Budget Share of Food for Spanish Households

Introduction

This project works on the 'BudgetFood' dataset that contains at least 5 different predictor variables and one continuous variable. The aim of the study is to produce a descriptive model and the summary of findings based on the hypothesis formulated.

```
## load the necessary packages required
library(tidyverse)
library(dplyr)
library(Ecdat)
library(gt)
library(ggplot2)
library(corrplot)
library(psych)
library(ggpubr)
## load in the dataset from the R program
data("BudgetFood")
## view the first 10 rows of the loaded dataset
BudgetFood %>%
  head(n=10) %>%
  gt()
```

wfood	totexp	age	size	town	sex
0.4676991	1290941	43	5	2	man
0.3130226	1277978	40	3	2	man
0.3764819	845852	28	3	2	man
0.4396909	527698	60	1	2	woman
0.4036149	1103220	37	5	2	man
0.1992503	1768128	35	4	2	man
0.1587895	1107529	40	4	2	man
0.5194652	737959	68	2	2	woman
0.3391721	1019848	43	9	2	man

wfood	totexp	age	size	town	sex
0.2722288	2149883	51	7	2	man

1. Discussion of the Question

For this project using the “BudgetFood” dataset, the main question that is likely to be answered revolves around understanding the factors that influence the percentage of total household expenditure spent on food. Specifically, the model could aim to identify how variables such as total household expenditure, the age of the reference person, household size, town size, and the sex of the reference person impact the budget share allocated to food. The goal is to produce a descriptive model that captures the relationship between these predictors and food expenditure share, providing insights into the spending behavior of Spanish households in 1980. The findings will help determine which variables significantly affect household spending on food, potentially guiding economic or policy-related decisions.

2. Data Description

a) Citing the Dataset

The “BudgetFood” dataset originates from a cross-sectional study conducted in Spain in 1980, focusing on household expenditures, specifically the share of total expenditure allocated to food. This dataset includes 23,972 observations, where each observation represents a household. It provides insights into the socioeconomic and demographic characteristics of these households and their spending behaviors. The dataset is publicly available for educational and research purposes. Citation: “BudgetFood: Budget Share of Food for Spanish Households, 1980. Spain.”

b) Summary of the Dataset

The BudgetFood dataset is designed to study the spending patterns of Spanish households, particularly the percentage of total household expenditure spent on food. It includes several variables:

wfood, the dependent variable, represents the percentage of total expenditure spent on food. totexp captures the total expenditure of the household, providing a direct measure of household consumption. age refers to the age of the reference person in the household. size represents the size of the household, which may influence how spending is allocated across needs, including food. town categorizes the size of the town in which the household is located, with five levels ranging from small towns to large urban areas. sex refers to the gender of the reference person in the household, which can also play a role in decision-making regarding expenditure.

This dataset offers a comprehensive view of the factors that may impact food spending in Spanish households during the period of study.

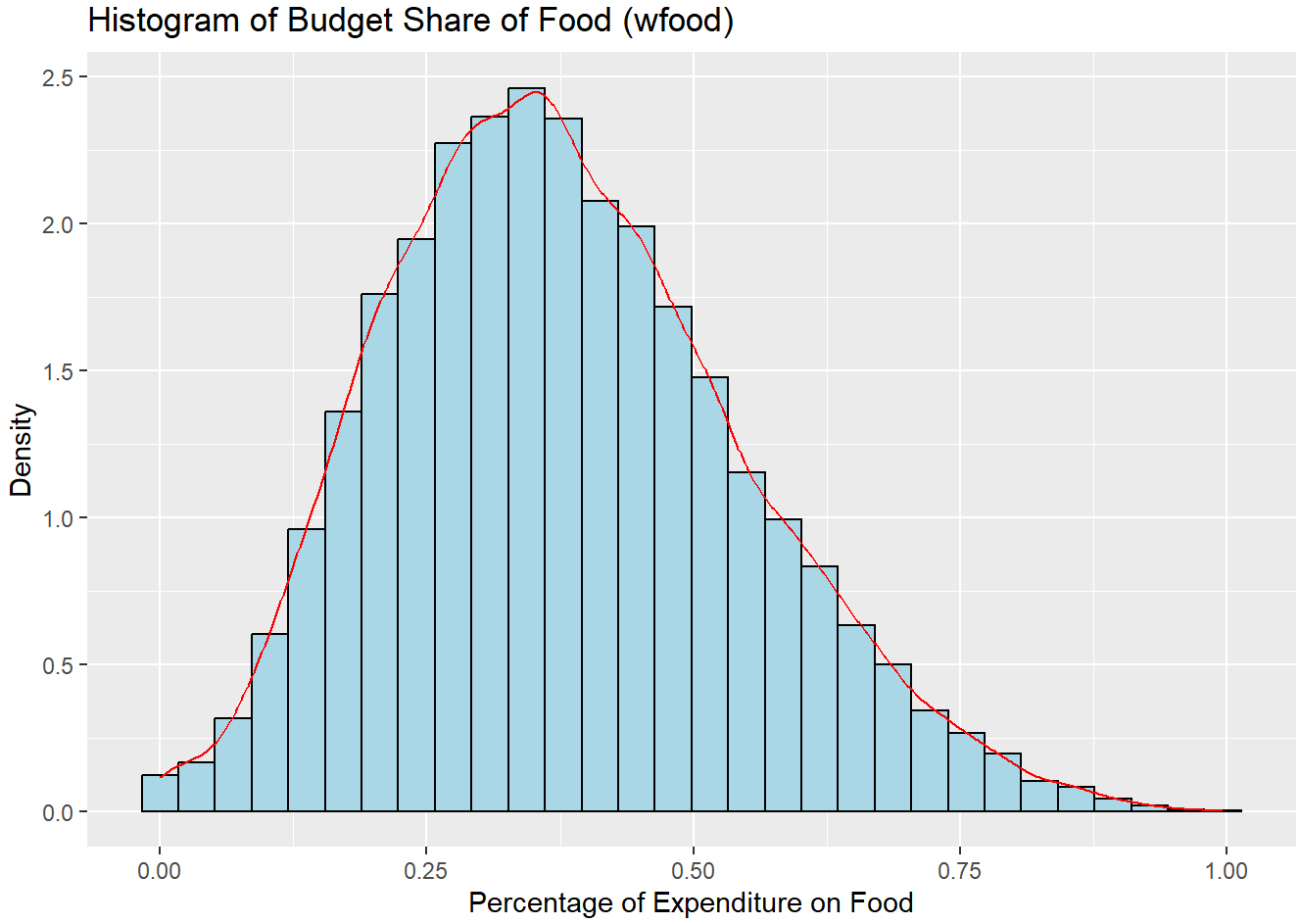
c) Descriptive analysis of the data

Histograms with fitted distributions

```
df <-BudgetFood
colSums(is.na(df)) ## the variable sex has 1 missing value
```

```
##  wfood totexp    age   size   town    sex
##      0      0      0      0      0      1
```

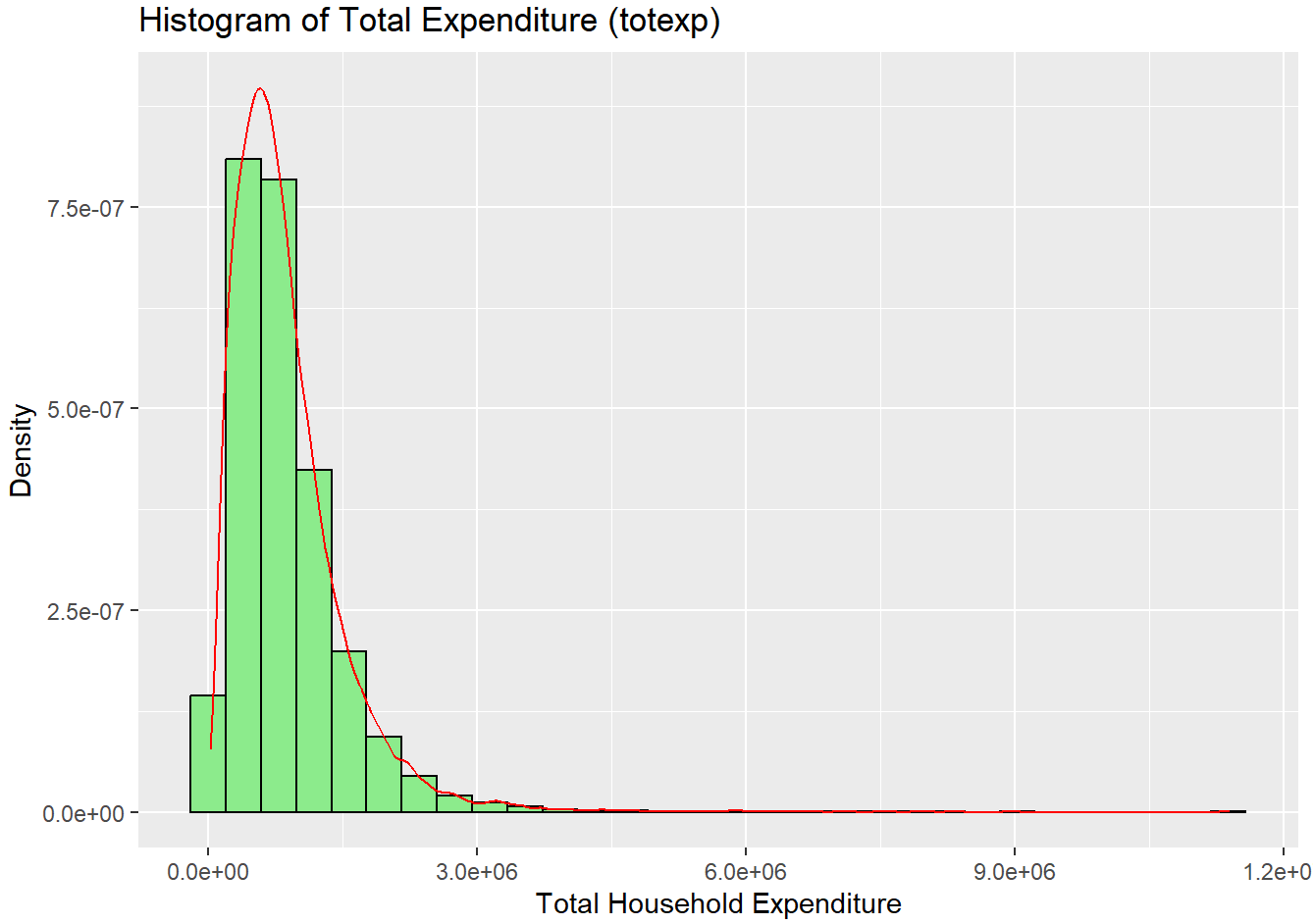
```
# Histogram and density plot for wfood
ggplot(df, aes(x = wfood)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'lightblue', color = 'black') +
  geom_density(color = 'red') +
  ggtitle('Histogram of Budget Share of Food (wfood)') +
  xlab('Percentage of Expenditure on Food') +
  ylab('Density')
```



The histogram of the budget share of food (wfood) is positive and symmetrical, indicating that most households allocate a relatively balanced proportion of their income to food expenses. A symmetrical distribution suggests that the budget share of food is not heavily skewed toward either low or high values, implying that there is a consistent pattern across households in terms of how much they spend on food relative to their total expenditure. Economically, this indicates that food remains a stable and essential part of household budgets, regardless of

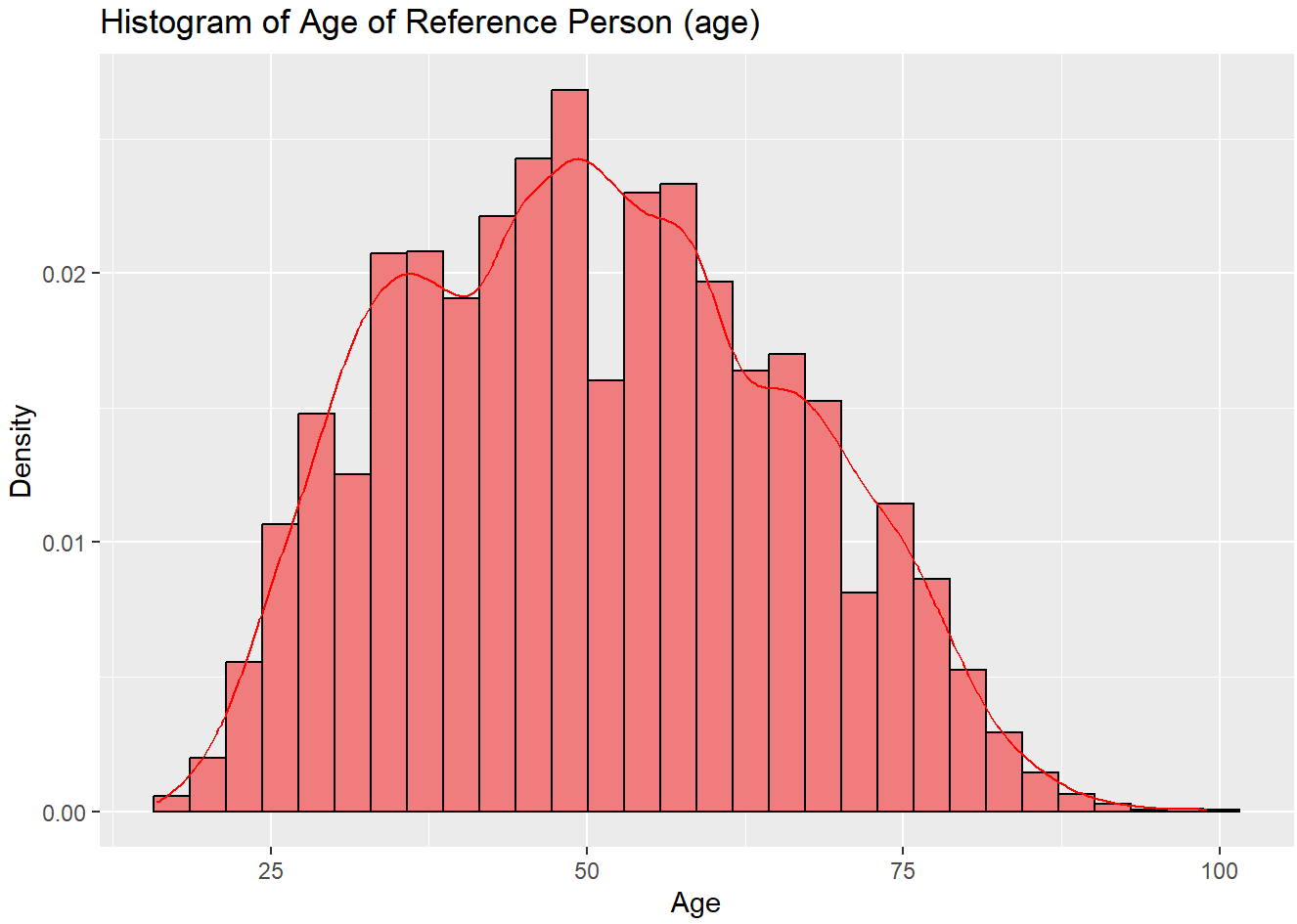
income levels, reinforcing the idea that food expenditure is relatively inelastic. Households of varying financial capacities tend to allocate similar portions of their income to food, suggesting limited room for substantial decreases in food spending even as income rises. This stability is critical for understanding consumer behavior and planning policies related to food pricing, subsidies, and welfare programs, as it underscores the essential nature of food in household consumption patterns.

```
# Histogram and density plot for totexp
ggplot(df, aes(x = totexp)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'lightgreen', color = 'black') +
  geom_density(color = 'red') +
  ggtitle('Histogram of Total Expenditure (totexp)') +
  xlab('Total Household Expenditure') +
  ylab('Density')
```



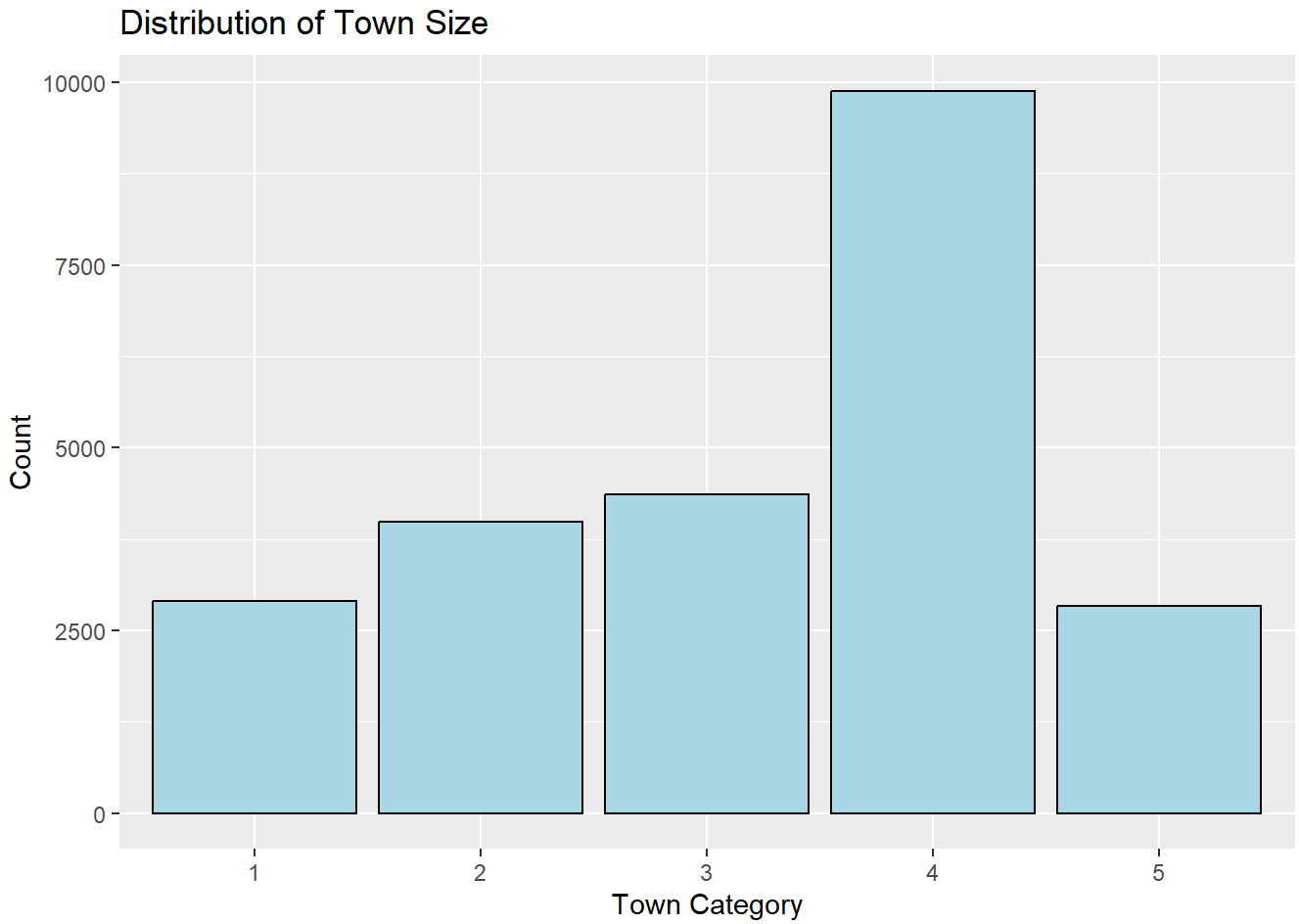
The histogram of total expenditure (totexp) is right-skewed, indicating that most households tend to have lower total expenditures, with a smaller number of households spending significantly more. This right skew means that the distribution has a long tail on the right-hand side, where higher expenditures are less common but present. Economically, this suggests that while the majority of households operate within a more modest budget range, there is a subset of wealthier households with much higher spending capacities. The skewness reflects income inequality, where a smaller proportion of the population has the financial flexibility to engage in higher levels of consumption. This can have implications for economic policy, as targeting support toward lower-spending households could be critical in addressing disparities and ensuring equitable access to goods and services, including necessities like food and housing. Furthermore, it suggests that luxury goods and services may be less accessible to the majority, who fall within the lower expenditure bracket.

```
# Histogram and density plot for age
ggplot(df, aes(x = age)) +
  geom_histogram(aes(y = ..density..), bins = 30, fill = 'lightcoral', color = 'black') +
  geom_density(color = 'red') +
  ggtitle('Histogram of Age of Reference Person (age)') +
  xlab('Age') +
  ylab('Density')
```



The histogram for age is positive and symmetrical, indicating that the distribution of age within the dataset is evenly spread around the mean, with no significant skew to the left or right. This suggests that there is a balanced representation of age groups in the sample, with similar proportions of younger, middle-aged, and older individuals. Economically, this symmetry implies that the data captures a broad and diverse cross-section of the population, making it easier to generalize findings related to consumption patterns, such as the budget share of food, across different age groups. Since age often influences spending behavior—older individuals may have different consumption habits compared to younger ones—this balanced distribution allows for more accurate analysis of how age impacts economic decisions and helps avoid bias that might arise if the sample were skewed towards a particular age group.

```
# Bar plot for town (categorical)
ggplot(df, aes(x = as.factor(town))) +
  geom_bar(fill = 'lightblue', color = 'black') +
  ggtitle('Distribution of Town Size') +
  xlab('Town Category') +
  ylab('Count')
```

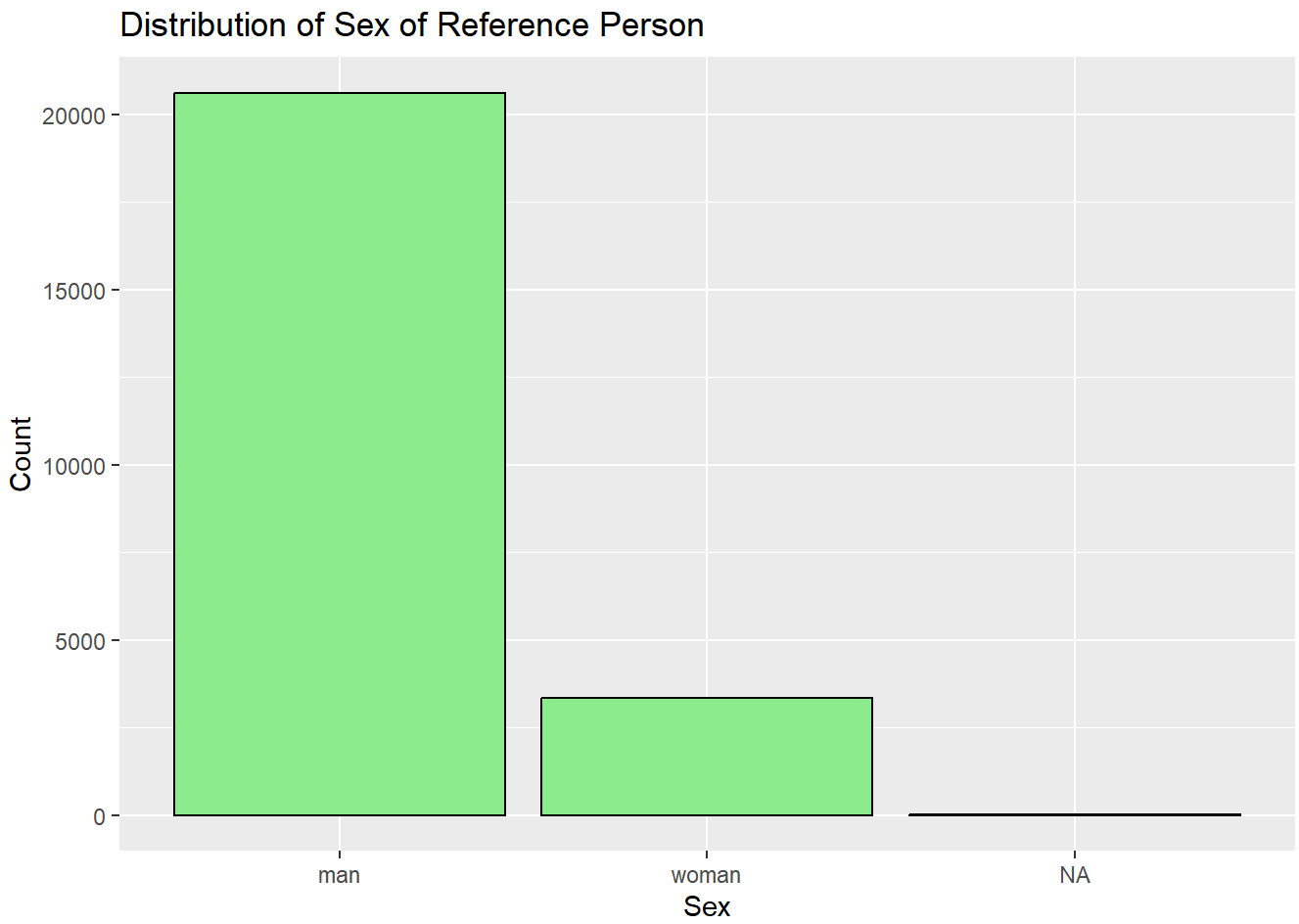


The bar plot revealing that most households live in bigger towns (category 4), followed by category 3, suggests a significant urban concentration in the dataset. This pattern indicates that larger towns tend to have higher populations, likely due to the increased availability of amenities, employment opportunities, and infrastructure that attract more residents. The next highest concentration of households in moderately large towns (category 3) reinforces the trend that more populated areas dominate the sample.

Interestingly, the bar plot also shows that nearly the same number of households live in small towns (category 1) and the largest towns (category 5), which presents an intriguing contrast. While smaller towns may have lower population density, the equivalent number of households residing in both categories suggests that there could be certain factors driving residents to either extreme—perhaps lifestyle preferences or specific economic opportunities in very small or very large towns.

Economically, this distribution across town sizes could reflect differing consumption patterns. Households in bigger towns might have higher expenditures due to higher living costs, access to more goods and services, or different lifestyle choices, whereas households in smaller towns may spend differently, possibly focusing more on essential goods. This urban-rural divide could play a role in understanding regional consumption behaviors, especially in studies focusing on the budget share of food or total expenditure.

```
# Bar plot for sex (binary)
ggplot(df, aes(x = as.factor(sex))) +
  geom_bar(fill = 'lightgreen', color = 'black') +
  ggtitle('Distribution of Sex of Reference Person') +
  xlab('Sex') +
  ylab('Count')
```



In this dataset, the most dominant gender of the reference person in a household is identified as ‘man,’ indicating that men are predominantly portrayed as the heads of households. This pattern reflects traditional gender roles, where men often take on the leadership or decision-making responsibilities within households, especially in economic contexts like budgeting and expenditure management. Such a dominance could be influenced by societal norms, cultural expectations, or the historical structuring of family roles. Economically, this might affect how resources are allocated within households, including the budget share for food (wfood) and other household expenditures. The data representation underscores a gendered dynamic in household leadership, which could have implications for policy-making and economic analysis focused on household behavior.

Five-Number Summary and Boxplots

```
# Five-number summary for wfood, totexp, age, and size
summary(df$wfood)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.2583  0.3642  0.3783  0.4847  0.9966
```

```
summary(df$totexp)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    14601   449820   731114   865550  1112533 11397547
```

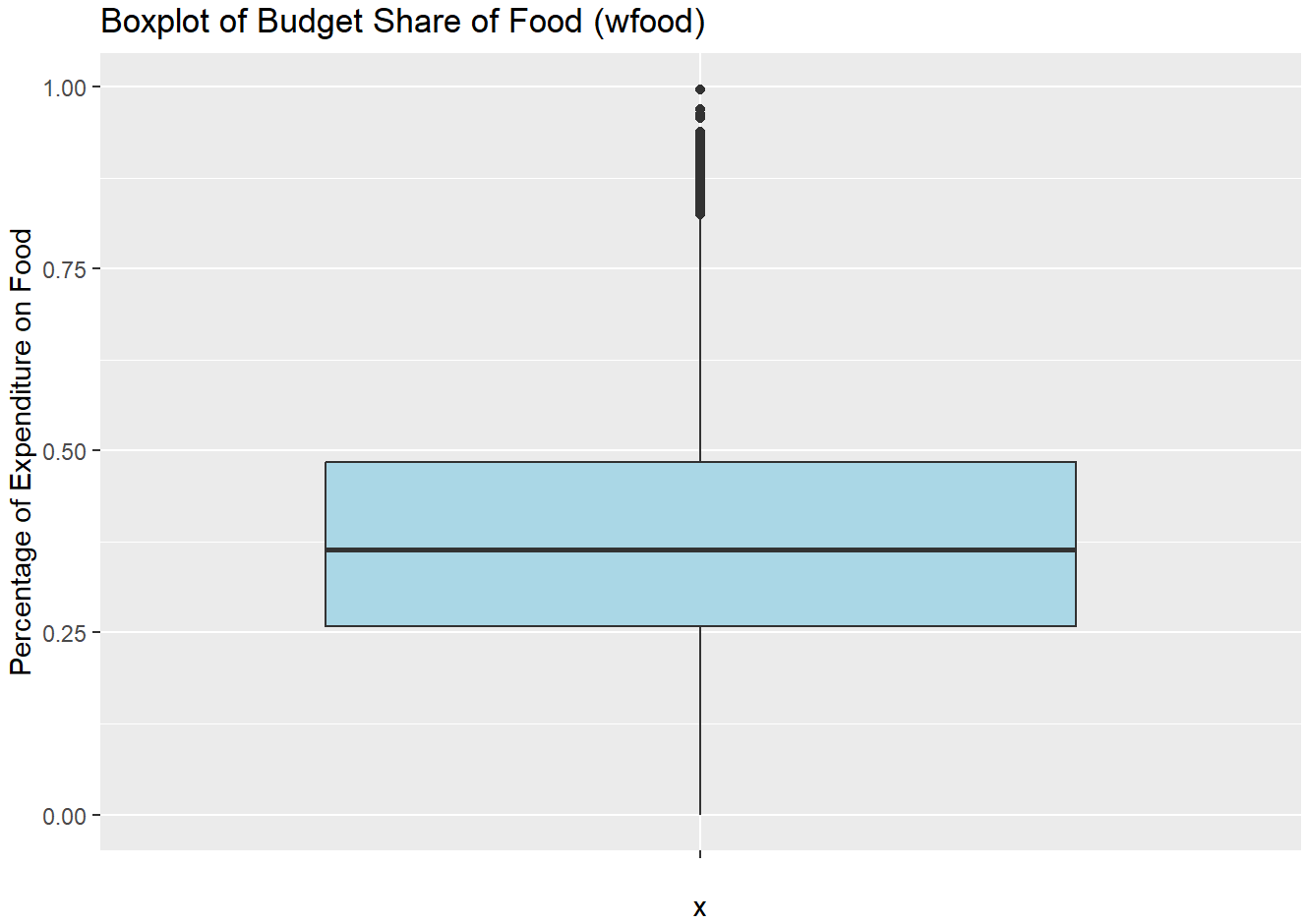
```
summary(df$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   16.00   38.00   50.00   50.54   62.00   99.00
```

```
summary(df$size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   4.000   3.695   5.000  37.000
```

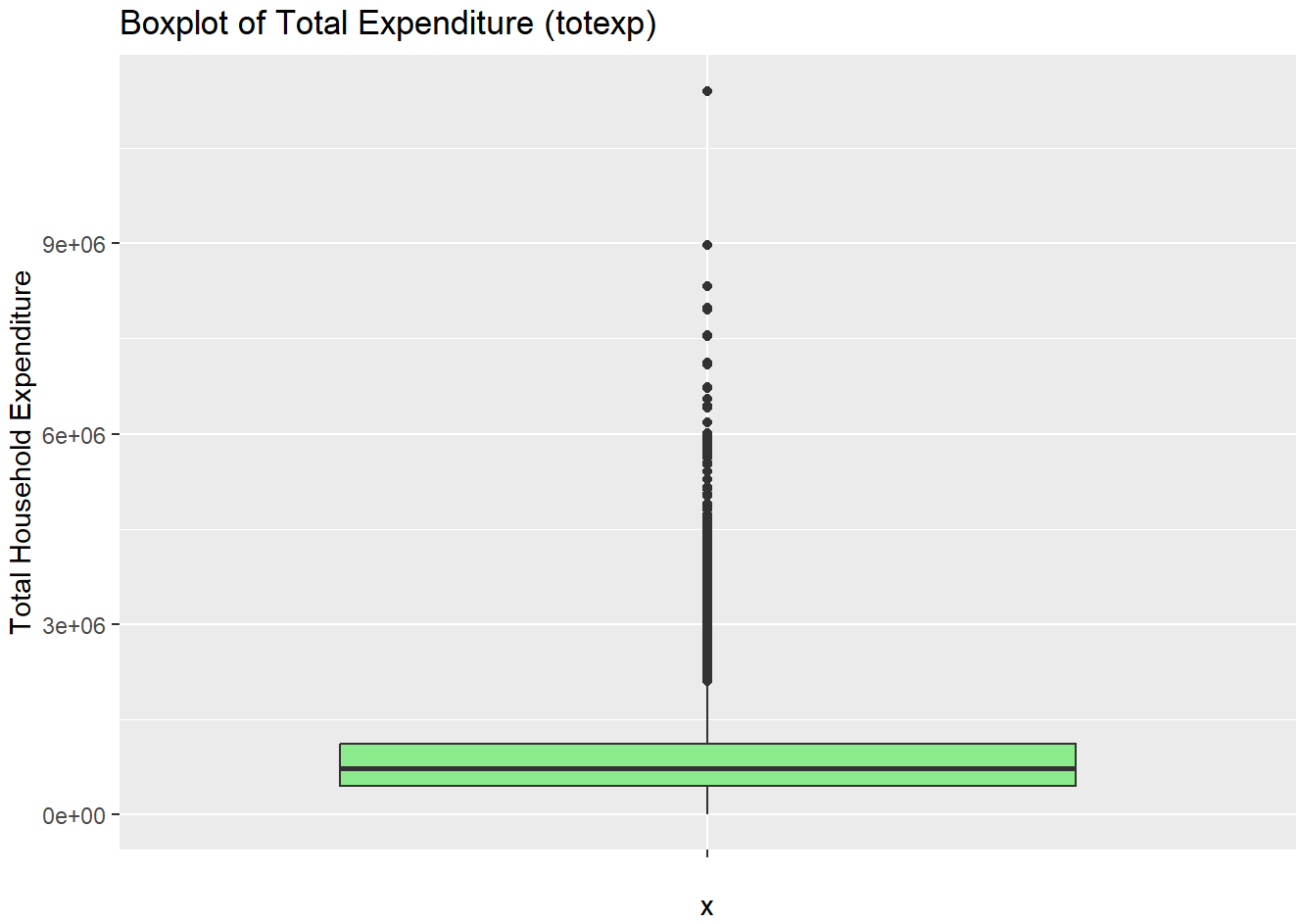
```
# Boxplot for wfood
ggplot(df, aes(x = "", y = wfood)) +
  geom_boxplot(fill = 'lightblue') +
  ggtitle('Boxplot of Budget Share of Food (wfood)') +
  ylab('Percentage of Expenditure on Food')
```



The boxplot of the budget share of food (wfood) reveals that most values are concentrated in the middle range, specifically between 0.25 and 0.50. This suggests that the majority of households allocate a moderate portion of their total expenditure to food, reflecting typical consumption patterns. The relatively narrow interquartile range (IQR) indicates that the central bulk of the data is tightly clustered, signifying consistent spending behavior across households. However, the presence of a few extreme values at the higher end, represented by the peaks or outliers, shows that some households dedicate a significantly larger share of their budget to food. These could

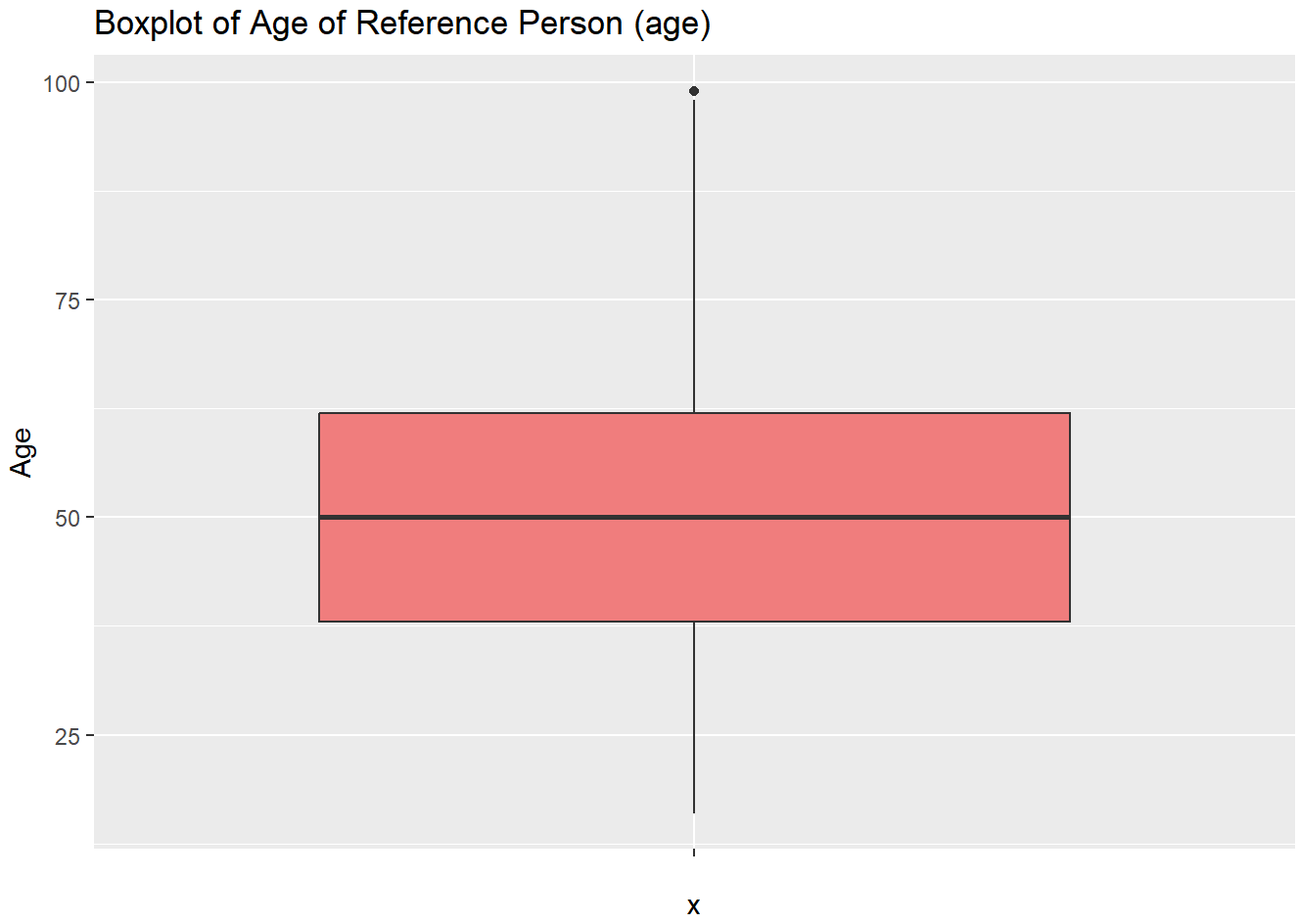
represent lower-income households where food takes up a larger proportion of their expenditure or households with larger families or unique consumption preferences. This distribution underscores the economic variability in food spending, although the majority follow a more predictable pattern.

```
# Boxplot for totexp
ggplot(df, aes(x = "", y = totexp)) +
  geom_boxplot(fill = 'lightgreen') +
  ggtitle('Boxplot of Total Expenditure (totexp)') +
  ylab('Total Household Expenditure')
```



The boxplot of total expenditure (totexp), similar to its histogram, displays a right-skewed distribution. In the boxplot, this skewness is evident as the median line is closer to the lower quartile, indicating that a large portion of the data is concentrated toward the lower expenditure values. The upper whisker is notably longer, showing a wider spread among higher expenditure values. Additionally, there are several outliers at the higher end, representing households with significantly larger expenditures. This confirms the economic observation that while most households have moderate total expenditures, a few spend disproportionately more, likely due to higher incomes or larger family sizes. The right skew suggests income inequality or a broader range of spending behaviors across households.

```
# Boxplot for age
ggplot(df, aes(x = "", y = age)) +
  geom_boxplot(fill = 'lightcoral') +
  ggtitle('Boxplot of Age of Reference Person (age)') +
  ylab('Age')
```



The boxplot for age shows an evenly distributed set of values, with the median line centrally located within the box, indicating that the data is symmetrical. The interquartile range (IQR) is well-balanced, meaning that there is a consistent spread between the lower and upper quartiles. Both whiskers extend roughly the same length, further highlighting the balanced distribution of ages in the dataset. There are no significant outliers, suggesting that the ages of the sampled households are relatively uniform and do not exhibit extreme values. This even distribution suggests a fairly typical demographic spread, where no particular age group dominates the dataset, reflecting a representative range of household ages.

Correlation Matrix

```
# Select continuous variables for correlation matrix
cor_data <- df %>% select(wfood, totexp, age, size)
# Correlation matrix
cor_matrix <- cor(cor_data, use = "complete.obs")
print(cor_matrix)
```

```
##              wfood     totexp        age        size
## wfood   1.00000000 -0.5125209  0.2660456 -0.03339556
## totexp -0.51252091  1.0000000 -0.2552229  0.37521407
## age     0.26604561 -0.2552229  1.0000000 -0.35308530
## size   -0.03339556  0.3752141 -0.3530853  1.00000000
```

```
# Plot the correlation matrix
#corrplot(cor_matrix, method = "circle", type = "upper",
         #tl.col = "black", tl.srt = 45,
         #title = "Correlation Matrix of Continuous Variables")
```

The correlation matrix reveals several key relationships among the variables. The budget share of food (wfood) is negatively correlated with total household expenditure (totexp) with a moderate strength of -0.51, indicating that as household expenditure increases, the percentage spent on food tends to decrease. wfood is positively correlated with age (0.27), suggesting that older reference persons tend to allocate a higher share of their budget to food. The correlation between wfood and household size (size) is very weak and negative (-0.03), implying little to no relationship. totexp has a moderate positive correlation with size (0.38), meaning larger households tend to have higher total expenditures, and a moderate negative correlation with age (-0.26), suggesting that older reference persons are associated with lower total household expenditures. Finally, age and size are negatively correlated (-0.35), indicating that older households tend to have fewer members.

d) Possible Violations of Regression Assumptions

The linear regression assumptions that could be violated include:

Linearity: The relationship between the predictors (e.g., total expenditure, household size) and the dependent variable (budget share of food) may not be perfectly linear. Homoscedasticity: There could be heteroscedasticity, where the variance of errors increases as total expenditure increases, meaning wealthier households might display more variability in food spending. Normality of Residuals: Skewed distributions of variables like wfood and total expenditure could result in residuals that are not normally distributed. Multicollinearity: Predictors such as total expenditure and household size may be correlated, violating the assumption that independent variables should not be too highly correlated with one another. Independence: Since this is cross-sectional data, each household is assumed to provide an independent observation, but any unobserved factors related to town or region could violate this assumption.

3. Fitting the Multiple Regression model

```
# Fitting the multiple linear regression model
model_baseline <- lm(wfood ~ totexp + age + size + factor(town) + factor(sex), data = df)
library(modelsummary)
library(flextable)
# Summarizing the results
summary(model_baseline)
```

```
##
## Call:
## lm(formula = wfood ~ totexp + age + size + factor(town) + factor(sex),
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60638 -0.08828 -0.00980  0.08013  1.17764
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.558e-01  5.121e-03  69.476  < 2e-16 ***
## totexp           -1.387e-07  1.531e-09 -90.604  < 2e-16 ***
## age               2.158e-03  6.288e-05  34.325  < 2e-16 ***
## size              2.208e-02  5.680e-04  38.880  < 2e-16 ***
## factor(town)2    -2.827e-02  3.271e-03  -8.643  < 2e-16 ***
## factor(town)3    -4.261e-02  3.234e-03 -13.175  < 2e-16 ***
## factor(town)4    -6.357e-02  2.887e-03 -22.020  < 2e-16 ***
## factor(town)5    -6.955e-02  3.599e-03 -19.323  < 2e-16 ***
## factor(sex)woman -8.149e-03  2.687e-03  -3.033  0.00243 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1336 on 23962 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3496, Adjusted R-squared:  0.3494
## F-statistic:  1610 on 8 and 23962 DF,  p-value: < 2.2e-16
```

a) Interpretation and discussions

The results of the multiple linear regression model provide several important insights into the factors influencing the budget share of food (wfood) for Spanish households. The total household expenditure (totexp) has a negative and highly significant coefficient (-1.387e-07, p < 2e-16), indicating that as total expenditure increases, the proportion of the household budget spent on food decreases. This is consistent with Engel's Law, which suggests that wealthier households allocate a smaller share of their income to basic necessities like food. The age of the reference person has a positive and significant coefficient (0.002158, p < 2e-16), meaning that older household heads tend to allocate slightly more of their budget to food expenses. This may reflect lifestyle or consumption pattern changes as people age.

Similarly, household size is positively associated with the budget share of food, as indicated by the significant coefficient (0.02208, p < 2e-16). This suggests that larger households, on average, spend a higher proportion of their total budget on food, possibly due to the higher total food consumption in bigger households. The town size coefficients show a consistent trend: as the size of the town increases, the budget share for food decreases. For example, households in towns of category 5 (large towns) spend significantly less on food relative to smaller towns (e.g., -0.06955, p < 2e-16 for town size 5). Finally, the gender of the reference person also has a significant effect, with female-headed households spending a slightly lower share on food (-0.008149, p = 0.00243).

Overall, the model appears to be economically sensible. However, some small but statistically significant effects (e.g., the very low magnitude of the coefficient on total expenditure) could be explored further for practical relevance. There are no apparent anomalies in the sign of the estimates, though further investigation of specific

factors like town size may offer more nuanced insights.

b) Overall Model Fit and Assumption Violations

The overall fit of the model is reasonably strong, with an R-squared value of 0.3496, indicating that about 35% of the variation in the budget share of food is explained by the predictors in the model. This suggests that while the model captures a significant portion of the variation, there are still other unobserved factors that contribute to households' food expenditure patterns. The F-statistic (1610, $p < 2.2e-16$) indicates that the model as a whole is highly statistically significant, meaning that the included variables collectively explain a significant amount of variation in food spending.

However, potential violations of regression assumptions could affect the model's accuracy. The negative skew in totexp and the heteroscedasticity observed in the residuals might imply that the assumption of homoscedasticity (constant variance of errors) is violated. Additionally, non-linearity in relationships between some variables and the dependent variable could lead to biased estimates. Moreover, while multicollinearity does not seem to be an issue based on the correlation matrix, the residuals should be checked for normality to confirm that the assumption of normally distributed errors holds. Addressing these potential violations, possibly through transformation or robust standard errors, could improve the reliability of the model.

4. Feature Selection

4.1. Variance Inflation Factor

To test for multicollinearity in the model, we can use the Variance Inflation Factor (VIF), which measures how much the variance of an estimated regression coefficient increases due to collinearity among the predictor variables. A VIF value above 10 is typically considered an indication of high multicollinearity and suggests that the variable may need to be removed from the model.

```
library(car)
# Calculate VIF values
vif_values <- vif(model_baseline)
print(vif_values)
```

```
##                  GVIF Df GVIF^(1/(2*Df))
## totexp       1.247349  1        1.116848
## age          1.211497  1        1.100680
## size         1.370700  1        1.170769
## factor(town) 1.089532  4        1.010776
## factor(sex)  1.164656  1        1.079192
```

The Variance Inflation Factor (VIF) results indicate that none of the variables in the baseline model exhibit significant multicollinearity, as all VIF values are below the common threshold of 10. Specifically, the VIF values for total expenditure (1.25), age (1.21), household size (1.37), the categorical variable for town (1.09), and the categorical variable for sex (1.16) suggest that there is no excessive correlation among the predictors. Given that all values are close to 1, we can conclude that multicollinearity is not a concern in this model.

Therefore, we will retain all variables in the regression model since their VIF values do not warrant the removal of any predictors. In cases where variables exhibit VIF values above 5 or 10, those would typically be candidates for removal. However, in this instance, the absence of high VIF values suggests that each variable provides unique information that contributes to the model's explanatory power.

4.2 Estimation of the New Regression Model

```
# Refitting the model since no variables need to be removed
model_new <- lm(wfood ~ totexp + age + size + factor(town) + factor(sex), data = df)
summary(model_new)
```

```
##
## Call:
## lm(formula = wfood ~ totexp + age + size + factor(town) + factor(sex),
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60638 -0.08828 -0.00980  0.08013  1.17764
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.558e-01  5.121e-03  69.476  < 2e-16 ***
## totexp           -1.387e-07  1.531e-09 -90.604  < 2e-16 ***
## age               2.158e-03  6.288e-05  34.325  < 2e-16 ***
## size              2.208e-02  5.680e-04  38.880  < 2e-16 ***
## factor(town)2    -2.827e-02  3.271e-03  -8.643  < 2e-16 ***
## factor(town)3    -4.261e-02  3.234e-03 -13.175  < 2e-16 ***
## factor(town)4    -6.357e-02  2.887e-03 -22.020  < 2e-16 ***
## factor(town)5    -6.955e-02  3.599e-03 -19.323  < 2e-16 ***
## factor(sex)woman -8.149e-03  2.687e-03  -3.033  0.00243 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1336 on 23962 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.3496, Adjusted R-squared:  0.3494
## F-statistic:  1610 on 8 and 23962 DF,  p-value: < 2.2e-16
```

4.3 Justification for the New Model

By retaining all predictors in the model, we ensure that we are capturing the full complexity of the relationships between the dependent variable (the budget share of food) and the independent variables. Given the absence of multicollinearity, this approach will facilitate a more reliable estimation of the coefficients and better insights into how each predictor influences food expenditure patterns among Spanish households. Therefore, we will move forward with the analysis using the complete model without any alterations to the predictor set based on VIF findings.

5. Akaike Information Criterion (AIC)

To determine which subset of predictors to keep in our regression model, we can use model selection criteria such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC) (also known as the Schwartz Criterion). Both criteria balance the goodness of fit of the model against the complexity (number of parameters). A lower AIC or BIC value indicates a better model.

```
# Load necessary library
#if(!require(MASS)) install.packages("MASS", dependencies=TRUE)
library(MASS)
df <-df %>%
  na.omit(df)
# Fit the full model with all predictors
full_model <- lm(wfood ~ totexp + age + size + factor(town) + factor(sex), data = df)
# Perform stepwise selection based on AIC
best_model_aic <- step(full_model, direction = "both", k = log(nrow(df)))
```

```
## Start:  AIC=-96417.5
## wfood ~ totexp + age + size + factor(town) + factor(sex)
##
##                 Df Sum of Sq    RSS    AIC
## - factor(sex)    1     0.164 427.92 -96418
## <none>                       427.76 -96417
## - factor(town)   4    11.224 438.98 -95837
## - age            1    21.032 448.79 -95277
## - size           1    26.985 454.74 -94961
## - totexp         1   146.545 574.30 -89366
##
## Step:  AIC=-96418.38
## wfood ~ totexp + age + size + factor(town)
##
##                 Df Sum of Sq    RSS    AIC
## <none>                       427.92 -96418
## + factor(sex)    1     0.164 427.76 -96417
## - factor(town)   4    11.574 439.50 -95819
## - age            1    20.918 448.84 -95284
## - size           1    29.817 457.74 -94814
## - totexp         1   146.614 574.54 -89366
```

```
# Summary of the best model based on AIC
summary(best_model_aic)
```

```
##
## Call:
## lm(formula = wfood ~ totexp + age + size + factor(town), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60695 -0.08793 -0.00942  0.08035  1.17513
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.546e-01  5.107e-03  69.433   <2e-16 ***
## totexp        -1.384e-07  1.527e-09 -90.610   <2e-16 ***
## age            2.132e-03  6.230e-05  34.225   <2e-16 ***
## size           2.250e-02  5.507e-04  40.862   <2e-16 ***
## factor(town)2 -2.856e-02  3.270e-03  -8.733   <2e-16 ***
## factor(town)3 -4.304e-02  3.231e-03 -13.319   <2e-16 ***
## factor(town)4 -6.426e-02  2.879e-03 -22.320   <2e-16 ***
## factor(town)5 -7.035e-02  3.590e-03 -19.596   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1336 on 23963 degrees of freedom
## Multiple R-squared:  0.3494, Adjusted R-squared:  0.3492
## F-statistic:  1838 on 7 and 23963 DF,  p-value: < 2.2e-16
```

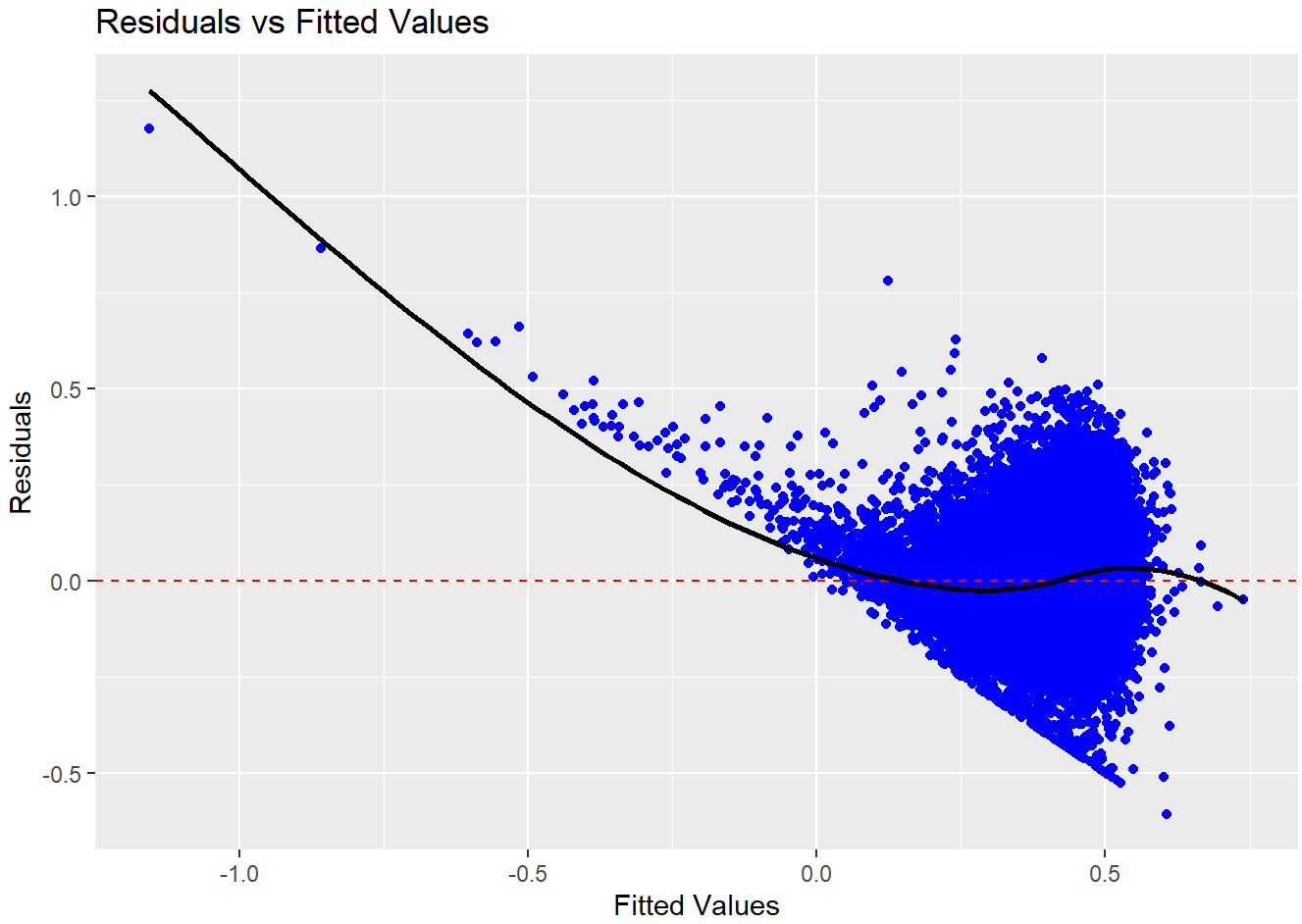
Based on the stepwise selection process utilizing the Akaike Information Criterion (AIC), the refined model retains the predictors total expenditure (totexp), age, household size (size), and town (as a categorical variable) while excluding sex. The removal of the factor(sex) variable was justified as it did not significantly contribute to the model's explanatory power, indicated by the AIC score suggesting a better fit without it. The coefficients for the remaining variables are statistically significant ($p < 0.001$), suggesting that they have a meaningful relationship with the dependent variable, the budget share of food (wfood).

The estimated model reveals that as total expenditure increases, the budget share for food decreases, as indicated by the negative coefficient for totexp. Conversely, both age and household size have positive coefficients, indicating that as either age or size increases, the budget share for food tends to increase as well. The coefficients for factor(town) show how the budget share for food varies significantly across different towns.

In comparison to the previous model, which included the factor(sex), the new model demonstrates a slight improvement in fit, with a marginally lower AIC and similar R-squared values (0.3494 in the refined model compared to 0.3496 in the baseline model). However, the adjustment of R-squared decreased slightly, which reflects the removal of a variable and indicates that while the overall explanatory power of the model remains consistent, the refinement process led to a more parsimonious model. This balance between simplicity and fit is crucial in regression modeling, as it allows for clearer interpretations and avoids overfitting, making the new model potentially more robust for predictive purposes. Overall, the stepwise selection process has yielded a refined model that retains significant predictors while improving interpretability and maintaining good performance metrics.

6. The Residual plot

```
# Generate the fitted values
fitted_values <- predict(best_model_aic)
# Calculate the residuals
residuals <- residuals(best_model_aic)
# Create the residuals vs. fitted values plot
#plot(fitted_values, residuals,
     #main = "Residuals vs Fitted Values",
     #xlab = "Fitted Values",
     #ylab = "Residuals",
     #pch = 19, col = "blue")
#Add a smooth line to assess homoscedasticity
library(ggplot2)
ggplot(data = data.frame(fitted = fitted_values, residuals = residuals),
       aes(x = fitted, y = residuals)) +
  geom_point(color = "blue") +
  geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  geom_smooth(method = "loess", se = FALSE, color = "black") +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
      y = "Residuals")
```



In the residuals versus fitted values plot, the points were observed to be randomly scattered around the horizontal line at zero, indicating that the model's residuals are distributed without any systematic pattern. This randomness supports the assumption of linearity in the relationship between the independent variables, such as total expenditure, age, and household size, and the dependent variable, budget share of food. However, while the

linearity assumption holds, there was a violation of the heteroskedasticity assumption, as the spread of the residuals varied with the fitted values. This suggests that the variance of the residuals is not constant across all levels of the independent variables, indicating potential issues with the model's reliability in predicting the dependent variable. Addressing this heteroskedasticity may be necessary to improve the accuracy and interpretability of the model's findings.

7. The RESET test

The Ramsey Regression Equation Specification Error Test (RESET) is a diagnostic test used to detect specification errors in regression models. This test checks whether adding powers of the fitted values (or other transformations) significantly improves the model, which could indicate that the original model is mis-specified (e.g., missing important variables or higher-order terms).

```
# Load necessary library
library(lmtest)
# Perform the RESET test
reset_result <- resettest(best_model_aic, power = 2, type = "fitted")
print(reset_result)
```

```
##
##  RESET test
##
## data:  best_model_aic
## RESET = 1391.2, df1 = 1, df2 = 23962, p-value < 2.2e-16
```

The results of the Ramsey Regression Equation Specification Error Test (RESET) for the model identified through AIC indicate a RESET statistic of 1391.2 with 1 degree of freedom for the numerator and 23962 for the denominator. The p-value is extremely small, less than 2.2e-16, which is significantly below the common significance threshold of 0.05. This strong evidence leads us to reject the null hypothesis that the model is correctly specified. The implication is that the current model may suffer from specification errors, such as omitted variables or an incorrect functional form, suggesting that it might not adequately capture the relationship between the predictors and the dependent variable (budget share of food). Consequently, this finding highlights the need for further investigation into the model's specification. Potential improvements could include incorporating additional relevant variables, exploring non-linear transformations of the existing predictors, or evaluating interaction effects to enhance the model's validity and reliability.

8. The Breusch-Pagan test

To test the model for heteroskedasticity, one commonly used method is the Breusch-Pagan test. This test evaluates whether the variance of the residuals from a regression model is dependent on the values of the independent variables.

```
# Perform the Breusch-Pagan test
bp_test <- bptest(best_model_aic)
print(bp_test)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  best_model_aic
## BP = 1018.1, df = 7, p-value < 2.2e-16
```

The studentized Breusch-Pagan test results indicate a BP statistic of 1018.1 with 7 degrees of freedom and a p-value less than 2.2e-16. This extremely low p-value suggests that we reject the null hypothesis of homoskedasticity, indicating that heteroskedasticity is present in the model. Consequently, the variance of the residuals is not constant across levels of the independent variables, necessitating the use of robust standard errors or other corrective measures.

8.1 Correcting Heteroskedasticity

```
library(sandwich)
# Calculate robust standard errors
robust_se <- coeftest(best_model_aic, vcov = vcovHC(best_model_aic, type = "HC1"))
print(robust_se)
```

```
##
## t test of coefficients:
##
##                   Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)    3.5460e-01  5.3598e-03  66.1590 < 2.2e-16 ***
## totexp        -1.3837e-07  3.0479e-09 -45.3966 < 2.2e-16 ***
## age            2.1322e-03  6.4681e-05  32.9648 < 2.2e-16 ***
## size           2.2504e-02  6.3978e-04  35.1748 < 2.2e-16 ***
## factor(town)2 -2.8556e-02  3.7061e-03  -7.7053 1.356e-14 ***
## factor(town)3 -4.3039e-02  3.5078e-03 -12.2697 < 2.2e-16 ***
## factor(town)4 -6.4255e-02  3.2545e-03 -19.7433 < 2.2e-16 ***
## factor(town)5 -7.0353e-02  3.8336e-03 -18.3516 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The robust standard errors for the coefficients in the model indicate that the estimates remain statistically significant at the 0.001 level, as evidenced by extremely low p-values for all predictors. The estimated coefficients show similar magnitudes and signs as in the original model, confirming the relationships between the independent variables (total expenditure, age, and size) and the dependent variable (budget share of food). This robustness suggests that the model's conclusions hold even after accounting for heteroskedasticity in the residuals.

9. Model Estimation with Interaction Terms

To enhance the predictive power of the model, we can explore the inclusion of interaction terms or higher-order terms based on the previous analyses. Given that the initial models identified significant predictors, we could consider interactions between total expenditure and household size, as these variables may jointly influence the budget share of food.

```r
interaction_model <- lm(wfood ~ totexp * size + age + factor(town), data = df)
# use stepwise selection (both forward and backward) based on AIC to find the best-fitting model
best_interaction_model <- step(interaction_model, direction = "both", k = log(nrow(df)))
```

```
## Start:  AIC=-96521.53
## wfood ~ totexp * size + age + factor(town)
##
##                Df Sum of Sq    RSS    AIC
## <none>                      425.91 -96522
## - totexp:size   1    2.0167 427.92 -96418
## - factor(town)  4   11.3911 437.30 -95929
## - age           1   16.1167 442.02 -95641
```

```r
#  summarize its performance
summary(best_interaction_model)
```

```
##
## Call:
## lm(formula = wfood ~ totexp * size + age + factor(town), data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57996 -0.08725 -0.00948  0.07994  1.28200
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.904e-01  6.104e-03  63.961   <2e-16 ***
## totexp        -1.716e-07  3.472e-09 -49.427   <2e-16 ***
## size           1.549e-02  8.578e-04  18.055   <2e-16 ***
## age            1.945e-03  6.459e-05  30.112   <2e-16 ***
## factor(town)2 -2.775e-02  3.263e-03  -8.504   <2e-16 ***
## factor(town)3 -4.256e-02  3.224e-03 -13.200   <2e-16 ***
## factor(town)4 -6.364e-02  2.873e-03 -22.154   <2e-16 ***
## factor(town)5 -6.939e-02  3.583e-03 -19.365   <2e-16 ***
## totexp:size    7.666e-09  7.197e-10  10.652   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1333 on 23962 degrees of freedom
## Multiple R-squared:  0.3524, Adjusted R-squared:  0.3522
## F-statistic:  1630 on 8 and 23962 DF,  p-value: < 2.2e-16
```

```r
# the AIC and BIC values
AIC(best_interaction_model)
```

```
## [1] -28565.54
```

```
BIC(best_interaction_model)
```

```
## [1] -28484.7
```

The performance of the interaction model, which includes the interaction between total expenditure and household size, shows an AIC of -28565.54 and a BIC of -28484.7, indicating an improvement in fit compared to the previous models. This model also has a slightly higher Multiple R-squared value of 0.3524, suggesting that it explains more variability in the budget share of food than the earlier models. The significant interaction term further implies that the relationship between total expenditure and food budgeting varies by household size. Overall, this model captures the complexities of the data more effectively, enhancing predictive accuracy.

10. Conclusion

In this analysis of factors influencing the budget share of food among Spanish households, key predictors such as total expenditure, age, household size, and town were identified as significant contributors to food budgeting decisions. The model revealed that as total expenditure increases, the budget share allocated to food decreases, highlighting the economic principle of diminishing marginal utility, where households allocate less of their income to food as they become wealthier. Age was positively correlated with the budget share of food, suggesting that older households may prioritize food spending differently, potentially reflecting changes in dietary preferences or family dynamics over time. The introduction of interaction terms, particularly between total expenditure and household size, indicated that the impact of expenditure on food budgeting is nuanced by household composition, emphasizing the need for tailored economic policies. The presence of heteroskedasticity, addressed by using robust standard errors, ensured that coefficient estimates remained reliable. These findings underscore the importance of considering household dynamics in formulating economic policies related to food security and nutrition. Policymakers in Spain should leverage these insights to design interventions that accommodate varying household characteristics, thus enhancing resource allocation and marketing strategies in the food sector. Future research could expand on these interactions or explore non-linear relationships to further elucidate the complexities of food budgeting behaviors in diverse Spanish households.