Sveučilište u Rijeci
Fakultet informatike
i digitalnih tehnologija

Sveučilišni prijediplomski studij Informatika

Zara Čubranić

Istraživačka analiza skupa podataka o bolestima srca

Završni rad

Mentor: prof. dr. sc. Maja Matetić

Rijeka, 25.08.2025.

Rijeka, 15. travanj 2025.

Zadatak za završni rad

Pristupnik/ica: Zara Čubranić

Naziv završnog rada: Istraživačka analiza skupa podataka o bolestima srca

Naziv završnog rada na engleskom jeziku: Exploratory analysis of a heart disease dataset

Sadržaj zadatka:

Zadatak završnog rada je izvršiti pripremu i istraživačku analizu skupa sa podacima o bolestima srca, s ciljem razumijevanja faktora koji utječu na srčane bolesti. Potrebno je primijeniti postupke iz područja dekriptivne i inferencijalne statistike, kao što su vizualizacija i linearna regresija. Potrebno je interpretirati dobivene rezultate.

Mentorica
Prof. dr. sc. Maja Matetić

_____

Voditelj za završne radove
Izv. prof. dr. sc. Miran Pobar

_____

Zadatak preuzet: 15. travanj 2025.

_____

(potpis pristupnika/ice)

Sažetak

Ovaj rad bavi se analizom podataka vezanih uz srčane bolesti s ciljem prepoznavanja i razumijevanja faktora rizika povezanih s pojavom srčanog udara. Korišten je Heart Attack Analysis & Prediction Dataset, koji sadrži kliničke podatke pacijenata, poput dobi, spola, krvnog tlaka, kolesterola i drugih medicinskih mjera. Metodologija rada obuhvaća učitavanje, čišćenje i uređivanje podataka, njihovu obradu uz pomoć paketa *dplyr* te vizualizaciju pomoću paketa *ggplot2*. Poseban naglasak stavljen je na primjenu osnovnih statističkih metoda i tehnika, kao što su distribucije, mjere centralne tendencije, korelacije te jednostruka i višestruka linearna regresija.

Dobiveni rezultati ukazuju na postojanje određenih obrazaca, primjerice povezanost između dobi i povišenog krvnog tlaka te dobi i razine kolesterola. Vizualizacije i regresijska analiza dodatno su pomogle u boljem razumijevanju odnosa među varijablama. Zaključci rada naglašavaju važnost pravovremenog prepoznavanja i praćenja čimbenika rizika za srčani udar, čime se može doprinijeti boljoj prevenciji i unapređenju zdravstvene skrbi.

Ključne riječi: srčane bolesti, analiza podataka, statistika, vizualizacija, linearna regresija, rizik, kolesterol, krvni tlak

Sadržaj

1. Uvod

Analiza i predviđanje srčanog udara od velike je važnosti jer doprinosi učinkovitijoj zdravstvenoj skrbi i boljem razumijevanju čimbenika rizika. Srčani udar jedan je od vodećih uzroka smrti u svijetu, stoga je nužno istražiti koje varijable najviše utječu na njegovo pojavljivanje. Prepoznavanje tih varijabli može pomoći u razvijanju preventivnih mjera i odgovarajućih intervencija.

Korištenjem dostupnih podataka moguće je istražiti odnose između različitih čimbenika poput dobi, spola, krvnog tlaka, razine kolesterola. Analiza takvih podataka otvara prostor za donošenje zaključaka temeljenih na dokazima i podržava donošenje odluka u zdravstvenom sustavu.

U ovom radu korišten je *Heart Attack Analysis & Prediction Dataset* [1], koji sadrži kliničke podatke pacijenata. Na temelju ovog skupa podataka provedene su različite metode obrade, statističke analize i vizualizacije, a u nastavku je u Tablici 1 prikazan manji uzorak redaka iz dataseta.

Tablica 1. Uzorak podataka iz dataseta

age	**sex**	**cp**	**trtbps**	**chol**	**fbs**	**restecg**	**thalachh**	**exng**	**oldpeak**	**slp**	**caa**	**thall**	**output**
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1
56	0	1	140	294	0	0	153	0	1.3	1	0	2	1
44	1	1	120	263	0	1	173	0	0	2	0	3	1
52	1	2	172	199	1	1	162	0	0.5	2	0	3	1
57	1	2	150	168	0	1	174	0	1.6	2	0	2	1

Varijable:

1. **Age (Dob):** Dob pacijenta.
2. **Sex (Spol):** Spol pacijenta.
3. **cp (Tip bolova u prsima):** Klasifikacija tipova bolova u prsima prema četiri vrijednosti:
    - Tipična angina
    - Atipična angina
    - Neanginalni bol
    - Asimptomatski
4. **trtbps (Krvni tlak u mirovanju):** Krvni tlak pacijenta u mirovanju (u mm Hg).
5. **chol (Kolesterol):** Razina kolesterola u krvi (mg/dl) dobivena putem BMI senzora.
6. **fbs (Post prandijalni šećer u krvi):** Informacija o tome je li post prandijalni šećer u krvi pacijenta veći od 120 mg/dl (1 = istina, 0 = laž).
7. **restecg (Elektrokardiografski rezultati u mirovanju):** Rezultati elektrokardiograma u mirovanju, klasificirani u tri vrijednosti:
    - Normalno
    - ST-T valna abnormalnost
    - Pokazuje vjerojatnu ili definitivnu hipertrofiju lijeve klijetke prema kriterijima Estes.
8. **thalachh (Maksimalni postignuti otkucaj srca):** Najviša postignuta brzina otkucaja srca.
9. **exng (Inducirana angina tijekom vježbanja):** Informacija o tome je li pacijent doživio induciranu anginu tijekom vježbanja (1 = da, 0 = ne).
10. **oldpeak (Prethodni vrh):** Numerička vrijednost koja označava prethodni vrh.
11. **slp (Nagib):** Nagib.
12. **caa (Broj glavnih krvnih žila):** Broj glavnih krvnih žila (0-3).
13. **thall (Thal rate):** Thal rate.
14. **output (Izlazna varijabla):** Ciljna varijabla koja označava rizik od srčanog udara (0 = manji rizik, 1 = veći rizik).

2. Metodologija i korišteni alati

2.1. Paket *dplyr*

```
library(dplyr)
```

Za obradu podataka korišten je paket *dplyr*, koji je dio šireg programskog okruženja *tidyverse* [2]. *Tidyverse* predstavlja skup međusobno povezanih R paketa razvijenih za rad s podacima u tzv. *tidy* formatu, gdje svaki stupac predstavlja varijablu, svaki redak predstavlja opažanje, a svaka ćelija jednu vrijednost. Ovakva struktura omogućuje veću preglednost i jednostavnije izvođenje analiza.

Unutar *tidyverse*-a, paket *dplyr* zauzima važno mjesto jer nudi intuitivne funkcije za filtriranje, odabir i transformaciju podataka [3]. Ključne funkcije uključuju:

- filter() za odabir redaka koji zadovoljavaju određene uvjete,
- select() za izdvajanje relevantnih stupaca,
- mutate() za stvaranje novih varijabli,
- summarise() za dobivanje sažetih statističkih mjera,
- group_by() za grupiranje podataka prije daljnjih izračuna.

Sve se funkcije najčešće povezuju operatorom cijevi (%>%), što omogućuje čitljiviji prikaz koraka obrade. Umjesto gniježđenja funkcija unutar drugih, svaka operacija se piše u novom retku, čime je analiza preglednija i razumljivija. Upravo zbog toga *dplyr* predstavlja standardan alat u modernim analizama podataka u R-u.

2.2. Paket ggplot2

Za izradu grafova u ovom radu korišten je paket *ggplot2*, koji je dio šireg okvira *tidyverse* [2]. Radi se o jednom od najpopularnijih alata za vizualizaciju u R-u jer omogućuje izradu estetski privlačnih i prilagodljivih grafova uz relativno jednostavnu sintaksu. Prednost *ggplot2* paketa je što se temelji na "gramatici grafike" (eng. Grammar of Graphics), što znači da grafovi nisu samo unaprijed zadani tipovi, nego se mogu fleksibilno graditi i kombinirati prema potrebama istraživača [3].

```
library(ggplot2)
```

3. Linearna regresija

Linearna regresija statistička je metoda koja modelira odnos između jedne zavisne (odgovorne) varijable i jedne ili više nezavisnih (objašnjavajućih) varijabli. Cilj je pronaći pravac (u slučaju jedne nezavisne varijable – jednostavna linearna regresija) koji najbolje opisuje taj odnos korištenjem linearne jednadžbe (1) poput:

$$Y = mX + b \tag{1}$$

gdje je:

Y zavisna varijabla,

X nezavisna varijabla,

m nagib regresijske linije, a

b presjek s osi y.

Nagib regresijske linije (2) i presjek s osi y (3) može se izračunati metodom najmanjih kvadrata.

$$m = \frac{n(\Sigma xy) - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} \tag{2}$$

i

$$b = \frac{\Sigma y - m\Sigma x}{n} \tag{3}$$

gdje je:

Σ suma svih vrijednosti, a

n broj uzoraka.

Metoda najmanjih kvadrata (engl. Ordinary Least Squares) koristi se za procjenu optimalnih vrijednosti β_0 i β_1 kako bi se dobio minimalni zbroj kvadrata pogrešaka. Cilj ove metode je odrediti parametre koji imaju najmanju vertikalnu udaljenost između predviđenih i stvarnih *y* vrijednosti [5].

Za izračun linearne regresije često se koristi i sljedeća formula (4):

$$y = \beta_0 + \beta_1 x + \epsilon \tag{4}$$

gdje je:

y zavisna varijabla (engl. dependent variable)

x nezavisna varijabla (engl. independent variable)

ϵ komponenta slučajne pogreške (engl. random error component)

β_0 presjek (engl. intercept)

β_1 koeficijent od x (engl. coefficient of *x*) [6]

4. Učitavanje podataka

Za učitavanje podataka korištena je funkcija read_csv() iz paketa readr, kojom je učitan skup podataka „heart.csv“. Nakon uspješnog učitavanja prikazani su prvi redci tablice kako bi se dobio uvid u strukturu i sadržaj podataka.

```
# Učitavanje paketa
library(readr)

# Učitavanje podataka
heart_data <- read_csv("heart.csv")

# Pregled prvih nekoliko redaka i strukture
head(heart_data)
str(heart_data)
```

A tibble: 6 × 14

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1

```
spc_tbl_ [303 × 14] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ age     : num [1:303] 63 37 41 56 57 57 56 44 52 57 ...
 $ sex     : num [1:303] 1 1 0 1 0 1 0 1 1 1 ...
 $ cp      : num [1:303] 3 2 1 1 0 0 1 1 2 2 ...
 $ trtbps  : num [1:303] 145 130 130 120 120 140 140 120 172 150 ...
 $ chol    : num [1:303] 233 250 204 236 354 192 294 263 199 168 ...
 $ fbs     : num [1:303] 1 0 0 0 0 0 0 0 1 0 ...
 $ restecg : num [1:303] 0 1 0 1 1 1 0 1 1 1 ...
 $ thalachh: num [1:303] 150 187 172 178 163 148 153 173 162 174 ...
 $ exng    : num [1:303] 0 0 0 0 1 0 0 0 0 0 ...
 $ oldpeak : num [1:303] 2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
 $ slp     : num [1:303] 0 0 2 2 2 1 1 2 2 2 ...
 $ caa     : num [1:303] 0 0 0 0 0 0 0 0 0 0 ...
 $ thall   : num [1:303] 1 2 2 2 2 1 2 3 3 2 ...
 $ output  : num [1:303] 1 1 1 1 1 1 1 1 1 1 ...
 - attr(*, "spec")=
  .. cols(
  ..   age = col_double(),
  ..   sex = col_double(),
  ..   cp = col_double(),
  ..   trtbps = col_double(),
  ..   chol = col_double(),
  ..   fbs = col_double(),
  ..   restecg = col_double(),
  ..   thalachh = col_double(),
  ..   exng = col_double(),
  ..   oldpeak = col_double(),
  ..   slp = col_double(),
  ..   caa = col_double(),
  ..   thall = col_double(),
  ..   output = col_double()
  .. )
```

Slika 1. Prvih nekoliko redaka i struktura dataseta

5. Reprezentacija datuma i vremenskih oznaka

Podaci o datumima i vremenima rođenja nisu bili dostupni u originalnom datasetu, stoga su oni generirani nasumično. Godine rođenja su izračunate na temelju starosti pacijenata. Nakon toga su generirani su slučajni datumi i vremena rođenja koje sam dodala u dataset. Ti podaci su zatim razdvojeni u zasebne stupce za datum i vrijeme rođenja. Razdvojeni stupci su na kraju spojeni u jedan za datum i vrijeme rođenja radi boljeg prikaza podataka.

```
library(dplyr)
library(lubridate)


# Generiranje godina rođenja na temelju dobi
godina_rodenja <- 2024 - heart_data$age


# Generiranje slučajnih datuma i vremena
set.seed(123)  # Postavljanje sjemena za reproduktivnost
datum_rodenja <- as.Date(paste0(godina_rodenja, "-01-01")) + sample(0:365, nrow(heart_data), replace = TRUE)
datum_vrijeme_rodenja <- as_datetime(paste0(datum_rodenja, " ", sample(0:23, nrow(heart_data), replace = TRUE), ":", sample(0:59, nrow(heart_data), replace = TRUE), ":", sample(0:59, nrow(heart_data), replace = TRUE)))


# Dodavanje generiranih podataka u dataset
heart_data$datum_rodenja <- datum_rodenja
heart_data$datum_vrijeme_rodenja <- datum_vrijeme_rodenja


head(heart_data[,c(1, 17, 18)])
```

A tibble: 6 × 3

age	datum_rodenja	datum_vrijeme_rodenja
<dbl>	<date>	<dttm>
63	1961-06-28	1961-06-28 10:47:48
37	1987-01-14	1987-01-14 00:07:05
41	1983-07-14	1983-07-14 18:52:24
56	1968-11-01	1968-11-01 11:02:41
57	1967-04-28	1967-04-28 01:43:25
57	1967-10-26	1967-10-26 20:52:23

Slika 2. Reprezentacija datuma i vremenskih oznaka

6. Uređivanje podataka

Podaci su uređeni kako bi se olakšala analiza. To uključuje pretvaranje vrijednosti u stupcima 'sex', 'fbs', 'exng' i 'output' u prikladnije kategorije. Također su dodani novi stupci za prilagođenu kategorizaciju kolesterola i krvnog tlaka.

```
# Pretvaranje vrijednosti u stupcima 'sex', 'fbs', 'exng' i 'output'
heart_data <- heart_data %>%
  mutate(sex = ifelse(sex == 1, "M", "F"),
         fbs = ifelse(fbs == 1, "T", "F"),
         exng = ifelse(exng == 1, "T", "F"),
         output = ifelse(output == 1, "X", "O"))
# Dodavanje kategotrija za stupac kolesterola
heart_data <- heart_data %>%
  mutate(chol_category = case_when(
    chol < 200 ~ "normalno",
    chol >= 200 & chol <= 239 ~ "visoko",
    chol >= 240 ~ "opasno",
  ))
# Dodavanje kategorija za stupac trtbps
heart_data <- heart_data %>%
  mutate(trtbps_category = case_when(
    trtbps < 120 ~ "normalno",
    trtbps >= 120 & trtbps < 140 ~ "visoko",
    trtbps >= 140 ~ "opasno"
  ))
# Uklanjanje redaka s nedostajućim vrijednostima
heart_dropna <- drop_na(heart_data)
# Provjera promjena
head(heart_data)
```

A tibble: 6 × 5

age	sex	fbs	exng	output
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
63	1	1	0	1
37	1	0	0	1
41	0	0	0	1
56	1	0	0	1
57	0	0	1	1
57	1	0	0	1

A tibble: 6 × 5

age	sex	fbs	exng	output
<dbl>	<chr>	<chr>	<chr>	<chr>
63	M	T	F	X
37	M	F	F	X
41	F	F	F	X
56	M	F	F	X
57	F	F	T	X
57	M	F	F	X

Slika 3 Uređivanje podataka – prije i poslije

Slika 3 prikazuje mogućnosti promjene formata podataka koji omogućuje lakše čitanje podataka i daljnju analizu.

A tibble: 6 × 4

trtbps	chol	chol_category	trtbps_category
<dbl>	<dbl>	<chr>	<chr>
145	233	visoko	opasno
130	250	opasno	visoko
130	204	visoko	visoko
120	236	visoko	visoko
120	354	opasno	visoko
140	192	normalno	opasno

Slika 3. Dodavanje stupaca za kategorije

Slika 4 prikazuje tablicu koja sadrži podatke o razinama krvnog tlaka („trtbps“) i kolesterola („chol“) i uz pripadajuće kategorije za krvni tlak („trtbps_category“) i kolesterol („chol_category“).

Funkcije separate i unite su korištene za obradu podataka o datumima i vremenima rođenja, čime su razdvojeni i ponovno spojeni kako bi se dobile jasne informacije o ovim varijablama.

```
# Provjera promjena
head(heart_data[,c(1, 17)])


# Razdvajanje datuma i vremena rođenja
heart_data <- heart_data %>%
  separate(col = datum_vrijeme_rodenja, into = c("datum_rodenja", "vrijeme_rodenja"), sep =
" ")


# Provjera promjena
head(heart_data[,c(1, 17, 18)])


# Spajanje datuma i vremena rođenja
heart_data <- heart_data %>%
  unite(col = "datum_vrijeme_rodenja", c("datum_rodenja", "vrijeme_rodenja"), sep = " ")


# Provjera promjena
head(heart_data[,c(1, 17)])
```

A tibble: 6 × 2

age	datum_vrijeme_rodenja
<dbl>	<chr>
63	1961-06-28 10:47:48
37	1987-01-14 00:07:05
41	1983-07-14 18:52:24
56	1968-11-01 11:02:41
57	1967-04-28 01:43:25
57	1967-10-26 20:52:23

A tibble: 6 × 3

age	datum_rodenja	vrijeme_rodenja
<dbl>	<chr>	<chr>
63	1961-06-28	10:47:48
37	1987-01-14	00:07:05
41	1983-07-14	18:52:24
56	1968-11-01	11:02:41
57	1967-04-28	01:43:25
57	1967-10-26	20:52:23

A tibble: 6 × 2

age	datum_vrijeme_rodenja
<dbl>	<chr>
63	1961-06-28 10:47:48
37	1987-01-14 00:07:05
41	1983-07-14 18:52:24
56	1968-11-01 11:02:41
57	1967-04-28 01:43:25
57	1967-10-26 20:52:23

Slika 4. Uređivanje podataka - prije, nakon separate, nakon unite

Slika 5 prikazuje proces transformacije podataka o datumu i vremenu ređenja. Prvi stupac „datum_vrijeme_rodenja“ je funkcijom *separate()* podijeljen na dva zasebna stupca, „datum_rodenja“ i „vrijeme_rodenja“. Nakon analize, ti su stupci ponovno spojeni u originalni format pomoću funkcije *unite(),* čime se osigurava urednost i konzistentnost skupa podataka te prikazuje mogućnosti promjene formata podataka u svrhu daljnje analize.

7. Priprema podatkovnih okvira

U prvom koraku iz izvornog podatkovnog okvira izdvojeni su samo zapisi koji se odnose na pacijente u riziku od srčanog udara. Budući da je cilj istraživanja bio usmjeren upravo na ovu skupinu, izdvajanje je omogućilo preciznije fokusiranje na relevantne slučajeve. Nakon toga odabrani su samo oni stupci koji su bili ključni za analizu: dob, spol, krvni tlak, kolesterol i izlazna varijabla.

```
# Filtriranje
heart_data_rizik <- heart_data %>%
  filter(output == "X")

# Odabir podskupa stupaca
heart_data_rizik_subset <- heart_data_rizik %>%
  select(age, sex, trtbps, chol, output)
```

Nakon filtriranja i odabira stupaca, kreirani su novi varijabilni stupci koji omogućuju lakšu interpretaciju i daljnju obradu podataka. Krvni tlak podijeljen je u tri kategorije: normalno (ispod 120 mmHg), visoko (od 120 do 139 mmHg) i opasno (140 mmHg i više). Na isti način klasificiran je i kolesterol: normalno (ispod 200 mg/dl), visoko (200–239 mg/dl) i opasno (240 mg/dl i više).

Dodatno, varijabla spola pretvorena je u numerički oblik, pri čemu je muški spol označen vrijednošću 1, a ženski vrijednošću 0. Ovakva transformacija nužna je jer se u daljnjim statističkim postupcima zahtijevaju numeričke varijable.

```
# Stvaranje novih stupaca
heart_data_rizik_subset <- heart_data_rizik_subset %>%
  mutate(
    trtbps_risk_category = case_when(
      trtbps < 120 ~ "normalno",
      trtbps >= 120 & trtbps < 140 ~ "visoko",
      trtbps >= 140 ~ "opasno"
    ),
    chol_risk_category = case_when(
      chol < 200 ~ "normalno",
      chol >= 200 & chol < 240 ~ "visoko",
      chol >= 240 ~ "opasno"
    ),
    sex_numeric = ifelse(sex == "M", 1, 0)
  )
```

Na kraju pripreme podataka izrađena je matrica korelacija između ključnih varijabli: dobi, krvnog tlaka, kolesterola i spola. Dobiveni rezultati omogućuju uvid u smjer i jačinu povezanosti među varijablama, što je korisno za kasniju interpretaciju potencijalnih obrazaca i poveznica između rizičnih čimbenika.

```
# Izračun korelacija
cor_matrix <- cor(heart_data_rizik_subset %>%
                    select(age, trtbps, chol, sex_numeric))
```

8. Deskriptivna statistika i analiza podataka

Nakon što su podaci pripremljeni i kategorizirani, provedena je osnovna deskriptivna analiza. Cilj ovog dijela istraživanja bio je dobiti uvid u osnovne karakteristike podataka i prepoznati potencijalne obrasce unutar populacije pacijenata u riziku od srčanog udara.

Najprije je izračunata prosječna dob pacijenata u riziku, što pruža okvirnu sliku o životnoj dobi u kojoj se najčešće javljaju simptomi.

```
# Prosječna dob pacijenata u riziku
rizik_udara <- heart_data %>%
  filter(output == "1") %>%
  summarise(prosjek_dobi = mean(age))
print("Prosječna dob pacijenata u najvecem riziku:")
print(rizik_udara)
```

Dalje su izračunati srednja vrijednost, medijan, najmanja i najveća vrijednost krvnog tlaka. Ovi pokazatelji pružaju osnovne informacije o distribuciji vrijednosti unutar uzorka i omogućuju uočavanje odstupanja.

```
# Izračun srednje vrijednosti i medijana
mean_trtbps <- mean(heart_data$trtbps)
median_trtbps <- median(heart_data$trtbps)
min_trtbps <- min(heart_data$trtbps)
max_trtbps <- max(heart_data$trtbps)
```

Nakon toga podaci o krvnom tlaku i kolesterolu razvrstani su u tri kategorije: normalno, visoko i opasno. Ova kategorizacija omogućuje jasniju interpretaciju i lakšu vizualnu prezentaciju.

```
# Postavljanje redoslijeda kategorija za krvni tlak i kolesterol
heart_data$trtbps_category <- factor(heart_data$trtbps_category,
                                     levels = c("normalno", "visoko", "opasno"))
heart_data$chol_category <- factor(heart_data$chol_category,
                                   levels = c("normalno", "visoko", "opasno"))
```

Za prikaz distribucije vrijednosti izračunata je učestalost svake od kategorija.

```
# Distribucija kategorija krvnog tlaka
trtbps_category_distribution <- heart_data %>%
  count(trtbps_category)

# Distribucija kategorija kolesterola
chol_category_distribution <- heart_data %>%
  count(chol_category)
```

Rezultati su ispisani u obliku tablica koje prikazuju prosječne vrijednosti i učestalost pojedinih kategorija.

```
# Ispis rezultata
cat("Srednja vrijednost krvnog tlaka:", mean_trtbps, "\n")
cat("Medijan krvnog tlaka:", median_trtbps, "\n")
cat("Najmanja vrijednost krvnog tlaka:", min_trtbps, "\n")
cat("Najveća vrijednost krvnog tlaka:", max_trtbps, "\n\n")

cat("Distribucija kategorija krvnog tlaka:\n")
print(trtbps_category_distribution)
cat("\nDistribucija kategorija kolesterola:\n")
print(chol_category_distribution)
```

Na Slici 6. prikazan je rezultat deskriptivne statistike i kategorije krvnog tlaka i kolesterola. Ovaj dio analize predstavlja polaznu točku za daljnja statistička istraživanja.

```
[1] "Prosječna dob pacijenata u najvecem riziku:"
# A tibble: 1 × 1
  prosjek_dobi
         <dbl>
1         52.5
Srednja vrijednost krvnog tlaka: 131.6238
Medijan krvnog tlaka: 130
Najmanja vrijednost krvnog tlaka: 94
Najveća vrijednost krvnog tlaka: 200

Distribucija kategorija krvnog tlaka:
# A tibble: 3 × 2
  trtbps_category     n
  <fct>           <int>
1 normalno           60
2 visoko            146
3 opasno             97

Distribucija kategorija kolesterola:
# A tibble: 3 × 2
  chol_category     n
  <fct>         <int>
1 normalno         50
2 visoko           98
3 opasno          155
```

Slika 6. Deskriptivna statistika i distribucija krvnog tlaka i kolesterola

Utvrđeno je da je prosječna dob pacijenta u najvećem riziku 52,5 godina što je u skladu s općim saznanjima da se srčani problemi javljaju najčešće u srednjoj i starijoj dobi. Analiza centralne tendencije pokazuje da je srednja vrijednost krvnoga tlaka 131,62 mmHg, dok je medijan 130 mmHg što upućuje na prilično simetričnu raspodjelu vrijednosti. Izračunate su i krajnje vrijednosti, najmanja vrijednost prosječnog krvnog tlaka iznosi 94, a najveća 200.

Detaljnija analiza distribucije po kategorijama otkrila je jasne obrasce unutar uzorka. Kod krvnog tlaka, najveći broj pacijenata nalazi se u kategoriji "visoko", dok je najmanji broj pacijenata s "normalnim" vrijednostima. Sličan obrazac uočen je i kod kolesterola, gdje je najveći broj pacijenata u "opasnoj" kategoriji, a najmanji broj u "normalnoj". Sveukupno, rezultati deskriptivne statistike služe kao čvrst temelj za daljnje istraživanje odnosa između ključnih varijabli.

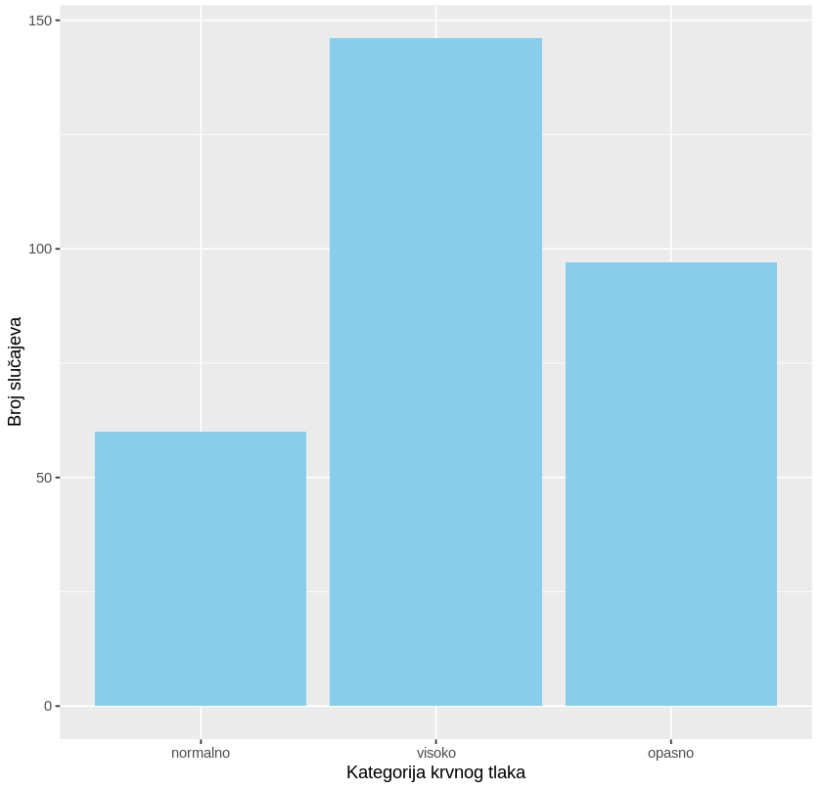
9. Vizualizacija podataka

Vizualizacija podataka jedan je od najvažnijih koraka u procesu analize jer omogućuje jednostavnije razumijevanje i interpretaciju rezultata. Često je puno lakše uočiti obrasce, trendove i odnose među varijablama kada se podaci prikažu grafički, nego kada ih promatramo isključivo kroz tablice i brojeve. Na taj način dobivamo intuitivniji uvid u strukturu skupa podataka, što može olakšati donošenje zaključaka i daljnje odluke u analizi.

9.1. Graf distribucije kategorija krvnog tlaka i kolesterola

Za analizu distribucije krvnog tlaka korišten je stupčasti graf prikazan na Slici 7. Ovakav graf odabran je zato što omogućuje brzi pregled učestalosti pacijenata unutar različitih kategorija krvnog tlaka: normalno, visoko i opasno. Na taj način možemo vizualno uočiti u kojoj se kategoriji nalazi najviše pacijenata, što je važna informacija za medicinsku interpretaciju podataka.

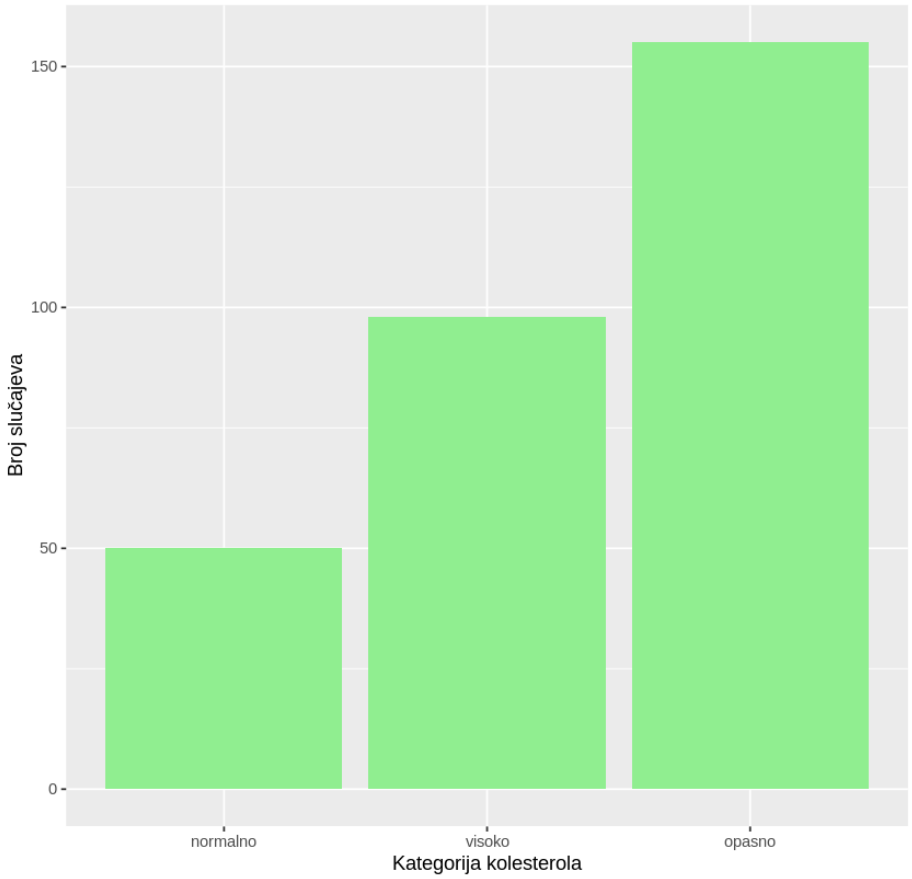
```
# Graf distribucije kategorija krvnog tlaka
ggplot(trtbps_category_distribution, aes(x = trtbps_category, y = n)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Distribucija kategorija krvnog tlaka",
       x = "Kategorija krvnog tlaka",
       y = "Broj slučajeva")
```



Slika 7. Stupčasti dijagram: kategorije krvnog tlaka

Slično prethodnom primjeru, i ovdje se koristi stupčasti graf, ali za kategorije kolesterola, prikazan na slici 8. Njegova svrha je slična – omogućiti jasan pregled strukture pacijenata prema kategorijama razine kolesterola: normalno, visoko i opasno.

```
# Graf distribucije kategorija kolesterola
ggplot(chol_category_distribution, aes(x = chol_category, y = n)) +
  geom_bar(stat = "identity", fill = "lightgreen") +
  labs(title = "Distribucija kategorija kolesterola",
       x = "Kategorija kolesterola",
       y = "Broj slučajeva")
```



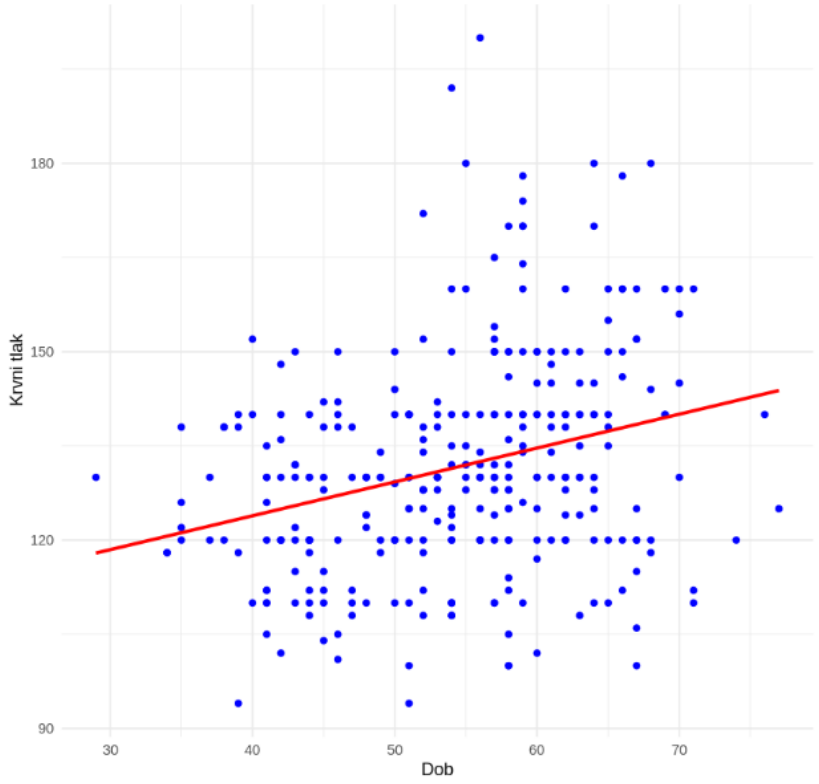
Slika 8. Stupčasti dijagram: kategorije kolesterola

Iz Slike 7 i Slike 8 jasno se vidi da je najzastupljenija kategorija za krvni tlak „visoko“, a ka kolesterol „opasno“.

9.2. Dijagram raspršivanja prosječnog krvnog tlaka i kolesterola prema dobi

Dijagram raspršivanja(engl. scatter plot) prikazan na Slici 9 prikazuje odnos između dobi pacijenata i njihovog krvnog tlaka. Ovaj graf odabran je jer jednostavno ilustrira povezanost dviju varijabli (dob i tlak), a uz dodatak regresijske linije omogućuje i lakše uočavanje općeg obrasca.

```
# Graf prosječnog krvnog tlaka prema dobi
ggplot(heart_data, aes(x = age, y = trtbps)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Prosječni krvni tlak prema dobi",
       x = "Dob",
       y = "Krvni tlak") +
  theme_minimal()
```

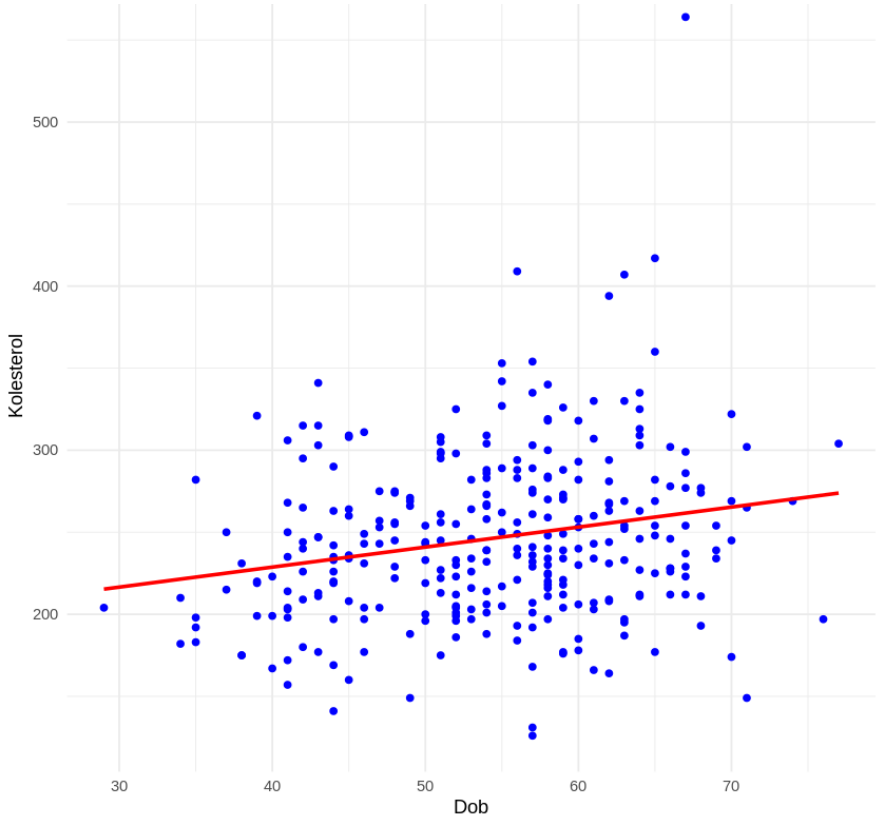


Slika 9. Dijagram raspršivanja: krvni tlak vs. Dob

Dijagram sa Slike 9 otkriva pozitivan linearni odnos između prosječne razine krvnoga tlaka i dobi ispitanika. Podatkovne točkice imaju veliko raspršenje što sugerira na slabiju, ali ipak postojeću povezanost dvaju faktora, Crvena linija koja na dijagramu predstavlja linearnu regresiju prikazuje da porastom dobi ispitanika raste i prosječna razina krvnog tlaka. Ovaj dijagram potvrđuje da je dob značajan faktor povezan s razinom krvnog tlaka.

Na isti način kao i prethodni primjer, napravljen je i dijagram raspršivanja na Slici 10 koji prikazuje odnos između kolesterola i dobi ispitanika. Cilj izrade ovog dijagrama je istražiti postoji li slična povezanost kolesterola u odnosu na dob kao što je zaključeno da postoji za krvni tlak.

```
# Grafikon kolesterola prema dobi
ggplot(heart_data, aes(x = age, y = chol)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Kolesterol prema dobi",
       x = "Dob",
       y = "Kolesterol") +
  theme_minimal()
```



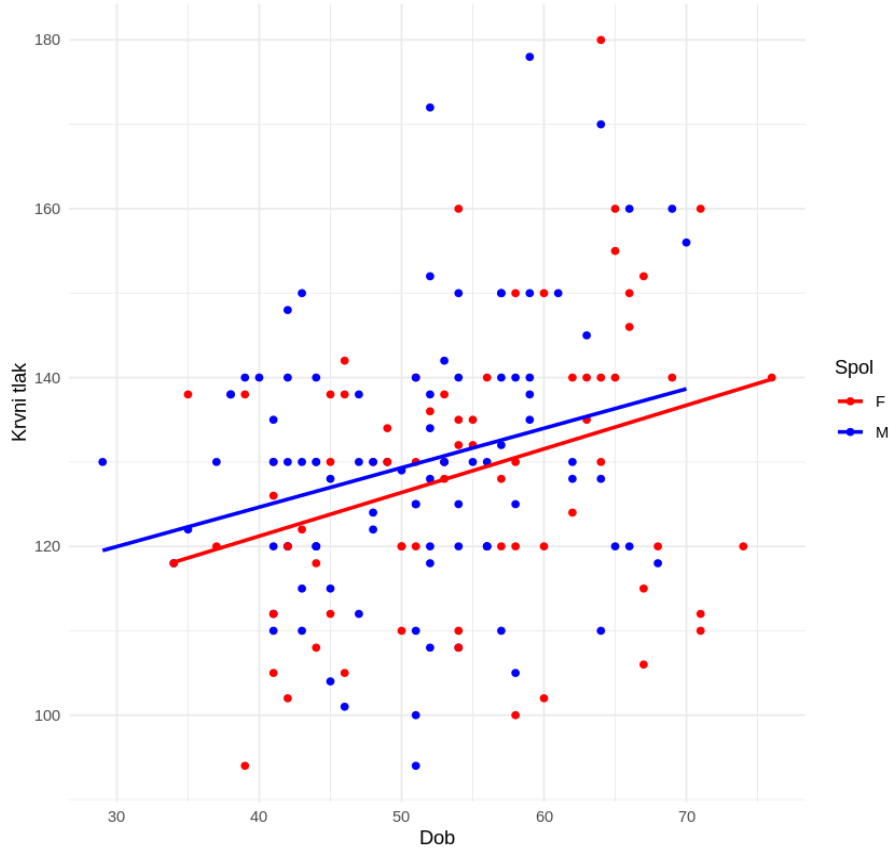
Slika 10. Dijagram raspršivanja: kolesterol vs. dob

Dijagram sa Slike 10 također pokazuje pozitivan linearni trend. Iako je raspršenje točaka šire nego kod krvnog tlaka, regresijska linija ipak ima pozitivan nagib što ukazuje na to da korelacija između tih dvaju faktora postoji. Zaključivo je da porastom dobi, u prosjeku raste i razina kolesterola.

9.3. Dijagram raspršivanja krvnog tlaka i kolesterola prema dobi (s bojanjem po spolu)

Nakon vizualne potvrde da postoji pozitivna korelacija dobi u odnosu na krvni tlak i kolesterol, sljedeći je korak istražiti postoje li kakve razlike s obzirom na spol ispitanika. U nastavku slijede dijagrami raspršivanja uz dodatno bojanje podatkovnih točaka ovisno o spolu. Dodane su i regresijske linije za svaki spol što omogućuje brzu i jednostavnu usporedbu trendova za oba spola.

```
# Scatter plot: dob vs. krvni tlak (s bojama po spolu)
p2 <- ggplot(heart_data_rizik_subset, aes(x = age, y = trtbps, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_manual(values = c("F" = "red", "M" = "blue")) +
  labs(title = "Dob vs. Krvni tlak",
       x = "Dob",
       y = "Krvni tlak",
       color = "Spol") +
  theme_minimal()
```

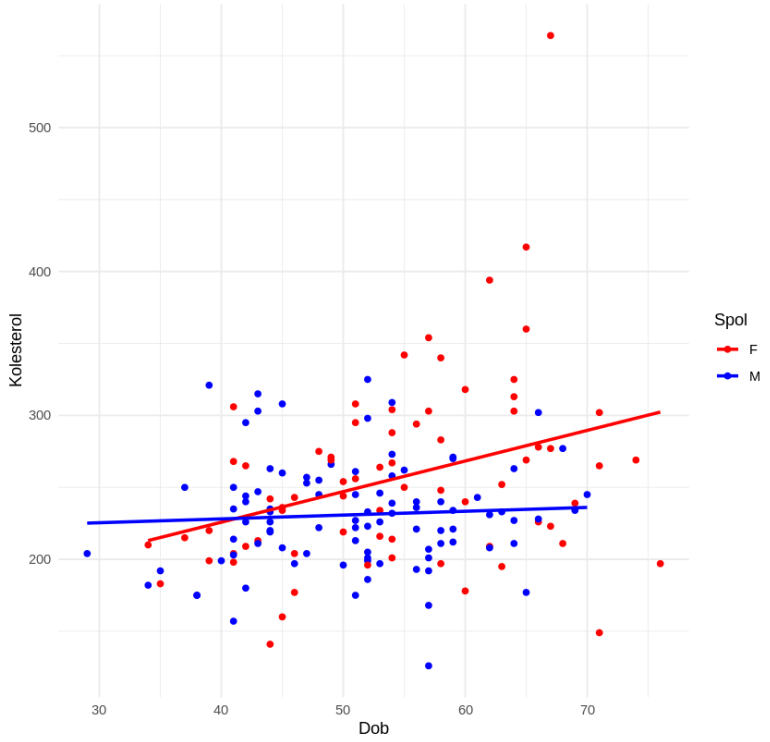


Slika 11. Dijagram raspršivanja: krvni tlak vs. dob F/M

Analiza odnosa krvnog tlaka i dobi, prikazana na Slici 11 prikazuje pozitivan linearni trend za oba spola, no treba uzeti u obzir veliko raspršenje podatkovnih točkica. Regresijske linije za oba spola imaju pozitivan nagib, ali regresijska linija za muškarce nalazi se iznad regresijske linije za žene što upućuje na to na muškarci u prosjeku imaju viši krvni tlak od žena te da prosječni krvni tlak raste ovisno o dobi kod oba spola.

U sljedećem primjeru također je korišten dijagram raspršivanja uz bojanje prema spolu ispitanika, ali ovdje je prikazana razina kolesterola u odnosu na dob.

```
# Scatter plot: dob vs. kolesterol (s bojama po spolu)
p3 <- ggplot(heart_data_rizik_subset, aes(x = age, y = chol, color = sex)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_manual(values = c("F" = "red", "M" = "blue")) +
  labs(title = "Dob vs. Kolesterol",
       x = "Dob",
       y = "Kolesterol",
       color = "Spol") +
  theme_minimal()
```



Slika 12. Dijagram raspršivanja: kolesterol vs. dob F/M

Analiza dijagrama sa Slike 12 prikazuje jasan pozitivan trend za žensku populaciju, ali ne i za mušku. Linija za muškarce gotovo je vodoravna što znači da razina kolesterola ne raste

značajno s dobi. S druge strane, linija za žene vrlo je slična liniji za žene u Slici 11 što bi značilo da kod žena rastu razine i kolesterola i krvnoga tlaka u odnosu na dob

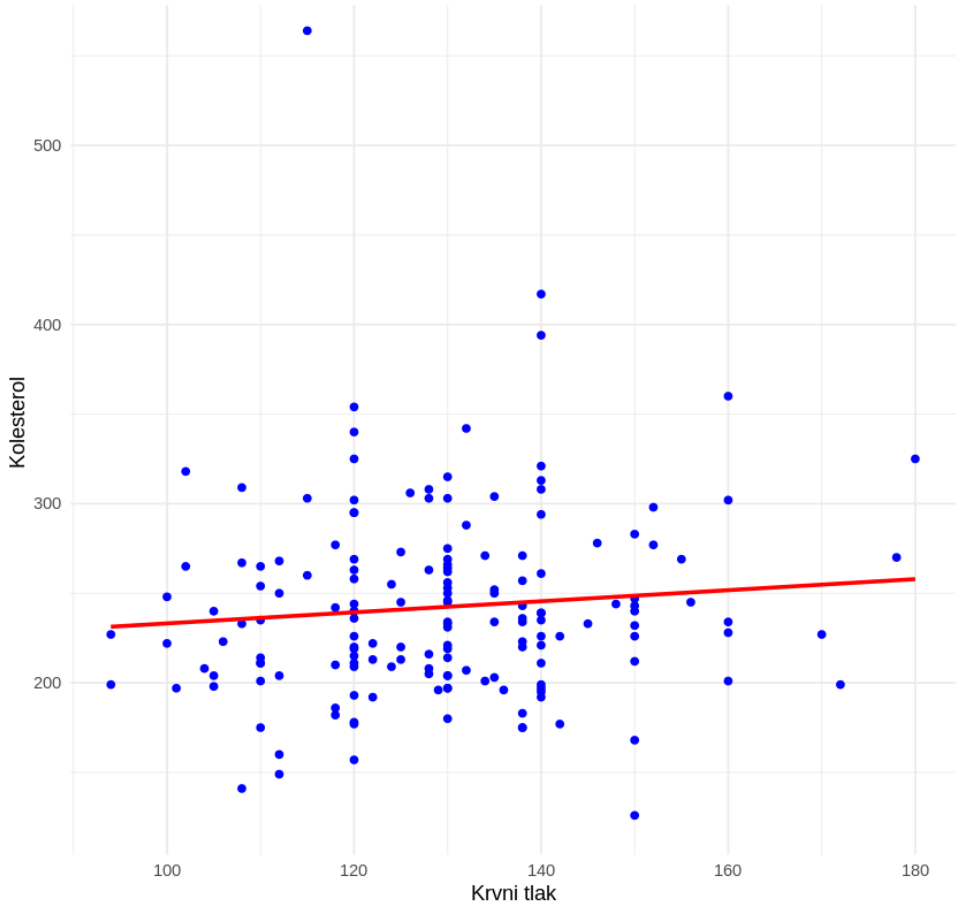
Vizualizacije na Slici 11 i Slici 12 dodatno potvrđuju važnost dobi kao faktora rizika za kardiovaskularne bolesti, ali i ističu ulogu spola u tim odnosima. Iako krvni tlak i kolesterol općenito rastu s godinama za oba spola, uočene razlike u regresijskim linijama ukazuju na specifične obrasce.

Ova vizualna analiza s podjelom po spolu važna je jer pruža precizniji uvid u varijabilnost podataka i potvrđuje da je spol važna varijabla koju treba uključiti u složenije statističke modele, poput višestruke linearne regresije, kako bi se dobila potpunija slika o faktorima rizika.

9.4. Dijagram raspršivanja: krvni tlak vs. kolesterol

U ovom dijelu analizira se direktan odnos između krvnog tlaka i razine kolesterola ispitanika pomoću dijagrama raspršivanja i regresijske linije.

```
# Scatter plot: krvni tlak vs. Kolesterol tlak
p1 <- ggplot(heart_data_rizik_subset, aes(x = trtbps, y = chol)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Krvni tlak vs. Kolesterol",
       x = "Krvni tlak",
       y = "Kolesterol") +
  theme_minimal()
```



Slika 13. Dijagram raspršivanja: krvni tlak vs. kolesterol

Na dijagramu se primjećuje da su točke vrlo raštrkane po cijelom grafikonu, što ukazuje na to da nema jasne, snažne povezanosti između krvnog tlaka i kolesterola. Iako je crvena regresijska linija blago nagnuta prema gore, njezin nagib nije jasan niti strm. Linija je gotovo vodoravna, što znači da povećanje krvnog tlaka ne dovodi nužno do značajnog povećanja razine kolesterola.

Ovaj zaključak je važan zato što pokazuje da se ove dvije varijable ponašaju neovisno jedna o drugoj unatoč sličnostima uočenim da prethodnim dijagramima

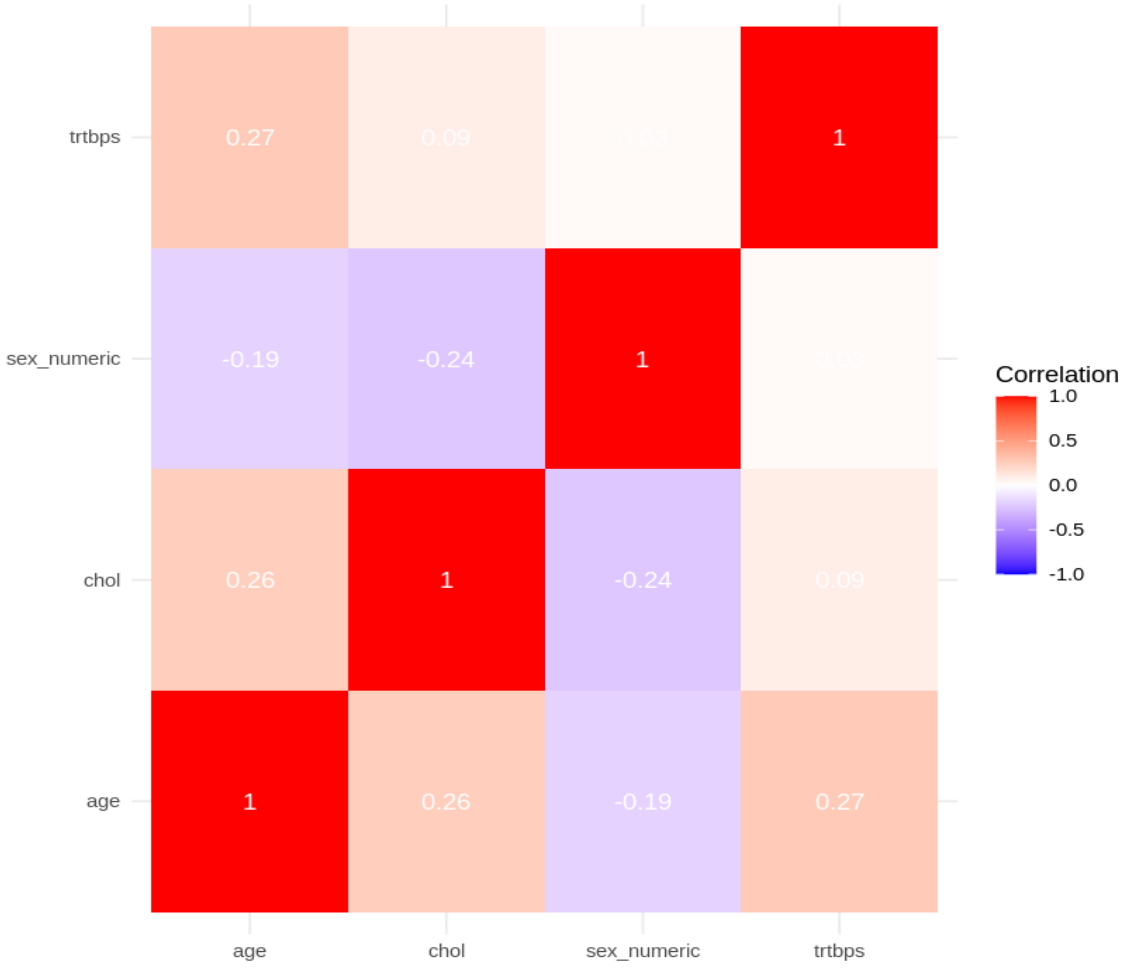
9.5. Heatmap za prikaz korelacija

Nakon vizualnog ispitivanja pojedinačnih odnosa između odabranih varijabli, heatmap korelacija na Slici 14 pruža sveobuhvatan pregled svih međusobnih odnosa unutar skupa podataka. Ova je metoda izabrana zbog mogućnosti prikazivanja složenije matrice korelacija.

Intenzitet i nijansa boje u svakoj ćeliji ukazuju na jačinu povezanosti. Vrijednosti blizu +1 (crvena boja) predstavljaju snažnu pozitivnu korelaciju, dok vrijednosti blizu -1 (plava boja) predstavljaju snažnu negativnu korelaciju. Vrijednosti blizu 0 (bijela boja) ukazuju na odsutnost korelacije između ta dva faktora.

```r
# Heatmap za prikaz korelacija
cor_df <- as.data.frame(cor_matrix)
cor_df$Variable <- rownames(cor_df)
cor_melt <- gather(cor_df, key = "Variable2", value = "Correlation", -Variable)

p4 <- ggplot(cor_melt, aes(x = Variable, y = Variable2, fill = Correlation)) +
  geom_tile() +
  geom_text(aes(label = round(Correlation, 2)), color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1)) +
  labs(title = "Heatmap Korelacija",
       x = "",
       y = "") +
  theme_minimal()
```



Slika 14. Heatmap korelacija

Analizom heatmapa možemo potvrditi zaključke iz prethodnih analiza:

- Dob i krvni tlak (age i trtbps): Postoji umjerena pozitivna korelacija (vrijednost korelacije je 0.27), što potvrđuje da se krvni tlak povećava s dobi
- Dob i kolesterol (age i chol): Postoji umjerena pozitivna korelacija (vrijednost korelacije je 0.26). To potvrđuje porast razine kolesterola s dobi
- Kolesterol i krvni tlak (chol i trtbps): Korelacija je vrlo slaba i gotovo nepostojeća (vrijednost korelacije je 0.03). To potvrđuje zaključak iz prethodnog dijagrama raspršivanja da krvni tlak i kolesterol nisu snažno povezani što ističe važnost uključivanja drugih faktora u istraživanje.

Sveukupno, heatmapa služi kao snažan dokaz da je za potpunu analizu odnosa među varijablama korisno koristiti višestruku linearnu regresiju, jer jednostavna korelacija ne može u potpunosti objasniti složene interakcije koje utječu na rizik od srčanih bolesti.

10. Višestruka linearna regresija

Dok se jednostavna linearna regresija fokusira na odnos između jedne nezavisne i jedne zavisne varijable, višestruka linearna regresija proširuje taj koncept tako šta uključuje nekoliko nezavisnih varijabli u model istovremeno. To je posebno primjenjivo kada se predmet istraživanja ne može objasniti koristeći samo jedan faktor već je za to potrebna kombinacija više različitih faktora.

Formula (4) koju koristimo za izradu višestruke linearne regresije glasi:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (4)$$

gdje je:

Y zavisna varijabla

$X_1, X_2, \ldots, X_n$ nezavisne varijable

β_0 presjek

$\beta_1, \beta_2, \ldots, \beta_n$ koeficijent regresije

ε komponenta slučajne pogreške

U kontekstu skupa podataka o srčanim bolestima, varijable poput dobi, spola, krvnog tlaka i kolesterola potencijalni su prediktori koji mogu utjecati na vjerojatnost srčane bolesti. Njihovim zajedničkim uključivanjem u jedan regresijski model može se dobiti realnija slika o tome kako ti faktori rizika međusobno djeluju.

U praksi to znači da umjesto da se prikazuje razina kolesterola samo kroz krvni tlak (kao u jednostavnoj regresiji), može se izgraditi model u kojem se kolesterol istovremeno objašnjava krvnim tlakom, dobi i spolom.

U ovom istraživanju, model višestruke regresije pomaže u razumijevanju kako nekoliko prediktora djeluje zajedno. Na primjer, moguće je da je sama dob slab prediktor kolesterola, ali kada se kombinira s krvnim tlakom, model postaje jači. Analiza koeficijenata također nam omogućuje da vidimo koja varijabla ima najjači utjecaj i je li njezin učinak pozitivan ili negativan.

Output modela uključuje vrijednosti poput R-kvadrata (koji pokazuje koliki dio varijance u zavisnoj varijabli objašnjavaju neovisne) i p-vrijednosti za svaki koeficijent (koje testiraju je li učinak svakog prediktora statistički značajan).

Općenito, višestruka linearna regresija pruža složeniji, ali i točniji prikaz podataka. Nadopunjuje rezultate jednostavne regresijske i korelacijske analize pokazujući zajednički učinak više faktora rizika na ishode povezane sa srčanim bolestima.

10.1. Višestruka linearna regresija - kolesterol

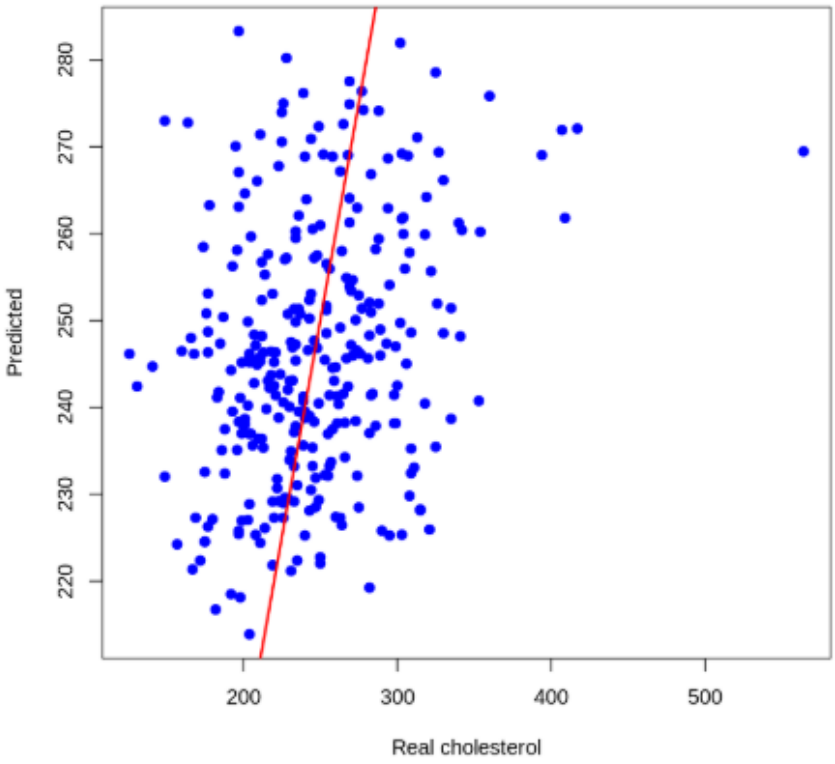
Kako bi se testiralo utječe li nekoliko neovisnih varijabli zajedno na razinu kolesterola, stvoren je model višestruke linearne regresije s dobi, krvnim tlakom i spolom kao prediktorima. Model nam omogućuje da vidimo kako te varijable zajedno objašnjavaju varijancu u vrijednostima kolesterola. Taj model prikazuje Slika 15.

U ovom grafu svaka plava točka predstavlja jednog pacijenta, gdje horizontalna x-os prikazuje stvarnu razinu kolesterola, a vertikalna y-os prikazuje vrijednost predviđenu regresijskim modelom. Crvena dijagonalna linija predstavlja idealan slučaj u kojem bi predviđene vrijednosti savršeno odgovarale stvarnim. Raspršenost točaka oko linije ukazuje na pogrešku modela. Vidljivo je da predviđanja prate opći trend podataka, iako je raspršenost široka, što sugerira da model objašnjava dio varijacije, ali ne cijelu.

```
#visestruka linearna regresija - kolesterol
chol_model <- lm(chol ~ age + trtbps + sex, data=heart_data)

# rezultat
summary(chol_model)

# graf - pred vs. real
pred_vals = predict(chol_model)
plot(heart_data$chol, pred_vals,
     xlab="Real cholesterol", ylab="Predicted", col="blue", pch=19)
abline(0,1,col="red", lwd=2)
```



Slika 15. Višestruka linearna regresija; kolesterol

```
Call:
lm(formula = chol ~ age + trtbps + sex, data = heart_data)

Residuals:
     Min       1Q   Median       3Q      Max
-123.980  -34.188   -3.829   27.617  294.536

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 179.6747    25.5094   7.043 1.29e-11 ***
age           1.0193     0.3307   3.082  0.00225 **
trtbps        0.1869     0.1707   1.095  0.27447
sex         -19.6572     6.1994  -3.171  0.00168 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 49.94 on 299 degrees of freedom
Multiple R-squared:  0.08093,   Adjusted R-squared:  0.07171
F-statistic: 8.777 on 3 and 299 DF,  p-value: 1.351e-05
```

Slika 16. Numerički output višestruke linearne regresije; kolesterol

U modelu višestruke regresije, prikazanom na Slici 16, presjek (engl. intercept) je iznosio 179.67. To je vrijednost kolesterola koju model predviđa kada su sve varijable (dob, krvni tlak i spol) jednake nuli. Iako u stvarnosti takva situacija nema puno smisla, presjek je važan jer služi kao početna točka u jednadžbi.

Varijabla dob imala je koeficijent 1.02 i bila je statistički značajna (p = 0.002). To znači da se za svaku dodatnu godinu života očekuje porast kolesterola za oko 1 mg/dl, uz pretpostavku da ostale varijable ostaju iste. Statistička značajnost ovdje znači da je vrlo mala vjerojatnost da je ovakav rezultat slučajan, pa možemo reći da dob stvarno ima utjecaj na kolesterol.

Za spol je koeficijent bio -19.65 i također statistički značajan (p = 0.0017). Negativan znak pokazuje da muškarci u ovom skupu podataka u prosjeku imaju niži kolesterol od žena. Razlika iznosi otprilike 20 mg/dl, kad se kontrolira za dob i krvni tlak.

Krvni tlak imao je koeficijent 0.19, ali rezultat nije bio značajan (p = 0.27). To znači da nema dovoljno dokaza da krvni tlak utječe na kolesterol kada već znamo dob i spol pacijenta.

Standardna pogreška modela iznosila je 49.94, a R^2 vrijednost bila je 0.08. To znači da model objašnjava samo oko 8% ukupnih razlika u kolesterolu među pacijentima, što je relativno malo i pokazuje da na kolesterol utječu i brojni drugi faktori koji nisu uključeni u ovaj model.

Višestruka regresijska analiza pokazuje da na razinu kolesterola jače utječu dob i spol nego krvni tlak u mirovanju. Iako model ne obuhvaća svu varijabilnost (što se vidi po relativno niskom R-kvadratu), ipak pokazuje da kombiniranje prediktora daje jasniju sliku od jednostavne regresije. Ovi rezultati ističu važnost zajedničkog razmatranja nekoliko čimbenika rizika, budući da fokusiranje na samo jedan od njih može dovesti do pogrešnih zaključaka.

10.2. Višestruka linearna regresija – krvni tlak

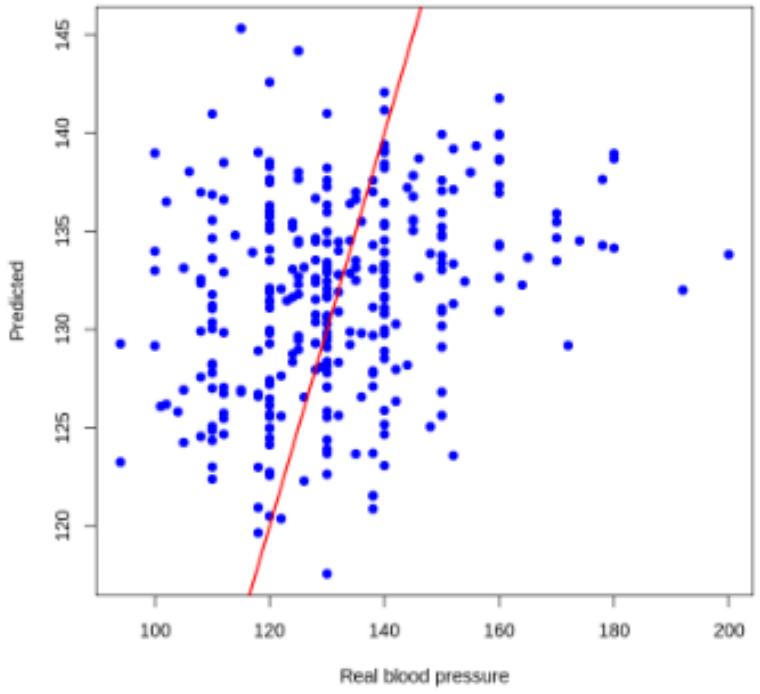
Nakon analize kolesterola, isti postupak primijenjen je na krvni tlak. Cilj je bio testirati mogu li varijable poput dobi, spola i razine kolesterola objasniti razlike u krvnom tlaku među pacijentima. Krvni tlak jedan je od ključnih pokazatelja kardiovaskularnog rizika, stoga je korisno vidjeti kako se mijenja ovisno o tim prediktorima.

Grafikon raspršenja na Slici 17 prikazuje predviđene u odnosu na stvarne vrijednosti krvnog tlaka. Svaka točka predstavlja jednog pacijenta, pri čemu horizontalna os prikazuje stvarni krvni tlak, a vertikalna os vrijednost predviđenu regresijskim modelom. Crvena linija predstavlja idealan scenarij u kojem bi predviđanja bila savršeno točna. Slično modelu kolesterola, točke su raspršene po liniji, što znači da model objašnjava dio varijacije, ali i dalje postoji mnogo šuma.

```
#visestruka linearna regresija - krvni tlak
bp_model <- lm(trtbps ~ age + sex + chol, data=heart_data)

# rezultat
summary(bp_model)

# graf - pred vs. real
bp_pred <- predict(bp_model)
plot(heart_data$trtbps, bp_pred,
     xlab="Real blood pressure", ylab="Predicted", col="blue", pch=19)
abline(0,1,col="red", lwd=2)
```



Slika 17. Višestruka linearna regresija; krvni tlak

```
Call:
lm(formula = trtbps ~ age + sex + chol, data = heart_data)

Residuals:
    Min      1Q  Median      3Q     Max
-38.975 -11.431  -1.296   9.717  66.182

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 99.11042    7.33898  13.505  < 2e-16 ***
age          0.50992    0.10968   4.649 5.01e-06 ***
sex         -0.68789    2.13040  -0.323    0.747
chol         0.02136    0.01951   1.095    0.274
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.88 on 299 degrees of freedom
Multiple R-squared:  0.08258,   Adjusted R-squared:  0.07337
F-statistic: 8.971 on 3 and 299 DF,  p-value: 1.043e-05
```

Slika 18. Numerički output višestruke linearne regresije; krvni tlak

Slika 18 prikazuje numerički output višestruke linearne regresije koja pokušava predvidjeti kretanje krvnog tlaka koristeći druge poznate faktore. U ovom modelu presjek iznosio je 99.11. To znači da bi očekivani krvni tlak pacijenta bio oko 99 mmHg kada su sve ostale varijable (dob, spol i kolesterol) jednake nuli. Kao i kod većine regresijskih modela, presjek sam po sebi nema praktičnu medicinsku interpretaciju, ali je nužan za pravilno formiranje jednadžbe modela.

Varijabla dob imala je koeficijent 0.59 i bila je statistički značajna ($p < 0.001$). To znači da se s porastom dobi pacijenta krvni tlak u prosjeku povećava za otprilike 0.6 mmHg po godini, uz pretpostavku da su ostale varijable konstantne. Budući da je p-vrijednost vrlo mala, možemo biti sigurni da je ovaj učinak stvaran, a ne posljedica slučajnosti.

Za spol je procijenjen koeficijent -0.69, ali rezultat nije bio značajan ($p = 0.747$). To upućuje na to da spol u ovom datasetu ne igra veliku ulogu u objašnjavanju krvnog tlaka – odnosno, ne možemo tvrditi da se krvni tlak muškaraca i žena statistički razlikuje kada se kontrolira dob i kolesterol.

Varijabla kolesterol imala je pozitivan koeficijent 0.02, no i taj rezultat nije bio značajan ($p = 0.274$). To znači da nema dovoljno dokaza da kolesterol utječe na krvni tlak kada su već uračunate ostale varijable.

Standardna pogreška modela bila je 16.88, a R^2 vrijednost 0.08. To pokazuje da model objašnjava otprilike 8% ukupnih razlika u krvnom tlaku. Dakle, iako dob ima jasan i značajan utjecaj, postoje mnogi drugi faktori koji nisu uključeni u ovaj model, a koji u stvarnosti imaju velik utjecaj na krvni tlak (npr. stres, prehrana, tjelesna težina ili genetika).

11. Zaključak

U ovom radu analizirani su podaci vezani uz faktore rizika srčanih bolesti, s naglaskom na dob, spol, prosječni krvni tlak i razinu kolesterola.

Ti su podaci analizirani koristeći programski jezik R i pakete *dplyr* i *ggplot2* iz okruženja *tidyverse*. Paket *dplyr* korišten je radni njegovih intuitivnih funkcija za filtriranje, odabir, grupiranje i transformaciju podataka dok je paket *ggplot2* omogućio prikaz svih potrebnih dijagrama i vizualizacija.

Od statističkim metoda korištene su deskriptivna statistika, korelacijska analiza, vizualizacija podataka te jednostavni i višestruki linearni regresijski model. Deskriptivna statistika koristila je za izračun osnovnih mjera kao što su srednja vrijednost, najmanja i najveća vrijednost te medijan za dob, krvni tlak i kolesterol. Otkriveno je da je prosječna dob osobe pod najvećim rizikom oboljenja od srčane bolesti 52,5 i da najviše ispitanika ima krvni tlak kategorije „visoko“ i kolesterol kategorije „opasno“. Korelacijskom analizom i linearnom regresijom ispitani su odnosi između varijabli te je potvrđeno da dob ima najveći utjecaj na razinu kolesterola i krvnog tlaka. Istom analizom s dodatnim filterom za spol zaključeno je da dob ima sličan utjecaj na porast krvnoga tlaka kod oba spola te da muškarci u prosjeku imaju nešto viši krvni tlak od žena, ali zato žene imaju viši kolesterol od muškaraca i kod njih je vidljiv porast razine kolesterola u odnosu na dob dok je kod muškaraca razina kolesterola ostala ista povećanjem dobi. Sve navedene analize prikazane su odgovarajućim dijagramima – distribucija kategorija prikazana je stupčastim dijagramima, a korelacije su prikazane dijagramima raspršivanja i regresijskim linijama.

Istražena je i povezanost razine kolesterola s krvnim tlakom. Iako su korelacije između varijabli bile relativno slabe, ipak se mogu uočiti određeni obrasci koji upućuju na važnost redovitog praćenja ovih pokazatelja.

Može se zaključiti da pravovremeno prepoznavanje povišenih vrijednosti krvnog tlaka i kolesterola ima veliku ulogu u prevenciji srčanog udara. Iako je korišten relativno mali skup podataka, provedena analiza pokazala je kako se uz pomoć jednostavnih statističkih metoda i vizualizacije mogu dobiti korisni uvidi.

Literatura

[1] „Heart Attack Analysis & Prediction Dataset“, kaggle.com

https://www.kaggle.com/datasets/sonialikhan/heart-attack-analysis-and-prediction-dataset

[2] „Use Tidyverse“, learn.microsoft.com

https://learn.microsoft.com/en-us/fabric/data-science/r-use-tidyverse

(pristupljeno 22.08.2025.)

[3] „Introduction to dplyr“, cran.r-project.org

https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html

(pristupljeno 22.08.2025.)

[4] „*ggplot2*“, *ggplot2*.tidyverse.org

https://*ggplot2*.tidyverse.org/

(pristupljeno 23.08.2025.)

[5] Ken Stewart, „linear regression“, britannica.com

https://www.britannica.com/topic/linear-regression

(pristupljeno 25.08.2025.)

[6] Evelyn Clarke, „R Stepwise & Multiple Linear Regression“, guru99.com

https://www.guru99.com/r-simple-multiple-linear-regression.html

(pristupljeno 06.09.2025.)

[7] „Višestruka linearna regresija“, old.matf.bg.ac.rs/

http://old.matf.bg.ac.rs/p/files/1432029947-63-Visestruka_linearna_regresija.pdf

(pristupljeno 09.09.2025.)

Popis tablica

Popis slika

Popis priloga

Prilog 1: Cjelovit programski Kod

```r
# Učitavanje potrebnih paketa
library(readr)
library(tidyverse)
library (dplyr)
library (ggplot2)

# Učitavanje podataka
heart_data <- read_csv("heart.csv")

# Pregled prvih nekoliko redaka
head(heart_data)
str(heart_data)

# Generiranje godina rođenja na temelju dobi
godina_rodenja <- 2025 - heart_data$age

# Generiranje slučajnih datuma i vremena
set.seed(123)  # Postavljanje sjemena za reproduktivnost
datum_rodenja <- as.Date(paste0(godina_rodenja, "-01-01")) +
                 sample(0:365, nrow(heart_data), replace = TRUE)

datum_vrijeme_rodenja <- as.POSIXct(
  paste0(datum_rodenja, " ",
         sample(0:23, nrow(heart_data), replace = TRUE), ":",
         sample(0:59, nrow(heart_data), replace = TRUE), ":",
         sample(0:59, nrow(heart_data), replace = TRUE))
)

# Dodavanje generiranih podataka u dataset
heart_data$datum_rodenja <- datum_rodenja
heart_data$datum_vrijeme_rodenja <- datum_vrijeme_rodenja

# Pogledaj nove kolone (15 i 16)
head(heart_data[, c(1, 15, 16)])



# Dodavanje godine rođenja na temelju dobi
heart_data <- heart_data %>%
  mutate(godina_rodenja = 2025 - age)

# Pregled prvih par redaka s novim stupcem
head(heart_data[, c("age", "godina_rodenja")])

# Dodavanje kategotrija za stupac kolesterola
heart_data <- heart_data %>%
  mutate(chol_category = case_when(
    chol < 200 ~ "normalno",
    chol >= 200 & chol <= 239 ~ "visoko",
    chol >= 240 ~ "opasno",
  ))
# Dodavanje kategorija za stupac trtbps
heart_data <- heart_data %>%
  mutate(trtbps_category = case_when(
    trtbps < 120 ~ "normalno",
    trtbps >= 120 & trtbps < 140 ~ "visoko",
    trtbps >= 140 ~ "opasno"
  ))
# Uklanjanje redaka s nedostajućim vrijednostima
heart_dropna <- drop_na(heart_data)
# Provjera promjena
head(heart_data)
```

```r
# Provjera promjena
head(heart_data[,c(1, 17)])

# Razdvajanje datuma i vremena rođenja
heart_data <- heart_data %>%
  separate(col = datum_vrijeme_rodenja, into = c("datum_rodenja", "vrijeme_rodenja"), sep =
" ")

# Provjera promjena
head(heart_data[,c(1, 17, 18)])

heart_data <- heart_data %>%
  unite(col = "datum_vrijeme_rodenja", c("datum_rodenja", "vrijeme_rodenja"), sep = " ")

# Provjera promjena
head(heart_data[,c(1, 17)])



# Filtriranje opservacija
heart_data_rizik_subset <- heart_data %>%
  filter(output == "X")
head(heart_data_rizik_subset)

# Odabir podskupa stupaca
heart_data_rizik_subset <- heart_data_rizik_subset %>%
  select(age, sex, trtbps, chol, output)
head(heart_data_rizik_subset)

# Stvaranje novih stupaca
heart_data_rizik_subset <- heart_data_rizik_subset %>%
  mutate(
    trtbps_risk_category = case_when(
      trtbps < 120 ~ "normalno",
      trtbps >= 120 & trtbps < 140 ~ "visoko",
      trtbps >= 140 ~ "opasno"
    ),
    chol_risk_category = case_when(
      chol < 200 ~ "normalno",
      chol >= 200 & chol < 240 ~ "visoko",
      chol >= 240 ~ "opasno"
    ),
    sex_numeric = ifelse(sex == "M", 1, 0)
  )

head(heart_data_rizik_subset)


cor_matrix <- cor(heart_data_rizik_subset %>% select(age, trtbps, chol, sex_numeric))
cat("Korelacijska matrica:\n")
print(round(cor_matrix, 2))   # zaokruženo na 2 decimale



# Prosječna dob pacijenata u riziku
rizik_udara <- heart_data %>%
  filter(output == "1") %>%
  summarise(prosjek_dobi = mean(age))
print("Prosječna dob pacijenata u najvecem riziku:")
print(rizik_udara)


# Izračun srednje vrijednosti i medijana
mean_trtbps <- mean(heart_data$trtbps)
median_trtbps <- median(heart_data$trtbps)
min_trtbps <- min(heart_data$trtbps)
```

```r
max_trtbps <- max(heart_data$trtbps)



# Postavljanje redoslijeda kategorija za krvni tlak i kolesterol
heart_data$trtbps_category <- factor(heart_data$trtbps_category, levels = c("normalno",
"visoko", "opasno"))
heart_data$chol_category <- factor(heart_data$chol_category, levels = c("normalno",
"visoko", "opasno"))

# Distribucija kategorija krvnog tlaka
trtbps_category_distribution <- heart_data %>%
  count(trtbps_category)

# Distribucija kategorija kolesterola
chol_category_distribution <- heart_data %>%
  count(chol_category)

# Ispis rezultata
cat("Srednja vrijednost krvnog tlaka:", mean_trtbps, "\n")
cat("Medijan krvnog tlaka:", median_trtbps, "\n")
cat("Najmanja vrijednost krvnog tlaka:", min_trtbps, "\n")
cat("Najveća vrijednost krvnog tlaka:", max_trtbps, "\n\n")

cat("Distribucija kategorija krvnog tlaka:\n")
print(trtbps_category_distribution)
cat("\nDistribucija kategorija kolesterola:\n")
print(chol_category_distribution)


# Grafovi
# Grafikon prosječnog krvnog tlaka prema dobi
ggplot(heart_data, aes(x = age, y = trtbps)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(title = "Prosječni krvni tlak prema dobi",
       x = "Dob",
       y = "Krvni tlak") +
  theme_minimal()


# Grafikon distribucije kategorija krvnog tlaka
ggplot(trtbps_category_distribution, aes(x = trtbps_category, y = n)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  labs(title = "Distribucija kategorija krvnog tlaka",
       x = "Kategorija krvnog tlaka",
       y = "Broj slučajeva")

# Grafikon distribucije kategorija kolesterola
ggplot(chol_category_distribution, aes(x = chol_category, y = n)) +
  geom_bar(stat = "identity", fill = "lightgreen") +
  labs(title = "Distribucija kategorija kolesterola",
       x = "Kategorija kolesterola",
       y = "Broj slučajeva")



# Filtriranje opservacija
heart_data_rizik <- heart_data %>%
  filter(output == 1)

# Odabir podskupa stupaca
heart_data_rizik_subset <- heart_data_rizik %>%
  select(age, sex, trtbps, chol, output)

# Stvaranje novih stupaca
```

```r
heart_data_rizik_subset <- heart_data_rizik_subset %>%
  mutate(
    trtbps_risk_category = case_when(
      trtbps < 120 ~ "normalno",
      trtbps >= 120 & trtbps < 140 ~ "visoko",
      trtbps >= 140 ~ "opasno"
    ),
    chol_risk_category = case_when(
      chol < 200 ~ "normalno",
      chol >= 200 & chol < 240 ~ "visoko",
      chol >= 240 ~ "opasno"
    )
  )

# Izračun korelacija
cor_matrix <- cor(
  heart_data_rizik_subset %>%
    select(age, trtbps, chol, sex)
)

# Scatter plot: krvni tlak vs. kolesterol
p1 <- ggplot(heart_data_rizik_subset, aes(x = trtbps, y = chol)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(
    title = "Krvni tlak vs. Kolesterol",
    x = "Krvni tlak",
    y = "Kolesterol"
  ) +
  theme_minimal()

# Scatter plot: dob vs. krvni tlak (s bojanjem po spolu, 0 = žena, 1 = muškarac)
p2 <- ggplot(heart_data_rizik_subset, aes(x = age, y = trtbps, color = factor(sex))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_manual(
    values = c("0" = "red", "1" = "blue"),
    labels = c("0" = "Žena", "1" = "Muškarac")
  ) +
  labs(
    title = "Dob vs. Krvni tlak",
    x = "Dob",
    y = "Krvni tlak",
    color = "Spol"
  ) +
  theme_minimal()

# Scatter plot: dob vs. kolesterol (s bojanjem po spolu)
p3 <- ggplot(heart_data_rizik_subset, aes(x = age, y = chol, color = factor(sex))) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  scale_color_manual(
    values = c("0" = "red", "1" = "blue"),
    labels = c("0" = "Žena", "1" = "Muškarac")
  ) +
  labs(
    title = "Dob vs. Kolesterol",
    x = "Dob",
    y = "Kolesterol",
    color = "Spol"
  ) +
  theme_minimal()

# Heatmap za prikaz korelacija
cor_df <- as.data.frame(cor_matrix)
cor_df$Variable <- rownames(cor_df)
```

```r
cor_melt <- gather(cor_df, key = "Variable2", value = "Correlation", -Variable)

p4 <- ggplot(cor_melt, aes(x = Variable, y = Variable2, fill = Correlation)) +
  geom_tile() +
  geom_text(aes(label = round(Correlation, 2)), color = "white") +
  scale_fill_gradient2(
    low = "blue", high = "red", mid = "white", midpoint = 0,
    limit = c(-1, 1)
  ) +
  labs(
    title = "Heatmap Korelacija",
    x = "",
    y = ""
  ) +
  theme_minimal()

# Prikaz grafova
print(p1)
print(p2)
print(p3)
print(p4)
```

```r
#visestruka linearna regresija - kolesterol
chol_model <- lm(chol ~ age + trtbps + sex, data=heart_data)

# rezultat
summary(chol_model)

# graf - pred vs. real
pred_vals = predict(chol_model)
plot(heart_data$chol, pred_vals,
     xlab="Real cholesterol", ylab="Predicted", col="blue", pch=19)
abline(0,1,col="red", lwd=2)


#visestruka linearna regresija - krvni tlak
bp_model <- lm(trtbps ~ age + sex + chol, data=heart_data)

# rezultat
summary(bp_model)

# graf - pred vs. real
bp_pred <- predict(bp_model)
plot(heart_data$trtbps, bp_pred,
     xlab="Real blood pressure", ylab="Predicted", col="blue", pch=19)
abline(0,1,col="red", lwd=2)
```