USDA Nutritional Data Analysis

Created by Christopher Wong: 8-4-22

Last updated: 1-19-25

Data Preprocessing Steps

Read and merged the datasets on a common key, handled missing values, converted non-numeric data into numeric formats, created a cleaned dataset for analysis.

```
# Read the data files into data frames
macroNut <- read.csv("./USDA_Macronutrients.csv", header = TRUE)
microNut <- read.csv("./USDA_Micronutrients.csv", header = TRUE)

# Merge the data frames using the variable "ID"
USDA <- merge(macroNut, microNut, by = "ID")

# Delete commas from the sodium recs, then make data types numeric instead
USDA$Sodium <- gsub(",", "", USDA$Sodium)
USDA$Sodium <- as.numeric(USDA$Sodium)

# Delete commas from the potassium recs, then make data types numeric instead
USDA$Potassium <- gsub(",", "", USDA$Potassium)
USDA$Potassium <- as.numeric(USDA$Potassium)

# Remove records (rows) with missing values in more than 6 attributes (columns)
# get the rows with 4 or less missing values
USDA <- USDA[rowSums(is.na(USDA)) <= 4, ]

# Replace missing values for Sugar, Vitamin E and Vitamin D, with the mean value for the
respective variable
USDA$Sugar[is.na(USDA$Sugar)]       <- mean(USDA$Sugar[!is.na(USDA$Sugar)])
USDA$VitaminE[is.na(USDA$VitaminE)] <- mean(USDA$VitaminE[!is.na(USDA$VitaminE)])
USDA$VitaminD[is.na(USDA$VitaminD)] <- mean(USDA$VitaminD[!is.na(USDA$VitaminD)])

#remove all remaining records with missing values, save to new data frame
USDAclean <- USDA[complete.cases(USDA), ]
```

Sodium Content Analysis

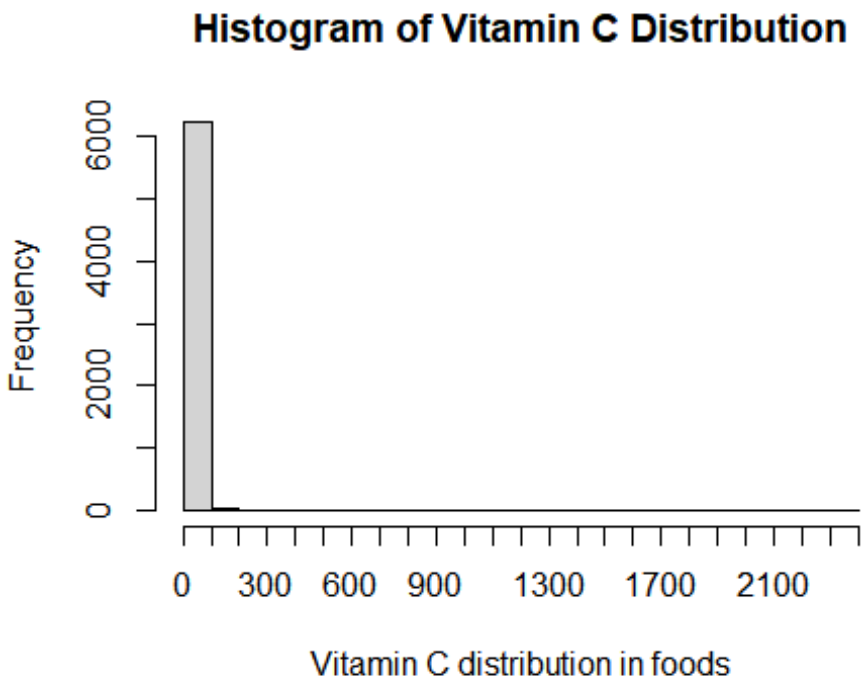
Consuming foods with such high sodium levels can lead to health issues such as hypertension, heart disease, and stroke. It is important to monitor and limit the intake of these foods, especially for individuals with health conditions that require low sodium diets. Below is a list of the top 20 foods and ingredients with the highest sodium levels.

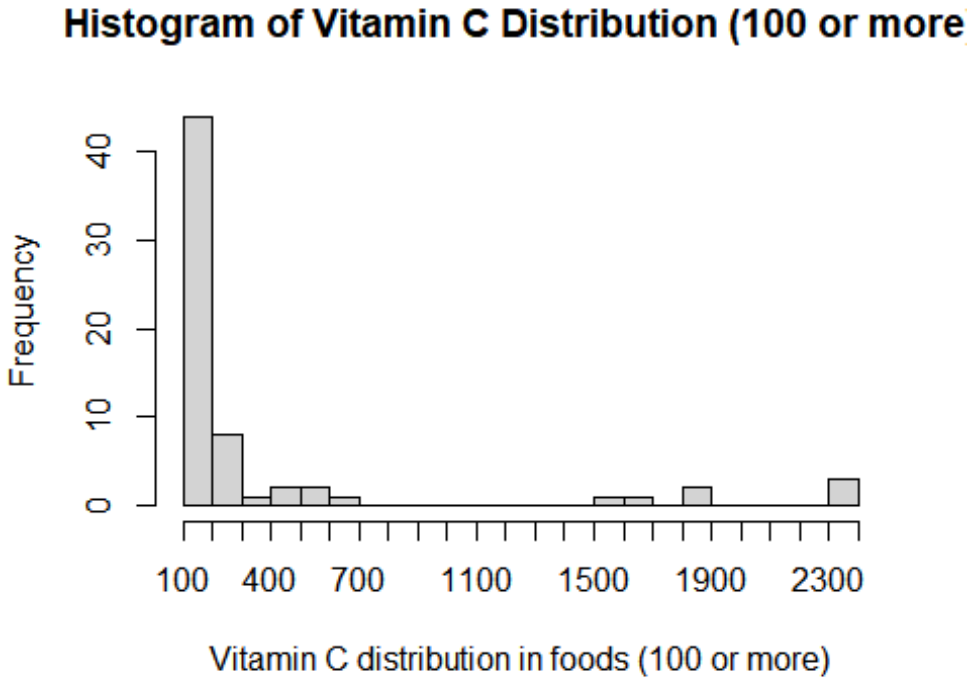
```
##         ID                                                Description Calories
## 265   2047                                                 SALT,TABLE        0
## 5324 18372                                LEAVENING AGENTS,BAKING SODA        0
## 5698 19225                             DESSERTS,RENNIN,TABLETS,UNSWTND       84
```

```
## 922   6075                          SOUP,BF BROTH OR BOUILLON,PDR,DRY      213
## 923   6076                                  SOUP,BEEF BROTH,CUBED,DRY      170
## 926   6081                                SOUP,CHICK BROTH CUBES,DRY       198
## 925   6080                           SOUP,CHICK BROTH OR BOUILLON,DRY      267
## 1303  6979                                                ADOBO FRESCO      271
## 938   6115                                             GRAVY,AU JUS,DRY      313
## 5321 18369 LEAVENING AGENTS,BAKING PDR,DOUBLE-ACTING,NA AL SULFATE       53
## 7024 43497                                       JELLYFISH,DRIED,SALTED       36
## 924   6077                                       SOUP,BF NOODLE,DRY,MIX      324
## 930   6094                                           SOUP,ONION,DRY,MIX      293
## 5322 18370  LEAVENING AGENTS,BAKING PDR,DOUBLE-ACTING,STRAIGHT PO4       51
## 977   6179                                    SAUCE,FISH,READY-TO-SERVE       35
## 4160 15018                                    COD,ATLANTIC,DRIED&SALTED      290
## 932   6099                                      SOUP,TOMATO VEG,DRY,MIX      325
## 945   6122                                       GRAVY,MUSHROOM,DRY,PDR      328
## 950   6127                                        GRAVY,UNSPEC TYPE,DRY      344
## 4515 16125                       SOY SAU MADE FROM HYDROLYZED VEG PROT       40
##      Protein TotalFat Carbohydrate Sodium Cholesterol     Sugar Calcium  Iron
## 265     0.00     0.00         0.00  38758           0  0.000000      24  0.33
## 5324    0.00     0.00         0.00  27360           0  0.000000       0  0.00
## 5698    1.00     0.10        19.80  26050           0  8.229355    3733  7.07
## 922    15.97     8.89        17.40  26000          10 16.710000      60  1.00
## 923    17.30     4.00        16.10  24000           4 14.510000      60  2.23
## 926    14.60     4.70        23.50  24000          13  0.000000     190  1.87
## 925    16.66    13.88        18.01  23875          13 17.360000     187  1.03
## 1303    2.00    20.90        18.60  17152           0  2.030000     123  3.20
## 938     9.20     9.63        47.49  11588           4  8.229355     140  9.30
## 5321    0.00     0.00        27.70  10600           0  0.000000    5876 11.02
## 7024    5.50     1.40         0.00   9690           5  0.000000       2  2.27
## 924    17.93     6.39        48.64   8408          13  5.100000      48  2.68
## 930     7.48     0.34        65.07   8031           0  4.650000     143  1.25
## 5322    0.10     0.00        24.10   7893           0  0.000000    7364 11.27
## 977     5.06     0.01         3.64   7851           0  3.640000      43  0.78
## 4160   62.82     2.37         0.00   7027         152  0.000000     160  2.50
## 932    11.73     5.09        59.90   6722           3  3.770000      46  3.71
## 945    10.00     4.00        64.66   6580           3  3.260000     230  1.00
## 950    13.00     8.00        58.00   5730           4  8.229355     150  1.00
## 4515    2.43     0.08         7.73   5689           0  1.700000       5  1.49
```

Vitamin C distribution in foods

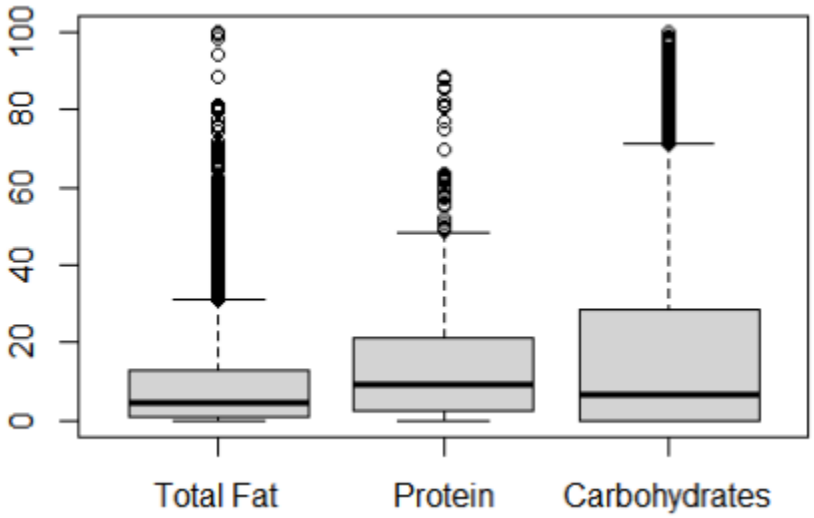
Vitamin C is an essential nutrient that plays a key role in maintaining overall health and immunity. Below are histograms showing the distribution of Vitamin C content in foods. We see that the vast majority of foods have low to moderate levels of Vitamin C. For a more detailed view, we also created a histogram for foods with Vitamin C levels of 100 or more.





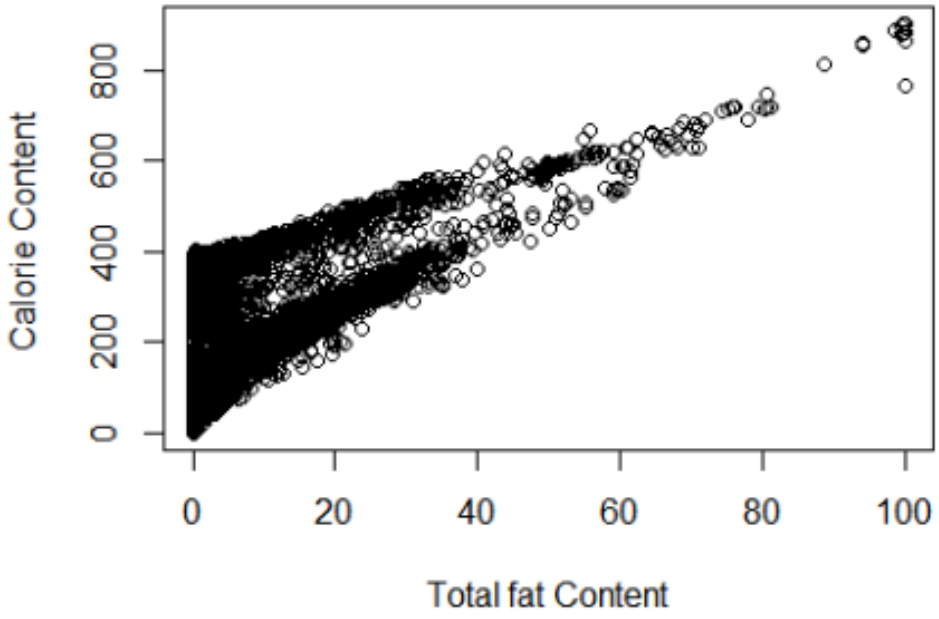
Distribution of values for TotalFat, Protein and Carbohydrate

The boxplot below visualizes the distribution of Total Fat, Protein, and Carbohydrates in the dataset. This visualization aids in comparing the central tendencies and variability of these macronutrients, identifying outliers, and gaining insights into the nutritional composition of the foods in the dataset.



Relationship between a food's TotalFat content and its Calorie content

The graph below indicates that there is a positive relationship between Total Fat content and Calorie content in foods. This means that foods with higher fat content tend to have higher calorie content. This relationship is expected because fat is a dense source of energy, providing 9 calories per gram, compared to 4 calories per gram for protein and carbohydrates. Therefore, as the fat content in foods increases, the overall calorie content also increases.



Adding high nutrient indicators and foods with high sodium and fat

The analysis reveals that there are 1,827 foods in the dataset that have both high sodium and high fat content. This indicates a significant number of foods that may pose health risks due to their high levels of these two nutrients.

By adding binary indicators, we provide a clear way to identify and categorize foods based on their nutrient levels. This can be useful for further nutritional analysis and dietary planning.

```
# Make new columns and check if value is greater than the avg. if true: set it equal to 1
USDAclean$HSodium   <- 0
USDAclean$HSodium[USDAclean$Sodium     > mean(USDAclean$Sodium)]   <- 1

USDAclean$HCalories <- 0
USDAclean$HCalories[USDAclean$Calories > mean(USDAclean$Calories)] <- 1

USDAclean$HProtein  <- 0
USDAclean$HProtein[USDAclean$Protein   > mean(USDAclean$Protein)]  <- 1

USDAclean$HSugar    <- 0
USDAclean$HSugar[USDAclean$Sodium      > mean(USDAclean$Sodium)]   <- 1

USDAclean$HFat      <- 0
USDAclean$HFat[USDAclean$Sodium        > mean(USDAclean$Sodium)]   <- 1

# Number of foods have both high sodium and high fat
sum(apply(USDAclean[c("HSodium", "HFat")], 1, function(x) all(x == 1)))
```

```
## [1] 1827
```

```
# 1827 foods have both high sodium and high fat
```

Average amount of iron for high and low protein foods

Iron is an essential nutrient that plays a crucial role in oxygen transport and energy production. Understanding the relationship between protein and iron content in foods can help in creating balanced diets that support overall health and well-being.

Low Protein Foods: The average amount of iron is approximately 2.70 mg.

High Protein Foods: The average amount of iron is approximately 3.07 mg.

High protein foods tend to have a slightly higher average iron content compared to low protein foods. This information is important for nutritional analysis and dietary planning, helping to ensure adequate intake of both protein and iron.

```
# average amount of iron, for high and low protein foods
meanIron.lowProtein  <- mean(USDAclean$Iron[USDAclean$HProtein == 0])
meanIron.lowProtein
```

```
## [1] 2.696634
```

```
meanIron.highProtein <- mean(USDAclean$Iron[USDAclean$HProtein == 1])
meanIron.highProtein
```

```
## [1] 3.069541
```

Custom HealthCheck program to detect unhealthy foods

This section implements a custom HealthCheck program to classify foods as either “Pass” or “Fail” based on their sodium, sugar, and fat content. The results show that 1,827 foods fail the HealthCheck, highlighting the importance of monitoring and improving the nutritional quality of foods to support better health outcomes.

```
HealthCheck <- function(sodium, sugar, fat) {
  if      (sodium == 0) return("Pass")

  else if (sugar == 0)  return("Pass")
  else if (fat == 0)    return("Pass")

  else                  return("Fail") }
#input each type of food into the function
USDAclean$HealthCheck <- apply(USDAclean[, c("HSodium", "HSugar", "HFat")], 1,
                function(x) HealthCheck(x["HSodium"], x["HSugar"], x["HFat"]))
#number of failed foods
length(which(USDAclean$HealthCheck == "Fail"))
```

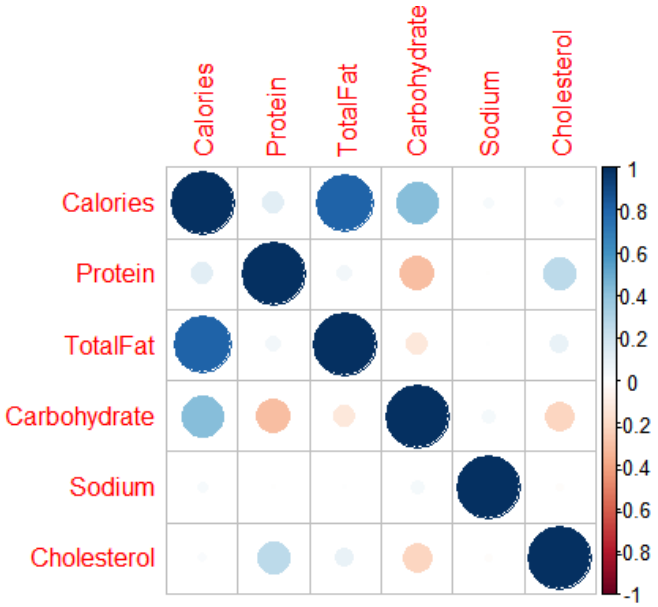
```
## [1] 1827
```

Correlation among Calories, Protein, Total Fat, Carbohydrate, Sodium and Cholesterol

The plot below provides a visual representation of the correlation between Calories, Protein, Total Fat, Carbohydrate, Sodium, and Cholesterol.

- Positive Correlations: Strong positive correlations are indicated by darker colors and larger circles. For example, Calories show a strong positive correlations with Total Fat.
- Negative Correlations: Negative correlations, are indicated by different colors and smaller circles.

  The size and color intensity of the circles indicate the strength of the correlations, with larger and darker circles representing stronger correlations.



Statistical significance of the correlation between Calories & Total Fat

The correlation between Calories and Total Fat is statistically significant (p-value < 2.2e-16), indicating a strong relationship between these two variables.

```
# Find the pvalue using the function
cor.test(USDAclean$Calories, USDAclean$TotalFat)
##  Pearson's product-moment correlation
##
## data:  USDAclean$Calories and USDAclean$TotalFat
## t = 107.58, df = 6308, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7956139 0.8130305
## sample estimates:
##      cor
## 0.804495

# Set alpha = 0.05.
# Now we compare it to the pvalue <  2.2 E-16
# Since the pvalue is less than alpha, the correlation between Calories and Total Fat
statistically significant.
```

Linear Regression Model for Calories based on Protein, Total Fat, Carbohydrate, Sodium, and Cholesterol

This model is used to predict Calories based on the independent variables: Protein, Total Fat, Carbohydrate, Sodium, and Cholesterol. This analysis helps in understanding the contribution of each macronutrient to the calorie content of foods. Sodium has a p-value of 0.055, so it does not significantly contribute to the prediction of Calories.

```
nutLm <- lm (Calories ~ Protein + TotalFat + Carbohydrate + Sodium + Cholesterol,
             data = USDAclean)
summary(nutLm)
## Residuals:
##      Min       1Q   Median       3Q      Max
## -191.087   -3.832    0.426    5.147  291.011
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.9882753  0.4832629   8.253  < 2e-16 ***
## Protein      3.9891994  0.0233550 170.807  < 2e-16 ***
## TotalFat     8.7716980  0.0143291 612.158  < 2e-16 ***
## Carbohydrate 3.7432001  0.0091404 409.522  < 2e-16 ***
## Sodium       0.0003383  0.0002189   1.545    0.122
## Cholesterol  0.0110138  0.0019861   5.545 3.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.92 on 6304 degrees of freedom
## Multiple R-squared:  0.9877, Adjusted R-squared:  0.9877
## F-statistic: 1.009e+05 on 5 and 6304 DF,  p-value: < 2.2e-16
```

```r
# Sodium has a p-value of 0.05500, making it a non-significant variable.
```

New model using only significant independent variables.

```r
# Create a new model by using only the significant independent variables.
sigVarLm <- (lm(Calories ~ Protein + TotalFat + Carbohydrate + Cholesterol,
                data = USDAclean))
summary(sigVarLm)
```

```
##
## Call:
## lm(formula = Calories ~ Protein + TotalFat + Carbohydrate + Cholesterol,
##     data = USDAclean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -191.220   -3.787    0.464    5.104  290.922
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.077907   0.479822   8.499  < 2e-16 ***
## Protein      3.989679   0.023355 170.824  < 2e-16 ***
## TotalFat     8.771904   0.014330 612.131  < 2e-16 ***
## Carbohydrate 3.743859   0.009131 409.996  < 2e-16 ***
## Cholesterol  0.010980   0.001986   5.528 3.36e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.93 on 6305 degrees of freedom
## Multiple R-squared:  0.9877, Adjusted R-squared:  0.9876
## F-statistic: 1.261e+05 on 4 and 6305 DF,  p-value: < 2.2e-16
```

Using the new model to predict number of Calories for a new product.

The new model predicts that a product with 0.1g Protein, 37g Total Fat, 400g Carbohydrate, and 75mg Cholesterol will have approximately 1827.4 Calories.

```r
# New product data
new_product <- data.frame(Protein = 0.1, TotalFat = 37, Carbohydrate = 400, Cholesterol =
75)

# Predict the value for Calories using the new model
predicted_calories <- predict(sigVarLm, newdata = new_product)

# Display the predicted value
predicted_calories
```

```
##        1
## 1827.405
```

```r
# 1827.405 is the predicted value for Calories
```

If the new product's Carbohydrate count increases by 10000%, how much change will occur on Calories?

The Calories would increase by approximately 1083.33% when the Carbohydrate amount increases from 400 to 40000. This large increase is due to the direct proportional relationship between Carbohydrate and Calories as indicated by the linear regression model.

```
# Extract the coefficient for Carbohydrate from the model
carb_coef <- coef(sigVarLm)["Carbohydrate"]

# Calculate the change in Carbohydrate
carb_change <- 40000 - 400

# Calculate the change in Calories
calories_change <- carb_coef * carb_change

# Calculate the initial predicted Calories for Carbohydrate = 400
initial_calories <- predict(sigVarLm, newdata = data.frame(Protein = 0.1, TotalFat = 37,
Carbohydrate = 400, Cholesterol = 75))

# Calculate the new predicted Calories for Carbohydrate = 40000
new_calories <- predict(sigVarLm, newdata = data.frame(Protein = 0.1, TotalFat = 37,
Carbohydrate = 40000, Cholesterol = 75))

# Calculate the percentage change in Calories
percent_change <- ((new_calories - initial_calories) / initial_calories) * 100

# Display the results
initial_calories
```

```
##        1
## 1827.405
```

```
new_calories
```

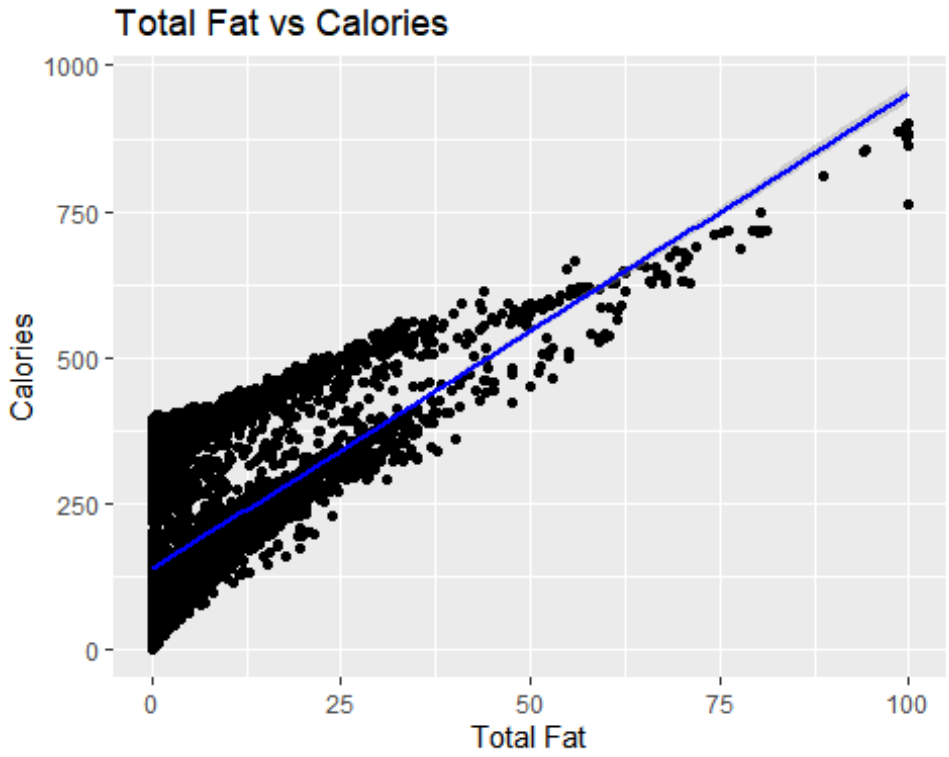
```
##        1
## 150084.2
```

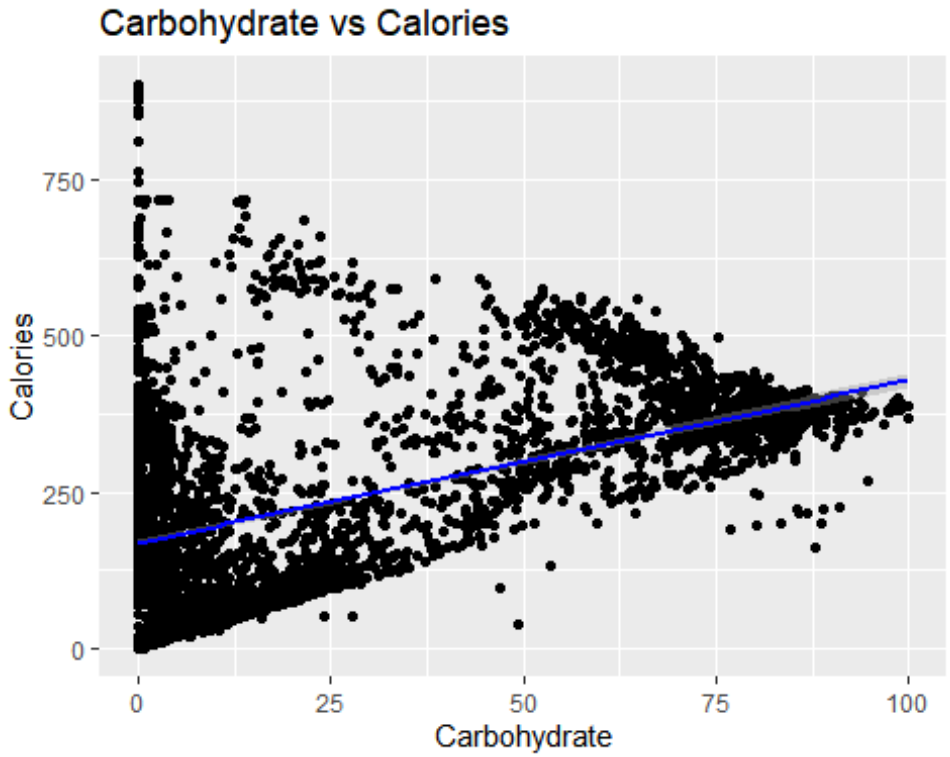
```
percent_change
```

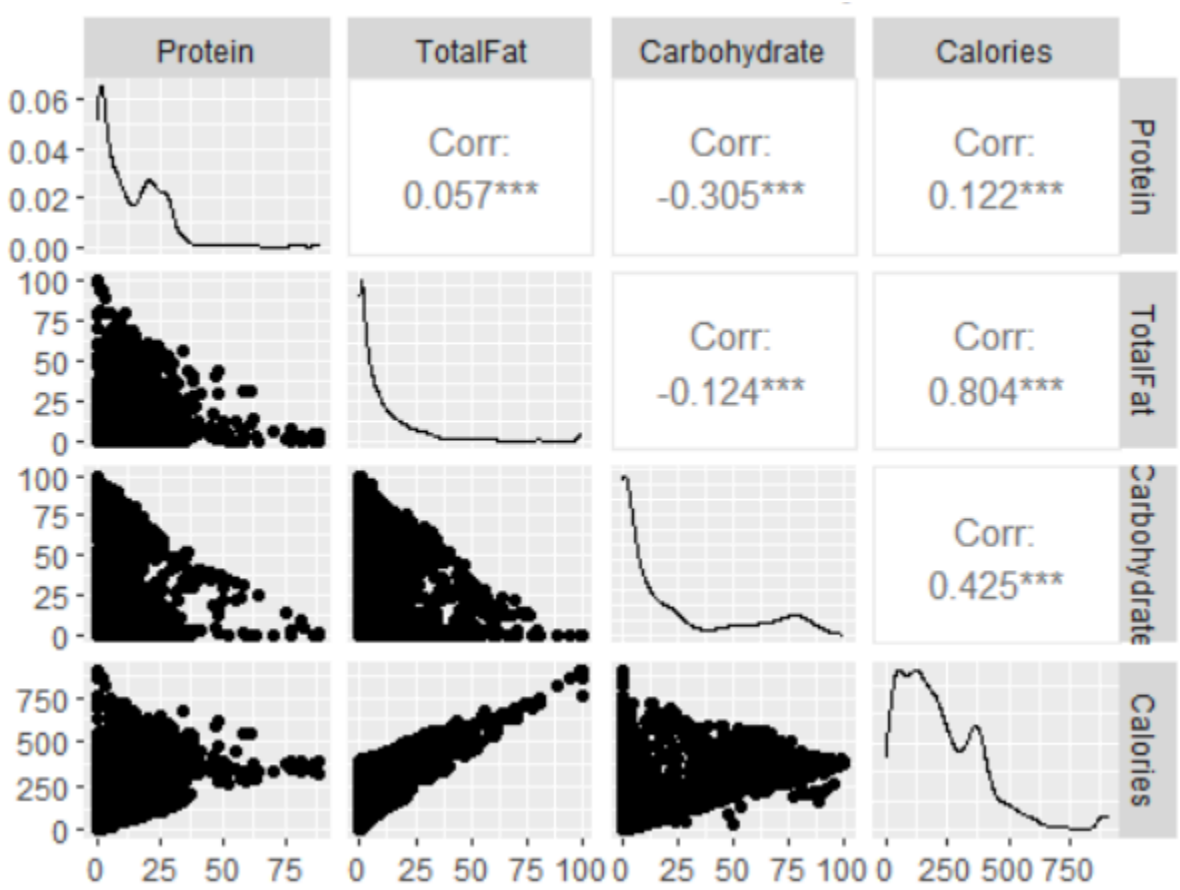
```
##        1
## 8112.973
```

How do Protein, Total Fat, and Carbohydrate content relate to the Calorie content in foods?

The scatter plots and pair plot below will help us understand this relationship. The linear regression lines in the scatter plots will indicate the direction and strength of the relationships. The pair plot will provide a comprehensive view of the relationships between all four attributes.




Carbohydrate vs Calories
Calories
750
500
250
0
0
25
50
75
100
Carbohydrate


Protein
TotalFat
Carbohydrate
Calories
Corr:
0.057***
Corr:
-0.305***
Corr:
0.122***
Corr:
-0.124***
Corr:
0.804***
Corr:
0.425***
Protein
TotalFat
Carbohydrate
Calories
0.06
0.04
0.02
0.00
100
75
50
25
0
750
500
250
0
0 25 50 75
0 25 50 75 100
0 25 50 75 100
0 250 500 750