Nutrition\_data visualization

Racheal\_wanyoike

2022-06-26

Obtained data

nutrition\_data <- read.csv ("https://raw.githubusercontent.com/fivethirtyeight/data/master/nutrition-studies/raw\_anonymized\_data.csv")

Next, I cleaned the data by selecting 12 variables that may have an association to cancer.

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.1.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

cancerlinktofoods <- select(nutrition\_data,cancer,BEEFPORKDISHFREQ,DRIEDFRUITFREQ,OTHERFRESHFRUITFREQ,CORNFREQ,SUGARINCOFFEE,SALTFREQ,HOTTEAFREQ,RAWTOMATOESFREQ,MILKFREQ,BUTTERFREQ,OTHERBREADSFREQ, BROCCOLIFREQ)  
cancerlinktofoods

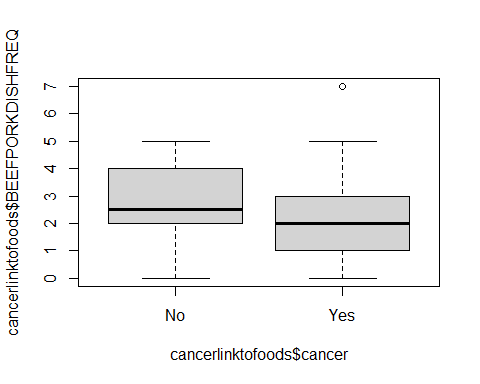
## cancer BEEFPORKDISHFREQ DRIEDFRUITFREQ OTHERFRESHFRUITFREQ CORNFREQ  
## 1 Yes 4 1 1 1  
## 2 No 3 5 5 6  
## 3 Yes 2 4 4 4  
## 4 No 3 7 7 3  
## 5 Yes 3 3 3 3  
## 6 No 0 3 6 6  
## 7 Yes 0 2 3 2  
## 8 Yes 1 5 5 3  
## 9 Yes 3 4 2 2  
## 10 Yes 2 2 2 2  
## 11 Yes 4 6 4 4  
## 12 No 2 1 2 3  
## 13 Yes 0 2 5 4  
## 14 No 2 4 8 5  
## 15 No 4 1 1 1  
## 16 No 5 5 4 3  
## 17 No 5 1 6 4  
## 18 Yes 2 1 2 2  
## 19 Yes 1 2 2 1  
## 20 No 2 2 6 4  
## 21 Yes 4 4 2 2  
## 22 No 2 1 2 3  
## 23 No 2 7 2 3  
## 24 No 4 2 4 2  
## 25 No 1 1 7 1  
## 26 No 4 4 5 3  
## 27 No 2 7 4 3  
## 28 No 3 4 4 3  
## 29 Yes 2 5 7 3  
## 30 No 1 2 5 2  
## 31 Yes 5 1 4 4  
## 32 Yes 3 2 4 3  
## 33 No 3 1 4 2  
## 34 Yes 2 6 1 2  
## 35 No 3 4 7 3  
## 36 No 2 2 4 4  
## 37 Yes 5 2 7 3  
## 38 Yes 7 2 3 4  
## 39 No 1 1 4 7  
## 40 No 1 1 2 2  
## 41 No 4 1 1 1  
## 42 Yes 2 1 3 6  
## 43 Yes 3 1 5 2  
## 44 Yes 3 2 2 1  
## 45 No 4 4 3 5  
## 46 Yes 1 2 2 2  
## 47 No 4 2 6 4  
## 48 Yes 2 2 2 4  
## 49 Yes 1 1 4 4  
## 50 Yes 1 2 8 4  
## 51 Yes 1 2 6 2  
## 52 Yes 3 4 3 4  
## 53 Yes 3 3 2 5  
## 54 No 1 3 3 2  
## SUGARINCOFFEE SALTFREQ HOTTEAFREQ RAWTOMATOESFREQ MILKFREQ BUTTERFREQ  
## 1 1 1 3 4 1 7  
## 2 1 5 1 4 1 3  
## 3 1 5 2 6 5 7  
## 4 1 5 9 3 2 2  
## 5 1 6 8 5 1 3  
## 6 1 2 6 8 1 3  
## 7 2 7 1 4 1 4  
## 8 1 8 2 5 3 7  
## 9 1 6 6 6 1 3  
## 10 1 1 9 6 1 3  
## 11 1 3 1 7 1 4  
## 12 2 3 4 2 1 3  
## 13 1 1 7 7 1 1  
## 14 1 3 9 7 2 9  
## 15 1 1 1 1 1 1  
## 16 1 9 7 3 2 9  
## 17 2 8 5 4 5 7  
## 18 1 2 7 2 1 7  
## 19 1 1 1 2 1 1  
## 20 1 6 7 4 9 3  
## 21 1 6 2 2 1 2  
## 22 2 2 4 2 8 1  
## 23 1 3 6 8 1 4  
## 24 1 4 2 4 5 2  
## 25 1 6 1 6 1 1  
## 26 1 4 6 6 5 5  
## 27 1 2 4 3 1 1  
## 28 1 9 6 5 2 7  
## 29 2 2 7 4 6 9  
## 30 1 6 8 6 1 4  
## 31 2 2 1 5 2 7  
## 32 2 4 7 7 2 6  
## 33 1 9 7 6 2 8  
## 34 1 1 1 1 1 1  
## 35 1 2 9 5 2 6  
## 36 1 9 1 4 8 7  
## 37 1 1 9 7 1 6  
## 38 1 5 4 4 6 9  
## 39 2 1 1 1 1 1  
## 40 1 4 2 3 1 4  
## 41 1 9 2 2 7 9  
## 42 1 2 1 1 9 9  
## 43 1 2 1 4 5 6  
## 44 1 1 2 8 1 5  
## 45 1 5 9 5 3 5  
## 46 1 7 1 6 1 2  
## 47 1 9 8 6 1 5  
## 48 1 9 7 9 1 4  
## 49 1 6 1 6 2 6  
## 50 1 7 7 4 1 2  
## 51 1 1 7 5 6 5  
## 52 2 2 4 7 9 5  
## 53 1 1 9 7 1 1  
## 54 1 2 8 2 4 4  
## OTHERBREADSFREQ BROCCOLIFREQ  
## 1 1 5  
## 2 2 4  
## 3 5 4  
## 4 3 4  
## 5 3 4  
## 6 8 5  
## 7 4 4  
## 8 5 5  
## 9 9 3  
## 10 9 3  
## 11 4 5  
## 12 8 4  
## 13 8 5  
## 14 7 6  
## 15 9 6  
## 16 7 5  
## 17 7 3  
## 18 3 6  
## 19 4 9  
## 20 9 3  
## 21 8 2  
## 22 2 6  
## 23 5 6  
## 24 5 3  
## 25 7 5  
## 26 5 4  
## 27 7 3  
## 28 7 4  
## 29 1 5  
## 30 2 2  
## 31 5 3  
## 32 7 5  
## 33 7 4  
## 34 1 2  
## 35 5 5  
## 36 7 4  
## 37 5 6  
## 38 9 7  
## 39 8 2  
## 40 5 4  
## 41 1 5  
## 42 9 1  
## 43 8 4  
## 44 7 6  
## 45 5 4  
## 46 9 5  
## 47 5 5  
## 48 4 7  
## 49 4 1  
## 50 4 6  
## 51 7 4  
## 52 4 4  
## 53 9 5  
## 54 5 4

To understand the numeric variables and in my cancerlinktofoods dataset, I plotted a summary box plot.

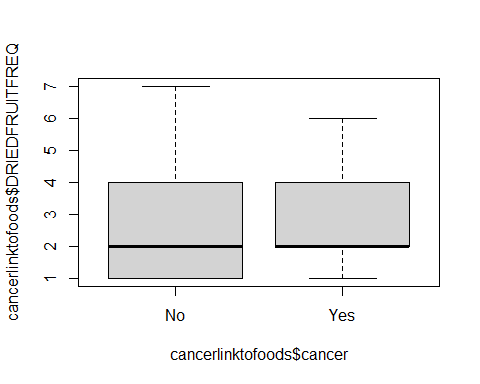
cancerlinktofoods$cancer = factor(cancerlinktofoods$cancer)  
summary(cancerlinktofoods)

## cancer BEEFPORKDISHFREQ DRIEDFRUITFREQ OTHERFRESHFRUITFREQ CORNFREQ   
## No :26 Min. :0.000 Min. :1.000 Min. :1.000 Min. :1.000   
## Yes:28 1st Qu.:1.250 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:2.000   
## Median :2.000 Median :2.000 Median :4.000 Median :3.000   
## Mean :2.556 Mean :2.778 Mean :3.889 Mean :3.111   
## 3rd Qu.:3.750 3rd Qu.:4.000 3rd Qu.:5.000 3rd Qu.:4.000   
## Max. :7.000 Max. :7.000 Max. :8.000 Max. :7.000   
## SUGARINCOFFEE SALTFREQ HOTTEAFREQ RAWTOMATOESFREQ  
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:1.250 1st Qu.:3.000   
## Median :1.000 Median :4.000 Median :4.500 Median :5.000   
## Mean :1.167 Mean :4.222 Mean :4.648 Mean :4.648   
## 3rd Qu.:1.000 3rd Qu.:6.000 3rd Qu.:7.000 3rd Qu.:6.000   
## Max. :2.000 Max. :9.000 Max. :9.000 Max. :9.000   
## MILKFREQ BUTTERFREQ OTHERBREADSFREQ BROCCOLIFREQ   
## Min. :1.000 Min. :1.000 Min. :1.00 Min. :1.00   
## 1st Qu.:1.000 1st Qu.:2.250 1st Qu.:4.00 1st Qu.:4.00   
## Median :1.000 Median :4.000 Median :5.00 Median :4.00   
## Mean :2.759 Mean :4.556 Mean :5.63 Mean :4.37   
## 3rd Qu.:4.750 3rd Qu.:7.000 3rd Qu.:7.75 3rd Qu.:5.00   
## Max. :9.000 Max. :9.000 Max. :9.00 Max. :9.00

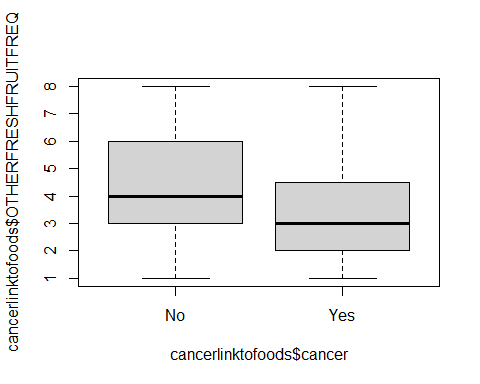
boxplot(cancerlinktofoods$BEEFPORKDISHFREQ~cancerlinktofoods$cancer)



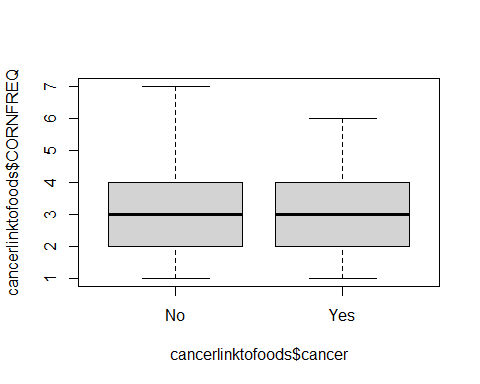
boxplot(cancerlinktofoods$DRIEDFRUITFREQ~cancerlinktofoods$cancer)



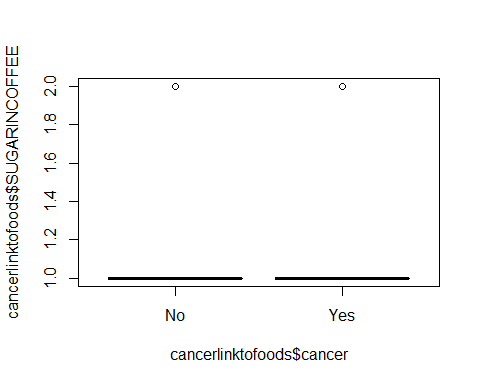
boxplot(cancerlinktofoods$OTHERFRESHFRUITFREQ~cancerlinktofoods$cancer)



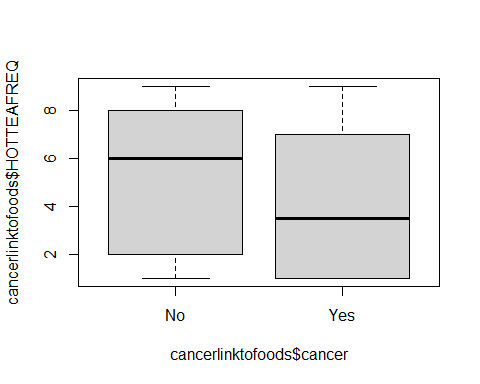
boxplot(cancerlinktofoods$CORNFREQ~cancerlinktofoods$cancer)



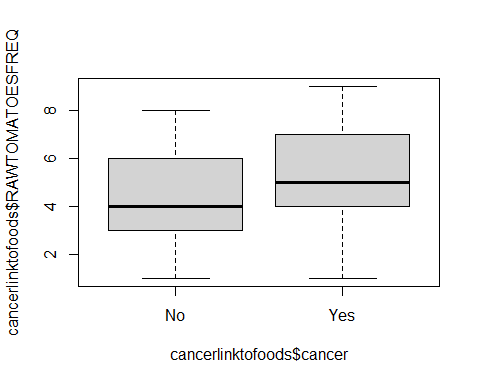
boxplot(cancerlinktofoods$SUGARINCOFFEE~cancerlinktofoods$cancer)



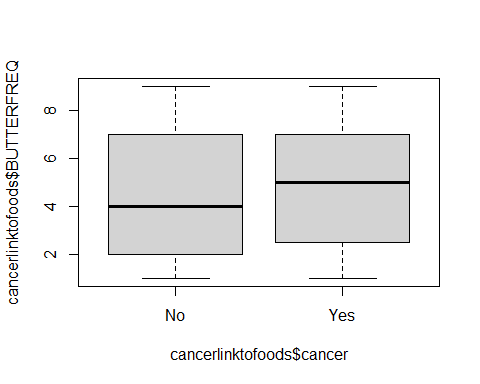
boxplot(cancerlinktofoods$HOTTEAFREQ~cancerlinktofoods$cancer)



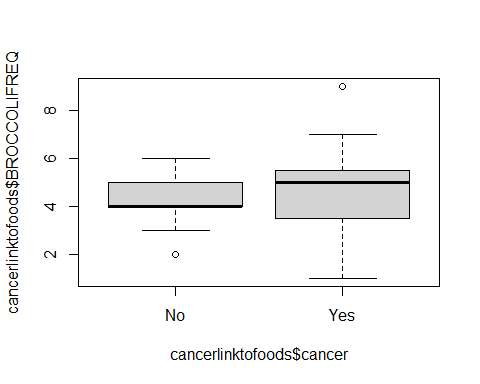
boxplot(cancerlinktofoods$RAWTOMATOESFREQ~cancerlinktofoods$cancer)



boxplot(cancerlinktofoods$BUTTERFREQ~cancerlinktofoods$cancer)



boxplot(cancerlinktofoods$BROCCOLIFREQ~cancerlinktofoods$cancer)



The boxplot chart helps visualize the summary statistics grouped by the factor(cancer).We can also see outliers.

The percentage of respondents with cancer and those without:

cancerlinktofoods

## cancer BEEFPORKDISHFREQ DRIEDFRUITFREQ OTHERFRESHFRUITFREQ CORNFREQ  
## 1 Yes 4 1 1 1  
## 2 No 3 5 5 6  
## 3 Yes 2 4 4 4  
## 4 No 3 7 7 3  
## 5 Yes 3 3 3 3  
## 6 No 0 3 6 6  
## 7 Yes 0 2 3 2  
## 8 Yes 1 5 5 3  
## 9 Yes 3 4 2 2  
## 10 Yes 2 2 2 2  
## 11 Yes 4 6 4 4  
## 12 No 2 1 2 3  
## 13 Yes 0 2 5 4  
## 14 No 2 4 8 5  
## 15 No 4 1 1 1  
## 16 No 5 5 4 3  
## 17 No 5 1 6 4  
## 18 Yes 2 1 2 2  
## 19 Yes 1 2 2 1  
## 20 No 2 2 6 4  
## 21 Yes 4 4 2 2  
## 22 No 2 1 2 3  
## 23 No 2 7 2 3  
## 24 No 4 2 4 2  
## 25 No 1 1 7 1  
## 26 No 4 4 5 3  
## 27 No 2 7 4 3  
## 28 No 3 4 4 3  
## 29 Yes 2 5 7 3  
## 30 No 1 2 5 2  
## 31 Yes 5 1 4 4  
## 32 Yes 3 2 4 3  
## 33 No 3 1 4 2  
## 34 Yes 2 6 1 2  
## 35 No 3 4 7 3  
## 36 No 2 2 4 4  
## 37 Yes 5 2 7 3  
## 38 Yes 7 2 3 4  
## 39 No 1 1 4 7  
## 40 No 1 1 2 2  
## 41 No 4 1 1 1  
## 42 Yes 2 1 3 6  
## 43 Yes 3 1 5 2  
## 44 Yes 3 2 2 1  
## 45 No 4 4 3 5  
## 46 Yes 1 2 2 2  
## 47 No 4 2 6 4  
## 48 Yes 2 2 2 4  
## 49 Yes 1 1 4 4  
## 50 Yes 1 2 8 4  
## 51 Yes 1 2 6 2  
## 52 Yes 3 4 3 4  
## 53 Yes 3 3 2 5  
## 54 No 1 3 3 2  
## SUGARINCOFFEE SALTFREQ HOTTEAFREQ RAWTOMATOESFREQ MILKFREQ BUTTERFREQ  
## 1 1 1 3 4 1 7  
## 2 1 5 1 4 1 3  
## 3 1 5 2 6 5 7  
## 4 1 5 9 3 2 2  
## 5 1 6 8 5 1 3  
## 6 1 2 6 8 1 3  
## 7 2 7 1 4 1 4  
## 8 1 8 2 5 3 7  
## 9 1 6 6 6 1 3  
## 10 1 1 9 6 1 3  
## 11 1 3 1 7 1 4  
## 12 2 3 4 2 1 3  
## 13 1 1 7 7 1 1  
## 14 1 3 9 7 2 9  
## 15 1 1 1 1 1 1  
## 16 1 9 7 3 2 9  
## 17 2 8 5 4 5 7  
## 18 1 2 7 2 1 7  
## 19 1 1 1 2 1 1  
## 20 1 6 7 4 9 3  
## 21 1 6 2 2 1 2  
## 22 2 2 4 2 8 1  
## 23 1 3 6 8 1 4  
## 24 1 4 2 4 5 2  
## 25 1 6 1 6 1 1  
## 26 1 4 6 6 5 5  
## 27 1 2 4 3 1 1  
## 28 1 9 6 5 2 7  
## 29 2 2 7 4 6 9  
## 30 1 6 8 6 1 4  
## 31 2 2 1 5 2 7  
## 32 2 4 7 7 2 6  
## 33 1 9 7 6 2 8  
## 34 1 1 1 1 1 1  
## 35 1 2 9 5 2 6  
## 36 1 9 1 4 8 7  
## 37 1 1 9 7 1 6  
## 38 1 5 4 4 6 9  
## 39 2 1 1 1 1 1  
## 40 1 4 2 3 1 4  
## 41 1 9 2 2 7 9  
## 42 1 2 1 1 9 9  
## 43 1 2 1 4 5 6  
## 44 1 1 2 8 1 5  
## 45 1 5 9 5 3 5  
## 46 1 7 1 6 1 2  
## 47 1 9 8 6 1 5  
## 48 1 9 7 9 1 4  
## 49 1 6 1 6 2 6  
## 50 1 7 7 4 1 2  
## 51 1 1 7 5 6 5  
## 52 2 2 4 7 9 5  
## 53 1 1 9 7 1 1  
## 54 1 2 8 2 4 4  
## OTHERBREADSFREQ BROCCOLIFREQ  
## 1 1 5  
## 2 2 4  
## 3 5 4  
## 4 3 4  
## 5 3 4  
## 6 8 5  
## 7 4 4  
## 8 5 5  
## 9 9 3  
## 10 9 3  
## 11 4 5  
## 12 8 4  
## 13 8 5  
## 14 7 6  
## 15 9 6  
## 16 7 5  
## 17 7 3  
## 18 3 6  
## 19 4 9  
## 20 9 3  
## 21 8 2  
## 22 2 6  
## 23 5 6  
## 24 5 3  
## 25 7 5  
## 26 5 4  
## 27 7 3  
## 28 7 4  
## 29 1 5  
## 30 2 2  
## 31 5 3  
## 32 7 5  
## 33 7 4  
## 34 1 2  
## 35 5 5  
## 36 7 4  
## 37 5 6  
## 38 9 7  
## 39 8 2  
## 40 5 4  
## 41 1 5  
## 42 9 1  
## 43 8 4  
## 44 7 6  
## 45 5 4  
## 46 9 5  
## 47 5 5  
## 48 4 7  
## 49 4 1  
## 50 4 6  
## 51 7 4  
## 52 4 4  
## 53 9 5  
## 54 5 4

cancertb <- table(cancerlinktofoods$cancer)  
cancerprop <- prop.table(cancertb)  
cancerdf <- as.data.frame(cancerprop)  
names(cancerdf) <- c("cancer", "Frequency")  
  
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.1.3

ggplot(data=cancerdf, mapping=aes(x=cancer, y=Frequency)) +   
 geom\_col(fill="darkgreen", alpha=0.5) +  
 scale\_y\_continuous(labels = scales::percent)

