Spojeno

Grgur

12/16/2020

Uvod

Navike svakog čovjeka mogu imati pozitivan ili negativan utjecaj na njegovo zdravlje. U moderno doba uobičajeno je da čovjek iz raznih izvora saznaje razne informacije o utjecaju pojedinih akcija na njegovo zdravlje. U moru informacija ponekad je, međutim, teško razlučiti bitno od nebitnog, istinito od neistintog i odrediti koje navike imaju stvarni utjecaj na zdravlje i koliki taj utjecaj zapravo jest.

Cilj ovog projekta je istražiti preventivne mjere i zdravstvene tegobe koje imaju ljudi u raznim američkim gradovima, postoji li razlika u navikama ljudi u različitim gradovima i potencijalno pronaći vezu između pojedinih navika i njihovih utjecaja na zdravlje.

Učitavanje podataka

Učitavanje i upoznavanje s podatcima

Prvi korak je učitavanje i osnovno upoznavanje s podatcima.

```
health_data = read.csv("data_health_and_prevention.csv")
dim(health_data)
```

```
## [1] 16000    10
```

Podatci se sastoje od 16000 redaka i 10 stupaca. Svaki redak izražava udio stanovnika nekog američkog grada koji se pridržava određene preventivne mjere ili ima određeno zdravstveno stanje.

Tablice mogućih mjera i zdravstvenih stanja i njihov skraćen oblik dane su ovdje:

Table 1: Prevention

Short_Question_Text	Measure
Health Insurance	Current lack of health insurance among adults aged 18â€“64 Years
Taking BP Medication	Taking medicine for high blood pressure control among adults aged >=18 Years with high blood pressure
Annual Checkup	Visits to doctor for routine checkup within the past Year among adults aged >=18 Years
Cholesterol Screening	Cholesterol screening among adults aged >=18 Years

Table 2: Health Outcomes

Short_Question_Text	Measure
Arthritis	Arthritis among adults aged >=18 Years
High Blood Pressure	High blood pressure among adults aged >=18 Years
Cancer (except skin)	Cancer (excluding skin cancer) among adults aged >=18 Years
Current Asthma	Current asthma among adults aged >=18 Years
Coronary Heart Disease	Coronary heart disease among adults aged >=18 Years
COPD	Chronic obstructive pulmonary disease among adults aged >=18 Years
Diabetes	Diagnosed diabetes among adults aged >=18 Years
High Cholesterol	High cholesterol among adults aged >=18 Years who have been screened in the past 5 Years
Chronic Kidney Disease	Chronic kidney disease among adults aged >=18 Years
Mental Health	Mental health not good for >=14 days among adults aged >=18 Years
Physical Health	Physical health not good for >=14 days among adults aged >=18 Years
Stroke	Stroke among adults aged >=18 Years

Manipulacija podatcima

Za lakšu obradu podataka pretvaramo sljedeće stupce u faktorske varijable:

```
health_data$StateDesc = as.factor(health_data$StateDesc)
health_data$CityName = as.factor(health_data$CityName)
health_data$Category = as.factor(health_data$Category)
health_data$Measure = as.factor(health_data$Measure)
health_data$DataValueTypeID = as.factor(health_data$DataValueTypeID)
health_data$Short_Question_Text = as.factor(health_data$Short_Question_Text)
```

Svi podatci u datasetu izraženi su u dvije varijante: kao sirova stopa (Crude Rate) i kao dobno prilagođena stopa (Age-Adjusted Rate). Za razliku od sirove stope, dobno prilagođena uzima u obzir razlike u dobnoj raspodjeli stanovništva u različitim gradovima. S obzirom da države i gradove koje ćemo uspoređivati imaju različitu dobnu raspodjelu stanovništva, odlučili smo koristiti dobno prilagođene podatke.

```
health_data_adj = health_data[health_data$DataValueTypeID== "AgeAdjPrv",]
```

U pomoćne varijable dodajemo podatke o populaciji i broju gradova za svaku saveznu državu i statistike po pojedinim saveznim državama.

```
state_data <- health_data_adj %>% group_by(StateDesc) %>% summarise(
  City.count = n_distinct(CityName),
  Population.count = sum(unique(PopulationCount))
)

per_state_summary <- health_data_adj %>%
  group_by(StateDesc, Category, Measure, Short_Question_Text) %>% summarise(
  Total.percentage = sum(Data_Value*PopulationCount)/sum(PopulationCount),
  Population = sum(PopulationCount),
  Population.affected = round(sum(Data_Value*PopulationCount)/100)
)
```

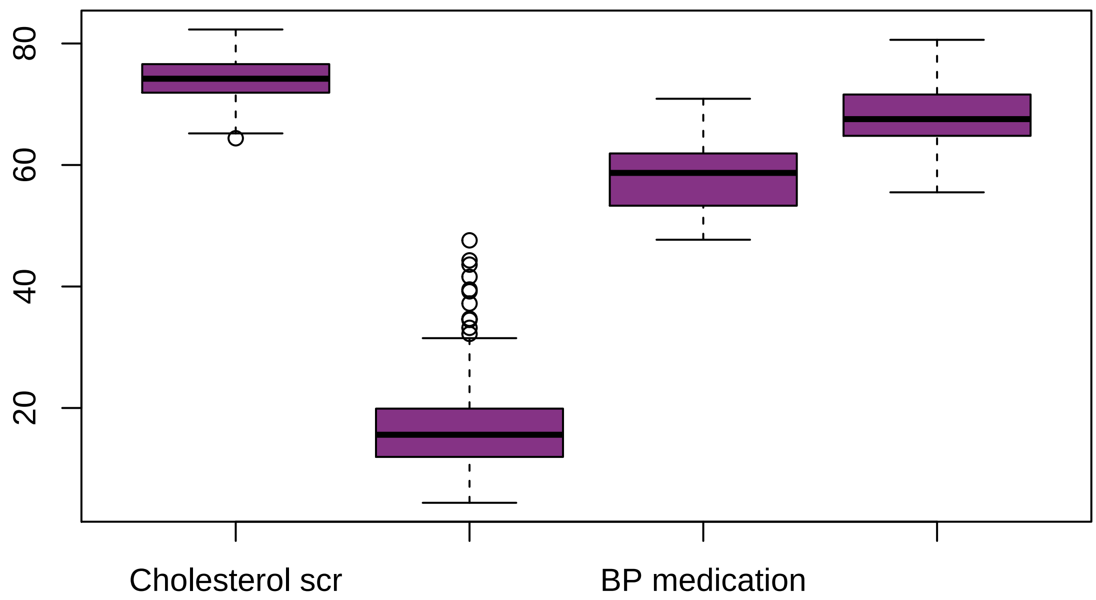
Za daljnji rad u dataset dodajemo nove stupce za postotak u svom mjerenom stanovništvu i ukupan broj ljudi zahvaćenih određenom mjerom ili zdravstvenim stanjem.

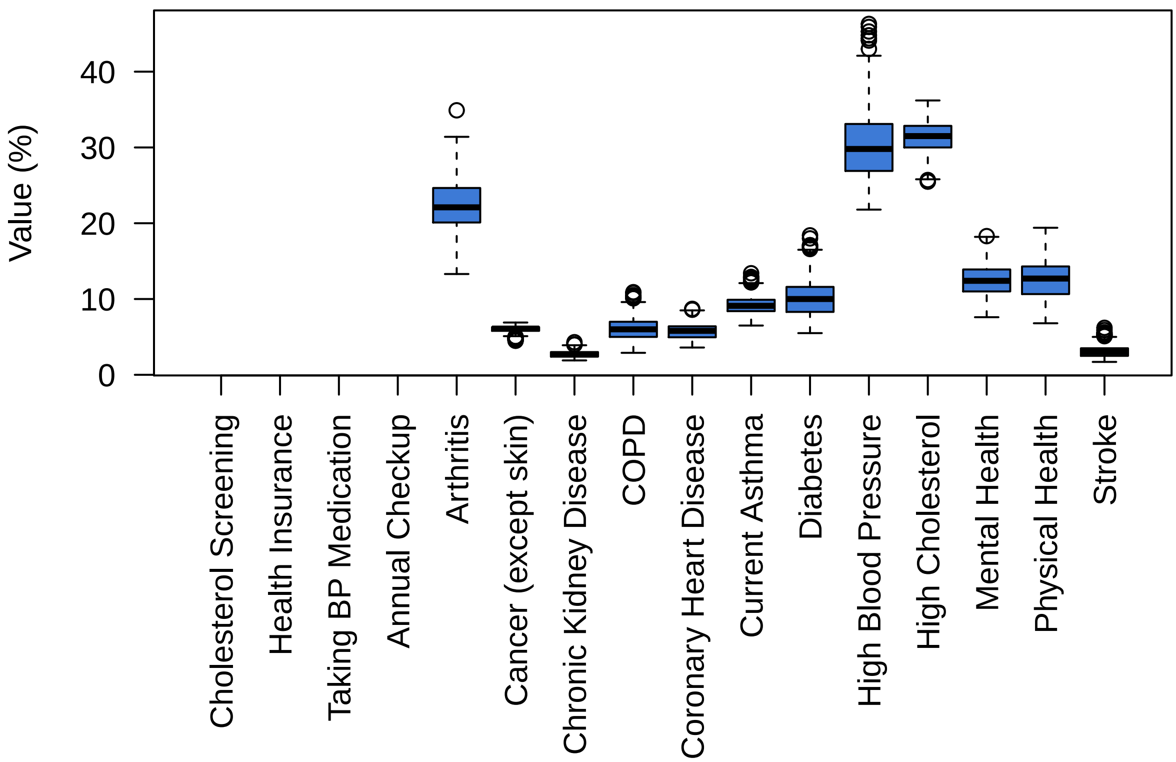
```
health_data_adj$Percentage_in_Total =
  health_data_adj$Data_Value*health_data_adj$PopulationCount/sum(state_data$Population.count)

health_data_adj$Affected_population =
  round( health_data_adj$Data_Value*health_data_adj$PopulationCount*0.01)
```

Deskriptivna statistika

Ukupni podatci

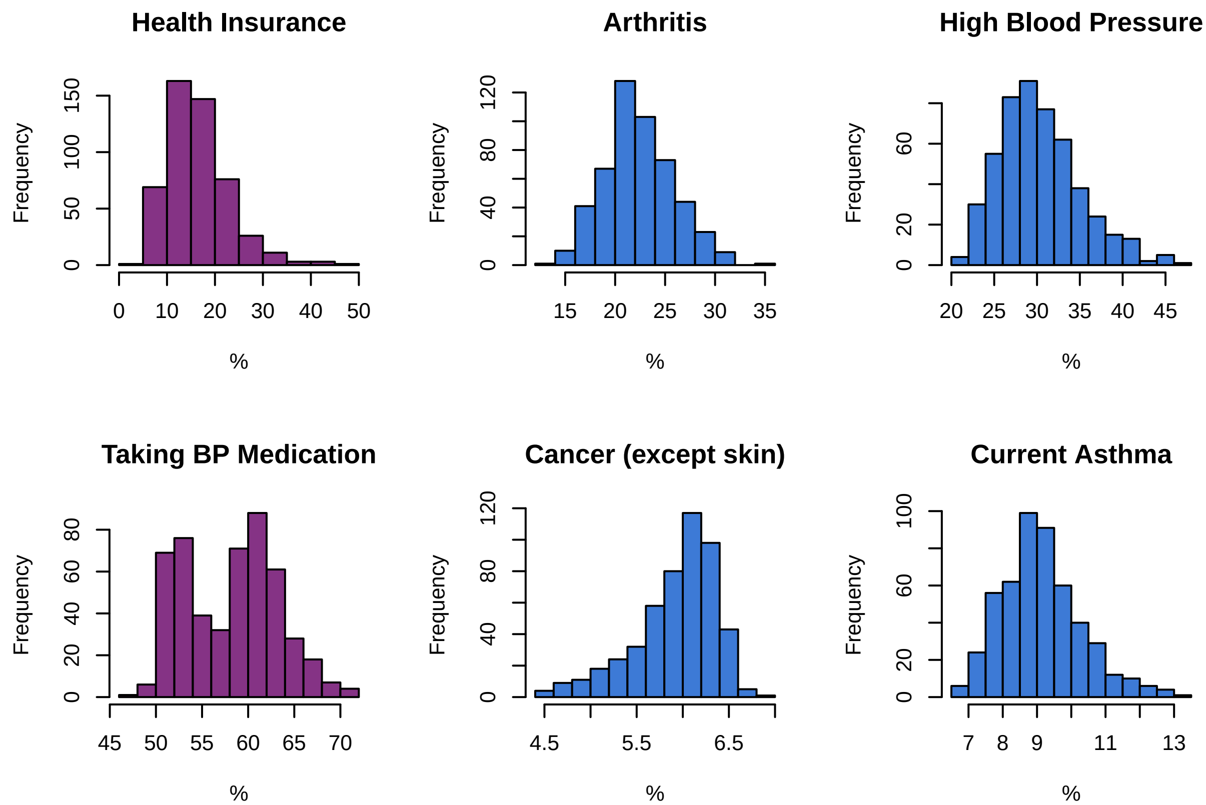
Prikaz raspodjele udjela građana koji primjenjuju pojedine preventivne mjere i imaju pojedina zdravstvena stanja:

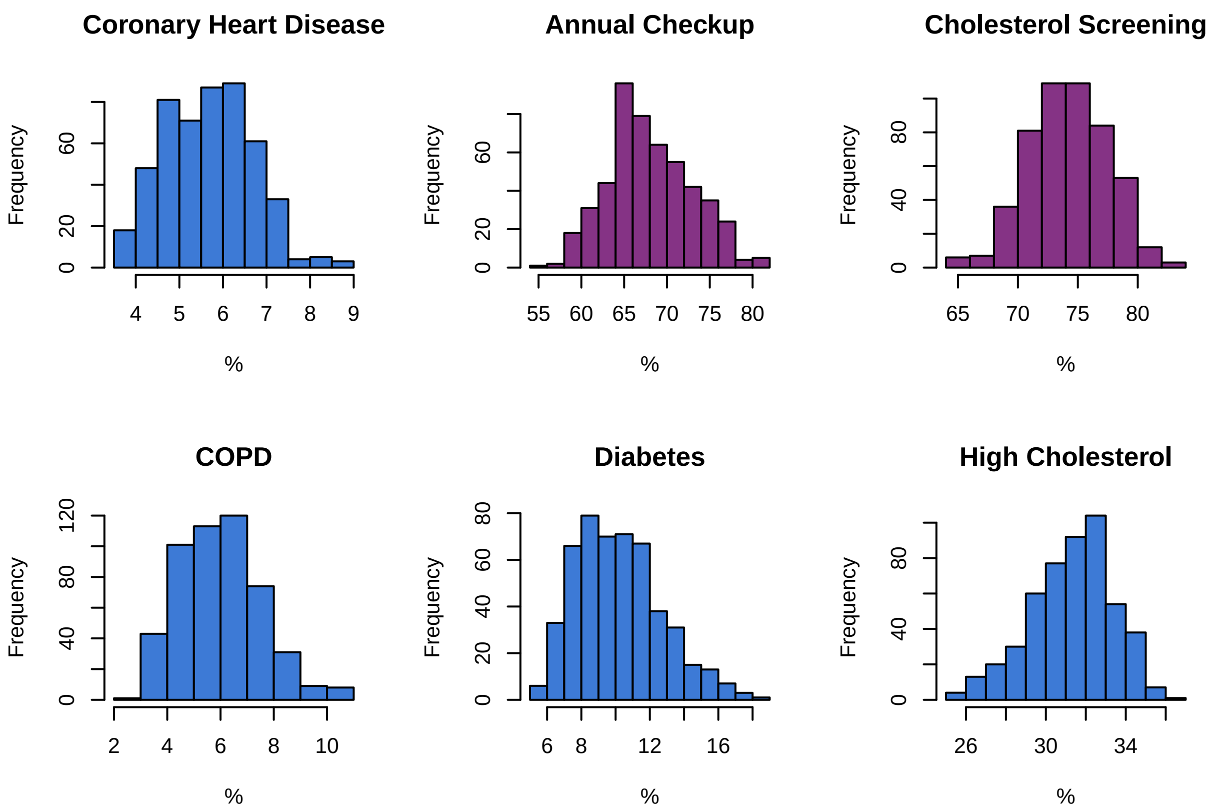


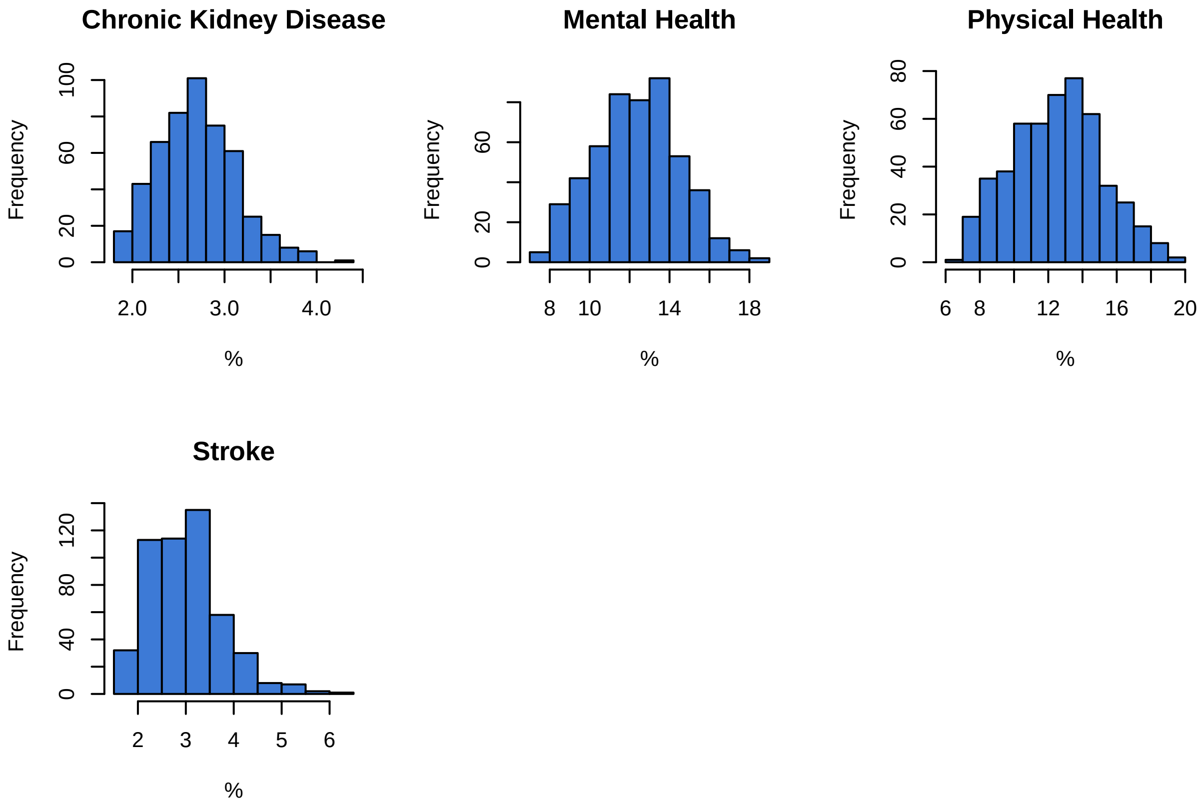


Value (%)
0
10
20
30
40
Cholesterol Screening
Health Insurance
Taking BP Medication
Annual Checkup
Arthritis
Cancer (except skin)
Chronic Kidney Disease
COPD
Coronary Heart Disease
Current Asthma
Diabetes
High Blood Pressure
High Cholesterol
Mental Health
Physical Health
Stroke

Pregledom histograma za svaku mjeru, primjećujemo da ih većina prati približno normalnu razdiobu, uz iznimku BP Medication koji izgleda bimodalno:

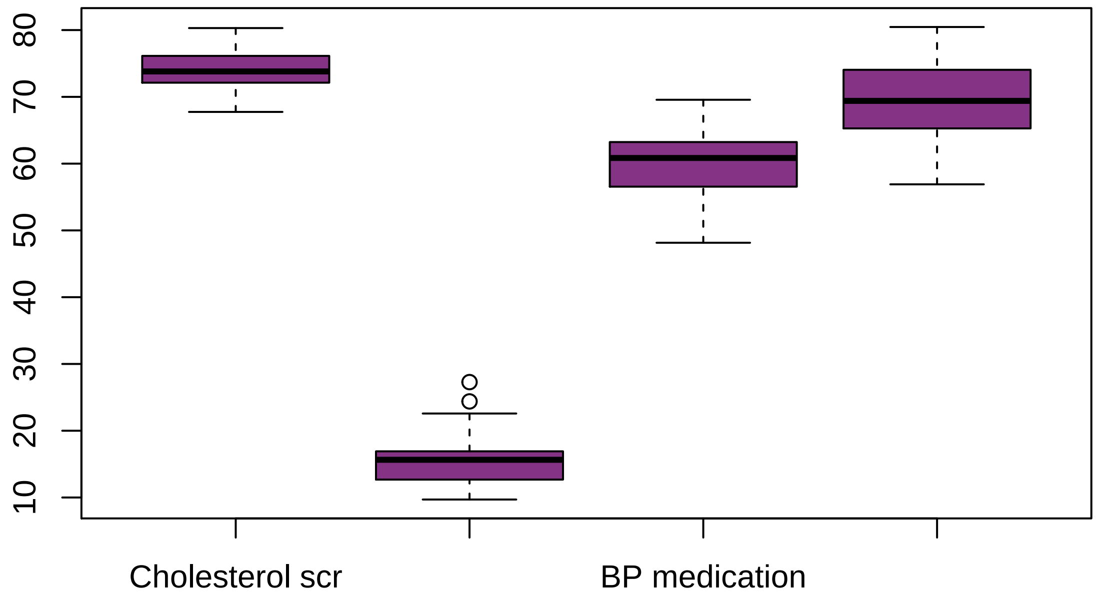


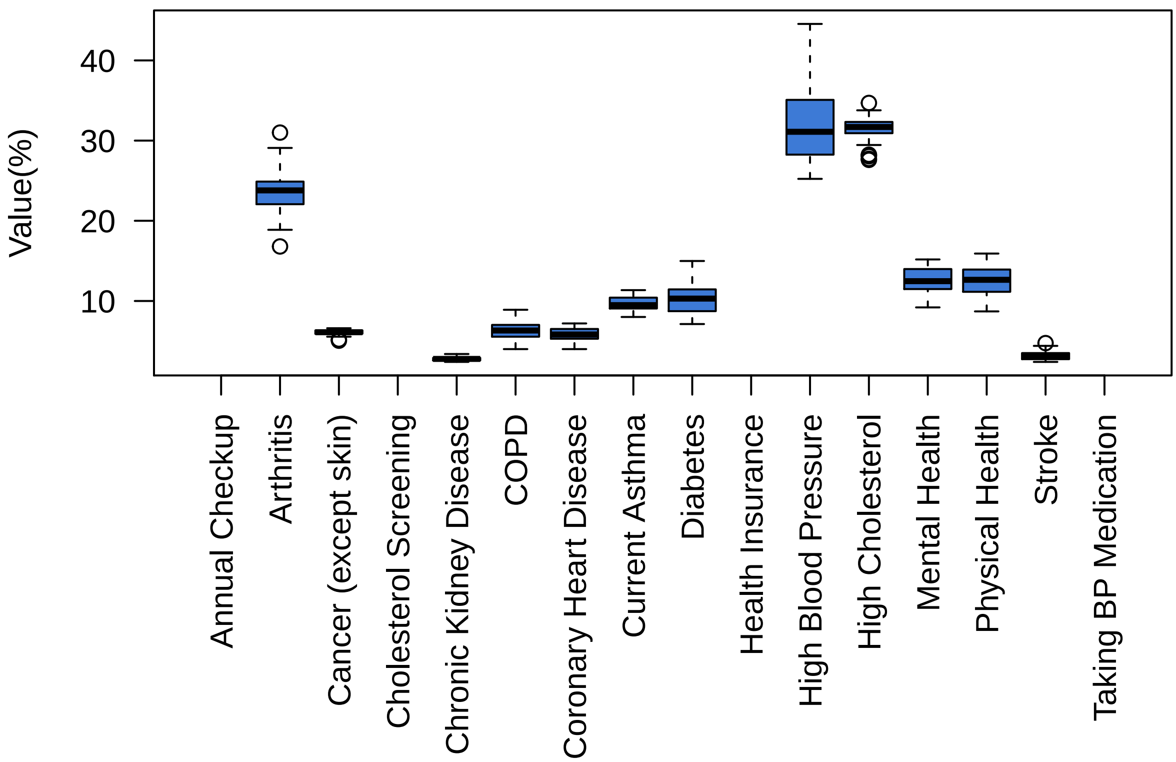

Coronary Heart Disease
Annual Checkup
Cholesterol Screening
COPD
Diabetes
High Cholesterol
Frequency
%


Chronic Kidney Disease
Frequency
0
20
60
100
2.0
3.0
4.0
%
Mental Health
Frequency
0
20
60
8
10
14
18
%
Physical Health
Frequency
0
20
40
60
80
6
8
12
16
20
%
Stroke
Frequency
0
40
80
120
2
3
4
5
6
%

Podaci grupirani po saveznim državama

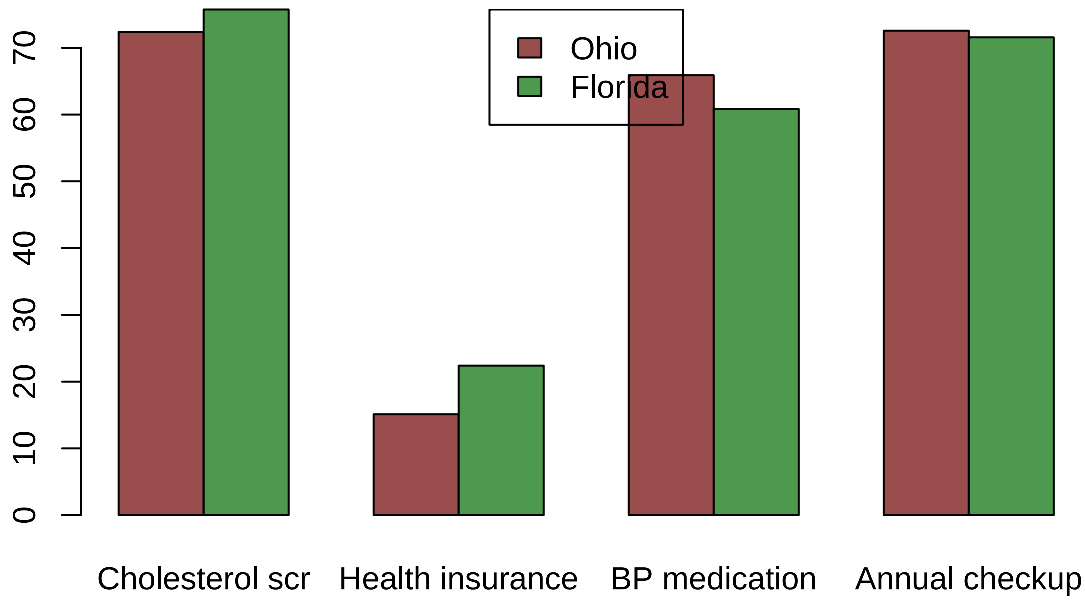
Prikaz raspodjele udjela građana po državama koji primjenjuju pojedine preventivne mjere i koji imaju pojedina zdravstvena stanja:




Value(%)
40
30
20
10
Annual Checkup
Arthritis
Cancer (except skin)
Cholesterol Screening
Chronic Kidney Disease
COPD
Coronary Heart Disease
Current Asthma
Diabetes
Health Insurance
High Blood Pressure
High Cholesterol
Mental Health
Physical Health
Stroke
Taking BP Medication

Statistike - Ohio i Florida

Prikaz udjela stanovnika koji se pridržavaju pojedinih mjera za Ohio i Floridu:



Hi-kvadrat testovi proporcija za Ohio i Floridu:

Prvi test uspoređuje udio cholesterol screening-a u Ohiu i Floridi. Hipoteze: H0 - udjeli su jednaki H1 - udio u Floridi je veći nego udio u Ohiu Dobivamo ekstremno malu p-vrijednost pa možemo odbaciti H0 u korist H1

Drugi test uspoređuje udio heart insurance-a u Ohiu i Floridi. Hipoteze: H0 - udjeli su jednaki H1 - udio u Floridi je veći nego udio u Ohiu Dobivamo ekstremno malu p-vrijednost pa možemo odbaciti H0 u korist H1

Treći test uspoređuje udio Uzimanja lijekova za visoki krvni tlak u Ohiu i Floridi. Hipoteze: H0 - udjeli su jednaki H1 - udio u Ohiu je veći nego udio u Floridi Dobivamo ekstremno malu p-vrijednost pa možemo odbaciti H0 u korist H1

Četvrti test uspoređuje udio godišnjih pregleda u Ohiu i Floridi. Hipoteze: H0 - udjeli su jednaki H1 - udio u Ohiu je veći nego udio u Floridi Dobivamo ekstremno malu p-vrijednost pa možemo odbaciti H0 u korist H1

Zbog velikih uzoraka u hi-kvadrat testu proporcija uvijek ćemo dobiti male p-vrijednosti pa i jako male razlike u proporcijama ispadaju statistički značajne.

```
#Hi-kvadrat testovi proporcije za Ohio i Floridu

res1 <- prop.test(c(Ohio[Ohio$Short_Question_Text == "Cholesterol Screening",]$Population.affected, Flo

res1
```

```
##
##  2-sample test for equality of proportions with
##  continuity correction
##
## data:  c(Ohio[Ohio$Short_Question_Text == "Cholesterol Screening", ]$Population.affected, Florida[Flo
## X-squared = 9463.3, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.03279826
## sample estimates:
##    prop 1    prop 2
## 0.7240165 0.7573880
```

```
res2 <- prop.test(c(Ohio[Ohio$Short_Question_Text == "Health Insurance",]$Population.affected, Florida[F

res2
```

```
##
##  2-sample test for equality of proportions with
##  continuity correction
##
## data:  c(Ohio[Ohio$Short_Question_Text == "Health Insurance", ]$Population.affected, Florida[Florida$
## X-squared = 53176, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.00000000 -0.07247731
## sample estimates:
##    prop 1    prop 2
## 0.1510326 0.2240000
```

```
res3 <- prop.test(c(Ohio[Ohio$Short_Question_Text == "Taking BP Medication",]$Population.affected, Flor

res3
```

```
##
##  2-sample test for equality of proportions with
##  continuity correction
##
## data:  c(Ohio[Ohio$Short_Question_Text == "Taking BP Medication", ]$Population.affected, Florida[Flor
## X-squared = 17389, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.04977949 1.00000000
## sample estimates:
##    prop 1    prop 2
## 0.6588348 0.6084339
```

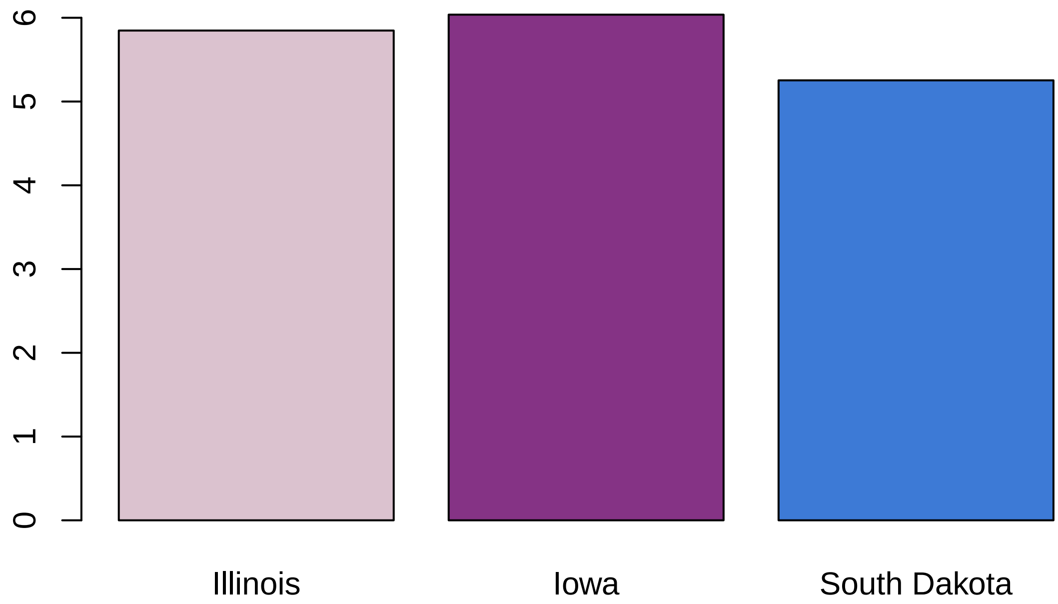
```
res4 <- prop.test(c(Ohio[Ohio$Short_Question_Text == "Annual Checkup",]$Population.affected, Florida[Fl

res4
```

```
##
```

```
##  2-sample test for equality of proportions with
##  continuity correction
##
## data:  c(Ohio[Ohio$Short_Question_Text == "Annual Checkup", ]$Population.affected, Florida[Florida$Sh
## X-squared = 803.75, df = 1, p-value < 2.2e-16
## alternative hypothesis: greater
## 95 percent confidence interval:
##  0.009477296 1.000000000
## sample estimates:
##    prop 1    prop 2
## 0.7256914 0.7156327
```

Statistike - Illinois, Iowa i South Dakota

Prikaz udjela stanovništva koje boluje od kroničnih plućnih bolesti (COPD) u državama Illinois, Iowa i South Dakota:



Hi-kvadrat test za proporcije također smo koristili da pronađemo razlike za COPD u državama Illinois, Iowa i South Dakota. Hipoteze: H0 - udjeli su jednaki H1 - udjeli su različiti Dobili smo malu p-vrijednost pa sukladno tome odbacujemo H0 u korist H1.

Sukladno prijašnjim hi-kvadrat testovima, zbog velikih uzoraka čak i male razlike u proporcijama imaju veliku značajnost.

```
#Hi-kvadrat test proporcije za COPD u odabranim drzavama
res5 <- prop.test(c(Illinois_COPD$Population.affected, Iowa_COPD$Population.affected, S_Dakota_COPD$Popu

res5
```

```
##
##  3-sample test for equality of proportions without
##  continuity correction
##
## data:  c(Illinois_COPD$Population.affected, Iowa_COPD$Population.affected, S_Dakota_COPD$Population.a
## X-squared = 184.77, df = 2, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##     prop 1     prop 2     prop 3
## 0.05847321 0.06037360 0.05253241
```

Utjecaj metoda prevencije na bolesti

Napravimo multivarijantnu linearnu regresiju kako bismo perliminarno vidjeli na koje bolesti naše mjere prevencije imaju značajni učinak. Za svaku bolest odredit ćemo model oblika: Očekivan postotak bolesti = SUM(koeficijent_i* postotak_prevencije_i), na razini čitave države.

```
per_city_data <- health_data_adj %>% group_by(CityName, PopulationCount) %>% summarise(
  checkup = Data_Value[Short_Question_Text == "Annual Checkup"],
  insurance = 100.0 - Data_Value[Short_Question_Text == "Health Insurance"],
  bp_med = Data_Value[Short_Question_Text == "Taking BP Medication"],
  chol_screen = Data_Value[Short_Question_Text == "Cholesterol Screening"],
  arthritis = Data_Value[Short_Question_Text == "Arthritis"],
  cancer_noskin = Data_Value[Short_Question_Text == "Cancer (except skin)"],
  copd = Data_Value[Short_Question_Text == "COPD"],
  coronary_heart_disease = Data_Value[Short_Question_Text == "Coronary Heart Disease"],
  asthma = Data_Value[Short_Question_Text == "Current Asthma"],
  diabetes = Data_Value[Short_Question_Text == "Diabetes"],
  high_bp = Data_Value[Short_Question_Text == "High Blood Pressure"],
  high_col = Data_Value[Short_Question_Text == "High Cholesterol"],
  mental_health = Data_Value[Short_Question_Text == "Mental Health"],
  physical_health = Data_Value[Short_Question_Text == "Physical Health"],
  stroke = Data_Value[Short_Question_Text == "Stroke"],
  ckd = Data_Value[Short_Question_Text == "Chronic Kidney Disease"]
)
```

```
## `summarise()` regrouping output by 'CityName' (override with `.groups` argument)
```

#Utjecaj metoda prevencije na bolesti

Napravimo multivarijantnu linearnu regresiju kako bismo perliminarno vidjeli na koje bolesti naše mjere prevencije imaju značajni učinak.

```
formula <- cbind(arthritis, cancer_noskin, copd, coronary_heart_disease, asthma, diabetes, high_bp, high
fit <- lm(formula, data=per_city_data)
summary(fit)
```

```
## Response arthritis :
##
## Call:
## lm(formula = arthritis ~ checkup + insurance + bp_med + chol_screen,
##     data = per_city_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5926 -1.4670 -0.0341  1.5781  7.9313
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.35705    2.65139   6.924 1.37e-11 ***
## checkup      0.14870    0.04312   3.448 0.000612 ***
## insurance    0.17649    0.02125   8.305 9.61e-16 ***
## bp_med       0.41849    0.03680  11.372  < 2e-16 ***
## chol_screen -0.60788    0.04886 -12.440  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.26 on 495 degrees of freedom
## Multiple R-squared:  0.5788, Adjusted R-squared:  0.5754
## F-statistic:   170 on 4 and 495 DF,  p-value: < 2.2e-16
##
##
## Response cancer_noskin :
##
## Call:
## lm(formula = cancer_noskin ~ checkup + insurance + bp_med + chol_screen,
##     data = per_city_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.33974 -0.19104  0.01975  0.21865  0.72618
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.079909   0.383119   2.819  0.00501 **
## checkup     -0.042077   0.006231  -6.753 4.08e-11 ***
## insurance    0.047660   0.003071  15.521  < 2e-16 ***
## bp_med       0.053188   0.005317  10.003  < 2e-16 ***
## chol_screen  0.009261   0.007061   1.312  0.19023
## ---
## Signif. codes:
```

```
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3266 on 495 degrees of freedom
## Multiple R-squared:  0.4598, Adjusted R-squared:  0.4554
## F-statistic: 105.3 on 4 and 495 DF,  p-value: < 2.2e-16
##
##
## Response copd :
##
## Call:
## lm(formula = copd ~ checkup + insurance + bp_med + chol_screen,
##     data = per_city_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0959 -0.6259  0.0340  0.6273  2.9116
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.615599   1.069040   9.930  < 2e-16 ***
## checkup      0.101451   0.017387   5.835 9.75e-09 ***
## insurance    0.020059   0.008568   2.341   0.0196 *
## bp_med       0.130565   0.014838   8.800  < 2e-16 ***
## chol_screen -0.279126   0.019702 -14.168  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9113 on 495 degrees of freedom
## Multiple R-squared:  0.6337, Adjusted R-squared:  0.6308
## F-statistic: 214.1 on 4 and 495 DF,  p-value: < 2.2e-16
##
##
## Response coronary_heart_disease :
##
## Call:
## lm(formula = coronary_heart_disease ~ checkup + insurance + bp_med +
##     chol_screen, data = per_city_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.11692 -0.28471  0.00739  0.31111  1.43599
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.937874   0.529241  24.446  < 2e-16 ***
## checkup      0.062551   0.008608   7.267 1.44e-12 ***
## insurance   -0.031627   0.004242  -7.456 4.02e-13 ***
## bp_med       0.070402   0.007346   9.584  < 2e-16 ***
## chol_screen -0.173960   0.009754 -17.835  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4511 on 495 degrees of freedom
## Multiple R-squared:  0.7954, Adjusted R-squared:  0.7938
## F-statistic: 481.2 on 4 and 495 DF,  p-value: < 2.2e-16
##
##
## Response asthma :
##
## Call:
## lm(formula = asthma ~ checkup + insurance + bp_med + chol_screen,
##     data = per_city_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4552 -0.4706 -0.0178  0.5055  2.8011
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.892471   1.004087  11.844   <2e-16 ***
## checkup      0.145758   0.016331   8.925   <2e-16 ***
## insurance    0.084690   0.008048  10.524   <2e-16 ***
## bp_med       0.033116   0.013936   2.376   0.0179 *
## chol_screen -0.291332   0.018505 -15.744   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8559 on 495 degrees of freedom
## Multiple R-squared:  0.4648, Adjusted R-squared:  0.4605
## F-statistic: 107.5 on 4 and 495 DF,  p-value: < 2.2e-16
##
##
## Response diabetes :
##
## Call:
## lm(formula = diabetes ~ checkup + insurance + bp_med + chol_screen,
##     data = per_city_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6114 -0.8395 -0.0332  0.7585  4.2460
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.07719    1.31211  22.923   <2e-16 ***
## checkup      0.21770    0.02134  10.201   <2e-16 ***
## insurance   -0.19507    0.01052 -18.549   <2e-16 ***
## bp_med       0.04351    0.01821   2.389   0.0173 *
## chol_screen -0.28114    0.02418 -11.626   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.118 on 495 degrees of freedom
## Multiple R-squared:  0.7965, Adjusted R-squared:  0.7948
```

```
## F-statistic: 484.3 on 4 and 495 DF,  p-value: < 2.2e-16
##
##
## Response high_bp :
##
## Call:
## lm(formula = high_bp ~ checkup + insurance + bp_med + chol_screen,
##     data = per_city_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5478 -1.5775 -0.1655  1.4384  7.3778
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.33484    2.78267   8.745  < 2e-16 ***
## checkup      0.25948    0.04526   5.733 1.72e-08 ***
## insurance   -0.11260    0.02230  -5.049 6.27e-07 ***
## bp_med       0.50406    0.03862  13.051  < 2e-16 ***
## chol_screen -0.42402    0.05128  -8.268 1.26e-15 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.372 on 495 degrees of freedom
## Multiple R-squared:  0.7453, Adjusted R-squared:  0.7432
## F-statistic: 362.1 on 4 and 495 DF,  p-value: < 2.2e-16
##
##
## Response high_col :
##
## Call:
## lm(formula = high_col ~ checkup + insurance + bp_med + chol_screen,
##     data = per_city_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8745 -0.7365  0.0184  0.8569  4.0089
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.28749    1.60560  24.469  < 2e-16 ***
## checkup      0.02267    0.02611   0.868    0.386
## insurance   -0.10849    0.01287  -8.431 3.77e-16 ***
## bp_med       0.17479    0.02228   7.844 2.71e-14 ***
## chol_screen -0.14235    0.02959  -4.811 2.00e-06 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.369 on 495 degrees of freedom
## Multiple R-squared:  0.5589, Adjusted R-squared:  0.5553
## F-statistic: 156.8 on 4 and 495 DF,  p-value: < 2.2e-16
##
```

```
##
## Response mental_health :
##
## Call:
## lm(formula = mental_health ~ checkup + insurance + bp_med + chol_screen,
##     data = per_city_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5016 -0.9224  0.1401  1.0034  3.7979
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.93820    1.67705  18.448  < 2e-16 ***
## checkup      0.25760    0.02728   9.444  < 2e-16 ***
## insurance   -0.04115    0.01344  -3.062  0.00232 **
## bp_med      -0.01598    0.02328  -0.687  0.49271
## chol_screen -0.42645    0.03091 -13.798  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.43 on 495 degrees of freedom
## Multiple R-squared:  0.5479, Adjusted R-squared:  0.5443
## F-statistic:   150 on 4 and 495 DF,  p-value: < 2.2e-16
##
##
## Response physical_health :
##
## Call:
## lm(formula = physical_health ~ checkup + insurance + bp_med +
##     chol_screen, data = per_city_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1254 -0.8191  0.1541  0.9010  4.0753
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 44.22260    1.62654  27.188   <2e-16 ***
## checkup      0.26819    0.02645  10.137   <2e-16 ***
## insurance   -0.13749    0.01304 -10.547   <2e-16 ***
## bp_med      -0.02359    0.02258  -1.045    0.296
## chol_screen -0.49886    0.02998 -16.642   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.387 on 495 degrees of freedom
## Multiple R-squared:  0.7223, Adjusted R-squared:  0.7201
## F-statistic: 321.9 on 4 and 495 DF,  p-value: < 2.2e-16
##
##
## Response stroke :
```

```
##
## Call:
## lm(formula = stroke ~ checkup + insurance + bp_med + chol_screen,
##     data = per_city_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04859 -0.27732  0.01754  0.23382  1.82329
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.821149   0.476669  14.310  < 2e-16 ***
## checkup      0.054646   0.007753   7.049 6.10e-12 ***
## insurance   -0.016936   0.003820  -4.433 1.15e-05 ***
## bp_med       0.049533   0.006616   7.487 3.25e-13 ***
## chol_screen -0.120620   0.008785 -13.731  < 2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4063 on 495 degrees of freedom
## Multiple R-squared:    0.7,  Adjusted R-squared:  0.6976
## F-statistic: 288.8 on 4 and 495 DF,  p-value: < 2.2e-16
##
##
## Response ckd :
##
## Call:
## lm(formula = ckd ~ checkup + insurance + bp_med + chol_screen,
##     data = per_city_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48607 -0.13366  0.00186  0.12571  0.82723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.718127   0.229579  33.619   <2e-16 ***
## checkup      0.037444   0.003734  10.028   <2e-16 ***
## insurance   -0.026724   0.001840 -14.523   <2e-16 ***
## bp_med       0.002156   0.003186   0.677    0.499
## chol_screen -0.072804   0.004231 -17.207   <2e-16 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1957 on 495 degrees of freedom
## Multiple R-squared:  0.7868, Adjusted R-squared:  0.785
## F-statistic: 456.6 on 4 and 495 DF,  p-value: < 2.2e-16
```

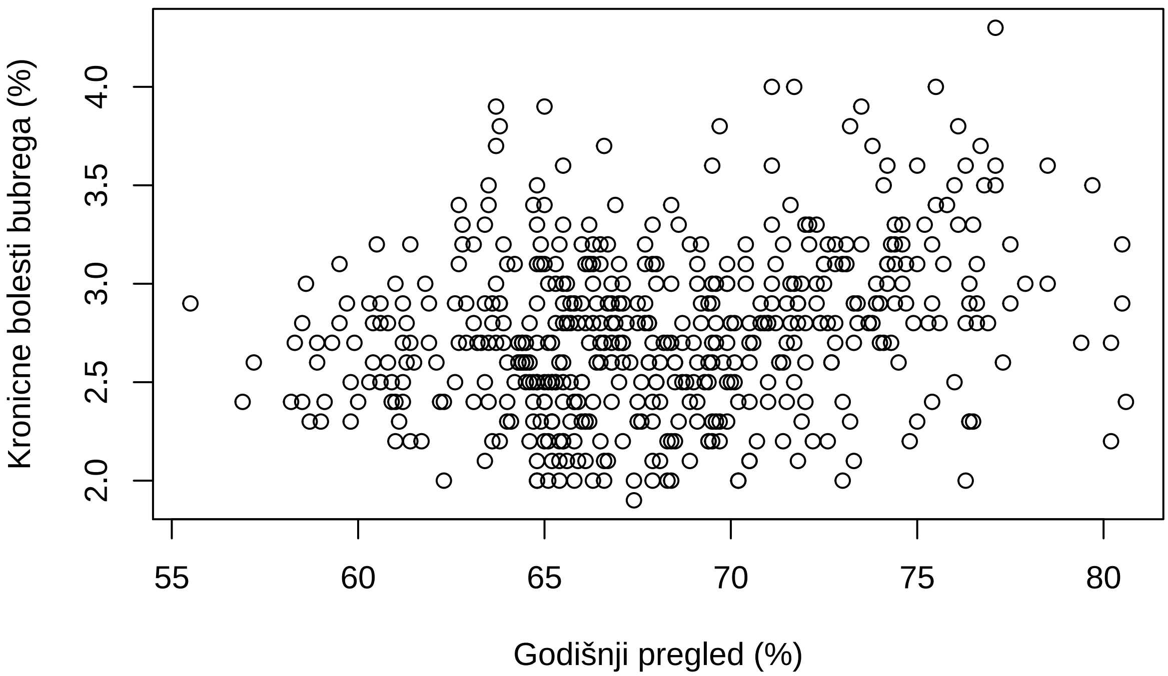
Rezultati kronične bubrežne bolesti ističu se kao zanimljivi jer ih relativno dobro predviđamo linearnom regresijom, a također čini se kao da je jedan regresor nepotreban.

Kronične Bubrežne bolesti

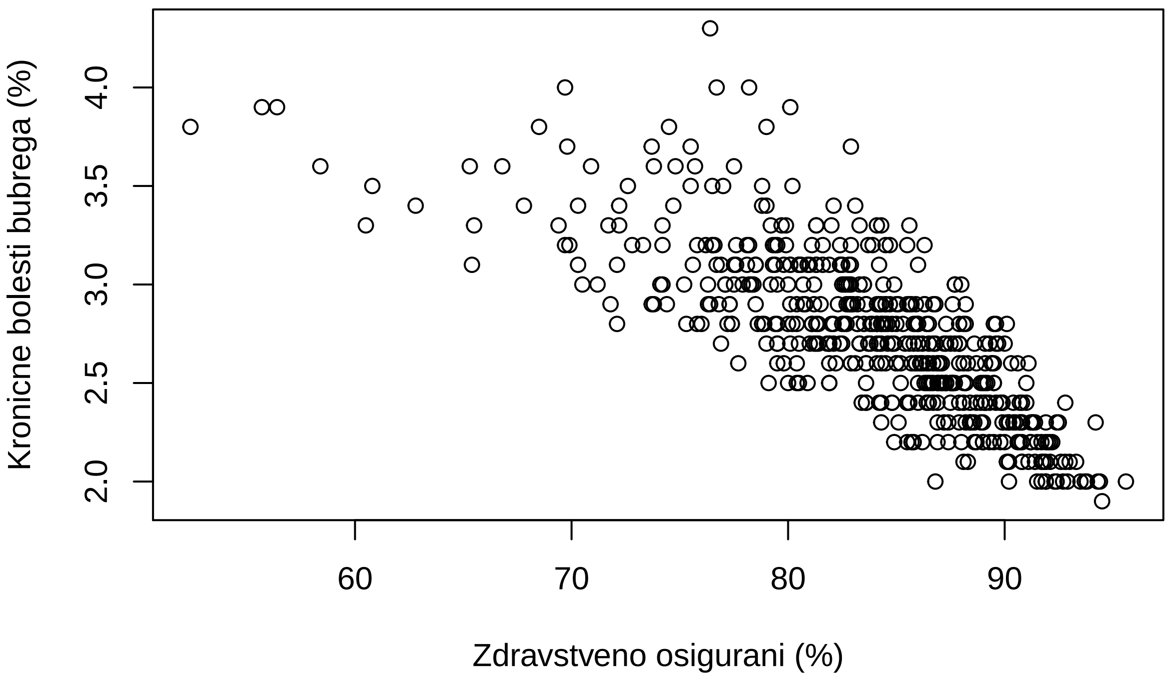
U ovom potpoglavlju istražit ćemo vezu između ove četiri mjere prevencije i kroničnih bubrežnih bolesti (KBB). Tu vezu pokušat ćemo objasniti metodom linearne regresije, koju ćemo obaviti na razini cijele države.

Prvo pogledajmo grafove koje prikazuju pojedinačne veze između metode prevencija i KBB, na sljedećim grafovima svaka točka predstavlja jedan grad.

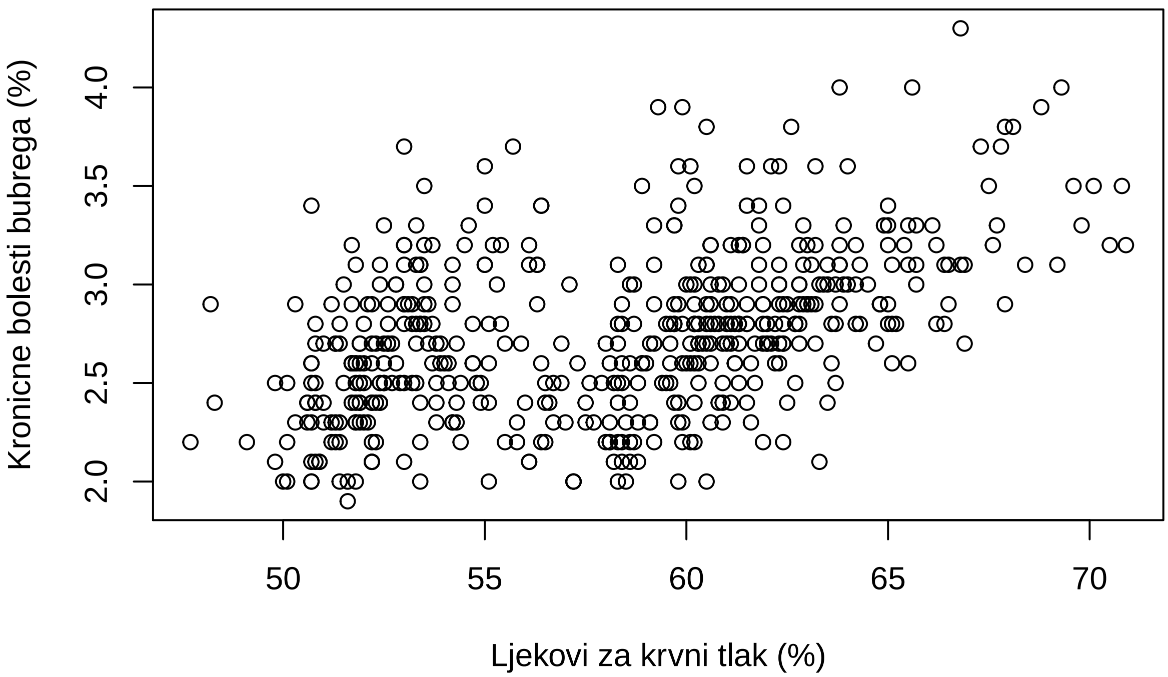
```
#Chronic Kidney Disease data
CKD_data <- per_city_data[, c("checkup", "insurance", "bp_med", "chol_screen", "ckd", "CityName", "Popul

plot(CKD_data$checkup, CKD_data$ckd, xlab="Godišnji pregled (%)", ylab="Kronične bolesti bubrega (%)")
```



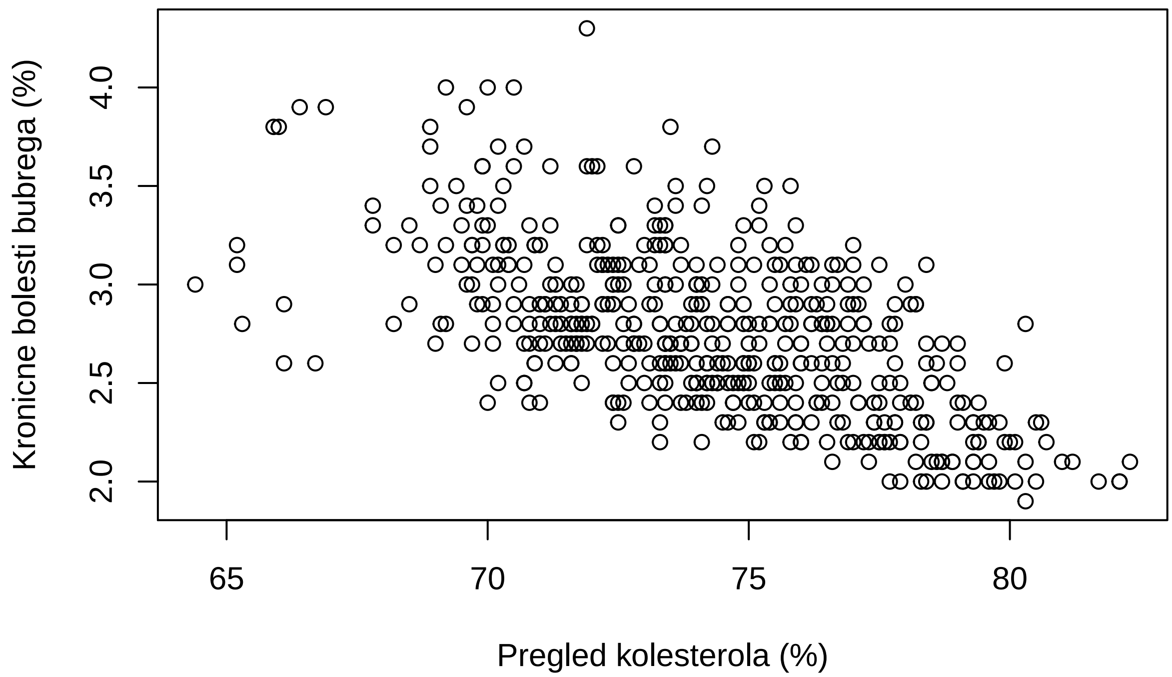

```
plot(CKD_data$insurance, CKD_data$ckd, xlab="Zdravstveno osigurani (%)", ylab="Kronične bolesti bubrega
```




```
plot(CKD_data$bp_med, CKD_data$ckd, xlab="Ljekovi za krvni tlak (%)", ylab="Kronične bolesti bubrega (%
```




```
plot(CKD_data$chol_screen, CKD_data$ckd, xlab="Pregled kolesterola (%)", ylab="Kronične bolesti bubrega
```



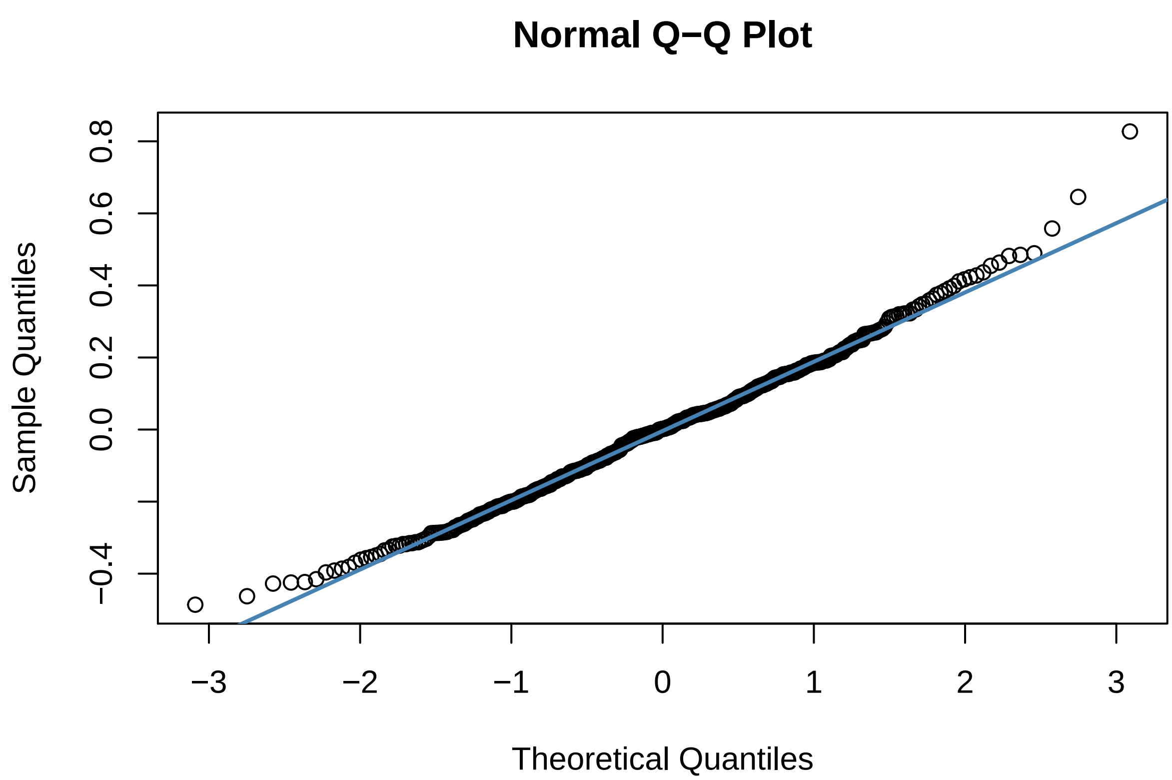
Primjećujemo da postoji jak utjecaj zdravstvenog osiguranja te učestalosti testiranja kolesterola na KBB, no grafovi su previše raspršeni da bi ijedan od njih u potpunosti objasnio fenomen. Iz grafova godišnjih pregleda i uzimanja ljekova za krvni tlak ne možemo previše zaključiti.

```
fit1 <- lm(CKD_data$ckd ~ CKD_data$insurance + CKD_data$chol_screen + CKD_data$checkup + CKD_data$bp_med
summary(fit1)
```

```
##
## Call:
## lm(formula = CKD_data$ckd ~ CKD_data$insurance + CKD_data$chol_screen +
##     CKD_data$checkup + CKD_data$bp_med)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48607 -0.13366  0.00186  0.12571  0.82723
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           7.718127   0.229579  33.619   <2e-16
## CKD_data$insurance   -0.026724   0.001840 -14.523   <2e-16
## CKD_data$chol_screen -0.072804   0.004231 -17.207   <2e-16
## CKD_data$checkup      0.037444   0.003734  10.028   <2e-16
## CKD_data$bp_med       0.002156   0.003186   0.677    0.499
##
## (Intercept)          ***
## CKD_data$insurance   ***
```

```
## CKD_data$chol_screen ***
## CKD_data$checkup     ***
## CKD_data$bp_med
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1957 on 495 degrees of freedom
## Multiple R-squared:  0.7868, Adjusted R-squared:  0.785
## F-statistic: 456.6 on 4 and 495 DF,  p-value: < 2.2e-16
```

```
qqnorm(fit1$residuals)
qqline(fit1$residuals, col = "steelblue", lwd = 2)
```



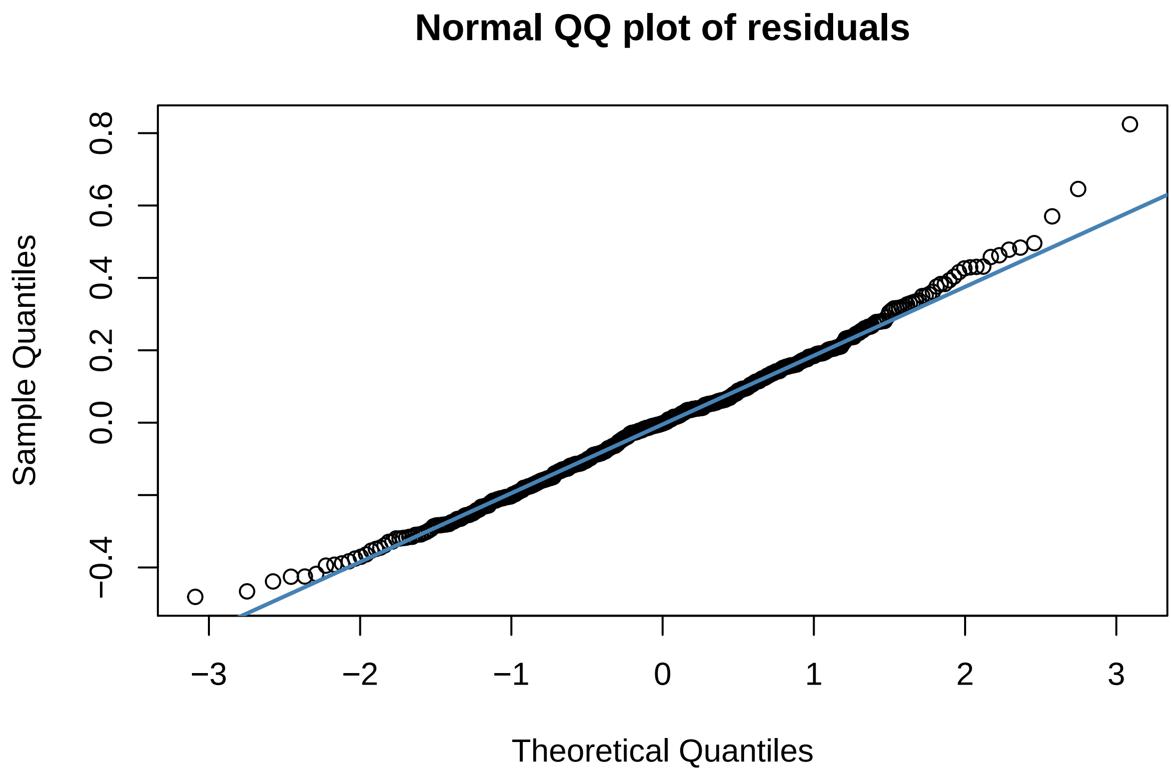
Reziduali regresije prate normalnu distribuciju dovoljno dobro da možemo opravdati pretpostavku normalnosti greške. U zadnjem stupcu rezultata regresije "Pr(>|t|)", za svaki parametar možemo vidjeti p-vrijednost testa o regresijskim koeficijentima. Iz tog stupca možemo očitati da su faktori zdravstvenog osiguranja, pregleda kolesterola, te godišnjih pregleda signifikatni čak i pri jako malim vrijednostima alfa. Isto ne možemo reći i za utjecaj uzimanja ljekova za krvni tlak, tako da ćemo regresiju provesti još jednom, ali ćemo izbaciti taj regresor.

```
fit2 <- lm(CKD_data$ckd ~ CKD_data$insurance + CKD_data$chol_screen + CKD_data$checkup)
summary(fit2)
```

```
##
## Call:
```

```
## lm(formula = CKD_data$ckd ~ CKD_data$insurance + CKD_data$chol_screen +
##     CKD_data$checkup)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48121 -0.13286 -0.00182  0.12352  0.82450
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             7.788684   0.204415   38.10   <2e-16
## CKD_data$insurance     -0.026816   0.001834  -14.62   <2e-16
## CKD_data$chol_screen   -0.073831   0.003947  -18.70   <2e-16
## CKD_data$checkup        0.039481   0.002208   17.89   <2e-16
##
## (Intercept)            ***
## CKD_data$insurance     ***
## CKD_data$chol_screen   ***
## CKD_data$checkup       ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1956 on 496 degrees of freedom
## Multiple R-squared:  0.7866, Adjusted R-squared:  0.7853
## F-statistic: 609.3 on 3 and 496 DF,  p-value: < 2.2e-16
```

```r
qqnorm(fit2$residuals, main="Normal QQ plot of residuals")
qqline(fit2$residuals, col = "steelblue", lwd = 2)
```



Konačni model linearne regresije objašnjava 78.66% varijacije KBB. Rezultati testova o regresijskim koeficijentima kao i u prošlom primjeru javlja jako male p-vrijednosti što nam omogućuje da odbacimo hipotezu da je neki od koeficijenata zapravo jednak nuli. Normalni qq graf reziduala opravdava pretpostavku normalnosti pogreške.

Iz iznosa koeficijenata regresije možemo zaključiti da veće stope zdravstvene osiguranosti te pregleda kolesterola imaju poželjan utjecaj na postotak kroničnih bubrežnih bolesti. Te od ova dva faktora, pregled kolesterola možemo izdvojiti kao značajnijeg u suzbijanju kroničnih bubrežnih bolesti. Iznenađujuć rezultat ove analize je činjenica da godišnji pregledi naizgled imaju negativan utjecaj na kronične bolesti bubrega, to jest postoji trend da u populacijama u kojima više ljudi ide na godišnje pregleda ima i više kroničnih bubrežnih bolesti. Ta činjenica bi se mogla objasniti trećom skrivenom varijablom, koja utječe na obje varijable. Na primjer moguće je da u gradovima sa starijim stanovništvom ljudi više oboljevaju od bolesti, ali iz istog razloga češće idu na preglede. Ovu hipotezu ipak ne možemo istražiti jer nemamo podatke o starosti stanovništva.