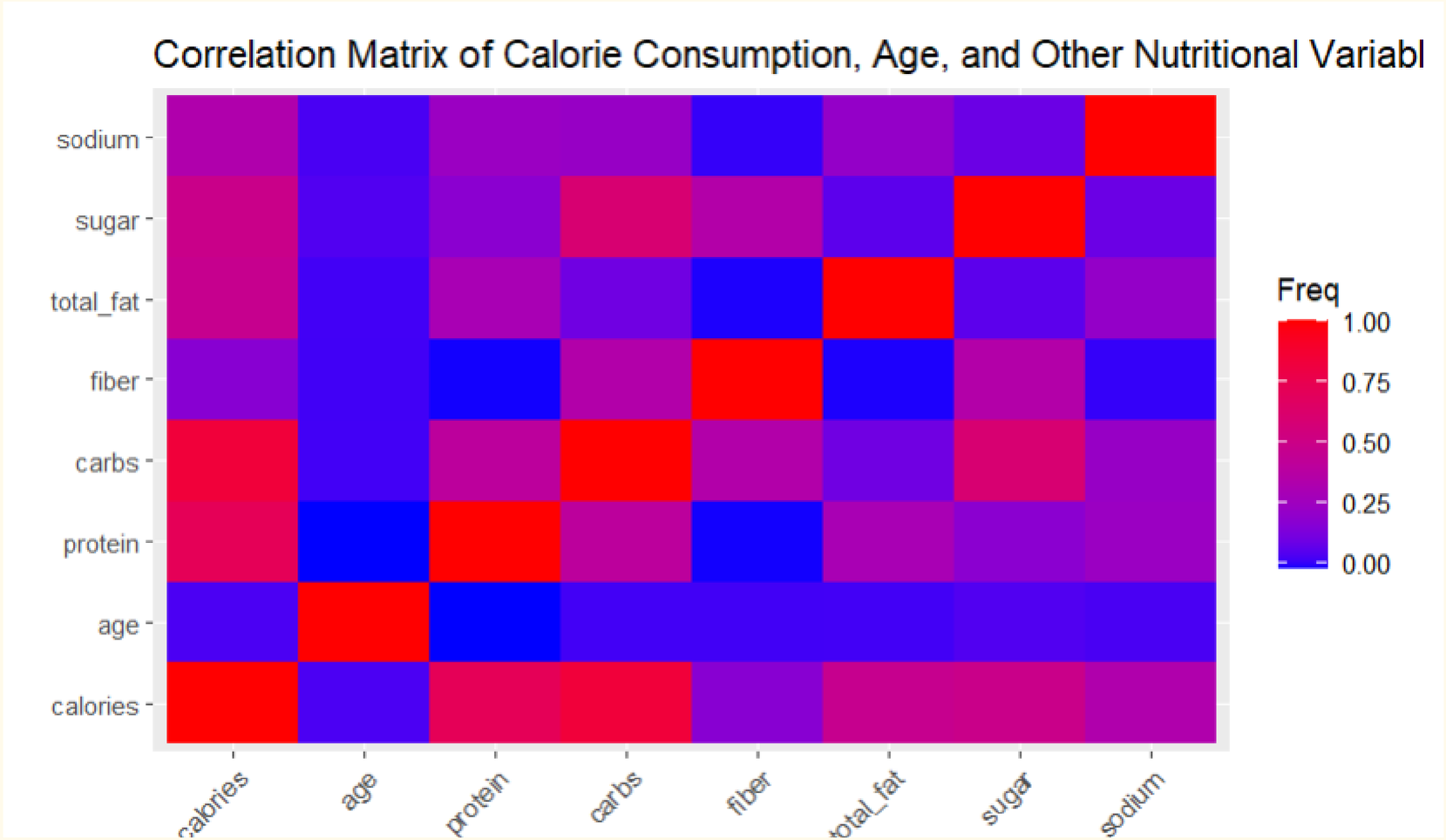


Add a little bit of body text



Add a little bit of body text

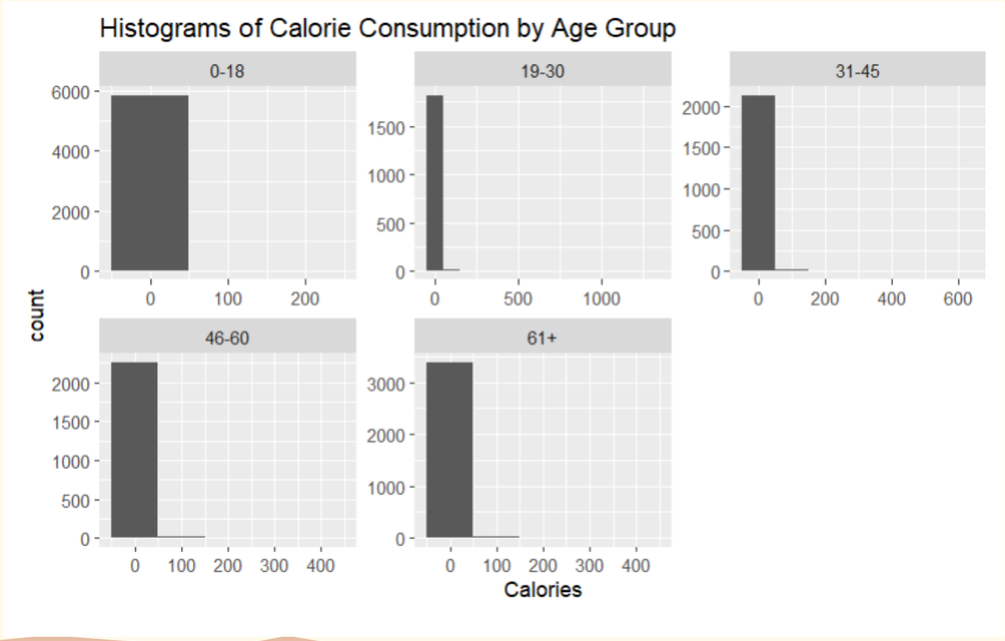
CORRELATION ANALYSIS

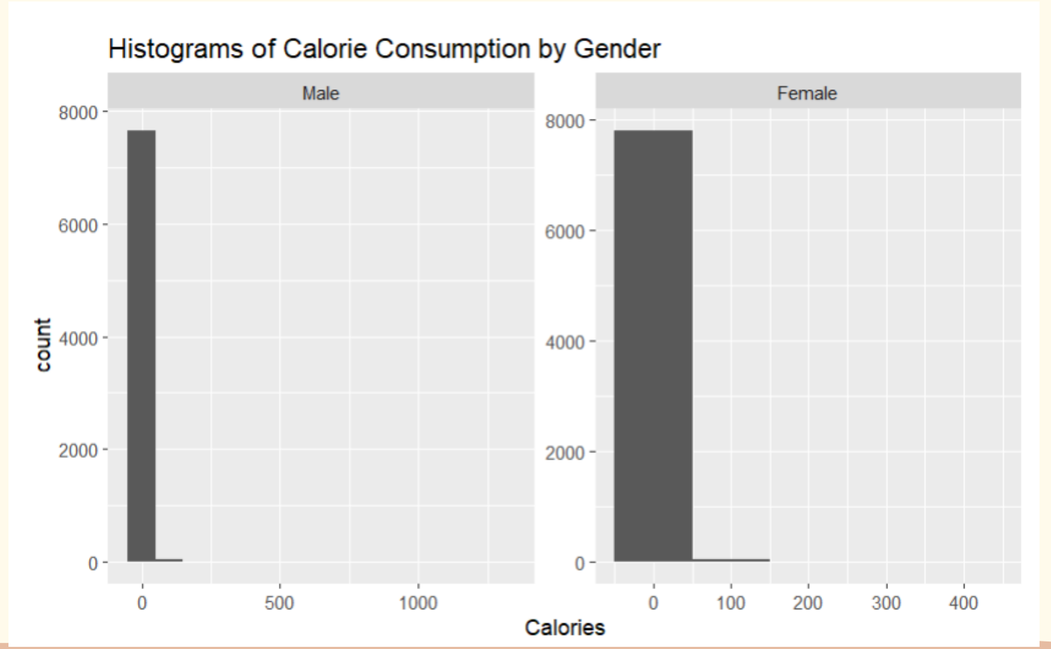


- The correlation matrix reveals several notable correlations:
- Calories are strongly positively correlated with total fat (0.80), carbs (0.67), and protein (0.54). This means that calorie intake is mainly determined by the consumption of these macronutrients.
- Age has weak negative correlations with most nutritional factors, suggesting that calorie intake tends to decrease slightly with age.
- Sugar intake is moderately positively correlated with carbs (0.55) and calories (0.47), indicating that higher sugar intake is associated with higher carbohydrate and overall calorie consumption.
- Fiber intake has modest positive correlations with carbs (0.38), protein (0.29), and calories (0.28), suggesting that foods higher in fiber tend to be higher in these macronutrients and contribute to overall calorie intake.
- There are moderately strong positive inter-correlations among protein, carbs, and total fat (0.42 to 0.55), indicating that foods high in one macronutrient tend to be higher in the others as well.
- Sodium intake appears relatively independent of calorie consumption and the intake of other nutrients in this population.

VISUALIZATION OF DATA DISTRIBUTION

- Histograms are used to visualize calorie consumption distribution by age group and gender.
- The data on calorie consumption across different age groups and gender does not appear to be normally distributed. The histograms display skewed or non-symmetric distributions.





STATISTICAL ANALYSIS

Kruskal-Wallis test:

The Kruskal-Wallis test is a non-parametric statistical test used to determine whether there are significant differences between three or more independent groups.

```
        Kruskal-Wallis rank sum test

data:  calories by age
Kruskal-Wallis chi-squared = 44.103, df =
4, p-value = 6.107e-09

The distributions of calorie consumption across age groups are significantly
different.
```

STATISTICAL ANALYSIS

Dunn’s Test:

Dunn's test is a post-hoc test which compares pairwise differences between groups.

```
data: x and group
Kruskal-Wallis chi-squared = 44.103, df = 4, p-value = 0


                          Comparison of x by group
                                (Bonferroni)
Col Mean-|
Row Mean |       0-18      19-30      31-45      46-60
---------+----------------------------------------------
   19-30 |  -5.553401
         |    0.0000*
         |
   31-45 |  -4.565414   1.051877
         |    0.0000*    1.0000
         |
   46-60 |  -3.624121   1.880345   0.850150
         |    0.0014*    0.3003     1.0000
         |
     61+ |  -3.072601   2.847194   1.778870   0.864471
         |    0.0106*    0.0221*     0.3763     1.0000

alpha = 0.05
Reject Ho if p <= alpha/2
$chi2
[1] 44.10297
```

```
$Z
 [1] -5.5534013 -4.5654149  1.0518773 -3.6241215
 [5]  1.8803455  0.8501509 -3.0726017  2.8471950
 [9]  1.7788707  0.8644712

$P
 [1] 1.400820e-08 2.492537e-06 1.464279e-01
 [4] 1.449726e-04 3.003050e-02 1.976206e-01
 [7] 1.061008e-03 2.205317e-03 3.763048e-02
[10] 1.936645e-01

$P.adjusted
 [1] 1.400820e-07 2.492537e-05 1.000000e+00
 [4] 1.449726e-03 3.003050e-01 1.000000e+00
 [7] 1.061008e-02 2.205317e-02 3.763048e-01
[10] 1.000000e+00

$comparisons
 [1] "0-18 - 19-30"  "0-18 - 31-45"
 [3] "19-30 - 31-45" "0-18 - 46-60"
 [5] "19-30 - 46-60" "31-45 - 46-60"
 [7] "0-18 - 61+"    "19-30 - 61+"
 [9] "31-45 - 61+"   "46-60 - 61+"
```

LINEAR REGRESSION

```
Call:
lm(formula = calories ~ age + gender, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
 -16.72   -0.87   -0.08    0.43 1253.18

Coefficients:
             Estimate Std. Error t value
(Intercept)   14.8929     0.2426  61.396
age19-30       1.9320     0.4258   4.537
age31-45       1.5497     0.4016   3.859
age46-60       1.3014     0.3934   3.308
age61+         1.1852     0.3427   3.458
genderFemale  -0.9131     0.2556  -3.572
             Pr(>|t|)
(Intercept)   < 2e-16 ***
age19-30     5.75e-06 ***
age31-45     0.000114 ***
age46-60     0.000942 ***
age61+       0.000545 ***
genderFemale 0.000356 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.93 on 15554 degrees of freedom
Multiple R-squared:  0.002838,  Adjusted R-squared:  0.002517
F-statistic: 8.853 on 5 and 15554 DF,  p-value: 2.105e-08
```

LOGISTIC REGRESSION

```
Call:
glm(formula = calories ~ age + gender, data = data)

Coefficients:
             Estimate Std. Error t value
(Intercept)   14.8929     0.2426  61.396
age19-30       1.9320     0.4258   4.537
age31-45       1.5497     0.4016   3.859
age46-60       1.3014     0.3934   3.308
age61+         1.1852     0.3427   3.458
genderFemale  -0.9131     0.2556  -3.572
             Pr(>|t|)
(Intercept)   < 2e-16 ***
age19-30     5.75e-06 ***
age31-45     0.000114 ***
age46-60     0.000942 ***
age61+       0.000545 ***
genderFemale 0.000356 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 253.8538)

    Null deviance: 3959679  on 15559  degrees of freedom
Residual deviance: 3948443  on 15554  degrees of freedom
AIC: 130317

Number of Fisher Scoring iterations: 2
```

RESULTS

Data Distribution Visualization:

- Histograms and density plots depict the distribution of calorie consumption within age groups and genders.
- Boxplots illustrate the variability in calorie intake across different demographic categories.

Statistical Analysis Findings:

- The Kruskal-Wallis test reveals significant differences in calorie consumption across age groups.
- Dunn's test identifies specific pairwise differences between age groups.
- Linear regression indicates the influence of age and gender on calorie consumption.
- Logistic regression highlights the likelihood of high calorie intake based on demographic factors.

INTERPRETATION

Age Effects:

- The Kruskal-Wallis test indicates that there are significant differences in calorie consumption across different age groups.
- This suggests that age plays a significant role in determining calorie intake, with individuals in different age brackets exhibiting varying dietary habits.

Gender Effects:

- Gender-related factors such as metabolic rates, body composition, and dietary preferences may influence calorie consumption patterns.

Interaction Effects:

- The linear regression model confirms age and gender significantly influence calorie intake.
- Age group coefficients show their impact compared to a reference group. Females consume fewer calories than males, considering age.
- Interaction between age and gender also affects consumption, highlighting dietary complexity across demographics.

APPLICATIONS AND FUTURE DIRECTIONS

Applications:

- Health authorities can develop personalized dietary guidelines based on age and gender demographics.
- Nutritional programs can be designed to address specific dietary needs of diverse population segments.

Future Directions:

- Future research could investigate other factors influencing dietary behavior, such as socioeconomic status or cultural practices.
- Longitudinal studies could explore changes in dietary habits over time and their impact on health outcomes.

CONCLUSION

- Overall, the results suggest that age and gender are important determinants of calorie consumption among individuals in the studied population.
- Younger age groups tend to have higher calorie needs, while males generally consume more calories than females.
- However, the relationship between age, gender, and calorie intake is multifaceted, with potential interactions between these factors influencing dietary habits.
- These findings underscore the need for tailored dietary interventions and public health policies that consider age and gender differences to promote healthy eating behaviors and prevent diet-related diseases.
- Based on the statistical analyses and models, we reject the null hypothesis and accept the alternative hypothesis that age and gender are significantly associated with calorie intake patterns in the diverse study population.

THANK YOU