Ćwiczenia 1: Analiza danych

Kroki Analizy

1. Zdobycie wiedzy o zjawisku, danych.
2. Analiza danych w trakcie prreprocessingu, Preprocessing danych (uznajemy, że został wykonany): brak błędów grubych, wartości ekstremalnych i braków danych.
3. Analiza danych przed przystąpieniem do wykonania modelu.
4. Wybór zmiennych do modelu.
5. Model.
6. Ocena jakości modelu.
7. Wnioski (wdrożenie).

Przykład 1

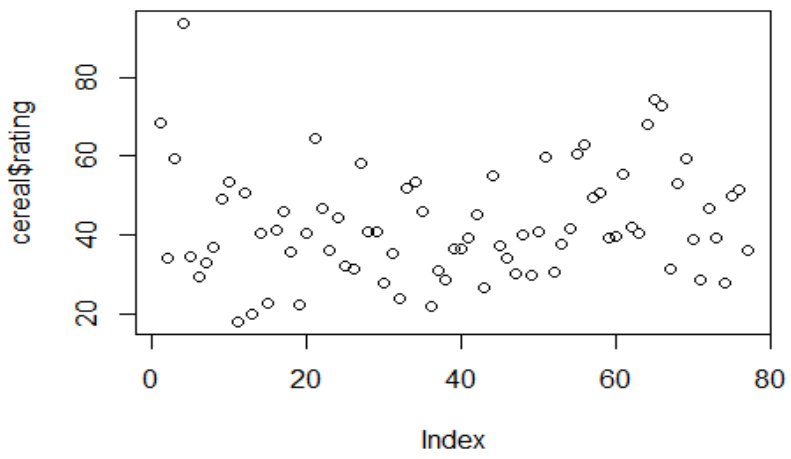
Celem jest analiza danych oraz wybór zmiennych do modelu regresyjnego. Zmienna, którą chcemy zamodelować to **rating** z danych cereal z pakietu liver.

Przed wykonaniem tutorialu sprawdź informacje o danych.

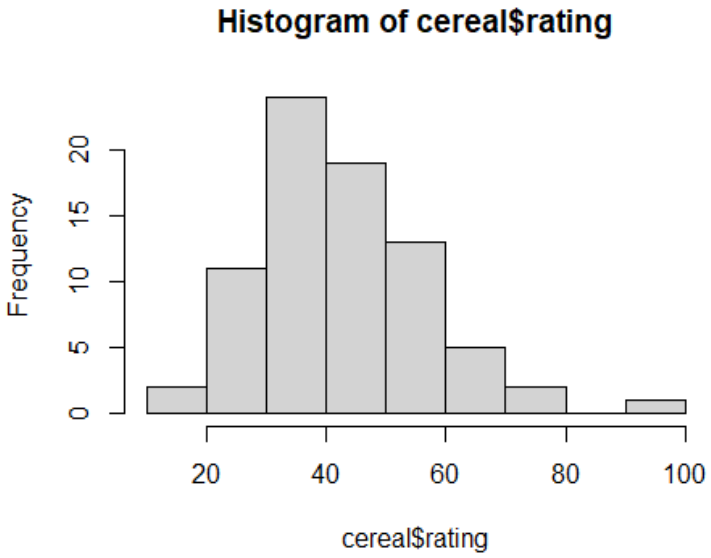
1. Dane nie wymagają preprocesingu
2. Analiza danych tabelarycznych:
    - Sprawdzenie statystyk opisowych,
    - Analiza rozkładu danych,
    - Wykresy ramka-wąsy ze zmienną grupującą,
    - Macierze korelacji liniowej, nieparametrycznej, wykresy korelacyjne,
    - Wykresy rozrzutu,
    - Wybór zmiennych do planowanego modelu.

```
library(liver)

data("cereal")

str(cereal)

## 'data.frame':    77 obs. of  16 variables:
##  $ name    : Factor w/ 77 levels "100% Bran","100% Natural Bran",..: 1 2 3
4 5 6 7 8 9 10 ...
##  $ manuf   : Factor w/ 7 levels "A","G","K","N",..: 4 6 3 3 7 2 3 2 7 5 ..
.
##  $ type    : Factor w/ 2 levels "cold","hot": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##  $ calories: int  70 120 70 50 110 110 110 130 90 90 ...
##  $ protein : int  4 3 4 4 2 2 2 3 2 3 ...
##  $ fat     : int  1 5 1 0 2 2 0 2 1 0 ...
##  $ sodium  : int  130 15 260 140 200 180 125 210 200 210 ...
##  $ fiber   : num  10 2 9 14 1 1.5 1 2 4 5 ...
##  $ carbo   : num  5 8 7 8 14 10.5 11 18 15 13 ...
##  $ sugars  : int  6 8 5 0 8 10 14 8 6 5 ...
##  $ potass  : int  280 135 320 330 -1 70 30 100 125 190 ...
##  $ vitamins: int  25 0 25 25 25 25 25 25 25 25 ...
##  $ shelf   : int  3 3 3 3 3 1 2 3 1 3 ...
##  $ weight  : num  1 1 1 1 1 1 1 1.33 1 1 ...
##  $ cups    : num  0.33 1 0.33 0.5 0.75 0.75 1 0.75 0.67 0.67 ...
##  $ rating  : num  68.4 34 59.4 93.7 34.4 ...

summary(cereal)

##                          name    manuf    type       calories
##  100% Bran                 : 1   A: 1   cold:74   Min.   : 50.0
##  100% Natural Bran         : 1   G:22   hot : 3   1st Qu.:100.0
##  All-Bran                  : 1   K:23             Median :110.0
##  All-Bran with Extra Fiber: 1   N: 6             Mean   :106.9
##  Almond Delight            : 1   P: 9             3rd Qu.:110.0
##  Apple Cinnamon Cheerios   : 1   Q: 8             Max.   :160.0
##  (Other)                   :71   R: 8
##     protein           fat            sodium          fiber
##  Min.   :1.000   Min.   :0.000   Min.   :  0.0   Min.   : 0.000
##  1st Qu.:2.000   1st Qu.:0.000   1st Qu.:130.0   1st Qu.: 1.000
##  Median :3.000   Median :1.000   Median :180.0   Median : 2.000
##  Mean   :2.545   Mean   :1.013   Mean   :159.7   Mean   : 2.152
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:210.0   3rd Qu.: 3.000
##  Max.   :6.000   Max.   :5.000   Max.   :320.0   Max.   :14.000
##
##      carbo          sugars           potass          vitamins
##  Min.   :-1.0   Min.   :-1.000   Min.   : -1.00   Min.   :  0.00
##  1st Qu.:12.0   1st Qu.: 3.000   1st Qu.: 40.00   1st Qu.: 25.00
##  Median :14.0   Median : 7.000   Median : 90.00   Median : 25.00
##  Mean   :14.6   Mean   : 6.922   Mean   : 96.08   Mean   : 28.25
##  3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00
##  Max.   :23.0   Max.   :15.000   Max.   :330.00   Max.   :100.00
##
##      shelf           weight          cups           rating
##  Min.   :1.000   Min.   :0.50   Min.   :0.250   Min.   :18.04
##  1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.670   1st Qu.:33.17
##  Median :2.000   Median :1.00   Median :0.750   Median :40.40
##  Mean   :2.208   Mean   :1.03   Mean   :0.821   Mean   :42.67
##  3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:50.83
##  Max.   :3.000   Max.   :1.50   Max.   :1.500   Max.   :93.70
##

plot(cereal$rating)
```

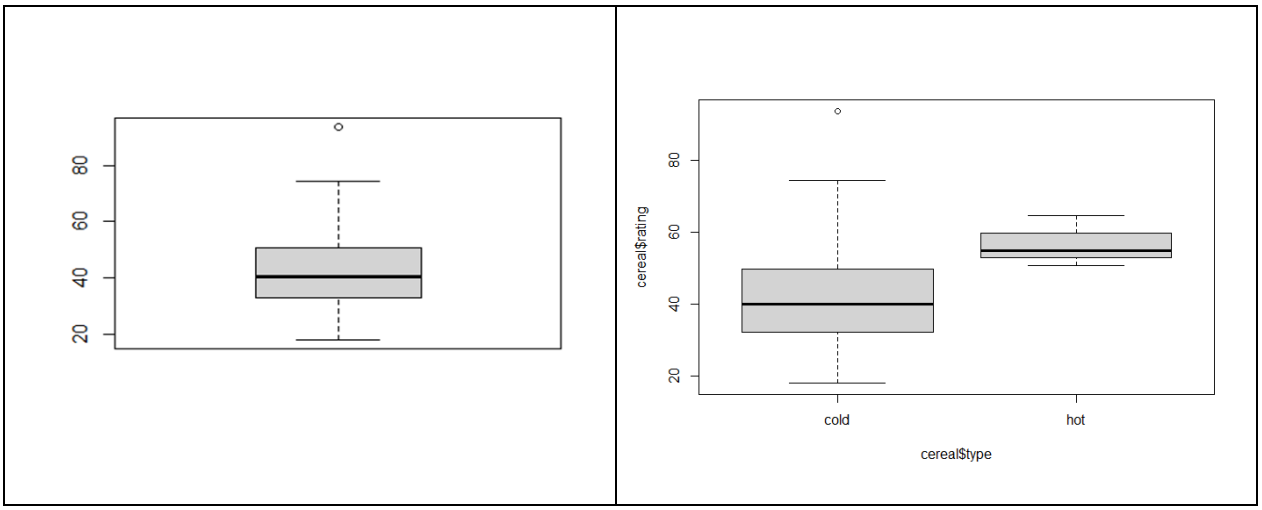


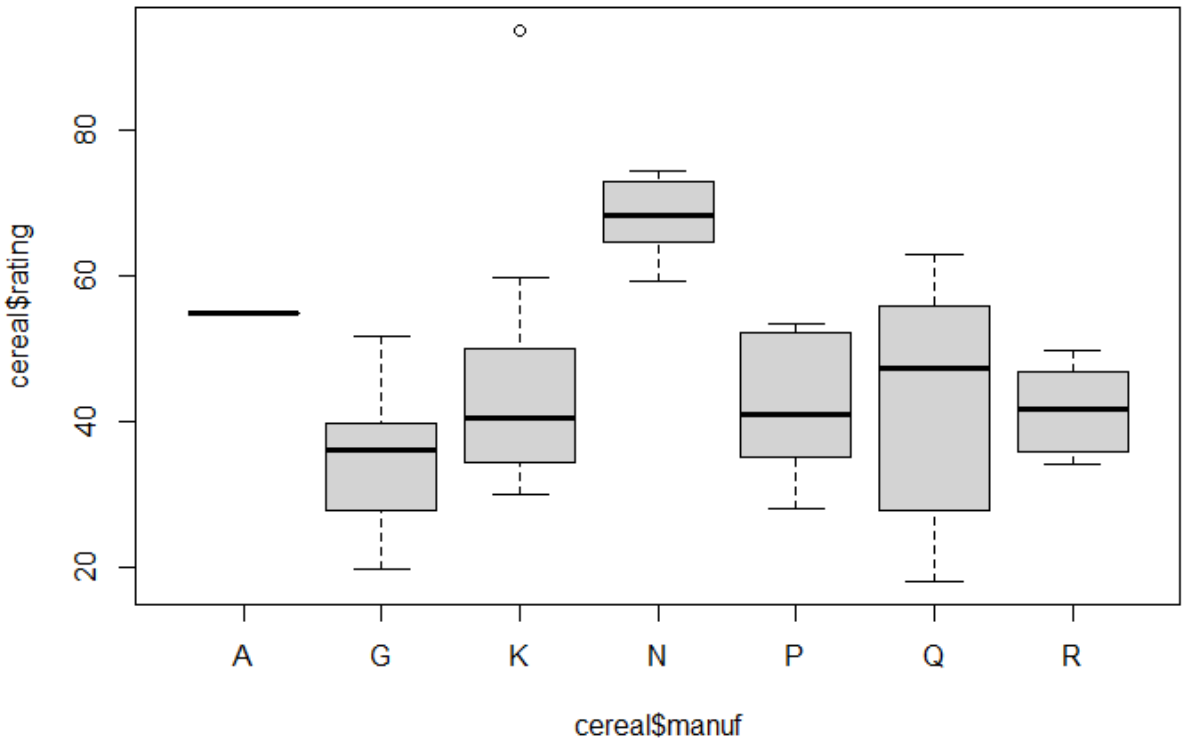

```
hist(cereal$rating)
```




```
boxplot(cereal$rating)
```

```
boxplot(cereal$rating~cereal$type)
```





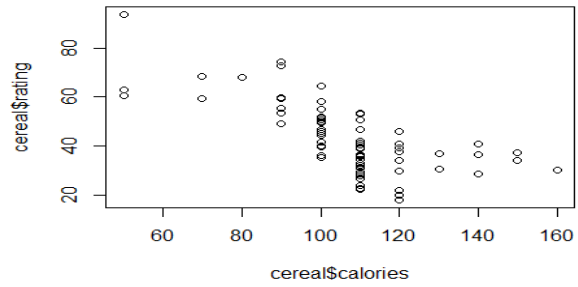

```
library(corrplot)

cor_matrixp<-round(cor(cereal[4:16]),2)
```

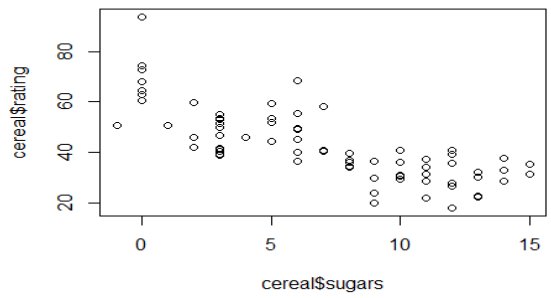
```
          calories protein   fat sodium fiber carbo sugars potass vitamins shelf weight  cups rating
calories     1.00    0.02  0.50   0.30 -0.29  0.25   0.56  -0.07     0.27  0.10   0.70  0.09  -0.69
protein      0.02    1.00  0.21  -0.05  0.50 -0.13  -0.33   0.55     0.01  0.13   0.22 -0.24   0.47
fat          0.50    0.21  1.00  -0.01  0.02 -0.32   0.27   0.19    -0.03  0.26   0.21 -0.18  -0.41
sodium       0.30   -0.05 -0.01   1.00 -0.07  0.36   0.10  -0.03     0.36 -0.07   0.31  0.12  -0.40
fiber       -0.29    0.50  0.02  -0.07  1.00 -0.36  -0.14   0.90    -0.03  0.30   0.25 -0.51   0.58
carbo        0.25   -0.13 -0.32   0.36 -0.36  1.00  -0.33  -0.35     0.26 -0.10   0.14  0.36   0.05
sugars       0.56   -0.33  0.27   0.10 -0.14 -0.33   1.00   0.02     0.13  0.10   0.45 -0.03  -0.76
potass      -0.07    0.55  0.19  -0.03  0.90 -0.35   0.02   1.00     0.02  0.36   0.42 -0.50   0.38
vitamins     0.27    0.01 -0.03   0.36 -0.03  0.26   0.13   0.02     1.00  0.30   0.32  0.13  -0.24
shelf        0.10    0.13  0.26  -0.07  0.30 -0.10   0.10   0.36     0.30  1.00   0.19 -0.34   0.03
weight       0.70    0.22  0.21   0.31  0.25  0.14   0.45   0.42     0.32  0.19   1.00 -0.20  -0.30
cups         0.09   -0.24 -0.18   0.12 -0.51  0.36  -0.03  -0.50     0.13 -0.34  -0.20  1.00 -0.20
rating      -0.69    0.47 -0.41  -0.40  0.58  0.05  -0.76   0.38    -0.24  0.03  -0.30 -0.20  1.00
```

```
cor_matrixs<-round(cor(cereal[4:16], method="spearman"),2)
```

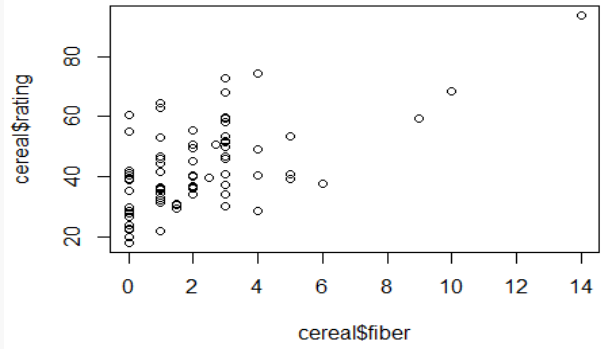
	Calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
calories	1.00	-0.07	0.54	0.29	-0.14	0.08	0.60	-0.01	0.31	0.15	0.62	0.05	-0.71
protein	-0.07	1.00	0.23	-0.11	0.68	0.00	-0.29	0.71	-0.01	0.28	0.29	-0.36	0.51
fat	0.54	0.23	1.00	0.03	0.12	-0.30	0.33	0.32	0.12	0.24	0.29	-0.25	-0.44
sodium	0.29	-0.11	0.03	1.00	-0.17	0.38	-0.01	-0.12	0.44	-0.15	0.26	0.15	-0.24
fiber	-0.14	0.68	0.12	-0.17	1.00	-0.15	-0.11	0.85	-0.04	0.32	0.35	-0.51	0.49
carbo	0.08	0.00	-0.30	0.38	-0.15	1.00	-0.46	-0.20	0.19	-0.11	0.12	0.32	0.24
sugars	0.60	-0.29	0.33	-0.01	-0.11	-0.46	1.00	-0.01	0.30	0.07	0.45	-0.06	-0.81
potass	-0.01	0.71	0.32	-0.12	0.85	-0.20	-0.01	1.00	0.01	0.36	0.43	-0.54	0.31
vitamins	0.31	-0.01	0.12	0.44	-0.04	0.19	0.30	0.01	1.00	0.27	0.39	0.08	-0.33
shelf	0.15	0.28	0.24	-0.15	0.32	-0.11	0.07	0.36	0.27	1.00	0.26	-0.31	0.07
weight	0.62	0.29	0.29	0.26	0.35	0.12	0.45	0.43	0.39	0.26	1.00	-0.29	-0.28
cups	0.05	-0.36	-0.25	0.15	-0.51	0.32	-0.06	-0.54	0.08	-0.31	-0.29	1.00	-0.16
rating	-0.71	0.51	-0.44	-0.24	0.49	0.24	-0.81	0.31	-0.33	0.07	-0.28	-0.16	1.00

`plot(cereal$rating~cereal$calories)`	`plot(cereal$rating~cereal$sugars)`



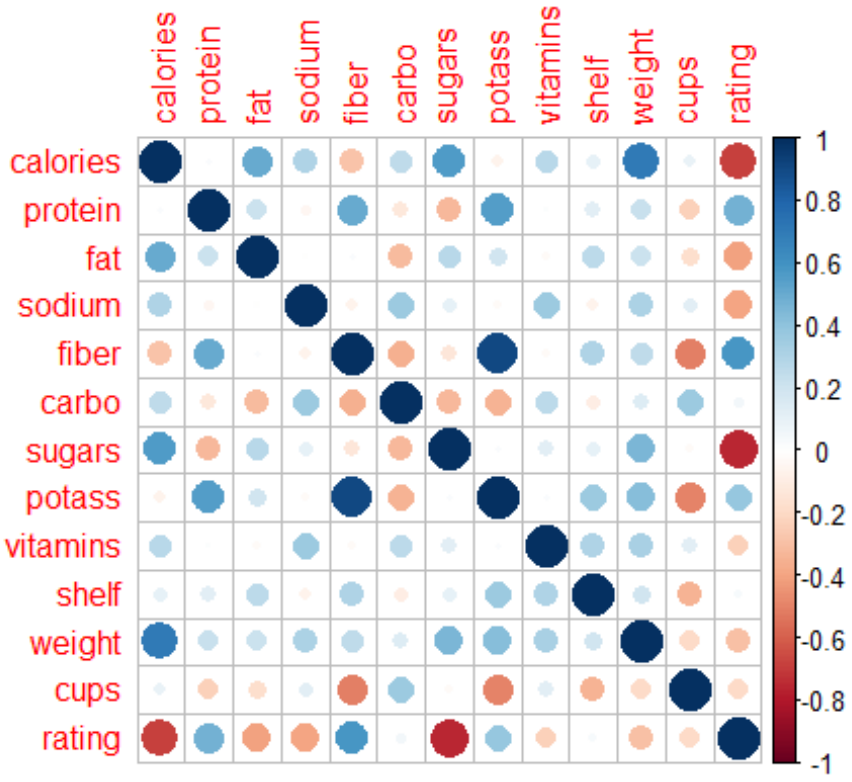



```
plot(cereal$rating~cereal$fiber)
```



Wykonaj dodatkowo korelację rang Spearmana i wykres korelacyjny tej dla macierzy korelacji

```
corrplot(cor_matrix)
```



korelacje- kilka najsilniejszych korelacji, które mogą się przydać w modelach dla rating: calories -0.69; protein 0.47; fat -0.41; sodium -0.40; fiber 0.58; sugars -0.76; potass 0.38, a z pozostałych wykresów zmienna type.

```
#podział na zbiór uczący i testowy

Wybór celowy
train<-cereal[-c(5, 15, 25, 35, 55),]
test<-cereal[c(5, 15, 25, 35, 55),]

Losowy
#sets <- sample(1:nrow(cereals), 0.9 * nrow(cereals))
#train2<- cereals [sets,]
#test2<- cereals[-sets,]
```

Zadanie 1

Sprawdź kroki metodyki CRISP-dm. W zadaniu postępuj zgodnie z jej etapami.

Wykonaj analizę danych dla **real_estate** (dane dostępne na kanale ogólnym zajęć). Sprawdź czego dotyczą dane. Oceń na podstawie analizy danych czy dane wymagają preprocessingu, jakie występują zależności między zmiennymi. Pomyśl jaki model można wykonać: co będzie zmienną zależną, jakie zmienne będą niezależne.

Dane dotyczą cen domów za m2 w zależności od lokalizacji domu i jego wieku. Do zbioru testowego do predykcji przenieś 5% obserwacji (w celu wykonania prognoz i ich oceny).

Zadanie 2

Dla danych Series_G wykonaj analizę danych. Na podstawie analizy danych wprowadź dodatkowe zmienne do arkusza z danymi w celu wykonania modelu endogennego.

Podpowiedź: Najpierw sprawdź występowanie trendu i sezonowości (wykresy ACF, PACF, wykresy szeregów czasowych). Dodaj odpowiednie zmienne do danych informujące o trendzie i sezonowości jeśli one występują (dummies variables). Sprawdź jakie pakiety pozwalają na tworzenie takich zmiennych.