MULTIVARIJANTNA I DUBINSKA ANALIZA PODATAKA

ZDRAVSTVENI POKAZATELJI DIJABETESA

CDC DIABETES HEALTH INDICATORS

JOSHUA LEE FLETCHER, NOA MIDŽIĆ
MENTORICA: PROF. DR. SC. JASMINKA DOBŠA
FAKULTET ORGANIZACIJE I INFORMATIKE
INFORMACIJSKO I PROGRAMSKO INŽENJERSTVO 1.4

SADRŽAJ

Dataset i deskriptivna statistika

Provedene obrade

Hi-Kvadrat test

Wilcoxon rank-sum test

Zaključak

PODACI O DATASETU

- Behavioral risk factor surveillance system (BRFSS) – godišnja anketa CDC-a preko telefona
- 70,692 pojedinaca odgovorilo
  - 21 varijabli (podaci su odgovori ili izračunati s obzirom na odgovore)
  - 50% nema dijabetes
- Diabetes_binary
  - 0 – nema dijabetes, 1 - ima preddijabetes ili dijabetes

Varijable koje ćemo analizirati

Prikupljeni podaci za različite rizične i druge faktore za dijabetes – kolesterol, fizičko i mentalno zdravlje, pokriće zdravstvenim osiguranjem, ekonomski i socijalni status…

Varijabla	Značenje	Vrsta varijable	Vrijednosti koje poprima
HighChol	Razina kolesterola	Kvalitativna, nominalna	0 (nizak), 1 (visok)
CholCheck	Pregledan kolesterol u zadnjih 5 godina?	Kvalitativna, nominalna	0 (jest), 1 (nije)
BMI_Group	Indeks tjelesne mase	Numerička, diskretna	12-98
HvyAlcoholConsump	Žene >= 7 pića tjedno Muškarci >= 14 pića tjedno	Kvalitativna, nominalna	0 (ne), 1 (da)
AnyHealthcare	Ima li zdravstveno osiguranje?	Kvalitativna, nominalna	0 (ne), 1 (da)
NoDocbcCost	U zadnjih godinu dana, nemogućnost odlaska doktoru zbog cijene?	Kvalitativna, nominalna	0 (ne), 1 (da)
Sex	Spol	Kvalitativna, nominalna	0 (žena), 1 (muškarac)

Tablica 1. Varijable koje ćemo analizirati

Varijable koje ćemo analizirati

Varijabla	Značenje	Vrsta varijable	Vrijednosti koje poprima
HighChol	Razina kolesterola	Kvalitativna, nominalna	0 (nizak), 1 (visok)
CholCheck	Pregledan kolesterol u zadnjih 5 godina?	Kvalitativna, nominalna	0 (jest), 1 (nije)
BMI_Group	Indeks tjelesne mase	Numerička, diskretna	12-98
HvyAlcoholConsump	Žene >= 7 pića tjedno Muškarci >= 14 pića tjedno	Kvalitativna, nominalna	0 (ne), 1 (da)
AnyHealthcare	Ima li zdravstveno osiguranje?	Kvalitativna, nominalna	0 (ne), 1 (da)
NoDocbcCost	U zadnjih godinu dana, nemogućnost odlaska doktoru zbog cijene?	Kvalitativna, nominalna	0 (ne), 1 (da)
Sex	Spol	Kvalitativna, nominalna	0 (žena), 1 (muškarac)

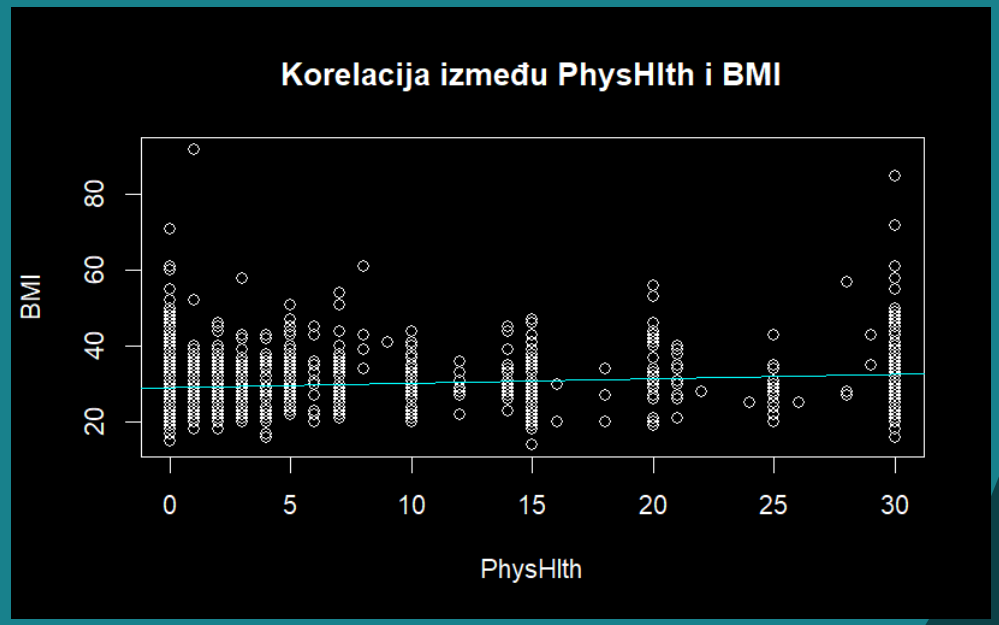
Tablica 1. Varijable koje ćemo analizirati

Varijable koje ćemo analizirati

Varijabla	Značenje	Vrsta varijable	Vrijednosti koje poprima
Age_Group	AGE5GYR skala dobi: 1 = 18-24, 2 = 25-29, 3 = 30-34, 4 = 35-39, 5 = 40-44, 6 = 45-49, 7 = 50-54, 8 = 55-59, 9 = 60-64, 10 = 65-69, 11 = 70-74, 12 = 75-79, 13 = 80+ godina	Kvalitativna, redoslijedna	1-13
Education_Group	EDUCA skala obrazovanja: 1 = samo vrtić, 2 = osnovna, 3 = nešto srednje škole, 4 = srednja škola, 5 = fakultet 1-3 godine, 6 = fakultet 4 godine ili više	Kvalitativna, redoslijedna	1-6
Income	INCOME2 skala zarade: 1 = manje od 10.000 dolara, 2 = 10.000-15.000 dolara, 3 = 15.000-20.000 dolara, 4 = 20.000-25.000, 5 = 25.000-35.000, 6 = 35.000-50.000, 7 = 50.000-75.000, 8 = 75.000 ili više dolara	Kvalitativna, redoslijedna	1-8

Tablica 1. Varijable koje ćemo analizirati

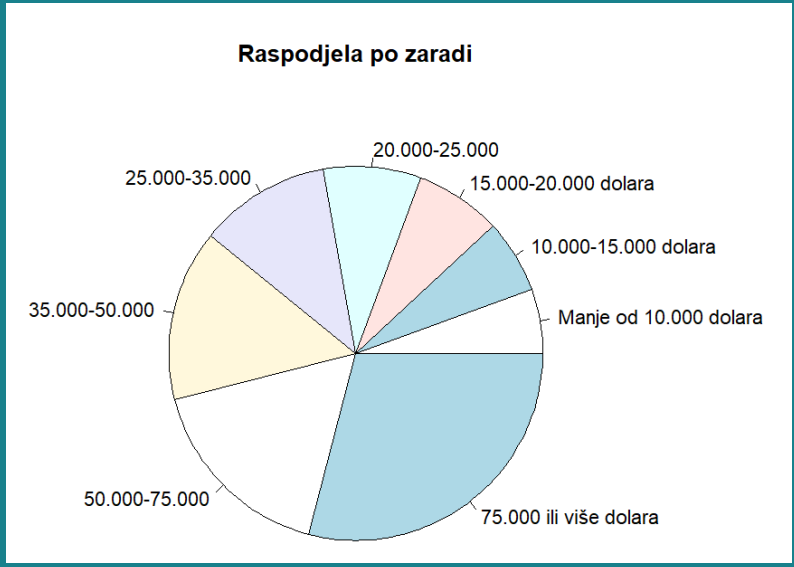
DESKRIPTIVNA STATISTIKA



Slika 1. Korelacija PhysHlth i BMI

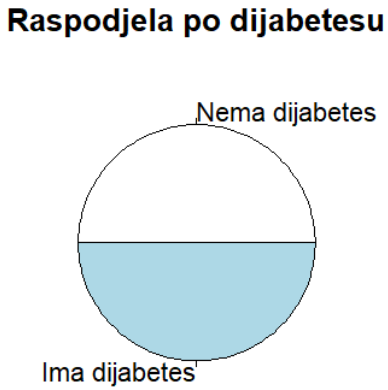
```
> # Deskriptivna statistika - numericka varijabla (BMI)
> median_bmi <- median(data$BMI)
> print(median_bmi)
[1] 29

> mean_bmi <- mean(data$BMI)
> print(mean_bmi)
[1] 29.77

> standard_deviation <- sd(data$BMI)
> print(standard_deviation)
[1] 6.806707

> variance <- var(data$BMI)
> print(variance)
[1] 46.33127

> quantiles <- quantile(data$BMI, probs = c(0.25, 0.5, 0.75))
> print(quantiles)
25% 50% 75%
 25  29  33

> correlation <- cor(data$PhysHlth, data$BMI)
> print(correlation)
[1] 0.1499764
```

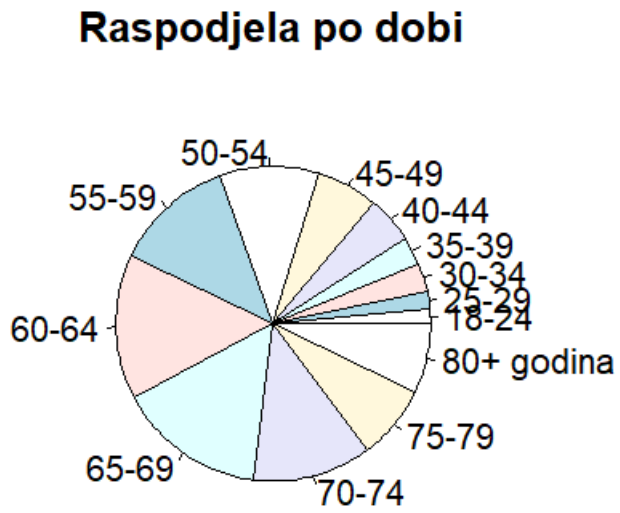
DESKRIPTIVNA STATISTIKA

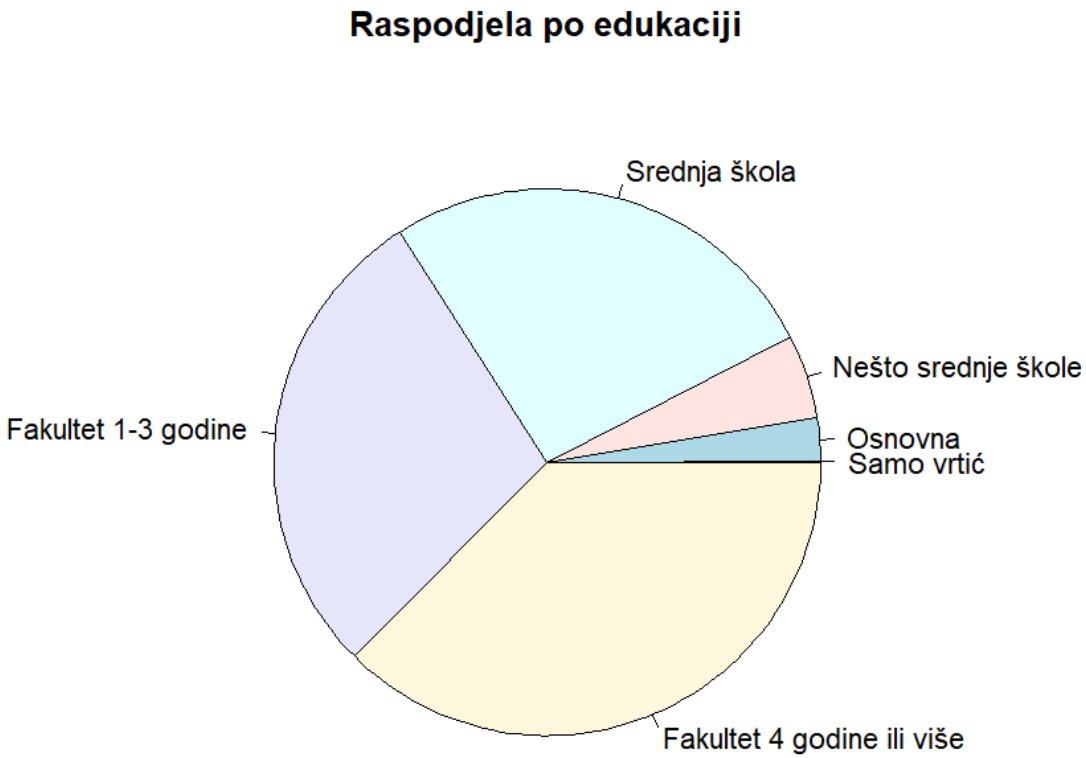


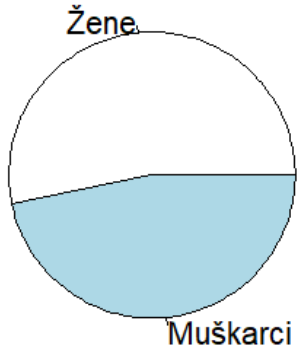
Slika 2. Raspodjela kvalitativnih varijabli

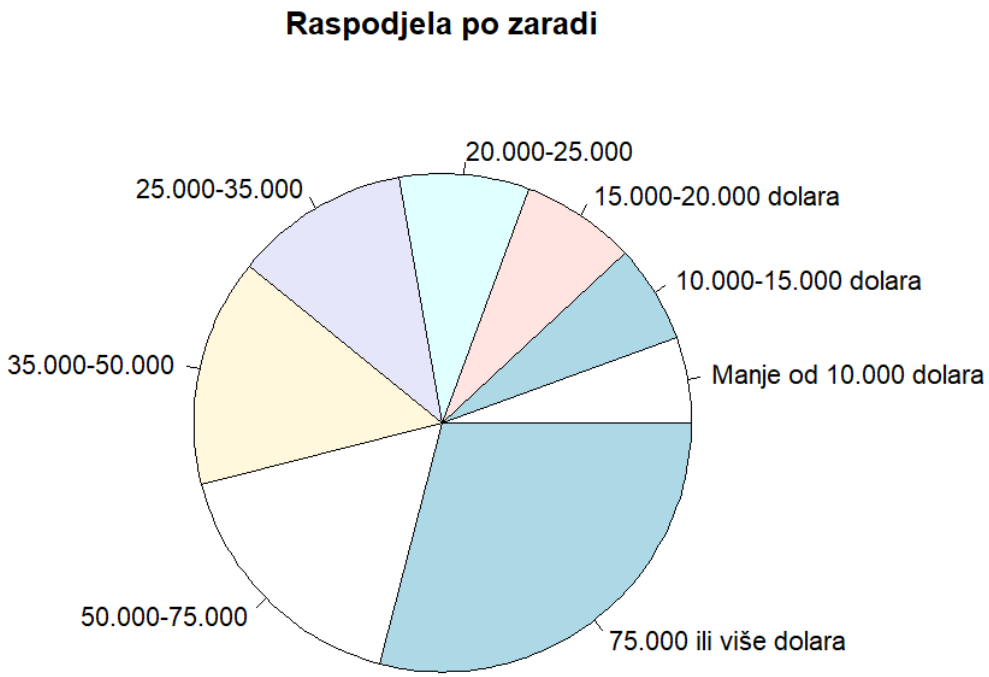
```r
# Deskriptivna statistika - kvalitativne varijable
promatranja <- table(data$Diabetes_binary)
pie(promatranja, labels = c("Nema dijabetes", "Ima dijabetes"), main =
"Raspodjela po dijabetesu")

promatranja <- table(data$Sex)
pie(promatranja, labels = c("Žene", "Muškarci"), main = "Raspodjela spola")

promatranja <- table(data$HighChol)
pie(promatranja, labels = c("Visok kolesterol", "Nizak kolesterol"),
                            main =  "Raspodjela po kolesterolu")

promatranja <- table(data$Age)
pie(promatranja, labels = c("18-24", "25-29", "30-34", "35-39", "40-44",
                            "45-49", "50-54", "55-59", "60-64", "65-69",
                            "70-74", "75-79", "80+ godina"),
    main = "Raspodjela po dobi")

promatranja <- table(data$Education)
pie(promatranja, labels = c("Samo vrtić", "Osnovna", "Nešto srednje škole",
                            "Srednja škola", "Fakultet 1-3 godine",
                            "Fakultet 4 godine ili više"),
    main = "Raspodjela po edukaciji")

promatranja <- table(data$Income)
pie(promatranja, labels = c("Manje od 10.000 dolara", "10.000-15.000 dolara",
                            "15.000-20.000 dolara", "20.000-25.000",
                            "25.000-35.000", "35.000-50.000", "50.000-75.000",
                            "75.000 ili više dolara"),
    main = "Raspodjela po zaradi")
```











Slika 2. Raspodjela kvalitativnih varijabli

PRIPREMA PODATAKA

- diabetes_binary_5050split_health_indicators_BRFSS2015.csv
- Python skripta
  - Odabir 2.000 opservacija na nasumičan način
  - 50% ispitanika nema dijabetes (0) i 50% ima preddijabetes ili dijabetes (1)

```
data = pd.read_csv("diabetes_binary_
    5050split_health_indicators_BRFSS2015.csv")

diabetes_0 = data[data['Diabetes_binary'] == 0]
diabetes_1 = data[data['Diabetes_binary'] == 1]

num_per_class = 2000 // 2

selected_0 = diabetes_0.sample(num_per_class)
selected_1 = diabetes_1.sample(num_per_class)

selected_data = pd.concat(
    [selected_0, selected_1])

selected_data = selected_data.sample(
    frac = 1).reset_index(drop = True)

selected_data.to_csv(
    "selected_data.csv", index = False)
```

PROVEDENE OBRADE

- 2 nezavisna uzorka
  - dijabetičari vs ne-dijabetičari
- Odabrane varijable (kontinuirane vrijednosti)
  - BMI: x ε [12, 98], Age: x ε [18, 80+]
- Provjera uvjeta za parametarski test:
  - Normalna distribucija (Shapiro-Wilks)
  - Jednakost varijanci skupova
    - S normalnom distribucijom skupova: F-test
    - Inače: Leveneov & Bartlettov test

Hi-kvadrat testovi

Shapiro-Wilksov, F-test, Leveneov i Bartlettov

Parametarski: T-test
Neparametarski: MWW (Wilcoxon rank-sum)

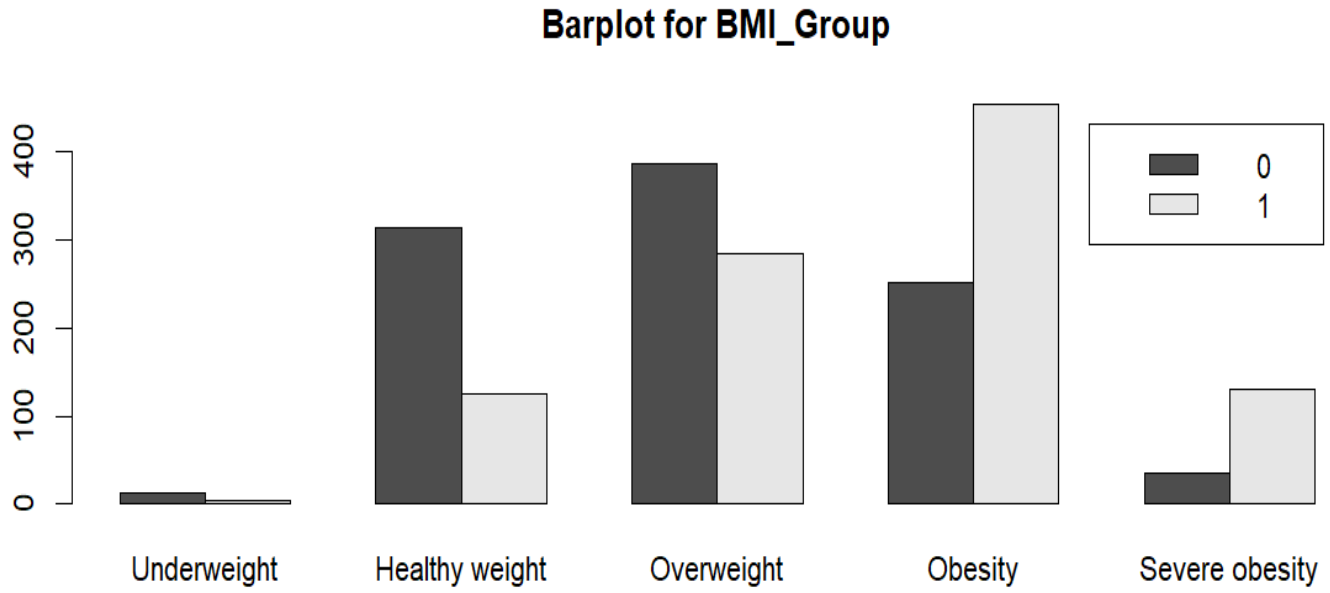
HI-KVADRAT TESTOVI NA RAZINI SIGNIFIKANTNOSTI 1%

- HighChol – H1
- CholCheck – H1
- BMI_Group – H1
- HvyAlcoholConsump – H1
- AnyHealthcare – H0
- NoDocbcCost – H0
- Sex – H0
- Age_Group – H1
- Education_Group – H1
- Income – H1

HI-KVADRAT TESTOVI BMI GRUPE

- Grupiranje podataka za primjenu hi-kvadrat testa

```
data$BMI_Group <- cut(
    data$BMI,
    breaks = c(
        0, 18.5, 24.9,
        29.9, 39.9, Inf),
    labels = c("Underweight“,
        "Healthy weight",
        "Overweight",
        "Obesity",
        "Severe obesity"))
```



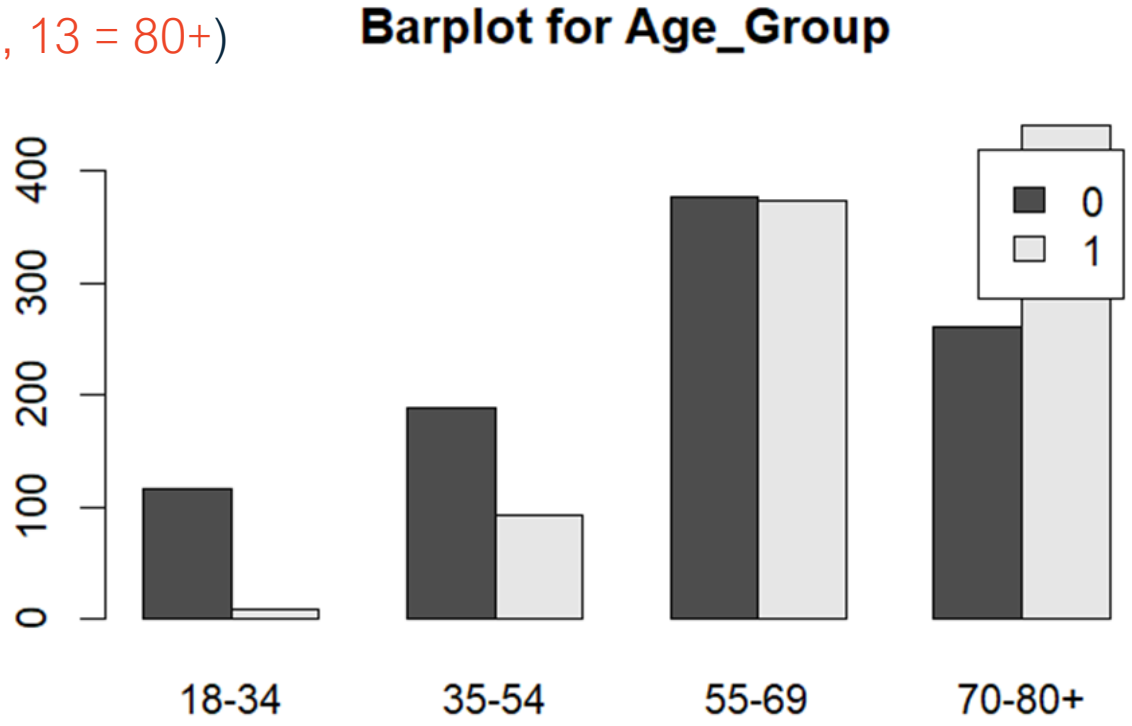
Slika 3. Barplot za BMI grupe

```
        Pearson's Chi-squared test

data:  Tab
X-squared = 216.08, df = 4, p-value < 2.2e-16


Reject H0: There is a significant association between BMI_Group and Diabetes_binary
```

HI-KVADRAT TESTOVI GRUPE GODINA

- Dataset koristi AGEG5YR (1 = 18-24, 9 = 60-64, 13 = 80+)

```
data$Age_Group <- cut(
    data$Age,
    breaks = c(
        seq(0, 13, by = 3)),
        labels = c(
            "18-34", "35-54",
            "55-69", "70-80+"))
```



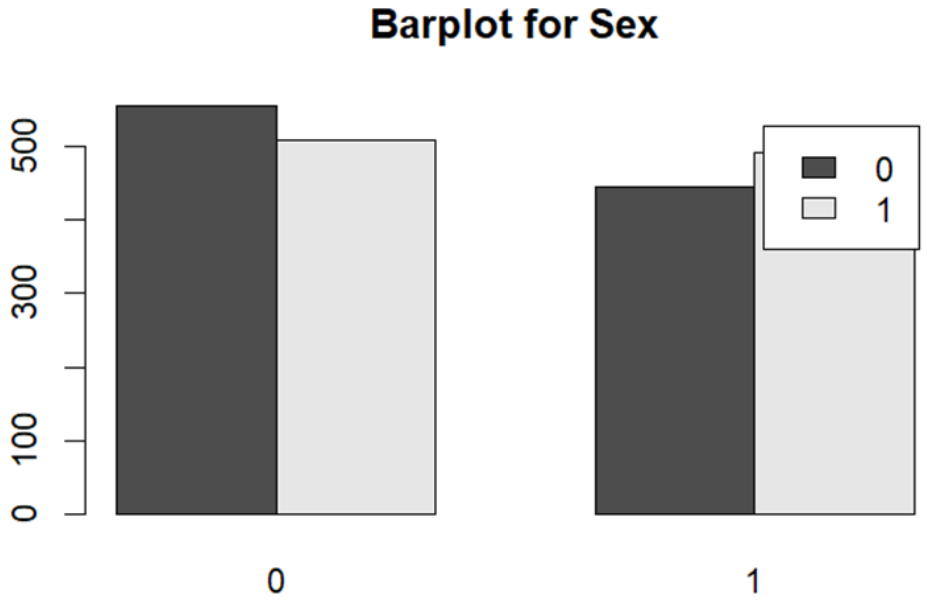
Slika 4. Barplot za grupe godina

```
        Pearson's Chi-squared test

data:  Tab
X-squared = 170.93, df = 3, p-value < 2.2e-16
```

```
Reject H0: There is a significant association between Age_Group and Diabetes_binary
```

HI-KVADRAT TESTOVI SPOL

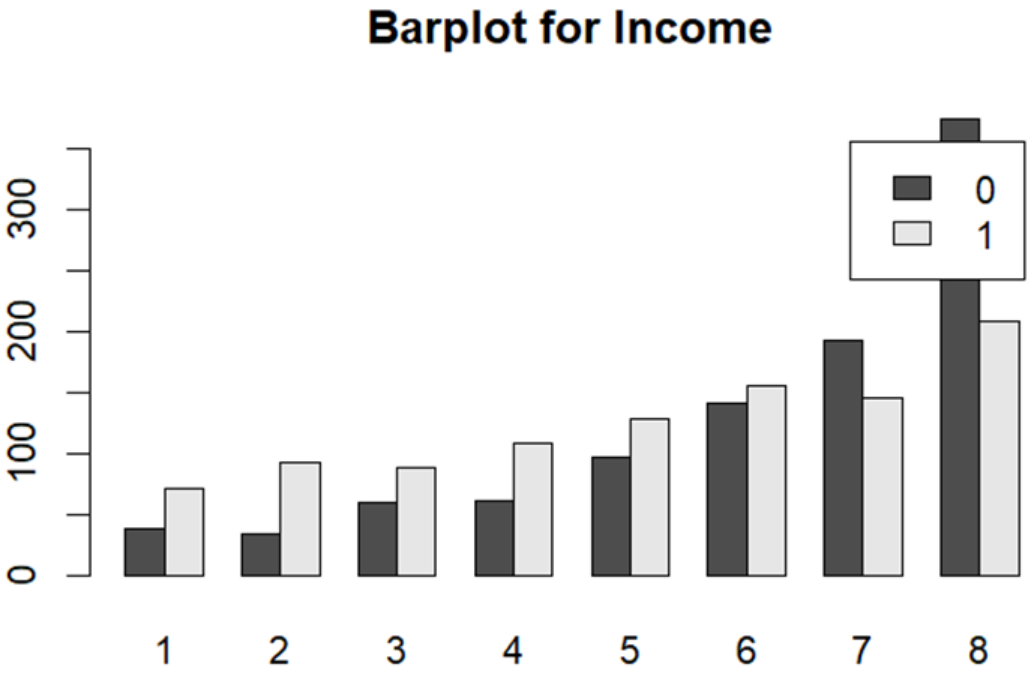
- Na razini signifikantnosti 1% - odbačena je nul-hipoteza
- Na razini signifikantnosti 5% - muškarci imaju veću tendenciju biti dijabetičari



Slika 5. Barplot za spol

```
Table for Sex :

      0   1
  0 555 445
  1 508 492

H0: The Sex is independent of Diabetes_binary
H1: The Sex is not independent of Diabetes_binary


        Pearson's Chi-squared test with Yates' continuity correction

data:  Tab
X-squared = 4.2489, df = 1, p-value = 0.03928


Fail to reject H0: There is no significant association between Sex and Diabetes_binary
```

HI-KVADRAT TESTOVI ZARADA



Slika 6. Barplot za zaradu

```
        Pearson's Chi-squared test

data:  Tab
X-squared = 113.86, df = 7, p-value < 2.2e-16


Reject H0: There is a significant association between
           Income and Diabetes_binary
```

PROVJERA UVJETA ZA T-TESTOVE SHAPIRO-WILKSOVI TESTOVI

```
Shapiro-Wilk test for normality for BMI :
Diabetic Group:

        Shapiro-Wilk normality test

data:  data[[var]][data$Diabetes_binary == 1]
W = 0.90346, p-value < 2.2e-16


Non-Diabetic Group:

        Shapiro-Wilk normality test

data:  data[[var]][data$Diabetes_binary == 0]
W = 0.90176, p-value < 2.2e-16
```

```
Shapiro-Wilk test for normality for Age :
Diabetic Group:

        Shapiro-Wilk normality test

data:  data[[var]][data$Diabetes_binary == 1]
W = 0.96189, p-value = 1.72e-15


Non-Diabetic Group:

        Shapiro-Wilk normality test

data:  data[[var]][data$Diabetes_binary == 0]
W = 0.96524, p-value = 1.039e-14
```

PROVJERA UVJETA F-TEST, LEVENEOV & BARTLETTOV TEST

```
F-test for equality of variances for BMI :

        F test to compare two variances

data:  data[[var]] by data$Diabetes_binary
F = 0.61225, num df = 999, denom df = 999, p-value = 1.229e-14
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.5408032 0.6931449
sample estimates:
ratio of variances
          0.612254


Levene's test for BMI :
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    1  34.387 5.274e-09 ***
      1998
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Bartlett's test for BMI :

        Bartlett test of homogeneity of variances

data:  data[[var]] by data$Diabetes_binary
Bartlett's K-squared = 59.491, df = 1, p-value = 1.229e-14
```

```
F-test for equality of variances for Age :

        F test to compare two variances

data:  data[[var]] by data$Diabetes_binary
F = 1.9681, num df = 999, denom df = 999, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.738397 2.228095
sample estimates:
ratio of variances
          1.968073


Levene's test for Age :
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    1  98.985 < 2.2e-16 ***
      1998
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Bartlett's test for Age :

        Bartlett test of homogeneity of variances

data:  data[[var]] by data$Diabetes_binary
Bartlett's K-squared = 112.31, df = 1, p-value < 2.2e-16
```

PROVJERA UVJETA ZA T-TESTOVE ZAKLJUČAK

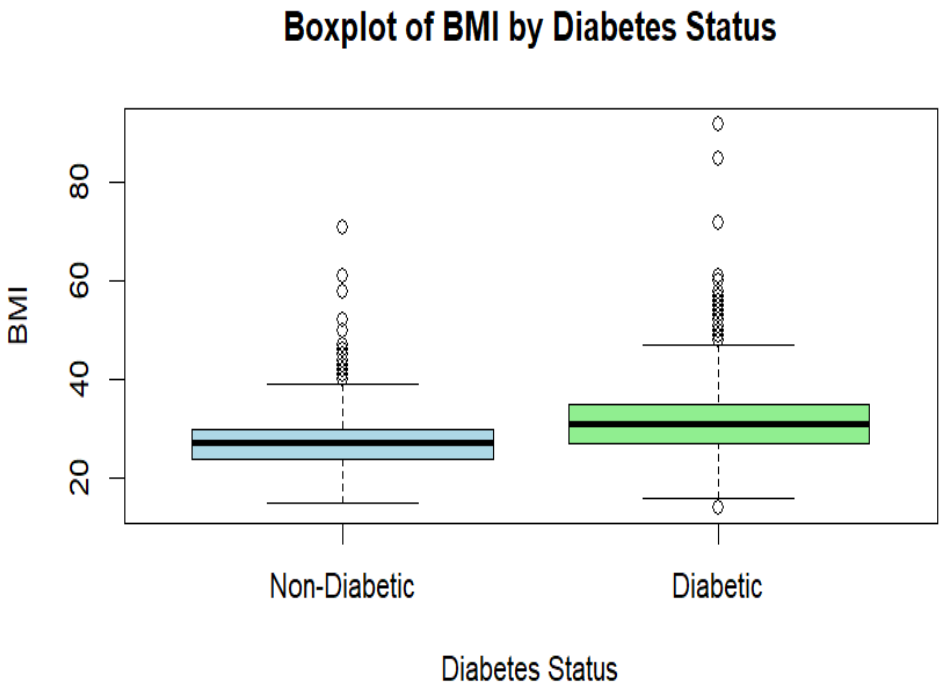
- Shapiro-Wilks – nema normalne distribucije za BMI ili Age ni u grupi dijabetičara ni u grupi ne-dijabetičara
    - *koristimo Leveneove i Bartlettove testove za potvrdu F-testa*
- F-test, Levene i Bartlett – varijance između skupova nisu jednake ni za BMI ni za Age
- **Nisu ispunjeni uvjeti za parametarski test**
- Umjesto T-testova, koristimo MWW

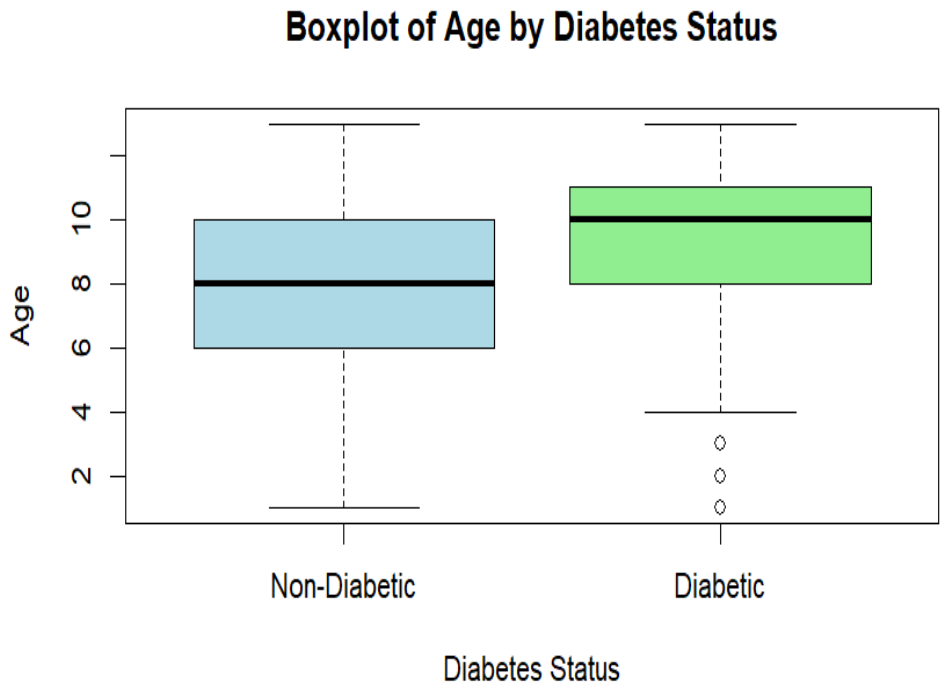
MWW TEST NAD BMI

```
At least one group does not follow a normal distribution.
Performing nonparametric Mann-Whitney U test (Wilcoxon rank sum test)...
Wilcoxon rank sum test for BMI :

        Wilcoxon rank sum test with continuity correction

data:  BMI by Diabetes_binary
W = 297212, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0


Reject H0: There is a significant difference in  BMI  between diabetic and non-diabetic groups.
```

MWW TEST NAD AGE

```
At least one group does not follow a normal distribution.
Performing nonparametric Mann-Whitney U test (Wilcoxon rank sum test)...
Wilcoxon rank sum test for Age :

        Wilcoxon rank sum test with continuity correction

data:  Age by Diabetes_binary
W = 345808, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0


Reject H0: There is a significant difference in  Age  between diabetic and non-diabetic groups.
```





Slika 7. Boxplotovi analize

```
Median for BMI in Diabetic group: 31
Median for BMI in Non-Diabetic group: 27

Median for Age in Diabetic group: 10
Median for Age in Non-Diabetic group: 8
```

ZAKLJUČAK

- Više dijabetičara kod većih BMI-jeva
- Više dijabetičara u starijim dobnim skupinama (70+ godina)
- Manje dijabetičara s višim stopama edukacije i zarade
- Na 1% razini signifikantnosti nema korelacije između dijabetesa i spola, odnosno imanja zdravstvenog osiguranja
- Nedostatak u datasetu: nije jasno navedeno je li uzet u obzir samo dijabetes tipa II ili oba tipa – u drugom slučaju se ovdje pronađene ovisnosti ne mogu generalizirati za sve tipove dijabetesa

testovi.R

GitHub poveznica na ovu datoteku u trenutku izrade dokumentacije.

```r
install.packages("car")
require("car")

data <- read.csv(file.choose())

data$Diabetes_binary <- as.factor(data$Diabetes_binary)

data$BMI_Group <- cut(data$BMI, breaks = c(0, 18.5, 24.9, 29.9, 39.9, Inf),
                      labels = c("Underweight", "Healthy weight", "Overweight", "Obesity", "Severe obesity"))
data$Age_Group <- cut(data$Age, breaks = c(seq(0, 13, by = 3)), labels = c("18-34", "35-54", "55-69", "70-80+"))
data$Education_Group <- cut(data$Education, breaks = c(seq(0, 6, by = 2)), labels = FALSE)

vars <- c("HighChol", "CholCheck", "BMI_Group", "HvyAlcoholConsump", "AnyHealthcare", "NoDocbcCost",
          "Sex", "Age_Group", "Education_Group", "Income")

for (var in vars) {
  tab <- table(data$Diabetes_binary, data[[var]])
  cat("Table for", var, ":\n")
  print(tab)
  barplot(tab, beside = TRUE, legend = TRUE, main = paste("Barplot for", var))

  c_test <- chisq.test(tab, correct = TRUE)
  cat("\nH0: The", var, "is independent of Diabetes_binary\n")
  cat("H1: The", var, "is not independent of Diabetes_binary\n\n")
  print(c_test)

  p_value <- c_test$p.value

  if (p_value <= 0.01) {
    cat("\nReject H0: There is a significant association between", var, "and Diabetes_binary\n\n")
  } else {
    cat("\nFail to reject H0: There is no significant association between", var, "and Diabetes_binary\n\n")
  }
}

vars_to_test <- c("BMI", "Age")

for (var in vars_to_test) { # nolint
  sw_test_diabetic <- shapiro.test(data[[var]][data$Diabetes_binary == 1])
  sw_test_nondiabetic <- shapiro.test(data[[var]][data$Diabetes_binary == 0])

  cat("\n\nShapiro-Wilk test for normality for", var, ":\n")
  cat("Diabetic Group:\n")
  print(sw_test_diabetic)
  cat("\nNon-Diabetic Group:\n")
  print(sw_test_nondiabetic)

  f_test <- var.test(data[[var]] ~ data$Diabetes_binary)

  cat("\nF-test for equality of variances for", var, ":\n")
  print(f_test)

  levene_test <- leveneTest(data[[var]], data$Diabetes_binary)

  cat("\nLevene's test for", var, ":\n")
  print(levene_test)

  bartlett_test <- bartlett.test(data[[var]] ~ data$Diabetes_binary)

  cat("\nBartlett's test for", var, ":\n")
  print(bartlett_test)


  if (sw_test_diabetic$p.value > 0.01 && sw_test_nondiabetic$p.value > 0.01) {
    cat("Both groups follow a normal distribution.\n")

   if (f_test$p.value > 0.01 && levene_test$p.value && bartlett_test$p.value) {
      cat("Variances of both groups are equal.\n")
      cat("Performing parametric two-sample t-test...\n")

      t_test <- t.test(data[[var]] ~ data$Diabetes_binary, conf.level = 0.99)
      print(t_test)

      if (t_test$p.value <= 0.01) {
        cat("\nReject H0: There is a significant difference in ", var, " between diabetic and non-diabetic groups.\n")
      } else {
        cat("\nFail to reject H0: There is no significant difference in ", var, " between diabetic and non-diabetic
groups.\n")
      }

    } else {
      cat("Variances of both groups are not equal.\n")
      cat("Performing nonparametric Mann-Whitney U test (Wilcoxon rank sum test)...\n")

      wilcox_test <- wilcox.test(as.formula(paste(var, "~ Diabetes_binary")), data = data)

      cat("Wilcoxon rank sum test for", var, ":\n")
      print(wilcox_test)

      if (wilcox_test$p.value <= 0.01) {
        cat("\nReject H0: There is a significant difference in ", var, " between diabetic and non-diabetic groups.\n")
      } else {
        cat("\nFail to reject H0: There is no significant difference in ", var, " between diabetic and non-diabetic
groups.\n")
      }

    }
  } else {
    cat("At least one group does not follow a normal distribution.\n")
    cat("Performing nonparametric Mann-Whitney U test (Wilcoxon rank sum test)...\n")

    wilcox_test <- wilcox.test(as.formula(paste(var, "~ Diabetes_binary")), data = data)

    cat("Wilcoxon rank sum test for", var, ":\n")
    print(wilcox_test)

    if (wilcox_test$p.value <= 0.01) {
      cat("\nReject H0: There is a significant difference in ", var, " between diabetic and non-diabetic groups.\n")
    } else {
      cat("\nFail to reject H0: There is no significant difference in ", var, " between diabetic and non-diabetic
groups.\n")
    }
  }
}
```

```r
for (var in vars_to_test) {
  median_diabetic <- median(data[[var]][data$Diabetes_binary == 1])
  median_nondiabetic <- median(data[[var]][data$Diabetes_binary == 0])

  cat("Median for", var, "in Diabetic group:", median_diabetic, "\n")
  cat("Median for", var, "in Non-Diabetic group:", median_nondiabetic, "\n\n")
}

vars_to_plot <- c("BMI", "Age")

par(mfrow = c(1, length(vars_to_plot)))

for (var in vars_to_plot) {
  boxplot(data[[var]] ~ data$Diabetes_binary,
          xlab = "Diabetes Status",
          ylab = var,
          main = paste("Boxplot of", var, "by Diabetes Status"),
          col = c("lightblue", "lightgreen"),
          names = c("Non-Diabetic", "Diabetic"))
}

par(mfrow = c(1, 1))


# Deskriptivna statistika - numericka varijabla (BMI)
median_bmi <- median(data$BMI)
print(median_bmi)

mean_bmi <- mean(data$BMI)
print(mean_bmi)

standard_deviation <- sd(data$BMI)
print(standard_deviation)

variance <- var(data$BMI)
print(variance)

quantiles <- quantile(data$BMI, probs = c(0.25, 0.5, 0.75))
print(quantiles)

correlation <- cor(data$PhysHlth, data$BMI)
print(correlation)

plot(data$PhysHlth, data$BMI,
     xlab = "PhysHlth",
     ylab = "BMI",
     main = "Korelacija između PhysHlth i BMI")

fit <- lm(data$BMI ~ data$PhysHlth)

abline(fit, col = "red")
```

GitHub poveznica na ovu datoteku u trenutku izrade dokumentacije.

```r
# Deskriptivna statistika - kvalitativne varijable
promatranja <- table(data$Diabetes_binary)
pie(promatranja, labels = c("Nema dijabetes", "Ima dijabetes"), main = "Raspodjela po dijabetesu")

promatranja <- table(data$Sex)
pie(promatranja, labels = c("Žene", "Muškarci"), main = "Raspodjela spola")

promatranja <- table(data$HighChol)
pie(promatranja, labels = c("Visok kolesterol", "Nizak kolesterol"), main = "Raspodjela po kolesterolu")

promatranja <- table(data$Age)
pie(promatranja, labels = c("18-24", "25-29", "30-34", "35-39", "40-44",
                            "45-49", "50-54", "55-59", "60-64", "65-69",
                            "70-74", "75-79", "80+ godina"),
    main = "Raspodjela po dobi")

promatranja <- table(data$Education)
pie(promatranja, labels = c("Samo vrtić", "Osnovna", "Nešto srednje škole",
                            "Srednja škola", "Fakultet 1-3 godine",
                            "Fakultet 4 godine ili više"),
    main = "Raspodjela po edukaciji")

promatranja <- table(data$Income)
pie(promatranja, labels = c("Manje od 10.000 dolara", "10.000-15.000 dolara",
                            "15.000-20.000 dolara", "20.000-25.000", "25.000-35.000",
                            "35.000-50.000", "50.000-75.000", "75.000 ili više dolara"),
    main = "Raspodjela po zaradi")
```

descriptive-statistics.R

GitHub poveznica na ovu datoteku u trenutku izrade dokumentacije.

```r
data <- read.csv(file.choose())

# Deskriptivna statistika
- numericka varijabla (BMI)
median_bmi <- median(data$BMI)
print(median_bmi)

mean_bmi <- mean(data$BMI)
print(mean_bmi)

standard_deviation <- sd(data$BMI)
print(standard_deviation)

variance <- var(data$BMI)
print(variance)

quantiles <- quantile(data$BMI, probs = c(0.25,
0.5, 0.75))
print(quantiles)

correlation <- cor(data$PhysHlth, data$BMI)
print(correlation)

plot(data$PhysHlth, data$BMI, xlab = "PhysHlth",
ylab = "BMI", main = "Korelacija između PhysHlth
i BMI")

fit <- lm(data$BMI ~ data$PhysHlth)

abline(fit, col = "red")
```

descriptive-statistics-pie-charts.R

GitHub poveznica na ovu datoteku u trenutku izrade dokumentacije.

```r
data <- read.csv(file.choose())

# Deskriptivna statistika - kvalitativne varijable
promatranja <- table(data$Diabetes_binary)
pie(promatranja, labels = c("Nema dijabetes", "Ima dijabetes"), main =
"Raspodjela po dijabetesu")

promatranja <- table(data$Sex)
pie(promatranja, labels = c("Žene", "Muškarci"), main = "Raspodjela
spola")

promatranja <- table(data$HighChol)
pie(promatranja, labels = c("Visok kolesterol", "Nizak kolesterol"), main
= "Raspodjela po kolesterolu")


promatranja <- table(data$Age)
pie(promatranja, labels = c("18-24", "25-29", "30-34", "35-39", "40-44",
                            "45-49", "50-54", "55-59", "60-64", "65-69",
                            "70-74", "75-79", "80+ godina"),
    main = "Raspodjela po dobi")

promatranja <- table(data$Education)
pie(promatranja, labels = c("Samo vrtić", "Osnovna", "Nešto srednje škole",
                            "Srednja škola", "Fakultet 1-3 godine",
                            "Fakultet 4 godine ili više"),
    main = "Raspodjela po edukaciji")

promatranja <- table(data$Income)
pie(promatranja, labels = c("Manje od 10.000 dolara",
"10.000-15.000 dolara",
                            "15.000-20.000 dolara",
"20.000-25.000", "25.000-35.000",
                            "35.000-50.000", "50.000-
75.000", "75.000 ili više dolara"),
    main = "Raspodjela po zaradi")
```

LITERATURA

Sav programski kod se može pronaći na javnom GitHub repozitoriju ovog rada: https://github.com/jfletcher20/diabetes-health-indicators-analysis

- *AGE5GYR - Variable Home Page* (bez dat.). Preuzeto na: https://www.icpsr.umich.edu/web/NAHDAP/studies/34085/datasets/0001/variables/AGEG5YR?archive=NAHDAP (pristupano: 5.5.2024.).
- *Body mass index (BMI) | NHS inform* (bez dat.). Preuzeto na: https://www.nhsinform.scot/healthy-living/food-and-nutrition/healthy-eating-and-weight-loss/body-mass-index-bmi/ (pristupano: 5.5.2024.).
- *Diabetes Health Indicators Dataset* (bez dat.). Preuzeto na: https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv (pristupano: 5.5.2024.).
- *EDUCA - Variable Home Page* (bez dat.). Preuzeto na: https://www.icpsr.umich.edu/web/NAHDAP/studies/34085/datasets/0001/variables/EDUCA?archive=nahdap (pristupano: 5.6.2024.).
- *INCOME2 - Variable Home Page* (bez dat.). Preuzeto na: https://www.icpsr.umich.edu/web/NAHDAP/studies/34085/datasets/0001/variables/INCOME2?archive=NAHDAP (pristupano: 5.6.2024.).

HVALA VAM NA PAŽNJI

JOSHUA LEE FLETCHER, NOA MIDŽIĆ