Projekt - Analiza tržišta nekretnina

Grupa FerovkeiFerovac

Marko Dodik, Silvija Gojević, Lucija Mičić, Antonia Žaja

15.01.2023.

Opis projekta:

Ovaj projekt obavezni je dio izbornog kolegija Statistička analiza podataka na Fakultetu elektrotehnike i računarstva. Svrha projekta je primjena teorijskih temelja stečenih na predavanjima na skup podataka iz stvarnog svijeta. Kao pomoć u izradi projekta korišten je programski jezik R koji je pružio potporu za izvođenje testiranja i bolju vizualizaciju podataka, te programski paket RStudio.

Opis problema:

Kupci često traže nekretnine sa određenim kriterijima (npr. određeni broj soba, veličina dvorišta), no takve "luksuze" ne žele preplatiti. Također, cijene nekretnina zbog razinih razloga znaju biti napuhane, dok je bankama u interesu objektivno procijeniti vrijednost nekretnine za potrebe kreditiranja klijenta. Upravo zato se prikupljaju podaci o prodanim nekretninama Cilj projektnog zadatka je analizirati te podatke i analizirati uspješnost prodaje nekretnina ovisno o značajkama koje ona sadrži.

```r
#učitavanje podataka
data=read.csv('preprocessed_data.csv')
```

Skup podataka:

Skup podataka koji se koristi u ovom projektu predstavlja informacije o prodanim nekretninama u gradu Ames (Iowa, Sjedinjenje Američke Države). Odnosi se na prodane nekretnine u sljedećim godinama: 2006., 2007., 2008., 2009. i 2010. Svaka nekretnina opisana je s 81 značajkom. Neke od značajki su kvadratura (LotArea), naziv susjedstva u kojem se nekretnina nalazi (Neighborhood), veličina podruma (TotalBsmtSF), tip krova (RoofStyle), broj spavaćih soba (Bedroom - nisu uračunate sobe u podrumu), lokacija garaže (GarageType) i slično. Ukupno je prikupljeno 1460 zapisa.

```r
#prikaz svih značajki
names(data)
```

```
##  [1] "Id"            "MSSubClass"    "MSZoning"      "LotFrontage"
##  [5] "LotArea"       "Street"        "Alley"         "LotShape"
##  [9] "LandContour"   "Utilities"     "LotConfig"     "LandSlope"
## [13] "Neighborhood"  "Condition1"    "Condition2"    "BldgType"
## [17] "HouseStyle"    "OverallQual"   "OverallCond"   "YearBuilt"
## [21] "YearRemodAdd"  "RoofStyle"     "RoofMatl"      "Exterior1st"
## [25] "Exterior2nd"   "MasVnrType"    "MasVnrArea"    "ExterQual"
## [29] "ExterCond"     "Foundation"    "BsmtQual"      "BsmtCond"
## [33] "BsmtExposure"  "BsmtFinType1"  "BsmtFinSF1"    "BsmtFinType2"
## [37] "BsmtFinSF2"    "BsmtUnfSF"     "TotalBsmtSF"   "Heating"
## [41] "HeatingQC"     "CentralAir"    "Electrical"    "X1stFlrSF"
## [45] "X2ndFlrSF"     "LowQualFinSF"  "GrLivArea"     "BsmtFullBath"
## [49] "BsmtHalfBath"  "FullBath"      "HalfBath"      "BedroomAbvGr"
## [53] "KitchenAbvGr"  "KitchenQual"   "TotRmsAbvGrd"  "Functional"
## [57] "Fireplaces"    "FireplaceQu"   "GarageType"    "GarageYrBlt"
## [61] "GarageFinish"  "GarageCars"    "GarageArea"    "GarageQual"
## [65] "GarageCond"    "PavedDrive"    "WoodDeckSF"    "OpenPorchSF"
## [69] "EnclosedPorch" "X3SsnPorch"    "ScreenPorch"   "PoolArea"
## [73] "PoolQC"        "Fence"         "MiscFeature"   "MiscVal"
## [77] "MoSold"        "YrSold"        "SaleType"      "SaleCondition"
## [81] "SalePrice"
```

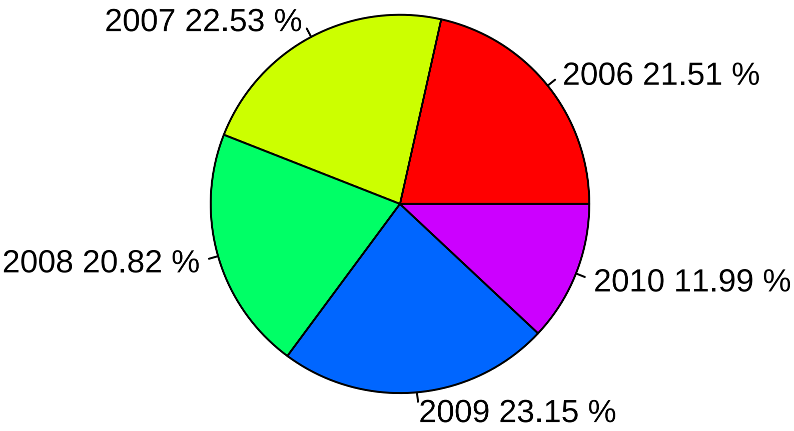
```
#ukupni broj zapisa
nrow(data)
```

```
## [1] 1460
```

Deskriptivna statistika skupa podataka:

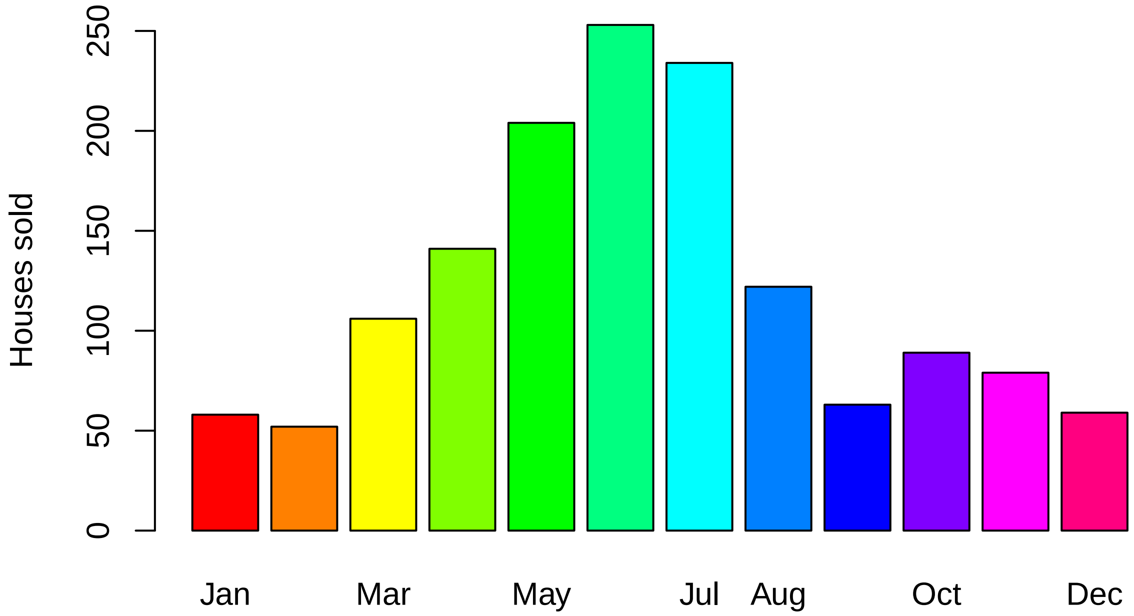
Proučavamo tržište nekretnina u godinama od periodu od 2006. do 2010. godine (uključivo). Na sljedećem dijagramu, prikazani su udjeli broja prodanih nekretnina po godinama. Iz njega vidimo kako je 2009. godine prodan najveći broj nekretnina (23.15%). Za prikaz ovih podataka je odabran strukturni krug.

```
#računanje broja prodanih nekretnina po godinama korištenjem značajke YrSold
values=c(sum(data$YrSold=='2006'),sum(data$YrSold=='2007'),sum(data$YrSold=='2008'),
         sum(data$YrSold=='2009'),sum(data$YrSold=='2010'))
labels=c("2006", "2007", "2008", "2009", "2010")
pct = round(values/sum(values)*100, digits = 2)
labels = paste(labels, pct)
labels = paste(labels,"%")
pie(values, labels=labels, col=rainbow(length(labels)))
```



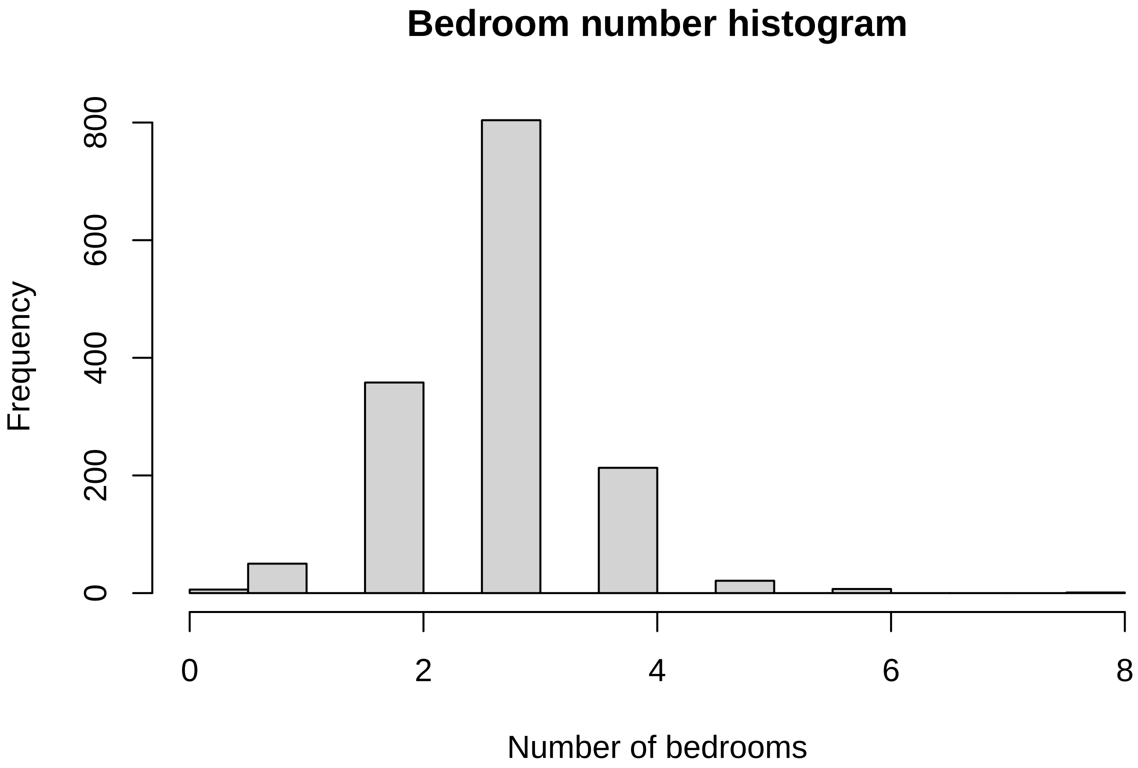
Zanimljiva informacija koja se može saznati iz danih podataka je u kojim mjesecima se tijekom godina najviše nekretnina prodalo. Iz sljedećeg stupičastog dijagrama vidimo kako je to u kasnim proljetnim i ljetnim mjesecima.

```
#računanje broja prodanih nekretnina po mjesecima korištenjem značajke MoSold
values=c(sum(data$MoSold=='1'),sum(data$MoSold=='2'),sum(data$MoSold=='3'),
         sum(data$MoSold=='4'),sum(data$MoSold=='5'),sum(data$MoSold=='6'),
         sum(data$MoSold=='7'),sum(data$MoSold=='8'),sum(data$MoSold=='9'),
         sum(data$MoSold=='10'),sum(data$MoSold=='11'),sum(data$MoSold=='12'))
labels=c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep",
         "Oct", "Nov", "Dec")
barplot(values, ylab = "Houses sold", names.arg = labels, col = rainbow(length(labels)))
```



Sljedeći dijagram prikazuje broj spavaćih soba prodanih nekretnina. Iz njega možemo zaključiti da prosječni broj spavaćih soba prodanih nekretnina iznosi 3. Najmanji broj spavaćih soba među prikupljenim podacima je 0, a najveći 8. Za prikaz ovih podataka odabran je histogram, broj razreda je 25.

```
#histogram kreiran korištenjem značajke BedroomAbvGr
hist(data$BedroomAbvGr,main='Bedroom number histogram',xlab='Number of bedrooms',
     ylab='Frequency', breaks=25)
```



```r
#prosječan broj spavaćih soba
mean(data$BedroomAbvGr)
```

```
## [1] 2.866438
```

S obzirom da je najveći broj prodanih nekretnina bio 2009. godine, zanima nas prosječni broj spavaćih soba te godine u usporedbi s ostalim godinama.

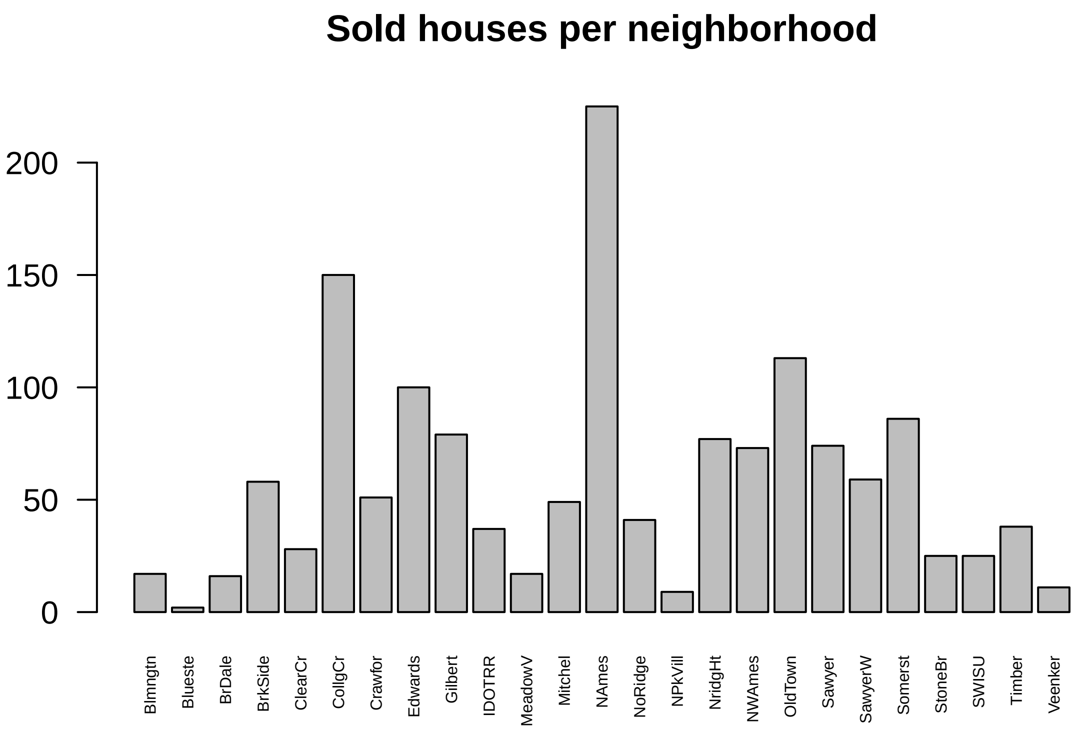
```r
#grupiranje nekretnina pomoću značajke YrSold
houses_sold_2006 = data[data$YrSold == "2006",]
houses_sold_2007 = data[data$YrSold == "2007",]
houses_sold_2008 = data[data$YrSold == "2008",]
houses_sold_2009 = data[data$YrSold == "2009",]
houses_sold_2010 = data[data$YrSold == "2010",]

df <- data.frame(year = c("2006", "2007", "2008", "2009", "2010"),
                 bedrooms =c(mean(houses_sold_2006$BedroomAbvGr),
                             mean(houses_sold_2007$BedroomAbvGr),
                             mean(houses_sold_2008$BedroomAbvGr),
                             mean(houses_sold_2009$BedroomAbvGr),
                             mean(houses_sold_2010$BedroomAbvGr)))
```

```r
#prikaz podataka u obliku tablice radi preglednosti
as.data.frame(t(df)) %>% kable(col.names = NULL)
```

year	2006	2007	2008	2009	2010
bedrooms	2.872611	2.951368	2.855263	2.784024	2.874286

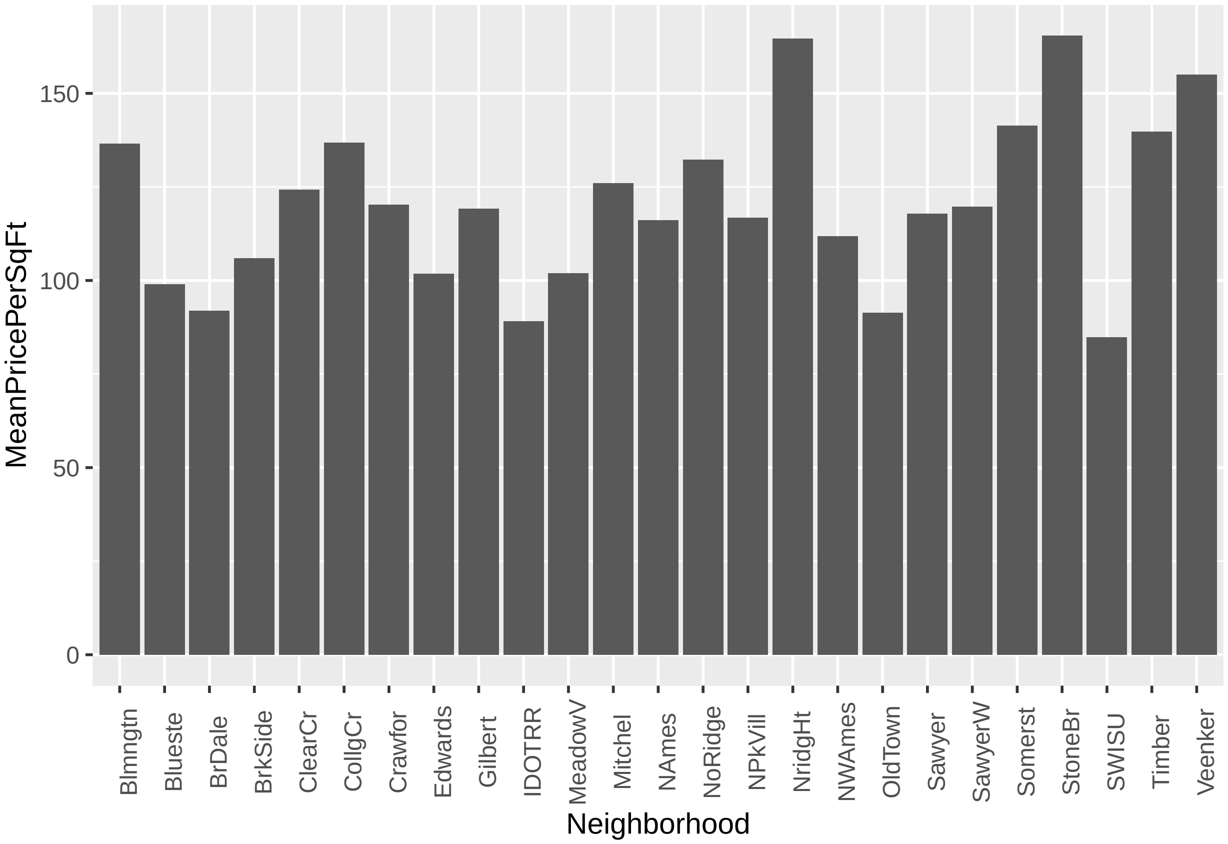
Sljedeći dijagram prikazuje distribuciju prodanih nekretnina u ovisnosti o kvartu u kojem se nalaze. Iz dijagrama je vidljivo da se najveći broj prodanih nekretnina nalazi u kvartu North Ames. Za prikaz podataka korišten je stupičasti dijagram.

```r
blmngtn = which(data$Neighborhood=='Blmngtn')/1460*100
blueste = which(data$Neighborhood=='Blueste')/1460*100
brdale = which(data$Neighborhood=='BrDale')/1460*100
brkside = which(data$Neighborhood=='BrkSide')/1460*100
clearcr = which(data$Neighborhood=='ClearCr')/1460*100
collgcr = which(data$Neighborhood=='CollgCr')/1460*100
crawfor = which(data$Neighborhood=='Crawfor')/1460*100
edwards = which(data$Neighborhood=='Edwards')/1460*100
gilbert = which(data$Neighborhood=='Gilbert')/1460*100
IDOTRR = which(data$Neighborhood=='IDOTRR')/1460*100
barplot(table(data$Neighborhood),las=2,cex.names=.5,main='Sold houses per neighborhood')
```



Nadalje, mogli bi se zapitati koja je prosječna cijena kvadrata po kvartu, kako bi dobili bolji uvid u poželjnije kvartove za život. Na temelju stupičastog dijagrama, zaključujemo kako su cijene nekretnina najpovoljnije u kvartu South & West of Iowa State University, a najskuplje u Stone Brook kvartu.

```
data$PricePerSqFt <- data$SalePrice / data$GrLivArea
data_by_neighborhood <- data %>% group_by(Neighborhood) %>%
  summarize(MeanPricePerSqFt = mean(PricePerSqFt))

ggplot(data_by_neighborhood, aes(x = Neighborhood, y = MeanPricePerSqFt)) +
  geom_bar(stat = "identity")+ theme(axis.text.x = element_text(angle = 90))
```



Statističko zaključivanje

Ovisnost broja katova nekretnine o obliku zemljišne čestice

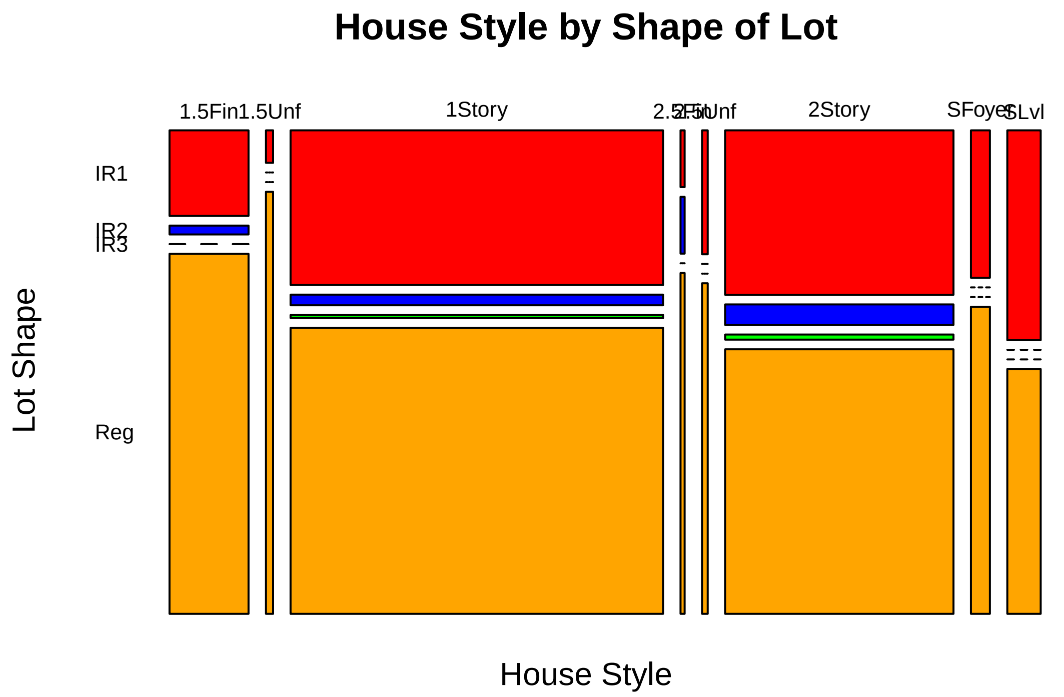
Svaka nekretnina ima određeni oblik (IR1, IR2, IR3, Reg). Zanima nas razlikuje li se broj katova nekretnine obzirom na njen oblik, odnosno želimo provjeriti imaju li nekretnine određenog oblika veći broj katova nego ostale. Kako bi provjerili postoji li veza koja bi objasnila ovisnost tih dvaju atributa, provodimo hi-kvadrat test. Test nezavisnosti χ^2 test u programskom paketu R implementiran je u funkciji `chisq.test()` koja kao ulaz prima kontingencijsku tablicu podataka koje testiramo na nezavisnost. Testom utvrđujemo p-vrijednost, koja je manja od 0.05, stoga odbacujemo nultu hipotezu na razini znacajnosti 5%, dakle x i y su zavisne.

H0: X i Y su nezavisne
H1: X i Y su zavisne
Pri čemu je X=LotShape, a Y=HouseStyle

```
tableHSLS <- table(data$HouseStyle, data$LotShape)
chi_squared_test <- chisq.test(tableHSLS,simulate.p.value = T)
chi_squared_test
```

```
##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  tableHSLS
## X-squared = 44.472, df = NA, p-value = 0.02399
```

```
#p_value < 0.05 pa odbacijemo H0 na razini znacajnosti od 5%
colors <- c("red", "blue", "green", "orange")
mosaicplot(tableHSLS,
           col=colors,
           xlab="House Style", ylab = "Lot Shape" ,main = "House Style by Shape of Lot",
           las=1)
```



Ovisnost cijene kvadrata nekretnine o broju spavaćih soba

Za utvrđivanje postoji li ovisnost između cijene kvadrata i broja spavaćih soba pokušat ćemo provest ANOVA test. ANOVA (analiza varijance) je statistički test koji se koristi za usporedbu srednjih vrijednosti više od dvije grupe. Često je dobro rješenje kada želimo utvrditi postoji li značajna razlika između srednjih vrijednosti više od dvije grupe, jer nam omogućuje testiranje više grupa odjednom. Za početak potrebno je provjeriti homogenost i normalnost, međutim Bartlettovim testom utvrđujemo kako uzorak nema homogene varijance, a ni normalnu razdiobu po grupama, što je provjereno Lillieforsovom inačicom Kolmogorov-Smirnov testa. Stoga ćemo provesti Kruskal-Wallis test za provjeru uvjetuje li broj soba cijenu kvadrata. P-vrijednost dobivenog rezultata je manja od 0.05, pa na razini značajnosti 5% odbacujemo hipotezu da su varijable PricePerSqFt i BedroomAbvGr nezavisne, stoga zaključujemo da broj soba uvjetuje cijenu kvadrata.

```
library(dplyr)
data$BedroomAbvGr <- as.factor(data$BedroomAbvGr)
data$PricePerSqFt <- as.numeric(data$SalePrice) / as.numeric(data$GrLivArea)
```

```
#provjera normalnosti
require(nortest)
```

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

```
lillie.test(data$PricePerSqFt[data$BedroomAbvGr=='0'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
## data:  data$PricePerSqFt[data$BedroomAbvGr == "0"]
## D = 0.15724, p-value = 0.9133
```

```
lillie.test(data$PricePerSqFt[data$BedroomAbvGr=='1'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$PricePerSqFt[data$BedroomAbvGr == "1"]
## D = 0.071617, p-value = 0.7527
```

```
lillie.test(data$PricePerSqFt[data$BedroomAbvGr=='2'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$PricePerSqFt[data$BedroomAbvGr == "2"]
## D = 0.032814, p-value = 0.4581
```

```
lillie.test(data$PricePerSqFt[data$BedroomAbvGr=='3'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$PricePerSqFt[data$BedroomAbvGr == "3"]
## D = 0.0363, p-value = 0.01398
```

```
lillie.test(data$PricePerSqFt[data$BedroomAbvGr=='4'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$PricePerSqFt[data$BedroomAbvGr == "4"]
## D = 0.078838, p-value = 0.002637
```

```
lillie.test(data$PricePerSqFt[data$BedroomAbvGr=='5'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$PricePerSqFt[data$BedroomAbvGr == "5"]
## D = 0.12237, p-value = 0.5661
```

```
lillie.test(data$PricePerSqFt[data$BedroomAbvGr=='6'])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data$PricePerSqFt[data$BedroomAbvGr == "6"]
## D = 0.16271, p-value = 0.8318
```

Za provjeru normalnosti koristimo Lillieforsovu inačicu KS testa, već nakon provjera za nekretnine s 3 sobe utvrđujemo da pretpotavka normalnosti nije zadovoljena. Iako već sad znamo da ne možemo provesti ANOVA test, provjerit ćemo i homogenost kao u auditornim vježbama. Nekretnina sa sedam soba nema, a s osam postoji samo jedna.

```
bartlett.test(
  list(data$PricePerSqFt[data$BedroomAbvGr=='0'],
       data$PricePerSqFt[data$BedroomAbvGr=='1'],
       data$PricePerSqFt[data$BedroomAbvGr=='2'],
       data$PricePerSqFt[data$BedroomAbvGr=='3'],
       data$PricePerSqFt[data$BedroomAbvGr=='4'],
       data$PricePerSqFt[data$BedroomAbvGr=='5'],
       data$PricePerSqFt[data$BedroomAbvGr=='6']
  )
)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  list(data$PricePerSqFt[data$BedroomAbvGr == "0"], data$PricePerSqFt[data$BedroomAbvGr == "1"]
## Bartlett's K-squared = 59.644, df = 6, p-value = 5.316e-11
```

```
mean(data$PricePerSqFt[data$BedroomAbvGr=='0'])
```

```
## [1] 155.5282
```

```
mean(data$PricePerSqFt[data$BedroomAbvGr=='1'])
```

```
## [1] 148.1292
```

```
mean(data$PricePerSqFt[data$BedroomAbvGr=='2'])
```

```
## [1] 128.6018
```

```
mean(data$PricePerSqFt[data$BedroomAbvGr=='3'])
```

```
## [1] 121.5211
```

```
mean(data$PricePerSqFt[data$BedroomAbvGr=='4'])
```

```
## [1] 102.1886
```

```
mean(data$PricePerSqFt[data$BedroomAbvGr=='5'])
```

```
## [1] 76.58501
```

```
mean(data$PricePerSqFt[data$BedroomAbvGr=='6'])
```

```
## [1] 74.42921
```

Buduci da uzorak nema homogene varijance, probati cemo problem rijesiti s Kruskal-Wallis koji to ne pretpostavlja:

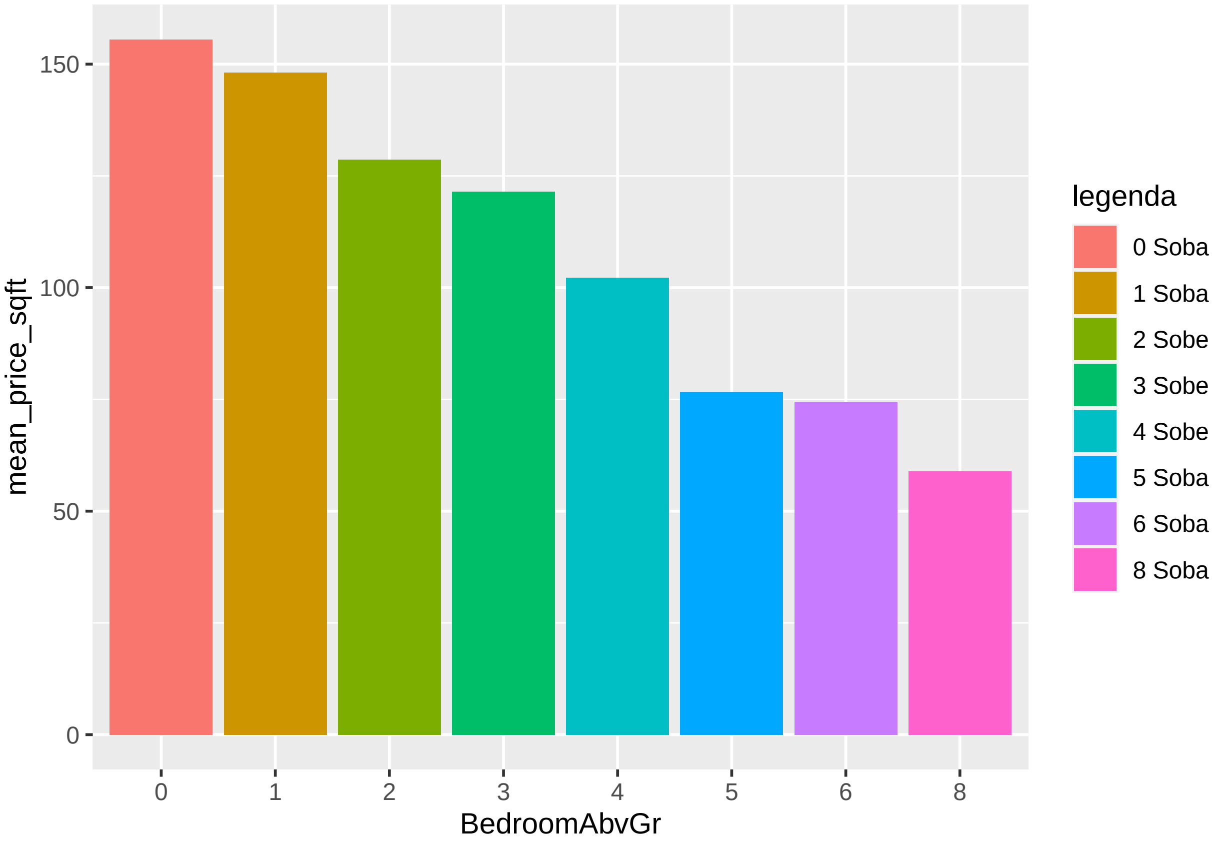
```
kruskal.test(data$PricePerSqFt~data$BedroomAbvGr)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data$PricePerSqFt by data$BedroomAbvGr
## Kruskal-Wallis chi-squared = 199.97, df = 7, p-value < 2.2e-16
```

P vrijednost je manja od 0.05, što znači da odbacujemo hipotezu da cijena po kvadratu ne ovisi o broju soba.

```
data_by_bedrooms <- data %>%
  group_by(BedroomAbvGr)%>%
  summarize(mean_price_sqft =mean(SalePrice/GrLivArea),
            sd_price_sqft=sd(SalePrice/GrLivArea))
legenda = c("0 Soba", "1 Soba", "2 Sobe", "3 Sobe", "4 Sobe", "5 Soba", "6 Soba", "8 Soba")
```

```
ggplot(data = data_by_bedrooms,
       aes(x = BedroomAbvGr, y = mean_price_sqft,fill = legenda)) + geom_bar(stat = "identity")
```



Ovisi li veličina podruma o kvartu u gradu

Svaka prodana nekretnina nalazi se u određenom naselju i ima određenu veličinu podruma. Zanima nas razlikuju li se uspješnosti prodaje nekretnina u određenom naselju s obzirom na veličinu podruma. Također, ovdje provodimo ANOVA test te provjeravamo njegove početne pretpostavke - normalnost i homogenost podataka.

Prvo utvrđujemo normalnost preko Lillieforsove inačice Kolmogorov-Smirnov testa. U ovom slučaju, razmatramo veličinu podruma (TotalBsmtSF) kao zavisnu veličinu, a kvart (Neighborhood) kao varijablu koja određuje grupe odnosno populacije.

```
#H0: Velicina podruma i kvart su nezavisni
#H1: Velicina podruma ovisi o kvartu
data_by_neighborhood <- data %>% group_by(data$Neighborhood)
#provjera normalnosti podataka
require(nortest)
lillie.test(data_by_neighborhood$TotalBsmtSF)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF
## D = 0.075952, p-value < 2.2e-16
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Blmngtn"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "Blmngtn"]
## D = 0.18973, p-value = 0.1066
```

```
#lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Blueste"])
#U kvartu Bluestem je premalo podataka stoga je test za normalnost zakomentiran
#kako ne bi ometao provedbu ostalih testova
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="BrDale"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "BrDale"]
## D = 0.26583, p-value = 0.003593
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="BrkSide"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "BrkSide"]
## D = 0.1589, p-value = 0.0008974
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="ClearCr"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "ClearCr"]
## D = 0.14172, p-value = 0.1592
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="CollgCr"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "CollgCr"]
## D = 0.15775, p-value = 8.8e-10
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Crawfor"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "Crawfor"]
## D = 0.17186, p-value = 0.0006647
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Edwards"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "Edwards"]
## D = 0.21311, p-value = 3.906e-12
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Gilbert"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "Gilbert"]
## D = 0.16864, p-value = 8.033e-06
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="IDOTRR"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "IDOTRR"]
## D = 0.11698, p-value = 0.2264
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="MeadowV"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "MeadowV"]
## D = 0.30813, p-value = 0.0001433
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Mitchel"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "Mitchel"]
## D = 0.14861, p-value = 0.00853
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NAmes"])
```

```
##
```

```
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "NAmes"]
## D = 0.14381, p-value = 3.079e-12
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NoRidge"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "NoRidge"]
## D = 0.22641, p-value = 1.397e-05
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NPkVill"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "NPkVill"]
## D = 0.31359, p-value = 0.011
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NridgHt"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "NridgHt"]
## D = 0.071105, p-value = 0.437
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NWAmes"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "NWAmes"]
## D = 0.13355, p-value = 0.002529
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="OldTown"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "OldTown"]
## D = 0.09805, p-value = 0.009496
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="SWISU"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "SWISU"]
## D = 0.14048, p-value = 0.2296
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Sawyer"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "Sawyer"]
## D = 0.12877, p-value = 0.003949
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="SawyerW"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "SawyerW"]
## D = 0.11631, p-value = 0.04556
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Somerst"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "Somerst"]
## D = 0.10825, p-value = 0.01444
```

```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="StoneBr"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "StoneBr"]
## D = 0.097434, p-value = 0.7822
```

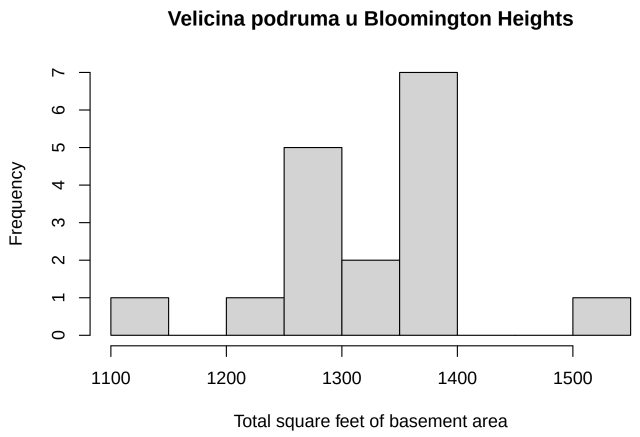
```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Timber"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "Timber"]
## D = 0.13166, p-value = 0.09458
```

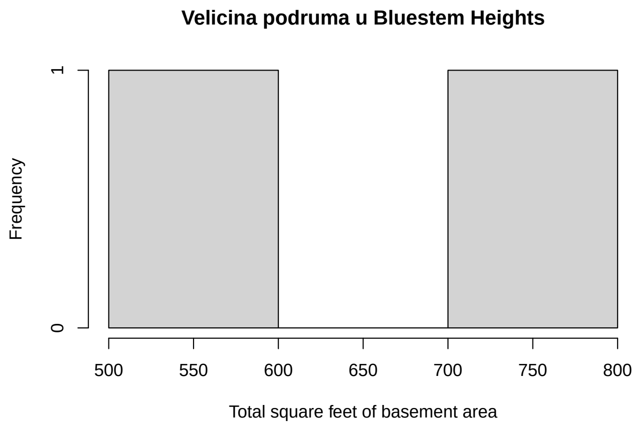
```
lillie.test(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Veenker"])
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood ==     "Veenker"]
## D = 0.2605, p-value = 0.03557
```

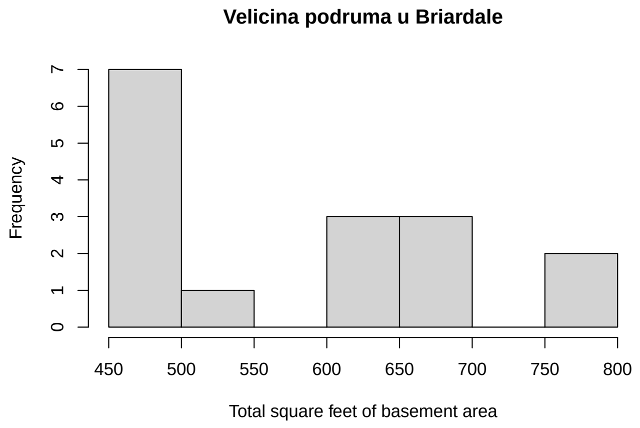
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Blmngtn"],
     main = "Velicina podruma u Bloomington Heights",
     xlab = "Total square feet of basement area")
```



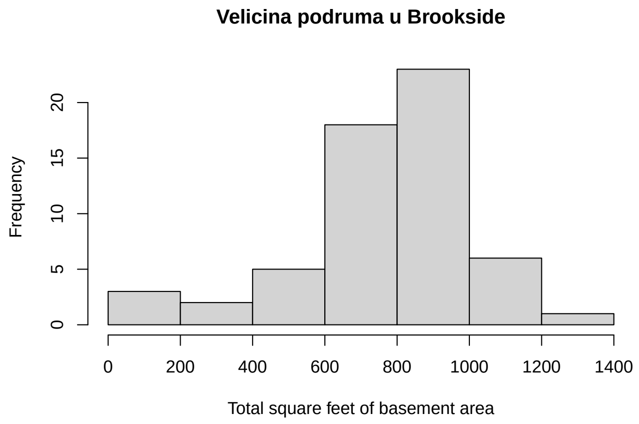
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Blueste"],
     main = "Velicina podruma u Bluestem Heights",
     xlab = "Total square feet of basement area")
```



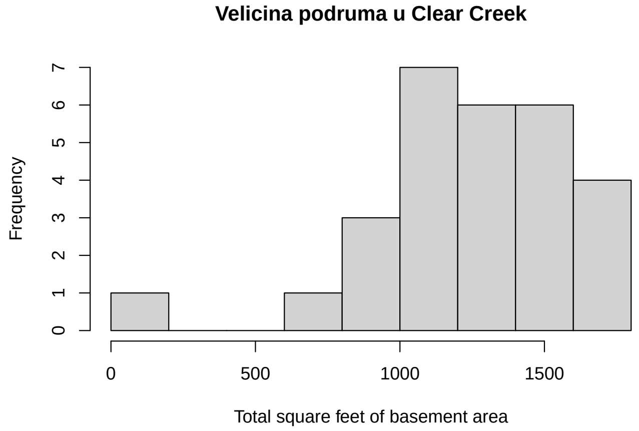
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="BrDale"],
     main = "Velicina podruma u Briardale",
     xlab = "Total square feet of basement area")
```



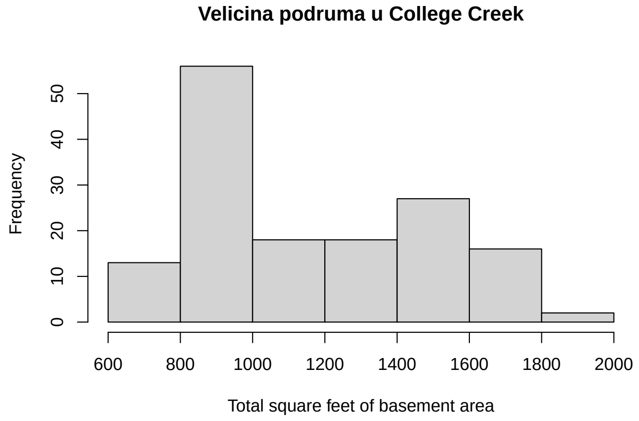
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="BrkSide"],
     main = "Velicina podruma u Brookside",
     xlab = "Total square feet of basement area")
```



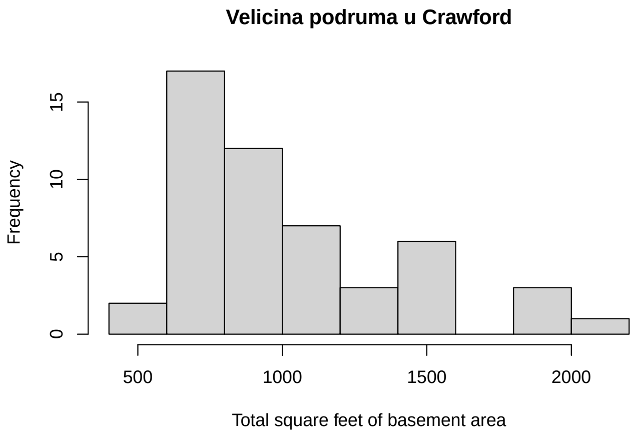
```r
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="ClearCr"],
     main = "Velicina podruma u Clear Creek",
     xlab = "Total square feet of basement area")
```



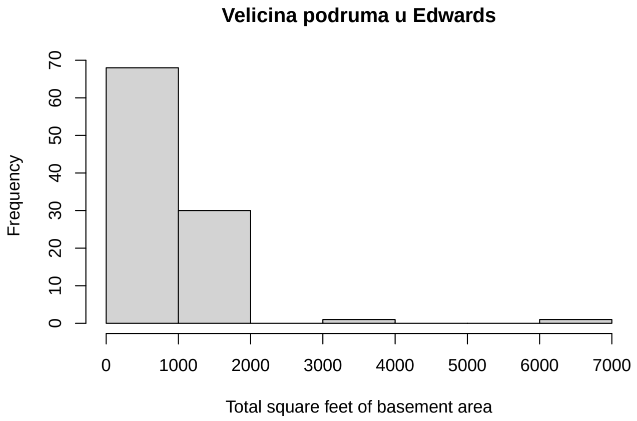
```r
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="CollgCr"],
     main = "Velicina podruma u College Creek",
     xlab = "Total square feet of basement area")
```



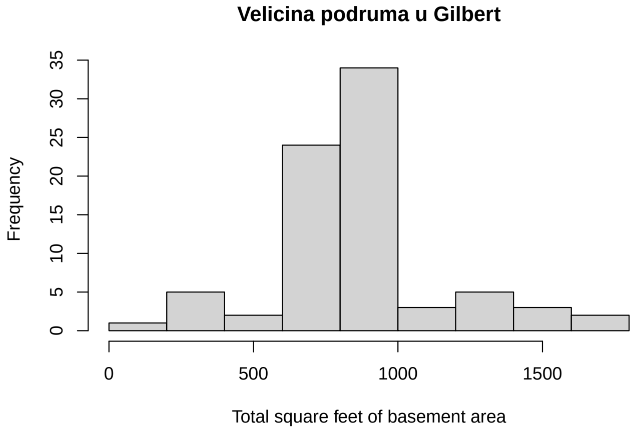
```r
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Crawfor"],
     main = "Velicina podruma u Crawford",
     xlab = "Total square feet of basement area")
```



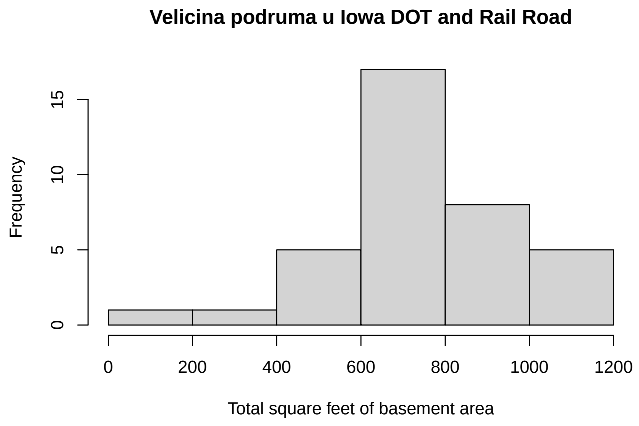
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Edwards"],
     main = "Velicina podruma u Edwards",
     xlab = "Total square feet of basement area")
```



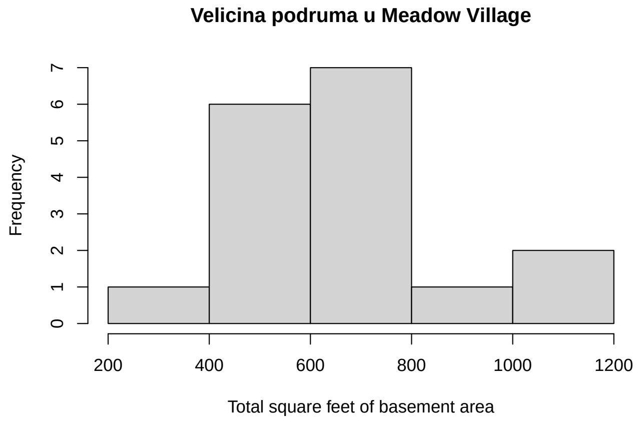
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Gilbert"],
     main = "Velicina podruma u Gilbert",
     xlab = "Total square feet of basement area")
```



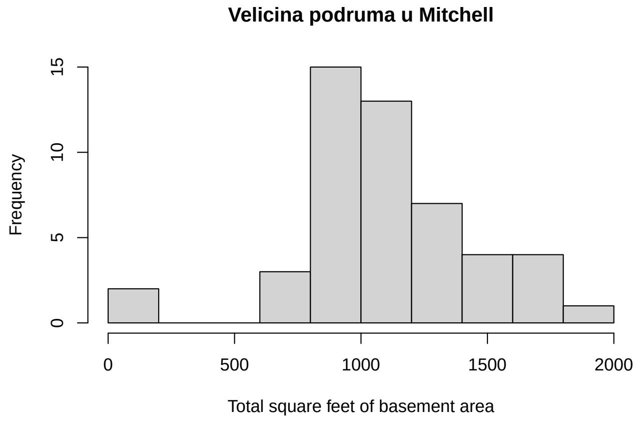
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="IDOTRR"],
     main = "Velicina podruma u Iowa DOT and Rail Road",
     xlab = "Total square feet of basement area")
```



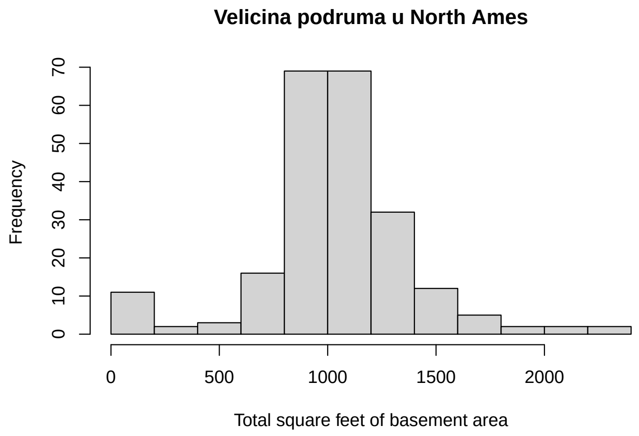
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="MeadowV"],
     main = "Velicina podruma u Meadow Village",
     xlab = "Total square feet of basement area")
```



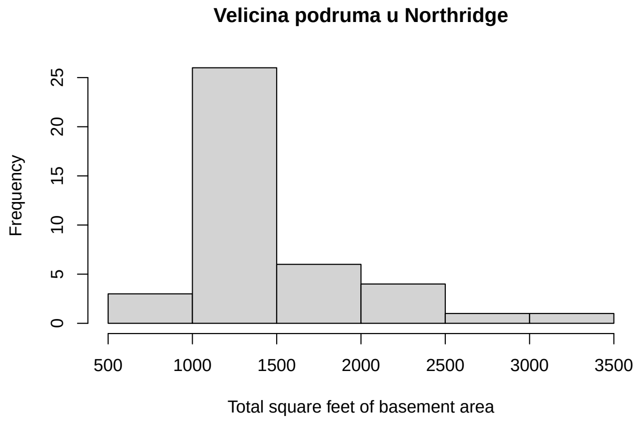
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Mitchel"],
     main = "Velicina podruma u Mitchell",
     xlab = "Total square feet of basement area")
```



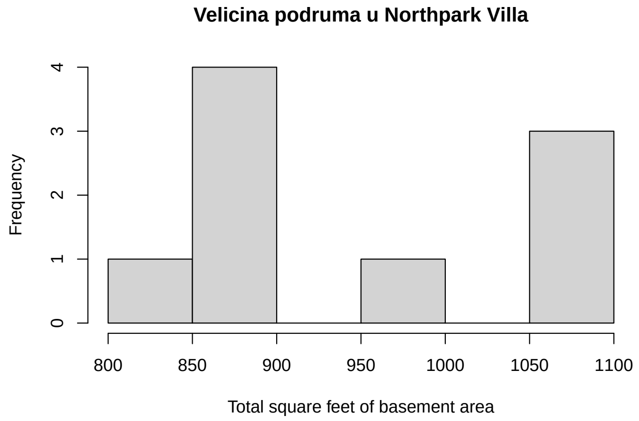
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NAmes"],
     main = "Velicina podruma u North Ames",
     xlab = "Total square feet of basement area")
```



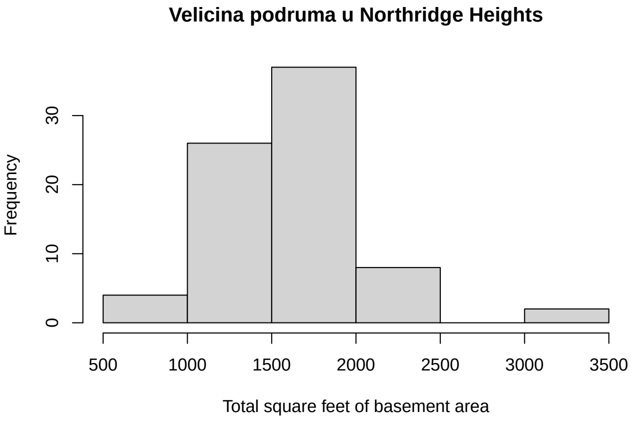
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NoRidge"],
     main = "Velicina podruma u Northridge",
     xlab = "Total square feet of basement area")
```



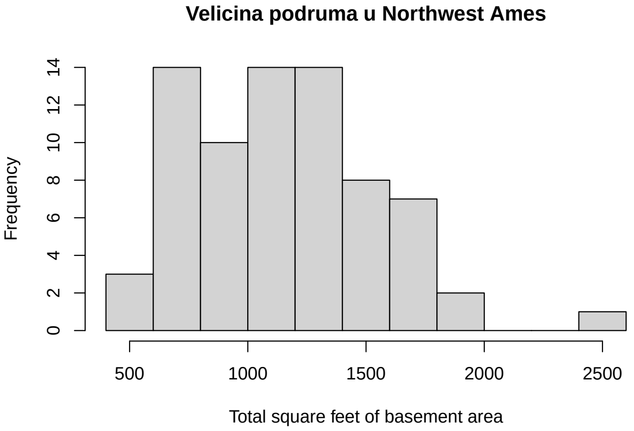
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NPkVill"],
     main = "Velicina podruma u Northpark Villa",
     xlab = "Total square feet of basement area")
```



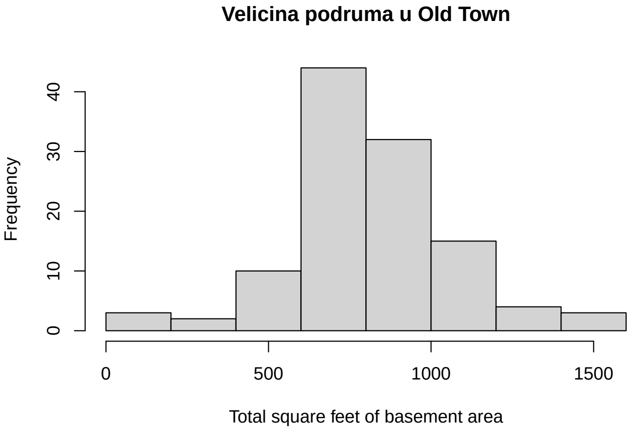
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NridgHt"],
     main = "Velicina podruma u Northridge Heights",
     xlab = "Total square feet of basement area")
```



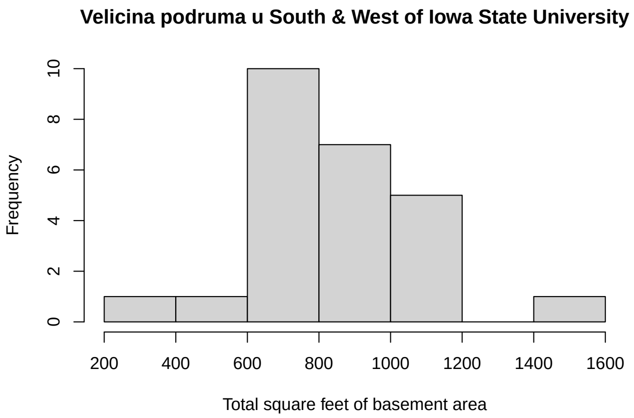
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NWAmes"],
     main = "Velicina podruma u Northwest Ames",
     xlab = "Total square feet of basement area")
```



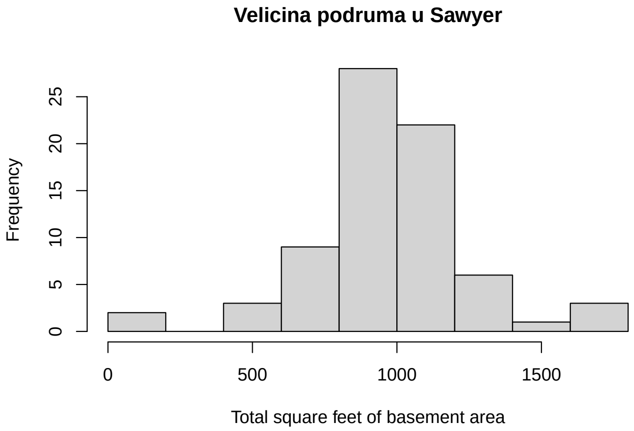
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="OldTown"],
     main = "Velicina podruma u Old Town",
     xlab = "Total square feet of basement area")
```



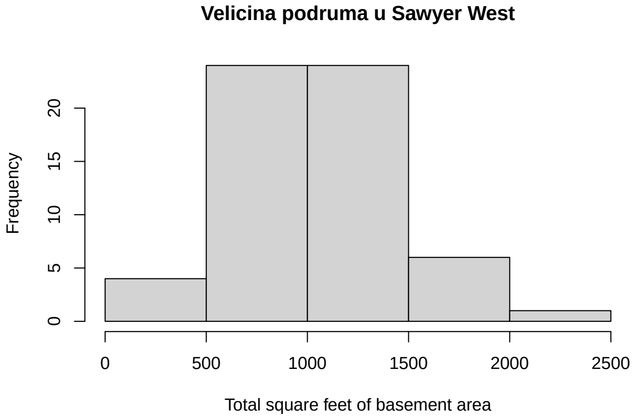
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="SWISU"],
     main = "Velicina podruma u South & West of Iowa State University",
     xlab = "Total square feet of basement area")
```



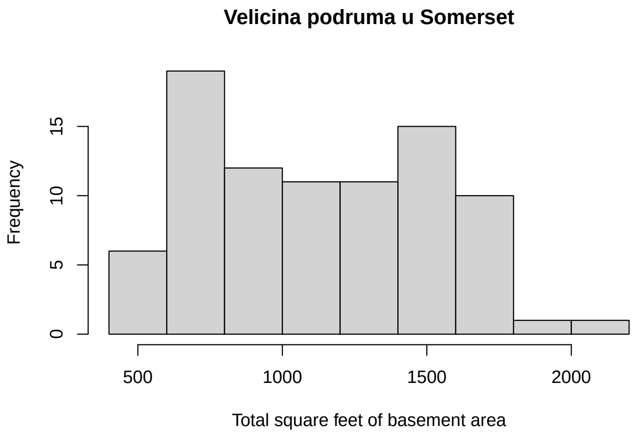
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Sawyer"],
     main = "Velicina podruma u Sawyer",
     xlab = "Total square feet of basement area")
```



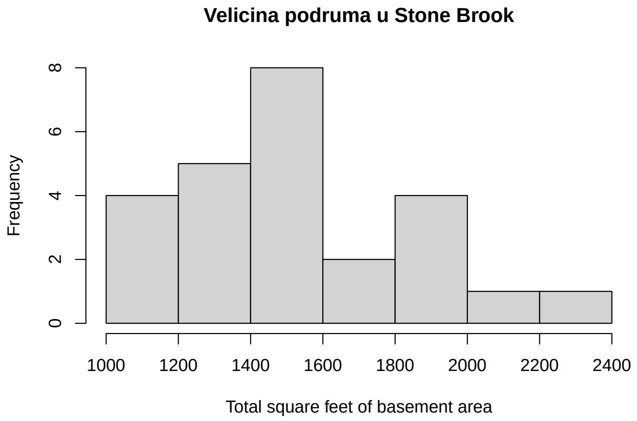
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="SawyerW"],
     main = "Velicina podruma u Sawyer West",
     xlab = "Total square feet of basement area")
```



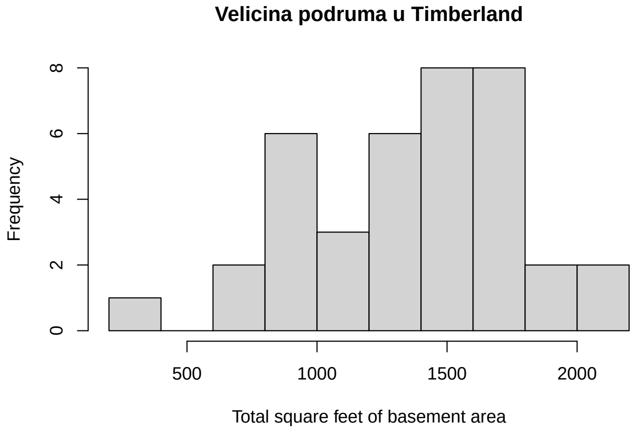
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Somerst"],
     main = "Velicina podruma u Somerset",
     xlab = "Total square feet of basement area")
```



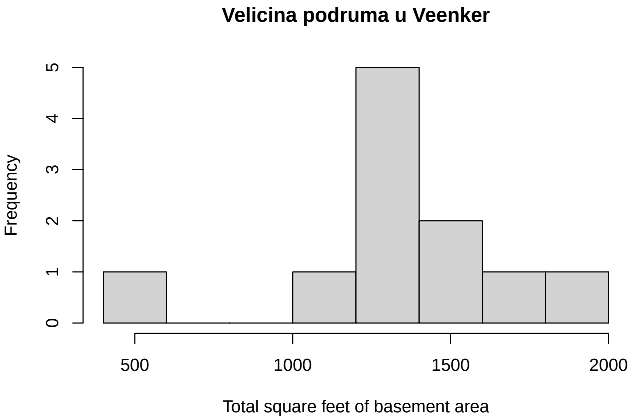
```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="StoneBr"],
     main = "Velicina podruma u Stone Brook",
     xlab = "Total square feet of basement area")
```



```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Timber"],
     main = "Velicina podruma u Timberland",
     xlab = "Total square feet of basement area")
```



```
hist(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Veenker"],
     main = "Velicina podruma u Veenker",
     xlab = "Total square feet of basement area")
```



Za kvart Bluestem ne postoji dovoljno zapisa (ima ih samo 2) da se provjeri normalnost poadataka. Vidimo da Lillieforsova inačica KS testa govori da podaci uglavnom ne dolaze iz normalne distribucije, no ukoliko pogledamo histograme vidimo (kada izuzmemo stršeće vrijednosti) distribucije nalik normalnoj. ANOVA je relativno robusna metoda na blaga odstupanja od pretpostavke normalnosti kada su veličine grupa podjednake, stoga nastavljamo dalje s provjerom pretpostavke homogenosti podataka.

Što se tiče homogenosti varijanci različitih populacija, potrebno je testirati:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$$
$$H_1 : \text{barem dvije varijance nisu iste.}$$

Navedenu hipotezu ćemo testirati Bartlettovim testom.

```
bartlett.test(data_by_neighborhood$TotalBsmtSF ~ data_by_neighborhood$Neighborhood)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  data_by_neighborhood$TotalBsmtSF by data_by_neighborhood$Neighborhood
## Bartlett's K-squared = 282.14, df = 24, p-value < 2.2e-16
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Blmngtn"])
```

```
## [1] 7044.632
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Blueste"])
```

```
## [1] 12012.5
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="BrDale"])
```

```
## [1] 11332.65
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="BrkSide"])
```

```
## [1] 71214.7
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="ClearCr"])
```

```
## [1] 133217.7
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="CollgCr"])
```

```
## [1] 102206.2
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Crawfor"])
```

```
## [1] 141672.3
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Edwards"])
```

```
## [1] 474460.2
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Gilbert"])
```

```
## [1] 84038.51
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="IDOTRR"])
```

```
## [1] 46982.47
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="MeadowV"])
```

```
## [1] 45855.37
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Mitchel"])
```

```
## [1] 137851.3
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NAmes"])
```

```
## [1] 134922
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NoRidge"])
```

```
## [1] 246306.6
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NPkVill"])
```

```
## [1] 12452.36
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NridgHt"])
```

```
## [1] 193546.7
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="NWAmes"])
```

```
## [1] 142499.5
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="OldTown"])
```

```
## [1] 66447.66
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="SWISU"])
```

```
## [1] 51548.33
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Sawyer"])
```

```
## [1] 91153.92
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="SawyerW"])
```

```
## [1] 176793.2
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Somerst"])
```

```
## [1] 155278.7
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="StoneBr"])
```

```
## [1] 101834
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Timber"])
```

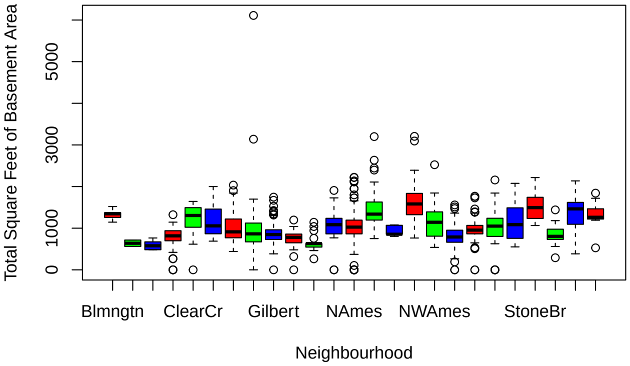
```
## [1] 158768.3
```

```
var(data_by_neighborhood$TotalBsmtSF[data_by_neighborhood$Neighborhood=="Veenker"])
```

```
## [1] 114813.2
```

Provjeravamo zatim postoje li razlike u veličinama podruma za različite kvartove u Amesu.

```
boxplot(data$TotalBsmtSF[data$Neighborhood != "<undefined>"]
        ~ data$Neighborhood[data$Neighborhood != "<undefined>"],
        ylab= "Total Square Feet of Basement Area",
        xlab= "Neighbourhood",
        col=rainbow(3))
```



Grafički prikaz pokazuje kako postoji jasna razlika između kvartova u veličinama podruma. Provedimo sada ANOVA test kako bismo potvrdili tu razliku:

```
res.aov <- aov(data_by_neighborhood$TotalBsmtSF ~ data_by_neighborhood$Neighborhood)
summary(res.aov)
```

```
##                                     Df    Sum Sq Mean Sq F value Pr(>F)
## data_by_neighborhood$Neighborhood   24  74574228 3107259   21.62 <2e-16 ***
## Residuals                         1435 206228358  143713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
kruskal.test(data_by_neighborhood$TotalBsmtSF ~ data_by_neighborhood$Neighborhood)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  data_by_neighborhood$TotalBsmtSF by data_by_neighborhood$Neighborhood
## Kruskal-Wallis chi-squared = 467.6, df = 24, p-value < 2.2e-16
```

P- vrijednost prilikom provođenja ANOVA testa (i Kruskal-Wallis testa kao dodatna potvrda koja ne pretpostavlja homogenost i normalnost) manja je od 0.01, stoga na razini značajnosti 1% odbacujemo hipotezu da su veličina podruma i kvart u gradu nezavisni.

Predviđanje cijene nekretnine

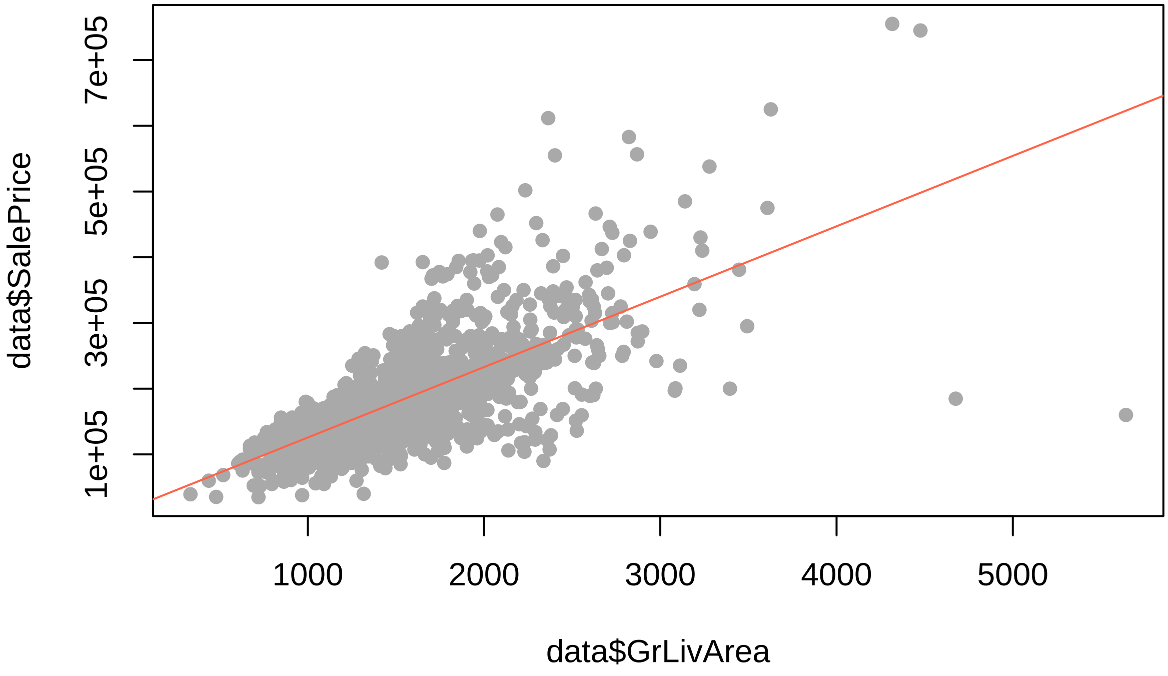
Kako bi znali predvidjeti cijenu nekretnine, možemo ispitati različite varijable koje bi mogle utjecati na cijenu:

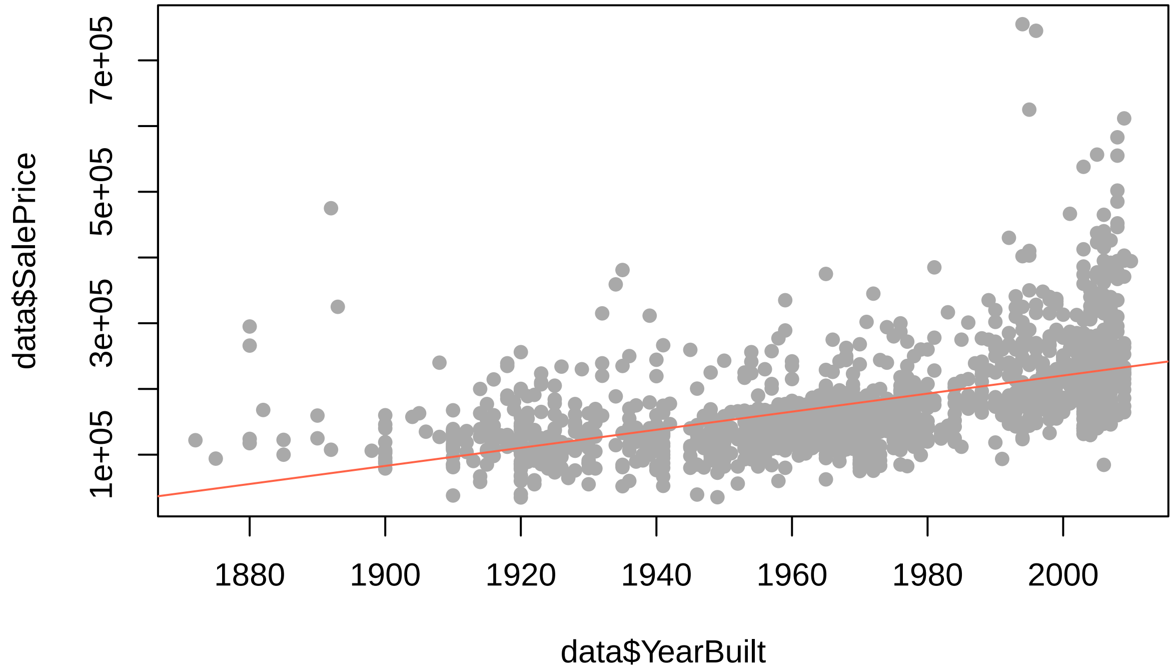
- veličina nekretnine
- godina izgradnje
- broj soba

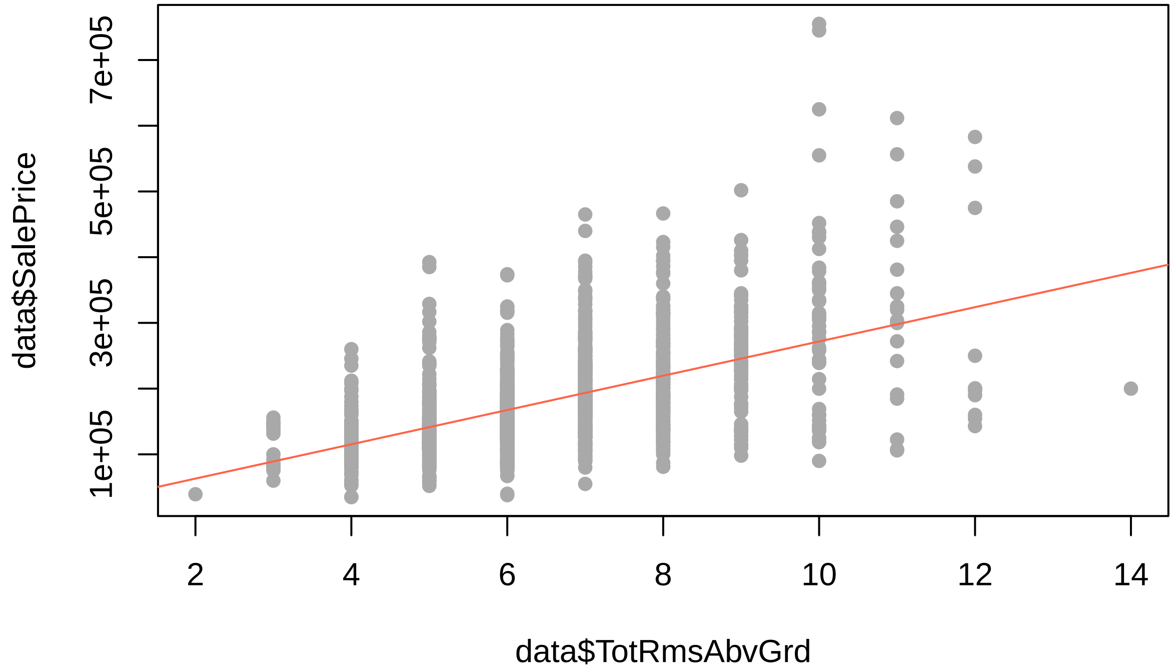
Ovisnost cijene o kvadraturi nekretnine

Kad promatramo utjecaj samo jedne nezavisne varijable X na neku zavisnu varijablu Y, grafički je moguće dobiti jako dobar dojam o njihovom odnosu - tu je najčešće od pomoći scatter plot.

```
linear_graph <- function(){
  plot(data$GrLivArea,data$SalePrice, col="darkgrey", pch = 16) #prosjecna kvadratura vs cijena
  model <- lm(SalePrice~GrLivArea,data=data)
  abline(model, col="tomato")
  plot(data$YearBuilt,data$SalePrice, col="darkgrey", pch = 16) #prosjecni godina izgradnje vs cijena
  model <- lm(SalePrice~YearBuilt,data=data)
  abline(model, col="tomato")
  plot(data$TotRmsAbvGrd,data$SalePrice, col="darkgrey", pch = 16) #prosjecna broj soba vs cijena
  model <- lm(SalePrice~TotRmsAbvGrd,data=data)
  abline(model, col="tomato")
}
linear_graph()
```







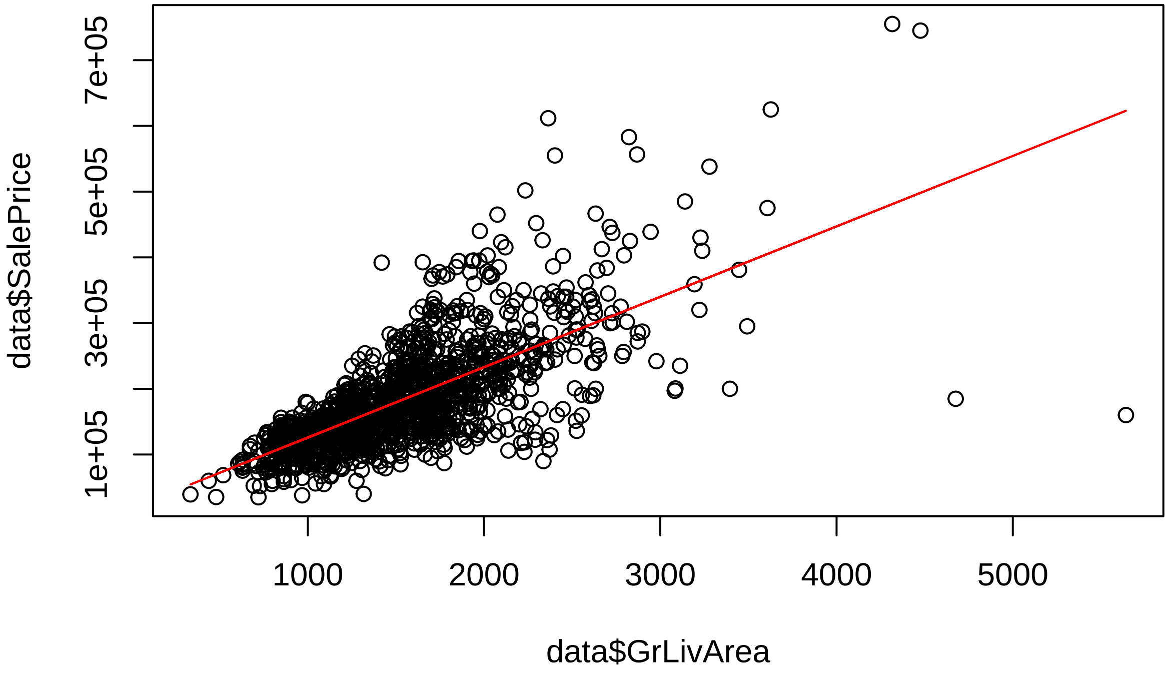
Očito je da kvadratura (i godina izgradnje i broj soba) ima izražen (i to pozitivan) utjecaj na izlaznu varijablu.

Kako bi ispitali pojedinačni utjecaj ovih varijabli, procijenit ćemo model jednostavne regresije - po jedan za svaku nezavisnu varijablu (uz cnt - broj iznajmljenih bicikala - kao zavisnu varijablu).

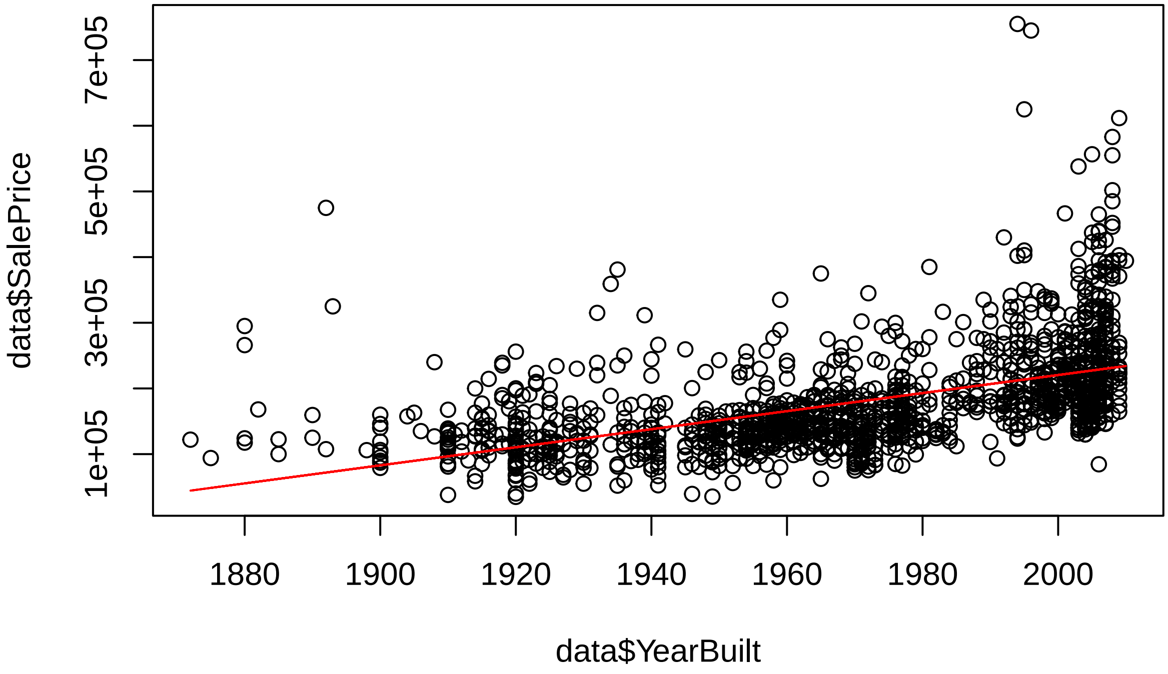
Regresijski model procjenjuje se funkcijom lm() koja kao parametre prima zavisne i nezavisne varijable, odnosno data.frame sa svim varijablama i definiciju varijabli u modelu.

```
#linearni model cijene nekretnina (SalePrice) i kvadratura nekretnine (GrLivArea)
fit.kvadratura = lm(data$SalePrice~data$GrLivArea,data=data)
#linearni model cijene nekretnina (SalePrice) i godinu izgradnje nekretnine (YearBuilt)
fit.godinaIzgradnje = lm(data$SalePrice~data$YearBuilt,data=data)
#linearni model cijene nekretnina (SalePrice) i broj soba u nekretnin (TotRmsAbvGrd)
```

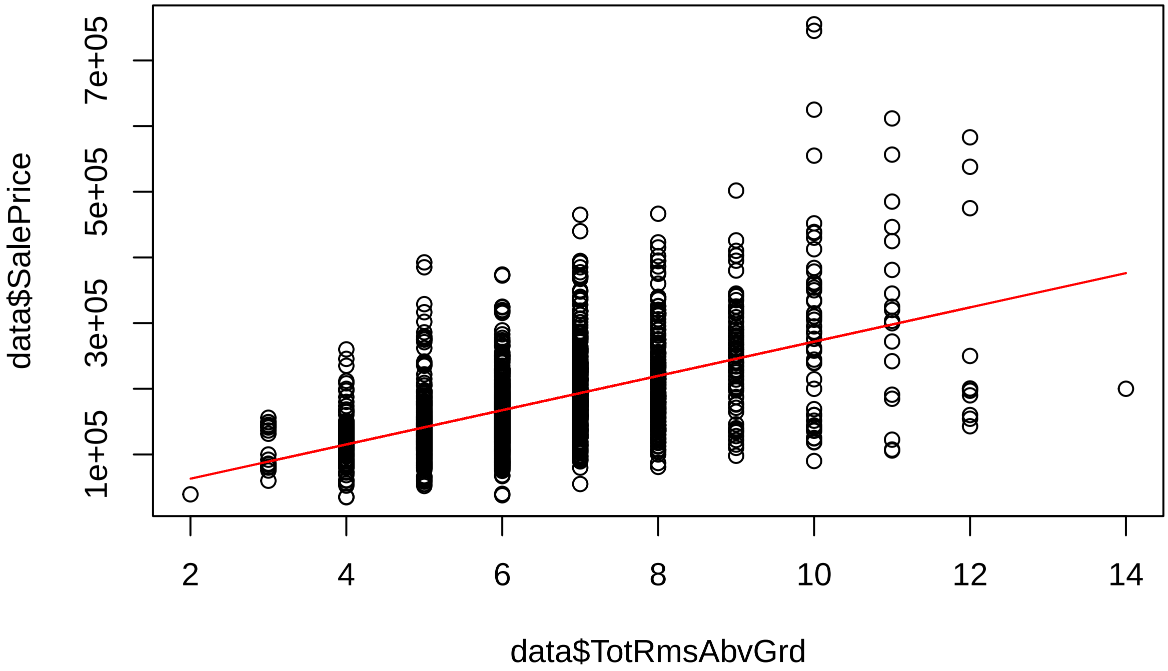
```r
fit.brojSoba = lm(data$SalePrice~data$TotRmsAbvGrd,data=data)
#graficki prikaz podataka
plot(data$GrLivArea,data$SalePrice)
#graficki prikaz procijenjenih vrijednosti iz modela
lines(data$GrLivArea,fit.kvadratura$fitted.values,col='red')
```



```r
#graficki prikaz podataka
plot(data$YearBuilt,data$SalePrice)
#graficki prikaz procijenjenih vrijednosti iz modela
lines(data$YearBuilt,fit.godinaIzgradnje$fitted.values,col='red')
```



```
#graficki prikaz podataka
plot(data$TotRmsAbvGrd,data$SalePrice)
#graficki prikaz procijenjenih vrijednosti iz modela
lines(data$TotRmsAbvGrd,fit.brojSoba$fitted.values,col='red')
```

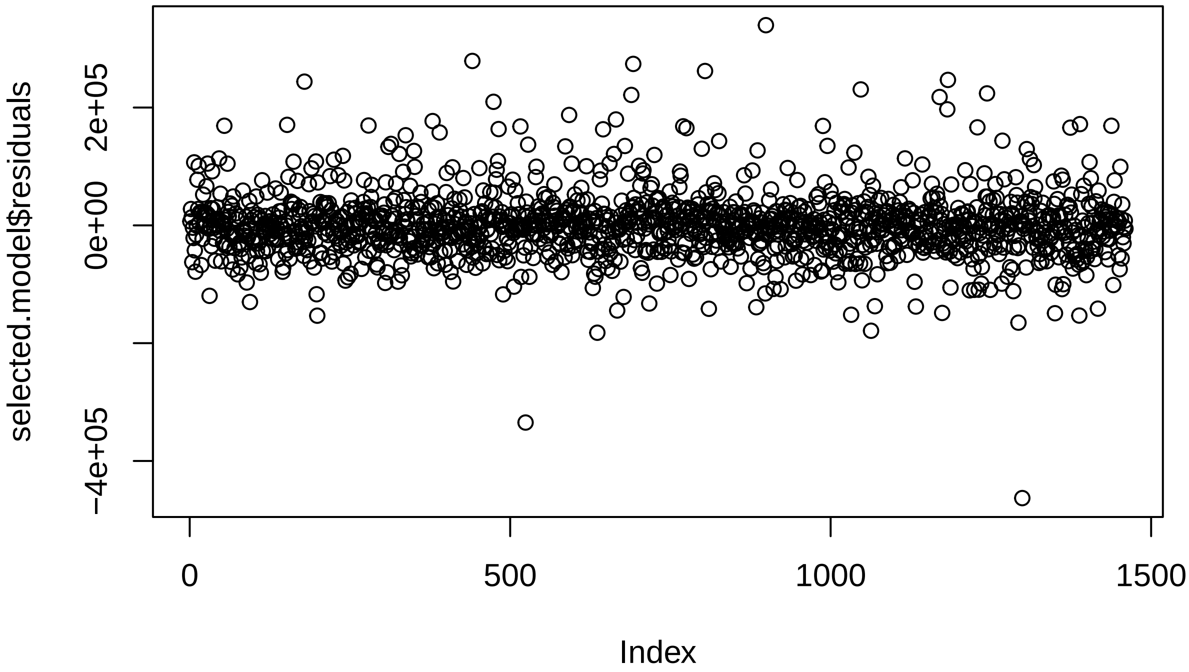


Nagibi pravaca linearne regresije potvrđuju tvrdnje o efektima pojedinih razmatranih varijabli na izlaznu varijablu. Kako bi se dobiveni modeli analizirali i usporedili, prvo je potrebno provjeriti da pretpostavke modela nisu (jako) narušene. Pritom su najbitnije pretpostavke o regresorima (u multivarijatnoj regresiji regresori ne smiju biti međusobno jako korelirani) i o rezidualima (normalnost reziduala i homogenost varijance).

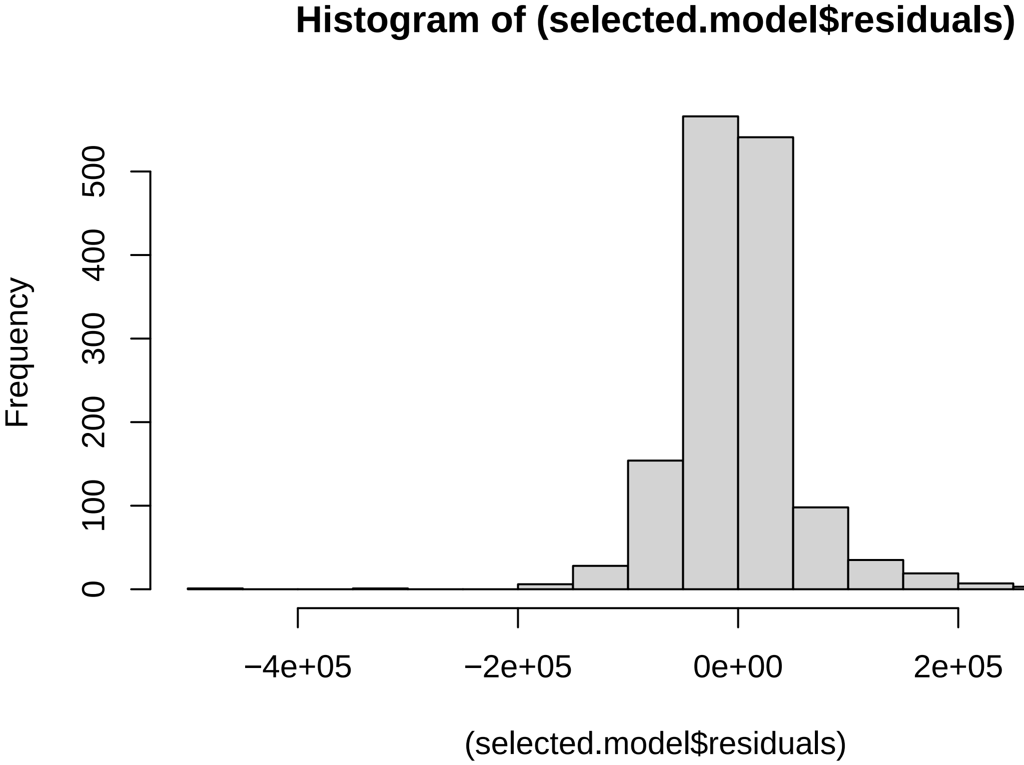
Normalnost reziduala i homogenost varijance

Normalnost reziduala moguće je provjeriti grafički, pomoću kvantil-kvantil plota (usporedbom s linijom normalne razdiobe), te statistički pomoću Anderson-Darlingovog testa.

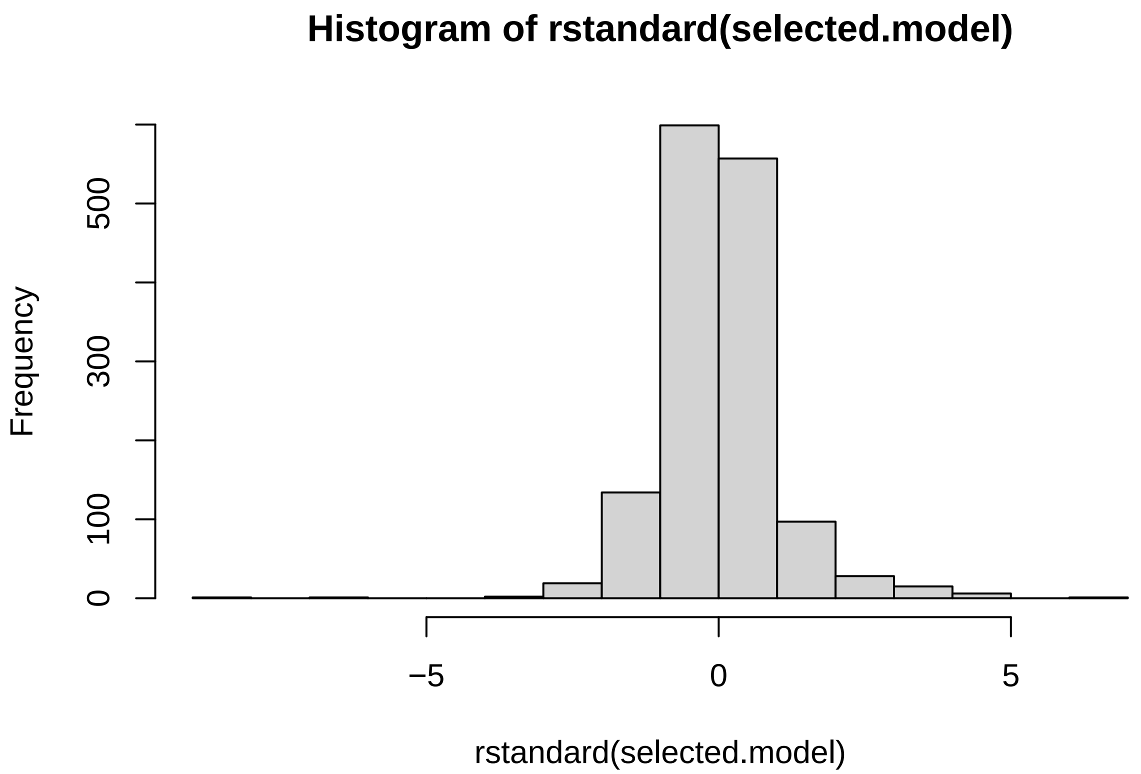
```
selected.model = fit.kvadratura
plot(selected.model$residuals) #gledajuci reziduale na ovaj nacin tesko je suditi o normalnosti
```



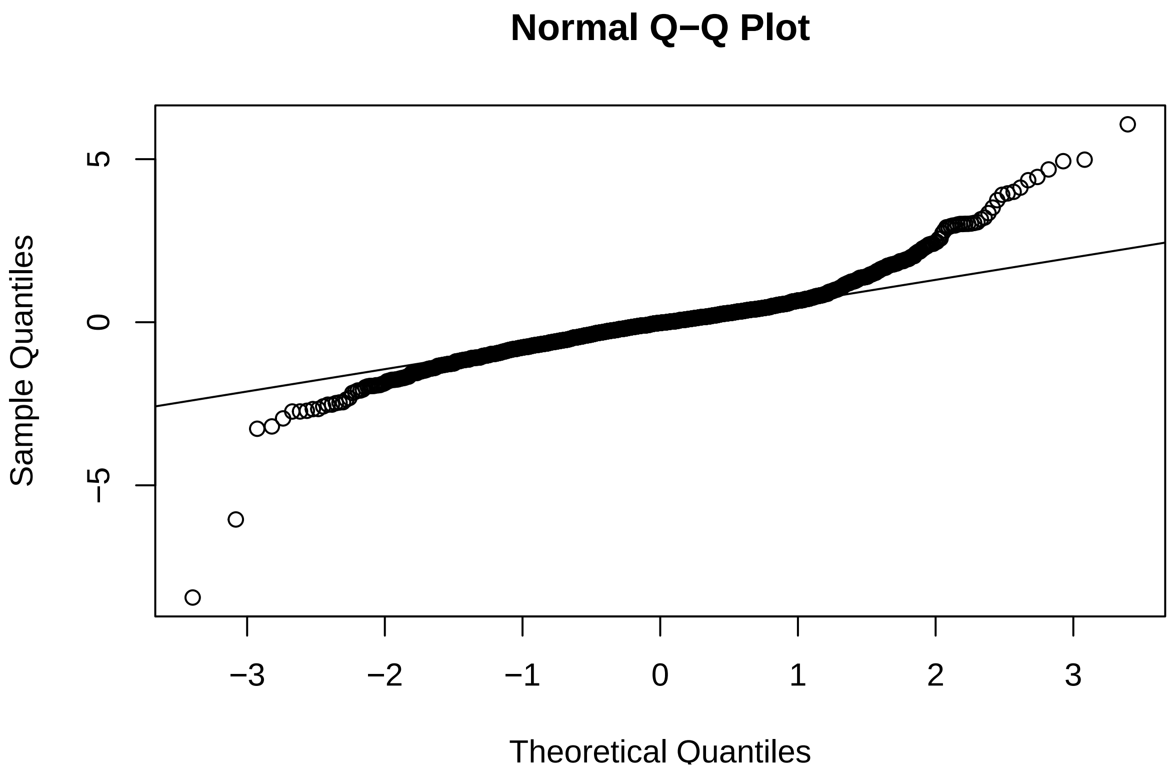
```
#histogram je vrlo interpretativan
hist((selected.model$residuals))
```



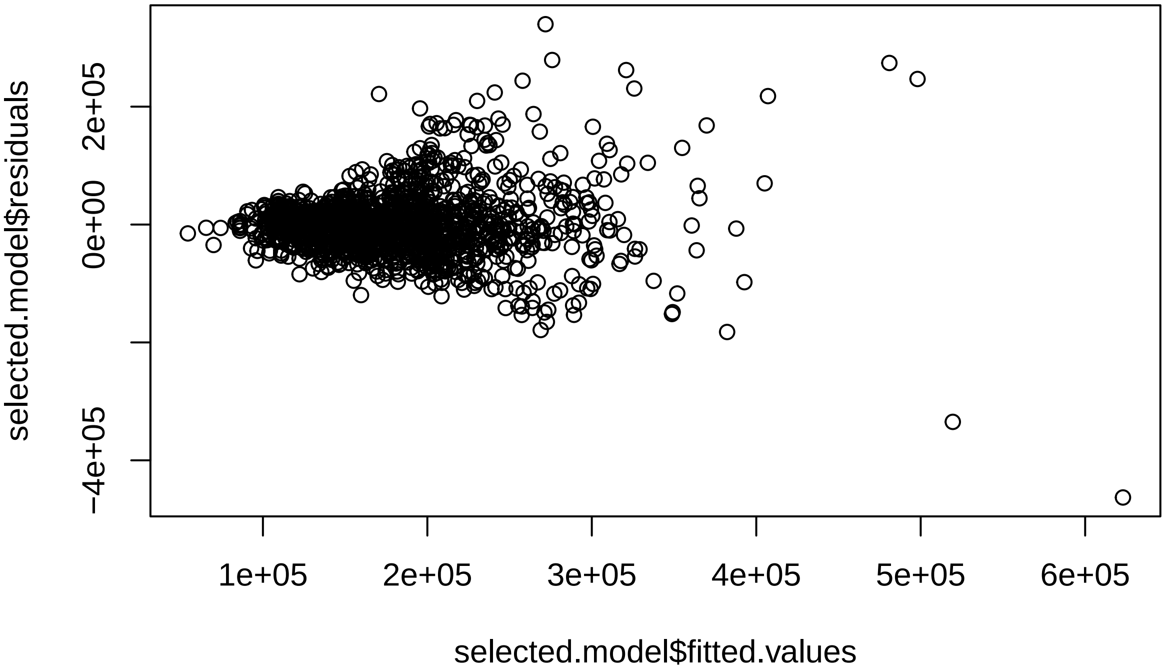
```
hist(rstandard(selected.model))
```



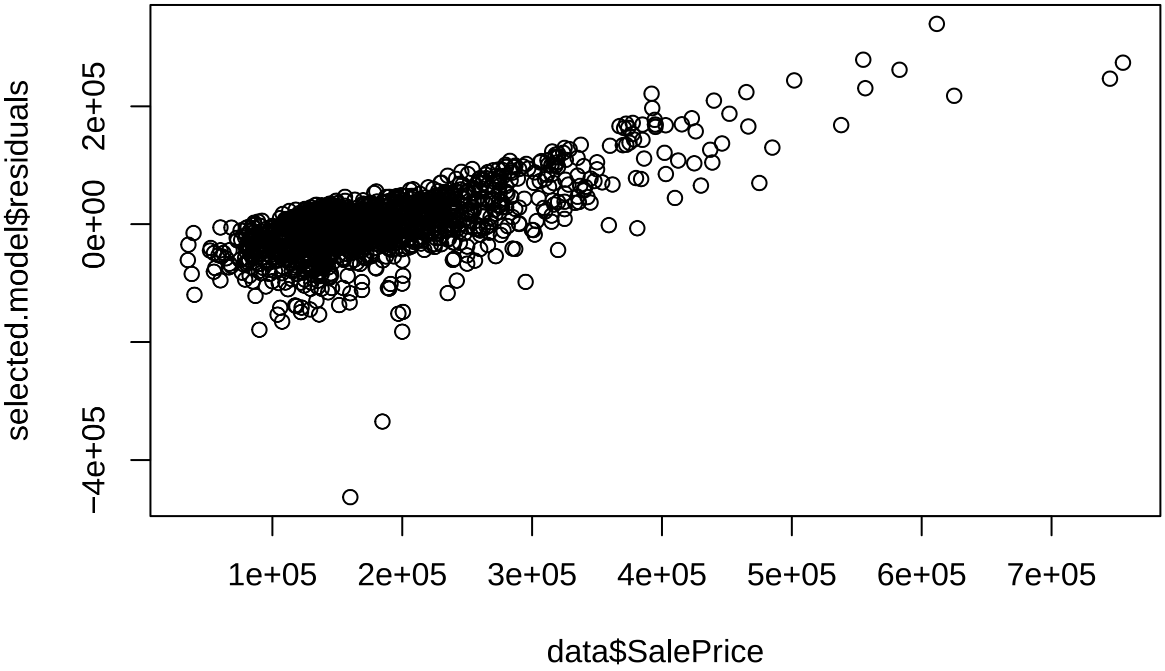
```
#q-q plot reziduala s linijom normalne distribucije
qqnorm(rstandard(selected.model))
qqline(rstandard(selected.model))
```



```
plot(selected.model$fitted.values,selected.model$residuals)
```



```
#reziduale je dobro prikazati u ovisnosti o procjenama modela
plot(data$SalePrice,selected.model$residuals)
```



```
#a ponekad i u ovisnosti o nekim drugim varijablama koje je mozda
#tesko modelirati kao nezavisne varijable s linearnim efektom na izlaz
#AD test na normalnost
ad.test(rstandard(fit.kvadratura))
```

```
##
##  Anderson-Darling normality test
```

```
##
## data:  rstandard(fit.kvadratura)
## A = 27.834, p-value < 2.2e-16
```

```
#KS test should only contain unique values
ks.test(rstandard(fit.kvadratura),'pnorm')
```

```
## Warning in ks.test.default(rstandard(fit.kvadratura), "pnorm"): ties should not
## be present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  rstandard(fit.kvadratura)
## D = 0.1084, p-value = 2.554e-15
## alternative hypothesis: two-sided
```

```
require(nortest)
lillie.test(rstandard(fit.kvadratura))
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  rstandard(fit.kvadratura)
## D = 0.10873, p-value < 2.2e-16
```

Grafički prikaz reziduala samo po indeksu po kojem su dani u podatcima rijetko kad može dati potpunu sliku o njihovoj prirodi - doduše, u ovom slučaju su podatci poredani po cijeni - koji svjedoči o određenoj zavisnosti podataka.

Histogram je vrlo lako čitljiv i interpretativan način prikazivanja ovakvih varijabli, te se lako može zaključiti nešto o općenitom obliku distribucije reziduala - u ovom slučaju, ta distribucija donekle nalikuje normalnoj (što otprilike pokazuje i q-q plot), te nije previše zakrivljena.

Također je jako bitno da u ovisnosti o predviđanjima modela sami reziduali ne pokazuju heterogenost varijance (ne "šire" se s povećanjem $\hat{y}$). No, u ovisnosti o cijeni postoji određena dinamika reziduala (ne "izgledaju" potpuno slučajno) koju model ne objašnjava. Takve vremenske zavisnosti se najčešće modeliraju tzv. autoregresivnim modelima (ARMA, ARIMA, ARIMAX, itd.).

```
summary(fit.kvadratura)
```

```
##
## Call:
## lm(formula = data$SalePrice ~ data$GrLivArea, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -462999  -29800   -1124   21957  339832
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    18569.026   4480.755   4.144 3.61e-05 ***
## data$GrLivArea   107.130      2.794  38.348  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56070 on 1458 degrees of freedom
## Multiple R-squared:  0.5021, Adjusted R-squared:  0.5018
## F-statistic:  1471 on 1 and 1458 DF,  p-value: < 2.2e-16
```

```
summary(fit.godinaIzgradnje)
```

```
## 
## Call:
## lm(formula = data$SalePrice ~ data$YearBuilt, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -144191  -40999  -15464   22685  542814 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    -2.530e+06  1.158e+05  -21.86   <2e-16 ***
## data$YearBuilt  1.375e+03  5.872e+01   23.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 67740 on 1458 degrees of freedom
## Multiple R-squared:  0.2734, Adjusted R-squared:  0.2729 
## F-statistic: 548.7 on 1 and 1458 DF,  p-value: < 2.2e-16
```

```
summary(fit.brojSoba)
```

```
## 
## Call:
## lm(formula = data$SalePrice ~ data$TotRmsAbvGrd, data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -191844  -35500   -8620   28543  483242 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          10896       7271   1.499    0.134    
## data$TotRmsAbvGrd    26086       1082  24.099   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 67200 on 1458 degrees of freedom
## Multiple R-squared:  0.2849, Adjusted R-squared:  0.2844 
## F-statistic: 580.8 on 1 and 1458 DF,  p-value: < 2.2e-16
```

Zaključak

U gradu Amesu, mnogo varijabli vezanih uz prodane nekretnine u periodu od 2006. do 2010. godine su međusobno zavisne. Tako smo u ovom radu pokazali kako je broj katova nekretnine određen oblikom zemljišne čestice na kojoj se nalazi. Tu zavisnost dokazali smo χ^2 testom. Zatim smo korištenjem ANOVA testa saznali kako veličina podruma nekretnine ovisi o kvartu u kojem se nalazi, te kako cijena kvadrata nekretnine ovisi o broju spavaćih soba u njoj. Za kraj, linearnom regresijom pokazali smo kako se cijene nekretnine može predvidjeti ako se pogledaju značajke kvadrature, godine izgradnje te broja spavaćih soba.