Analiza potrošačkih navika kupaca

WhatSAP

Tara Baće, Sara Gašpar, Mirta Hrnčić, Tin Salopek

2024-12-16

Motivacija i opis problema

Analiza potrošačkih navika kupaca omogućuje poduzećima prilagodbu ponude proizvoda različitim segmentima potrošača, s ciljem zadržavanja što većeg broja kupaca i poboljšanja prodajnih rezultata. Razumijevanje potrošačkih preferencija povećava učinkovitost marketinških kampanja kroz ciljano oglašavanje, optimizira asortiman te pospješuje upravljanje zalihama proizvoda.

Učitavanje i uređivanje podataka

Učitavamo podatke i pohranjujemo ih u varijablu "data".

```
# Učitavanje podataka
data <- read.csv("data.csv")
```

Prikazujemo sažeti prikaz podataka kako bi dobili uvid u građu i sadržaj.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
glimpse(data)
```

```
## Rows: 2,240
## Columns: 27
## $ ID                 <int> 5524, 2174, 4141, 6182, 5324, 7446, 965, 6177, 485~
## $ Year_Birth         <int> 1957, 1954, 1965, 1984, 1981, 1967, 1971, 1985, 19~
```

```
## $ Education           <chr> "Graduation", "Graduation", "Graduation", "Graduat~
## $ Marital_Status      <chr> "Single", "Single", "Together", "Together", "Marri~
## $ Income              <dbl> 58138, 46344, 71613, 26646, 58293, 62513, 55635, 3~
## $ Kidhome             <int> 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1,~
## $ Teenhome            <int> 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1,~
## $ Dt_Customer         <chr> "2012-09-04", "2014-03-08", "2013-08-21", "2014-02~
## $ Recency             <int> 58, 38, 26, 26, 94, 16, 34, 32, 19, 68, 11, 59, 82~
## $ Complain            <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ MntWines            <int> 635, 11, 426, 11, 173, 520, 235, 76, 14, 28, 5, 6,~
## $ MntFruits           <int> 88, 1, 49, 4, 43, 42, 65, 10, 0, 0, 5, 16, 61, 2, ~
## $ MntMeatProducts     <int> 546, 6, 127, 20, 118, 98, 164, 56, 24, 6, 6, 11, 4~
## $ MntFishProducts     <int> 172, 2, 111, 10, 46, 0, 50, 3, 3, 1, 0, 11, 225, 3~
## $ MntSweetProducts    <int> 88, 1, 21, 3, 27, 42, 49, 1, 3, 1, 2, 1, 112, 5, 1~
## $ MntGoldProds        <int> 88, 6, 42, 5, 15, 14, 27, 23, 2, 13, 1, 16, 30, 14~
## $ NumWebPurchases     <int> 8, 1, 8, 2, 5, 6, 7, 4, 3, 1, 1, 2, 3, 6, 1, 7, 3,~
## $ NumCatalogPurchases <int> 10, 1, 2, 0, 3, 4, 3, 0, 0, 0, 0, 0, 4, 1, 0, 6, 0~
## $ NumStorePurchases   <int> 4, 2, 10, 4, 6, 10, 7, 4, 2, 0, 2, 3, 8, 5, 3, 12,~
## $ NumWebVisitsMonth   <int> 7, 5, 4, 6, 5, 6, 6, 8, 9, 20, 7, 8, 2, 6, 8, 3, 8~
## $ NumDealsPurchases   <int> 3, 2, 1, 2, 5, 2, 4, 2, 1, 1, 1, 1, 1, 3, 1, 1, 3,~
## $ AcceptedCmp1        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
## $ AcceptedCmp2        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ AcceptedCmp3        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ AcceptedCmp4        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ AcceptedCmp5        <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
## $ Response            <int> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0,~
```

Provjeravamo postoje li nedostajeći podatci.

```
sum(!complete.cases(data))
```

```
## [1] 24
```

```
for (name in names(data)) {
  if (sum(is.na(data[[name]])) > 0) {
    print(name)
  }
}
```

```
## [1] "Income"
```

S obzirom na mali udio nedostojećih podataka te kako bismo olakšali kasniju analizu zavisnosti prihoda, iz podataka uklanjamo nedostajeće vrijednosti.

```
data.full = data[complete.cases(data),]
```

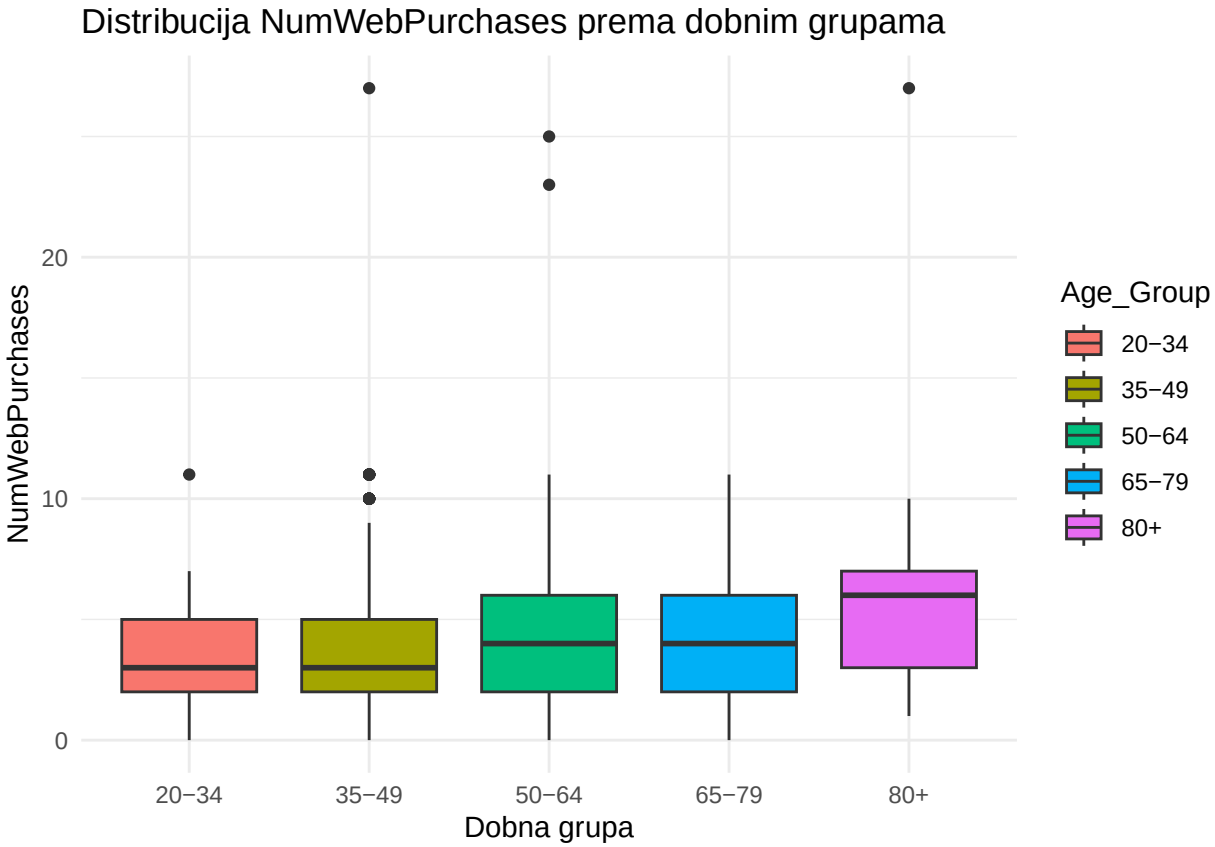
Zavisnost izmedu dobi klijenta i najčešće korištenih prodajnih kanala

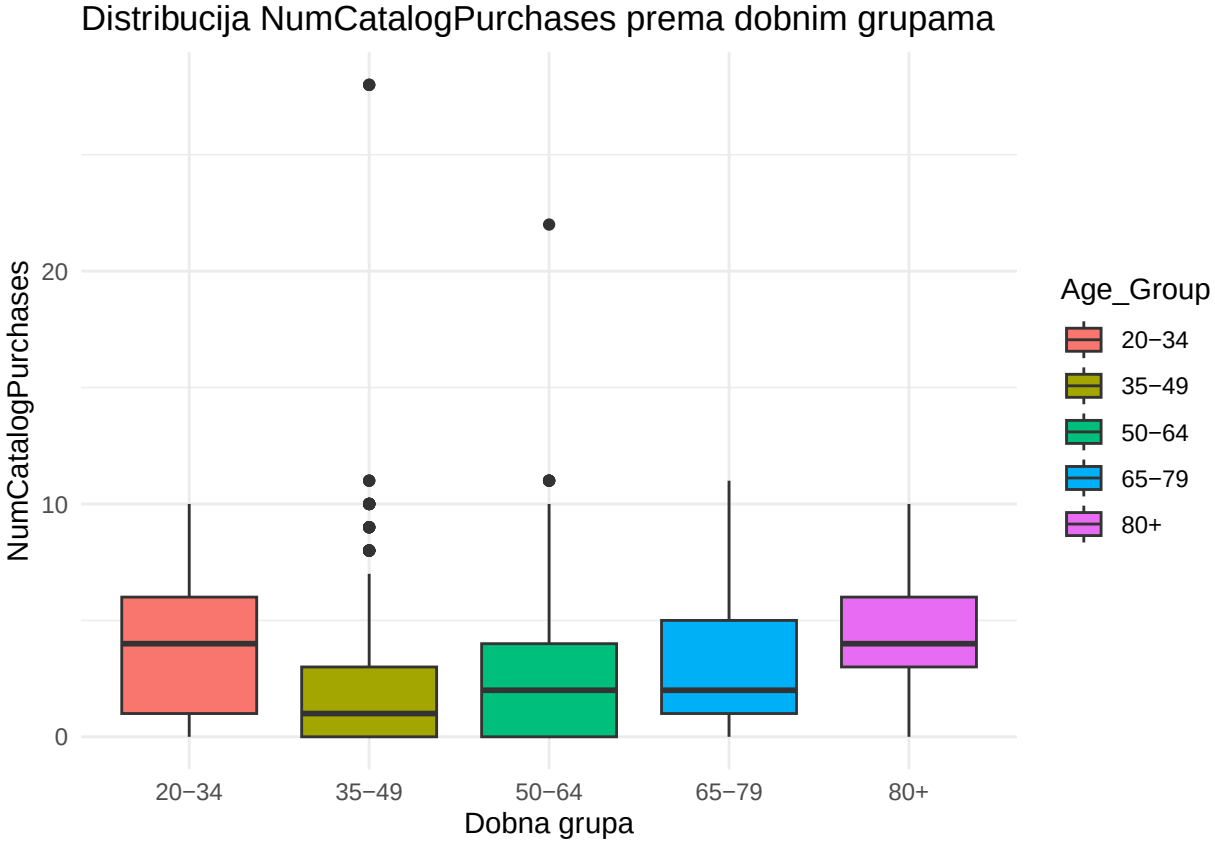
Cilj ove analize bio je ispitati postoji li zavisnost između dobi klijenata i njihovih preferencija za određene prodajne kanale, uključujući kupovine putem interneta, kataloga i u trgovinama, kao i učestalost posjeta web stranici.

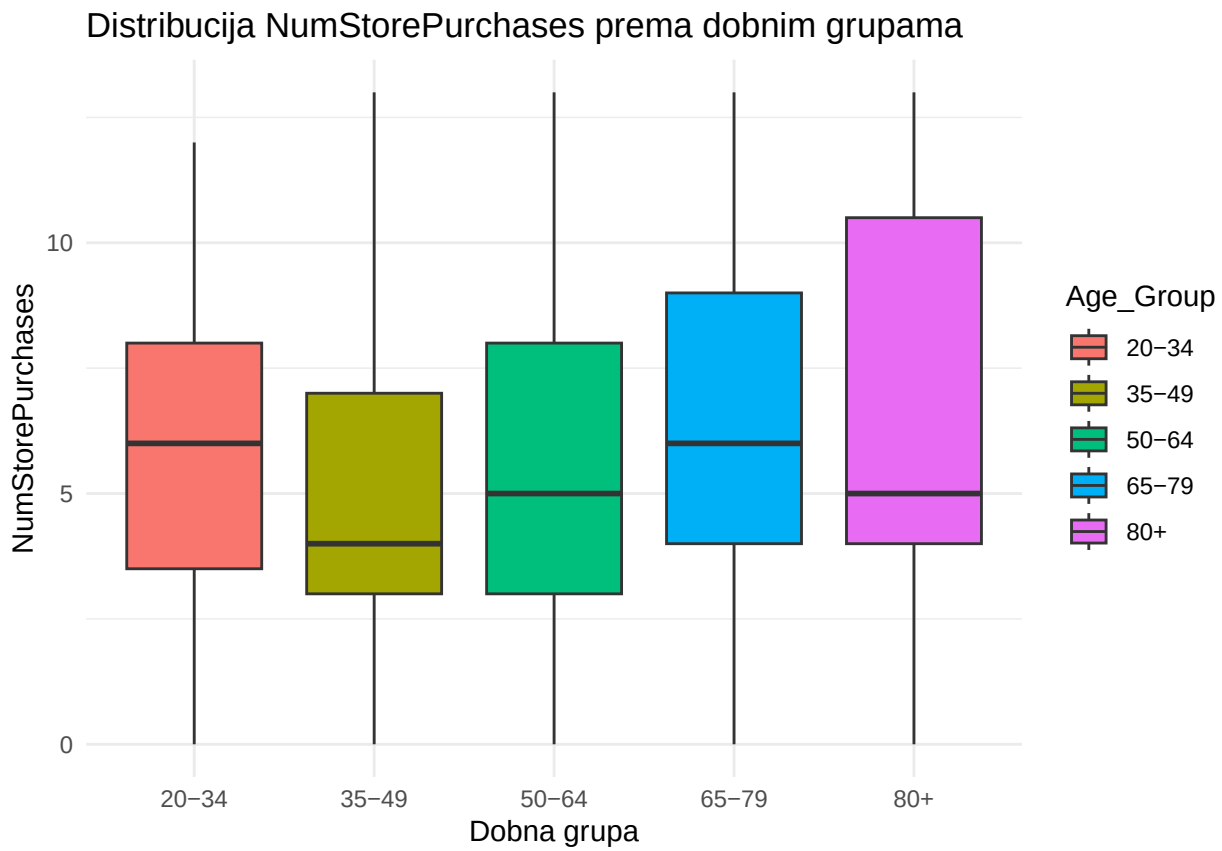
Kako bismo zaključili nešto o zavisnosti između ove dvije varijable, prvo je bilo potrebno prilagoditi podatke koje imamo. Dobi klijenata raspoređene su u starosne skupine na temelju godine rođenja koja je jedan od stupaca u dobivenim podacima. Intervali dobnih skupina koje su definirane su 20-34, 35-49, 50-64, 65-79 i više od 80 godina.

Nakon toga, podaci su analizirani pomoću vizualizacija i statističkih testova.

Vizualizacija - boxplot grafovi







Za svaki prodajni kanal (kupovine putem interneta, kataloga, trgovina i posjete web stranici) napravljeni su kutijasti dijagrami (boxplots) kako bi se prikazala distribucija podataka po dobnim grupama. Rezultati su pokazali:

Kupovine putem interneta (NumWebPurchases): Mlađe grupe imaju veću sklonost ka ovom kanalu, dok stariji klijenti rjeđe koriste internet za kupovinu. Kupovine putem kataloga (NumCatalogPurchases): Ovaj kanal je popularniji među starijim klijentima, dok mlađe grupe imaju znatno manje kupovina na ovaj način. Kupovine u trgovinama (NumStorePurchases): Ovaj kanal pokazuje slične obrasce među svim dobnim grupama, s blagim porastom kod srednjih i starijih grupa. ## ANOVA test

```
## $NumWebPurchases
##               Df Sum Sq Mean Sq F value   Pr(>F)
## Age_Group      4    374   93.43   12.35 6.23e-10 ***
## Residuals   2235  16914    7.57
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## $NumCatalogPurchases
##               Df Sum Sq Mean Sq F value   Pr(>F)
## Age_Group      4    557  139.29   16.76 1.52e-13 ***
## Residuals   2235  18574    8.31
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## $NumStorePurchases
##               Df Sum Sq Mean Sq F value  Pr(>F)
## Age_Group      4    437  109.37   10.52 1.9e-08 ***
```

```
## Residuals   2235  23226   10.39
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Rezultati ANOVA testa za sva četiri kanala pokazuju statistički značajne razlike među dobnim grupama (p-vrijednosti < 0.05). To znači da postoji zavisnost između dobi klijenata i preferiranih prodajnih kanala.

X2- test

```
##
##         NumCatalogPurchases NumStorePurchases NumWebPurchases
##   20-45                  26               295             103
##   46-64                  55               802             340
##   65+                    56               404             159
```

Na kraju smo proveli hi-kvadrat test:

```
chi_test <- chisq.test(contingency_table)

print(chi_test)
```

```
##
##  Pearson's Chi-squared test
##
## data:  contingency_table
## X-squared = 16.339, df = 4, p-value = 0.002596
```

Kada smo dobili kontigencijsku tablicu, mogli smo napraviti test kojime ćemo utvrditi zavisnost varijabli. Hi-kvadrat test je statistički test koji se koristi za utvrđivanje postoji li značajna povezanost između kategoričkih varijabli. Uspoređuje opažene frekvencije pojavljivanja u različitim kategorijama s frekvencijama koje se očekuju ako ne postoje povezanosti između varijabli. Uvjet za pouzdanost je da je očekivana frekvencija u svakoj ćeliji kontigencijske tablice veća od 5. To smo provjerili:

```
min(chi_test$expected)
```

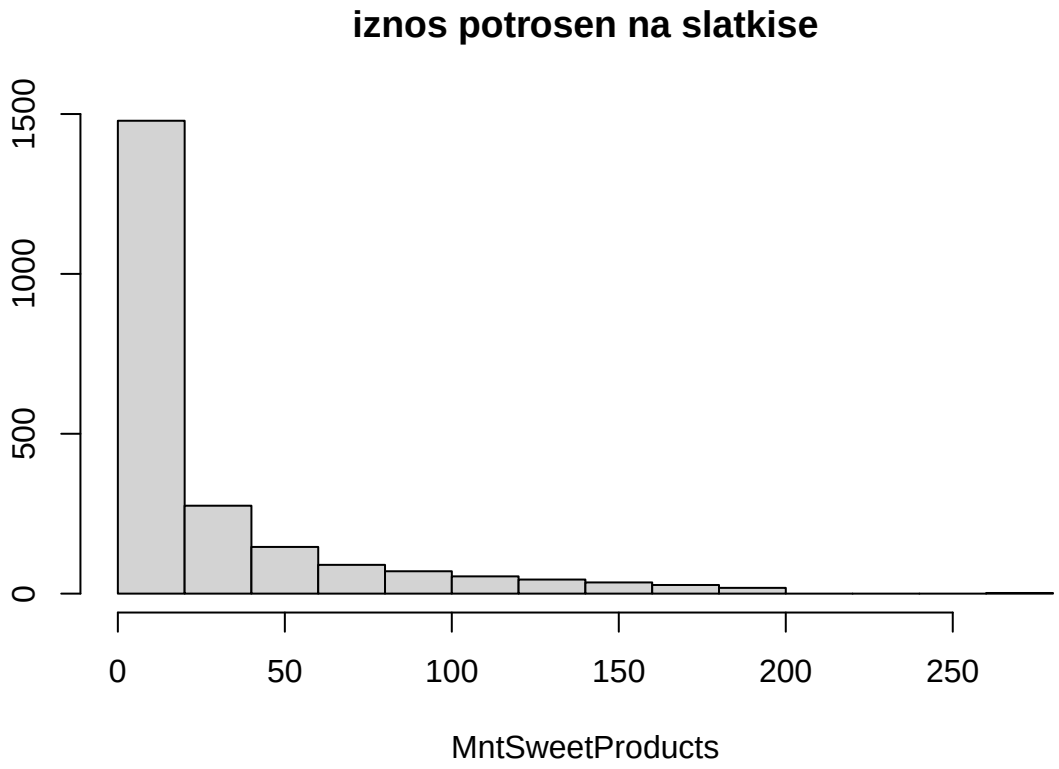
```
## [1] 25.93214
```

Za hi-test vrijedi ako je p-vrijednost manja od 0.05, ispitivane varijable su zavisne. Na temelju p-vrijednosti zaključujemo da su dob klijenta i najčešće korišten prodajni kanal zavisne varijable.

Usporedba iznosa potrošenog na slatkiše kod osoba s djecom i osoba bez djece

Potrebno je odgovoriti na pitanje troše li kupci s djecom više na slatkiše nego kupci bez djece. Promatrana varijabla je MntSweetProducts koja određuje iznos potrošen na slatkiše u posljedne dvije godine. Ovdje su prikazane mjere centralne tendencije i histogram varijable MntSweetProducts:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    1.00    8.00   27.06   33.00  263.00
```

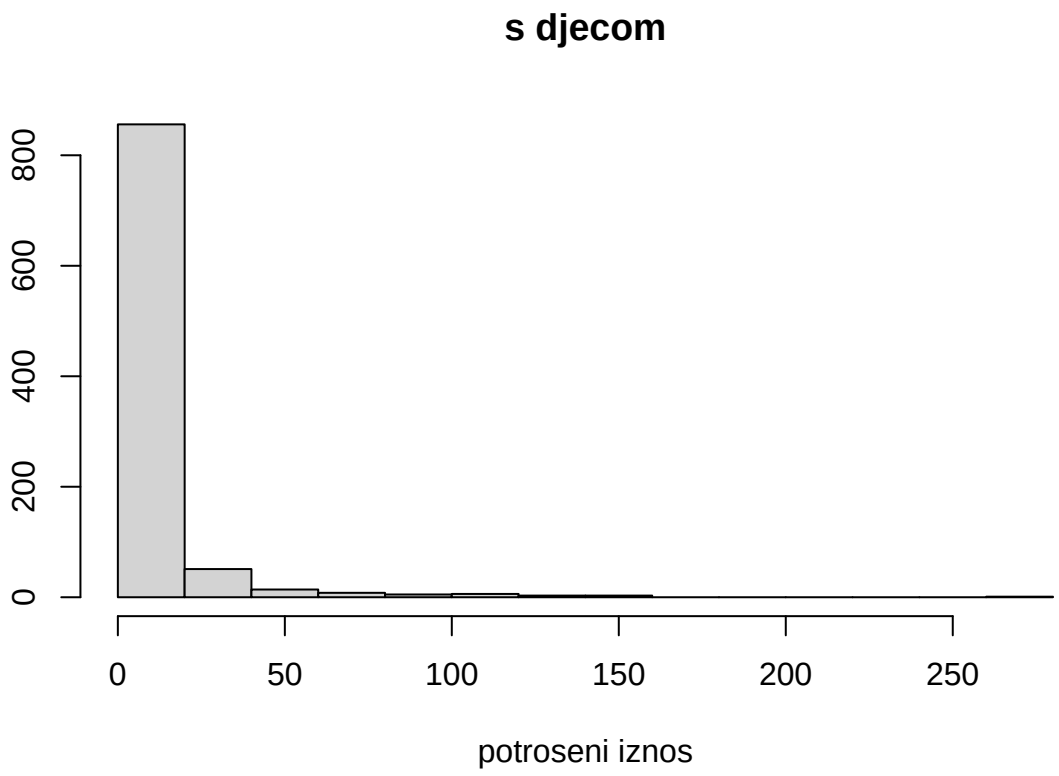


Iz deskriptivne statistike vidi se da distribucija varijable MntSweetProducts nije simetrična i ima težak desni rep. Također, moguće je pretpostaviti da distribucija nije normalna. Bilo bi moguće i formalnije pokazati je li ova distribucija normalna, ali to nije potrebno jer se kasnije koriste samo neparametarski testovi koji ne zahtijevaju normalnost.

Prije provedbe bilo kojih testova potrebno je podijeliti skup podataka na podskup kupaca s djecom i podskup bez djece. U daljnjoj analizi promatra se varijabla MntSweetProducts kod kupaca s djecom i varijabla MntSweetProducts kod kupaca bez djece.

Histogram za iznos potrošen na slatkiše za osobe s djecom:

```
with_kids = data[data$Kidhome>=1,]
hist(with_kids$MntSweetProducts,main='s djecom', xlab='potroseni iznos', ylab='')
```




```
summary(with_kids$MntSweetProducts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   3.000   8.718   8.000 263.000
```

Histogram za iznos potrošen na slatkiše za osobe bez djece:

```
without_kids = data[data$Kidhome == 0,]
hist(without_kids$MntSweetProducts,main='bez djece', xlab='potrosen iznos', ylab='')
```


```
summary(without_kids$MntSweetProducts)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     5.0    22.0    40.5    59.0   262.0
```

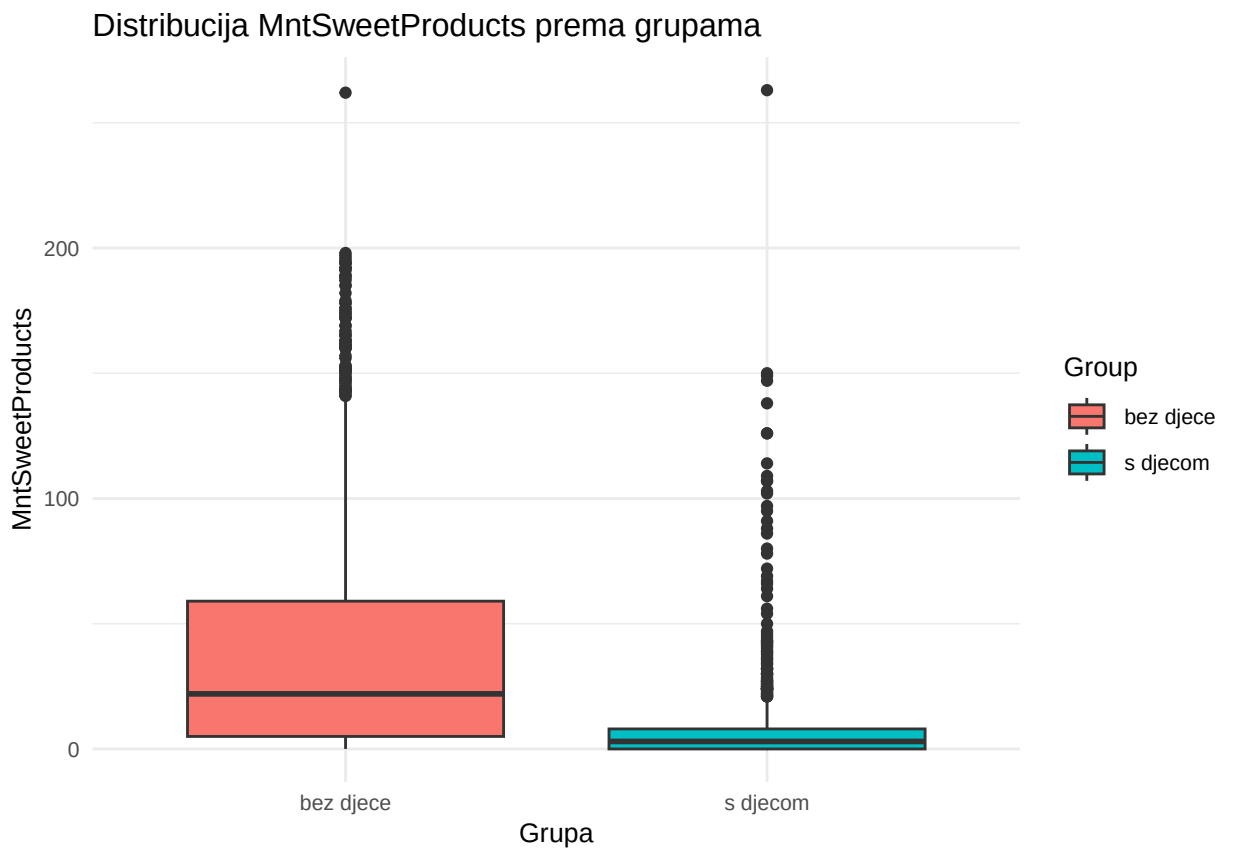
Medijani varijable MntSweetProducts za ove dvije skupine:

```
median(with_kids$MntSweetProducts)
```

```
## [1] 3
```

```
median(without_kids$MntSweetProducts)
```

```
## [1] 22
```



Na temelju box-plotova moguće je vidjeti da je medijan varijable MntSweetProducts veći u uzorku bez djece. Također postoji mnogo potencijalnih outliera u oba uzorka, ali za većinu njih nije moguće zaključiti da se radi o pogreškama te ih zato nema smisla uklanjati iz uzorka. Vidi se i da su vrijednosti u uzorku bez djece raspršenije (veći IQR) nego u uzorku s djecom.

Mann-Whitney-Wilcoxonov test (U-test)
Ovaj test koristi se za testiranje jednakosti medijana dviju populacija ili za testiranje jednakosti distribucija s alternativom da jedna stohastički dominira. Prva interpretacija testa može se koristiti pod uvjetom da znamo da distribucije slučajnih varijabli imaju isti oblik a razlikuju se samo u pomaku (F1(x)=F2(x+a)). Zbog toga što dvije varijable koje mi promatramo nisu simetrične i jedna je više raspršena od druge potrebno je interpretirati U-test na drugi način.

Objašnjenja parametara funkcije wilcox.text():
exact=FALSE znači da se može korsititi normalna aproksimacija U-testa jer n1>8 i n2>8, paired=FALSE znači da se koristi test gdje podatci iz dva uzorka moraju biti nezavisni.

H0: Kupci s djecom troše jednako puno na slatkiše kao kupci bez djece tj. distribucije su jednake

H1: Kupci bez djece troše više na slatkiše od kupaca s djecom. tj. distribucija varijable MntSweetProducts kod kupaca bez djece stohastički dominira

```
result = wilcox.test(with_kids$MntSweetProducts,without_kids$MntSweetProducts,exact=FALSE,paired=FALSE,a
print(result)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  with_kids$MntSweetProducts and without_kids$MntSweetProducts
```

```
## W = 301012, p-value < 2.2e-16
## alternative hypothesis: true location shift is less than 0
```

Na razini značajnosti 5% moguće je odbaciti hipotezu H0 i zaključiti da kupci bez djece troše više na slatkiše nego kupci s djecom.

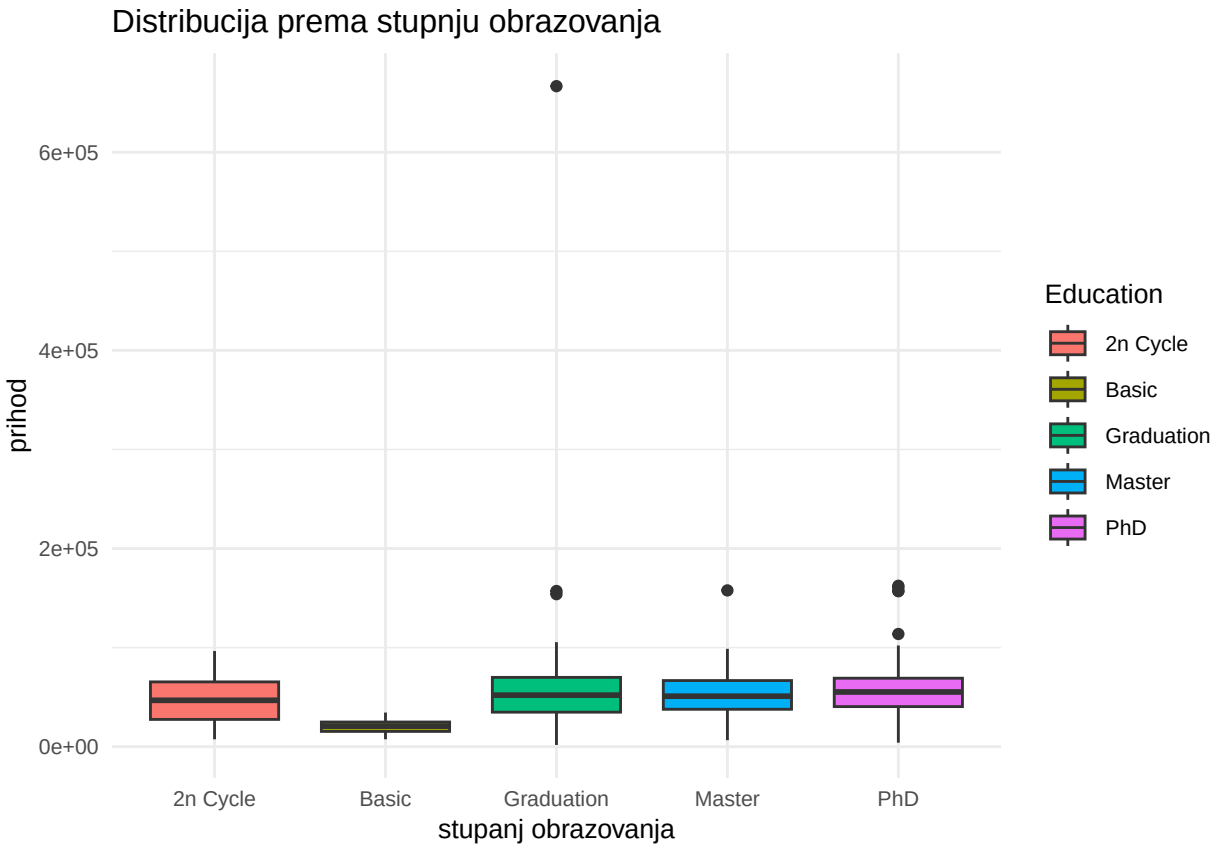
Razlika u prihodima između skupina različitog stupnja obrazovanja

Potrebno je utvrditi postoji li statistički značajna razlika u prihodima između skupina različitog stupnja obrazovanja. Razlikujemo pet stupnjeva obrazovanja: "2n Cycle", "Basic", "Graduation", "Master" i "PhD". Testiranje ćemo provesti usporedbom sredina prihoda za različite stupnjeve obrazovanja.

Vizualizacija i obrada podataka

Prije provedbe samog testiranja, vizulazirat ćemo podatke kako bi dobili uvid u problem koji modeliramo.

```
#Prikaz podataka box-plot dijagramom
p <- ggplot(data=data.full, aes(x = Education, y = Income, fill = Education)) +
    geom_boxplot() +
    labs(
      title = paste("Distribucija prema stupnju obrazovanja"),
      x = "stupanj obrazovanja",
      y = "prihod"
    ) +
    theme_minimal()
  p <- p + theme(
  text = element_text(size = 10),     # Veličina fonta
  axis.text = element_text(size = 8), # Veličina teksta osi
  plot.title = element_text(size = 12) # Veličina naslova
  )
  print(p)
```



Analiziramo graf i uviđamo nepravilnost u podatcima, odnosno prisutnost stršeće vrijednosti. Stršeće vrijednosti unose šum u podatke te tako otežavaju analizu. Ukoliko takvi podatci ne nose puno informacija možemo ih ukloniti iz našeg skupa podataka.

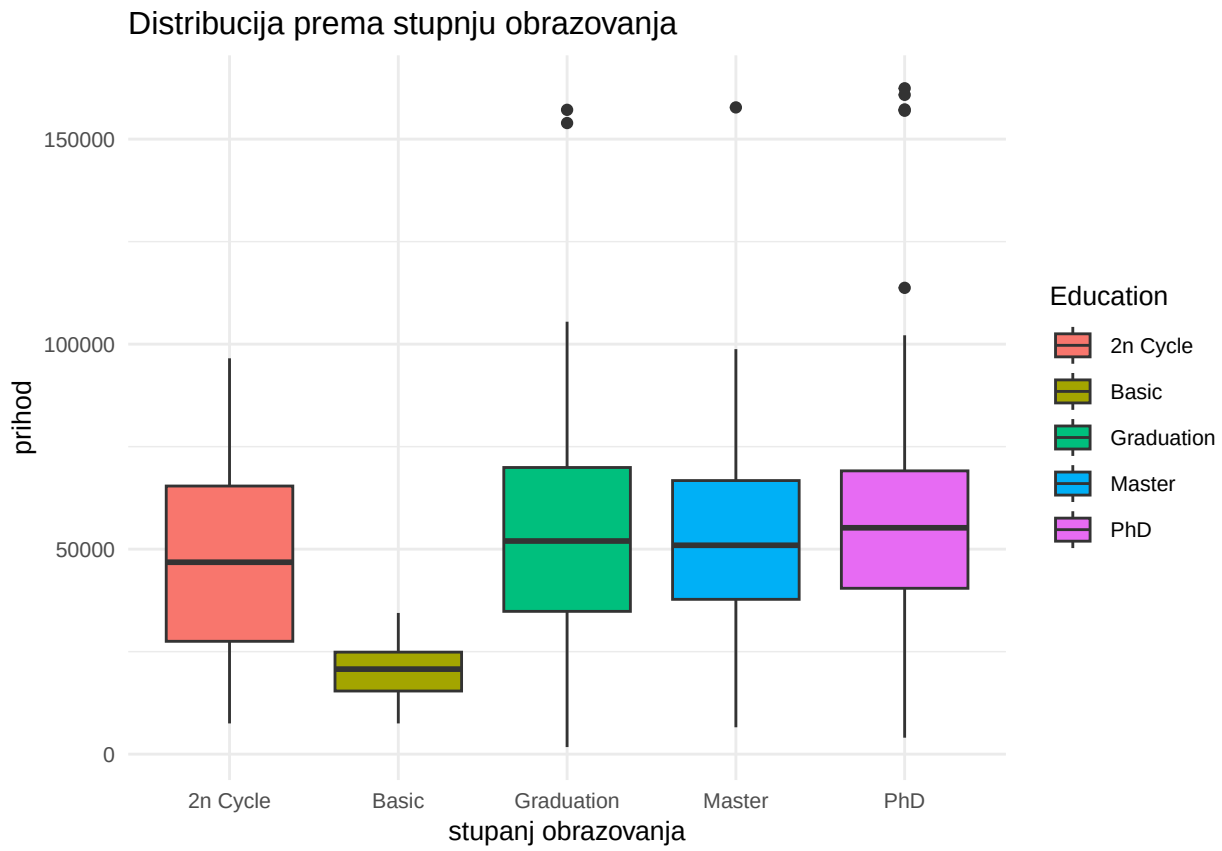
```
#tražimo točan redak outliera
ind = which(data.full$Income > 2e+05)

#prikaz stršeće vrijednosti u podatcima
data[ind,]
```

```
##        ID Year_Birth  Education Marital_Status Income Kidhome Teenhome
## 2210 6168       1963 Graduation       Divorced  45146       1        1
##      Dt_Customer Recency Complain MntWines MntFruits MntMeatProducts
## 2210  2013-07-15      28        0       33         0               5
##      MntFishProducts MntSweetProducts MntGoldProds NumWebPurchases
## 2210               0                0           15               1
##      NumCatalogPurchases NumStorePurchases NumWebVisitsMonth NumDealsPurchases
## 2210                   1                 2                 4                 2
##      AcceptedCmp1 AcceptedCmp2 AcceptedCmp3 AcceptedCmp4 AcceptedCmp5 Response
## 2210            0            0            0            0            0        0
##      Age Age_Group Age_Group_big preferred_channel
## 2210  62     50-64         46-64 NumStorePurchases
```

```
#uklanjamo outlier iz podataka
data.cleaned = data.full[-ind,]
```

Ponovno iscrtavamo graf.

```r
p <- ggplot(data.cleaned, aes(x = Education, y = Income, fill = Education)) +
   geom_boxplot() +
   labs(
     title = paste("Distribucija prema stupnju obrazovanja"),
     x = "stupanj obrazovanja",
     y = "prihod"
   ) +
   theme_minimal()
 p <- p + theme(
 text = element_text(size = 10),     # Veličina fonta
 axis.text = element_text(size = 8), # Veličina teksta osi
 plot.title = element_text(size = 12) # Veličina naslova
 )
 print(p)
```



Ovakav prikaz podataka lako je čitljiv te se iz njega da naslutiti da postoji razlika u primanjima između stupnja obrazovanja "Basic" u odnosu na ostale. Prelazimo na usporedbu srednjih prihoda za različite skupine stupnja obrazovanja. U tu svrhu primjenit ćemo analizu varijance (ANOVA).

ANOVA

Pretpostavke ANOVA-e su: nezavisnost pojedinih podataka u uzorku, normalna raspodjela podataka te homogenost varijanci među podatcima.

Pretpostavku normalnosti podataka provjerit ćemo Lillieforsovom inačicom Kolmogorov-Smirnov testa. Test je oblikovan tako da je stupanj obrazovanja varijabla koja određuje populacije i prihod zavisna varijabla.

Pretpostavke su slijedeće:

$H0:$ *Podatci dolaze iz normalne distribucije*

$H1:$ *Podatci ne dolaze iz normalne distribucije*

```
require(nortest)
```

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

```
lillie.test(data.cleaned$Income[data.cleaned$Education=='2n Cycle'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data.cleaned$Income[data.cleaned$Education == "2n Cycle"]
## D = 0.076178, p-value = 0.006682
```

```
lillie.test(data.cleaned$Income[data.cleaned$Education=='Basic'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data.cleaned$Income[data.cleaned$Education == "Basic"]
## D = 0.10836, p-value = 0.1177
```

```
lillie.test(data.cleaned$Income[data.cleaned$Education=='Graduation'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data.cleaned$Income[data.cleaned$Education == "Graduation"]
## D = 0.053904, p-value = 4.473e-08
```

```
lillie.test(data.cleaned$Income[data.cleaned$Education=='Master'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data.cleaned$Income[data.cleaned$Education == "Master"]
## D = 0.056481, p-value = 0.007081
```

```
lillie.test(data.cleaned$Income[data.cleaned$Education=='PhD'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data.cleaned$Income[data.cleaned$Education == "PhD"]
## D = 0.039737, p-value = 0.06673
```

Na razini značajnosti $\alpha = 0.05$ odbacujemo pretpostavke o normalnosti za grupe s obrazovanjem "2n Cycle", "Graduation" i "Master". Provođenjem logaritamske transformacije nad podatcima ne dobivamo znatno poboljšanje u normalnosti podataka. S obzirom na odbacivanje pretpostavke o normalnosti distribucije koristit ćemo **Kruskal-Wallisov test** kao neparametarsku inačicu ANOVA-e. Sukladno tome, nije nužno provoditi provjeru jednakosti varijanci jer neparametarskt test nema pretpostavki o distribuciji podataka, no ovdje ćemo u svrhu analize provesti i tu provjeru. Za testiranje homogenosti varijanci poslužit ćemo se Bartlettovim testom.

Pretpostavke:

$H0: \quad \sigma_{\text{2n cycle}} = \sigma_{\text{basic}} = \sigma_{\text{graduation}} = \sigma_{\text{master}} = \sigma_{\text{phd}}$

$H1: \quad Postoje \ \ barem \ \ dvije \ \ \sigma_{\text{i}} \ \ koje \ \ se \ \ razlikuju$

```
# Testiranje homogenosti varijance uzoraka Bartlettovim testom
bartlett.test(data.cleaned$Income ~ data.cleaned$Education)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data.cleaned$Income by data.cleaned$Education
## Bartlett's K-squared = 83.169, df = 4, p-value < 2.2e-16
```

Zbog vrlo niske p vrijednosti na razini značajnosti $\alpha = 0.05$ možemo odbaciti nul-hipotezu i zaključiti da postoji par varijanci koji se razlikuje. Ovime potvrđujemo uporabu neparametarskog **Kruskal-Wallisovog** testa s obzirom na manjak pretpostavki o distribuciji podataka.

Pretpostavke:

$H0: \quad \mu_{\text{2n cycle}} = \mu_{\text{basic}} = \mu_{\text{graduation}} = \mu_{\text{master}} = \mu_{\text{phd}}$

$H1: \quad barem \ \ dva \ \ \mu_{\text{i}} \ \ nisu \ \ jednaka.$

Provodimo testiranje.

```
kruskal.test(Income ~ Education, data = data.cleaned)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Income by Education
## Kruskal-Wallis chi-squared = 142.25, df = 4, p-value < 2.2e-16
```

Na nivou značajnosti od $\alpha = 0.05$ možemo odbaciti nul-hipotezu o jednakosti srednjih vrijednosti zbog izrazito niske p vrijednosti.

Nastavljamo daljnju analizu kako bismo provjerili sumnju u razlikovanje prihoda "Basic" razine obrazovanja u odnosu na ostale. Koristit ćemo Dunnov test za testiranje parova koji se koristi kad podatci nisu normalno distribuirani ili kada se ne mogu pretpostaviti jednakosti varijanci među grupama.

```r
library(dunn.test)

dunn.test(data.cleaned$Income, g = data.cleaned$Education, method = "bonferroni")
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 142.2475, df = 4, p-value = 0
##
##
##                            Comparison of x by group
##                                  (Bonferroni)
## Col Mean-|
## Row Mean |   2n Cycle      Basic   Graduati     Master
## ---------+-----------------------------------------
##    Basic |   8.380529
##          |    0.0000*
##          |
## Graduati |  -2.692991  -10.70779
##          |     0.0354    0.0000*
##          |
##   Master |  -2.671355  -10.42667  -0.467908
##          |     0.0378    0.0000*     1.0000
##          |
##      PhD |  -4.457823  -11.56844  -3.084600  -2.017553
##          |    0.0000*    0.0000*    0.0102*     0.2182
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Sukladno pretpostavci, prihod grupe s "Basic" razinom obrazovanja razlikuje se u odnosu na srednje vrijednosti ostalih grupa. Postoji još značajna razlika grupe "PhD" u odnosu na "2n Cycle" i "Graduation", dok razlika između stupnja "PhD" i "Master" nije statistički značajna.

```r
library(dplyr)
```

Mogu li dostupne varijable predvidjeti ukupnu potrošnju kupca?

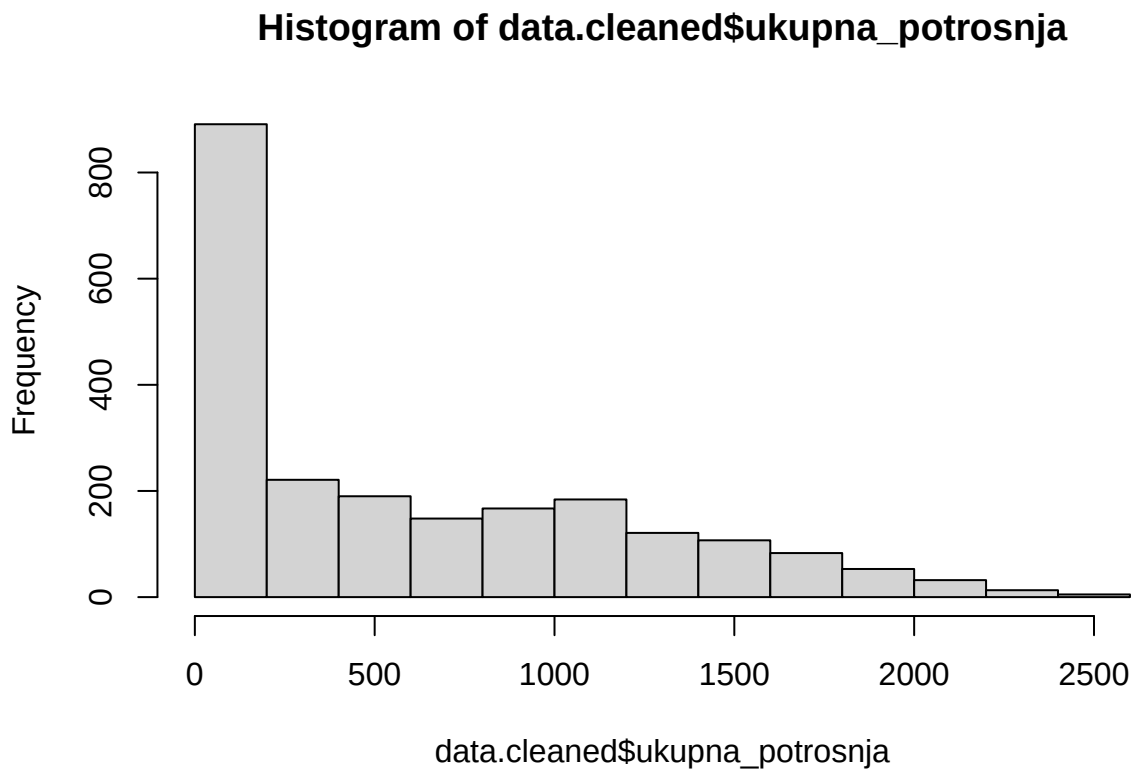
Vizualizacija i uređivanje podataka

U našem datasetu ne postoji unaprijed definirana varijabla koja predstavlja ukupnu potrošnju kupca, već su dostupne pojedinačne potrošnje po kategorijama (vina, voća, mesa, ribe, slatkiša i zlata). Kako bismo modelirali ponašanje ukupne potrošnje, prvo je potrebno definirati novu varijablu – ukupnu potrošnju – kao sumu svih pojedinačnih potrošnji.

```r
# Definiranje varijable 'ukupna_potrosnja'
data.cleaned$ukupna_potrosnja <- data.cleaned$MntWines +
                                 data.cleaned$MntFruits +
                                 data.cleaned$MntMeatProducts +
                                 data.cleaned$MntFishProducts +
                                 data.cleaned$MntSweetProducts +
                                 data.cleaned$MntGoldProds

# Prikaz prvih nekoliko vrijednosti
head(data.cleaned$ukupna_potrosnja)
```

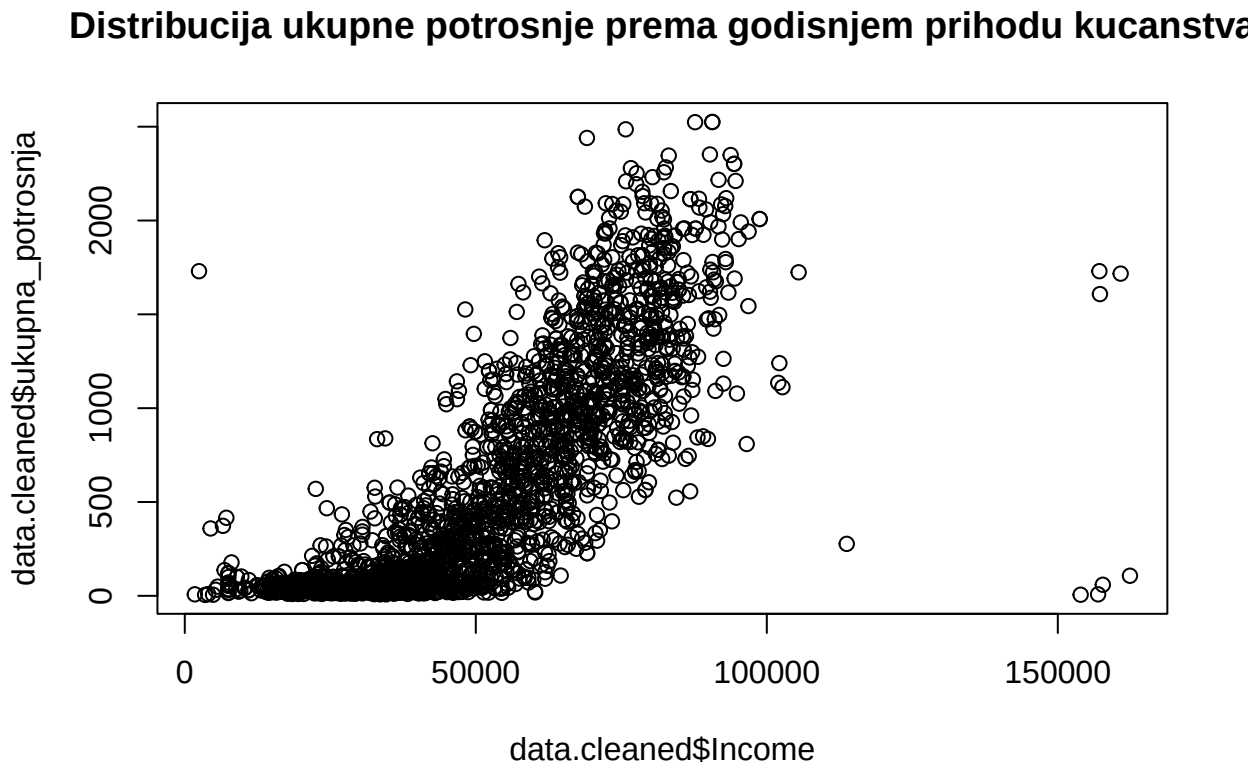
```
## [1] 1617   27  776   53  422  716
```

```r
hist(data.cleaned$ukupna_potrosnja)
```



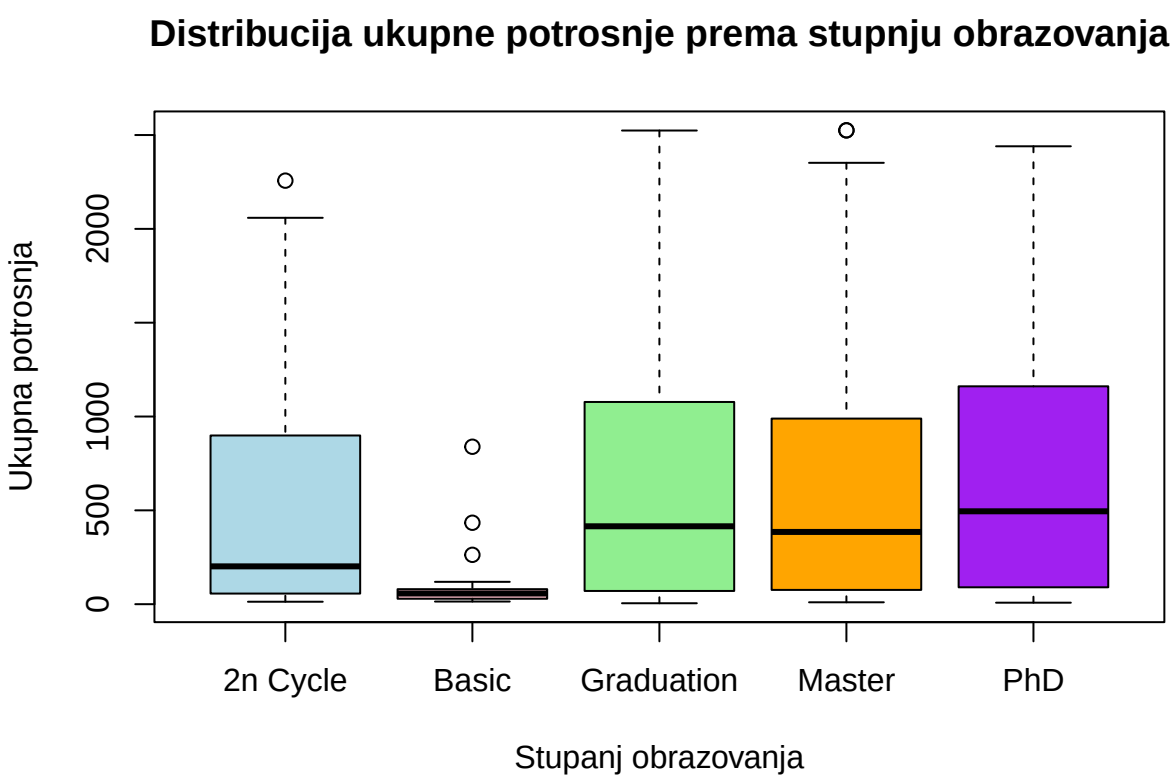
Sada kada imamo varijablu čije buduće ponašanje želimo predvidjeti, ispitati ćemo različite nezavisne varijable koje bi mogle utjecati na njeno ponašanje. Kada promatramo utjecaj samo jedne nezavisne varijable na zavisnu, *scatter plot* nam može pomoći s vizualizacijom. Varijable mogu pozitivno i negativno utjecati na zavisnu varijablu. Očekujemo da će godišnji prihod kućanstva imati izražen pozitivan utjecaj na ukupnu potrošnju.

```r
plot(data.cleaned$Income, data.cleaned$ukupna_potrosnja,
     main = "Distribucija ukupne potrosnje prema godisnjem prihodu kucanstva")
```



Pogledajmo i jednog slabijeg kandidata za regresora, primjerice stupanj obrazovanja kupca. Kako je *Education* kategorijska varijabla, koristiti ćemo *box plot*.

```r
boxplot(ukupna_potrosnja ~ Education, data = data.cleaned,
        xlab = "Stupanj obrazovanja",
        ylab = "Ukupna potrosnja",
        col = c("lightblue", "pink", "lightgreen", "orange", "purple"),
        main = "Distribucija ukupne potrosnje prema stupnju obrazovanja")
```



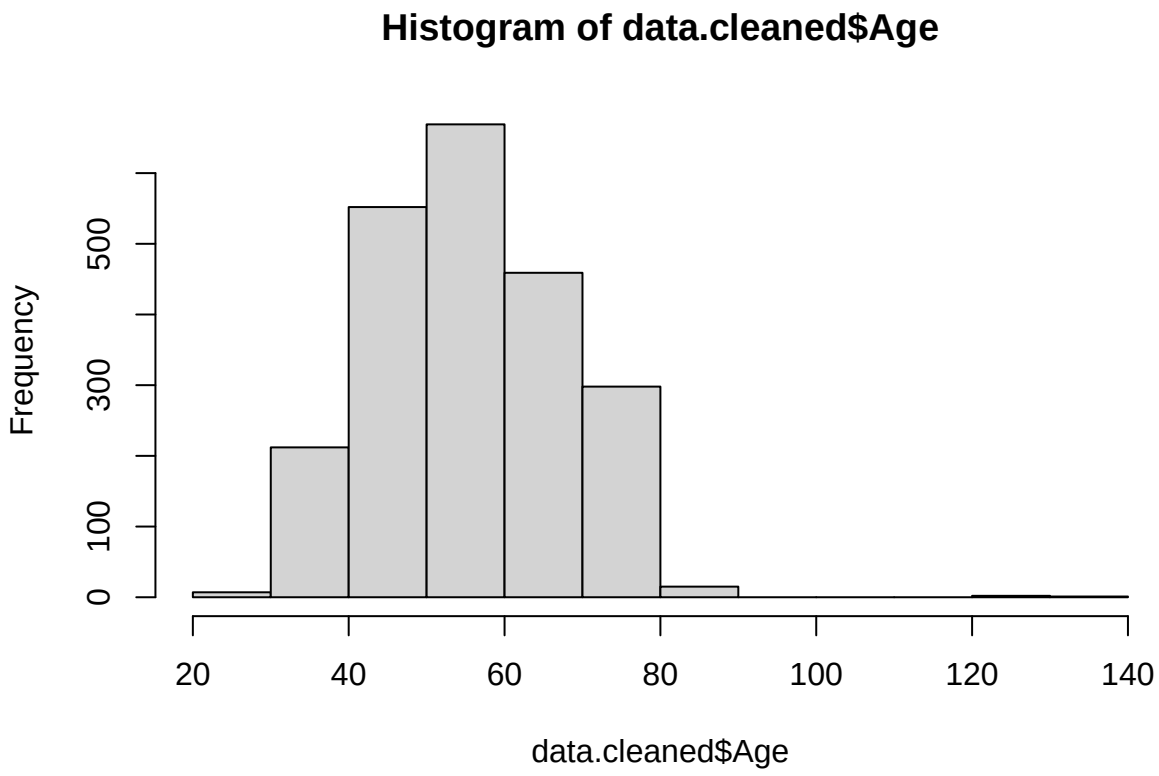
Iz prikazanog grafa je vidljivo kako stupanj obrazovanja ne govori puno o ukupnoj potrošnji kupca, odnosno unutar nekoliko različitih kategorija obrazovnog stupnja, kupci troše približno jednako.

Dodati ćemo i varijablu Age, koju ćemo koristiti umjesto godine rođenja.

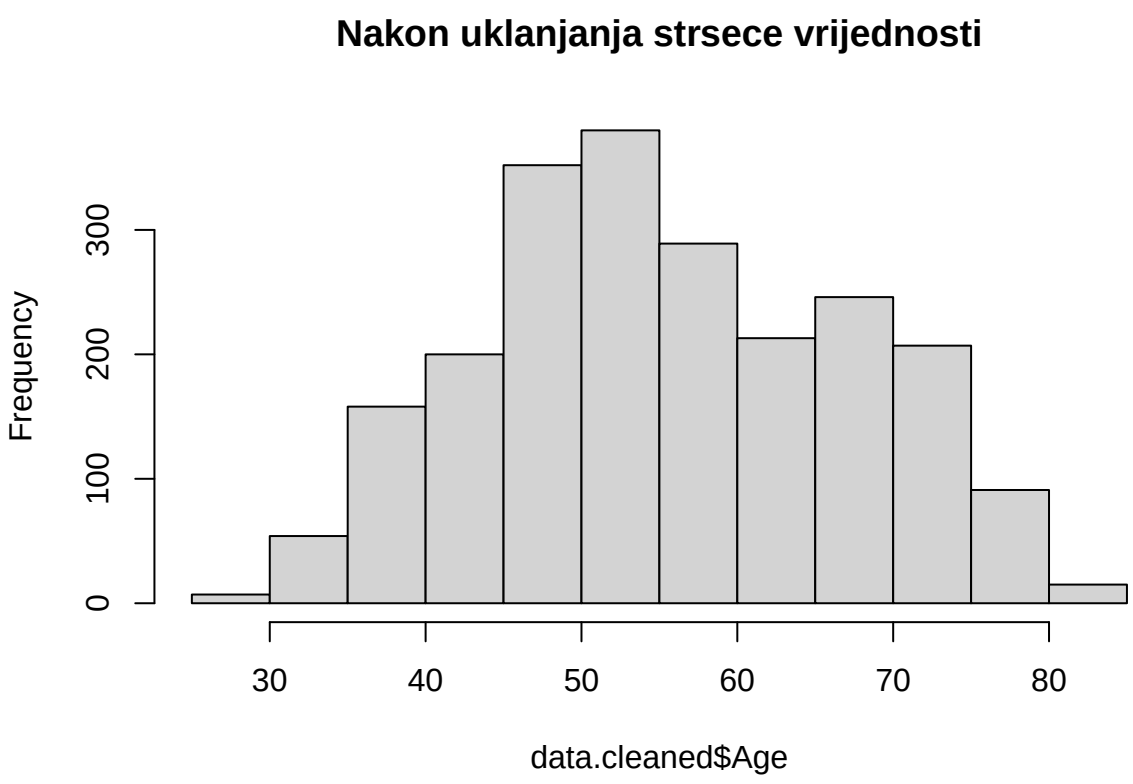
```
data.cleaned$Age <- 2025 - data.cleaned$Year_Birth
head(data.cleaned$Age)
```

```
## [1] 68 71 60 41 44 58
```

```
hist(data.cleaned$Age)
```




```r
#maknemo strsece vrijednosti
data.cleaned <- subset(data.cleaned, Age <= 100)
hist(data.cleaned$Age, main = "Nakon uklanjanja strsece vrijednosti")
```



Modeli linearne regresije

Kako bi pronašli najbolji model, potrebno testirati više modela linearne regresije odabirom različitih regresora. Za pojedine modele moramo izračunati R^2 vrijednost, koja nam govori koliko dobro model objašnjava varijabilnost podataka. Model s najvećom R^2 vrijednosti je najbolji model.

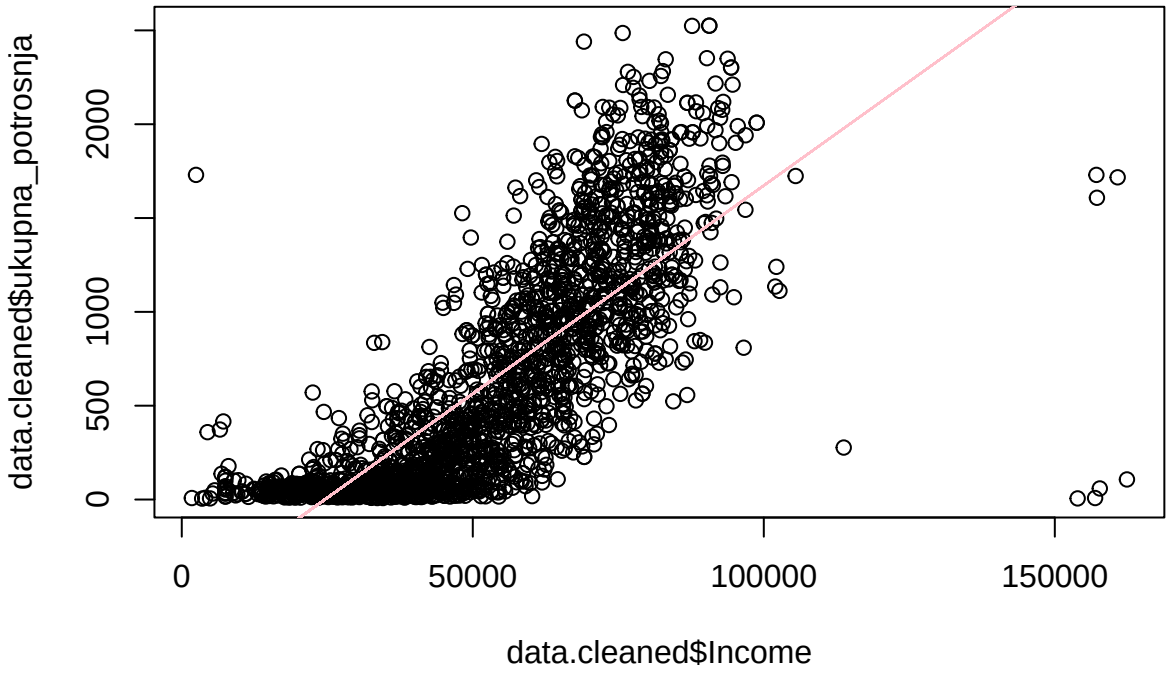
Za početak pogledati ćemo vrlo jednostavne modele, koji koriste samo jedan regresor.

```
fit.income = lm(ukupna_potrosnja ~ Income, data = data.cleaned)
summary(fit.income)
```

```
##
## Call:
## lm(formula = ukupna_potrosnja ~ Income, data = data.cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2950.61  -225.19   -37.16   206.54  2221.27
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.456e+02  2.041e+01  -26.73   <2e-16 ***
## Income       2.219e-02  3.629e-04   61.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 367.4 on 2210 degrees of freedom
## Multiple R-squared:  0.6284, Adjusted R-squared:  0.6283
## F-statistic:  3738 on 1 and 2210 DF,  p-value: < 2.2e-16
```

```
plot(data.cleaned$Income,data.cleaned$ukupna_potrosnja)
lines(data.cleaned$Income,fit.income$fitted.values,col='pink')
```




```
r_squared1 <- summary(fit.income)$r.squared


fit.education = lm(ukupna_potrosnja ~ Education, data = data.cleaned)
summary(fit.education)
```

```
##
## Call:
## lm(formula = ukupna_potrosnja ~ Education, data = data.cleaned)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -666.3 -533.9 -189.2  426.7 1915.2
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           499.49      42.33  11.800  < 2e-16 ***
```

```
## EducationBasic       -417.69      91.44  -4.568  5.2e-06 ***
## EducationGraduation   122.70      45.94   2.671 0.007616 **
## EducationMaster       110.28      52.57   2.098 0.036053 *
## EducationPhD          174.79      50.31   3.474 0.000522 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 595.6 on 2207 degrees of freedom
## Multiple R-squared:  0.02444,    Adjusted R-squared:  0.02267
## F-statistic: 13.82 on 4 and 2207 DF,  p-value: 3.876e-11
```

```
r_squared2 <-summary(fit.education)$r.squared

cat("R^2 za Income kao regresor:", r_squared1 %>% round(4),
"\nR^2 za Education kao regresor:", r_squared2 %>% round(4), "\n")
```

```
## R^2 za Income kao regresor: 0.6284
## R^2 za Education kao regresor: 0.0244
```

Sada smo pokazali pomoću R^2 vrijednosti kako je godišnji prihod zaista bolji regresor od stupnja obrazovanja. No, kako bi model bio bolje prilagođen podatcima, potrebno je dodati još regresora. Stoga ćemo napraviti nekoliko modela višestruke regresije.

```
#varijable vezane uz "obiteljske podatke kao regresori
    #~ Marital_Status, Teenhome, Kidhome
model1 = lm(ukupna_potrosnja ~ Marital_Status + Teenhome + Kidhome, data = data.cleaned)
model_summary1 <- summary(model1)
r_squared1 <- model_summary1$r.squared

#dob, stupanj obrazovanja i zarada kao regresori
    #~ Age, Education, Income
model2 = lm(ukupna_potrosnja ~ Age + Education + Income, data = data.cleaned)
model_summary2 <- summary(model2)
r_squared2 <- model_summary2$r.squared

#broj dana od zadnje kupnje, zalbe u protekle 2 godine i broj kupnji sa popustom, prihvacene kamapanje
    #~ Recency, Complain, NumDealsPurchases, AcceptedCmp1-5, Response
model3 = lm(ukupna_potrosnja ~ Recency + Complain + NumDealsPurchases + AcceptedCmp1 +
              AcceptedCmp2 + AcceptedCmp3 + AcceptedCmp4 + AcceptedCmp5 + Response,
              data = data.cleaned)
model_summary3 <- summary(model3)
r_squared3 <- model_summary3$r.squared

#broj obavljenih kupnji putem svih prodajnih kanala kao regresori
    #~ NumWebPurchases, NumCatalogPurchases, NumStorePurchases

model4 = lm(ukupna_potrosnja ~ NumWebPurchases + NumCatalogPurchases + NumStorePurchases, data = data.cl
model_summary4 <- summary(model4)
r_squared4 <- model_summary4$r.squared

cat("R^2 za model1:",
    r_squared1 %>% round(4),
  "\nR^2 za model2:",
```

```
    r_squared2 %>% round(4),
  "\nR^2 za model3:",
    r_squared3 %>% round(4),
  "\nR^2 za model4:",
    r_squared4 %>% round(4),"\n")
```

```
## R^2 za model1: 0.3375
## R^2 za model2: 0.6328
## R^2 za model3: 0.2779
## R^2 za model4: 0.7268
```

Iz danih modela, vidimo kako vrlo visoku R^2 postižu oni koji kao regresore koriste Income i broj kupnji obavljenih putem svakog pojedinog prodajnog kanala. Te ćemo regresore svakako uključiti u naš finalni model jer vrlo dobro objašnjavaju ponašanje varijable koje želimo predvidjeti.

Prije nego napravimo model višestruke regresije, bitno je osigurati se da parovi varijabli nisu pretjerano korelirani. Visoka koreliranost varijabli daje nestabilne rezultate. Varijable koje bi mogle biti korelirane su broj posjeta web stranici i broj obavljenih kupnji putem web stranice, broj kupnji obavljenih putem kataloga i web stranice te dob i prihod.

```
koef_korelacije = cor(data.cleaned$NumWebPurchases, data.cleaned$NumWebVisitsMonth)
cat("Koeficijent korelacije između varijabli NumWebVisitsMonth i NumWebStorePurchases:",
    koef_korelacije %>% round(4),"\n")
```

```
## Koeficijent korelacije između varijabli NumWebVisitsMonth i NumWebStorePurchases: -0.0516
```

```
koef_korelacije2 = cor(data.cleaned$NumCatalogPurchases, data.cleaned$NumWebPurchases)
cat("Koeficijent korelacije između varijabli NumCatalogPurchases i NumWebStorePurchases:",
    koef_korelacije2 %>% round(4),"\n")
```

```
## Koeficijent korelacije između varijabli NumCatalogPurchases i NumWebStorePurchases: 0.3865
```

```
koef_korelacije3 = cor(data.cleaned$Age, data.cleaned$Income)
cat("Koeficijent korelacije između varijabli Age i Income:",
    koef_korelacije3 %>% round(4),"\n")
```

```
## Koeficijent korelacije između varijabli Age i Income: 0.2
```

Ipak, vrijednosti koeficijenata korelacije nisu visoke, stoga je dozvoljeno sve varijable koristiti u modelu.

U nekim situacijama poželjno je primijeniti transformacije na ulazne varijable kako bi se bolje odrazila njihova nelinearna priroda. Uz pretpostaku da djeca i stariji ljudi troše manje od odraslih, dodati ćemo i kvadrat varijable Age u model kako bismo modelirali nelinearan pad potrošnje s godinama. Ova modifikacija omogućuje bolju prilagodbu modela pretpostavljenim obrascima ponašanja korisnika.

```
model_age1 = lm(ukupna_potrosnja ~ Age + Education + Income, data = data.cleaned)
model_summary1 <- summary(model_age1)
r_squared1 <- model_summary1$r.squared
```

```r
model_age2 = lm(ukupna_potrosnja ~ I(Age^2) + Age + Education + Income, data = data.cleaned)
model_summary2 <- summary(model_age2)
r_squared2 <- model_summary2$r.squared

cat("R^2 za model sa linearnom varijablom Age:",
    r_squared1 %>% round(4),
    "\nR^2 za model sa linearnom i kvadratnom varijablom Age:",
     r_squared2 %>% round(4),"\n")
```

```
## R^2 za model sa linearnom varijablom Age: 0.6328
## R^2 za model sa linearnom i kvadratnom varijablom Age: 0.6357
```

Sada možemo napraviti naš finalni model. S obzirom da u njemu uključujemo mnogo regresora, uz R^2 uzeti ćemo u obzir i adjusted R^2. Adjusted R^2 penalizira dodavanje parametara u model, čime sprječava nepotrebno povećanje složenosti. Kako preferiramo odabrati jednostavniji model, pod uvjetom da daje jednako dobre rezultate kao složeniji modeli, adjusted R^2 nam pomaže donijeti odluku o tome koji model najbolje balansira preciznost i složenost.

```r
model = lm(ukupna_potrosnja ~ I(Age^2) + Age + Education + Income + NumWebPurchases + NumWebVisitsMonth
model_summary <- summary(model)
r_squared <- model_summary$r.squared
r_squared_adj <- model_summary$adj.r.squared

#izbacujemo: Education i NumWebVisitsMonth
model2 = lm(ukupna_potrosnja ~ I(Age^2) + Age + Income + NumWebPurchases + NumCatalogPurchases + NumSto
model_summary2 <- summary(model2)
r_squared2 <- model_summary2$r.squared
r_squared_adj2 <- model_summary2$adj.r.squared

cat("R^2: ",r_squared %>% round(4),
    "\nR^2 adjusted:",r_squared_adj %>% round(4),"\n")
```

```
## R^2:  0.8
## R^2 adjusted: 0.7988
```

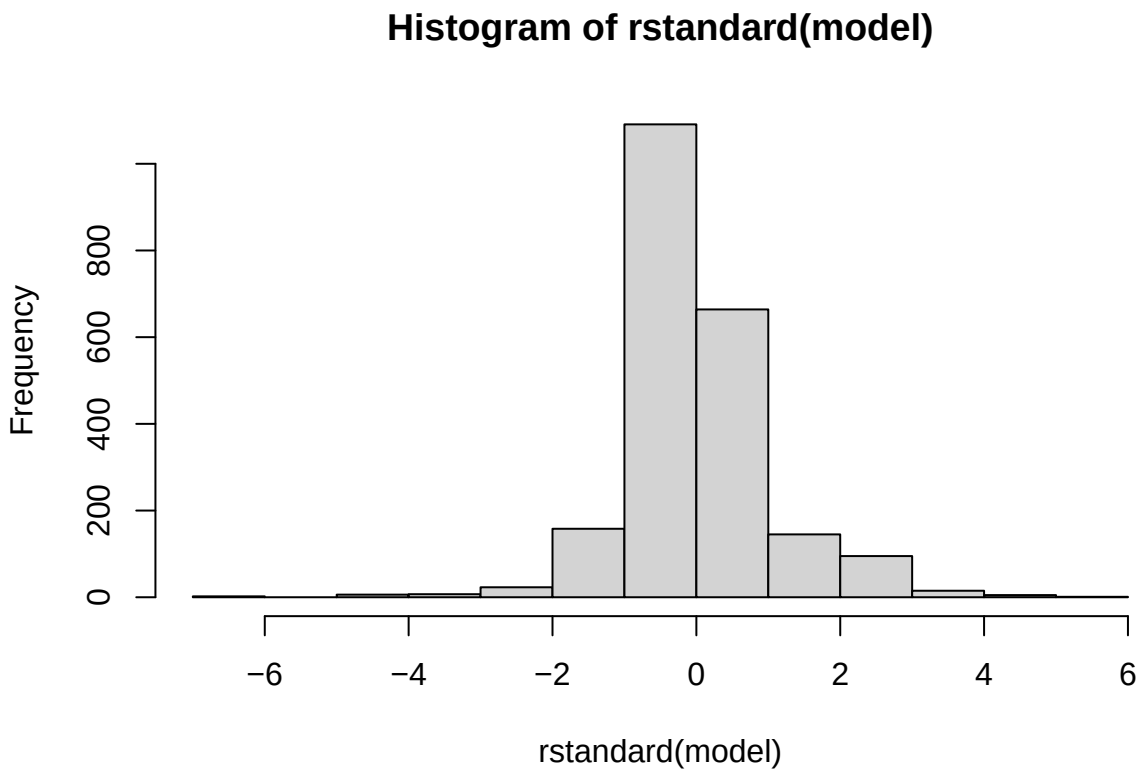
```r
cat("R^2 nakon izbacivanja Education i NumWebVisitsMonth: ",r_squared2 %>% round(4),
    "\nR^2 adjusted nakon izbacivanja Education i NumWebVisitsMonth:",r_squared_adj2 %>% round(4),"\n")
```

```
## R^2 nakon izbacivanja Education i NumWebVisitsMonth:  0.7959
## R^2 adjusted nakon izbacivanja Education i NumWebVisitsMonth: 0.7952
```

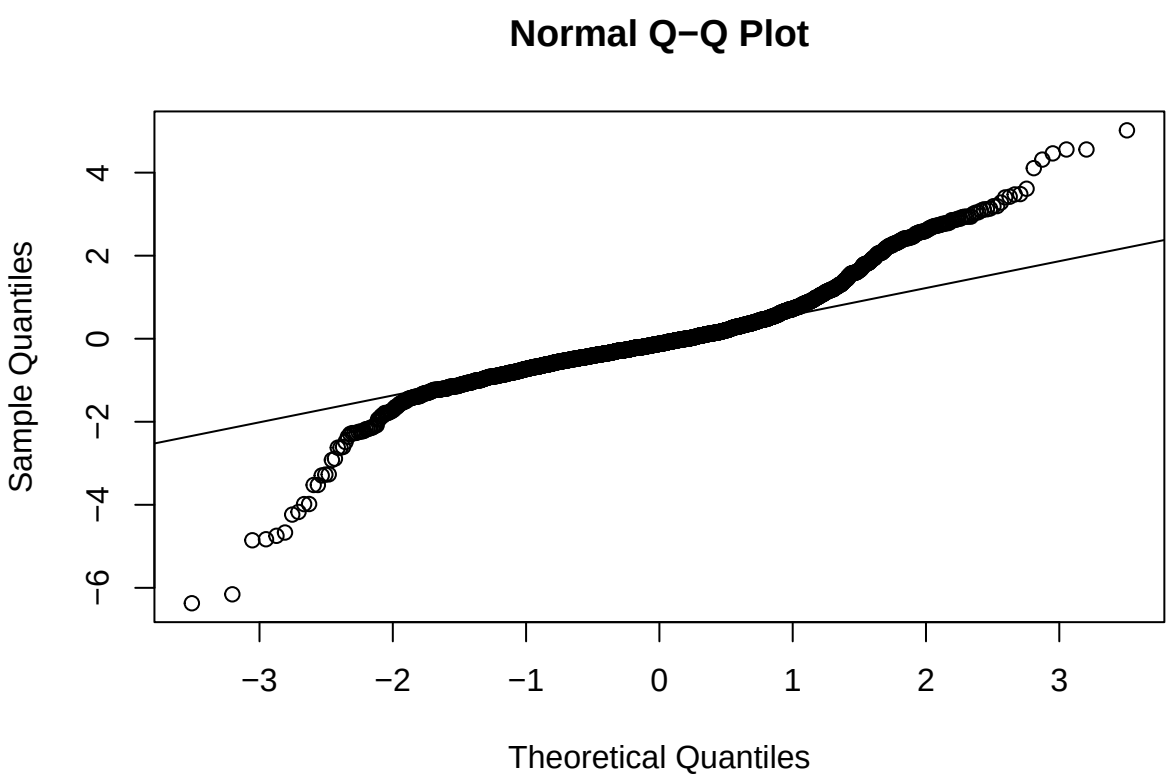
Rezultati upućuju na to da varijable Education i NumWebVisitsMonth ipak daju korisne informacije modelu, čak i kada pogledamo adjusted R^2.

Provjeriti ćemo još da nije narušena bitna pretpostavka linearne regresije: normalnost reziduala, pomoću histograma te *Q-Q plota*.

```r
hist(rstandard(model))
```




```
qqnorm(rstandard(model))
qqline(rstandard(model))
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
−6
−4
−2
0
2
4
−3
−2
−1
0
1
2
3