Projekt Zaliczeniowy nr 1

Mateusz Kapusta

2022-05-05

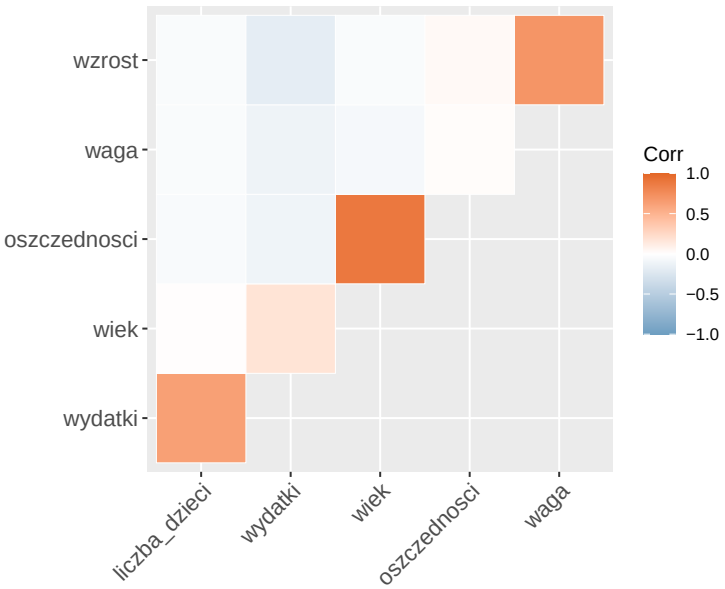
1

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.

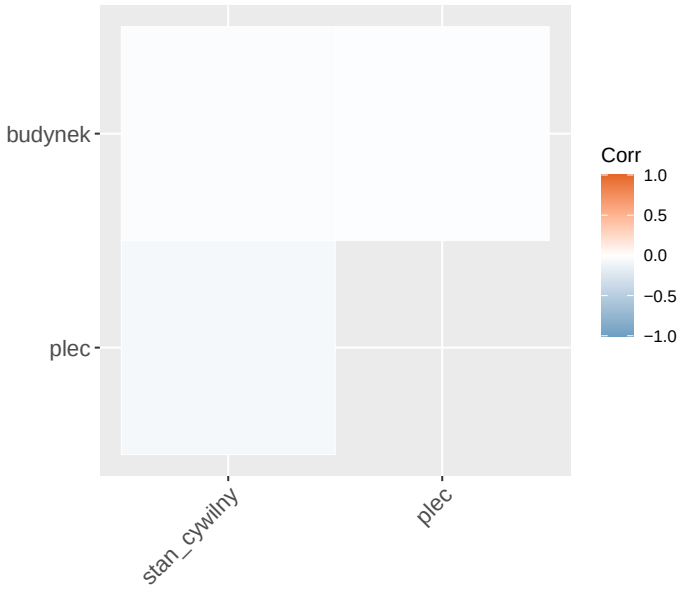
```
data_ilo<-data[unlist(num)]
cor_matrix<-cor(data_ilo)
ggcorrplot(cor_matrix, hc.order = TRUE, type = "upper",
   outline.col = "white",
   ggtheme = ggplot2::theme_grey,
   colors = c("#6D9EC1", "white", "#E46726"),
   insig = "blank")
```



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)))))
```

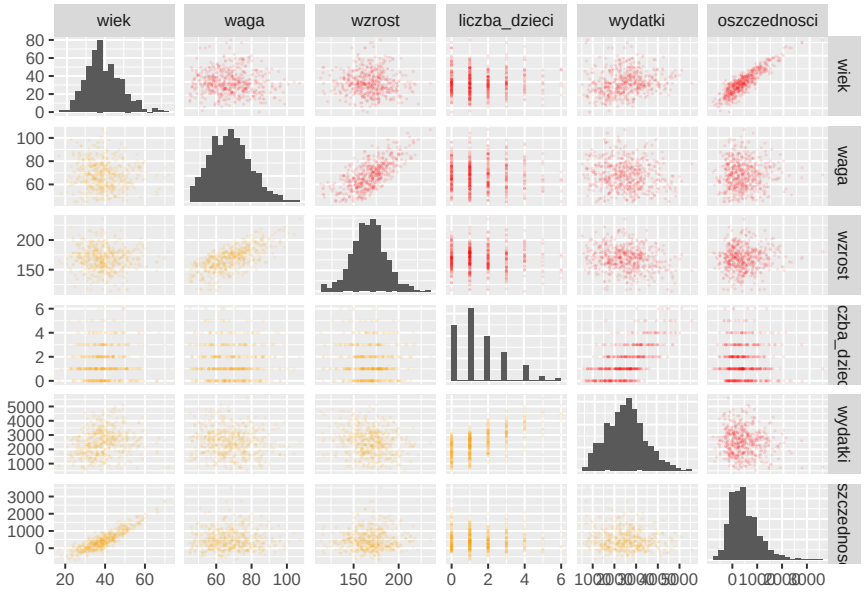
```
ggcorrplot(cor_matrix_jako, hc.order = TRUE, type = "upper",
   outline.col = "white",
   ggtheme = ggplot2::theme_grey,
   colors = c("#6D9EC1", "white", "#E46726"),
   insig = "blank")
```



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

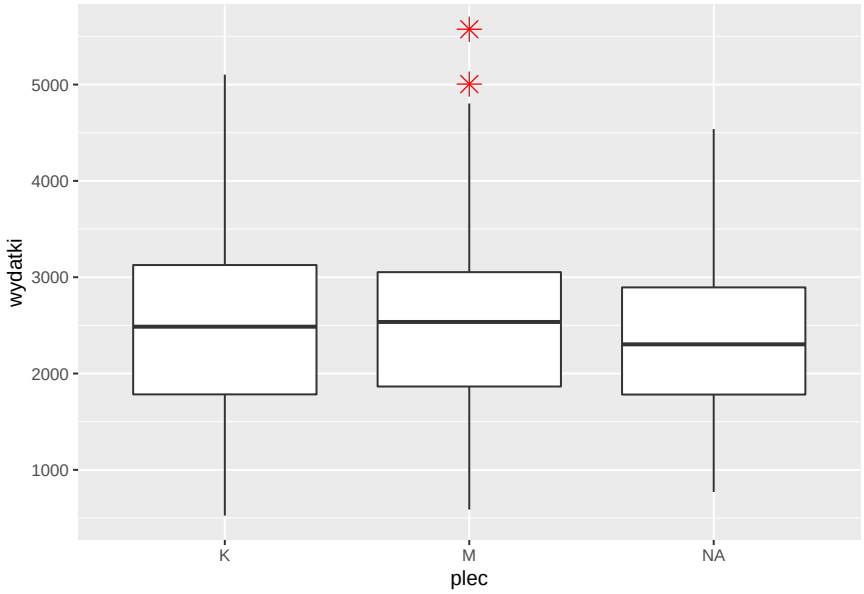
2

```
ggpairs(data_ilo,
        upper= list(continuous = wrap("points",color="red",size=0.1,alpha=1/10), combo = "box_no_facet"
        lower=list(continuous=wrap("points",color="orange",size=0.1,alpha=1/10)),
        diag=list(continuous=wrap("barDiag",bins=20))
        )
```



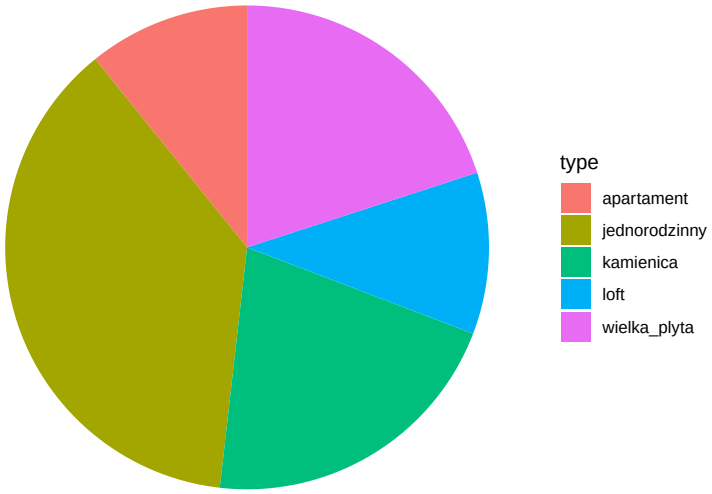
Wykres przedstawiający wydatki respondentów ze względu na płeć:

```
ggplot(data)+geom_boxplot(aes(x=plec,y=wydatki),outlier.colour="red", outlier.shape=8,
                outlier.size=4)
```



Widzimy, że męszczyźni średnio wydają więcej jednakże u kobiet mamy do czynienia z większym rozrzutem danych. Wykres kołowy przedstawiający rozkład osób miekszających w różnego typu budynkach:

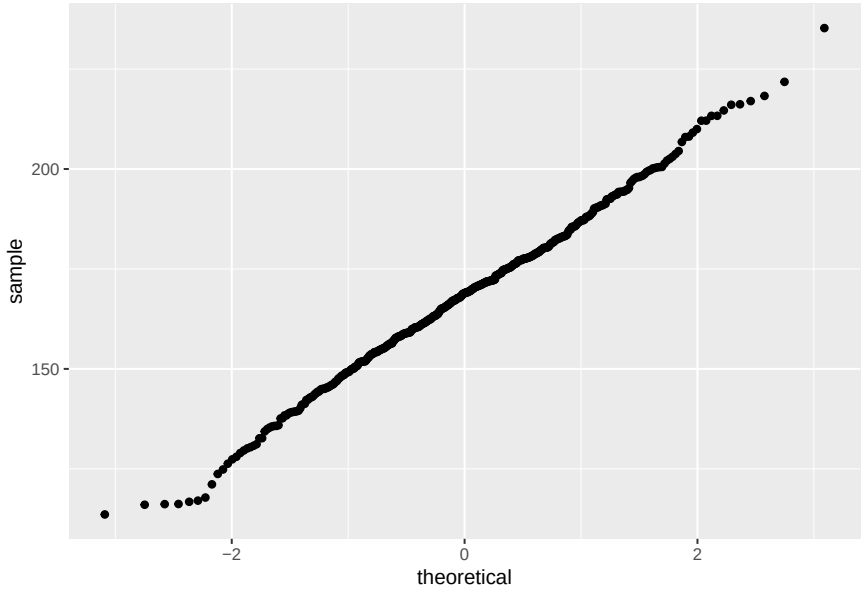
```
data$budynek=as.factor(data$budynek)
temp=count(data$budynek)
d=data.frame("val"=temp$freq,"type"=temp$x)
ggplot(d,aes(x="",y=val,fill=type))+geom_bar(stat="identity",width=1)+coord_polar("y",start=0)+theme_vo
```



3

Rozważmy teraz jaka jest p-wartość dla hipotezy, że średnia wzrostu to $m = 170$ cm. Wpierw zobaczymy jak rozkłada się wzrost wśród danych przy pomocy wykresu kwantylowego.

```
ggplot(data)+stat_qq(aes(sample=wzrost))+theme_grey()
```



Widzimy więc, że z bardzo dobrym przybliżeniem dane pochodzą z rozkładu normalnego. Do sprawdzenia hipotezy zerowej wystarczy wykorzystać test t-studenta.

```
mu_hip<-170 #średnia wartość wzrostu według hipotezy zerowej
med_hip<-165 #mediana wzorsty według hipotezy zerowej
x<-t.test(data$wzrost,mu=mu_hip,alternative="less")
```

Widzimy, że p-wartość wynosi 0.019487 a więc na poziomie istotności $0,05$ hipotezę zerową należy odrzucić. Aby przetestować medianę wykrozystamy test χ^2 Pearsona. Podzielmy wszystkie obserwację na populację

mniejsza od mediany i mniejszą od mediany. W takim układzie przy N obserwacjach oraz prawdziwej hipotezie zerowej że mediana rozkładu to m statystyka

$$a = \frac{(2N_1 - N)^2}{2N} + \frac{(2N_2 - N)^2}{2N} \tag{1}$$

ma rozkład χ^2 z 2 stopniami swobody gdzie N_1 to liczba obserwacji poniżej mediany a N_2 to liczba obserwacji powyżej mediany. Stąd w naszym przypadku

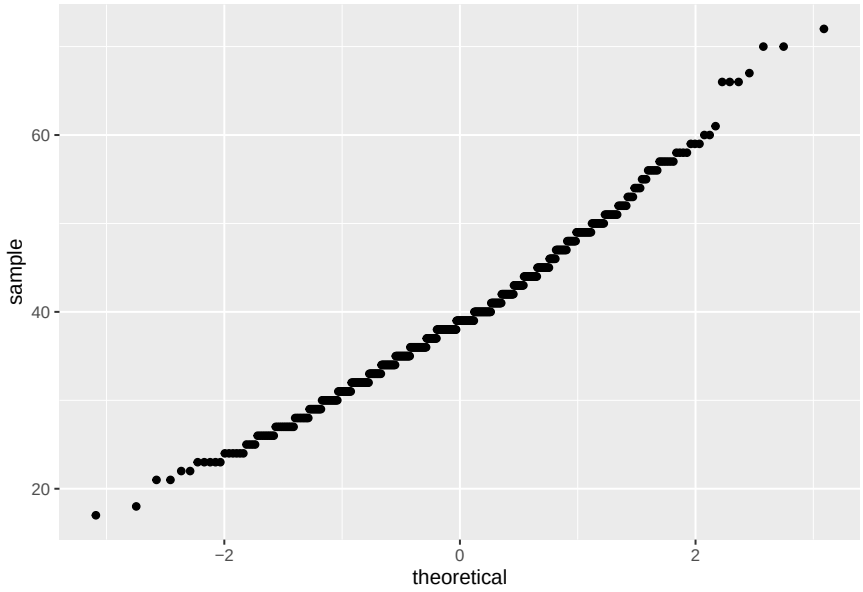
```
N1 <- sum(data$wzrost<=med_hip)
N2 <- sum(data$wzrost>med_hip)
N <- N1+N2
t <- (2*N1-N)^2/(2*N)+(2*N2-N)^2/(2*N)
k <- 1-pchisq(t,df=2)
```

Wartość statystyki testowej to 12.8 natomiast odpowiadająca jej p-wartość to 0.0016616. Stąd ponownie hipotezę zerową o medianie na poziomie istotności $0,05$ należy odrzucić. O ile do testu t-studenta potrzebujemy założenia o normalności rozkładu co jak widzimy z wykresu kwantylowego z dużą dokładnością jest spełnione o tyle w przypadku testu χ^2 Pearsona nie potrzebujemy tego założenia.

4

Przejdźmy teraz do obliczenia przedziałów ufności dla parametrów na poziomie $0,99$. W przypadku średniej przedział ufności. Zanim przejdziemy na wzozy szybko rzućmy okiem na rozkład kwantylowy danych.

```
ggplot(data)+stat_qq(aes(sample=wiek))+theme_grey()
```



Dane pochodzą z grubsza z rozkładu normalnego. W przypadku średniej i danych z rozkładu normalnego wiemy, że

$$T = \frac{X - \mu}{S\sqrt{\sigma}} \tag{2}$$

ma rozkład t-studenta (X oznacza średnią populacji a S odchylenie standardowe uzyksane estymatorem nieobciążonym). Chcemy zbadać, jaki jest przedział ufności dla statystyki T. wykorzystujac funckje R mamy, że

```
a<-0.99
c<-qt(1-a/2,df=length(data$wiek)-1)
```

Jeżeli T mieści się pomiędzy c a $-c$ to μ musi się mieścić pomiędzy $X - \frac{cS}{\sqrt{N}}$ oraz $X + \frac{cS}{\sqrt{N}}$.

```
up<-mean(data$wiek)+c*sd(data$wiek)/sqrt(length(data$wiek))
down<-mean(data$wiek)-c*sd(data$wiek)/sqrt(length(data$wiek))
```

Stąd przedział ufności dla μ to 39.4890339 do 39.4789661. W celu wyznaczenia przedziłów ufności dla wariancji wykorzystamy podobną metodę z tą różnicą, że zamiast wykrozystywać statystykę t studenta wykorzystamy statystykę χ^2. Wiemy albowiem, że statystyka

$$\frac{(N-1)S}{\sigma^2} \tag{3}$$

ma rozkład χ^2 o $N-1$ stopniach swobody. Stąd analogicznie znajdujemy wartości przedziałów dla statystyki

```
p<-qchisq(1-a/2,df=length(data$wiek)-1)
l<-qchisq(a/2,df=length(data$wiek)-1)
```

i po transformacjach znajdujemy jakie są przedziały ufności dla wariancji

```
N<-length(data$wiek)
lv<-(N-1)/p*var(data$wiek)
pv<-(N-1)/l*var(data$wiek)
```

Przedział ufności dla odchylenia standardowego to pierwiastek z tych granic a więc rozprzestrzenia się od 8.9787883 do 8.9859196.

5

1

2

3

Zbadajmy, czy stan cywilny jest niezależny od liczby dzieci. W tym celu wykorzystamy dokładny test Fishera. Wpierw musimy znaleźć macierz mówiącą nam, ile razy sklasyfikowane zostały poszczególne obserwacje razy.

```
a<-nrow(data[data$stan_cywilny==TRUE & data$plec=="M",]) # żonaci mężczyźni
b<-nrow(data[data$stan_cywilny==FALSE & data$plec=="M",]) #mężczyźni singlowie
c<-nrow(data[data$stan_cywilny==FALSE & data$plec=="K",]) # samotne kobiety
d<-nrow(data[data$stan_cywilny==TRUE & data$plec=="K",]) # zamężne kobiety
mat<-matrix(c(a,b,d,c),ncol=2,byrow=TRUE)
x<-fisher.test(mat)
```

p-wartość naszego testu to 0.1471494 a więc na żądanym poziomie istotności nie można stwierdzić, że istnieje zależność pomiędzy płcią a stanem cywilnym.