

# Projekt

Stella Balić, Jan Kolić, Marko Peroš, Jana Perak

2023-01-06

## Uvod

U ovom dokumentu bavit ćemo se analizom potrošnje kućanstava. Ova analiza koristi podatke sakupljene na temelju ankete koju je proveo španjolski Institut za statistiku (Instituto Nacional de Estadística ili INE) u svrhu uskladivanja životnog standarda u državi. Koristit ćemo skup podataka iz 2019. i 2020. godine.

## Učitavanje

Učitajmo potrebne pakete.

```
library(dplyr)
```

Učitajmo podatke.

```
kucanstva19 = read.csv("datasets/hogar_epf_2019.csv", header = TRUE)
kucanstva20 = read.csv("datasets/hogar_epf_2020.csv", header = TRUE)
#potrosnja19 = read.csv("datasets/gastos_epf_2019.csv", header = TRUE)
#potrosnja20 = read.csv("datasets/gastos_epf_2020.csv", header = TRUE)
dim(kucanstva19)
```

```
## [1] 20817    223
```

```
dim(kucanstva20)
```

```
## [1] 19170    223
```

```
#dim(potrosnja19)
#dim(potrosnja20)
```

### 1. Postoji li zavisnost izmedu broja članova kućanstva i energenta koji se upotrebljava za dobivanje tople vode?

Varijabla NMIEMB govori o broju članova kućanstva, a varijabla FUENAGUA govori o energentu koje kućanstvo upotrebljava za grijanje vode. Slijedeći opis varijabli, NMIEMB može poprimiti vrijednosti 0-19, a FUENAGUA: 1 Electricity, 2 Natural gas, 3 Liquefied gas, 4 Other liquid fuels, 5 Solid fuels, 6 Others, b Not applicable, -9 Not stated. Iz ovoga slijedi da ćemo promatrati samo one retke kojima je varijabla FUENAGUA iz skupa {1, 2, 3, 4, 5, 6}. Koristit ćemo  $\chi^2$  test provjere nezavisnosti.

Prvo ćemo provjeriti koje su sve moguće vrijednosti varijabli u našim podacima.

```
unique(kucanstva19$FUENAGUA)
```

```
## [1] 3 2 1 5 4 6 NA 9
```

```
unique(kucanstva19$NMIEMB)
```

```
## [1] 3 1 2 4 5 8 0 7 6 9
```

Vrijednosti varijable NMIEMB nam odgovaraju, a iz varijable FUENAGUA ćemo uzeti samo one koje označavaju energet.

```
kucanstva19_final = subset(kucanstva19, FUENAGUA %in% c(1,2,3,4,5,6))  
dim(kucanstva19_final)
```

```
## [1] 20772 223
```

Možemo prikazati kontigencijsku tablicu i vizualizirati kako su raspoređene vrijednosti i što bismo mogli pretpostaviti što će test pokazati.

```
f_tbl = table(kucanstva19_final$NMIEMB, kucanstva19_final$FUENAGUA) #tablica frekvencija  
f_tbl
```

```
##  
##      1     2     3     4     5     6  
##  0    0     1     2     0     0     0  
##  1 1187 1609 889 406 20 27  
##  2 1712 2804 1371 840 46 75  
##  3 1121 1943 990 468 30 65  
##  4  893 1854 818 386 28 66  
##  5   213 298 209 82 6 15  
##  6    61 58 69 16 4 0  
##  7    25 20 20 2 1 2  
##  8     4 2 5 0 1 0  
##  9     3 3 2 0 0 0
```

```
# Radimo histograme:
```

```
b = seq(min(kucanstva19_final$NMIEMB)-1,max(kucanstva19_final$NMIEMB) , 1)  
  
h1 = hist(kucanstva19_final[kucanstva19_final[ "FUENAGUA" ] == 1,]$NMIEMB,  
          breaks=b,  
          plot=FALSE)  
h2 = hist(kucanstva19_final[kucanstva19_final[ "FUENAGUA" ] == 2,]$NMIEMB,  
          breaks=b,  
          plot=FALSE)  
h3 = hist(kucanstva19_final[kucanstva19_final[ "FUENAGUA" ] == 3,]$NMIEMB,  
          breaks=b,  
          plot=FALSE)  
h4 = hist(kucanstva19_final[kucanstva19_final[ "FUENAGUA" ] == 4,]$NMIEMB,  
          breaks=b,  
          plot=FALSE)
```

```

h5 = hist(kucanstva19_final[kucanstva19_final["FUENAGUA"] == 5,]$NMIEMB,
          breaks=b,
          plot=FALSE)
h6 = hist(kucanstva19_final[kucanstva19_final["FUENAGUA"] == 6,]$NMIEMB,
          breaks=b,
          plot=FALSE)

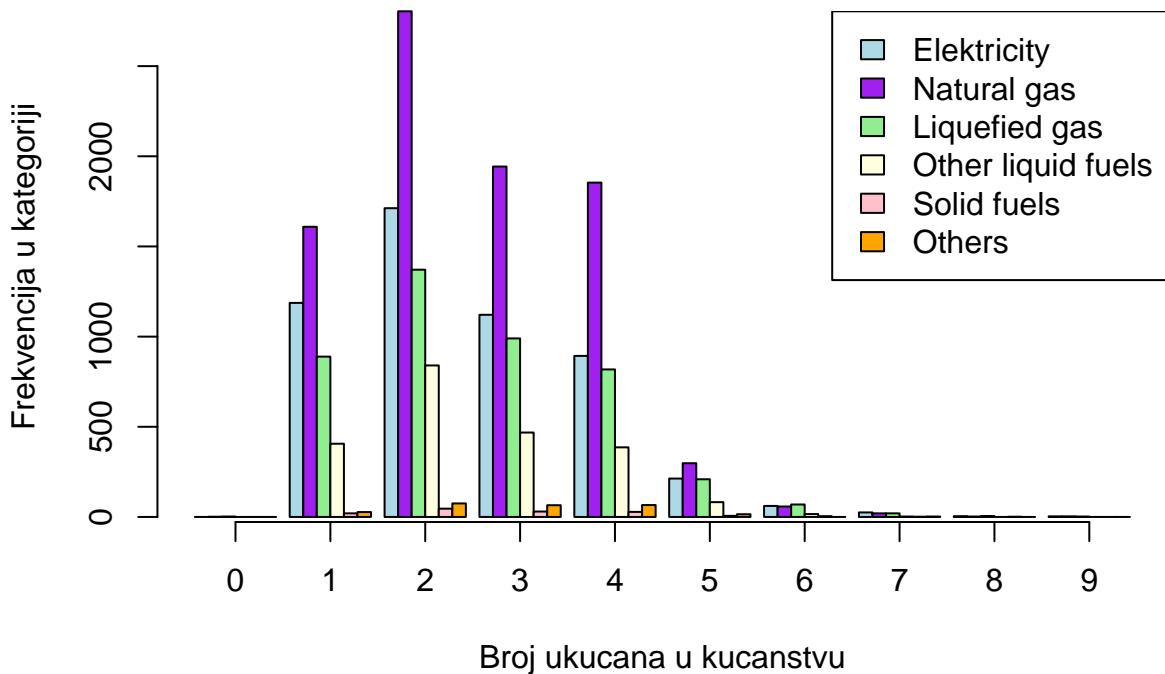
data <- t(cbind(h1$counts,h2$counts,h3$counts,h4$counts,h5$counts,h6$counts))

#Njihove counts koristimo za barplot
barplot(data,beside=TRUE, col=c("lightblue", "purple", "lightgreen", "lightyellow", "pink", "orange" ),

a = seq(0,9, 1)
c=seq(4,67,7)
axis(1, at = c, labels = a)

legend("topright",c("Elektricity","Natural gas","Liquefied gas", "Other liquid fuels", "Solid fuels", "Others"))

```



Sada ćemo krenuti s testom za što prvo trebamo provjeriti vrijednosti očekivanih frekvencija, pretpostavka testa je da su sve veće od 5. U slučaju da nisu, vršit ćemo grupaciju.

Kontigencijska tablica

```

c_table = addmargins(f_tbl)
c_table #kontigencijska tablica

```

```

##          1   2   3   4   5   6   Sum
## 0      0   1   2   0   0   0    3
## 1  1187 1609 889 406 20  27 4138
## 2  1712 2804 1371 840 46  75 6848
## 3  1121 1943 990 468 30  65 4617
## 4    893 1854 818 386 28  66 4045
## 5   213  298 209  82  6  15 823
## 6    61   58  69  16  4   0 208
## 7    25   20  20   2  1   2  70
## 8     4    2   5   0  1   0  12
## 9     3    3   2   0  0   0   8
## Sum 5219 8592 4375 2200 136  250 20772

```

Provjera očekivanih frekvencija

```

counter=0
for (col_names in colnames(c_table)){
  for (row_names in rownames(c_table)){
    if (!(row_names == 'Sum' | col_names == 'Sum')){
      cat('Očekivane frekvencije za razred ',col_names,' - ',row_names,': ',(c_table[row_names,'Sum'] * c_table['Sum',col_names]) / c_table['Sum','Sum'])
      if (c_table[row_names,'Sum'] * c_table['Sum',col_names] / c_table['Sum','Sum'] <=5){
        counter=counter+1
      }
    }
  }
}
}

## Očekivane frekvencije za razred 1 - 0 : 0.7537551
## Očekivane frekvencije za razred 1 - 1 : 1039.679
## Očekivane frekvencije za razred 1 - 2 : 1720.572
## Očekivane frekvencije za razred 1 - 3 : 1160.029
## Očekivane frekvencije za razred 1 - 4 : 1016.313
## Očekivane frekvencije za razred 1 - 5 : 206.7801
## Očekivane frekvencije za razred 1 - 6 : 52.26035
## Očekivane frekvencije za razred 1 - 7 : 17.58762
## Očekivane frekvencije za razred 1 - 8 : 3.01502
## Očekivane frekvencije za razred 1 - 9 : 2.010013
## Očekivane frekvencije za razred 2 - 0 : 1.240901
## Očekivane frekvencije za razred 2 - 1 : 1711.616
## Očekivane frekvencije za razred 2 - 2 : 2832.564
## Očekivane frekvencije za razred 2 - 3 : 1909.747
## Očekivane frekvencije za razred 2 - 4 : 1673.148
## Očekivane frekvencije za razred 2 - 5 : 340.4206
## Očekivane frekvencije za razred 2 - 6 : 86.03582
## Očekivane frekvencije za razred 2 - 7 : 28.95436
## Očekivane frekvencije za razred 2 - 8 : 4.963605
## Očekivane frekvencije za razred 2 - 9 : 3.30907
## Očekivane frekvencije za razred 3 - 0 : 0.6318602
## Očekivane frekvencije za razred 3 - 1 : 871.5458
## Očekivane frekvencije za razred 3 - 2 : 1442.326
## Očekivane frekvencije za razred 3 - 3 : 972.4328
## Očekivane frekvencije za razred 3 - 4 : 851.9582
## Očekivane frekvencije za razred 3 - 5 : 173.3403

```

```

## Očekivane frekvencije za razred 3 - 6 : 43.80897
## Očekivane frekvencije za razred 3 - 7 : 14.7434
## Očekivane frekvencije za razred 3 - 8 : 2.527441
## Očekivane frekvencije za razred 3 - 9 : 1.684961
## Očekivane frekvencije za razred 4 - 0 : 0.3177354
## Očekivane frekvencije za razred 4 - 1 : 438.263
## Očekivane frekvencije za razred 4 - 2 : 725.284
## Očekivane frekvencije za razred 4 - 3 : 488.9948
## Očekivane frekvencije za razred 4 - 4 : 428.4132
## Očekivane frekvencije za razred 4 - 5 : 87.16541
## Očekivane frekvencije za razred 4 - 6 : 22.02966
## Očekivane frekvencije za razred 4 - 7 : 7.413826
## Očekivane frekvencije za razred 4 - 8 : 1.270942
## Očekivane frekvencije za razred 4 - 9 : 0.8472944
## Očekivane frekvencije za razred 5 - 0 : 0.01964183
## Očekivane frekvencije za razred 5 - 1 : 27.09262
## Očekivane frekvencije za razred 5 - 2 : 44.83574
## Očekivane frekvencije za razred 5 - 3 : 30.22877
## Očekivane frekvencije za razred 5 - 4 : 26.48373
## Očekivane frekvencije za razred 5 - 5 : 5.388407
## Očekivane frekvencije za razred 5 - 6 : 1.361833
## Očekivane frekvencije za razred 5 - 7 : 0.4583093
## Očekivane frekvencije za razred 5 - 8 : 0.0785673
## Očekivane frekvencije za razred 5 - 9 : 0.0523782
## Očekivane frekvencije za razred 6 - 0 : 0.0361063
## Očekivane frekvencije za razred 6 - 1 : 49.80262
## Očekivane frekvencije za razred 6 - 2 : 82.41864
## Očekivane frekvencije za razred 6 - 3 : 55.56759
## Očekivane frekvencije za razred 6 - 4 : 48.68332
## Očekivane frekvencije za razred 6 - 5 : 9.905161
## Očekivane frekvencije za razred 6 - 6 : 2.50337
## Očekivane frekvencije za razred 6 - 7 : 0.8424803
## Očekivane frekvencije za razred 6 - 8 : 0.1444252
## Očekivane frekvencije za razred 6 - 9 : 0.09628346

```

```
cat("\nBroj očekivanih frekvencija <= 5 je ", counter)
```

```
##
## Broj očekivanih frekvencija <= 5 je 22
```

Uočavamo da imamo više očekivanih frekvencija manjih od 5 pa ćemo broj ukućana grupirati.

Napravit ćemo dvije grupacije. Prvo ćemo grupirat u grupe 0-2 (kućanstvo s malim brojem ukućana), 3-4 (kućanstvo sa srednjim brojem ukućana) i 5-9 (kućanstvo s velikim brojem ukućana). Zatim u grupe 0-1, 2, 3, 4-9 kako bismo pokušali dobiti što sličniji broj kućanstava u svakoj grupi. Ovo radimo kako bismo pokazali da način grupiranja ne utječe na konačan rezultat testa.

Kako bi izgledao prikaz s bar plotom?

```
# Grupirajmo na 1. nacin
kucanstva19_final$NMIEMB_group2 <- cut(kucanstva19_final$NMIEMB, breaks = c(0, 2, 4, 9), labels = c("0-2",
"3-4", "5-9"))

c_table2 <- table(kucanstva19_final$FUENAGUA, kucanstva19_final$NMIEMB_group2)

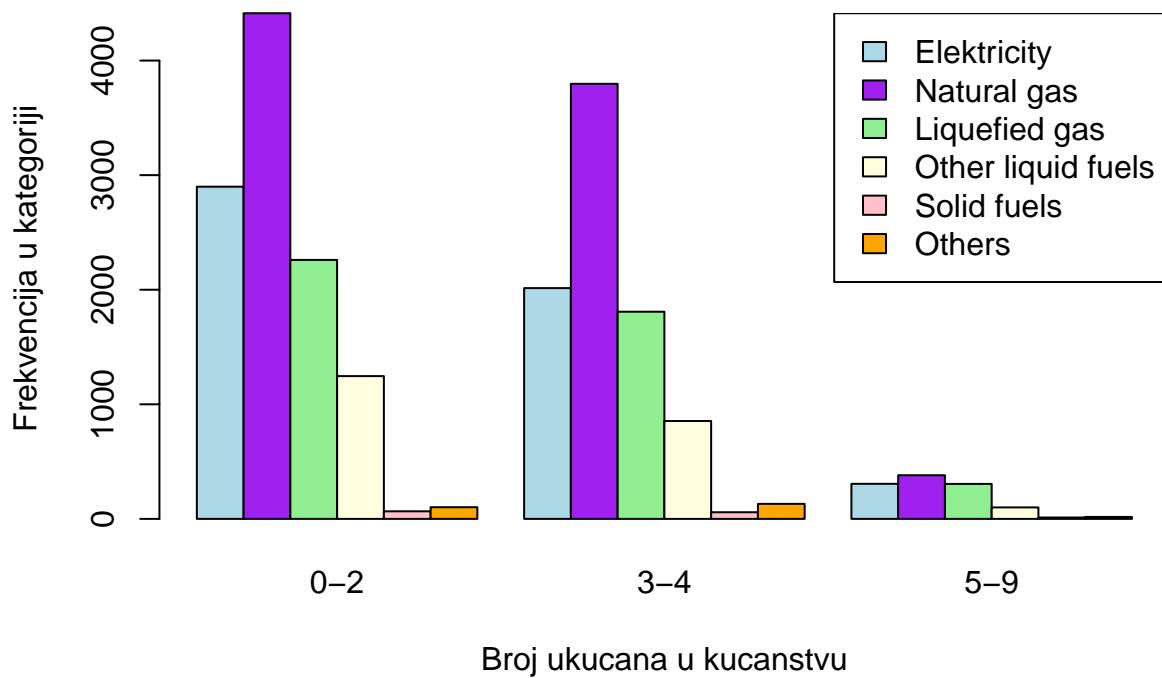
c_table2
```

```

##          0-2   3-4   5-9
## 1 2899 2014 306
## 2 4413 3797 381
## 3 2260 1808 305
## 4 1246  854 100
## 5   66   58  12
## 6  102  131  17

barplot(c_table2, beside = TRUE, xlab = "Broj ukućana u kućanstvu", ylab = 'Frekvencija u kategoriji', c
legend("topright", c("Elektricity", "Natural gas", "Liquefied gas", "Other liquid fuels", "Solid fuels"

```



```

# Grupirajmo na 2. nacin
kucanstva19_final$NMIEMB_group3 <- cut(kucanstva19_final$NMIEMB, breaks = c(0, 1, 2, 3, 9), labels = c(
c_table3 <- table(kucanstva19_final$FUEENAGUA, kucanstva19_final$NMIEMB_group3)

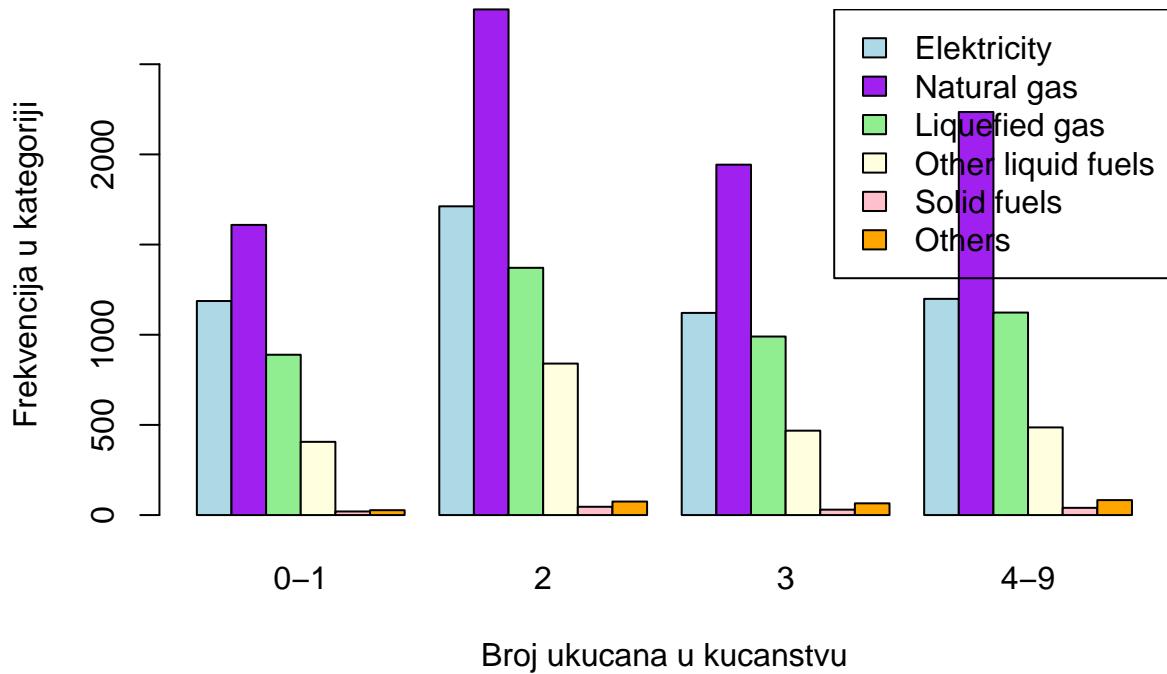
c_table3

##          0-1      2      3    4-9
## 1 1187 1712 1121 1199
## 2 1609 2804 1943 2235
## 3  889 1371  990 1123
## 4   406   840   468   486

```

```
##   5   20   46   30   40
##   6   27   75   65   83
```

```
barplot(c_table3, beside = TRUE, xlab = "Broj ukućana u kućanstvu", ylab = 'Frekvencija u kategoriji', c
legend("topright", c("Elektricity", "Natural gas", "Liquefied gas", "Other liquid fuels", "Solid fuels"
"Others"))
```



```
c_table_2 <- table(kucanstva19_final$NMIEMB_group2, kucanstva19_final$FUENAGUA)
c_table_3 <- table(kucanstva19_final$NMIEMB_group3, kucanstva19_final$FUENAGUA)
```

Kontigencijska tablica za prvu grupaciju:

```
c_table2 = addmargins(c_table_2)
c_table2 #kontigencijska tablica
```

```
##
##          1      2      3      4      5      6    Sum
## 0-2  2899  4413  2260  1246   66  102 10986
## 3-4  2014  3797  1808   854   58  131  8662
## 5-9   306   381   305   100   12   17  1121
## Sum   5219  8591  4373  2200  136  250 20769
```

Kontigencijska tablica za drugu grupaciju:

```
c_table3 = addmargins(c_table_3)
c_table3 #kontigencijska tablica
```

```
##
##          1     2     3     4     5     6   Sum
## 0-1  1187  1609   889   406   20   27  4138
## 2    1712  2804  1371   840   46   75  6848
## 3    1121  1943   990   468   30   65  4617
## 4-9  1199  2235  1123   486   40   83  5166
## Sum  5219  8591  4373  2200  136   250 20769
```

Provjera očekivanih frekvencija za prvu grupaciju:

```
counter = 0
for (col_names in colnames(c_table2)){
  for (row_names in rownames(c_table2)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za razred ',col_names,' - ',row_names,': ',(c_table2[row_names,'Sum'] * c_table2['Sum',col_names] / c_table2['Sum','Sum'])<=5){
        counter=counter+1
      }
    }
  }
}

## Očekivane frekvencije za razred 1 - 0-2 : 2760.65
## Očekivane frekvencije za razred 1 - 3-4 : 2176.656
## Očekivane frekvencije za razred 1 - 5-9 : 281.6938
## Očekivane frekvencije za razred 2 - 0-2 : 4544.308
## Očekivane frekvencije za razred 2 - 3-4 : 3582.996
## Očekivane frekvencije za razred 2 - 5-9 : 463.6964
## Očekivane frekvencije za razred 3 - 0-2 : 2313.148
## Očekivane frekvencije za razred 3 - 3-4 : 1823.82
## Očekivane frekvencije za razred 3 - 5-9 : 236.0312
## Očekivane frekvencije za razred 4 - 0-2 : 1163.715
## Očekivane frekvencije za razred 4 - 3-4 : 917.5406
## Očekivane frekvencije za razred 4 - 5-9 : 118.7443
## Očekivane frekvencije za razred 5 - 0-2 : 71.93875
## Očekivane frekvencije za razred 5 - 3-4 : 56.72069
## Očekivane frekvencije za razred 5 - 5-9 : 7.340556
## Očekivane frekvencije za razred 6 - 0-2 : 132.2404
## Očekivane frekvencije za razred 6 - 3-4 : 104.266
## Očekivane frekvencije za razred 6 - 5-9 : 13.49367

cat("\nBroj očekivanih frekvencija <= 5 je ", counter)

## 
## Broj očekivanih frekvencija <= 5 je  0
```

Provjera očekivanih frekvencija za drugu grupaciju:

```

counter = 0
for (col_names in colnames(c_table3)){
  for (row_names in rownames(c_table3)){
    if (!(row_names == 'Sum' | col_names == 'Sum') ){
      cat('Očekivane frekvencije za razred ',col_names,'-',row_names,': ',(c_table3[row_names,'Sum'] * c_table3['Sum',col_names] / c_table3['Sum','Sum'])<=5)
      counter=counter+1
    }
  }
}
}

## Očekivane frekvencije za razred 1 - 0-1 : 1039.83
## Očekivane frekvencije za razred 1 - 2 : 1720.82
## Očekivane frekvencije za razred 1 - 3 : 1160.197
## Očekivane frekvencije za razred 1 - 4-9 : 1298.154
## Očekivane frekvencije za razred 2 - 0-1 : 1711.664
## Očekivane frekvencije za razred 2 - 2 : 2832.643
## Očekivane frekvencije za razred 2 - 3 : 1909.801
## Očekivane frekvencije za razred 2 - 4-9 : 2136.892
## Očekivane frekvencije za razred 3 - 0-1 : 871.2732
## Očekivane frekvencije za razred 3 - 2 : 1441.875
## Očekivane frekvencije za razred 3 - 3 : 972.1287
## Očekivane frekvencije za razred 3 - 4-9 : 1087.723
## Očekivane frekvencije za razred 4 - 0-1 : 438.3264
## Očekivane frekvencije za razred 4 - 2 : 725.3888
## Očekivane frekvencije za razred 4 - 3 : 489.0654
## Očekivane frekvencije za razred 4 - 4-9 : 547.2194
## Očekivane frekvencije za razred 5 - 0-1 : 27.09654
## Očekivane frekvencije za razred 5 - 2 : 44.84222
## Očekivane frekvencije za razred 5 - 3 : 30.23314
## Očekivane frekvencije za razred 5 - 4-9 : 33.82811
## Očekivane frekvencije za razred 6 - 0-1 : 49.80981
## Očekivane frekvencije za razred 6 - 2 : 82.43055
## Očekivane frekvencije za razred 6 - 3 : 55.57562
## Očekivane frekvencije za razred 6 - 4-9 : 62.18402

cat("\nBroj očekivanih frekvencija <= 5 je ", counter)

##
## Broj očekivanih frekvencija <= 5 je  0

```

Sada ćemo provesti test nezavisnosti za prvu grupaciju:

```

chisq.test(c_table_2,correct=F)

##
## Pearson's Chi-squared test
##
## data:  c_table_2
## X-squared = 105.36, df = 10, p-value < 2.2e-16

```

Sada ćemo provesti test nezavisnosti za drugu grupaciju:

```
chisq.test(c_table_3,correct=F)

##
##  Pearson's Chi-squared test
##
## data: c_table_3
## X-squared = 97.565, df = 15, p-value = 3.769e-14
```

S obzirom da oba testa daju p-vrijednosti jako male (manje od 0.05), odbacujemo H<sub>0</sub> i prihvaćamo alternativu, tj da su broj ukućana i emergent koji se korist za grijanje vode zavisne varijable.

Dakle, zavisnost ove dvije varijable je kompleksnija od onoga što je vizualno bilo vidljivo na bar plotu, gdje se možda pogrešno moglo naslutiti da su nezavisne.

## 2. Implicitiraju li visoka mjesečna primanja dostupnost druge nekretnine?

Koristit ćemo varijablu IMPEXAC. IMPEXAC predstavlja točnu količinu ukupnog mjesečnog neto prihoda pojedinog kućanstva. Raspon vrijednosti je između 0-99999.

Želimo provjeriti nedostaju li nam podatci:

```
sum(is.na(kucanstva19$IMPEXAC))
```

```
## [1] 0
```

Osnovna deskriptivna statistika:

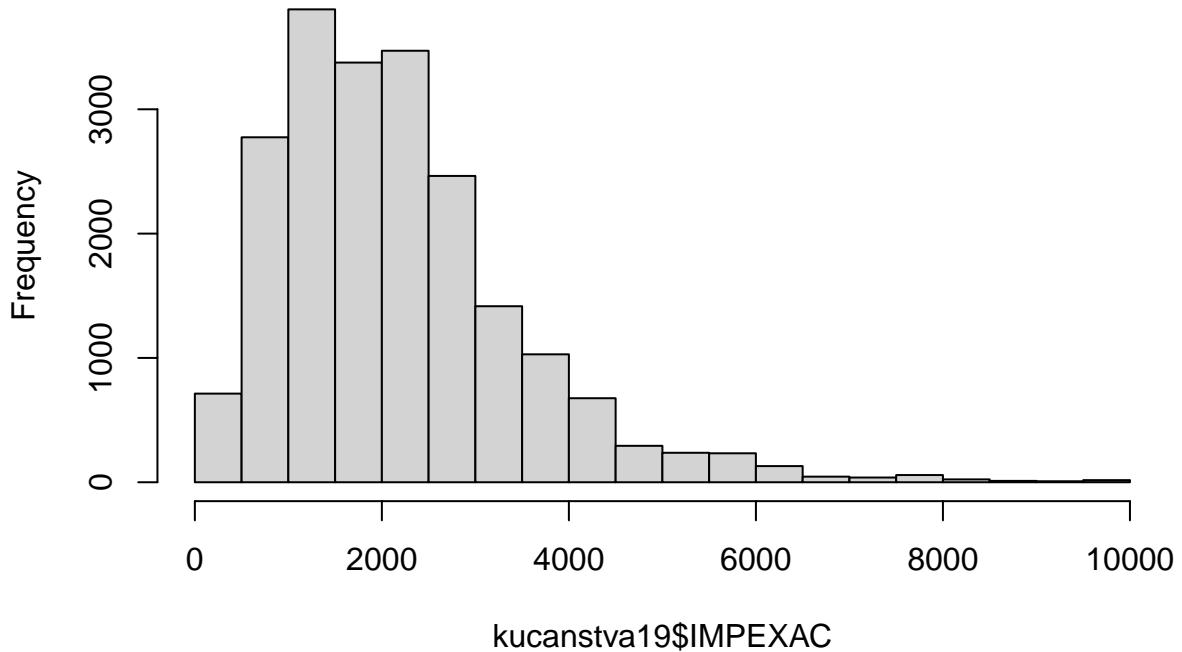
```
summary(kucanstva19$IMPEXAC)
```

```
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##          0    1247   1950    2181    2742    9991
```

Sada ćemo prikazati podatke o neto mjesečnim primanjima uz pomoć histograma:

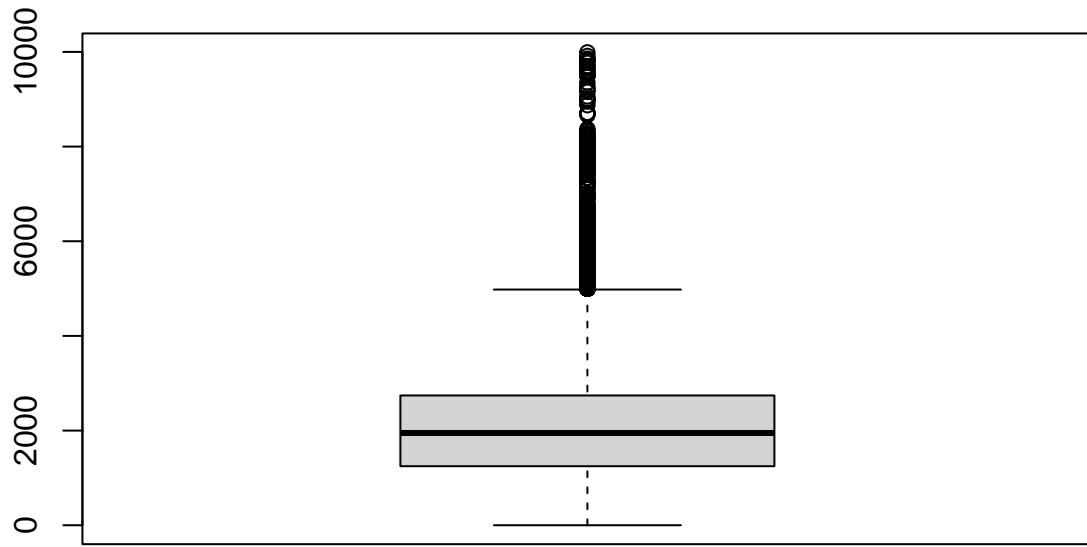
```
hist(kucanstva19$IMPEXAC)
```

### Histogram of kucanstva19\$IMPEXAC



Sada ćemo prikazati podatke uz pomoć box-plota:

```
boxplot(kucanstva19$IMPEXAC)
```



Također imamo velik broj stršećih vrijednosti povezanih s ukupnim mjesecnim neto prihodom pojedinog kućanstva, što pokazuje i box-plot Zanima nas koliko gornjih stršećih vrijednosti imamo:

```
IRQ19 = 1.5 * (quantile(kucanstva19$IMPEXAC, 0.75) - quantile(kucanstva19$IMPEXAC, 0.25))
IRQ19

##      75%
## 2242.5

ind19 = which(kucanstva19$IMPEXAC > IRQ19 + quantile(kucanstva19$IMPEXAC, 0.75))
length(kucanstva19$IMPEXAC[ind19]) #broj stršećih vrijednosti

## [1] 828

length(kucanstva19$IMPEXAC) #ukupan broj podataka

## [1] 20817

length(kucanstva19$IMPEXAC[ind19]) / length(kucanstva19$IMPEXAC) * 100 #% stršećih vrijednosti

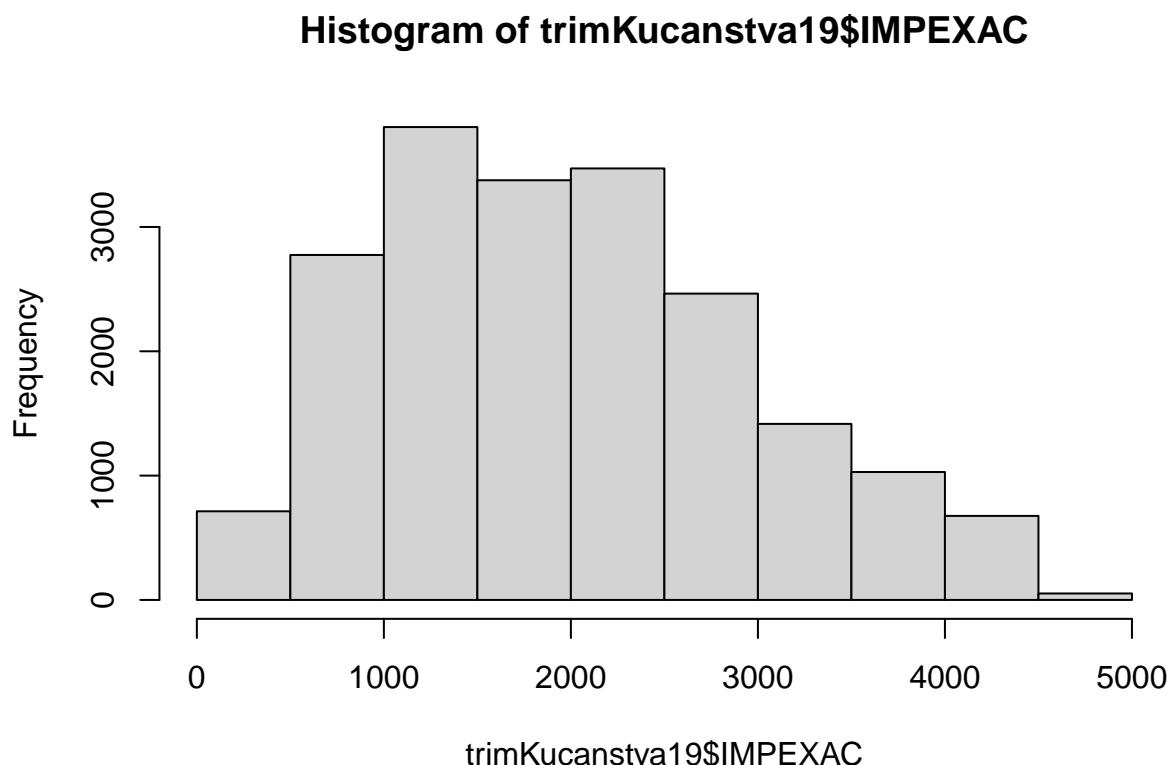
## [1] 3.977518
```

S obzirom na broj stršećih vrijednosti podrezat ćemo gornjih 5% podataka:

```
trimKucanstva19 = kucanstva19[kucanstva19$IMPEXAC < quantile(kucanstva19$IMPEXAC, 0.95), ]  
dim(trimKucanstva19)
```

```
## [1] 19776 223
```

```
hist(trimKucanstva19$IMPEXAC)
```



Želimo provjeriti odnos primanja pojedinog kućanstva i dostupnost druge nekretnine kućanstvu. Za pokazivanje navedenog dovoljno je provjeriti postojanje vrijednosti varijable REGTENV1, koja govori o načinu korištenja stambenog prostora, odnosno, ako vrijednost varijable REGTENV1 postoji onda kućanstvo ima dostupnu drugu nekretninu uz glavni stambeni prostor, inače kućanstvo nema dostupnu drugu nekretninu uz glavni stambeni prostor.

```
unique(trimKucanstva19$REGTENV1)
```

```
## [1] NA 1 2 3 6 5 4
```

Izdvojiti podatke varijabli IMPEXAC i REGTENV1.

```
kucPrimDostupno19 = data.frame(  
  IMPEXAC = trimKucanstva19$IMPEXAC,  
  REGTENV1 = trimKucanstva19$REGTENV1  
)
```

```
head(kucPrimDostupno19)
```

```

##    IMPEXAC REGTEENV1
## 1      1716      NA
## 2      2184      1
## 3      558      NA
## 4     2254      NA
## 5     1000      NA
## 6     3527      NA

```

Ova tablica služit će nam kako bi prikazali dostupnost druge nekretnine (DOSTUPNO = 1) ili nedostupnost druge nekretnine (DOSTUPNO = 0)

```

dostupno19 = data.frame(
  IMPEXAC = trimKucanstva19$IMPEXAC,
  DOSTUPNO = rep(0, nrow(trimKucanstva19))
)

head(dostupno19)

```

```

##    IMPEXAC DOSTUPNO
## 1      1716      0
## 2      2184      0
## 3      558      0
## 4     2254      0
## 5     1000      0
## 6     3527      0

```

Označavamo koja kućanstva imaju dostupnu drugu nekretninu za godinu 2019.

```

for(i in 1:nrow(kucPrimDostupno19)){
  for(j in 2:ncol(kucPrimDostupno19)){
    if(!is.na(kucPrimDostupno19[i, j]) && kucPrimDostupno19[i, j] != 0){
      dostupno19[i, 2] = 1
    }
  }
}
dim(dostupno19)

```

```

## [1] 19776      2

head(dostupno19)

```

```

##    IMPEXAC DOSTUPNO
## 1      1716      0
## 2      2184      1
## 3      558      0
## 4     2254      0
## 5     1000      0
## 6     3527      0

```

Sada ćemo podijeliti podatke u dvije tablice, na one s dostupnom drugom nekretninom i nedostupnom drugom nekretninom.

```
dostupno19_da = dostupno19[dostupno19$DOSTUPNO == 1, ]  
dostupno19_ne = dostupno19[dostupno19$DOSTUPNO == 0, ]  
head(dostupno19_da)
```

```
##      IMPEXAC DOSTUPNO  
## 2      2184      1  
## 10     1700      1  
## 14     1574      1  
## 15     2215      1  
## 16     1000      1  
## 20     811       1
```

```
head(dostupno19_ne)
```

```
##      IMPEXAC DOSTUPNO  
## 1      1716      0  
## 3      558       0  
## 4      2254      0  
## 5      1000      0  
## 6      3527      0  
## 7      2791      0
```

```
dim(dostupno19_da)
```

```
## [1] 2929      2
```

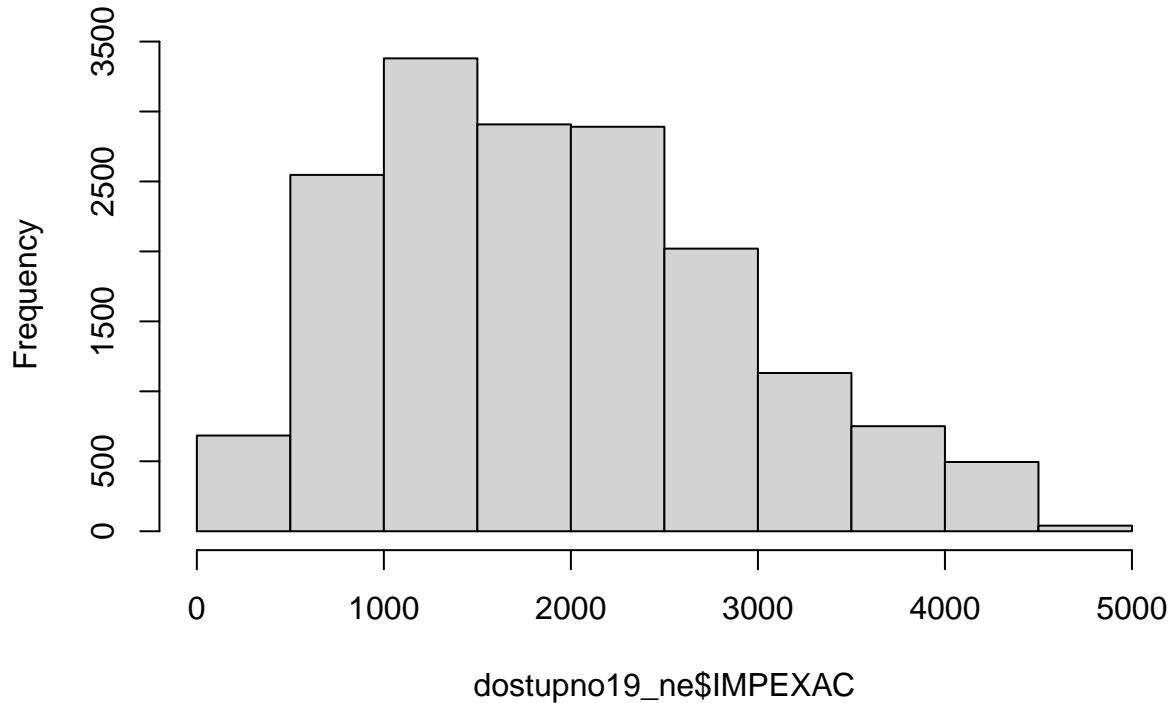
```
dim(dostupno19_ne)
```

```
## [1] 16847      2
```

Prikažimo sada pojedine skupine histogramom:

```
hist(dostupno19_ne$IMPEXAC)
```

### Histogram of dostupno19\_ne\$IMPEXAC

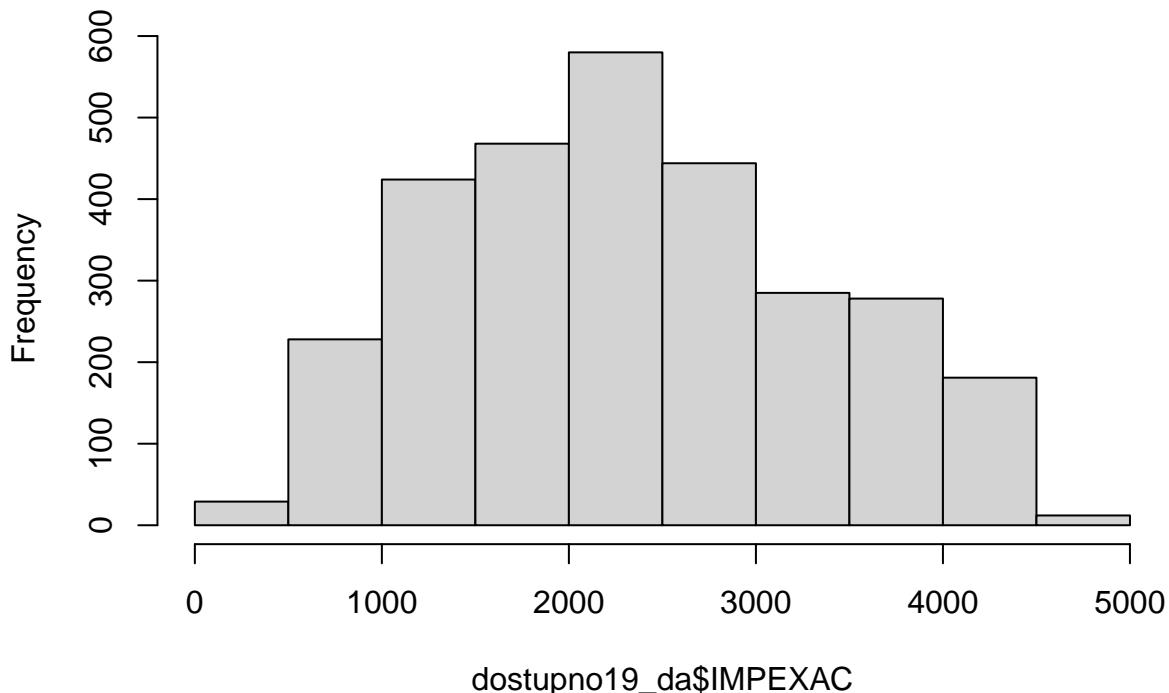


```
dim(dostupno19_ne)
```

```
## [1] 16847      2
```

```
hist(dostupno19_da$IMPEXAC)
```

### Histogram of dostupno19\_da\$IMPEXAC



```
dim(dostupno19_da)
```

```
## [1] 2929      2
```

Želimo provjeriti dolaze li podatci iz normalne distribucije:

```
require(nortest)
```

```
## Loading required package: nortest
```

```
lillie.test(dostupno19_ne$IMPEXAC)
```

```
##  
##  Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data:  dostupno19_ne$IMPEXAC  
## D = 0.083445, p-value < 2.2e-16
```

```
lillie.test(dostupno19_da$IMPEXAC)
```

```
##  
##  Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data:  dostupno19_da$IMPEXAC  
## D = 0.062654, p-value < 2.2e-16
```

Iz ovoga možemo zaključiti da podatci ne dolaze iz normalne distribucije.

S obzirom na to da podatci nisu normalno distribuirani ali imamo velik uzorak podataka, koristit ćemo t-test prilikom usporedbe srednjih vrijednosti dviju skupina zato što je t-test robustan na odstupanja od pretpostavke normalnosti prilikom velikih uzoraka.

$$H_0 : \mu_{dostupno19ne} = \mu_{dostupno19da}$$

$$H_1 : \mu_{dostupno19ne} < \mu_{dostupno19da}.$$

Sada ćemo usporediti srednje vrijednosti mjesecnih neto primanja pojedinih grupa uz pomoć t-testa.

```
t.test(dostupno19_ne$IMPEXAC, dostupno19_da$IMPEXAC, alternative = "less")
```

```
##  
## Welch Two Sample t-test  
##  
## data: dostupno19_ne$IMPEXAC and dostupno19_da$IMPEXAC  
## t = -21.169, df = 3932.6, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
##       -Inf -393.1399  
## sample estimates:  
## mean of x mean of y  
## 1924.087 2350.356
```

Možemo odbaciti  $H_0$  u korist  $H_1$ , odnosno zaključujemo da visoka mjesecna primanja impliciraju dostupnost druge nekretnine.

```
#Logistička regresija
```

```
dim(dostupno19)
```

```
## [1] 19776      2
```

```
logistickaRegresija19 = glm(dostupno19$DOSTUPNO ~ dostupno19$IMPEXAC, data = dostupno19, family = binomial)  
summary(logistickaRegresija19)
```

```
##  
## Call:  
## glm(formula = dostupno19$DOSTUPNO ~ dostupno19$IMPEXAC, family = binomial(),  
##       data = dostupno19)  
##  
## Deviance Residuals:  
##       Min        1Q    Median        3Q       Max  
## -0.8847  -0.5885  -0.5141  -0.4364   2.3263  
##  
## Coefficients:  
##                               Estimate Std. Error z value Pr(>|z|)  
## (Intercept)      -2.637e+00  4.915e-02 -53.65  <2e-16 ***  
## dostupno19$IMPEXAC 4.161e-04  1.969e-05   21.13  <2e-16 ***  
## ---
```

```

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 16589  on 19775  degrees of freedom
## Residual deviance: 16142  on 19774  degrees of freedom
## AIC: 16146
##
## Number of Fisher Scoring iterations: 4

```

```
logistickaRegresija19
```

```

##
## Call:  glm(formula = dostupno19$DOSTUPNO ~ dostupno19$IMPEXAC, family = binomial(),
##           data = dostupno19)
##
## Coefficients:
##             (Intercept)  dostupno19$IMPEXAC
##             -2.6365774          0.0004161
##
## Degrees of Freedom: 19775 Total (i.e. Null);  19774 Residual
## Null Deviance:      16590
## Residual Deviance: 16140      AIC: 16150

```

```
yHat19 = logistickaRegresija19$fitted.values > 0.2
```

```
tab19 = table(dostupno19$DOSTUPNO, yHat19)
tab19
```

```

##   yHat19
##   FALSE  TRUE
##   0 14433 2414
##   1  2174   755

```

```

accuracy19 = sum(diag(tab19)) / sum(tab19)
precision19 = tab19[2,2] / sum(tab19[,2])
recall19 = tab19[2,2] / sum(tab19[2,])
specificity19 = tab19[1,1] / sum(tab19[,1])

```

```
accuracy19 #true positive / (all)
```

```
## [1] 0.7680016
```

```
precision19 #true positive / (true positive + false positive) -> true positive / total predicted positive
```

```
## [1] 0.2382455
```

```
recall19 #true positive / (true positive + false negative)
```

```
## [1] 0.2577672
```

```
specificity19 #
```

```
## [1] 0.8690913
```

Granicu koja klasificira naše podatke kao TRUE ili FALSE postavili smo na vrijednost od 0,2. Možemo primijetiti kako naš model ima visoku točnost (u 76,8% slučajeva točno predviđa ishod) i specifičnost(ispravno klasificira 86,9% netočnih podataka od ukupno netočnih podataka), ali ne i visoku preciznost(u 86,9% slučajeva klasificira podatke kao točne kada oni uistinu jesu točni) i odziv(u 25,7% slučajeva ispravno klasificira točne podatke od svih točnih podataka). Ovime možemo zaključiti kako model nije idealan.

### 3. Usporedba potrošnje kućanstava s regijom u kojoj se nalazi.

Varijabla CCAA govori o regijama u kojima se nalaze ispitivana kućanstva, a varijabla GASTOT govori o ukupnom iznosu godišnjih izdataka kućanstva.

Provjerit ćemo koje vrijednosti mogu poprimiti ove varijable.

```
unique(kucanstva19$CCAA)
```

```
## [1] 1 3 5 6 8 0 7 9 4 2
```

```
min(kucanstva19$GASTOT)
```

```
## [1] 22235148
```

```
max(kucanstva19$GASTOT)
```

```
## [1] 38167962730
```

```
sum(is.na(kucanstva19$CCAA))
```

```
## [1] 0
```

```
sum(is.na(kucanstva19$GASTOT))
```

```
## [1] 0
```

Vidimo da varijabla CCAA može poprimiti vrijednosti:

0 1-Andalusia, 2-Aragon, 3-Asturias, Principality of, 4-Balears, Illes, 5-Canary Islands, 6-Cantabria, 7-Castilla y León, 8-Castilla - La Mancha, 9-Catalonia

Varijabla GASTOT može poprimiti numeričke vrijednosti od 22235148 do 38167962730.

Također vidimo da nema nedostajućih vrijednosti, ali u varijabli CCAA postoji vrijednost 0 koja ne označava ništa pa ćemo nju izbaciti.

```
kucanstva19.cleaned <- filter(kucanstva19, CCAA != 0)
unique(kucanstva19.cleaned$CCAA)
```

```
## [1] 1 3 5 6 8 7 9 4 2
```

```
#izbacit čemo nedostajuće vrijednosti nastale izbacivanjem 0
kucanstva19 <- kucanstva19.cleaned[complete.cases(kucanstva19.cleaned$CCAA),]
sum(is.na(kucanstva19$CCAA))
```

```
## [1] 0
```

Želimo usporediti potrošnju kućanstava s obzirom na to u kojoj regiji se nalaze. ANOVA je metoda kojom testiramo sredine više populacija. Koristit ćemo jednofaktorski ANOVA model gdje se razmatra utjecaj jednog faktora, u našem slučaju regija u kojoj se kućanstvo nalazi.

Analizom varijance testiramo:

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_1 : \text{barem dvije sredine nisu iste.}$$

Pripremit ćemo podatke.

```
kucanstva19$CCAA = factor(kucanstva19$CCAA, levels = c(1, 2, 3, 4, 5, 6, 7, 8, 9), labels = c('Andalusia'
```

```
summary(kucanstva19$CCAA)
```

	Andalusia	Aragon	Asturias, Principality of
##	3266	2311	2412
##	Balears, Illes	Canary Islands	Cantabria
##	1465	1588	2949
##	Castilla y León	Castilla - La Mancha	Catalonia
##	1980	1187	1981

Prije izvedbe testa, pobjerit ćemo jesu li pretpostavke ANOVA-e zadovoljene.

Pretpostavke ANOVA-e su nezavisnost pojedinih podataka u uzorcima, normalna razdioba podataka i homogenost varijanci među populacijama.

Kad su veličine grupa podjednake, ANOVA je relativno robusna metoda na blaga odstupanja od pretpostavke normalnosti i homogenosti varijanci. Vidimo da su u našim podacima veličina grupa približno podjednake, ali ipak ćemo provjerit koliko su odstupanja od pretpostavki velika.

Provjeru normalnosti možemo provjeriti Lillieforsovom inaćicom KS testa. Vizualno možemo provjeriti normalnost histogramom. U ovom slučaju razmatrat ćemo regije kao varijablu koja određuje grupe (populacije) i potrošnju kao zavisnu varijablu.

```
require(nortest)
```

```
lillie.test(kucanstva19$GASTOT)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: kucanstva19$GASTOT
## D = 0.16802, p-value < 2.2e-16
```

```
lillie.test(kucanstva19$GASTOT[kucanstva19$CCAA=='Andalusia'])
```

```

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: kucanstva19$GASTOT[kucanstva19$CCAA == "Andalusia"]
## D = 0.12471, p-value < 2.2e-16

lillie.test(kucanstva19$GASTOT[kucanstva19$CCAA=='Aragon'])

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: kucanstva19$GASTOT[kucanstva19$CCAA == "Aragon"]
## D = 0.13616, p-value < 2.2e-16

lillie.test(kucanstva19$GASTOT[kucanstva19$CCAA=='Asturias, Principality of'])

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: kucanstva19$GASTOT[kucanstva19$CCAA == "Asturias, Principality of"]
## D = 0.13418, p-value < 2.2e-16

lillie.test(kucanstva19$GASTOT[kucanstva19$CCAA=='Balears, Illes'])

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: kucanstva19$GASTOT[kucanstva19$CCAA == "Balears, Illes"]
## D = 0.1558, p-value < 2.2e-16

lillie.test(kucanstva19$GASTOT[kucanstva19$CCAA=='Canary Islands'])

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: kucanstva19$GASTOT[kucanstva19$CCAA == "Canary Islands"]
## D = 0.15118, p-value < 2.2e-16

lillie.test(kucanstva19$GASTOT[kucanstva19$CCAA=='Cantabria'])

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: kucanstva19$GASTOT[kucanstva19$CCAA == "Cantabria"]
## D = 0.10296, p-value < 2.2e-16

lillie.test(kucanstva19$GASTOT[kucanstva19$CCAA=='Castilla y León'])

```

```

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: kucanstva19$GASTOT[kucanstva19$CCAA == "Castilla y León"]
## D = 0.14726, p-value < 2.2e-16

lillie.test(kucanstva19$GASTOT[kucanstva19$CCAA=="Castilla - La Mancha"])

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: kucanstva19$GASTOT[kucanstva19$CCAA == "Castilla - La Mancha"]
## D = 0.15749, p-value < 2.2e-16

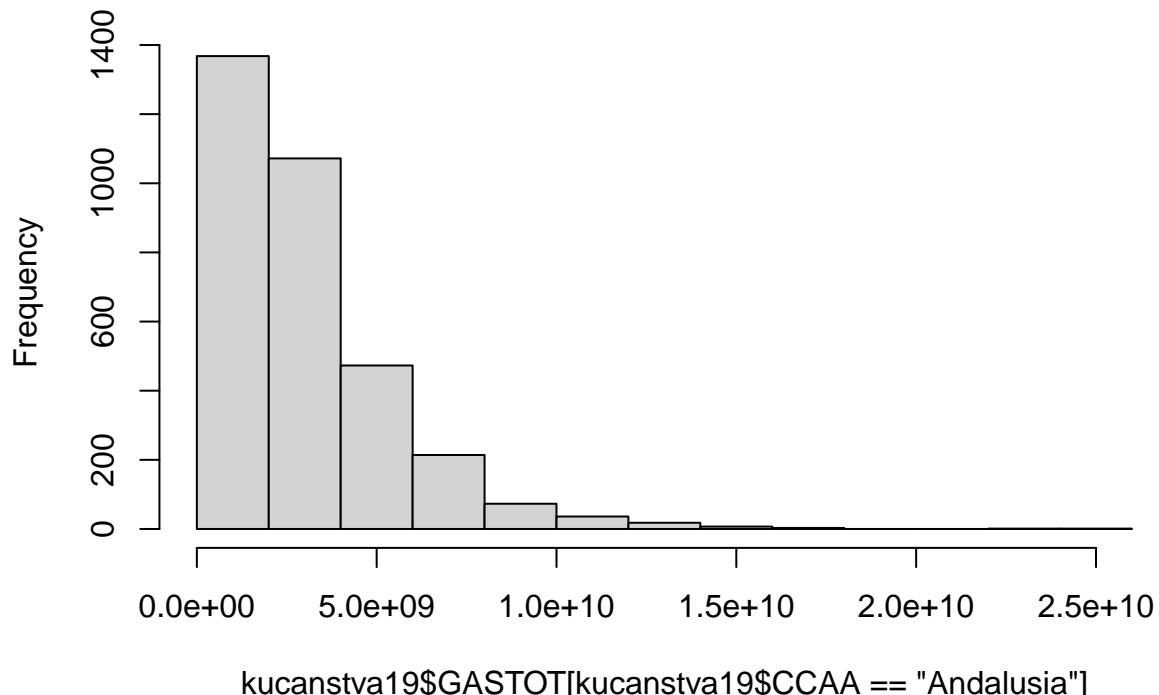
lillie.test(kucanstva19$GASTOT[kucanstva19$CCAA=="Catalonia"])

## 
## Lilliefors (Kolmogorov-Smirnov) normality test
## 
## data: kucanstva19$GASTOT[kucanstva19$CCAA == "Catalonia"]
## D = 0.1148, p-value < 2.2e-16

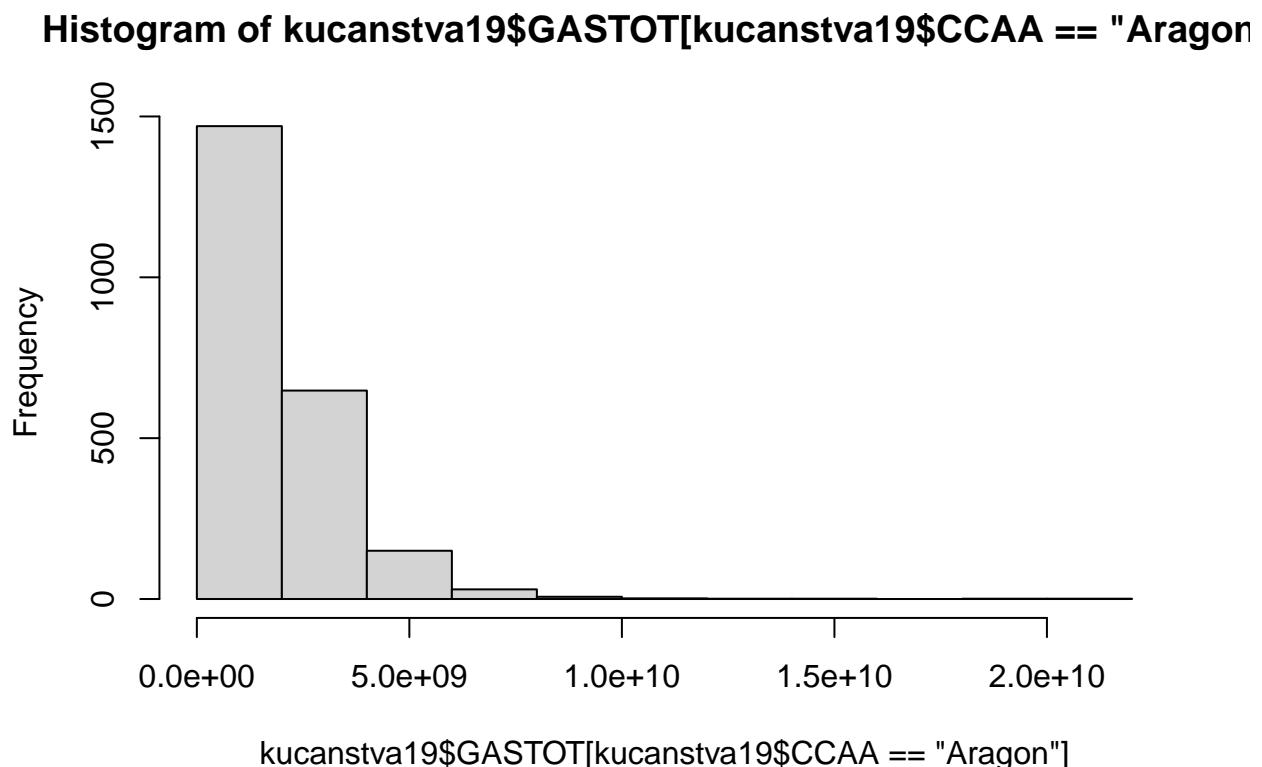
hist(kucanstva19$GASTOT[kucanstva19$CCAA=="Andalusia"])

```

## Histogram of kucanstva19\$GASTOT[kucanstva19\$CCAA == "Andalusia"]

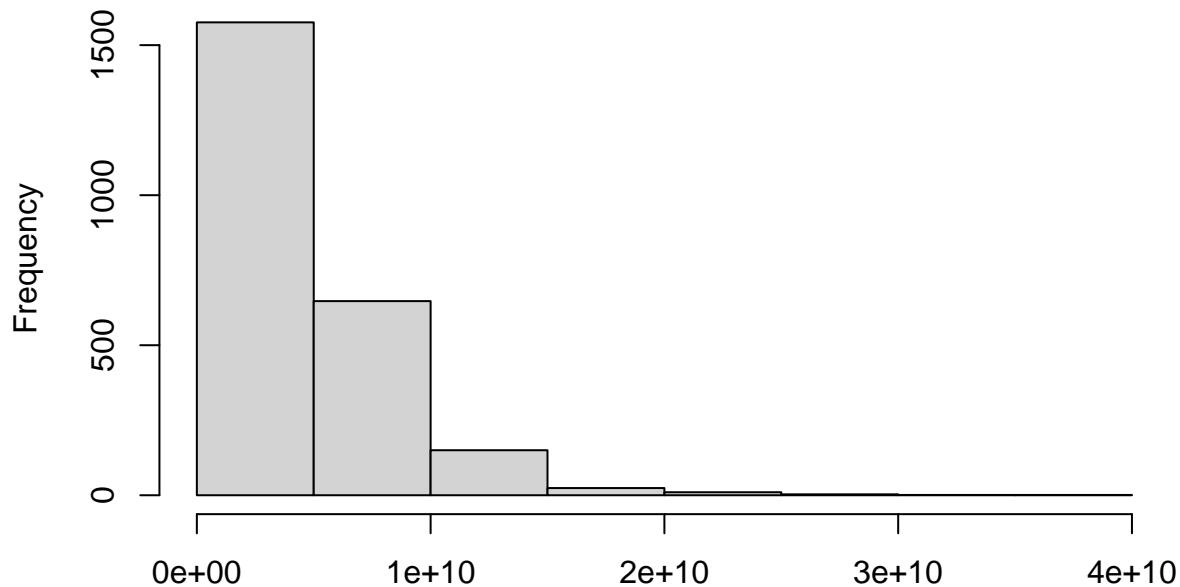


```
hist(kucanstva19$GASTOT[kucanstva19$CCAA=="Aragon"])
```



```
hist(kucanstva19$GASTOT[kucanstva19$CCAA=="Asturias, Principality of"])
```

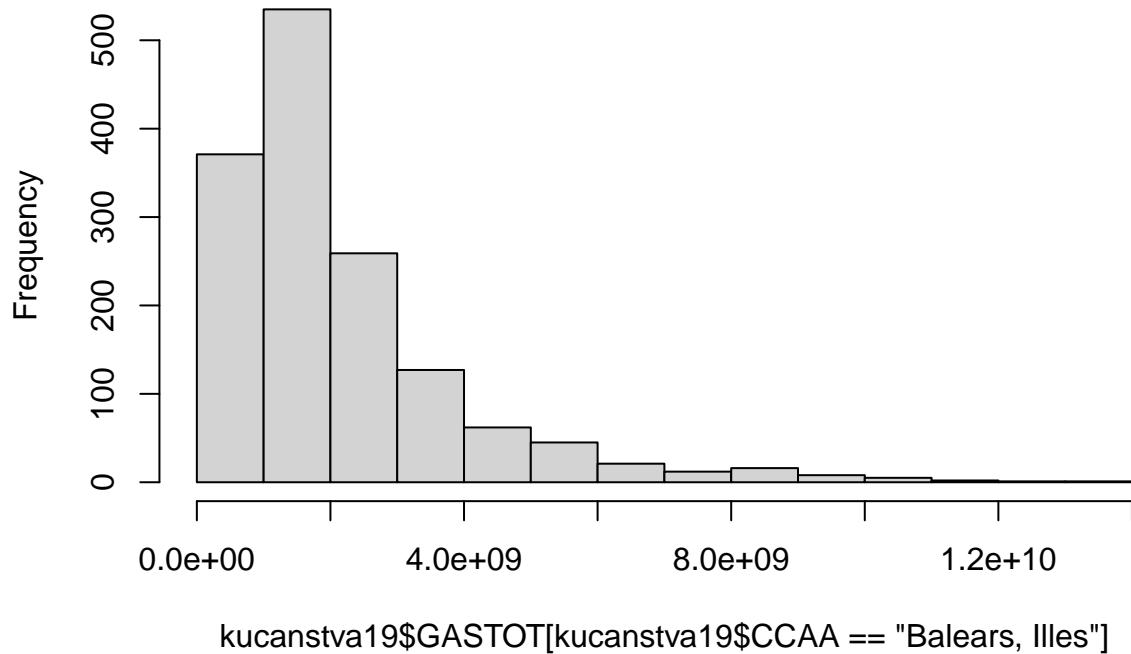
ogram of kucanstva19\$GASTOT[kucanstva19\$CCAA == "Asturias, Principality of"]



kucanstva19\$GASTOT[kucanstva19\$CCAA == "Asturias, Principality of"]

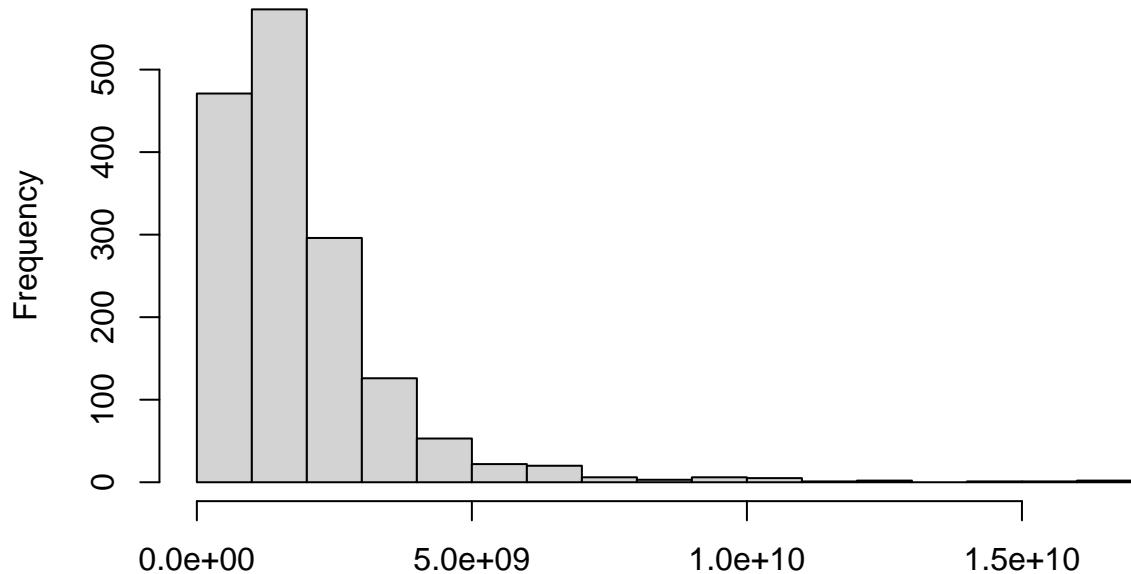
```
hist(kucanstva19$GASTOT[kucanstva19$CCAA=="Balears, Illes"])
```

histogram of kucanstva19\$GASTOT[kucanstva19\$CCAA == "Balears, Illes"]



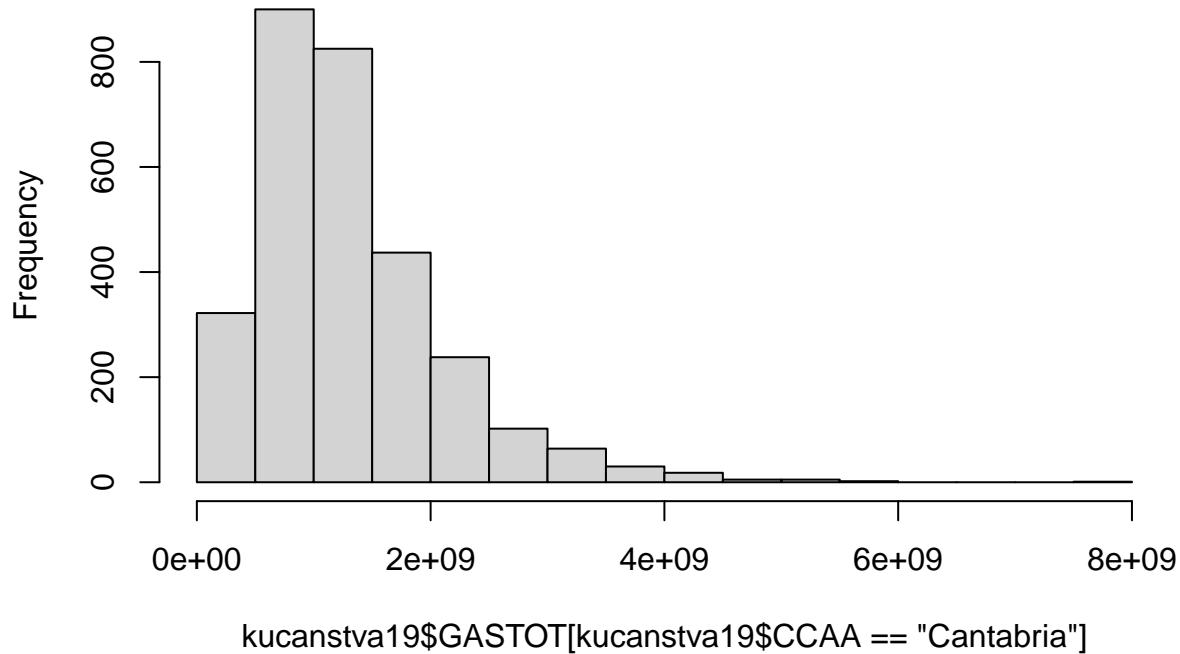
```
hist(kucanstva19$GASTOT[kucanstva19$CCAA=="Canary Islands"])
```

istogram of kucanstva19\$GASTOT[kucanstva19\$CCAA == "Canary Islands"]



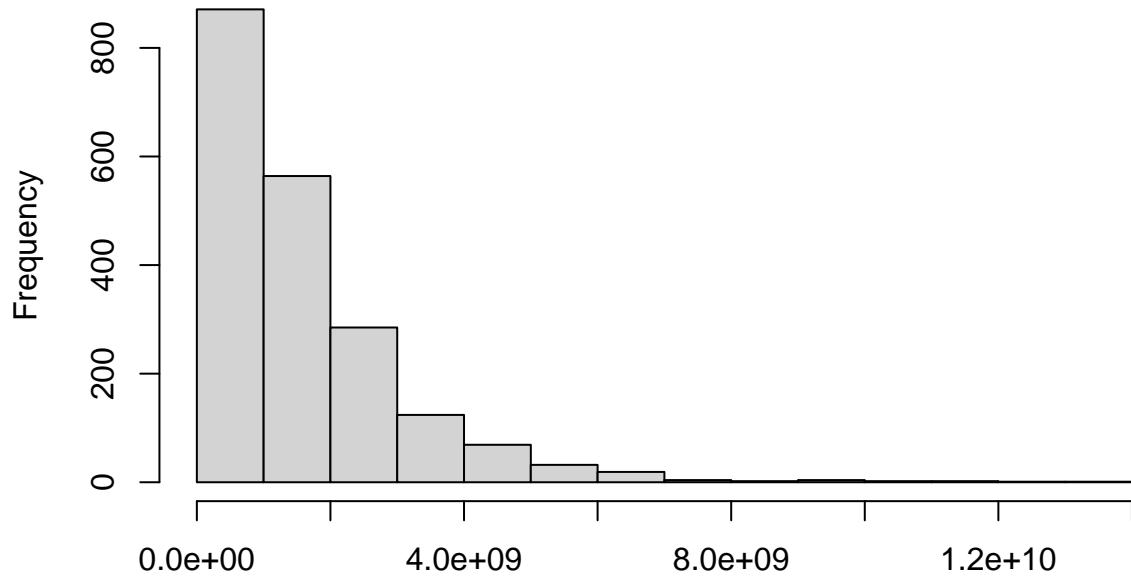
```
hist(kucanstva19$GASTOT[kucanstva19$CCAA=="Cantabria"])
```

## Histogram of kucanstv19\$GASTOT[kucanstv19\$CCAA == "Cantabria"]



```
hist(kucanstv19$GASTOT[kucanstv19$CCAA=="Castilla y León"])
```

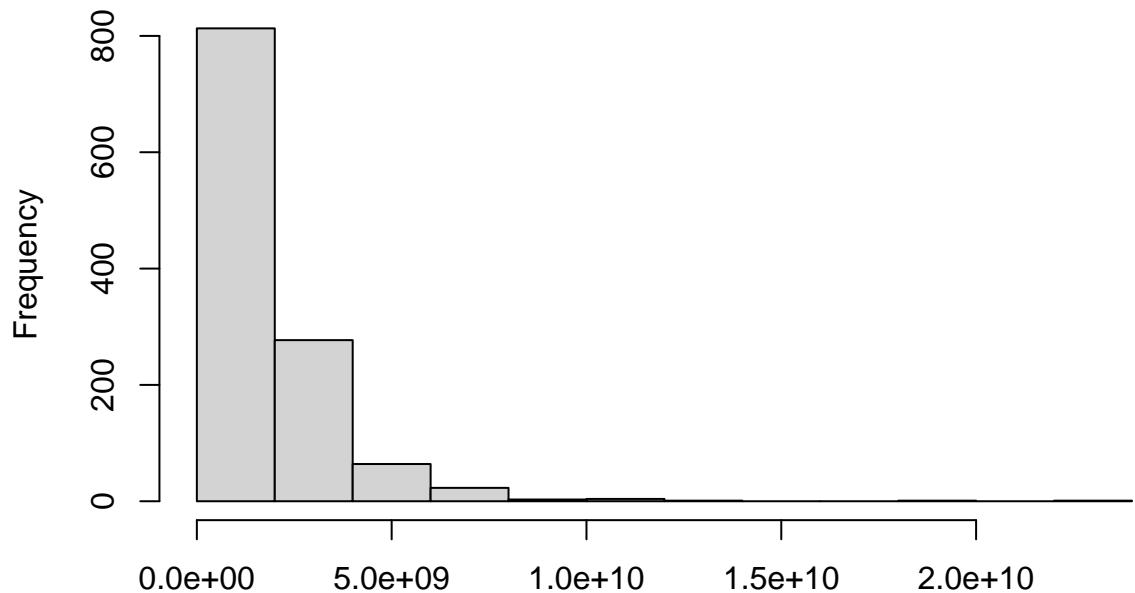
istogram of kucanstva19\$GASTOT[kucanstva19\$CCAA == "Castilla y L



kucanstva19\$GASTOT[kucanstva19\$CCAA == "Castilla y León"]

```
hist(kucanstva19$GASTOT[kucanstva19$CCAA=="Castilla - La Mancha"])
```

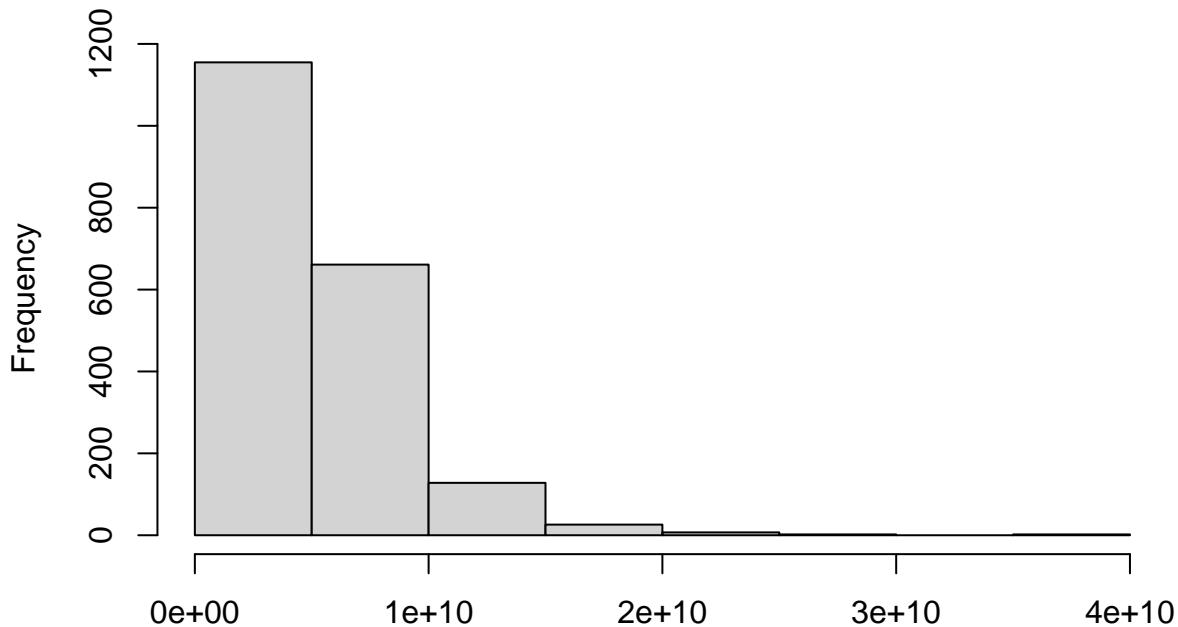
ogram of kucanstva19\$GASTOT[kucanstva19\$CCAA == "Castilla – La Mancha"]



kucanstva19\$GASTOT[kucanstva19\$CCAA == "Castilla – La Mancha"]

```
hist(kucanstva19$GASTOT[kucanstva19$CCAA=="Catalonia"])
```

## Histogram of kucanstva19\$GASTOT[kucanstva19\$CCAA == "Cataloni



kucanstva19\$GASTOT[kucanstva19\$CCAA == "Catalonia"]

Pokušat ćemo log transformacijom približiti podatke normalnoj distribuciji.

```
lillie.test(log(kucanstva19$GASTOT))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(kucanstva19$GASTOT)  
## D = 0.011307, p-value = 1.146e-05
```

```
lillie.test(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Andalusia']))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(kucanstva19$GASTOT[kucanstva19$CCAA == "Andalusia"])  
## D = 0.044721, p-value < 2.2e-16
```

```
lillie.test(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Aragon']))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(kucanstva19$GASTOT[kucanstva19$CCAA == "Aragon"])  
## D = 0.017034, p-value = 0.107
```

```
lillie.test(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Asturias, Principality of']))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(kucanstva19$GASTOT[kucanstva19$CCAA == "Asturias, Principality of"])  
## D = 0.061138, p-value < 2.2e-16
```

```
lillie.test(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Balears, Illes']))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(kucanstva19$GASTOT[kucanstva19$CCAA == "Balears, Illes"])  
## D = 0.021492, p-value = 0.1036
```

```
lillie.test(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Canary Islands']))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(kucanstva19$GASTOT[kucanstva19$CCAA == "Canary Islands"])  
## D = 0.016741, p-value = 0.3462
```

```
lillie.test(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Cantabria']))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(kucanstva19$GASTOT[kucanstva19$CCAA == "Cantabria"])  
## D = 0.036132, p-value = 1.721e-09
```

```
lillie.test(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Castilla y León']))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(kucanstva19$GASTOT[kucanstva19$CCAA == "Castilla y León"])  
## D = 0.030141, p-value = 0.0002568
```

```
lillie.test(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Castilla - La Mancha']))
```

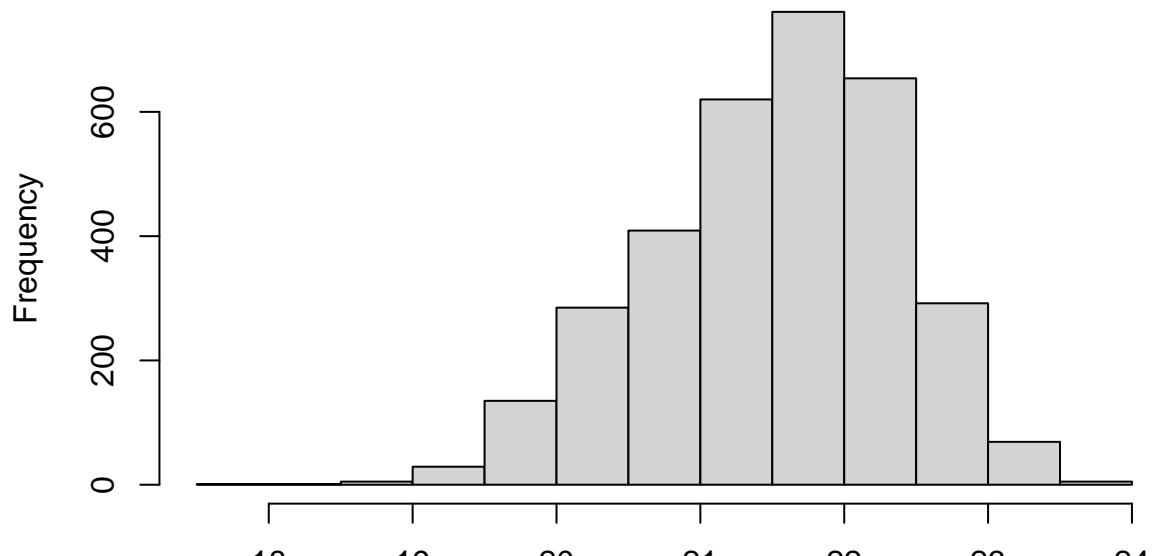
```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(kucanstva19$GASTOT[kucanstva19$CCAA == "Castilla - La Mancha"])  
## D = 0.031443, p-value = 0.007845
```

```
lillie.test(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Catalonia']))
```

```
##  
## Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data: log(kucanstva19$GASTOT[kucanstva19$CCAA == "Catalonia"])  
## D = 0.088233, p-value < 2.2e-16
```

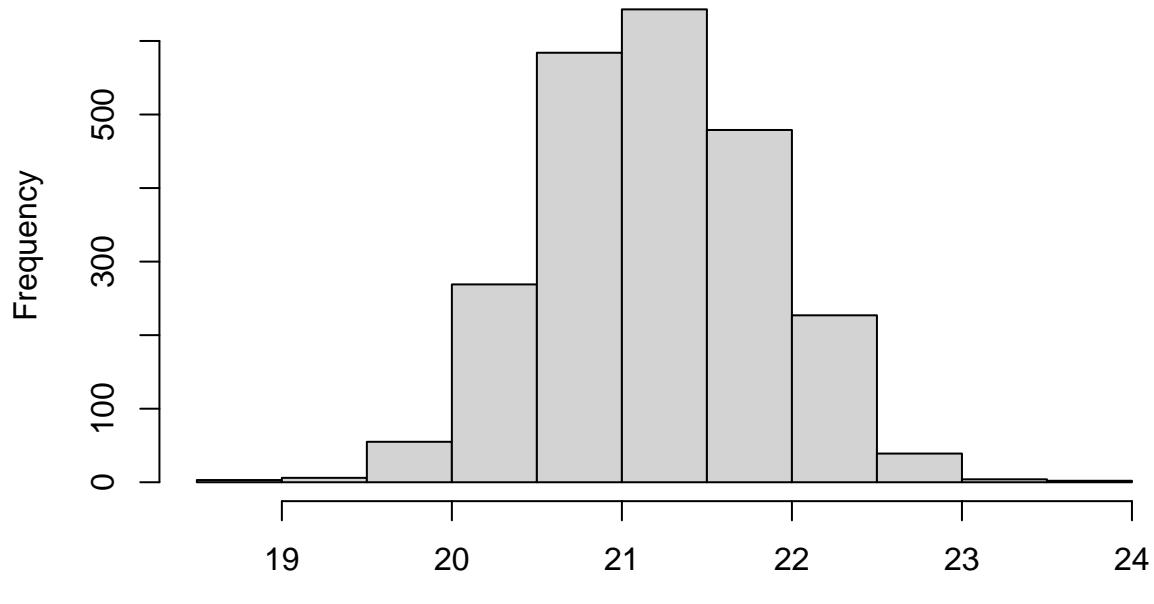
```
hist(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Andalusia']))
```

## istogram of log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Andalu



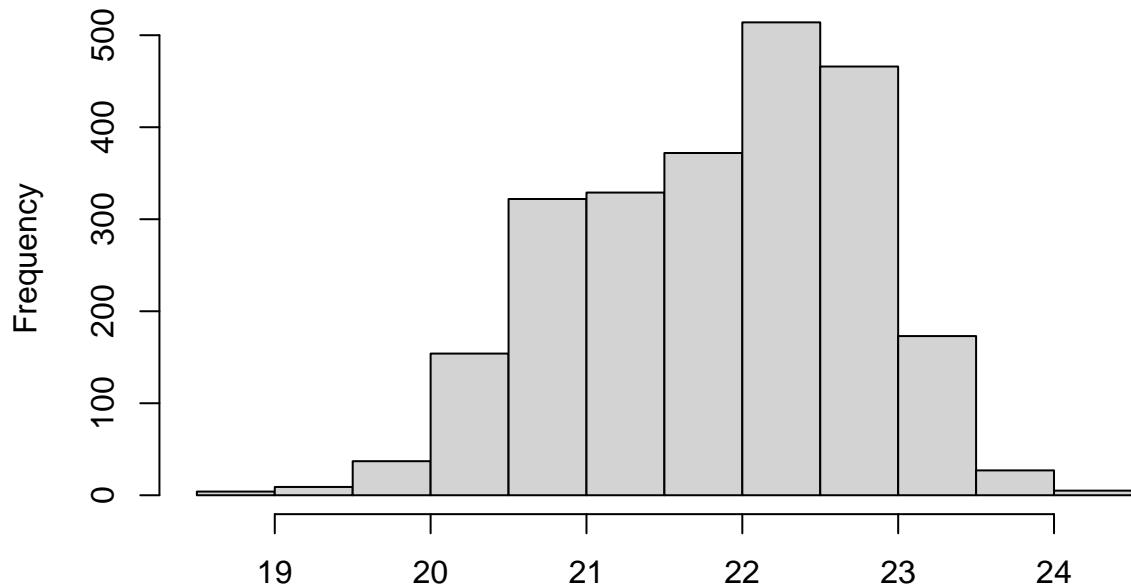
```
hist(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Aragon']))
```

## Histogram of log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Arago



```
hist(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Asturias, Principality of']))
```

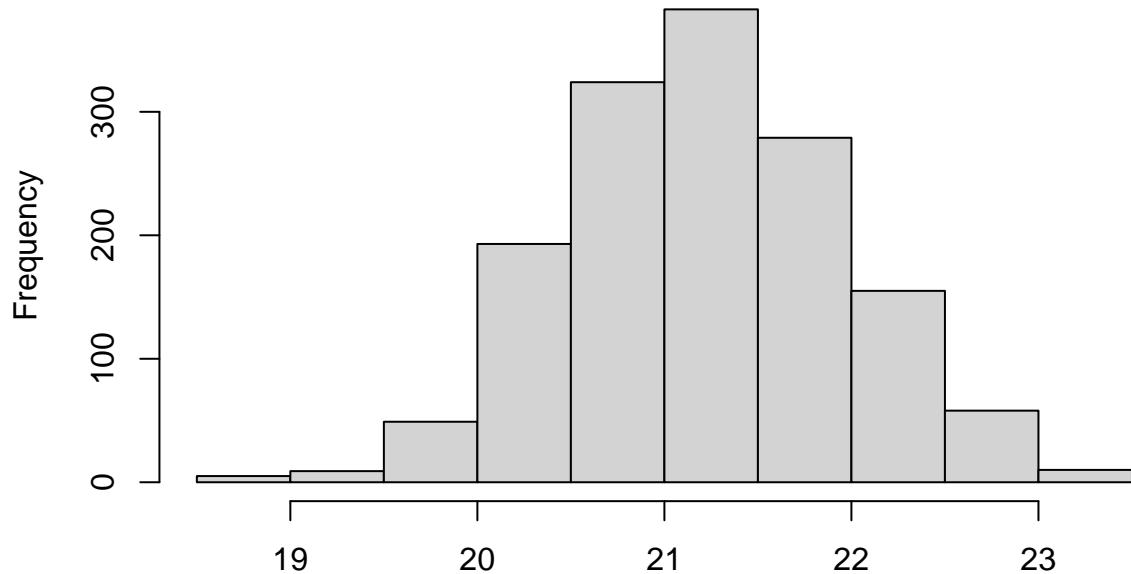
lm of log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Asturias, Principality of"])



log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Asturias, Principality of"])

```
hist(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Balears, Illes']))
```

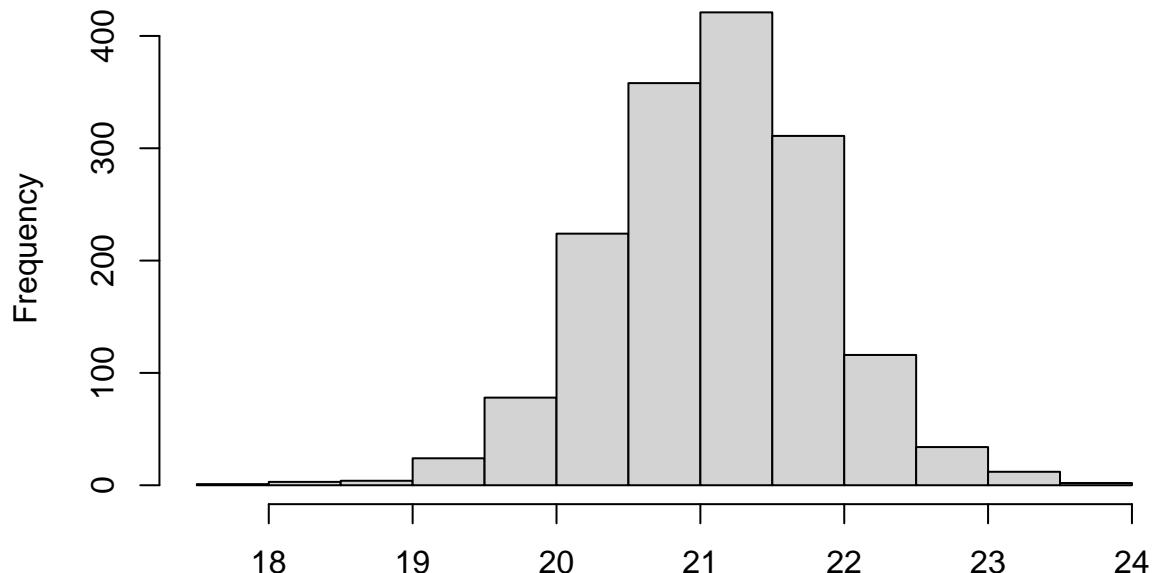
```
stogram of log(kucanstva19$GASTOT[kucanstva19$CCAA == "Balears,
```



```
log(kucanstva19$GASTOT[kucanstva19$CCAA == "Balears, Illes"])
```

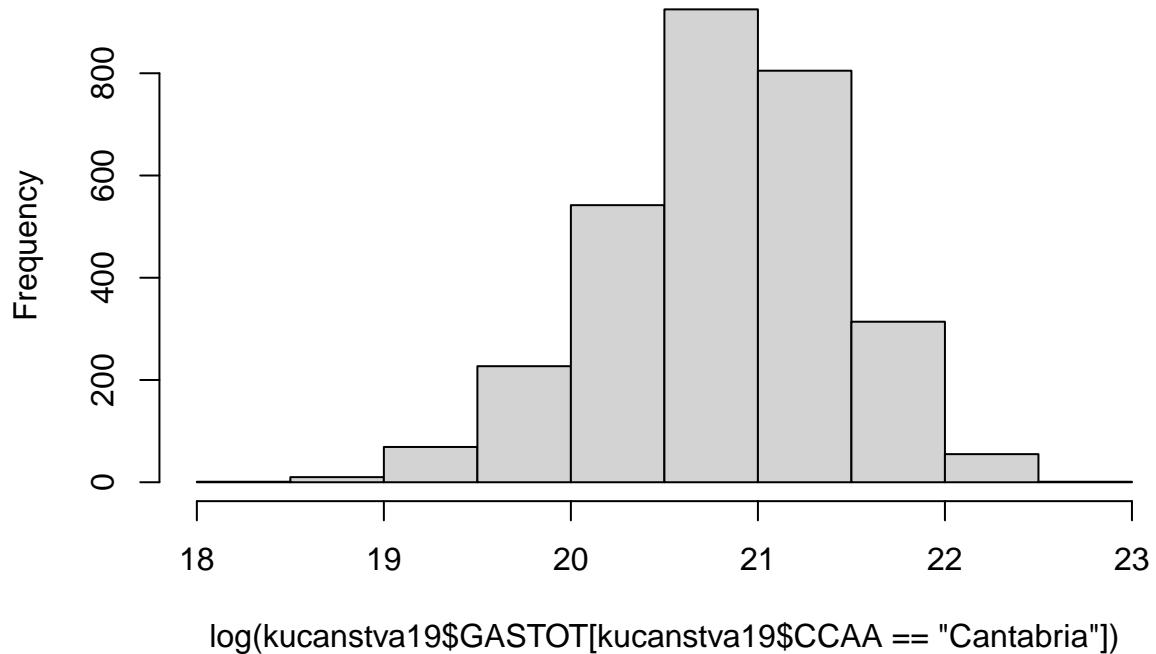
```
hist(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Canary Islands']))
```

ogram of log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Canary Is



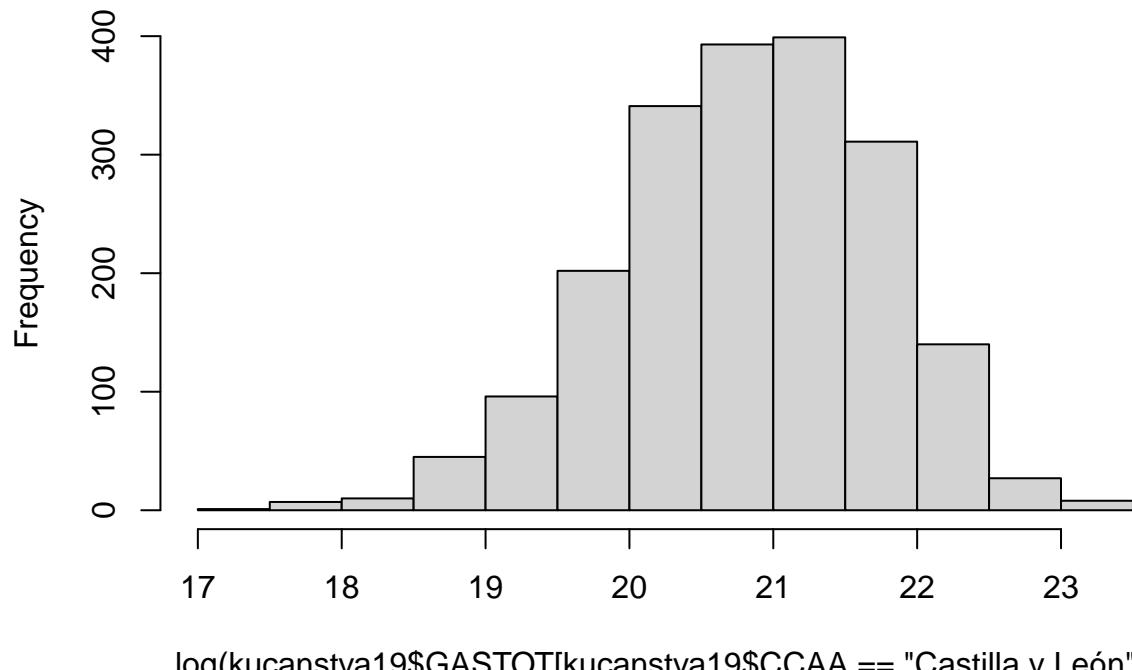
```
hist(log(kucanstva19$GASTOT [kucanstva19$CCAA=='Cantabria']))
```

histogram of log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Cantabria"])



```
hist(log(kucanstva19$GASTOT[kucanstva19$CCAA=="Castilla y León"]))
```

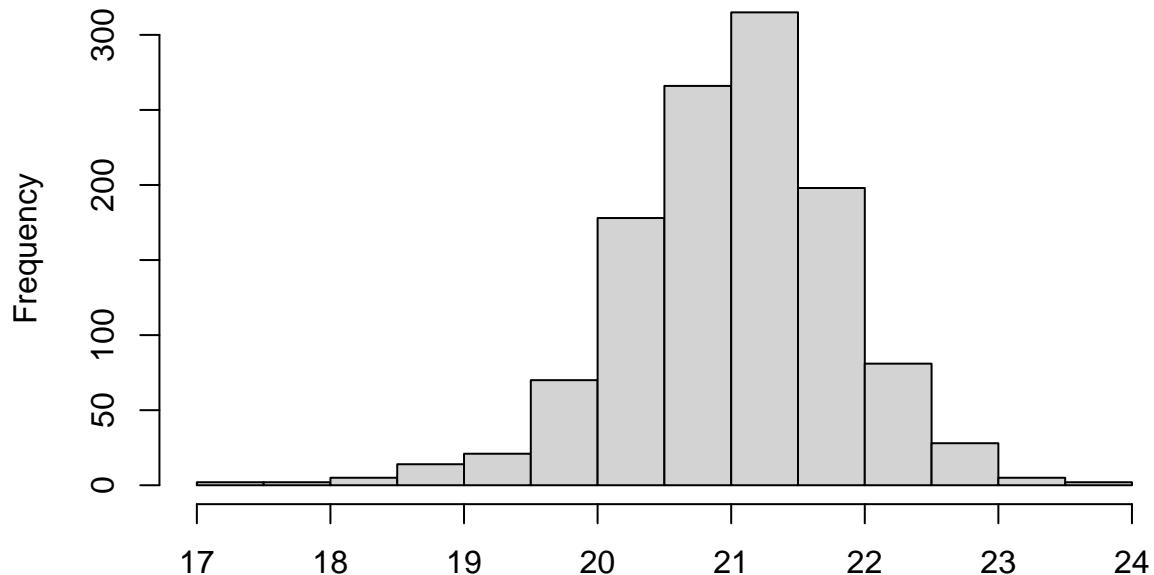
ogram of log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Castilla y



log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Castilla y León"])

```
hist(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Castilla - La Mancha']))
```

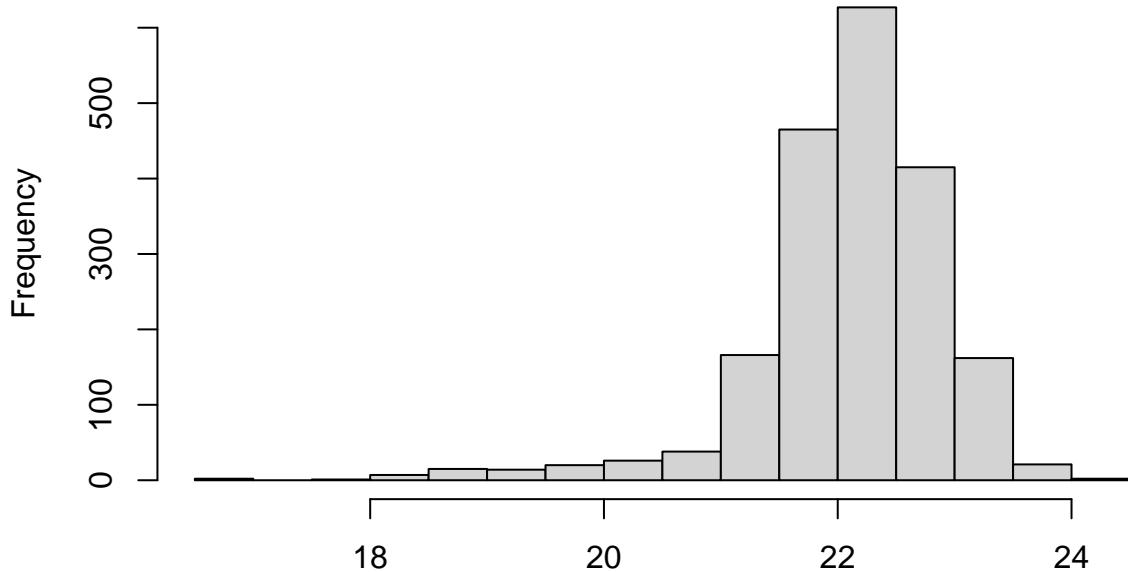
ram of  $\log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Castilla - La$



$\log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Castilla - La Mancha"])$

```
hist(log(kucanstva19\$GASTOT[kucanstva19\$CCAA=='Catalonia']))
```

## histogram of log(kucanstva19\$GASTOT[kucanstva19\$CCAA == "Catalonia"])



Sada vidimo da većina podataka ima normalnu razdiobu. Izbacit ćemo regije čija razdioba jako odstupa od normalne (Andalusia, Asturias, Principality of, Cantabria i Catalonia), te ćemo testiranje provesti na ostalim regijama.

```
kucanstva19 <- filter(kucanstva19, CCAA != "Andalusia")
kucanstva19 <- filter(kucanstva19, CCAA != "Asturias, Principality of")
kucanstva19 <- filter(kucanstva19, CCAA != "Cantabria")
kucanstva19 <- filter(kucanstva19, CCAA != "Catalonia")

kucanstva19$CCAA <- droplevels(kucanstva19$CCAA, "Andalusia")
kucanstva19$CCAA <- droplevels(kucanstva19$CCAA, "Asturias, Principality of")
kucanstva19$CCAA <- droplevels(kucanstva19$CCAA, "Cantabria")
kucanstva19$CCAA <- droplevels(kucanstva19$CCAA, "Catalonia")
```

```
summary(kucanstva19$CCAA)
```

```
##          Aragon           Balears, Illes           Canary Islands
##             2311                  1465                      1588
## Castilla y León Castilla - La Mancha
##                 1980                  1187
```

```
sum(is.na(kucanstva19$CCAA))
```

```
## [1] 0
```

Zanima nas koliko stršećih vrijednosti imamo:

```

IRQ19 = 1.5 * (quantile(kucanstva19$GASTOT, 0.75) - quantile(kucanstva19$GASTOT, 0.25))
ind19h = which(kucanstva19$GASTOT > IRQ19 + quantile(kucanstva19$GASTOT, 0.75))

cat(length(kucanstva19$GASTOT[ind19h]) / length(kucanstva19$GASTOT) * 100, "% gornjih stršećih vrijednosti")

## 5.438987 % gornjih stršećih vrijednosti

IRQ19 = 1.5 * (quantile(kucanstva19$GASTOT, 0.75) - quantile(kucanstva19$GASTOT, 0.25))
ind19l = which(kucanstva19$GASTOT < IRQ19 - quantile(kucanstva19$GASTOT, 0.75))

cat(length(kucanstva19$GASTOT[ind19l]) / length(kucanstva19$GASTOT) * 100, "% donjih stršećih vrijednosti")

## 0 % donjih stršećih vrijednosti

Postotak stršećih vrijednosti nije velik, pa ih možemo izbaciti.

kucanstva19.trimmed <- kucanstva19

cat("Minimalna i maksimalna vrijednost prije izbacivanja stršećih vrijednosti: \n", min(kucanstva19.trimmed),
    max(kucanstva19.trimmed))

## Minimalna i maksimalna vrijednost prije izbacivanja stršećih vrijednosti:
## 26272266
## 22580942571

kucanstva19.trimmed <- filter(kucanstva19.trimmed, GASTOT < 0.054*38167962730)

cat("\nMinimalna i maksimalna vrijednost nakon izbacivanja stršećih vrijednosti: \n", min(kucanstva19.trimmed),
    max(kucanstva19.trimmed))

## Minimalna i maksimalna vrijednost nakon izbacivanja stršećih vrijednosti:
## 26272266
## 2060124363

```

Homogenost varijanci testirat ćemo Bartlettovim testom. Naše hipoteze su:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$$

$$H_1 : \text{barem dvije varijance nisu iste.}$$

```
bartlett.test(log(kucanstva19$GASTOT) ~ kucanstva19$CCAA)
```

```

##
##  Bartlett test of homogeneity of variances
##
## data: log(kucanstva19$GASTOT) by kucanstva19$CCAA
## Bartlett's K-squared = 258.99, df = 4, p-value < 2.2e-16

var(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Aragon']))

```

```
## [1] 0.4302137
```

```

var(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Balears, Illes']))

## [1] 0.5531493

var(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Canary Islands']))

## [1] 0.5634273

var(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Castilla y León']))

## [1] 0.8426946

var(log(kucanstva19$GASTOT[kucanstva19$CCAA=='Castilla - La Mancha']))

## [1] 0.6755522

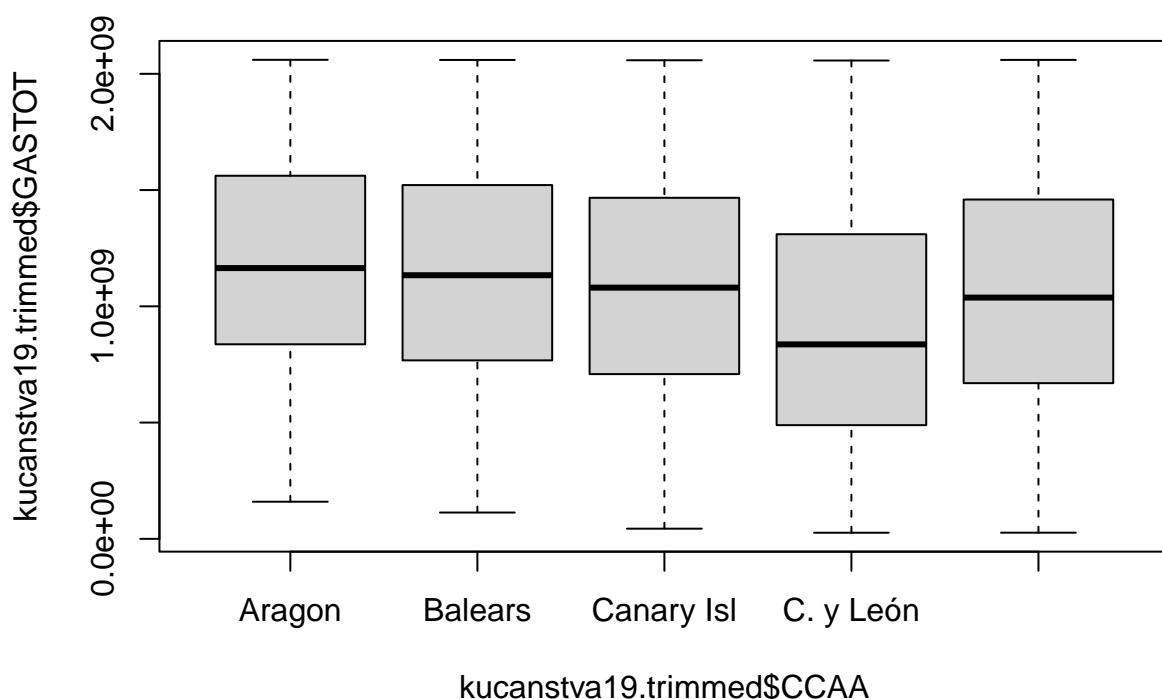
```

Provjerimo postoji li razlika u potrošnji za različite regije u kojima se nalaze kućanstva. Provest ćemo test.

```

boxplot(kucanstva19.trimmed$GASTOT ~ kucanstva19.trimmed$CCAA
       , names = c("Aragon", "Balears", "Canary Isl", "C. y León", "C. La Mancha"))

```



```

a = aov(kucanstva19.trimmed$GASTOT ~ kucanstva19.trimmed$CCAA)
summary(a)

##                                Df      Sum Sq   Mean Sq F value Pr(>F)
## kucanstva19.trimmed$CCAA     4 6.539e+19 1.635e+19   69.27 <2e-16 ***
## Residuals                  5791 1.367e+21 2.360e+17
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Grafički prikaz sugerira da postoji jasna razlika između grupa, pogotovo između regije Castilla y León i ostalih regija, što potvrđuje i ANOVA.

#### 4. Mogu li dostupne značajke predvidjeti ukupni mjesecni prihod kućanstva?

Varijabla koja označava ukupni mjesecni prihod kućanstva je IMPEXAC. Od ostalih varijabli, pokušat ćemo izdvojiti neke koje bi mogle značajnije određivati ukupni mjesecni prihod kućanstva, a onda ćemo ga pokušati i predvidjeti pomoću tih varijabli.

Skraćeni prikaz vrijednosti varijable IMPEXAC:

```

kucanstva19 = read.csv("datasets/hogar_epf_2019.csv", header = TRUE)
summary(kucanstva19$IMPEXAC)

```

```

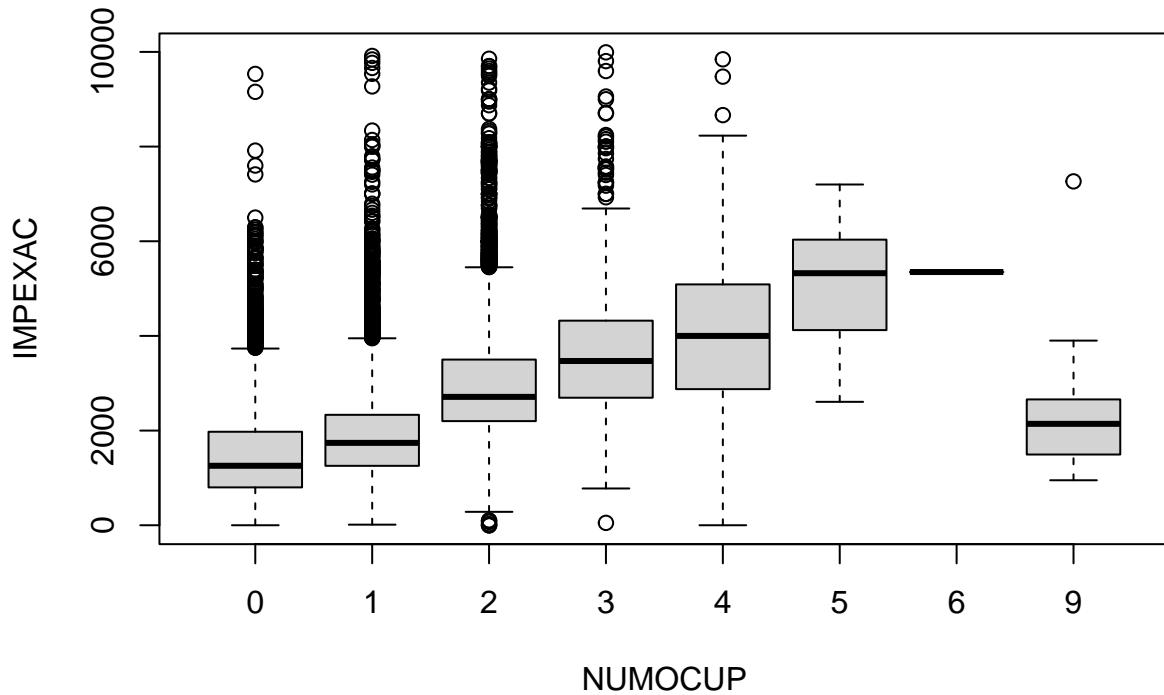
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##        0    1247   1950    2181    2742    9991

```

Logično se čini da bi ukupni prihod kućanstva mogao biti određen brojem članova tog kućanstva, a on je predstavljen varijablom NMIEMB. Još značajnija se čini varijabla NUMOCUP, koja govori koji je broj zaposlenih članova kućanstva.

Prikaz odnosa varijabli IMPEXAC (prihod) i NUMOCUP (broj zaposlenih):

```
boxplot(IMPEXAC~NUMOCUP, data=kucanstva19)
```



Vidimo da nam iskaču slučajevi kada je NUMOCUP == 9 i NUMOCUP == 6. Broj takvih slučajeva je:

```
sum(kucanstva19$NUMOCUP == 9)
```

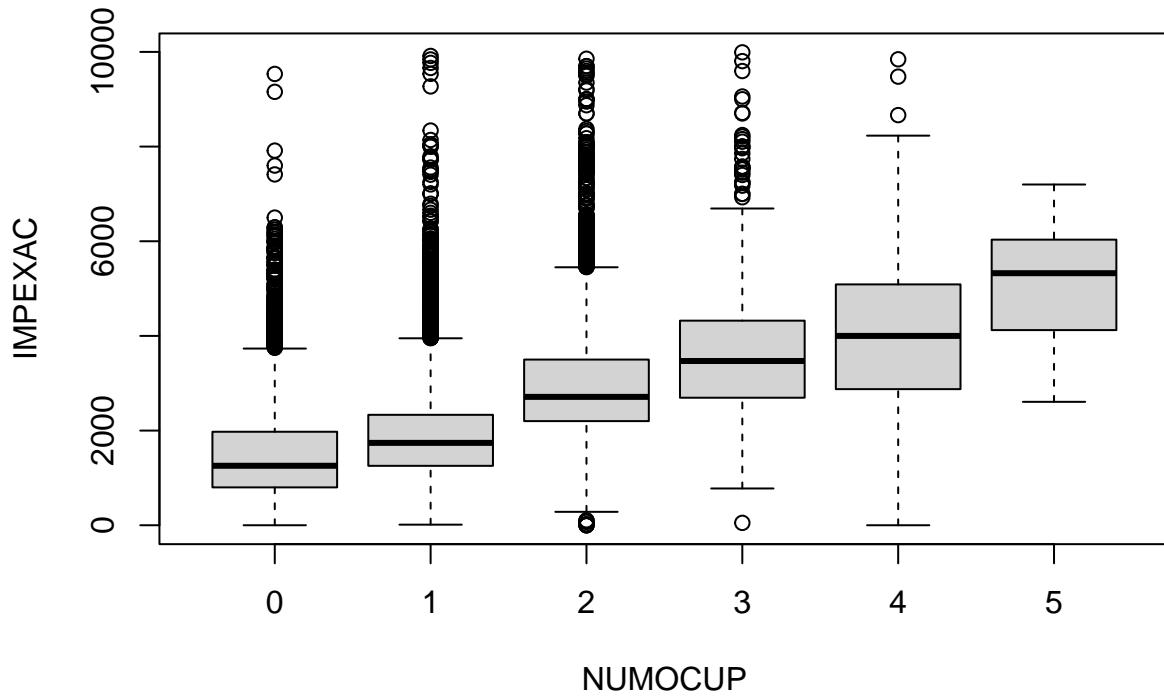
```
## [1] 16
```

```
sum(kucanstva19$NUMOCUP == 6)
```

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

Budući da je broj takvih slučajeva relativno mali, izbacit ćemo podatke s tim rubnim vrijednostima te ponovno prikazati boxplot:

```
kucanstva19.d = kucanstva19[kucanstva19$NUMOCUP != 9,]
kucanstva19.d = kucanstva19.d[kucanstva19.d$NUMOCUP != 6,]
boxplot(IMPEXAC~NUMOCUP, data=kucanstva19.d)
```

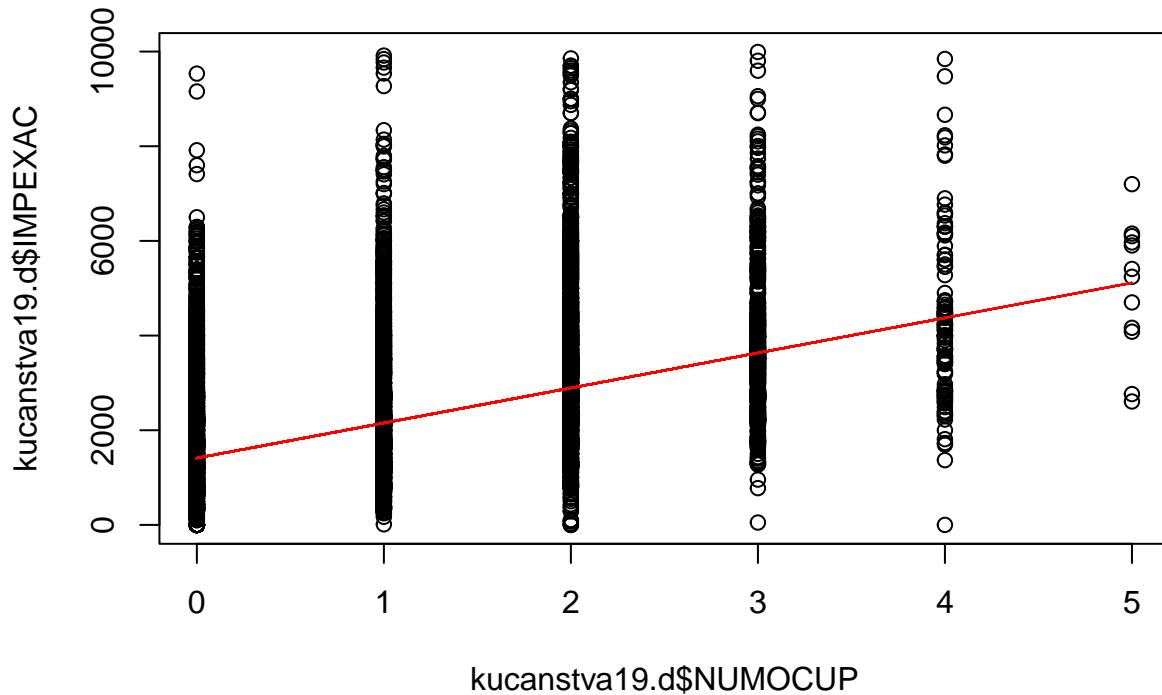


Možemo i izraditi model koji vrši predviđanja prihoda (IMPEXAC) samo na temelju broja zaposlenih u kućanstvu (NUMOCUP) te prikazati dobiveno grafom:

```
fit.numocup = lm(IMPEXAC~NUMOCUP, data=kucanstva19.d)
summary(fit.numocup)
```

```
##
## Call:
## lm(formula = IMPEXAC ~ NUMOCUP, data = kucanstva19.d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -4373.5  -707.0  -204.9   501.3  8123.1 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1411.939   11.807 119.59 <2e-16 ***
## NUMOCUP     740.381    8.553  86.57 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1122 on 20798 degrees of freedom
## Multiple R-squared:  0.2649, Adjusted R-squared:  0.2648 
## F-statistic: 7494 on 1 and 20798 DF,  p-value: < 2.2e-16
```

```
plot(kucanstva19.d$NUMOCUP, kucanstva19.d$IMPEXAC)
lines(kucanstva19.d$NUMOCUP, fit.numocup$fitted.values, col = 'red')
```

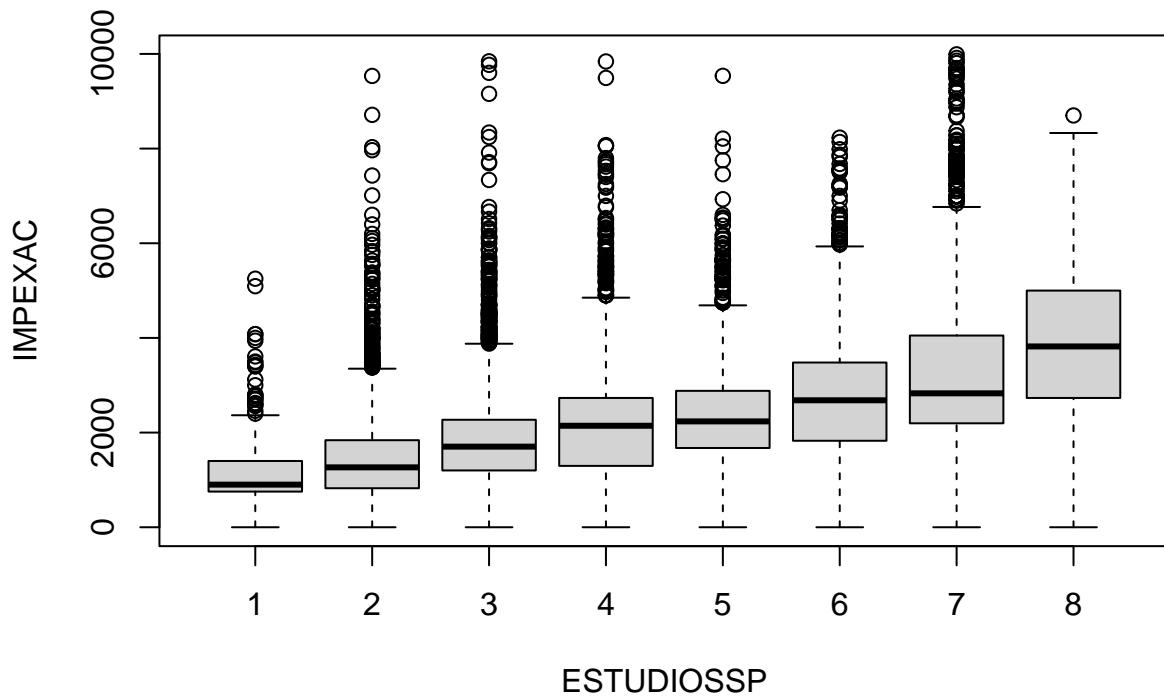


Sljedeća varijabla koja bi mogla odrediti ukupni prihod je obrazovanje primarnog financijskog opskrbljivača kućanstva (main breadwinner), a ono je sadržano u varijabli ESTUDIOSSP i to na sljedeći način:

1 Cannot read or write or went to school for less than 5 years  
 2 Completed primary education or went to school for at least 5 years  
 3 ESO, EGB or Elementary Baccalaureate (with degree or at least 3rd, 8th or 4th respectively)  
 certificates of Primary Studies, Schooling (prior to 1999), or Professionalism (levels 1 or 2) and similar.  
 4 Bachiller, BUP, COU, Bachiller Superior, FP de Grado Medio, FP Básica and other intermediate level studies (Certificate of Professionalism level 3, etc...).  
 5 Advanced Vocational Training, FPII and equivalents  
 6 Degree of 240 ECTS, Diploma, Architecture and Technical Engineering and equivalents.  
 7 Bachelor's degree with more than 240 ECTS, Bachelor's degree, Architecture, Engineering, Master's degree, specialization in Health Sciences and equivalent.  
 8 University doctorate.

Prikaz veze između te varijable i ukupnog prihoda:

```
boxplot(IMPEXAC~ESTUDIOSSP, data=kucanstva19.d)
```



Sljedeća odrednica ukupnog prihoda bi moglo biti zanimanje primarnog finansijskog opskrbljivača kućanstva, sadržano u varijabli OCUPA na sljedeći način:

1 Directors and managers 2 Scientific and intellectual technicians and professionals 3 Technicians and support professionals 4 Accounting, clerical, administrative and other clerical employees 5 Catering, personal, protection, and sales workers 6 Skilled workers in the agriculture, livestock, forestry and fishing sectors 7 Craftsmen and skilled workers in the manufacturing and construction industries (except plant and machinery operators) 8 Plant and machinery operators and assemblers 9 Elementary occupations 0 Military occupations b Not applicable (if WORK=6) -9 Not stated

Provjerimo koliko ima nepoznatih vrijednosti:

```
unique(kucanstva19.d$OCUPA)
```

```
## [1] 5 7 NA 9 8 1 4 2 3 6 0
```

```
sum(is.na(kucanstva19.d$OCUPA))
```

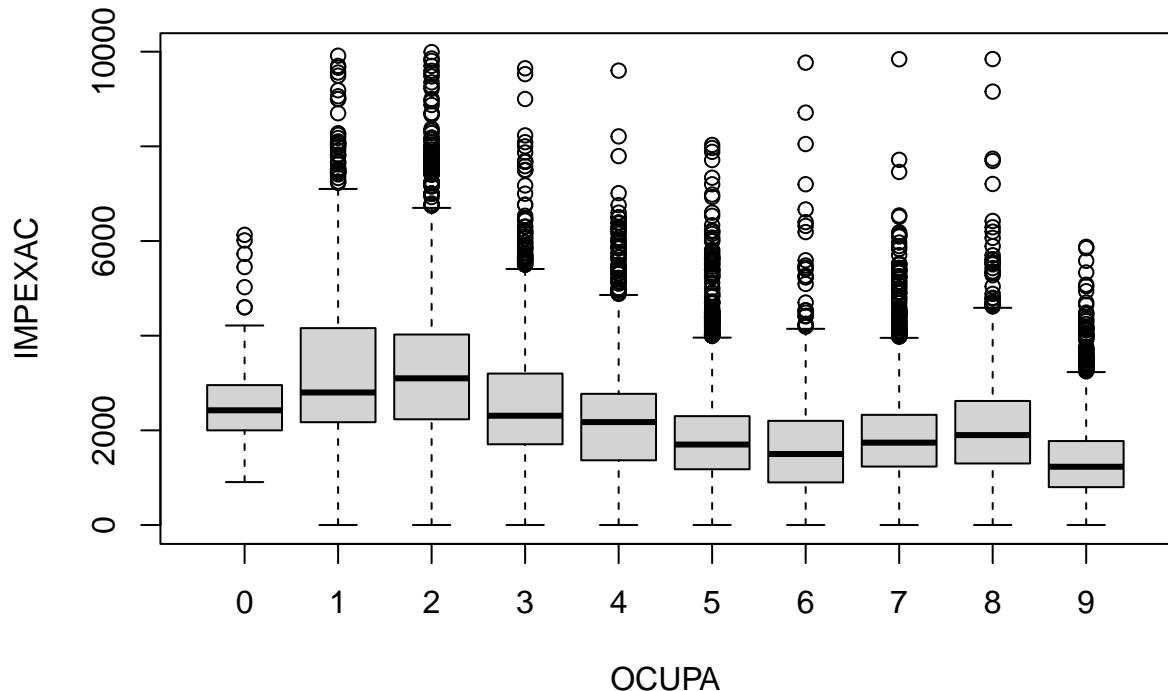
```
## [1] 726
```

U svrhu analize ćemo ih izbaciti, budući da će ostati dovoljno podataka:

```
kucanstva19.d = kucanstva19.d[!is.na(kucanstva19.d$OCUPA),]
```

Prikaz veze između te varijable i ukupnog prihoda:

```
boxplot(IMPEXAC~OCUPA, data=kucanstva19.d)
```

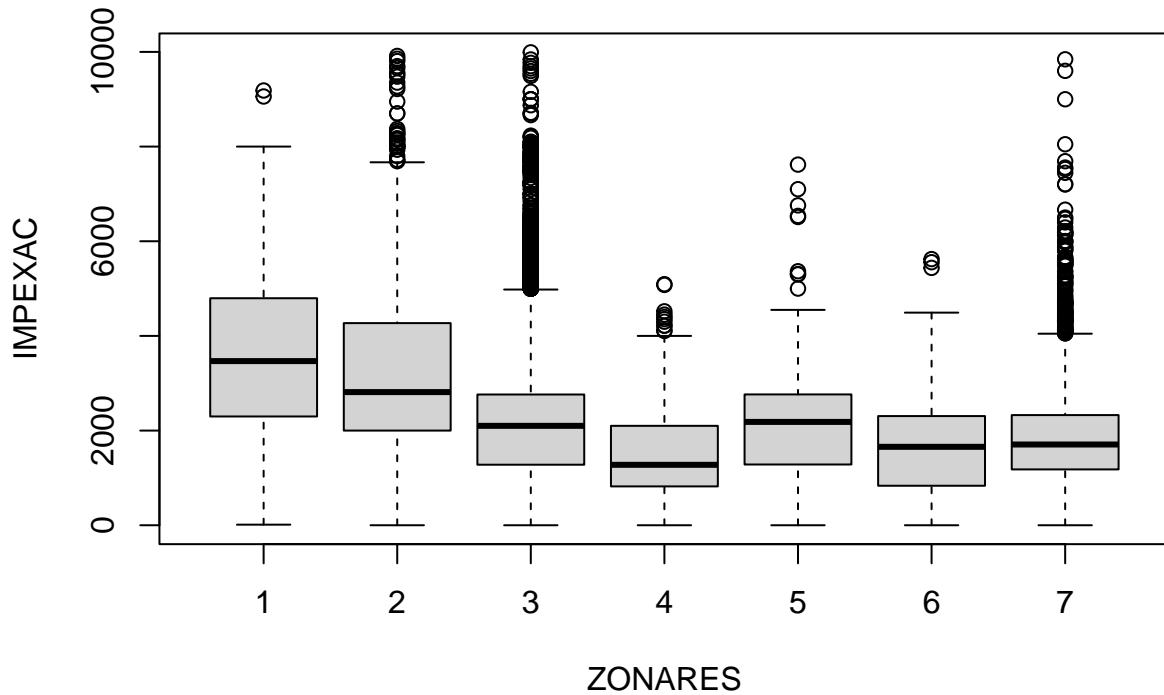


Još jedan kandidat za odrednicu ukupnog prihoda je varijabla ZONARES, koja je određena zonom u kojoj se nalazi stambeni prostor kućanstva. Moguće vrijednosti varijable su:

1 Urban luxury 2 High urban 3 Medium urban 4 Lower urban 5 Rural industrial 6 Rural fishing 7 Rural agricultural -9 Not stated

Odnos te varijable i ukupnog prihoda kućanstva:

```
boxplot(IMPEXAC~ZONARES, data=kucanstva19.d)
```



Dosad navedene varijable koje bi mogle biti značajne u određivanju ukupnog prihoda kućanstva:

NUMOCUP - broj zaposlenih članova kućanstva ESTUDIOSSP - razina obrazovanja primarnog finansijskog opskrbljivača OCUPA - zanimanje primarnog finansijskog opskrbljivača ZONARES - tip zone u kojoj je smješten stambeni prostor kućanstva

Varijable ESTUDIOSSP, OCUPA i ZONARES su kategorijske, pa ćemo za svaku izgraditi dummy varijable kako bismo ih mogli koristiti u modelu za previđanje ukupnog prihoda kućanstva:

```
require(fastDummies)

## Loading required package: fastDummies

kucanstva19.d = dummy_cols(kucanstva19.d, select_columns = c('ESTUDIOSSP', 'ZONARES', 'OCUPA'))
```

Provjerimo jesu li te varijable dovoljno nekorelirane kao preduvjet za korištenje modela linearne regresije:

```
cor(cbind(kucanstva19.d$NUMOCUP_0,kucanstva19.d$NUMOCUP_1,kucanstva19.d$NUMOCUP_2,kucanstva19.d$NUMOCUP_3))
```

```
##          [,1]      [,2]      [,3]      [,4]      [,5]
## [1,] 1.000000e+00 -0.065297496 -0.09742101 -0.0744560811 -0.053696173
## [2,] -6.529750e-02  1.000000000 -0.26533589 -0.2027886005 -0.146246910
## [3,] -9.742101e-02 -0.265335888  1.000000000 -0.3025517300 -0.218193999
## [4,] -7.445608e-02 -0.202788600 -0.30255173  1.0000000000 -0.166759408
## [5,] -5.369617e-02 -0.146246910 -0.21819400 -0.1667594078  1.0000000000
```

```

## [6,] -4.927458e-02 -0.134204255 -0.20022688 -0.1530276575 -0.110360356
## [7,] -6.371650e-02 -0.173538268 -0.25891151 -0.1978786342 -0.142705945
## [8,] -9.916431e-03 -0.016101454 -0.02643341 -0.0087990494 -0.007071808
## [9,] -1.906415e-02 -0.067047991 -0.09044968 -0.0292172557 -0.027649502
## [10,] -7.452147e-02 -0.063207175 -0.04006157 0.0618552088 0.048386362
## [11,] 7.335532e-02 0.056278428 0.01675308 -0.0099742598 -0.023904695
## [12,] -9.287147e-03 -0.015295887 0.02856204 0.0001855433 0.021963985
## [13,] -1.154583e-02 0.022965379 0.01418882 -0.0083360094 -0.006328348
## [14,] -1.269306e-02 -0.032862178 -0.02171908 0.0295016547 0.026905887
## [15,] -3.140725e-02 -0.057297500 -0.06580147 0.0057578882 -0.009841230
## [16,] -6.765382e-02 -0.177815133 -0.26286160 -0.1669974713 -0.098013393
## [17,] -4.822927e-02 -0.092866056 -0.08757568 0.0530077300 0.146107569
## [18,] -4.053317e-02 -0.075772888 -0.06994844 0.1121196618 0.072285315
## [19,] -2.352405e-02 -0.004837173 0.06711412 0.1143638479 0.002034527
## [20,] 5.777248e-02 0.131974949 0.05007585 -0.0439334906 -0.042610265
## [21,] 5.039012e-05 0.096270313 0.14140962 -0.0199601550 0.024931383
## [22,] -5.623604e-03 0.074052693 0.10613440 0.0073029030 -0.012269453
## [23,] 1.719751e-01 0.138229141 0.11575160 -0.0532006205 -0.077989405
## [6] [7] [8] [9] [10]
## [1,] -0.0492745788 -0.063716497 -0.009916431 -0.019064151 -0.07452147
## [2,] -0.1342042547 -0.173538268 -0.016101454 -0.067047991 -0.06320718
## [3,] -0.2002268833 -0.258911512 -0.026433412 -0.090449675 -0.04006157
## [4,] -0.1530276575 -0.197878634 -0.008799049 -0.029217256 0.06185521
## [5,] -0.1103603555 -0.142705945 -0.007071808 -0.027649502 0.04838636
## [6,] 1.00000000000 -0.130954869 0.003944194 0.062150562 0.02543133
## [7,] -0.1309548691 1.0000000000 0.060200859 0.178876227 0.01671819
## [8,] 0.0039441940 0.060200859 1.000000000 -0.015174145 -0.10824537
## [9,] 0.0621505616 0.178876227 -0.015174145 1.000000000 -0.40045704
## [10,] 0.0254313348 0.016718186 -0.108245367 -0.400457043 1.000000000
## [11,] -0.0332428588 -0.046836713 -0.010889043 -0.040284348 -0.28737000
## [12,] -0.0110087522 -0.024604385 -0.008860321 -0.032779027 -0.23383049
## [13,] -0.0144308910 -0.011595551 -0.004775594 -0.017667456 -0.12603150
## [14,] -0.0006829777 0.009726161 -0.005250113 0.002392787 0.02058310
## [15,] 0.0435483694 0.120952209 0.033543779 0.127067803 -0.03087528
## [16,] 0.3092094094 0.509288517 0.052910213 0.157332617 0.01716329
## [17,] 0.0465991893 0.008211084 -0.006343312 0.024629556 0.04622246
## [18,] 0.0170972965 -0.011140886 -0.010107543 -0.005355903 0.04388885
## [19,] -0.0689351306 -0.130744799 -0.010601380 -0.047517464 0.05166453
## [20,] -0.0579744966 -0.083142099 -0.010251398 -0.043261485 -0.18624862
## [21,] -0.1182544338 -0.169017716 -0.013856394 -0.069223990 0.01014004
## [22,] -0.0877672194 -0.124802789 -0.010230807 -0.049957868 0.00371391
## [23,] -0.1017579440 -0.141274337 -0.021689180 -0.071479925 -0.03907725
## [11] [12] [13] [14] [15]
## [1,] 0.0733553174 -9.287147e-03 -0.011545830 -0.0126930621 -0.031407253
## [2,] 0.0562784281 -1.529589e-02 0.022965379 -0.0328621776 -0.057297500
## [3,] 0.0167530841 2.856204e-02 0.014188816 -0.0217190785 -0.065801472
## [4,] -0.0099742598 1.855433e-04 -0.008336009 0.0295016547 0.005757888
## [5,] -0.0239046953 2.196398e-02 -0.006328348 0.0269058869 -0.009841230
## [6,] -0.0332428588 -1.100875e-02 -0.014430891 -0.0006829777 0.043548369
## [7,] -0.0468367129 -2.460439e-02 -0.011595551 0.0097261612 0.120952209
## [8,] -0.0108890432 -8.860321e-03 -0.004775594 -0.0052501127 0.033543779
## [9,] -0.0402843481 -3.277903e-02 -0.017667456 0.0023927869 0.127067803
## [10,] -0.2873700021 -2.338305e-01 -0.126031498 0.0205831015 -0.030875278
## [11,] 1.00000000000 -2.352240e-02 -0.012678256 0.0008704801 -0.023697538

```

```

## [12,] -0.0235223956 1.000000e+00 -0.010316187 0.0066851296 -0.006639133
## [13,] -0.0126782556 -1.031619e-02 1.0000000000 0.0021370465 -0.013455822
## [14,] 0.0008704801 6.685130e-03 0.002137047 1.0000000000 -0.018270997
## [15,] -0.0236975379 -6.639133e-03 -0.013455822 -0.0182709975 1.0000000000
## [16,] -0.0479960499 -2.244049e-02 -0.010608949 -0.0358183504 -0.097384131
## [17,] -0.0210676952 5.933973e-05 -0.014276011 -0.0278264591 -0.075655509
## [18,] -0.0107973555 -3.887646e-03 -0.011920022 -0.0240430968 -0.065369177
## [19,] 0.0022703923 -9.164403e-04 -0.005552629 -0.0373971990 -0.101676757
## [20,] -0.0082498047 -5.602247e-03 0.057782507 -0.0179148819 -0.048707581
## [21,] 0.0125374130 1.329974e-02 0.002875719 -0.0365514108 -0.099377200
## [22,] -0.0041265782 2.479422e-02 0.007684633 -0.0266826639 -0.072545721
## [23,] 0.0890360712 -1.819084e-03 0.002591678 -0.0307987377 -0.083736640
## [,16]      [,17]      [,18]      [,19]      [,20]
## [1,] -0.06765382 -4.822927e-02 -0.040533175 -0.0235240520 0.057772476
## [2,] -0.17781513 -9.286606e-02 -0.075772888 -0.0048371731 0.131974949
## [3,] -0.26286160 -8.757568e-02 -0.069948439 0.0671141167 0.050075853
## [4,] -0.16699747 5.300773e-02 0.112119662 0.1143638479 -0.043933491
## [5,] -0.09801339 1.461076e-01 0.072285315 0.0020345265 -0.042610265
## [6,] 0.30920941 4.659919e-02 0.017097296 -0.0689351306 -0.057974497
## [7,] 0.50928852 8.211084e-03 -0.011140886 -0.1307447994 -0.083142099
## [8,] 0.05291021 -6.343312e-03 -0.010107543 -0.0106013795 -0.010251398
## [9,] 0.15733262 2.462956e-02 -0.005355903 -0.0475174638 -0.043261485
## [10,] 0.01716329 4.622246e-02 0.043888852 0.0516645317 -0.186248618
## [11,] -0.04799605 -2.106770e-02 -0.010797356 0.0022703923 -0.008249805
## [12,] -0.02244049 5.933973e-05 -0.003887646 -0.0009164403 -0.005602247
## [13,] -0.01060895 -1.427601e-02 -0.011920022 -0.0055526292 0.057782507
## [14,] -0.03581835 -2.782646e-02 -0.024043097 -0.0373971990 -0.017914882
## [15,] -0.09738413 -7.565551e-02 -0.065369177 -0.1016767575 -0.048707581
## [16,] 1.00000000 -1.483146e-01 -0.128149330 -0.1993264864 -0.095486041
## [17,] -0.14831459 1.000000e+00 -0.099556290 -0.1548521987 -0.074180926
## [18,] -0.12814933 -9.955629e-02 1.000000000 -0.1337980657 -0.064095082
## [19,] -0.19932649 -1.548522e-01 -0.133798066 1.00000000000 -0.099695001
## [20,] -0.09548604 -7.418093e-02 -0.064095082 -0.0996950008 1.000000000
## [21,] -0.19481845 -1.513500e-01 -0.130772042 -0.2034059133 -0.097440264
## [22,] -0.14221818 -1.104861e-01 -0.095464070 -0.1484870627 -0.071131750
## [23,] -0.16415679 -1.275297e-01 -0.110190379 -0.1713927110 -0.082104550
## [,21]      [,22]      [,23]
## [1,] 5.039012e-05 -0.005623604 0.171975054
## [2,] 9.627031e-02 0.074052693 0.138229141
## [3,] 1.414096e-01 0.106134404 0.115751599
## [4,] -1.996015e-02 0.007302903 -0.053200620
## [5,] 2.493138e-02 -0.012269453 -0.077989405
## [6,] -1.182544e-01 -0.087767219 -0.101757944
## [7,] -1.690177e-01 -0.124802789 -0.141274337
## [8,] -1.385639e-02 -0.010230807 -0.021689180
## [9,] -6.922399e-02 -0.049957868 -0.071479925
## [10,] 1.014004e-02 0.003713910 -0.039077253
## [11,] 1.253741e-02 -0.004126578 0.089036071
## [12,] 1.329974e-02 0.024794224 -0.001819084
## [13,] 2.875719e-03 0.007684633 0.002591678
## [14,] -3.655141e-02 -0.026682664 -0.030798738
## [15,] -9.937720e-02 -0.072545721 -0.083736640
## [16,] -1.948184e-01 -0.142218182 -0.164156791
## [17,] -1.513500e-01 -0.110486060 -0.127529665

```

```

## [18,] -1.307720e-01 -0.095464070 -0.110190379
## [19,] -2.034059e-01 -0.148487063 -0.171392711
## [20,] -9.744026e-02 -0.071131750 -0.082104550
## [21,] 1.000000e+00 -0.145128827 -0.167516433
## [22,] -1.451288e-01 1.000000000 -0.122287611
## [23,] -1.675164e-01 -0.122287611 1.000000000

```

Vidimo kako nema značajnije koreliranosti između varijabli.

Iduće, razumno je pretpostaviti da će model imati poteškoća s predviđanjem gornjih rubnih vrijednosti ukupnog prihoda kućanstva, budući da bi ekstremniji prihodi mogli biti rezultat nekih specifičnosti osobnog života koji bi odredili zbog čega se netko toliko odvojio od medijana ukupnog prihoda kućanstva. Takve visoke vrijednosti bi nam mogle i predstavljati problem pri generiranju modela. Stoga ćemo iz uzorka maknuti gornjih 10% vrijednosti.

```

quant <- quantile(kucanstva19.d$IMPEXAC, c(0, 0.9)) #određivanje kvantila
kucanstva19.d = kucanstva19.d[kucanstva19.d$IMPEXAC > quant[1] & kucanstva19.d$IMPEXAC < quant[2], ] #os

```

Za predviđanje ukupnog prihoda kućanstva ćemo koristiti model multiple linearne regresije koja modelira vrijednost varijable IMPEXAC prema vrijednostima varijabli NUMOCUP, ESTUDIOSSP, OCUPA i ZONARES, s time da se za kategoriske varijable koriste njihove dummy varijable:

```

fit.multi = lm(IMPEXAC ~ NUMOCUP + ESTUDIOSSP_1 + ESTUDIOSSP_2 + ESTUDIOSSP_3 + ESTUDIOSSP_4 + ESTUDIOSS
summary(fit.multi)

```

```

##
## Call:
## lm(formula = IMPEXAC ~ NUMOCUP + ESTUDIOSSP_1 + ESTUDIOSSP_2 +
##     ESTUDIOSSP_3 + ESTUDIOSSP_4 + ESTUDIOSSP_5 + ESTUDIOSSP_6 +
##     ESTUDIOSSP_7 + ZONARES_1 + ZONARES_2 + ZONARES_3 + ZONARES_4 +
##     ZONARES_5 + ZONARES_6 + OCUPA_0 + OCUPA_1 + OCUPA_2 + OCUPA_3 +
##     OCUPA_4 + OCUPA_5 + OCUPA_6 + OCUPA_7 + OCUPA_8 + OCUPA_9,
##     data = kucanstva19.d)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -3410.6  -471.4   -76.0   430.0  2769.4 
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1325.689    70.767 18.733 < 2e-16 ***
## NUMOCUP     451.694     6.127 73.726 < 2e-16 ***
## ESTUDIOSSP_1 -457.538    75.190 -6.085 1.19e-09 ***
## ESTUDIOSSP_2 -391.800    69.015 -5.677 1.39e-08 ***
## ESTUDIOSSP_3 -384.824    68.204 -5.642 1.70e-08 ***
## ESTUDIOSSP_4 -292.030    68.146 -4.285 1.83e-05 ***
## ESTUDIOSSP_5 -256.047    68.741 -3.725 0.000196 ***
## ESTUDIOSSP_6 -174.989    67.713 -2.584 0.009766 ** 
## ESTUDIOSSP_7 -211.697    66.968 -3.161 0.001574 ** 
## ZONARES_1    313.836    94.729  3.313 0.000925 ***
## ZONARES_2    223.073    28.829  7.738 1.06e-14 ***
## ZONARES_3    70.929     14.479  4.899 9.74e-07 ***
## ZONARES_4   -176.246    31.593 -5.579 2.46e-08 ***

```

```

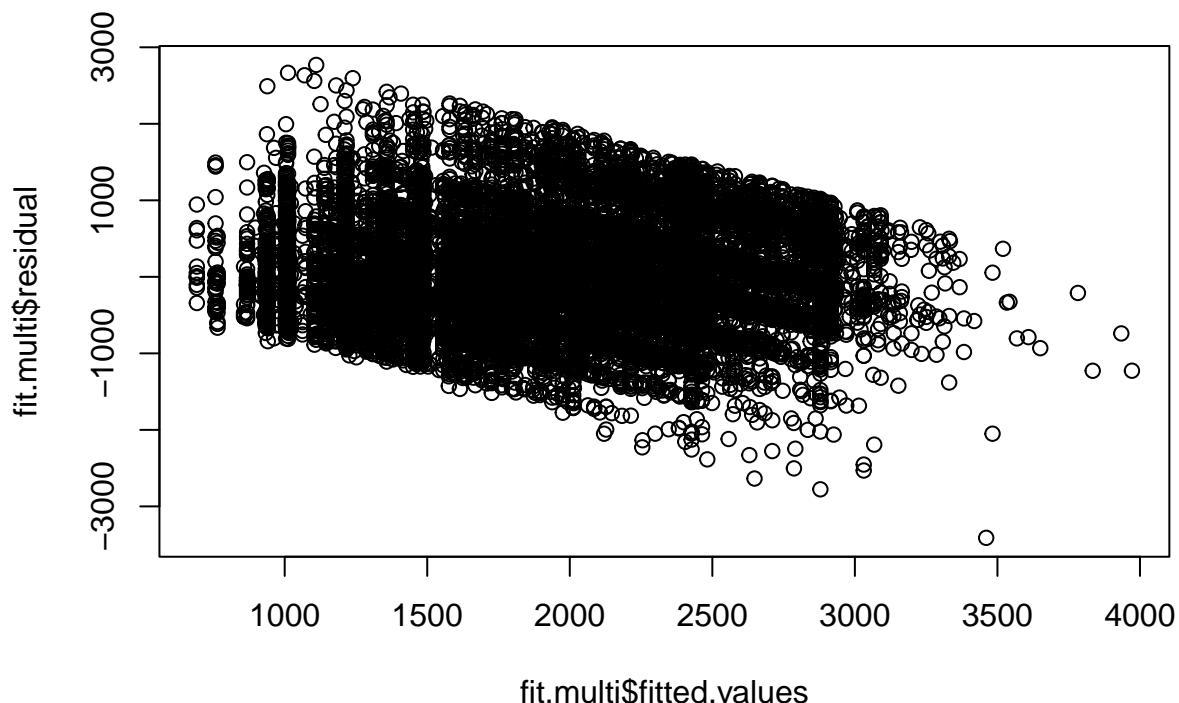
## ZONARES_5      80.771   38.296   2.109  0.034949 *
## ZONARES_6     -13.177   67.357  -0.196  0.844902
## OCUPA_0        810.544   64.038  12.657  < 2e-16 ***
## OCUPA_1        768.442   30.704  25.027  < 2e-16 ***
## OCUPA_2        790.629   25.358  31.178  < 2e-16 ***
## OCUPA_3        621.878   22.131  28.100  < 2e-16 ***
## OCUPA_4        474.090   23.421  20.242  < 2e-16 ***
## OCUPA_5        205.158   18.394  11.154  < 2e-16 ***
## OCUPA_6        239.783   26.988  8.885  < 2e-16 ***
## OCUPA_7        346.003   18.242  18.968  < 2e-16 ***
## OCUPA_8        471.157   20.890  22.554  < 2e-16 ***
## OCUPA_9          NA       NA       NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 667.8 on 17925 degrees of freedom
## Multiple R-squared:  0.3957, Adjusted R-squared:  0.395
## F-statistic: 510.4 on 23 and 17925 DF,  p-value: < 2.2e-16

```

Kako bi se modeli mogli analizirati i uspoređivati, bitno je provjeriti jesu li zadovoljene pretpostavke o rezidualima modela.

Prvo ćemo provjeriti homogenost varijance:

```
plot(fit.multi$residual ~ fit.multi$fitted.values)
```

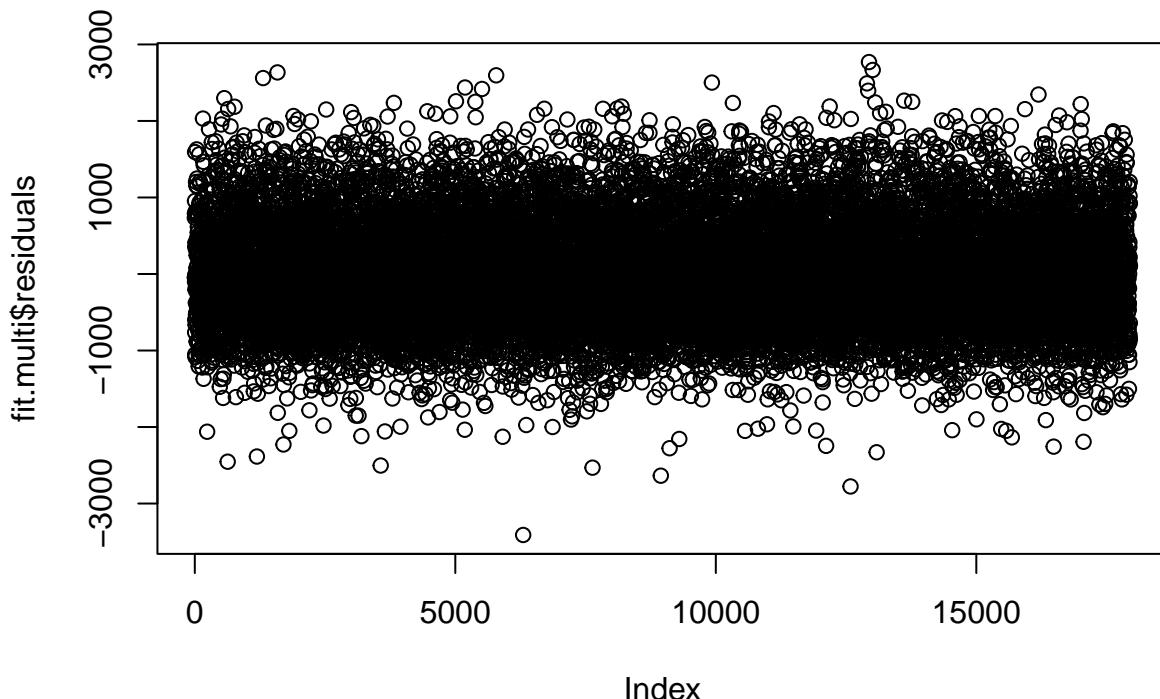


Čini se da ne postoji značajnija heterogenost varijanci.

Treba provjeriti i je li zadovoljena normalnost reziduala.

Prikaz reziduala dobivenog modela:

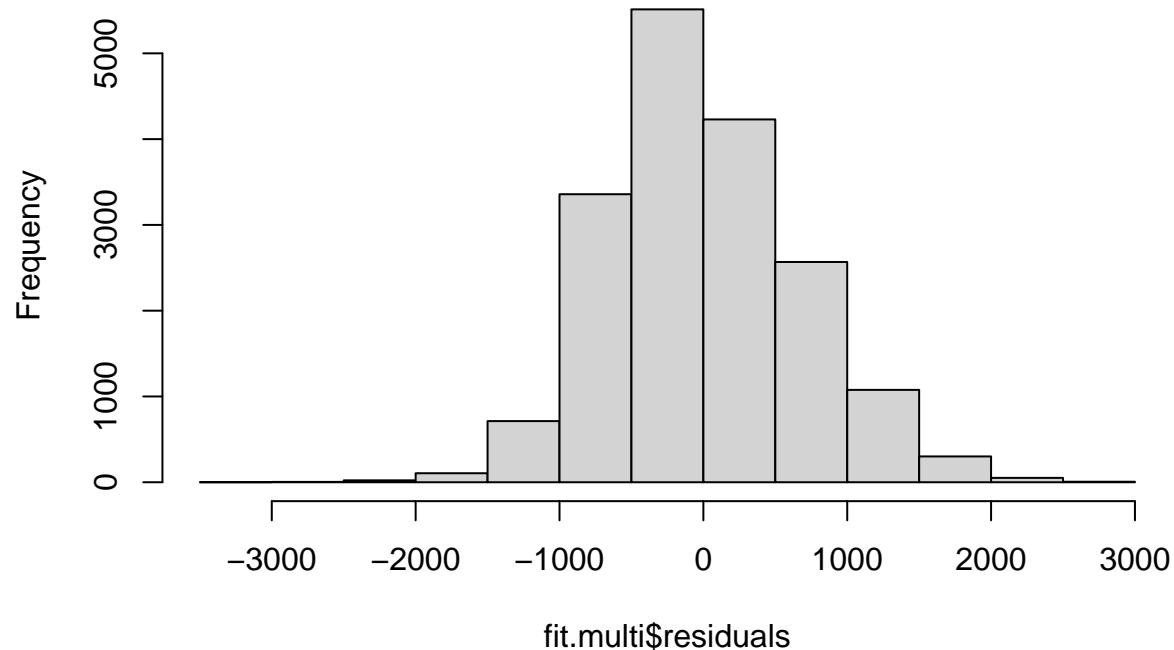
```
plot(fit.multi$residuals)
```



Bolji je prikaz histogramom:

```
hist(fit.multi$residuals)
```

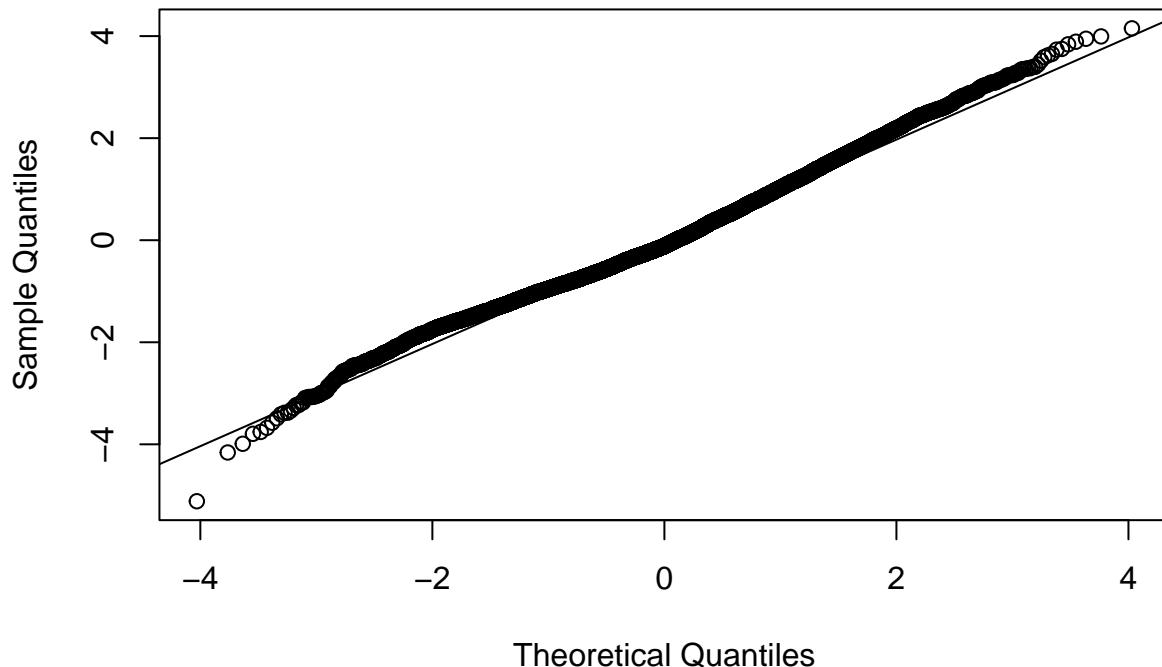
### Histogram of fit.multi\$residuals



q-q plot reziduala s linijom normalne distribucije:

```
qqnorm(rstandard(fit.multi))
qqline(rstandard(fit.multi))
```

## Normal Q-Q Plot



Sada provodimo i testove normalnosti:

```
ks.test(unique(rstandard(fit.multi)), 'pnorm')

##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data: unique(rstandard(fit.multi))
## D = 0.033743, p-value = 2.22e-16
## alternative hypothesis: two-sided

require(nortest)
lillie.test(rstandard(fit.multi))

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data: rstandard(fit.multi)
## D = 0.04568, p-value < 2.2e-16
```

Iz q-q plota i histograma se čini da je distribucija prihvatljiva i da bi se dobiveni model mogao koristiti kao legitimno rješenje.

Za razliku od toga, Kolmogorov-Smirnovljev i Lillieforsov test nam govore da distribucija nije normalna. Takvi rezultati testova tj. mali p-value su posljedica velike veličine uzorka, ali iz q-q plota možemo ipak zaključiti da odstupanje nije toliko da se model ne bi mogao koristiti kao legitimno rješenje.

ZAKLJUČAK: Dobiveni model objašnjava oko 40% varijance u podatcima i s obzirom na to da je riječ o društvenom/ekonomskom području, možemo procijeniti da je to zadovoljavajući rezultat.