```r
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(dplyr)
library(ggplot2)
```

Četvrto istraživačko pitanje koje proučavamo glasi: "Postoji li razlika u broju izostanaka iz matematike između učenika koji dolaze iz manjih i većih obitelji"?

Na početku koristeći deskriptivnu statistiku i analizu pokušavamo dobiti početni uvid u podatke - provjeravamo veličine uzoraka, distribucije, stršeće vrijednosti itd.

```r
data <- read.csv("student_data.csv")
d4 <- data[c('famsize', 'absences_mat')]

print("Sažetak 5 brojeva za absences_mat")
```

```
## [1] "Sažetak 5 brojeva za absences_mat"
```

```r
print(summary(d4$absences_mat))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   4.000   5.381   8.000  75.000
```

```r
d4_GT3 <- d4[d4$famsize == 'GT3',c('absences_mat')]  # 266 primjera == veća obitelj
d4_LE3 <- d4[d4$famsize == 'LE3', c('absences_mat')]  # 104 primjera == manja obitelj
```

Stvorili smo dva odvojena uzorka: **d4_GT3** je uzorak izostanaka učenika iz većih obitelji, a **d4_LE3** učenika iz manjih obitelji. Uočavamo neravnomjernu raspodjelu - više od 70% učenika dolazi iz većih obitelji. Gledajući sažetak 5 brojeva vidimo da 75% svih učenika ima do 8 izostanaka.

Odmah ćemo izbaciti sve stršeće vrijednosti pa provjeriti sažetak 5 brojeva na odvojenim uzorcima.

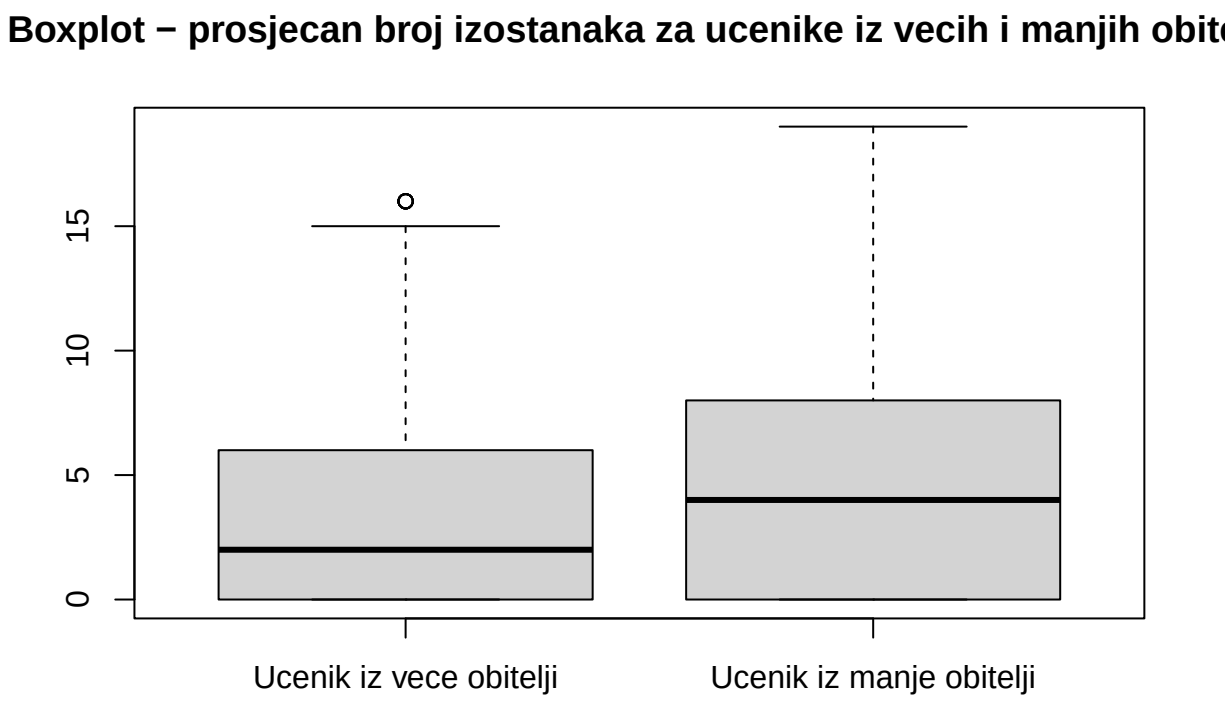
```r
# izbacujem outliere

granica_GT3 <- 1.5*IQR(d4_GT3)+quantile(d4_GT3, 0.75)  # granica = 16.875 ~ 17
granica_LE3 <- 1.5*IQR(d4_LE3)+quantile(d4_LE3, 0.75)  # granica = 19.5 ~ 20

d4_GT3 <- d4_GT3[d4_GT3 < round(granica_GT3)]
d4_LE3 <- d4_LE3[d4_LE3 < round(granica_LE3)]

# granica_GT3
# granica_LE3
# length(d4_GT3[d4_GT3 > granica_GT3]) - 14 outlier-a
# length(d4_LE3[d4_LE3 > granica_LE3]) - 3 outlier-a
```

```
boxplot(d4_GT3, d4_LE3,
        names = c('Učenik iz veće obitelji','Učenik iz manje obitelji'),
        main='Boxplot - prosječan broj izostanaka za učenike iz većih i manjih obitelji')
```




```
print("Sažetak 5 brojeva za absences_mat - veće obitelji")
```

```
## [1] "Sažetak 5 brojeva za absences_mat - veće obitelji"
```

```
print(summary(d4_GT3))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.000   3.869   6.000  16.000
```

```
print("Sažetak 5 brojeva za absences_mat - manje obitelji")
```

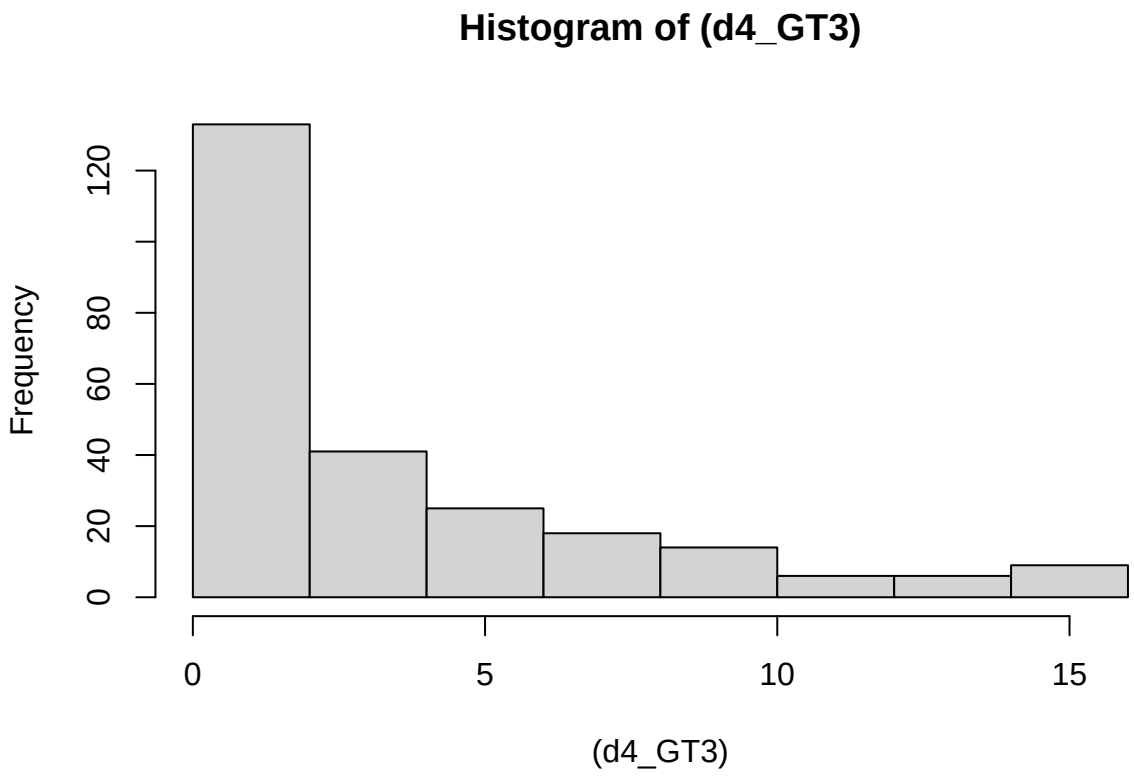
```
## [1] "Sažetak 5 brojeva za absences_mat - manje obitelji"
```

```
print(summary(d4_LE3))
```

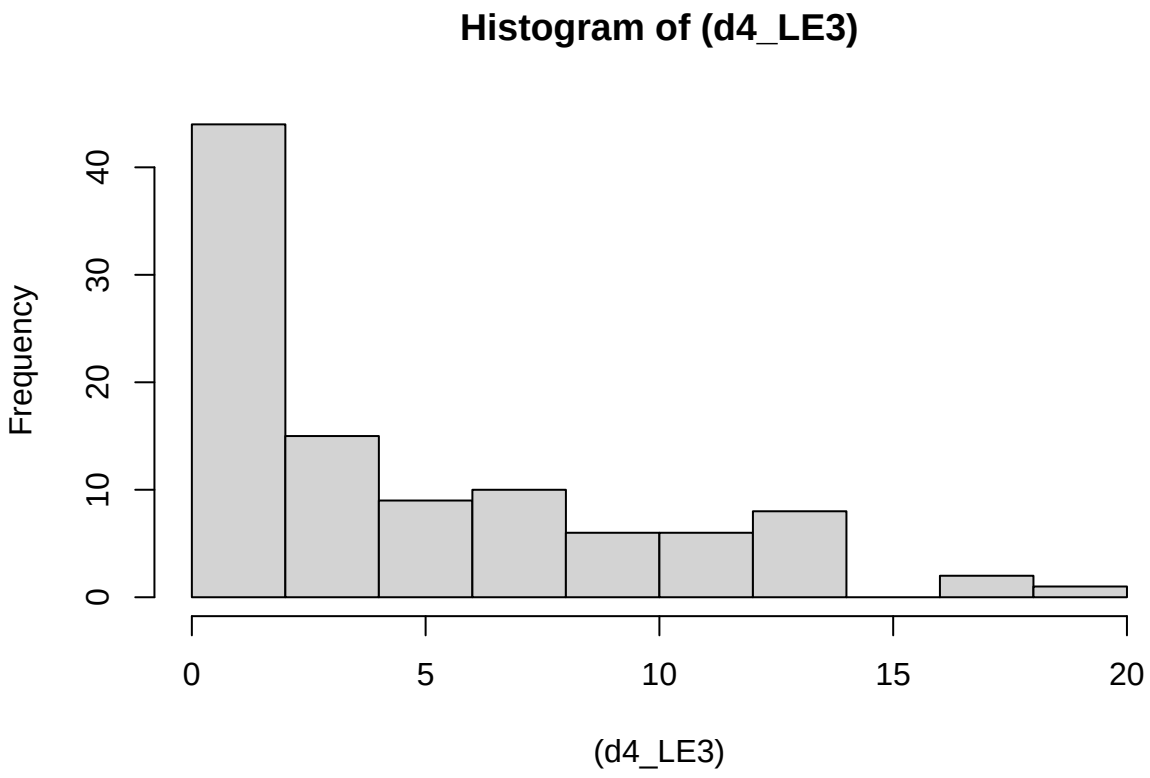
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   4.000   5.119   8.000  19.000
```

Nakon odbacivanja stršećih vrijednosti, vidimo da je srednja vrijednost izostanaka veća kod učenika iz manjih obitelji. Značajnost ove razlike provjerit ćemo statističkim testom.

```
hist((d4_GT3), breaks = 9)
```

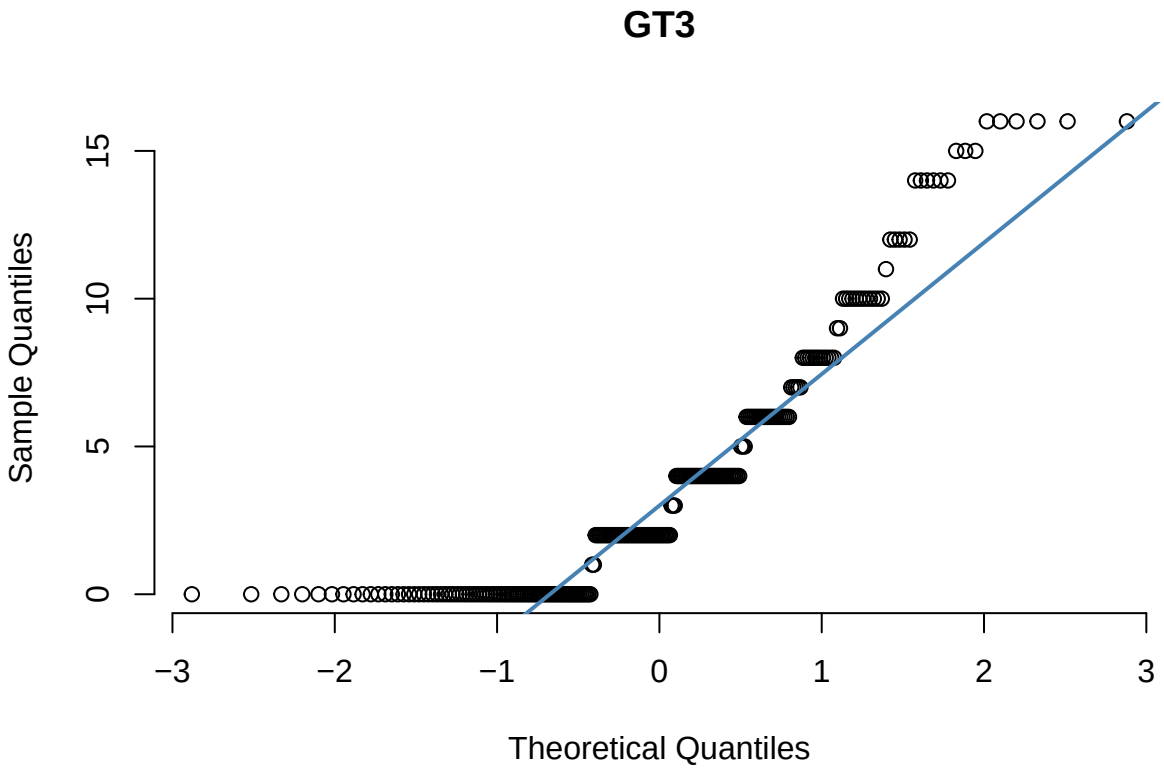



```
hist((d4_LE3), breaks = 7)
```

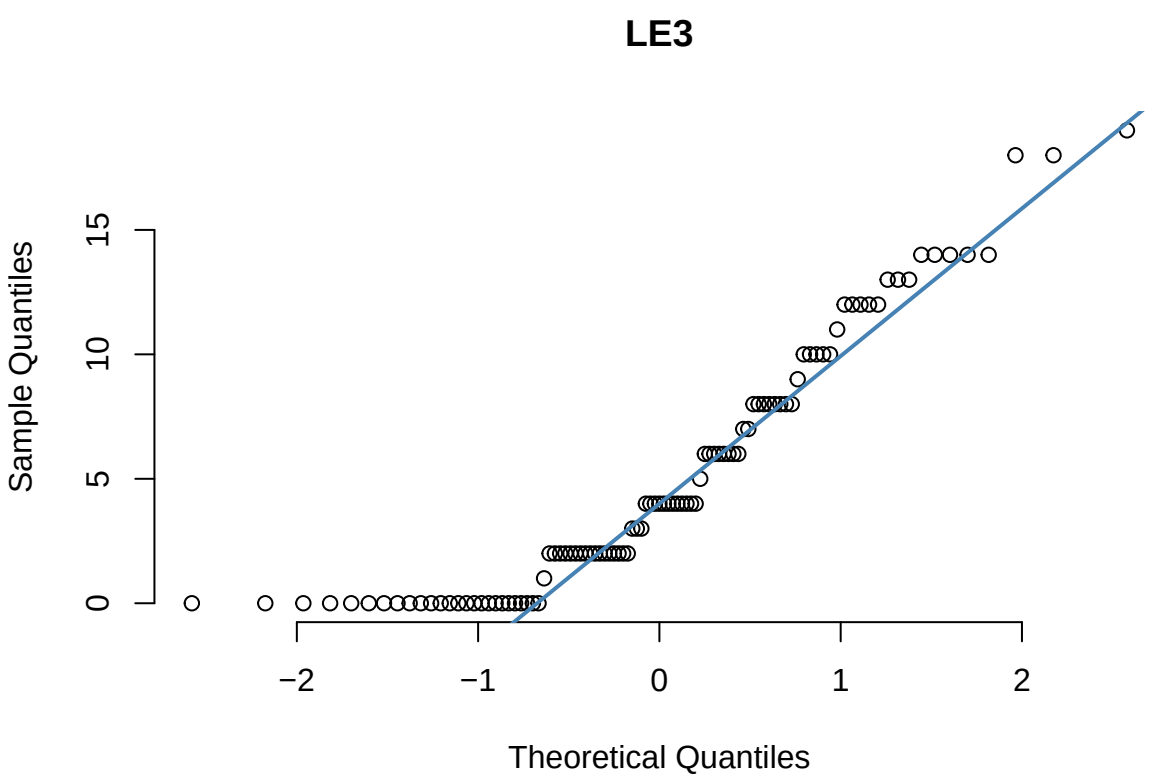



```
# varijable nisu normalno distribuirane (dominiraju studenti s malo izostanaka)
# length(d4_GT3[d4_GT3 <= 5]) # 177 / 252
# length(d4_LE3[d4_LE3 <= 5]) # 60 / 101
```

```
qqnorm(d4_GT3, pch = 1, frame = FALSE,main='GT3')
qqline(d4_GT3, col = "steelblue", lwd = 2)
```



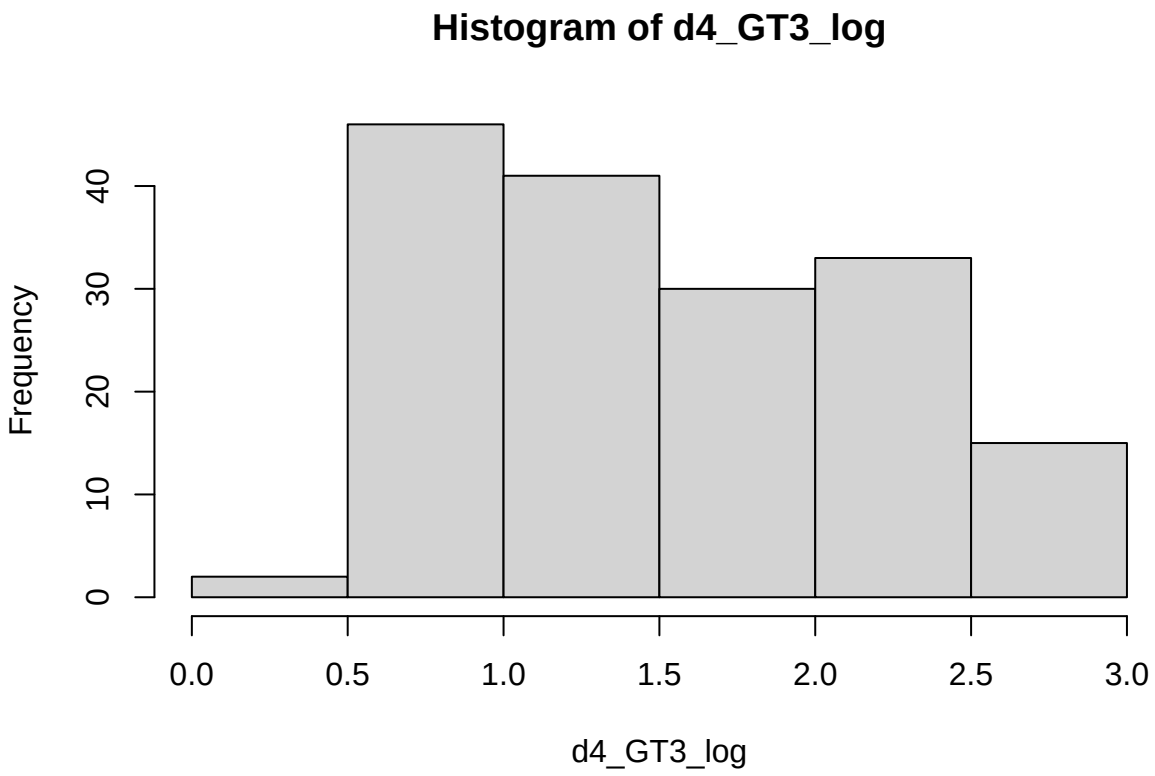

```
qqnorm(d4_LE3, pch = 1, frame = FALSE,main='LE3')
qqline(d4_LE3, col = "steelblue", lwd = 2)
```



Iz histograma i QQ-plotova vidimo da nemamo normalnu distribuciju. Razlog tome je što velik broj učenika nije imalo niti jedan izostanak. Možemo pokušati napraviti log-transformaciju da dobijemo nešto što je bliže normalnoj distribuciji, no nećemo dobiti ništa bolje.

```
d4_GT3_log <- log(d4_GT3)
d4_LE3_log <- log(d4_LE3)

hist(d4_GT3_log, breaks = 9)
```



Kako nam ključna pretpostavka t-testa nije zadovoljena, moramo pribjeći neparametarskim postupcima. Neparametarski testovi ne pretpostavljaju distribuciju, pa možemo koristiti podatke kojima raspolažemo bez transformacija.

Neparametarska alternativa t-testa jest Mann-Whitney U-test. Test se radi uz pretpostavke da su uzorci iz istih distribucija (što smo vidjeli na histogramu) te da su uzorci nezavisni (što je odrađeno dizajnom eksperimenta).

Nulta hipoteza jest da nema razlike između broja izostanaka učenika iz većih i manjih obitelji. Alternativna hipoteza jest da učenici iz većih obitelji imaju manji broj izostanaka (na što je upućivala srednja vrijednost uzoraka).

```
wilcox.test(d4_GT3, d4_LE3, alternative ="less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  d4_GT3 and d4_LE3
## W = 10947, p-value = 0.01806
## alternative hypothesis: true location shift is less than 0
```

Budući da je dobivena p-vrijednost manja od 0.05, vidimo da je razlika u srednjim vrijednostima statistički značajna, pa možemo odbaciti nultu hipotezu u korist hipoteze da učenici iz većih obitelji imaju manji broj izostanaka.