Projekt - Analiza tržišta nekretnina

Grupa FerovkeiFerovac

Marko Dodik, Silvija Gojević, Lucija Mičić, Antonia Žaja

19.12.2022.

Opis projekta:

Ovaj projekt obavezni dio je 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.

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

Skup podataka:

Skup podataka koji se koristi u ovom projektu predstavljaju 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.

```
#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"
```

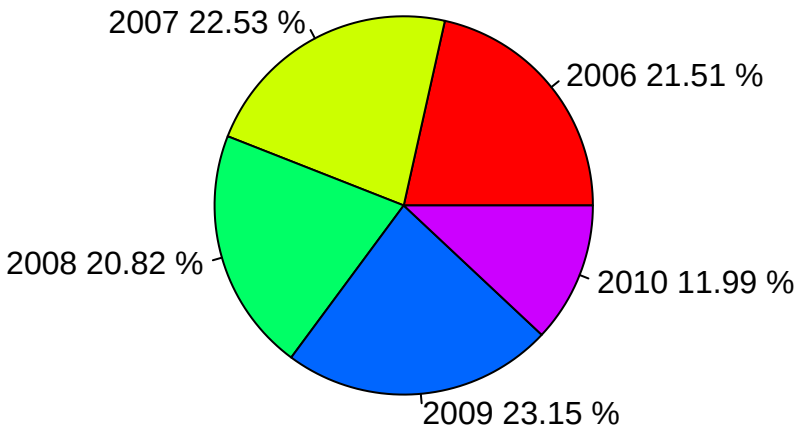
```
#ukupni broj zapisa
nrow(data)
```

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

Deskriptivna statistika skupa podataka:

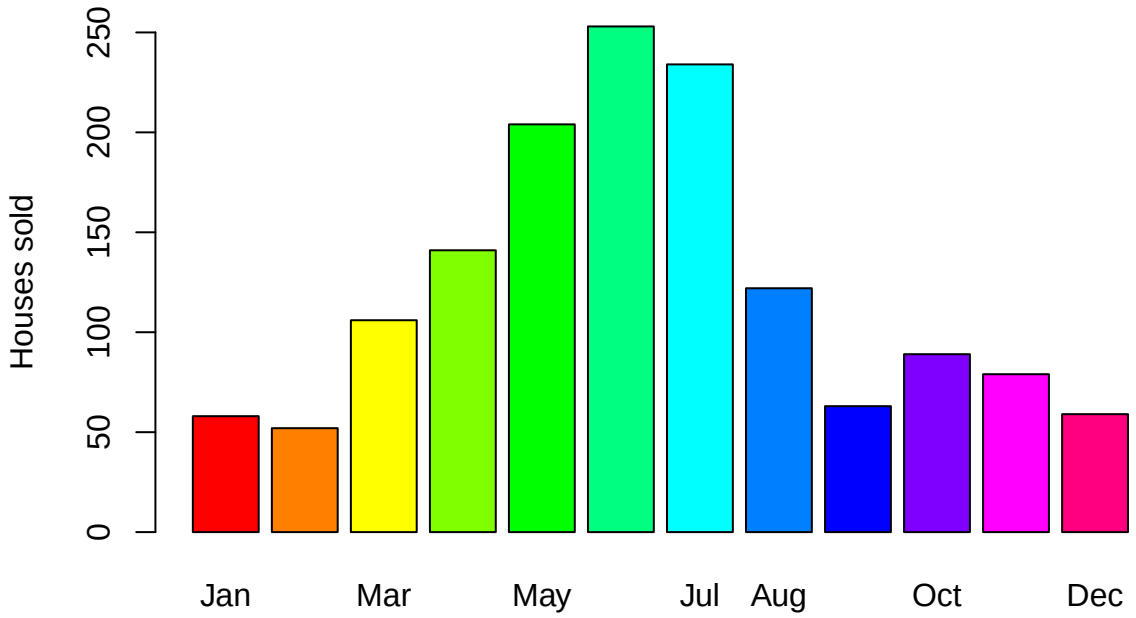
Proučavamo tržište nekretnina u godinama od periodu od 2006. do 2010. godine (uključivo). Na sljedećem dijagramu, prikazani su udjeli broja prodanih nekretnina po godinama. Iz njega vidimo kako je 2009. godine prodan najveći broj nekretnina (23.15%). Za prikaz ovih podataka je odabran strukturni krug.

```
#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)))
```



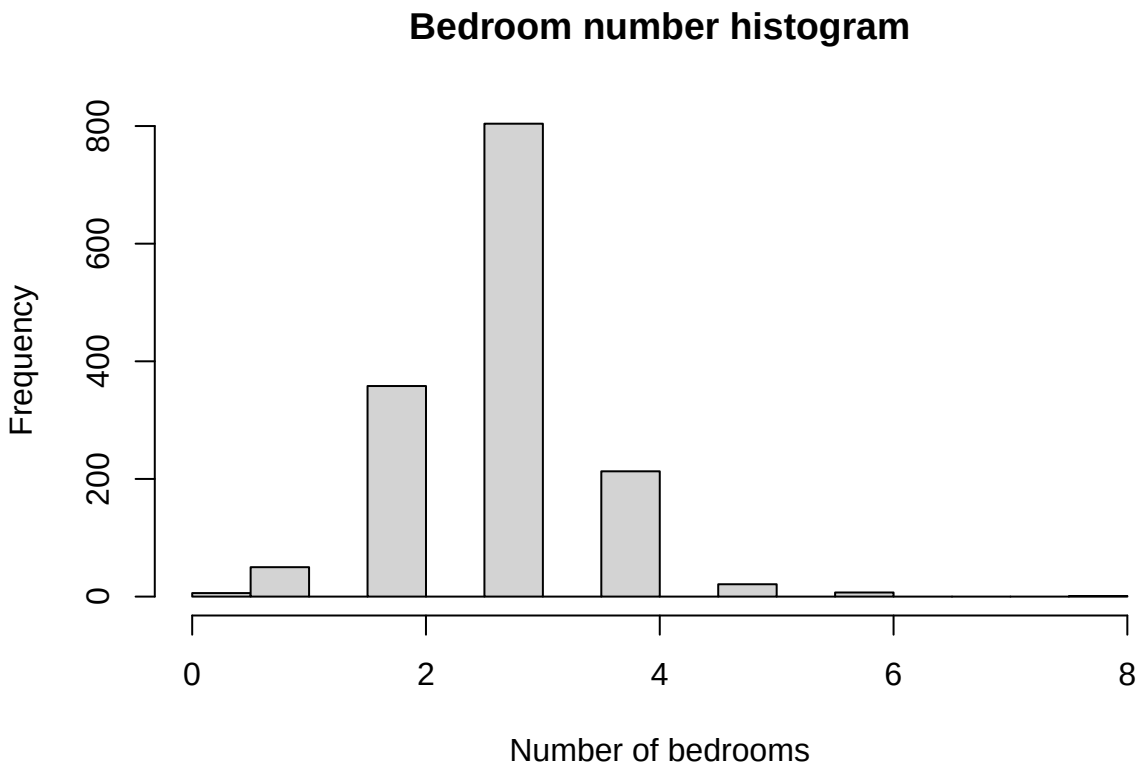
Zanimljiva informacija koja se može saznati iz danih podataka je u kojim mjesecima se tijekom godina najviše nekretnina prodalo. Iz sljedećeg stupičastog dijagrama vidimo kako je to u kasnim proljetnim i ljetnim mjesecima.

```
#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)
```



```
#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.

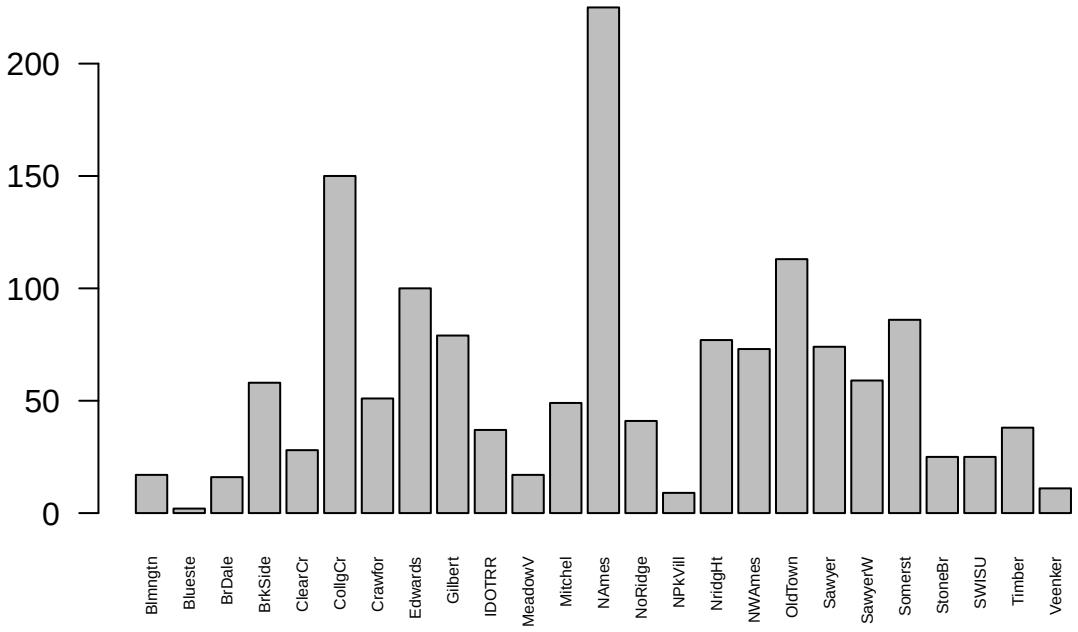
```
#grupiranje nekretnina pomoću značajke YrSold
houses_sold_2006 = data[data$YrSold == "2006",]
houses_sold_2007 = data[data$YrSold == "2007",]
houses_sold_2008 = data[data$YrSold == "2008",]
houses_sold_2009 = data[data$YrSold == "2009",]
houses_sold_2010 = data[data$YrSold == "2010",]

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)))
```

```
#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

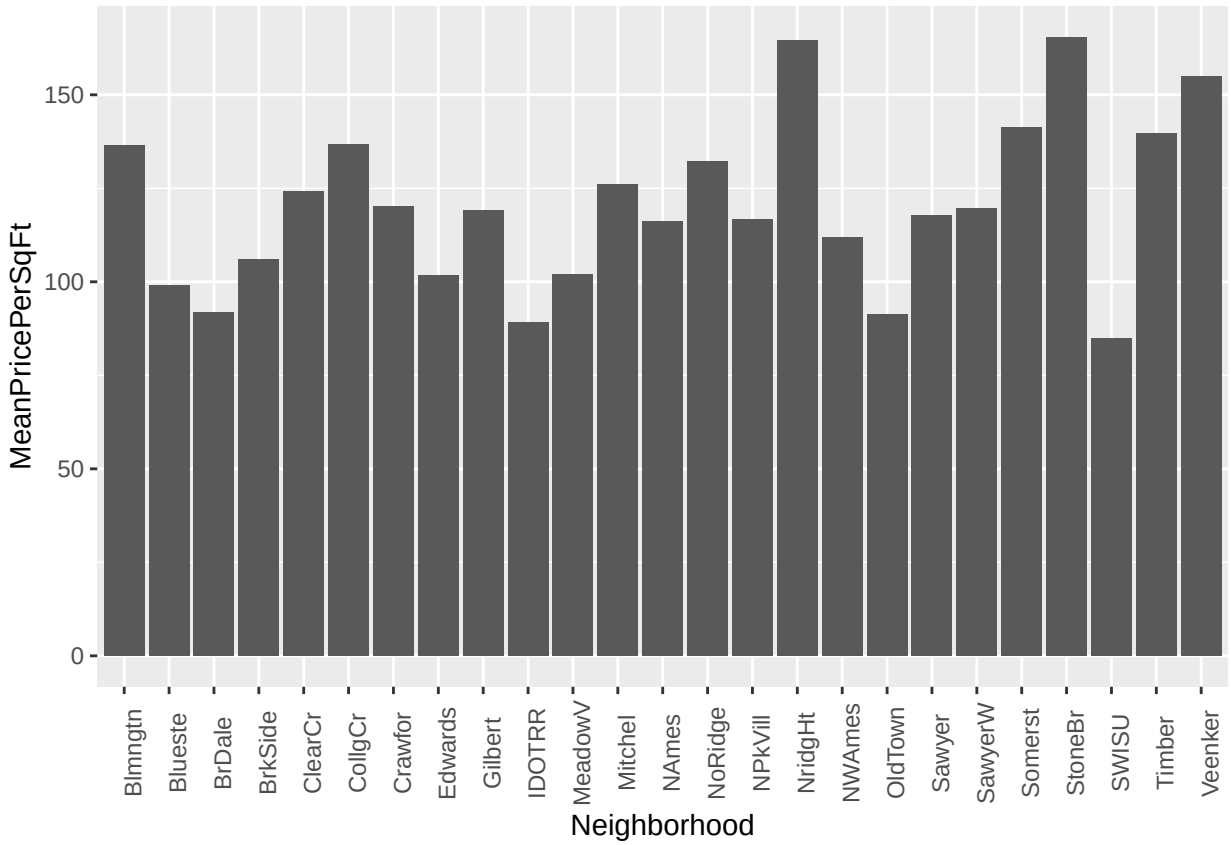
Sljedeći dijagram prikazuje distribuciju prodanih nekretnina u ovisnosi o kvartu u kojem se nalaze. Iz dijagrama je vidljivo da se najveći broj prodanih nekretnina nalazi u kvartu North Ames. Za prikaz podataka korišten je stupičasti dijagram.

```
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. Testom utvrđujemo p-vrijednost, koja je manja od 0.05, stoga zaključujemo da broj katova nekretnine ne ovisi o obliku čestice.

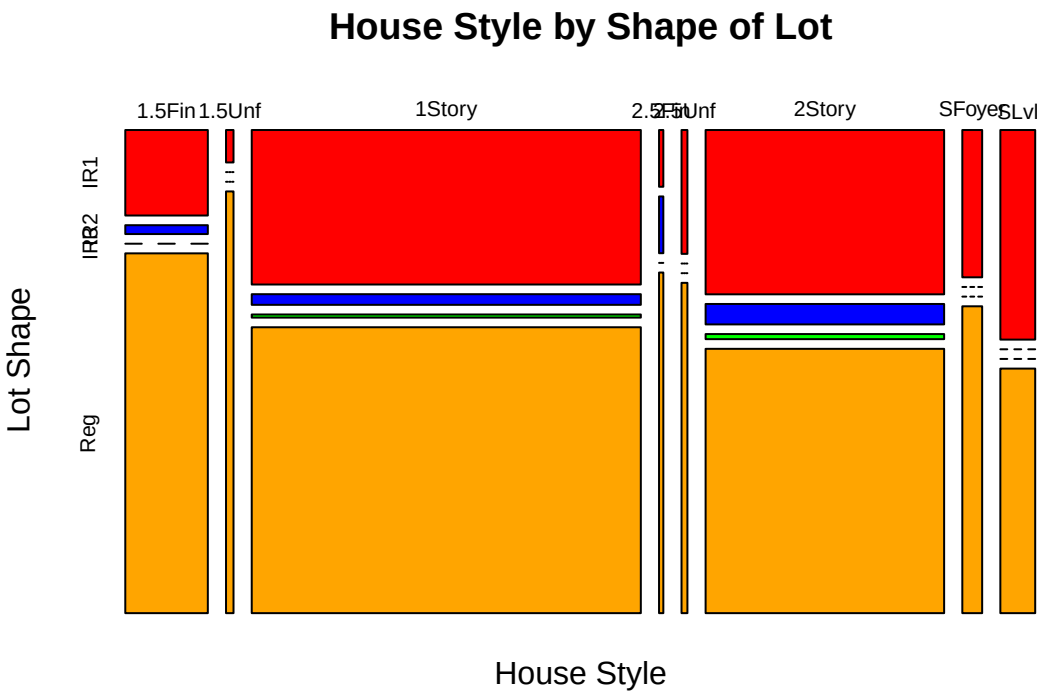
```
tableHSLS <- table(data$HouseStyle, data$LotShape)
chi_squared_test <- chisq.test(tableHSLS)
```

```
## Warning in chisq.test(tableHSLS): Chi-squared approximation may be incorrect
```

```
chi_squared_test
```

```
##
##  Pearson's Chi-squared test
##
## data:  tableHSLS
## X-squared = 44.472, df = 21, p-value = 0.002031
```

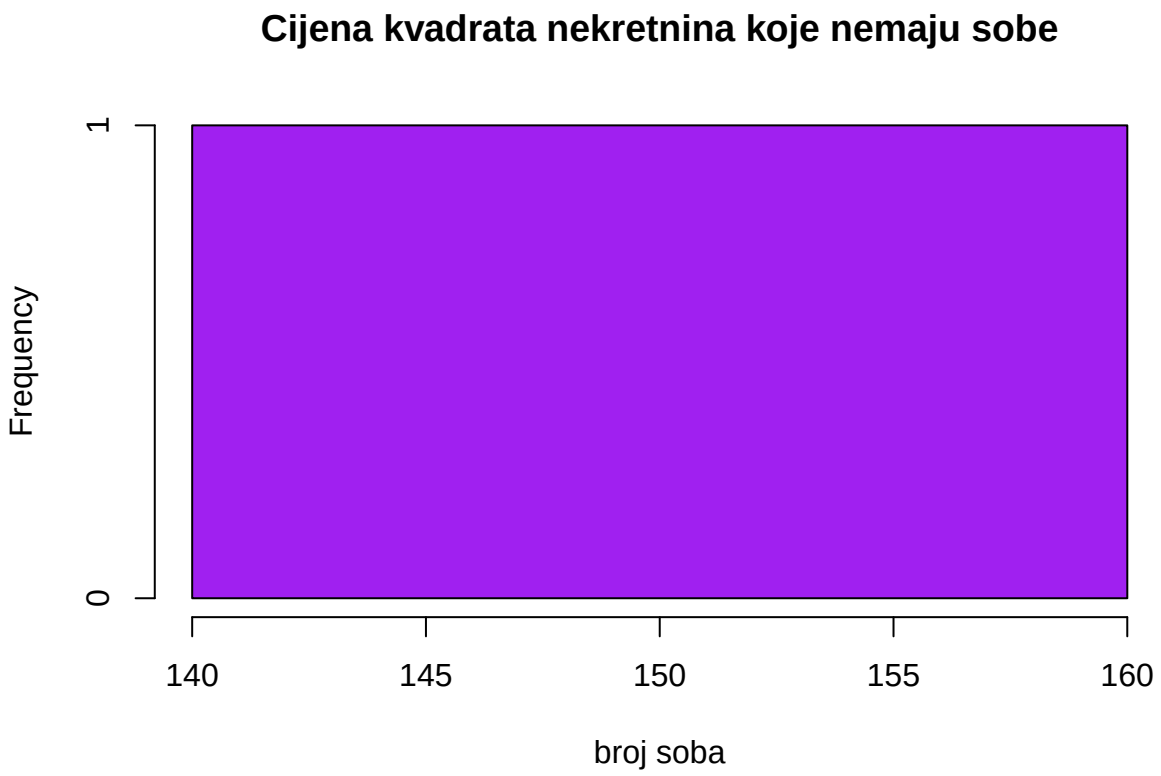
```
colors <- c("red", "blue", "green", "orange")

mosaicplot(tableHSLS,
           col=colors,
           xlab="House Style", ylab = "Lot Shape" ,main = "House Style by Shape of Lot")
```



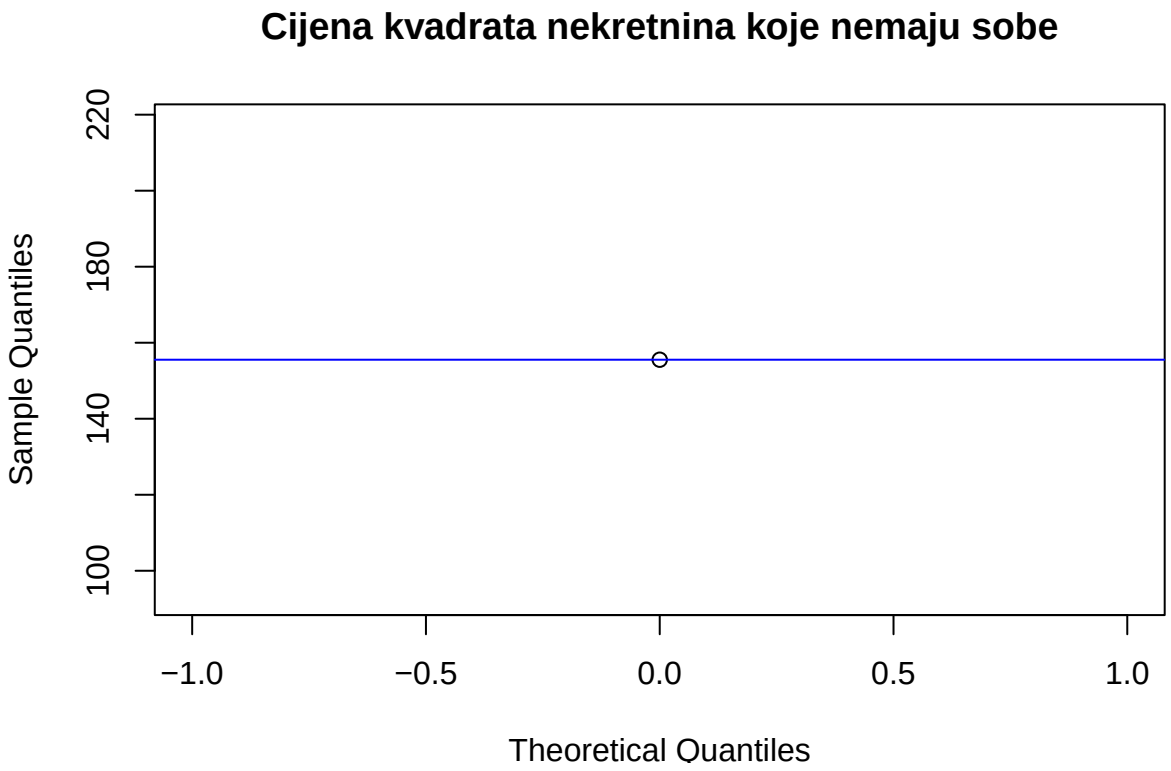
Ovisnost cijene kvadrata nekretnine o broju spavaćih soba

Za utvrđivanje postoji li ovisnost između cijene kvadrata i broja spavaćih soba provest ćemo 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. U kontekstu zadanog skupa podataka, ANOVA je dobro rješenje jer želimo usporediti srednje vrijednosti cijene po kvadratnom metru za nekretnine s različitim brojem spavaćih soba. P-vrijednost za ovaj test iznosi zanemarivih <2e-16, što nam govori da postoji značajna razlika između srednjih vrijednosti grupa.

```
data_room0 <- data[data$BedroomAbvGr == c("0"),]
hist(mean(data_room0$SalePrice/data_room0$GrLivArea),
     main = "Cijena kvadrata nekretnina koje nemaju sobe",
     col="purple", xlab="broj soba")
```



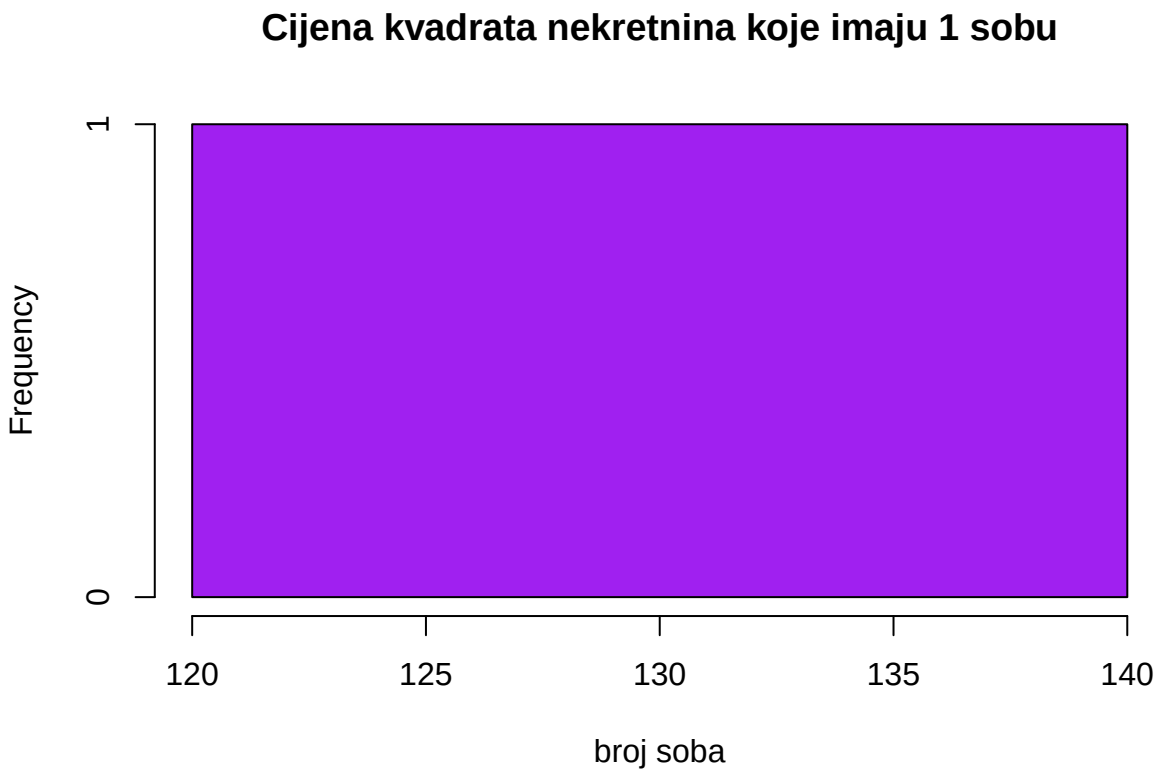

```
qqnorm(mean(data_room0$SalePrice/data_room0$GrLivArea),
       main="Cijena kvadrata nekretnina koje nemaju sobe")
qqline(mean(data_room0$SalePrice/data_room0$GrLivArea), col="blue")
```




```
data_room1 <- data[data$BedroomAbvGr == c("1"),]
```

```
hist(mean(data_room1$SalePrice/data_room0$GrLivArea),
     main = "Cijena kvadrata nekretnina koje imaju 1 sobu",
     col="purple", xlab="broj soba")
```

```
## Warning in data_room1$SalePrice/data_room0$GrLivArea: longer object length is
## not a multiple of shorter object length
```

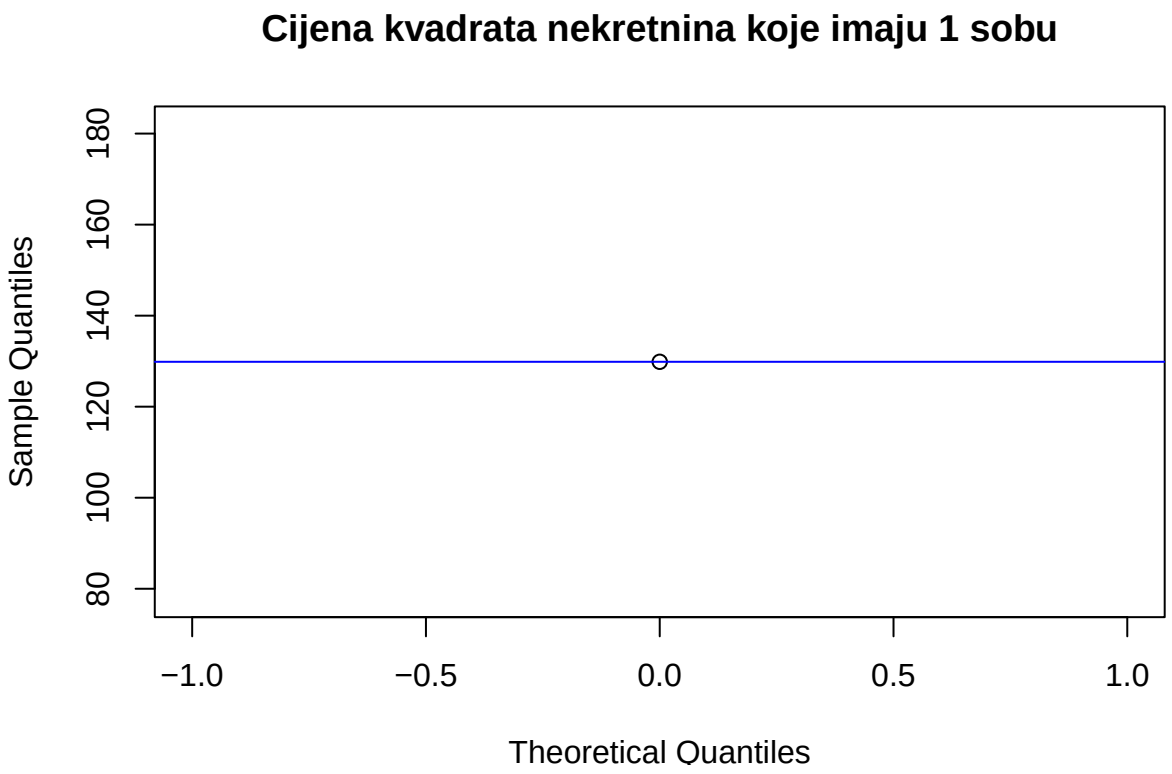



```
qqnorm(mean(data_room1$SalePrice/data_room0$GrLivArea),
       main="Cijena kvadrata nekretnina koje imaju 1 sobu")
```

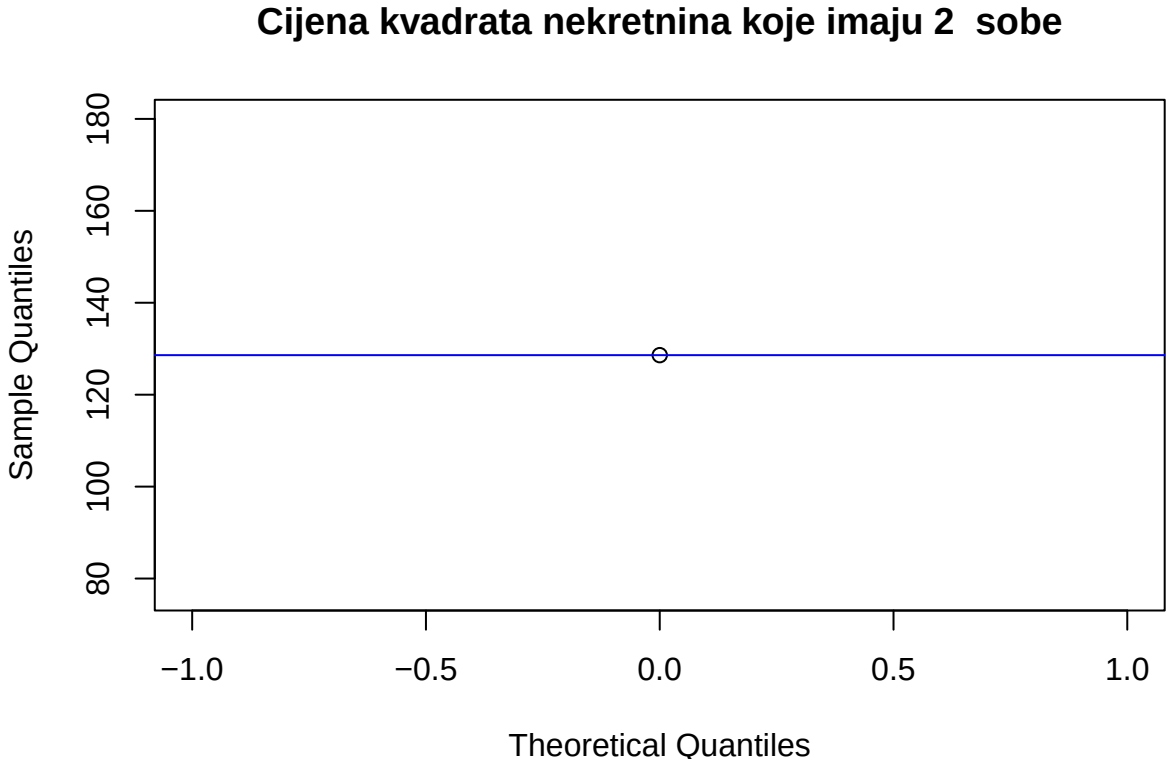
```
## Warning in data_room1$SalePrice/data_room0$GrLivArea: longer object length is
## not a multiple of shorter object length
```

```
qqline(mean(data_room1$SalePrice/data_room0$GrLivArea), col="blue")
```

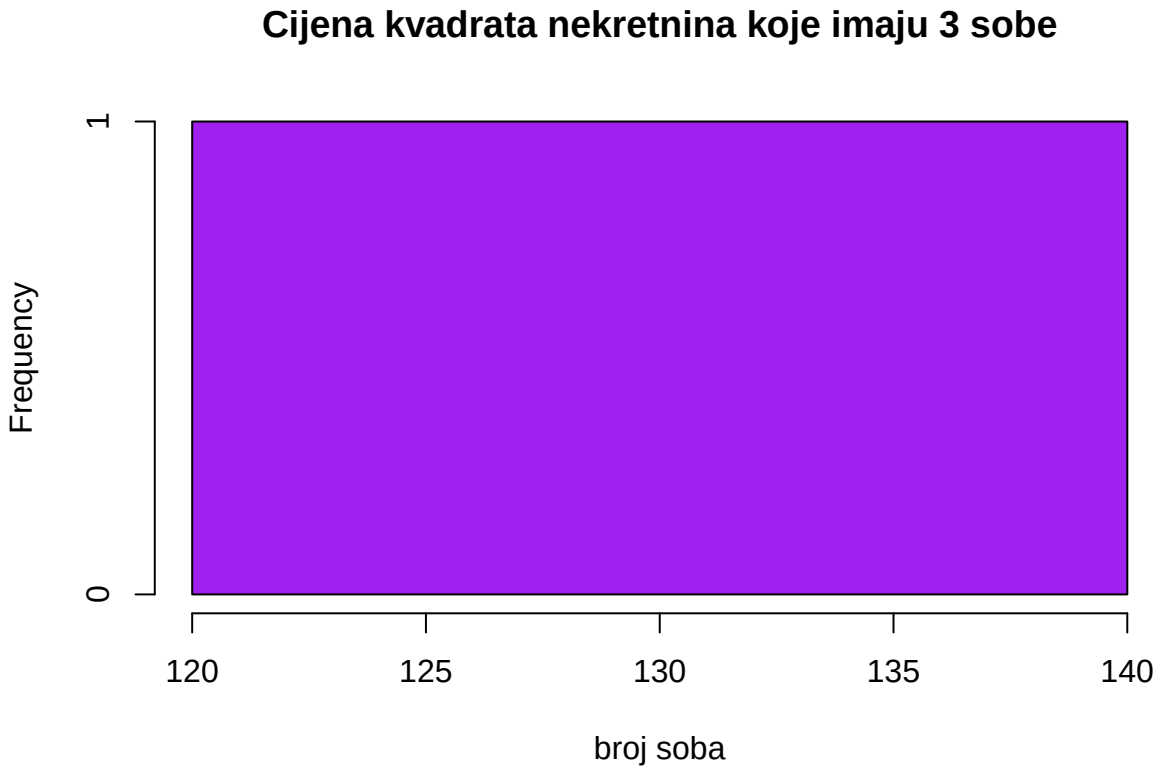
```
## Warning in data_room1$SalePrice/data_room0$GrLivArea: longer object length is
## not a multiple of shorter object length
```



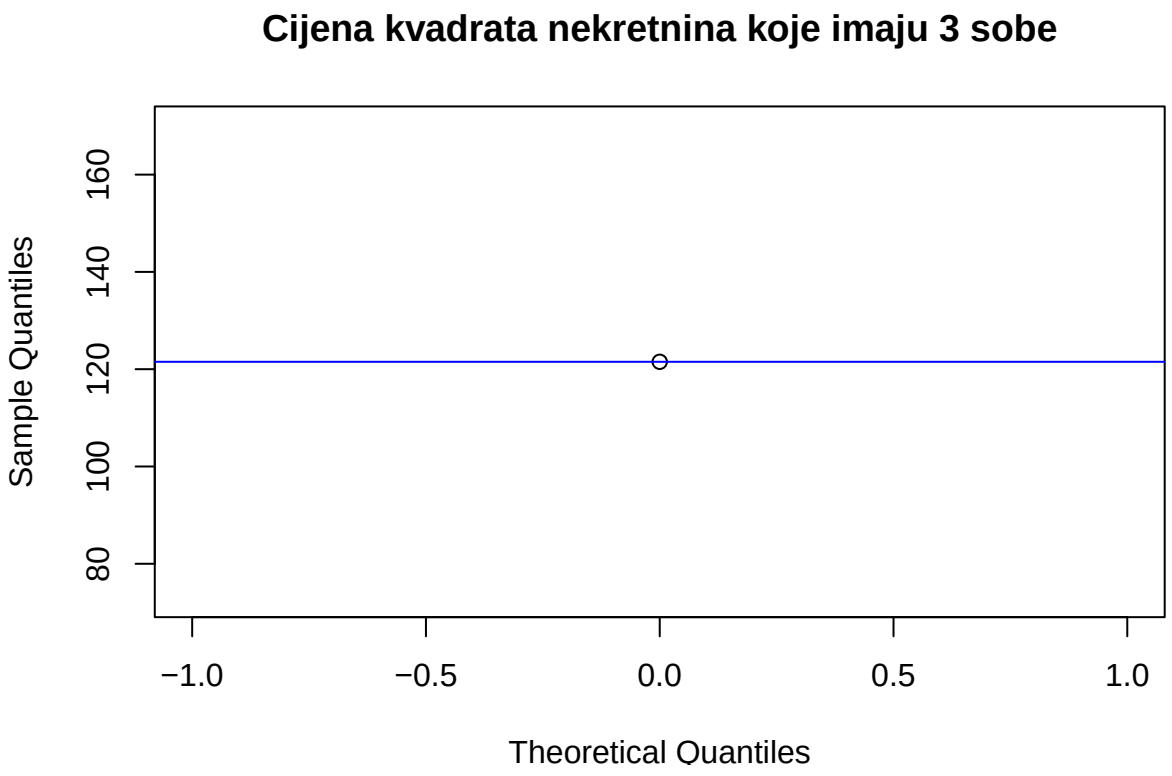

```
data_room2 <- data[data$BedroomAbvGr == c("2"),]

qqnorm(mean(data_room2$SalePrice/data_room2$GrLivArea),
       main="Cijena kvadrata nekretnina koje imaju 2  sobe")
qqline(mean(data_room2$SalePrice/data_room2$GrLivArea), col="blue")
```



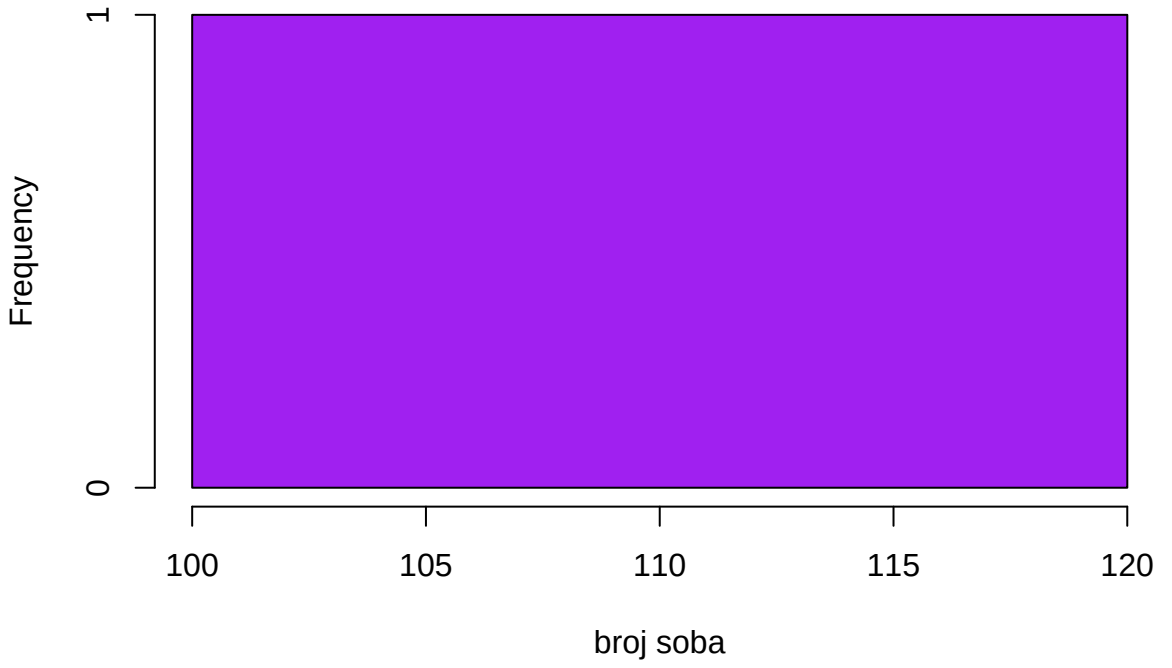
```
data_room3 <- data[data$BedroomAbvGr == c("3"),]
hist(mean(data_room3$SalePrice/data_room3$GrLivArea),
     main = "Cijena kvadrata nekretnina koje imaju 3 sobe",
     col="purple", xlab="broj soba")
```



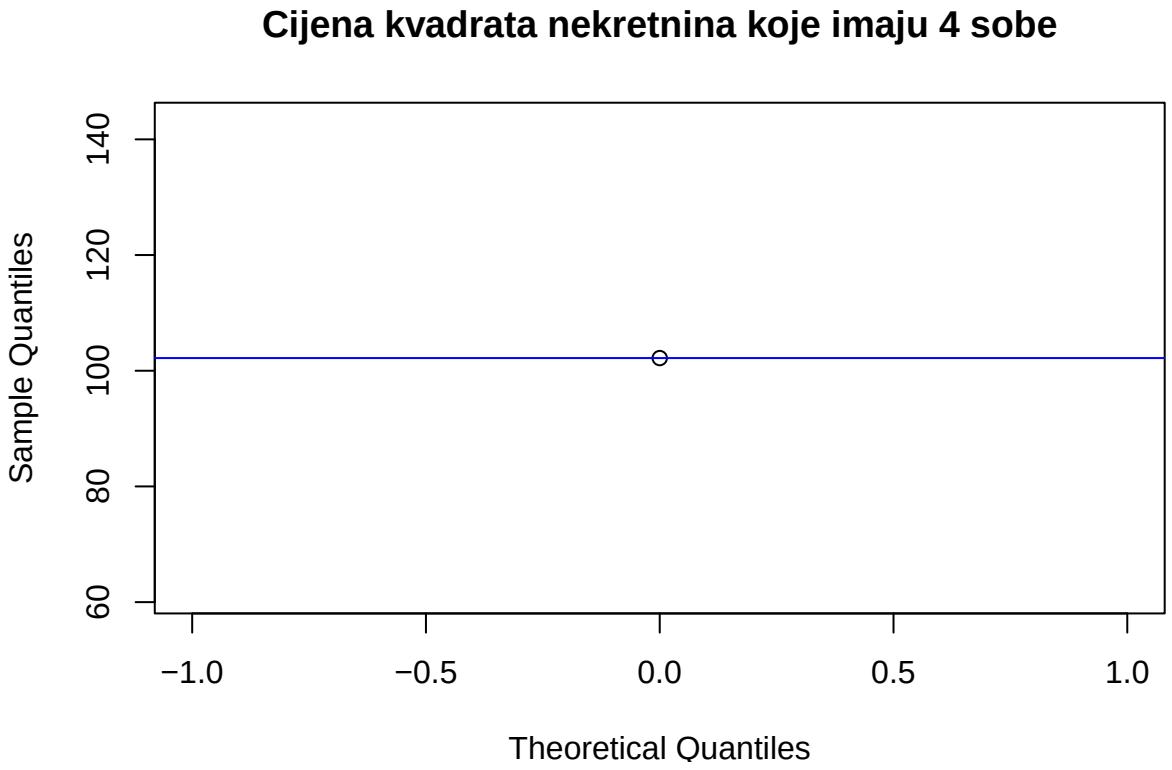

```
qqnorm(mean(data_room3$SalePrice/data_room3$GrLivArea),
       main="Cijena kvadrata nekretnina koje imaju 3 sobe")
qqline(mean(data_room3$SalePrice/data_room3$GrLivArea), col="blue")
```



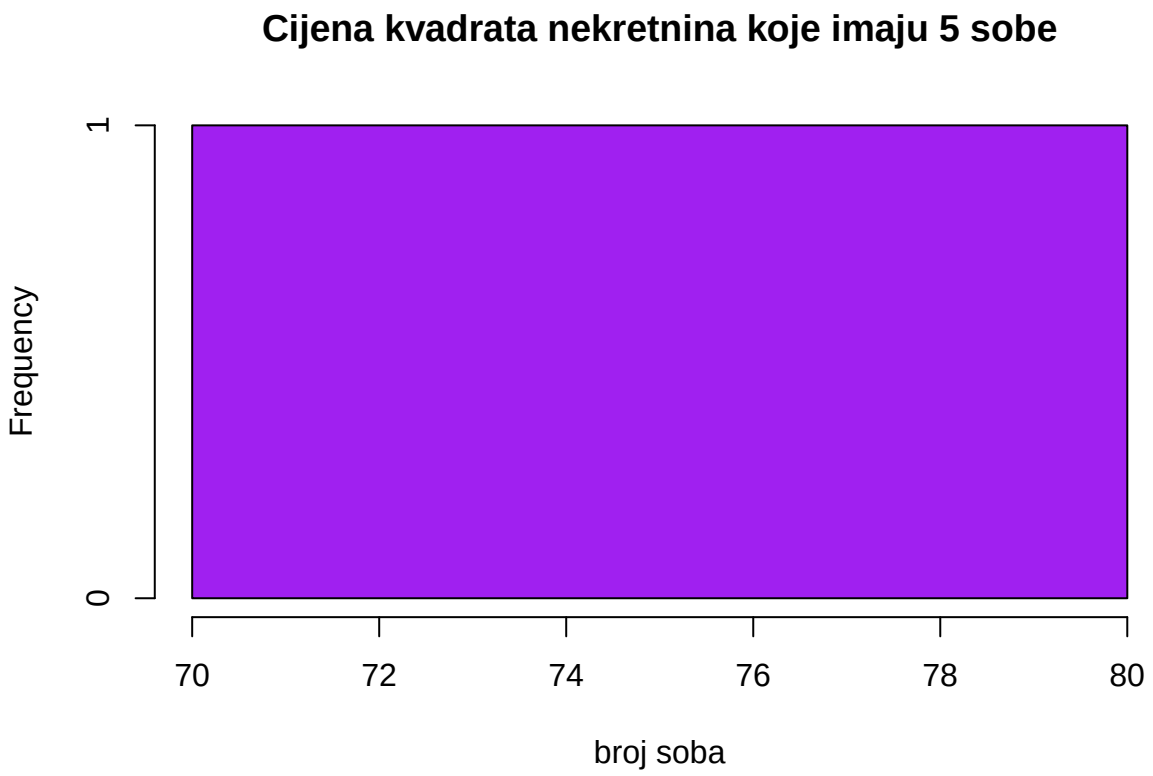

```
data_room4 <- data[data$BedroomAbvGr == c("4"),]
hist(mean(data_room4$SalePrice/data_room4$GrLivArea),
     main = "Cijena kvadrata nekretnina koje imaju 4 sobe",
     col="purple", xlab="broj soba")
```



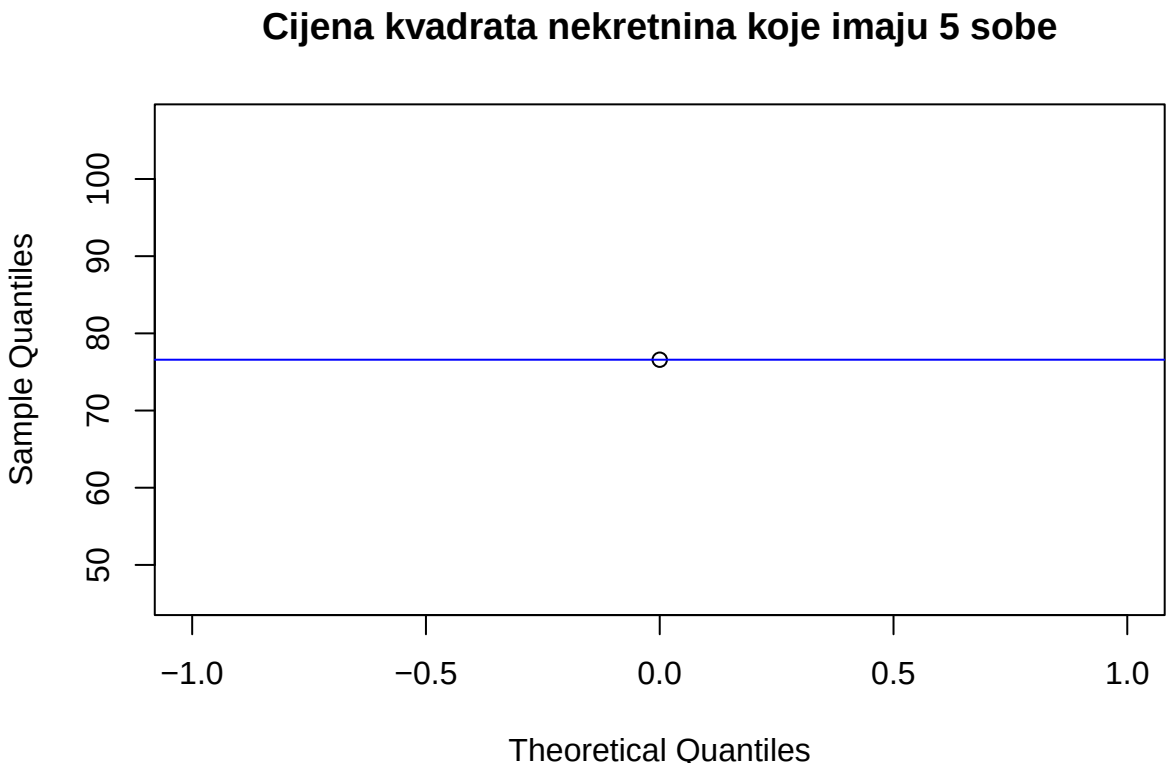
```
qqnorm(mean(data_room4$SalePrice/data_room4$GrLivArea),
       main="Cijena kvadrata nekretnina koje imaju 4 sobe")
qqline(mean(data_room4$SalePrice/data_room4$GrLivArea), col="blue")
```




```
data_room5 <- data[data$BedroomAbvGr == c("5"),]
hist(mean(data_room5$SalePrice/data_room5$GrLivArea),
     main = "Cijena kvadrata nekretnina koje imaju 5 sobe",
     col="purple", xlab="broj soba")
```

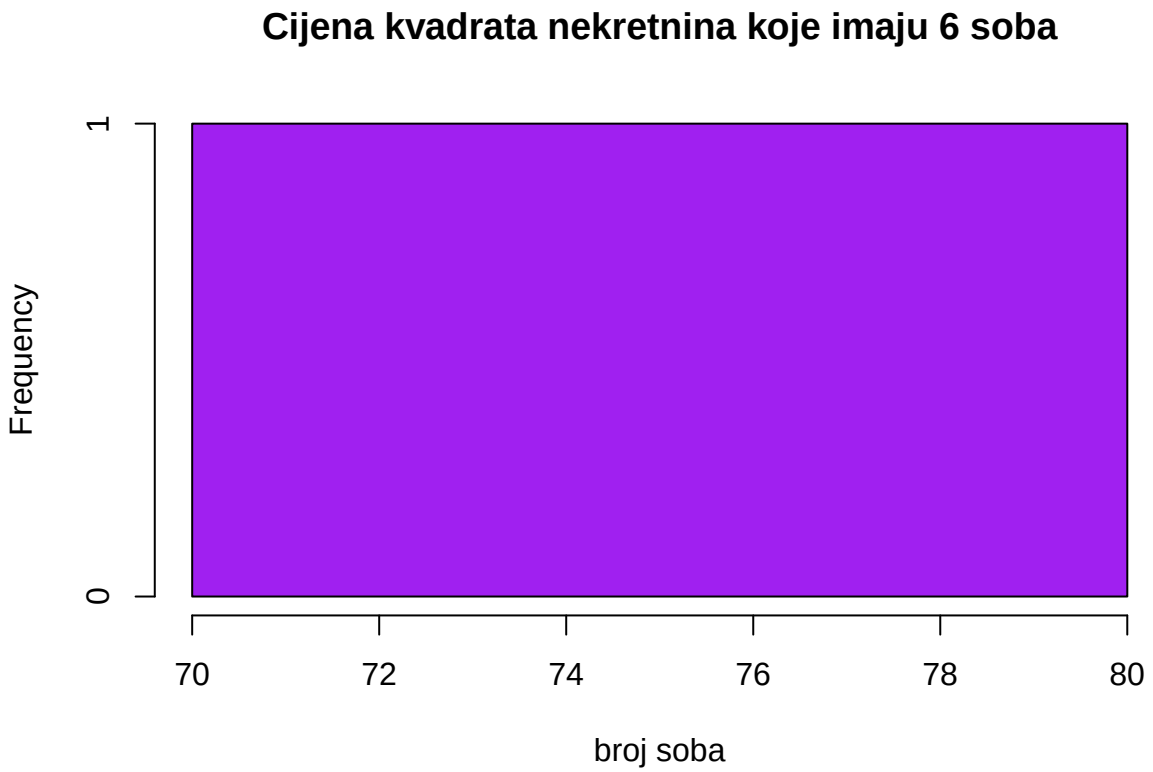



```
qqnorm(mean(data_room5$SalePrice/data_room5$GrLivArea),
       main="Cijena kvadrata nekretnina koje imaju 5 sobe")
qqline(mean(data_room5$SalePrice/data_room5$GrLivArea), col="blue")
```

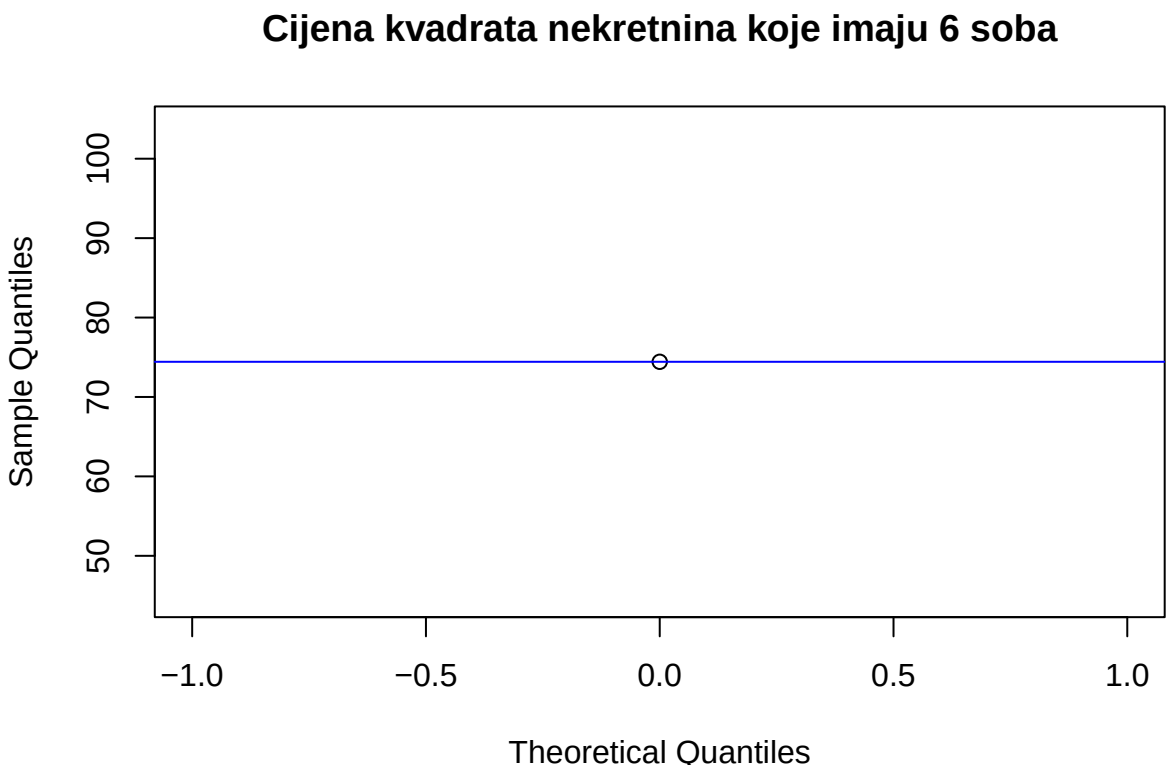



```
data_room6 <- data[data$BedroomAbvGr == c("6"),]
```

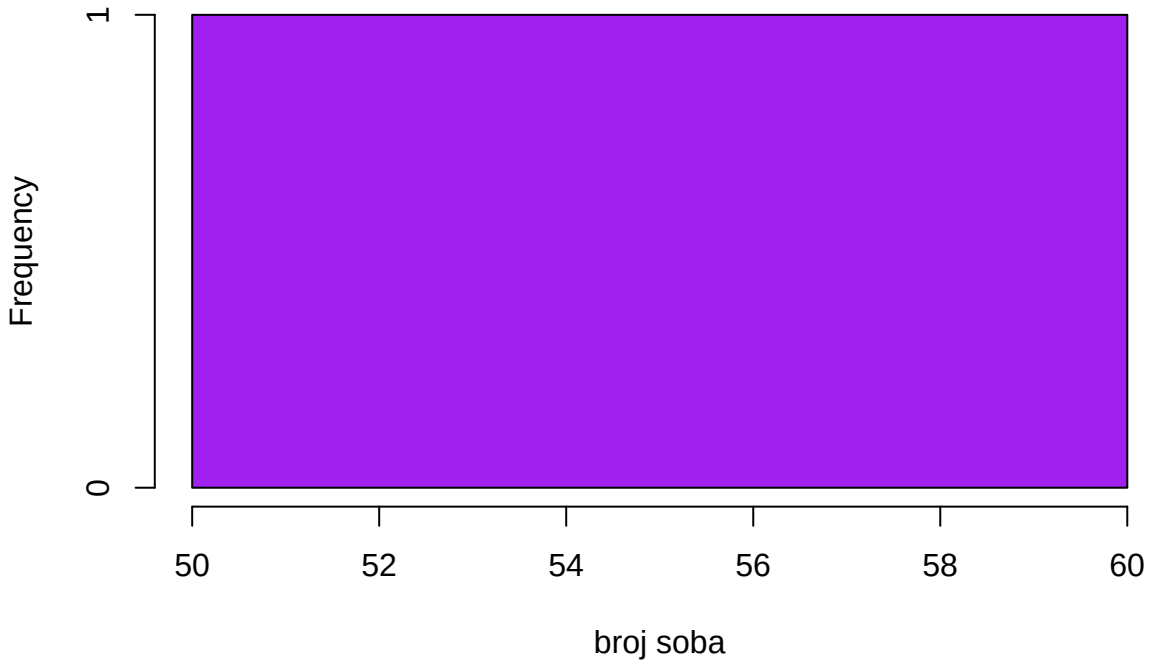
```
hist(mean(data_room6$SalePrice/data_room6$GrLivArea),
     main = "Cijena kvadrata nekretnina koje imaju 6 soba",
     col="purple", xlab="broj soba")
```



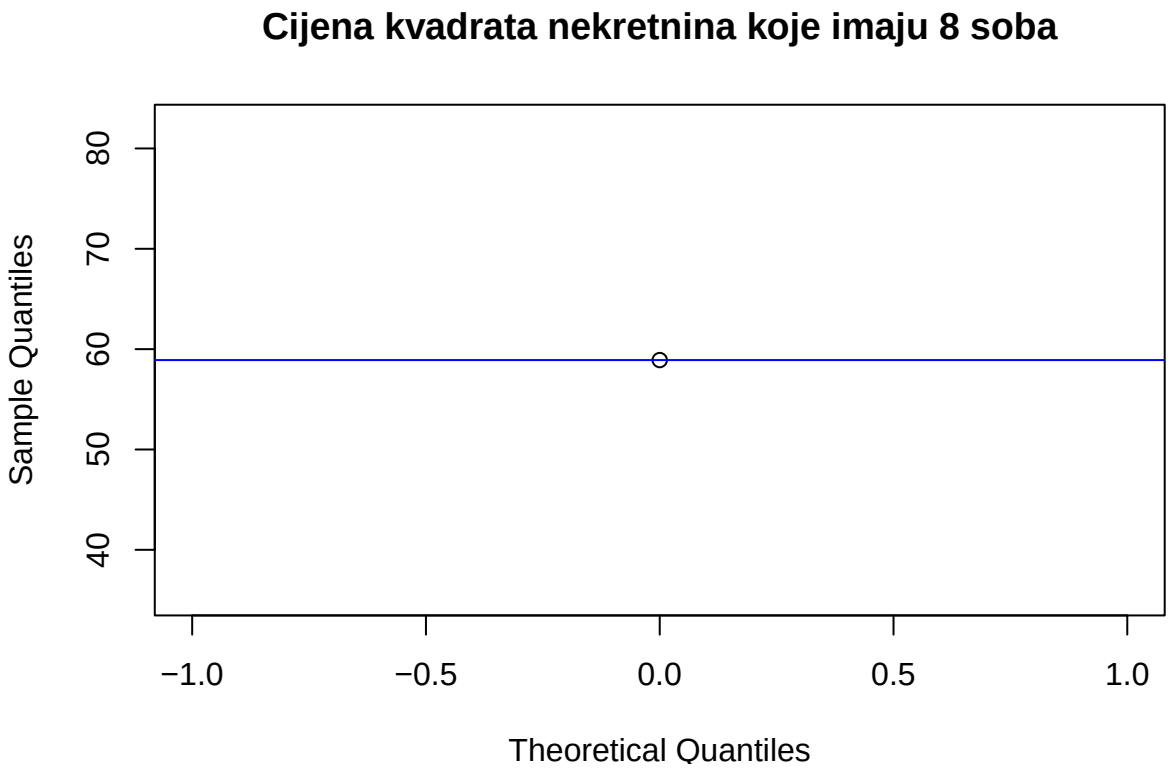

```
qqnorm(mean(data_room6$SalePrice/data_room6$GrLivArea),
       main="Cijena kvadrata nekretnina koje imaju 6 soba")
qqline(mean(data_room6$SalePrice/data_room6$GrLivArea), col="blue")
```




```
data_room8 <- data[data$BedroomAbvGr == c("8"),]
hist(mean(data_room8$SalePrice/data_room8$GrLivArea),
     main = "Cijena kvadrata nekretnina koje imaju 8 soba",
     col="purple", xlab="broj soba")
```



```
qqnorm(mean(data_room8$SalePrice/data_room8$GrLivArea),
       main="Cijena kvadrata nekretnina koje imaju 8 soba")
qqline(mean(data_room8$SalePrice/data_room8$GrLivArea), col="blue")
```




```
var(na.omit(mean(data_room0$SalePrice/data_room0$GrLivArea)))
```

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

```
var(na.omit(mean(data_room1$SalePrice/data_room1$GrLivArea)))
```

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

```
var(na.omit(mean(data_room2$SalePrice/data_room2$GrLivArea)))
```

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

```
var(na.omit(mean(data_room3$SalePrice/data_room3$GrLivArea)))
```

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

```
var(na.omit(mean(data_room4$SalePrice/data_room4$GrLivArea)))
```

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

```
var(na.omit(mean(data_room5$SalePrice/data_room5$GrLivArea)))
```

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

```
var(na.omit(mean(data_room6$SalePrice/data_room6$GrLivArea)))
```

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

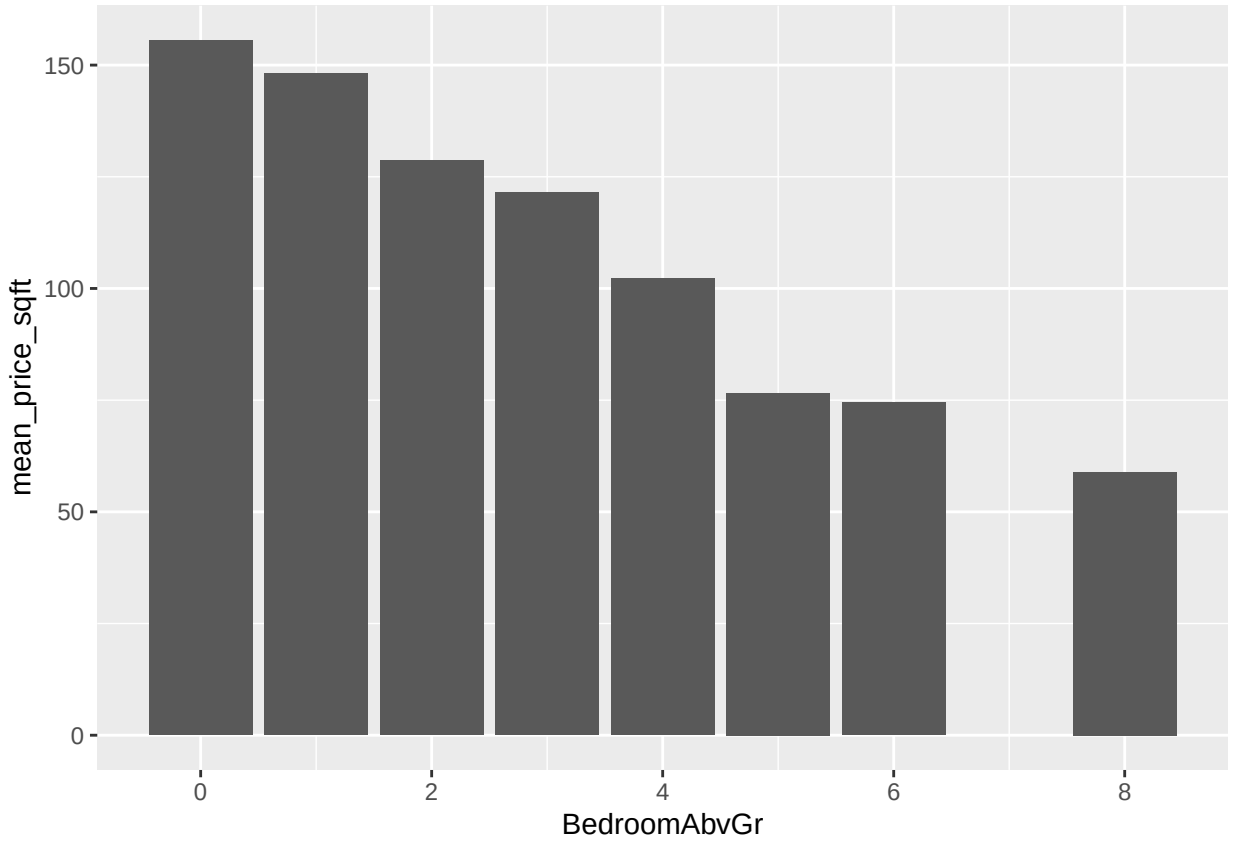
```
var(na.omit(mean(data_room8$SalePrice/data_room8$GrLivArea)))
```

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

```
res.aov <- aov( SalePrice/GrLivArea~ BedroomAbvGr, data = data)
summary(res.aov)
```

```
##                Df  Sum Sq Mean Sq F value Pr(>F)
## BedroomAbvGr    1  182177  182177   211.6 <2e-16 ***
## Residuals    1458 1255521     861
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
data_by_bedrooms <- data %>%
  group_by(BedroomAbvGr)%>%
  summarize(mean_price_sqft =mean(SalePrice/GrLivArea),
            sd_price_sqft=sd(SalePrice/GrLivArea))
 ggplot(data = data_by_bedrooms,
        aes(x = BedroomAbvGr, y = mean_price_sqft)) + geom_bar(stat = "identity")
```



ZAKLJUČAK: odbacujemo nultu hipotezu, to jest cijena kvadrata nekretnine ne ovisi o broju spavaćih soba

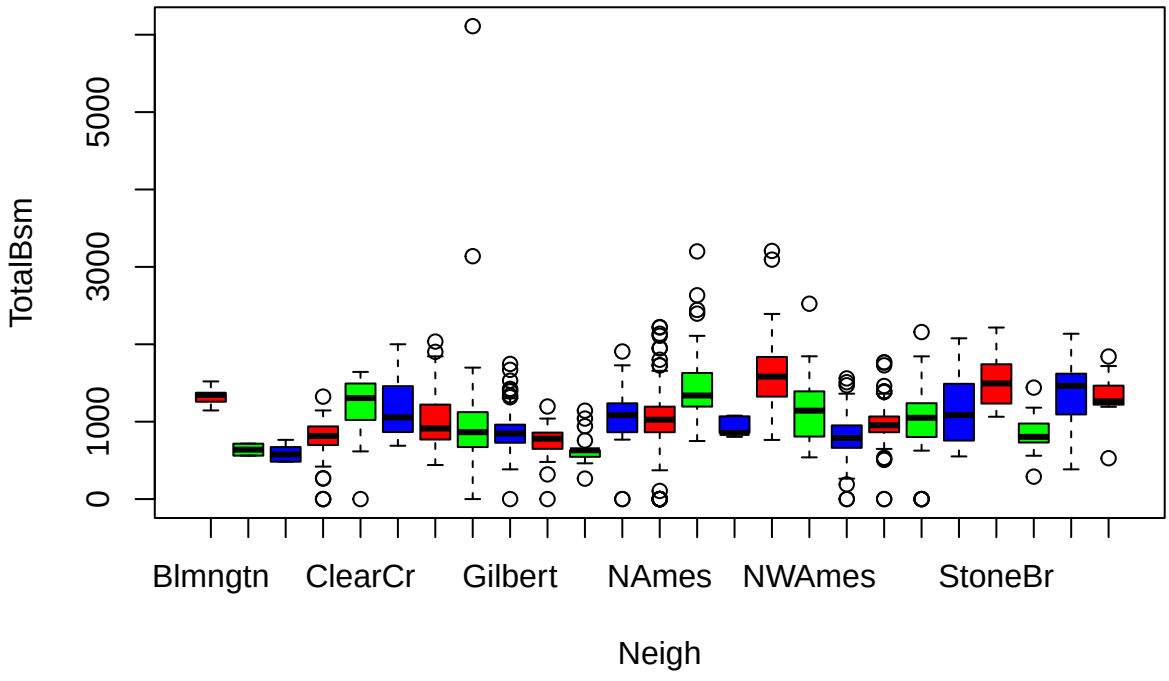
Ovisi li velicina podruma o kvartu u gradu

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.

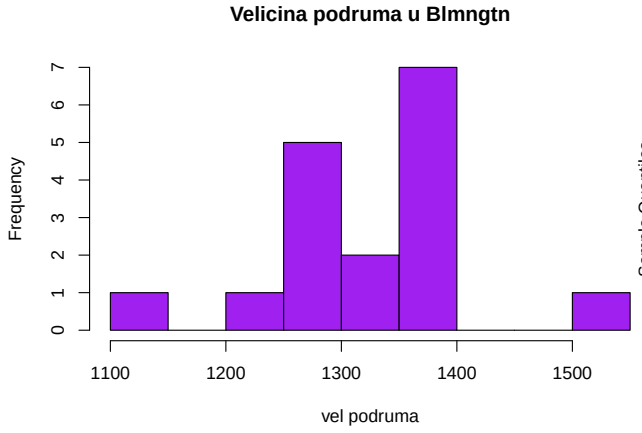
```
data_by_neighborhood <- data %>% group_by(data$Neighborhood)
print(data_by_neighborhood)
```

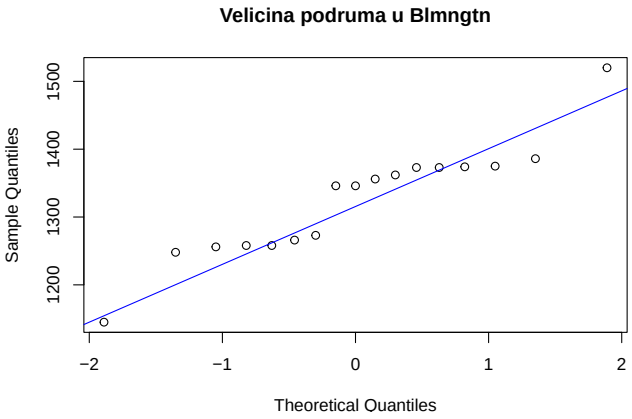
```
## # A tibble: 1,460 x 83
## # Groups:   data$Neighborhood [25]
##       Id MSSubClass MSZon~1 LotFr~2 LotArea Street Alley LotSh~3 LandC~4 Utili~5
##    <int>      <int> <chr>     <int>   <int> <chr>  <chr> <chr>   <chr>   <chr>
##  1     1         60 RL           65    8450 Pave   <NA>  Reg     Lvl     AllPub
##  2     2         20 RL           80    9600 Pave   <NA>  Reg     Lvl     AllPub
##  3     3         60 RL           68   11250 Pave   <NA>  IR1     Lvl     AllPub
##  4     4         70 RL           60    9550 Pave   <NA>  IR1     Lvl     AllPub
##  5     5         60 RL           84   14260 Pave   <NA>  IR1     Lvl     AllPub
##  6     6         50 RL           85   14115 Pave   <NA>  IR1     Lvl     AllPub
##  7     7         20 RL           75   10084 Pave   <NA>  Reg     Lvl     AllPub
##  8     8         60 RL           NA   10382 Pave   <NA>  IR1     Lvl     AllPub
##  9     9         50 RM           51    6120 Pave   <NA>  Reg     Lvl     AllPub
## 10    10        190 RL           50    7420 Pave   <NA>  Reg     Lvl     AllPub
## # ... with 1,450 more rows, 73 more variables: LotConfig <chr>,
## #   LandSlope <chr>, Neighborhood <chr>, Condition1 <chr>, Condition2 <chr>,
## #   BldgType <chr>, HouseStyle <chr>, OverallQual <int>, OverallCond <int>,
## #   YearBuilt <int>, YearRemodAdd <int>, RoofStyle <chr>, RoofMatl <chr>,
## #   Exterior1st <chr>, Exterior2nd <chr>, MasVnrType <chr>, MasVnrArea <int>,
## #   ExterQual <chr>, ExterCond <chr>, Foundation <chr>, BsmtQual <chr>,
## #   BsmtCond <chr>, BsmtExposure <chr>, BsmtFinType1 <chr>, ...
```

```
boxplot(data_by_neighborhood$TotalBsmtSF
        ~ data_by_neighborhood$Neighborhood,
        ylab = "TotalBsm",
        xlab = "Neigh", col=rainbow(3))
```

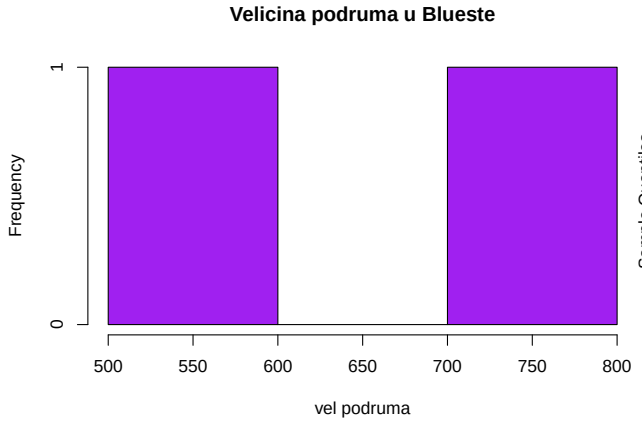


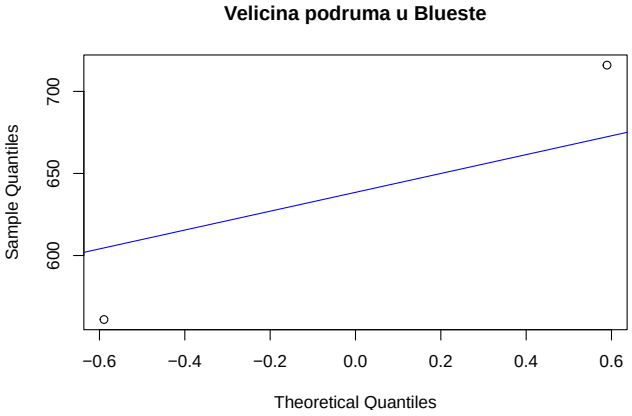
```
data_re1 <- data[data$Neighborhood == c("Blmngtn"),]
hist(data_re1$TotalBsm,
     main = "Velicina podruma u Blmngtn",
     col="purple", xlab="vel podruma")
qqnorm(data_re1$TotalBsmtSF, main="Velicina podruma u Blmngtn")
qqline(data_re1$TotalBsmtS, col="blue")
```



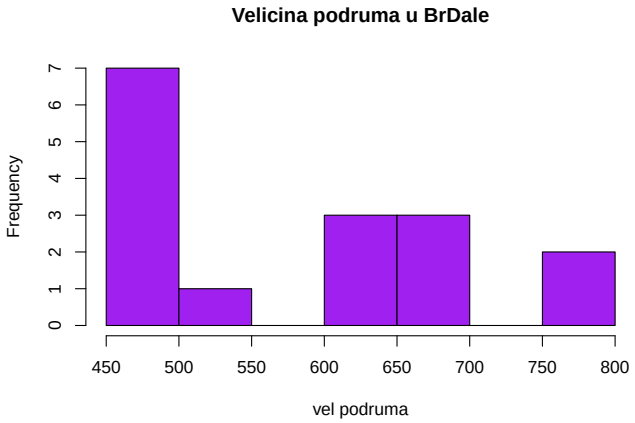


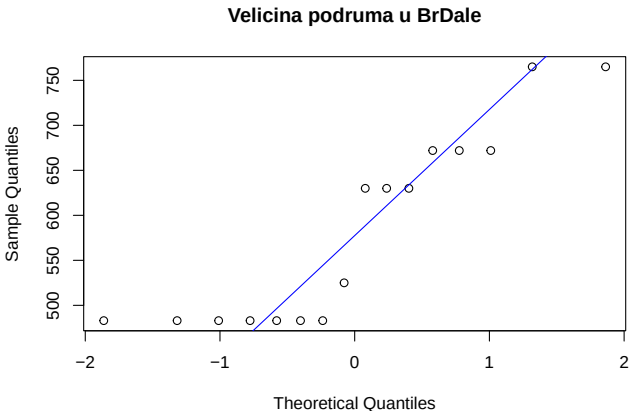

```
data_re2 <- data[data$Neighborhood == c("Blueste"),]
hist(data_re2$TotalBsm,
     main = "Velicina podruma u Blueste",
     col="purple", xlab="vel podruma")
qqnorm(data_re2$TotalBsmtSF, main="Velicina podruma u Blueste")
qqline(data_re2$TotalBsmtS, col="blue")
```



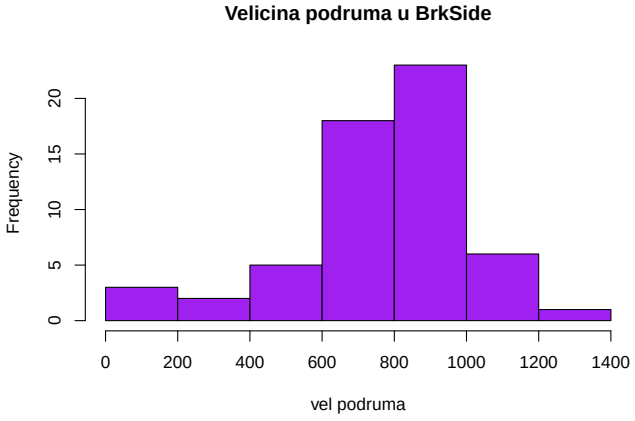


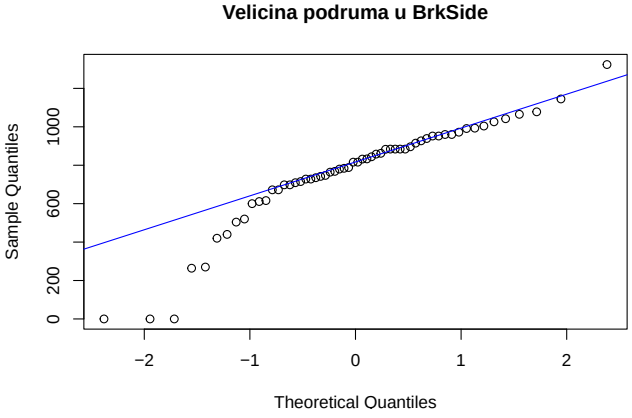

```
data_re3 <- data[data$Neighborhood == c("BrDale"),]
hist(data_re3$TotalBsm,
     main = "Velicina podruma u BrDale",
     col="purple", xlab="vel podruma")
qqnorm(data_re3$TotalBsmtSF, main="Velicina podruma u BrDale")
qqline(data_re3$TotalBsmtS, col="blue")
```



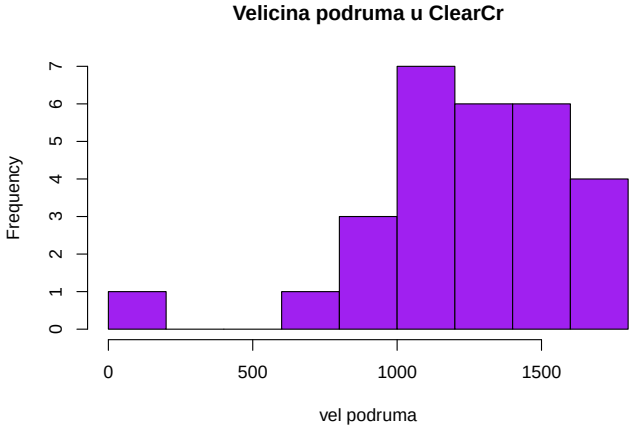


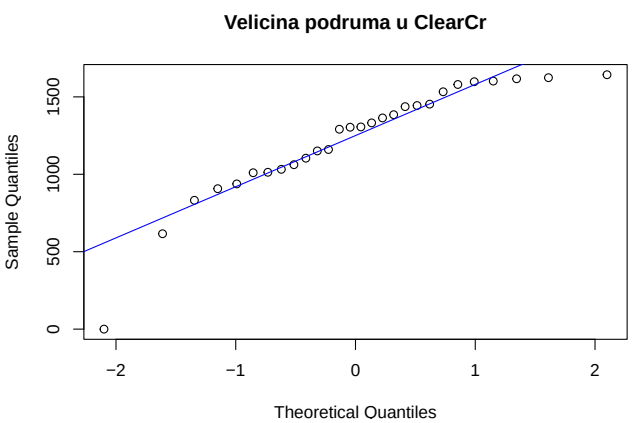

```
data_re4 <- data[data$Neighborhood == c("BrkSide"),]
hist(data_re4$TotalBsm,
     main = "Velicina podruma u BrkSide",
     col="purple", xlab="vel podruma")
qqnorm(data_re4$TotalBsmtSF, main="Velicina podruma u BrkSide")
qqline(data_re4$TotalBsmtS, col="blue")
```



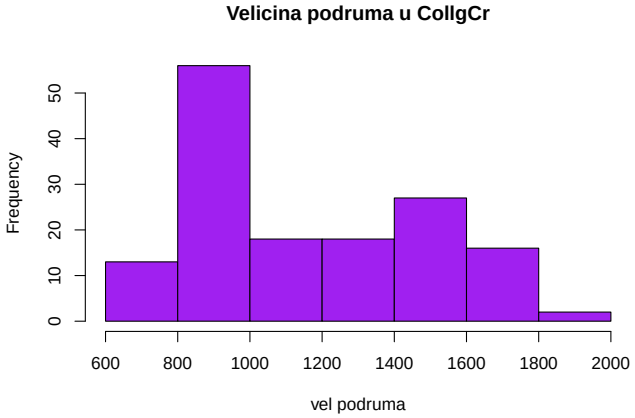


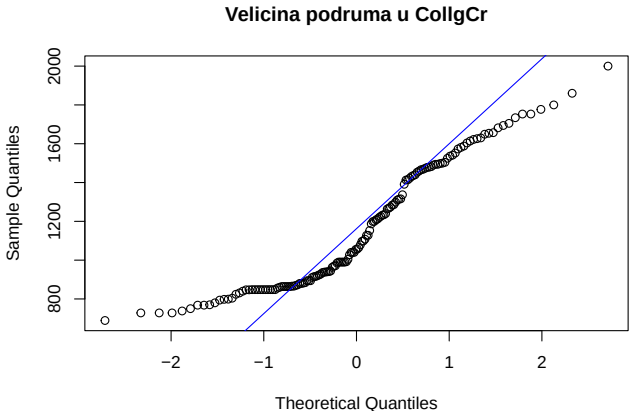

```
data_re5 <- data[data$Neighborhood == c("ClearCr"),]
hist(data_re5$TotalBsm,
     main = "Velicina podruma u ClearCr",
     col="purple", xlab="vel podruma")
qqnorm(data_re5$TotalBsmtSF, main="Velicina podruma u ClearCr")
qqline(data_re5$TotalBsmtS, col="blue")
```



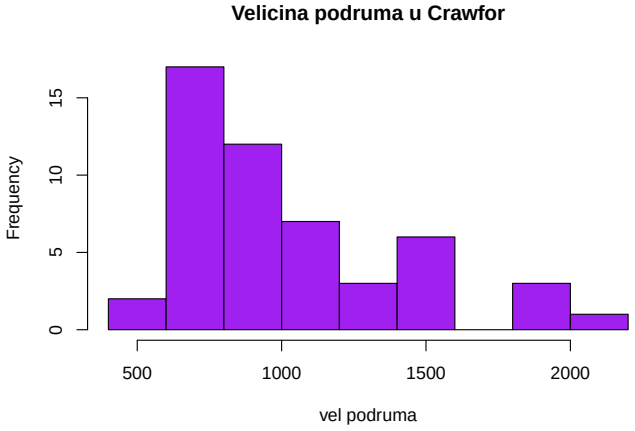


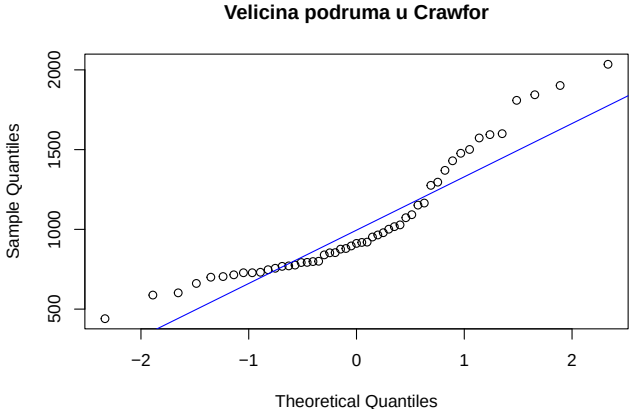
```
data_re6 <- data[data$Neighborhood == c("CollgCr"),]
hist(data_re6$TotalBsm,
     main = "Velicina podruma u CollgCr",
     col="purple", xlab="vel podruma")
qqnorm(data_re6$TotalBsmtSF, main="Velicina podruma u CollgCr")
qqline(data_re6$TotalBsmtS, col="blue")
```



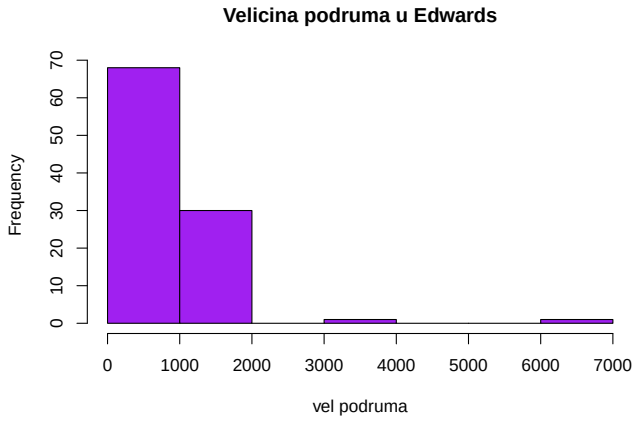


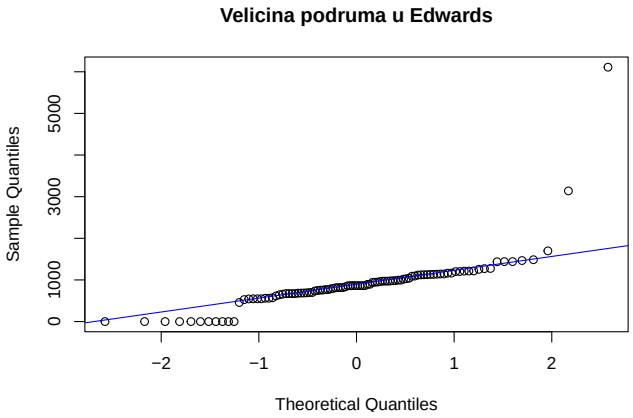

```
data_re7 <- data[data$Neighborhood == c("Crawfor"),]
hist(data_re7$TotalBsm,
     main = "Velicina podruma u Crawfor",
     col="purple", xlab="vel podruma")
qqnorm(data_re7$TotalBsmtSF, main="Velicina podruma u Crawfor")
qqline(data_re7$TotalBsmtS, col="blue")
```



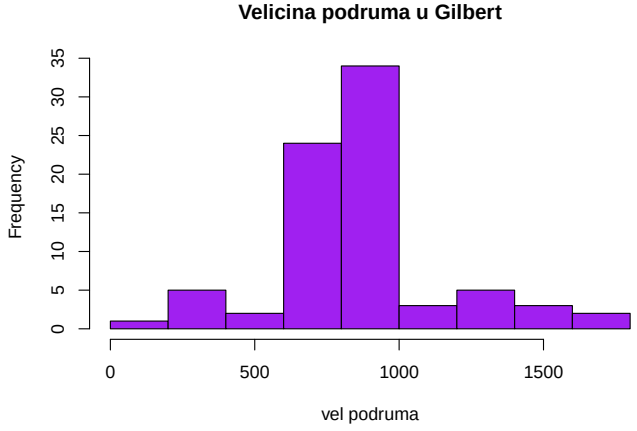


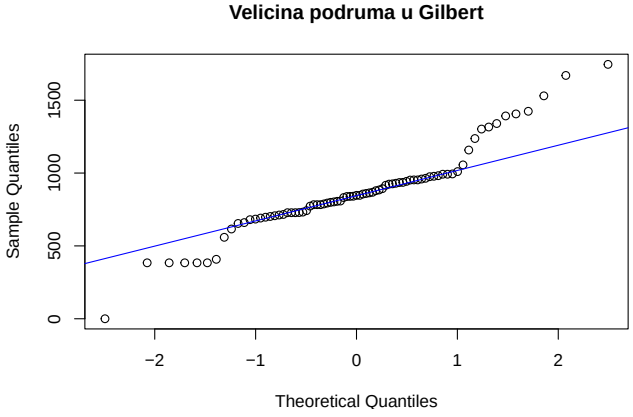

```
data_re8 <- data[data$Neighborhood == c("Edwards"),]
hist(data_re8$TotalBsm,
     main = "Velicina podruma u Edwards",
     col="purple", xlab="vel podruma")
qqnorm(data_re8$TotalBsmtSF, main="Velicina podruma u Edwards")
qqline(data_re8$TotalBsmtS, col="blue")
```



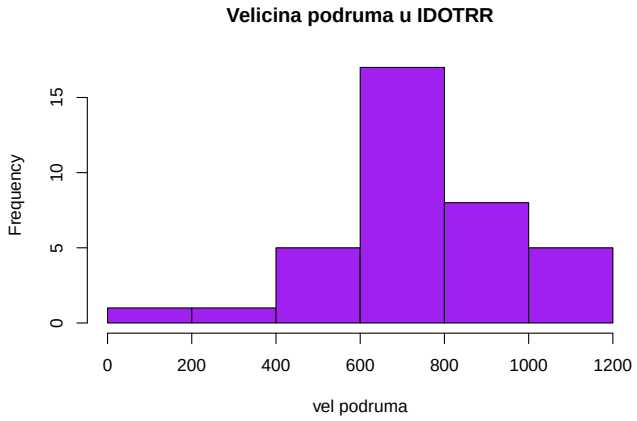


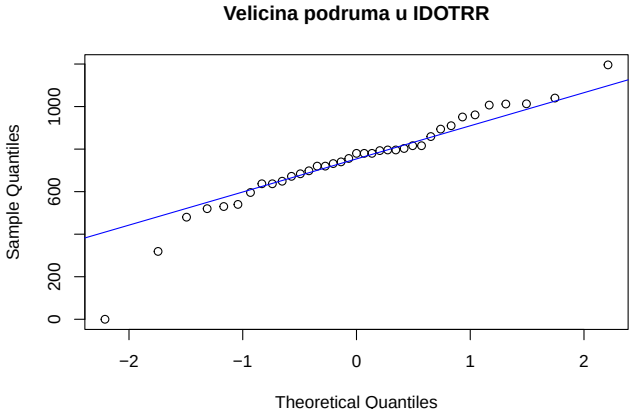

```
data_re9 <- data[data$Neighborhood == c("Gilbert"),]
hist(data_re9$TotalBsm,
     main = "Velicina podruma u Gilbert",
     col="purple", xlab="vel podruma")
qqnorm(data_re9$TotalBsmtSF, main="Velicina podruma u Gilbert")
qqline(data_re9$TotalBsmtS, col="blue")
```



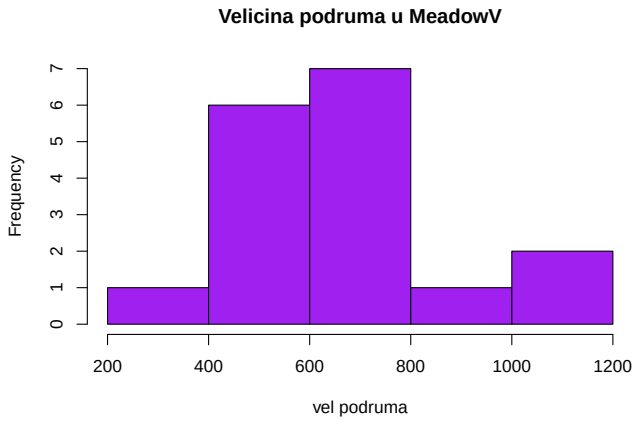


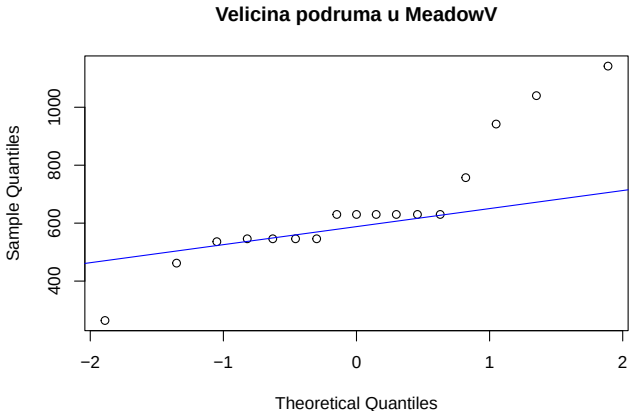

```
data_re10 <- data[data$Neighborhood == c("IDOTRR"),]
hist(data_re10$TotalBsm,
     main = "Velicina podruma u IDOTRR",
     col="purple", xlab="vel podruma")
qqnorm(data_re10$TotalBsmtSF, main="Velicina podruma u IDOTRR")
qqline(data_re10$TotalBsmtS, col="blue")
```



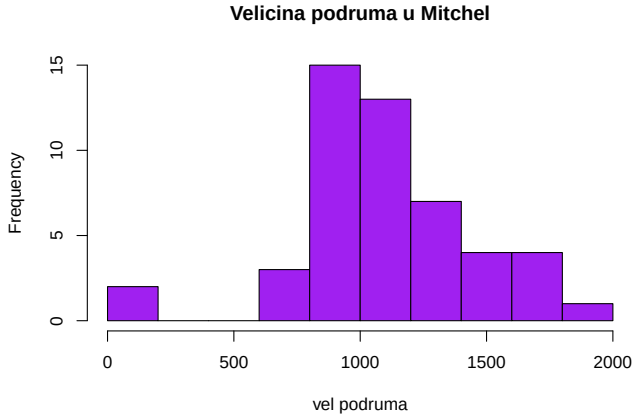


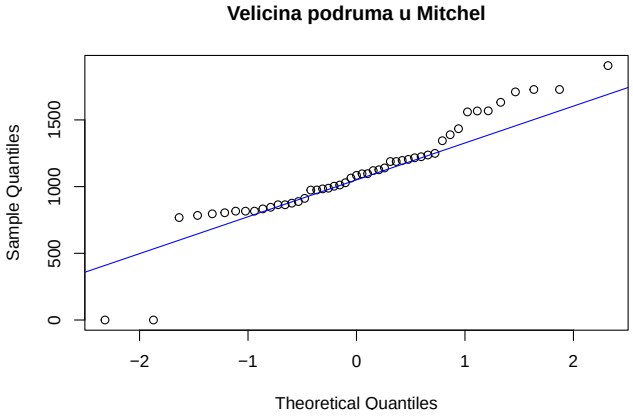
```
data_re11 <- data[data$Neighborhood == c("MeadowV"),]
hist(data_re11$TotalBsm,
     main = "Velicina podruma u MeadowV",
     col="purple", xlab="vel podruma")
qqnorm(data_re11$TotalBsmtSF, main="Velicina podruma u MeadowV")
qqline(data_re11$TotalBsmtS, col="blue")
```



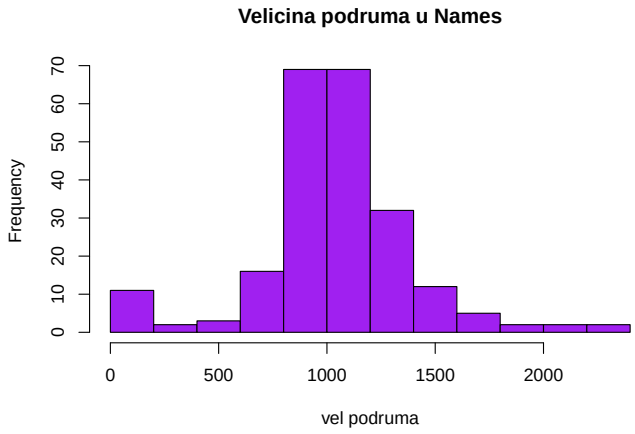


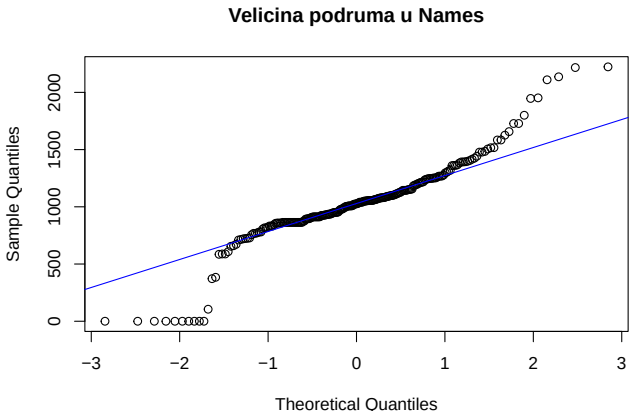

```
data_re12 <- data[data$Neighborhood == c("Mitchel"),]
hist(data_re12$TotalBsm,
     main = "Velicina podruma u Mitchel",
     col="purple", xlab="vel podruma")
qqnorm(data_re12$TotalBsmtSF, main="Velicina podruma u Mitchel")
qqline(data_re12$TotalBsmtS, col="blue")
```



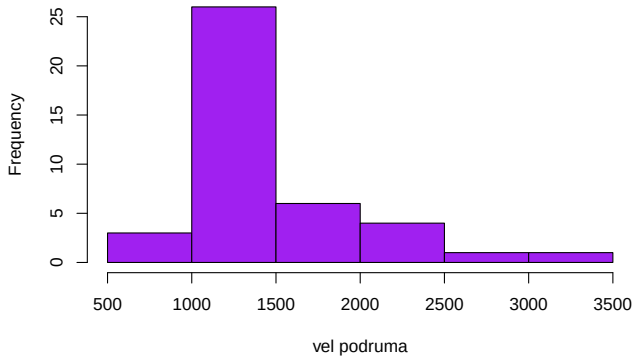


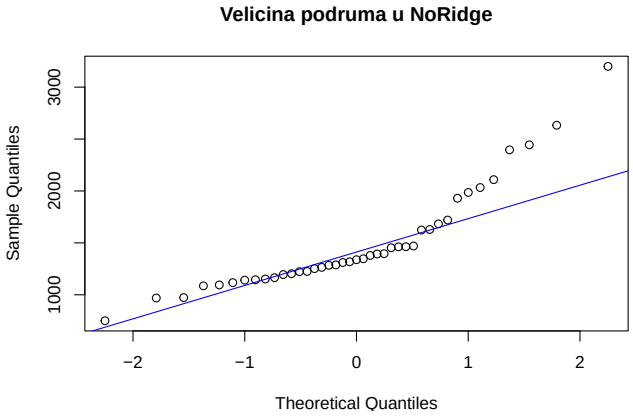

```
data_re13 <- data[data$Neighborhood == c("NAmes"),]
hist(data_re13$TotalBsm,
     main = "Velicina podruma u Names",
     col="purple", xlab="vel podruma")
qqnorm(data_re13$TotalBsmtSF, main="Velicina podruma u Names")
qqline(data_re13$TotalBsmtS, col="blue")
```



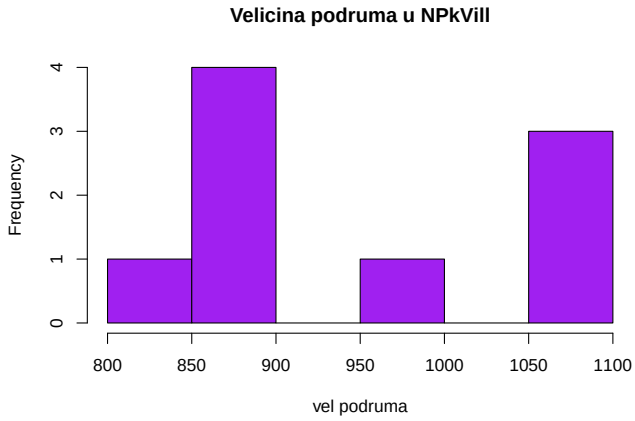


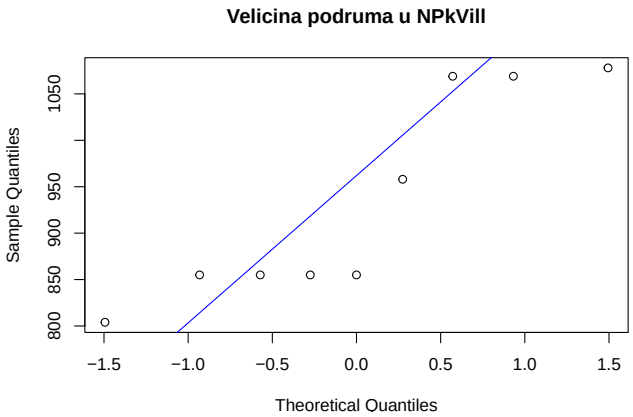

```
data_re14 <- data[data$Neighborhood == c("NoRidge"),]
hist(data_re14$TotalBsm,
     main = "Velicina podruma u NoRidge",
     col="purple", xlab="vel podruma")
qqnorm(data_re14$TotalBsmtSF, main="Velicina podruma u NoRidge")
qqline(data_re14$TotalBsmtS, col="blue")
```



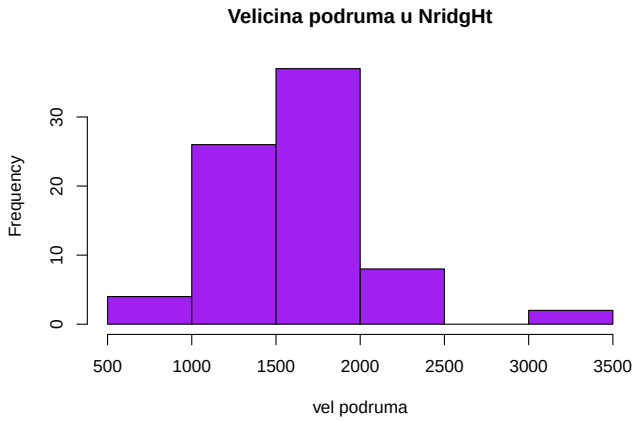


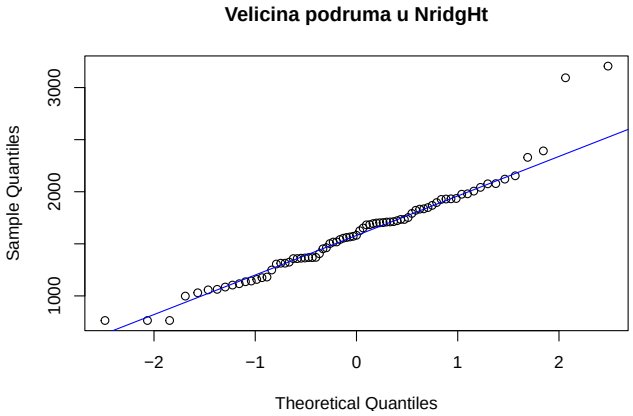

```
data_re15 <- data[data$Neighborhood == c("NPkVill"),]
hist(data_re15$TotalBsm,
     main = "Velicina podruma u NPkVill",
     col="purple", xlab="vel podruma")
qqnorm(data_re15$TotalBsmtSF, main="Velicina podruma u NPkVill")
qqline(data_re15$TotalBsmtS, col="blue")
```



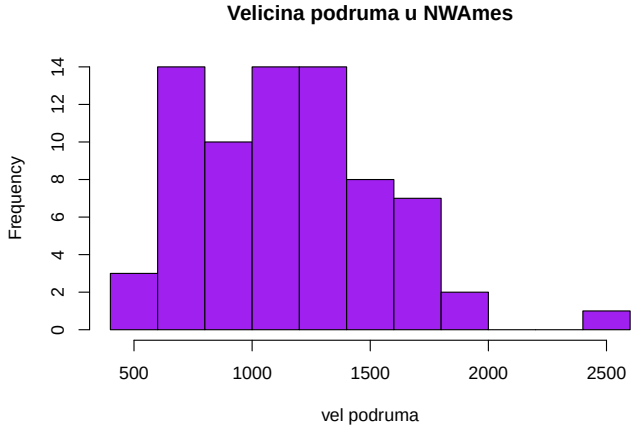


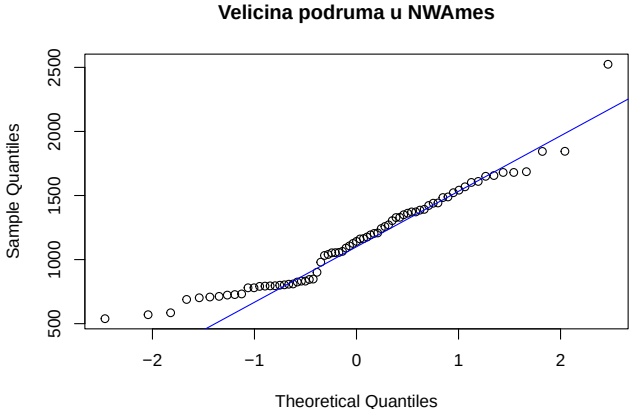
```
data_re16 <- data[data$Neighborhood == c("NridgHt"),]
hist(data_re16$TotalBsm,
     main = "Velicina podruma u NridgHt",
     col="purple", xlab="vel podruma")
qqnorm(data_re16$TotalBsmtSF, main="Velicina podruma u NridgHt")
qqline(data_re16$TotalBsmtS, col="blue")
```



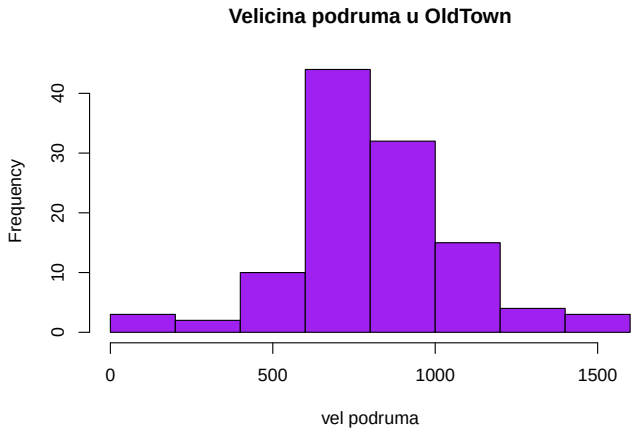


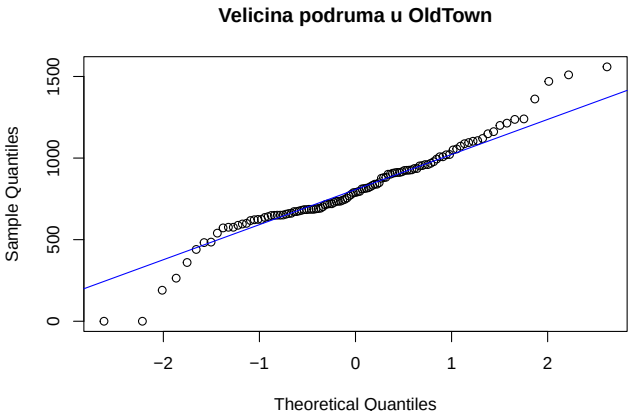

```
data_re17 <- data[data$Neighborhood == c("NWAmes"),]
hist(data_re17$TotalBsm,
     main = "Velicina podruma u NWAmes",
     col="purple", xlab="vel podruma")
qqnorm(data_re17$TotalBsmtSF, main="Velicina podruma u NWAmes")
qqline(data_re17$TotalBsmtS, col="blue")
```



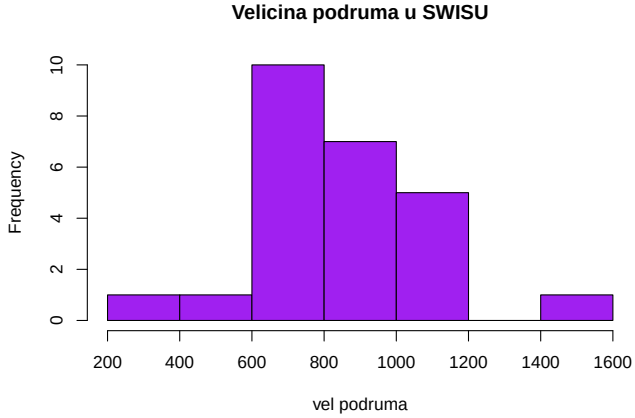


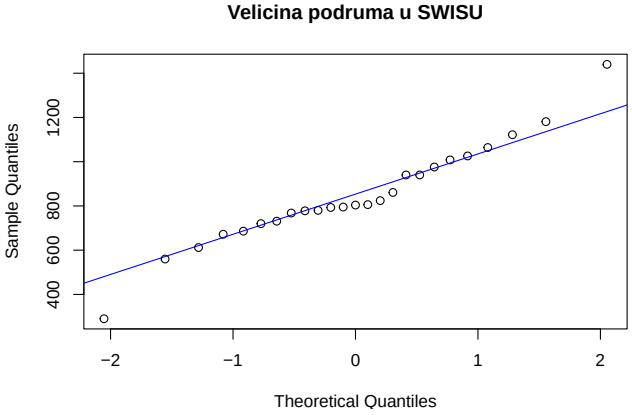

```
data_re18 <- data[data$Neighborhood == c("OldTown"),]
hist(data_re18$TotalBsm,
     main = "Velicina podruma u OldTown",
     col="purple", xlab="vel podruma")
qqnorm(data_re18$TotalBsmtSF, main="Velicina podruma u OldTown")
qqline(data_re18$TotalBsmtS, col="blue")
```



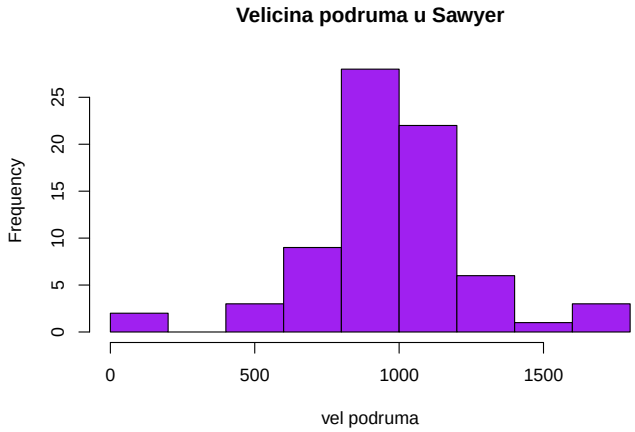


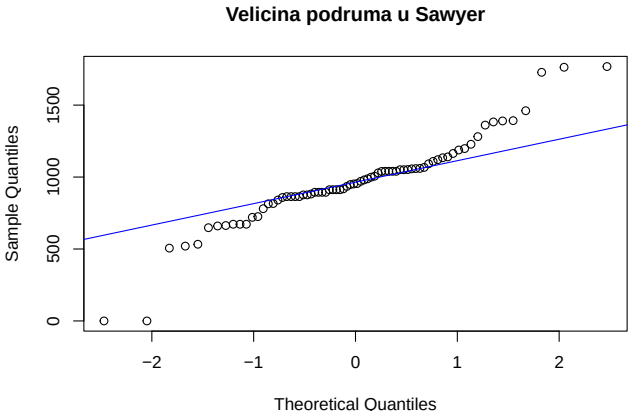

```
data_re19 <- data[data$Neighborhood == c("SWISU"),]
hist(data_re19$TotalBsm,
     main = "Velicina podruma u SWISU",
     col="purple", xlab="vel podruma")
qqnorm(data_re19$TotalBsmtSF, main="Velicina podruma u SWISU")
qqline(data_re19$TotalBsmtS, col="blue")
```



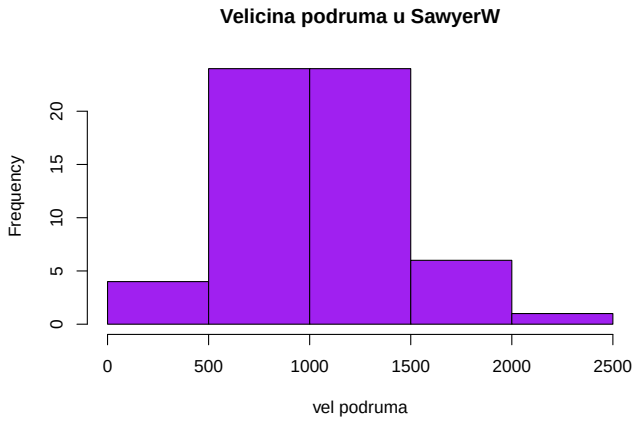


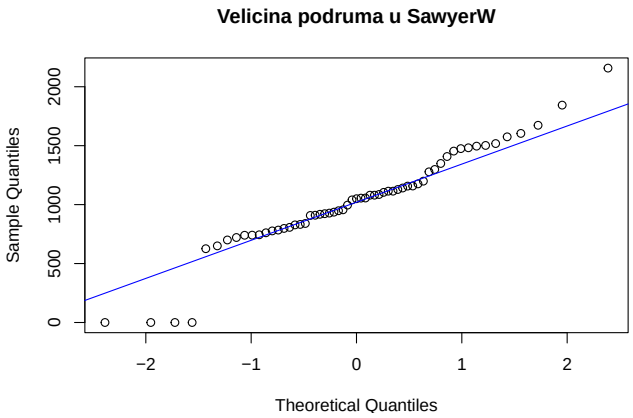

```
data_re20 <- data[data$Neighborhood == c("Sawyer"),]
hist(data_re20$TotalBsm,
     main = "Velicina podruma u Sawyer",
     col="purple", xlab="vel podruma")
qqnorm(data_re20$TotalBsmtSF, main="Velicina podruma u Sawyer")
qqline(data_re20$TotalBsmtS, col="blue")
```



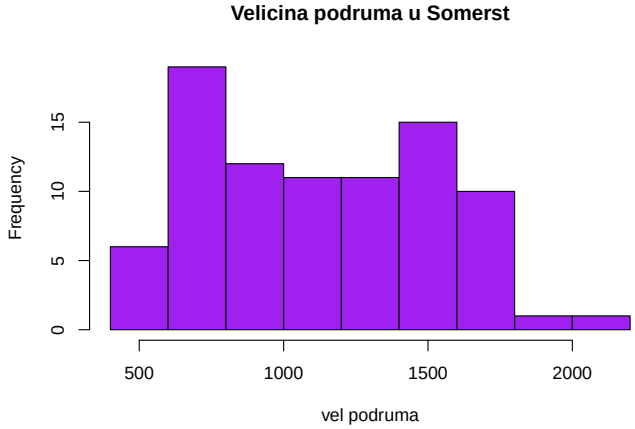


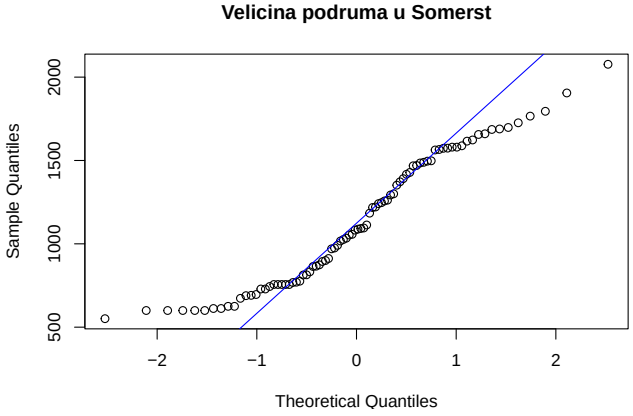
```
data_re21 <- data[data$Neighborhood == c("SawyerW"),]
hist(data_re21$TotalBsm,
     main = "Velicina podruma u SawyerW",
     col="purple", xlab="vel podruma")
qqnorm(data_re21$TotalBsmtSF, main="Velicina podruma u SawyerW")
qqline(data_re21$TotalBsmtS, col="blue")
```



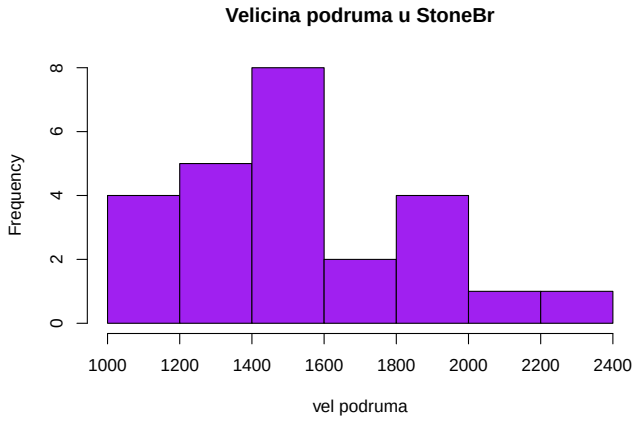


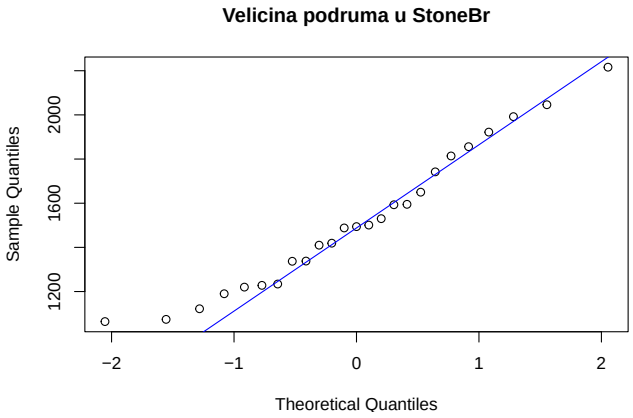

```
data_re22 <- data[data$Neighborhood == c("Somerst"),]
hist(data_re22$TotalBsm,
     main = "Velicina podruma u Somerst",
     col="purple", xlab="vel podruma")
qqnorm(data_re22$TotalBsmtSF, main="Velicina podruma u Somerst")
qqline(data_re22$TotalBsmtS, col="blue")
```



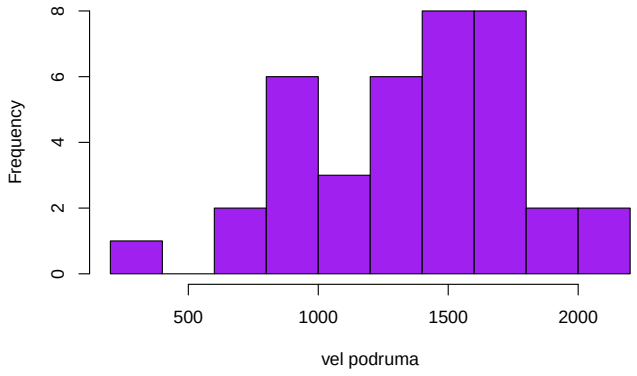


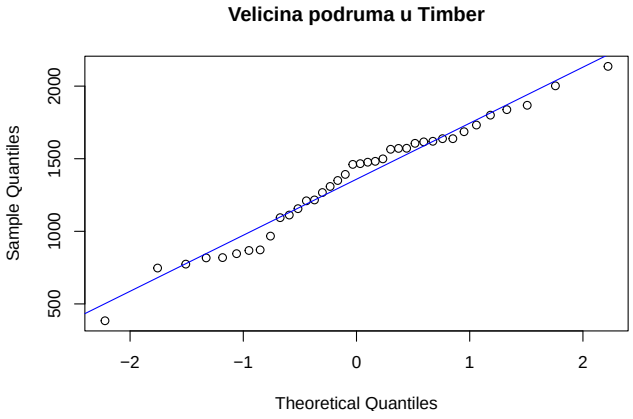

```
data_re23 <- data[data$Neighborhood == c("StoneBr"),]
hist(data_re23$TotalBsm,
     main = "Velicina podruma u StoneBr",
     col="purple", xlab="vel podruma")
qqnorm(data_re23$TotalBsmtSF, main="Velicina podruma u StoneBr")
qqline(data_re23$TotalBsmtS, col="blue")
```



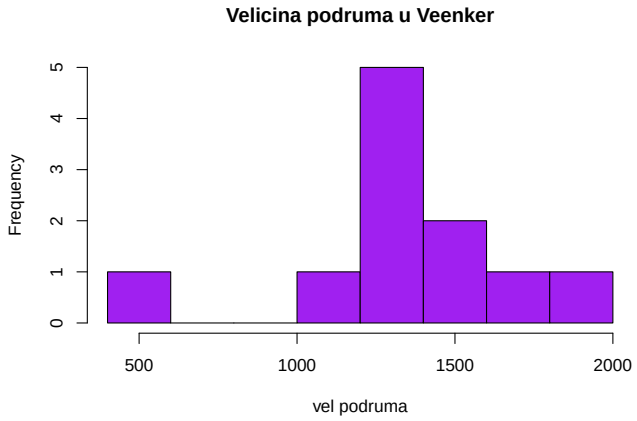


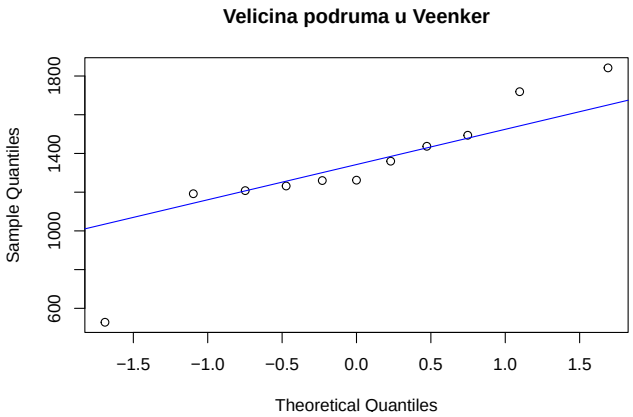

```
data_re24 <- data[data$Neighborhood == c("Timber"),]
hist(data_re24$TotalBsm,
     main = "Velicina podruma u Timber",
     col="purple", xlab="vel podruma")
qqnorm(data_re24$TotalBsmtSF, main="Velicina podruma u Timber")
qqline(data_re24$TotalBsmtS, col="blue")
```






```
data_re25 <- data[data$Neighborhood == c("Veenker"),]
hist(data_re25$TotalBsm,
     main = "Velicina podruma u Veenker",
     col="purple", xlab="vel podruma")
qqnorm(data_re25$TotalBsmtSF, main="Velicina podruma u Veenker")
qqline(data_re25$TotalBsmtS, col="blue")
```





Pretpostavke ANOVA testa su: populacije iz grupa međusobno su nezavisne i normalno distribuirane sa jednakim varijancama. Nezavisnost populacija teško možemo provjeriti stoga ćemo pretpostaviti da su one nezavisne.

U nastavku vidimo da se varijance populacija razlikuju.

```
var(na.omit(data_re1$TotalBsmtSF))
```

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

```
var(na.omit(data_re2$TotalBsmtSF))
```

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

```
var(na.omit(data_re3$TotalBsmtSF))
```

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

```
var(na.omit(data_re4$TotalBsmtSF))
```

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

```
var(na.omit(data_re5$TotalBsmtSF))
```

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

```
var(na.omit(data_re6$TotalBsmtSF))
```

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

```
var(na.omit(data_re7$TotalBsmtSF))
```

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

```
var(na.omit(data_re8$TotalBsmtSF))
```

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

```
var(na.omit(data_re9$TotalBsmtSF))
```

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

```
var(na.omit(data_re10$TotalBsmtSF))
```

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

```
var(na.omit(data_re11$TotalBsmtSF))
```

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

```
var(na.omit(data_re12$TotalBsmtSF))
```

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

```
var(na.omit(data_re13$TotalBsmtSF))
```

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

```
var(na.omit(data_re14$TotalBsmtSF))
```

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

```
var(na.omit(data_re15$TotalBsmtSF))
```

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

```
var(na.omit(data_re16$TotalBsmtSF))
```

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

```
var(na.omit(data_re17$TotalBsmtSF))
```

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

```
var(na.omit(data_re18$TotalBsmtSF))
```

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

```
var(na.omit(data_re19$TotalBsmtSF))
```

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

```
var(na.omit(data_re20$TotalBsmtSF))
```

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

```
var(na.omit(data_re21$TotalBsmtSF))
```

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

```
var(na.omit(data_re22$TotalBsmtSF))
```

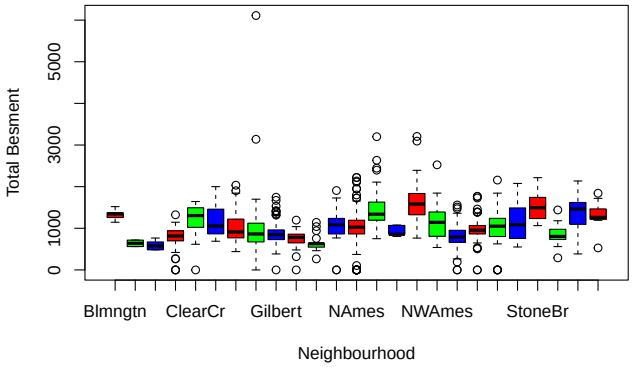
```
## [1] 155278.7
```

```
var(na.omit(data_re23$TotalBsmtSF))
```

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

U nastavku je prikazan pravokutni dijagram za sve grupe.

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



Pretpostavljamo da su sredine svih grupa jednake te uz gore navedene pretpostavke provodimo ANOVA test o jednakosti sredina. Nulta hipoteza je da su sredine za sve grupe jednake, a alternativna hipoteza je da se razlikuju.

```
res.aov <- aov( TotalBsmtSF~ factor(data$Neighborhood), data = data)
summary(res.aov)
```

```
##                               Df    Sum Sq Mean Sq F value Pr(>F)
## factor(data$Neighborhood)   24  74574228 3107259   21.62 <2e-16 ***
```

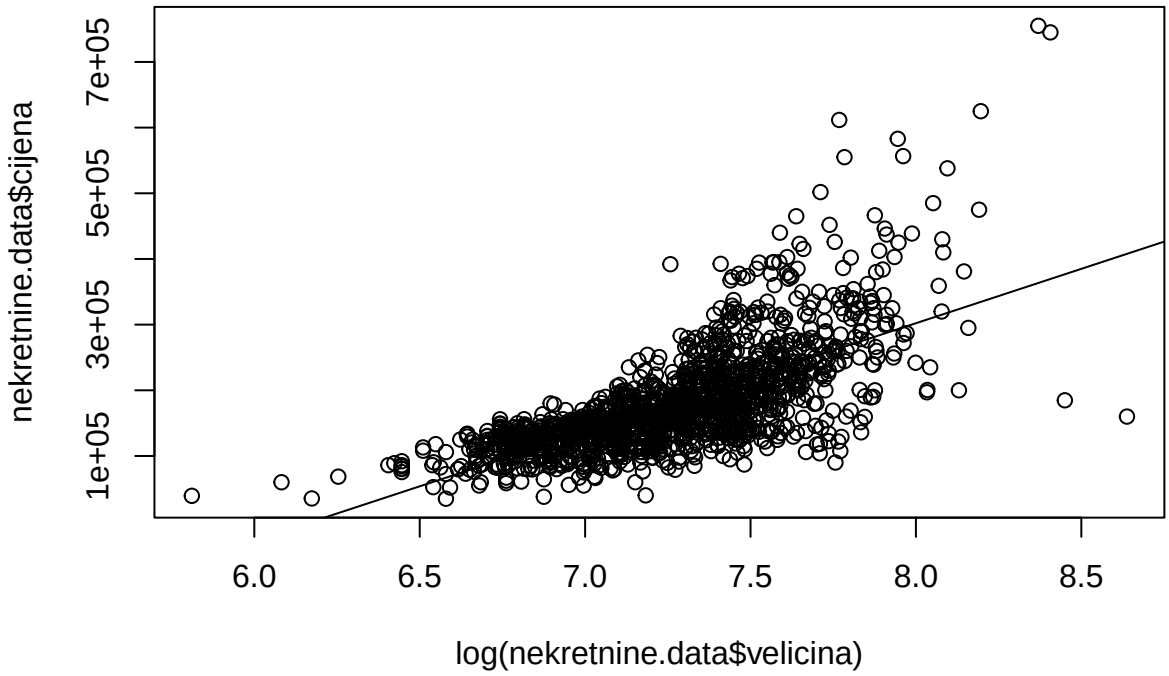
```
## Residuals                 1435 206228358  143713
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Iz rezultata ANOVA testa možemo zaključiti da sredine tih uzoraka nisu jednake te odbaciti nultu hipotezu u korist tvrdnje da su sredine različite. ANOVA nam samo govori da su sredine tih kategorija međusobno različite.

Ovisnost cijene o veličini nekretnine:

Provjeravamo linearnu zavisnost veličine nekretnine i cijene. Intuitivno bi se dalo naslutiti da će veće nekretnine imati veću cijenu.

```
nekretnine.data = data[,c("GrLivArea","SalePrice")]
colnames(nekretnine.data) = c("velicina", "cijena")
nekretnine.data = na.omit(nekretnine.data)

log_velicina = log(nekretnine.data$velicina)

plot(log(nekretnine.data$velicina), nekretnine.data$cijena)

fit.velicine = lm(nekretnine.data$cijena~log_velicina)

abline(fit.velicine)
```




```
summary(fit.velicine)
```

```
##
## Call:
## lm(formula = nekretnine.data$cijena ~ log_velicina)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -247772  -31767   -1680   24759  391583
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1022316      32624  -31.34   <2e-16 ***
## log_velicina   165558       4484   36.92   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 57130 on 1458 degrees of freedom
## Multiple R-squared:  0.4832, Adjusted R-squared:  0.4828
## F-statistic:  1363 on 1 and 1458 DF,  p-value: < 2.2e-16
```

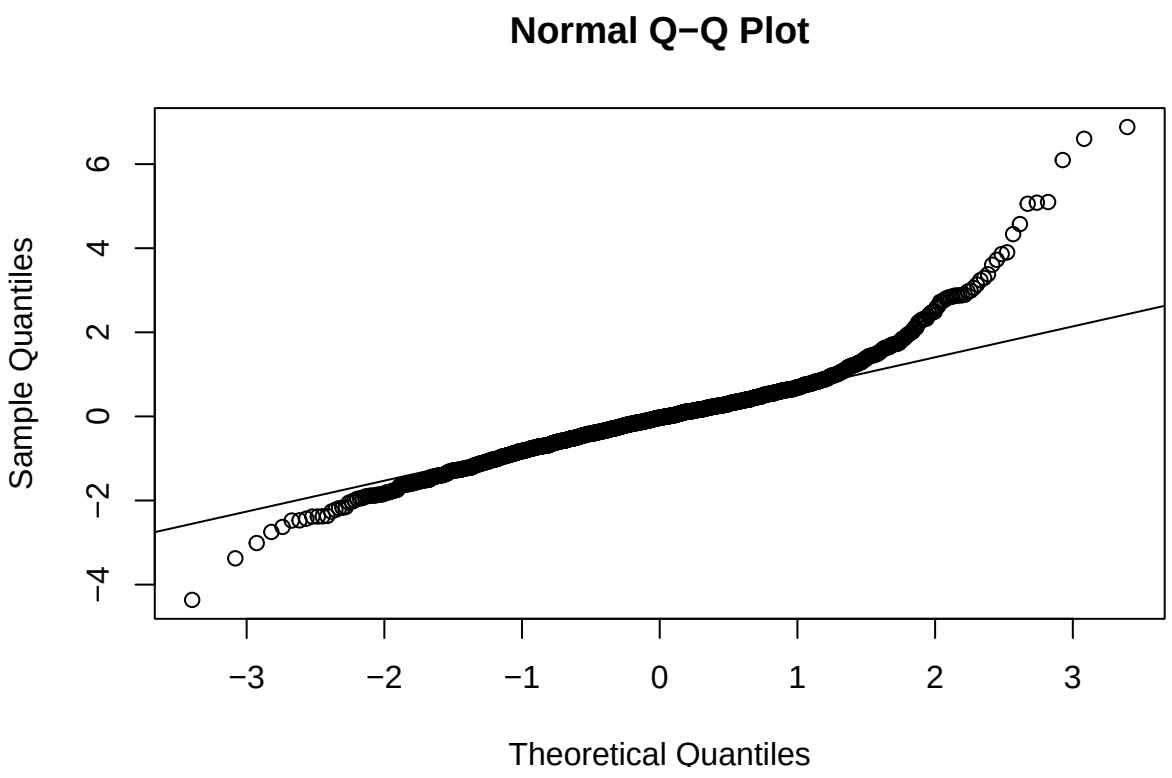
```
c("Pearson", cor(log_velicina, nekretnine.data$cijena, method = "pearson",
                 use = "complete.obs"))
```

```
## [1] "Pearson"           "0.695118068246328"
```

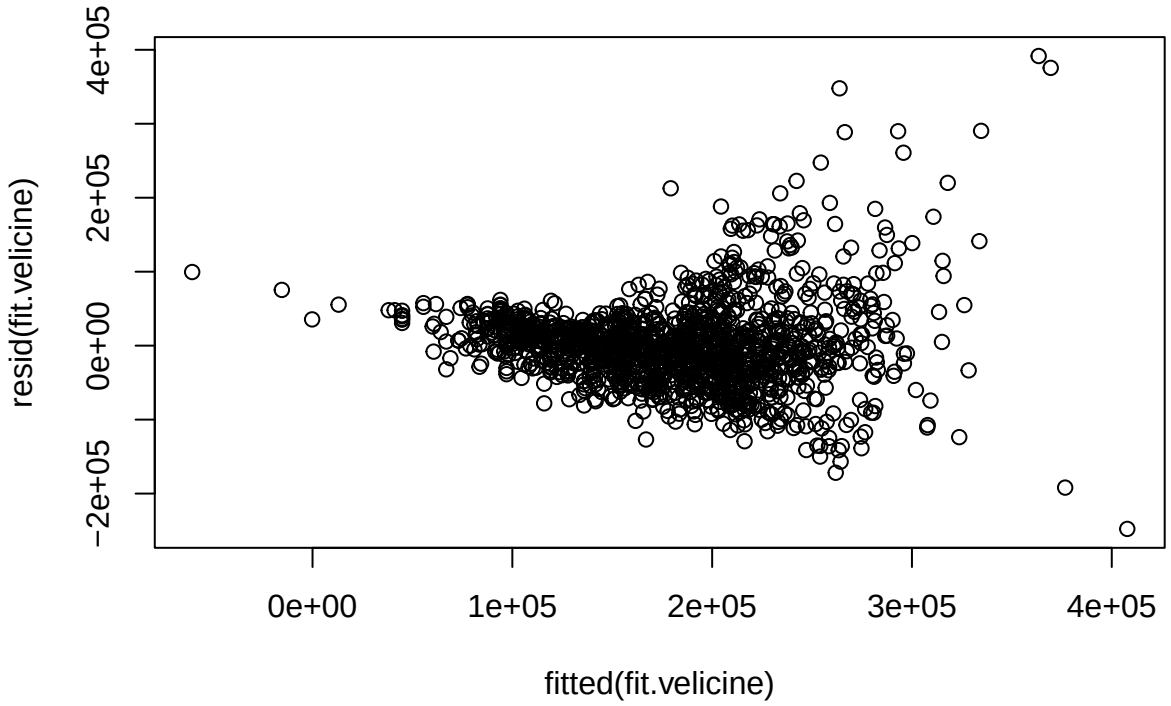
Prema Pearsonovom koeficijentu, vidimo da su varijable zavisne što je bilo i za očekivati.

U nastavku prikazujemo Q-Q plot i graf reziduala

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




```
plot(fitted(fit.velicine), resid(fit.velicine))
```




```
shapiro.test(data$GrLivArea)#odbacujemo nultu hipotezu da su podaci normalni
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$GrLivArea
## W = 0.92798, p-value < 2.2e-16
```

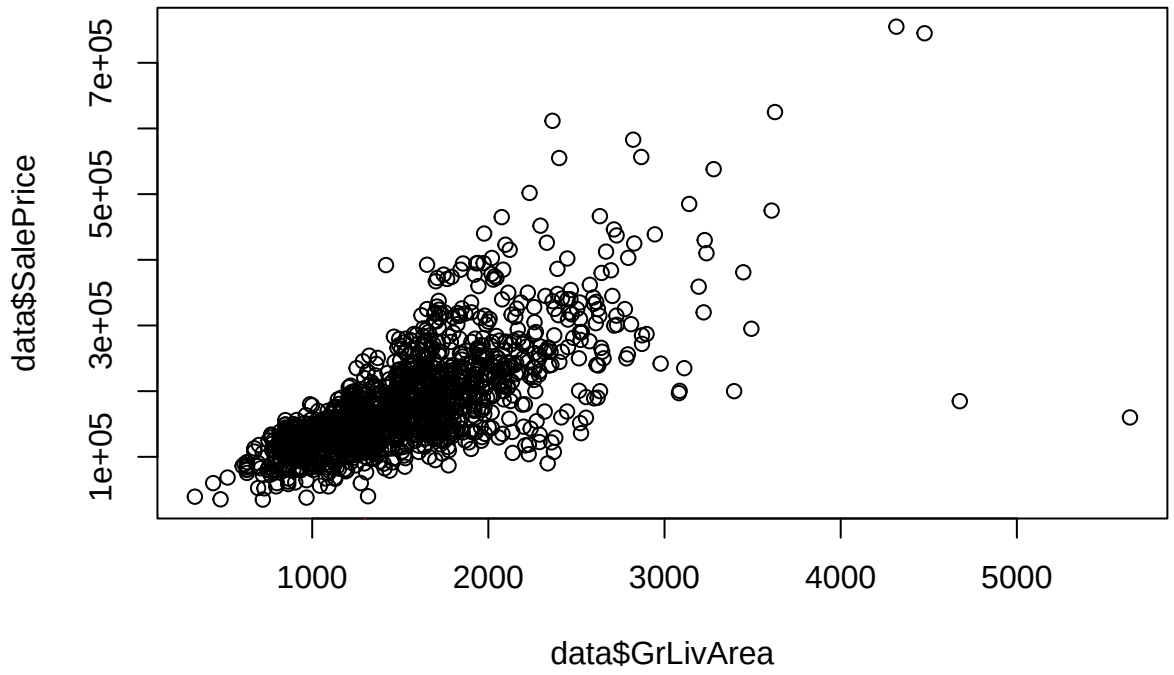
```
shapiro.test(data$SalePrice)#odbacujemo nultu hipotezu da su podaci normalni
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$SalePrice
## W = 0.86967, p-value < 2.2e-16
```

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

- veličina nekretnine
- godina izgradnje
- broj soba

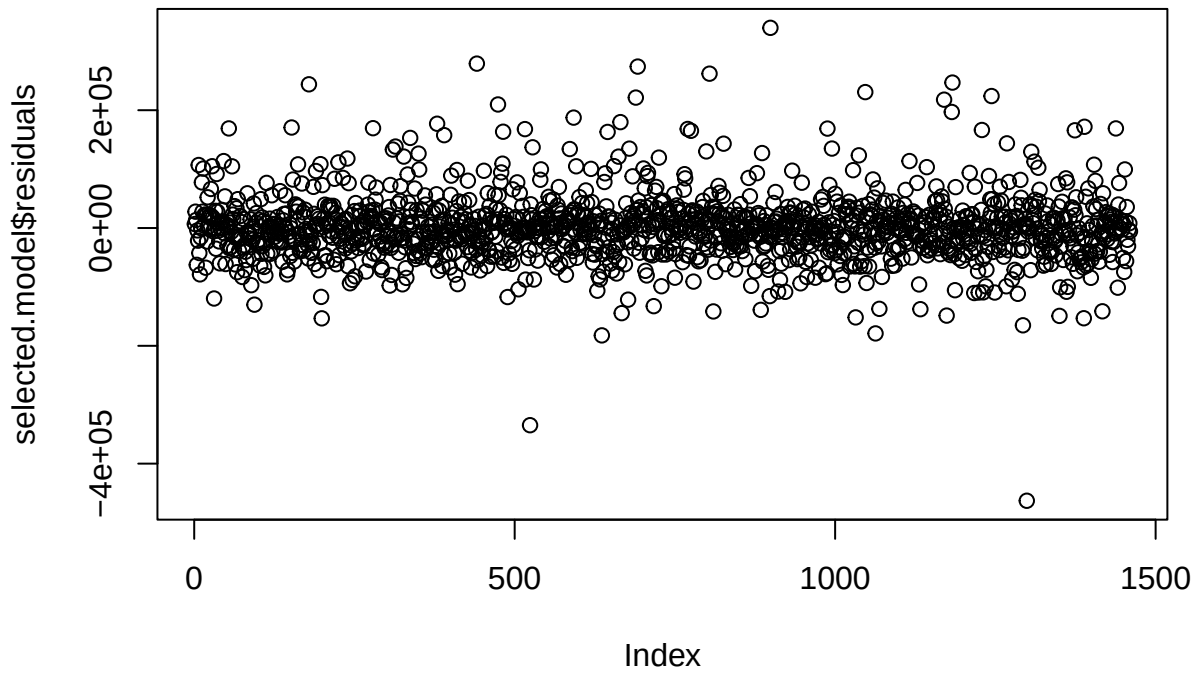
```
plot(data$GrLivArea,data$SalePrice) #kvadratura vs cijena

fit.livarea = lm(SalePrice~GrLivArea,data=data)

plot(data$GrLivArea,data$SalePrice)
lines(data$GrLivArea,fit.livarea$SalePrice,col='red')
```



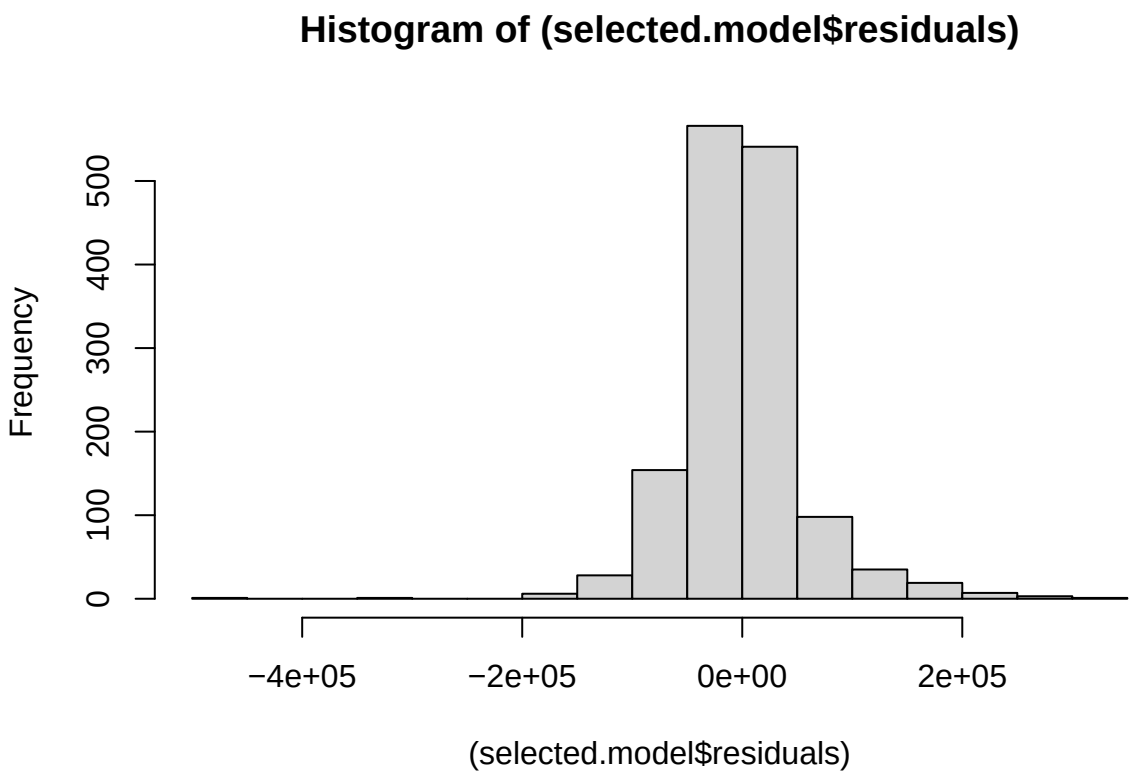
Normalnost reziduala i homogenost varijance

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

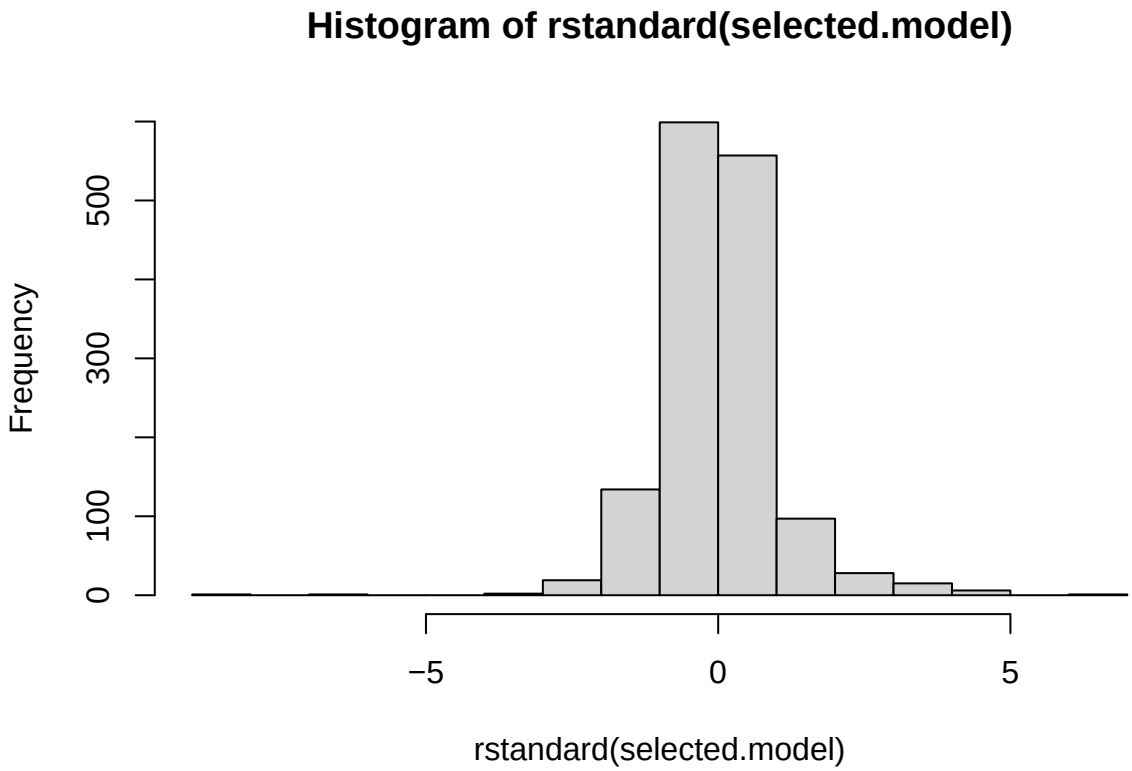
```
selected.model = fit.livarea
plot(selected.model$residuals)
```



```
hist((selected.model$residuals))
```

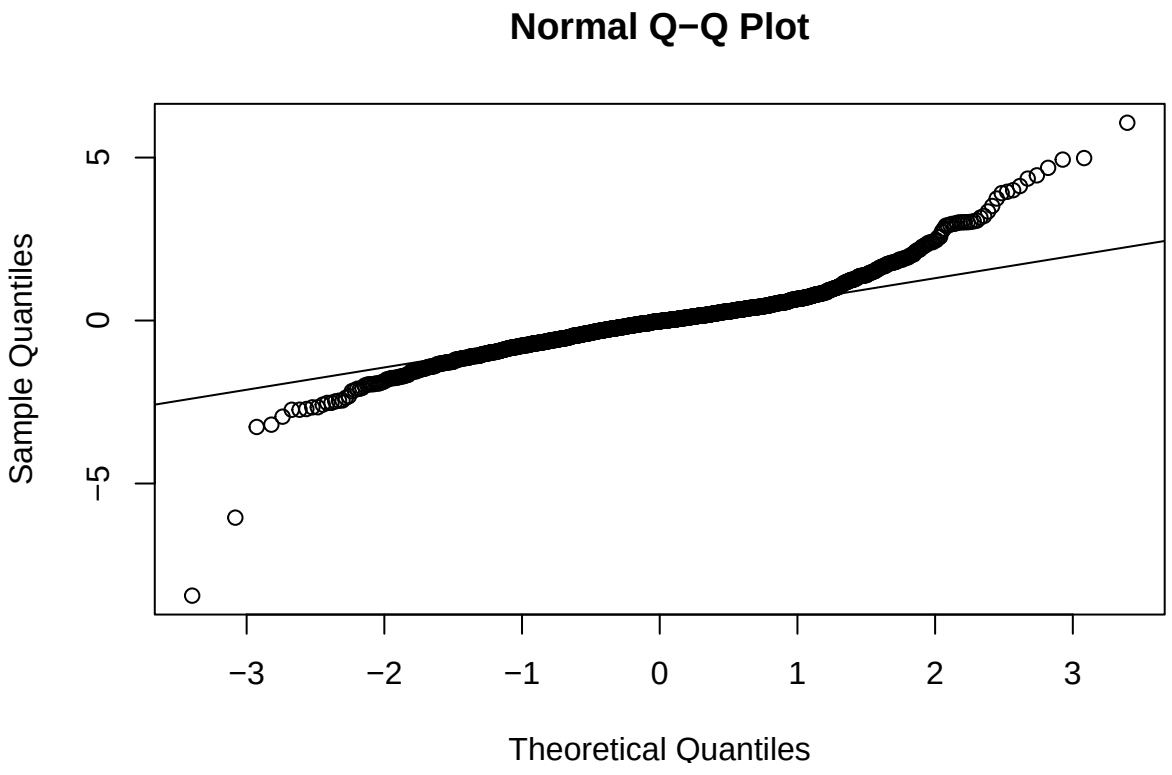



```
hist(rstandard(selected.model))
```

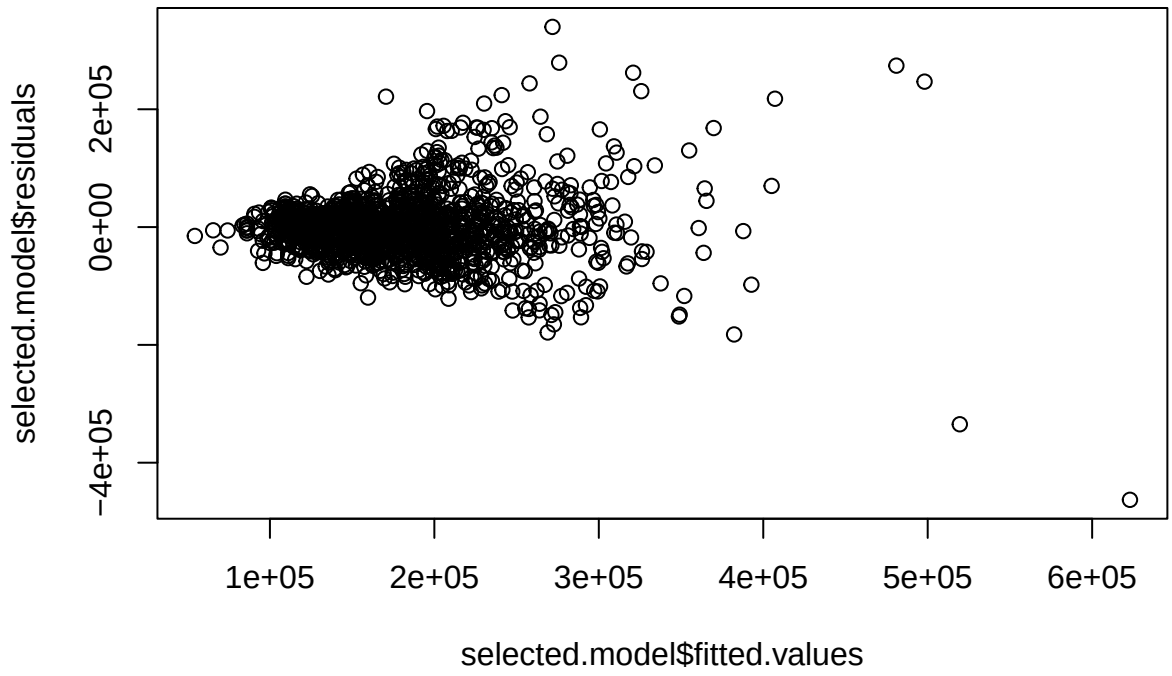


#q-q plot reziduala s linijom normalne distribucije

```
qqnorm(rstandard(selected.model))
qqline(rstandard(selected.model))
```




```
plot(selected.model$fitted.values,selected.model$residuals) #reziduale je dobro prikazati u ovisnosti o
```




```
#install.packages("nortest")
#library(nortest)
#require(nortest)
```

```
#lillie.test(rstandard(fit.livarea))
```

```
cor(data$YearBuilt,data$SalePrice)
```

```
## [1] 0.5228973
```

```
cor.test(data$YearBuilt,data$SalePrice)
```

```
##
##  Pearson's product-moment correlation
##
## data:  data$YearBuilt and data$SalePrice
## t = 23.424, df = 1458, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4845947 0.5591987
## sample estimates:
##       cor
## 0.5228973
```