Krzyżowska_Anna_Projekt1

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
projekt<-read.csv("E:/Users/Ania/Documents/DOKUMENTY/Studia/rintro-chapter7.csv", header=TRUE,
                  sep=",",dec=".", quote = NULL)
attach(projekt)
```

```
#Przyjmuję że, dystans podawany jest w [km], a punkty w poszczególnych
#kategoriach zostały przyznawane od <0,100>.
#Zmienna 'overall'=ogólnie, uznaje to za punkty satysfkacji ogólnie z pobytu w Parku
projekt %>% rename(weekend=X.weekend, liczba_dzieci=X..num.child.., dystans=X..distance..,
                   jazdy=X..rides.., gry=X..games.., czas=X..wait.., czystosc=X..clean.., sat_ogolna=X.
```

Opisz w sposób syntetyczny, ilościowy zbiór danych. Oceń jego jakość.

```
summary(projekt1)
```

```
##  weekend   liczba_dzieci      dystans             jazdy            gry
##  no :259   Min.   :0.000   Min.   :  0.5267   Min.   : 72.00   Min.   : 57.00
##  yes:241   1st Qu.:0.000   1st Qu.: 10.3181   1st Qu.: 82.00   1st Qu.: 73.00
##            Median :2.000   Median : 19.0191   Median : 86.00   Median : 78.00
##            Mean   :1.738   Mean   : 31.0475   Mean   : 85.85   Mean   : 78.67
##            3rd Qu.:3.000   3rd Qu.: 39.5821   3rd Qu.: 90.00   3rd Qu.: 85.00
##            Max.   :5.000   Max.   :239.1921   Max.   :100.00   Max.   :100.00
##       czas          czystosc       sat_ogolna
##  Min.   : 40.0   Min.   : 74.0   Min.   :  6.00
##  1st Qu.: 62.0   1st Qu.: 84.0   1st Qu.: 40.00
##  Median : 70.0   Median : 88.0   Median : 50.00
##  Mean   : 69.9   Mean   : 87.9   Mean   : 51.26
##  3rd Qu.: 77.0   3rd Qu.: 91.0   3rd Qu.: 62.00
##  Max.   :100.0   Max.   :100.0   Max.   :100.00
```

```r
is.null(projekt1) #nie ma pustych wartości
```

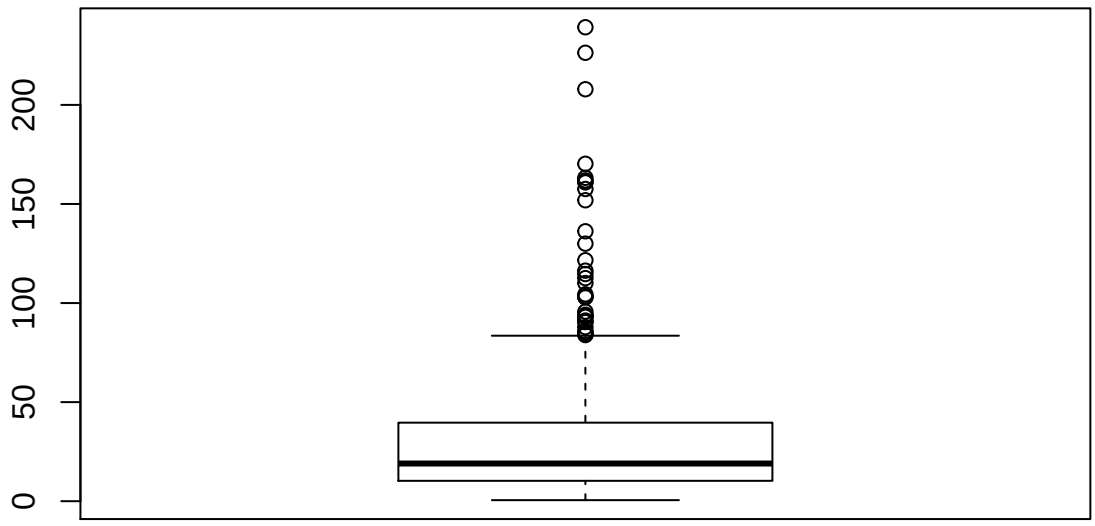
```
## [1] FALSE
```

```r
str(projekt1) #Występuje 500 obserwacji, jedna zmienna binarna
```

```
## 'data.frame':    500 obs. of  8 variables:
##  $ weekend      : Factor w/ 2 levels "no","yes": 2 2 1 2 1 1 2 1 1 2 ...
##  $ liczba_dzieci: int  0 2 1 0 4 5 1 0 0 3 ...
##  $ dystans      : num  114.6 27 63.3 25.9 54.7 ...
##  $ jazdy        : int  87 87 85 88 84 81 77 82 90 88 ...
##  $ gry          : int  73 78 80 72 87 79 73 70 88 86 ...
##  $ czas         : int  60 76 70 66 74 48 58 70 79 55 ...
##  $ czystosc     : int  89 87 88 89 87 79 85 83 95 88 ...
##  $ sat_ogolna   : int  47 65 61 37 68 27 40 30 58 36 ...
```

```r
#- weekend,zawiera 259 odpowiedzi "no" i 241"yes",
#pozostałe to zmienne numeryczne, oceny są całkowitoliczbowe.
#Uważam, że jakość zbioru jest dobra.
```

```r
boxplot(projekt1$dystans)
```

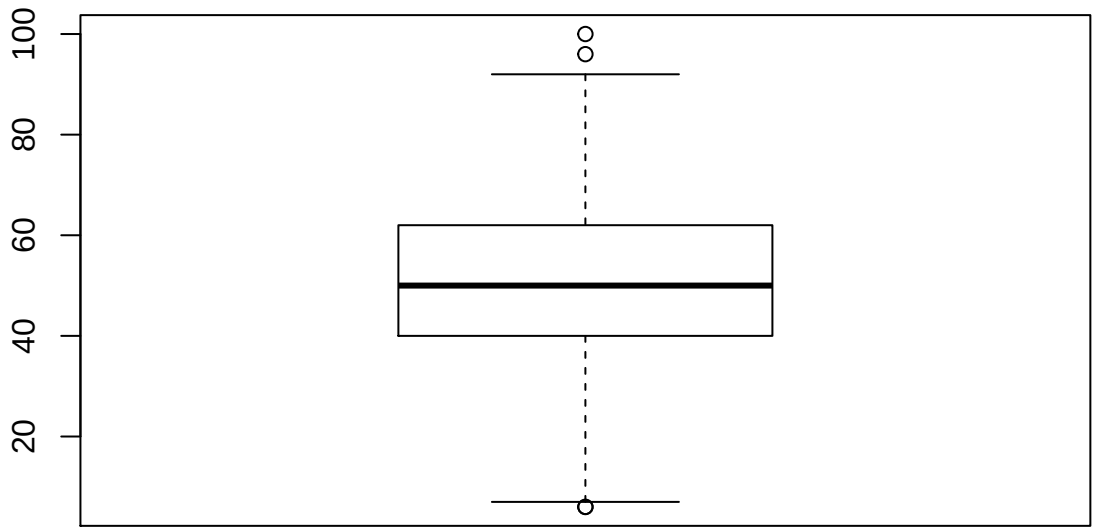


```r
projekt1 %>%
  summarise(sum(dystans),
            odchyleniedystans=sd(dystans),
            medianad=median(dystans),
            sredniad=mean(dystans))
```

```
##   sum(dystans) odchyleniedystans medianad sredniad
## 1     15523.75          33.14545 19.01909 31.04751
```

```r
#Występują outliers'y, jest to spowodowane tym, że zmienna dystans
#nie ma gónej granicy (ludzie mogą przyjechać a bardzo daleka)
#50% przebytego dystansu jest większe bądź równe 19,01[km], a pozostałe
#50% dystansu przebytego jest mniejszy, bądź równy 19,01[km]
#Łączdny dystnas jaki przebyli klienci do parku rozrywki wynosi 15523,75[km]
#Średnia przebyta droga przez klientów Parku wynosi 31.05[km]
#Dystans odchyla się średnio o 33.145[km] od średniej.
```

```r
boxplot(projekt1$sat_ogolna)
```




```r
projekt1 %>%
  summarise(sum(sat_ogolna),
            odchyleniesat=sd(sat_ogolna),
            medianas=median(sat_ogolna),
            srednias=mean(sat_ogolna))
```

```
##   sum(sat_ogolna) odchyleniesat medianas srednias
## 1           25629      15.87866       50   51.258
```

```
#występują dwie wartości odstające, w okolicach 90 i 100 [pkt].
#mediana jest na poziomie 48 [pkt]
#suma punktów uzyskanych przez park jako satysfakcja ogólna wynosi 25629
#50% uzyskanych punktów za satysfkacje z pobytu w parku jest większa bądź
#równa niż 50, a pozsotałe 50% jest mniejsza bądź równa niż 50pkt.
#Średnia liczba zdobytych punktów za satysfkacje wynosi 51,26pkt.
#punkty stysfakcji odchylają się się średnio o 15.88pkt od średniej
```

Sformułuj, zapisz i zweryfikuj hipotezę o niezależności zmiennych.

```
#Zamieniam sat_ogólna na skalę porządkową, przyjmuję ocene od <0-100>,
#dzielę na 4 oceny:
projekt1$ocena_kat[projekt1$sat_ogolna<=25] <- "zła"
projekt1$ocena_kat[projekt1$sat_ogolna>25 & projekt1$sat_ogolna<=50] <-"raczej zła"
projekt1$ocena_kat[projekt1$sat_ogolna>50 & projekt1$sat_ogolna<=75] <-"raczej dobra"
projekt1$ocena_kat[projekt1$sat_ogolna>75] <-"dobra"

#Czy zmienne weekend i ocena_kat są niezależne?
tab=table(projekt1$weekend, projekt1$ocena_kat)
tab
```

```
##
##       dobra raczej dobra raczej zła zła
##   no     15          118        117   9
##   yes    19           95        115  12
```

```
# H0: zmienne weekend i ocena_kat są niezależne
# H1: zmienne weekend i ocena_kat są zależne
chisq.test(tab)
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab
## X-squared = 2.7555, df = 3, p-value = 0.4309
```

```
# p>0,05 można sądzić,że zmienne weekend i ocena_kat są niezależne
#(brak podstaw do odrzucenia H0)
```

Sformułuj, zapisz i zweryfikuj hipotezę na równość średnich.

```
#Czy średni dystans jest taki sam dla każdego poziomu oceny satysfkacji?
#H0:Średnie w grupach są takie same
#H1:Średnie w grupach nie są takie same
#próba niezależna
group_by(projekt1, ocena_kat) %>%
summarise(
count = n(),
mean = mean(dystans, na.rm = TRUE),
sd = sd(dystans, na.rm = TRUE),
median=median(dystans,na.rm=TRUE))
```

```
## # A tibble: 4 x 5
##   ocena_kat    count  mean    sd median
##   <chr>        <int> <dbl> <dbl>  <dbl>
## 1 dobra           34  40.0  48.6   20.3
## 2 raczej dobra   213  32.3  34.0   20.7
## 3 raczej zła     232  29.2  30.3   17.0
## 4 zła             21  24.1  21.2   17.5
```

```
boxplot(projekt1$dystans~projekt1$ocena_kat, data = projekt1, varwidth=TRUE,  col="lightgrey",
        xlab="Ocena satysfkacji", ylab="Dystans")
#na wykresie widać, że mediany są na podobnym poziomie.

aa<-kruskal.test(dystans~factor(ocena_kat), data=projekt1)
aa
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  dystans by factor(ocena_kat)
## Kruskal-Wallis chi-squared = 2.5922, df = 3, p-value = 0.4589
```

```
#p-value >0,05, nie ma podstaw do odrzucenia H0
#średnie w grupach są takie same; nie istnieje przynajmniej jedna para, dla
#której różnica byłaby istotna

library(PMCMR)
```

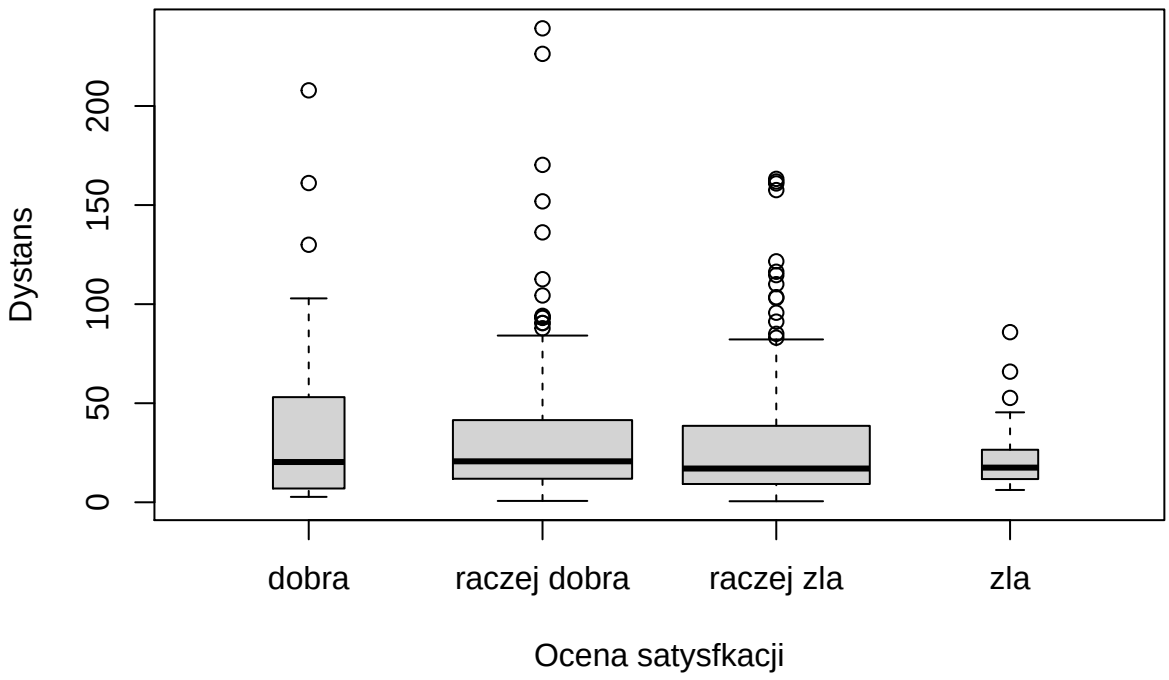
```
## Warning: package 'PMCMR' was built under R version 3.6.3
```

```
## PMCMR is superseded by PMCMRplus and will be no longer maintained. You may wish to install PMCMRplus
```

```
library(PMCMRplus)
```

```
## Warning: package 'PMCMRplus' was built under R version 3.6.3
```

```
## Registered S3 methods overwritten by 'PMCMRplus':
##   method        from
##   print.PMCMR   PMCMR
##   summary.PMCMR PMCMR
```




```
posthoc.kruskal.nemenyi.test(projekt1$dystans, factor(projekt1$ocena_kat), method="Tukey")
```

```
##
##  Pairwise comparisons using Tukey and Kramer (Nemenyi) test
##                    with Tukey-Dist approximation for independent samples

## data: projekt1$dystans and factor(projekt1$ocena_kat)

##              dobra raczej dobra raczej zła
## raczej dobra 0.98  -            -
## raczej zła   0.98  0.43         -
## zła          0.97  0.82         1.00

##
## P value adjustment method: none
```

```
#bardzo wysokie p-value, widać na macierzy, że nie isnieje na pewno ani jedna para, dla której
#różnica byłaby istotna
```