

SAP Projekt

Statistički Analizirani

2025-12-10

Motivacija i opis problema

Moždani udar je hitno medicinsko stanje koje nastaje nakon poremećaja cirkulacije u mozgu. U Hrvatskoj je drugi najveći uzrok smrti te prvi najveći uzrok invaliditeta. Rizik od moždanog udara kod pojedine osobe može ovisiti o više faktora pa se u medicini koriste razni modeli za njegovu procjenu. Nije strana ni uporaba alata zasnovanih na umjernoj inteligenciji. Predikcija moždanog udara bitna je za pravovremenu identifikaciju rizika koje omogućava pravovremene mjere prevencije i osvještavanje populacije.

Opis skupa podataka

Prikupljeni skup podataka sadrži kliničke podatke o pacijentima i informacije o moždanom udaru. Za svakog pacijenta navedena je vrijednost 12 značajki grupiranih u 4 kategorije: demografski podaci, zdravstveni podaci, fiziološki podaci i životne navike. Cilj je uočiti povezanost izmjerenih podataka i rizika od moždanog udara. Skup podataka sadržava 5,110 zapisa o pacijentima od kojih je njih 249 doživjelo moždani udar. Udio pacijenata koji su doživjeli moždani udar iznosi 4.87%

- id: jedinstveni identifikator pacijenta
- gender: spol pacijenta (Male, Female)
- age: dob pacijenta
- hypertension: oznaka koja daje informaciju o tome ima li pacijent visoki tlak (0, 1)
- heart_disease: oznaka koja daje informaciju ima li pacijent neku srčanu bolest (0, 1)
- ever_married: odgovara na pitanje je li pacijent ikad bio u braku (No, Yes) work_type: tip zaposlenja (children, Govt_job, Never_worked, Private, Self-employed)
- Residence_type: tip prebivališta u kojem živi pacijent (Rural, Urban)
- avg_glucose_lvl: prosječna razina glukoze u krvi (mg/dL)
- bmi: indeks tjelesne mase koji predstavlja odnos visine i težine pacijenta
- smoking_status: opis pacijentovog odnosa s pušenjem cigareta (formerly smoked, never smoked, smokes, Unknown)
- stroke: oznaka koja daje informaciju je li pacijent doživio moždani udar (0, 1)

Učitavanje i pregled podataka

```
data <- read.csv("data.csv")
head(data)
```

```
##      id gender age hypertension heart_disease ever_married work_type
## 1  9046  Male  67             0              1          Yes   Private
## 2 51676 Female  61             0              0          Yes Self-employed
## 3 31112  Male  80             0              1          Yes   Private
```

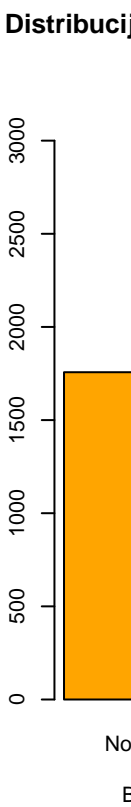
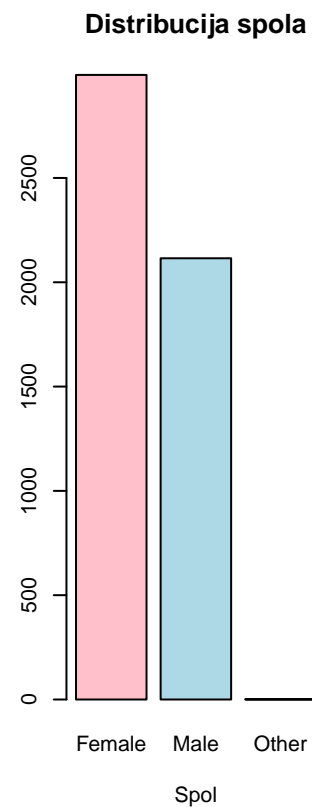
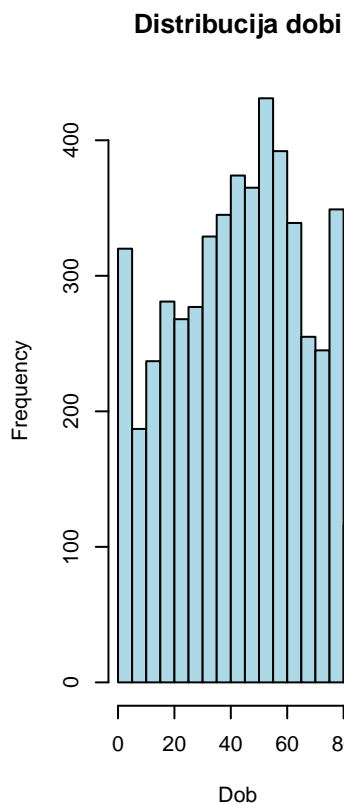
```
## 4 60182 Female 49          0          0      Yes      Private
## 5 1665 Female 79          1          0      Yes Self-employed
## 6 56669 Male 81          0          0      Yes      Private
## Residence_type avg_glucose_level bmi smoking_status stroke
## 1 Urban 228.69 36.6 formerly smoked 1
## 2 Rural 202.21 N/A never smoked 1
## 3 Rural 105.92 32.5 never smoked 1
## 4 Urban 171.23 34.4 smokes 1
## 5 Rural 174.12 24 never smoked 1
## 6 Urban 186.21 29 formerly smoked 1
```

```
str(data)
```

```
## 'data.frame': 5110 obs. of 12 variables:
## $ id : int 9046 51676 31112 60182 1665 56669 53882 10434 27419 60491 ...
## $ gender : chr "Male" "Female" "Male" "Female" ...
## $ age : num 67 61 80 49 79 81 74 69 59 78 ...
## $ hypertension : int 0 0 0 0 1 0 1 0 0 0 ...
## $ heart_disease : int 1 0 1 0 0 0 1 0 0 0 ...
## $ ever_married : chr "Yes" "Yes" "Yes" "Yes" ...
## $ work_type : chr "Private" "Self-employed" "Private" "Private" ...
## $ Residence_type : chr "Urban" "Rural" "Rural" "Urban" ...
## $ avg_glucose_level: num 229 202 106 171 174 ...
## $ bmi : chr "36.6" "N/A" "32.5" "34.4" ...
## $ smoking_status : chr "formerly smoked" "never smoked" "never smoked" "smokes" ...
## $ stroke : int 1 1 1 1 1 1 1 1 1 1 ...
```

Iz generiranih prikaza moguće je vidjeti neke od vrijednosti koje poprimaju pojedini atributi vezani uz pacijenta, ali i tip podataka koji prikazuju. Primjetljivo je kako se atribut bmi vodi kao niz znakova iako semantički predstavlja decimalni broj. Pri izradi dijagrama i u budućim računima potrebno je pripaziti da se bmi ne prikaže kao kategorični atribut.

```
par(mfrow = c(1,3))
hist(data$age , main = "Distribucija dobi", xlab = "Dob", col = "lightblue", border = "black")
barplot(table(data$gender), main = "Distribucija spola", xlab = "Spol", col = c("pink","lightblue","black"))
barplot(table(data$ever_married), main = "Distribucija bračnog statusa", xlab = "Bračni status", col = c("pink","lightblue","black"))
```



Prikaz demografskih podataka

```
summary(data$age)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      0.08  25.00   45.00   43.23  61.00   82.00
```

```
prop.table(table(data$gender, useNA = "ifany"))
```

```
##
##      Female      Male      Other
## 0.5859099804 0.4138943249 0.0001956947
```

```
prop.table(table(data$ever_married, useNA = "ifany"))
```

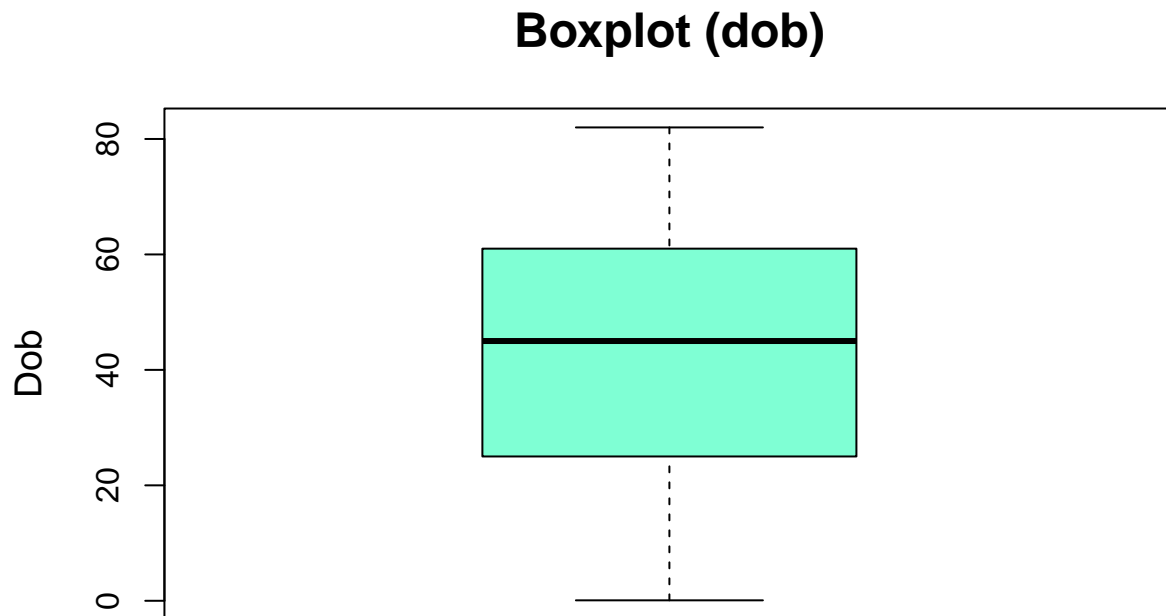
```
##
##      No      Yes
## 0.3438356 0.6561644
```

Distribucija dobi ispitanika blago je zakrivljena u lijevo. Prosječan broj godina pacijenta iznosi 43.23 godine, a njihov raspon je razlika između 82 godine i 0.08 godina. Dakle, zastupljene su gotovo sve dobne skupine te nema bitnih odstupanja.

Broj ženskih ispitanih pacijenata je neočekivano velik u odnosu na muške ispitanike. Pacijentice zauzimaju čak 58.6% skupa podataka.

Većina ispitanika je u svom životu bar jednom stupila u brak.

```
boxplot(data$age,
main = "Boxplot (dob)",
ylab = "Dob",
col = "aquamarine",
cex.main = 1.5, cex.lab = 1.2)
```



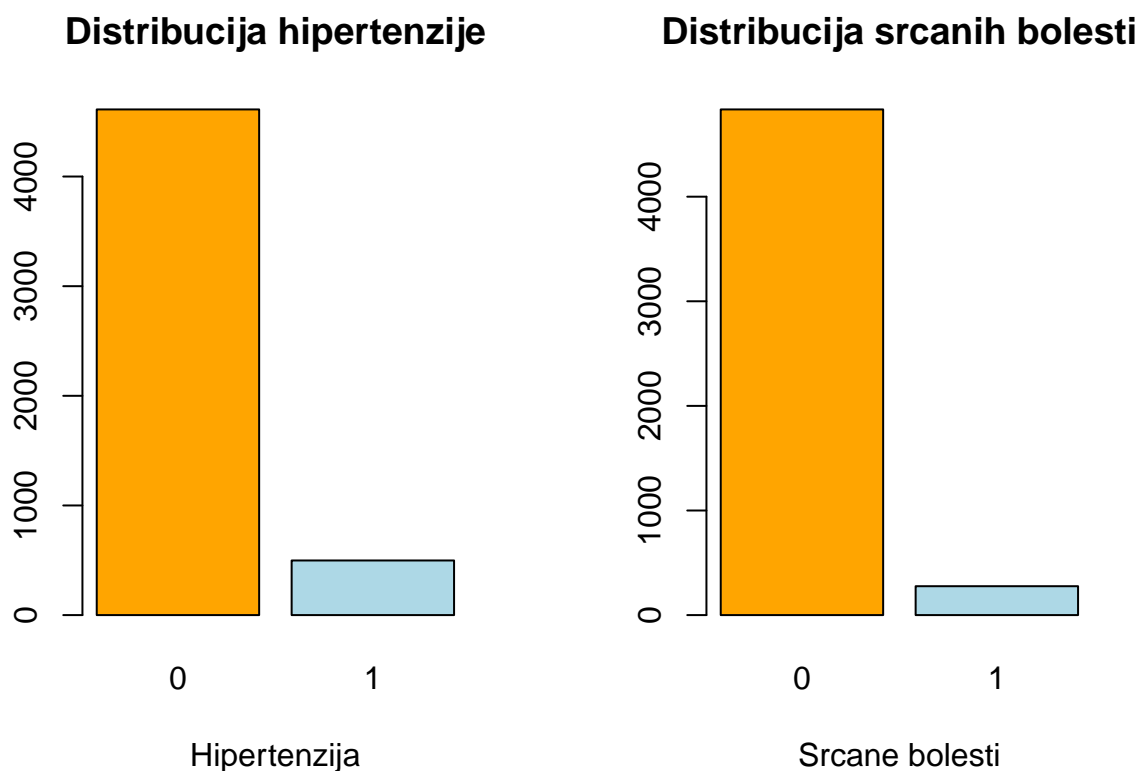
Prikaz zdravstvenih podataka Kao što je prethodno navedeno, potrebno je pripremiti atribut bmi.

```
bmi<-data$bmi
```

```
bmi[bmi == "N/A"] <- NA
bmi <- as.numeric(bmi)
```

```
par(mfrow = c(1,2))
```

```
barplot(table(data$hypertension), main = "Distribucija hipertenzije", xlab = "Hipertenzija", col = c("orange", "blue"))
barplot(table(data$heart_disease), main = "Distribucija srcanih bolesti", xlab = "Srcane bolesti", col = c("orange", "blue"))
```



```

young_with_hypertension <- data[data$age<=median(data$age),]$hypertension
old_with_hypertension <- data[data$age>=median(data$age),]$hypertension

list(
  overall_mean = mean(data$hypertension),
  young_mean   = mean(young_with_hypertension),
  old_mean     = mean(old_with_hypertension)
)

```

```

## $overall_mean
## [1] 0.09745597
##
## $young_mean
## [1] 0.02444614
##
## $old_mean
## [1] 0.1715173

```

U skupu podataka 9.7% ispitanika ima hipertenziju, međutim ona jako ovisi o dobi ispitanika. Ako uzmemo samo ispitanike starije od medijana skupa, ta brojka postane 17.1%

```

young_with_heart_disease <- data[data$age<=median(data$age),]$heart_disease
old_with_heart_disease <- data[data$age>=median(data$age),]$heart_disease

list(

```

```

overall_mean = mean(data$heart_disease),
young_mean   = mean(young_with_heart_disease),
old_mean     = mean(old_with_heart_disease)
)

```

```

## $overall_mean
## [1] 0.05401174
##
## $young_mean
## [1] 0.003055768
##
## $old_mean
## [1] 0.1043849

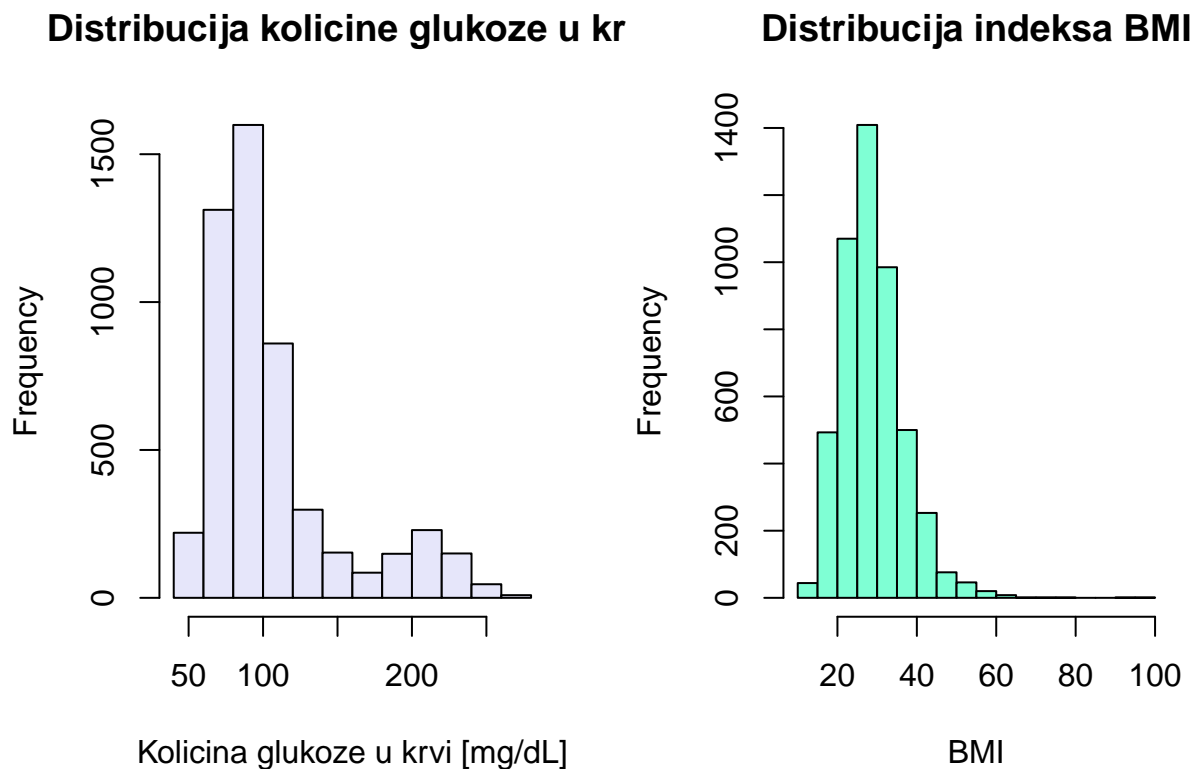
```

Sličan rezultat se dobije i sa srčanim bolestima, gdje u uzorku srčanu bolest ima oko 5.4% dok stariji dio uzorka ima 10.4%

```

par(mfrow = c(1,2))
hist(data$avg_glucose_level , main = "Distribucija kolicine glukoze u krvi", xlab = "Kolicina glukoze u krvi", col = "lightblue", border = "black")
hist(bmi , main = "Distribucija indeksa BMI", xlab = "BMI", col = "aquamarine", border = "black")

```



```
summary(data$avg_glucose_level)
```

```

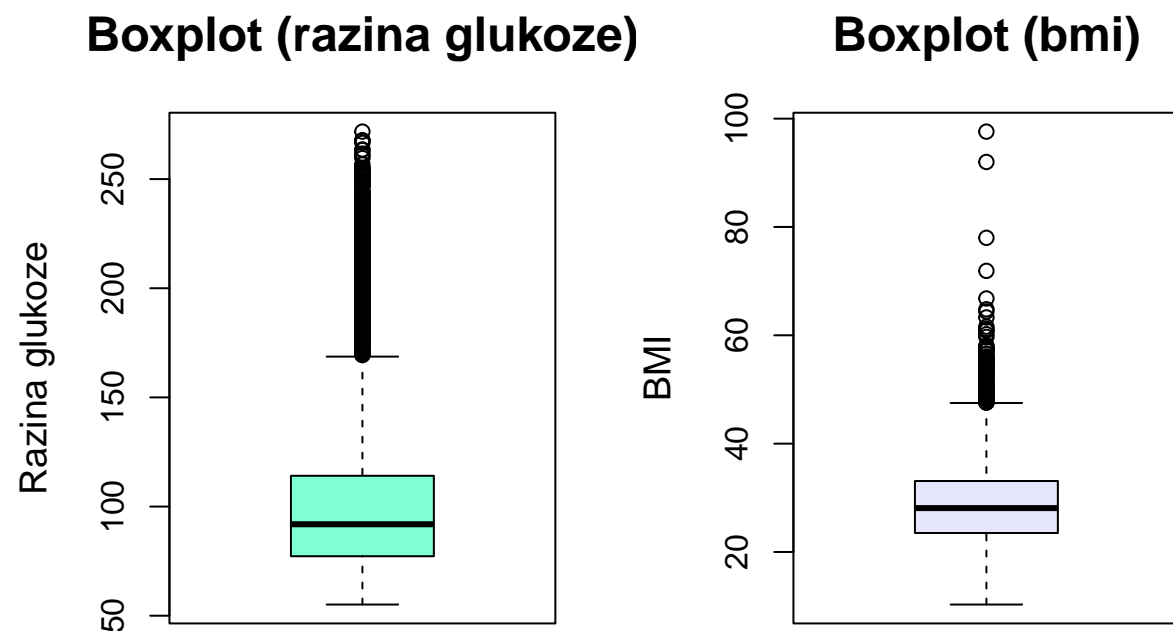
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  55.12   77.25   91.89  106.15  114.09  271.74

```

```
summary(bmi)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's  
##    10.30   23.50   28.10   28.89   33.10   97.60    201
```

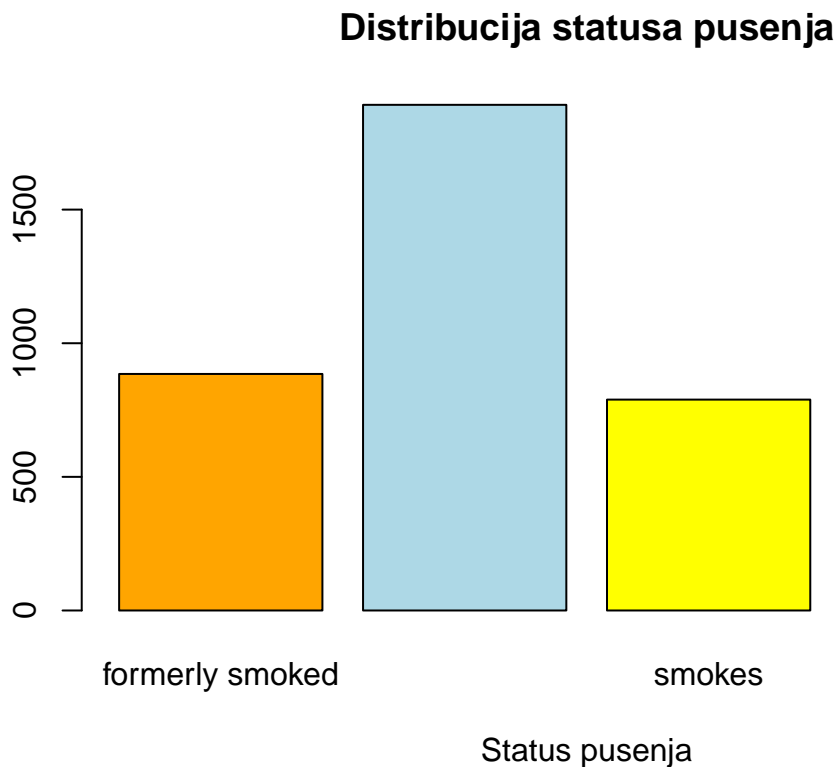
```
par(mfrow = c(1,2))  
boxplot(data$avg_glucose_level,  
main = "Boxplot (razina glukoze)",  
ylab = "Razina glukoze",  
col = "aquamarine",  
cex.main = 1.5, cex.lab = 1.2)  
boxplot(bmi,  
main = "Boxplot (bmi)",  
ylab = "BMI",  
col = "lavender",  
cex.main = 1.5, cex.lab = 1.2)
```



Distribucije atributa vezanih uz razinu glukoze u krvi te BMI imaju sličnost u tome da postoje stršće vrijednosti. Međutim, razina glukoze krvi je više zakrivljena i to se da očitati iz broja stršćih vrijednosti, ali i pomaka crte medijana u interkvartalnom intervalu.

Nijedna od ove dvije distribucije se ne može smatrati savršeno normalnom.

```
barplot(table(data$smoking_status), main = "Distribucija statusa pušenja", xlab = "Status pušenja", col
```



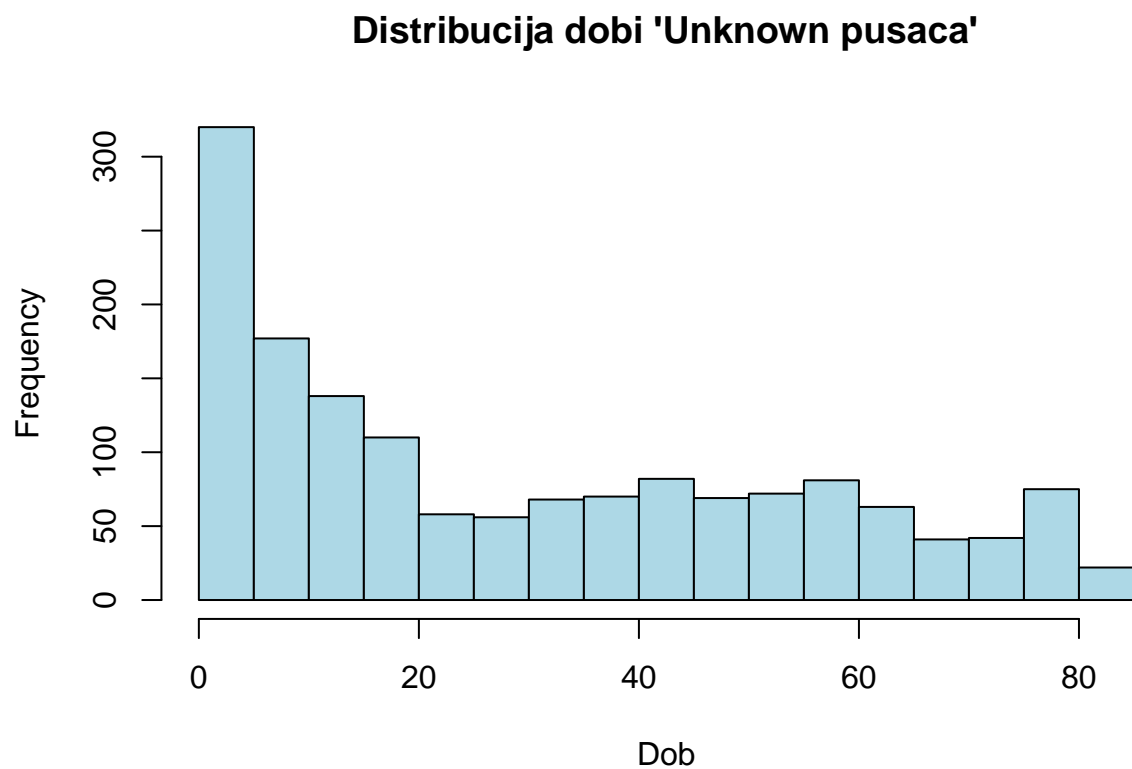
Prikaz podataka o životnim navikama

```
prop.table(table(data$smoking_status, useNA = "ifany"))
```

```
##
## formerly smoked    never smoked        smokes        Unknown
##      0.1731898      0.3702544      0.1544031      0.3021526
```

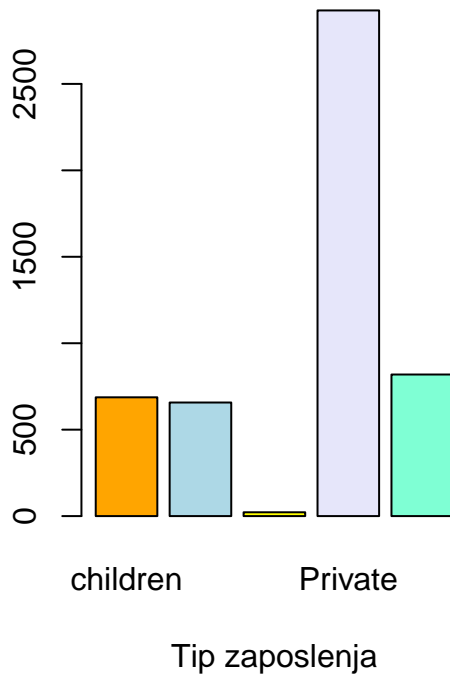
Vidljivo je da za veliki postotak ispitanika, čak preko 30%, nemamo informaciju o njihovom statusu pušenja. Važno je napomenuti da se izbacivanjem zapisa o pacijentima sa statusom pušenja “Unknown” zapravo izbacuju gotovo sva djeca iz skupa podataka. Ipak, u skupu pacijenata s “Unknown” statusom pušenja postoje predstavnici gotovo svih dobnih skupina. Sljedeći histogram prikazuje distribuciju dobi “Unknown” pušača.

```
hist(data[data$smoking_status == "Unknown",]$age , main = "Distribucija dobi 'Unknown pusaca'", xlab =
```

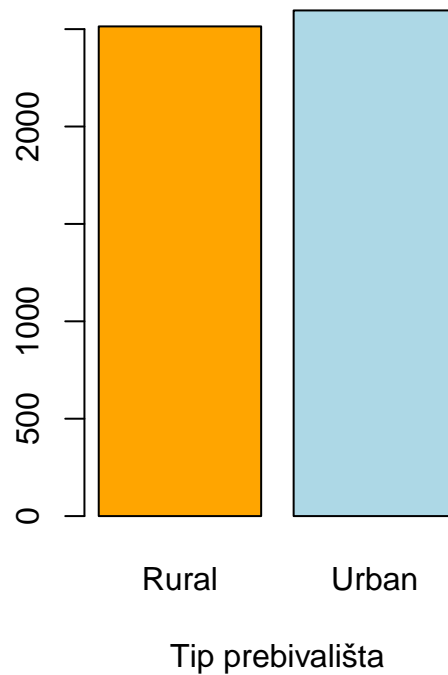


```
par(mfrow = c(1,2))
barplot(table(data$work_type), main = "Distribucija tipa zaposlenja", xlab = "Tip zaposlenja", col = c(
barplot(table(data$Residence_type), main = "Distribucija tipa prebivališta", xlab = "Tip prebivališta",
```

Distribucija tipa zaposlenja



Distribucija tipa prebivališta



```
prop.table(table(data$work_type, useNA = "ifany"))
```

```
##
##      children      Govt_job  Never_worked      Private Self-employed
## 0.134442270 0.128571429 0.004305284 0.572407045 0.160273973
```

```
prop.table(table(data$Residence_type, useNA = "ifany"))
```

```
##
##      Rural      Urban
## 0.4919765 0.5080235
```

Većina pacijenata je zaposlena u privatnom sektoru, dok 0,4% njih nikada nije radilo.

Raspodjela između urbane i ruralne sredine u kojoj živi pacijent je gotovo jednolika.

```
mean(data$stroke)
```

```
## [1] 0.04872798
```

U nastavku rada nastojat će se pokazati povezanost opisanih atributa s moždanim udarom. Uz pomoć statističkih testova nastojat će se zaključiti o tome kako i s kolikom uspješnosti je moguće prepoznati rizik od moždanog udara kod pojedine osobe.

Postoji li statistički značajna razlika u prosječnoj razini glukoze između pacijenata sa i bez moždanog udara?

Potrebno je odrediti postoji li značajna razlika u prosječnoj razini glukoze između pacijenata sa i bez moždanog udara. Započnimo s preciznim određivanjem hipoteza - H_0 : Ne postoji značajna razlika, odnosno: $\{\text{mean}(\text{glucose_stroke}) - \text{mean}(\text{glucose_no_stroke}) = 0\}$ - H_1 : Postoji značajna razlika: $\{\text{mean}(\text{glucose_stroke}) - \text{mean}(\text{glucose_no_stroke}) \neq 0\}$

```
glucose_stroke <- data[data$stroke == 1, ]$avg_glucose_level
glucose_no_stroke <- data[data$stroke == 0,]$avg_glucose_level
summary(glucose_stroke)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   56.11   79.79  105.22  132.54  196.71  271.74
```

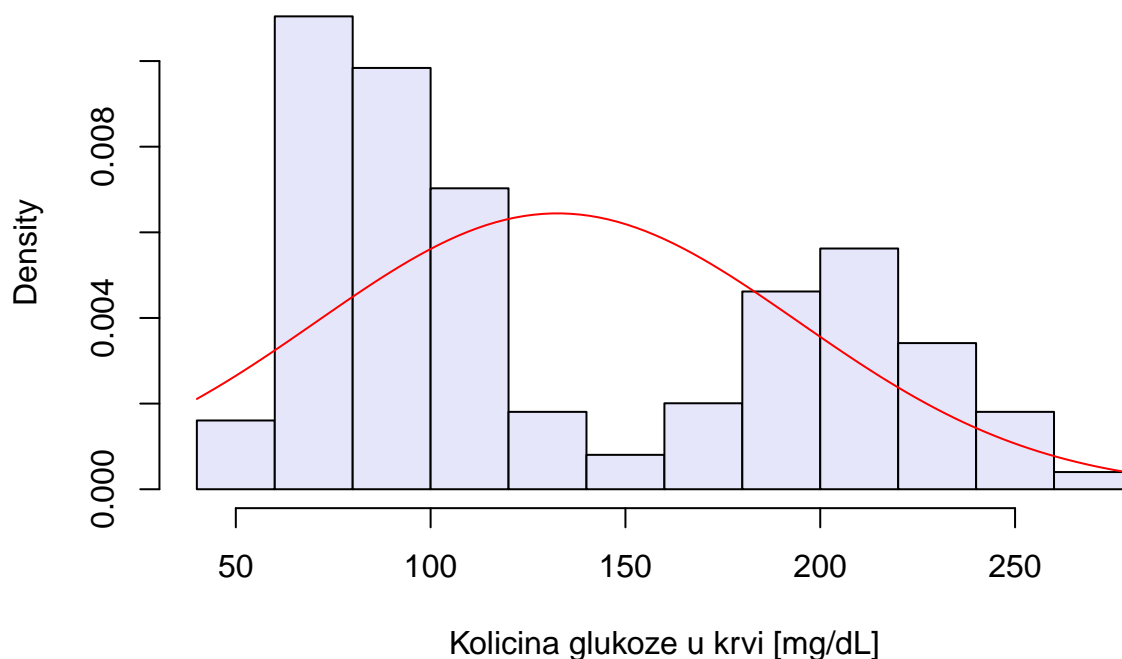
```
summary(glucose_no_stroke)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   55.12   77.12   91.47  104.80  112.83  267.76
```

Relativno velika razlika medijana i aritmetičke sredine govori o tome koliko je distribucija količine glukoze zakrivljena. Obe distribucije imaju težak desni rep, a distribucija kod ljudi koji su imali moždani udar se može čak smatrati i bimodalnom. Zbog toga se ne može smatrati normalnom, međutim centralni granični teorem omogućava da se za testiranje hipoteza koristi t-test kad su uzorci veliki. Bez obzira na to provest će se i neparametarska metoda da bi se usporedili rezultati.

```
hist(glucose_stroke , freq = FALSE, main = "Kolicina glukoze u krvi kod ljudi s mozdanim udarom", xlab = "Glucose level",
curve(dnorm(x, mean(glucose_stroke), sd(glucose_stroke)), add = TRUE, col = "red")
```

Kolicina glukoze u krvi kod ljudi s mozdanim udarom



```
result <- t.test(avg_glucose_level ~ stroke, data = data, var.equal = FALSE)
result
```

```
##
## Welch Two Sample t-test
##
## data: avg_glucose_level by stroke
## t = -6.9824, df = 260.89, p-value = 2.401e-11
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
## -35.57474 -19.92371
## sample estimates:
## mean in group 0 mean in group 1
## 104.7955 132.5447
```

Welchov t-test za dva uzorka s različitim varijancama odbacuje nultu hipotezu da su dvije sredine jednake, odnosno prihvaća alternativu gdje su sredine različite. To se može vidjeti iz p-vrijednosti koja je jako blizu nuli. Grupa 0 označava ljude koji nisu imali moždani udar i oni očekivano imaju manju prosječnu razinu glukoze u krvi, dok preostala grupa ima gotovo 28 mg/dL više glukoze u krvi. Ta se vrijednost nalazi i u izračunatom intervalu pouzdanosti.

Ukratko odbacuje se nulta hipoteza u korist alternativne hipoteze. Sada ćemo to još jednom provjeriti uz pomoć neparametarske metode bootstrap.

```

set.seed(67)

B <- 10000

boot_diff <- numeric(B)

for (i in 1:B) {
  boot_stroke <- sample(glucose_stroke, replace = TRUE)
  boot_no_stroke <- sample(glucose_no_stroke, replace = TRUE)
  boot_diff[i] <- mean(boot_stroke) - mean(boot_no_stroke)
}

# Observed mean difference
obs_diff <- mean(glucose_stroke) - mean(glucose_no_stroke)

# 95% bootstrap confidence interval
ci <- quantile(boot_diff, c(0.025, 0.975))

list(
  observed_mean_difference = obs_diff,
  ci_95 = ci
)

```

```

## $observed_mean_difference
## [1] 27.74923
##
## $ci_95
##      2.5%      97.5%
## 19.94809 35.74191

```

Vidljivo je da interval 95%-ne pouzdanosti odgovara onom izračunatom u Welchovom testu. Neparametarska metoda bootstrap također odbacuje nultu hipotezu u korist alternativne. Korištena je zato što ona ne zahtijeva da skup podataka koji testira bude neke određene distribucije.

Postoji li interakcijski učinak hipertenzije i srčanih bolesti na BMI?

U ovom istraživačkom pitanju proučavamo ovisi li indeks tjelesne mase (BMI) o prisutnosti hipertenzije i srčanih bolesti te postoji li interakcijski učinak ova dva čimbenika. Dakle, zanima nas utječe li hipertenzija na BMI jednako kod osoba sa i bez srčanih bolesti ili se taj učinak mijenja ovisno o postojanju srčane bolesti. Za potrebe analize koristimo dvostruku analizu varijance s hipertenzijom i srčanim bolestima kao čimbenicima te BMI-jem kao zavisnom varijablom.

Priprema podataka

Najprije pripremamo podatke. Brišemo nedostajuće vrijednosti BMI-ja, pretvaramo BMI u numeričku varijablu te kodiramo hipertenziju i srčane bolesti kao faktore s dvije razine.

```

data2 <- read.csv("data.csv", na.strings = "N/A", stringsAsFactors = FALSE)

```

```
data2$hypertension <- factor(data2$hypertension, levels = c(0,1), labels = c("no_hypertension", "has_hy
data2$heart_disease <- factor(data2$heart_disease, levels = c(0,1), labels = c("no_heart_disease", "has_l

data2$bmi <- as.numeric(data$bmi)
```

```
## Warning: NAs introduced by coercion
```

```
anova_data <- data2[, c("bmi", "hypertension", "heart_disease")]
```

```
anova_data <- na.omit(anova_data)
```

```
str(anova_data$bmi)
```

```
##  num [1:4909] 36.6 32.5 34.4 24 29 27.4 22.8 24.2 29.7 36.8 ...
```

```
summary(anova_data$bmi)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.30   23.50   28.10   28.89   33.10   97.60
```

```
table(anova_data$hypertension, anova_data$heart_disease)
```

```
##
##              no_heart_disease has_heart_disease
## no_hypertension             4273             185
##  has_hypertension             393              58
```

Broj opažanja nakon izbacivanja nedostajućih vrijednosti daje nam efektivnu veličinu uzorka za ovu analizu.

Ukupno je u analizi sudjelovalo 4909 ispitanika. Raspodjela prema prisutnosti hipertenzije i srčane bolesti je sljedeća:

- nema hipertenziju, nema srčanu bolest: **87.0 %**
- nema hipertenziju, ima srčanu bolest: **3.8 %**
- ima hipertenziju, nema srčanu bolest: **8.0 %**
- ima hipertenziju, ima srčanu bolest: **1.2 %**

Promatramo li postotke unutar razina hipertenzije:

- među osobama **bez hipertenzije**:
 - 95.9 % **nema** srčanu bolest
 - 4.1 % **ima** srčanu bolest
- među osobama **s hipertenzijom**:
 - 87.1 % **nema** srčanu bolest
 - 12.9 % **ima** srčanu bolest

Promatramo li postotke unutar razina srčane bolesti:

- među osobama **bez srčane bolesti**:
 - 91.6 % **nema** hipertenziju
 - 8.4 % **ima** hipertenziju
- među osobama **sa srčanom bolešću**:
 - 76.1 % **nema** hipertenziju
 - 23.9 % **ima** hipertenziju

```
head(anova_data)
```

```
##      bmi      hypertension      heart_disease
## 1 36.6   no_hypertension has_heart_disease
## 3 32.5   no_hypertension has_heart_disease
## 4 34.4   no_hypertension no_heart_disease
## 5 24.0 has_hypertension no_heart_disease
## 6 29.0   no_hypertension no_heart_disease
## 7 27.4 has_hypertension has_heart_disease
```

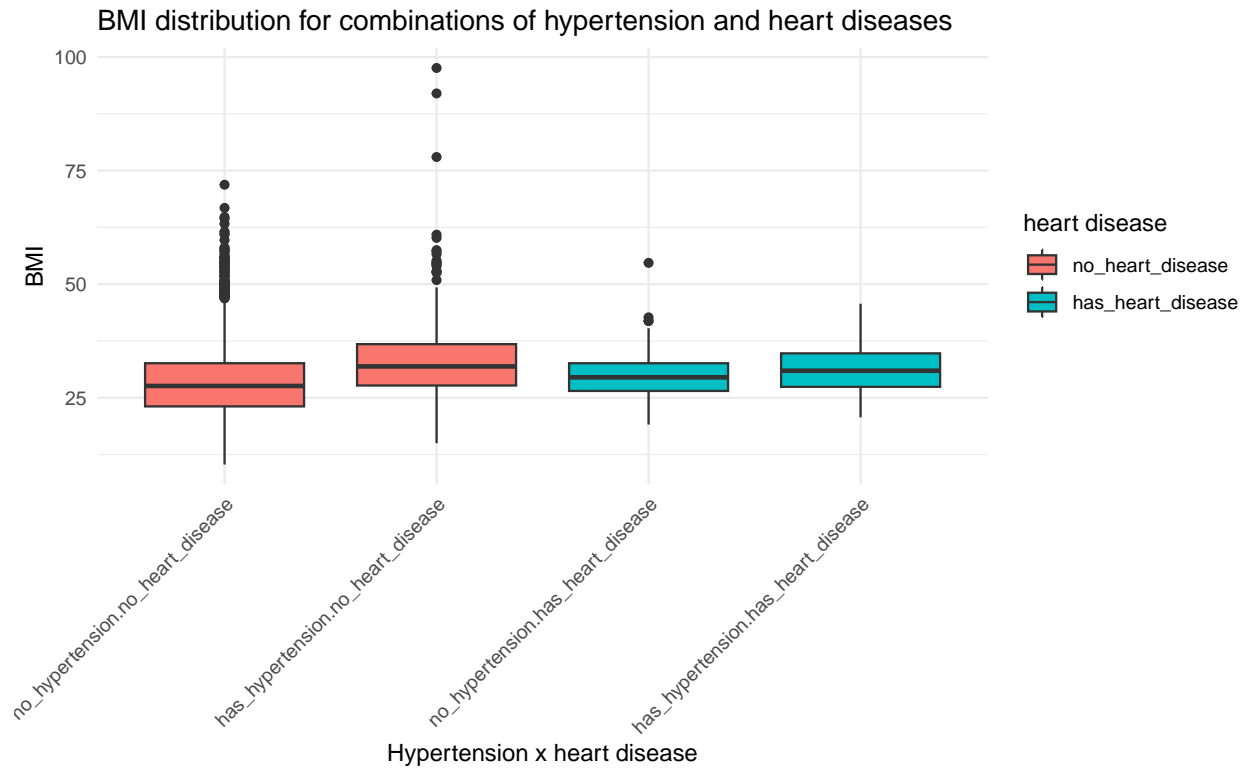
Sredine po grupama

Najprije izračunavamo prosječne vrijednosti BMI-ja po svakoj kombinaciji hipertenzije i srčane bolesti te ih vizualiziramo.

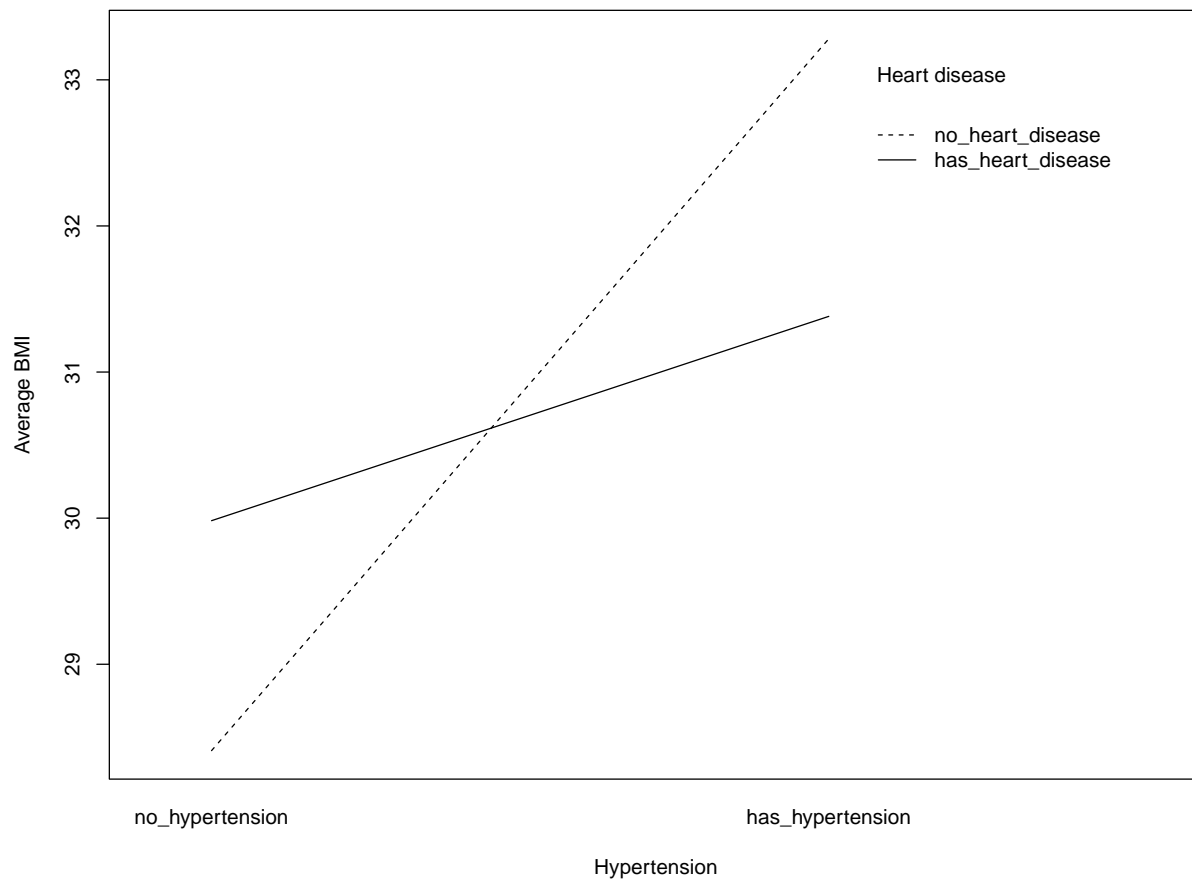
```
aggregate(bmi ~ hypertension + heart_disease,
          data = anova_data,
          FUN = mean)
```

```
##      hypertension      heart_disease      bmi
## 1 no_hypertension no_heart_disease 28.40875
## 2 has_hypertension no_heart_disease 33.28092
## 3 no_hypertension has_heart_disease 29.98270
## 4 has_hypertension has_heart_disease 31.38103
```

```
ggplot(anova_data,
       aes(x = interaction(hypertension, heart_disease),
           y = bmi,
           fill = heart_disease)) +
  geom_boxplot() +
  labs(x = "Hypertension x heart disease",
       y = "BMI",
       fill = "heart disease",
       title = "BMI distribution for combinations of hypertension and heart diseases") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



```
with(anova_data,  
  interaction.plot(hypertension, heart_disease, bmi,  
    fun = mean,  
    xlab = "Hypertension",  
    ylab = "Average BMI",  
    trace.label = "Heart disease"))
```



Iz tablice sredina i grafova vidimo kako se prosječni BMI razlikuje između grupa. Zbog toga nas zanima mijenja li se razlika u BMI-ju između osoba s i bez hipertenzije ovisno o tome imaju li srčanu bolest, što je indicacija mogućeg interakcijskog učinka.

Dvostruka ANOVA s interakcijom: formulacija hipoteza i primjena testa

Glavni učinak hipertenzije:

- H_0 : prosječni BMI je jednak kod osoba sa i bez hipertenzije
- H_1 : prosječni BMI se razlikuje između osoba sa i bez hipertenzije

Glavni učinak srčanih bolesti:

- H_0 : prosječni BMI je jednak kod osoba sa i bez srčane bolesti
- H_1 : prosječni BMI se razlikuje između osoba sa i bez srčane bolesti

Interakcijski učinak:

- H_0 : nema interakcijskog učinka hipertenzije i srčanih bolesti na BMI

- H1: postoji interakcijski učinak hipertenzije i srčanih bolesti na BMI

Za testiranje hipoteza koristimo dvostruku ANOVA-u s interakcijom:

```
test_anova <- aov(bmi ~ hypertension * heart_disease, data = anova_data)
summary(test_anova)
```

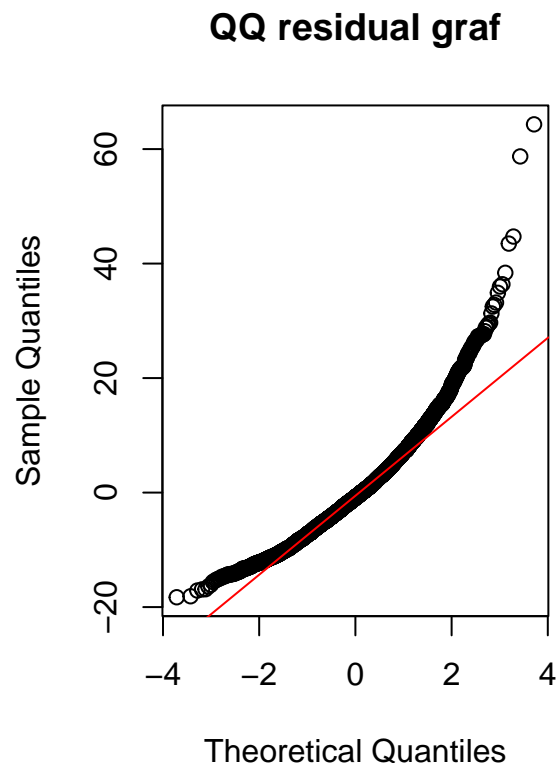
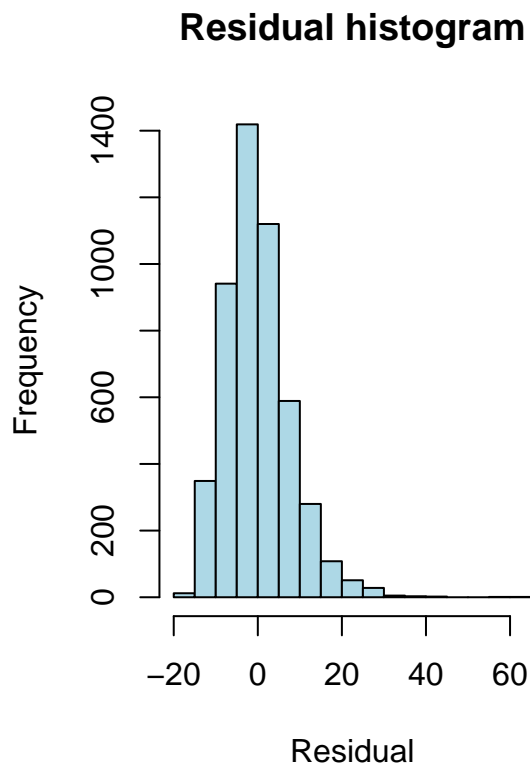
```
##              Df Sum Sq Mean Sq F value Pr(>F)
## hypertension      1    8526    8526 142.430 < 2e-16 ***
## heart_disease      1     147     147   2.457 0.11705
## hypertension:heart_disease 1     475     475   7.929 0.00488 **
## Residuals        4905 293609      60
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Provjera pretpostavki

Analiza varijance pretpostavlja približnu normalnost reziduala i sličnost varijanci između grupa.

```
rezid <- residuals(test_anova)

par(mfrow = c(1, 2))
hist(rezid, main = "Residual histogram", xlab = "Residual", col = "lightblue", border = "black")
qqnorm(rezid, main = "QQ residual graf")
qqline(rezid, col = "red")
```



```
par(mfrow = c(1, 1))
```

Reziduali pokazuju određena odstupanja, no zbog velikog uzorka ANOVA je relativno robusna na takva odstupanja.

Rezultati i interpretacija

Na temelju ANOVA tablice zaključujemo sljedeće:

- Glavni učinak hipertenzije na BMI je statistički značajan (F vrijednost je vrlo velika, $p < 0.001$), zbog čega možemo zaključiti da osobe s hipertenzijom u prosjeku imaju viši BMI od osoba bez hipertenzije.
- Glavni učinak srčanih bolesti na BMI nije statistički značajan (p-vrijednost je veća od 0.05), pa na ovoj razini značajnosti zaključujemo da prosječni BMI razlikuje između osoba sa i bez srčane bolesti kada se ostali čimbenici ne uzimaju u obzir.
- Interakcijski učinak hipertenzije i srčanih bolesti na BMI je statistički značajan ($p < 0.01$), zbog čega zaključujemo da se učinak hipertenzije na BMI razlikuje ovisno o tome ima li osoba srčanu bolest.

Iz tablice prosječnih BMI vrijednosti po grupama vidimo da je najviši BMI tipično prisutan kod osoba koje imaju i hipertenziju i srčanu bolest, dok je najniži BMI kod osoba koje nemaju ni hipertenziju ni srčanu bolest.

Postoji li povezanost između statusa pušenja i nastanka moždanog udara?

U analizi se ispituje postoji li povezanost između statusa pušenja i pojave moždanog udara. Prvo se varijable `smoking_status` i `stroke` pretvaraju u faktore kako bi se mogle ispravno koristiti u kontingencijskoj tablici i Chi-kvadrat testu.

Zatim se formira kontingencijska tablica za tri kategorije pušenja (bez Unknown) kako bi se izbjegla izobličenja u rezultatima.

Postotak moždanih udara po kategorijama pušenja

Kako bismo bolje razumjeli učestalost moždanog udara unutar svake skupine pušačkog statusa, izrađen je barplot koji prikazuje proporciju pacijenata s moždanim udarom u odnosu na ukupni broj osoba unutar svake kategorije pušenja.

Iz grafa se može jasno vidjeti u kojoj skupini je udio moždanih udara najveći, što dodatno nadopunjuje statistički rezultat Chi-kvadrat testa.

```
# Proporcije moždanih udara unutar svake kategorije pušenja
# Redovi: status pušenja, stupci: stroke (0/1)
data$smoking_status <- as.factor(data$smoking_status)
data$stroke <- as.factor(data$stroke)

tablica <- table(data$smoking_status, data$stroke)
tablica <- tablica[-4, ] # uklanjamo 'Unknown'

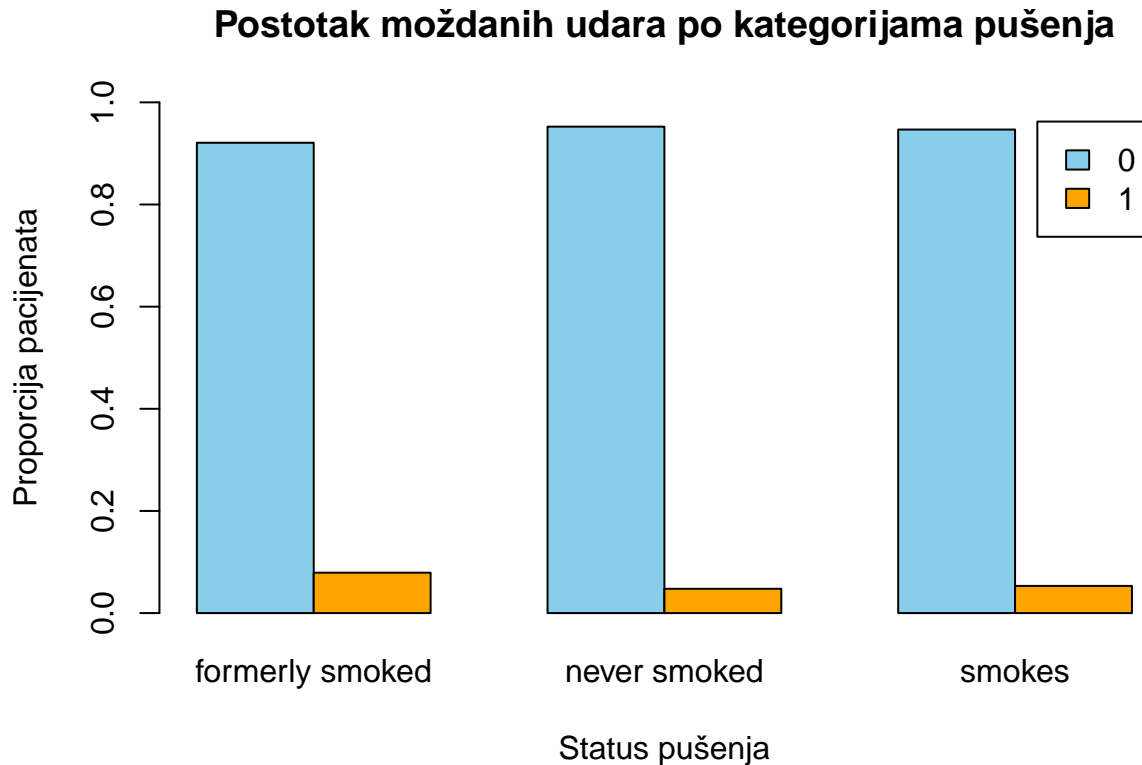
prop_tab <- prop.table(tablica, 1) # row-wise proportions

# Barplot - prikaz proporcija
```

```

barplot(t(prop_tab),
        beside = TRUE,                # stupci jedan do drugog
        legend = TRUE,                # legenda za stroke 0/1
        col = c("skyblue", "orange"), # boje za stroke 0 i 1
        main = "Postotak moždanih udara po kategorijama pušenja",
        xlab = "Status pušenja",
        ylab = "Proporcija pacijenata",
        ylim = c(0,1))                # proporcija od 0 do 1

```



Chi-kvadrat test – osnovna tablica

U ovom dijelu analiziramo odnos između statusa pušenja i pojave moždanog udara. Nakon čišćenja podataka kreirana je kontingencijska tablica koja uključuje tri kategorije pušenja: never smoked, formerly smoked i smokes (kategorija Unknown je uklonjena).

Na temelju ove tablice proveden je Chi-kvadrat test neovisnosti kako bi se provjerilo postoji li statistički značajna povezanost između statusa pušenja i moždanog udara. Test vraća vrijednost statistike, stupnjeve slobode i p-vrijednost. Ako je p-vrijednost manja od 0.05, odbacujemo hipotezu o neovisnosti i zaključujemo da je status pušenja povezan s pojavom moždanog udara.

Uz to se pregledavaju i očekivane frekvencije kako bismo provjerili jesu li zadovoljeni uvjeti za ispravno provođenje Chi-kvadrat testa (svaka očekivana frekvencija > 5). Ovaj test pruža osnovni uvid u to razlikuju li se skupine pušača u učestalosti moždanog udara.

Hipoteze:

- H0: Status pušenja nije povezan s pojavom moždanog udara (neovisne su varijable).
- H1: Status pušenja je povezan s pojavom moždanog udara (varijable nisu neovisne).

```
# Čišćenje: pretvaranje smoking_status u faktor
data$smoking_status <- as.factor(data$smoking_status)
data$stroke <- as.factor(data$stroke)
```

```
# Kontingencijska tablica
tablica <- table(data$smoking_status, data$stroke)
tablica <- tablica[-4, ]
tablica
```

```
##
##              0      1
##  formerly smoked 815   70
##  never smoked   1802  90
##  smokes          747  42
```

```
# Chi-square test
chi_rezultat <- chisq.test(tablica)
chi_rezultat
```

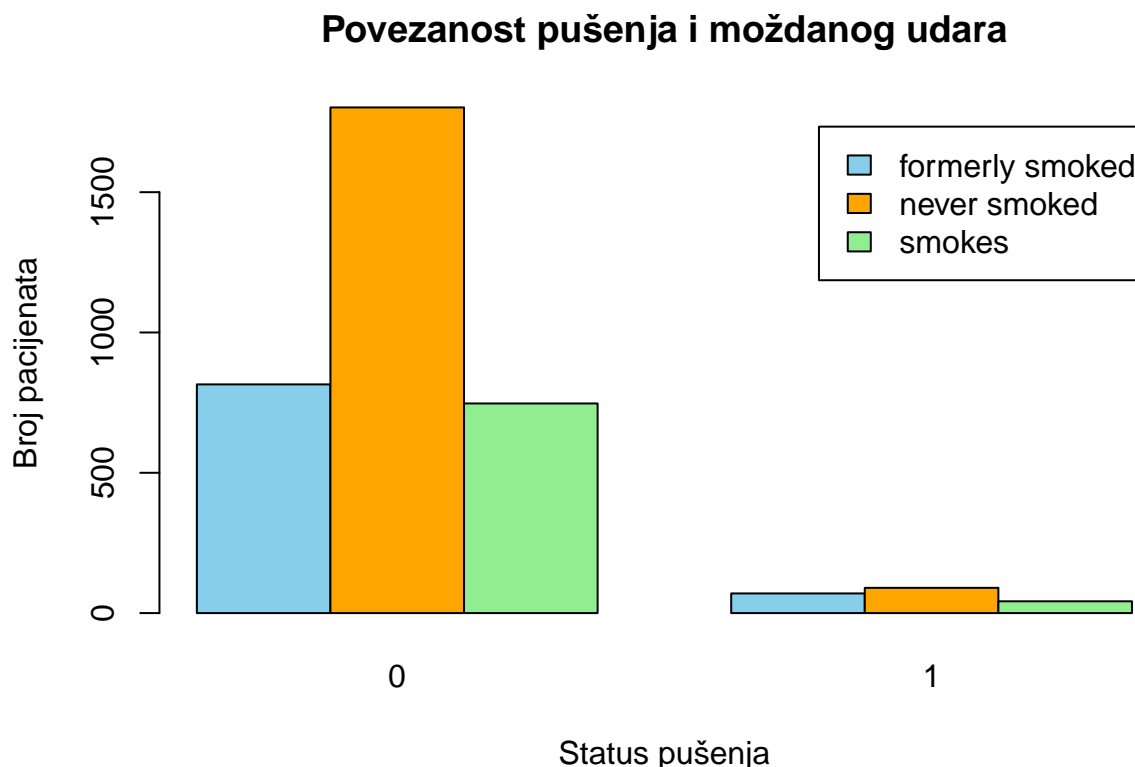
```
##
##  Pearson's Chi-squared test
##
## data:  tablica
## X-squared = 11.436, df = 2, p-value = 0.003285
```

```
chi_rezultat$expected
```

```
##
##              0      1
##  formerly smoked 834.8682 50.13180
##  never smoked   1784.8256 107.17443
##  smokes          744.3062  44.69377
```

U nastavku je prikazan barplot koji vizualno predstavlja povezanost između statusa pušenja i pojave moždanog udara. Svaka skupina pušačkog statusa (never smoked, formerly smoked, smokes) prikazana je s pripadnim brojem osoba koje su doživjele ili nisu doživjele moždani udar.

```
# Barplot - odnos pušenja i moždanog udara (osnovna tablica)
barplot(tablica,
        beside = TRUE,
        legend = TRUE,
        col = c("skyblue", "orange", "lightgreen"),
        main = "Povezanost pušenja i moždanog udara",
        xlab = "Status pušenja",
        ylab = "Broj pacijenata")
```



Kontingencijska tablica uključivala je tri kategorije pušenja: never smoked, formerly smoked i smokes. Chi-kvadrat test neovisnosti pokazao je p-vrijednost manju od 0.05, što znači da postoji statistički značajna povezanost između statusa pušenja i moždanog udara. Očekivane frekvencije su bile zadovoljavajuće (>5), što potvrđuje valjanost testa.

Spajanje kategorija – bivši i sadašnji pušači vs. nikad pušili

U drugom pristupu spajaju se kategorije formerly smoked i smokes u jednu zajedničku skupinu nazvanu „was smoking/smokes“. Time se formiraju dvije jasne skupine:

1. Osobe koje nikada nisu pušile (never smoked)
2. Osobe koje su pušile, bez obzira na to puše li trenutno ili su prestale (was smoking/smokes)

Ovakva podjela omogućuje ispitivanje temeljnog pitanja: Je li bilo kakva povijest pušenja povezana s pojavom moždanog udara?

Nakon formiranja nove tablice provodi se Chi-kvadrat test neovisnosti kako bi se provjerilo razlikuje li se učestalost moždanog udara između ove dvije skupine. Ako je dobivena p-vrijednost statistički značajna ($p < 0.05$), zaključujemo da postoji povezanost između povijesti pušenja i rizika moždanog udara.

Uz test izrađen je i barplot koji grafički prikazuje raspodjelu moždanog udara u dvjema skupinama. Vizualizacija dodatno olakšava uočavanje eventualnih razlika između osoba koje nikad nisu pušile i onih koje su pušile barem jednom u životu.

Hipoteze:

- H0: Povijest pušenja (nikad vs. ikad) nije povezana s pojavom moždanog udara.
- H1: Povijest pušenja (nikad vs. ikad) jest povezana s pojavom moždanog udara.

```
# Kreiranje nove tablice: spajamo kategorije 'formerly smoked' i 'smokes'
nova_tablica <- table(data$smoking_status, data$stroke)

# Uklanjanje kategoriju 'Unknown' jer nema korisne informacije
nova_tablica <- nova_tablica[-4, ]

# Spajamo redove: formerly smoked + smokes
# (to znači da gledamo osobe koje su ikada pušile - trenutno ili ranije)
nova_tablica[2,] <- nova_tablica[2, ] + nova_tablica[3, ]

# Brišemo sada već spojen treći red
nova_tablica <- nova_tablica[-3, ]

# Preimenujemo redove radi jasnijeg prikaza
rownames(nova_tablica)[1] <- "never smoked"           # nikad nisu pušili
rownames(nova_tablica)[2] <- "was smoking/smokes"     # ikad pušili (prije ili sada)

# Prikaz nove tablice nakon spajanja
nova_tablica

##
##              0      1
## never smoked    815   70
## was smoking/smokes 2549 132

# Chi-kvadrat test za provjeru povezanosti između pušenja (nikad vs. ikad) i moždanog udara
# Ako je p-vrijednost < 0.05 → postoji statistička povezanost između pušenja i moždanog udara
chi_rezultat <- chisq.test(nova_tablica)

# Ispis rezultata testa
chi_rezultat

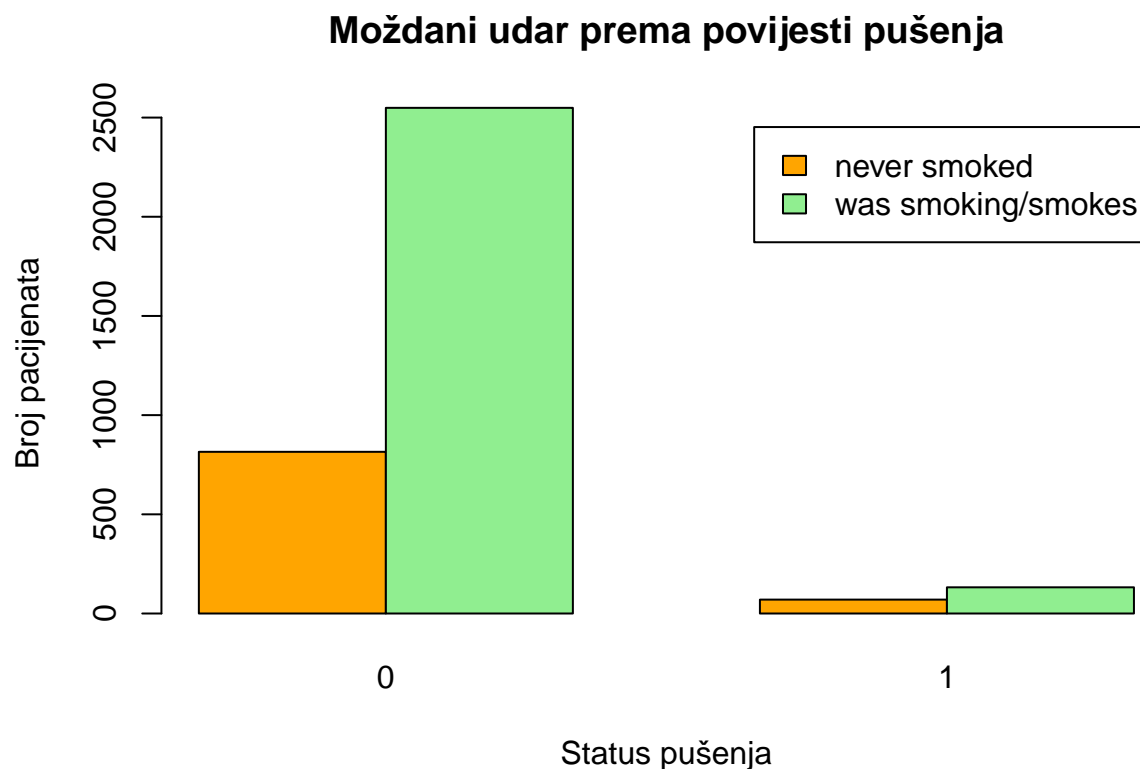
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data: nova_tablica
## X-squared = 10.551, df = 1, p-value = 0.001162

chi_rezultat$expected

##
##              0      1
## never smoked    834.8682 50.1318
## was smoking/smokes 2529.1318 151.8682

barplot(nova_tablica,
        beside = TRUE,
        legend = TRUE,
```

```
col = c("orange", "lightgreen"),
main = "Moždani udar prema povijesti pušenja",
xlab = "Status pušenja",
ylab = "Broj pacijenata")
```



Kombinirane su kategorije formerly smoked i smokes u jednu skupinu „was smoking/smokes“, a druga skupina su osobe koje nikada nisu pušile. Chi-kvadrat test na ovoj dvodijelnoj tablici također je dao p-vrijednost manju od 0.05, što potvrđuje statistički značajnu razliku u učestalosti moždanog udara između osoba koje nikada nisu pušile i onih koje su ikada pušile (trenutno ili prije).