Analiza cene nekretnina

Seminarski rad u okviru kursa
Uvod u teoriju uzoraka
Matematički fakultet

Petar Zečević
mi16169@alas.matf.bg.ac.rs

26. jul 2020.

Sažetak

U radu je prikazana analiza srednje vrednosti cene nekretnina i cene metra kvadartnog nekretnina u okrugu King County pomoću raznih metoda uzorkovanja. Te metode su: prost slučajan uzorak, regresiono ocenjivanje i stratifikovan uzorak. Pokazalo se da su za relativno malu veličinu uzorka sva tri metoda dala pristojne rezltate.

Sadržaj

1 Uvod

Analiziranje cene nekretnina je traženo iz razloga što bi ljudima olakšala kupovinu stanova. Želimo da nad relativno malom bazom podataka, prodate kuće u periodu od godinu dana u okrugu King County, nađemo dobru metodu za uzorkovanje da bismo to mogli da primenimo nad nekom većom bazom (na primer nad celom državom).

Pokušaćemo da ocenimo kvalitet ocene srednje vrednosti cene prodatih nekretnina kao i cene kvadratnog metra prodatih nekretnina. Prvo ćemo da opišemo bazu podataka i analiziramo potencijalno korisna obeležja, zatim ćemo da ukratko teorijski objasnimo uzorkovanja koja smo koristili. Onda sledi analiza rezltata koje smo dobili. U dodatku se nalazi kod u programskom jeziku R koji je korišćen u radu.

1.1 Opis baze podataka

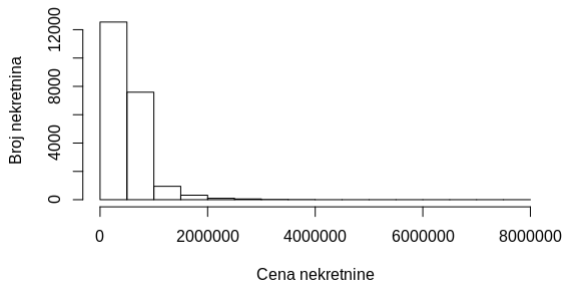
Baza predstavlja informacije o prodatim kućama u periodu od maja 2014. do maja 2015. godine u okrugu King County, SAD (uključujući Sietl) [1]. Sastoji se iz 21613 redova i 21 kolone. Kolone koje predstavljaju broj kvadratnih stopa su konvertovani u broj kvadratnih metara. Sledeći atributi se nalaze u bazi:

- **id** - jedinstveni identifikator reda. Nema nikakav značaj za istraživanje.
- **date** - datum prodaje. Izražen kao vremenski žig.
- **price** - cena po kojoj je nekretnina prodata. Ovo je ciljna promenljiva u našem istraživanju.
- **bedrooms** - broj soba. Postoje nekretnine bez soba što je verovatno greška
- **bathrooms** - broj kupatila. Ovaj atribut može uzeti vrednost na primer 2.5 jer treće kupatilo nema neki deo da bi bilo kompletno. Mi smo zaokruživali ove vrednosti jer je tako razumljivije.
- **sqft_living** - broj kvadratnih stopa u zatvorenom prostoru
- **sqft_lot** - broj kvadratnih stopa ukupno na placu
- **floors** - broj spratova. Može biti na primer 1.5 što znači da kuća ima jedan međusprat.
- **waterfronts** - indikator da li se nekretnina nalazi u blizini reke ili mora.
- **view** - ocena koliko dobar pogled ima
- **condition** - ocena u kakvom je stanju objekat
- **grade** - ocena koliko je kvalitetna izgradnja objekta
- **sqft_above** - broj kvadratnih stopa u zatvorenom prostoru ne računajući podrum
- **sqft_basement** - broj kvadratnih stopa podruma
- **yr_built** - godina izgradnje
- **yr_renovated** - godina poslednjeg renoviranja. Ako objekat nije renoviran onda je ova vrednost nula.
- **zipcode** - poštanski broj
- **lat** - geografska širina
- **long** - geografska dužina

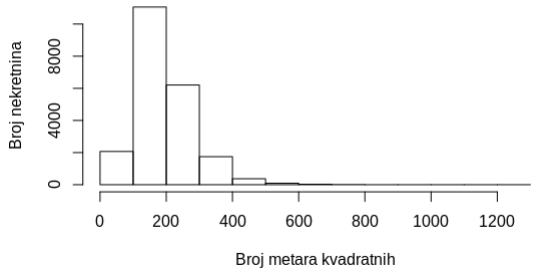
- **sqft_living15** - prosek broja kvadratnih stopa u zatvorenom prostoru za 15 najbližih kuća
- **sqft_lot15** - prosek broja kvadratnih stopa ukupno na placu za 15 najbližih kuća

2 Analiza baze podataka

U okrugu King County cene nekretnina se kreću od $77,000 do $7,700,000. 93% nekretnina koštaju ispod $1,000,000 što možemo videti na slici 1. Na slici 2 može se videti histogram obeležja broja metara kvadratnih u zatvorenom prostoru nekretnine. Ostala obeležja za broj metara kvadratnih nisu dobro korelisana sa cenom nekretnine tako da na njih nije posvećeno mnogo pažnje.



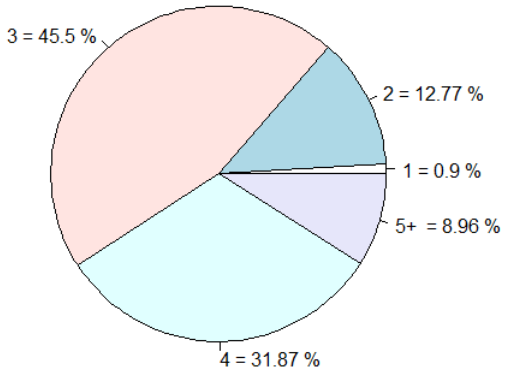
Slika 1: Histogram cena nekretnina



Slika 2: Histogram broja metara kvadratnih

Iz baze su izbačene nekretnine koje imaju nula soba ili kupatila jer verovatno predstavljaju neku grešku. Procenat nekretnina po broju soba može se videti na slici 3, a procenat nekretnina po broju kupatila na 4. Sve nekretnine sa više od 4 sobe su kategorisane kao 5+, a sa više od 3 kupatila u 4+ radi preglednijih slika. Vrednosti broja kupatila su zaokružena na

cele brojeve jer je nedefinisano šta bi na primer značilo da kuća ima 2.25 kupatila.



Slika 3: Procenat nekretnina po broju soba

Cena nekretnine ima najbolju korelaciju sa brojem metara kvadratnih u zatvorenom prostoru, 0.7. Graf korelacije je prikazan na slici 6. Procenat nekretnina po metrima kvadratnim u zatvorenom prostoru je prikazan na slici 5. Vrednosti su diskretizovane u intervale po $100m^2$.

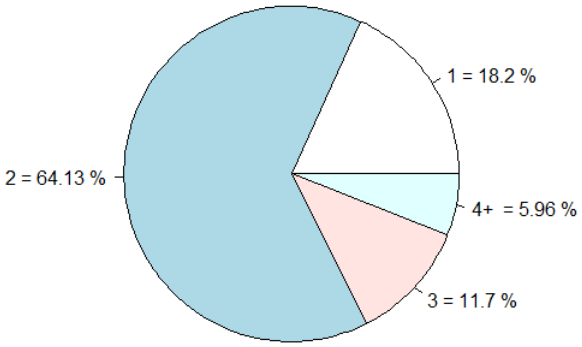
Prosečna cena metra kvadratnog u okrugu King County je 2843.03. Histogram je prikazan na slici 7.

Pošto imamo obeležja geografska širina i dužina, možemo aproksimativno da odredimo koja nekretnina se nalazi u gradu Sijetl, a koja van njega. Gledajući mapu King County okruga, uzimamo da se grad Sijetl nalazi na geografskoj širini u intervalu (47.501221, 47.735211) i geografskoj dužini (-122.443828, -122.238907). Procenat nekretnina u Sijetlu i van njega nalazi se na 8. Korelacija između ovog obeležja i cene nekretnina je samo 0.0955997 tako da ovo obeležje nećemo dalje koristiti. Korelacija između ovog obelžja i cene metra kvadratnog je bolja, 0.4337856, ali opet ne dovoljno dobra da bismo je iskoristili za regresiono ocenjivanje.

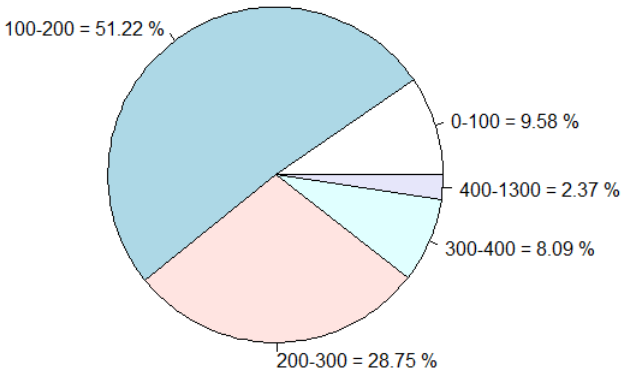
Po formuli iz [2] dobijamo da je dobra veličina uzorka 1411. To je 6.53% ukupne populacije.

3 Teorijski uvod

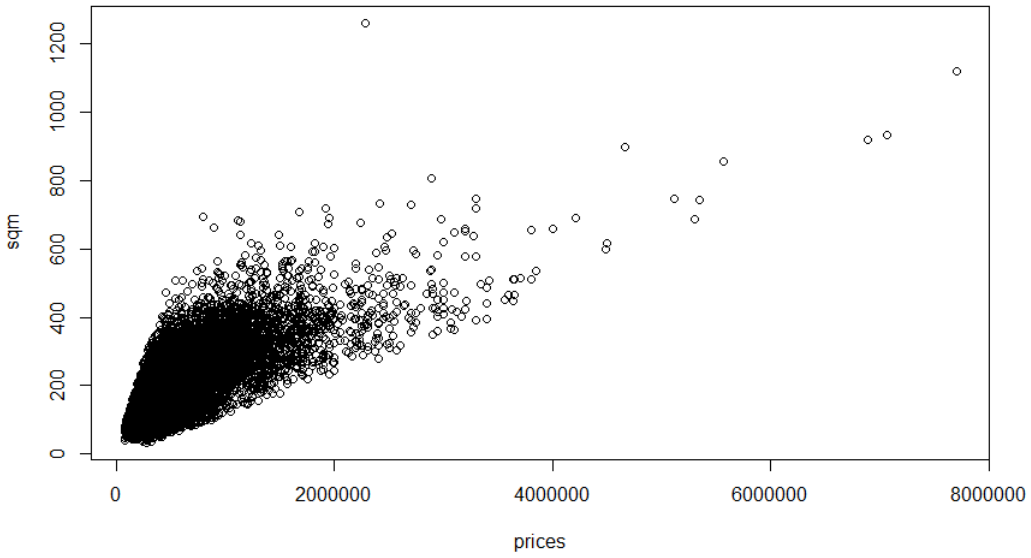
U radu ćemo koristiti prost slučajan i stratifikovan uzorak u kome se iz stratuma izvlače uzorci prostim slučajnim uzorkovanjem. Takođe cemo koristiti regresiono ocenjivanje prostog slučajnog uzorkovanja. U nastavku sledi teorijska podloga ovih uzorkovanja.



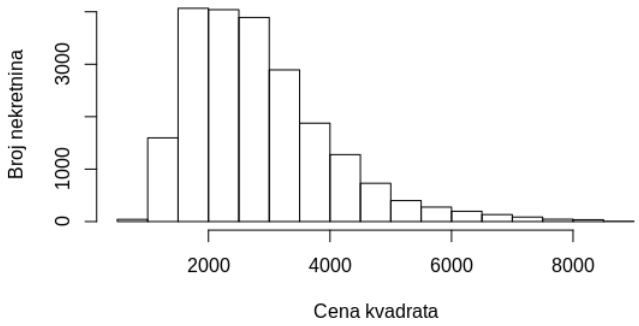
Slika 4: Procenat nekretnina po broju kupatila



Slika 5: Procenat nekretnina po kvadratnim metrima



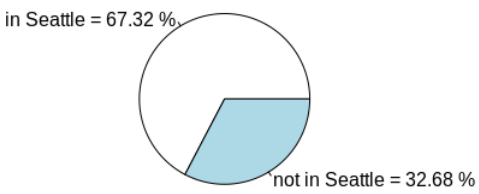
Slika 6: Korelacija između cene i broja kvadratnih metara



Slika 7: Histogram cena metra kvadratnog

3.1 Prost slučajan uzorak

Prost slučajan uzorak bez ponavljanja je plan izbora uzorka koji iz populacije od N jedinica, nasumično bira n različitih jedinica tako da svaka kombinacija ima jednaku verovatnoću [3]. Verovatnoća da određena



Slika 8: Procenat nekretnina u Sijetlu i van njega

kombinacija bude izabrana iznosi:

$$p = \begin{cases} \binom{N}{n}^{-1} & \text{ako je obim uzorka jednak n} \\ 0 & inace \end{cases}$$

Ocena populacijske srednje vrednosti iznosi:

$$\hat{m}_Y = \frac{1}{n} \sum_{k \in S} y_k$$

Ova ocena je nepristrasna. Disperzija ove ocene i ocena te disprezije su:

$$D(\hat{m}_Y) = \frac{\sigma^2}{n}(1 - \frac{n}{N})$$

$$D(\hat{\hat{m}}_Y) = \frac{S^2}{n}(1 - \frac{n}{N})$$

Kada su u pitanju intervalne ocene, aproksimativni $100(1-\alpha)\%$ interval poeverenja izgleda:

$$I_{\hat{m}_Y} = [\hat{m}_Y - z_{1-\frac{\alpha}{2}} \sqrt{D(\hat{\hat{m}}_Y)}, \hat{m}_Y + z_{1-\frac{\alpha}{2}} \sqrt{D(\hat{\hat{m}}_Y)}]$$

- m_Y - populacijska srednja vrednost
- S - prost slučajan uzorak
- σ^2 - populacijska disperzija
- S^2 - uzoračka disperzija
- $z_{1-\frac{\alpha}{2}}$ - vrednost $(1 - \frac{\alpha}{2})$ kvantila standardne normalne raspodele. Ukoliko je veličina uzorka manja od 30 onda se koristi studentova raspodela.

3.2 Regresiono ocenjivanje

Regresiono ocenjivanje je tehnika ocenjivanja, koja koristi neke dodatne informacije radi postizanja veće preciznosti ocena [3]. Koristi se u slučajevima kada se veza između glavnog i pomoćnog obeležja najbolje može opisati regresionom pravom koja ne prolazi kroz koordinatni početak, odnosno u situacijama kada se smatra da između obeležja postoji približno linearna veza, koja ne prolazi kroz koordinatni početak. Da bismo mogli da koristimo regresiono ocenjivanje, potrebno je da nam bude poznata suma ili srednja vrednost pomoćnog obeležja. Prvo definišemo koeficijent korelacije ciljnog i pomoćnog obelezja kao

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

i uzorački koeficijent korelacije kao

$$\hat{\rho} = \frac{\sum_{i \in S}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{(n-1)s_n(x)s_n(y)}$$

Ocena srednje vrednosti obeležja x je data sa

$$\hat{x}_{LR} = \hat{x}_n + \hat{b}(\bar{y} - \bar{y}_n)$$

gde je

$$\hat{b} = \hat{\rho}\frac{s_n(x)}{s_n(y)}$$

Disperzija ove ocene je

$$D(\hat{x}_{LR}) = \frac{\sigma_d^2}{n}(1 - \frac{n}{N}))$$

a ocena te disprezije:

$$D(\hat{\hat{x}}_{LR}) = \frac{\bar{s}_d^2}{n}(1 - \frac{n}{N}))$$

gde su

$$\sigma_d^2 = \frac{1}{N-1}\sum_{k=1}^{N}(x_k - (\bar{x} + b(y_k - \bar{y})))^2$$

$$\bar{s}_d^2 = \frac{1}{n-1}\sum_{k \in S}(x_k - (\bar{x}_n + b(y_k - \bar{y}_n)))^2$$

3.3 Straifikovano uzorkovanje

Stratifikacija podrazumeva podelu populacije na delove koji se nazivaju stratumi, pri čemu treba formirati relativno homogene, među sobom različite stratume, što znači da vrednosti obeležja, koje je predmet istraživanja, treba da budu približne na jedinkama unutar jednog stratuma, ali da se bitno razlikuju među stratumima. Dobijeni stratumi su disjunktni i treba da zadovoljavaju uslov pokrivenosti [3]. Odnosno treba da važi:

$$N_1 + N_2 + ... + N_l = N$$

gde je $N_h, \forall h = 1,\bar{}L$ veličina stratuma, a N ukupna veličina populacije. Uzorak veličine n dobijen iz stratifikovane populacije sadrži iz svakog stratuma uzorak veličine $n_h, \forall h = 1,\bar{}L$, gde važi da je:

$$n_1 + n_2 + ... + n_l = n$$

Ocena za uzoračku sredinu je

$$\bar{x}_s tr = \frac{1}{N} \sum_{h=1}^{L} \frac{N_h^2 s_h^2}{n_h} (1 - \frac{n_h}{N_h})$$

Disperzija te ocene i ocena te disperzije su:

$$D(\bar{x}_{str} = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 s_h^2}{n_h} (1 - \frac{n_h}{N_h})$$

$$D(\hat{\bar{x}}_{str} = \frac{1}{N^2} \sum_{h=1}^{L} \frac{N_h^2 s_{n_h}^2}{n_h} (1 - \frac{n_h}{N_h})$$

Nakon stratifikacije, iz svakog stratuma biraju se uzorci, unapred određenih obima. Uzorci se biraju međusobno nezavisno iz različitih stratuma i nije neophodno koristiti isti plan uzorkovanja za sve stratume.

Veličine uzoraka iz stratuma mogu da se uzimaju proporcionalno broju jedinki u tom stratumu, odnosno:

$$\frac{n_1}{N_1} = ... = \frac{n_L}{N_L} = \frac{n}{N} \Rightarrow n_h = \frac{n}{N} N_h, h = 1, ..., L$$

Drugi način za odabir veličine uzoraka je Nejmanov raspored. On uzima u obzir disperzije u stratumima:

$$n_h = \frac{N_h s_h}{\sum_{h=1}^{L} N_h s_h} n, h = 1, ..., L.$$

4 Rezultati uzorkovanja

Za ocenu srednje vrednosti nad ciljnim obeležjem cena nekretnine vršimo prosto slučajno uzorkovanje, regresiono ocenjivanje gde je pomoćna promenljiva broj metara kvadratnih i stratifikovano uzorkovanje gde su stratumi podeljeni po brojevima kupatila u objektu (ovde se takođe broj kupatila zaokružuje na ceo broj i svi objekti sa preko 3 kupatila se svrstavaju u isti stratum). Za ocenu srednje vrednosti nad ciljnim obeležljem cena metra kvadratnog nekretnine vršimo prosto slučajno uzorkovanje i stratifikovano uzorkovanje gde su stratumi podeljeni po brojevima kupatila u objektu.

Prava srednja vrednost ciljnog obeležja je 540259. Prema tome, dobili smo da je stratifikovano uzorkovanje dalo najpribližniju vrednost. Međutim, što se tiče ocene disperzije ocena, ovo uzorkovanje ima najgori rezltat. Najbolji rezltat daje ocena disperzije regresionog ocenjivanja. Što se tiče intervala poverenja, i ovde se regresiono ocenjivanje pokazuje kao najbolje mada ni stratifikovano uzorkovanje ne daje loš rezultat. Rezultati se mogu videti u tabeli 1.

Što se tiče prosečne cene metra kvadratnog, gde je prava srednja vrednost 2843.03, tu se bolje pokazalo prosto slučajno uzorkovanje mada stratifikovano uzorkovanje nije mnogo lošije. Rezultati se mogu videti u tabeli 2.

Tabela 1: Ocene srednje vrednosti, ocene disperzija tih ocena i intervali tih ocena za tri ocenjvianja.

plan	$\hat{m}_Y$	$\hat{D}(\hat{m}_Y)$	$I_{\hat{m}_Y}$
PSU	531920.8	69334651	[515600.7, 548240.9]
REG	544836.3	41271541	[532244.9, 557427.7]
STRAT	538549.5	86723407	[520297.2, 556801.7]

Tabela 2: Ocene srednje vrednosti, ocene disperzija tih ocena i intervali tih ocena za obeležje cena metra kvadratnog.

plan	$\hat{m}_Y$	$\hat{D}(\hat{m}_Y)$	$I_{\hat{m}_Y}$
PSU	2848.116	905.7124	[2789.130, 2907.101]
STRAT	2822.609	1116.006	[2757.133, 2888.085]

5 Zaključak

Možemo primetiti da svi načini uzorkovanja daju prihvatljive rezultate. Što znači da lako možemo oceniti koliko prosečno košta stan u nekom gradu, okrugu ili zemlji sa relativno malim uzorkom u odnosu na populaciju i možemo primeniti bilo koji od metoda primenjenih u radu.

U ovom konkretnom primeru (King County okrug) vidimo da nije velika razlika u cenama između velikog grada i okruga, što ne znači da u nekom drugom okrugu to ne bi bio slučaj. Da bi se ova hipoteza proverila potrebno je raditi na dosta više baza podataka.

Literatura

[1] harlfoxem. House Sales in King County, USA. on-line at: https://www.kaggle.com/harlfoxem/housesalesprediction.

[2] Robert V. Krejcije and Daryle V. Morgan. Determining sample size for research activities. on-line at: https://home.kku.ac.th/sompong/guest_speaker/KrejcieandMorgan_article.pdf.

[3] Astrea Camstra Reinder Banning and Paul Knottnerus. *Sampling Theory, Sampling design and estimation models.* Statistics Netherlands, 2012.

A Dodatak

```
install.packages("ggplot2")
library(ggplot2)
library(scales)

#otklanjamo e notifikaciju sa realnih brojeva
options(scipen=999)

#konvertujemo stope kvadratne u metre kvadtarne
f2_to_m2 = (0.3048)^2

N = length(kc_house_data$id)

#izbacujemo redove koji imaju 0 soba ili 0 kupatila jer su to verovatno greske
indexes_wo_ano = c()
for (i in 1:N){
  if (kc_house_data$bedrooms[i] > 0 && kc_house_data$bathrooms[i] > 0.5)
  {
    indexes_wo_ano = c(indexes_wo_ano, i)
  }
}
house_data_prep = kc_house_data[indexes_wo_ano,]

N = length(house_data_prep$id)

#pravimo kolone za metre kvadratne umesto kolona za stope kvadratne
house_data_prep$sqm_living = house_data_prep$sqft_living*f2_to_m2
house_data_prep$sqm_lot = house_data_prep$sqft_lot*f2_to_m2
house_data_prep$sqm_above = house_data_prep$sqft_above*f2_to_m2
house_data_prep$sqm_basement = house_data_prep$sqft_basement*f2_to_m2
house_data_prep$sqm_living15 = house_data_prep$sqft_living15*f2_to_m2
house_data_prep$sqm_lot15 = house_data_prep$sqft_lot15*f2_to_m2
house_data_prep = subset(house_data_prep, select=-c(sqft_living, sqft_lot, sqft_above, sqft_basement, sqft_living15, sqft_lot15))
#dimnames(house_data_prep)

#histogram cena
hist(house_data_prep$price, xlab="Cena nekretnine", ylab = "Broj nekretnina", main = NULL)
#ggplot(house_data_prep, aes(x=price)) + geom_histogram()

hist(house_data_prep$sqm_living, xlab="Broj metara kvadratnih", ylab = "Broj nekretnina", main = NULL)

#procenat nekretnina koje kostaju manje od 1,000,000 dolara
n_lt_1m = length(house_data_prep[house_data_prep$price<1000000,1])
perc_n_lt_1m = n_lt_1m/N*100
perc_n_lt_1m

#minimalna i maksimalna cena
min(house_data_prep$price)
max(house_data_prep$price)

#racunamo cene jednog kvadrata svake nekretnine kao i njihovu prosecnu ocenu
house_data_prep$price_of_m2 = house_data_prep$price/house_data_prep$sqm_living
house_data_prep$price_of_m2
hist(house_data_prep$price_of_m2, xlab = "Cena kvadrata", ylab = "Broj nekretnina", main = NULL)

#sve kuce sa 4 ili vise soba kategorisemo zajedno
bedrooms = house_data_prep$bedrooms
for (i in 1:length(bedrooms)){
  if (bedrooms[i] > 4){
    bedrooms[i] = 5
```

```
  }
}

#pravimo pie chart za broj soba
mytable = table(bedrooms)
lbls = paste(names(mytable), "=", round(mytable/sum(mytable)*100,
    digits = 2),"%", sep=" ")
lbls[5] = paste("5+ ", "=", round(mytable[5]/sum(mytable)*100,
    digits = 2),"%", sep=" ")
pie(mytable, labels = lbls,
    main=NULL)

#sve kuce sa 3 ili vise kupatila kategorisemo zajedno
bathrooms = house_data_prep$bathrooms
for (i in 1:length(bathrooms)){
  bathrooms[i] = round(bathrooms[i])
  if (bathrooms[i] > 3){
    bathrooms[i] = 4
  }
}

#pravimo pie chart za broj kupatila
mytable = table(bathrooms)
lbls = paste(names(mytable), "=", round(mytable/sum(mytable)*100,
    digits = 2),"%", sep=" ")
lbls[4] = paste("4+ ", "=", round(mytable[4]/sum(mytable)*100,
    digits = 2),"%", sep=" ")
pie(mytable, labels = lbls,
    main=NULL)

#diskretizujemo kolonu sa metrima kvadratnim
house_data_prep$sqm_living_bin = cut(house_data_prep$sqm_living,
    breaks = c(0,100,200,300,400,1300), labels = c("0-100", "
    100-200", "200-300", "300-400", "400-1300"))

#pravimo pie chart sa diskretizovanim podacima
bins = house_data_prep$sqm_living_bin
mytable = table(bins)
lbls = paste(names(mytable), "=", round(mytable/sum(mytable)*100,
    digits = 2),"%", sep=" ")
pie(mytable, labels = lbls,
    main=NULL)

#pravimo grafik korelacije izmedju cene i metara kvadratnih
    nekretnine
plot(house_data_prep$price, house_data_prep$sqm_living, xlab = "
    prices", ylab = "sqm")
cor(house_data_prep$price, house_data_prep$sqm_living)

#racunamo idealnu velicinu uzorka
n = ceiling(qchisq(0.95, 7)*N*0.5*0.5/(0.05^2*(N-1)) + qchisq
    (0.95, 7)*0.5*0.5)
n
procenat_velicine_uzorka = n/N*100

#postavljamo seed na fiksnu vrednost zbog vise pokretanja
set.seed(6)

#Vadjenje prostog slucajnog uzorka
sample_psu = sample(1:N, n)

#Ocena srednje vrednosti prostog slucajnog uzorka
xn_ocena_psu = mean(house_data_prep$price[sample_psu])

#Ocena disprezije ocene srednje vrednosti prostog slucajnog uzorka
s2_psu = var(house_data_prep$price[sample_psu])
d_ocena_psu = s2_psu/n*(1-n/N)

xn_ocena_psu
d_ocena_psu

#Disperzija srednje vrednosti prostog slucajnog uzorka
var(house_data_prep$price)/n*(1-n/N)

```

```
#Racunjanje intervala poverenja za psu
aplha = 0.05

#Posto je n>30 onda koristimo normalnu raspodelu
z = qnorm(1 - aplha/2)
interval = c(xn_ocena_psu - z*sqrt(d_ocena_psu), xn_ocena_psu + z*
    sqrt(d_ocena_psu))
interval

#Regresiono ocenjivanje gde je pomocna promenljiva broj metara
    kvadratnih zatvorenog prostora
uzorak_za_reg = sample(1:N, n)

#Racunanje pomocnih promenljivih ro i b
ro = cor(house_data_prep$price[uzorak_za_reg],
    house_data_prep$sqm_living[uzorak_za_reg])
b = ro*sqrt(var(house_data_prep$price[uzorak_za_reg]))/sqrt(var(
    house_data_prep$sqm_living[uzorak_za_reg]))

#Racunanje ocene srednje vrednosti regresionim ocenjivanjem
xn_lr = mean(house_data_prep$price[uzorak_za_reg])+b*(mean(
    house_data_prep$sqm_living[uzorak_za_reg] - mean(
    house_data_prep$sqm_living[uzorak_za_reg])))

#Racunanje ocene disperzije srednje vrednosti regresionim
    ocenjivanjem
se = 1/(n-1)*sum((house_data_prep$price[uzorak_za_reg] - (mean(
    house_data_prep$price[uzorak_za_reg]) + b*(
    house_data_prep$sqm_living[uzorak_za_reg] - mean(
    house_data_prep$sqm_living[uzorak_za_reg]))))^2)
d_ocena_reg = se/n*(1-n/N)

#Racunanje inervala poverenja za regresiono ocenjivanje
interval_reg = c(xn_lr - z*sqrt(d_ocena_reg), xn_lr + z*sqrt(
    d_ocena_reg))
interval_reg

xn_lr
d_ocena_reg

#Prava srednja vrednost
mean(house_data_prep$price)

#Pravimo stratume od obelezja bathrooms pa hocemo da
    diskretizujemo to obelezje
#Takodje sve nekretnine sa 4 ili vise kupatila svrstavamo u jedan
    stratum
house_data_prep$bathrooms_rounded = round(
    house_data_prep$bathrooms)
for (i in 1:N){
  if (house_data_prep$bathrooms_rounded[i] > 4){
    house_data_prep$bathrooms_rounded[i] = 4
  }
}
unique(house_data_prep$bathrooms_rounded)

#Pravimo listu stratuma
stratumi = list()
stratumi[[1]] = house_data_prep[house_data_prep$bathrooms_rounded
    == 1,]
stratumi[[2]] = house_data_prep[house_data_prep$bathrooms_rounded
    == 2,]
stratumi[[3]] = house_data_prep[house_data_prep$bathrooms_rounded
    == 3,]
stratumi[[4]] = house_data_prep[house_data_prep$bathrooms_rounded
    == 4,]
stratumi[[1]]

#Cuvamo disperzije svakog stratuma i velicine
disprezije_stratuma = c(sqrt(var(stratumi[[1]]$price)), sqrt(var(
    stratumi[[2]]$price)), sqrt(var(stratumi[[3]]$price)), sqrt(
    var(stratumi[[4]]$price)))
N_stratuma = c(length(stratumi[[1]]$price), length(stratumi[[2]]
    $price), length(stratumi[[3]]$price), length(stratumi[[4]]
```

```
    $price))

#Nejmanovom rapodelom odredjujemo velicine uzorka iz svakog
    stratuma
n_stratuma = round(n*(N_stratuma*disprezije_stratuma)/sum(
    N_stratuma*disprezije_stratuma))

#Posto smo odmah dobili zeljeni ukupni broj elemenata u uzoraku,
    ne moramo da dodajemo ili oduzimamo elemente
n_stratuma
sum(n_stratuma) == n

#Uzimamo prost slucajan uzorak iz svakog stratuma
indeksi_strat = c()
indeksi_strat[[1]] = sample(1:N_stratuma[1], n_stratuma[1],
    replace = F)
indeksi_strat[[2]] = sample(1:N_stratuma[2], n_stratuma[2],
    replace = F)
indeksi_strat[[3]] = sample(1:N_stratuma[3], n_stratuma[3],
    replace = F)
indeksi_strat[[4]] = sample(1:N_stratuma[4], n_stratuma[4],
    replace = F)

#Racunamo ocene srednje vrednosti svakog stratuma
xn_strat = c()
xn_strat[1] = mean(house_data_prep$price[indeksi_strat[[1]]])
xn_strat[2] = mean(house_data_prep$price[indeksi_strat[[2]]])
xn_strat[3] = mean(house_data_prep$price[indeksi_strat[[3]]])
xn_strat[4] = mean(house_data_prep$price[indeksi_strat[[4]]])

#Racunamo ocene uzoracke disperzije svakog stratuma
uzoracka_disperzija_stratuma = c()
uzoracka_disperzija_stratuma[1] = var(house_data_prep$price[
    indeksi_strat[[1]]])
uzoracka_disperzija_stratuma[2] = var(house_data_prep$price[
    indeksi_strat[[2]]])
uzoracka_disperzija_stratuma[3] = var(house_data_prep$price[
    indeksi_strat[[3]]])
uzoracka_disperzija_stratuma[4] = var(house_data_prep$price[
    indeksi_strat[[4]]])

#Racunamo ocenu srednje vrednosti dobijenu stratifikovanim
    uzorkovanjem
xn_st = 1/N*sum(N_stratuma*xn_strat)
xn_st

#Racunamo ocenu disperzije te ocene
d_ocena_st = 1/N^2*sum((N_stratuma^2*uzoracka_disperzija_stratuma)
    /n_stratuma*(1-n_stratuma/N_stratuma))
d_ocena_st

#Racunamo interval poverenja te ocene
interval_strat = c(xn_st - z*sqrt(d_ocena_st), xn_st + z*sqrt(
    d_ocena_st))
interval_strat

#Aproskimativno delimo nekretnine na one koje su u Sijetlu i one
    koje su van njega
lat_min = 47.501221
long_max = -122.238907
lat_max = 47.735211
long_min = -122.443828
house_data_prep$is_seattle = ifelse(house_data_prep$lat > lat_min
    & house_data_prep$lat < lat_max & house_data_prep$long >
    long_min & house_data_prep$long < long_max, 1, 0)
unique(house_data_prep$is_seattle)

mytable = table(house_data_prep$is_seattle)
lbls = paste(c("in Seattle", "not in Seattle"), "=", round(mytable
    /sum(mytable)*100, digits = 2),"%", sep=" ")
pie(mytable, labels = lbls,
    main=NULL)
```

```
#Zelimo da vidimo da li postoji korelacije izmedju cene i toga da li se nekretnina nalazi u Sijetlu
#Ispostavlja se da nema korelacije
cor(house_data_prep$price, house_data_prep$is_seattle)

#Isto to samo za obelezje cena metra kvadatnog
cor(house_data_prep$price_of_m2, house_data_prep$is_seattle)

#U ovom delu koda ocenjijemo srednju vrednost za obeleyje cena metra kvadratnog
#Koristimo PSU i straifikovano uzorkovanje po broju kupatila
#Veoma slicno kao do sad tako da nece biti komentara
sample_psu_2 = sample(1:N, n)

xn_ocena_psu_2 = mean(house_data_prep$price_of_m2[sample_psu_2])

s2_psu_2 = var(house_data_prep$price_of_m2[sample_psu_2])
d_ocena_psu_2 = s2_psu_2/n*(1-n/N)

xn_ocena_psu_2
d_ocena_psu_2
mean(house_data_prep$price_of_m2)

interval_psu_2 = c(xn_ocena_psu_2 - z*sqrt(d_ocena_psu_2), xn_ocena_psu_2 + z*sqrt(d_ocena_psu_2))
interval_psu_2

indeksi_strat_2 = c()
indeksi_strat_2[[1]] = sample(1:N_stratuma[1], n_stratuma[1], replace = F)
indeksi_strat_2[[2]] = sample(1:N_stratuma[2], n_stratuma[2], replace = F)
indeksi_strat_2[[3]] = sample(1:N_stratuma[3], n_stratuma[3], replace = F)
indeksi_strat_2[[4]] = sample(1:N_stratuma[4], n_stratuma[4], replace = F)

xn_strat_2 = c()
xn_strat_2[1] = mean(house_data_prep$price_of_m2[indeksi_strat_2[[1]]])
xn_strat_2[2] = mean(house_data_prep$price_of_m2[indeksi_strat_2[[2]]])
xn_strat_2[3] = mean(house_data_prep$price_of_m2[indeksi_strat_2[[3]]])
xn_strat_2[4] = mean(house_data_prep$price_of_m2[indeksi_strat_2[[4]]])

uzoracka_disperzija_stratuma_2 = c()
uzoracka_disperzija_stratuma_2[1] = var(house_data_prep$price_of_m2[indeksi_strat_2[[1]]])
uzoracka_disperzija_stratuma_2[2] = var(house_data_prep$price_of_m2[indeksi_strat_2[[2]]])
uzoracka_disperzija_stratuma_2[3] = var(house_data_prep$price_of_m2[indeksi_strat_2[[3]]])
uzoracka_disperzija_stratuma_2[4] = var(house_data_prep$price_of_m2[indeksi_strat_2[[4]]])

xn_st_2 = 1/N*sum(N_stratuma*xn_strat_2)
xn_st_2

d_ocena_st_2 = 1/N^2*sum((N_stratuma^2*uzoracka_disperzija_stratuma_2)/n_stratuma*(1-n_stratuma/N_stratuma))
d_ocena_st_2

interval_strat = c(xn_st_2 - z*sqrt(d_ocena_st_2), xn_st_2 + z*sqrt(d_ocena_st_2))
interval_strat
```

Listing 1: Kod korišćen u radu