2347265_CIA_C2.R

vishnu

2024-01-25

```
#R1

data <- read.csv("UScereal1.csv")
str(data)
```

```
## 'data.frame':    65 obs. of  12 variables:
##  $ Name     : chr  "100% Bran" "All-Bran" "All-Bran with Extra Fiber" "Apple Cinnamon Chee
rios" ...
##  $ mfr      : chr  "N" "K" "K" "G" ...
##  $ calories : num  212 212 100 147 110 ...
##  $ protein  : num  12.12 12.12 8 2.67 2 ...
##  $ fat      : num  3.03 3.03 0 2.67 0 2.67 1.49 0 2.67 NA ...
##  $ sodium   : num  394 788 280 240 125 ...
##  $ fibre    : num  30.3 27.3 28 2 1 ...
##  $ carbo    : num  15.2 21.2 16 14 11 ...
##  $ sugars   : num  18.2 15.2 0 13.3 14 ...
##  $ shelf    : int  3 3 3 1 2 3 1 3 2 1 ...
##  $ potassium: num  848.5 969.7 660 93.3 30 ...
##  $ vitamins : chr  "enriched" "enriched" "enriched" "enriched" ...
```

```
head(data)
```

```
##                       Name mfr calories protein  fat sodium fibre carbo sugars
## 1                  100% Bran   N   212.12   12.12 3.03 393.94 30.30 15.15  18.18
## 2                   All-Bran   K   212.12   12.12 3.03 787.88 27.27 21.21  15.15
## 3 All-Bran with Extra Fiber   K   100.00    8.00 0.00 280.00 28.00 16.00   0.00
## 4   Apple Cinnamon Cheerios   G   146.67    2.67 2.67 240.00  2.00 14.00  13.33
## 5                Apple Jacks   K   110.00    2.00 0.00 125.00  1.00 11.00  14.00
## 6                    Basic 4   G   173.33    4.00 2.67 280.00  2.67 24.00  10.67
##   shelf potassium vitamins
## 1     3    848.48 enriched
## 2     3    969.70 enriched
## 3     3    660.00 enriched
## 4     1     93.33 enriched
## 5     2     30.00 enriched
## 6     3    133.33 enriched
```

```
summary(data)
```

```
##      Name               mfr              calories        protein
##  Length:65          Length:65          Min.   : 50.0   Min.   : 0.750
##  Class :character   Class :character   1st Qu.:110.0   1st Qu.: 2.000
##  Mode  :character   Mode  :character   Median :137.2   Median : 3.000
##                                        Mean   :149.6   Mean   : 3.726
##                                        3rd Qu.:179.1   3rd Qu.: 4.480
##                                        Max.   :440.0   Max.   :12.120
##                                        NA's   :1       NA's   :1
##       fat           sodium          fibre            carbo
##  Min.   :0.00   Min.   :  0.0   Min.   : 0.000   Min.   :10.53
##  1st Qu.:0.00   1st Qu.:180.0   1st Qu.: 0.000   1st Qu.:14.92
##  Median :1.00   Median :235.4   Median : 2.000   Median :18.67
##  Mean   :1.42   Mean   :238.6   Mean   : 3.871   Mean   :20.01
##  3rd Qu.:2.00   3rd Qu.:290.0   3rd Qu.: 4.480   3rd Qu.:22.39
##  Max.   :9.09   Max.   :787.9   Max.   :30.300   Max.   :68.00
##  NA's   :1      NA's   :1                        NA's   :1
##      sugars          shelf         potassium        vitamins
##  Min.   : 0.00   Min.   :1.000   Min.   : 15.00   Length:65
##  1st Qu.: 3.75   1st Qu.:1.000   1st Qu.: 45.00   Class :character
##  Median :12.00   Median :2.000   Median : 94.96   Mode  :character
##  Mean   :10.07   Mean   :2.169   Mean   :158.69
##  3rd Qu.:14.00   3rd Qu.:3.000   3rd Qu.:220.00
##  Max.   :20.90   Max.   :3.000   Max.   :969.70
##  NA's   :1                       NA's   :1
```

```
data
```

```
##                                   Name mfr calories protein  fat sodium fibre
## 1                            100% Bran   N   212.12   12.12 3.03 393.94 30.30
## 2                             All-Bran   K   212.12   12.12 3.03 787.88 27.27
## 3             All-Bran with Extra Fiber   K   100.00    8.00 0.00 280.00 28.00
## 4               Apple Cinnamon Cheerios   G   146.67    2.67 2.67 240.00  2.00
## 5                           Apple Jacks   K   110.00    2.00 0.00 125.00  1.00
## 6                                Basic 4   G   173.33    4.00 2.67 280.00  2.67
## 7                             Bran Chex   R   134.33    2.99 1.49 298.51  5.97
## 8                           Bran Flakes   P       NA    4.48 0.00 313.43  7.46
## 9                         Cap'n'Crunch   Q   160.00    1.33 2.67 293.33  0.00
## 10                            Cheerios   G    88.00    4.80   NA 232.00  1.60
## 11               Cinnamon Toast Crunch   G   160.00    1.33 4.00 280.00  0.00
## 12                            Clusters   G   220.00    6.00 4.00 280.00  4.00
## 13                          Cocoa Puffs   G   110.00    1.00 1.00 180.00  0.00
## 14                            Corn Chex   R   110.00    2.00 0.00 280.00  0.00
## 15                          Corn Flakes   K   100.00    2.00 0.00 290.00  1.00
## 16                            Corn Pops   K   110.00      NA 0.00  90.00  1.00
## 17                        Count Chocula   G   110.00    1.00 1.00 180.00  0.00
## 18                   Cracklin' Oat Bran   K   220.00    6.00 6.00 280.00  8.00
## 19                              Crispix   K   110.00    2.00 0.00 220.00  1.00
## 20                Crispy Wheat & Raisins   G   133.33    2.67 1.33     NA  2.67
## 21                          Double Chex   R   133.33    2.67 0.00 253.33  1.33
## 22                          Froot Loops   K   110.00    2.00 1.00 125.00  1.00
## 23                       Frosted Flakes   K   146.67    1.33 0.00 266.67  1.33
## 24                   Frosted Mini-Wheats   K   125.00    3.75 0.00   0.00  3.75
## 25 Fruit & Fibre: Dates Walnuts and Oats   P   179.10    4.48 2.99 238.81  7.46
## 26                           Fruitful Bran   K   179.10    4.48 0.00 358.21  7.46
## 27                        Fruity Pebbles   P   146.67    1.33 1.33 180.00  0.00
## 28                          Golden Crisp   P   113.64    2.27 0.00  51.14  0.00
## 29                        Golden Grahams   G   146.67    1.33 1.33 373.33  0.00
## 30                   Grape Nuts Flakes   P   113.64    3.41 1.14 159.09  3.41
## 31                             Grape-Nuts   P   440.00   12.00 0.00 680.00 12.00
## 32                   Great Grains Pecan   P   363.64    9.09 9.09 227.27  9.09
## 33                     Honey Graham Ohs   Q   120.00    1.00 2.00 220.00  1.00
## 34                   Honey Nut Cheerios   G   146.67    4.00 1.33 333.33  2.00
## 35                             Honey-comb   P    82.71    0.75 0.00 135.34  0.00
## 36               Just Right Fruit & Nut   K   186.67    4.00 1.33 226.67  2.67
## 37                                  Kix   G    73.33    1.33 0.67 173.33  0.00
## 38                                 Life   Q   149.25    5.97 2.99 223.88  2.99
## 39                         Lucky Charms   G   110.00    2.00 1.00 180.00  0.00
## 40                 Mueslix Crispy Blend   K   238.81    4.48 2.99 223.88  4.48
## 41                 Multi-Grain Cheerios   G   100.00    2.00 1.00 220.00  2.00
## 42                     Nut&Honey Crunch   K   179.10    2.99 1.49 283.58  0.00
## 43             Nutri-Grain Almond-Raisin   K   208.96    4.48 2.99 328.36  4.48
## 44                  Oatmeal Raisin Crisp   G   260.00    6.00 4.00 340.00  3.00
## 45                Post Nat. Raisin Bran   P   179.10    4.48 1.49 298.51  8.96
## 46                             Product 19   K   100.00    3.00 0.00 320.00  1.00
## 47                          Puffed Rice   Q    50.00    1.00 0.00   0.00  0.00
## 48                    Quaker Oat Squares   Q   200.00    8.00 2.00 270.00  4.00
## 49                             Raisin Bran   K   160.00    4.00 1.33 280.00  6.67
## 50                       Raisin Nut Bran   G   200.00    6.00 4.00 280.00  5.00
## 51                        Raisin Squares   K   180.00    4.00 0.00   0.00  4.00
## 52                             Rice Chex   R    97.35    0.88 0.00 212.39  0.00
## 53                         Rice Krispies   K   110.00    2.00 0.00 290.00  0.00
## 54                Shredded Wheat 'n'Bran   N   134.33    4.48 0.00   0.00  5.97
```

```
## 55              Shredded Wheat spoon size   N   134.33    4.48 0.00   0.00  4.48
## 56                                 Smacks   K   146.67    2.67 1.33  93.33  1.33
## 57                              Special K   K   110.00    6.00 0.00 230.00  1.00
## 58                     Total Corn Flakes   G   110.00    2.00 1.00 200.00  0.00
## 59                     Total Raisin Bran   G   140.00    3.00 1.00 190.00  4.00
## 60                     Total Whole Grain   G   100.00    3.00 1.00 200.00  3.00
## 61                               Triples   G   146.67    2.67 1.33 333.33  0.00
## 62                                  Trix   G   110.00    1.00 1.00 140.00  0.00
## 63                            Wheat Chex   R   149.25    4.48 1.49 343.28  4.48
## 64                              Wheaties   G   100.00    3.00 1.00 200.00  3.00
## 65                    Wheaties Honey Gold   G   146.67    2.67 1.33 266.67  1.33
##    carbo sugars shelf potassium vitamins
## 1  15.15  18.18     3    848.48 enriched
## 2  21.21  15.15     3    969.70 enriched
## 3  16.00   0.00     3    660.00 enriched
## 4  14.00  13.33     1     93.33 enriched
## 5  11.00  14.00     2     30.00 enriched
## 6  24.00  10.67     3    133.33 enriched
## 7  22.39   8.96     1        NA enriched
## 8  19.40   7.46     3    283.58 enriched
## 9  16.00  16.00     2     46.67 enriched
## 10 13.60   0.80     1     84.00 enriched
## 11    NA  12.00     2     60.00 enriched
## 12 26.00  14.00     3    210.00 enriched
## 13 12.00  13.00     2     55.00 enriched
## 14 22.00   3.00     1     25.00 enriched
## 15 21.00   2.00     1     35.00 enriched
## 16 13.00  12.00     2     20.00 enriched
## 17 12.00  13.00     2     65.00 enriched
## 18 20.00  14.00     3    320.00 enriched
## 19 21.00   3.00     3     30.00 enriched
## 20 14.67  13.33     3    160.00 enriched
## 21 24.00   6.67     3    106.67 enriched
## 22 11.00  13.00     2     30.00 enriched
## 23 18.67  14.67     1     33.33 enriched
## 24 17.50     NA     2    125.00 enriched
## 25 17.91  14.93     3    298.51 enriched
## 26 20.90  17.91     3    283.58 enriched
## 27 17.33  16.00     2     33.33 enriched
## 28 12.50  17.05     1     45.45 enriched
## 29 20.00  12.00     2     60.00 enriched
## 30 17.05   5.68     3     96.59 enriched
## 31 68.00  12.00     3    360.00 enriched
## 32 39.39  12.12     3    303.03 enriched
## 33 12.00  11.00     2     45.00 enriched
## 34 15.33  13.33     1    120.00 enriched
## 35 10.53   8.27     1     26.32 enriched
## 36 26.67  12.00     3    126.67     100%
## 37 14.00   2.00     2     26.67 enriched
## 38 17.91   8.96     2    141.79 enriched
## 39 12.00  12.00     2     55.00 enriched
## 40 25.37  19.40     3    238.81 enriched
## 41 15.00   6.00     1     90.00 enriched
## 42 22.39  13.43     2     59.70 enriched
## 43 31.34  10.45     3    194.03 enriched
## 44 27.00  20.00     3    240.00 enriched
```

```
## 45 16.42  20.90     3    388.06 enriched
## 46 20.00   3.00     3     45.00     100%
## 47 13.00   0.00     3     15.00     none
## 48 28.00  12.00     3    220.00 enriched
## 49 18.67  16.00     2    320.00 enriched
## 50 21.00  16.00     3    280.00 enriched
## 51 30.00  12.00     3    220.00 enriched
## 52 20.35   1.77     1     26.55 enriched
## 53 22.00   3.00     1     35.00 enriched
## 54 28.36   0.00     1    208.96     none
## 55 29.85   0.00     1    179.10     none
## 56 12.00  20.00     2     53.33 enriched
## 57 16.00   3.00     1     55.00 enriched
## 58 21.00   3.00     3     35.00     100%
## 59 15.00  14.00     3    230.00     100%
## 60 16.00   3.00     3    110.00     100%
## 61 28.00   4.00     3     80.00 enriched
## 62 13.00  12.00     2     25.00 enriched
## 63 25.37   4.48     1    171.64 enriched
## 64 17.00   3.00     1    110.00 enriched
## 65 21.33  10.67     1     80.00 enriched
```

```
result <- aggregate(protein ~ mfr, data = data, FUN = max)
print(result)
```

```
##   mfr protein
## 1   G    6.00
## 2   K   12.12
## 3   N   12.12
## 4   P   12.00
## 5   Q    8.00
## 6   R    4.48
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
#R2

missing_values <- sapply(data, function(x) {
  sum(is.na(x))
})
print(missing_values)
```

```
##      Name       mfr  calories   protein       fat    sodium     fibre     carbo
##         0         0         1         1         1         1         0         1
##    sugars     shelf potassium  vitamins
##         1         0         1         0
```

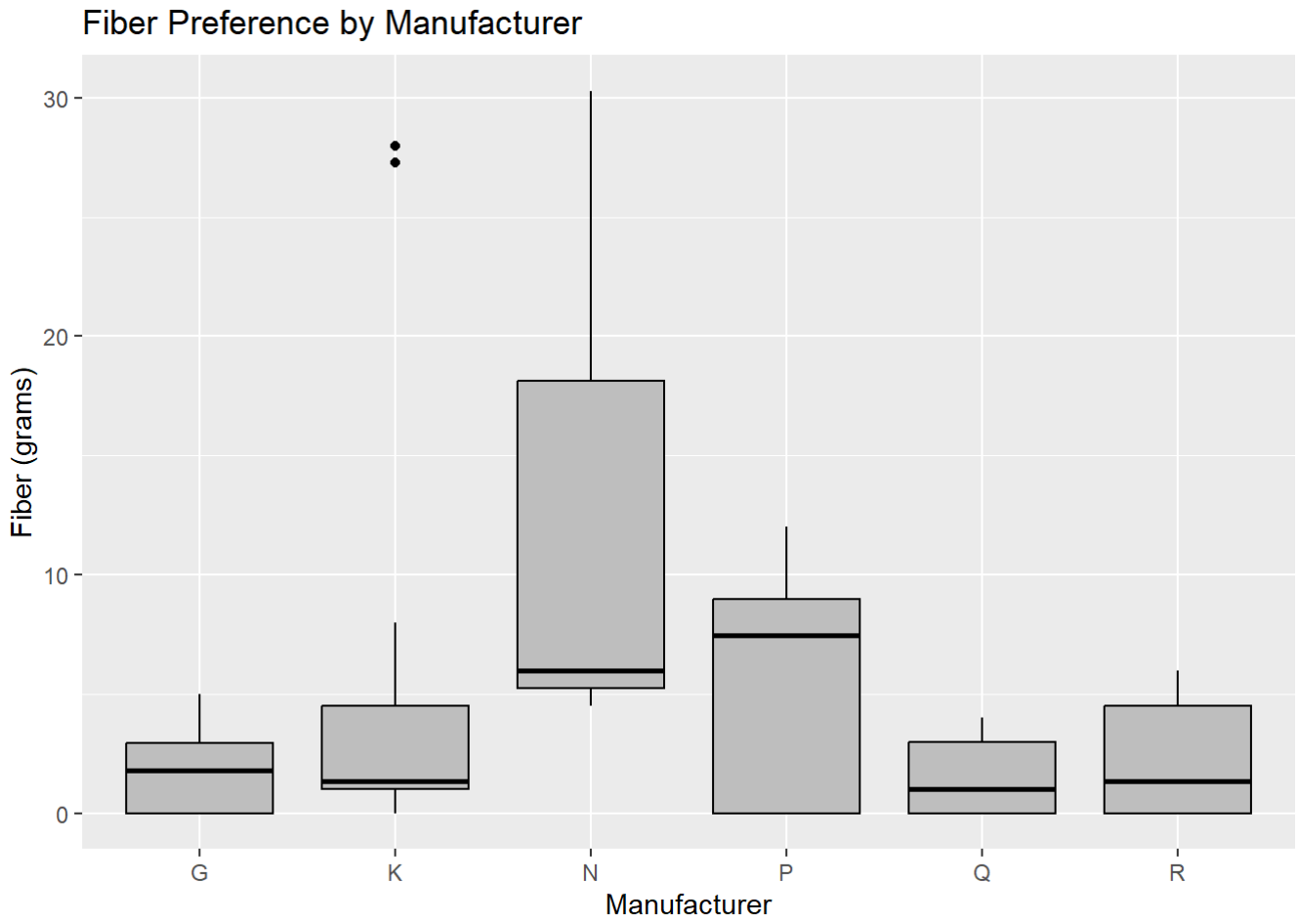
```
print(names(data)[which(missing_values > 0)])
```

```
## [1] "calories"  "protein"   "fat"       "sodium"    "carbo"     "sugars"
## [7] "potassium"
```

```
clean = na.omit(data)

summary(clean)
```

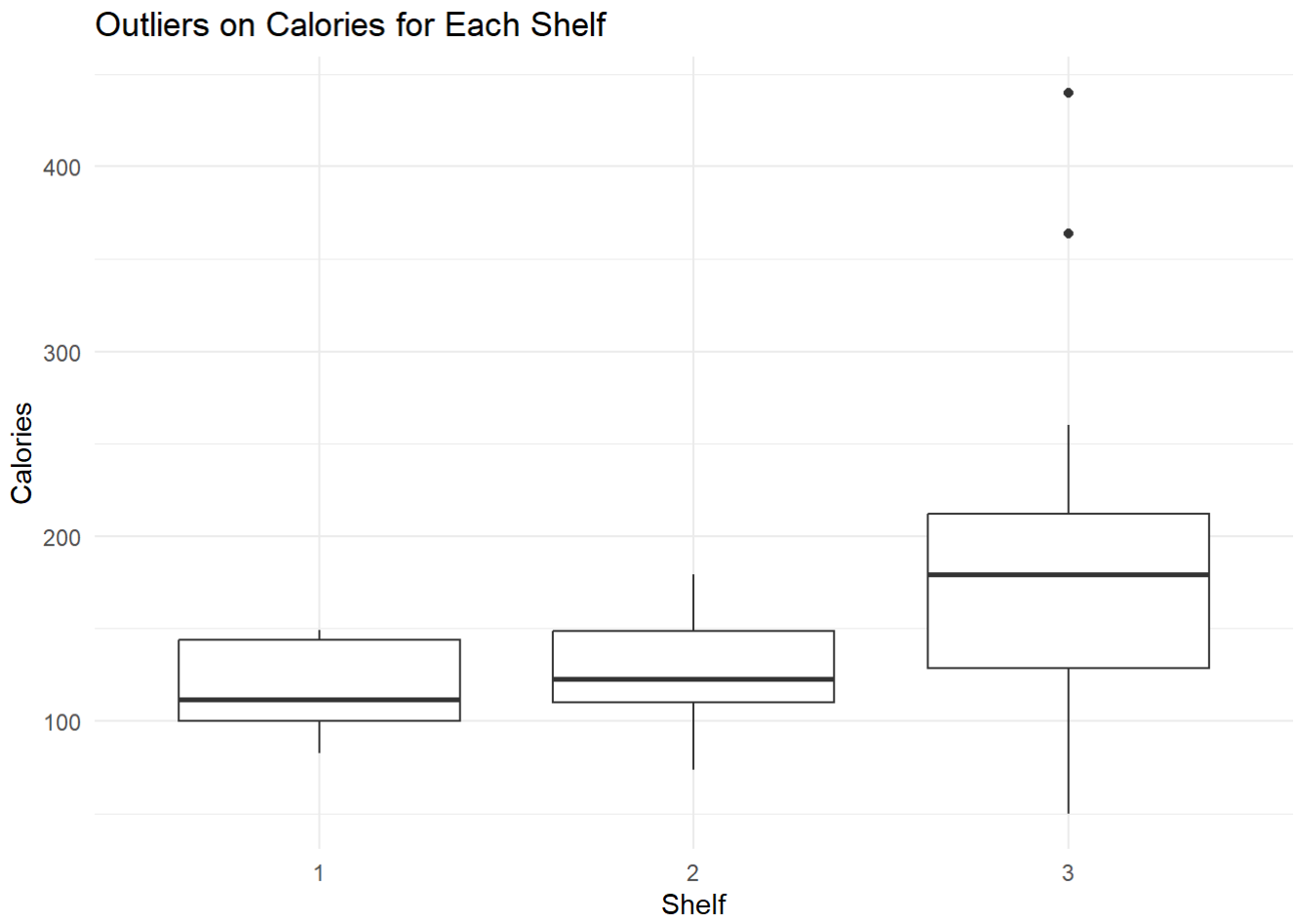
```
##      Name               mfr                calories        protein
##  Length:58          Length:58          Min.   : 50.0   Min.   : 0.750
##  Class :character   Class :character   1st Qu.:110.0   1st Qu.: 2.000
##  Mode  :character   Mode  :character   Median :146.7   Median : 3.000
##                                        Mean   :152.2   Mean   : 3.766
##                                        3rd Qu.:179.1   3rd Qu.: 4.480
##                                        Max.   :440.0   Max.   :12.120
##       fat            sodium          fibre            carbo
##  Min.   :0.000   Min.   :  0.0   Min.   : 0.000   Min.   :10.53
##  1st Qu.:0.000   1st Qu.:180.0   1st Qu.: 0.000   1st Qu.:15.04
##  Median :1.000   Median :234.4   Median : 2.000   Median :19.34
##  Mean   :1.449   Mean   :242.4   Mean   : 3.951   Mean   :20.34
##  3rd Qu.:2.000   3rd Qu.:288.4   3rd Qu.: 4.480   3rd Qu.:23.60
##  Max.   :9.090   Max.   :787.9   Max.   :30.300   Max.   :68.00
##      sugars          shelf        potassium        vitamins
##  Min.   : 0.00   Min.   :1.00   Min.   : 15.00   Length:58
##  1st Qu.: 3.25   1st Qu.:1.00   1st Qu.: 45.00   Class :character
##  Median :12.00   Median :2.00   Median : 94.96   Mode  :character
##  Mean   :10.17   Mean   :2.19   Mean   :162.48
##  3rd Qu.:14.00   3rd Qu.:3.00   3rd Qu.:220.00
##  Max.   :20.90   Max.   :3.00   Max.   :969.70
```

```
#R3


library(ggplot2)

# plot to analyze the spread of Fiber for each Manufacturer
ggplot(data, aes(x = mfr, y = fibre)) + geom_boxplot(fill="grey" , color="black") +  labs(tit
le = "Fiber Preference by Manufacturer",x = "Manufacturer",y = "Fiber (grams)")
```



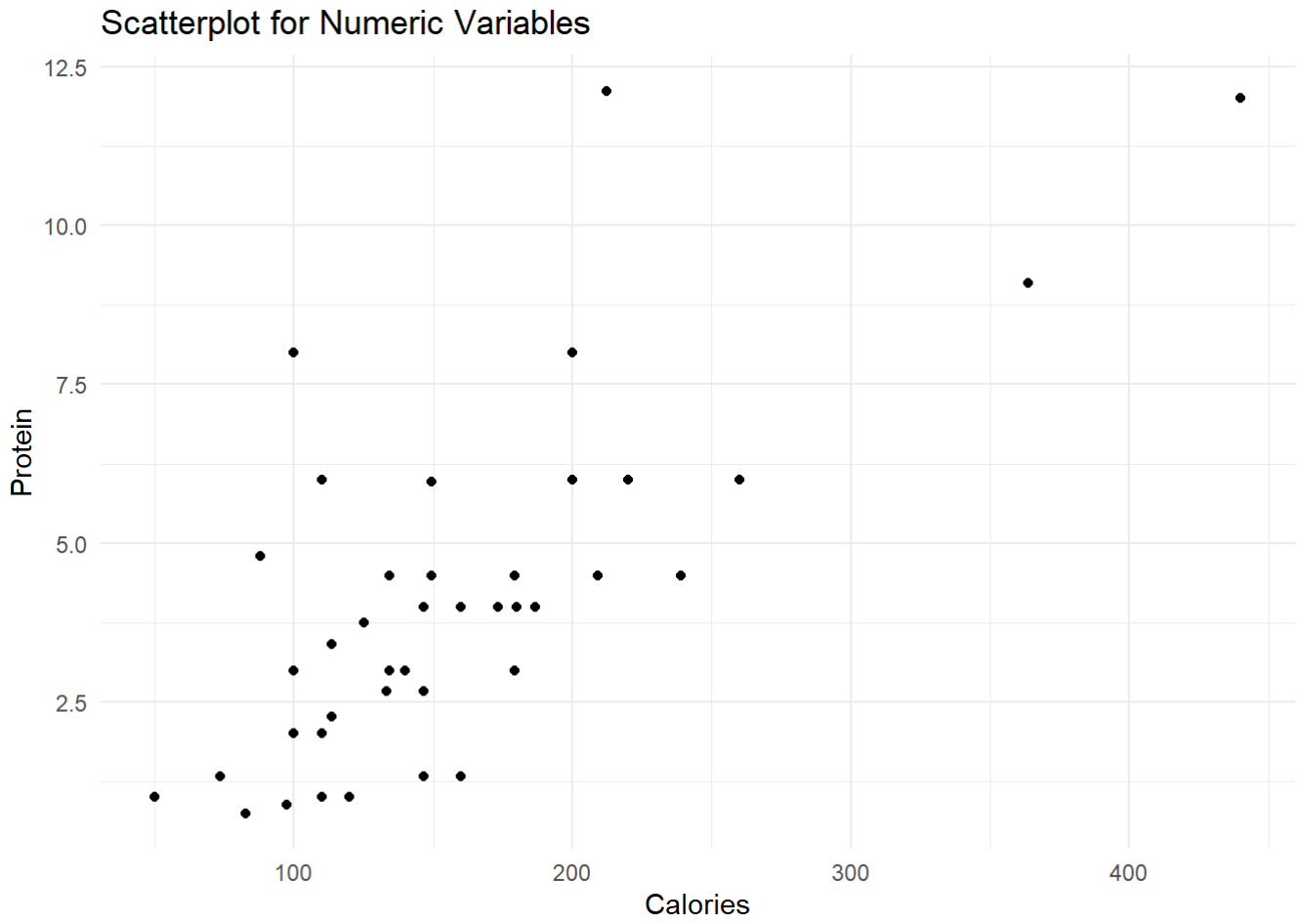

```
#Manufacturer N likes cereals with more fiber compared to the rest.
#This suggests that when it comes to fiber in cereals, Manufacturer N stands out for choosing
options with higher fiber content.

# Outlier on calories for each shelf

# Box plot to find outlines on calories for each shelf
ggplot(data, aes(x = as.factor(shelf), y = calories)) + geom_boxplot() + labs(title = "Outlie
rs on Calories for Each Shelf", x = "Shelf", y = "Calories") + theme_minimal()
```

```
## Warning: Removed 1 rows containing non-finite values (`stat_boxplot()`).
```




```
#Examining the box plot highlighting calorie outliers for each shelf, it becomes evident that shelf 3 stands out with the highest number of outliers in calorie ranges. Following closely, shelf 2 exhibits the next highest frequency of outliers.
#This observation suggests that when it comes to calorie content, shelf 3 contains a notable number of data points deviating from the overall trend, making it the shelf with the most diverse calorie values.s.

# Scatter plot matrix to explore all numeric variables
ggplot(data, aes(x = calories, y = protein)) + geom_point() + labs(title = "Scatterplot for Numeric Variables", x = "Calories", y = "Protein") + theme_minimal()
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```

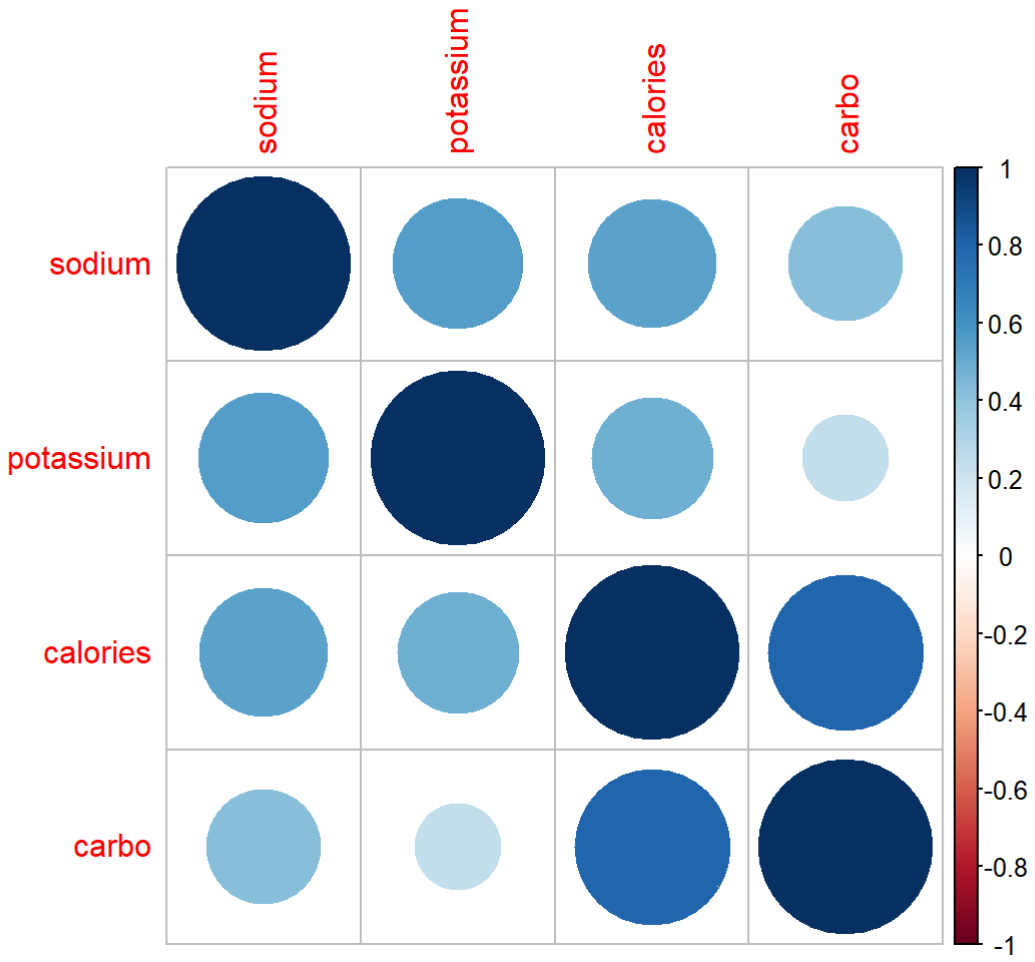



```
#finding
#Examining the scatter plot matrix focusing on calories and protein, a noticeable trend emerg
es. The protein-to-calories ratio appears higher in the 100-200 calorie range, indicating tha
t cereals within this calorie bracket tend to have more protein relative to their calorie con
tent.
#However, as we move into the 300-400 calorie range, the protein-to-calories ratio substantia
lly decreases.
#This finding suggests a potential inverse relationship between calorie content and protein-t
o-calories ratio in cereals, with higher-calorie cereals exhibiting a lower protein-to-calori
es ratio.


#R4

mean_values <- colMeans(data[, c("calories", "protein", "fat", "sodium", "fibre", "carbo", "s
ugars", "potassium")], na.rm = TRUE)
top_four_mean_variables <- names(sort(mean_values, decreasing = TRUE)[1:4])

GreaterMeanFour <- data[, c("Name", top_four_mean_variables)]

correlation_matrix <- cor(data[, top_four_mean_variables], use = "complete.obs")

library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.2
```

```
## corrplot 0.92 loaded
```

```
corrplot(correlation_matrix, method = "circle")
```

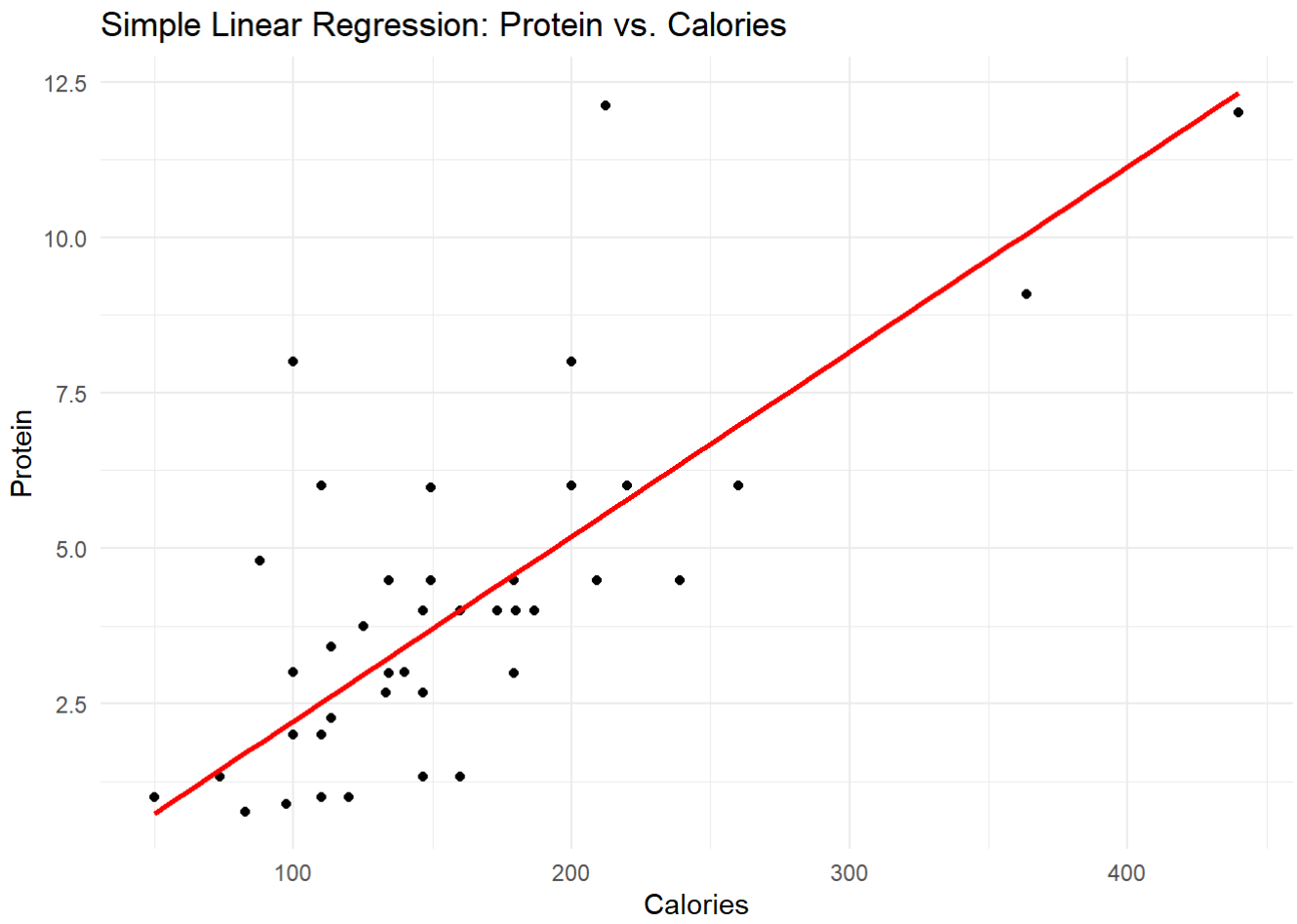



```
model <- lm(protein ~ calories, data = data)

ggplot(data, aes(x = calories, y = protein)) + geom_point() + geom_smooth(method = "lm", se = FALSE, color = "red") + labs(title = "Simple Linear Regression: Protein vs. Calories", x = "Calories", y = "Protein") + theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 2 rows containing non-finite values (`stat_smooth()`).
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```




```
predictions_before <- predict(model, newdata = data.frame(calories = 150))

outliers <- which(model$residuals > quantile(model$residuals, 0.975) | model$residuals < quan
tile(model$residuals, 0.025))
us_cereal_no_outliers <- data[-outliers, ]

model_no_outliers <- lm(protein ~ calories, data = us_cereal_no_outliers)

predictions_after <- predict(model_no_outliers, newdata = data.frame(calories = 150))

print(paste("Prediction Before Removing Outliers:", predictions_before))
```

```
## [1] "Prediction Before Removing Outliers: 3.70583695717782"
```

```
print(paste("Prediction After Removing Outliers:", predictions_after))
```

```
## [1] "Prediction After Removing Outliers: 3.43591615905386"
```