Projekt: Analiza uspjeha učenika

Fran Galić, Jana Gazdek, Vedran Maksić, Vjekoslav Gračaković

2025-01-27

Motivacija i opis problema

Podaci o učenicima prikupljeni su kako bi se istražili čimbenici koji utječu na njihov školski uspjeh. Cilj je razumjeti kako različiti aspekti, poput vremena provedenog u učenju, obiteljske podrške i izostanaka, doprinosi postignućima učenika. Ova analiza omogućuje bolji uvid u područja ključna za poboljšanje obrazovnih rezultata te za donošenje preporuka u svrhu optimizacije obrazovnog procesa.

Učitavanje i uređivanje podatkovnog skupa

Učitavanje i proučavanje podatkovnog skupa

Podatkovni skup učitavamo u varijablu `studentData` kako bismo ga mogli dalje analizirati.

```
studentData <- read_csv("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.
```

Prikaz prvih nekoliko redaka kako bismo vidjeli primjere vrijednosti u svakoj koloni.

```
head(studentData)
```

```
## # 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>
```

Kratak uvid u tipove podataka i strukturu podatkovnog okvira.

```
glimpse(studentData)
```

```
## 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~
```

Uređivanje podataka podatkovnog skupa

Provjera nedostajućih vrijednosti:

```
missing_vals <- sum(is.na(studentData))
cat("Ukupno nedostajućih vrijednosti:", missing_vals)
```

```
## Ukupno nedostajućih vrijednosti: 0
```

Pretvaranje nominalnih kategorija u faktore

Kako bismo podatke prilagodili za analizu, pojedine varijable pretvaramo u faktore. Razlikujemo nominalne faktore (nesortirane kategorije) i ordinalne faktore (kategorije s prirodnim poretkom). Ovakva priprema olakšava kasniju primjenu statističkih testova i modela.

```
convert_to_factor <- function(data, cols) {
  for (col in cols) {
    data[[col]] <- as.factor(data[[col]])
  }
  return(data)
}

nominal_factors <- c(
  "school", "sex", "address", "famsize", "Pstatus", "schoolsup", "famsup",
  "paid_mat", "paid_por", "activities", "nursery", "higher", "internet",
  "romantic", "Mjob", "Fjob", "reason", "guardian"
)
studentData <- convert_to_factor(studentData, nominal_factors)
```

Pretvaranje ordinalnih kategorija u faktore

```
ordinal_factors <- list(
  famrel       = 1:5,
  freetime     = 1:5,
  goout        = 1:5,
  Dalc         = 1:5,
  Walc         = 1:5,
  health       = 1:5,
  traveltime   = 1:4,
  studytime    = 1:4
)

for (col in names(ordinal_factors)) {
  studentData[[col]] <- factor(studentData[[col]], levels = ordinal_factors[[col]], ordered = TRUE)
}
```

Prikaz prvih nekoliko redaka nakon uređivanja podataka.

```r
head(studentData)
```

```
## # A tibble: 6 x 39
##   school sex     age address famsize Pstatus  Medu  Fedu Mjob     Fjob    reason
##   <fct>  <fct> <dbl> <fct>   <fct>   <fct>   <dbl> <dbl> <fct>    <fct>   <fct>
## 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 <fct>, traveltime <ord>, studytime <ord>,
## #   failures_mat <dbl>, failures_por <dbl>, schoolsup <fct>, famsup <fct>,
## #   paid_mat <fct>, paid_por <fct>, activities <fct>, nursery <fct>,
## #   higher <fct>, internet <fct>, romantic <fct>, famrel <ord>, freetime <ord>,
## #   goout <ord>, Dalc <ord>, Walc <ord>, health <ord>, absences_mat <dbl>,
## #   absences_por <dbl>, G1_mat <dbl>, G2_mat <dbl>, G3_mat <dbl>, G1_por <dbl>,
## #   G2_por <dbl>, G3_por <dbl>
```

Analiza podatkovnog skupa

Jesu li prosječne konačne ocjene iz matematike različite između spolova?

Jedan od ključnih čimbenika u analizi školskog uspjeha je pitanje postoje li razlike među spolovima u postignutim rezultatima. Iako postoje razni stereotipi, ovdje ćemo ispitati postoji li statistički značajna razlika u konačnim ocjenama iz matematike (G3_mat) između djevojaka (F) i dječaka (M).

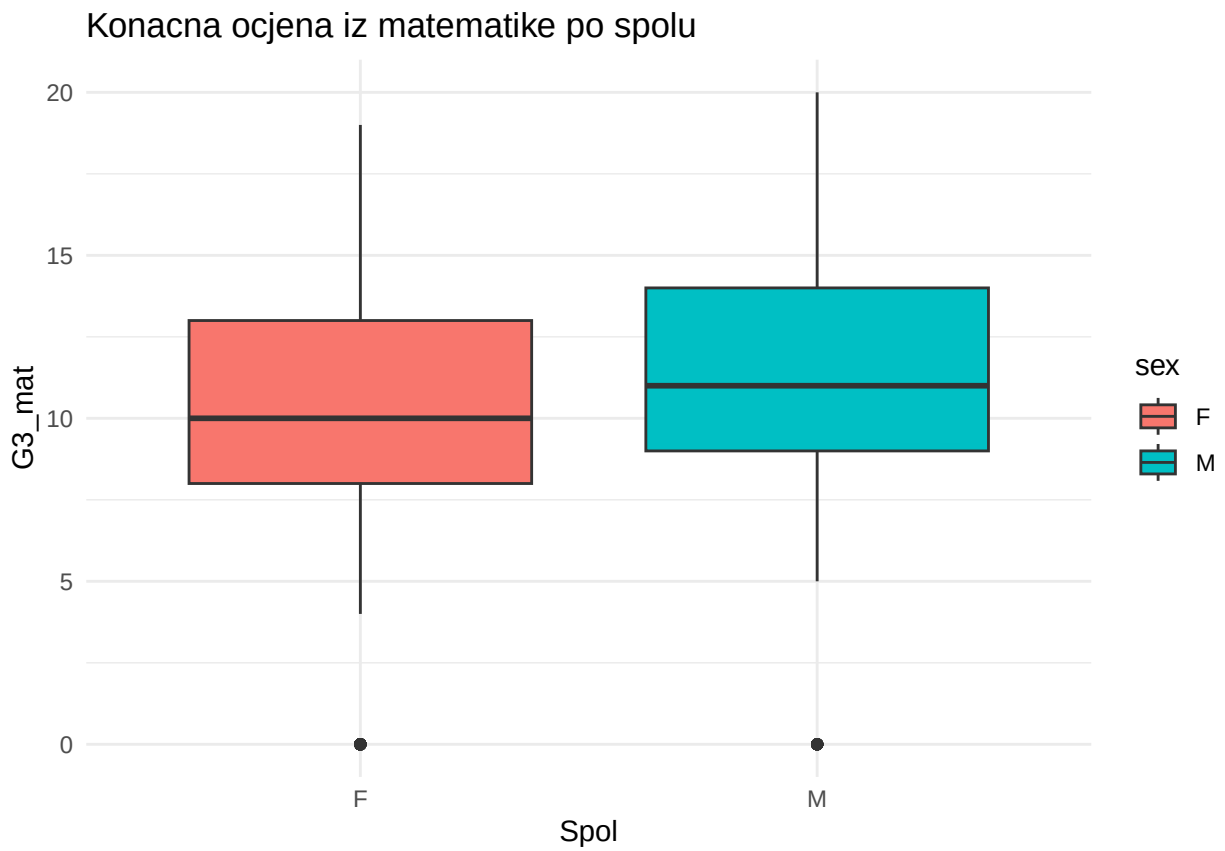
Postavka hipoteza

H_0 : $\mu_{\text{G3\_mat dječaka}} = \mu_{\text{G3\_mat djevojaka}}$

H_1 : $\mu_{\text{G3\_mat dječaka}} \neq \mu_{\text{G3\_mat djevojaka}}$

Boxplot i kratka interpretacija

```r
ggplot(studentData, aes(x = sex, y = G3_mat, fill = sex)) +
  geom_boxplot() +
  labs(
    title = "Konačna ocjena iz matematike po spolu",
    x = "Spol",
    y = "G3_mat"
  ) +
  theme_minimal()
```



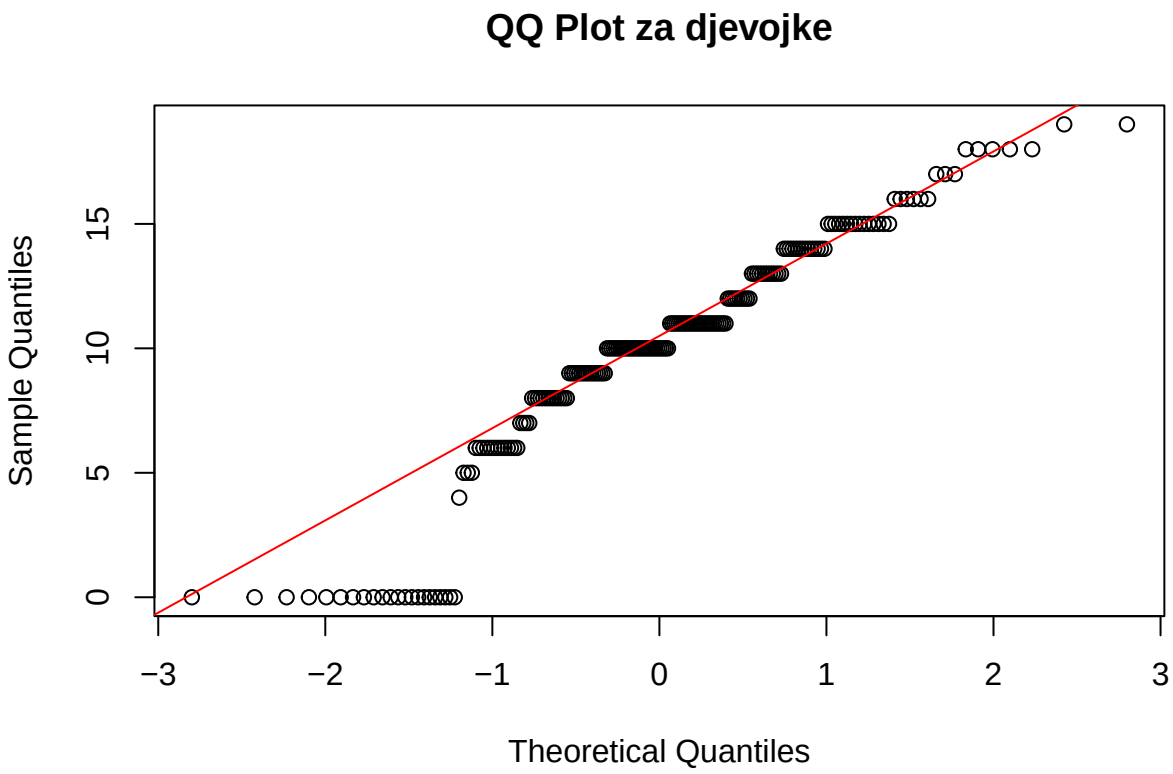
Vizualno, boxplot ne sugerira preveliku razliku između dviju srednjih vrijednosti, no tek formalnim testom možemo utvrditi je li ta razlika statistički značajna.

Provjera normalnosti podataka

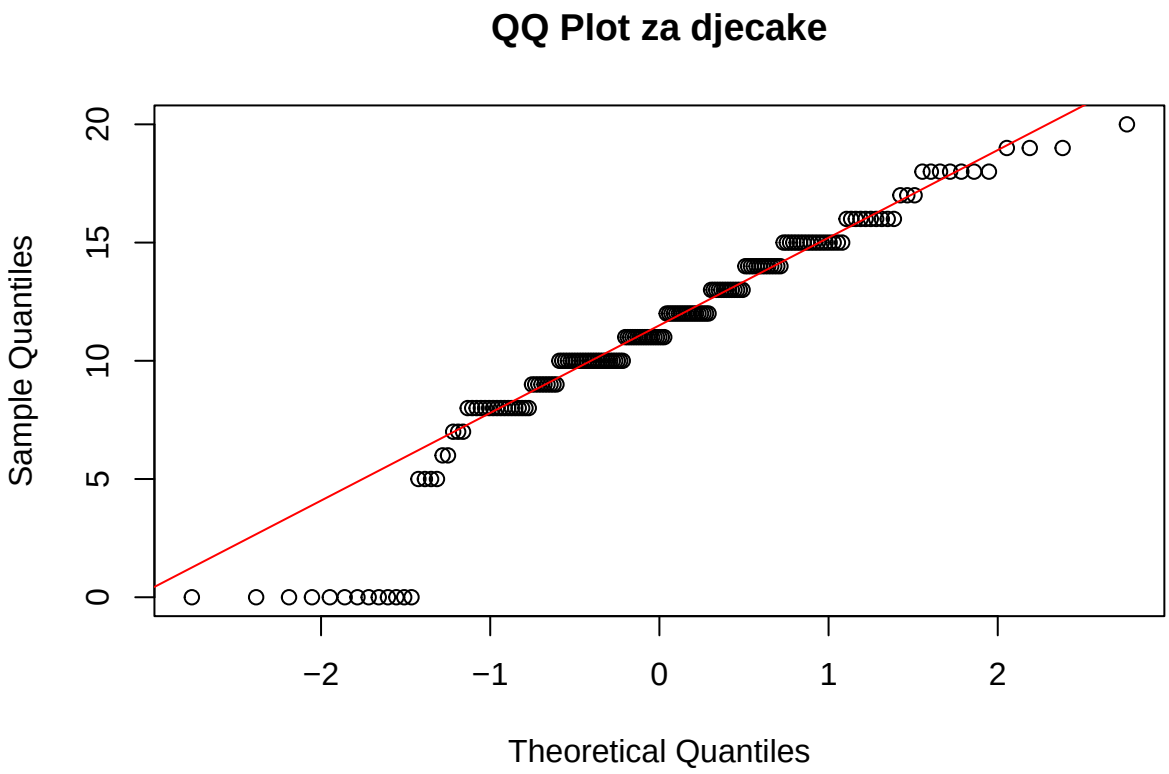
Za formalno testiranje razlika između srednjih vrijednosti odabrali smo t-test. T-test pretpostavlja (između ostalog) da su podaci u svakoj skupini približno normalno distribuirani. Provjerit ćemo to grafički (QQ-plot) i Shapiro-Wilkovim testom.

```
female_data <- studentData$G3_mat[studentData$sex == "F"]
male_data   <- studentData$G3_mat[studentData$sex == "M"]

qqnorm(female_data, main = "QQ Plot za djevojke")
qqline(female_data, col = "red")
```




```r
qqnorm(male_data, main = "QQ Plot za dječake")
qqline(male_data, col = "red")
```



Većina točaka je blizu dijagonale, no pri najnižim vrijednostima (0-2) vidimo veći otklon (floor efekt) - određen broj učenika ostvario je vrlo niske ocjene, pa su se "nagomilali" u donjem dijelu raspodjele. Najviše vrijednosti također pokazuju malo odvajanje od dijagonale. Sveukupno, raspodjele su "grubo" normalne s blagim odstupanjem u repovima.

Provodimo Shapiro-Wilkov test normalnosti:

```
shapiro.test(female_data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  female_data
## W = 0.92925, p-value = 4.058e-08
```

```
shapiro.test(male_data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  male_data
## W = 0.93103, p-value = 2.07e-07
```

Rezultati analize pokazuju da smo dobili nisku p-vrijednost, što nam omogućuje da pri razini značajnosti od 0,05 odbacimo nul-hipotezu i zaključimo da podaci „statistički“ odstupaju od normalne distribucije. Unatoč tome, s obzirom na robusnost t-testa u odnosu na pretpostavku normalnosti, kao i na dostatno velik uzorak, možemo nastaviti s primjenom t-testa bez potrebe za transformacijama podataka. Stoga se t-test i dalje može smatrati valjanim za ovu analizu.

Provjera jednakosti varijanci:

T-test također pretpostavlja (u klasičnoj varijanti) jednake varijance. Provjerit ćemo F-testom. No, F-test i sam pretpostavlja normalnost, pa je ponekad osjetljiv na odudaranja. Ipak, obavit ćemo ga informativno:

```
var.test(G3_mat ~ sex, data = studentData)
```

```
##
##  F test to compare two variances
##
## data:  G3_mat by sex
## F = 1.0998, num df = 194, denom df = 174, p-value = 0.5223
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8217022 1.4685075
## sample estimates:
## ratio of variances
##           1.099763
```

Rezultati F-testa za usporedbu varijanci završnih ocjena iz matematike pokazuju da nema statistički značajne razlike u varijancama (p = 0,5223). Budući da je p-vrijednost značajno veća od 0,05, ne odbacujemo nul-hipotezu i zaključujemo da su varijance ocjena slične za obje skupine.

Provođenje t-testa:

Provodimo t-test za nepoznate, ali jednake varijance:

```
t.test(G3_mat ~ sex, data = studentData, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  G3_mat by sex
## t = -2.5411, df = 368, p-value = 0.01146
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  -2.147348 -0.273751
## sample estimates:
## mean in group F mean in group M
##        9.892308       11.102857
```

Konačni zaključak i objašnjenje

Rezultati t-testa pokazuju statistički značajnu razliku u prosječnim konačnim ocjenama iz matematike između učenika i učenica.

Prosječna ocjena učenika iznosi 11,10, dok je prosječna ocjena učenica 9,89. Budući da je p-vrijednost manja od uobičajene razine značajnosti (0,05), možemo zaključiti da postoji značajna razlika u korist učenika.

Interval pouzdanosti od -2,15 do -0,27 dodatno potvrđuje da je prosječna ocjena učenika značajno viša od ocjene učenica.

Kratak osvrt na izbor testa usprkos narušenoj normalnosti Iako Shapiro-Wilk test možda sugerira da raspodjela nije "idealno normalna", t-test je razmjerno robustan - pogotovo uz veći broj ispitanika. Za ekstremnija odstupanja mogli bismo razmisliti o neparametrijskom testu (npr. Mann-Whitney), no ovdje zaključujemo da je odstupanje dovoljno blago da je t-test prihvatljiv, čime dobivamo odgovor na naše istraživačko pitanje.

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

Cilj je testirati postoji li statistički značajna razlika između ove dvije grupe te analizirati distribuciju ocjena. U podatkovnom skupu stupac G1_mat predstavlja prvu ocjenu iz matematika (između 0 i 20). Stupac address predstavlja mjesto stanovanja učenika, a poprima dvije vrijednosti U za (urbano područje) i R (ruralno područje).

Učitavanje i pregled podataka

funkcija za pronalaženje stršečih vrijednosti

```
find_outliers <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  outliers <- x[x < lower_bound | x > upper_bound]
  return(outliers)
}
```

U nastavku prikazujemo osnovne statističke mjere za ocjene G1_mat, uključujući minimum, prvi kvartil (Q1), medijan, srednju vrijednost, treći kvartil (Q3) i maksimum, kako za sve učenike u skupini, tako i zasebno za urbane i ruralne učenike.

```
table(studentData$address)
```

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

```
summary(studentData$G1_mat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    8.00   11.00   10.89   13.00   19.00
```

```
data_U <- subset(studentData, address == "U")
data_R <- subset(studentData, address == "R")
summary(data_U$G1_mat)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    4.00    8.00   11.00   11.01   14.00   19.00
```

```
summary(data_R$G1_mat)
```

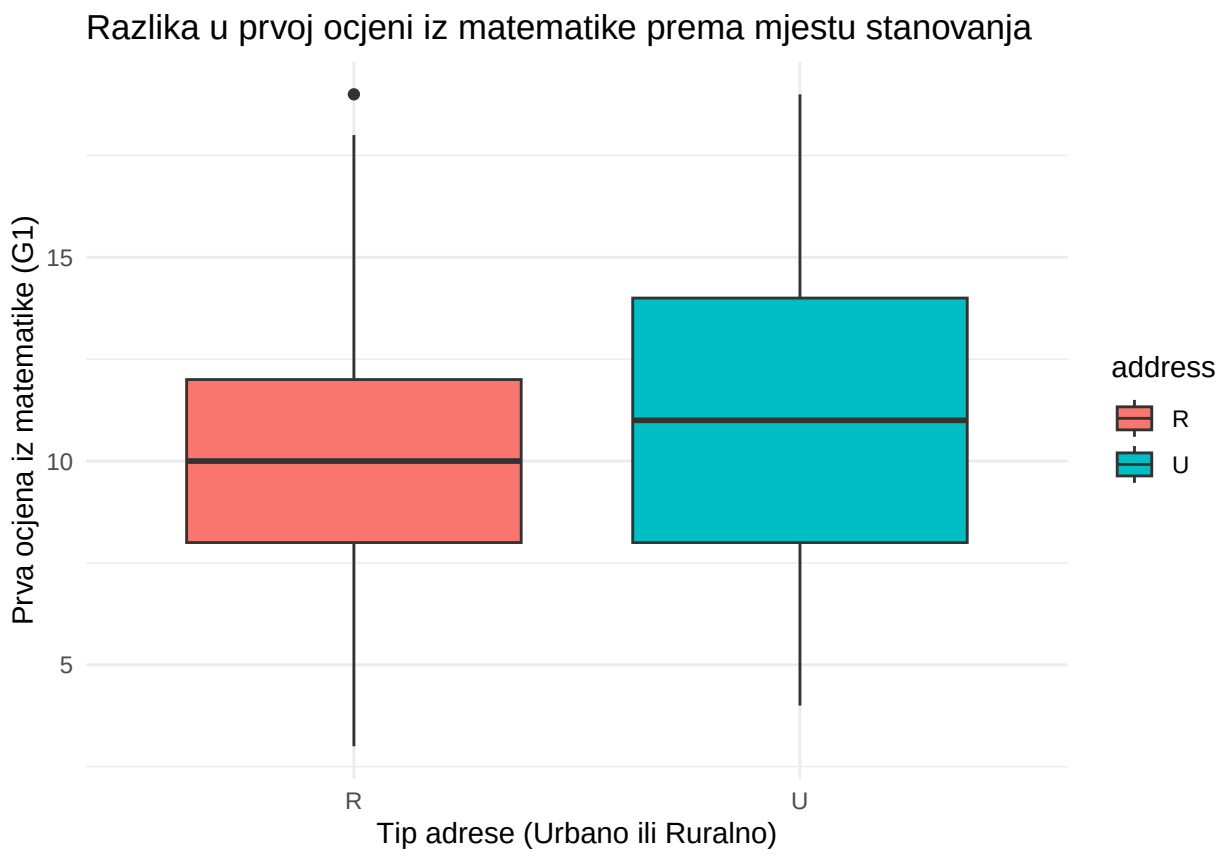
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3.00    8.00   10.00   10.46   12.00   19.00
```

Uočavamo da je broj urbanih učenika znatno veći u skupu podataka. Što se tiče deskriptivnih statističkih mjera, ne primjećuju se značajne razlike između ruralnih i urbanih skupina.

Vizualizacija podataka

Boxplot omogućuje pregled medijana, kvartila i outlinera za obje grupe.

```
ggplot(studentData, aes(x = address, y = G1_mat, fill = address)) +
  geom_boxplot() +
  labs(
    title = "Razlika u prvoj ocjeni iz matematike prema mjestu stanovanja",
    x = "Tip adrese (Urbano ili Ruralno)",
    y = "Prva ocjena iz matematike (G1)"
  ) +
  theme_minimal()
```



Na temelju grafa možemo procijeniti da su podaci simetrični i ne postoji značajna razlika u centralnim tendencijama između urbanih i ruralnih učenika.

Također možemo primjetiti da postoji jedan outliner (ocjena 19) za skup podataka kojeg čine učenici iz ruralnih područja. Za testiranje normalnosti razmotrit ćemo micanje outliera.

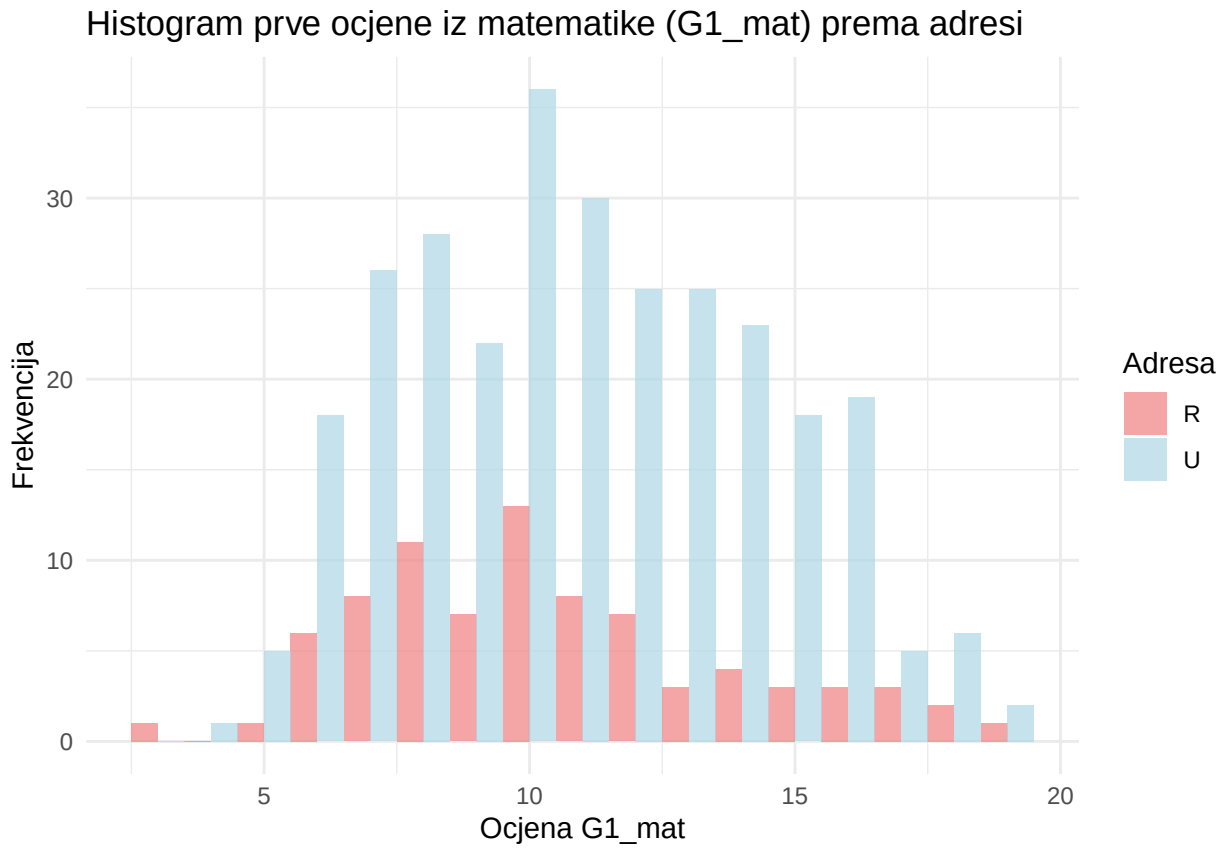
```
rural_outliers <- find_outliers(studentData$G1_mat[studentData$address == "R"])
print(rural_outliers)
```

```
## [1] 19
```

Histogram za detaljniji pregled distribucije

Histogram pruža uvid u raspodjelu podataka, omogućujući procjenu njihove približne normalnosti. Na temelju ovog pregleda, možemo odlučiti je li potrebna transformacija podataka ili primjena neparametarskih metoda.

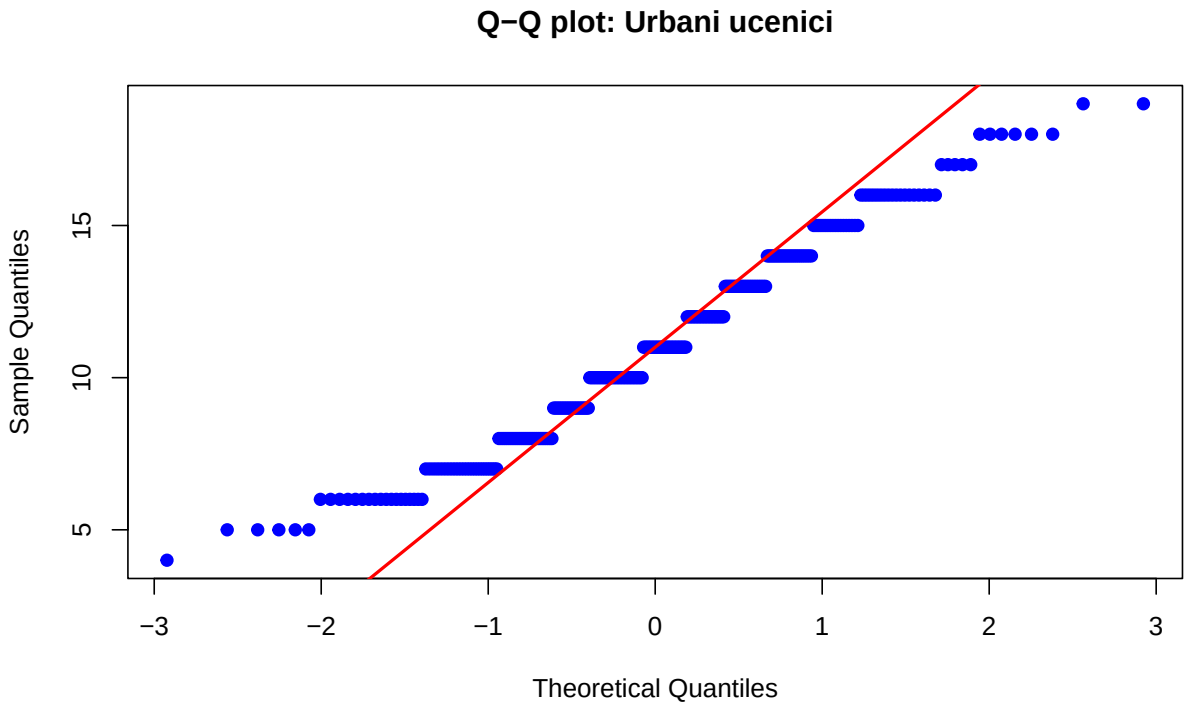
```
ggplot(studentData, aes(x = G1_mat, fill = address)) +
  geom_histogram(
    position = "dodge",
    binwidth = 1,  # Postavi širinu binova
    alpha = 0.7    # Postavi prozirnost za bolju vizualizaciju
  ) +
  labs(
    title = "Histogram prve ocjene iz matematike (G1_mat) prema adresi",
    x = "Ocjena G1_mat",
    y = "Frekvencija",
    fill = "Adresa"
  ) +
  theme_minimal() +
  scale_fill_manual(values = c("U" = "lightblue", "R" = "lightcoral"))
```



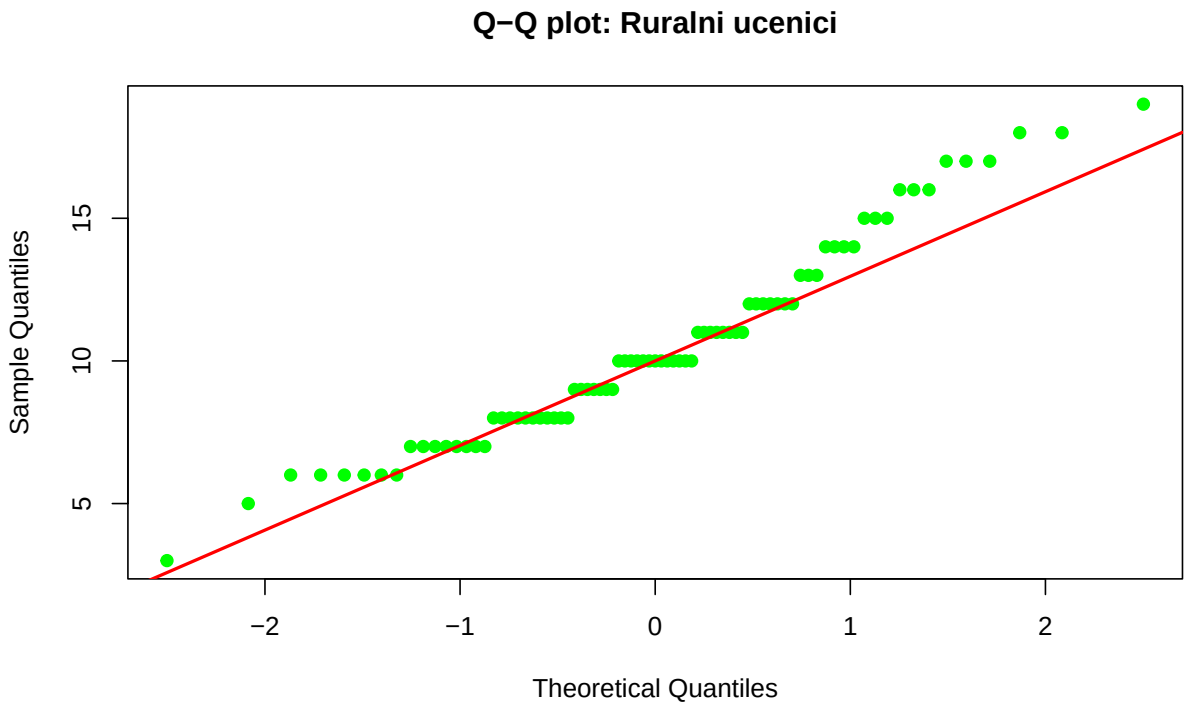
Grafovi upućuju na to da podaci mogu biti približno normalno distribuirani.

Q-Q plotovi za testiranje normalnosti podataka

```
qqnorm(studentData$G1_mat[studentData$address == "U"],
       main = "Q-Q plot: Urbani učenici", pch = 19, col = "blue")
qqline(studentData$G1_mat[studentData$address == "U"], col = "red", lwd = 2)
```




```
qqnorm(studentData$G1_mat[studentData$address == "R"],
       main = "Q-Q plot: Ruralni učenici", pch = 19, col = "green")
qqline(studentData$G1_mat[studentData$address == "R"], col = "red", lwd = 2)
```



Q-Q plot za urbane učenike: podaci vidljivo odstupaju od crvene linije, osobito u krajevima (repovima). Ovo sugerira da podaci za urbane učenike ne prate normalnu distribuciju.

Q-Q plot za ruralne učenike: podaci prate crvenu liniju puno bliže, s manjim odstupanjima u krajevima. Ovo sugerira da podaci za ruralne učenike mogu biti približno normalno distribuirani. ### Test normalnosti (Shapiro-Wilk test) za G1_mat ocjene za urbane i ruralne učenike Izvršit ćemo dva testa normalnosti. Prvi test će obuhvatiti G1_mat ocjene učenika koji žive u urbanim područjima, dok će drugi test obuhvatiti G1_mat ocjene učenika iz ruralnih područja.

Postavka hipoteza

H_0 : Podaci dolaze iz populacije koja je normalno distribuirana.

H_1 : Podaci dolaze iz populacije koja nije normalno distribuirana.

Shapiro-Wilk test ocjena za urbane učenike:

```
shapiro.test(studentData$G1_mat[studentData$address == "U"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  studentData$G1_mat[studentData$address == "U"]
## W = 0.97351, p-value = 3.486e-05
```

Shapiro-Wilk test ocjena za ruralne učenike:

```
shapiro_test_R <- shapiro.test(studentData$G1_mat[studentData$address == "R"])
```

```
shapiro_test_R_no_outliers <- shapiro.test(studentData$G1_mat[studentData$address == "R" & studentData$G
print(shapiro_test_R_no_outliers)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  studentData$G1_mat[studentData$address == "R" & studentData$G1_mat != 19]
## W = 0.96034, p-value = 0.01412
```

Rezultat testa normalnosti za urbane učenike pokazuje p-vrijednost od 3.486e-05, što je znatno niže od 0.05, stoga odbacujemo nul-hipotezu u korist alternativne i zaključujemo da G1_mat ocjene učenika koji žive na urbanim područjima nisu normalno distribuirane.

Za ruralne učenike, rezultat testa normalnosti daje p-vrijednost od 0.01059, koja je manja od 0.05, što također vodi do odbacivanja nul-hipoteze u korist alternativne, te zaključujemo da G1_mat ocjene učenika koji žive na ruralnim područjima nisu normalno distribuirane.

Kada uklonimo outliere iz skupa podataka za ruralne učenike, p-vrijednost ostaje ispod 0.05 i iznosi 0.01412, što ukazuje na to da ni nakon uklanjanja outliera, G1_mat ocjene za ruralne učenike i dalje nisu normalno distribuirane.

Rezultati testova normalnosti sugeriraju da podaci nisu normalno distribuirani, unatoč tome što histogrami upućuju na moguću normalnost. Q-Q plot za urbane učenike pokazuje značajna odstupanja u repovima, što dodatno potvrđuje nenormalnost distribucije podataka. Zbog niske p-vrijednosti i rezultata Q-Q plota, odbacujemo pretpostavku normalnosti za urbane učenike. S druge strane, za ruralne učenike pretpostavka normalnosti je nešto prihvatljivija, no s obzirom na p-vrijednost, odlučujemo odbaciti pretpostavku normalnosti i za ovu skupinu.

Iako t-test zahtijeva normalnost distribucije podataka, on je relativno robustan na odstupanja od normalnosti, osobito u slučajevima s većim brojem uzoraka. Međutim, kako bismo se odlučili za rigorozniji pristup koji se bolje uklapa u naš istraživački okvir, odlučujemo koristiti Wilcoxon test, koji je neparametarski i ne zahtijeva pretpostavku normalnosti. Ovaj pristup također omogućuje da se naš postupak razlikuje od postupka u prvom istraživačkom pitanju, čime smo spremni izabrati test koji je prikladniji za specifične uvjete našeg skupa podataka.

Wilcoxon test

Budući da podaci nisu normalno distribuirani, koristimo Wilcoxon test za ispitivanje razlika između dvije grupe.

Postavka hipoteza

H_0 : Ne postoji značajna razlika u distribuciji ocjena G1_mat između urbanih i ruralnih učenika.

H_1 : Postoji značajna razlika u distribuciji ocjena G1_mat između urbanih i ruralnih učenika.

```
wilcox_test <- wilcox.test(G1_mat ~ address, data = studentData)
print(wilcox_test)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  G1_mat by address
## W = 10464, p-value = 0.1432
## alternative hypothesis: true location shift is not equal to 0
```

Rezultat i zaključak:

P-vrijednost iznosi 0.1432, što je veće od 0.05, stoga ne odbacujemo nul-hipotezu na razini značajnosti od 0.05. Na temelju ovog uzorka, ne možemo donijeti zaključak o postojanju značajne razlike u ocjenama G1_mat između urbanih i ruralnih učenika.

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

Pitanje predviđanja uspjeha na završnom ispitu iz jezika sve više postaje važno u kontekstu obrazovne analize, jer može pomoći u razumijevanju faktora koji utječu na akademske rezultate. Korištenjem različitih statističkih metoda, nastojimo utvrditi koliko ove varijable mogu biti korisni prediktori u modeliranju vjerojatnosti prolaza na ispitu.

Kreiranje binarne varijable za prolaz

Za analizu prolaznosti na završnom ispitu iz jezika, kreiramo binarnu varijablu passed_por koja označava prolaz (1) ili neprolaz (0) na temelju ocjene iz ispita, gdje je prolaz definiran kao ocjena veća ili jednaka 10. Ova varijabla se koristi kao ovisna varijabla u modelu logističke regresije.

```
studentData <- studentData %>%
  mutate(passed_por = ifelse(G3_por >= 10, 1, 0))  # 1 = prošao, 0 = nije prošao
studentData$passed_por <- as.factor(studentData$passed_por)
```

Distribucija prolaza na završnom ispitu iz jezika

```
table(studentData$passed_por)
```

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

Na temelju varijable passed_por, 341 student (91.6%) uspješno je položio završni ispit iz jezika, dok je 29 studenata (8.4%) nije prošlo. Ova neravnoteža u distribuciji podataka može značajno utjecati na performanse modela, osobito u pogledu prepoznavanja manjinske klase.

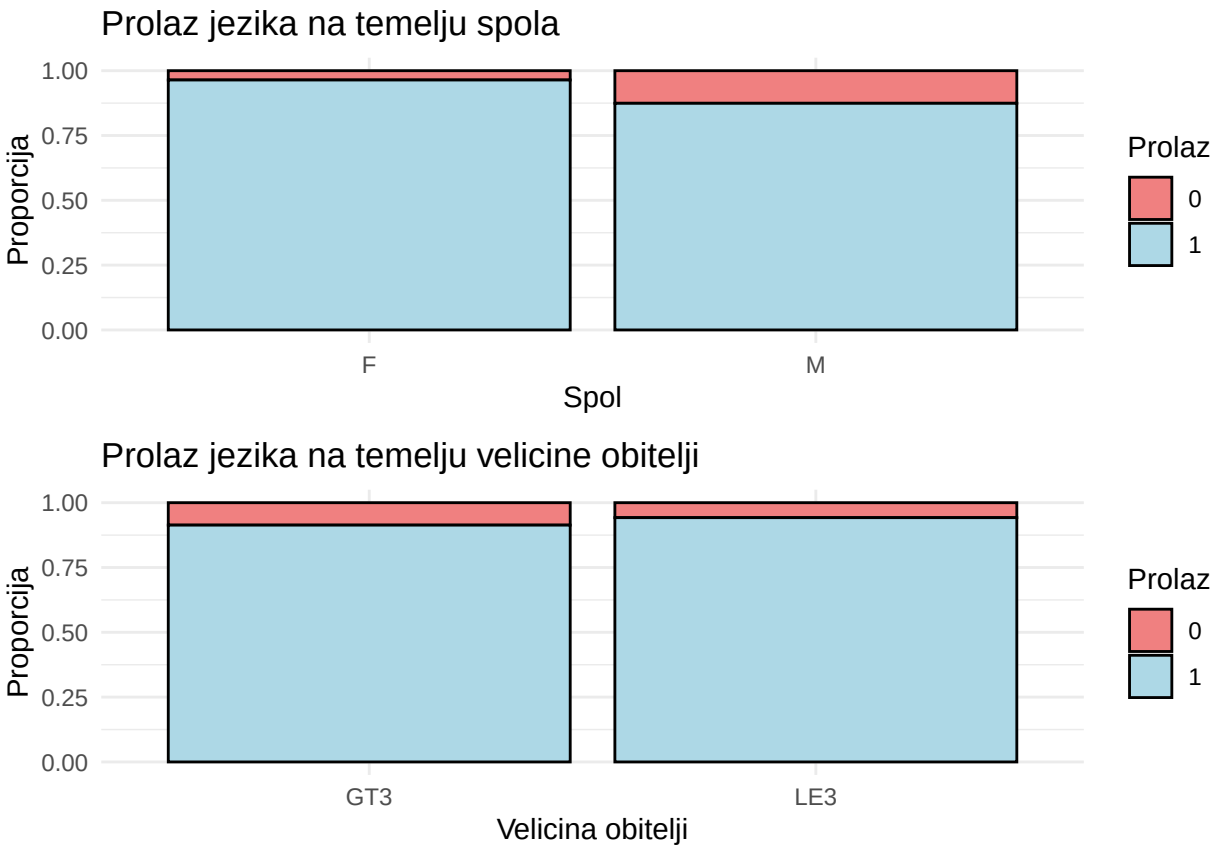
Vizualizacija prolaznosti na temelju spola i veličine obitelji

U nastavku su prikazane dvije vizualizacije prolaznosti na temelju sociodemografskih varijabli: spola i veličine obitelji. Prva vizualizacija prikazuje odnos između spola i prolaznosti, dok druga prikazuje povezanost između veličine obitelji i prolaznosti na završnom ispitu iz jezika.

```
plot1 <- ggplot(studentData, aes(x = sex, fill = passed_por)) +
  geom_bar(position = "fill", color = "black") +
  labs(
    title = "Prolaz jezika na temelju spola",
    x = "Spol",
    y = "Proporcija",
    fill = "Prolaz"
  ) +
  scale_fill_manual(values = c("lightcoral", "lightblue")) +
  theme_minimal()

plot2 <- ggplot(studentData, aes(x = famsize, fill = passed_por)) +
  geom_bar(position = "fill", color = "black") +
  labs(
    title = "Prolaz jezika na temelju veličine obitelji",
    x = "Veličina obitelji",
    y = "Proporcija",
    fill = "Prolaz"
  ) +
  scale_fill_manual(values = c("lightcoral", "lightblue")) +
  theme_minimal()

grid.arrange(plot1, plot2, ncol = 1)
```



Iz vizualizacije se može zaključiti da je veći postotak muških studenata koji nisu prošli ispit iz jezika u odnosu na ženske studente. Također, uočljivo je da studenti iz većih obitelji imaju nešto veću vjerojatnost za prolaz na završnom ispitu.

Vizualizacija prolaznosti na temelju obrazovanja majke i oca

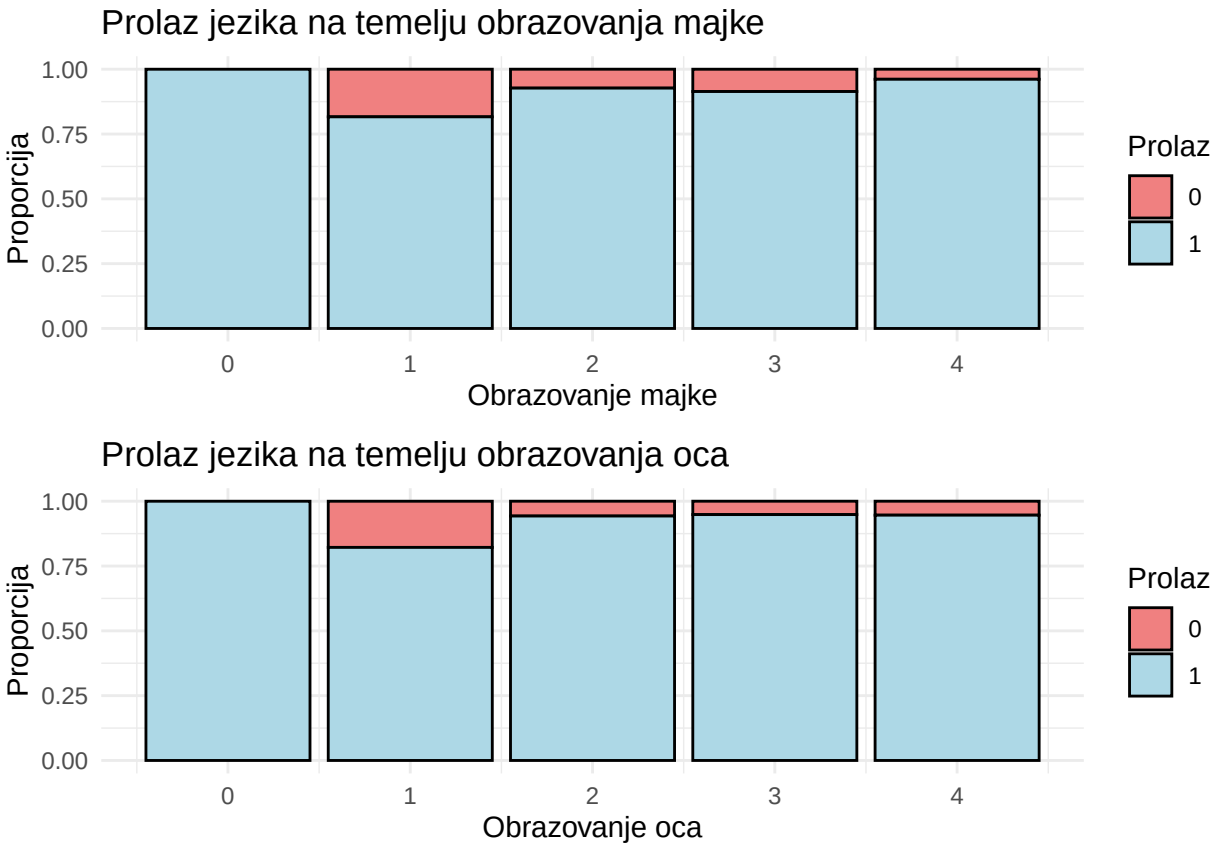
U nastavku su prikazane dvije vizualizacije koje prikazuju prolaznost na temelju obrazovanja majke i obrazovanja oca. Obje vizualizacije koriste barplotove za prikaz proporcija studenata koji su prošli ispit u odnosu na njihov sociodemografski faktor.

```
plot1 <- ggplot(studentData, aes(x = Medu, fill = passed_por)) +
  geom_bar(position = "fill", color = "black") +
  labs(
    title = "Prolaz jezika na temelju obrazovanja majke",
    x = "Obrazovanje majke",
    y = "Proporcija",
    fill = "Prolaz"
  ) +
  scale_fill_manual(values = c("lightcoral", "lightblue")) +
  theme_minimal()

plot2 <- ggplot(studentData, aes(x = Fedu, fill = passed_por)) +
  geom_bar(position = "fill", color = "black") +
  labs(
    title = "Prolaz jezika na temelju obrazovanja oca",
    x = "Obrazovanje oca",
```

```
    y = "Proporcija",
    fill = "Prolaz"
  ) +
  scale_fill_manual(values = c("lightcoral", "lightblue")) +
  theme_minimal()

grid.arrange(plot1, plot2, ncol = 1)
```



Vizualizacije upućuju na to da ni obrazovanje majke ni oca nemaju značajan utjecaj na prolaznost na završnom ispitu.

Model logističke regresije

Na temelju odabranih sociodemografskih faktora (spol, veličina obitelji, obrazovanje majke, obrazovanje oca, bračni status roditelja) izrađujemo logistički regresijski model s ciljem predviđanja prolaznosti na završnom ispitu iz jezika.

```
logistic_model <- glm(
  passed_por ~ sex + famsize + Medu + Fedu + Pstatus,
  data = studentData,
  family = binomial
)
```

Sažetak modela

Rezultati logističke regresije:

```
summary(logistic_model)
```

```
##
## Call:
## glm(formula = passed_por ~ sex + famsize + Medu + Fedu + Pstatus,
##     family = binomial, data = studentData)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.3731     1.2243   1.938 0.052583 .
## sexM         -1.5337     0.4621  -3.319 0.000902 ***
## famsizeLE3    0.6494     0.4970   1.307 0.191304
## Medu          0.3349     0.2296   1.459 0.144632
## Fedu          0.2935     0.2382   1.232 0.217926
## PstatusT     -0.7229     1.0587  -0.683 0.494737
## ---
## 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: 180.41  on 364  degrees of freedom
## AIC: 192.41
##
## Number of Fisher Scoring iterations: 6
```

Varijabla spola (sexM) ima značajan utjecaj na prolaznost, pri čemu negativni koeficijent za muški spol (-1.5337) ukazuje da muški studenti imaju nižu vjerojatnost prolaska u usporedbi sa ženskim studentima. P-vrijednost od 0.0009 potvrđuje statističku značajnost ovog efekta.

Studenti iz manjih obitelji (LE3) imaju nešto veću šansu za prolaz (OR = 0.6494), ali p-vrijednost od 0.1913 sugerira da ovaj rezultat nije statistički značajan.

Obrazovanje majke i oca nema statistički značajan utjecaj na prolaznost, s p-vrijednostima većim od 0.99 za obje varijable, što upućuje na to da obrazovanje roditelja nije ključno za predviđanje prolaznosti u ovom skupu podataka.

Bračni status roditelja (Pstatus): Koeficijent za bračni status roditelja (PstatusT) je negativan, ali nije značajan (p=0.494), što znači da bračni status roditelja nema značajan utjecaj na prolaznost.

Interpretacija koeficijenata u smislu omjera izgleda (Odds Ratio)

```
exp(coef(logistic_model))
```

```
## (Intercept)        sexM  famsizeLE3        Medu        Fedu    PstatusT
##  10.7309097   0.2157299   1.9144326   1.3978429   1.3410635   0.4853457
```

Omjer izgleda (Odds Ratio) za muški spol (sexM) iznosi 0.18, što znači da su muški studenti oko 82% manje vjerojatno da će položiti ispit u odnosu na ženske studente.

Predikcije

Na temelju izgrađenog logističkog modela, izračunavamo predviđene vjerojatnosti prolaza za svakog studenta koristeći funkciju predict(). Zatim, ove vjerojatnosti pretvaramo u binarne predikcije (1 = prošao, 0 = nije prošao) pomoću ifelse() funkcije, gdje studenti s vjerojatnostima većim ili jednakim 0.5 bivaju klasificirani kao prošli, dok ostali bivaju označeni kao neprolazni.

```
predicted_probabilities <- predict(logistic_model, type = "response")

# Kreiranje predikcija (1 = prošao, 0 = nije prošao)
predicted_classes <- ifelse(predicted_probabilities >= 0.5, 1, 0)
```

Confusion Matrix

```
confusionMatrix(
  factor(predicted_classes, levels = c(0, 1)),
  studentData$passed_por
)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0   0   0
##          1  29 341
##
##                Accuracy : 0.9216
##                  95% CI : (0.8894, 0.9469)
##     No Information Rate : 0.9216
##     P-Value [Acc > NIR] : 0.5492
##
##                   Kappa : 0
##
##  Mcnemar's Test P-Value : 1.999e-07
##
##             Sensitivity : 0.00000
##             Specificity : 1.00000
##          Pos Pred Value :     NaN
##          Neg Pred Value : 0.92162
##              Prevalence : 0.07838
##          Detection Rate : 0.00000
##    Detection Prevalence : 0.00000
##       Balanced Accuracy : 0.50000
##
##        'Positive' Class : 0
##
```

Model predviđa da svi studenti prolaze ispit, što je izravna posljedica neravnoteže u podacima, gdje je većina studenata prošla ispit. Zbog ove neravnoteže, model je sklon predviđanju većinske klase (prolaz), dok zanemaruje manjinsku klasu (neprolaz). Osjetljivost (sensitivity) modela iznosi 0%, što znači da model nije u stanju prepoznati niti jednog studenta koji nije prošao ispit.

ROC krivulja za procjenu performansi modela

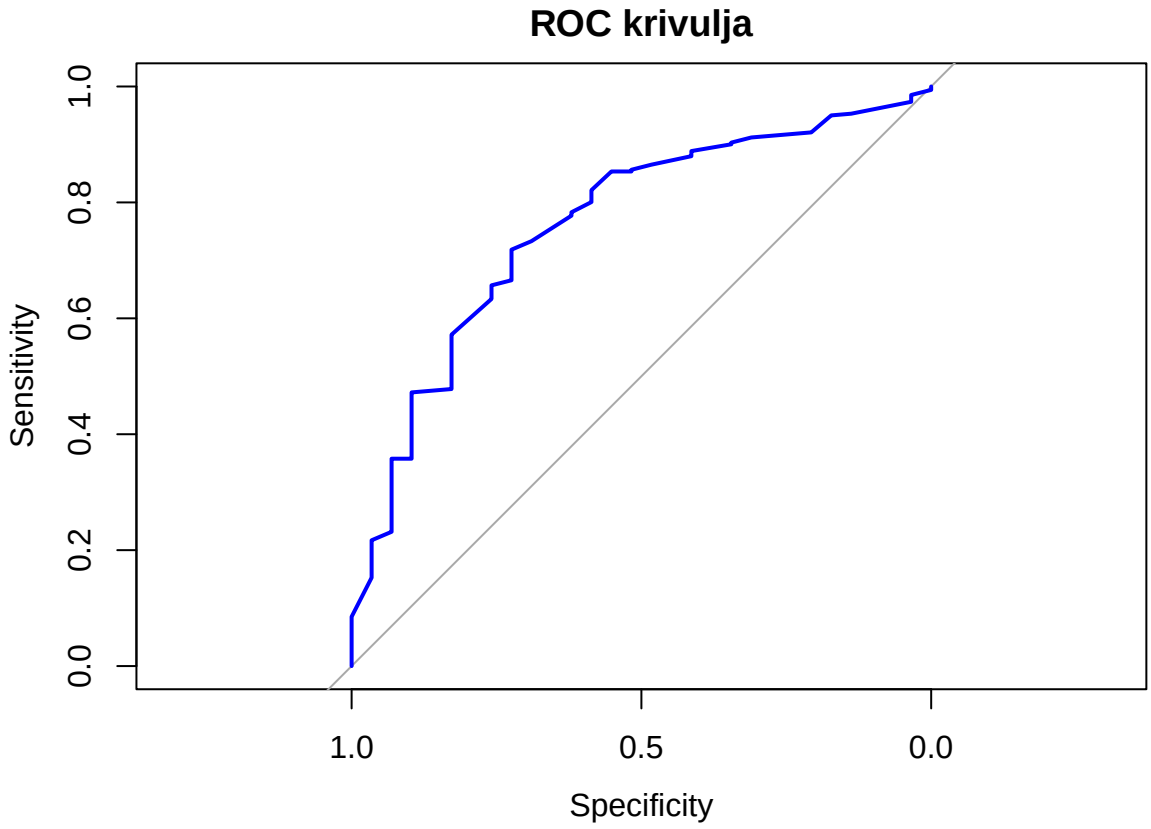
Za procjenu performansi modela koristimo ROC krivulju (Receiver Operating Characteristic), koja prikazuje odnos između osjetljivosti i specifičnosti modela pri različitim pragovima odluke.

```
roc_curve <- roc(studentData$passed_por, predicted_probabilities)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
# Plot ROC krivulje
plot(roc_curve, main = "ROC krivulja", col = "blue", lwd = 2)
```




```
cat("AUC:", auc(roc_curve), "\n")
```

```
## AUC: 0.7604915
```

Na temelju AUC (Area Under the Curve) vrijednosti, koja iznosi približno 0.7604915, možemo zaključiti da model ima dobru prediktivnu sposobnost.

Zaključci

- Spol je jedini statistički značajan prediktor u ovom modelu, pri čemu ženski studenti imaju veću vjerojatnost prolaska na ispitu u odnosu na muške studente.
- Model je pod utjecajem neravnoteže klasa (veći broj prolaznih slučajeva), što je očito iz confusion matrice, gdje model dominantno predviđa prolaz, zanemarujući studente koji nisu prošli.
- AUC vrijednost od 0.76 ukazuje na solidnu prediktivnu sposobnost modela, ali bilo bi potrebno dodatno razmotriti tehnike za uravnoteženje podataka (kao što su undersampling ili oversampling) kako bi model bolje prepoznavao i manjinsku klasu.
- Vizualizacije sugeriraju da spol ima značajan utjecaj na prolaznost, dok ostale sociodemografske varijable poput veličine obitelji i obrazovanja roditelja nemaju statistički značajan utjecaj na prolaznost u ovom modelu.

Postoji li razlika u broju izostanaka iz nastave?

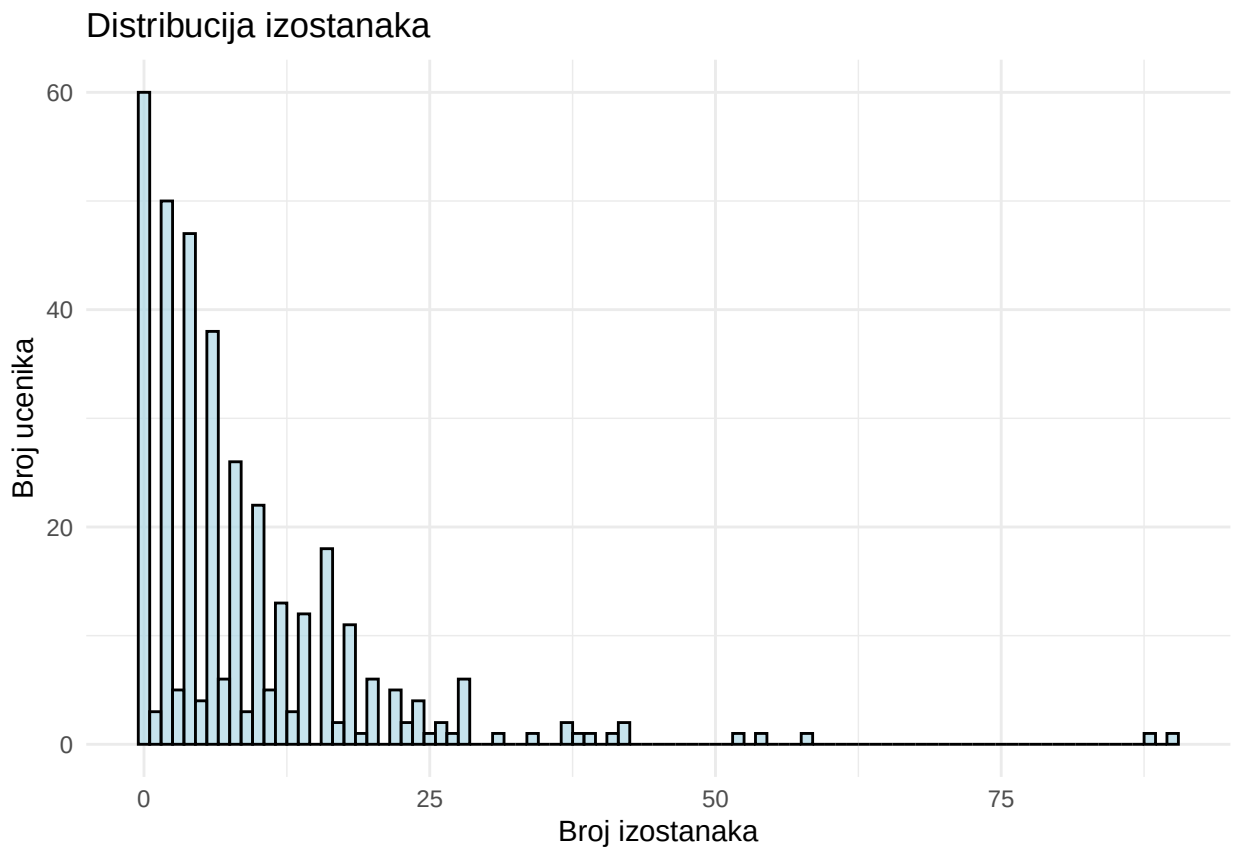
Sada ćemo se fokusirati na analizu broja izostanaka kod učenika i istražiti postoji li statistička povezanost između izostanaka i određenih sociodemografskih faktora. Za početak, izvršit ćemo nekoliko manipulacija podacima kako bismo ih pripremili za daljnju analizu.

Uređivanje podataka i provjera normalnosti

Za početak zbrajamo izostanke iz matematike i izostanke iz portugalskog jezika te gledamo ukupan broj izostanaka. Također provjeravamo normalnost podataka i vizualiziramo distribuciju izostanaka pomoću histograma.

```
studentData$total_absences <- studentData$absences_mat + studentData$absences_por

ggplot(studentData, aes(x = total_absences)) +
  geom_histogram(binwidth = 1, fill = "lightblue", color = "black", alpha = 0.7) +
  labs(
    title = "Distribucija izostanaka",
    x = "Broj izostanaka",
    y = "Broj učenika"
  ) +
  theme_minimal()
```



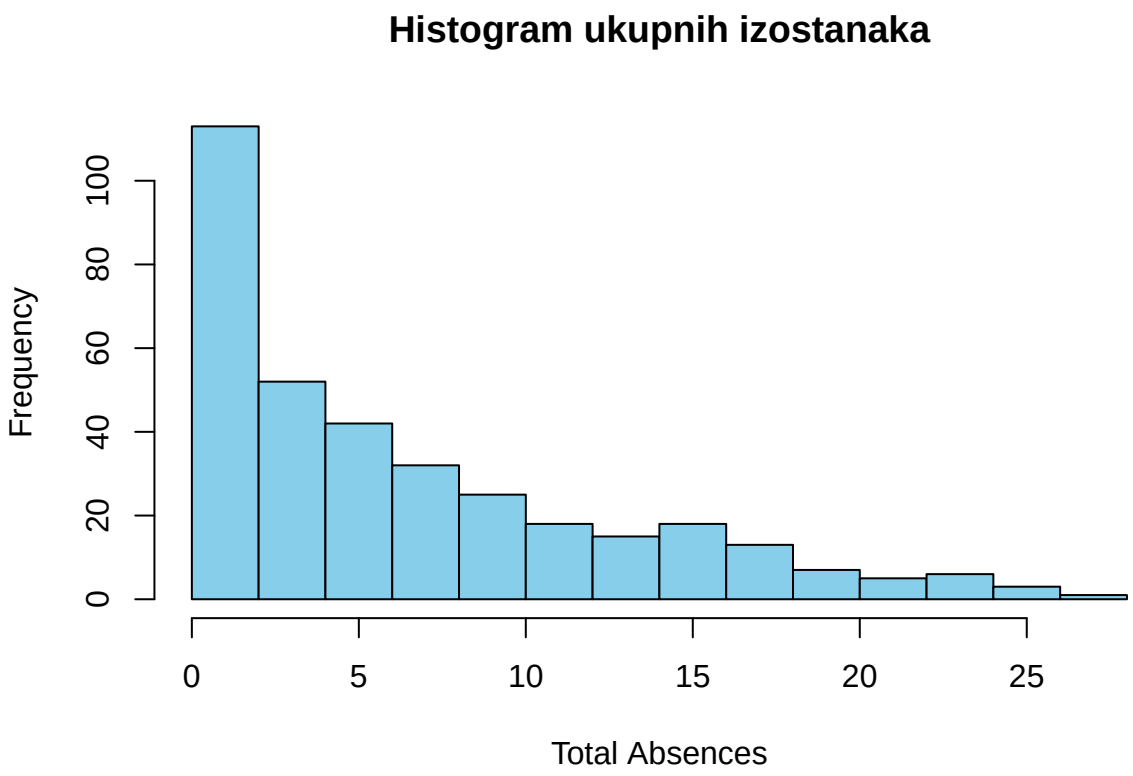
Zbog primjetne vidljivosti stršećih vrijednosti, radimo IQR metodu za odbacivanje stršećih vrijednosti u svrhu normaliziranja distribucije te prikazujemo histogram s podacima bez stršećih vrijednosti.

```
Q1 <- quantile(studentData$total_absences, 0.25)
Q3 <- quantile(studentData$total_absences, 0.75)
IQR <- Q3 - Q1

lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

cleaned_data <- studentData[studentData$total_absences >=
                            lower_bound & studentData$total_absences <= upper_bound, ]

hist(cleaned_data$total_absences, main = "Histogram ukupnih izostanaka",
     xlab = "Total Absences", col = "skyblue", border = "black")
```



Iako smo uspješno eliminirali stršeće vrijednosti, iz histograma je vidljivo da podaci nisu normalno distribuirani te za dodatnu provjeru normalnosti provodimo Shapiro-Wilk test nad podacima sa i bez stršećih vrijednosti.

```
shapiro.test(studentData$total_absences)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  studentData$total_absences
## W = 0.70904, p-value < 2.2e-16
```

```
shapiro.test(cleaned_data$total_absences)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cleaned_data$total_absences
## W = 0.89342, p-value = 6.393e-15
```

U oba slučaja, dobili smo jako malenu p-vrijednost te možemo potvrditi da podaci nisu normalno distribuirani s razinom značajnosti 0.05. Iz tog razloga, za daljnje analize koristit ćemo podatke koji uključuju stršeće vrijednosti jer želimo analizirati ponašanje svih učenika.

Odnos između izostanaka i mjesta stanovanja

Intuitivna ideja za ispitivanje razlike u broju izostanaka bila je ispitati postoji li statistička značajnost između ukupnog broja izostanaka i mjesta stanovanja. Na temelju ideje postavljamo zadane hipoteze:

H_0 : *Nema razlike u iznosu izostanaka između učenika iz ruralnih i urbanih područja*

H_1 : *Učenici iz ruralnih područja imaju veći broj izostanaka od učenika iz urbanih područja*

Provođenje testa:

S obzirom da podaci o izostancima nemaju normalnu distribuciju, koristit ćemo neparametarski Mann-Whitney U test, poznatiji i kao Wilcoxon rank sum test:

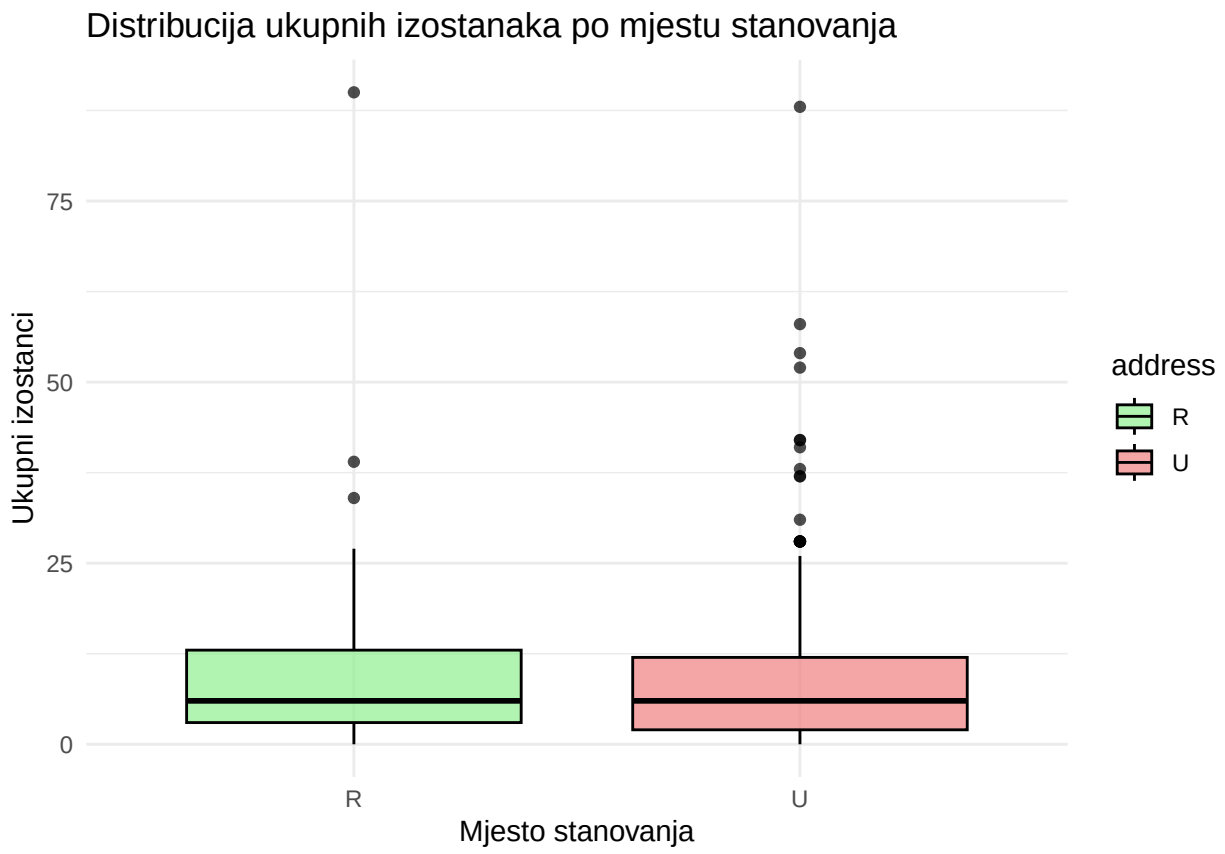
```
wilcox_result <- wilcox.test(total_absences ~ address, data = studentData)
print(wilcox_result)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  total_absences by address
## W = 12712, p-value = 0.2343
## alternative hypothesis: true location shift is not equal to 0
```

Zaključak

Dobivena p-vrijednost (gledamo p-vrijednost / 2 zbog korištenja jednostranog testa) veća je od 0.05 što znači da ne možemo odbaciti hipotezu H0, odnosno ne postoji statistički značajna razlika u broju izostanaka između učenika iz ruralnih i urbanih područja. Rezultat prikazujemo koristeći box plot.

```
ggplot(studentData, aes(x = address, y = total_absences, fill = address)) +
  geom_boxplot(color = "black", alpha = 0.7) +
  labs(
    title = "Distribucija ukupnih izostanaka po mjestu stanovanja",
    x = "Mjesto stanovanja",
    y = "Ukupni izostanci"
  ) +
  scale_fill_manual(values = c("lightgreen", "lightcoral")) +
  theme_minimal()
```



Utjecaj konzumiranja alkohola na broj izostanaka

Analizirat ćemo kako konzumacija alkohola tijekom tjedna utječe na razliku u broju izostanaka iz nastave. Za ovu analizu koristit ćemo varijablu Walc (weekend alcohol consumption), koja je numerička i mjeri učestalost konzumacije alkohola tijekom vikenda, s vrijednostima od 1 (vrlo niska) do 5 (vrlo visoka). Na temelju ovoga postavljamo sljedeće hipoteze:

H_0 : *Ne postoji razlika u distribuciji između različitih grupa konzumacije alkohola tijekom tjedna*

H_1 : *Postoji razlika u distribuciji između različitih grupa konzumacije alkohola tijekom tjedna*

Provođenje testa:

Budući da podaci o izostancima nisu normalno distribuirani koristimo neparametarsku metodu kako bismo provjerili postoji li statistički značajna razlika u distribuciji izostanaka između grupa. Provodimo Kruskal-Wallis test za razliku u broju izostanaka iz nastave u odnosu na konzumaciju alkohola tijekom tjedna.

```
kruskal_test_alcohol <- kruskal.test(total_absences ~ Dalc, data = studentData)
print(kruskal_test_alcohol)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  total_absences by Dalc
## Kruskal-Wallis chi-squared = 15.942, df = 4, p-value = 0.003098
```

Budući da je p-vrijednost < 0.05 odbacujemo nultu hipotezu te možemo zaključiti da postoji statistički značajna razlika u izostancima između grupa. Ovaj test ne daje izravno odgovor na to je li viša konzumacija alkohola povezana s većim brojem izostanaka, ali daje pokazatelj da razlike u izostancima postoje među grupama.

Post-hoc analiza

Nakon što smo dokazali postojanje razlika pomoću Kruskal-Wallis testa napravit ćemo post-hoc analizu kako bismo razjasnili koje se grupe međusobno razlikuju. Za to ćemo koristiti Pairwise Wilcoxon rank sum test koji uzima svaki par grupa iz podataka o konzumaciji alkohola i uspoređuje ih statistički koristeći Wilcoxon rank sum test. Bitno je napomenuti da smo postavili ograničenje na egzaktno računanje p-vrijednosti s postavljanjem varijable $exact = FALSE$ budući da podaci koje računamo sadrže puno ponovljenih vrijednosti. Hipoteze koje testiramo s Pairwise Wilcoxon rank sum testom su sljedeće:

H_0 : *Ne postoji razlika u distribuciji izostanaka između dviju grupa*

H_1 : *Postoji razlika u distribuciji izostanaka između dviju grupa*

Provođenje testa:

```
pairwise_wilcox_alcohol <- pairwise.wilcox.test(studentData$total_absences,
                                                studentData$Dalc,
                                                p.adjust.method = "BH", exact = FALSE)

print(pairwise_wilcox_alcohol)
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test with continuity correction
##
## data:  studentData$total_absences and studentData$Dalc
##
##   1     2     3     4
## 2 1.000 -     -     -
## 3 0.039 0.081 -     -
## 4 0.095 0.134 1.000 -
## 5 0.081 0.095 1.000 1.000
##
## P value adjustment method: BH
```

Analiza rezultata:

Na temelju p-vrijednosti između grupa zaključujemo da postoji statistički značajna razlika u broju izostanak između grupe 1 i grupe 3 jer je p-vrijednost manja od 0.05 te odbacujemo hipotezu H0, koja tvrdi suprotno. Rezultate analize prikazujemo koristeći box plot.

```
ggplot(studentData, aes(x = factor(Dalc), y = total_absences, fill = factor(Dalc))) +
  geom_boxplot() +
  labs(
    title = "Box plot izostanaka u odnosu na konzumaciju alkohola tjedno",
    x = "Konzumacija alkohola tjedno",
    y = "Izostanci"
```

```
) +
scale_fill_brewer(palette = "Set3", labels = c("Vrlo rijetko",
                                "Rijetko", "Umjereno", "Često", "Vrlo često")) +
theme_minimal() +
theme(legend.position = "none")
```

Ova razlika može ukazivati na to da povećanje konzumacije alkohola između "vrlo rijetko" i "umjereno" može imati neki utjecaj na broj izostanaka. Dakle, studenti u grupi 3 (umjereno) mogu imati više izostanaka od onih u grupi 1 (vrlo rijetko). Kakogod, primjećujemo da nema statistički značajnih razlika između grupa 1 i 4 ili grupa 1 i 5. To može biti zato što su distribucije podataka unutar tih grupa sličnije nego što je to slučaj između grupa 1 i 3.

Zaključak

Ukratko, iako postoji statistički značajna razlika u izostancima između grupa 1 i 3, odnosno povećana konzumacija alkohola tijekom tjedna može biti povezana s većim brojem izostanaka, ta povezanost nije dovoljno jaka kako bismo mogli dovesti zaključak nad cijelim skupom podataka.

Zaključak

Pokazali smo da muški studenti imaju višu prosječnu završnu ocjenu iz matematike u odnosu na žene, dok nije utvrđena značajna razlika u početnim ocjenama iz matematike između studenata iz urbanih i ruralnih područja. Također smo pokušali izgraditi model logističke regresije kako bismo predvidjeli prolaz na završnom ispitu iz jezika temeljen na sociodemografskim faktorima. Rezultati modela pokazali su da je spol jedini statistički značajan prediktor prolaznosti, pri čemu ženske studente imaju veću vjerojatnost prolaska na ispitu u odnosu na muške. Model je bio pod utjecajem neravnoteže klasa, budući da većina studenata prolazi ispit, što je vidljivo u confusion matrici gdje model dominantno predviđa prolaz. AUC vrijednost od 0.76 pokazuje solidnu prediktivnu sposobnost modela, no kako bi model bolje prepoznao manjinsku klasu, preporučuje se daljnja primjena tehnika uravnoteženja podataka poput undersampling ili oversampling.

Na kraju, istražena je povezanost između broja izostanaka i mjesta stanovanja, kao i konzumacije alkohola tijekom tjedna. Rezultati nisu pokazali značajnu razliku u broju izostanaka između studenata iz urbanih i ruralnih područja, ali su ukazali na to da povećana konzumacija alkohola može biti povezana s većim brojem izostanaka. Iako je analiza konzumacije alkohola sugerirala postojanje povezanosti, potrebna su daljnja istraživanja kako bi se donijeli čvrsti zaključci.
