CEREALS_DATASET_R

MARIA BLOQUERT

Dataset: 80 Cereals

https://www.kaggle.com/crawford/80-cereals/code

Загружаем библиотеки для работы

```
library(tidyverse)
library(ggplot2)
library(readr)
library(dplyr)
```

Читаем файл, создаем переменную и изучаем данные

```
cereal <- read.csv("/Users/mariabloquert/cereal.csv")
View(cereal)
head(cereal)
summary(cereal)
```

```
     name               mfr                type              calories    
 Length:77          Length:77          Length:77          Min.   : 50.0  
 Class :character   Class :character   Class :character   1st Qu.:100.0  
 Mode  :character   Mode  :character   Mode  :character   Median :110.0  
                                                          Mean   :106.9  
                                                          3rd Qu.:110.0  
                                                          Max.   :160.0  
    protein            fat              sodium            fiber             carbo       
 Min.   :1.000   Min.   :0.000   Min.   :  0.0   Min.   : 0.000   Min.   :-1.0  
 1st Qu.:2.000   1st Qu.:0.000   1st Qu.:130.0   1st Qu.: 1.000   1st Qu.:12.0  
 Median :3.000   Median :1.000   Median :180.0   Median : 2.000   Median :14.0  
 Mean   :2.545   Mean   :1.013   Mean   :159.7   Mean   : 2.152   Mean   :14.6  
 3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:210.0   3rd Qu.: 3.000   3rd Qu.:17.0  
 Max.   :6.000   Max.   :5.000   Max.   :320.0   Max.   :14.000   Max.   :23.0  
     sugars            potass          vitamins          shelf           weight     
 Min.   :-1.000   Min.   : -1.00   Min.   :  0.00   Min.   :1.000   Min.   :0.50  
 1st Qu.: 3.000   1st Qu.: 40.00   1st Qu.: 25.00   1st Qu.:1.000   1st Qu.:1.00  
 Median : 7.000   Median : 90.00   Median : 25.00   Median :2.000   Median :1.00  
 Mean   : 6.922   Mean   : 96.08   Mean   : 28.25   Mean   :2.208   Mean   :1.03  
 3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00   3rd Qu.:3.000   3rd Qu.:1.00  
 Max.   :15.000   Max.   :330.00   Max.   :100.00   Max.   :3.000   Max.   :1.50  
      cups           rating      
 Min.   :0.250   Min.   :18.04  
 1st Qu.:0.670   1st Qu.:33.17  
 Median :0.750   Median :40.40  
 Mean   :0.821   Mean   :42.67  
 3rd Qu.:1.000   3rd Qu.:50.83  
 Max.   :1.500   Max.   :93.70  
```

Можно выбрать несколько ячеек и взглянуть на данные

```
cereal_some_cols <- cereal %>%
  select(mfr, type, sugars, calories, sodium, rating)
cereal_some_cols
```

	mfr	type	sugars	calories	sodium	rating
1	N	C	6	70	130	68.40297
2	Q	C	8	120	15	33.98368
3	K	C	5	70	260	59.42551
4	K	C	0	50	140	93.70491
5	R	C	8	110	200	34.38484
6	G	C	10	110	180	29.50954
7	K	C	14	110	125	33.17409
8	G	C	8	130	210	37.03856
9	R	C	6	90	200	49.12025
10	P	C	5	90	210	53.31381

Посмотрим на колонку «type» с числовыми значениями

```
cereal %>%
  distinct(type)
```

```
  type
1    C
2    H
```

```
cereal %>%
  count(type)
```

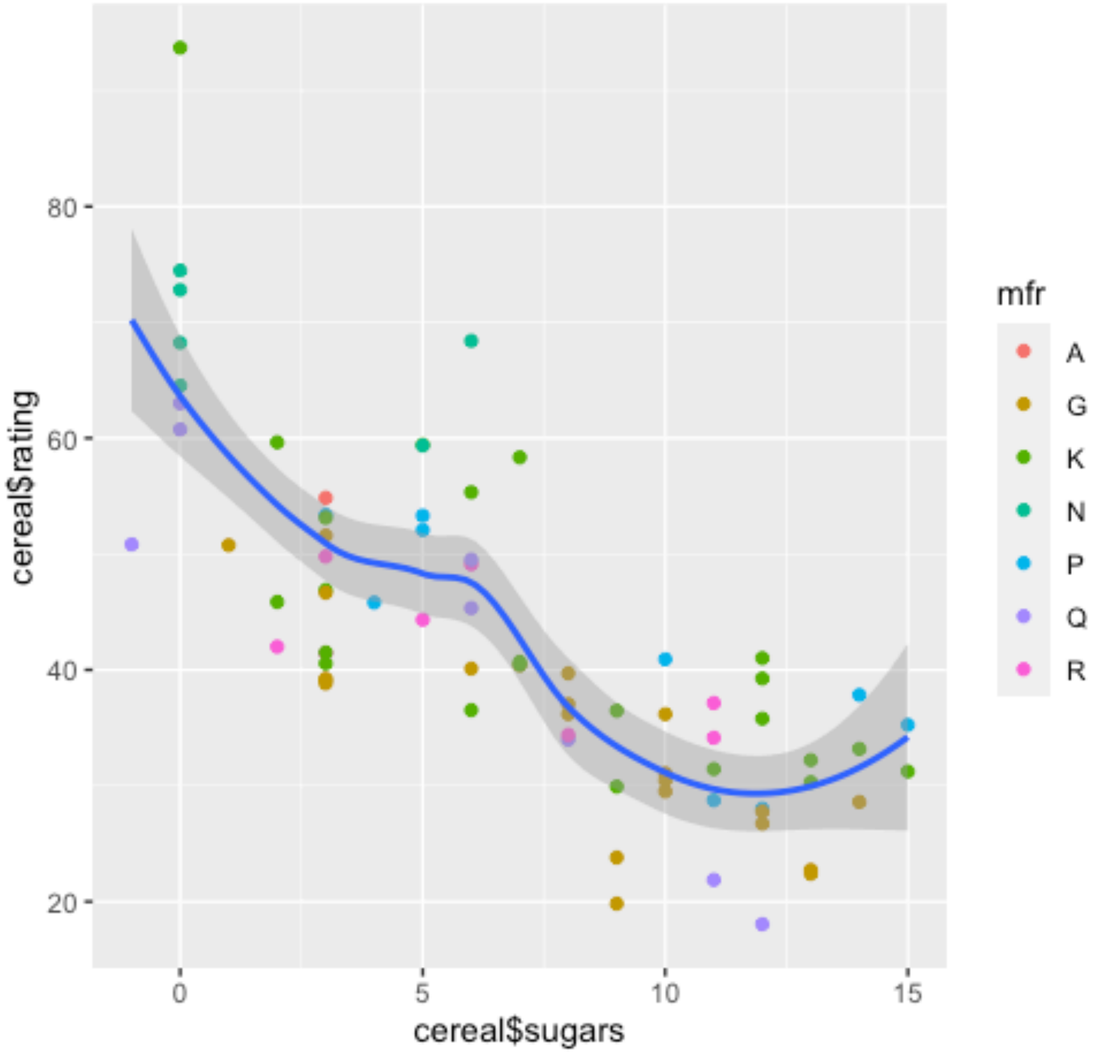
```
  type  n
1    C 74
2    H  3
>
```

В колонке всего 2 значения:
H - для Hot
C - для Cold
Соотношение этих типов очень не симметричны и скорее всего не показательны для исследования

А будет ли сахар влиять на рейтинг продукта?
чем < Сахара, тем > Рейтинг?

```
# Рисуем scatter plot

ggplot(cereal, aes(x = cereal$sugars, y =
cereal$rating)) +
  geom_point(aes(color = mfr)) +
  stat_smooth()
```



Посмотрим на коэффициент корреляции Пирсона

```
cor(cereal$sugars, cereal$rating)
```

```
-0.7596747
```

Вывод: отрицательная корреляция означает, что зависимость между рейтингом и уровнем сахара есть и она значительная.

Провдем тестирование уровня значимости для нулевой гипотезы корреляции

```
# H0: ρ = 0 (нет корреляции между уровнем рейтинга и кол-вом
сахара)
# Ha: ρ ≠ 0 есть корреляция между уровнем рейтинга и кол-вом
сахара)

test <- cor.test(cereal$sugars, cereal$rating)
test
Pearson's product-moment correlation

data:  cereal$sugars and cereal$rating
t = -10.117, df = 75, p-value = 1.153e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.8406210 -0.6455341
sample estimates:
       cor
-0.7596747
#Вывод: очень высокий p-value (1.153e-15) означает высокую вероятность
отклонения нулевой гипотезы
```

Как популярность хлопьев объясняется содержанием сахара?
Чтобы понять, построим модель линейной ренгресии

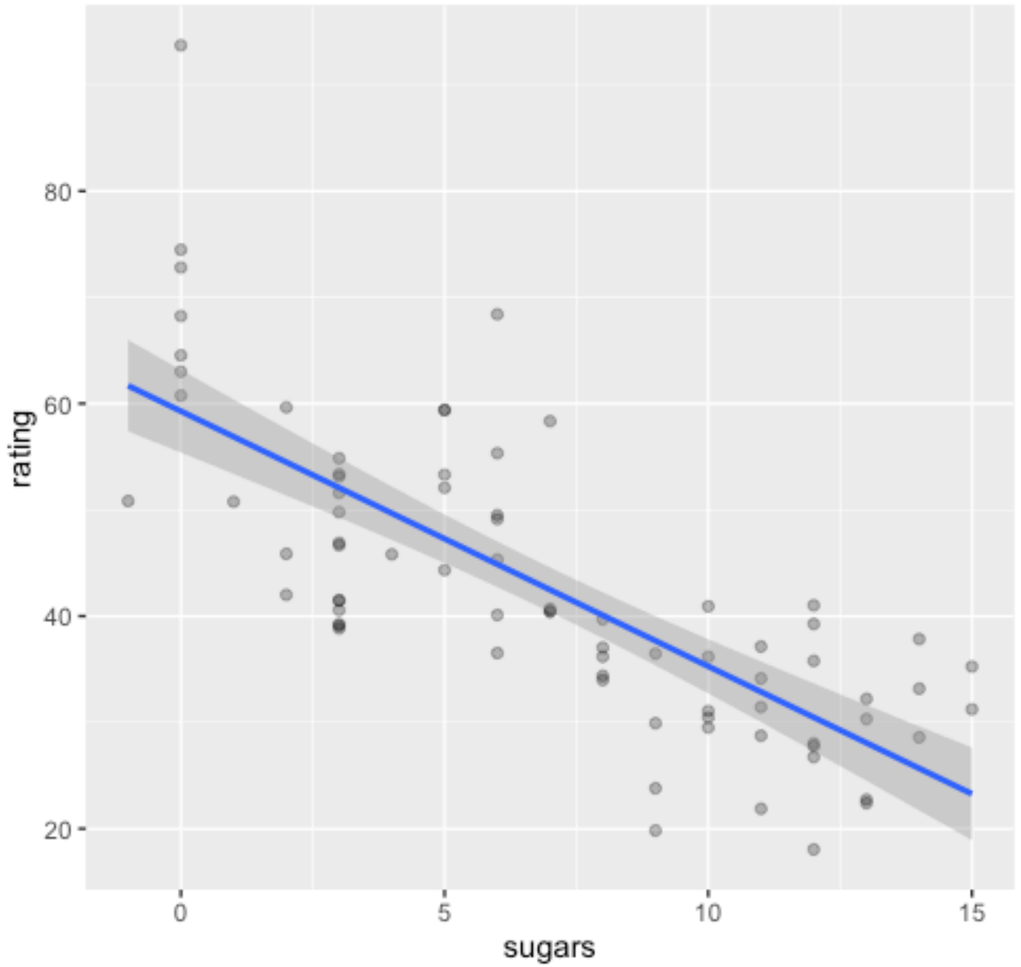
```
model <- lm(cereal$sugars ~ cereal$rating, data = cereal)
model

Call:
lm(formula = cereal$sugars ~ cereal$rating, data = cereal)

Coefficients:
  (Intercept)  cereal$rating
      17.1780        -0.2404
```

Сделаем визуализацию модели

```
ggplot(data = cereal,aes(x =
sugars, y = rating)) +
  geom_point(alpha = 0.3) +
  stat_smooth(method = lm)
```



Вывод параметров и коэффициентов линейной регрессии

summary(model)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   17.17800    1.06661   16.11  < 2e-16 ***
cereal$rating -0.24038    0.02376  -10.12 1.15e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.91 on 75 degrees of freedom
Multiple R-squared:  0.5771,                    Adjusted R-squared:  0.5715
F-statistic: 102.3 on 1 and 75 DF,  p-value: 1.153e-15
```

Мы допускаем, что в реальности есть некие настоящие коэффициенты линейной регрессии, и каждый раз собирая новые данные, они будут посчитаны как немного разные.