SAP - Analiza uspjeha učenika u školi

Lucija Burić, Barbara Cvitanović, Hana Ćerić, Borna Odobašić

2025-01-26

Uvod

U ovom projektu analiziramo podatke učenika dviju portugalskih škola. Korištenjem statističkih metoda i vizualizacije podataka pokušat ćemo otkriti veze između uspjeha učenika i varijabli poput spola, socioekonomskog statusa, navika učenja, prisutnosti i sl. Prije samog početka potrebno je učitati i proučiti podatkovni skup i prilagoditi ga da može biti na ispravan način korišten u analizi. Osim samog opisa varijabli i njihovog značenja, dobro je pogledati tipove varijabli podatkovnog skupa te nekoliko redaka kako bi se dobio bolji uvid u podatke s kojima se radi.

```
dataset <- read_csv("data/student_data.csv")
```

```
## Rows: 370 Columns: 39
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (18): school, sex, address, famsize, Pstatus, Mjob, Fjob, reason, guardi...
## dbl (21): age, Medu, Fedu, traveltime, studytime, failures_mat, failures_por...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(dataset)
```

```
## Rows: 370
## Columns: 39
## $ school       <chr> "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP", "GP~
## $ sex          <chr> "F", "F", "F", "F", "F", "M", "M", "F", "M", "M", "F", "F~
## $ age          <dbl> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15, 1~
## $ address      <chr> "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U", "U~
## $ famsize      <chr> "GT3", "GT3", "LE3", "GT3", "GT3", "LE3", "LE3", "GT3", "~
## $ Pstatus      <chr> "A", "T", "T", "T", "T", "T", "T", "A", "A", "T", "T", "T~
## $ Medu         <dbl> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, 3, ~
## $ Fedu         <dbl> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, 2, ~
## $ Mjob         <chr> "at_home", "at_home", "at_home", "health", "other", "serv~
## $ Fjob         <chr> "teacher", "other", "other", "services", "other", "other"~
## $ reason       <chr> "course", "course", "other", "home", "home", "reputation"~
## $ guardian     <chr> "mother", "father", "mother", "mother", "father", "mother~
## $ traveltime   <dbl> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, 1, ~
## $ studytime    <dbl> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, 1, ~
## $ failures_mat <dbl> 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, ~
## $ failures_por <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, ~
```

```
## $ schoolsup    <chr> "yes", "no", "yes", "no", "no", "no", "no", "yes", "no", ~
## $ famsup       <chr> "no", "yes", "no", "yes", "yes", "yes", "no", "yes", "yes~
## $ paid_mat     <chr> "no", "no", "yes", "yes", "yes", "yes", "no", "no", "yes"~
## $ paid_por     <chr> "no", "no", "no", "no", "no", "no", "no", "no", "no", "no~
## $ activities   <chr> "no", "no", "no", "yes", "no", "yes", "no", "no", "no", "~
## $ nursery      <chr> "yes", "no", "yes", "yes", "yes", "yes", "yes", "yes", "y~
## $ higher       <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes", "~
## $ internet     <chr> "no", "yes", "yes", "yes", "no", "yes", "yes", "no", "yes~
## $ romantic     <chr> "no", "no", "no", "yes", "no", "no", "no", "no", "no", "n~
## $ famrel       <dbl> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, 5, ~
## $ freetime     <dbl> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, 5, ~
## $ goout        <dbl> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, 5, ~
## $ Dalc         <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, ~
## $ Walc         <dbl> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, 4, ~
## $ health       <dbl> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, 5, ~
## $ absences_mat <dbl> 6, 4, 10, 2, 4, 10, 0, 6, 0, 0, 0, 4, 2, 2, 0, 4, 6, 4, 1~
## $ absences_por <dbl> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 10, 2, 2,~
## $ G1_mat       <dbl> 5, 5, 7, 15, 6, 15, 12, 6, 16, 14, 10, 10, 14, 10, 14, 14~
## $ G2_mat       <dbl> 6, 5, 8, 14, 10, 15, 12, 5, 18, 15, 8, 12, 14, 10, 16, 14~
## $ G3_mat       <dbl> 6, 6, 10, 15, 10, 15, 11, 6, 19, 15, 9, 12, 14, 11, 16, 1~
## $ G1_por       <dbl> 0, 9, 12, 14, 11, 12, 13, 10, 15, 12, 14, 10, 12, 12, 14,~
## $ G2_por       <dbl> 11, 11, 13, 14, 13, 12, 12, 13, 16, 12, 14, 12, 13, 12, 1~
## $ G3_por       <dbl> 11, 11, 12, 14, 13, 13, 13, 13, 17, 13, 14, 13, 12, 13, 1~
```

```
head(dataset)
```

```
## # A tibble: 6 x 39
##   school sex     age address famsize Pstatus  Medu  Fedu Mjob     Fjob    reason
##   <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl> <dbl> <chr>    <chr>   <chr>
## 1 GP     F        18 U       GT3     A           4     4 at_home  teacher course
## 2 GP     F        17 U       GT3     T           1     1 at_home  other   course
## 3 GP     F        15 U       LE3     T           1     1 at_home  other   other
## 4 GP     F        15 U       GT3     T           4     2 health   servic~ home
## 5 GP     F        16 U       GT3     T           3     3 other    other   home
## 6 GP     M        16 U       LE3     T           4     3 services other   reput~
## # i 28 more variables: guardian <chr>, traveltime <dbl>, studytime <dbl>,
## #   failures_mat <dbl>, failures_por <dbl>, schoolsup <chr>, famsup <chr>,
## #   paid_mat <chr>, paid_por <chr>, activities <chr>, nursery <chr>,
## #   higher <chr>, internet <chr>, romantic <chr>, famrel <dbl>, freetime <dbl>,
## #   goout <dbl>, Dalc <dbl>, Walc <dbl>, health <dbl>, absences_mat <dbl>,
## #   absences_por <dbl>, G1_mat <dbl>, G2_mat <dbl>, G3_mat <dbl>, G1_por <dbl>,
## #   G2_por <dbl>, G3_por <dbl>
```

Veliki broj varijabli stigao je u obliku znakovnog niza, no puno korisniji je u obliku faktora, odnosno kategorizirane vrste podataka.

```
dataset$school <- factor(dataset$school)
dataset$sex <- factor(dataset$sex)
dataset$address <- factor(dataset$address)
dataset$famsize <- factor(dataset$famsize)
dataset$Pstatus <- factor(dataset$Pstatus)
dataset$schoolsup <- factor(dataset$schoolsup)
dataset$famsup <- factor(dataset$famsup)
```

```
dataset$paid_mat <- factor(dataset$paid_mat)
dataset$paid_por <- factor(dataset$paid_por)
dataset$activities <- factor(dataset$activities)
dataset$nursery <- factor(dataset$nursery)
dataset$higher <- factor(dataset$higher)
dataset$internet <- factor(dataset$internet)
dataset$romantic <- factor(dataset$romantic)
```

```
missing_counts <- sapply(data, function(col) sum(is.na(col)))
```

```
## Warning in is.na(col): is.na() applied to non-(list or vector) of type 'symbol'

## Warning in is.na(col): is.na() applied to non-(list or vector) of type
## 'language'
## Warning in is.na(col): is.na() applied to non-(list or vector) of type
## 'language'

## Warning in is.na(col): is.na() applied to non-(list or vector) of type 'symbol'

## Warning in is.na(col): is.na() applied to non-(list or vector) of type
## 'language'
```

```
missing_counts
```

```
##       ...      list   package   lib.loc   verbose     envir overwrite
##         0         0         0         0         0         0         0         0
```

Vidimo da u našim podacima nemamo nedefiniranih vrijednosti, pa ne trebamo dodatno popunjavati ili čistiti dataset.

Rudimentarna deskriptivna analiza cijelog skupa dobro je pokrivena metodom *summary*. Pomoću te metode može se steći inicijalan uvod u vrijednosti i raspršenost podataka.

```
summary(dataset)
```

```
##  school   sex          age        address famsize   Pstatus      Medu
##  GP:331   F:195   Min.   :15.00   R: 81   GT3:266   A: 38   Min.   :0.0
##  MS: 39   M:175   1st Qu.:16.00   U:289   LE3:104   T:332   1st Qu.:2.0
##                   Median :17.00                             Median :3.0
##                   Mean   :16.58                             Mean   :2.8
##                   3rd Qu.:17.00                             3rd Qu.:4.0
##                   Max.   :22.00                             Max.   :4.0
##       Fedu           Mjob              Fjob              reason
##  Min.   :0.000   Length:370         Length:370         Length:370
##  1st Qu.:2.000   Class :character   Class :character   Class :character
##  Median :3.000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2.557
##  3rd Qu.:3.750
##  Max.   :4.000
##    guardian          traveltime      studytime      failures_mat
##  Length:370         Min.   :1.000   Min.   :1.000   Min.   :0.0000
```

```
##  Class :character   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:0.0000
##  Mode  :character   Median :1.000   Median :2.000   Median :0.0000
##                     Mean   :1.446   Mean   :2.043   Mean   :0.2784
##                     3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:0.0000
##                     Max.   :4.000   Max.   :4.000   Max.   :3.0000
##   failures_por    schoolsup famsup    paid_mat  paid_por  activities nursery
##  Min.   :0.0000   no :321   no :139   no :196   no :345   no :179    no : 72
##  1st Qu.:0.0000   yes: 49   yes:231   yes:174   yes: 25   yes:191    yes:298
##  Median :0.0000
##  Mean   :0.1324
##  3rd Qu.:0.0000
##  Max.   :3.0000
##  higher    internet  romantic      famrel         freetime         goout
##  no : 16   no : 57   no :251   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  yes:354   yes:313   yes:119   1st Qu.:4.000   1st Qu.:3.000   1st Qu.:2.000
##                                Median :4.000   Median :3.000   Median :3.000
##                                Mean   :3.935   Mean   :3.224   Mean   :3.116
##                                3rd Qu.:5.000   3rd Qu.:4.000   3rd Qu.:4.000
##                                Max.   :5.000   Max.   :5.000   Max.   :5.000
##       Dalc            Walc           health       absences_mat
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:3.000   1st Qu.: 0.000
##  Median :1.000   Median :2.000   Median :4.000   Median : 4.000
##  Mean   :1.484   Mean   :2.295   Mean   :3.562   Mean   : 5.381
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.: 8.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :75.000
##   absences_por        G1_mat          G2_mat          G3_mat
##  Min.   : 0.000   Min.   : 3.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.: 8.00   1st Qu.: 9.00   1st Qu.: 8.00
##  Median : 2.000   Median :11.00   Median :11.00   Median :11.00
##  Mean   : 3.632   Mean   :10.89   Mean   :10.75   Mean   :10.46
##  3rd Qu.: 6.000   3rd Qu.:13.00   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :32.000   Max.   :19.00   Max.   :19.00   Max.   :20.00
##      G1_por          G2_por          G3_por
##  Min.   : 0.00   Min.   : 5.00   Min.   : 0.00
##  1st Qu.:10.00   1st Qu.:11.00   1st Qu.:11.00
##  Median :12.00   Median :12.00   Median :13.00
##  Mean   :12.14   Mean   :12.27   Mean   :12.55
##  3rd Qu.:14.00   3rd Qu.:14.00   3rd Qu.:14.00
##  Max.   :19.00   Max.   :19.00   Max.   :19.00
```

1. Jesu li prosječne konačne ocjene iz matematike različite između spolova?

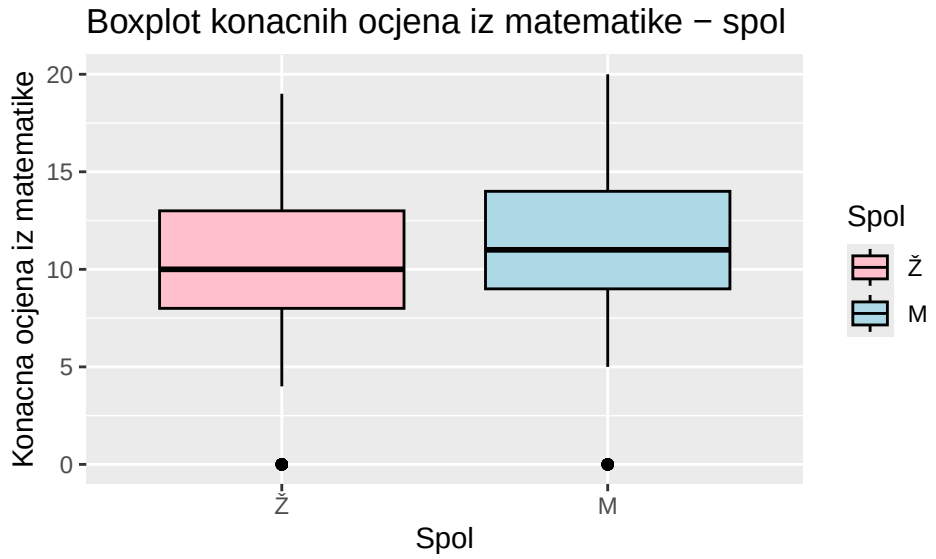
Prvo istraživačko pitanje ovog projekta bavi se razlikom prosječnih konačnih ocjena iz matematike između učenika i učenica, odnosno pitanjem postoji li između njih ikakva razlika. Prikladan test za ovakvo pitanje je t-test o jednakosti dviju sredina za dva promatrana uzorka: uzorak učenika i uzorak učenica, naravno ako možemo pretpostaviti uzorci dolaze iz normalne distribucije i da su nezavisni.

Na samom početku istraživačkog pitanja, postavit ćemo hipoteze.
H_0: Konačne ocjene iz matematike kod učenika i učenica su jednake, $\mu_{ocjeneM} = \mu_{ocjeneF}$.
H_1: Konačne ocjene iz matematike kod učenika i učenica nisu jednake, $\mu_{ocjeneM} \neq \mu_{ocjeneF}$.
Test ćemo provoditi na razini značajnosti od 5%.

Podatke značajne za samo pitanje treba vizualizirati. Promatramo vrijednosti varijable G3_mat koja je u predlošku o podacima definirana kao konačna ocjena iz matematike. Prvo ćemo prikazati *box-plot* potrebnih

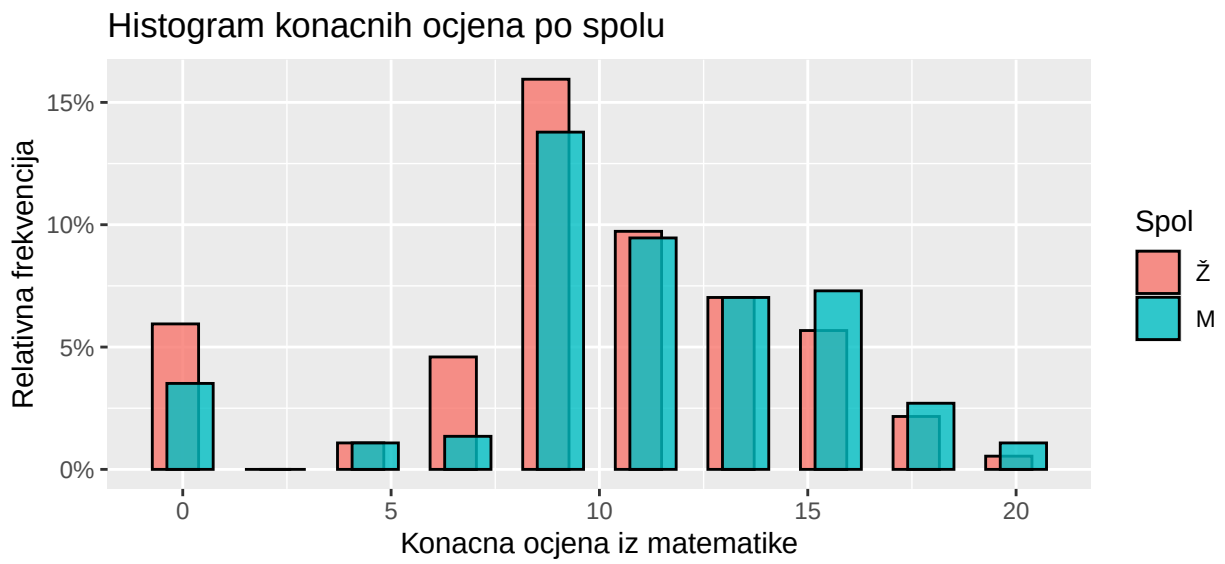
podataka, odnosno konačnih ocjena iz matematike ovisno o spolu kako bismo dobili inicijalan uvid u razdiobu podataka te potencijalnih outliera.

```
ggplot(dataset, aes(x = factor(sex), y = G3_mat, fill = factor(sex))) +
  geom_boxplot(color = "black") +
  scale_y_continuous(name = "Konačna ocjena iz matematike") +
  scale_x_discrete(name = "Spol", labels = c("Ž", "M")) +
  labs(title = "Boxplot konačnih ocjena iz matematike - spol") +
  scale_fill_manual(
    name = "Spol",
    labels = c("Ž", "M"),
    values = c("F" = "pink", "M" = "lightblue")
    )
```



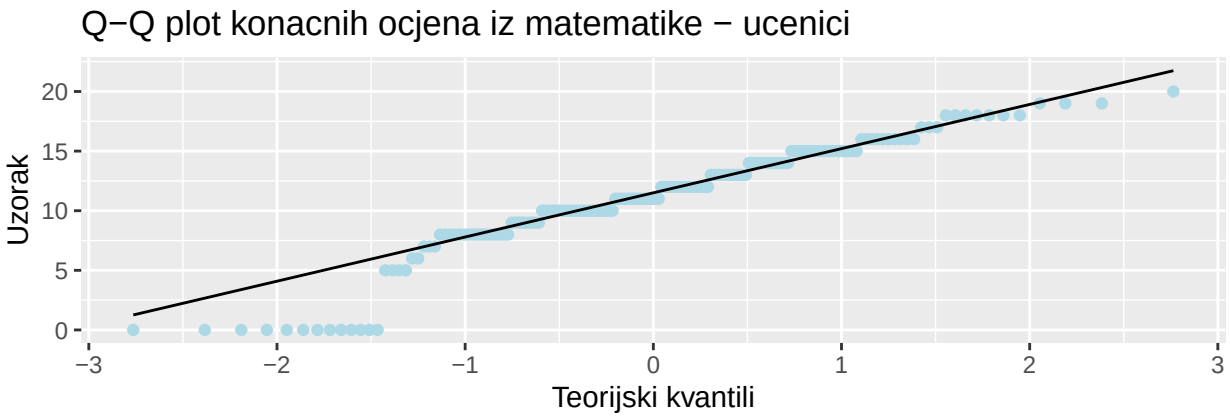
Pomoću box-plotova dobili smo kvartile i outliere. U oba spola problem predstavljaju ocjene vrijednosti 0. Ono što nas zanima je sam broj ocjena 0, odnosno jesu li ovi outlieri zanemarivi. Zato ćemo prikazati podatke histogramom s relativnim frekvencijama.

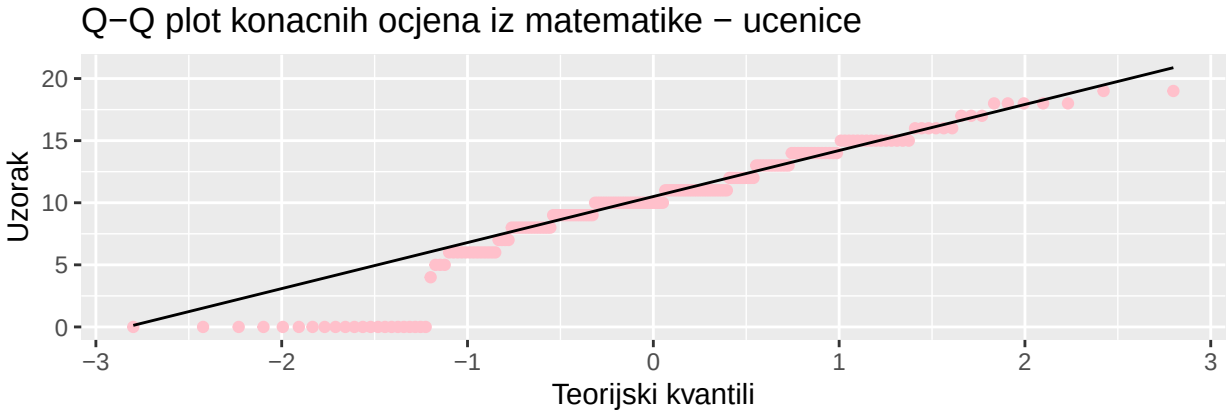
```
ggplot(dataset, aes(G3_mat, fill = sex)) +
  geom_histogram(
    aes(y = after_stat(count) / sum(after_stat(count))),
    bins = 10,
    colour = "black",
    position = position_dodge(width = 0.7),
    alpha = 0.8
  ) +
  scale_x_continuous(name = "Konačna ocjena iz matematike") +
  scale_y_continuous(
    name = "Relativna frekvencija",
    labels = scales::percent_format(),
    breaks = scales::pretty_breaks(5)
  ) +
  labs(title = "Histogram konačnih ocjena po spolu") +
  scale_fill_discrete(name = "Spol", labels = c("Ž", "M"))
```



Iz ovih histograma možemo vidjeti slične stvari kao i kod *box-plota*, ali isto tako i da je poveći broj ocjena 0 (kod učenica i preko 5%). Kako je uvjet provedbe ovakvog testa normalnost podataka, napravit ćemo i *Q-Q plot* pomoću kojeg bismo mogli donijeli odluku o tome pripadaju li podaci normalnoj distribuciji.

```
g1qq1 <- ggplot(dataset |> filter(sex == "M"), aes(sample = G3_mat)) +
  geom_qq(colour = "lightblue") +
  geom_qq_line(colour = "black") +
  scale_y_continuous(name = "Uzorak") +
  scale_x_continuous(name = "Teorijski kvantili") +
  labs(title = "Q-Q plot konačnih ocjena iz matematike - učenici")

g1qq2 <- ggplot(dataset |> filter(sex == "F"), aes(sample = G3_mat)) +
  geom_qq(colour = "pink") +
  geom_qq_line(colour = "black") +
  scale_y_continuous(name = "Uzorak") +
  scale_x_continuous(name = "Teorijski kvantili") +
  labs(title = "Q-Q plot konačnih ocjena iz matematike - učenice")

grid.arrange(g1qq1, g1qq2)
```





Po *Q-Q plotovima* možemo zaključiti kako je broj ocjena 0 prevelik i narušava normalnost podataka.

Još jedan način provjere normalnosti je i Lillieforsova inačica Kolmogorov-Smirnovljevog testa.

```
#učenici
lillie.test(dataset |> filter(sex == "M") |> pull(G3_mat))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pull(filter(dataset, sex == "M"), G3_mat)
## D = 0.12805, p-value = 2.172e-07
```

```
#učenice
lillie.test(dataset |> filter(sex == "F") |> pull(G3_mat))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pull(filter(dataset, sex == "F"), G3_mat)
## D = 0.13483, p-value = 3.273e-09
```

Na temelju Lillieforsove inačice KS testa vidimo da su p-vrijednosti iznimno male i možemo zaključiti da distribucija ne podilazi normalnoj. S obzirom na sve dosad prikazano, pokušat ćemo ponovno uz izbacivanje svih ocjena 0. To ćemo napraviti s opravdanjem da su te vrijednosti OUTLIERI u našim boxplotovima.

```
dataset |> filter(sex == "M") |> count(G3_mat == 0)
```

```
## # A tibble: 2 x 2
##   `G3_mat == 0`     n
##   <lgl>         <int>
## 1 FALSE           162
## 2 TRUE             13
```

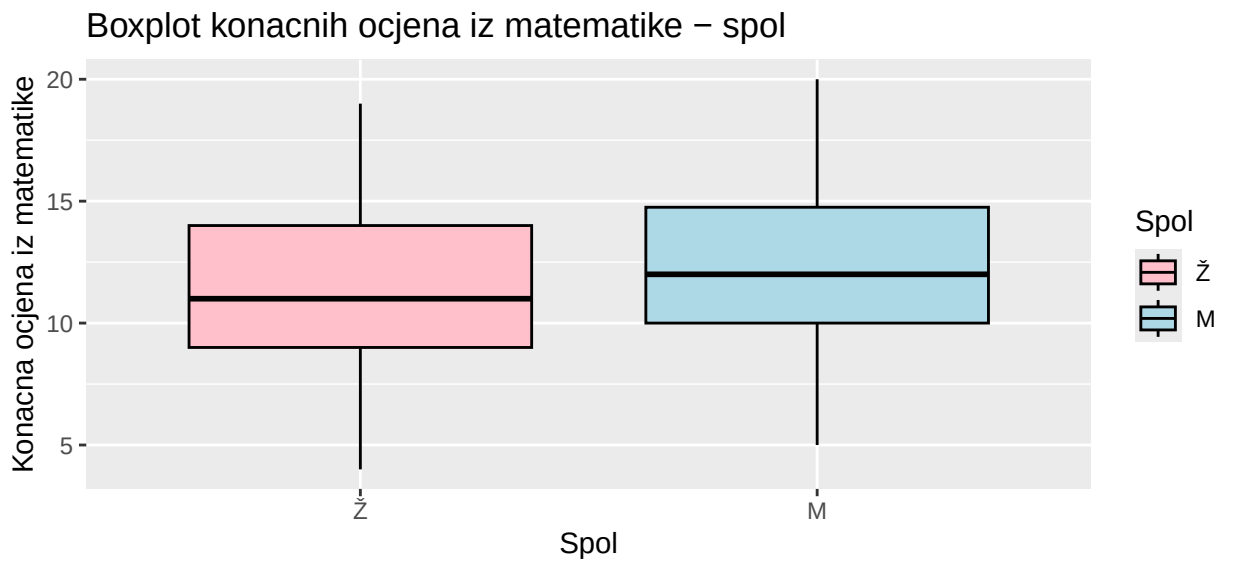
```
dataset |> filter(sex == "F") |> count(G3_mat == 0)
```

```
## # A tibble: 2 x 2
##   `G3_mat == 0`     n
##   <lgl>         <int>
## 1 FALSE           173
## 2 TRUE             22
```

Vidimo da kod učenica čak 22 imaju ocjenu 0 iz završnog ispita iz matematike, dok je kod učenika taj broj 13. To ćemo imati na umu kod konačnog rezultata ukoliko je potrebno.

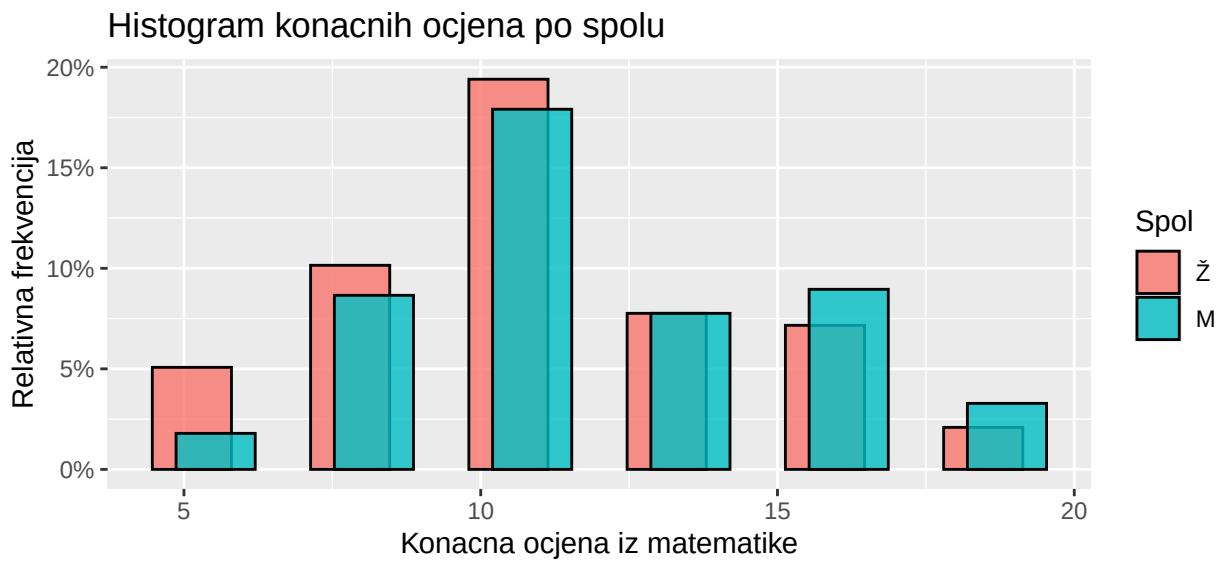
Pogledajmo sada *box-plot* ocjena bez nula.

```
df1 <- dataset
df1 <- df1 |> filter(G3_mat != 0)

ggplot(df1, aes(x = factor(sex), y = G3_mat, fill = factor(sex))) +
  geom_boxplot(color = "black") +
  scale_y_continuous(
    name = "Konačna ocjena iz matematike",
    breaks = scales::pretty_breaks()
    ) +
  scale_x_discrete(name = "Spol", labels = c("Ž", "M")) +
  labs(title = "Boxplot konačnih ocjena iz matematike - spol") +
  scale_fill_manual(
    name = "Spol",
    labels = c("Ž", "M"),
    values = c("F" = "pink", "M" = "lightblue")
    )
```



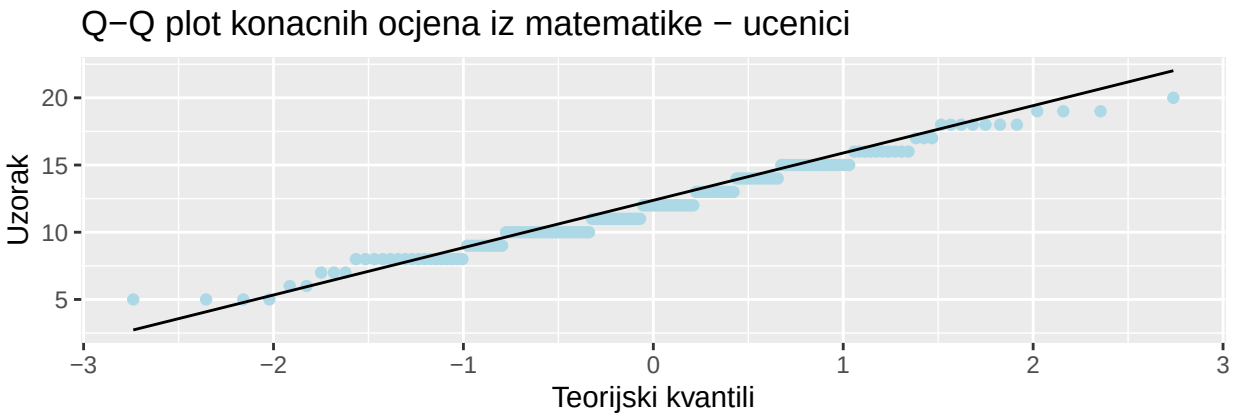
Box-plotovi su sada eliminirali ocjene 0 kao outliere, dapače, više nemamo ouliere i čini se da je ovaj rezultat nešto bliži traženom. Također, naizgled se oblik box-plota nije promijenio, pa se čini kao da nismo previše narušili podatke izbacivanjem 0.

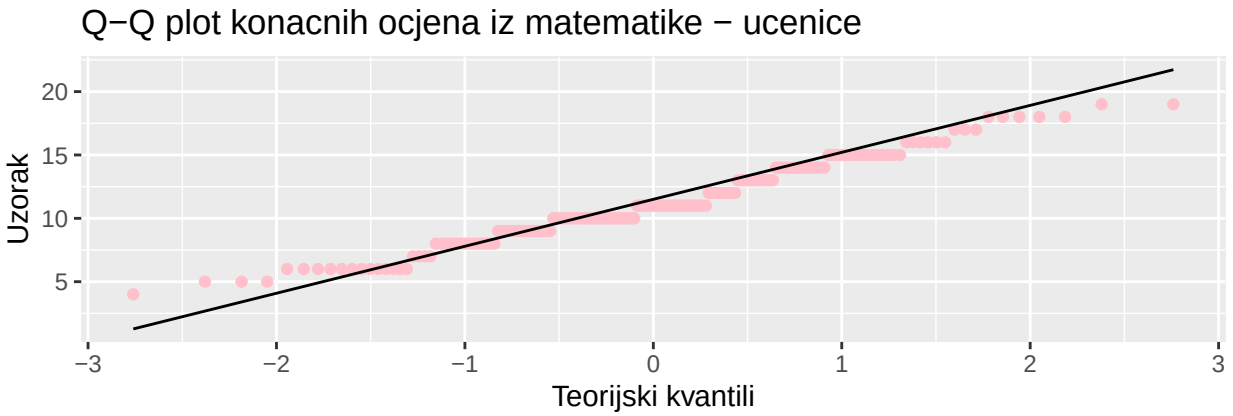
```
ggplot(df1, aes(G3_mat, fill = sex)) +
  geom_histogram(
    aes(y = after_stat(count) / sum(after_stat(count))),
    bins = 7,
    colour = "black",
    position = position_dodge(width = 0.8),
    alpha = 0.8
  ) +
  scale_x_continuous(name = "Konačna ocjena iz matematike") +
  scale_y_continuous(
    name = "Relativna frekvencija",
    labels = scales::percent_format(),
    breaks = scales::pretty_breaks(5)
  ) +
  labs(title = "Histogram konačnih ocjena po spolu") +
  scale_fill_discrete(name = "Spol", labels = c("Ž", "M"))
```



Histogrami sada pokazuju puno ljepšu distribuciju, ali ćemo dodatno napraviti i *Q-Q plotove*.

```
gq1 <- ggplot(df1 |> filter(sex == "M"), aes(sample = G3_mat)) +
  geom_qq(colour = "lightblue") +
  geom_qq_line(colour = "black") +
  scale_y_continuous(name = "Uzorak") +
  scale_x_continuous(name = "Teorijski kvantili") +
  labs(title = "Q-Q plot konačnih ocjena iz matematike - učenici")

gq2 <- ggplot(df1 |> filter(sex == "F"), aes(sample = G3_mat)) +
  geom_qq(colour = "pink") +
  geom_qq_line(colour = "black") +
  scale_y_continuous(name = "Uzorak") +
  scale_x_continuous(name = "Teorijski kvantili") +
  labs(title = "Q-Q plot konačnih ocjena iz matematike - učenice")

grid.arrange(gq1, gq2)
```





Izbacivanjem nula podaci su puno bliže *Q-Q liniji* što je bitno za uvjet normalnosti kako bi se proveo *t-test* nad ovim podacima. Ponovno ćemo provesti Lillieforsov test.

```
#učenici
lillie.test(df1 |> filter(sex == "M") |> pull(G3_mat))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pull(filter(df1, sex == "M"), G3_mat)
## D = 0.098916, p-value = 0.0005437
```

```
#učenice
lillie.test(df1 |> filter(sex == "F") |> pull(G3_mat))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  pull(filter(df1, sex == "F"), G3_mat)
## D = 0.13116, p-value = 1.11e-07
```

Vidimo da se Lillieforsovim testom p-vrijednost povećala za nekoliko redova veličine, iako je p-vrijednost mala, provest ćemo t-test na temelju prethodnih histograma i *Q-Q plota* i time pretpostavljamo normalnost naših uzoraka. Lillieforsova inačica KS test-a vrlo je strog statistički test koji će vrlo često odbaciti pretpostavku normalnosti, čak i za najmanja odstupanja od normalne razdiobe.

Zadnja stvar koju treba napraviti prije samog *t-testa* je provjeriti jednakost varijanci jer se sama provedba testa razlikuje oko pretpostavke jednakosti ili nejednakosti varijanci.

Za provođenje koristit ćemo *F-test* na razini značajnosti od 5%.
Prvo ćemo izračunati varijance oba uzorka, a zatim provesti *F-test*.

Hipoteze *F-testa*:
H_0: Varijance konačnih ocjena kod oba spola su jednake
H_1: Varijance ocjena kod oba spola nisu jednake

```
maleGrades <- df1 |> filter(sex == "M") |> pull(G3_mat)
femaleGrades <- df1 |> filter(sex == "F") |> pull(G3_mat)

var.test(maleGrades, femaleGrades, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  F test to compare two variances
##
## data:  maleGrades and femaleGrades
## F = 1.0171, num df = 161, denom df = 172, p-value = 0.9118
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7503026 1.3808785
## sample estimates:
## ratio of variances
##           1.017087
```

Dobiveni omjer varijanci je ~1.017, a 95%-tni interval je između 0.75 i 1.38. Upadanjem u sami interval i dobivenom p-vrijednosti od 0.9118, nećemo odbaciti H_0. S tim saznanjem možemo provesti *t-test* o jednakosti sredina kod oba uzorka s jednakim varijancama (uz pretpostavku jednakih varijanci).

```
t.test(maleGrades, femaleGrades, alternative = "two.sided", var.equal = T, conf.level = 0.95)
```

```
##
##  Two Sample t-test
##
## data:  maleGrades and femaleGrades
## t = 2.3647, df = 333, p-value = 0.01862
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.1418203 1.5452560
## sample estimates:
## mean of x mean of y
##  11.99383  11.15029
```

Nakon provedenog testa, vidimo da je dobivena vrijednost 2.36 i pripadna p-vrijednost 0.0186. Isto tako možemo vidjeti i značajno veću sredinu uzorka učenika. S tim podacima jasno je da se odbacuje H_0 hipoteza na razini značajnosti od 5%. Ne smijemo zaboraviti kako smo kod odstranjivanja ocjena nula odbacili više ocjena kod učenica. Time nam ovi rezultati još više idu u prilog te gotovo sigurno možemo reći da su učenici uspješniji od učenica.

2. Postoji li razlika u prvoj ocjeni iz matematike s obzirom na mjesto stanovanja učenika?

U drugom istraživačkom pitanju bavimo se utvrđivanjem postoji li razlika u prvoj ocjeni iz matematike obzirom na mjesto stanovanja. Varijabla *address* sadrži podatak o tome živi li učenik u urbanom ili ruralnom području.

```
addr <- dataset |> select(address)
#broj učenika iz ruralnih/urbanih područja - vidljivo u tablici
table(addr)
```

```
## address
##   R   U
##  81 289
```

Sada ćemo kategorizirati ocjene. U Portugalu se u sustavu ocjenjivanja koristi skala od 0 do 20, ali rezultatima se pridjeljuju određene ocjene: *Mau*, *Mediocre*, *Suficiente*, *Bom*, *Muito Bom*, *Excelente* i *Muito bom con distincao e louvor*. Zato ćemo rezultatima pridijeliti ocjene na skali od 1 do 7 kako bismo napravili analizu. Napomena: ocjenu *Excelente* i *Muito bom con distincao e louvor* spojit ćemo u jednu najvišu ocjenu jer najviša ocjena ne sadrži učenike u toj kategoriji.

```
df2 <- dataset

df2 <- df2 |> mutate(Grade = ifelse(G1_mat < 7, 1, ifelse(G1_mat < 10, 2 , ifelse(G1_mat < 14, 3, ifels
```

Za ovakav zadatak htjeli bismo provesti *Hi-kvadrat test nezavisnosti*. Hipoteze testa:
H_0: Razlike u prvoj ocjeni iz matematike s obzirom na mjesto stanovanja ne postoje, tj. prva ocjena iz matematike ne ovisi o mjestu stanovanja učenika.
H_1: Postoje razlike u prvoj ocjeni iz matematike s obzirom na mjesto stanovnja učenika, tj. prva ocjena iz matematike ovisi o mjestu stanovanja.
Test ćemo provesti na razini značajnosti od 5%.

Izrađujemo kontingencijsku tablicu.

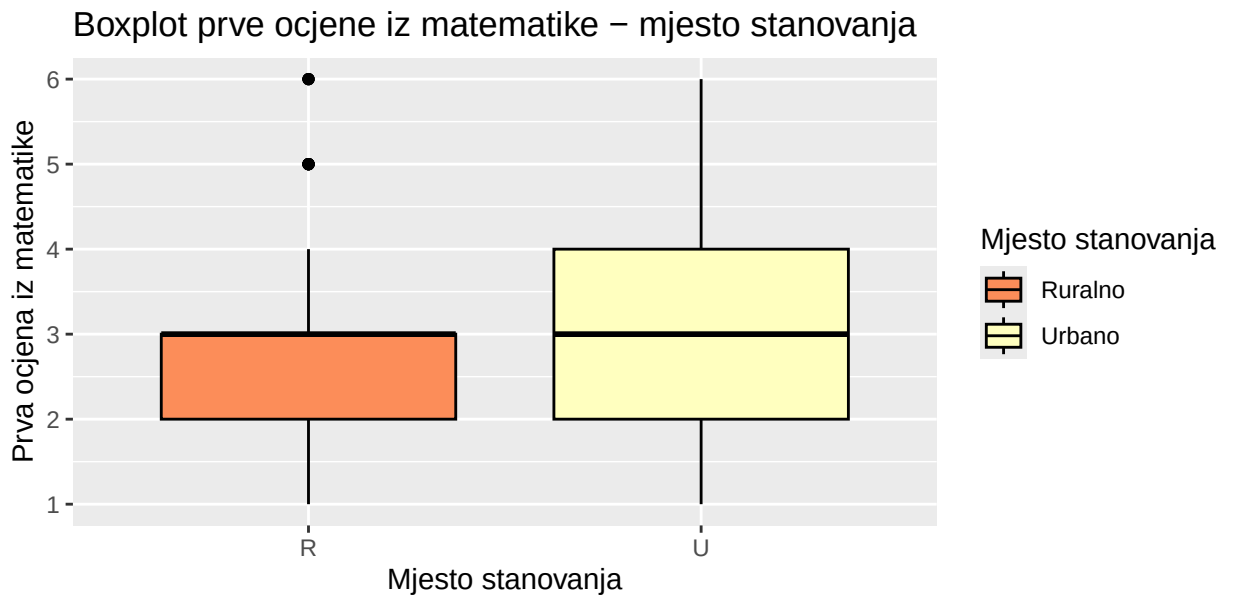
```
grade_table <- table(df2$address, df2$Grade)
added_margins_tbl = addmargins(grade_table)
added_margins_tbl
```

```
##
##         1   2   3   4   5   6 Sum
##   R     8  26  31   7   6   3  81
##   U    24  76 116  41  24   8 289
##   Sum  32 102 147  48  30  11 370
```

Ponovno ćemo prvo napraviti *box-plot* dijagrame kako bismo dobili osnovni pogled na distribuciju.

```
ggplot(df2, aes(x = factor(address), y = Grade, fill = factor(address))) +
  geom_boxplot(color = "black") +
  scale_y_continuous(name = "Prva ocjena iz matematike",
  breaks = scales::pretty_breaks()
  ) +
  labs(
    title = "Boxplot prve ocjene iz matematike - mjesto stanovanja"
    ) +
```

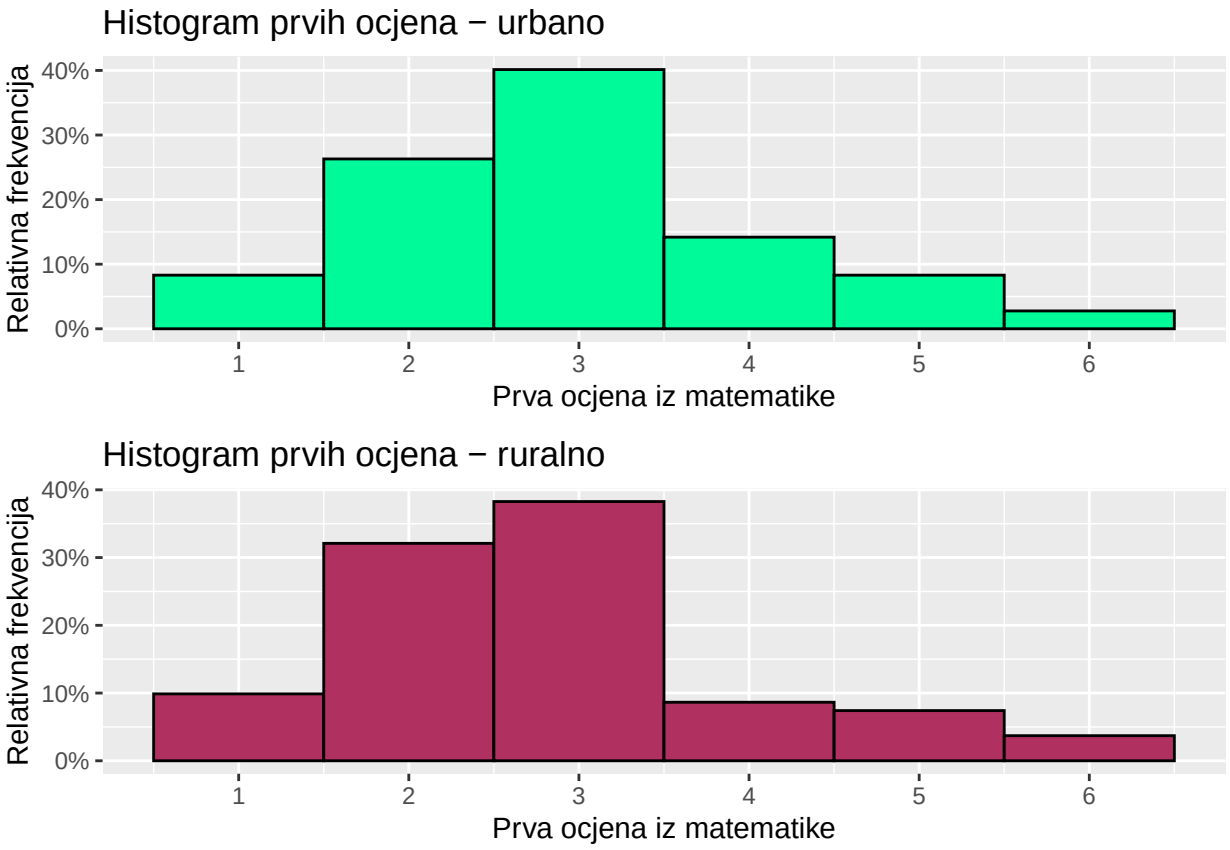
```
  scale_x_discrete(name = "Mjesto stanovanja") +
  scale_fill_brewer(
    name = "Mjesto stanovanja",
    labels = c("Ruralno", "Urbano"),
    palette = "Spectral"
    )
```



Iz ovih *box-plotova* možemo odmah vidjeti da je raspršenje ocjena veće kod učenika iz urbanih područja, dok kod učenika iz ruralnih područja vidimo dosta skupljene rezultate, tako da su i generirale nešto više outliera. Koliko su značajni, ponovno ćemo provjeriti pomoću histograma.

```
g2hist1 <- ggplot(df2 |> filter(address == "U"), aes(Grade)) +
  geom_histogram(
    aes(y = after_stat(count) / sum(after_stat(count))),
    bins = 6,
    colour = "black",
    fill = "mediumspringgreen"
  ) +
  scale_x_continuous(
    name = "Prva ocjena iz matematike",
    breaks = scales::pretty_breaks()
    ) +
  scale_y_continuous(
    name = "Relativna frekvencija",
    labels = scales::percent_format(),
    breaks = scales::pretty_breaks()
  ) +
  labs(title = "Histogram prvih ocjena - urbano")


g2hist2 <- ggplot(df2 |> filter(address == "R"), aes(Grade)) +
  geom_histogram(
    aes(y = after_stat(count) / sum(after_stat(count))),
    bins = 6,
```

```
    colour = "black",
    fill = "maroon"
  ) +
  scale_x_continuous(
    name = "Prva ocjena iz matematike",
    breaks = scales::pretty_breaks()
    ) +
  scale_y_continuous(
    name = "Relativna frekvencija",
    labels = scales::percent_format(),
    breaks = scales::pretty_breaks()
  ) +
  labs(title = "Histogram prvih ocjena - ruralno")

grid.arrange(g2hist1, g2hist2)
```



Iz priloženih histograma, dalo bi se naslutiti da su relativni udjeli podataka prilično slični, no prije provedbe samog testa, potrebno je još izračunati očekivane frekvencije. Uvjet za provedbu *Hi-kvadrat testa nezavisnosti* je da su očekivane frekvencije **svake kategorije** barem 5.

```
row_totals <- margin.table(grade_table, 1)
col_totals <- margin.table(grade_table, 2)
grand_total <- sum(grade_table)

expected_frequencies <- outer(row_totals, col_totals) / grand_total
expected_frequencies
```

```
##
##             1        2         3        4         5        6
##   R  7.005405 22.32973  32.18108 10.50811  6.567568 2.408108
##   U 24.994595 79.67027 114.81892 37.49189 23.432432 8.591892
```

Prema navedenim frekvencijama, izvjesno je da nemaju sve kategorije očekivanu frekvenciju veću ili jednaku 5. U tom slučaju koristit ćemo *Fischer-Irwin egzaktni test.*

```
fisher.test(grade_table, conf.level = 0.95)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  grade_table
## p-value = 0.7117
## alternative hypothesis: two.sided
```

Prema provedenom Fisherovom egazaktnom testu, dobili smo p-vrijednost od čak 0.71. Dakako, to znači da se H_0 ne može odbaciti pa su prve ocjene iz matematike neovisne o tome odakle učenik dolazi, što je samo dodatna potvrda viđenog na histogramu s relativnim frekvencijama.

3. Možemo li predvidjeti prolaz iz završnog ispita iz jezika na temelju sociodemografskih varijabli poput spola, obrazovanja roditelja i veličine obitelji?

U ovom istraživačkom pitanju zanima nas predikcija prolaza iz portugalskog jezika na temelju određenih varijabli, stoga ćemo koristiti model logističke regresije koji ćemo trenirati nad našim podacima. Podaci su na početku već prilagođeni za testiranje, jedino kako ne postoji zapis o samom prolazu (ocjena veća od 10), moramo u tablicu dodati još podatak o tome je li učenik prošao ili pao na završnom ispitu iz jezika. Upravo ti podaci bit će ključni za učenje modela, metodom procjene najveće izglednosti.

```
df3 <- dataset |> mutate(pass_por = ifelse(G3_por >= 10, 1 , 0)) |>
  mutate(pass_por = factor(pass_por, levels = c(0, 1)))
```

```
sum(df3$pass_por == 1)
```

```
## [1] 341
```

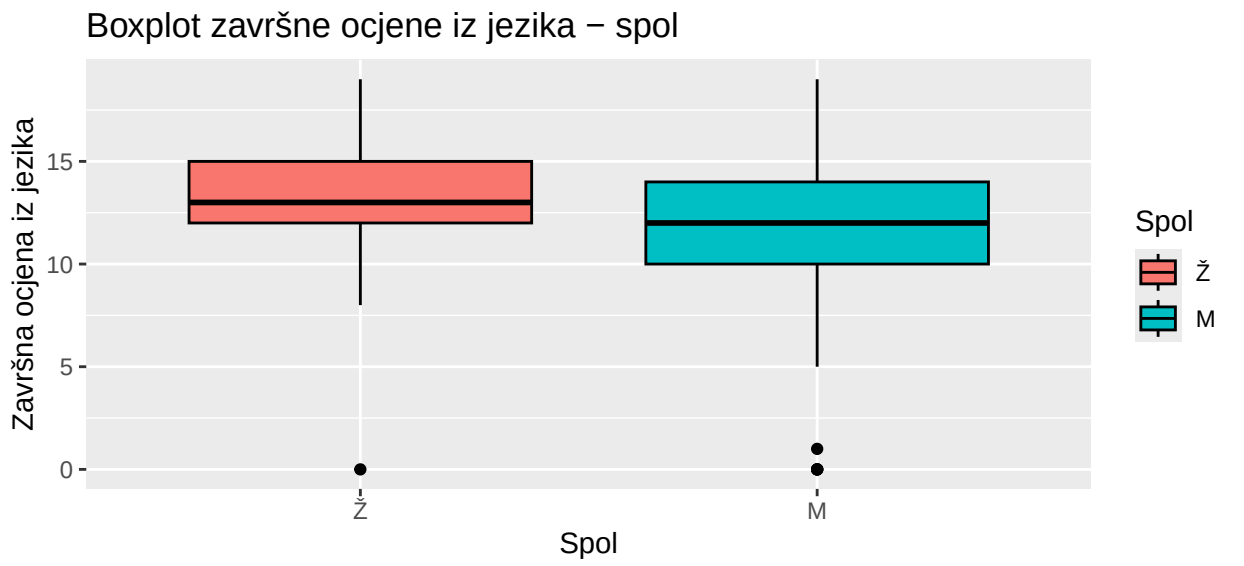
```
sum(df3$pass_por == 0)
```

```
## [1] 29
```

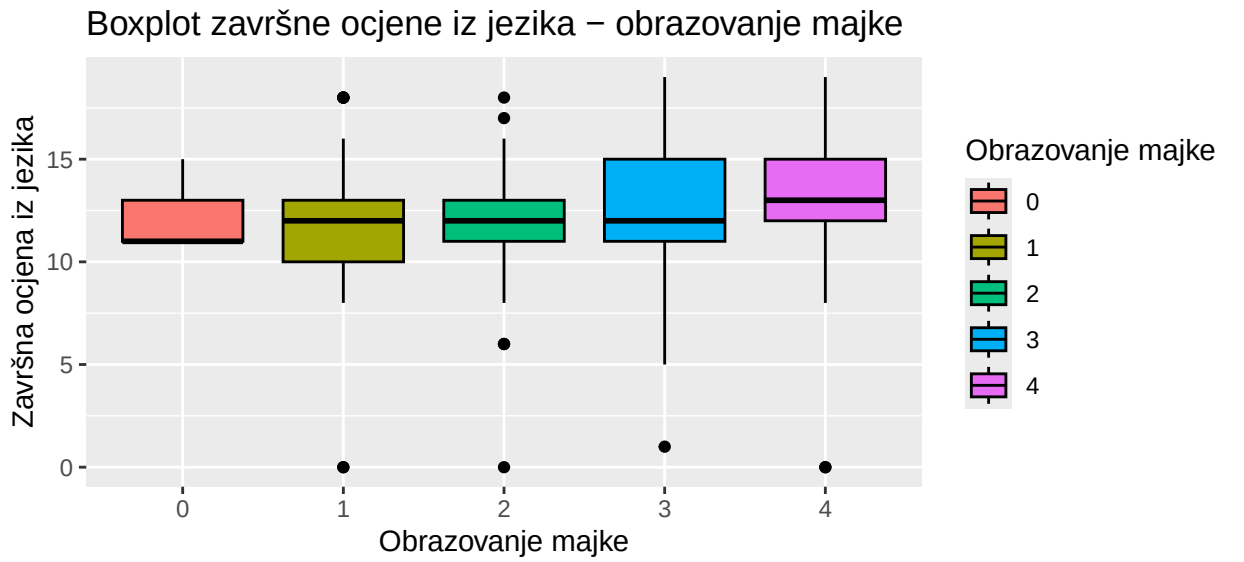
Moramo imati na umu moguću pristranost modela kojeg treniramo zbog nebalansiranosti podataka nad kojima prilagođavamo model. Naime, preko 90% učenika je prošlo portugalski jezik. To bi nam moglo dati iskrivljene podatke.

Prije same prilagodbe modela podacima, viuzalizirat ćemo podatke prema prije navedenim sociodemografskim varijablama, kako bismo lakše interpretirali rezultate za trenirani model. (Napomena: Kako smo prolaz označili faktorom, dobiveni box-plotovi bili bi beznačajni pa ćemo se za potrebe grafova koristiti ocjenama koje su učenici dobili u završnom ispitu.)

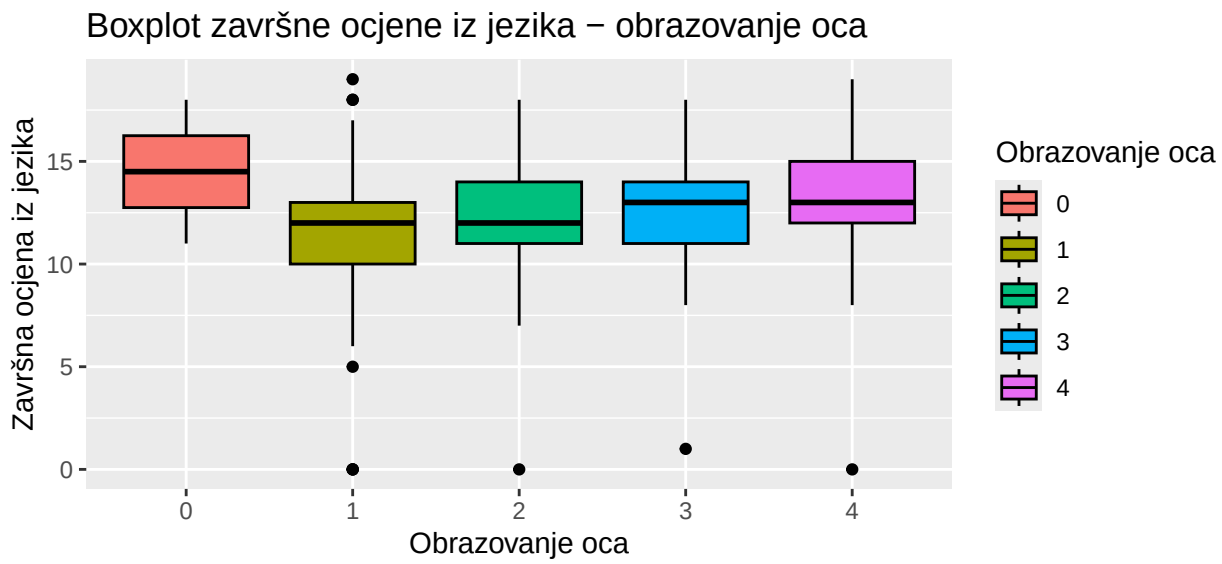
```r
g1 <- ggplot(df3 , aes(x = sex, y = G3_por, fill = sex)) +
  geom_boxplot(color = "black") +
  scale_x_discrete(
    name = "Spol",
    labels = c("Ž", "M")
  ) +
  scale_y_continuous(
    name = "Završna ocjena iz jezika",
    breaks = seq(0, 20, 5)
    ) +
  labs(
    title = "Boxplot završne ocjene iz jezika - spol"
    ) +
  scale_fill_discrete(name = "Spol", labels = c("Ž", "M"))

g2 <- ggplot(df3 , aes(x = factor(Medu), y = G3_por, fill = factor(Medu))) +
  geom_boxplot(color = "black") +
  scale_y_continuous(
    name = "Završna ocjena iz jezika",
    breaks = seq(0, 20, 5)
    ) +
  scale_x_discrete(name = "Obrazovanje majke") +
  labs(
    title = "Boxplot završne ocjene iz jezika - obrazovanje majke"
    ) +
  scale_fill_discrete(name = "Obrazovanje majke")

g3 <- ggplot(df3 , aes(x = factor(Fedu), y = G3_por, fill = factor(Fedu))) +
  geom_boxplot(color = "black") +
  scale_y_continuous(
    name = "Završna ocjena iz jezika",
    breaks = seq(0, 20, 5)
    ) +
  scale_x_discrete(name = "Obrazovanje oca") +
  labs(
    title = "Boxplot završne ocjene iz jezika - obrazovanje oca"
    ) +
  scale_fill_discrete(name = "Obrazovanje oca")

g4 <- ggplot(df3 , aes(x = factor(famsize), y = G3_por, fill = factor(famsize))) +
  geom_boxplot(color = "black") +
  scale_y_continuous(
    name = "Završna ocjena iz jezika",
    breaks = seq(0, 20, 5)
    ) +
  scale_x_discrete(name = "Veličina obitelji") +
  labs(
    title = "Boxplot završne ocjene iz jezika - veličina obitelji"
    ) +
  scale_fill_discrete(name = "Veličina obitelji")

g1
```


Boxplot završne ocjene iz jezika − spol
Završna ocjena iz jezika
15
10
5
0
Ž
M
Spol
Spol
Ž
M

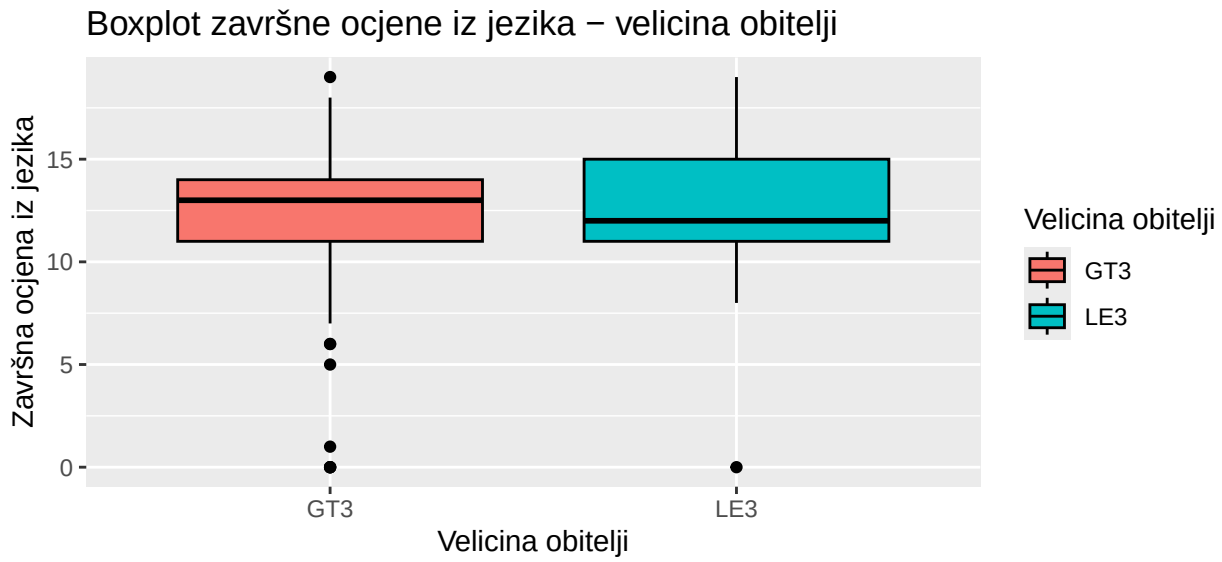
```
g2
```


Boxplot završne ocjene iz jezika − obrazovanje majke
Završna ocjena iz jezika
15
10
5
0
0
1
2
3
4
Obrazovanje majke
Obrazovanje majke
0
1
2
3
4

```
g3
```



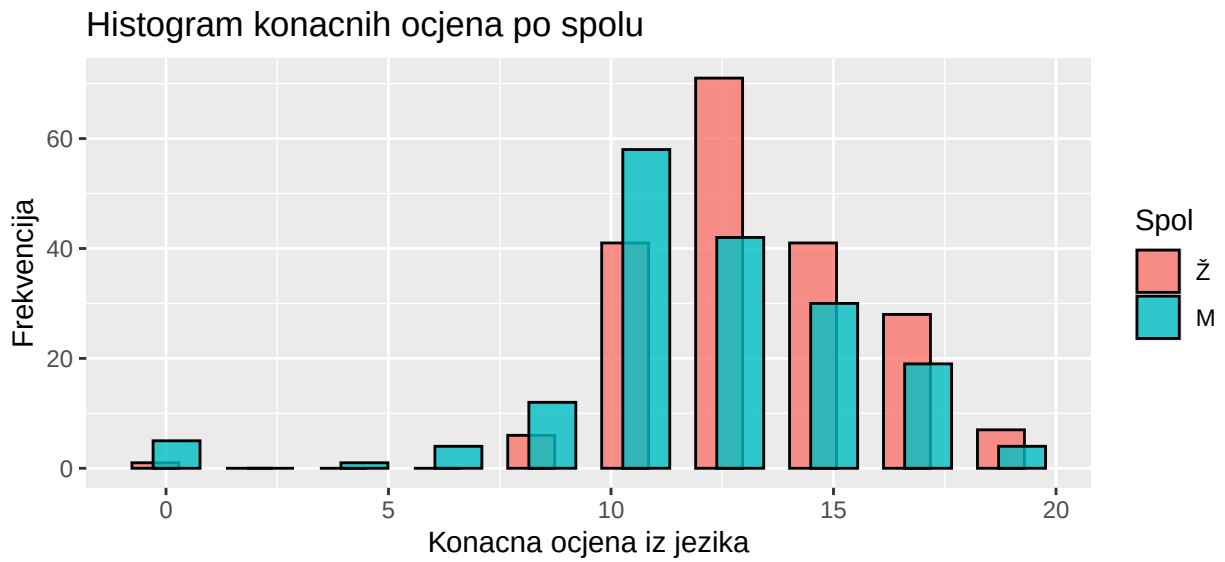

```
g4
```



Iz ovih box-plotova mogli bismo izvući razne zaključke, no treba imati na umu da se testira prolaz iz određenog predmeta, tako da je ključno gledati ocjene veće ili jednake 10. Npr. iz box-plota o obrazovanju majke moglo bi se reći da je većim obrazovanjem veća i ocjena, ali u smislu samog prolaza predmeta praktički se ništa ne mijenja. Jedino vrijedno spomena moglo bi biti to što je prvi kvartil muških učenika točno na granici prolaza pa bi se eventualno tu mogla vidjeti odstupanja. Sada ćemo prikazati histogram, samo za varijablu spol (jer će nam ona kasnije biti značajna).

```
ggplot(df3, aes(G3_por, fill = sex)) +
  geom_histogram(
    bins = 10,
    colour = "black",
    position = position_dodge(width = 0.95),
    alpha = 0.8
  ) +
```

```
  scale_x_continuous(name = "Konačna ocjena iz jezika") +
  scale_y_continuous(
    name = "Frekvencija",
    breaks = scales::pretty_breaks(5)
  ) +
  labs(title = "Histogram konačnih ocjena po spolu") +
  scale_fill_discrete(name = "Spol", labels = c("Ž", "M"))
```



Za daljnje zaključke konačno ćemo trenirati model logističke regresije nad našim podacima, ocijeniti model na temelju matrice zabune (engl. confusion matrix) i interpretirati rezultate iz statističkih rezultata iz treniranja modela.

```
#treniranje modela uzimajuci u obzir sve sociodemografske varijable
model <- glm(pass_por ~ sex + Medu + Fedu + famsize,
             data = df3,
             family = binomial)

summary(model)
```

```
##
## Call:
## glm(formula = pass_por ~ sex + Medu + Fedu + famsize, family = binomial,
##     data = df3)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.6410     0.5939   2.763 0.005728 **
## sexM         -1.5582     0.4616  -3.376 0.000737 ***
## Medu          0.3574     0.2286   1.564 0.117895
## Fedu          0.2934     0.2372   1.237 0.216126
## famsizeLE3    0.7037     0.4923   1.429 0.152894
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 203.35  on 369  degrees of freedom
## Residual deviance: 180.97  on 365  degrees of freedom
## AIC: 190.97
##
## Number of Fisher Scoring iterations: 6
```

```
df3 <- df3 |>
  mutate(predicted_prob = predict(model, type = "response"))
```

Interpretiramo dobiveni model logističke regresije: - spol ima utjecaj na predikciju, takav da za spol=M vrijedi da će negativno utjecati na predikciju o čemu govori mala p-vrijednost i to na razini značajnosti 0.001 - učenici iz manjih obitelji (famSizeLE3) imaju veće vrijednosti procijenjenih logitsa, pa time sugeriraju mogući prolaz na ispitu, ali rezultat nije statistički značajan.

No zaustavimo se na trenutak i promotrimo matricu zabune da vidimo kako model donosi odluku.

```
threshold <- 0.5

df3 <- df3 %>%
  mutate(predicted_class = ifelse(predicted_prob > threshold, 1, 0))

table(df3$predicted_class, df3$pass_por)
```

```
##
##       0   1
##   1  29 341
```

```
accuracy <- mean(df3$pass_por == df3$predicted_class)
precision <- sum(
  df3$pass_por == 1 & df3$predicted_class == 1) / sum(df3$predicted_class == 1
                                                      )
recall <- sum(df3$pass_por == 1 & df3$predicted_class == 1) / sum(df3$pass_por == 1)
f1_score <- 2 * (precision * recall) / (precision + recall)

accuracy
```

```
## [1] 0.9216216
```

```
precision
```

```
## [1] 0.9216216
```

```
recall
```

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

```
f1_score
```

```
## [1] 0.9592124
```

Rezultat iz matrice zabune sugerira nam da možemo odbaciti ovaj trenirani model, jer se on ne prilagođava našim ulaznim podacima iz sociodemografskih varijabli već, zbog preko 90% oznaka 1 od ukupnog broja oznaka, za svaki ulaz daje predikciju 1 tj. prolazak ispita. Takav model ne generalizira dobro. To ćemo pokušati popraviti kreiranjem balansiranijeg dataseta jer empirijski gledano, najbolje je imati otprilike jednak broj oznaka 0 - pao i 1 - prošao, ali mi ćemo uzeti nešto veći uzorak učenika koji su prošli. Time pretpostavljamo da je naš slučajni poduzorak učenika koji su prošli dobro reprezentira početni veći uzorak.

```
set.seed(-123456)

dfSampled <- df3 |> filter(pass_por == 0)
dfHelp <- df3 |> filter(pass_por == 1) |> slice_sample(n = 58)

dfSampled <- rbind(dfSampled, dfHelp)
```

Odabrali smo 58 učenika koji su prošli da nam predstavljaju poduzorak, tako da naš skup nad kojim ćemo ponovno provesti test ima ukupno 87 učenika. Kako bismo se uvjerili da naš slučajni poduzorak učenika koji su prošli dobro reprezentira početni veći uzorak, ponovili smo prethodnu deskriptivnu analizu nad ovim smanjenim datasetom i uvjerili se da se box plotovi i histogrami nisu značajno promijenili (ne nalazi se u izvještaju zbog sažetosti).

```
model2 <- glm(pass_por ~ sex + Medu + Fedu + famsize,
              data = dfSampled,
              family = binomial)

summary(model2)
```

```
##
## Call:
## glm(formula = pass_por ~ sex + Medu + Fedu + famsize, family = binomial,
##     data = dfSampled)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.11012    0.70518   0.156 0.875907
## sexM        -2.03687    0.57753  -3.527 0.000421 ***
## Medu         0.62293    0.28777   2.165 0.030410 *
## Fedu        -0.04761    0.27044  -0.176 0.860267
## famsizeLE3   1.01717    0.63247   1.608 0.107779
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 110.753  on 86  degrees of freedom
## Residual deviance:  90.237  on 82  degrees of freedom
## AIC: 100.24
##
## Number of Fisher Scoring iterations: 4
```

Promotrimo rezultate prilagodbe modela podacima. Vidimo da na razini značajnosti 0.001 možemo zaključiti da spol igra ulogu u predikciji prolaza, tj. da je za muški spol manja vjerojatnost prolaza predmeta, iako ne možemo znati koliko točno. Možemo izvući jedino zaključak o pozitivnom ili negativnom utjecaju. Također prema rezultatu modela Medu varijabla također ima pozitivan utjecaj na prolaz učenika na ispitu. Također

ovi rezultati upućuju i na pozitivni utjecaj varijable Medu, dok ostale varijable nemaju statistički značajan utjecaj, no sigurno na neki način pridonose prilagodbi modela. Ono što nam je najzanimljivije promotriti je ocjena statističkog testa koja se bazira na maximum likehood estimation metodi, devijancu koju nam je izbacio R. Rezidualna devijanca modela nad našim podacima za odabrani poduzorak znatno je manja od od prve rezidualne devijance, što ide u prilog da se drugi model bolje prilagodio podacima kada smo smanjili skup za treniranje.

Još jednom provjeravamo confusion matrix kako bismo provjerili ispravnost modela.

```
dfSampled <- dfSampled |>
  mutate(predicted_prob = predict(model2, type = "response"))

threshold <- 0.5

dfSampled <- dfSampled %>%
  mutate(predicted_class = ifelse(predicted_prob > threshold, 1, 0))

table(dfSampled$predicted_class, dfSampled$pass_por)
```

```
##
##      0  1
##   0 17  5
##   1 12 53
```

```
accuracy2 <- mean(dfSampled$pass_por == dfSampled$predicted_class)
precision2 <- sum(
  dfSampled$pass_por == 1 & dfSampled$predicted_class == 1) / sum(dfSampled$predicted_class == 1)


recall2 <- sum(dfSampled$pass_por == 1 & dfSampled$predicted_class == 1) / sum(dfSampled$pass_por == 1)
f1_score2 <- 2 * (precision2 * recall2) / (precision2 + recall2)


accuracy2
```

```
## [1] 0.8045977
```

```
precision2
```

```
## [1] 0.8153846
```

```
recall2
```

```
## [1] 0.9137931
```

```
f1_score2
```

```
## [1] 0.8617886
```

Sada vidimo da unatoč malo lošijem accuracyju, se naš model bolje prilagođava podacima iz balansiranijeg dataseta. Dakle visok accuracy ne znači nužno bolji model!

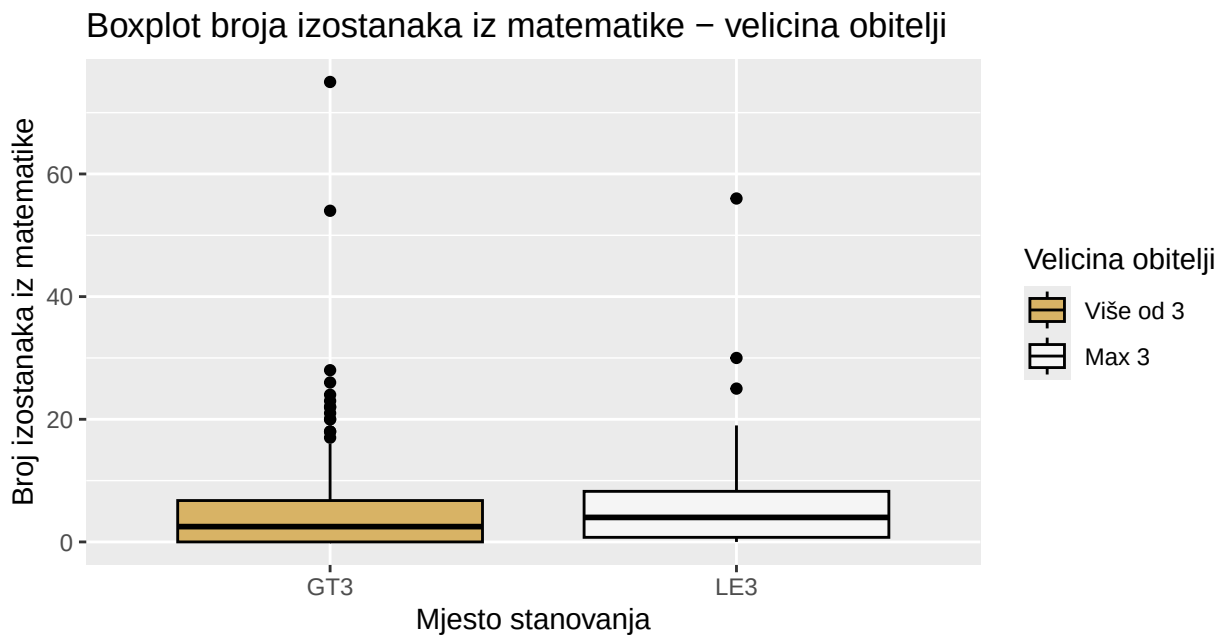
Zaključujemo da s vjerojatnošću 80.45% (vrijednost accuracy) možemo predvidjeti prolaz na ispitu na temelju sociodemografskih varijabli: spol, obrazovanje roditelja i veličina obitelji, na temelju uzoraka kojima smo prethodno balansirali podatke.

4. Postoji li razlika u broju izostanaka iz matematike izmedu učenika koji dolaze iz manjih i većih obitelji?

U podatkovnom skupu postoji varijabla *famsize* koja definira manje i veće obitelji, vrijednostima *LE3* i *GT3*,odnosno je li veličina obitelji manja ili jednaka 3, ili je veća od toga.

Zanima nas razlika u broju izostanaka, pa pogledajmo distribuciju podataka kroz varijable broj izostanaka u ovisnosti o veličini obitelji. Prvo ćemo pogledati *box-plotove* ovisno o veličini obitelji.

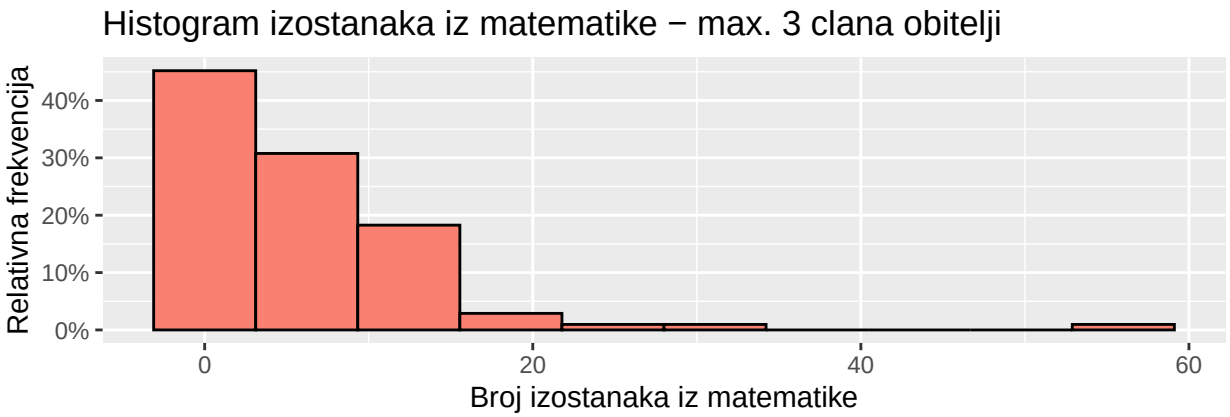
```
ggplot(dataset, aes(x = factor(famsize), y = absences_mat, fill = factor(famsize))) +
  geom_boxplot(color = "black") +
  scale_y_continuous(name = "Broj izostanaka iz matematike",
  breaks = scales::pretty_breaks()
  ) +
  labs(
    title = "Boxplot broja izostanaka iz matematike - veličina obitelji"
    ) +
  scale_x_discrete(name = "Mjesto stanovanja") +
  scale_fill_brewer(
    name = "Veličina obitelji",
    labels = c("Više od 3", "Max 3"),
    palette = "BrBG"
    )
```

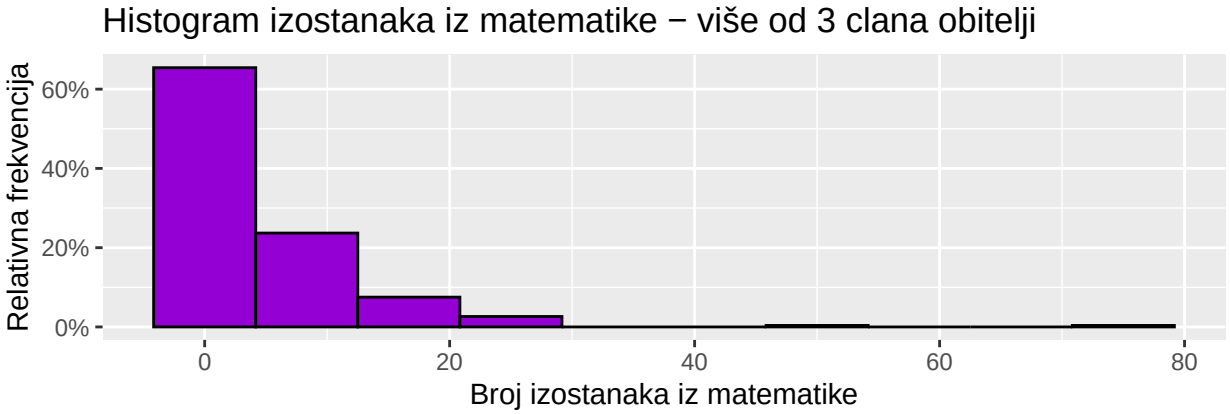


Box-plotovi su prikazali donekle očekivano, većina učenika ima iznimno malen broj izostanaka, no postoje outlieri među učenicima koji najviše izostaju. Za bolje razumijevanje distribucije, napravit ćemo histograme.

```
g4hist1 <- ggplot(dataset |> filter(famsize == "LE3"), aes(absences_mat)) +
  geom_histogram(
    aes(y = after_stat(count) / sum(after_stat(count))),
    bins = 10,
```

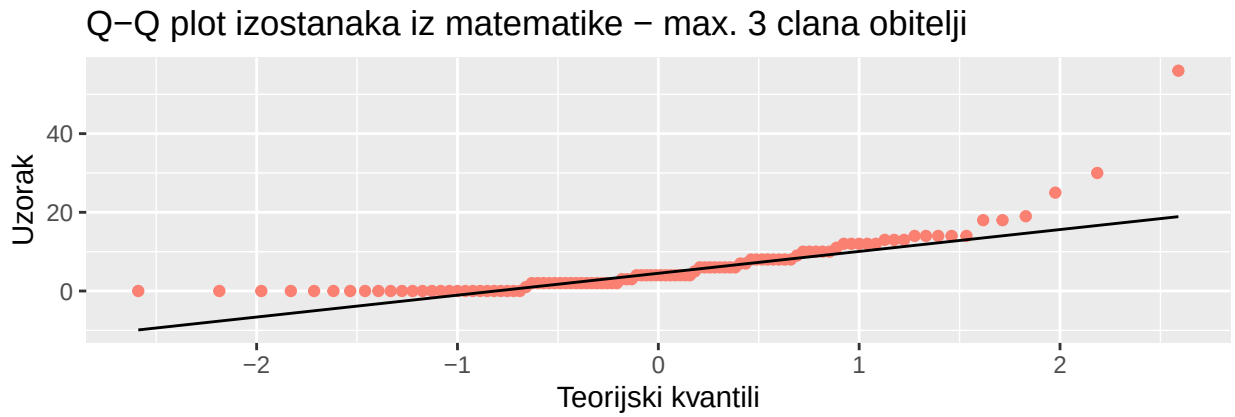
```
    colour = "black",
    fill = "salmon"
  ) +
  scale_x_continuous(name = "Broj izostanaka iz matematike") +
  scale_y_continuous(
    name = "Relativna frekvencija",
    labels = scales::percent_format(),
    breaks = scales::pretty_breaks(5)
  ) +
  labs(title = "Histogram izostanaka iz matematike - max. 3 člana obitelji")


g4hist2 <- ggplot(dataset |> filter(famsize == "GT3"), aes(absences_mat)) +
  geom_histogram(
    aes(y = after_stat(count) / sum(after_stat(count))),
    bins = 10,
    colour = "black",
    fill = "darkviolet"
  ) +
  scale_x_continuous(name = "Broj izostanaka iz matematike") +
  scale_y_continuous(
    name = "Relativna frekvencija",
    labels = scales::percent_format(),
    breaks = scales::pretty_breaks(5)
  ) +
  labs(title = "Histogram izostanaka iz matematike - više od 3 člana obitelji")

grid.arrange(g4hist1, g4hist2)
```

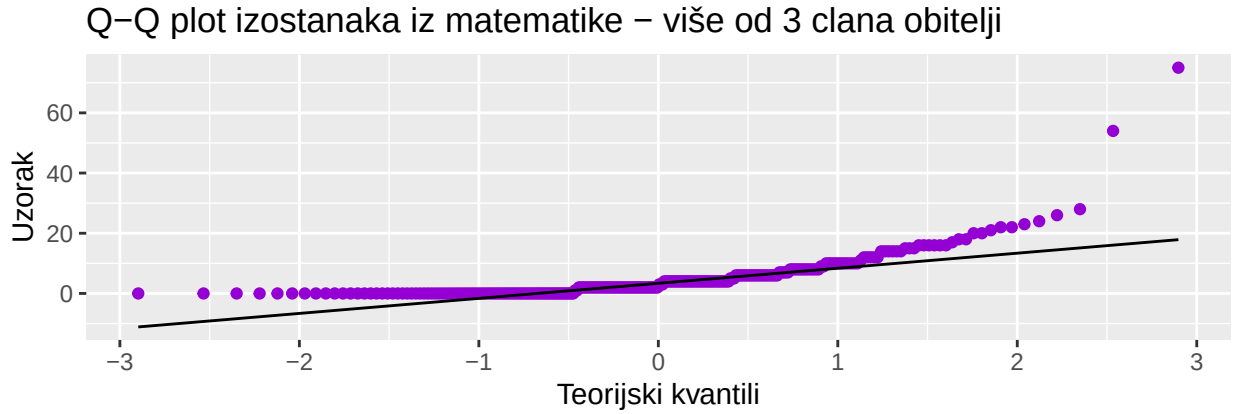




Histogrami najbolje dočaravaju koncentraciju učenika s malim brojem izostanaka. Dodatno ćemo ovu situaciju prikazati *Q-Q plotom* koja bi samo trebala potvrdili da ovdje nema riječi o normalnosti podataka.

```
g4qq1 <- ggplot(dataset |> filter(famsize == "LE3"), aes(sample = absences_mat)) +
  geom_qq(colour = "salmon") +
  geom_qq_line(colour = "black") +
  scale_y_continuous(name = "Uzorak") +
  scale_x_continuous(name = "Teorijski kvantili") +
  labs(title = "Q-Q plot izostanaka iz matematike - max. 3 člana obitelji")

g4qq2 <- ggplot(dataset |> filter(famsize == "GT3"), aes(sample = absences_mat)) +
  geom_qq(colour = "darkviolet") +
  geom_qq_line(colour = "black") +
  scale_y_continuous(name = "Uzorak") +
  scale_x_continuous(name = "Teorijski kvantili") +
  labs(title = "Q-Q plot izostanaka iz matematike - više od 3 člana obitelji")

grid.arrange(g4qq1, g4qq2)
```





U oba slučaja podaci su preteški u lijevom repu, *right-skewed*, a zatim outlieri koje smo vidjeli u *box-plotovima* su preveliki da bi ih se moglo smatrati dijelom normalne distribucije. Za vlastite potrebe pokušali smo transformirati podatke logaritamskom transformacijom kako bismo raspršili podatke u lijevom repu, ali nismo uspjeli dobiti distribuciju ni približnu normalnoj.

Kako s ovakvim podacima ne možemo pristupiti parametarskom *t-testu*, koristit ćemo neparametarski test koji ima manju snagu testa, ali ga možemo provesti neovisno o distribuciji.

Jedina pretpostavka koju imamo, koja je i u potpunosti opravdana, jest da su dva uzorka (učenika iz manjih i učenika iz većih obitelji) nezavisna.

Odabir prikladnog testa je *Mann-Whitney-Wilcoxonov test* koji za nezavisne uzorke testira jednakost dviju distribucija: H0 : $M_{LE3} = M_{GT3}$. Nema razlike u broju izostanaka između učenika koji dolaze iz manjih i većih obitelji. H1: $M_{LE3} \neq M_{GT3}$. Postoji razlika u broju izostanaka između učenika koji dolaze iz manjih i većih obitelji.

```
le3absences <- dataset |> filter(famsize == "LE3") |> pull(absences_mat)
gt3absences <- dataset |> filter(famsize == "GT3") |> pull(absences_mat)

median(le3absences)
```

```
## [1] 4
```

```
median(gt3absences)
```

```
## [1] 2.5
```

```r
wilcox.test(le3absences, gt3absences, alternative = "two.sided", conf.level = 0.95)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  le3absences and gt3absences
## W = 15303, p-value = 0.1055
## alternative hypothesis: true location shift is not equal to 0
```

Prije same provedbe testa, dodatno su ispisani medijani za obje grupe, i moguće je vidjeti da je medijan izostanaka na zadanom skupu nešto veći kod učenika čije obitelji imaju najviše tri člana. No, provedbom Mann-Whitney-Wilcoxonovog testa dobivamo p-vrijednost od 0.1055. To znači da na razini značajnosti od 5% ne možemo odbaciti H_0, odnosno ostajemo pri tome da je broj izostanaka neovisan o broju članova obitelji, iako treba imati na umu relativno malenu p-vrijednost. Također ovaj test je neparametarski i zbog manje snage testa, ubuduće bismo se zapitali kako provesti test s većom snagom i kakve bismo tada rezultate i odluke mogli donijeti.