Analiza trzista nekretnina_808

808: Borna Budimir-Bekan, Kristo Palić, Timoteja Piveta, Josipa Vujević

2023-01-15

```
r = getOption("repos")
r["CRAN"] = "http://cran.us.r-project.org"
options(repos = r)
library(dplyr)
```

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

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

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

```
library(tinytex)
#install.packages("aov", repos = "http://cran.us.r-project.org")
#install.packages("car", repos = "http://cran.us.r-project.org")
```

1. Uvjetuje li broj spavaćih soba cijenu kvadrata nekretnine?

U ovom dijelu istražujemo imaju li stanovi različitog broja spavaćih soba statistički značajno različitu cijenu kvadrtata.

```
data <- read.csv("preprocessed_data.csv", header = T, sep = ',')

# it will remove first column (unique index - X)
head(data)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         60       RL          65    8450   Pave  <NA>      Reg         Lvl
## 2  2         20       RL          80    9600   Pave  <NA>      Reg         Lvl
## 3  3         60       RL          68   11250   Pave  <NA>      IR1         Lvl
## 4  4         70       RL          60    9550   Pave  <NA>      IR1         Lvl
## 5  5         60       RL          84   14260   Pave  <NA>      IR1         Lvl
## 6  6         50       RL          85   14115   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
```

```
## 1    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 2    AllPub       FR2       Gtl      Veenker      Feedr       Norm     1Fam
## 3    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 4    AllPub    Corner       Gtl      Crawfor       Norm       Norm     1Fam
## 5    AllPub       FR2       Gtl      NoRidge       Norm       Norm     1Fam
## 6    AllPub    Inside       Gtl      Mitchel       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     2Story           7           5      2003         2003     Gable  CompShg
## 2     1Story           6           8      1976         1976     Gable  CompShg
## 3     2Story           7           5      2001         2002     Gable  CompShg
## 4     2Story           7           5      1915         1970     Gable  CompShg
## 5     2Story           8           5      2000         2000     Gable  CompShg
## 6     1.5Fin           5           5      1993         1995     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     VinylSd     VinylSd    BrkFace        196        Gd        TA      PConc
## 2     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 3     VinylSd     VinylSd    BrkFace        162        Gd        TA      PConc
## 4     Wd Sdng     Wd Shng       None          0        TA        TA     BrkTil
## 5     VinylSd     VinylSd    BrkFace        350        Gd        TA      PConc
## 6     VinylSd     VinylSd       None          0        TA        TA       Wood
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           No          GLQ        706          Unf
## 2       Gd       TA           Gd          ALQ        978          Unf
## 3       Gd       TA           Mn          GLQ        486          Unf
## 4       TA       Gd           No          ALQ        216          Unf
## 5       Gd       TA           Av          GLQ        655          Unf
## 6       Gd       TA           No          GLQ        732          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       150         856    GasA        Ex          Y      SBrkr
## 2          0       284        1262    GasA        Ex          Y      SBrkr
## 3          0       434         920    GasA        Ex          Y      SBrkr
## 4          0       540         756    GasA        Gd          Y      SBrkr
## 5          0       490        1145    GasA        Ex          Y      SBrkr
## 6          0        64         796    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1       856       854            0      1710            1            0        2
## 2      1262         0            0      1262            0            1        2
## 3       920       866            0      1786            1            0        2
## 4       961       756            0      1717            1            0        1
## 5      1145      1053            0      2198            1            0        2
## 6       796       566            0      1362            1            0        1
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        1            3            1          Gd            8        Typ
## 2        0            3            1          TA            6        Typ
## 3        1            3            1          Gd            6        Typ
## 4        0            3            1          Gd            7        Typ
## 5        1            4            1          Gd            9        Typ
## 6        1            1            1          TA            5        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          0        <NA>     Attchd        2003          RFn          2
## 2          1          TA     Attchd        1976          RFn          2
## 3          1          TA     Attchd        2001          RFn          2
## 4          1          Gd     Detchd        1998          Unf          3
## 5          1          TA     Attchd        2000          RFn          3
```

```
## 6           0        <NA>     Attchd        1993          Unf          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1        548         TA         TA          Y          0          61
## 2        460         TA         TA          Y        298           0
## 3        608         TA         TA          Y          0          42
## 4        642         TA         TA          Y          0          35
## 5        836         TA         TA          Y        192          84
## 6        480         TA         TA          Y         40          30
##   EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1             0          0           0        0   <NA>  <NA>        <NA>
## 2             0          0           0        0   <NA>  <NA>        <NA>
## 3             0          0           0        0   <NA>  <NA>        <NA>
## 4           272          0           0        0   <NA>  <NA>        <NA>
## 5             0          0           0        0   <NA>  <NA>        <NA>
## 6             0        320           0        0   <NA> MnPrv        Shed
##   MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1       0      2   2008       WD        Normal    208500
## 2       0      5   2007       WD        Normal    181500
## 3       0      9   2008       WD        Normal    223500
## 4       0      2   2006       WD       Abnorml    140000
## 5       0     12   2008       WD        Normal    250000
## 6     700     10   2009       WD        Normal    143000
```

Gledamo koji različiti brojevi spavaćih soba postoje te koliko je stanova u pojedinim određenim brojem.

```
n_distinct(unique(data$BedroomAbvGr))
```

```
## [1] 8
```

```
NumerOfBedrooms = unlist(data$BedroomAbvGr)
table(NumerOfBedrooms)
```

```
## NumerOfBedrooms
##   0   1   2   3   4   5   6   8
##   6  50 358 804 213  21   7   1
```

Vidimo da imamo 8 različitih kategorija stanova, od 0 do 8 spavaćih soba, bez 7. Zbog broja podatada različitih kategorija odlučujemo grupirati stanove sa 0 ili 1 sobom grupirat ćemo u kategoriju zvanu maxOne, a one sa 5, 6 ili 8 soba u kategoriju zvanu fiveSixEight.

S obzirom da se ovdje bavimo statističkim zaključivanjem na više od dva uzorka, koristit cemo ANOVA test.

ANOVA (ANalysis Of VAriance) je metoda kojom testiramo sredine više populacija. U analizi varijance pretpostavlja se da je ukupna varijabilnost u podatcima posljedica varijabilnosti podataka unutar svakog pojedine grupe (populacije) i varijabilnosti između različitih grupa. Varijabilnost unutar pojedinog uzorka je rezultat slučajnosti, a ukoliko postoje razlike u srednimana populacija, one će biti odražene u varijabilnosti među grupama. Jedan od glavnih ciljeva analize varijance je ustanoviti jesu li upravo te razlike između grupa samo posljedica slučajnosti ili je statistički značajna.

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

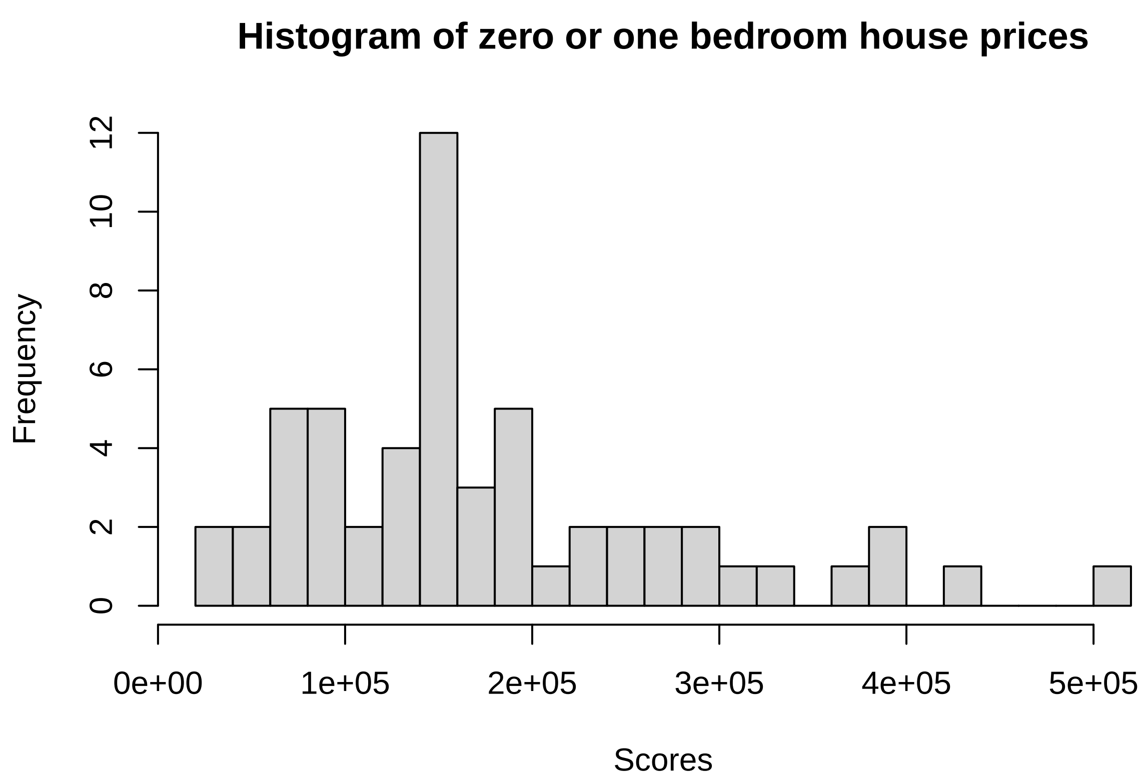
Nezavisnot podataka pretpostavljamo na temelju različitih uzoraka nad kojima se provodi ispitivanje, svaki uzorak reprezentiran je različitim brojem spavaćih soba.

Provjeru **normalnosti podataka** radit ćemo preko histograma, a testiranje **homogenosti varijance** uzoraka Bartletovim testom.

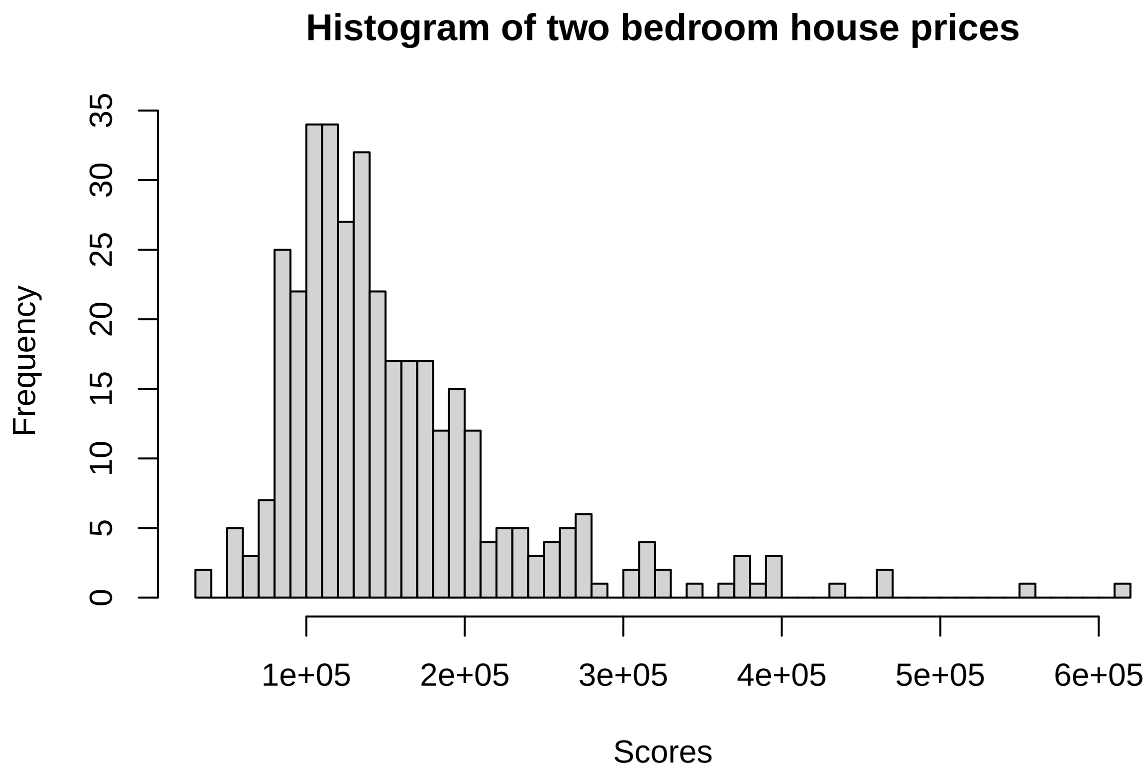
```
head(data$BedroomAbvGr)
```

```
## [1] 3 3 3 3 4 1
```

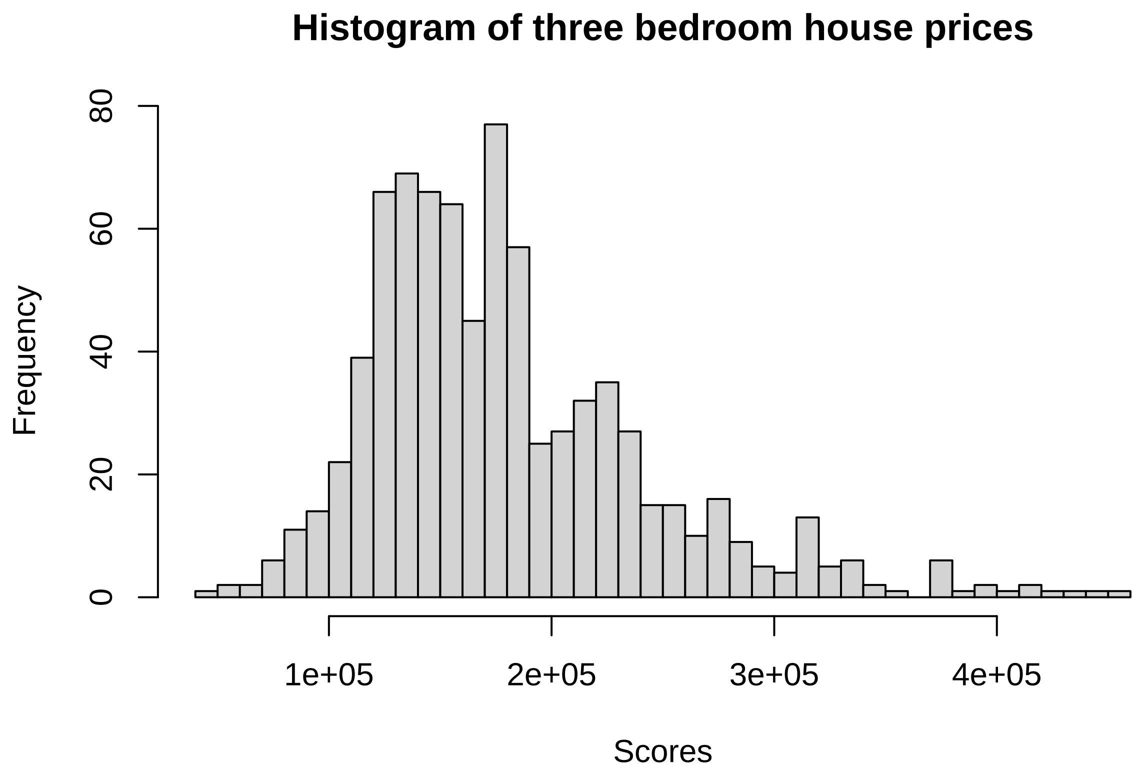
```
maxOne <- subset(data, data$BedroomAbvGr == 0 | data$BedroomAbvGr == 1)
two <- subset(data, data$BedroomAbvGr == 2)
three <- subset(data, data$BedroomAbvGr == 3)
four <- subset(data, data$BedroomAbvGr == 4)
fiveSixEight <- subset(data, data$BedroomAbvGr == 5 | data$BedroomAbvGr == 6 | data$BedroomAbvGr == 8)

hist(as.double(maxOne$SalePrice),
     breaks=25,
     main='Histogram of zero or one bedroom house prices',
     xlab='Scores')
```



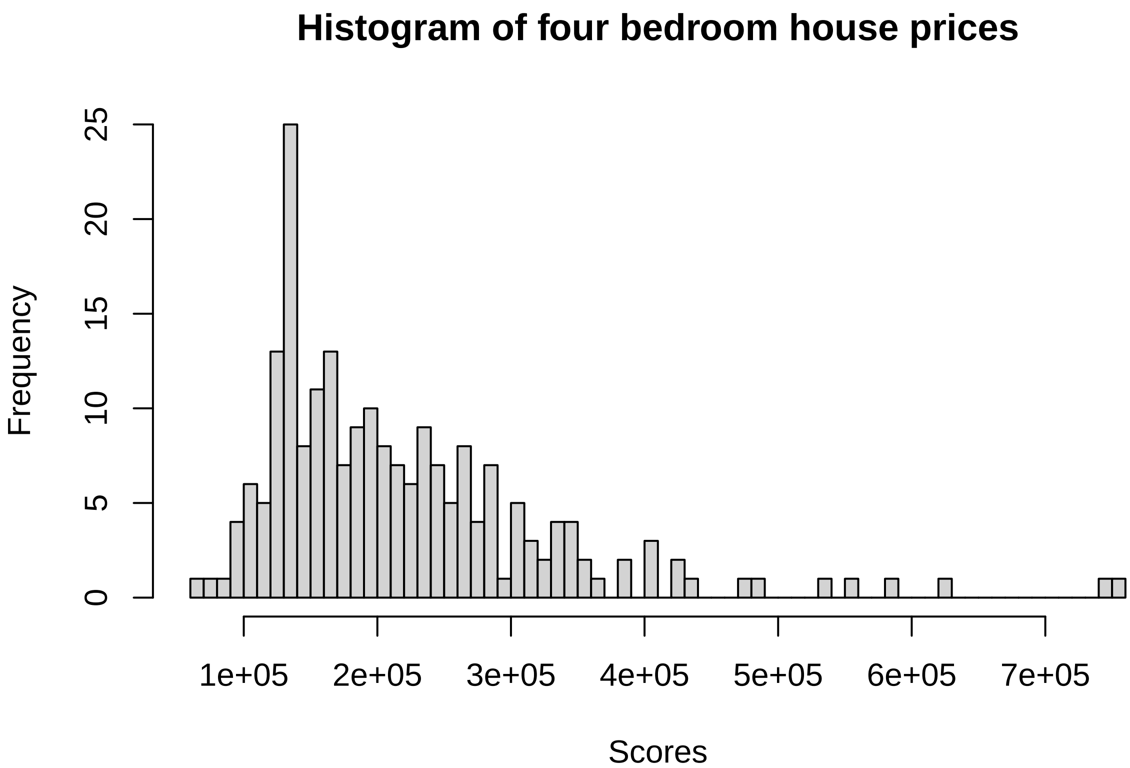

```
hist(as.double(two$SalePrice),
     breaks=50,
     main='Histogram of two bedroom house prices',
     xlab='Scores')
```




```r
hist(as.double(three$SalePrice),
     breaks=50,
     main='Histogram of three bedroom house prices',
     xlab='Scores')
```




```r
hist(as.double(four$SalePrice),
     breaks=50,
     main='Histogram of four bedroom house prices',
     xlab='Scores')
```




```r
hist(as.double(fiveSixEight$SalePrice),
     breaks=15,
     main='Histogram of five, six or eight bedroom house prices',
     xlab='Scores')
```

Razdiobe izgledaju normalno.

```r
df1 <- data.frame(group = 'maxOne', price = maxOne$SalePrice)
df2 <- data.frame(group = 'two', price = two$SalePrice)
df3 <- data.frame(group = 'three', price = three$SalePrice)
df4 <- data.frame(group = 'four', price = four$SalePrice)
df5 <- data.frame(group = 'fiveSixEight', price = fiveSixEight$SalePrice)

dataMerged = rbind(df1, df2, df3, df4, df5)
head(dataMerged)
```

```
##    group  price
## 1 maxOne 143000
## 2 maxOne  68500
## 3 maxOne 239686
## 4 maxOne 385000
## 5 maxOne 180000
## 6 maxOne 235000
```

Nadalje radimo provjeru homogenosti varijance:

Testiramo tezu H0: sve varijance su jednake dok alternativna hipoteza H1 opovrgava H0.

```r
bartlett.test(price ~ group, data = dataMerged)
```

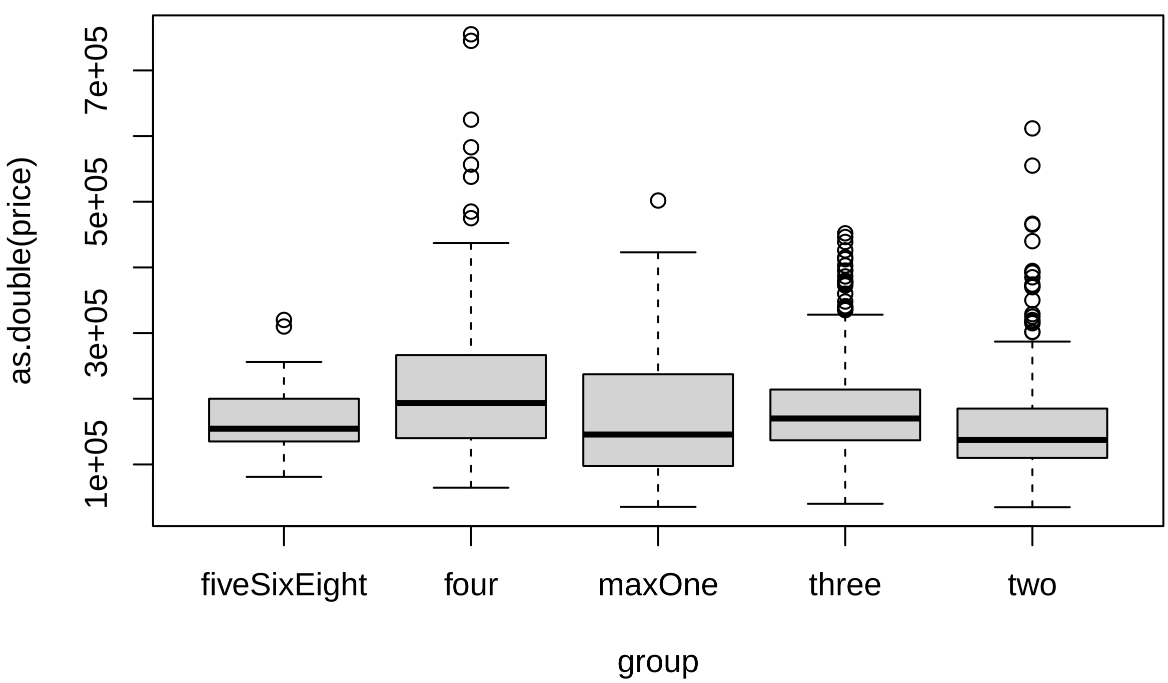
```
##
```

```
##  Bartlett test of homogeneity of variances
##
## data:  price by group
## Bartlett's K-squared = 131.54, df = 4, p-value < 2.2e-16
```

Rezultat testa nam daje p-vrijednost manju od 2.2e-16 što nam govori da je vjerojatnost da smo uočili takvu testnu statistiku da su varijance jednake uz istinitost H0, jako mala – dakle **odbacujemo hipotezu** $H0$ o tome da su varijance jednake.

Provjerimo postoje li razlike u cijenama za različiti broj spavaćih soba.

```
# Graficki prikaz podataka
boxplot(as.double(price) ~ group, data = dataMerged)
```




```
# Test
a = aov(price ~ group, data = dataMerged)
summary(a)
```

```
##                Df    Sum Sq   Mean Sq F value Pr(>F)    
## group           4 5.196e+11 1.299e+11   21.75 <2e-16 ***
## Residuals    1455 8.688e+12 5.971e+09                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Grafički prikaz sugerira da postoji razlika u cijenama među brojem spavaćih doba, što potvrđuje i ANOVA.

2. Određuje li oblik zemljišne čestice broj katova kuće?

Grupiramo podatke po obliku zemljišta i broju katova kuće u kontigencijsku tablicu u kojoj su retci brojevi katova kuće, a stupci oblik zemljišta. Katova ima jedan ili dva (ne brojimo podrum), a četiri su različita oblika zemljišta.

Nad tablicom koristimo hi-kvadrat test kako bismo dosli do zaključka odudaraju li očitane vrijednosti previše od očekivanih vrijednosti. Ukoliko vrijednosti ne odudaraju previše, varijable su homogene.

Testiramo tezu H0: varijable su homogene Alternativna hipoteza H1 opovrgava H0.

```
data <- read.csv("preprocessed_data.csv", header = T, sep = ',')

# radimo praznu 2 x 4 matricu
mat1 <- matrix(, nrow = 2, ncol = 4)

colnames(mat1) <- c("Reg", "IR1", "IR2", "IR3")
rownames(mat1) <- c(1, 2)

# imamo 4 lot shapea
lotShapes <- unique(data$LotShape)

# upisi u matricu
for (i in (1:length(lotShapes))) {
  mat1[1, i] = nrow(data[which(data$LotShape == unique(data$LotShape)[i] & data$X2ndFlrSF == 0),])
  mat1[2, i] = nrow(data[which(data$LotShape == unique(data$LotShape)[i] & data$X2ndFlrSF != 0),])
  print(lotShapes[i])
}
```

```
## [1] "Reg"
## [1] "IR1"
## [1] "IR2"
## [1] "IR3"
```

```
mat1
```

```
##   Reg IR1 IR2 IR3
## 1 523 284  17   5
## 2 402 200  24   5
```

```
chisq.test(mat1)
```

```
## Warning in chisq.test(mat1): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat1
## X-squared = 4.8387, df = 3, p-value = 0.184
```

P-vrijednost nije dovoljno mala da odbacimo H0, što znači da zaključujemo da su varijable homogene, odnosno da broj katova kuće ne ovisi o obliku zemljišta.

3. Ovisi li veličina podruma o kvartu u gradu?

Gledamo koji kvartovi postoje te u kolikom broju podataka se pojavljuju.

```
n_distinct(unique(data$Neighborhood))
```

```
## [1] 25
```

```
Neighborhood = unlist(data$Neighborhood)
table(Neighborhood)
```

```
## Neighborhood
## Blmngtn Blueste  BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert  IDOTRR
##      17       2      16      58      28     150      51     100      79      37
## MeadowV Mitchel   NAmes NoRidge NPkVill NridgHt  NWAmes OldTown  Sawyer SawyerW
##      17      49     225      41       9      77      73     113      74      59
## Somerst StoneBr   SWISU  Timber Veenker
##      86      25      25      38      11
```

Imamo 25 različitih kvartova. S obzirom da ne želimo grupirati kvartove, radit ćemo t-test nad svim parovima kvartova.

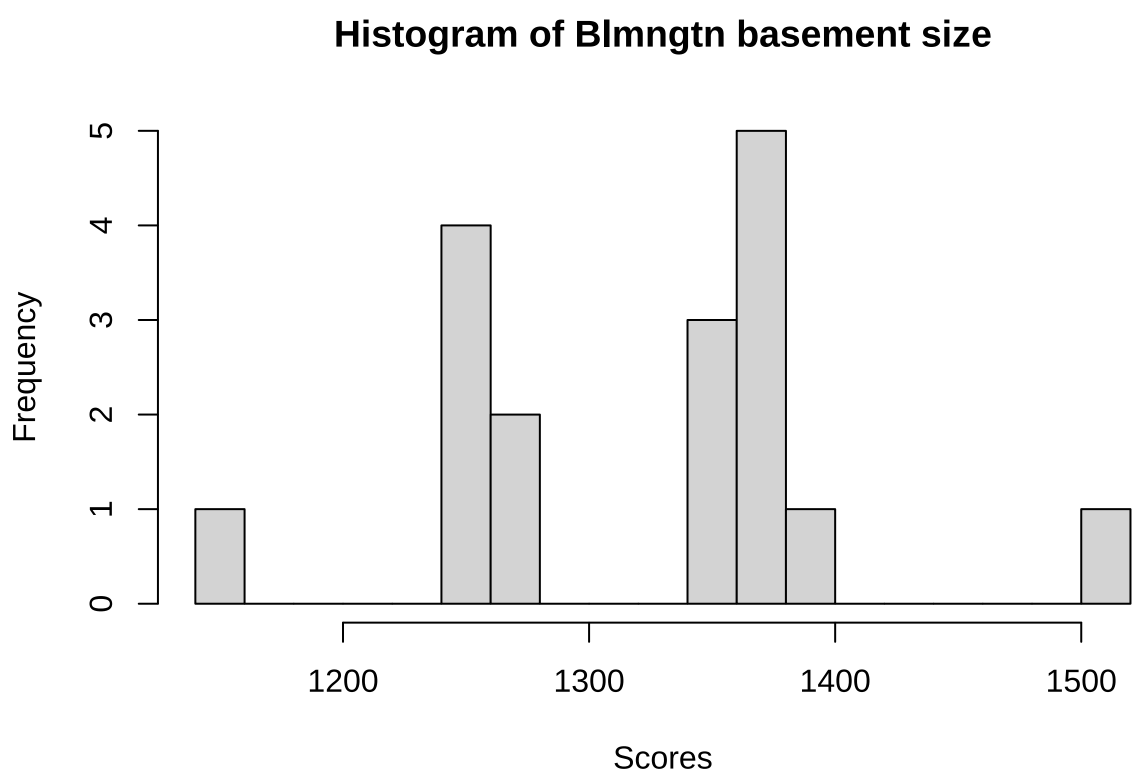
Prvo ćemo provjeriti neke početne značajke podataka, nezavisnost i normalnost podataka.

Nezavisnot podataka pretpostavljamo na temelju dvaju različitih uzoraka nad kojima se provodi ispitivanje, svaki uzorak pripada određenom kvartu.

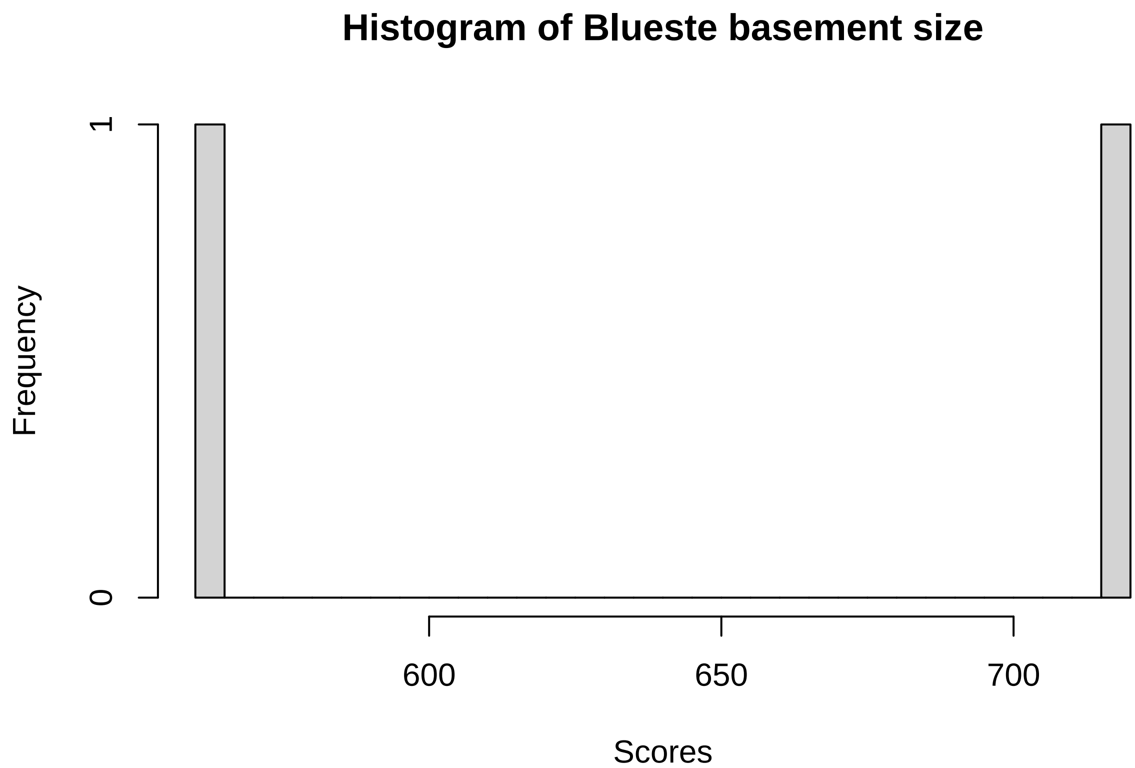
Nadalje ispitujemo **normalnost podataka** koju ćemo provjeriti pomoću histograma.

```
Blmngtn <- subset(data, data$Neighborhood == "Blmngtn")
Blueste <- subset(data, data$Neighborhood == "Blueste")
BrDale <- subset(data, data$Neighborhood == "BrDale")
BrkSide <- subset(data, data$Neighborhood == "BrkSide")
ClearCr <- subset(data, data$Neighborhood == "ClearCr")
CollgCr <- subset(data, data$Neighborhood == "CollgCr")
Crawfor <- subset(data, data$Neighborhood == "Crawfor")
Edwards <- subset(data, data$Neighborhood == "Edwards")
Gilbert <- subset(data, data$Neighborhood == "Gilbert")
IDOTRR <- subset(data, data$Neighborhood == "IDOTRR")
MeadowV <- subset(data, data$Neighborhood == "MeadowV")
Mitchel <- subset(data, data$Neighborhood == "Mitchel")
NAmes <- subset(data, data$Neighborhood == "NAmes")
NoRidge <- subset(data, data$Neighborhood == "NoRidge")
NPkVill <- subset(data, data$Neighborhood == "NPkVill")
NridgHt <- subset(data, data$Neighborhood == "NridgHt")
NWAmes <- subset(data, data$Neighborhood == "NWAmes")
OldTown <- subset(data, data$Neighborhood == "OldTown")
Sawyer <- subset(data, data$Neighborhood == "Sawyer")
SawyerW <- subset(data, data$Neighborhood == "SawyerW")
Somerst <- subset(data, data$Neighborhood == "Somerst")
StoneBr <- subset(data, data$Neighborhood == "StoneBr")
SWISU <- subset(data, data$Neighborhood == "SWISU")
Timber <- subset(data, data$Neighborhood == "Timber")
```

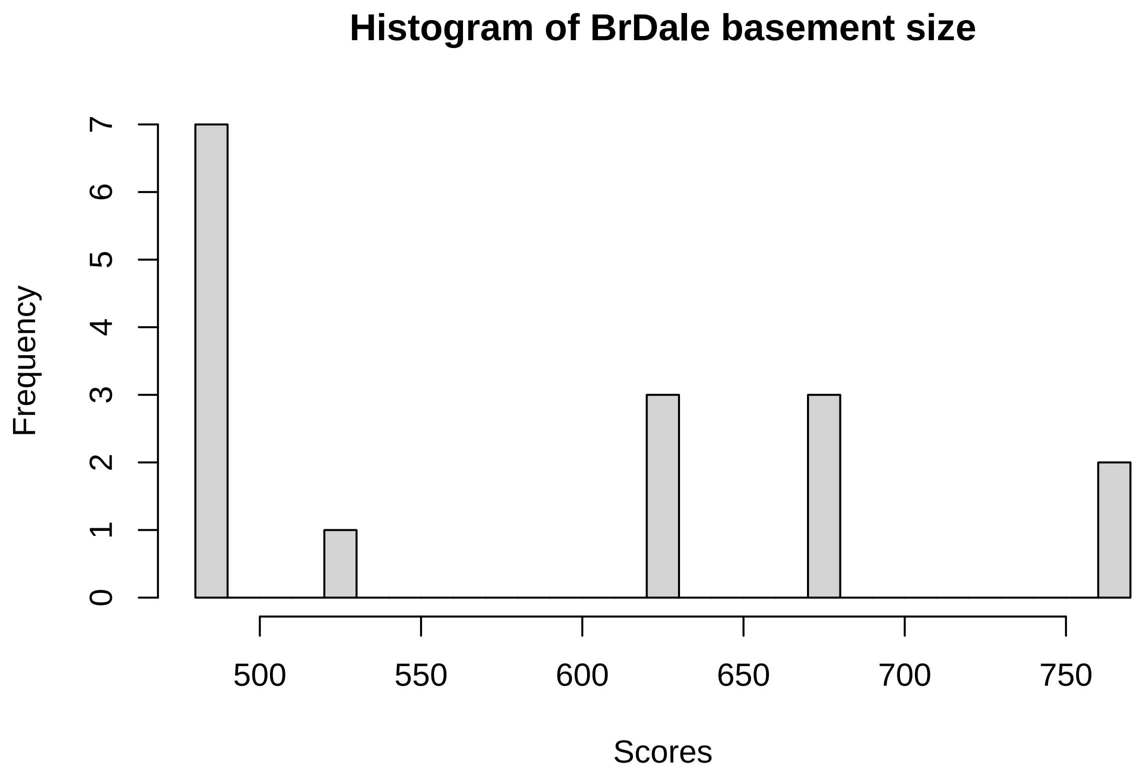
```
Veenker <- subset(data, data$Neighborhood == "Veenker")

hist(as.double(Blmngtn$TotalBsmtSF),
     breaks=25,
     main='Histogram of Blmngtn basement size',
     xlab='Scores')
```



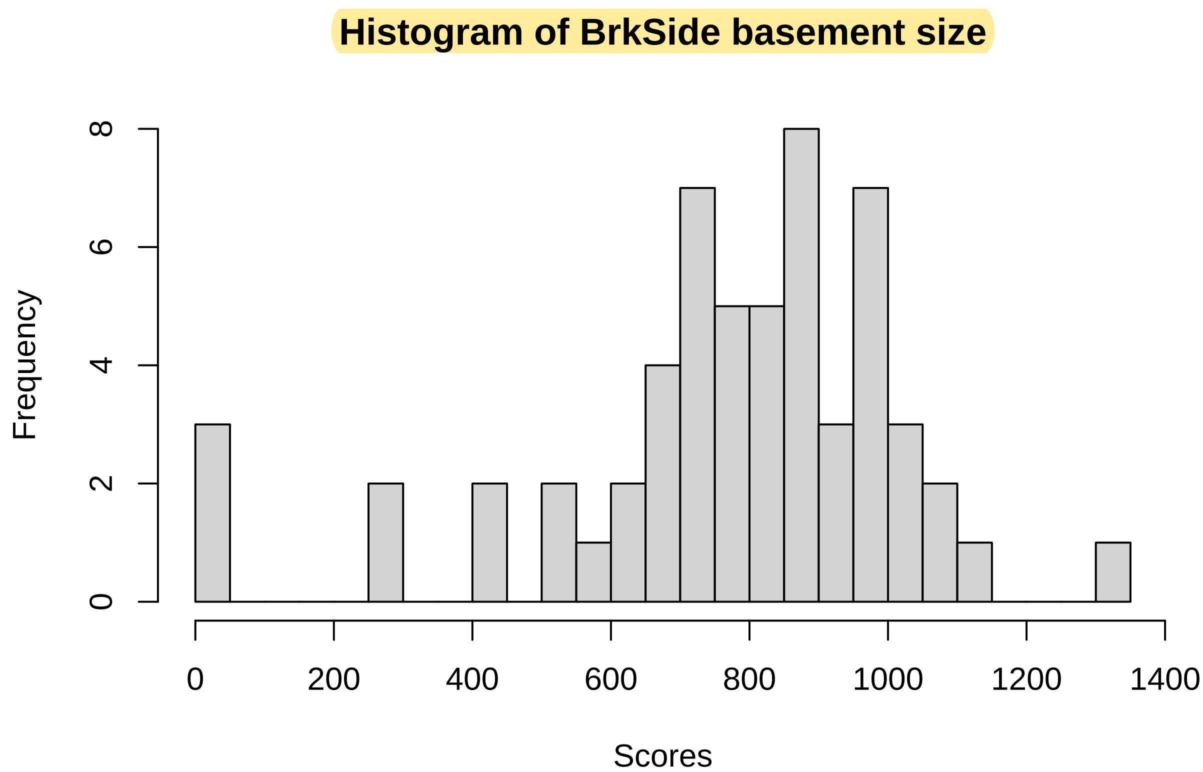

```
hist(as.double(Blueste$TotalBsmtSF),
     breaks=25,
     main='Histogram of Blueste basement size',
     xlab='Scores')
```



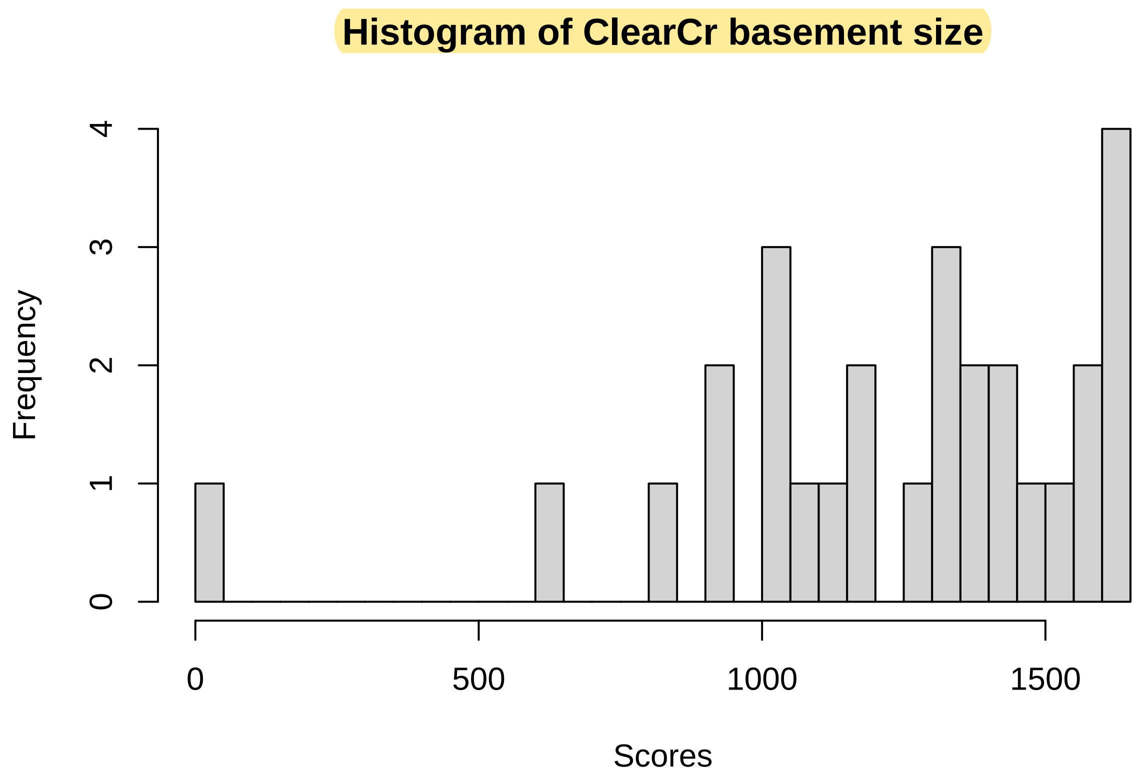

```r
hist(as.double(BrDale$TotalBsmtSF),
     breaks=25,
     main='Histogram of BrDale basement size',
     xlab='Scores')
```



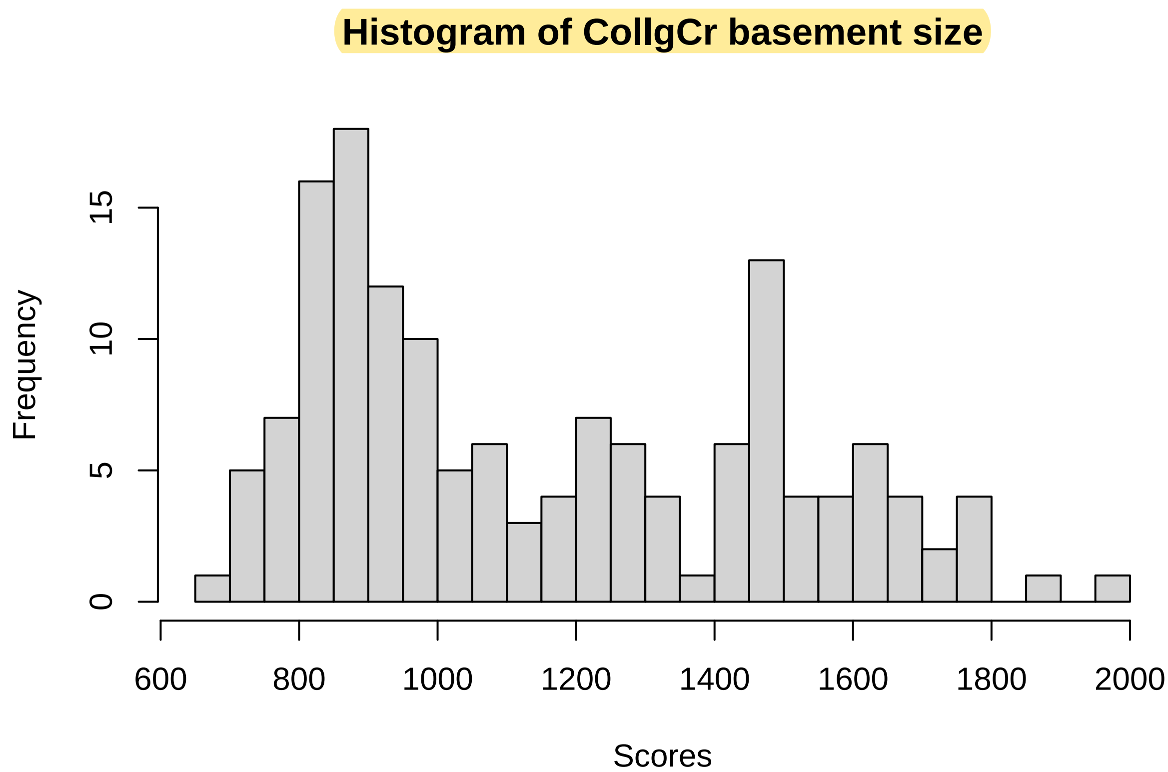

```
hist(as.double(BrkSide$TotalBsmtSF),
     breaks=25,
     main='Histogram of BrkSide basement size',
     xlab='Scores')
```



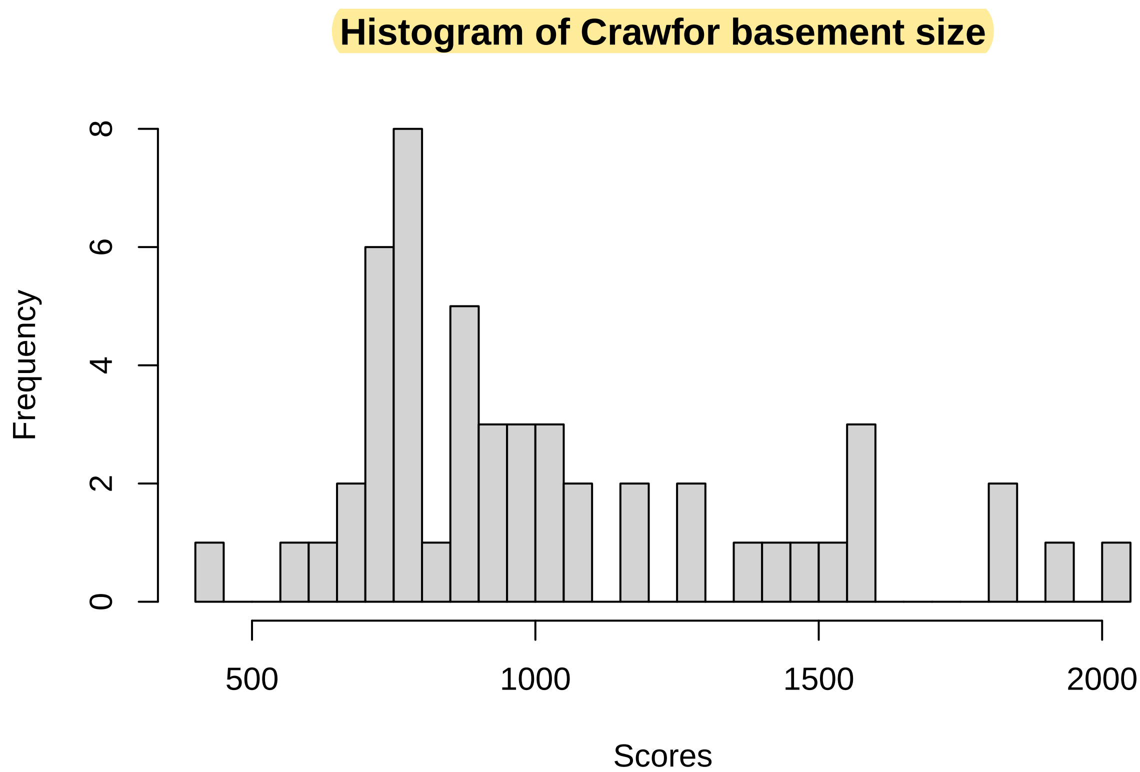

```
hist(as.double(ClearCr$TotalBsmtSF),
     breaks=25,
     main='Histogram of ClearCr basement size',
     xlab='Scores')
```



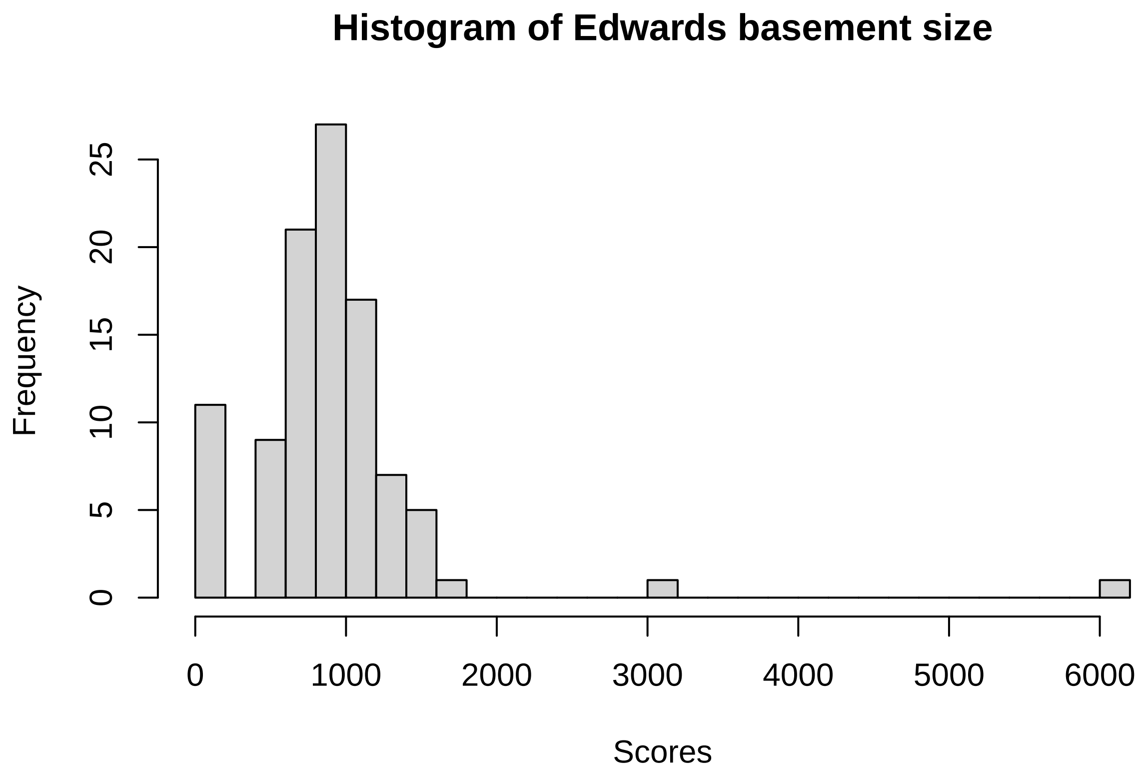

```
hist(as.double(CollgCr$TotalBsmtSF),
     breaks=25,
     main='Histogram of CollgCr basement size',
     xlab='Scores')
```



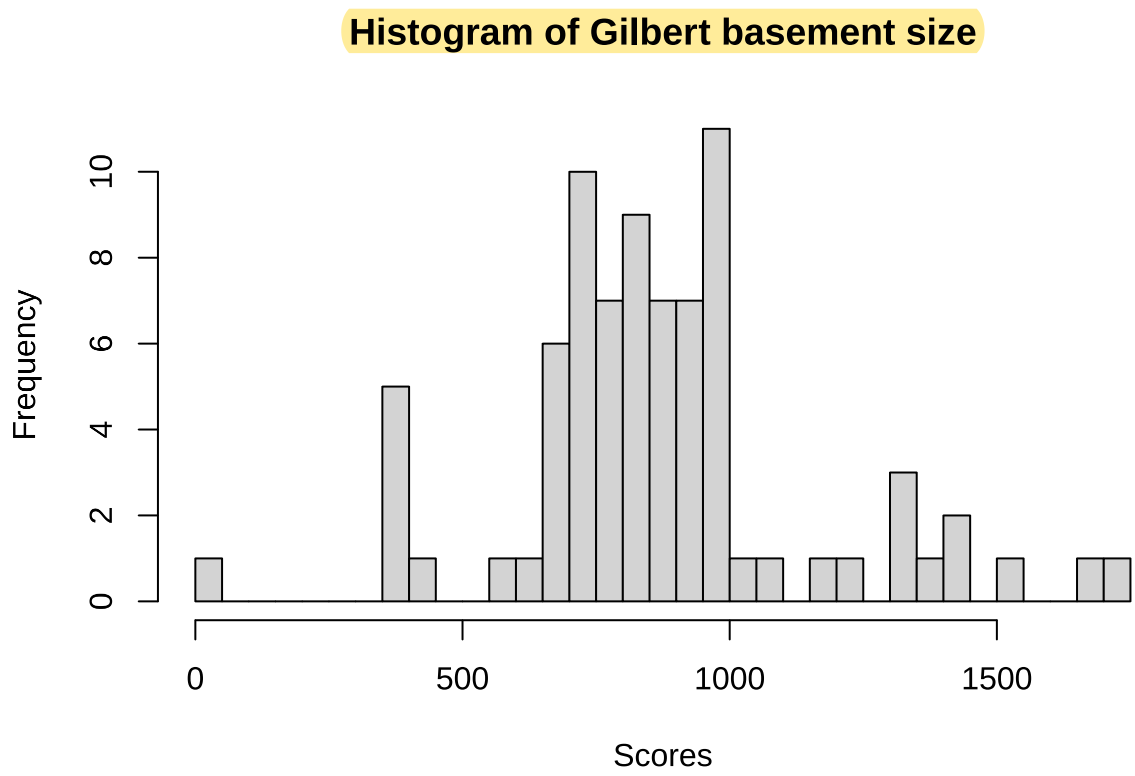

```
hist(as.double(Crawfor$TotalBsmtSF),
     breaks=25,
     main='Histogram of Crawfor basement size',
     xlab='Scores')
```



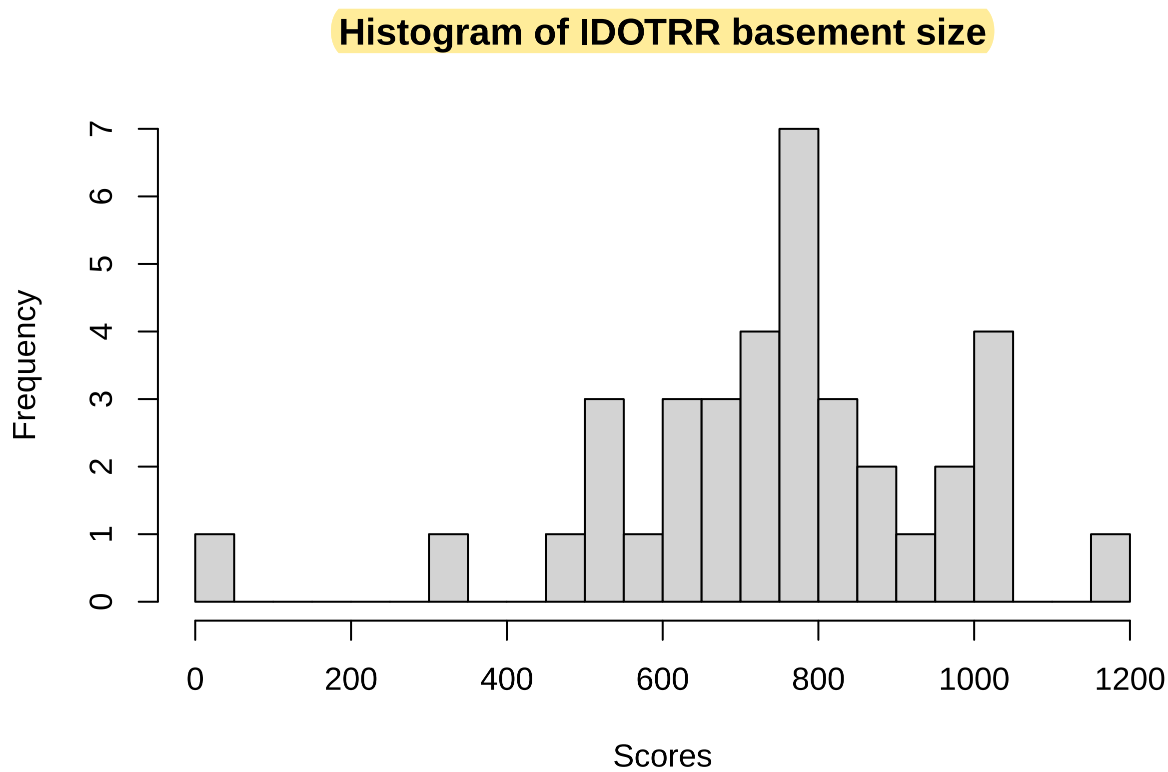

```
hist(as.double(Edwards$TotalBsmtSF),
     breaks=25,
     main='Histogram of Edwards basement size',
     xlab='Scores')
```



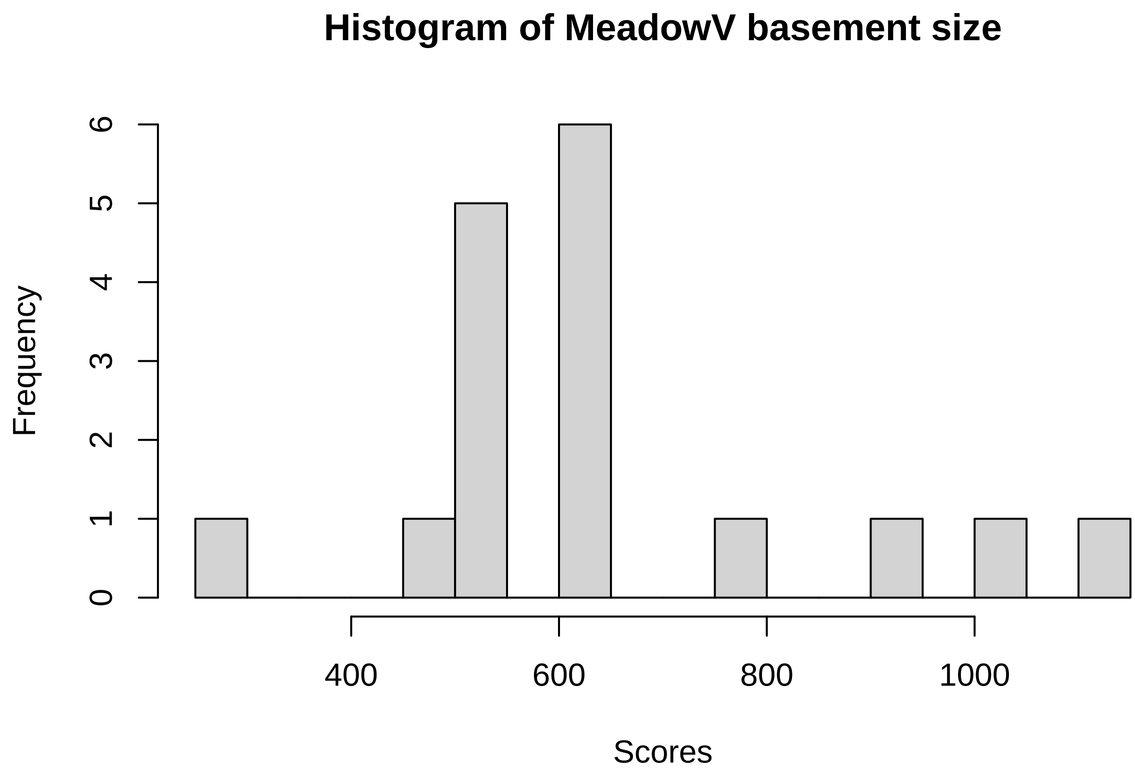

```
hist(as.double(Gilbert$TotalBsmtSF),
     breaks=25,
     main='Histogram of Gilbert basement size',
     xlab='Scores')
```



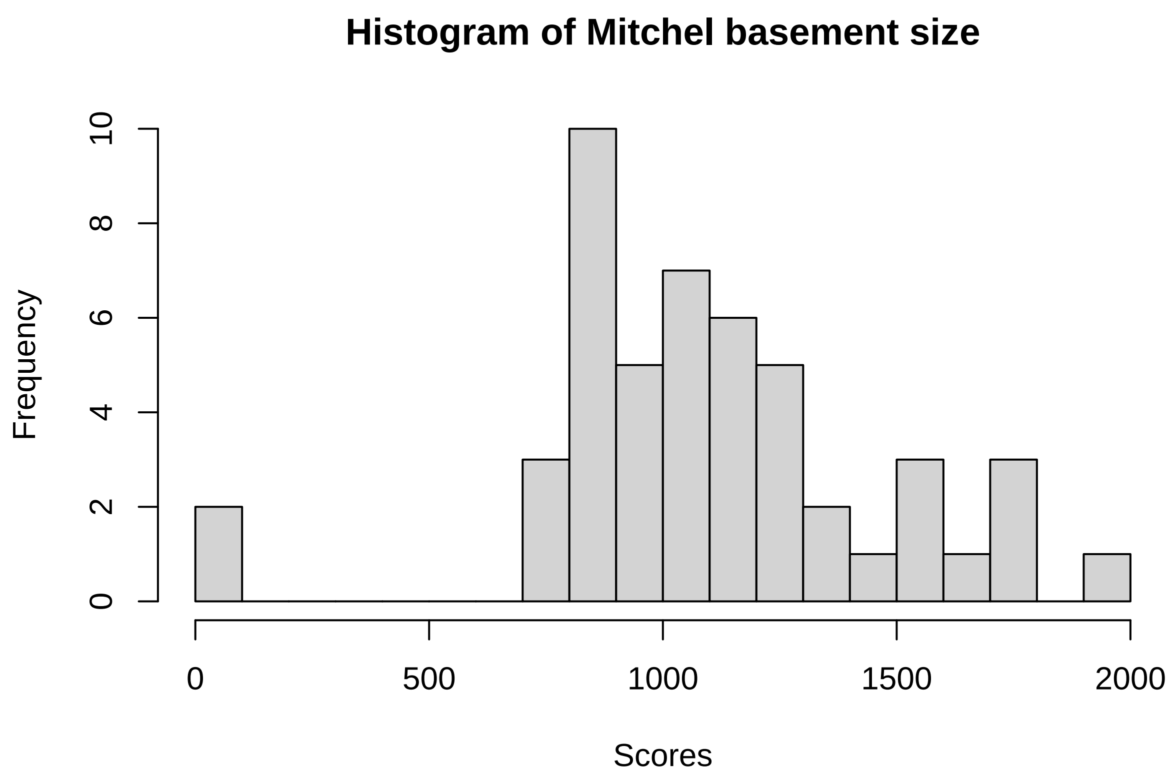

```
hist(as.double(IDOTRR$TotalBsmtSF),
     breaks=25,
     main='Histogram of IDOTRR basement size',
     xlab='Scores')
```



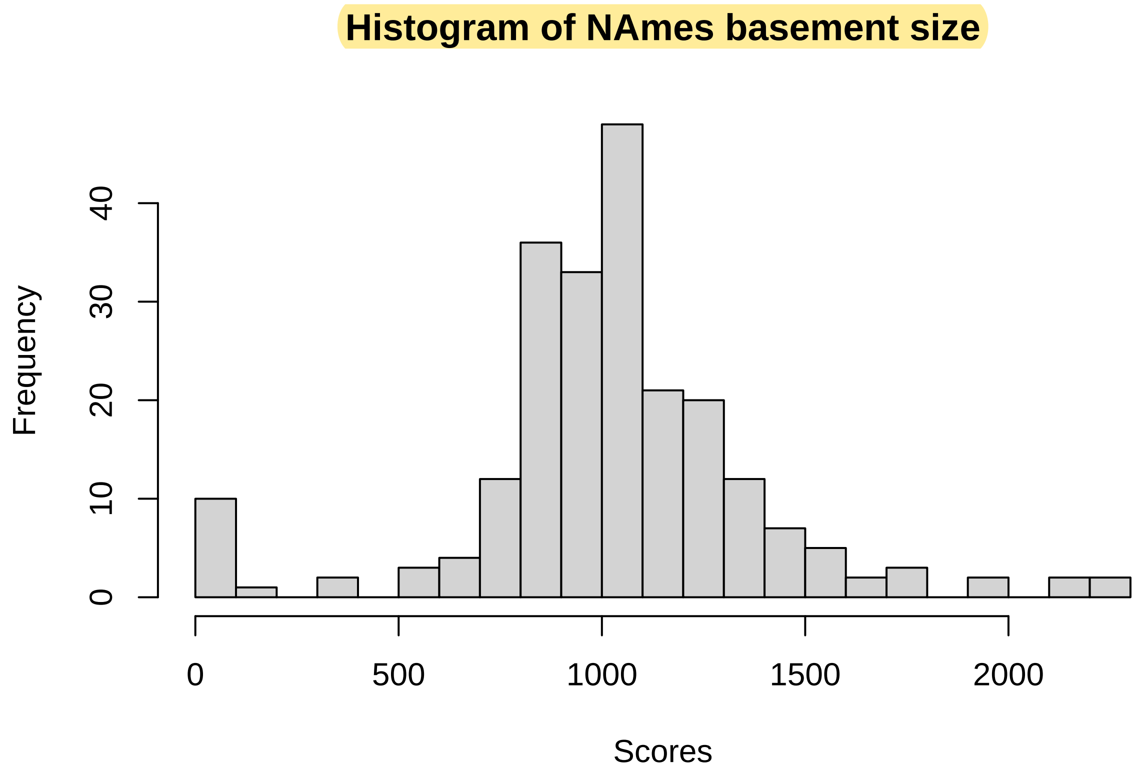

```
hist(as.double(MeadowV$TotalBsmtSF),
     breaks=25,
     main='Histogram of MeadowV basement size',
     xlab='Scores')
```



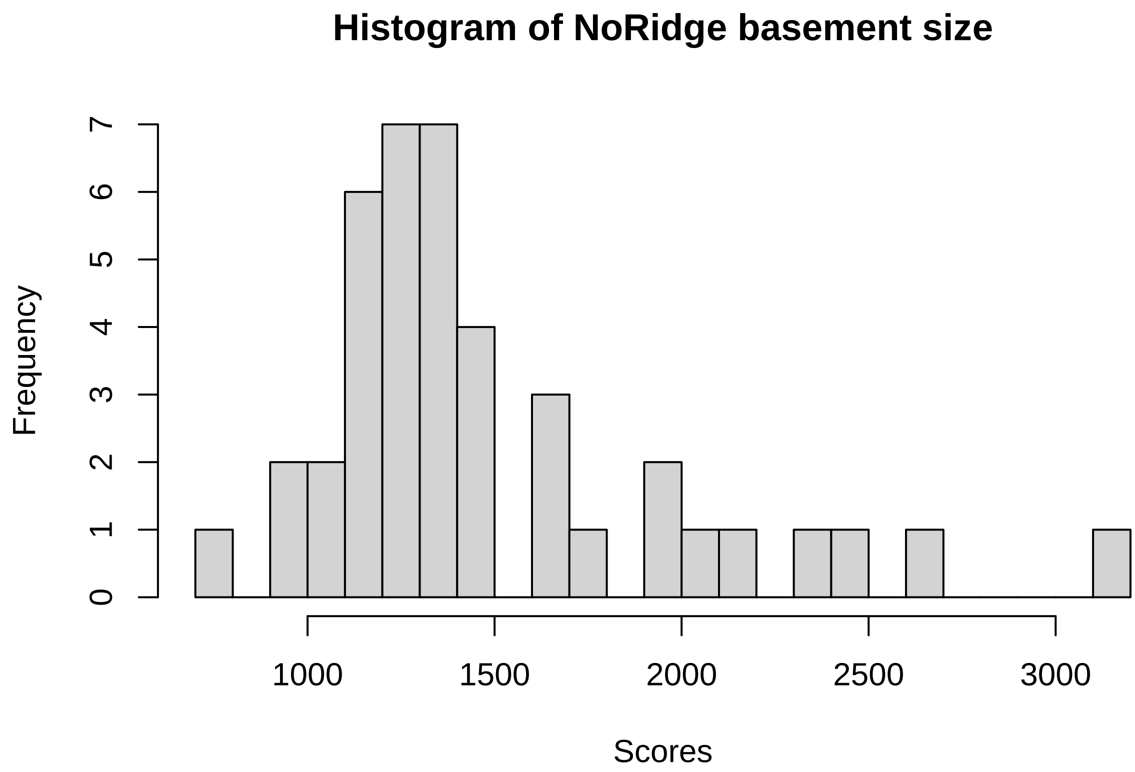

```r
hist(as.double(Mitchel$TotalBsmtSF),
     breaks=25,
     main='Histogram of Mitchel basement size',
     xlab='Scores')
```



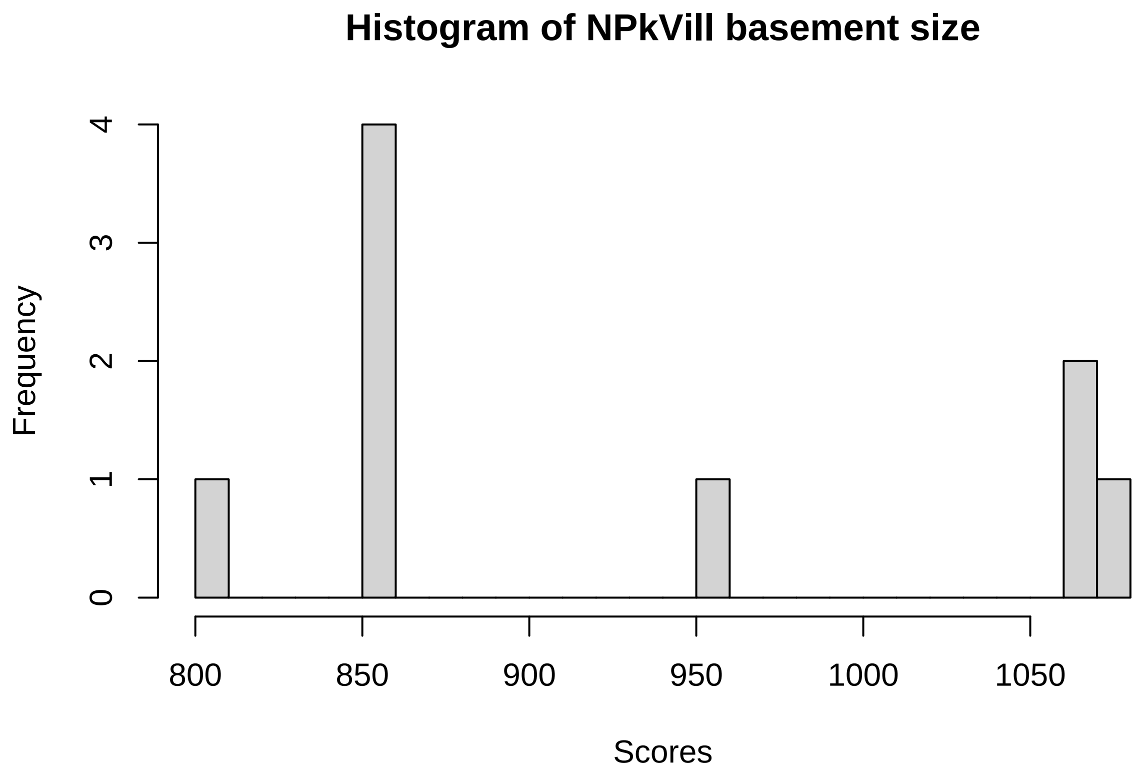

```r
hist(as.double(NAmes$TotalBsmtSF),
     breaks=25,
     main='Histogram of NAmes basement size',
     xlab='Scores')
```



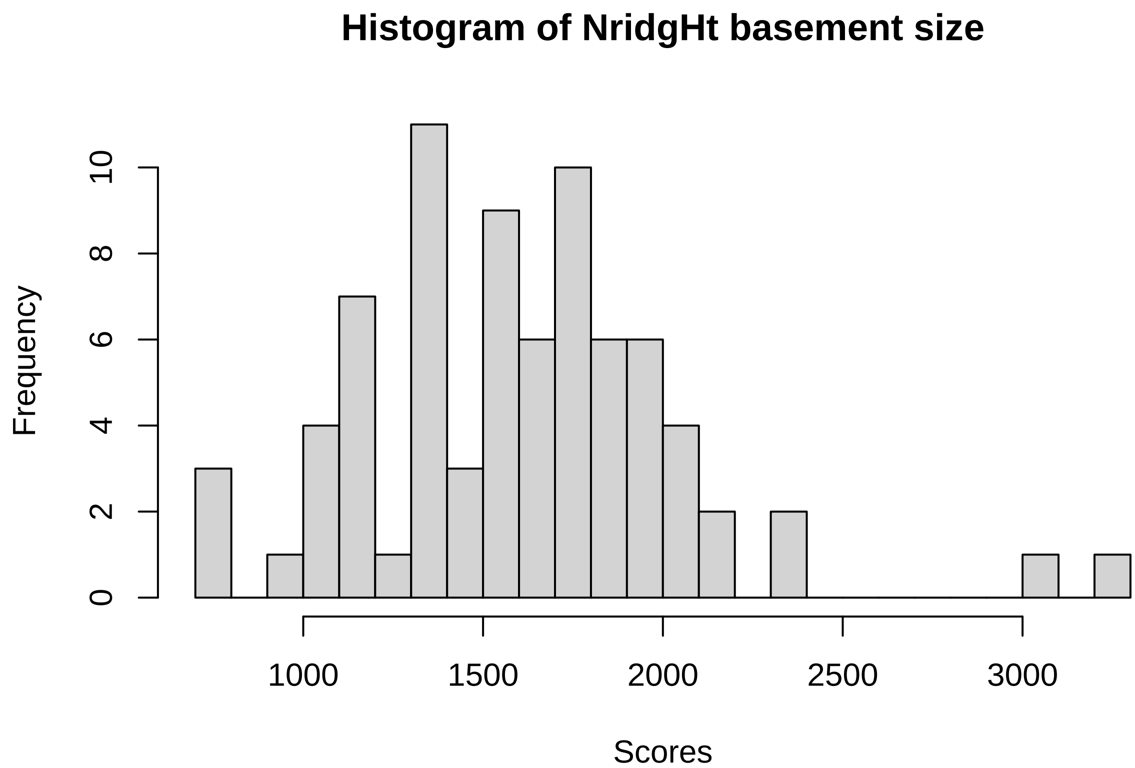

```
hist(as.double(NoRidge$TotalBsmtSF),
     breaks=25,
     main='Histogram of NoRidge basement size',
     xlab='Scores')
```



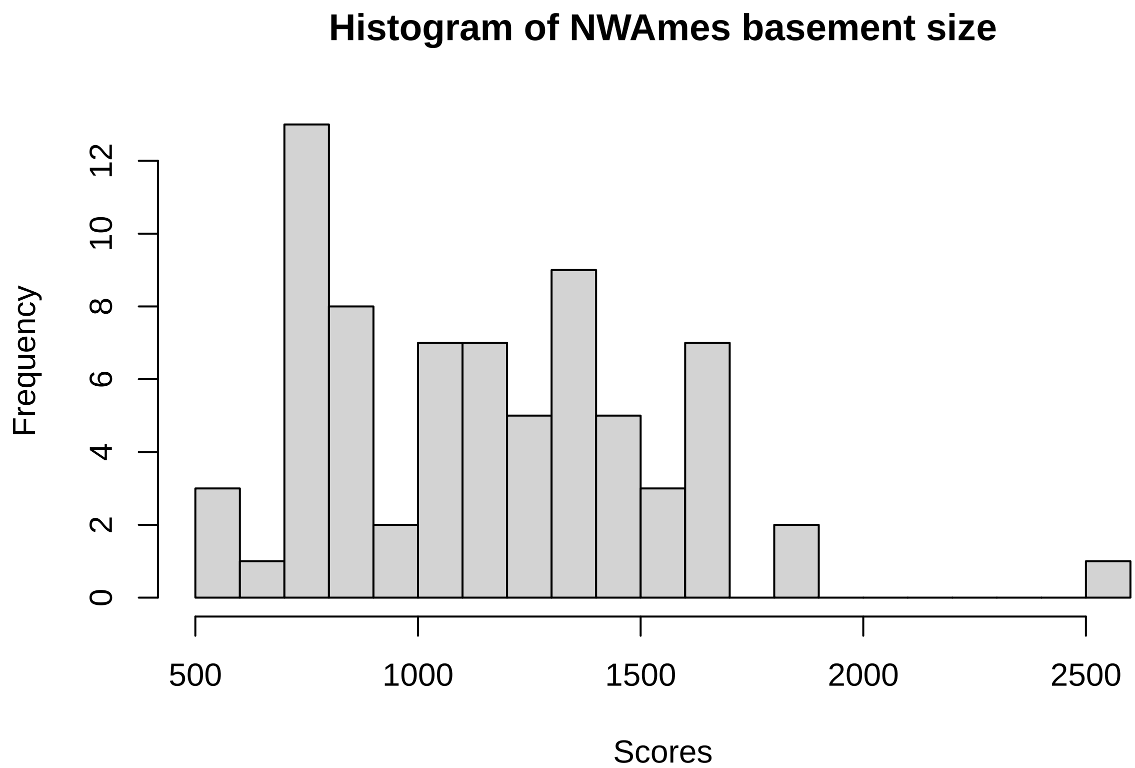

```
hist(as.double(NPkVill$TotalBsmtSF),
     breaks=25,
     main='Histogram of NPkVill basement size',
     xlab='Scores')
```



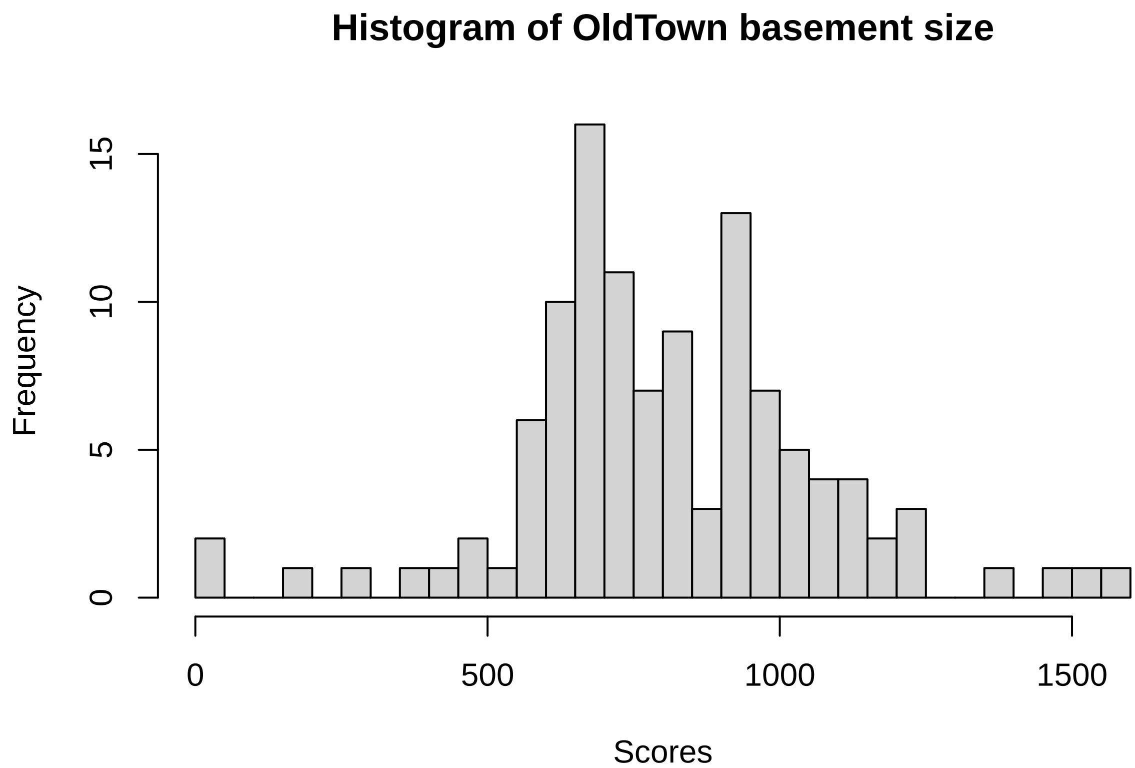

```
hist(as.double(NridgHt$TotalBsmtSF),
     breaks=25,
     main='Histogram of NridgHt basement size',
     xlab='Scores')
```



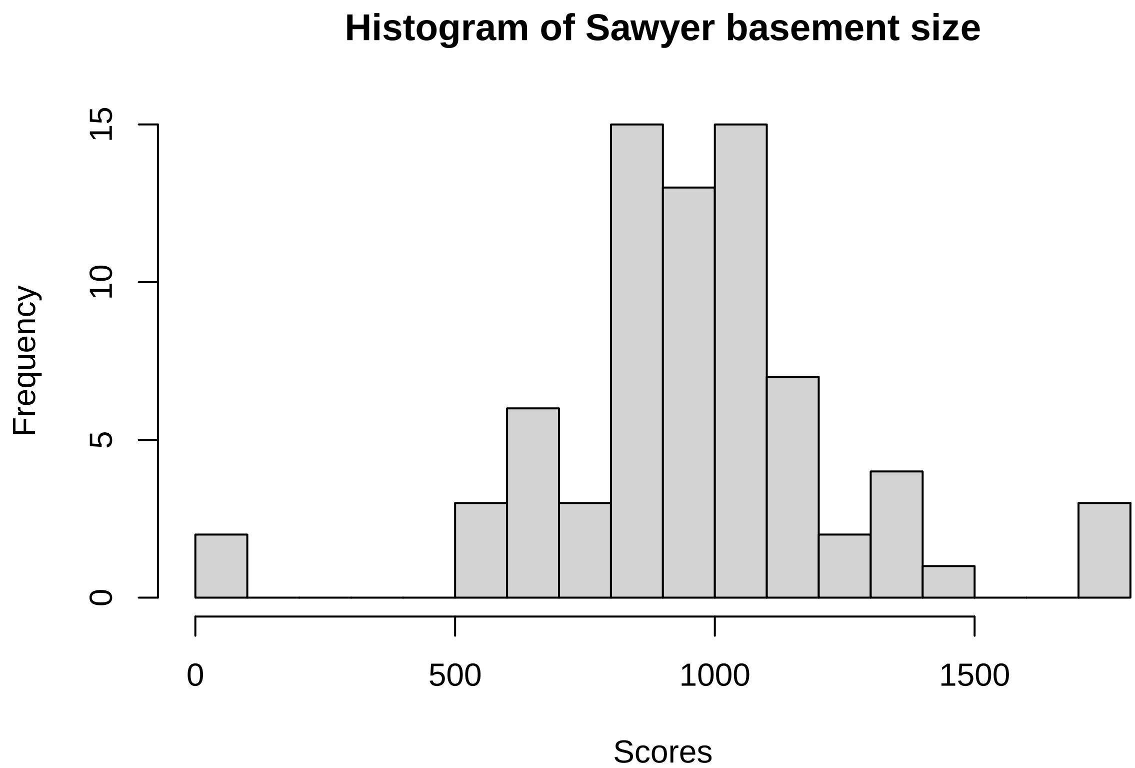

```
hist(as.double(NWAmes$TotalBsmtSF),
     breaks=25,
     main='Histogram of NWAmes basement size',
     xlab='Scores')
```



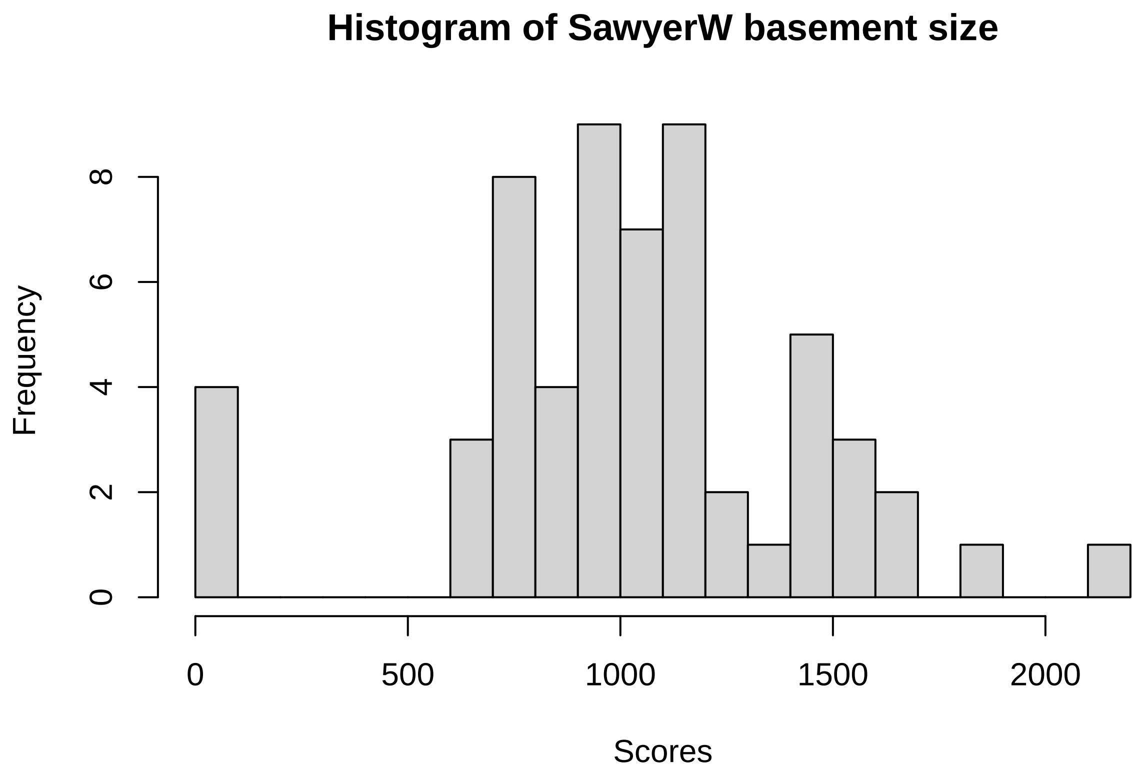

```
hist(as.double(OldTown$TotalBsmtSF),
     breaks=25,
     main='Histogram of OldTown basement size',
     xlab='Scores')
```




```
hist(as.double(Sawyer$TotalBsmtSF),
     breaks=25,
     main='Histogram of Sawyer basement size',
     xlab='Scores')
```

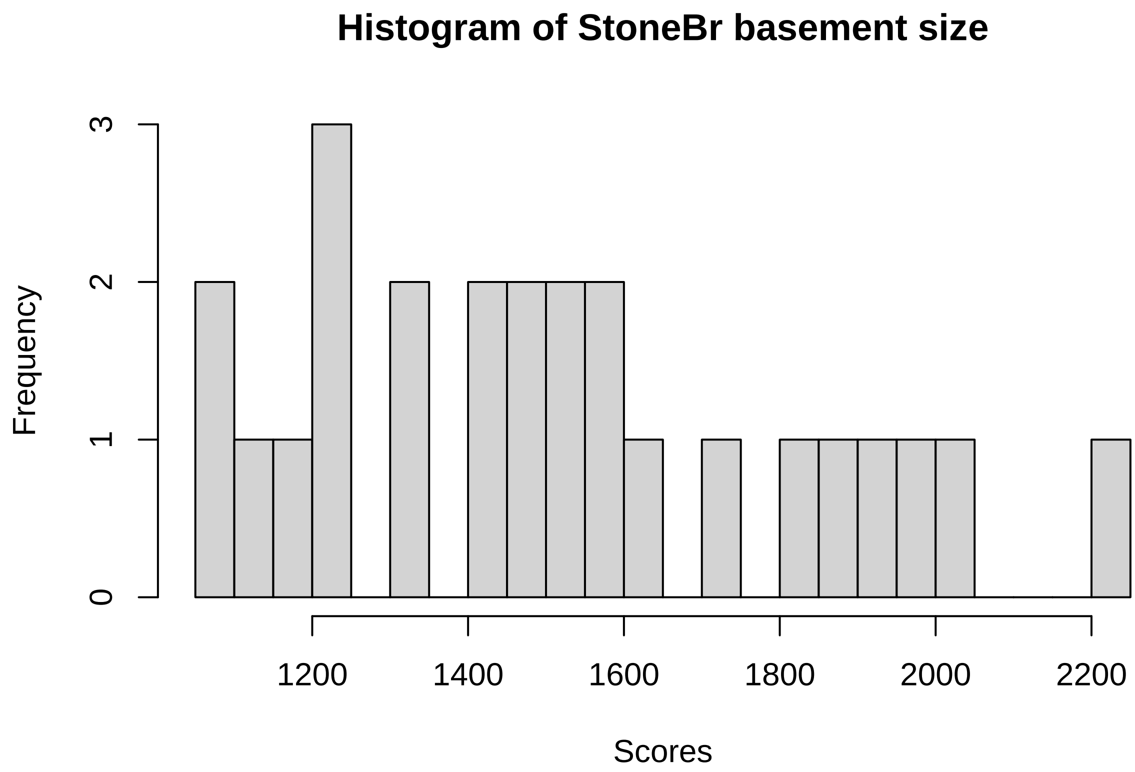



```
hist(as.double(SawyerW$TotalBsmtSF),
     breaks=25,
     main='Histogram of SawyerW basement size',
     xlab='Scores')
```

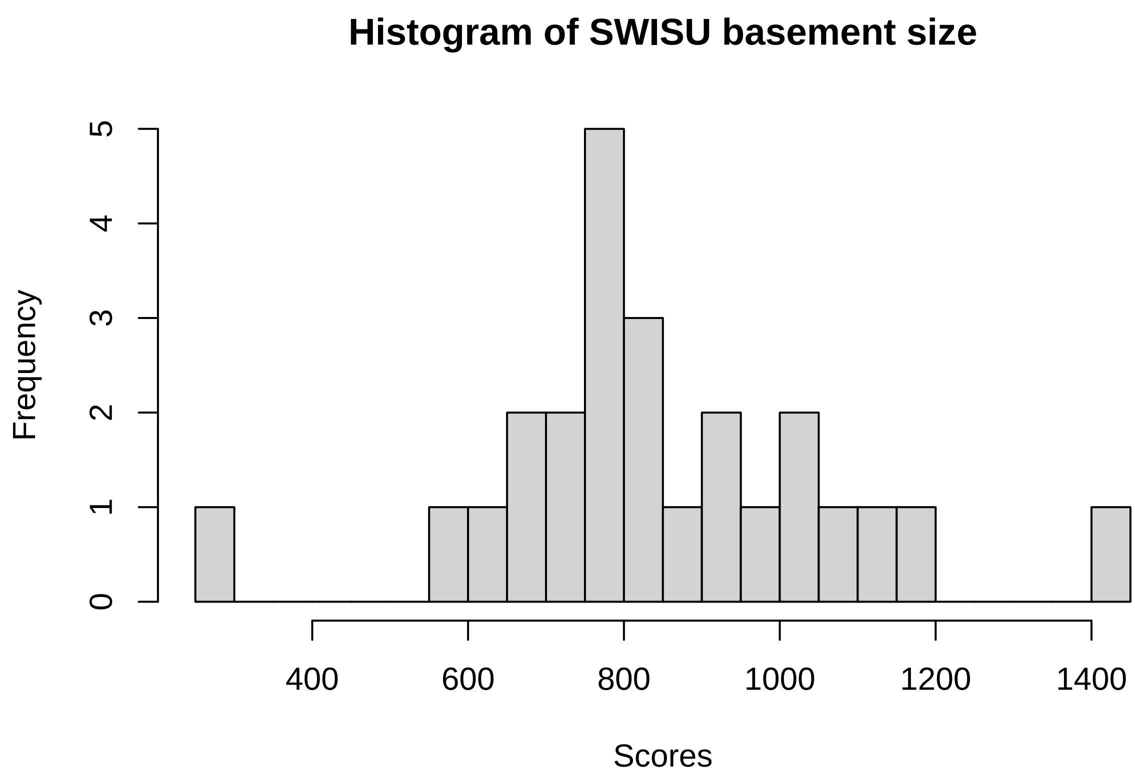



```
hist(as.double(Somerst$TotalBsmtSF),
     breaks=25,
     main='Histogram of Somerst basement size',
     xlab='Scores')
```

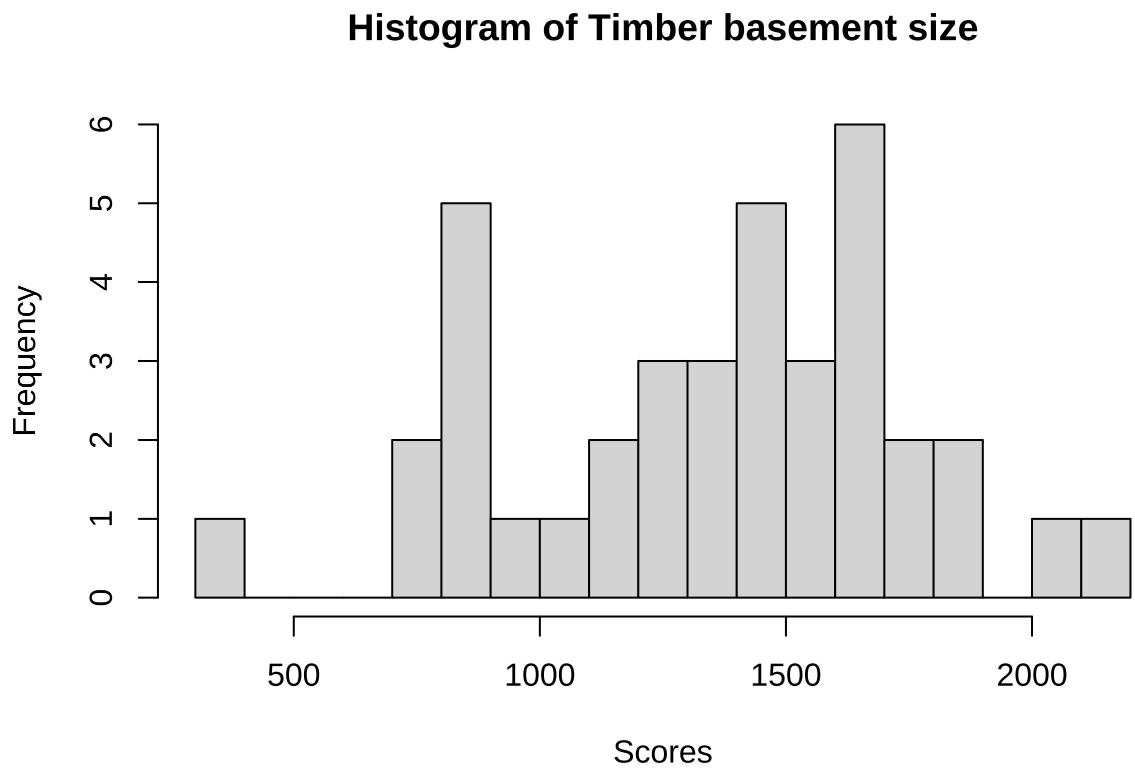

```
hist(as.double(StoneBr$TotalBsmtSF),
     breaks=25,
     main='Histogram of StoneBr basement size',
     xlab='Scores')
```



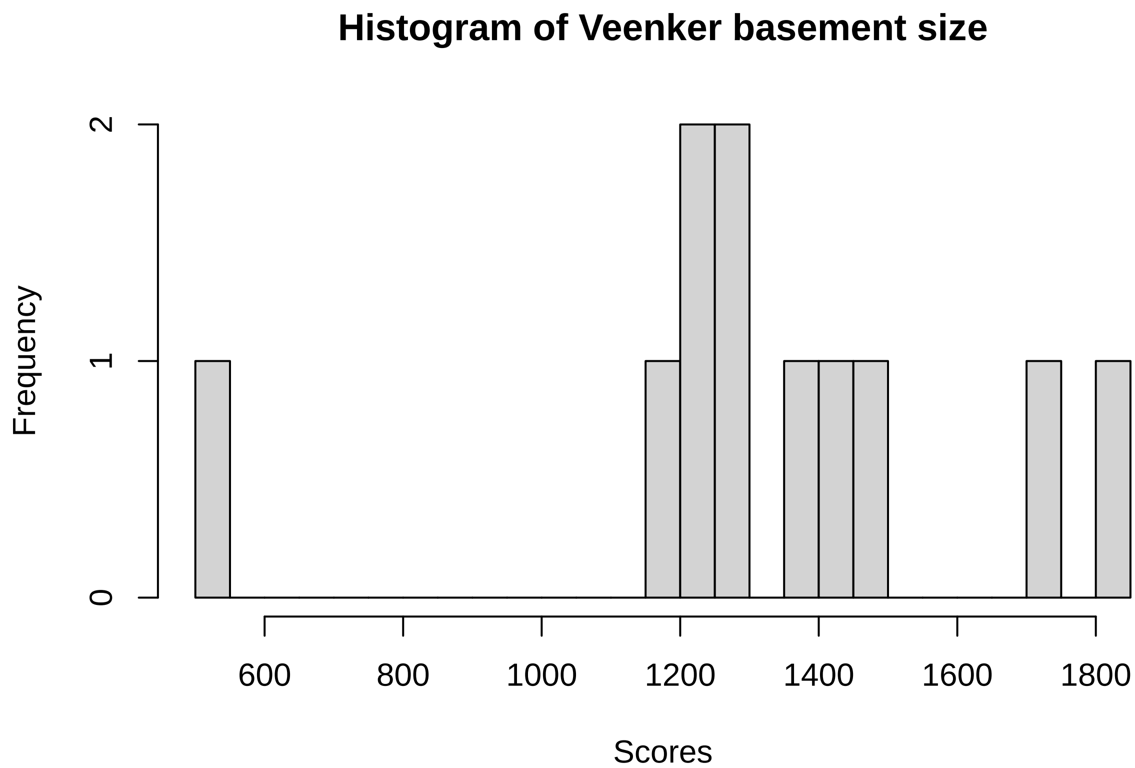

```
hist(as.double(SWISU$TotalBsmtSF),
     breaks=25,
     main='Histogram of SWISU basement size',
     xlab='Scores')
```




```
hist(as.double(Timber$TotalBsmtSF),
     breaks=25,
     main='Histogram of Timber basement size',
     xlab='Scores')
```




```
hist(as.double(Veenker$TotalBsmtSF),
     breaks=25,
     main='Histogram of Veenker basement size',
     xlab='Scores')
```



Podatci izgledaju normalno. Sada možemo raditi t-test test.

```
# Grupiramo podatke po četvrtima
grouped_data <- group_by(data, data$Neighborhood)
n_distinct(unique(data$Neighborhood))
```

```
## [1] 25
```

```
grouped_data
```

```
## # A tibble: 1,460 x 82
## # 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, 72 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>, ...
```

```
# Stvaramo praznu listu u koju ćemo spremati p-vrijednosti
p_values <- list()

# Prolazimo kroz sve četvrti
for (i in 1:(length(unique(data$Neighborhood))-1)) {
  for (j in (i+1):length(unique(data$Neighborhood))) {
    # Radimo t-test za svaki par četvrti
    test_result <- t.test(TotalBsmtSF ~ Neighborhood, data = data, subset = Neighborhood %in% c(unique(d
    # Spremamo p-vrijednost u listu
    p_values[[paste0(unique(data$Neighborhood)[i], "-", unique(data$Neighborhood)[j])]] <- test_result$p
  }
}

# Prilagođavamo razinu značanosti Bonferronijevom korekcijom
alpha <- 0.05
bonferroni_alpha <- alpha / length(p_values)

# Uspoređujemo p-vrijednosti prilagođenom razinom značanosti
significant_tests <- which(p_values < bonferroni_alpha)

print(length(significant_tests))
```

```
## [1] 141
```

Na kraju smo dobili broj testova u kojima je p-vrijednost manja od bonferroni alphe. Taj broj je 141. S obzirom da imamo 25 cetvrti, napravljeno je ukupno 25*24/2 = 300 testova, te je 141 statisticki značajan broj testova. Zbog toga zakljucujemo da velicina podruma ovisio kvartu.

4. Mogu li dostupne značajke predvidjeti cijenu nekretnine?

```
buildings_unfiltered <- read.csv("preprocessed_data.csv", header=TRUE, numerals="no.loss")
```

```
# Ucitamo podatke
buildings_unfiltered <- read.csv("preprocessed_data.csv", header=TRUE, numerals="no.loss")

# Izbacujemo one podatke gdje nema sale price
ind = which(buildings_unfiltered$SalePrice >= 0)

# Zelimo redove koji imaju sale price
data_outliers = buildings_unfiltered[ind,]

remove_outliers <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs=c(.25, .75), na.rm = na.rm, ...)
```

```r
  H <- 1.5 * IQR(x, na.rm = na.rm)
  y <- x
  y[x < (qnt[1] - H)] <- NA
  y[x > (qnt[2] + H)] <- NA
  y
}

# Odabiremo 10 znacajki koje zelimo provjeriti kako predvidaju cijenu

#MSSubClass
data_outliers$MSSubClass <-  remove_outliers(data_outliers$MSSubClass)
#MSZoning
msZoning_map <- c("A" = 1,"C" = 2, "FV" = 3, "I" = 4, "RH" = 5, "RL" = 6, "RP" = 7, "RM" = 8)
data_outliers$MSZoning <-  as.numeric(msZoning_map[data_outliers$MSZoning])
data_outliers$MSZoning <-  remove_outliers(as.numeric(data_outliers$MSZoning))
#OverallQual
data_outliers$OverallQual <-  remove_outliers(data_outliers$OverallQual)
#OverallCond
data_outliers$OverallCond <- remove_outliers(data_outliers$OverallCond)
#YearBuilt
data_outliers$YearBuilt <- remove_outliers(data_outliers$YearBuilt)
#YearRemodAdd
data_outliers$YearRemodAdd <- remove_outliers(data_outliers$YearRemodAdd)
#ExterQual
extQual_map <- c("Ex" = 5, "Gd" = 4, "TA" = 3, "Fa" = 2, "Po" = 1)
data_outliers$ExterQual <- as.numeric(extQual_map[data_outliers$ExterQual])
#TotalBsmtSF
data_outliers$TotalBsmtSF <- remove_outliers(data_outliers$TotalBsmtSF)
#SQFT
sqft <- remove_outliers(data_outliers$X1stFlrSF + data_outliers$X2ndFlrSF)
#SaleType
saleType_map <- c("Oth" = 0, "ConLD" = 1, "ConLI" = 2, "ConLw" = 3, "Con" = 4, "COD" = 5, "New" = 6, "VI
data_outliers$SaleType <- as.numeric(saleType_map[data_outliers$SaleType])

#saleprice

model <- lm(data_outliers$SalePrice ~ data_outliers$MSSubClass + data_outliers$MSZoning + data_outliers$
summary(model)
```

```
##
## Call:
## lm(formula = data_outliers$SalePrice ~ data_outliers$MSSubClass +
##     data_outliers$MSZoning + data_outliers$OverallQual + data_outliers$OverallCond +
##     data_outliers$YearBuilt + data_outliers$YearRemodAdd + data_outliers$ExterQual +
##     sqft + data_outliers$TotalBsmtSF + data_outliers$SaleType)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -135612  -16966    -585   14616  140193
##
## Coefficients: (1 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -8.652e+05  1.226e+05  -7.056 3.26e-12 ***
```

```
## data_outliers$MSSubClass   -1.294e+02  3.271e+01  -3.957 8.14e-05 ***
## data_outliers$MSZoning             NA         NA      NA       NA
## data_outliers$OverallQual   1.707e+04  1.231e+03  13.869  < 2e-16 ***
## data_outliers$OverallCond   9.241e+03  1.357e+03   6.810 1.72e-11 ***
## data_outliers$YearBuilt     2.815e+02  6.331e+01   4.447 9.73e-06 ***
## data_outliers$YearRemodAdd  8.331e+01  7.088e+01   1.175   0.2401
## data_outliers$ExterQual     1.840e+04  2.604e+03   7.064 3.11e-12 ***
## sqft                        5.851e+01  2.574e+00  22.730  < 2e-16 ***
## data_outliers$TotalBsmtSF   3.649e+01  3.174e+00  11.498  < 2e-16 ***
## data_outliers$SaleType     -1.413e+03  7.271e+02  -1.943   0.0523 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27410 on 964 degrees of freedom
##   (486 observations deleted due to missingness)
## Multiple R-squared:  0.8273, Adjusted R-squared:  0.8257
## F-statistic: 513.2 on 9 and 964 DF,  p-value: < 2.2e-16
```

Izradjen je model predvidjanja cijene nekretnine s obzirom na ovih 10 znacajki. Na temelju ispisa modela vidimo da su odabrane znacajke jako dobre gdje Pearsonov koeficijent korelacije iznosi 0.8273. Takodjer vidimo da mozemo pretpostaviti cijenu na temelju zadanog modela.

KRAJ