Projekt Zaliczeniowy nr 1

Mateusz Kapusta

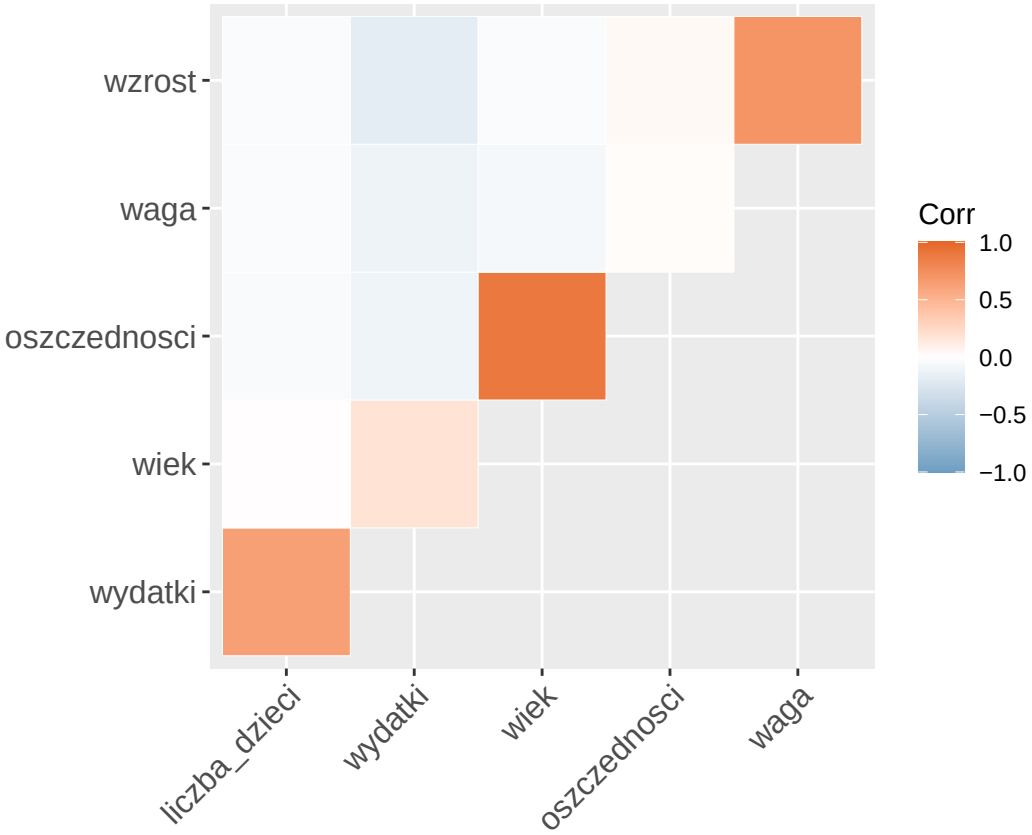
2022-04-27

Wpierw na samym początku wczytujemy dane, które posłużą nam do wykonania modelu.

```
data<-read.csv("people_tab.csv",sep="\t")
num<-lapply(data,is.numeric)
```

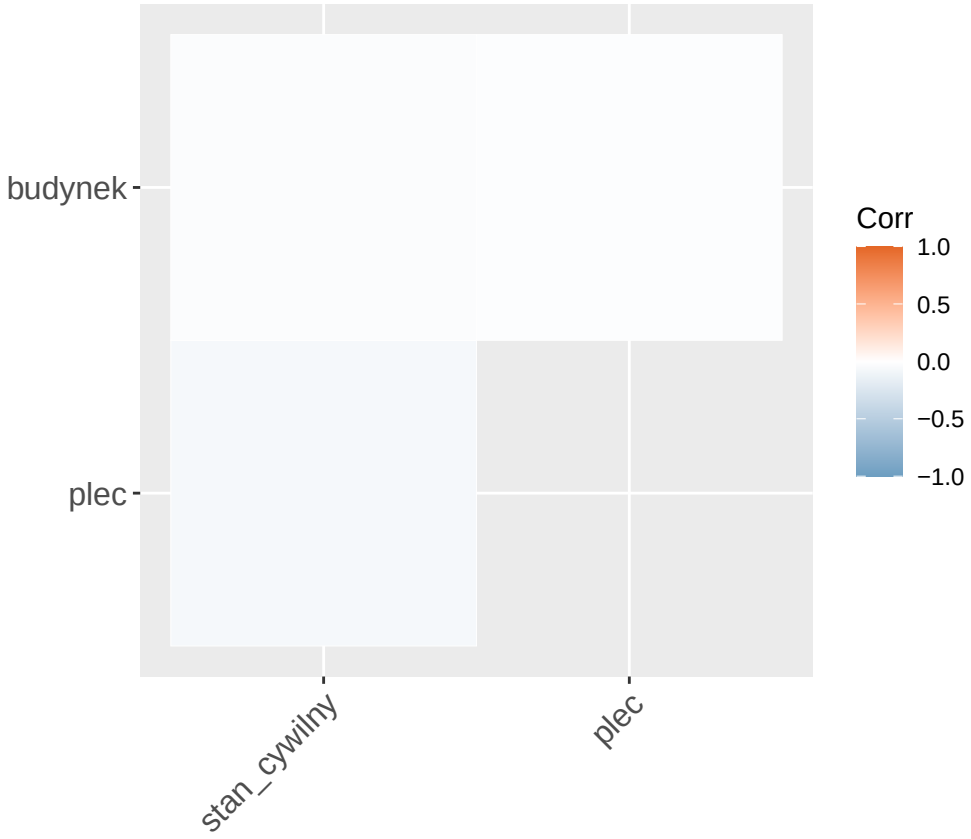
Dane składają się z 500 obserwacji natomiast każda obserwacja liczy sobie 9 parametrów z czego $ncol(data) - sum(num)$ to parametry jakościowe. W celu zbadaniu korelacji pomiędzy zmiennymi znajdujemy macierz korelacji metodą Pearsona.

```
data_ilo<-data[unlist(num)]
cor_matrix<-cor(data_ilo)
ggcorrplot(cor_matrix, hc.order = TRUE, type = "upper",
   outline.col = "white",
   ggtheme = ggplot2::theme_grey,
   colors = c("#6D9EC1", "white", "#E46726"),
   insig = "blank")
```



Widzimy, że największe dodatnie korelacje zachodzą pomiędzy wiekiem a oszczędnościami, wydatkami a liczbą dzieci oraz wzrostem a wagą. Brakuje natomiast nam silnych ujemnych korelacji pomiędzy danymi. Teraz zbadajmy korelacje pomiędzy zmiennymi jakościowymi.

```
data_jako<-na.omit(data[!unlist(num)]) #usuwamy pola bez wartości
cor_matrix_jako<-cor(data.frame(lapply(data_jako,function (x) as.integer(as.factor(x)))))
ggcorrplot(cor_matrix_jako, hc.order = TRUE, type = "upper",
   outline.col = "white",
   ggtheme = ggplot2::theme_grey,
   colors = c("#6D9EC1", "white", "#E46726"),
   insig = "blank")
```



Widzimy więc, że korelacje pomiędzy zmiennymi jakościowymi są bardzo małe. W przypadku zmiennej płeć mamy braki w danych, które na samym początku usuwamy.