Курсов проект по R

Гьокан Неждетов Сюлейманов ФН 62117, СИ 3-ти курс

1. Преглед на данните

Данните използвани в този проект са изтеглени от https://www.statcrunch.com/app/index.php?dataid=2323899

Това са хранителни данни за бургерите предлагани в различни вериги за бързо хранене в САЩ през 2017г.

```
# Прочитаме данните от food_data.csv
food_data = read.table("food_data.csv", header=TRUE, sep=";")

# Разгледаме първите няколко реда
head(food, n=15)
```

```
> head(food_data, n=15)
    restaurant servingSize calories totalFat saturatedFat transFat sodiumInMg carbs sugars protein
1    McDonalds          98      240        8            3      0.0        480    32      6      12
2    McDonalds         113      290       11            5      0.5        680    33      7      15
3    McDonalds         211      530       27           10      1.0        960    47      9      24
4    McDonalds         202      520       26           12      1.5       1100    41     10      30
5    McDonalds         270      720       40           15      1.5       1470    51     14      39
6    McDonalds         283      750       43           19      2.5       1280    42     10      48
7  Burger King          93      220        8            3      0.5        380    26      6      11
8  Burger King         104      270       12            5      0.5        560    27      7      13
9  Burger King         260      630       38           11      1.5        810    49     11      26
10 Burger King         279      750       49           16      2.0       1260    46      8      33
11 Burger King         341      850       54           18      2.5        870    49     11      43
12 Burger King         327     1040       69           28      2.5       1900    48     10      57
13      Wendys         102      250       10            4      1.0        600    25      5      14
14      Wendys         113      290       13            6      1.0        800    26      6      16
15      Wendys         247      580       31           13      2.0       1220    42     10      30
```

Данните представляват подробен хранителен анализ на бургери в ралични ресторанти.

Колоните в таблицата са:

- restaurant -> *McDonalds, Burger King, Wendys, Jack in the Box, Sonic, Dairy Queen, Carls Jr., Hardees, White Castle, Whataburger,In-N-Out Burger*
- servingSize -> тегло на бургера в грамове
- calories -> калории на бургера
- totalFat -> общо мазнини в грамове
- saturatedFat -> наситените мазнини в грамове
- transFat -> транс мазнини в грамове
- sodiumInMg -> сол в милиграм
- carbs -> въглехидрати в грамове

- sugars -> захари в грамове
- protein -> белтъчини в грамове

Тъй като представените данни са в грамове би било по-полезно да трансофрмираме колоните totalFat, carbs и protein в % калории.

Прието е, че 1грам мазнини са 9 калории, 1 грам въглехидрати и белтъци са 4 калории.

За целта ще създадем идентичен нов data-frame който ще съдържа % съотношение между трите съставки.

```
fd <- food_data[,c(4, 8,10)] # Избираме съответните номера на колони
fd$totalFat <- fd$totalFat * 9
fd$carbs <- fd$carbs * 4
fd$protein <- fd$protein * 4
percentage <- round(fd/rowSums(fd), 2) #   Процентното разпределение
```

2. Анализ на данните

- Структура на променливите

```
str(food_data)
```

```
'data.frame':   69 obs. of  10 variables:
 $ restaurant  : Factor w/ 11 levels "Burger King",..: 7 7 7 7 7 7 1 1 1 1 ...
 $ servingSize : int  98 113 211 202 270 283 93 104 260 279 ...
 $ calories    : int  240 290 530 520 720 750 220 270 630 750 ...
 $ totalFat    : num  8 11 27 26 40 43 8 12 38 49 ...
 $ saturatedFat: num  3 5 10 12 15 19 3 5 11 16 ...
 $ transFat    : num  0 0.5 1 1.5 1.5 2.5 0.5 0.5 1.5 2 ...
 $ sodiumInMg  : int  480 680 960 1100 1470 1280 380 560 810 1260 ...
 $ carbs       : num  32 33 47 41 51 42 26 27 49 46 ...
 $ sugars      : num  6 7 9 10 14 10 6 7 11 8 ...
 $ protein     : num  12 15 24 30 39 48 11 13 26 33 ...
```

Тук имаме 1 категорийна променлива с 11 нива и 9 числови променливи

- **Обща статистика на данните**

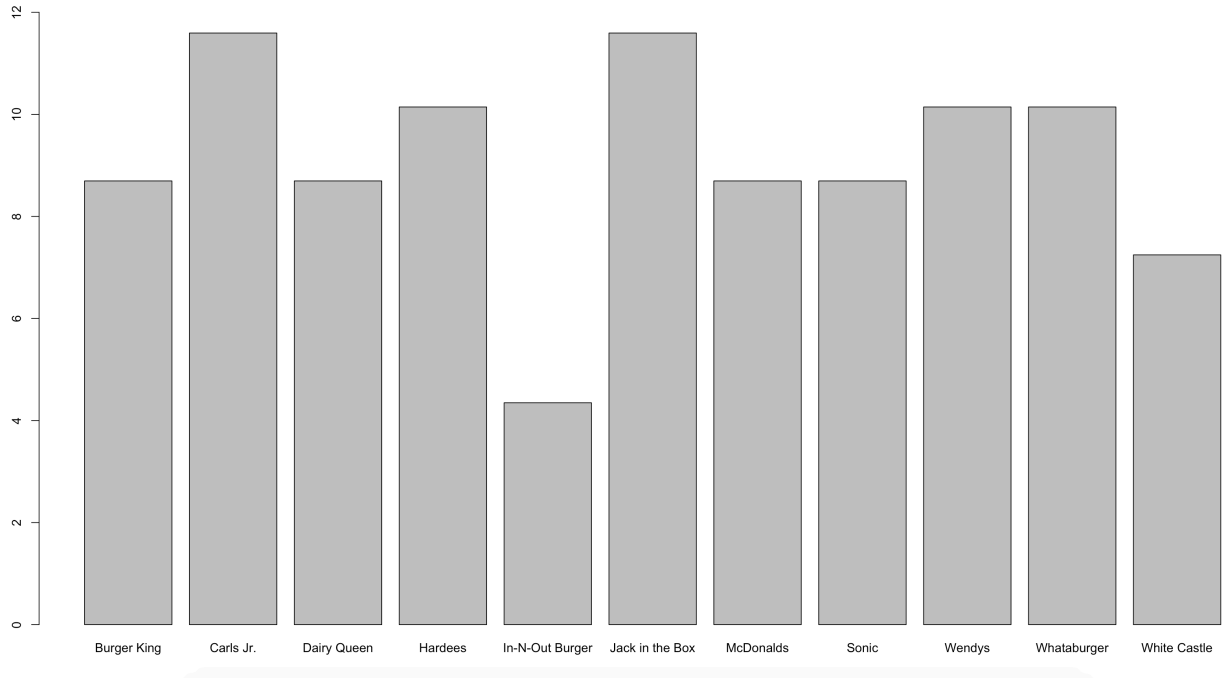
```
summary(food_data)
```

```
          restaurant  servingSize       calories         totalFat      saturatedFat
Carls Jr.       : 8   Min.   : 44.0   Min.   : 140.0   Min.   : 6.00   Min.   : 2.50
Jack in the Box: 8   1st Qu.:153.0   1st Qu.: 390.0   1st Qu.:19.00   1st Qu.: 7.00
Hardees         : 7   Median :247.0   Median : 591.0   Median :34.00   Median :13.00
Wendys          : 7   Mean   :241.8   Mean   : 620.2   Mean   :35.91   Mean   :13.56
Whataburger     : 7   3rd Qu.:321.0   3rd Qu.: 820.0   3rd Qu.:49.00   3rd Qu.:19.00
Burger King     : 6   Max.   :467.0   Max.   :1240.0   Max.   :87.00   Max.   :35.00
(Other)         :26
   transFat      sodiumInMg       carbs           sugars          protein
Min.   :0.00   Min.   : 360   Min.   :13.00   Min.   : 1.000   Min.   : 7.00
1st Qu.:0.50   1st Qu.: 840   1st Qu.:34.00   1st Qu.: 6.000   1st Qu.:17.70
Median :1.00   Median :1170   Median :43.00   Median : 9.000   Median :30.00
Mean   :1.31   Mean   :1209   Mean   :42.72   Mean   : 8.888   Mean   :31.46
3rd Qu.:2.00   3rd Qu.:1540   3rd Qu.:52.00   3rd Qu.:11.000   3rd Qu.:41.00
Max.   :4.00   Max.   :2460   Max.   :75.00   Max.   :20.000   Max.   :69.00
```

Тук виждаме минимални, максимални, средна стойност и мода на различните променливи. Например: най-високо калоричния бургер е 1240 калории, средната стойност на транс-мазнините е 1.31 грама, Средно един бургер има 1200милиграма сол и 620 калории.

- **Процентно разпределение на ресторантите по броя на предлагани бургери**

```
t_restaurants <- table(food_data$restaurant)
barplot(prop.table(t_restaurants)*100, 2, ylim=c(0,12))
```

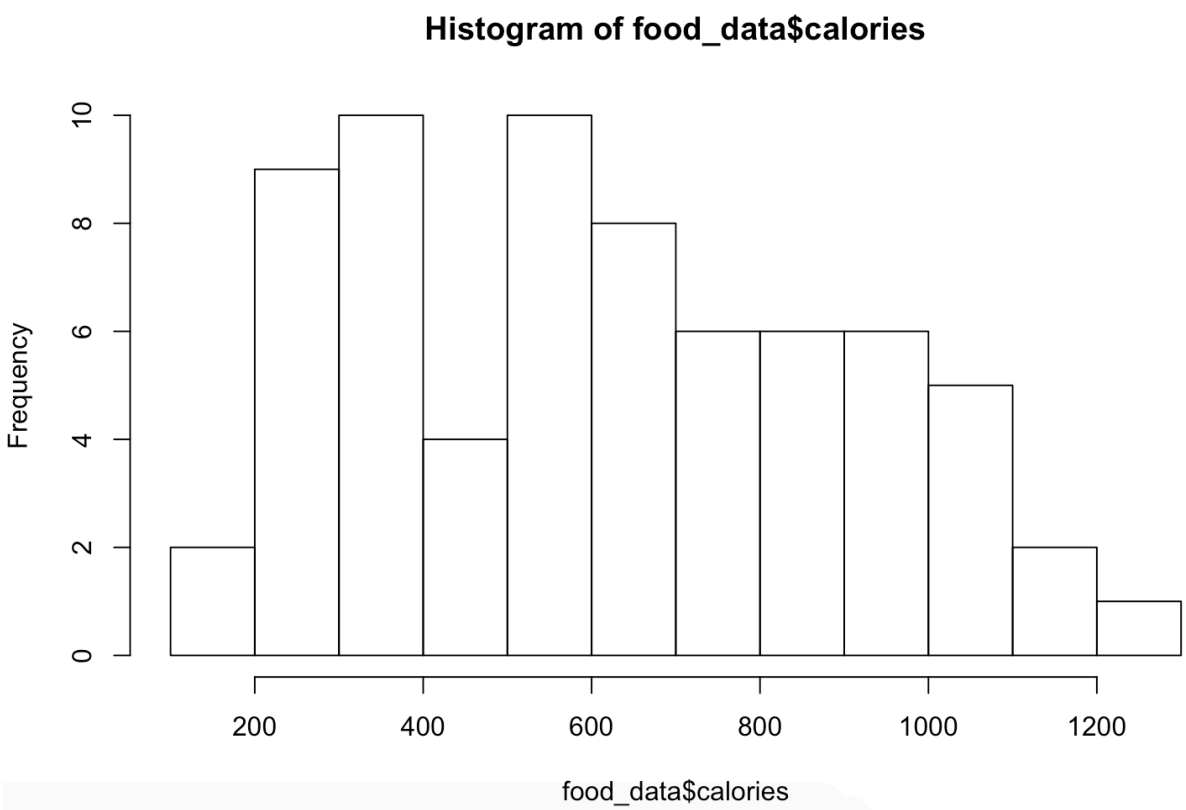


Както се вижда от хистограмата, различните ресторанти предлагат подобно разнообразие от бургери с изключение на *In-N-Out Burger*. Най-голямо разнообразие предлагат *Carls Jr.* и *Jack in the Box*.

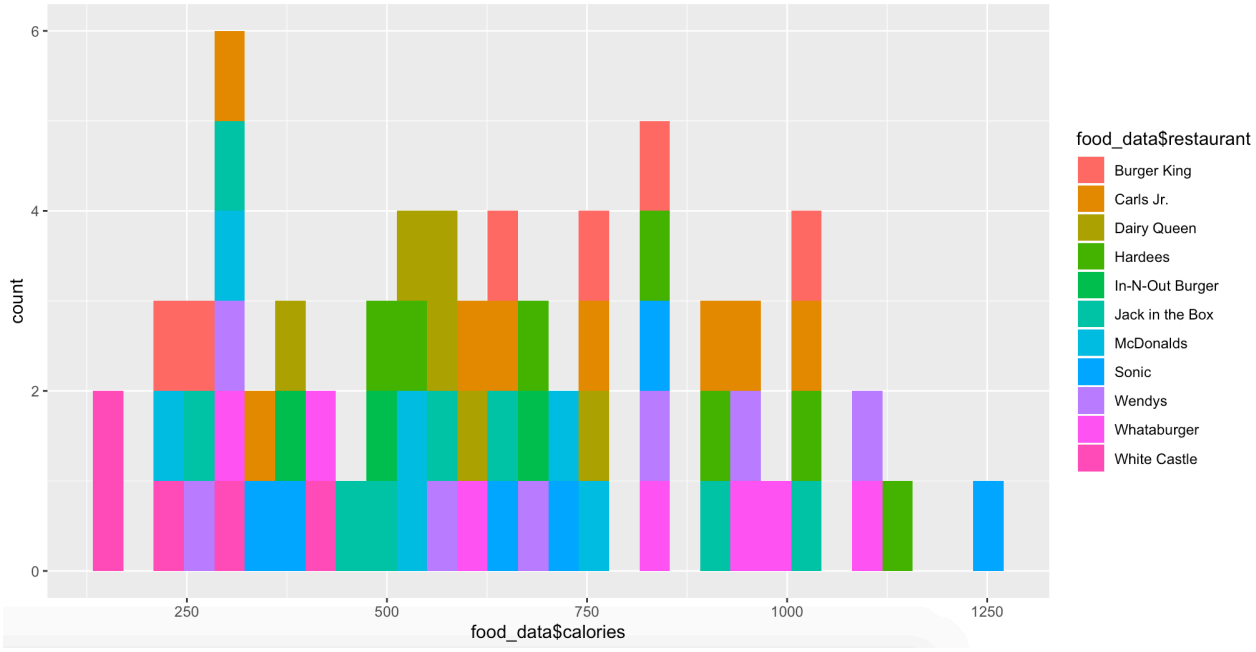
- **Графично представяне на данните**

  Визуално разпределение на бургери по калории:

```
hist(food_data$calories)
```



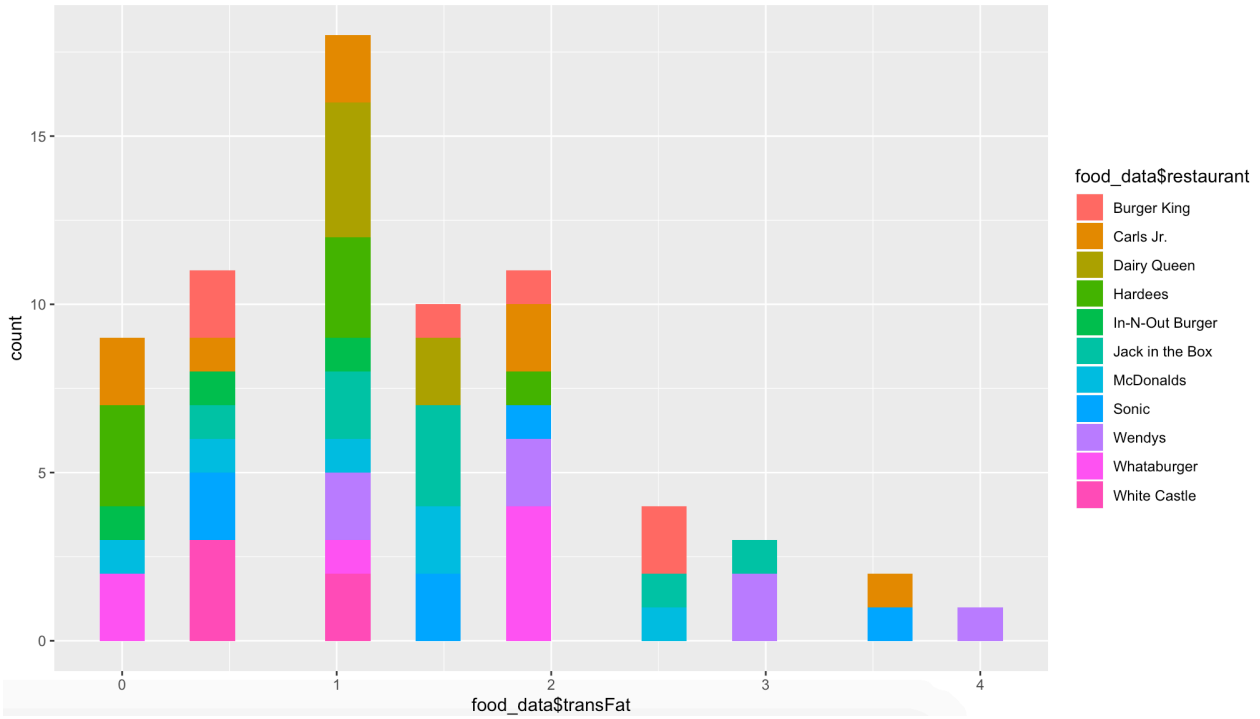

```
library(ggplot2)
ggplot(food_data,aes(x=food_data$calories, fill=food_data$restaurant)) +
geom_histogram()
```



Тук разпределението е доста смесено и не можем да направим някакви изводи освен, че *WhiteCastle* предлагат най-малко калорични бургери, а *Sonic* най-калорични.

Визуално разпределение спрямо транс-мазнини в бургерите:

```
ggplot(food_data,aes(x=food_data$transFat, fill=food_data$restaurant)) +
geom_histogram()
```

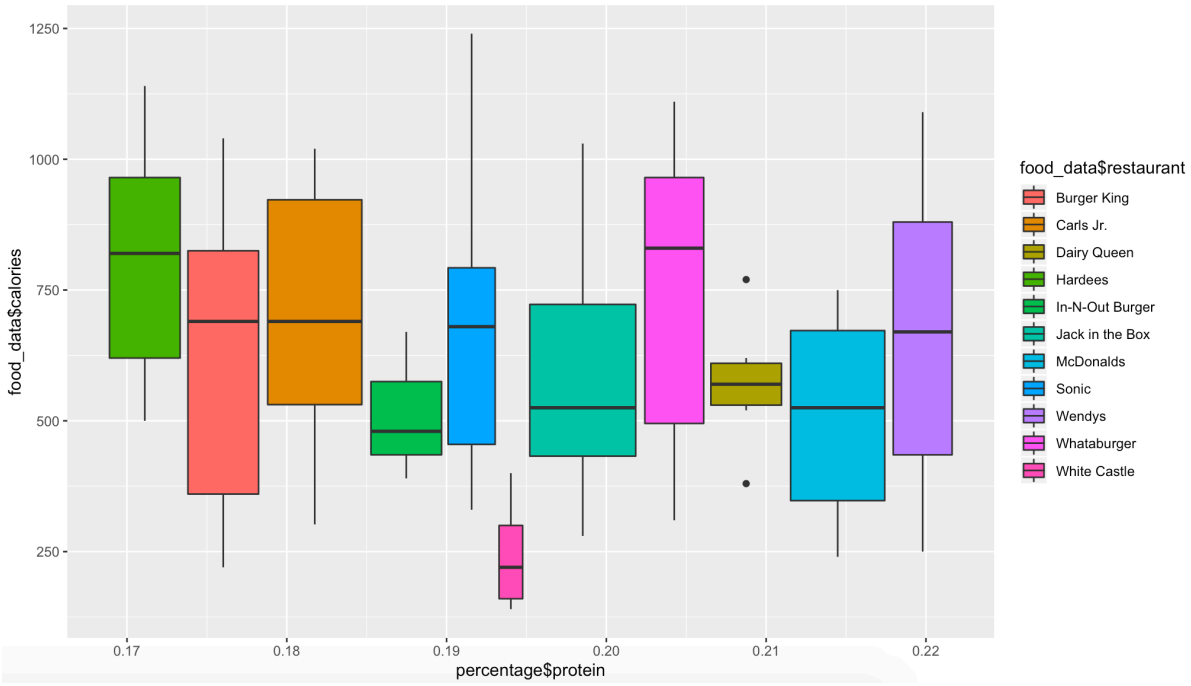


С най-високо съдържание на транс-мазнини са бургери, предлагани от *Wednys, Sonic, Carls Jr.*

Отново тук не бихме могли да направим някакво заключение, защото дистрибуцията не е разпределна. Можем да избягваме да се храним от *Wednys, Sonic, Carls Jr.* , които са първенци по транс-мазнини.

- Boxplot на данните - връзка между числова и категорийна променлива

  ```
  ggplot(food_data, aes(x=percentage$protein, y = food_data$calories, fill =
  food_data$restaurant)) + geom_boxplot()
  ```



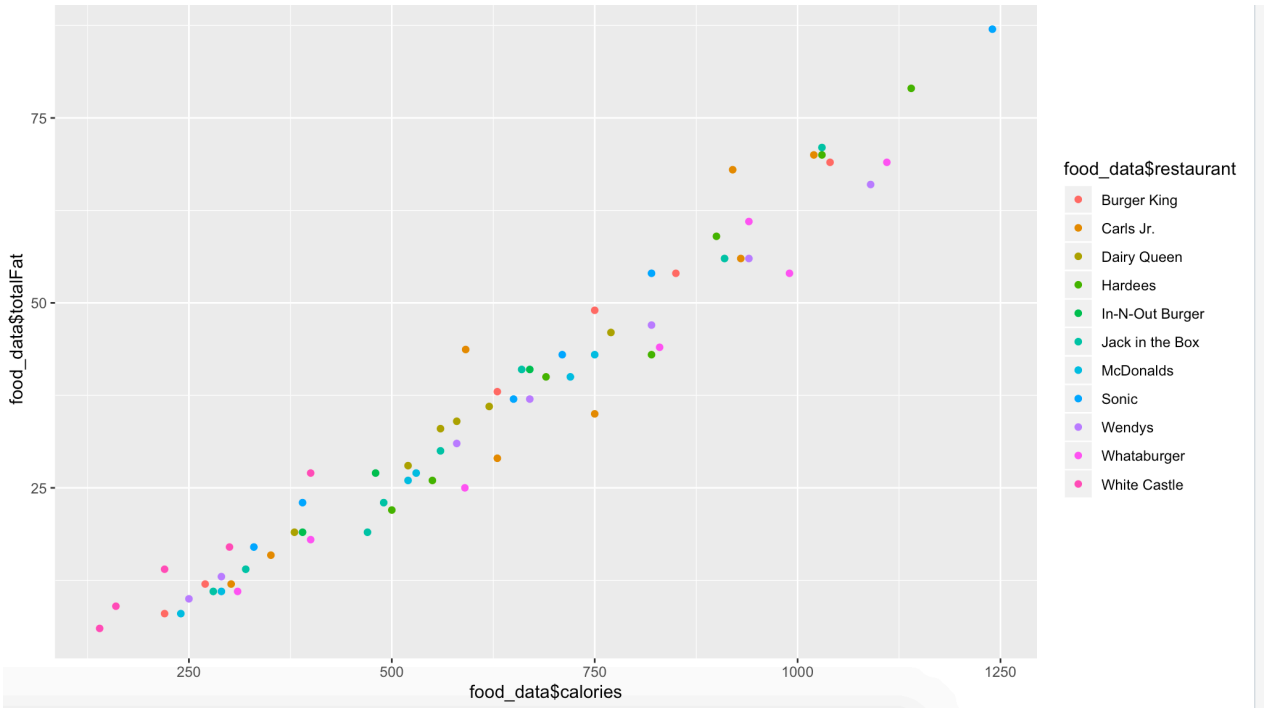
Най-богатите на % протеин бургери са предлагани от *Wednys, McDonalds, Dairy Queen* с над 20%.

Най-ниско калорични и богати на протеин бургери се предлагат от *Whataburger*

Най-ниски на белтъци и високи на калории, са бургерите в *Hardees*

- Корелация между променливи

```
ggplot(food_data,aes(food_data$calories, food_data$totalFat, colour =
food_data$restaurant)) + geom_point()
```



Тук виждаме логична пряка корелация между грамове мазнини и калории. Спрямо типовете ресторанти не може да направим извод, защото нямаме струпване. Може би, защото повечето верига магазини предлагат голямо разнобразие от типове на бургери.

```
cor(food_data[,2:10])
```

```
             servingSize  calories  totalFat saturatedFat  transFat sodiumInMg     carbs    sugars   protein
servingSize    1.0000000 0.9383381 0.9009281    0.8593496 0.6463626  0.8272979 0.7766999 0.7902987 0.8651828
calories       0.9383381 1.0000000 0.9777498    0.9543523 0.7666487  0.8933407 0.7133347 0.6749705 0.9481034
totalFat       0.9009281 0.9777498 1.0000000    0.9716487 0.7664865  0.8606982 0.6127573 0.5932260 0.9202488
saturatedFat   0.8593496 0.9543523 0.9716487    1.0000000 0.8275962  0.8538272 0.5168593 0.5491691 0.9495214
transFat       0.6463626 0.7666487 0.7664865    0.8275962 1.0000000  0.6437838 0.3039944 0.3664050 0.8278487
sodiumInMg     0.8272979 0.8933407 0.8606982    0.8538272 0.6437838  1.0000000 0.6457932 0.6268372 0.8601028
carbs          0.7766999 0.7133347 0.6127573    0.5168593 0.3039944  0.6457932 1.0000000 0.8345948 0.5836979
sugars         0.7902987 0.6749705 0.5932260    0.5491691 0.3664050  0.6268372 0.8345948 1.0000000 0.5759143
protein        0.8651828 0.9481034 0.9202488    0.9495214 0.8278487  0.8601028 0.5836979 0.5759143 1.0000000
```

3. Заключение

Може да се заключи, че различните верига магазини предлагат бургери с подобен хранителен състав. Трудно можем да определим някоя от тях като най-здравословен или най-нездравословен.