HOMEWORK 2 GRUPPO G

2024-06-01

Importazione dei dataset

```
case <- read.csv("C:\\Users\\giova\\Downloads\\train.csv", stringsAsFactors =
TRUE)
banche <- read.csv("C:\\Users\\giova\\Downloads\\credit_card.csv")
```

Librerie necessarie

```
library(moments)
library(ggplot2)
library(dplyr)

##
## Caricamento pacchetto: 'dplyr'

## I seguenti oggetti sono mascherati da 'package:stats':
##
##     filter, lag

## I seguenti oggetti sono mascherati da 'package:base':
##
##     intersect, setdiff, setequal, union
```

Funzioni usate nel progetto

Funzioni per Analisi Univariata

```
# Funzione per le Variabili Quantitative
display_summary_and_var <- function(variabile){
  c(summary(variabile),
  var = var(variabile, na.rm = T),
  sd = sd(variabile, na.rm = T),
  sk = skewness(variabile, na.rm = T))
}

# Funzione per le Variabili Qualitative
display_table <- function(variabile, titolo){
  DistAs <- table(variabile)
  DistRe <- prop.table(table(variabile))
  barplot(prop.table(table(variabile)), main = titolo)
  print(rbind(DistAs, DistRe))
}
```

Funzioni per Analisi Bivariata

```
# Funzione per le Variabili Quantitative
calcolo_cov_cor <- function(variabile_numerica) {
  c(cov = cov(variabile_numerica, case$SalePrice, use = "complete.obs"), cor =
cor(variabile_numerica, case$SalePrice, use = "complete.obs"))
```

```
}
# Funzione per le Variabili Qualitative
calcola_devianza <- function(numerical_var, categorical_var) {
  # Creiamo un dataframe temporaneo per facilitare i calcoli
  data <- data.frame(numerical_var, categorical_var)
  # Togliamo le righe contenenti valori NA
  data <- na.omit(data)
  # Calcoliamo la media generale
  mean_total <- mean(data$numerical_var, na.rm = T)

  # Devianza totale
  devianza_totale <- sum((data$numerical_var - mean_total)^2)

  # Calcolo della devianza tra i gruppi
  devianza_tra_gruppi <- 0
  livelli <- levels(data$categorical_var)
  for (livello in livelli) {
    gruppo <- data[data$categorical_var == livello, ]
    n <- nrow(gruppo)
    mean_gruppo <- mean(gruppo$numerical_var, na.rm = T)
    devianza_tra_gruppi <- devianza_tra_gruppi + n * (mean_gruppo - mean_total)^2
  }

  # Calcolo della devianza entro i gruppi
  devianza_entro_gruppi <- 0
  for (livello in livelli) {
    gruppo <- data[data$categorical_var == livello,]
    mean_gruppo <- mean(gruppo$numerical_var, na.rm=T)
    devianza_entro_gruppi <- devianza_entro_gruppi + sum((gruppo$numerical_var -
mean_gruppo)^2, na.rm=T)
  }

  return(list(devianza_totale = devianza_totale, devianza_tra_gruppi =
devianza_tra_gruppi, devianza_entro_gruppi = devianza_entro_gruppi, eta2 =
(devianza_tra_gruppi/devianza_totale)))
}
```

Dataset - Credit card customers

Analisi univariata

Variabile Attrition_Flag

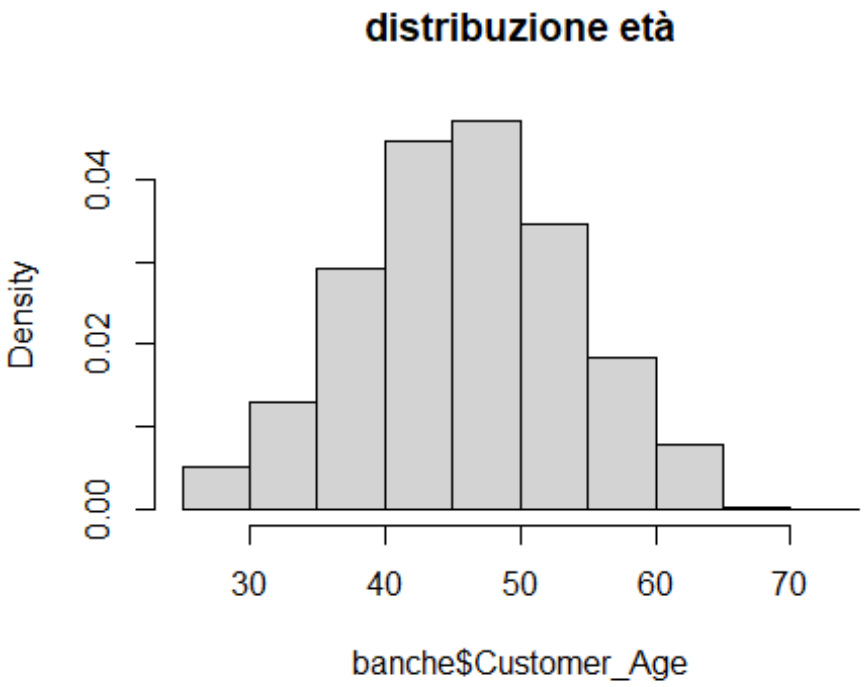
```
banche$Attrition_Flag <- factor(banche$Attrition_Flag)
display_table(banche$Attrition_Flag, "Attrition_Flag")
```


```
##        Attrited Customer Existing Customer
## DistAs      1627.0000000      8500.0000000
## DistRe         0.1606596         0.8393404
```

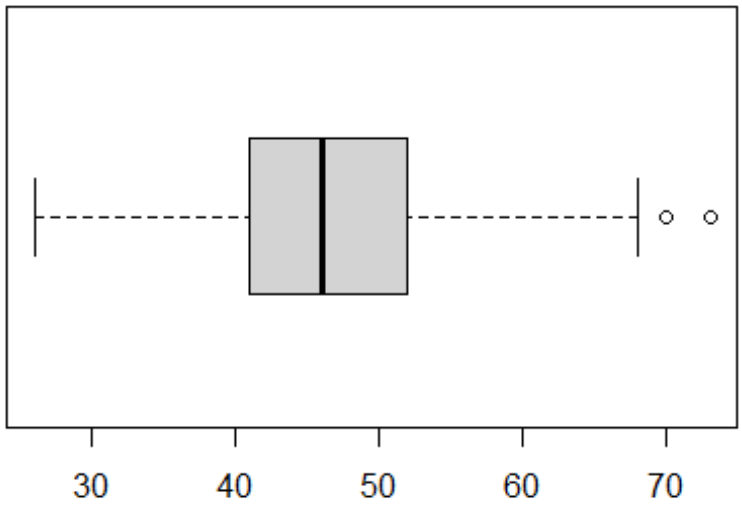
Si nota che l'83.9% dei dati riguarda clienti esistenti della banca mentre il 16.1% dei dati riguarda clienti persi / clienti passati.

Variabile Customer Age

```
display_summary_and_var(banche$Customer_Age)

##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## 26.00000000 41.00000000 46.00000000 46.32596030 52.00000000 73.00000000
##         var          sd          sk
## 64.26930723  8.01681403 -0.03360004

hist(banche$Customer_Age, freq = F, main = "distribuzione età")
```




```r
boxplot(banche$Customer_Age, horizontal = T)
```

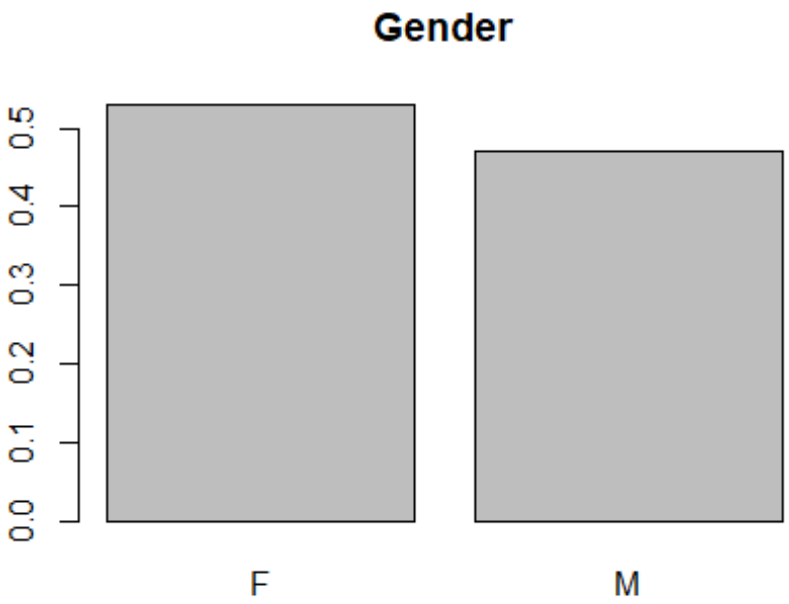


La variabile giguardante l'età dei clienti è una variabile quantitativa, si nota che l'età media del campione è 46.33 e la mediana è 46.00, i valori sono abbastanza vicini infati se si calcola l'indice

di asimmetria attraverso la fnuzione skewness() si ottiene un numero prossimo allo zero. Infine dal boxplot si evince che 2 clienti hanno come età dei valori outliers

Variabile Gender

```
banche$Gender <- factor(banche$Gender)
display_table(banche$Gender, "Gender")
```

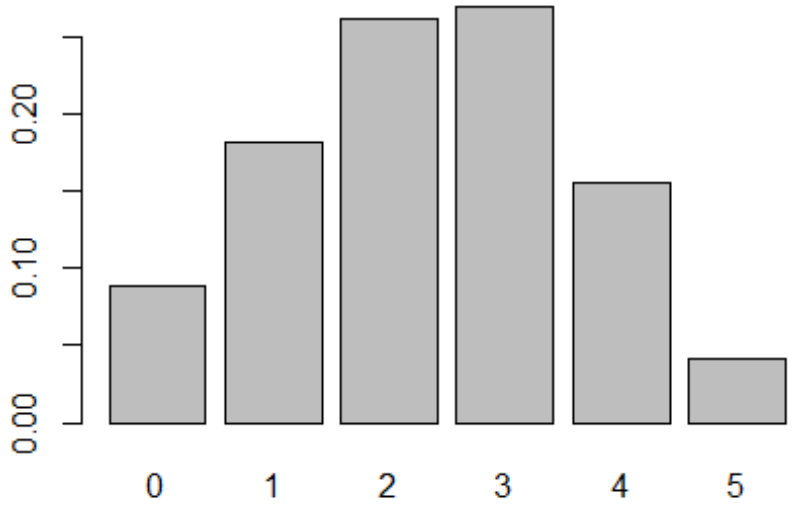



```
##                    F            M
## DistAs 5358.0000000 4769.0000000
## DistRe    0.5290807    0.4709193
```

La variabile Gender è una variabile qualitativa che presenta le seguenti frequenze relative: si osserva che la frequenza delle donne è leggermente maggiore

Variabile Dependent_count

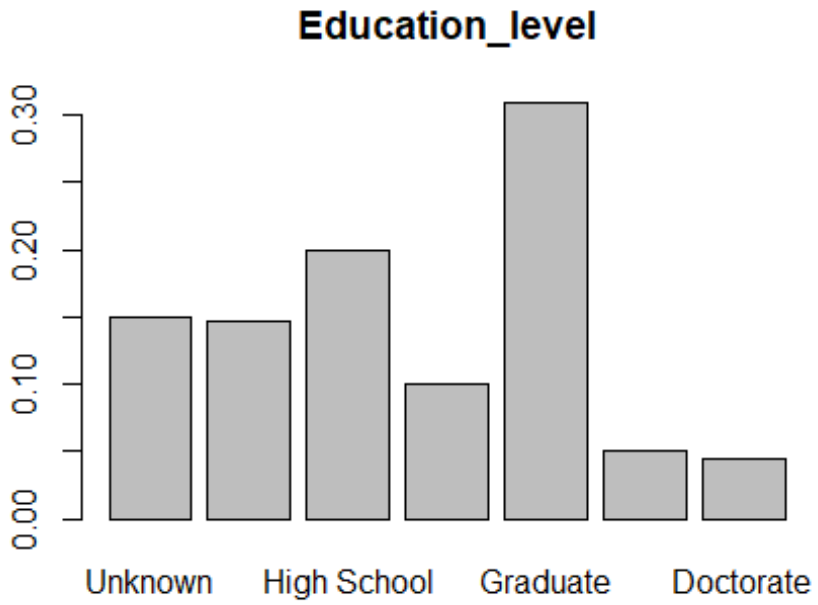
```
display_summary_and_var(banche$Dependent_count)

##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##  0.00000000  1.00000000  2.00000000  2.34620322  3.00000000  5.00000000
##         var          sd          sk
##  1.68716290  1.29890835 -0.02082245

barplot(prop.table(table(banche$Dependent_count)))
```



E’ una variabile quantitativa discreta, che ha come minimo 0 e come massimo 5 la media dei valori è 2.346

Variabile Education_Level

```
banche$Education_Level <- factor(banche$Education_Level, levels =
c("Unknown","Uneducated","High School", "College", "Graduate", "Post-Graduate",
"Doctorate") )
display_table(banche$Education_Level, "Education_level")
```

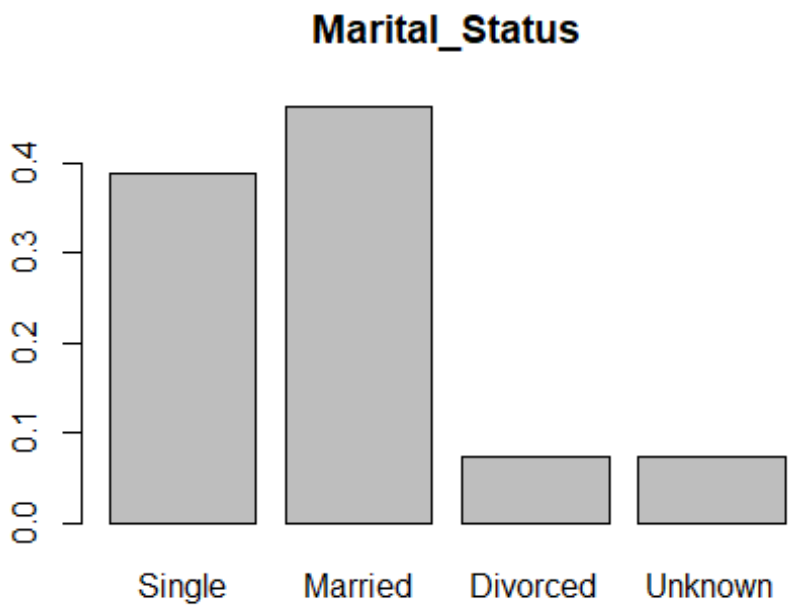



```
##             Unknown   Uneducated  High School      College     Graduate
## DistAs 1519.0000000 1487.0000000 2013.0000000 1013.0000000 3128.0000000
## DistRe    0.1499951    0.1468352    0.1987756    0.1000296    0.3088773
##        Post-Graduate    Doctorate
## DistAs   516.0000000 451.00000000
## DistRe     0.0509529   0.04453441
```

La variabile Gender è una variabile qualitativa che presenta le seguenti frequenze relative: si osserva che la frequenza maggiore è quella relativa ai clienti "Graduate", del 30.8% e si osserva inoltre che non si conoscono i dati di circa il 15% del nostro campione

Variabile Marital_Status

```
banche$Marital_Status <- factor(banche$Marital_Status)
banche$Marital_Status <- ordered(banche$Marital_Status, levels = c("Single",
"Married", "Divorced", "Unknown"))
display_table(banche$Marital_Status,"Marital_Status")
```

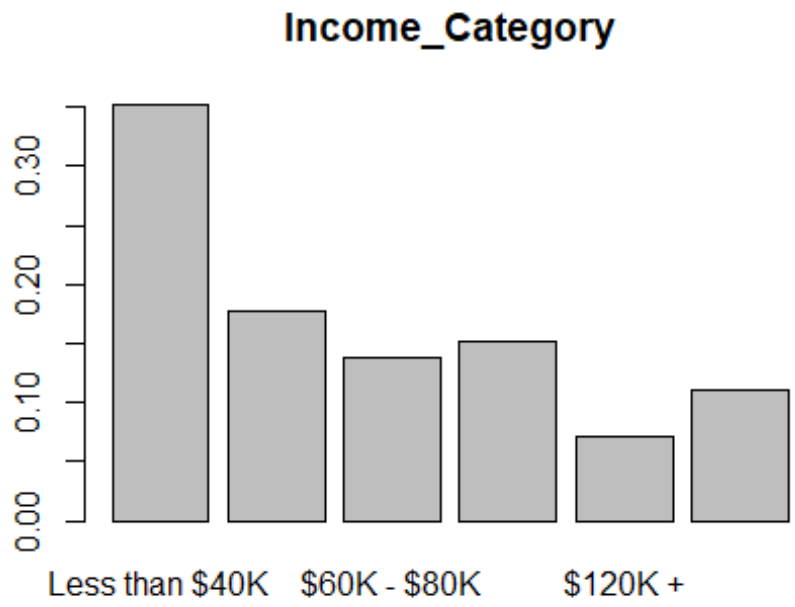



```
##              Single      Married     Divorced     Unknown
## DistAs 3943.0000000 4687.0000000 748.00000000 749.0000000
## DistRe    0.3893552    0.4628222   0.07386195   0.0739607
```

Variabile Character che indica lo stato di relazione della persona che possiede il conto in banca. Vediamo che le categorie più popolari sono “Merried” e “single”

Variabile Income_Category

```
banche$Income_Category <- factor(banche$Income_Category)
banche$Income_Category <- ordered(banche$Income_Category, levels = c("Less than
$40K","$40K - $60K","$60K - $80K","$80K - $120K","$120K +","Unknown"))
display_table(banche$Income_Category, "Income_Category")
```

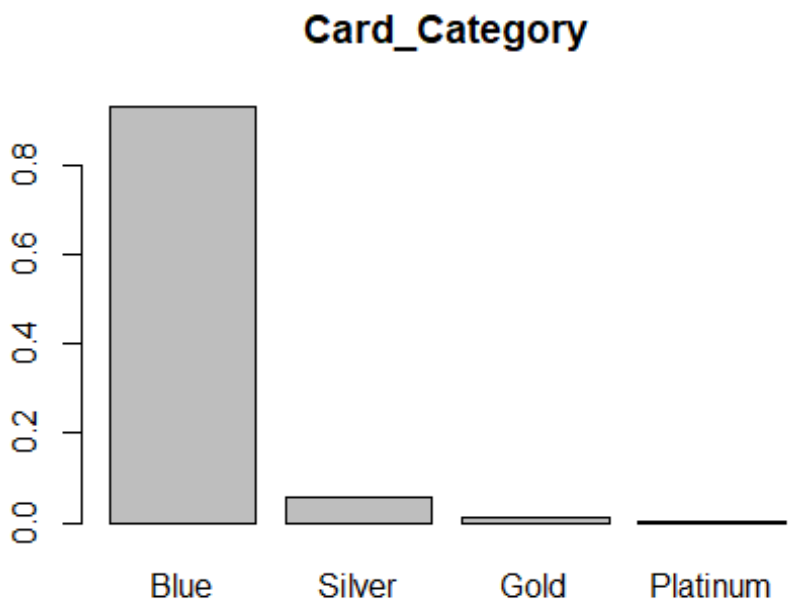



```
##        Less than $40K  $40K - $60K  $60K - $80K $80K - $120K      $120K +
## DistAs   3561.0000000 1790.0000000 1402.0000000  1535.000000 727.00000000
## DistRe      0.3516342    0.1767552    0.1384418     0.151575   0.07178829
##             Unknown
## DistAs 1112.0000000
## DistRe    0.1098055
```

Anche se la variabile potrebbe essere vista come una variabile Numerica che rappresenta il reddito diviso in classi, questa ci viene invece fornita come variabile Character che verrà ordinata a mano. La "Classe" più comune è "Less than $40K" e la più rara "$120K +".

Variabile Card_Category

```
banche$Card_Category <- factor(banche$Card_Category)
banche$Card_Category <- ordered(banche$Card_Category, levels =
c("Blue","Silver","Gold","Platinum"))
display_table(banche$Card_Category, "Card_Category")
```




```
##                Blue       Silver         Gold     Platinum
## DistAs 9436.0000000 555.00000000 116.00000000 20.000000000
## DistRe    0.9317666   0.05480399   0.01145453  0.001974919
```

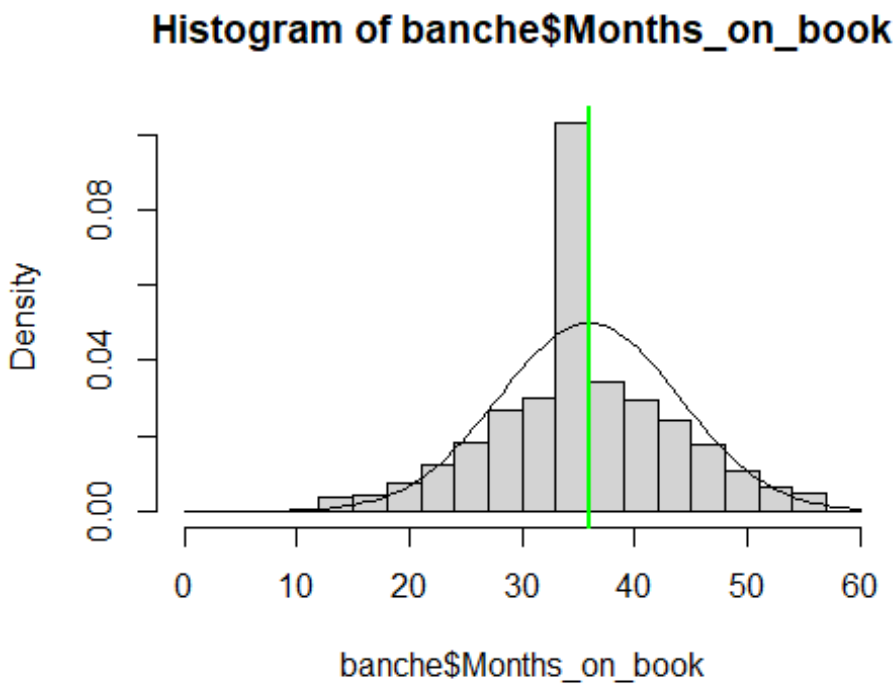
E' una variabile Character che indica la categoria di carta del della persona che possiede il conto. Considerando che i "livelli" di una carta di credito si spostano generalmente da bronzo fino a platino si riordiniamo i livelli prima di continuare le osservazioni (nel nostro caso il "Bronzo" corrisponderà al "Blue") Le carte rilasciate sono quasi esclusivamente del tipo "Blue".

Variabile Months_on_book

```
display_summary_and_var(banche$Months_on_book)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.        var
## 13.0000000 31.0000000 36.0000000 35.9284092 40.0000000 56.0000000 63.7828458
##         sd         sk
##  7.9864163 -0.1065496
```

```
hist(banche$Months_on_book, breaks = c(3*0:20), probability = T)
curve(dnorm(x,mean(banche$Months_on_book, na.rm = T), sd(banche$Months_on_book,
na.rm = T)),add = T)
abline(v = median(banche$Months_on_book, na.rm = T),lwd = 2, col = "red")
abline(v = mean(banche$Months_on_book, na.rm = T),lwd = 2, col = "green")
```



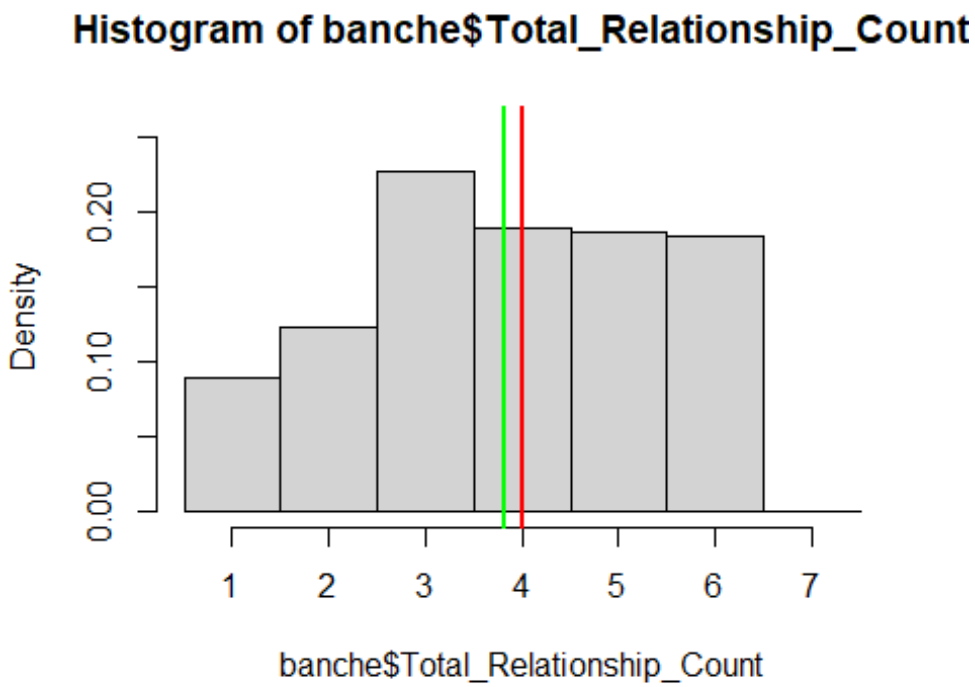
Variabile di tipo Quantitativo che indica il numero di mesi che che un cliente ha passato come cliente della Banca Vediamo che è presente un picco tra 33 e 36 che va ben sopra la gaussiana costruita con Media e SD della distribuzione. Moda e Mediana sono estremamente vicine.

Variabile Total_Relationship_Count

```
display_summary_and_var(banche$Total_Relationship_Count)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.        var
##  1.0000000  3.0000000  4.0000000  3.8125802  5.0000000  6.0000000  2.4161838
##         sd         sk
##  1.5544079 -0.1624284
```

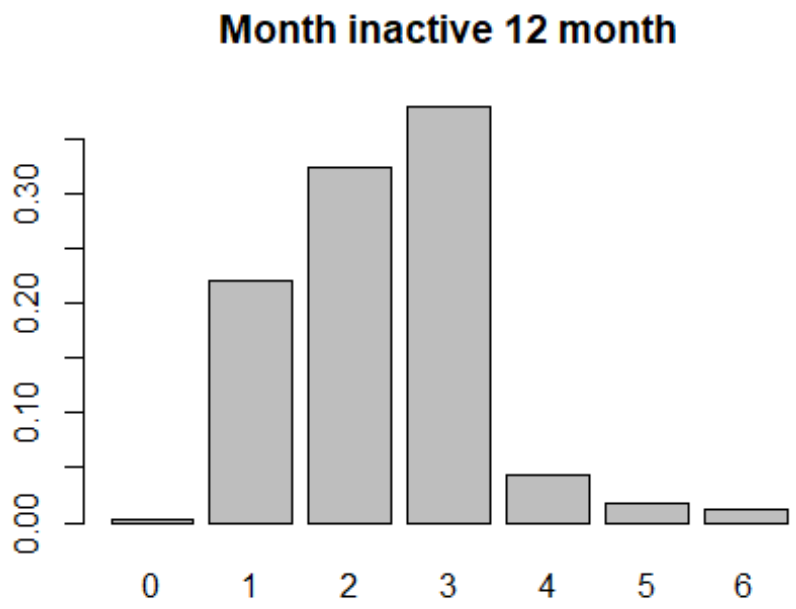
```
hist(banche$Total_Relationship_Count, breaks = c(0:7)+0.5,ylim= c(0,0.26),
probability = T)
abline(v = median(banche$Total_Relationship_Count, na.rm = T),lwd = 2, col =
"red")
abline(v = mean(banche$Total_Relationship_Count, na.rm = T),lwd = 2, col =
"green")
```



E' una variabile numerica che rappresenta il numero totale di prodotti della banca posseduto dal utente. Raramente i clienti possiedono solo una o due carte e si nota che, anche se la moda è 3, sia media che mediana sono vicine al 4.

Variabile month_inactive

```
banche$Months_Inactive_12_mon <-factor(banche$Months_Inactive_12_mon)
display_table(banche$Months_Inactive_12_mon, titolo = 'Month inactive 12 month')
```




```
##                    0            1            2            3            4
## DistAs 29.000000000 2233.0000000 3282.0000000 3846.0000000 435.00000000
## DistRe  0.002863632    0.2204997    0.3240841    0.3797768   0.04295448
##                    5            6
## DistAs 178.00000000 124.00000000
## DistRe   0.01757677   0.01224449
```

Si osserva una maggiore concentrazione di valori in 2 e 3 presenta solamente valori da 1 a 6 quindi verrà trattata come una variabile categoriale

Variabile contacts count 12 month

```
banche$Contacts_Count_12_mon <- factor(banche$Contacts_Count_12_mon)
display_table(banche$Contacts_Count_12_mon, titolo = 'contacts 12 month')
```


```
##                    0            1            2            3            4
## DistAs 399.00000000 1499.0000000 3227.0000000 3380.0000000 1392.0000000
## DistRe   0.03939962    0.1480201    0.3186531    0.3337612    0.1374543
##                    5           6
## DistAs 176.00000000 54.00000000
## DistRe   0.01737928  0.00533228
```

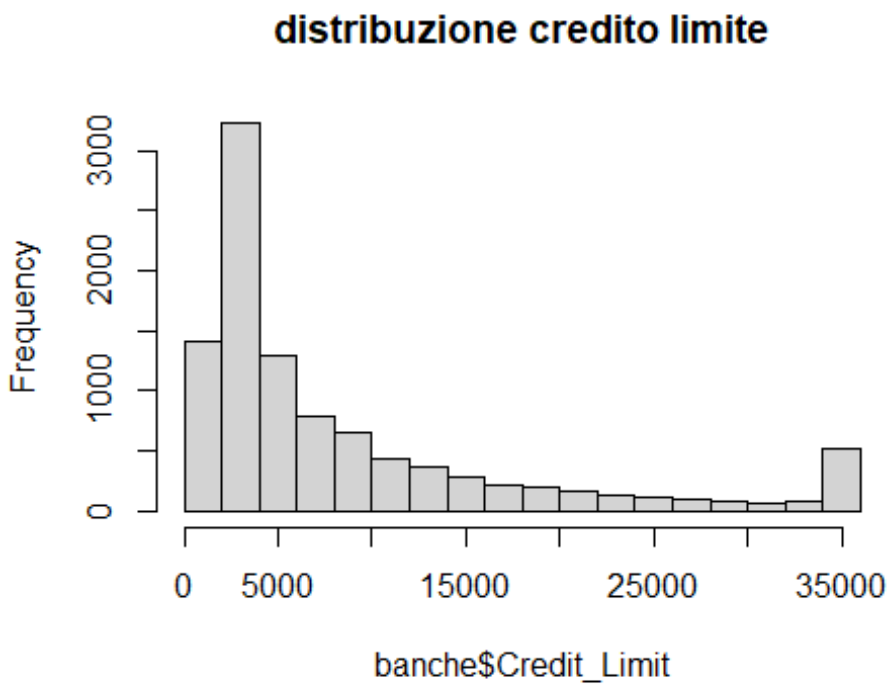
Anche qui si osserva una maggiore concentrazione di valori maggiore in 2 e 3, questa variabile è meno assimetrica della precedente presenta solamente valori da 1 a 6 quindi verrà trattata come una variabile categoriale

Variabile credit Limit

```
display_summary_and_var(banche$Credit_Limit)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 1.438300e+03 2.555000e+03 4.549000e+03 8.631954e+03 1.106750e+04 3.451600e+04
##          var           sd           sk
## 8.260586e+07 9.088777e+03 1.666479e+00
```

```
hist(banche$Credit_Limit, freq = T, main = "distribuzione credito limite")
```



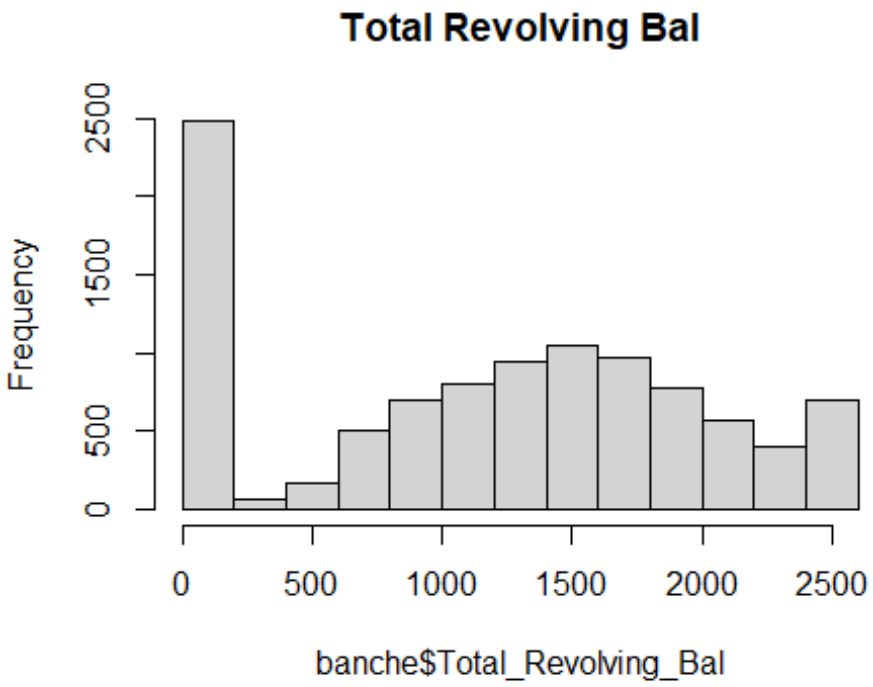
Si osserva una forte asimetria nella distribuzione, la media è di circa 8631 mentre la mediana di 4549

Variabile total revolving bal

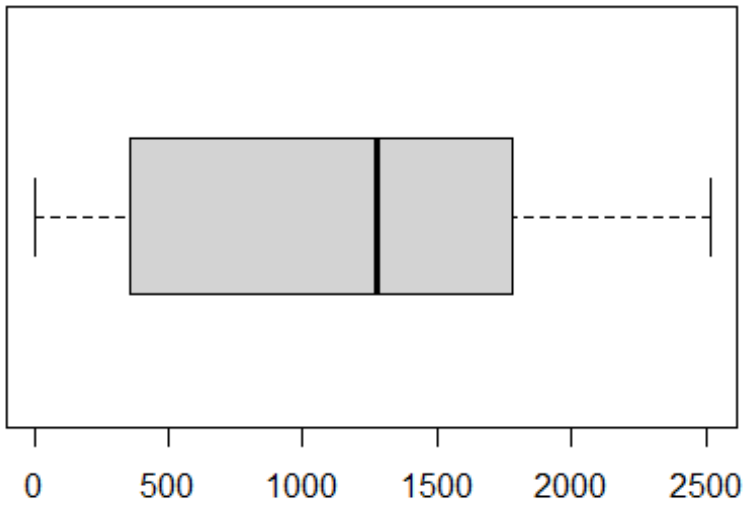
```
display_summary_and_var(banche$Total_Revolving_Bal)
```

```
##          Min.       1st Qu.        Median          Mean       3rd Qu.
##  0.000000e+00  3.590000e+02  1.276000e+03  1.162814e+03  1.784000e+03
##          Max.           var            sd            sk
##  2.517000e+03  6.642044e+05  8.149873e+02 -1.488152e-01
```

```
hist(banche$Total_Revolving_Bal, freq = T, main = "Total Revolving Bal")
```



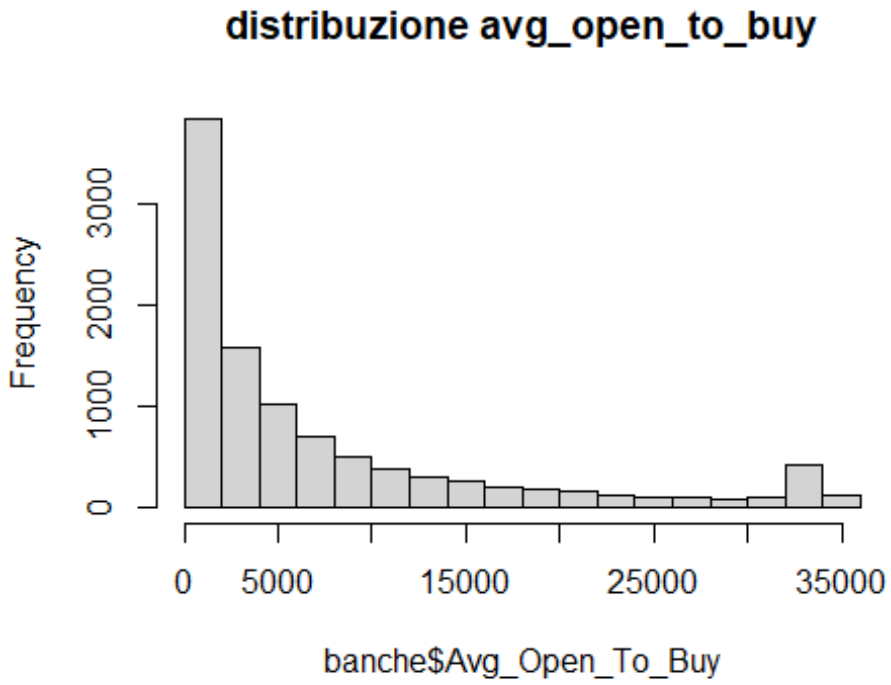

```r
boxplot(banche$Total_Revolving_Bal, horizontal = T)
```



Si osserva che la moda è 0

Variabile avg open to buy

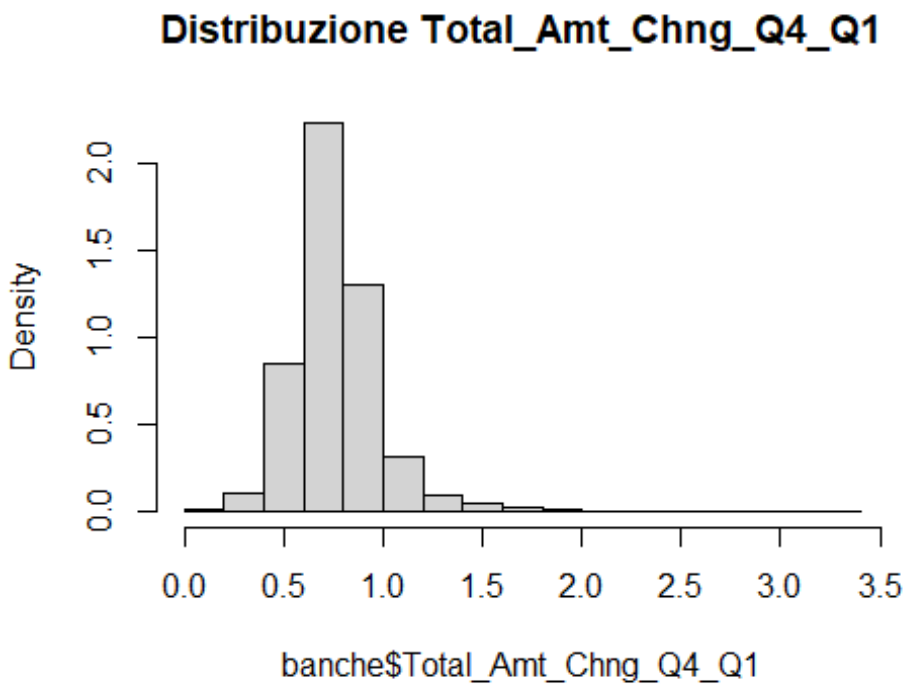
```
display_summary_and_var(banche$Total_Revolving_Bal)

##          Min.       1st Qu.        Median          Mean       3rd Qu.
##  0.000000e+00  3.590000e+02  1.276000e+03  1.162814e+03  1.784000e+03
##          Max.           var            sd            sk
##  2.517000e+03  6.642044e+05  8.149873e+02 -1.488152e-01

hist(banche$Avg_Open_To_Buy ,freq = T, main = "distribuzione avg_open_to_buy")
```



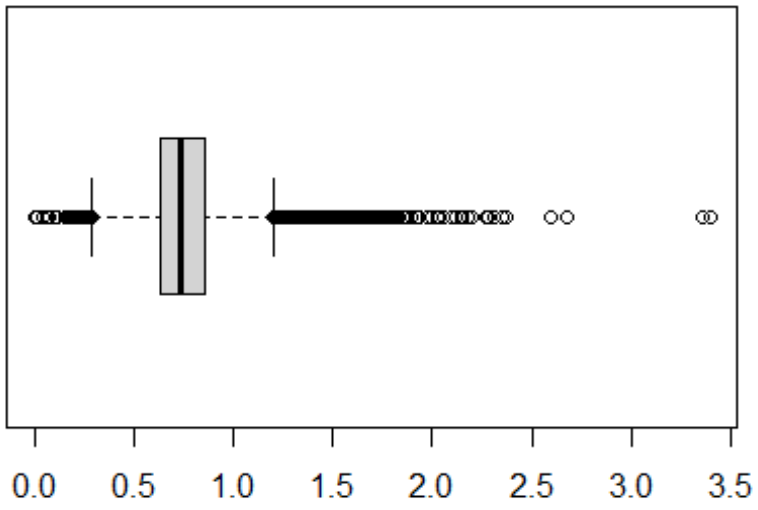
distribuzione asimetrica concentrata a sinistra

Variabile Total_Amt_Chng_Q4_Q1

```
display_summary_and_var(banche$Total_Amt_Chng_Q4_Q1)

##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.        var
## 0.00000000 0.63100000 0.73600000 0.75994065 0.85900000 3.39700000 0.04805161
##         sd         sk
## 0.21920677 1.73180685

hist(banche$Total_Amt_Chng_Q4_Q1, freq = F, main = "Distribuzione
Total_Amt_Chng_Q4_Q1")
```




```
boxplot(banche$Total_Amt_Chng_Q4_Q1, horizontal = T)
```



La variabile Total_Amt_Chng_Q4_Q1 è quantitativa e rappresenta il cambiamento nell'importo totale delle transazioni tra il quarto trimestre e il primo trimestre. La distribuzione ha una

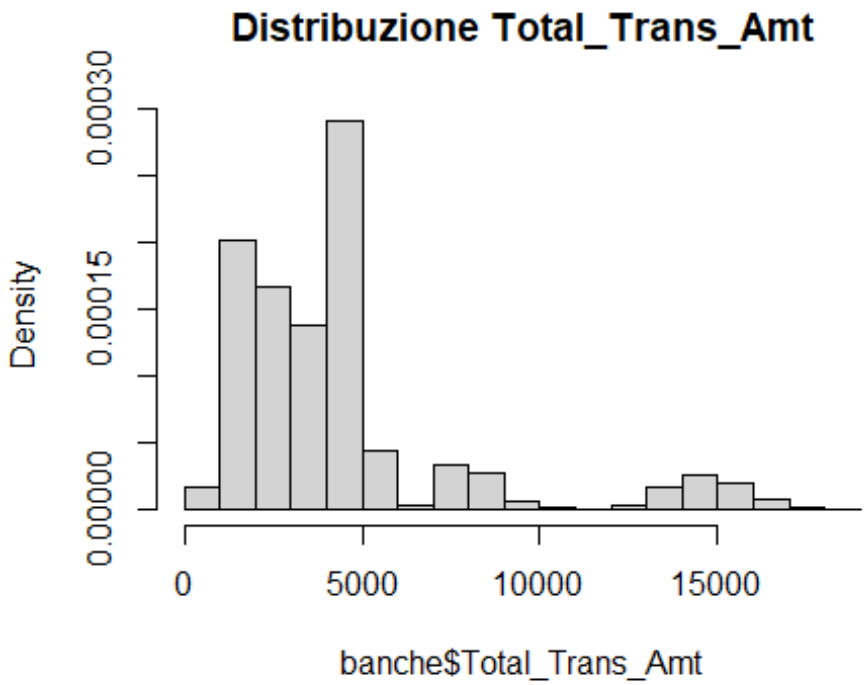
skewness positiva, indicando una coda lunga a destra. Questo significa che la maggior parte dei clienti ha avuto cambiamenti minori, con pochi clienti che hanno avuto grandi aumenti nell'importo delle transazioni.

Variabile Total_Trans_Amt

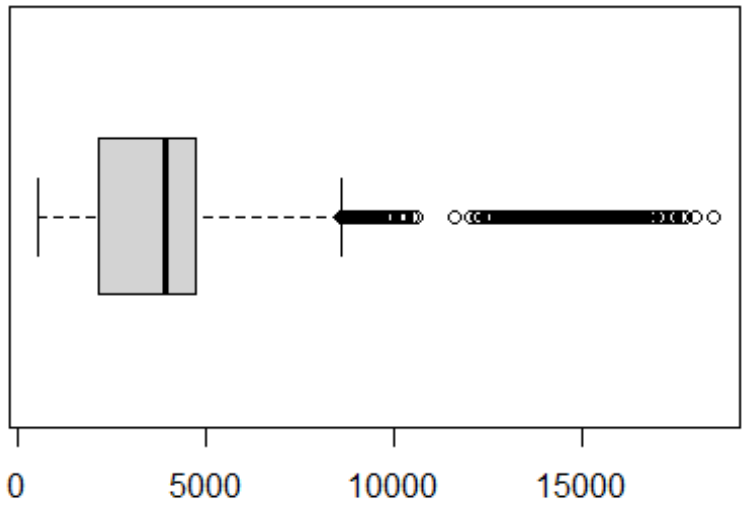
```
display_summary_and_var(banche$Total_Trans_Amt)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 5.100000e+02 2.155500e+03 3.899000e+03 4.404086e+03 4.741000e+03 1.848400e+04
##          var           sd           sk
## 1.154049e+07 3.397129e+03 2.040701e+00
```

```
hist(banche$Total_Trans_Amt, freq = F, main = "Distribuzione Total_Trans_Amt")
```




```
boxplot(banche$Total_Trans_Amt, horizontal = T)
```



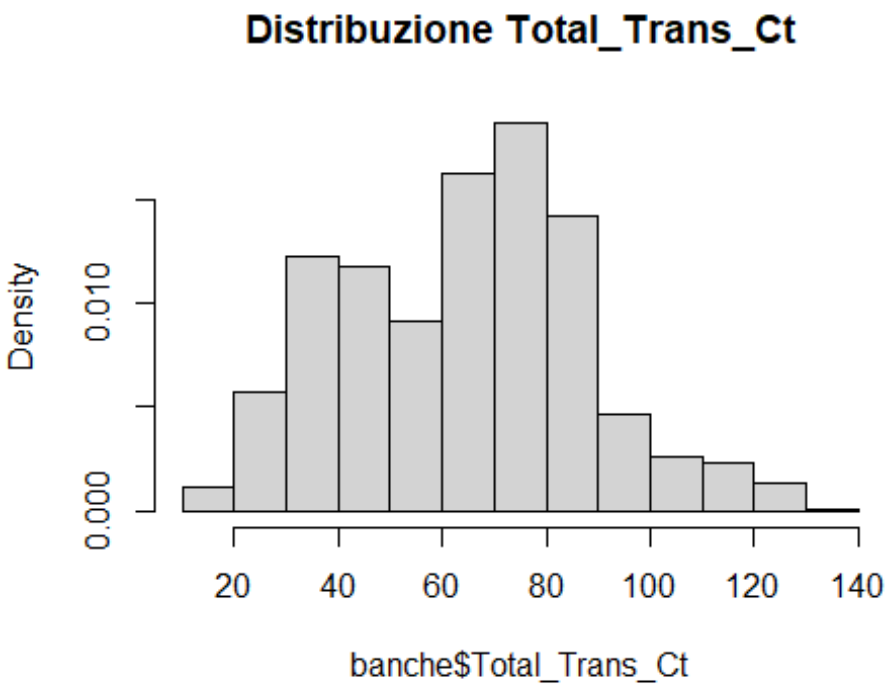
La variabile Total_Trans_Amt è quantitativa e rappresenta l'importo totale delle transazioni. Anche questa variabile ha una skewness positiva, indicando che la maggior parte dei clienti ha un importo totale delle transazioni relativamente basso, con alcuni che hanno importi significativamente più alti.

Variabile Total_Trans_Ct

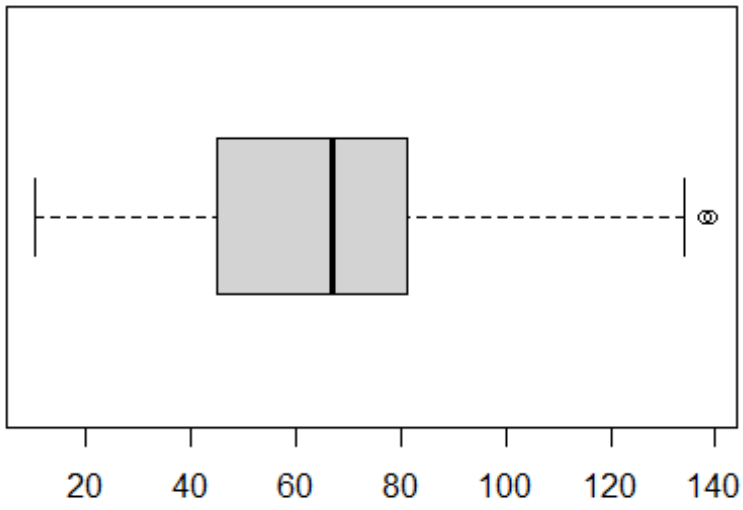
```
display_summary_and_var(banche$Total_Trans_Ct)
```

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##  10.0000000  45.0000000  67.0000000  64.8586946  81.0000000 139.0000000
##         var          sd          sk
## 550.9615635  23.4725704   0.1536503
```

```
hist(banche$Total_Trans_Ct, freq = F, main = "Distribuzione Total_Trans_Ct")
```




```r
boxplot(banche$Total_Trans_Ct, horizontal = T)
```



La variabile Total_Trans_Ct è quantitativa e rappresenta il numero totale delle transazioni. La distribuzione mostra una skewness positiva, indicando che la maggior parte dei clienti effettua

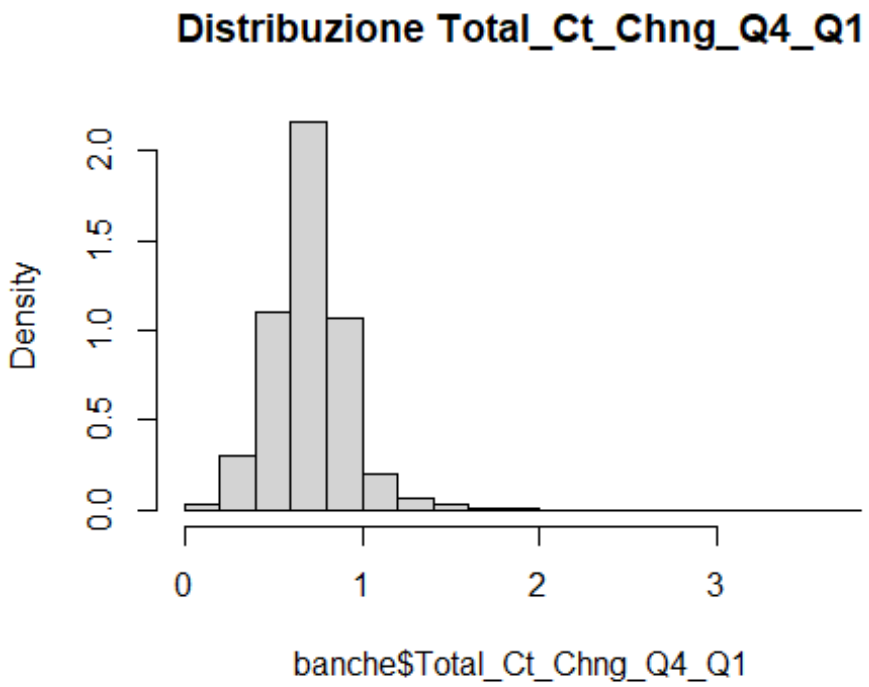
un numero relativamente basso di transazioni, mentre pochi clienti effettuano un numero molto alto di transazioni.

Variabile Total_Ct_Chng_Q4_Q1

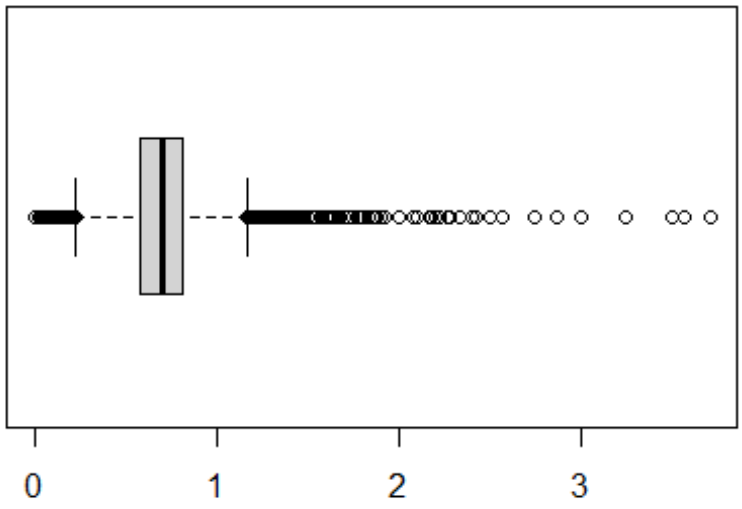
```
display_summary_and_var(banche$Total_Ct_Chng_Q4_Q1)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.        var
## 0.00000000 0.58200000 0.70200000 0.71222238 0.81800000 3.71400000 0.05668499
##         sd         sk
## 0.23808609 2.06372483
```

```
hist(banche$Total_Ct_Chng_Q4_Q1, freq = F, main = "Distribuzione
Total_Ct_Chng_Q4_Q1")
```




```
boxplot(banche$Total_Ct_Chng_Q4_Q1, horizontal = T)
```



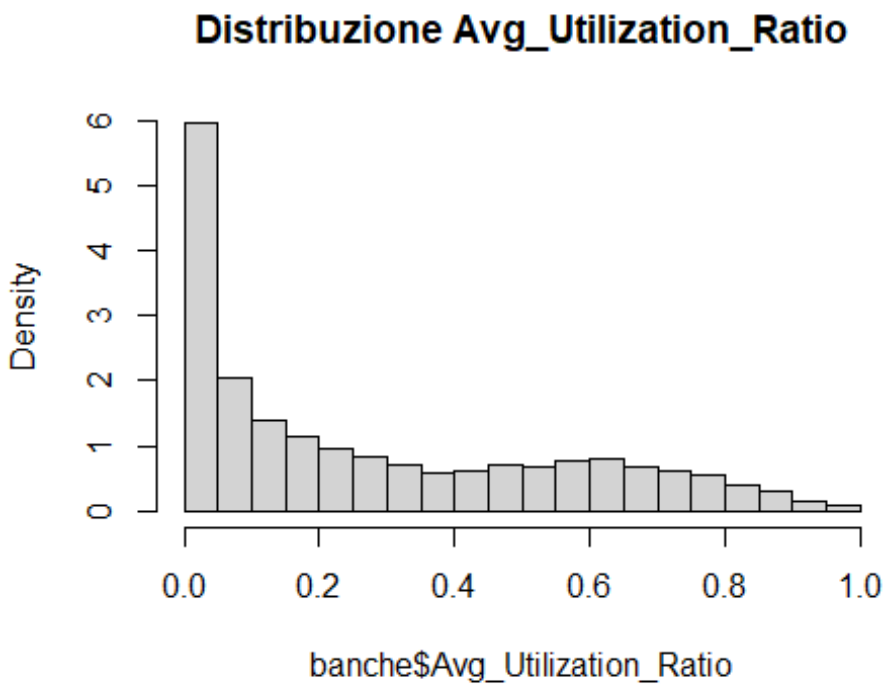
La variabile Total_Ct_Chng_Q4_Q1 rappresenta il cambiamento nel numero totale delle transazioni tra il quarto trimestre e il primo trimestre. Anche questa variabile mostra una skewness positiva, indicando una coda lunga a destra. La maggior parte dei clienti ha avuto cambiamenti minori nel numero di transazioni, con pochi che hanno avuto grandi aumenti.

Variabile Avg_Utilization_Ratio

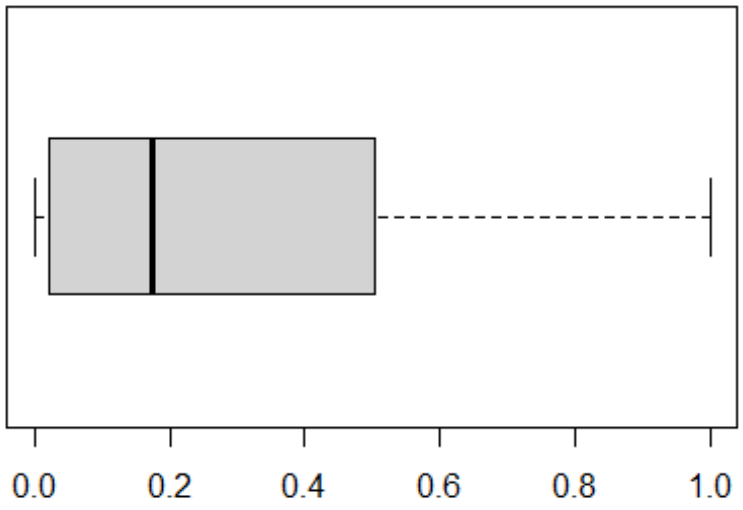
```
display_summary_and_var(banche$Avg_Utilization_Ratio)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.        var
## 0.00000000 0.02300000 0.17600000 0.27489355 0.50300000 0.99900000 0.07600579
##         sd         sk
## 0.27569147 0.71790164
```

```
hist(banche$Avg_Utilization_Ratio, freq = F, main = "Distribuzione
Avg_Utilization_Ratio")
```




```r
boxplot(banche$Avg_Utilization_Ratio, horizontal = T)
```



La variabile Avg_Utilization_Ratio è quantitativa e rappresenta il rapporto medio di utilizzo della carta di credito. La distribuzione ha una skewness positiva, indicando che la maggior parte dei

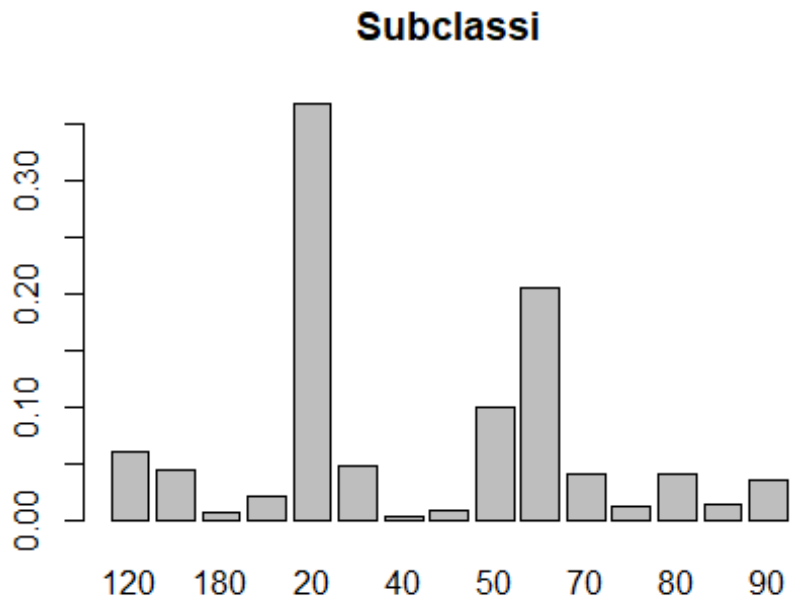
clienti utilizza una piccola porzione del proprio credito disponibile, con pochi che utilizzano una grande porzione.

Dataset - House Prices

Analisi univariata

Variabile MSSubClass

```
case$MSSubClass <- factor(replace(case$MSSubClass, is.na(case$MSSubClass), "Non Presente"))
display_table(case$MSSubClass, "Subclassi")
```

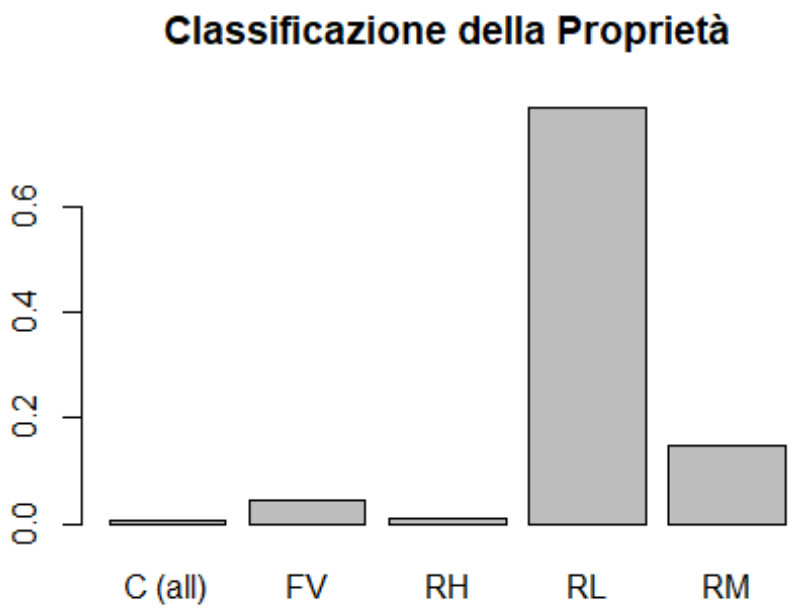



```
##                120         160          180         190          20          30
## DistAs 87.00000000 63.00000000 10.000000000 30.00000000 536.0000000 69.00000000
## DistRe  0.05958904  0.04315068  0.006849315  0.02054795   0.3671233  0.04726027
##                 40           45           50          60          70          75
## DistAs 4.000000000 12.000000000 144.00000000 299.0000000 60.00000000 16.0000000
## DistRe 0.002739726  0.008219178   0.09863014   0.2047945  0.04109589  0.0109589
##                 80          85          90
## DistAs 58.00000000 20.00000000 52.00000000
## DistRe  0.03972603  0.01369863  0.03561644
```

Una Variabile Qualitativa che descrive il tipo di abitazione della proprietà in vendita. Questi tipi di abitazione sono in totale 16 e vengono indicati, per brevità, usando dei numeri. I numeri delle classi non possono essere quindi usati per calcolare medie o mediane. La variabile non è ben distribuita, infatti solo le classi “20” e “60” comprendono il 57.20% delle osservazioni totali. La classe meno presente è invece “40” con solo 4 osservazioni totali (<0.3%).

Variabile MSZoning

```
case$MSZoning <- factor(replace(case$MSZoning, is.na(case$MSZoning), "Non
Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

display_table(case$MSZoning,"Classificazione della Proprietà")
```



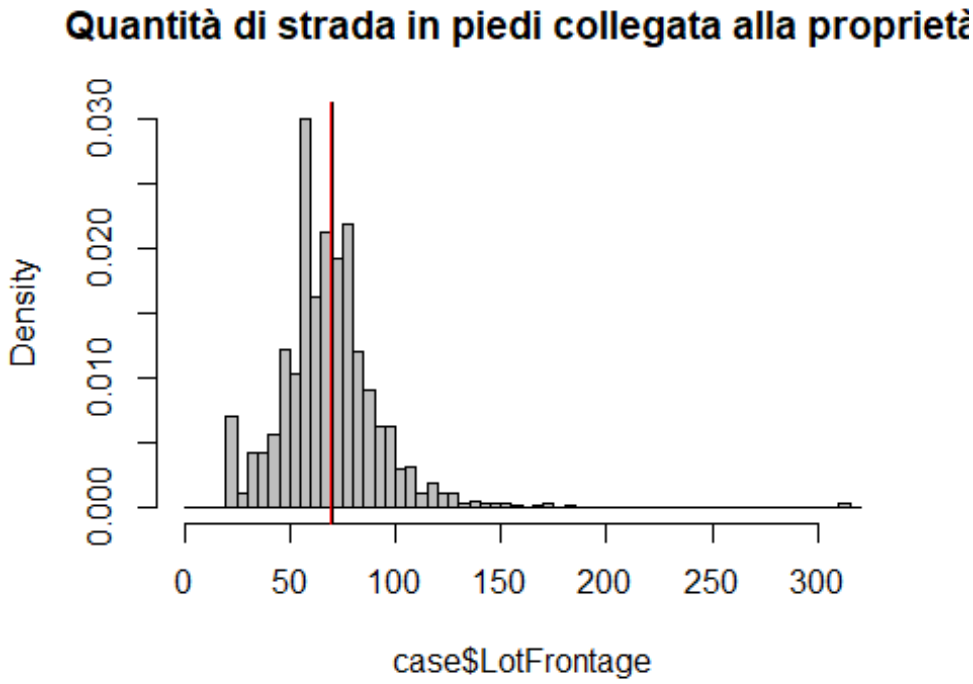

```
##             C (all)          FV         RH           RL          RM
## DistAs 10.000000000 65.00000000 16.0000000 1151.0000000 218.0000000
## DistRe  0.006849315  0.04452055  0.0109589    0.7883562   0.1493151
```

Una Variabile Quantitativa che indica il tipo di classificazione della proprietà in vendita. Sono presenti un totale di 5 diversi tipi indicati con una sigla di al più 2 lettere. La variabile non è distribuita uniformemente con "RL" (Residenziale a bassa densità) che comprende il 78.73% delle osservazioni totali. La classe meno presente è "C" (Commerciale) che viene vista sole 10 volte (<0.7%).

Variabile LotFrontage

```
display_summary_and_var(case$LotFrontage)

##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.       NA's
##  21.000000  59.000000  69.000000  70.049958  80.000000 313.000000 259.000000
##        var         sd         sk
## 589.749169  24.284752   2.160866

hist(case$LotFrontage, probability = T, breaks =c(5*0:64),col = "gray", main =
"Quantità di strada in piedi collegata alla proprietà")
```

```
abline(v = median(case$LotFrontage, na.rm = T),lwd = 1, col = "red")
abline(v = mean(case$LotFrontage, na.rm = T),lwd = 1)
```



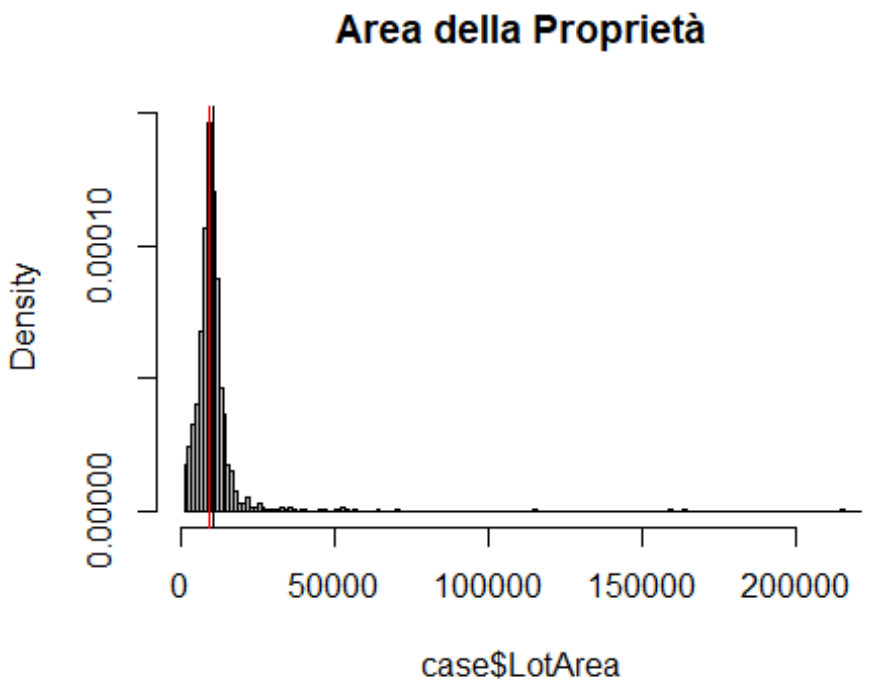
Variabile Quantitativa che salva la quantità di strada a contatto con la proprietà. Sono presenti in questo caso 259 valori mancanti. I valori di essa vanno da un minimo di 21 ad un massimo di 313. Con questi ultimi che sono valori estremi per questa variabile essendo ogni altro valore assunto da essa minore di 200. Media e Mediana sono molto vicine, entrambe a circa 70. Si nota infatti che la Skewness è in questo caso molto bassa a 2.1 . Dal grafico vediamo che il picco di valori si trova nel range 55-80, che contiene sia il primo che terzo Quantile.

Variabile LotArea

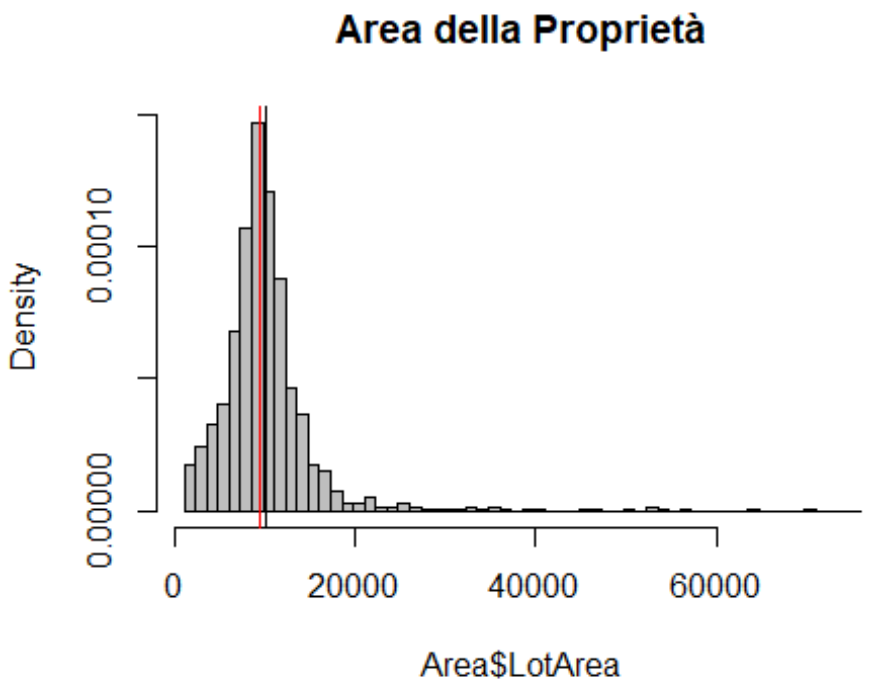
```
display_summary_and_var(case$LotArea)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max. 
## 1.300000e+03 7.553500e+03 9.478500e+03 1.051683e+04 1.160150e+04 2.152450e+05 
##          var           sd           sk 
## 9.962565e+07 9.981265e+03 1.219514e+01
```

```
hist(case$LotArea, probability = T, breaks = c(1250*0:176)+1000,col = "gray", main
= "Area della Proprietà")
abline(v = median(case$LotArea, na.rm = T),lwd = 1, col = "red")
abline(v = mean(case$LotArea, na.rm = T),lwd = 1)
```



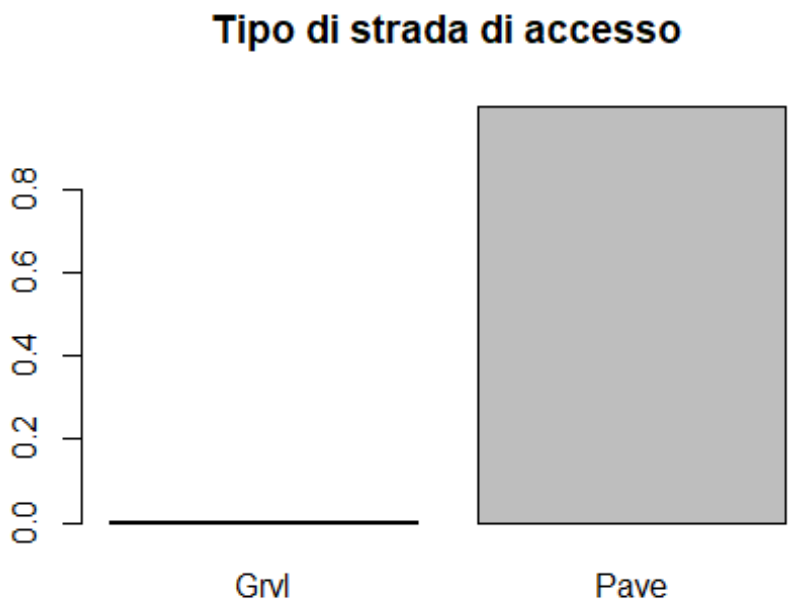

```r
#LotArea: Senza valori estremi
Area <- case[case$LotArea < 100000,]
hist(Area$LotArea, probability = T, breaks = c(1250*0:60)+1000,col = "gray", main
= "Area della Proprietà")
abline(v = median(Area$LotArea, na.rm = T),lwd = 1, col = "red")
abline(v = mean(Area$LotArea, na.rm = T),lwd = 1)
```



Variabile Quantitativa che indica l'area della proprietà. Ha un range di valori molto alto dovuto principalmente alla presenza di valori estremi che arrivano ad un massimo di 2.152450e+05. La gran parte delle osservazioni rimane sotto i 25000 metri quadrati con un alta concentrazione tra 7500 e 11600. Media e mediana molto vicine tra loro e skewness, infatti, molto bassa soprattutto rispetto ai valori che assume "LotArea". Si nota che i valori estremi che va variabile assume sono riflessi in una varianza e deviazione standard elevati.

Variabile Street

```
case$Street <- factor(replace(case$Street, is.na(case$Street), "Non Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

display_table(case$Street, "Tipo di strada di accesso")
```

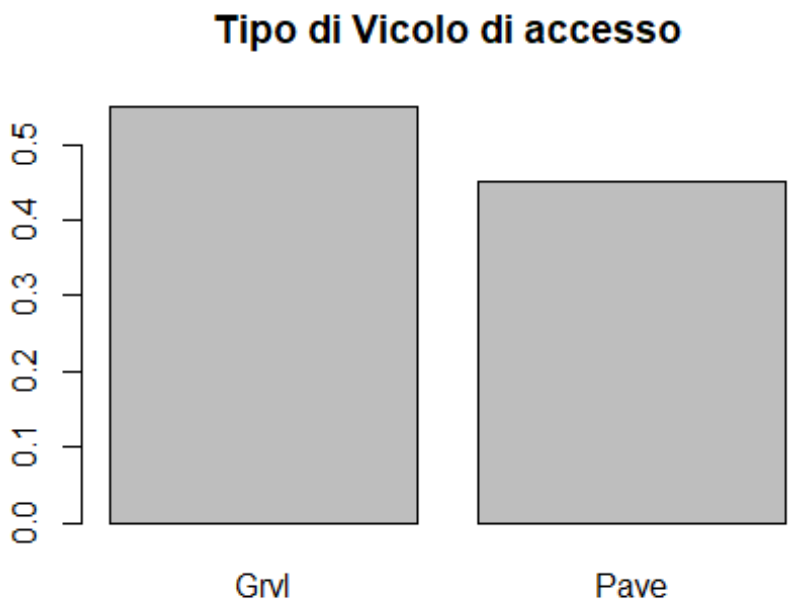



```
##                Grvl         Pave
## DistAs 6.000000000 1454.0000000
## DistRe 0.004109589    0.9958904
```

Variabile Qualitativa che indica il tipo di strada di accesso alla proprietà. Sono presenti solo due tipi di Strada di accesso e il 99.59% delle osservazioni è del tipo “Pave”.

Variabile Alley

```
case$Alley <- factor(replace(case$Alley, is.na(case$Alley), "Non Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

display_table(case$Alley, "Tipo di Vicolo di accesso")
```

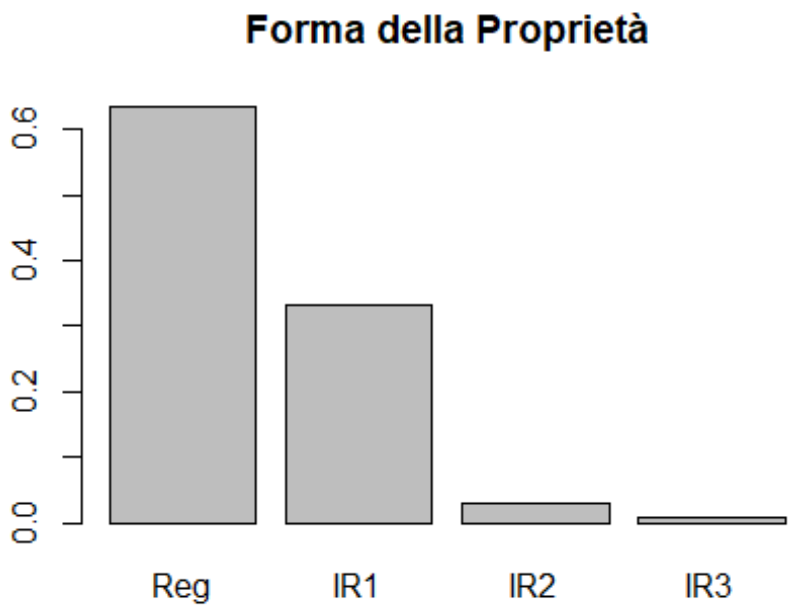



```
##              Grvl       Pave
## DistAs 50.0000000 41.0000000
## DistRe  0.5494505  0.4505495
```

Variabile Qualitativa che indica il tipo del vicolo di accesso alla proprietà. Vediamo che il 93.77% delle abitazioni non presenta un vicolo di accesso mentre le restanti lo hanno o pavimentato o in ghiaia.

Variabile LotShape

```
case$LotShape <- factor(replace(case$LotShape, is.na(case$LotShape), "Non
Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

case$LotShape <- ordered(case$LotShape,levels = c("Reg","IR1","IR2","IR3"))
display_table(case$LotShape, "Forma della Proprietà")
```

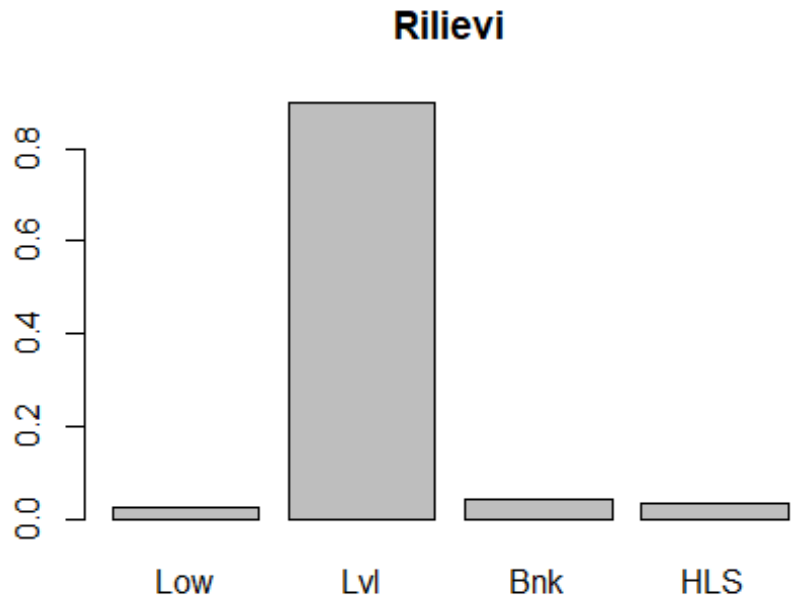



```
##                Reg         IR1         IR2          IR3
## DistAs 925.0000000 484.0000000 41.00000000 10.000000000
## DistRe   0.6335616   0.3315068  0.02808219  0.006849315
```

Variabile Qualitativa che descrive la forma generale della proprietà. Sono presenti 4 tipi di forma che vanno da quella "Regolare" a "IR3", ovvero altamente irregolare. La forma più comune è quella "Regolare" che viene osservata più spesso delle tre forme Irregolari combinate.

Variabile LandContour

```
case$LandContour <- factor(replace(case$LandContour, is.na(case$LandContour), "Non Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

case$LandContour <- ordered(case$LandContour,levels = c("Low","Lvl","Bnk","HLS"))
display_table(case$LandContour,"Rilievi")
```

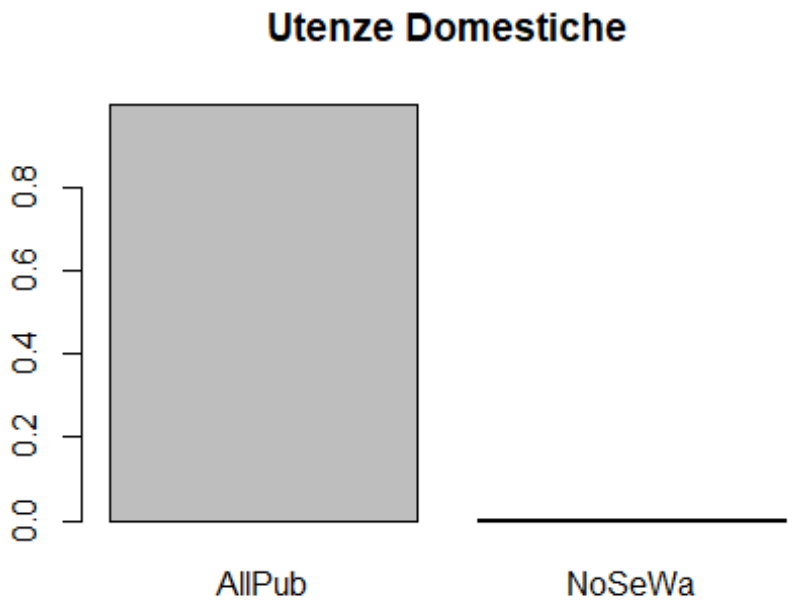



```
##                Low          Lvl         Bnk         HLS
## DistAs 36.00000000 1311.0000000 63.00000000 50.00000000
## DistRe  0.02465753    0.8979452  0.04315068  0.03424658
```

Variabile Qualitativa che descrive i rilievi presenti sulla proprietà dividendoli in 4 possibili categorie. Questo è un alto caso in cui una di queste categorie comprende un numero estremamente elevato di osservazioni rispetto alle altre. “Lvl”, ovvero “a livello con il terreno”, comprende quasi il 90% delle proprietà osservate.

Variabile Utilities

```
case$Utilities <- factor(replace(case$Utilities, is.na(case$Utilities), "Non
Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

display_table(case$Utilities, "Utenze Domestiche")
```

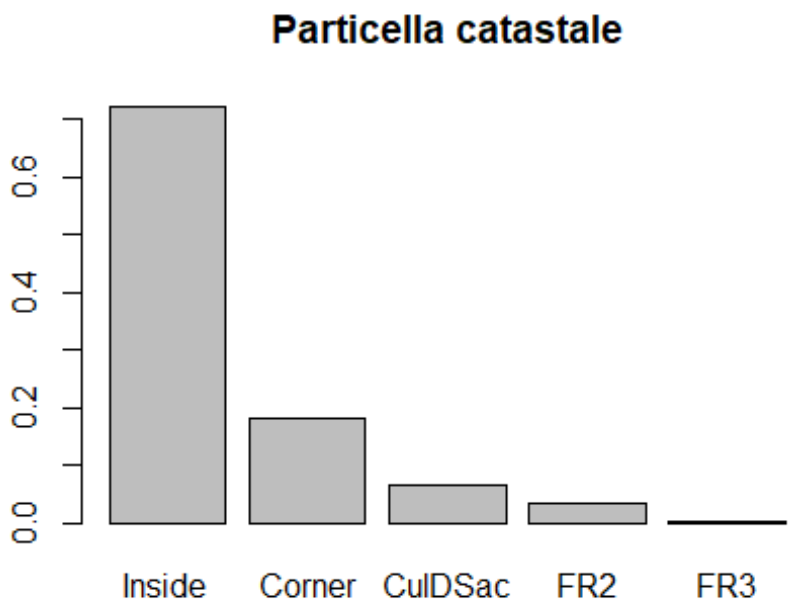



```
##               AllPub       NoSeWa
## DistAs 1459.0000000 1.0000000000
## DistRe    0.9993151 0.0006849315
```

Variabile qualitativa che descrive le utenze domestiche presenti nella proprietà.
Dei 4 valori della variabile possibili sono solo presenti “AllPub” e “NoSeWa”. “NoSeWa” è stata inoltre osservata solo una volta (<0.1%)

Variabile LotConfig

```
case$LotConfig <- factor(replace(case$LotConfig, is.na(case$LotConfig), "Non
Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

case$LotConfig <- ordered(case$LotConfig, levels =
c("Inside","Corner","CulDSac","FR2","FR3"))
display_table(case$LotConfig, "Particella catastale")
```

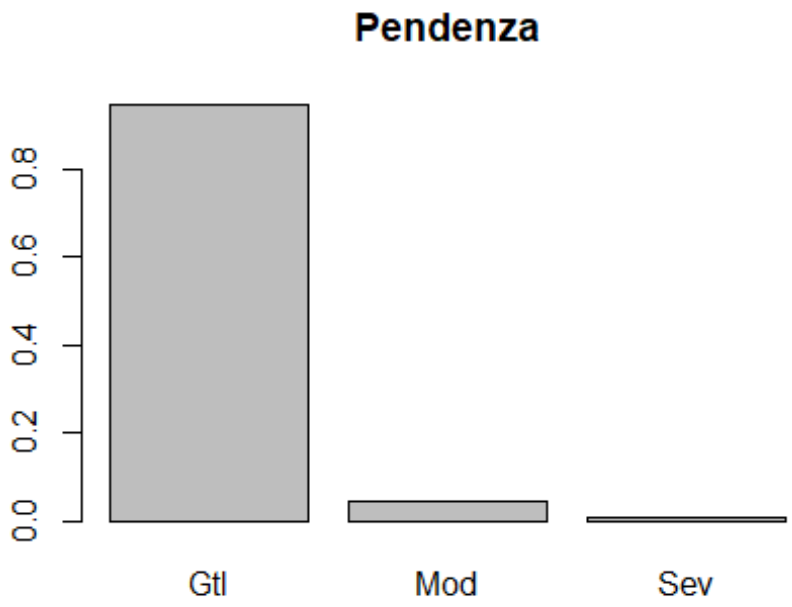



```
##              Inside     Corner     CulDSac         FR2         FR3
## DistAs 1052.0000000 263.000000 94.00000000 47.00000000 4.000000000
## DistRe    0.7205479   0.180137  0.06438356  0.03219178 0.002739726
```

Variabile Qualitativa che descrive il tipo della particella catastale. “Inside” è il valore più comune della variabile e corrisponde al 72% delle osservazioni totali. “FR3”, invece, è presente solo 4 volte.

Variabile LandSlope

```
case$LandSlope <- factor(replace(case$LandSlope, is.na(case$LandSlope), "Non
Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

display_table(case$LandSlope, "Pendenza")
```

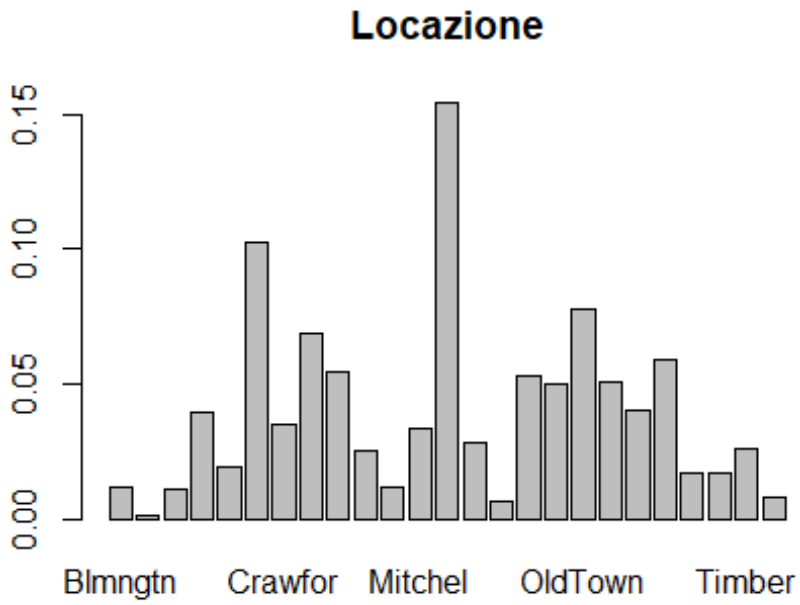



```
##                  Gtl         Mod         Sev
## DistAs 1382.0000000 65.00000000 13.00000000
## DistRe    0.9465753  0.04452055  0.00890411
```

Variabile Qualitativa che descrive la pendenza del terreno su cui è costruita la proprietà dividendola in 3 possibili valori dal Gentile ("Gtl") al Severo ("Sev"). Il valore più visto è "Gtl" con il 94.65% delle osservazioni.

Variabile Neighborhood

```
case$Neighborhood <- factor(replace(case$Neighborhood, is.na(case$Neighborhood),
"Non Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

display_table(case$Neighborhood,"Locazione")
```

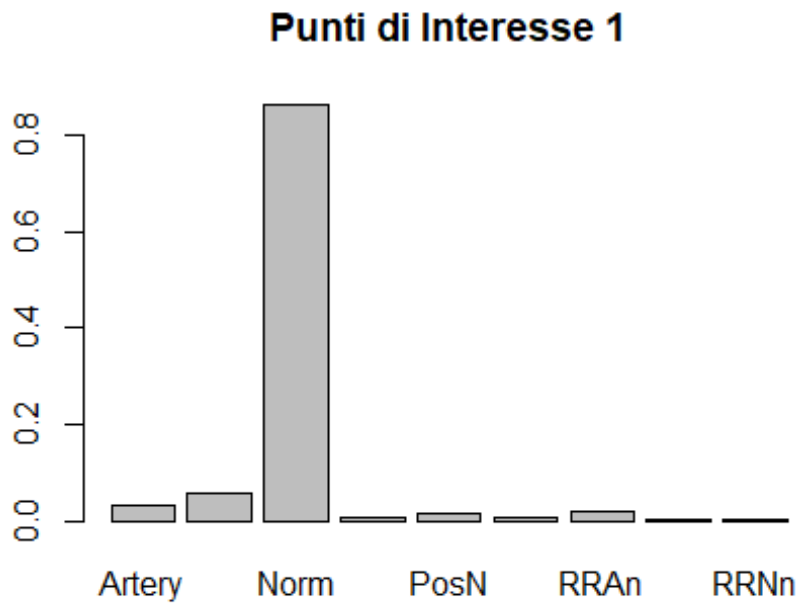



```
##            Blmngtn     Blueste     BrDale     BrkSide     ClearCr     CollgCr
## DistAs 17.00000000 2.000000000 16.0000000 58.00000000 28.00000000 150.0000000
## DistRe  0.01164384 0.001369863  0.0109589  0.03972603  0.01917808   0.1027397
##            Crawfor      Edwards     Gilbert      IDOTRR     MeadowV     Mitchel
## DistAs 51.00000000 100.00000000 79.00000000 37.00000000 17.00000000 49.00000000
## DistRe  0.03493151   0.06849315  0.05410959  0.02534247  0.01164384  0.03356164
##              NAmes     NoRidge     NPkVill     NridgHt NWAmes      OldTown
## DistAs 225.0000000 41.00000000 9.000000000 77.00000000  73.00 113.00000000
## DistRe   0.1541096  0.02808219 0.006164384  0.05273973   0.05   0.07739726
##             Sawyer     SawyerW     Somerst     StoneBr       SWISU     Timber
## DistAs 74.00000000 59.00000000 86.00000000 25.00000000 25.00000000 38.0000000
## DistRe  0.05068493  0.04041096  0.05890411  0.01712329  0.01712329  0.0260274
##             Veenker
## DistAs 11.000000000
## DistRe  0.007534247
```

Variabile Qualitativa che descrive la locazione della proprietà in Ames city. In questo caso abbiamo 25 possibili valori per questa variabile con NAmes il più frequente a 15.43% delle osservazioni e Blueste il meno a sole 2 ossecrazioni (<0.2%).

Variabile Condition1

```
case$Condition1 <- factor(replace(case$Condition1, is.na(case$Condition1), "Non
Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

display_table(case$Condition1, "Punti di Interesse 1")
```

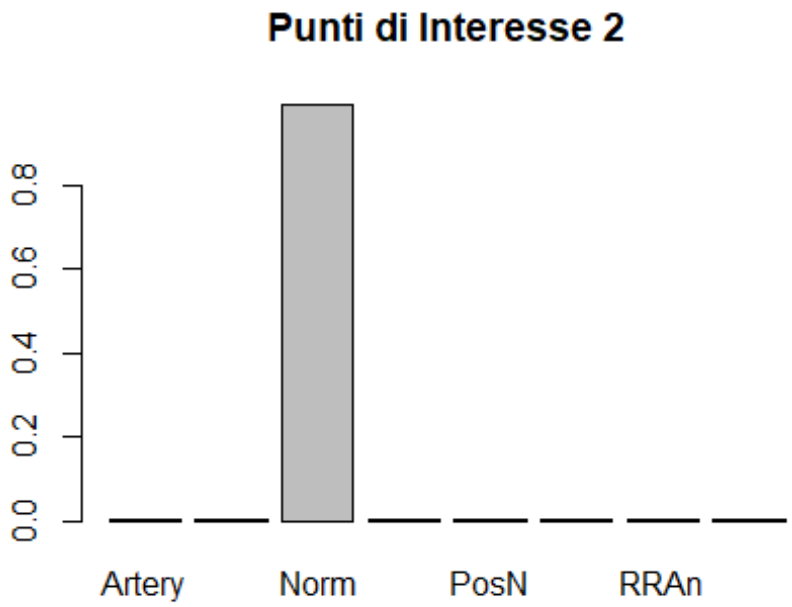



```
##             Artery       Feedr         Norm        PosA       PosN         RRAe
## DistAs 48.00000000 81.00000000 1260.0000000 8.000000000 19.0000000 11.000000000
## DistRe  0.03287671  0.05547945    0.8630137 0.005479452  0.0130137  0.007534247
##               RRAn        RRNe        RRNn
## DistAs 26.00000000 2.000000000 5.000000000
## DistRe  0.01780822 0.001369863 0.003424658
```

Variabile Qualitativa che indica se la proprietà è vicina a uno tra 8 tipi di punti di interesse o se è lontana da tutti essi, questo è il caso "Norm". Il valore più comune è "Norm" con l'86.3% dei dati osservati.

Variabile Condition2

```
case$Condition2 <- factor(replace(case$Condition2, is.na(case$Condition2), "Non
Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

display_table(case$Condition2, "Punti di Interesse 2")
```

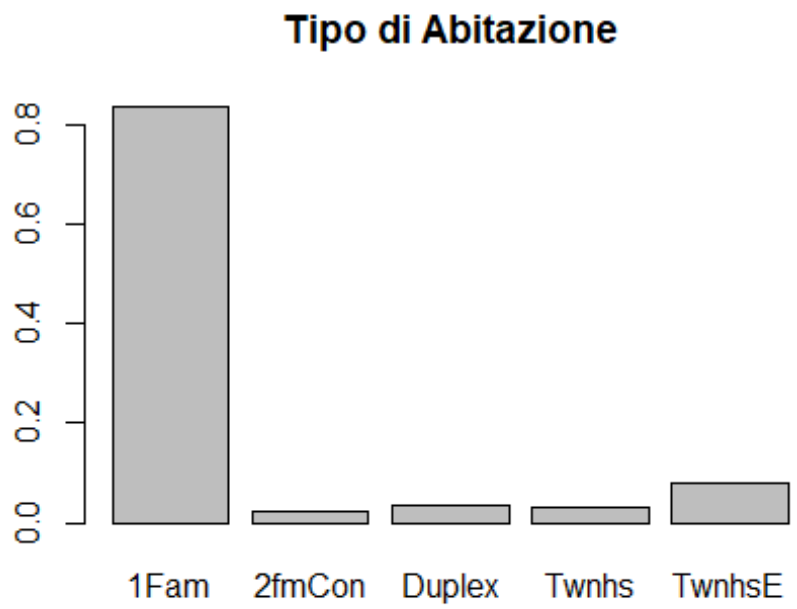



```
##              Artery       Feedr        Norm         PosA        PosN
## DistAs 2.000000000 6.000000000 1445.000000 1.0000000000 2.000000000
## DistRe 0.001369863 0.004109589    0.989726 0.0006849315 0.001369863
##                RRAe         RRAn        RRNn
## DistAs 1.0000000000 1.0000000000 2.000000000
## DistRe 0.0006849315 0.0006849315 0.001369863
```

Variabile Qualitativa che indica la vicinanza della abitazione a ulteriori punti di interesse. “Norm” rimane il valore della variabile più comune con adesso il 98.97% di tutte le osservazioni mentre gli altri valori ne hanno, in totale, solo 15. Essedo presenti solo 2 variabili che mi indicano la vicinanza a punti di interesse,non possiamo sapere se le 15 abitazioni che sono vicine a 2 punti di interesse non siano vicine anche a 3 o più di essi.

Variabile BldgType

```
case$BldgType <- factor(replace(case$BldgType, is.na(case$BldgType), "Non
Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

display_table(case$BldgType, "Tipo di Abitazione")
```

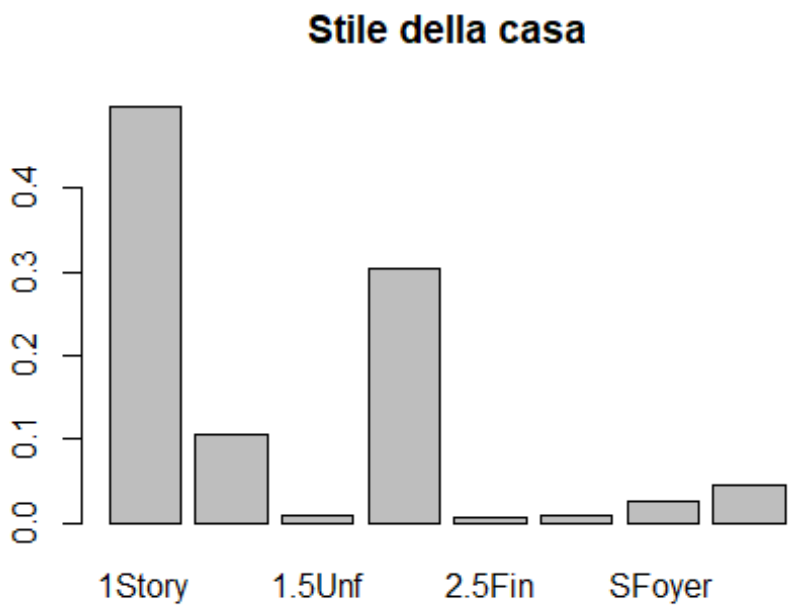



```
##                1Fam      2fmCon      Duplex       Twnhs       TwnhsE
## DistAs 1220.0000000 31.00000000 52.00000000 43.00000000 114.00000000
## DistRe    0.8356164  0.02123288  0.03561644  0.02945205   0.07808219
```

Variabile Qualitativa che descrive che tipo di abitazione è quella in vendita. Il gruppo più comune è quello delle abitazioni da Singola Famiglia ("1Fam") con l'83.56% delle osservazioni totali

Variabile HouseStyle

```
case$HouseStyle <- factor(replace(case$HouseStyle, is.na(case$HouseStyle), "Non
Presente"))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

case$HouseStyle <- ordered(case$HouseStyle, levels =
c("1Story","1.5Fin","1.5Unf","2Story","2.5Fin","2.5Unf","SFoyer","SLvl"))
display_table(case$HouseStyle, "Stile della casa")
```




```
##             1Story      1.5Fin       1.5Unf      2Story      2.5Fin
## DistAs 726.0000000 154.0000000 14.000000000 445.0000000 8.000000000
## DistRe   0.4972603   0.1054795  0.009589041   0.3047945 0.005479452
##              2.5Unf      SFoyer        SLvl
## DistAs 11.000000000 37.00000000 65.00000000
## DistRe  0.007534247  0.02534247  0.04452055
```

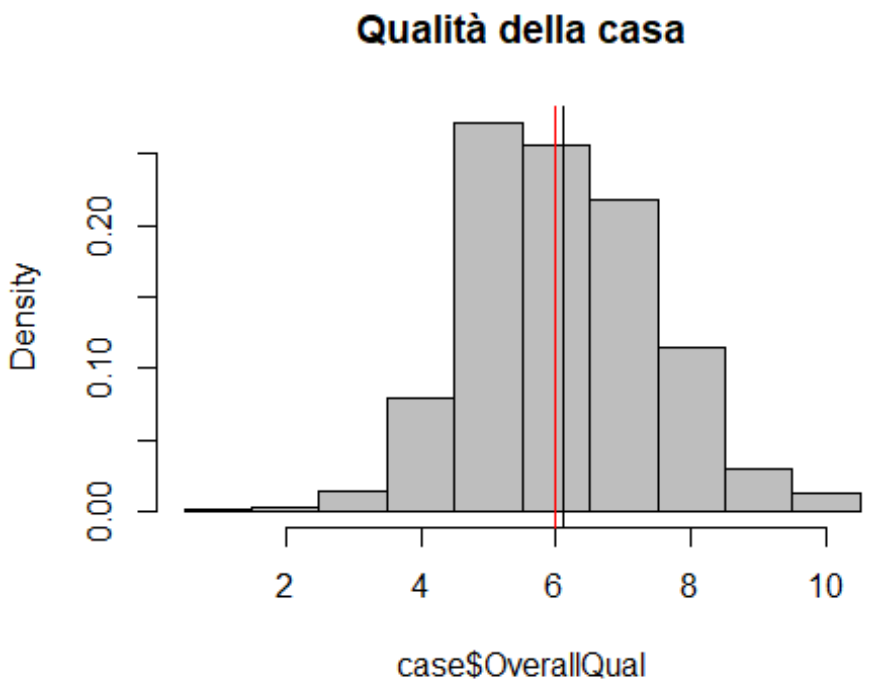
Variabile Qualitativa che descrive lo stile o tipo dell'edificio. Vediamo che edifici con mezzi piani sono molto più rari e che gli edifici con meno piani sono più numerosi. La moda è infatti il gruppo "1Story", al 49.72% delle osservazioni totali, seguito da "2Story", al 30.47%. "2.5Fin" e "2.5Unf" sono invece i più rari.

Variabile OverallQual

```
display_summary_and_var(case$OverallQual)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.       var        sd
##  1.000000  5.000000  6.000000  6.099315  7.000000 10.000000  1.912679  1.382997
##        sk
##  0.216721
```

```
hist(case$OverallQual, probability = T,breaks= c(0:10)+0.5,col = "gray", main =
"Qualità della casa")
abline(v = median(case$OverallQual, na.rm = T),lwd = 1, col = "red")
abline(v = mean(case$OverallQual, na.rm = T),lwd = 1)
```



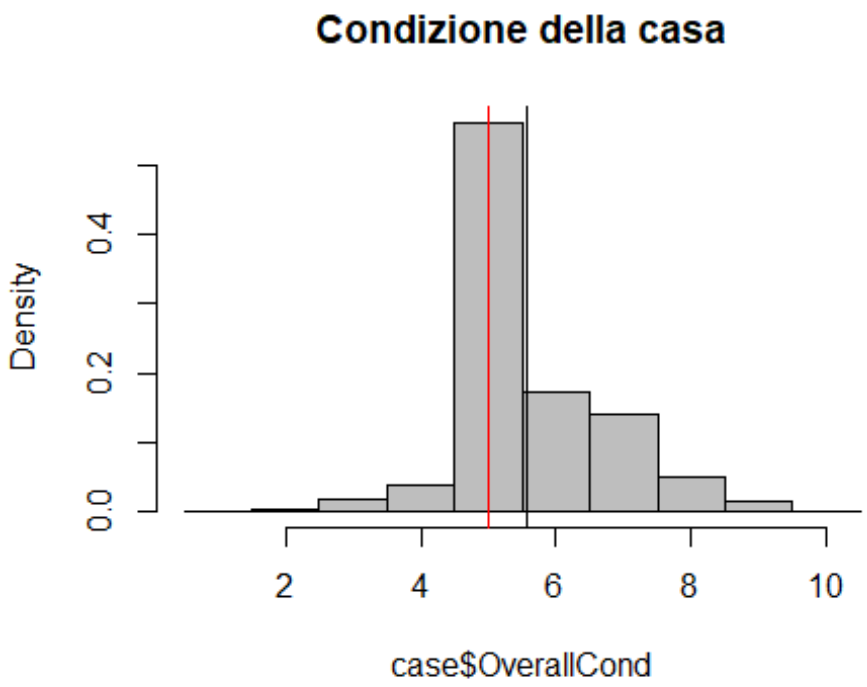
Variabile Quantitativa che descrive la qualità dei materiali e della finitura della casa in una scala di interi che va da 1 a 10. Può anche essere vista come una variabile Qualitativa ordinabile divisa in 10 classi. Si nota che la maggior parte delle case si trova della fascia che va da "Avarege" (5) a "Good" (7). Con media e mediana entrambe vicino a "Above Avarege" (6). La Skewness è infatti bassa a 0.216721.

Variabile OverallCond

```
display_summary_and_var(case$OverallCond)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.       var        sd
## 1.0000000 5.0000000 5.0000000 5.5753425 6.0000000 9.0000000 1.2383224 1.1127993
##        sk
## 0.6923552
```

```
hist(case$OverallCond, probability = T,breaks= c(0:10)+0.5,col = "gray", main =
"Condizione della casa")
abline(v = median(case$OverallCond, na.rm = T),lwd = 1, col = "red")
abline(v = mean(case$OverallCond, na.rm = T),lwd = 1)
```



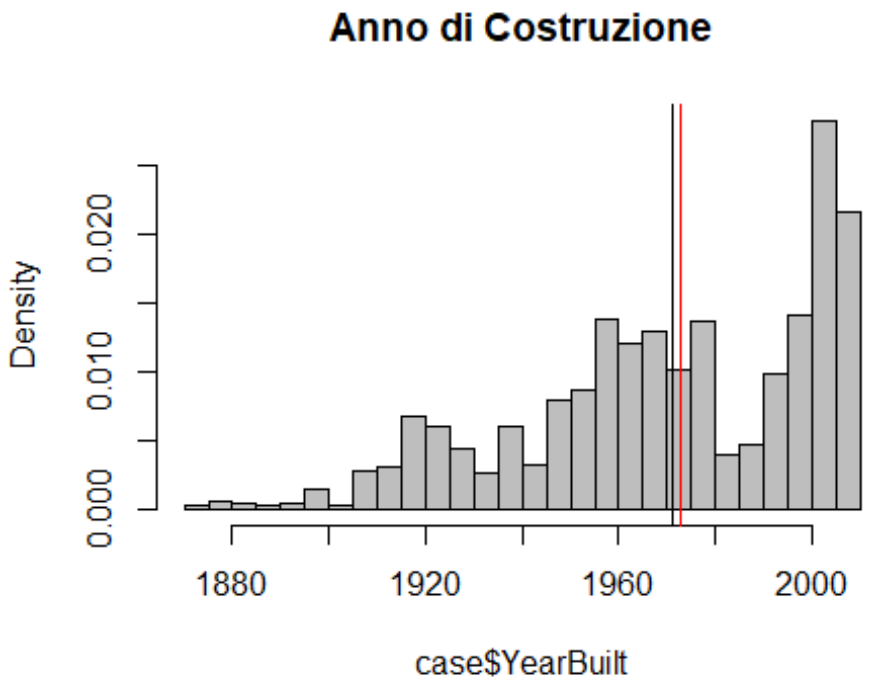
"OverallCond" è una Variabile Quantitativa che descrive le condizioni in cui si trova l'abitazione in una scala di interi che va da 1 a 10. Come "OverallQual", può anche essere vista come una variabile Qualitativa ordinabile divisa in 10 classi. Si vede anche dal grafico che la moda è "5" con più del 50%delle osservazioni. Si nota che media e mediana non sono veramente vicine, infatti la Skewness è 0.6923552 che, considerato che il range di valori è 1-10, non è bassa.

Variabile YearBuilt

```
display_summary_and_var(case$YearBuilt)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 1872.0000000 1954.0000000 1973.0000000 1971.2678082 2000.0000000 2010.0000000
##          var           sd           sk
##  912.2154126   30.2029040   -0.6128307
```

```
hist(case$YearBuilt, probability = T,breaks= c(5*0:28)+1870,col = "gray", main =
"Anno di Costruzione")
abline(v = median(case$YearBuilt, na.rm = T),lwd = 1, col = "red")
abline(v = mean(case$YearBuilt, na.rm = T),lwd = 1)
```



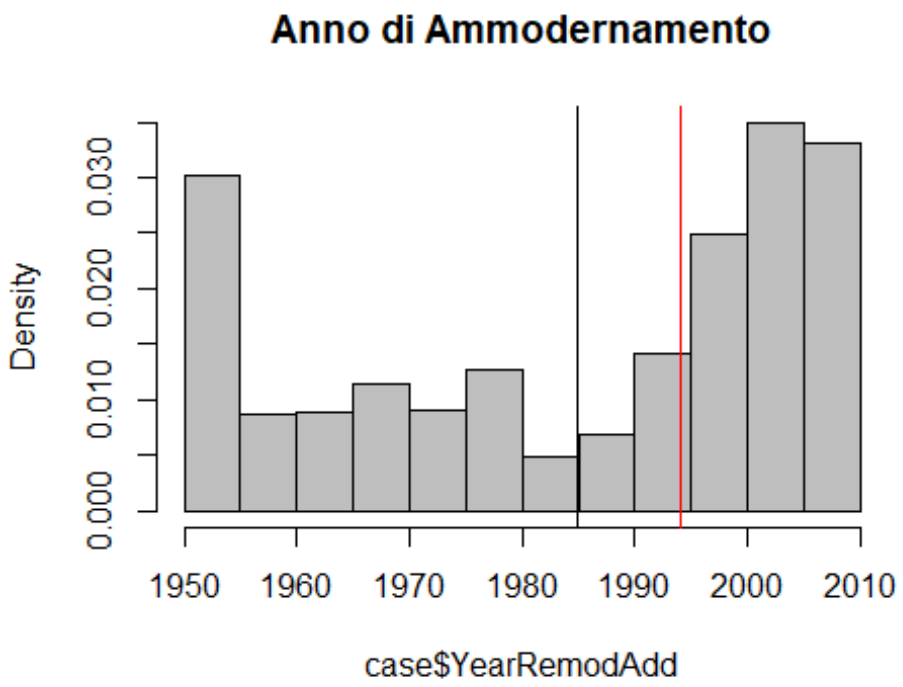
Variabile Quantitativa che descrive l'anno in cui le abitazioni sono state costruite. Il range va dal 1872 fino il 2010 con una quantità di case costruite maggiore all'avanzare degli anni. Si nota che il picco è avvenuto nei primi anni del 2000 e che tra il 1980 e 1990, durante la recessione globale, è avvenuto un forte calo nella costruzione di nuovi edifici. Questo anche maggiore dei cali avvenuti durante le guerre mondiali. Media e Mediana sono vicine mentre la Deviazione Standard e Varianza non sono basse.

Variabile YearRemodAdd

```
display_summary_and_var(case$YearRemodAdd)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 1950.0000000 1967.0000000 1994.0000000 1984.8657534 2004.0000000 2010.0000000
##          var           sd           sk
##  426.2328223   20.6454068   -0.5030445
```

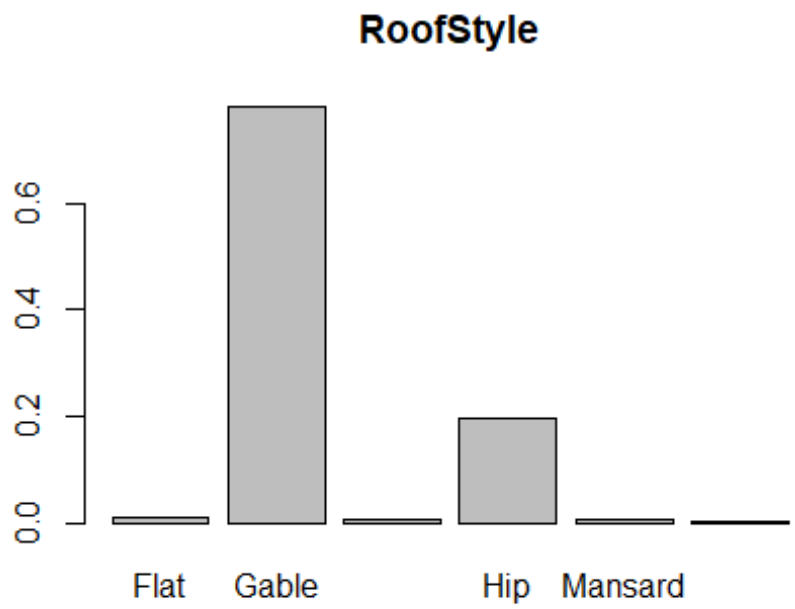
```
hist(case$YearRemodAdd, probability = T,col = "gray", main = "Anno di
Ammodernamento")
abline(v = median(case$YearRemodAdd, na.rm = T),lwd = 1, col = "red")
abline(v = mean(case$YearRemodAdd, na.rm = T),lwd = 1)
```



La Variabile è quantitativa e indica l'anno di ammodernamento dell'abitazione si nota che la media e la varianza sono diverse con una differenza di 10 anni

Variabile RoofStyle

```
case$RoofStyle <- factor(case$RoofStyle)
display_table(case$RoofStyle, "RoofStyle")
```

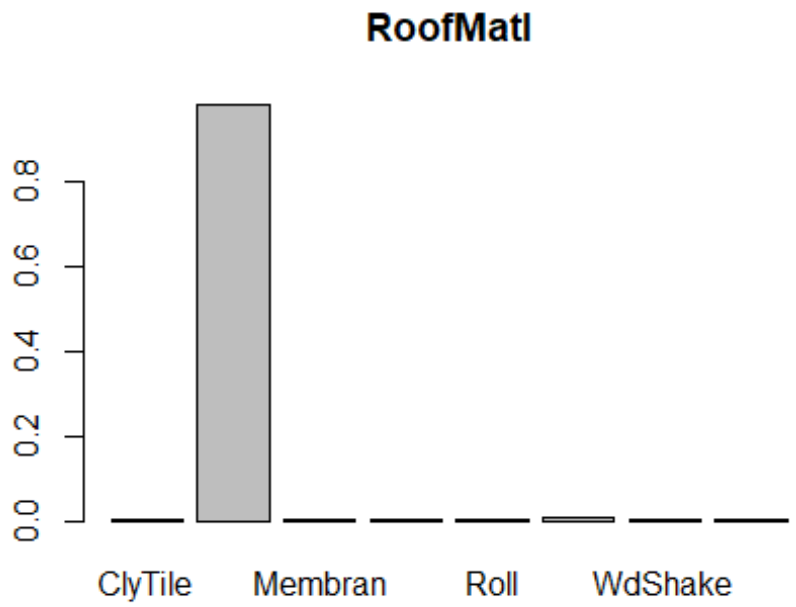



```
##                Flat        Gable      Gambrel         Hip     Mansard
## DistAs 13.00000000 1141.0000000 11.000000000 286.0000000 7.000000000
## DistRe  0.00890411    0.7815068  0.007534247   0.1958904 0.004794521
##                Shed
## DistAs 2.000000000
## DistRe 0.001369863
```

Variabile qualitativa categoriale con 6 diverse categorie, la maggior parte delle case nel campione presenta un tetto di tipo “Gable”

Variabile RoofMatl

```
case$RoofMatl <- factor(case$RoofMatl)
display_table(case$RoofMatl, "RoofMatl")
```

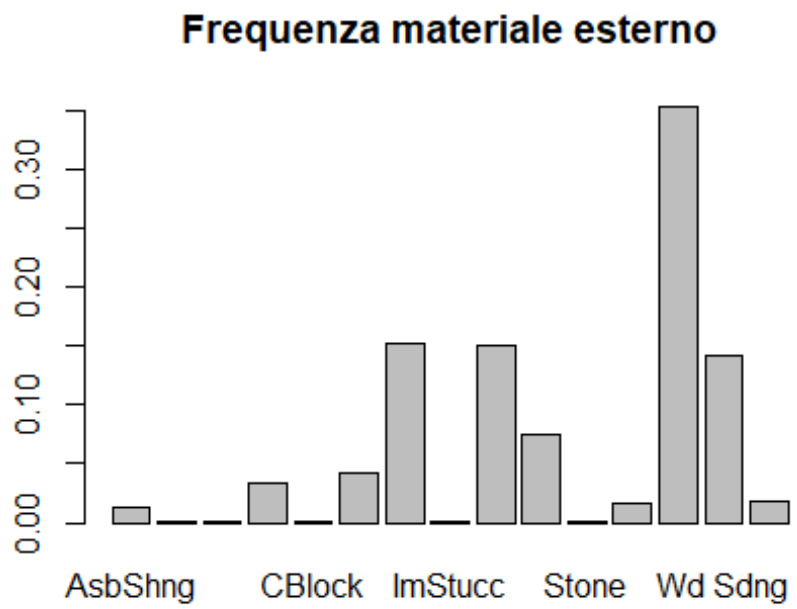



```
##             ClyTile      CompShg      Membran        Metal         Roll
## DistAs 1.0000000000 1434.0000000 1.0000000000 1.0000000000 1.0000000000
## DistRe 0.0006849315    0.9821918 0.0006849315 0.0006849315 0.0006849315
##             Tar&Grv     WdShake     WdShngl
## DistAs 11.000000000 5.000000000 6.000000000
## DistRe  0.007534247 0.003424658 0.004109589
```

Variabile qualitativa categoriale con 8 diverse categorie, la maggior parte delle case (ben il 98.2%) nel campione presenta il tetto in materiale "CompShg" con le frequenze percentuali degli altri materiali che non raggiungono nemmeno l'1%

Variabile Exterior1st

```
case$RoofMatl <- factor(case$Exterior1st)
display_table(case$Exterior1st, "Frequenza materiale esterno")
```

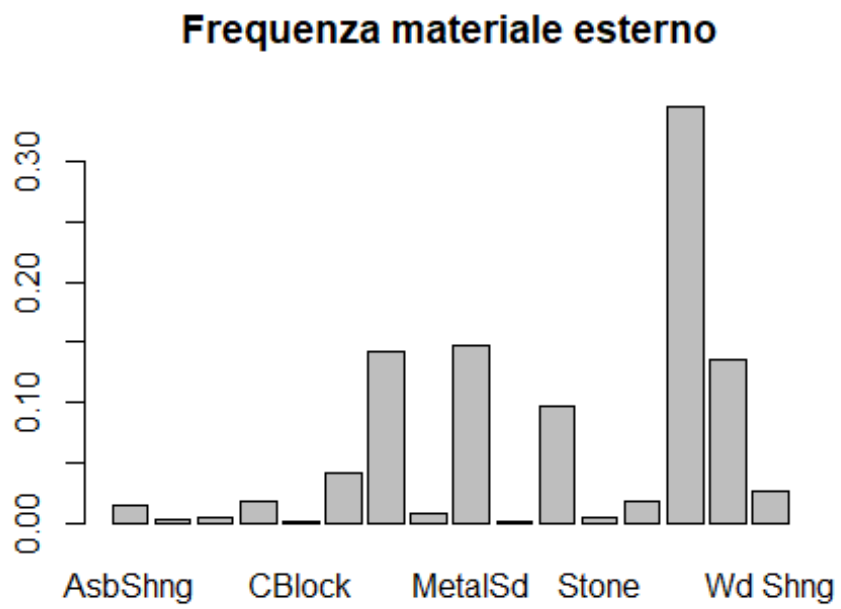



```
##            AsbShng      AsphShn     BrkComm     BrkFace       CBlock
## DistAs 20.00000000 1.0000000000 2.000000000 50.00000000 1.0000000000
## DistRe  0.01369863 0.0006849315 0.001369863  0.03424658 0.0006849315
##            CemntBd     HdBoard      ImStucc     MetalSd     Plywood       Stone
## DistAs 61.00000000 222.0000000 1.0000000000 220.0000000 108.0000000 2.000000000
## DistRe  0.04178082   0.1520548 0.0006849315   0.1506849   0.0739726 0.001369863
##             Stucco     VinylSd     Wd Sdng     WdShing
## DistAs 25.00000000 515.0000000 206.0000000 26.00000000
## DistRe  0.01712329   0.3527397   0.1410959  0.01780822
```

Variabile qualitativa categoriale con 15 diverse categorie, la maggior parte delle case nel campione presenta una finitura esterna in vinile che ha una frequenza di 35.3%

Variabile Exterior2nd

```
case$RoofMatl <- factor(case$Exterior2nd)
display_table(case$Exterior2nd, "Frequenza materiale esterno")
```

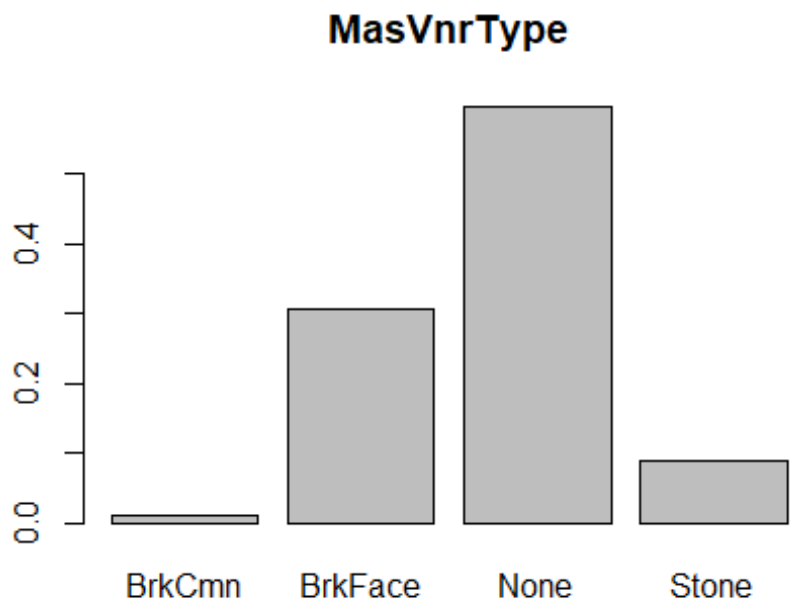



```
##            AsbShng     AsphShn     Brk Cmn     BrkFace       CBlock     CmentBd
## DistAs 20.00000000 3.000000000 7.000000000 25.00000000 1.0000000000 60.00000000
## DistRe  0.01369863 0.002054795 0.004794521  0.01712329 0.0006849315  0.04109589
##            HdBoard      ImStucc     MetalSd        Other      Plywood
## DistAs 207.0000000 10.000000000 214.0000000 1.0000000000 142.00000000
## DistRe   0.1417808  0.006849315   0.1465753 0.0006849315   0.09726027
##              Stone      Stucco     VinylSd     Wd Sdng    Wd Shng
## DistAs 5.000000000 26.00000000 504.0000000 197.0000000 38.0000000
## DistRe 0.003424658  0.01780822   0.3452055   0.1349315  0.0260274
```

Variabile qualitativa categoriale con 16 diverse categorie, la maggior parte delle case nel campione presenta una seconda finitura esterna in vinile che ha una frequenza di 34.5%

Variabile MasVnrType

```
case$RoofMatl <- factor(case$MasVnrType)
display_table(case$MasVnrType, "MasVnrType")
```




```
##             BrkCmn     BrkFace        None        Stone
## DistAs 15.00000000 445.0000000 864.0000000 128.00000000
## DistRe  0.01033058   0.3064738   0.5950413   0.08815427
```

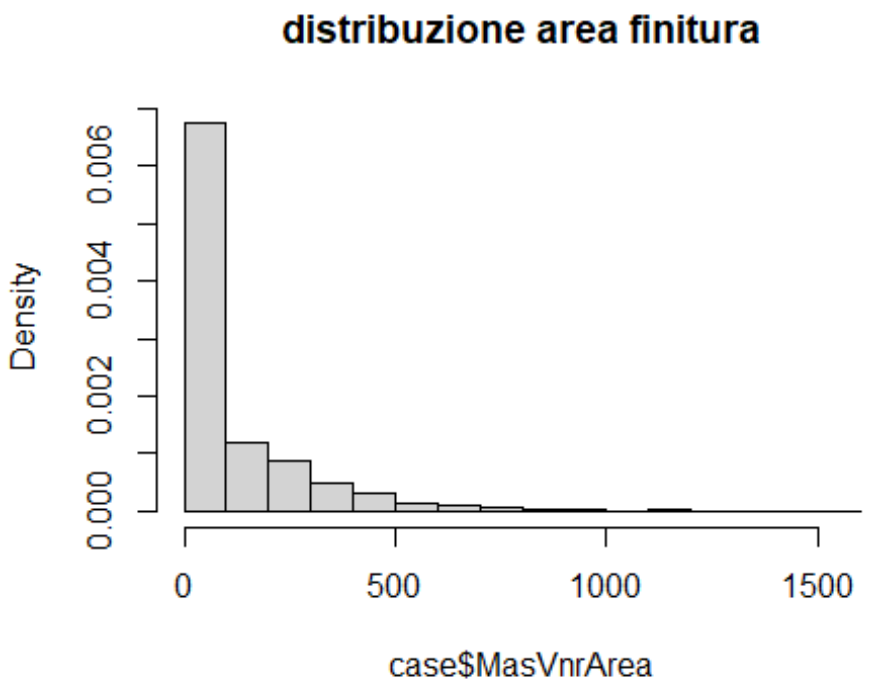
Variabile qualitativa categoriale con 4 diverse categorie, la maggior parte delle case nel campione (59.5%) non ha nessun tipo di rivestimento in muratura esterno mentre il materiale più utilizzato è il Brick Face

Variabile MasVnrArea

```
display_summary_and_var(case$MasVnrArea)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##     0.000000     0.000000     0.000000   103.685262   166.000000  1600.000000
##         NA's          var           sd           sk
##     8.000000 32784.971168   181.066207     2.666326
```

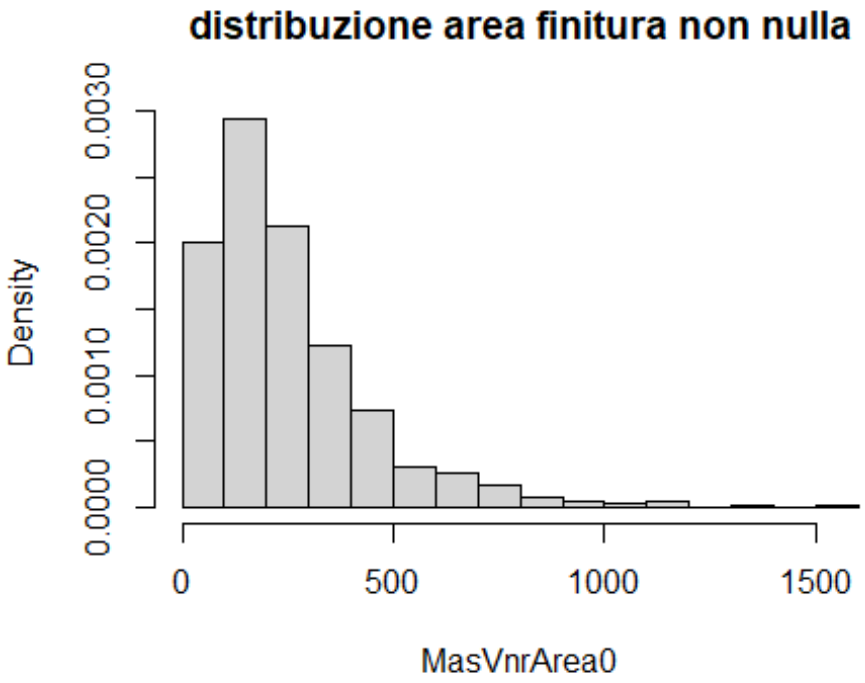
```
hist(case$MasVnrArea, freq = F, main = "distribuzione area finitura")
```



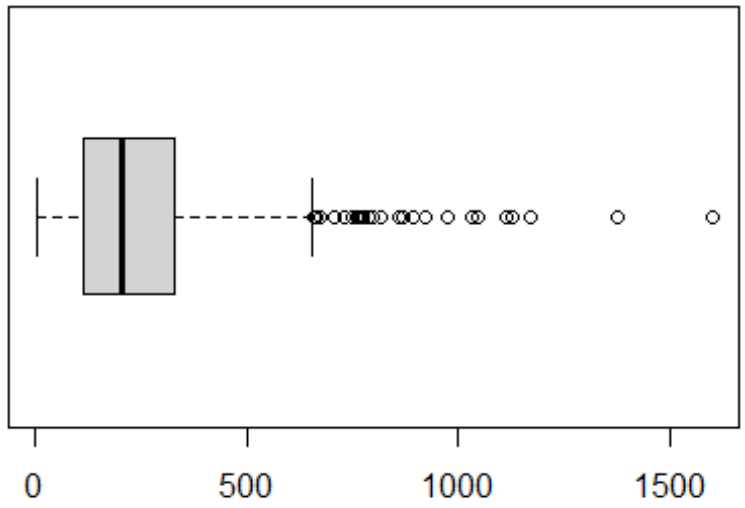

```r
# senza gli zeri
MasVnrArea0 <- na.omit(case[case$MasVnrArea > 0, "MasVnrArea"])
display_summary_and_var(MasVnrArea0)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##     1.000000   113.000000   203.000000   254.739425   330.500000  1600.000000
##          var           sd           sk
## 42084.131985   205.144174     2.088559
```

```r
hist(MasVnrArea0, breaks = 16, freq = F, main = "distribuzione area finitura non
nulla")
```




```
boxplot(MasVnrArea0, horizontal = T)
```

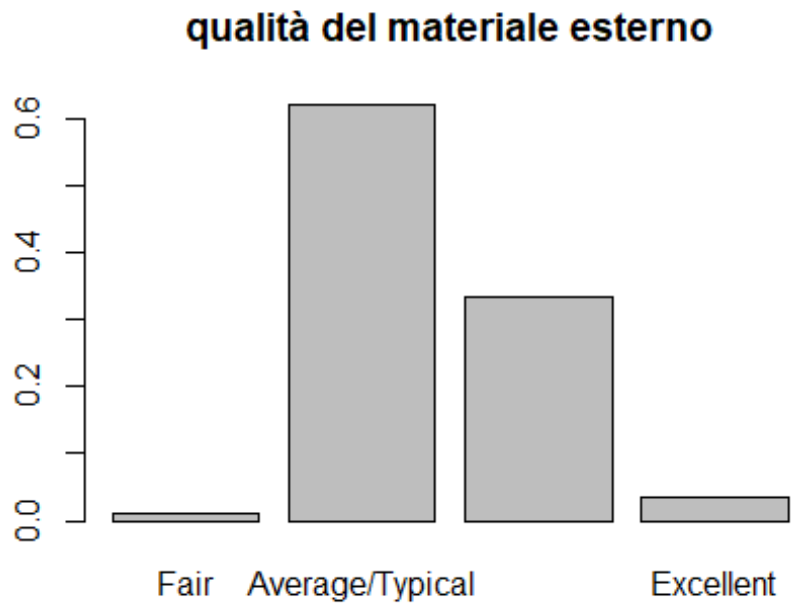


Si nota che la maggior parte delle case non ha una finitura esterna i seguenti grafici indicano la distribuzione dell'area del MasVnr delle case che effettivamente hanno una MasVnr. Si nota una

coda destra piuttosto lunga, con un numero valori outlier elevato infatti l'indice di asimmetria è di 2.088.

Variabile ExterQual

```
case$ExterQual <- factor(case$ExterQual, levels = c("Fa","TA","Gd","Ex"))
levels(case$ExterQual) <- c("Fair","Average/Typical","Good","Excellent")
display_table(case$ExterQual, "qualità del materiale esterno")
```

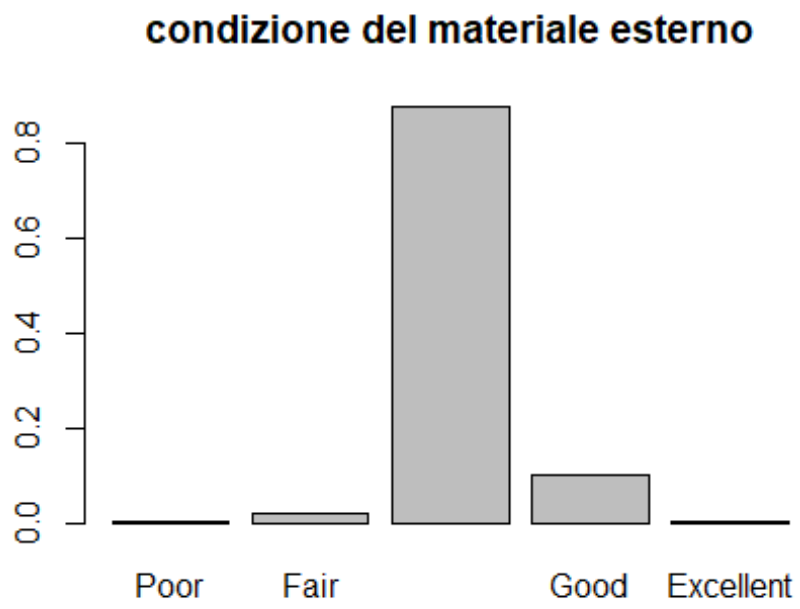



```
##                Fair Average/Typical        Good   Excellent
## DistAs 14.000000000     906.0000000 488.0000000 52.00000000
## DistRe  0.009589041       0.6205479   0.3342466  0.03561644
```

variabile qualitativa categoriale con 4 diverse categorie, si è scelto di riordinare i fattori nel seguente ordine: "Fair","Average/Typical","Good","Excellent" la maggior parte delle case nel campione (62%) ha qualità dei materiali esterni Average/Typical

Variabile ExterCond

```
case$ExterCond <- factor(case$ExterCond, levels = c("Po","Fa","TA","Gd","Ex"))
levels(case$ExterCond) <- c("Poor","Fair","Average/Typical","Good","Excellent")
display_table(case$ExterCond, "condizione del materiale esterno")
```

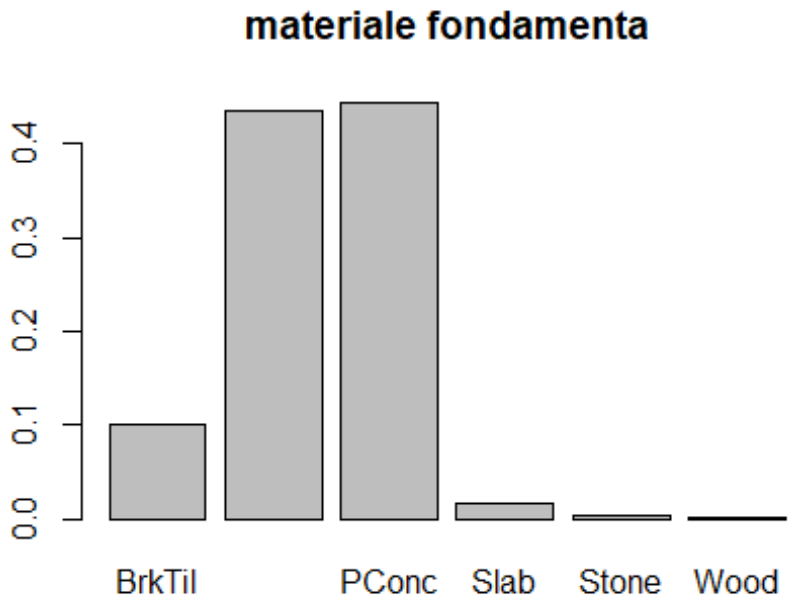



```
##                Poor        Fair Average/Typical  Good   Excellent
## DistAs 1.0000000000 28.00000000    1282.0000000 146.0 3.000000000
## DistRe 0.0006849315  0.01917808       0.8780822   0.1 0.002054795
```

Analogamente alle variabili precedenti, variabile qualitativa categoriale con 5 diverse categorie, si è scelto di riordinare i levels nel seguente ordine:
"Poor","Fair","Average/Typical","Good","Excellent" la maggior parte delle case nel campione (87.8%) ha condizione dei materiali esterni Average/Typical

Variabile Foundation

```
case$Foundation <- factor(case$Foundation)
display_table(case$Foundation, "materiale fondamenta")
```

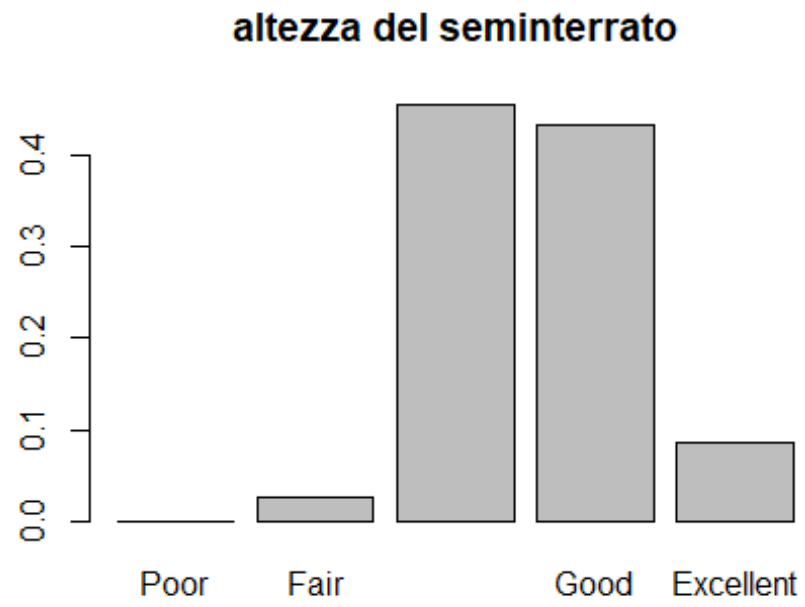



```
##         BrkTil      CBlock       PConc        Slab       Stone        Wood
## DistAs  146.0 634.0000000 647.0000000 24.00000000 6.000000000 3.000000000
## DistRe    0.1   0.4342466   0.4431507  0.01643836 0.004109589 0.002054795
```

Variabile qualitativa categoriale con 6 diverse categorie, la maggior parte delle case nel campione (44.3%) ha le fondamenta in Poured Contrete

Variabile BsmtQual

```
case$BsmtQual <- factor(case$BsmtQual, levels = c("Po","Fa","TA","Gd","Ex"))
levels(case$BsmtQual) <- c("Poor","Fair","Average/Typical","Good","Excellent")
display_table(case$BsmtQual, "altezza del seminterrato")
```

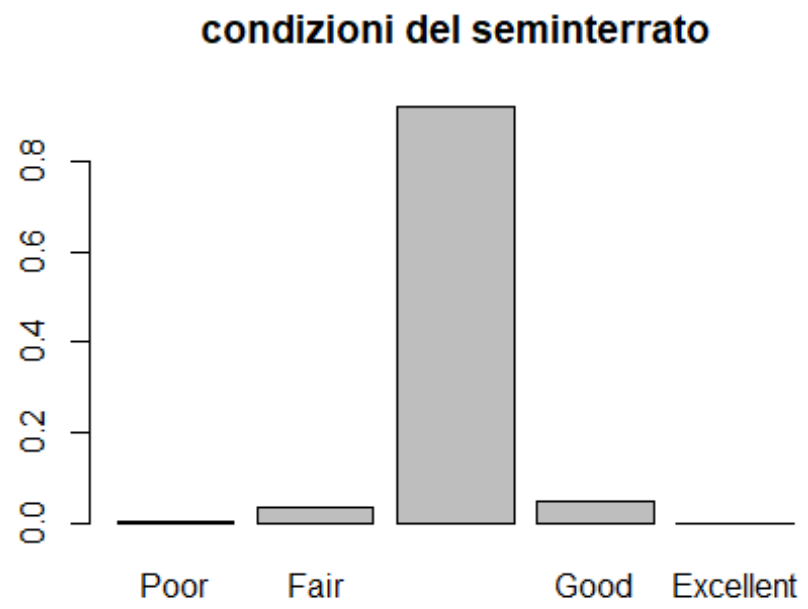



```
##         Poor         Fair Average/Typical         Good    Excellent
## DistAs    0 35.00000000     649.0000000 618.0000000 121.00000000
## DistRe    0  0.02459592       0.4560787   0.4342937   0.08503162
```

Analogamente alle variabili precedenti, variabile qualitativa categoriale con 5 diverse categorie, si è scelto di riordinare i levels nel seguente ordine:
"Poor","Fair","Average/Typical","Good","Excellent" la maggior parte delle case nel campione (45.6%) ha altezza del seminterrato Average/Typical

Variabile BsmtCond

```
case$BsmtCond <- factor(case$BsmtCond, levels = c("Po","Fa","TA","Gd","Ex"))
levels(case$BsmtCond) <- c("Poor","Fair","Average/Typical","Good","Excellent")
display_table(case$BsmtCond, "condizioni del seminterrato")
```

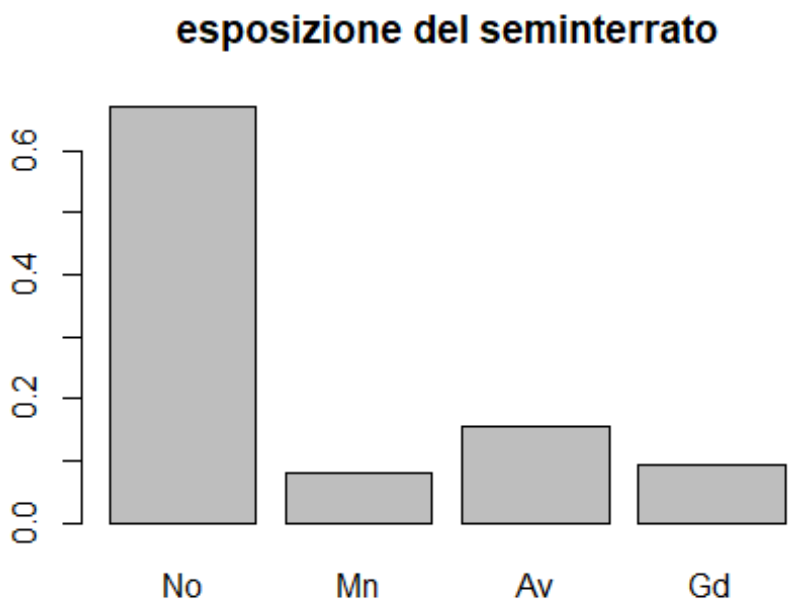



```
##                Poor        Fair Average/Typical        Good Excellent
## DistAs 2.000000000 45.00000000     1311.000000 65.00000000         0
## DistRe 0.001405481  0.03162333        0.921293  0.04567814         0
```

Analogamente alle variabili precedenti, variabile qualitativa categoriale con 5 diverse categorie, si è scelto di riordinare i levels nel seguente ordine:
"Poor","Fair","Average/Typical","Good","Excellent" la maggior parte delle case nel campione (ben il 92.1%) ha condizione del seminterrato Average/Typical

Variabile BsmtExposure

```
case$BsmtExposure <- factor(case$BsmtExposure, levels = c("No", "Mn", "Av", "Gd"))
display_table(case$BsmtExposure, "esposizione del seminterrato")
```

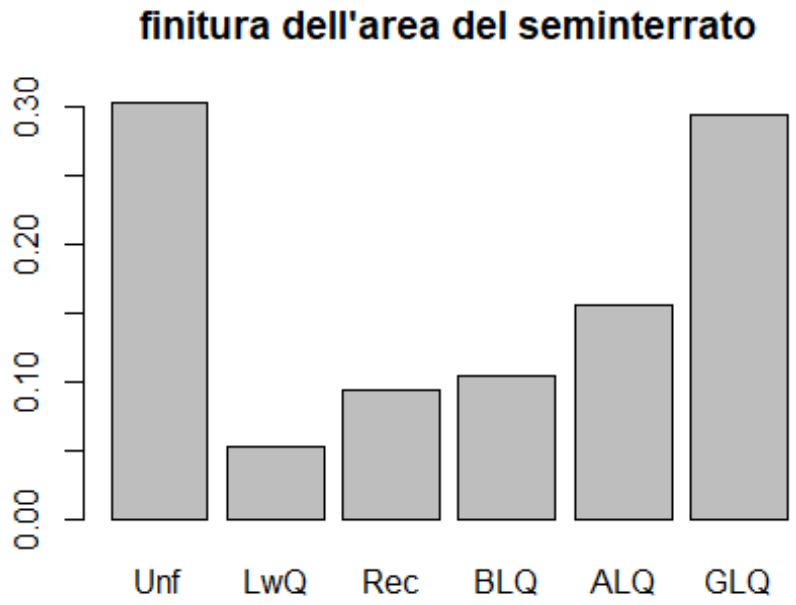



```
##                  No           Mn          Av           Gd
## DistAs 953.0000000 114.00000000 221.0000000 134.00000000
## DistRe   0.6701828   0.08016878   0.1554149   0.09423347
```

Variabile qualitativa categoriale con 4 diverse categorie, la maggior parte delle case nel campione (67%) non ha il seminterrato esposte all'esterno

Variabile BsmtFinType1

```
case$BsmtFinType1 <- factor(case$BsmtFinType1, levels = c("Unf", "LwQ", "Rec",
"BLQ", "ALQ", "GLQ"))
display_table(case$BsmtFinType1, "finitura dell'area del seminterrato")
```




```
##                Unf         LwQ          Rec         BLQ         ALQ         GLQ
## DistAs 430.0000000 74.00000000 133.00000000 148.0000000 220.000000 418.0000000
## DistRe   0.3021785  0.05200281   0.09346451   0.1040056   0.154603   0.2937456
```

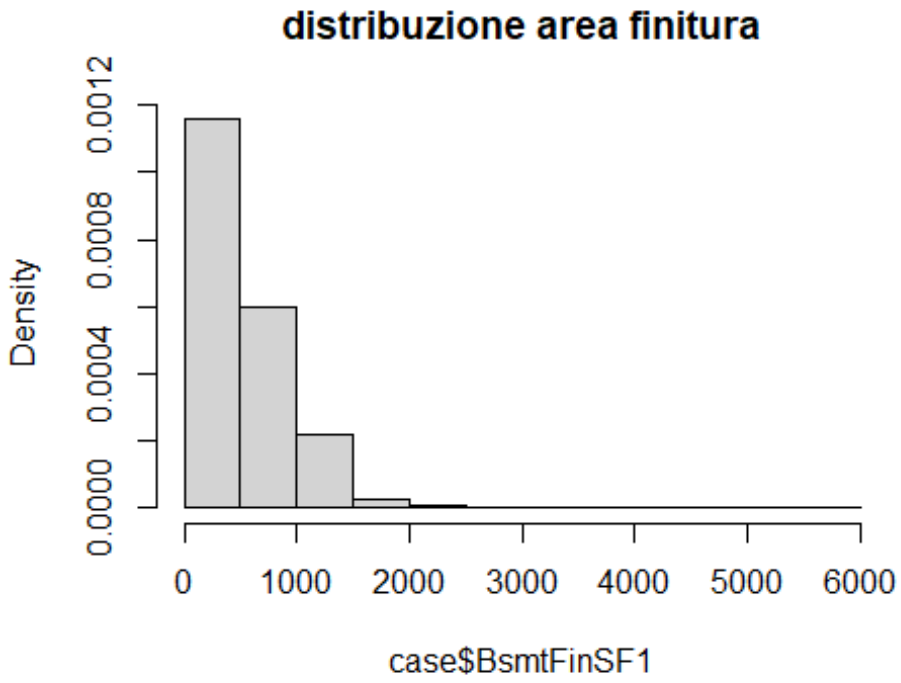
Variabile qualitativa categoriale con 6 diverse categorie, la maggior parte delle case nel campione (30.2%) ha il piano seminterrato non finito (grezzo) seguito da una percentuale del 29.4% di case che hanno il seminterrato abitabile con una Buona qualità.

Variabile BsmtFinSF1

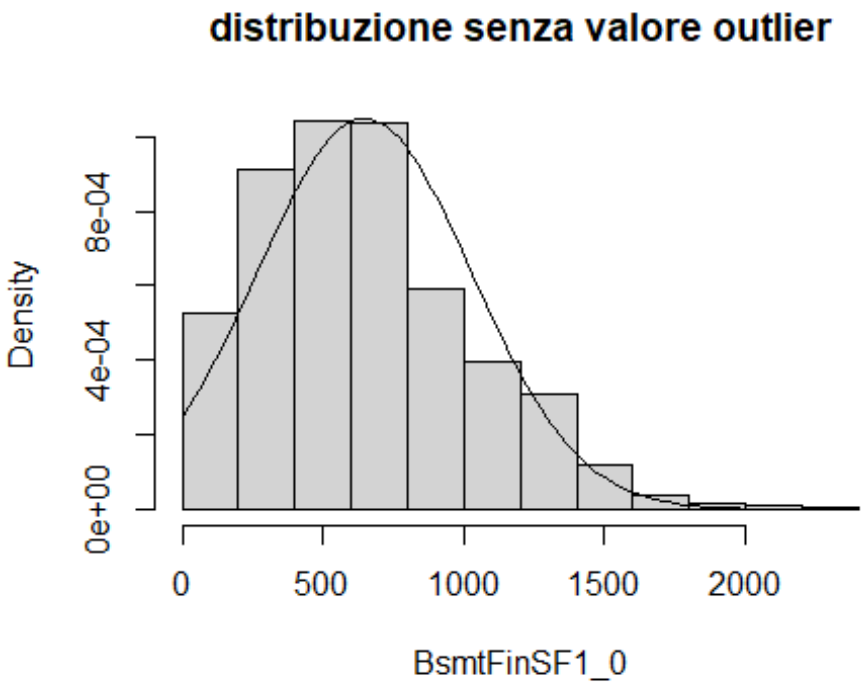
```
display_summary_and_var(case$BsmtFinSF1)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 0.000000e+00 0.000000e+00 3.835000e+02 4.436397e+02 7.122500e+02 5.644000e+03
##          var           sd           sk
## 2.080255e+05 4.560981e+02 1.683771e+00
```

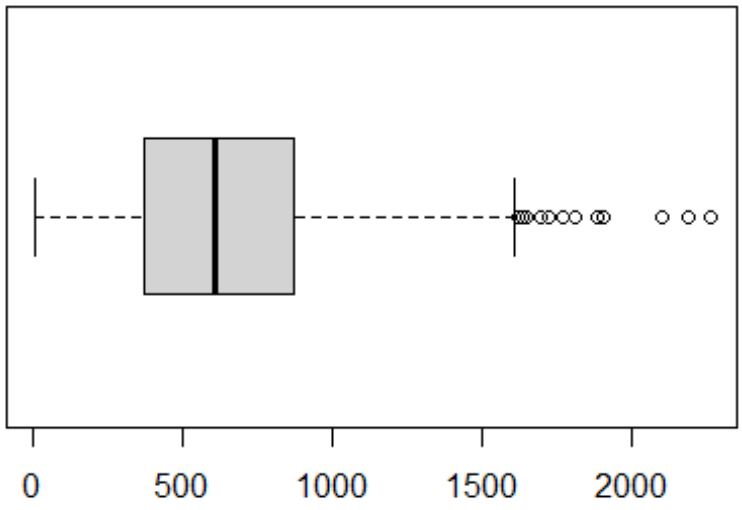
```
hist(case$BsmtFinSF1, freq = F, main = "distribuzione area finitura")
```




```
## si nota che la maggior parte delle case ha il seminterrato incompleto
# i seguenti grafici indicano la distribuzione dell'area di seminterrato delle
case che effettivamente hanno il seminterrato completo

BsmtFinSF1_0 <- na.omit(case[case$BsmtFinSF1 > 0 & case$BsmtFinSF1 <
max(case$BsmtFinSF1), "BsmtFinSF1"])
display_summary_and_var(BsmtFinSF1_0)

##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 2.000000e+00 3.707500e+02 6.040000e+02 6.472480e+02 8.662500e+02 2.260000e+03
##          var           sd           sk
## 1.447301e+05 3.804341e+02 6.813904e-01

hist(BsmtFinSF1_0, freq = F, main = "distribuzione senza valore outlier")
curve(dnorm(x,mean(BsmtFinSF1_0), sd(BsmtFinSF1_0)),add = T)
```



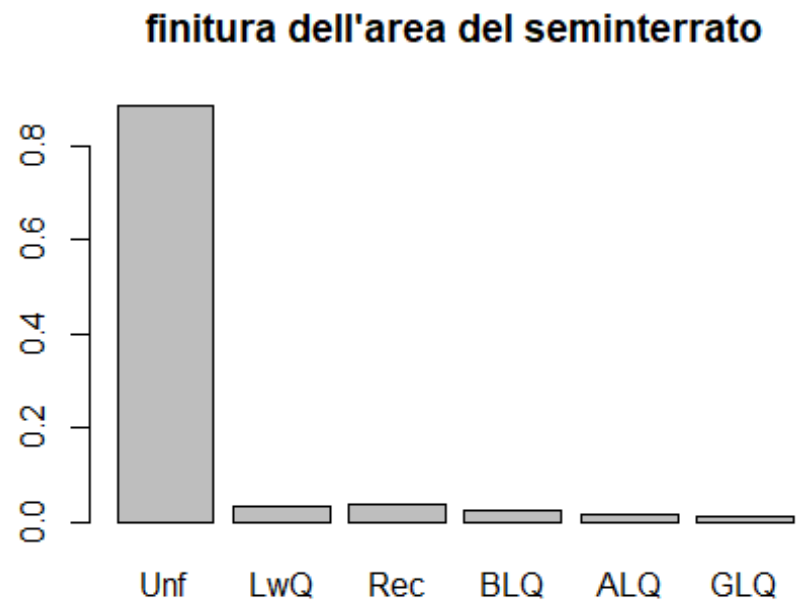

```
boxplot(BsmtFinSF1_0, horizontal = T)
```



Si nota una coda destra piuttosto lunga, con un numero valori outlier elevato infatti l'indice di asimmetria è di 2.298795 in particolare un valore massimo è di molto superiore alla media togliendo il valore elevato si comprende che la distribuzione segue l'andmento di una gaussiana

Variabile BsmtFinType2

```
case$BsmtFinType2 <- factor(case$BsmtFinType2, levels = c("Unf", "LwQ", "Rec",
"BLQ", "ALQ", "GLQ"))
display_table(case$BsmtFinType2, "finitura dell'area del seminterrato")
```

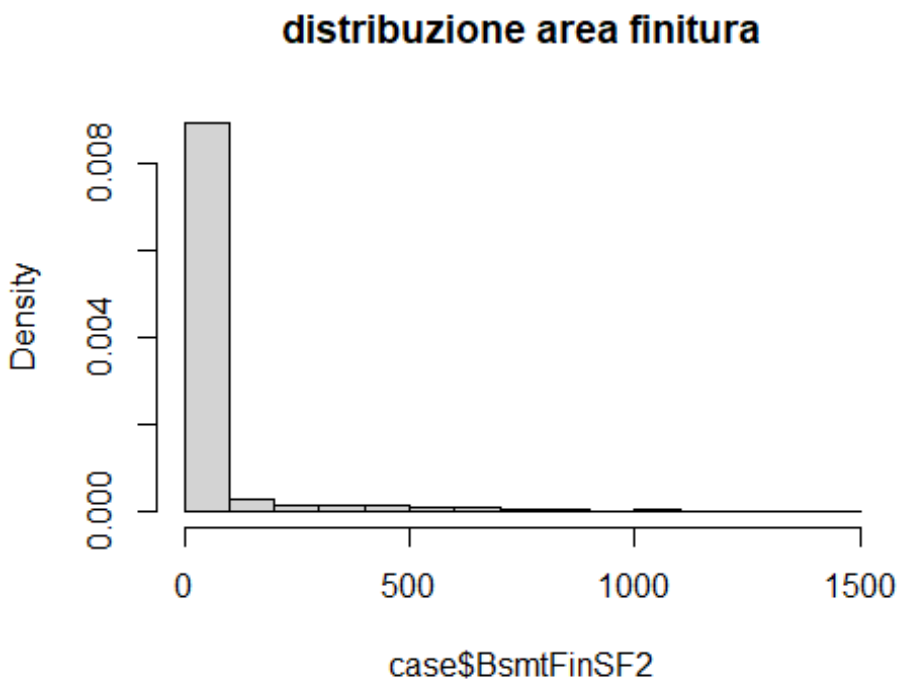



```
##                Unf        LwQ         Rec         BLQ         ALQ          GLQ
## DistAs 1256.000000 46.0000000 54.00000000 33.00000000 19.00000000 14.000000000
## DistRe    0.883263  0.0323488  0.03797468  0.02320675  0.01336146  0.009845288
```

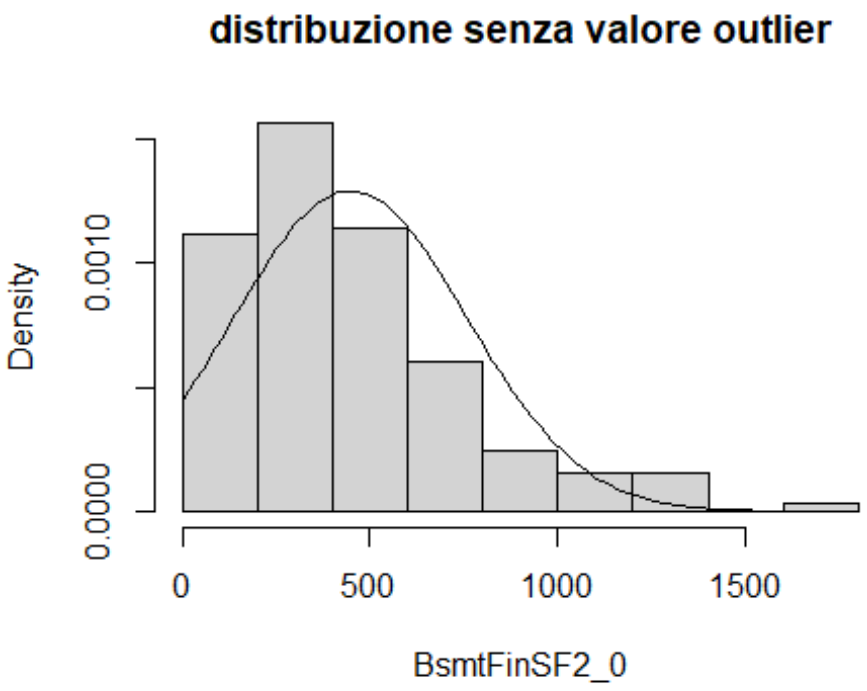
Variabile qualitativa categoriale con 6 diverse categorie, analogamente alla precedente

Variabile BsmtFinSF2

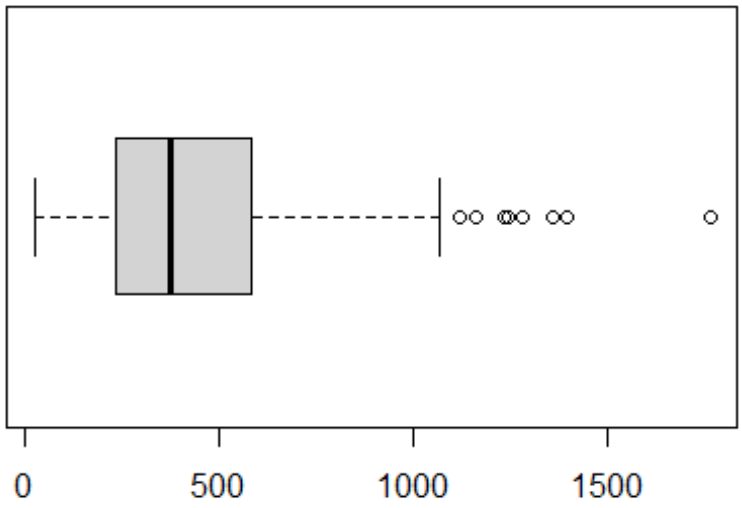
```
display_summary_and_var(case$BsmtFinSF2)

##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##     0.000000     0.000000     0.000000    46.549315     0.000000  1474.000000
##          var           sd           sk
## 26023.907779   161.319273     4.250888

hist(case$BsmtFinSF2, freq = F, main = "distribuzione area finitura")
```




```
## si nota che la maggior parte delle case ha il seminterrato incompleto
# i seguenti grafici indicano la distribuzione dell'area di seminterrato delle
case che effettivamente hanno il seminterrato completo

BsmtFinSF2_0 <- na.omit(case[case$BsmtFinSF2 > 0 & case$BsmtFinSF2 <
max(case$BsmtFinSF2), "BsmtFinSF1"])
display_summary_and_var(BsmtFinSF2_0)

##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##    25.000000   234.000000   377.000000   448.240964   581.750000  1767.000000
##          var           sd           sk
## 95328.971888   308.753902     1.312235

hist(BsmtFinSF2_0, freq = F, main = "distribuzione senza valore outlier")
curve(dnorm(x,mean(BsmtFinSF2_0), sd(BsmtFinSF2_0)),add = T)
```




```
boxplot(BsmtFinSF2_0, horizontal = T)
```



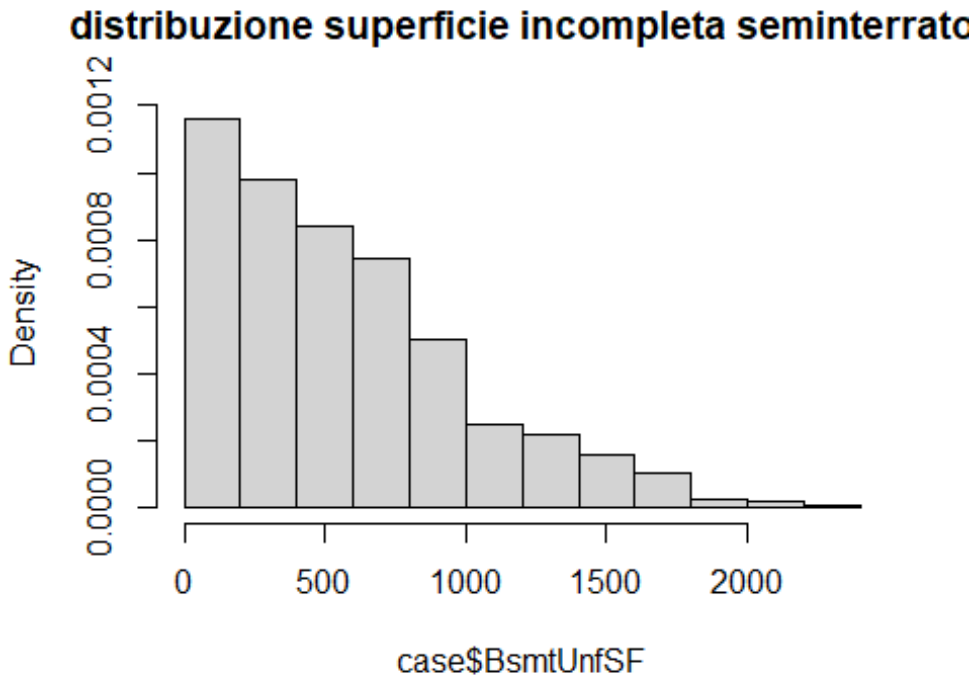
Variabile analoga alla precedente

Variabile BsmtUnfSF

```
display_summary_and_var(case$BsmtUnfSF)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 0.000000e+00 2.230000e+02 4.775000e+02 5.672404e+02 8.080000e+02 2.336000e+03
##          var           sd           sk
## 1.952464e+05 4.418670e+02 9.193227e-01
```

```
hist(case$BsmtUnfSF, freq = F, main = "distribuzione superficie incompleta
seminterrato")
```



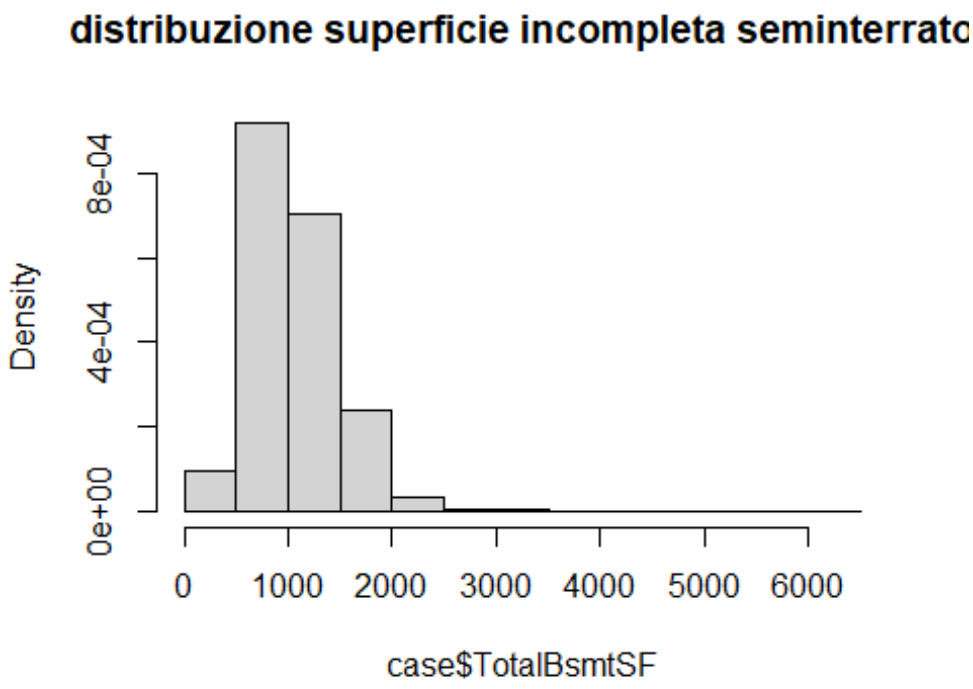
Si nota una coda destra leggermente allungata, con un numero valori outlier elevato infatti l'indice di asimmetria è di 0.919

Variabile TotalBsmtSF

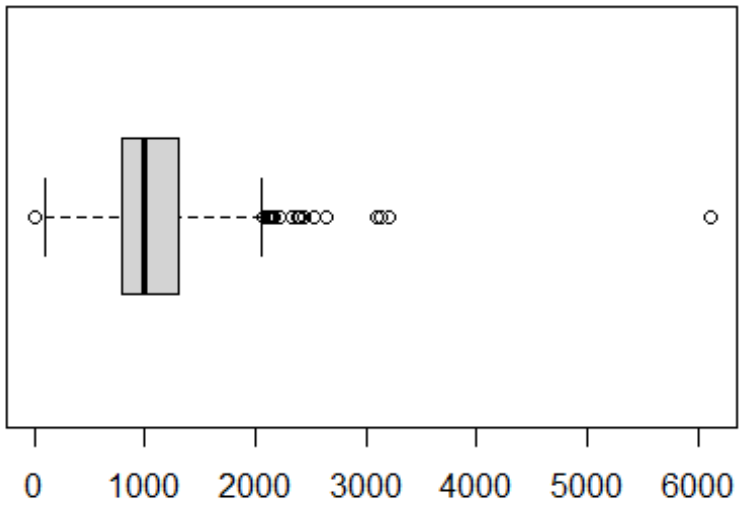
```
display_summary_and_var(case$TotalBsmtSF)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 0.000000e+00 7.957500e+02 9.915000e+02 1.057429e+03 1.298250e+03 6.110000e+03
##          var           sd           sk
## 1.924624e+05 4.387053e+02 1.522688e+00
```

```
hist(case$TotalBsmtSF, freq = F, main = "distribuzione superficie incompleta
seminterrato")
```




```r
boxplot(case$TotalBsmtSF, horizontal = T)
```

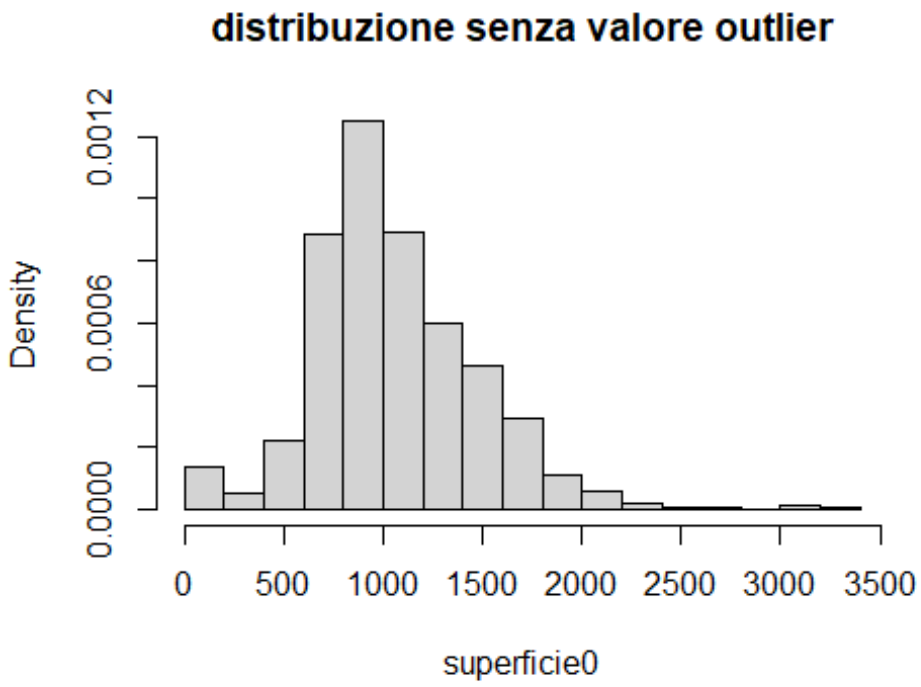



```r
#
superficie0 <- na.omit(case[case$TotalBsmtSF < max(case$TotalBsmtSF),
```

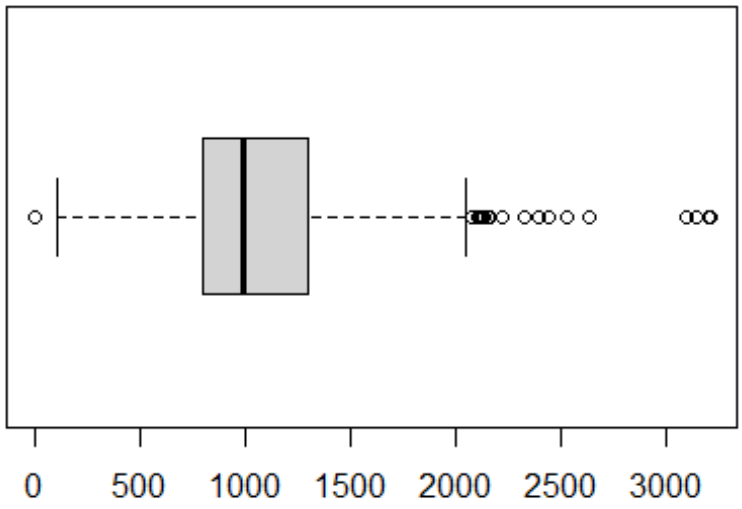
```
"TotalBsmtSF"])
display_summary_and_var(superficie0)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 0.000000e+00 7.955000e+02 9.910000e+02 1.053966e+03 1.297500e+03 3.206000e+03
##          var           sd           sk
## 1.750731e+05 4.184174e+02 5.730470e-01
```

```
hist(superficie0, freq = F, main = "distribuzione senza valore outlier")
```



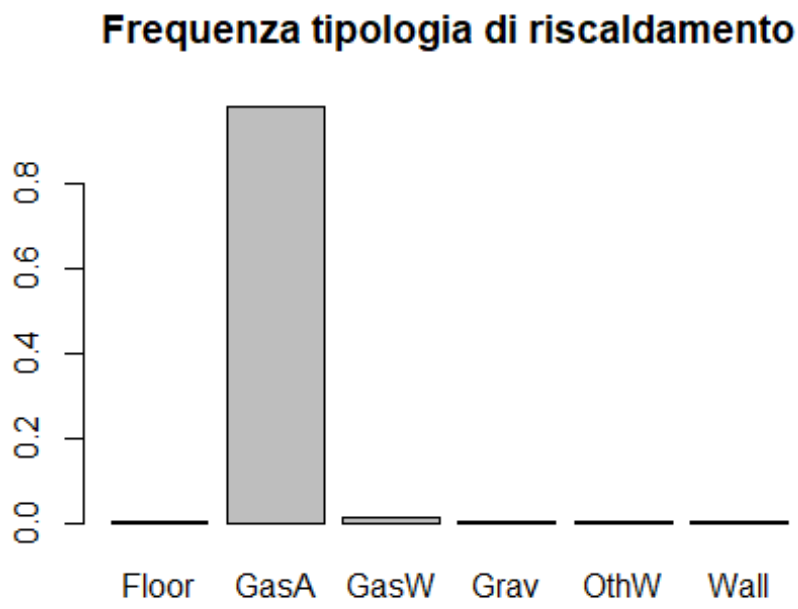

```
boxplot(superficie0, horizontal = T)
```



Si nota una coda destra piuttosto lunga, con un numero valori outlier elevato, in particolare un valore massimo molto elevato e distante dalla media per una migliore visualizzaœone del grafico rimuovo questo valore

Variabile Heating

```
case$Heating <- factor(case$Heating)
display_table(case$Heating, "Frequenza tipologia di riscaldamento")
```

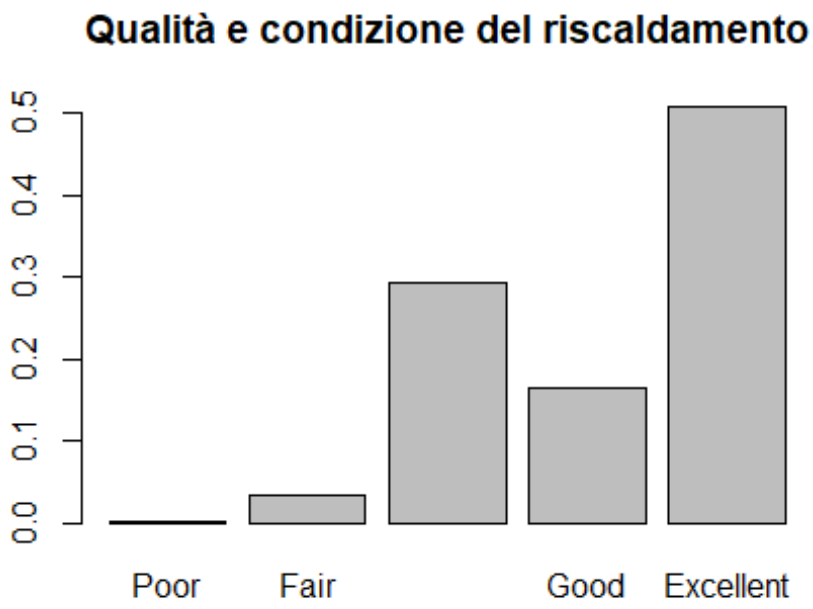



```
##                Floor         GasA        GasW        Grav        OthW
## DistAs 1.0000000000 1428.0000000 18.00000000 7.000000000 2.000000000
## DistRe 0.0006849315    0.9780822  0.01232877 0.004794521 0.001369863
##                Wall
## DistAs 4.000000000
## DistRe 0.002739726
```

Variabile qualitativa categoriale con 6 diverse categorie, la maggior parte delle case nel campione (ben il 97.8%) ha utilizza un sistema di riscaldamento di tipo “Gas forced warm air furnace”

Variabile HeatingQC

```
case$HeatingQC <- factor(case$HeatingQC, levels = c("Po","Fa","TA","Gd","Ex"))
levels(case$HeatingQC) <- c("Poor","Fair","Average/Typical","Good","Excellent")
display_table(case$HeatingQC, "Qualità e condizione del riscaldamento")
```

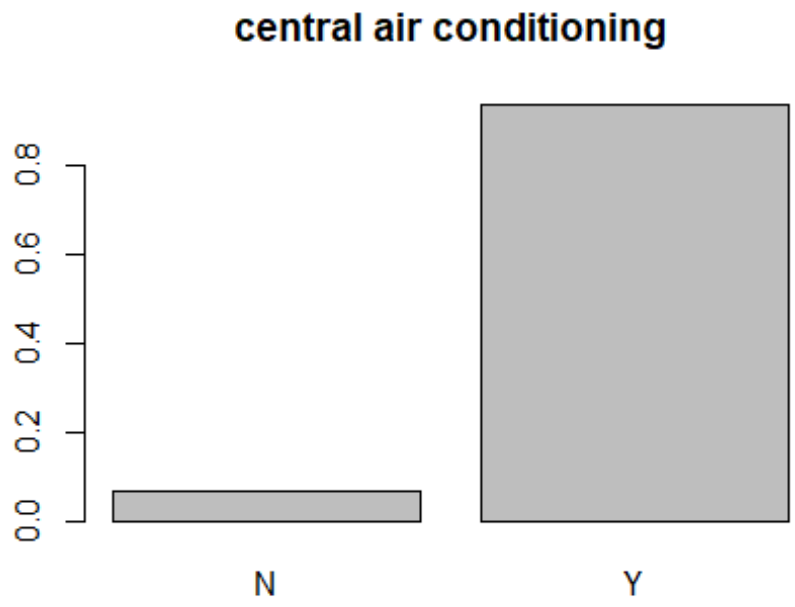



```
##                 Poor        Fair Average/Typical        Good   Excellent
## DistAs 1.0000000000 49.00000000     428.0000000 241.0000000 741.0000000
## DistRe 0.0006849315  0.03356164       0.2931507   0.1650685   0.5075342
```

analogamente alle variabili precedenti, variabile qualitativa categoriale con 5 diverse categorie, si è scelto di riordinare i levels nel seguente ordine:
"Poor","Fair","Average/Typical","Good","Excellent" la maggioranza delle case nel campione (50.7%) ha una eccellente qualità e condizione del riscaldamento

Variabile CentralAir

```
case$CentralAir <- factor(case$CentralAir)
display_table(case$CentralAir, titolo = 'central air conditioning')
```

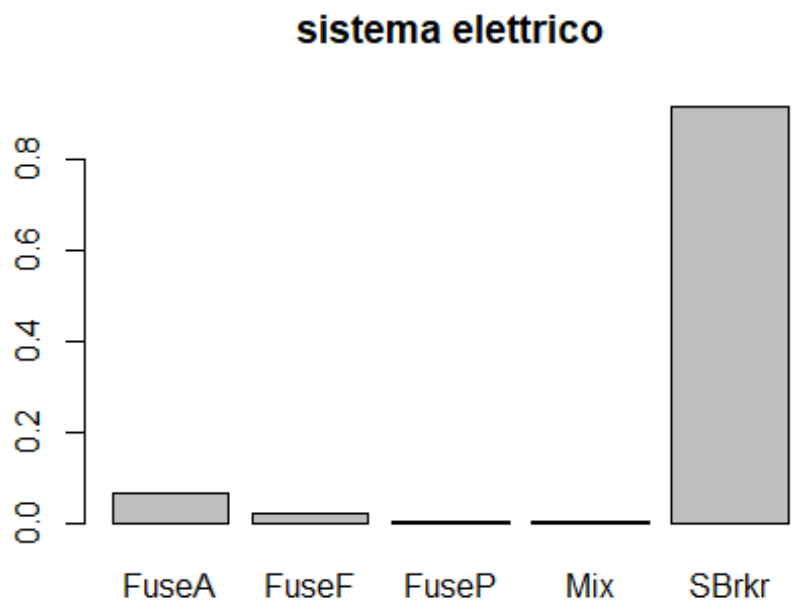



```
##                   N            Y
## DistAs 95.00000000 1365.0000000
## DistRe  0.06506849    0.9349315
```

Si osserva che il 93% circa delle case ha il central air conditioning.

Variabile Electrical

```
case$Electrical <- factor(case$Electrical)
display_table(case$Electrical, titolo = 'sistema elettrico')
```




```
##              FuseA       FuseF       FuseP         Mix        SBrkr
## DistAs 94.00000000 27.00000000 3.000000000 1.000000000 1334.0000000
## DistRe  0.06442769  0.01850583 0.002056203 0.000685401    0.9143249
```

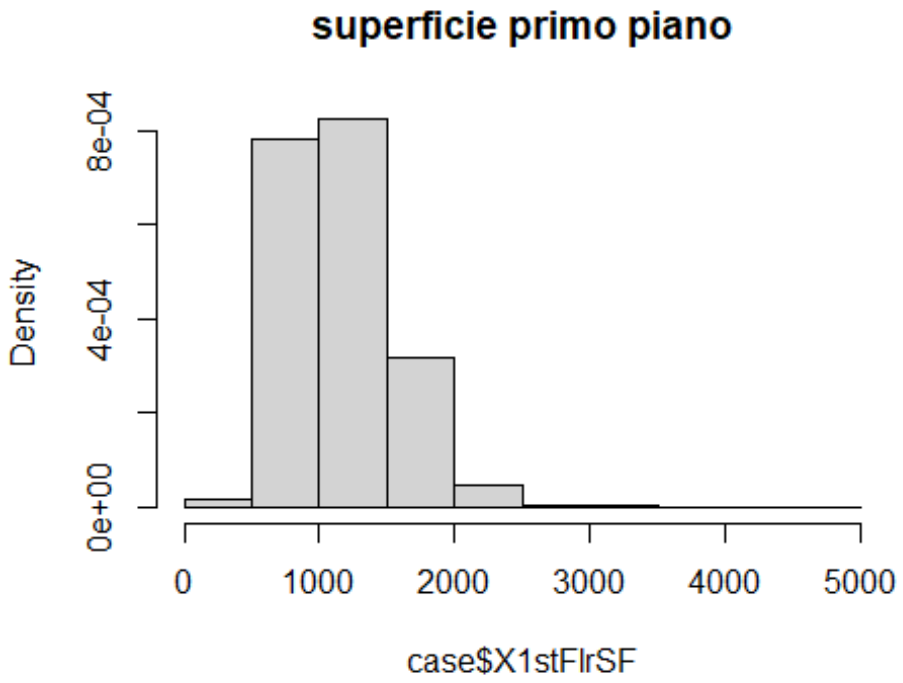
Vi sono 4 tipi diversi, il 91% delle case ha il circuito standard, il restante 9% si dovide negli altri tre tipi. il 6% è rappresentato dal tipo FuseA

Variabile X1stFlrSF

```
display_summary_and_var(case$X1stFlrSF)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 3.340000e+02 8.820000e+02 1.087000e+03 1.162627e+03 1.391250e+03 4.692000e+03
##          var           sd           sk
## 1.494501e+05 3.865877e+02 1.375342e+00
```

```
hist(case$X1stFlrSF, probability = T, main = 'superficie primo piano')
```



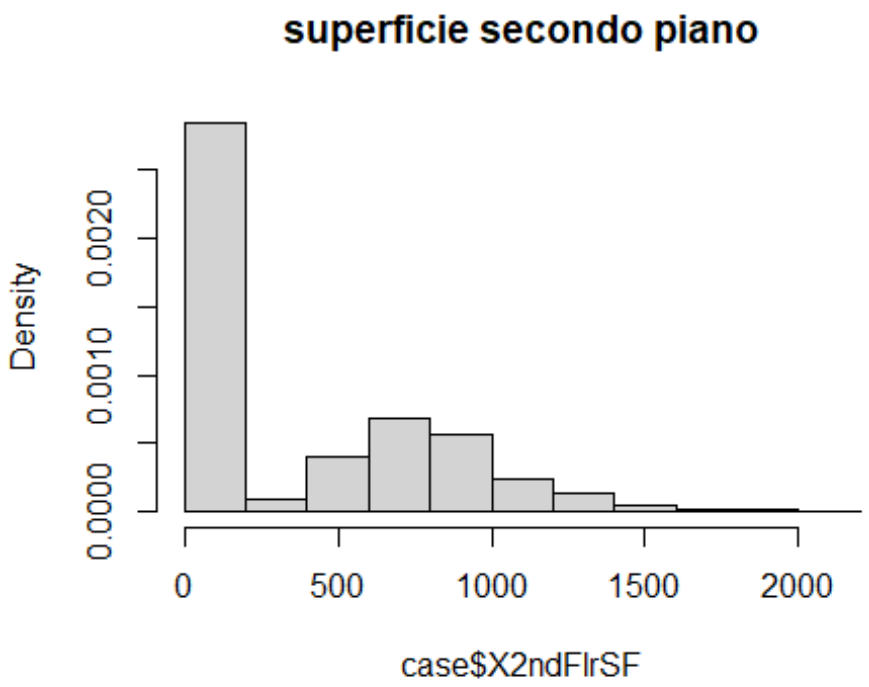
La maggior parte delle case si trova nel range 800-1200 piedi quadrati

Variabile X2ndFlrSF

```
display_summary_and_var(case$X2ndFlrSF)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max. 
## 0.000000e+00 0.000000e+00 0.000000e+00 3.469925e+02 7.280000e+02 2.065000e+03 
##          var           sd           sk 
## 1.905571e+05 4.365284e+02 8.121943e-01
```

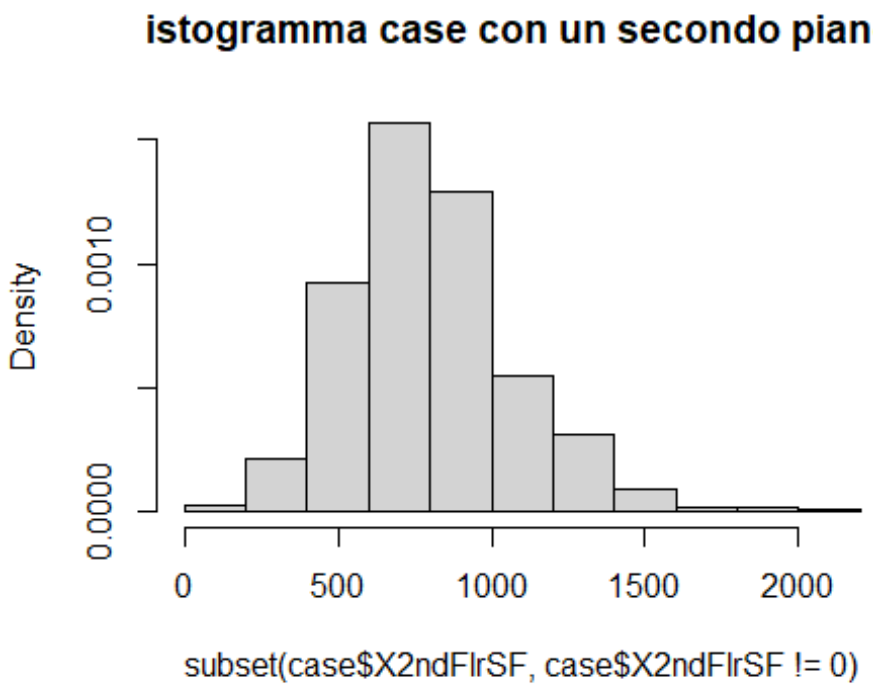
```
hist(case$X2ndFlrSF, probability = T, main = 'superficie secondo piano')
```




```r
# solo le case che hanno solo il secondo piano:
display_summary_and_var(subset(case$X2ndFlrSF, case$X2ndFlrSF != 0))
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 1.100000e+02 6.250000e+02 7.760000e+02 8.028669e+02 9.265000e+02 2.065000e+03
##          var           sd           sk
## 7.471856e+04 2.733470e+02 7.011031e-01
```

```r
hist(subset(case$X2ndFlrSF, case$X2ndFlrSF != 0), main = 'istogramma case con un secondo piano', probability = T)
```



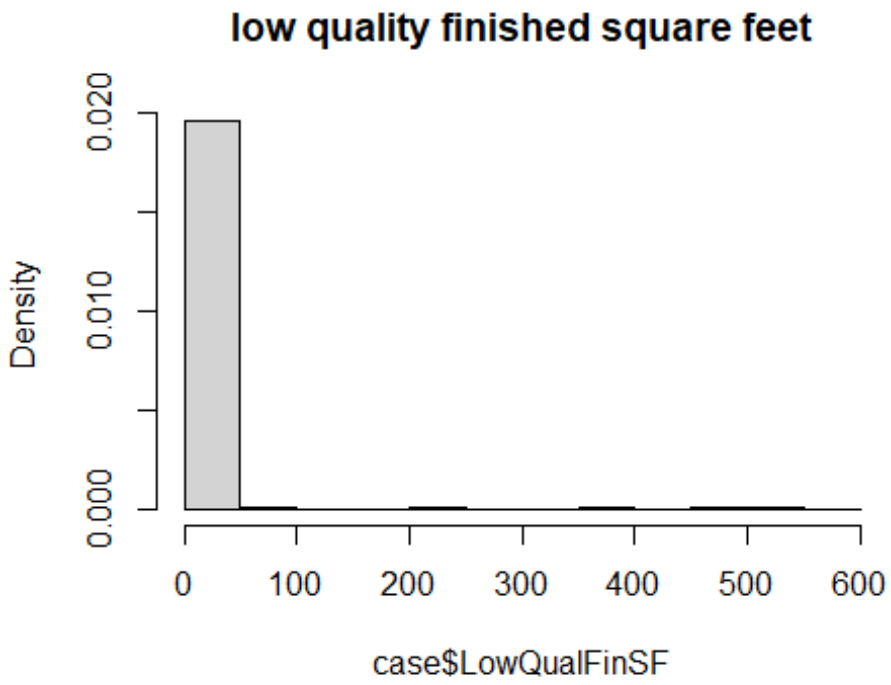
La maggior parte delle case (il 56 %) non possiede un secondo piano, prendendo in considerazione solo quelle che lo hanno osserviamo una media di 776 ft^2 e una mediana di 776 ft^2

Variabile LowQualFinSF

```
display_summary_and_var(case$LowQualFinSF)
```

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000    0.000000    0.000000    5.844521    0.000000  572.000000
##         var          sd          sk
## 2364.204048   48.623081    9.002080
```

```
hist(case$LowQualFinSF, prob = T, main = 'low quality finished square feet')
```

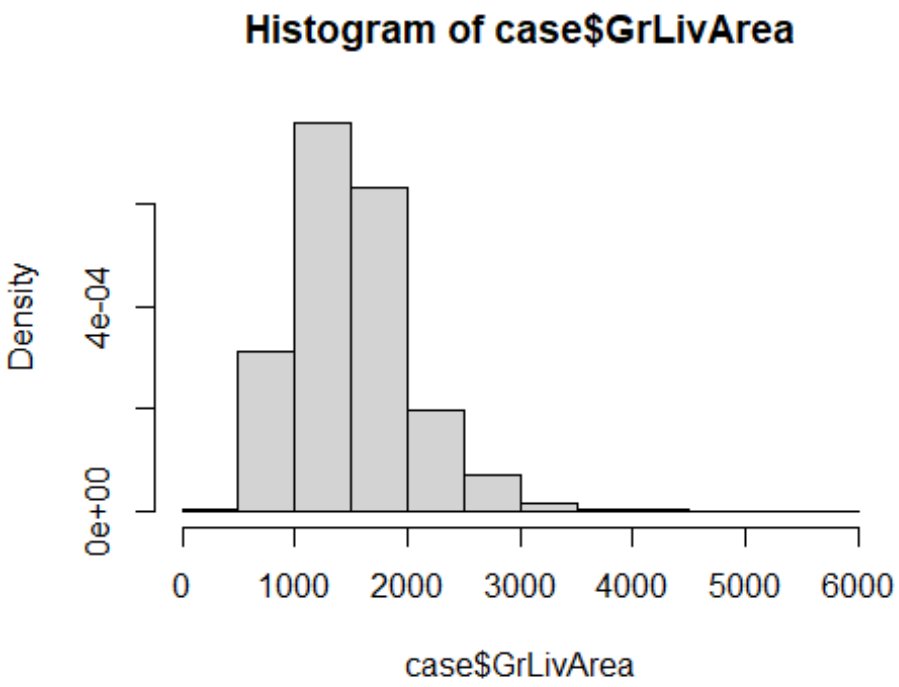



```
length(case$LowQualFinSF[case$LowQualFinSF == 0])

## [1] 1434
```

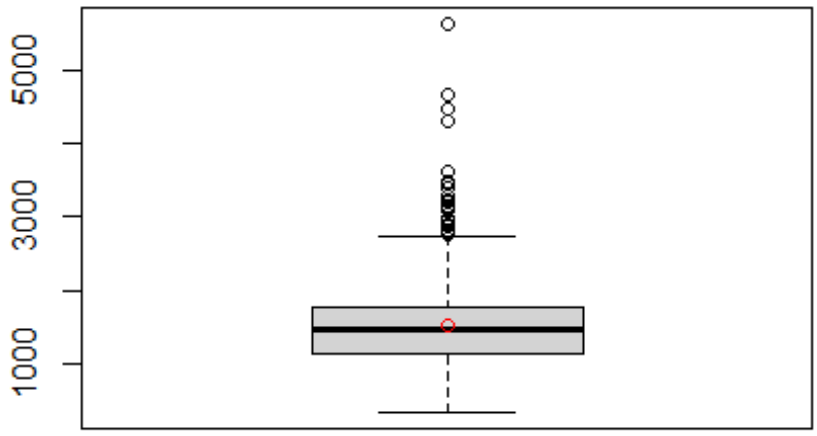
Si osserva che 1434 case su 1460 hanno 0 low quality finished square feet

Variabile GrLivArea

```
display_summary_and_var(case$GrLivArea)

##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 3.340000e+02 1.129500e+03 1.464000e+03 1.515464e+03 1.776750e+03 5.642000e+03
##          var           sd           sk
## 2.761296e+05 5.254804e+02 1.365156e+00

hist(case$GrLivArea, probability = T)
```



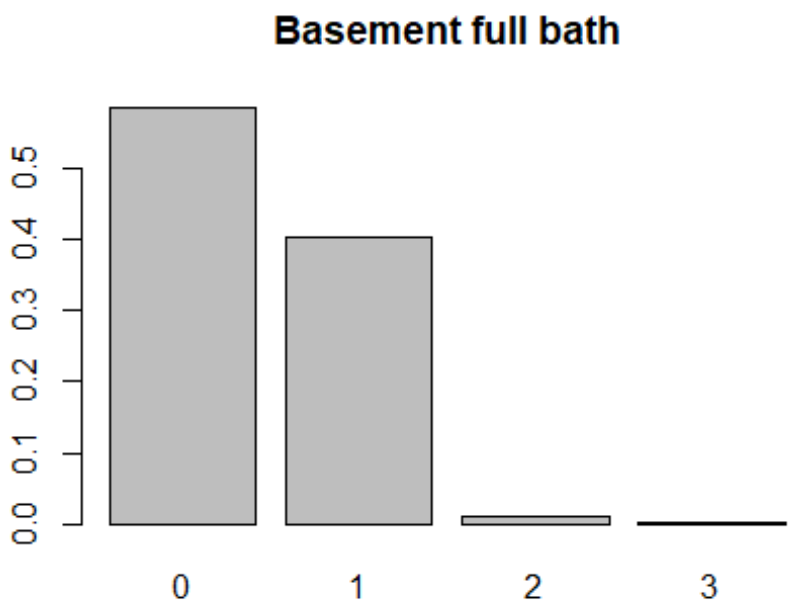

```r
boxplot(case$GrLivArea)
points(mean(case$GrLivArea), col = 'red')
```



Usando il boxplot vediamo che la media è vicina alla mediana (intorno a 1500) e che vi è una bassa varianza, tuttavia sono presenti molti outliers che hanno più 3000 piedi quadrati di superficie abitablie sopra il suolo

Variabile BsmtFullBath

```
case$BsmtFullBath <- factor(case$BsmtFullBath)
display_table(case$BsmtFullBath, titolo = 'Basement full bath')
```

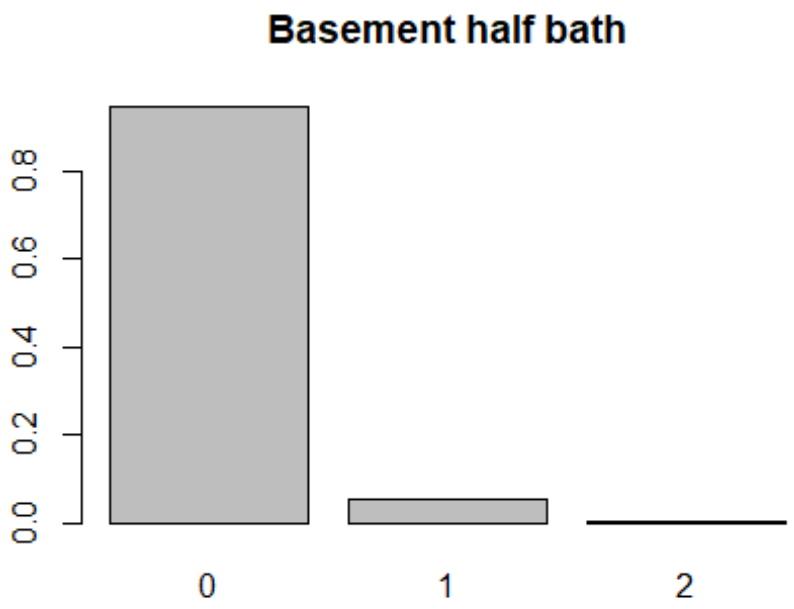



```
##                   0           1           2            3
## DistAs 856.0000000 588.0000000 15.00000000 1.0000000000
## DistRe   0.5863014   0.4027397  0.01027397 0.0006849315
```

Presenta solo i caratteri 0, 1 , 2 e 3. il 59% delle case ha 0, il 40% 1

Variabile BsmtHalfBath

```
case$BsmtHalfBath <- factor(case$BsmtHalfBath)
display_table(case$BsmtHalfBath, titolo = 'Basement half bath')
```

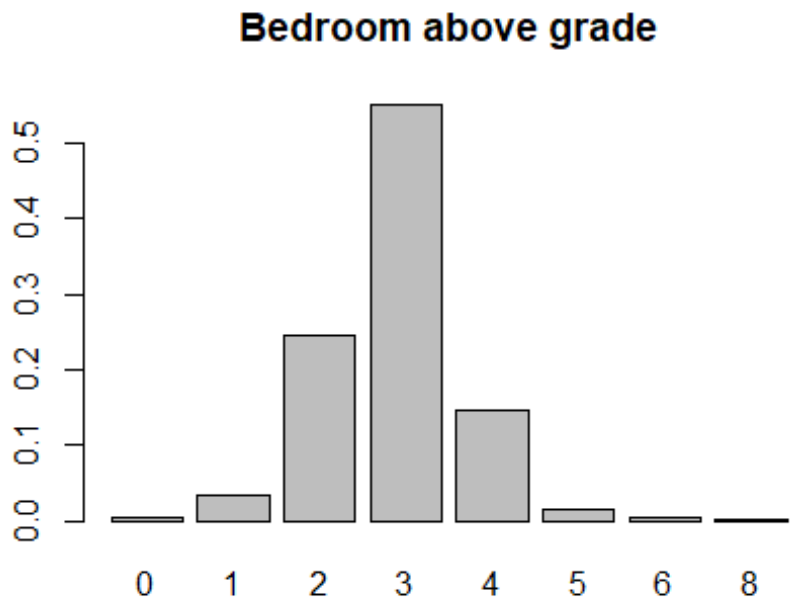



```
##                    0           1           2
## DistAs 1378.0000000 80.00000000 2.000000000
## DistRe    0.9438356  0.05479452 0.001369863
```

Presenta i caratteri 0, 1 e 2, il 94% delle case ha 0, il 5% 1

Variabile BedroomAbvGr

```
case$BedroomAbvGr <- factor(case$BedroomAbvGr)
display_table(case$BedroomAbvGr, titolo = 'Bedroom above grade')
```

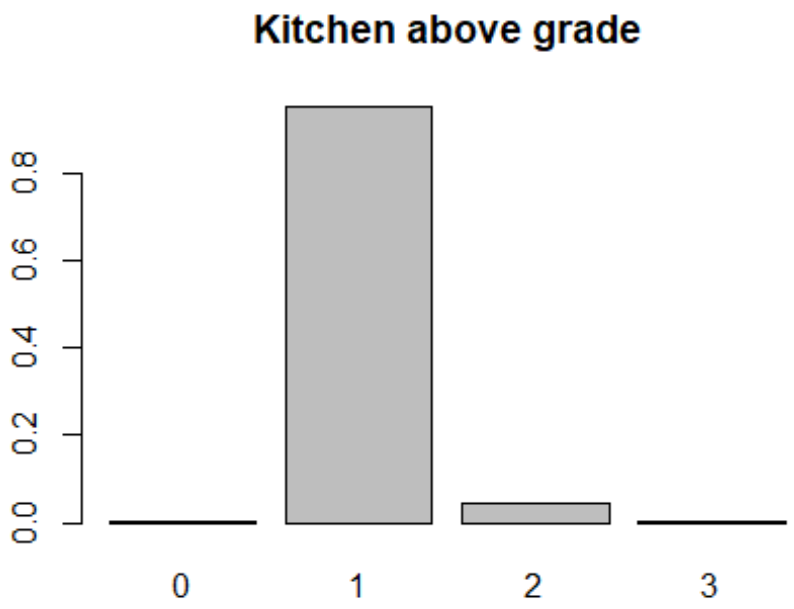



```
##                   0           1           2           3           4           5
## DistAs 6.000000000 50.00000000 358.0000000 804.0000000 213.0000000 21.00000000
## DistRe 0.004109589  0.03424658   0.2452055   0.5506849   0.1458904  0.01438356
##                   6           8
## DistAs 7.000000000 1.0000000000
## DistRe 0.004794521 0.0006849315
```

Numero camere sopra il livello del suolo, non include le camerdel basement. Il 55% delle case ne ha 3.

Variabile KitchenAbvGr

```
case$KitchenAbvGr <- factor(case$KitchenAbvGr)
display_table(case$KitchenAbvGr, 'Kitchen above grade')
```

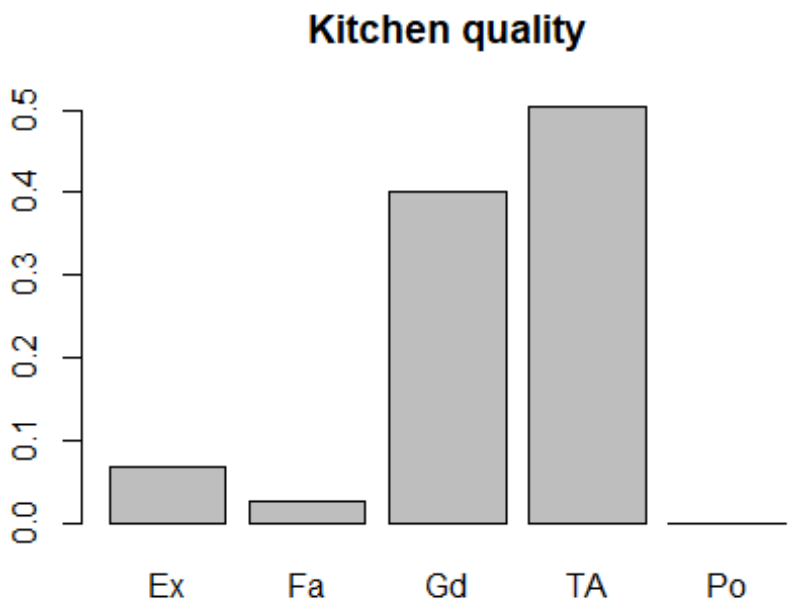



```
##                    0            1           2           3
## DistAs 1.0000000000 1392.0000000 65.00000000 2.000000000
## DistRe 0.0006849315    0.9534247  0.04452055 0.001369863
```

Numero cucine sopra il livello del suolo il 95% delle case ne ha 1

Variabile KitchenQual

```
case$KitchenQual <- factor(case$KitchenQual, levels = c('Ex', 'Fa', 'Gd', 'TA',
'Po'))
display_table(case$KitchenQual, titolo = 'Kitchen quality')
```




```
##                   Ex          Fa          Gd          TA Po
## DistAs 100.00000000 39.00000000 586.0000000 735.0000000  0
## DistRe   0.06849315  0.02671233   0.4013699   0.5034247  0
```

Qualità delle cucine. Il 50% delle cucine è a un livello Average e il 40% a un livelo good

Variabile TotRmsAbvGrd

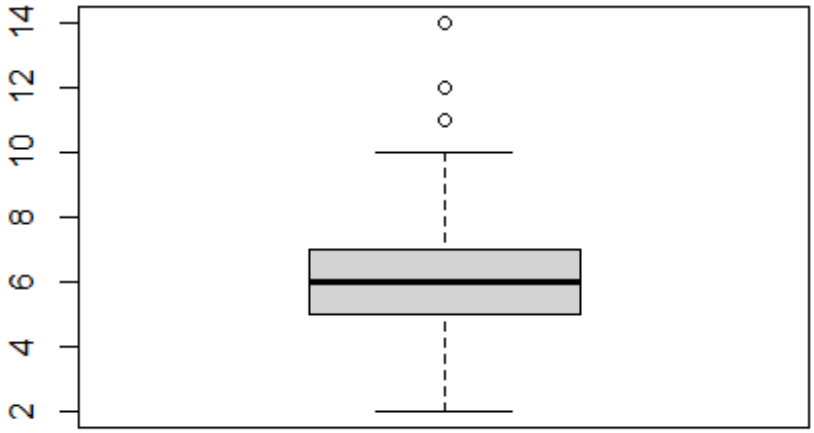
```
display_summary_and_var(case$TotRmsAbvGrd)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.        var
##  2.0000000  5.0000000  6.0000000  6.5178082  7.0000000 14.0000000  2.6419033
##         sd         sk
##  1.6253933  0.6756458
```

```
rbind(tot = table(case$TotRmsAbvGrd), prop = prop.table(table(case$TotRmsAbvGrd)))
```

```
##                  2           3           4           5           6           7
## tot  1.0000000000 17.00000000 97.00000000 275.0000000 402.0000000 329.0000000
## prop 0.0006849315  0.01164384  0.06643836   0.1883562   0.2753425   0.2253425
##                  8           9          10          11          12          14
## tot  187.0000000 75.00000000 47.00000000 18.00000000 11.000000000 1.0000000000
## prop   0.1280822  0.05136986  0.03219178  0.01232877  0.007534247 0.0006849315
```

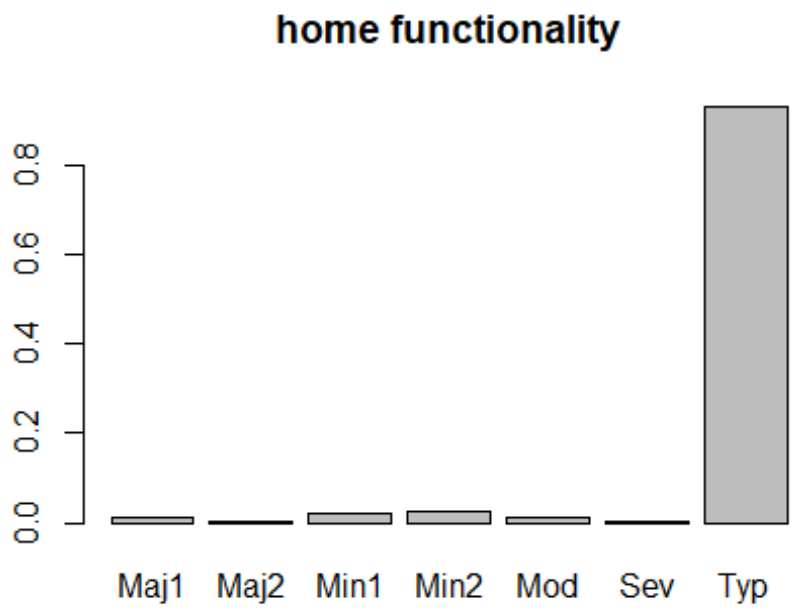
```
boxplot(case$TotRmsAbvGrd)
```



Totale stanze above grade, non include i bagni Il 28% ha 6 stanze

Variabile Functional

```
case$Functional <- factor(case$Functional)
display_table(case$Functional, titolo = 'home functionality')
```

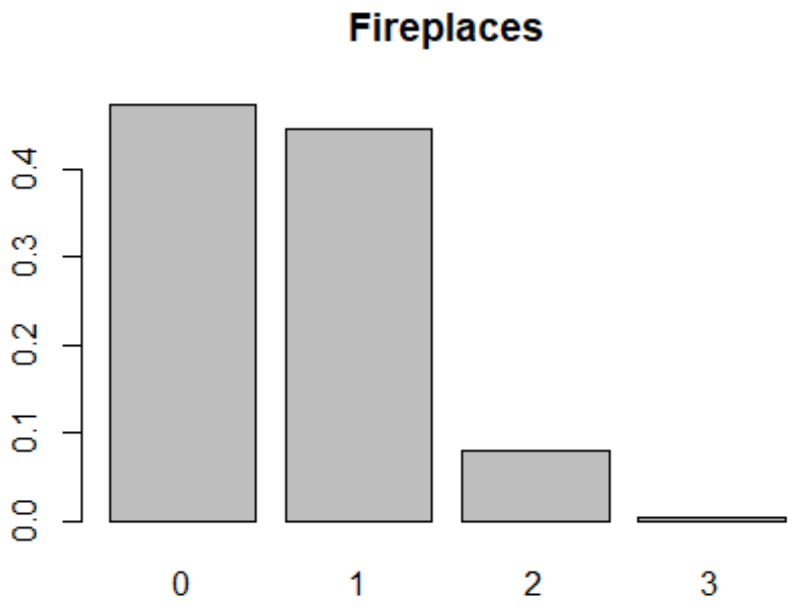



```
##                Maj1        Maj2        Min1        Min2         Mod
## DistAs 14.000000000 5.000000000 31.00000000 34.00000000 15.00000000
## DistRe  0.009589041 0.003424658  0.02123288  0.02328767  0.01027397
##                Sev          Typ
## DistAs 1.0000000000 1360.0000000
## DistRe 0.0006849315    0.9315068
```

Home functionality Il 93% delle case presenta typical functionality.

Variabile Fireplaces

```
case$Fireplaces <- factor(case$Fireplaces)
display_table(case$Fireplaces, titolo = 'Fireplaces')
```

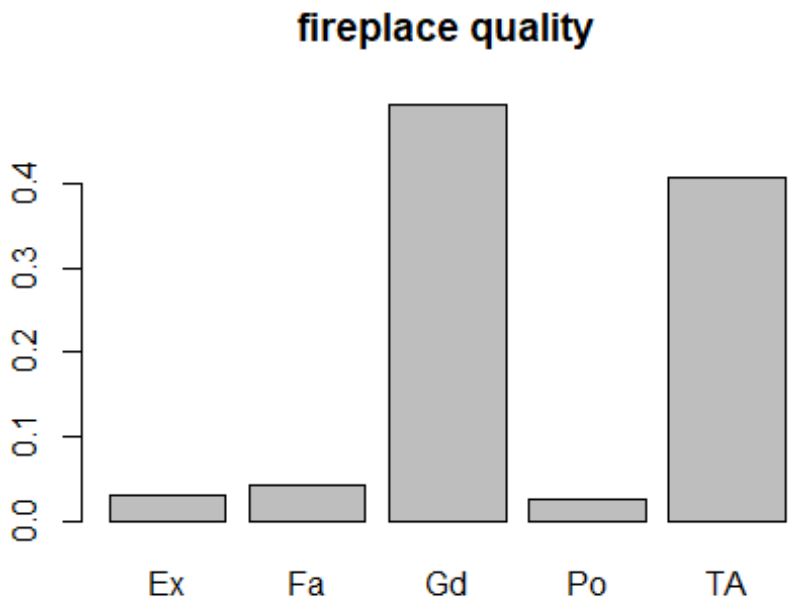



```
##                   0           1            2           3
## DistAs 690.0000000 650.0000000 115.00000000 5.000000000
## DistRe   0.4726027   0.4452055   0.07876712 0.003424658
```

Il 47% delle case ha 0, il 44% ha 1

Variabile FireplaceQu

```
case$FireplaceQu <- factor(case$FireplaceQu)
display_table(case$FireplaceQu, titolo = 'fireplace quality')
```

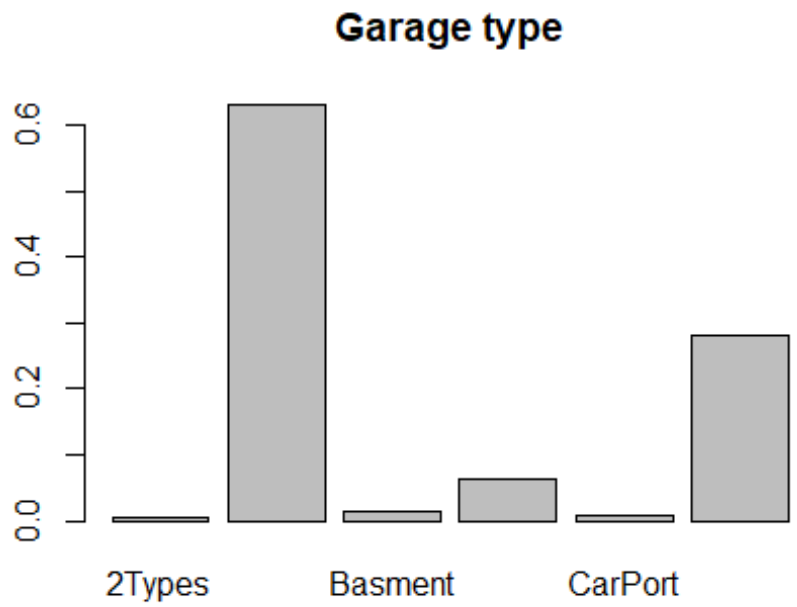



```
##                  Ex          Fa          Gd          Po          TA
## DistAs 24.00000000 33.00000000 380.0000000 20.00000000 313.0000000
## DistRe  0.03116883  0.04285714   0.4935065  0.02597403   0.4064935
```

Qualità dei fire places Il 41% ha una qualità average, il 49% ha una qualità good

Variabile GarageType

```
case$GarageType <- factor(case$GarageType)
display_table(case$GarageType, titolo = 'Garage type')
```

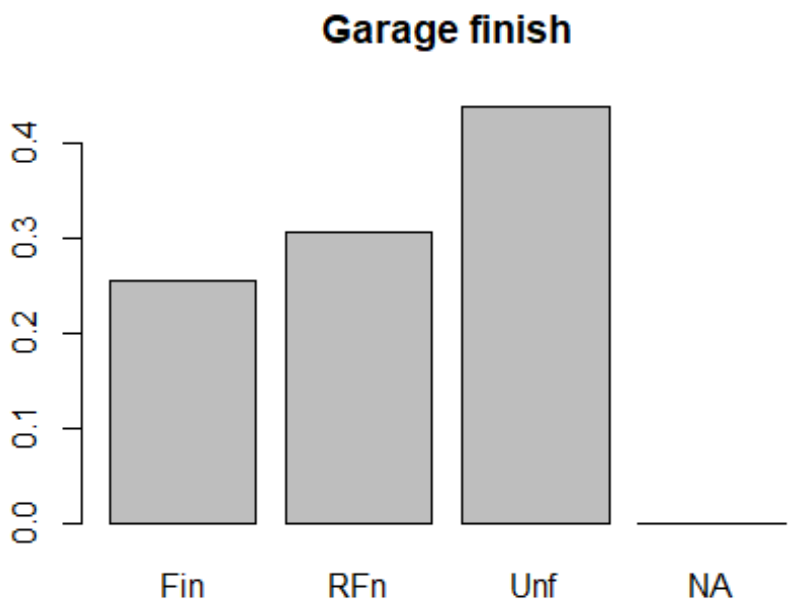



```
##             2Types     Attchd    Basment     BuiltIn     CarPort      Detchd
## DistAs 6.000000000 870.000000 19.0000000 88.00000000 9.000000000 387.0000000
## DistRe 0.004350979   0.630892  0.0137781  0.06381436 0.006526468   0.2806381
```

Posizione garage Il 63% ha il garage affiancato alla casa

Variabile GarageFinish

```
case$GarageFinish <- factor(replace(case$GarageFinish, is.na(case$MSSubClass),
"Non Presente"), levels = c('Fin', 'RFn', 'Unf', 'NA'))

## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA

display_table(case$GarageFinish, titolo = 'Garage finish')
```

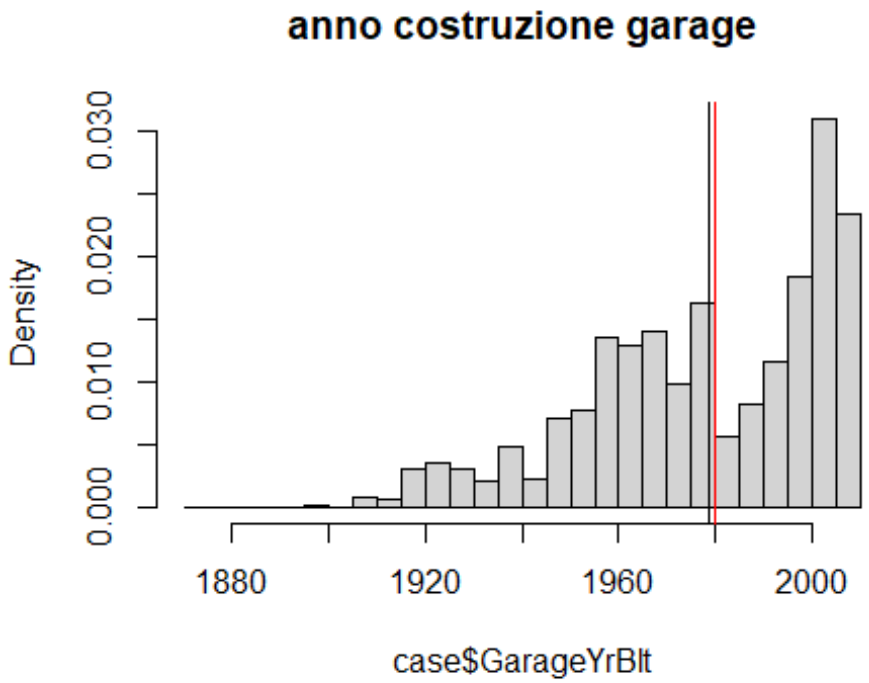



```
##                 Fin         RFn         Unf NA
## DistAs 352.0000000 422.0000000 605.0000000  0
## DistRe   0.2552574   0.3060189   0.4387237  0
```

Indica lo stato del garage Il 43% delle case ha il garage non finito

Variabile GarageYrBlt

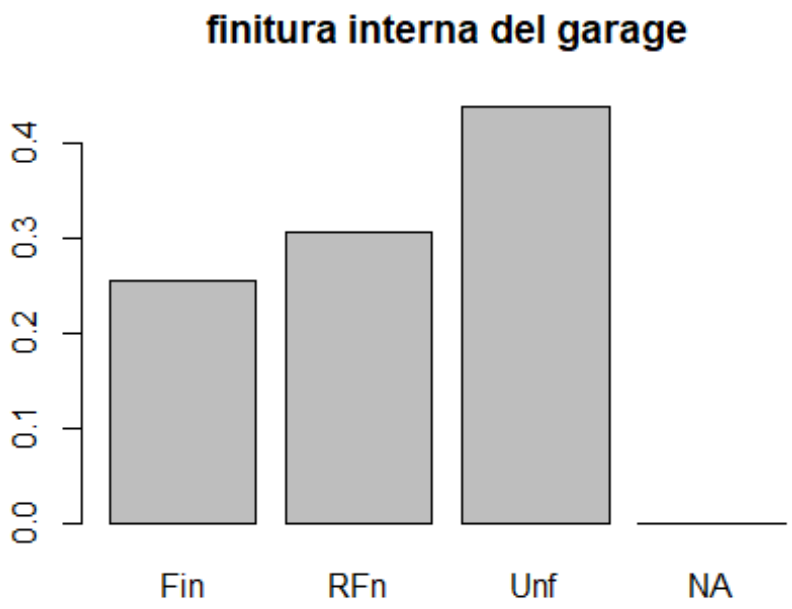
```
hist(case$GarageYrBlt, probability = T,breaks= c(5*0:28)+1870,  main = 'anno
costruzione garage')
abline(v = median(case$GarageYrBlt, na.rm = T),lwd = 1, col = "red")
abline(v = mean(case$GarageYrBlt, na.rm = T),lwd = 1)
```



Indica l'anno di costruzione del garage Si nota un calo tra il 1980 e il 1990 e un picco nei primi anni del 2000

Variabile GarageFinis

```
display_table(case$GarageFinis, "finitura interna del garage")
```




```
##                Fin         RFn         Unf NA
## DistAs 352.0000000 422.0000000 605.0000000  0
## DistRe   0.2552574   0.3060189   0.4387237  0
```

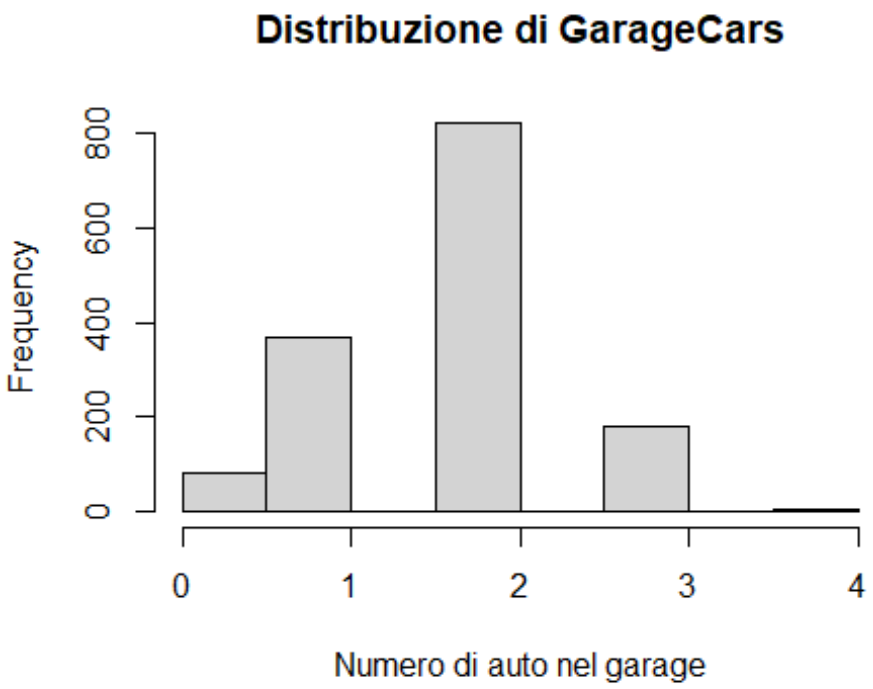
GarageFinis è una variabile categoriale che rappresenta la finitura interna del garage si nota che la maggior parte dei garage nel campione è di tipo Unfinished che rappresenta il 43%

Variabile GarageCars

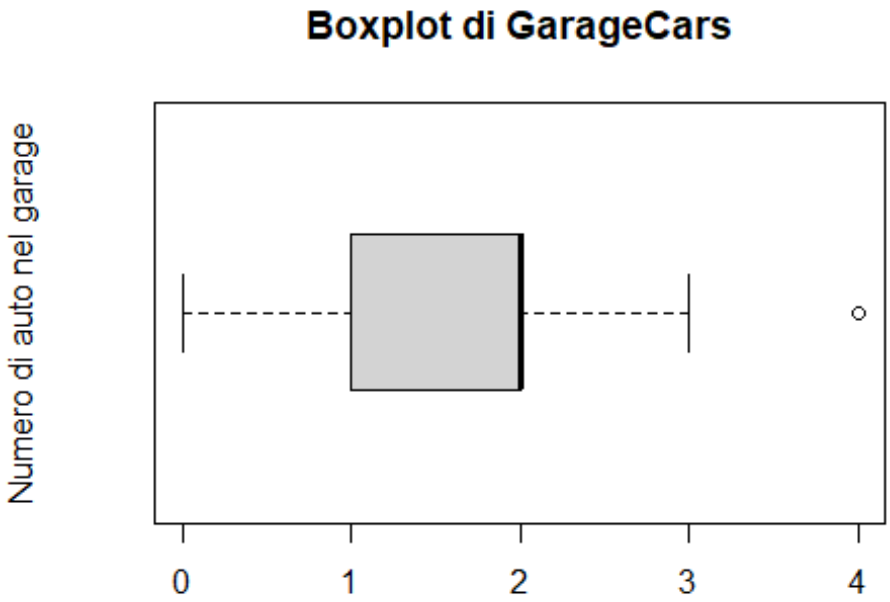
```
display_summary_and_var(case$GarageCars)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.        var
##  0.0000000  1.0000000  2.0000000  1.7671233  2.0000000  4.0000000  0.5584797
##         sd         sk
##  0.7473150 -0.3421969
```

```
hist(case$GarageCars, main = "Distribuzione di GarageCars", xlab = "Numero di auto
nel garage")
```




```r
boxplot(case$GarageCars, main = "Boxplot di GarageCars", ylab = "Numero di auto nel garage", horizontal = T)
```



La variabile GarageCars rappresenta il numero di auto che possono essere alloggiate nel garage. Il boxplot mostra che la maggior parte delle case ha uno o due posti auto nel garage, con poche

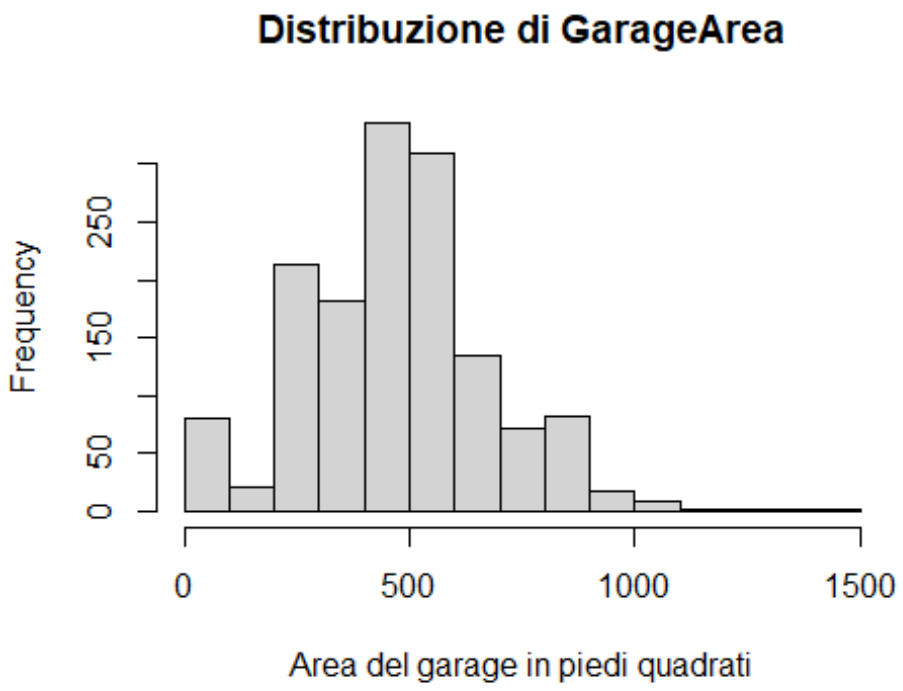
eccezioni che hanno più di due posti auto. L'istogramma mostra una distribuzione simile, con una concentrazione intorno ai valori bassi e qualche outlier con valori più alti. L'analisi della skewness suggerisce una leggera coda a destra nella distribuzione, indicando una maggior concentrazione di case con un numero inferiore di posti auto nel garage.

Variabile GarageArea

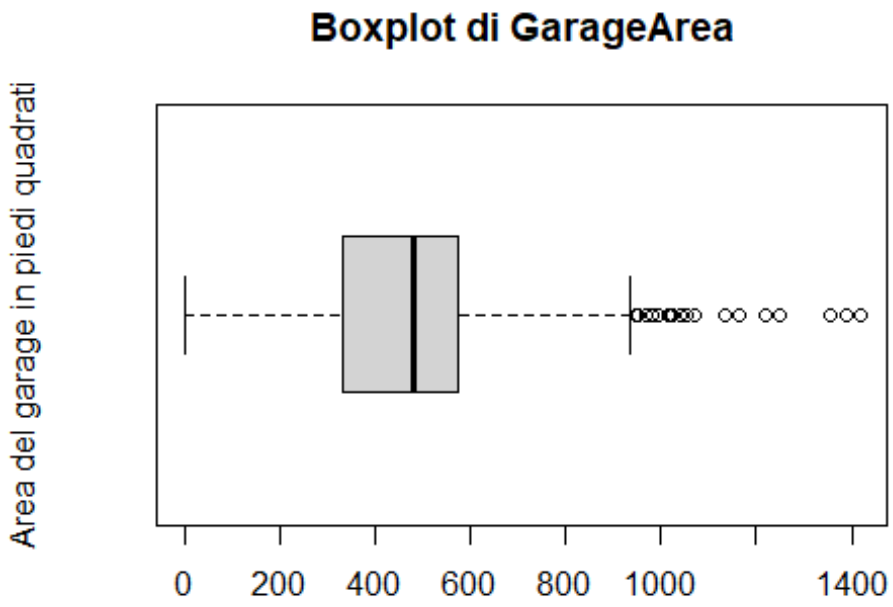
```
display_summary_and_var(case$GarageArea)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 0.000000e+00 3.345000e+02 4.800000e+02 4.729801e+02 5.760000e+02 1.418000e+03
##          var           sd           sk
## 4.571251e+04 2.138048e+02 1.797959e-01
```

```
hist(case$GarageArea, main = "Distribuzione di GarageArea", xlab = "Area del
garage in piedi quadrati")
```



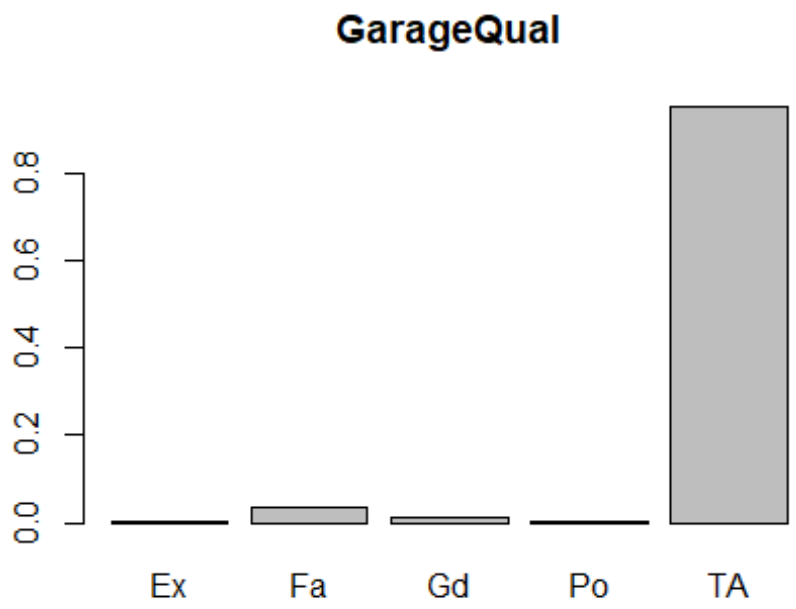

```
boxplot(case$GarageArea, main = "Boxplot di GarageArea", ylab = "Area del garage
in piedi quadrati", horizontal = T)
```



La variabile GarageArea rappresenta l'area del garage in piedi quadrati. Il boxplot mostra una vasta gamma di aree del garage, con alcune case che hanno garage molto grandi rispetto alla media. L'istogramma e la densità dei dati indicano una distribuzione asimmetrica, con una coda a destra e un picco intorno alle aree più basse. Ciò suggerisce che la maggior parte delle case ha garage di dimensioni moderate, ma ci sono alcune case con garage molto grandi.

Variabile GarageQual

```
display_table(case$GarageQual, "GarageQual")
```

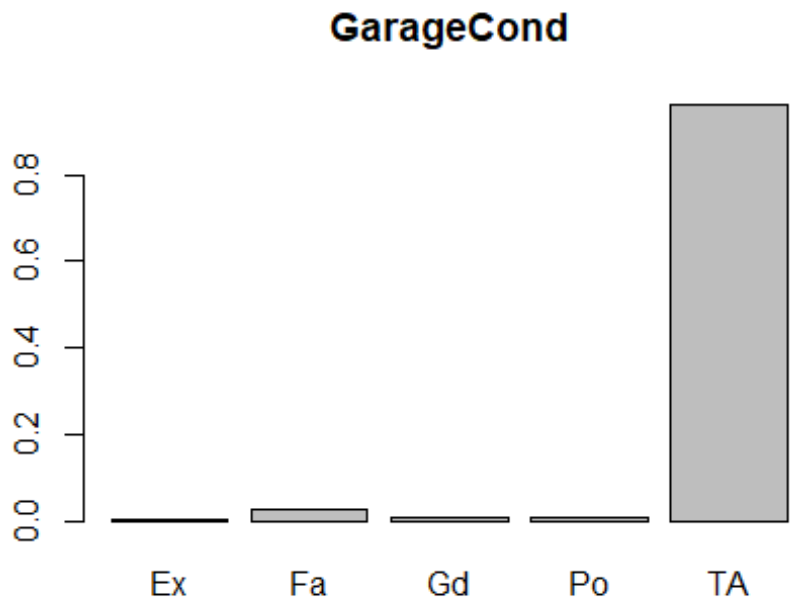



```
##                  Ex          Fa          Gd          Po           TA
## DistAs 3.000000000 48.00000000 14.00000000 3.000000000 1311.0000000
## DistRe 0.002175489  0.03480783  0.01015228 0.002175489    0.9506889
```

La variabile GarageQual rappresenta la qualità del garage. Il barplot mostra la distribuzione delle diverse categorie di qualità del garage. La maggior parte delle case ha una qualità media o tipica del garage, con poche eccezioni che hanno una qualità eccellente o scarsa.

Variabile GarageCond

```
display_table(case$GarageCond, "GarageCond")
```

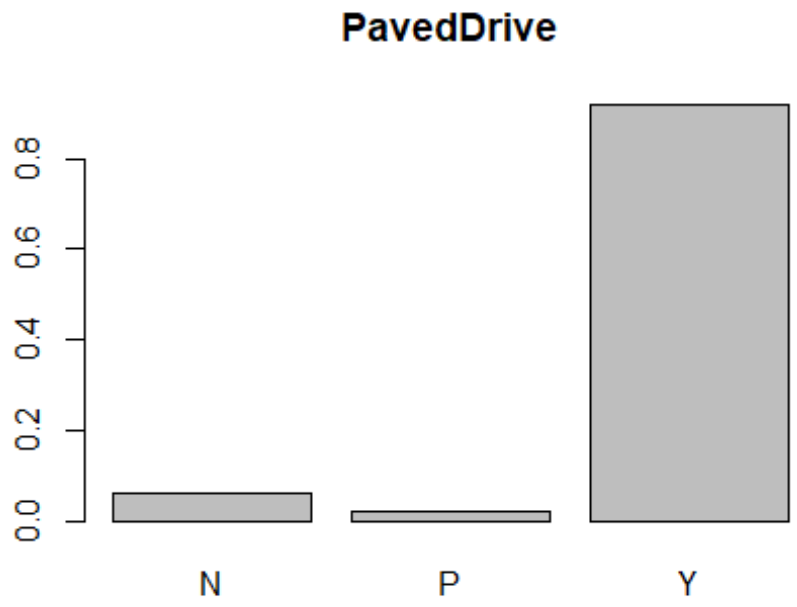



```
##                  Ex          Fa          Gd          Po           TA
## DistAs 2.000000000 35.00000000 9.000000000 7.000000000 1326.0000000
## DistRe 0.001450326  0.02538071 0.006526468 0.005076142    0.9615664
```

La variabile GarageCond rappresenta le condizioni del garage. Il barplot mostra la distribuzione delle diverse categorie di condizioni del garage. La maggior parte delle case ha condizioni medie o tipiche del garage, con poche eccezioni che hanno condizioni eccellenti o pessime.

Variabile PavedDrive

```
display_table(case$PavedDrive, "PavedDrive")
```




```
##                   N           P            Y
## DistAs 90.00000000 30.00000000 1340.0000000
## DistRe  0.06164384  0.02054795    0.9178082
```

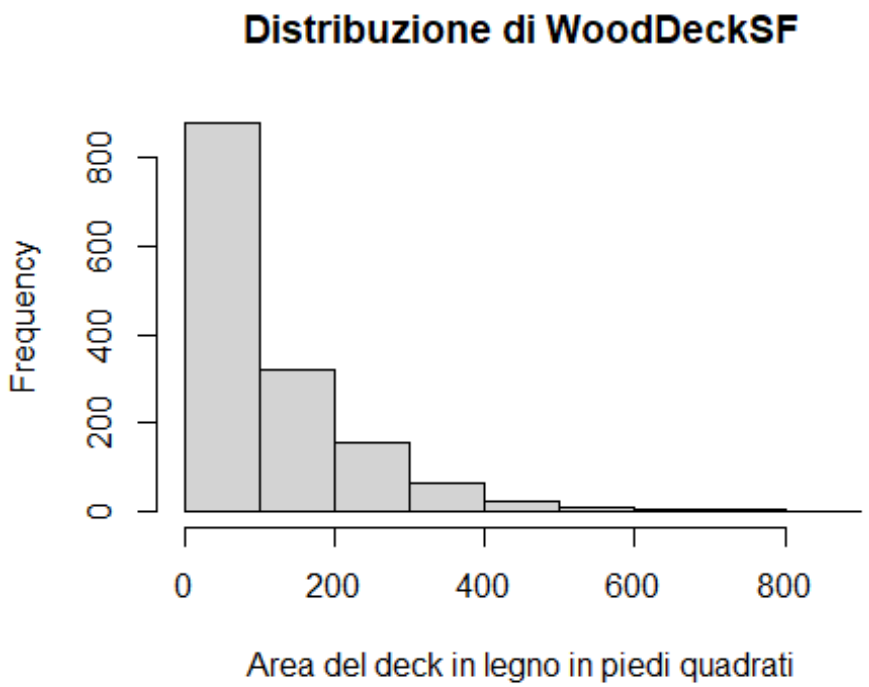
La variabile PavedDrive indica se la via di accesso è pavimentata. Il barplot mostra la distribuzione delle diverse categorie di tipi di via di accesso. La maggior parte delle case ha una via di accesso pavimentata, con poche eccezioni che hanno un accesso parzialmente pavimentato o non pavimentato.

Variabile WoodDeckSF

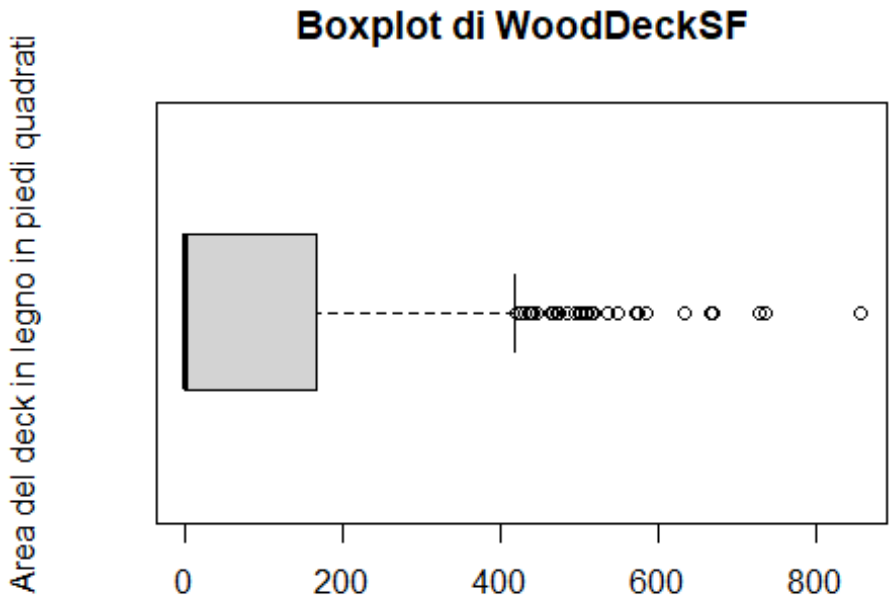
```
display_summary_and_var(case$WoodDeckSF)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##     0.000000     0.000000     0.000000    94.244521   168.000000   857.000000
##          var           sd           sk
## 15709.813370   125.338794     1.539792
```

```
hist(case$WoodDeckSF, main = "Distribuzione di WoodDeckSF", xlab = "Area del deck
in legno in piedi quadrati")
```




```r
boxplot(case$WoodDeckSF, main = "Boxplot di WoodDeckSF", ylab = "Area del deck in legno in piedi quadrati", horizontal = T)
```



La variabile WoodDeckSF rappresenta l'area del deck in legno in piedi quadrati. Il boxplot mostra una vasta gamma di dimensioni del deck in legno, con una concentrazione intorno ai valori bassi

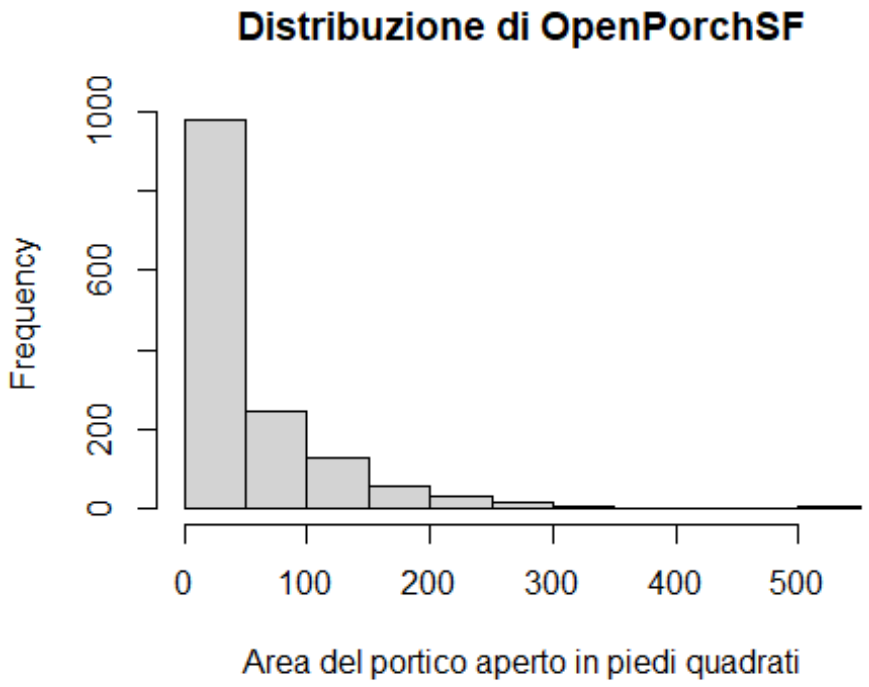
e alcuni outlier con dimensioni più grandi. L'istogramma e la densità dei dati indicano una distribuzione asimmetrica, con una coda a destra e un picco intorno alle dimensioni più basse. Questo suggerisce che la maggior parte delle case ha deck in legno di dimensioni moderate, ma ci sono alcune case con deck molto grandi. L'analisi della skewness mostra una leggera coda a destra nella distribuzione, indicando una maggiore concentrazione di case con dimensioni più basse del deck in legno.

Variabile OpenPorchSF

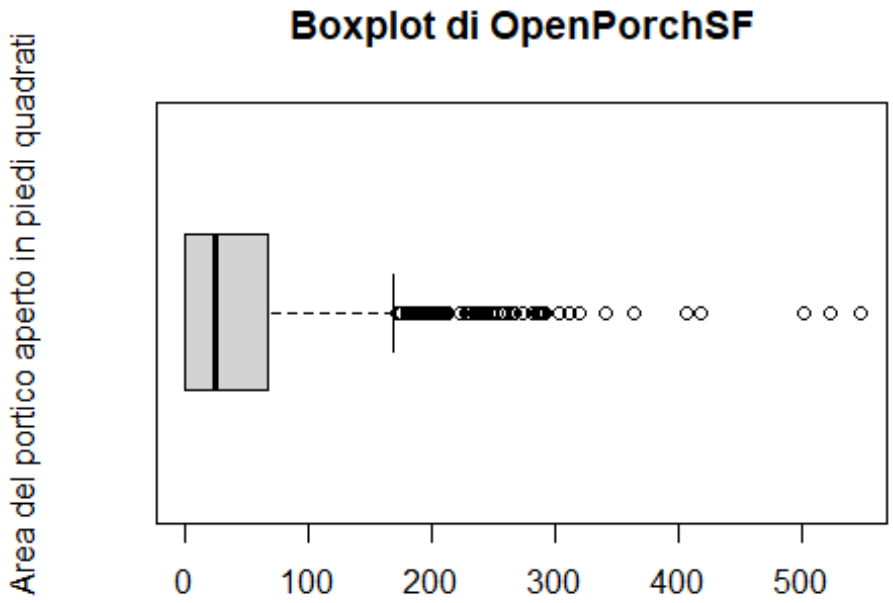
```
display_summary_and_var(case$OpenPorchSF)
```

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000    0.000000   25.000000   46.660274   68.000000  547.000000
##         var          sd          sk
## 4389.861203   66.256028    2.361912
```

```
hist(case$OpenPorchSF, main = "Distribuzione di OpenPorchSF", xlab = "Area del
portico aperto in piedi quadrati")
```




```
boxplot(case$OpenPorchSF, main = "Boxplot di OpenPorchSF", ylab = "Area del
portico aperto in piedi quadrati", horizontal = T)
```



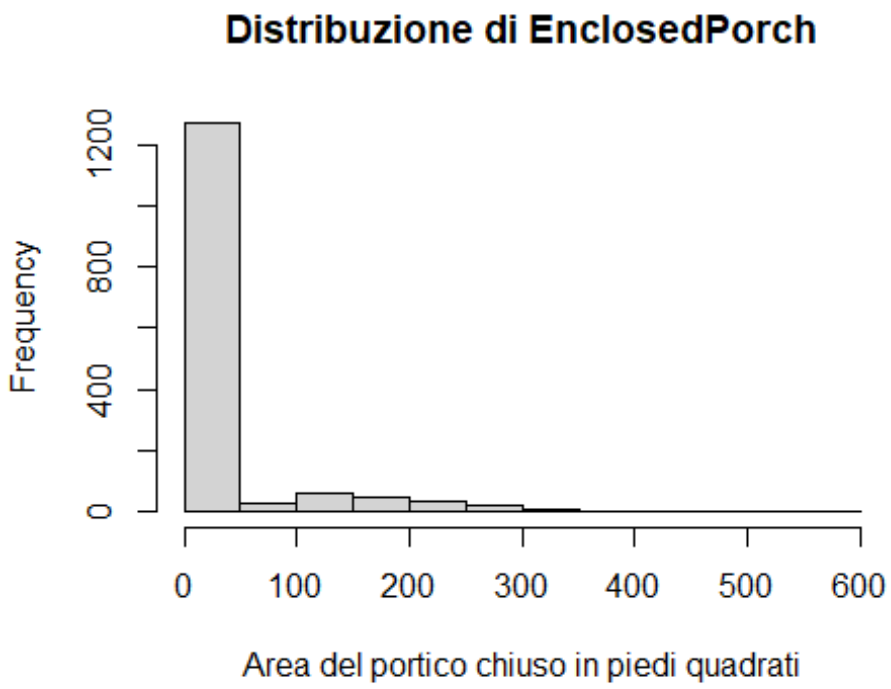
La variabile OpenPorchSF rappresenta l'area del portico aperto in piedi quadrati. Il boxplot mostra una vasta gamma di dimensioni del portico aperto, con una concentrazione intorno ai valori bassi e alcuni outlier con dimensioni più grandi. L'istogramma e la densità dei dati indicano una distribuzione asimmetrica, con una coda a destra e un picco intorno alle dimensioni più basse. Questo suggerisce che la maggior parte delle case ha portici aperti di dimensioni moderate, ma ci sono alcune case con portici molto grandi. L'analisi della skewness mostra una coda a destra nella distribuzione, indicando una maggiore concentrazione di case con dimensioni più basse del portico aperto.

Variabile EnclosedPorch

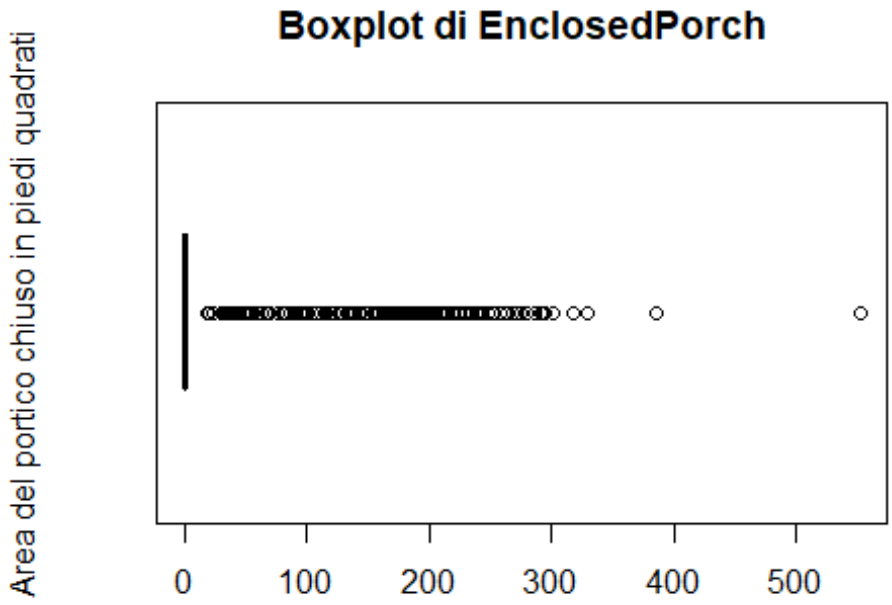
```
display_summary_and_var(case$EnclosedPorch)
```

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000    0.000000    0.000000   21.954110    0.000000  552.000000
##         var          sd          sk
## 3735.550326   61.119149    3.086696
```

```
hist(case$EnclosedPorch, main = "Distribuzione di EnclosedPorch", xlab = "Area del
portico chiuso in piedi quadrati")
```




```r
boxplot(case$EnclosedPorch, main = "Boxplot di EnclosedPorch", ylab = "Area del portico chiuso in piedi quadrati", horizontal = T)
```



La variabile EnclosedPorch rappresenta l'area del portico chiuso in piedi quadrati. Il boxplot mostra una vasta gamma di dimensioni del portico chiuso, con una concentrazione intorno ai

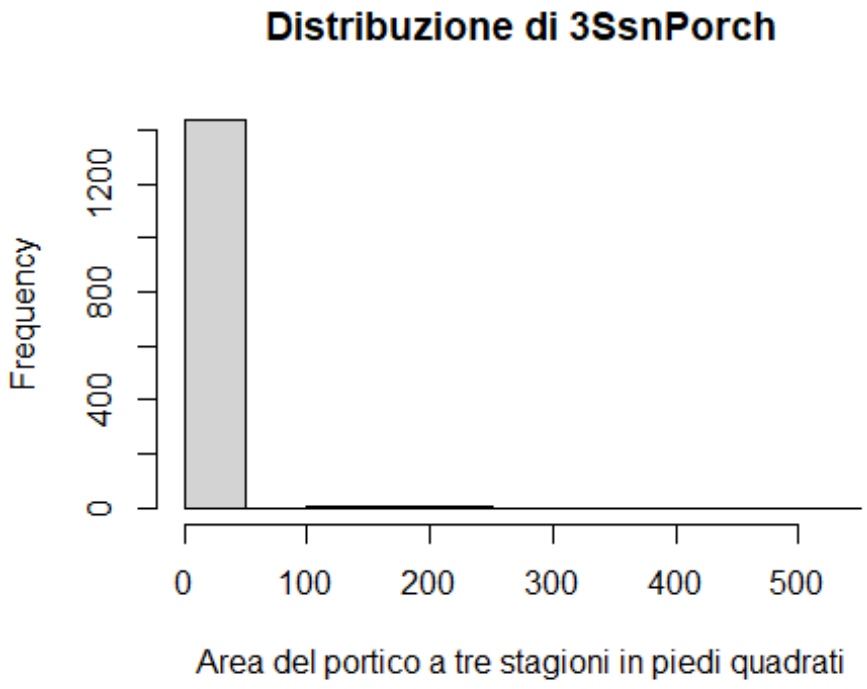
valori bassi e alcuni outlier con dimensioni più grandi. L'istogramma e la densità dei dati indicano una distribuzione asimmetrica, con una coda a destra e un picco intorno alle dimensioni più basse. Questo suggerisce che la maggior parte delle case ha portici chiusi di dimensioni moderate, ma ci sono alcune case con portici molto grandi. L'analisi della skewness mostra una coda a destra nella distribuzione, indicando una maggiore concentrazione di case con dimensioni più basse del portico chiuso.

Variabile X3SsnPorch

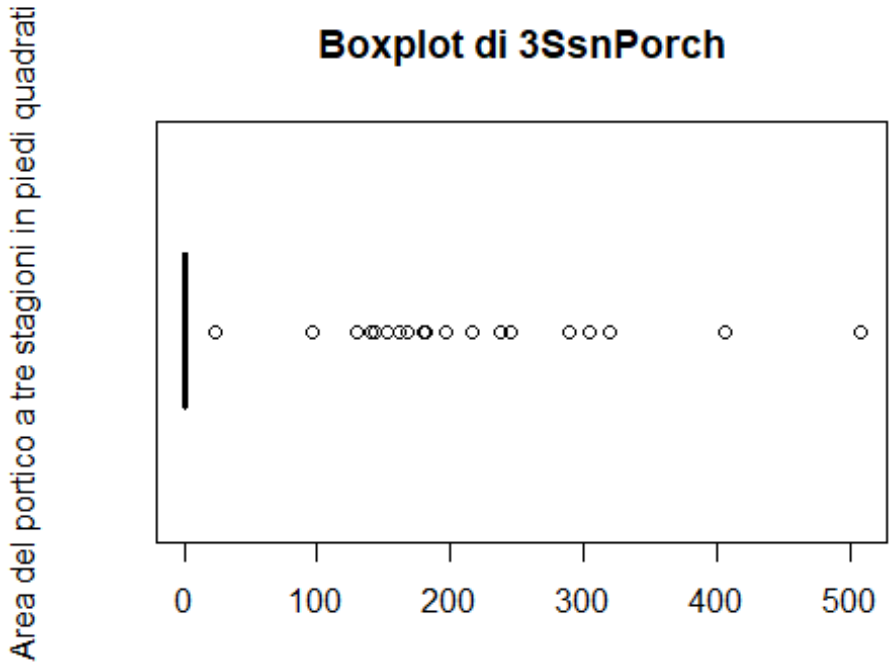
```
variabile <- case$X3SsnPorch
display_summary_and_var(case$X3SsnPorch)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.        var
##   0.000000   0.000000   0.000000   3.409589   0.000000 508.000000 859.505871
##         sd         sk
##  29.317331  10.293752
```

```
hist(case$X3SsnPorch, main = "Distribuzione di 3SsnPorch", xlab = "Area del
portico a tre stagioni in piedi quadrati")
```




```
boxplot(case$X3SsnPorch, main = "Boxplot di 3SsnPorch", ylab = "Area del portico a
tre stagioni in piedi quadrati", horizontal = T)
```



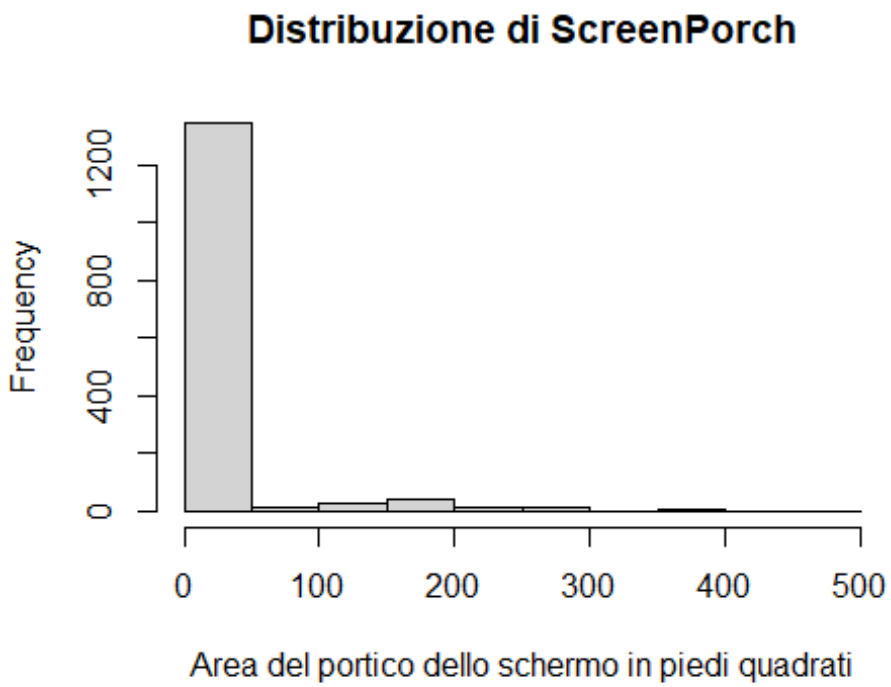
La variabile 3SsnPorch rappresenta l'area del portico a tre stagioni in piedi quadrati. Il boxplot mostra una vasta gamma di dimensioni del portico a tre stagioni, con una concentrazione intorno ai valori bassi e alcuni outlier con dimensioni più grandi. L'istogramma e la densità dei dati indicano una distribuzione asimmetrica, con una coda a destra e un picco intorno alle dimensioni più basse. Questo suggerisce che la maggior parte delle case ha portici a tre stagioni di dimensioni moderate, ma ci sono alcune case con portici molto grandi. L'analisi della skewness mostra una leggera coda a destra nella distribuzione, indicando una maggiore concentrazione di case con dimensioni più basse del portico a tre stagioni.

Variabile ScreenPorch

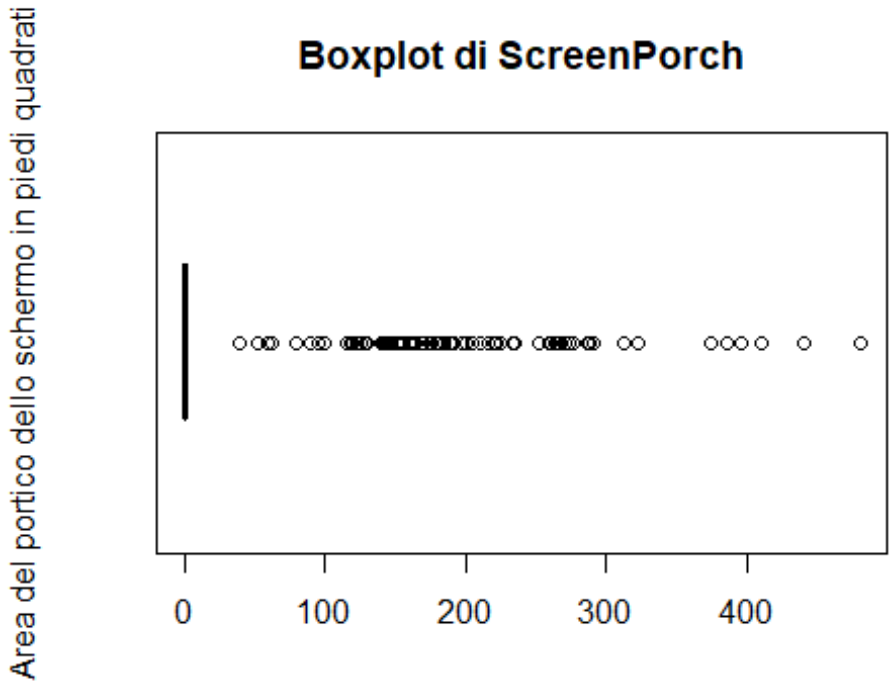
```
variabile <- case$ScreenPorch
display_summary_and_var(case$ScreenPorch)
```

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000    0.000000    0.000000   15.060959    0.000000  480.000000
##         var          sd          sk
## 3108.889359   55.757415    4.117977
```

```
hist(case$ScreenPorch, main = "Distribuzione di ScreenPorch", xlab = "Area del
portico dello schermo in piedi quadrati")
```




```r
boxplot(case$ScreenPorch, main = "Boxplot di ScreenPorch", ylab = "Area del
portico dello schermo in piedi quadrati", horizontal = T)
```



La variabile ScreenPorch rappresenta l'area del portico dello schermo in piedi quadrati. Il boxplot mostra una vasta gamma di dimensioni del portico dello schermo, con una concentrazione

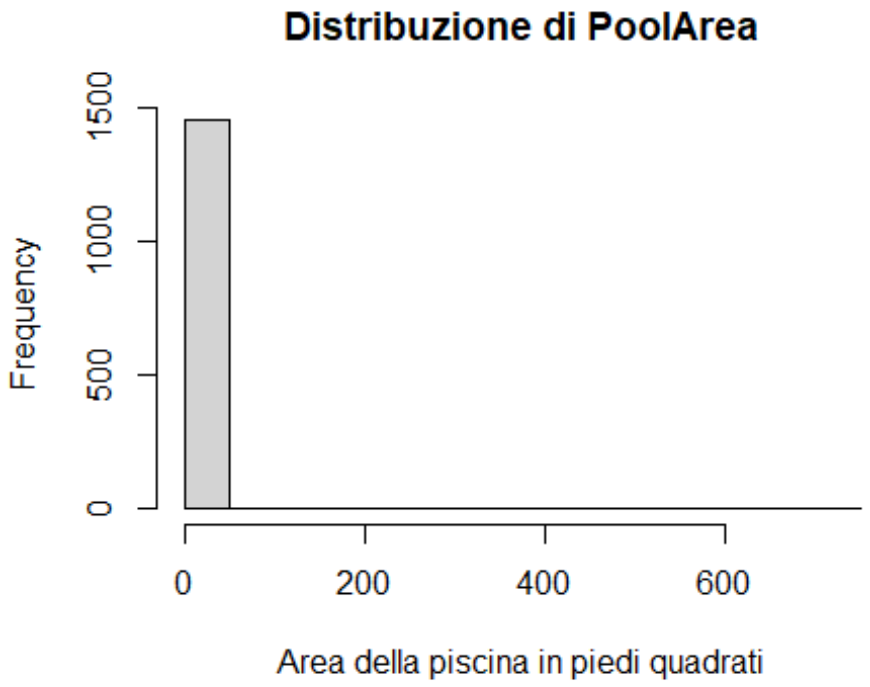
intorno ai valori bassi e alcuni outlier con dimensioni più grandi. L'istogramma e la densità dei dati indicano una distribuzione asimmetrica, con una coda a destra e un picco intorno alle dimensioni più basse. Questo suggerisce che la maggior parte delle case ha portici dello schermo di dimensioni moderate, ma ci sono alcune case con portici molto grandi. L'analisi della skewness mostra una coda a destra nella distribuzione, indicando una maggiore concentrazione di case con dimensioni più basse del portico dello schermo.

Variabile PoolArea

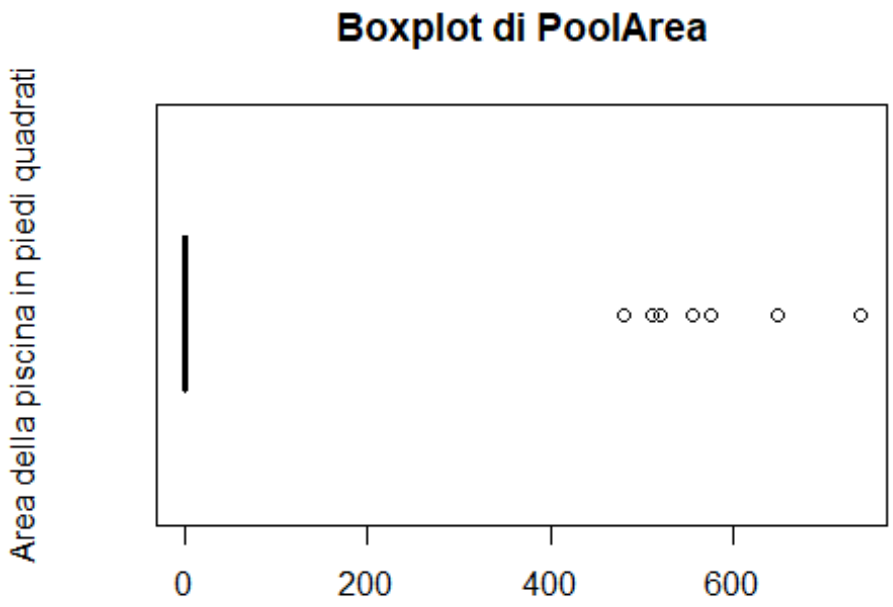
```
display_summary_and_var(case$PoolArea)
```

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000    0.000000    0.000000    2.758904    0.000000  738.000000
##         var          sd          sk
## 1614.215993   40.177307   14.813135
```

```
hist(case$PoolArea, main = "Distribuzione di PoolArea", xlab = "Area della piscina
in piedi quadrati")
```



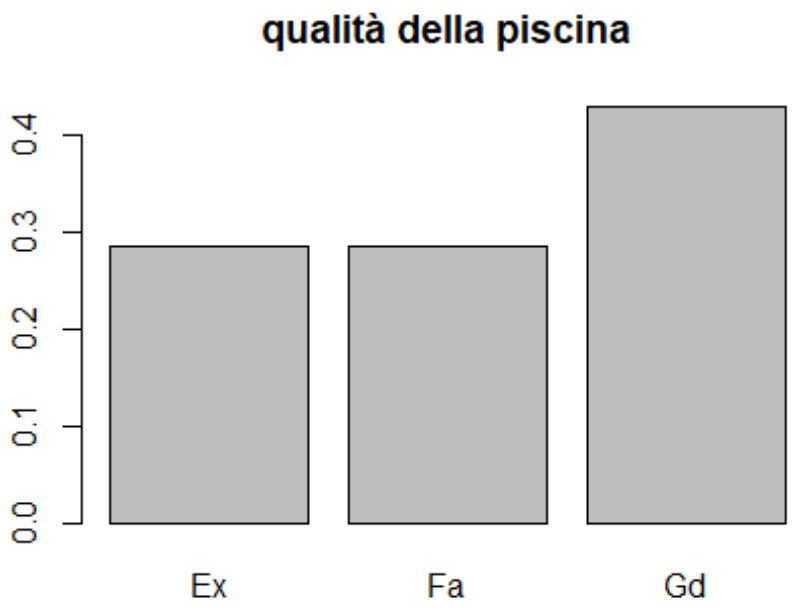

```
boxplot(case$PoolArea, main = "Boxplot di PoolArea", ylab = "Area della piscina in
piedi quadrati", horizontal = T)
```



La variabile PoolArea rappresenta l'area della piscina in piedi quadrati. Il boxplot mostra una vasta gamma di dimensioni della piscina, con una concentrazione intorno ai valori bassi e alcuni outlier con dimensioni più grandi. L'istogramma e la densità dei dati indicano una distribuzione asimmetrica, con una coda a destra e un picco intorno alle dimensioni più basse. Questo suggerisce che la maggior parte delle case ha piscine di dimensioni moderate, ma ci sono alcune case con piscine molto grandi. L'analisi della skewness mostra una coda a destra nella distribuzione, indicando una maggiore concentrazione di case con dimensioni più basse della piscina.

Variabile PoolQC

```
display_table(case$PoolQC, "qualità della piscina")
```

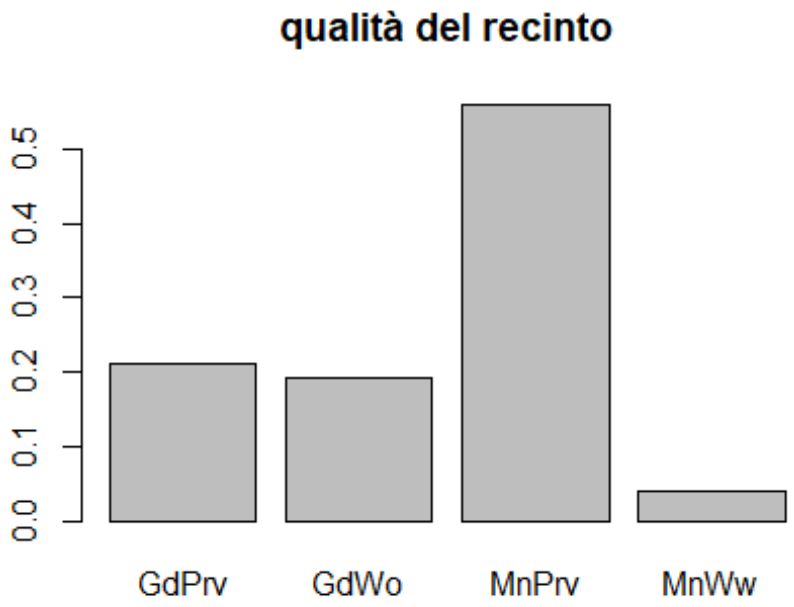



```
##                Ex        Fa        Gd
## DistAs 2.0000000 2.0000000 3.0000000
## DistRe 0.2857143 0.2857143 0.4285714
```

La variabile PoolQC rappresenta la qualità della piscina. Il barplot mostra la distribuzione delle diverse categorie di qualità della piscina. La maggior parte delle case non ha una piscina, con poche eccezioni che hanno piscine di alta qualità.

Variabile Fence

```
display_table(case$Fence, "qualità del recinto")
```

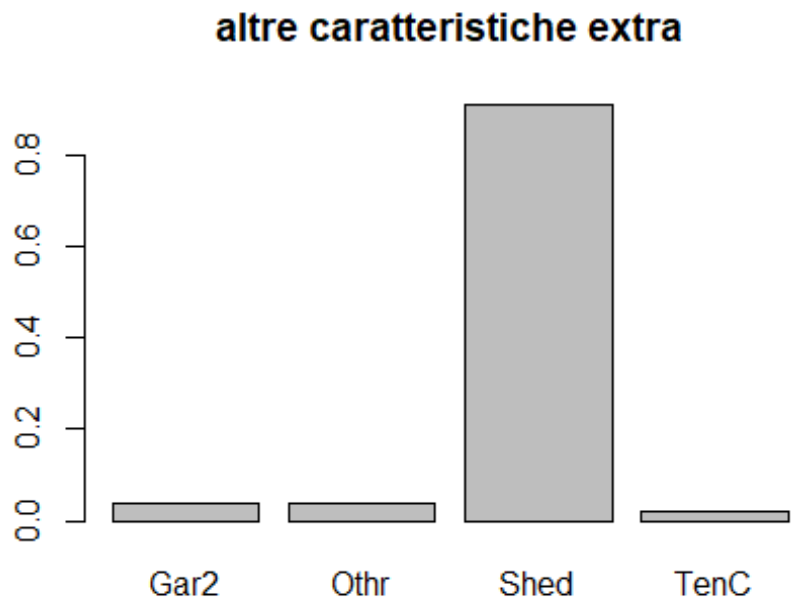



```
##             GdPrv       GdWo       MnPrv        MnWw
## DistAs 59.0000000 54.0000000 157.0000000 11.00000000
## DistRe  0.2099644  0.1921708   0.5587189  0.03914591
```

La variabile Fence rappresenta la qualità del recinto. Il barplot mostra la distribuzione delle diverse categorie di qualità del recinto. La maggior parte delle case non ha un recinto, mentre una piccola parte ha recinti di qualità variabile.

Variabile MiscFeature

```
display_table(case$MiscFeature, "altre caratteristiche extra")
```




```
##               Gar2       Othr       Shed       TenC
## DistAs 2.00000000 2.00000000 49.0000000 1.00000000
## DistRe 0.03703704 0.03703704  0.9074074 0.01851852
```

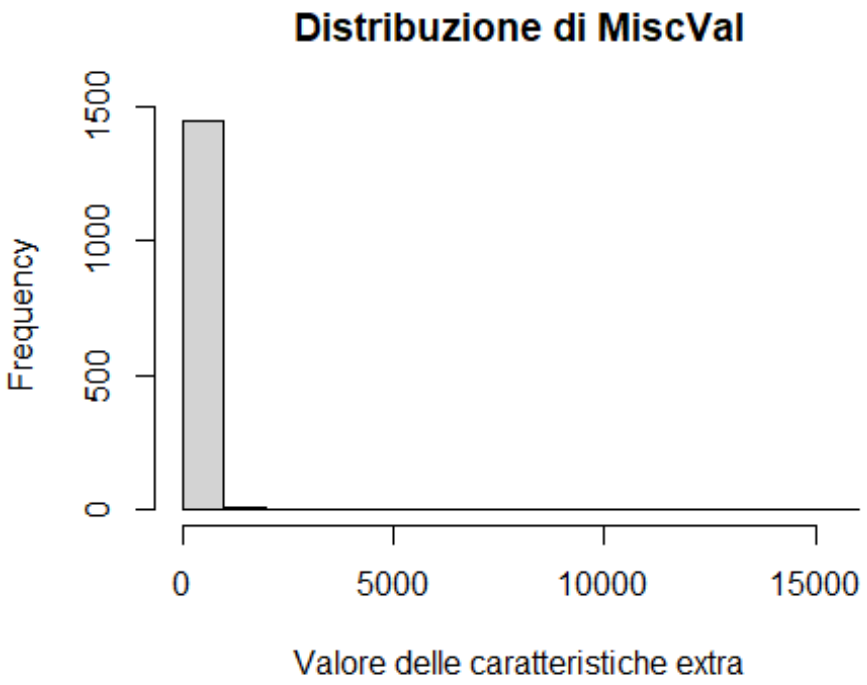
La variabile MiscFeature rappresenta altre caratteristiche extra presenti nelle proprietà. Il barplot mostra la distribuzione delle diverse categorie di caratteristiche extra. La maggior parte delle case non ha caratteristiche extra, con poche eccezioni che includono strutture come ripostigli, recinti e altri.

Variabile MiscVal

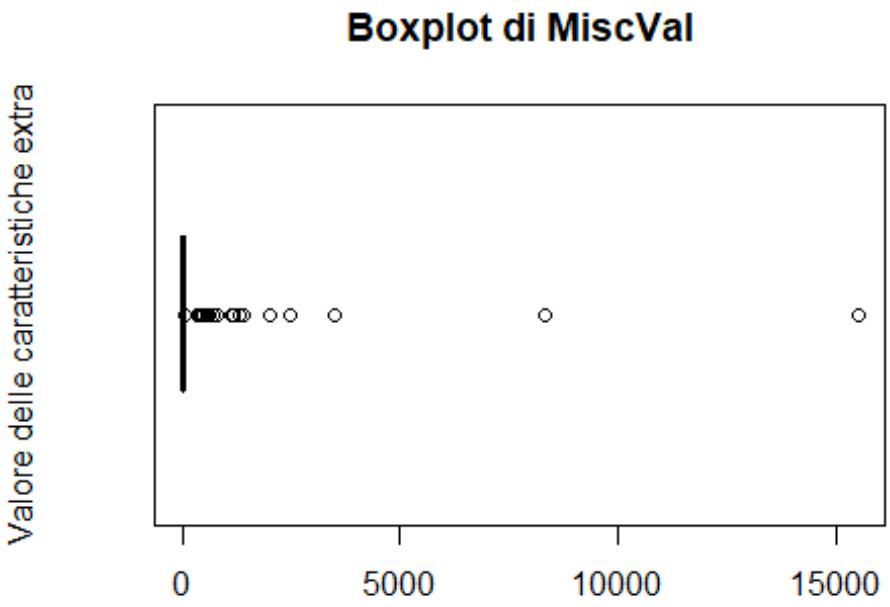
```
print(display_summary_and_var(case$MiscVal))
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##      0.00000      0.00000      0.00000     43.48904      0.00000  15500.00000
##          var           sd           sk
## 246138.05540    496.12302     24.45164
```

```
hist(case$MiscVal, main = "Distribuzione di MiscVal", xlab = "Valore delle
caratteristiche extra")
```




```r
boxplot(case$MiscVal, main = "Boxplot di MiscVal", ylab = "Valore delle caratteristiche extra", horizontal = T)
```



La variabile MiscVal rappresenta il valore delle caratteristiche extra. Il boxplot mostra che la maggior parte delle case ha un valore extra basso, con alcune eccezioni che hanno un valore extra

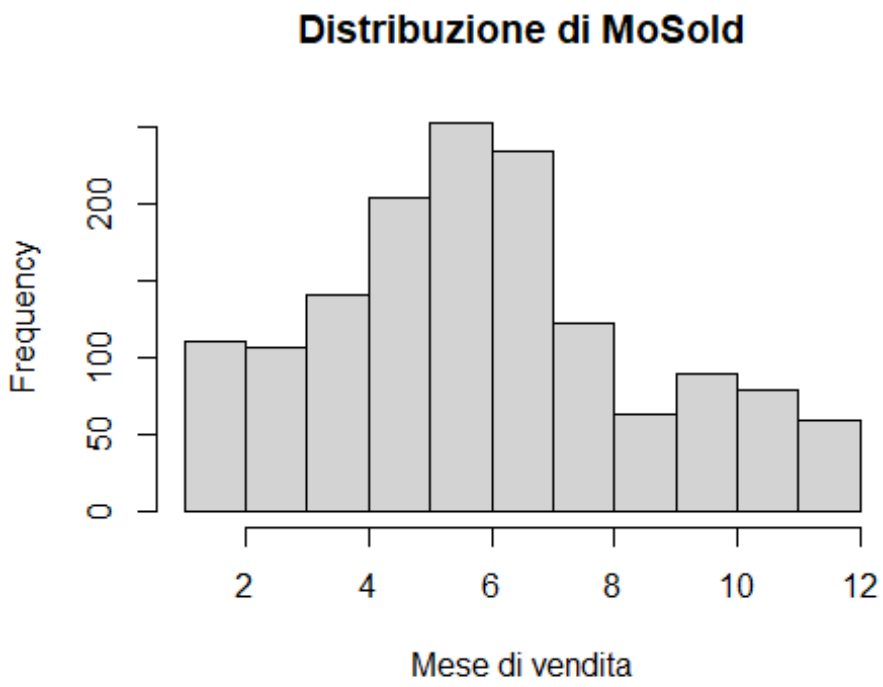
più alto. L'istogramma e la densità dei dati indicano una distribuzione asimmetrica, con una coda a destra. Questo suggerisce che la maggior parte delle case ha valori extra di basso valore, ma ci sono alcune case con valori extra molto alti.

Variabile MoSold

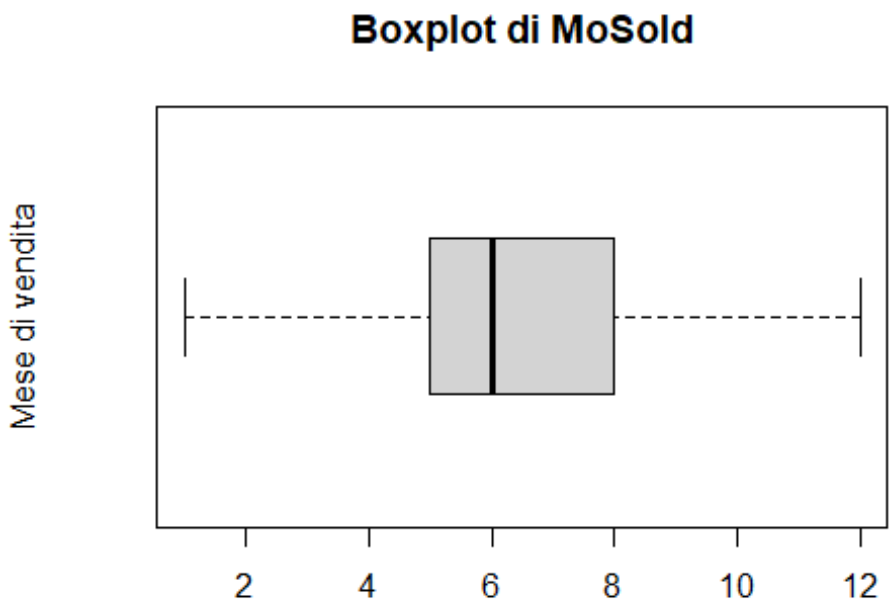
```
display_summary_and_var(case$MoSold)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.        var
##  1.0000000  5.0000000  6.0000000  6.3219178  8.0000000 12.0000000  7.3095947
##         sd         sk
##  2.7036262  0.2118351
```

```
hist(case$MoSold, main = "Distribuzione di MoSold", xlab = "Mese di vendita")
```




```
boxplot(case$MoSold, main = "Boxplot di MoSold", ylab = "Mese di vendita",
horizontal = T)
```



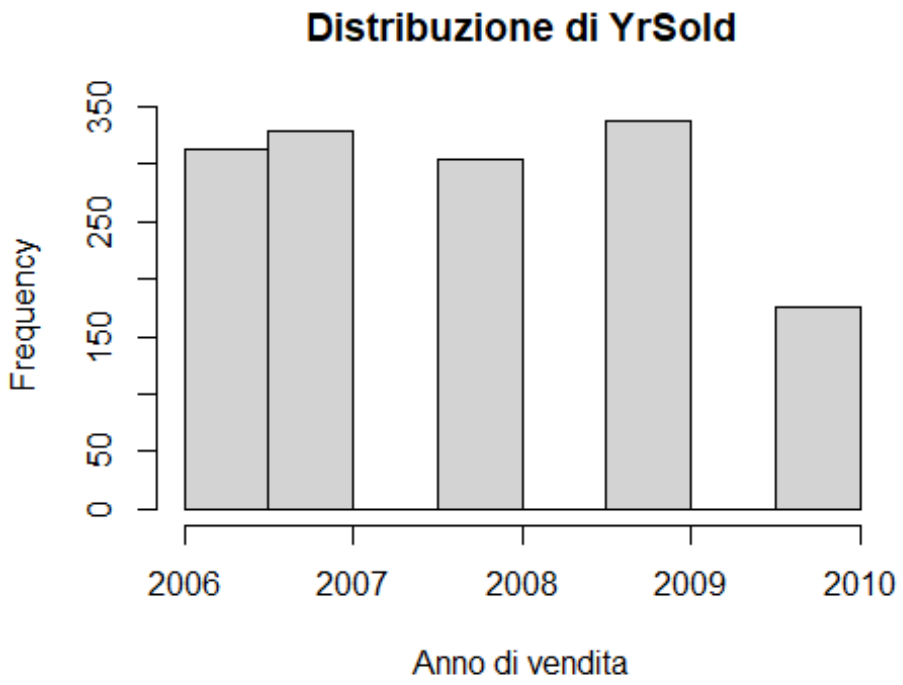
La variabile MoSold rappresenta il mese in cui la casa è stata venduta. Il boxplot mostra che le vendite sono distribuite durante tutto l'anno, con picchi nei mesi di primavera e estate. L'istogramma mostra che i mesi con il maggior numero di vendite sono giugno, luglio e agosto. L'analisi della skewness indica una distribuzione abbastanza uniforme delle vendite durante l'anno.

Variabile YrSold

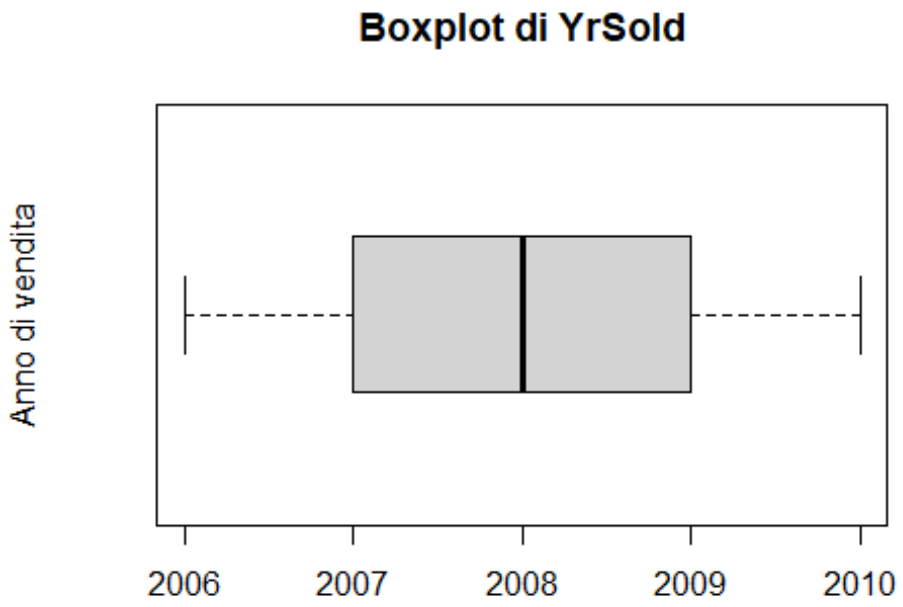
```
display_summary_and_var(case$YrSold)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 2.006000e+03 2.007000e+03 2.008000e+03 2.007816e+03 2.009000e+03 2.010000e+03
##          var           sd           sk
## 1.763837e+00 1.328095e+00 9.616958e-02
```

```
hist(case$YrSold, main = "Distribuzione di YrSold", xlab = "Anno di vendita")
```




```r
boxplot(case$YrSold, main = "Boxplot di YrSold", ylab = "Anno di vendita",
horizontal = T)
```

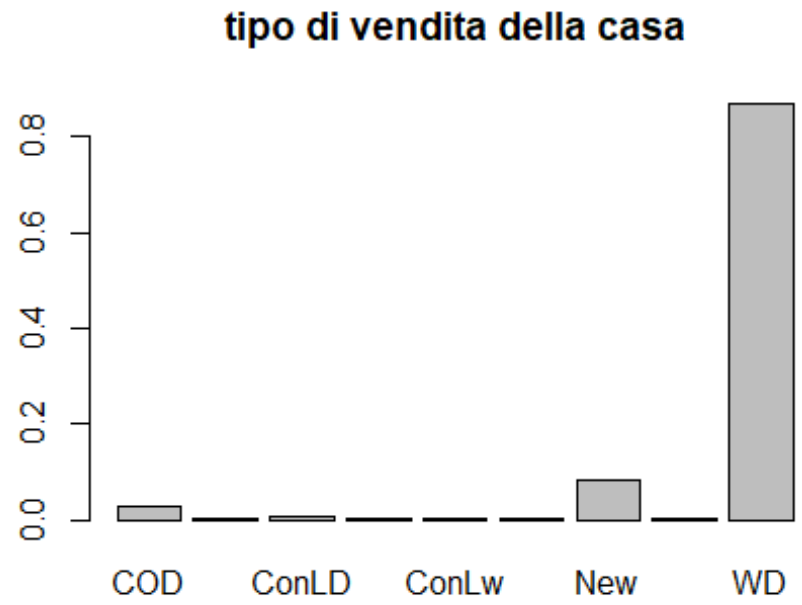


La variabile YrSold rappresenta l'anno in cui la casa è stata venduta. Il boxplot mostra che le vendite sono distribuite uniformemente tra gli anni presenti nel dataset. L'istogramma mostra

che c'è una leggera diminuzione del numero di vendite negli anni più recenti. L'analisi della skewness indica una distribuzione relativamente uniforme delle vendite negli anni considerati.

Variabile SaleType

```
display_table(case$SaleType, "tipo di vendita della casa")
```

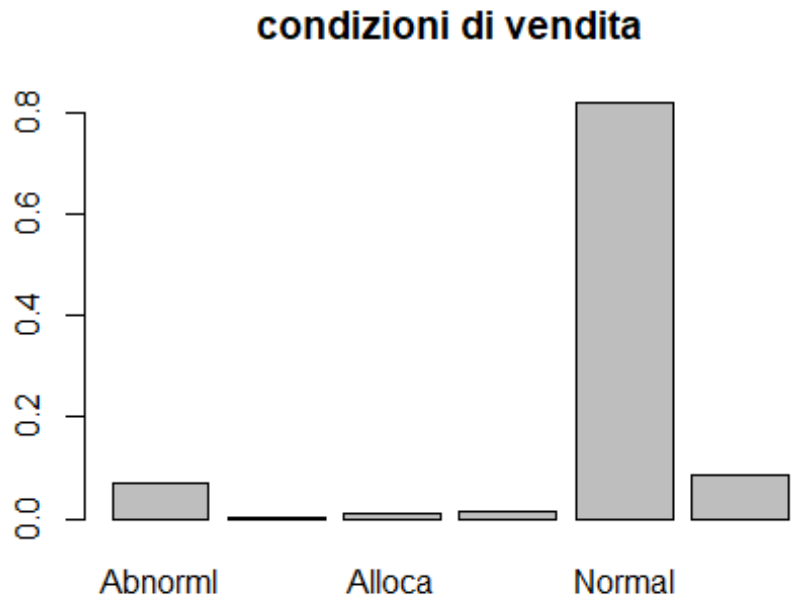



```
##                COD          Con       ConLD       ConLI       ConLw          CWD
## DistAs 43.00000000 2.000000000 9.000000000 5.000000000 5.000000000 4.000000000
## DistRe  0.02945205 0.001369863 0.006164384 0.003424658 0.003424658 0.002739726
##                New          Oth           WD
## DistAs 122.00000000 3.000000000 1267.0000000
## DistRe   0.08356164 0.002054795    0.8678082
```

La variabile SaleType rappresenta il tipo di vendita della casa. # Il barplot mostra la distribuzione delle diverse categorie di tipi di vendita. # La maggior parte delle case è venduta con una vendita normale, con altre categorie come vendita in contanti, vendita con finanziamento, e altre meno comuni.

Variabile SaleCondition

```
display_table(case$SaleCondition, "condizioni di vendita")
```




```
##             Abnorml     AdjLand       Alloca      Family       Normal
## DistAs 101.00000000 4.000000000 12.000000000 20.00000000 1198.0000000
## DistRe   0.06917808 0.002739726  0.008219178  0.01369863    0.8205479
##             Partial
## DistAs 125.00000000
## DistRe   0.08561644
```

La variabile SaleCondition rappresenta le condizioni di vendita della casa. Il barplot mostra la distribuzione delle diverse categorie di condizioni di vendita. La maggior parte delle case è venduta in condizioni normali, con altre categorie come vendite anomale, vendite parziali e altre meno comuni.

Dataset - House Prices

Analisi bivariata

Variabile MSSubClass

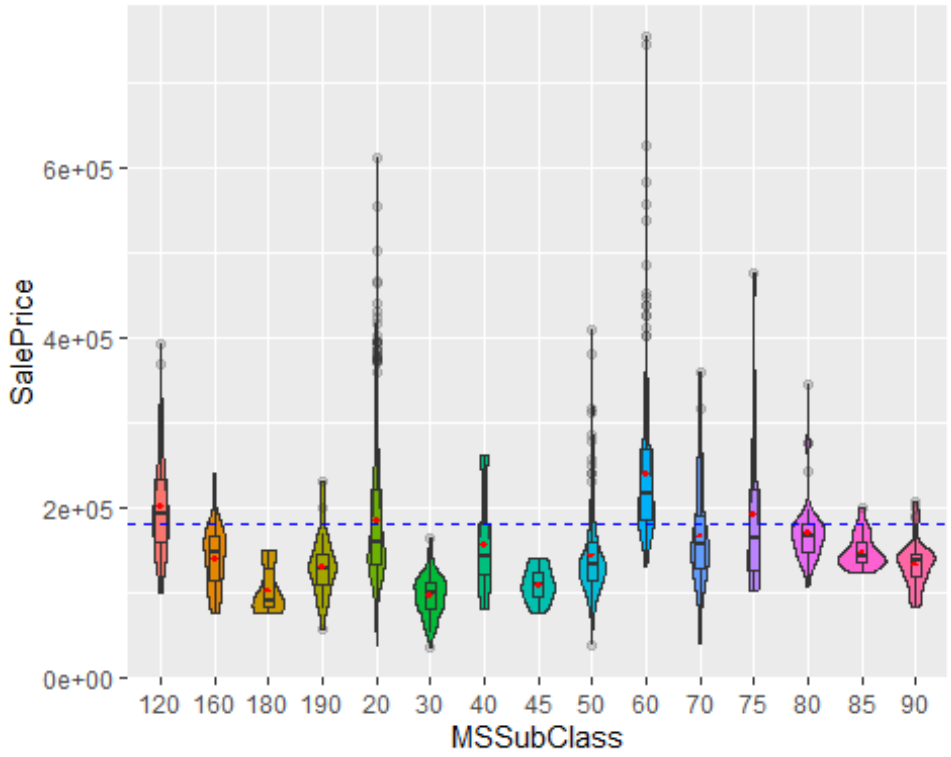
```
calcola_devianza(case$SalePrice, factor(case$MSSubClass))
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 2.268056e+12
##
## $devianza_entro_gruppi
## [1] 6.939856e+12
##
```

```
## $eta2
## [1] 0.246316
```

```
ggplot(case, aes(x = factor(MSSubClass), y = SalePrice, fill =
factor(MSSubClass))) + geom_violin() + geom_boxplot(width=0.2, alpha=1/5) +
guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape = 18, size =
1, color = "red") +  geom_hline(yintercept = mean(case$SalePrice), linetype =
"dashed", color = "blue") + labs(x = "MSSubClass")
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none"
instead as
## of ggplot2 3.3.4.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Visto che stiamo lavorando con la variabile "SalePrice" come target avremo in generale un valore molto alto per la devianza. La devianza totale è infatti 9.207911e+12 che in questo caso vediamo essere principalmente Devianza "Entro". Eta Quadro è infatti 0.246 il che mi conferma che la variazione delle medie tra i gruppi è meno marcata della variazione dei valori nei gruppi stessi. Il grafico ci mostra infatti che tutte le medie dei gruppi sono vicine alla media generale mentre classi come la "60" e la "20" contengono anche valori di "SalePrice" decisamente elevati.

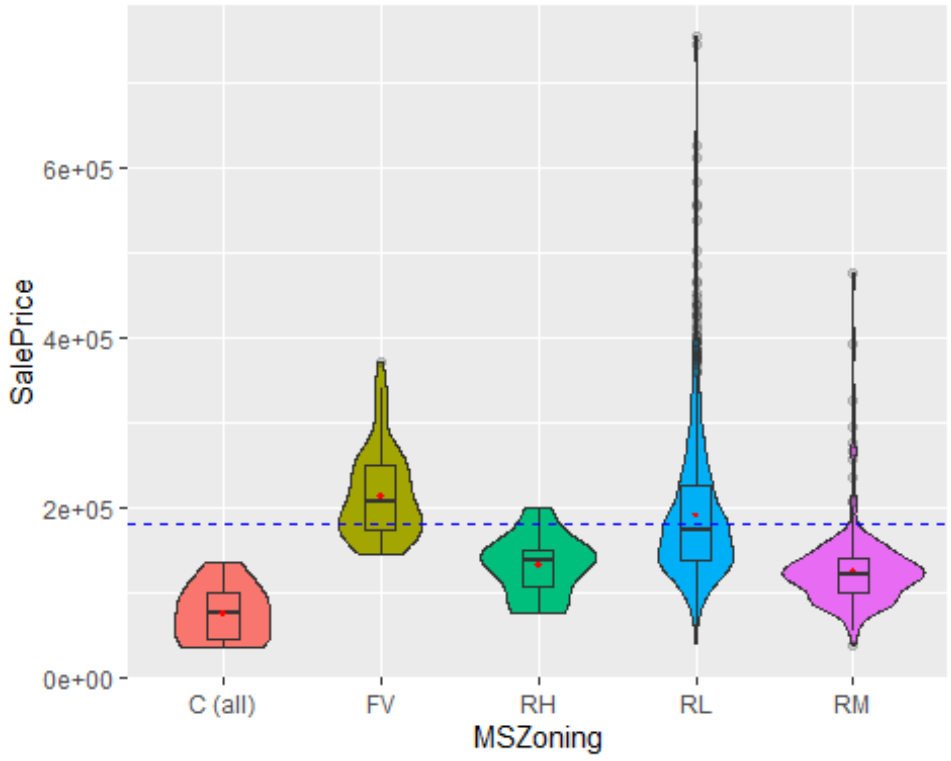
Variabile MSZoning

```
calcola_devianza(case$SalePrice, factor(case$MSZoning))
```

```
## $devianza_totale
## [1] 9.207911e+12
```

```
##
## $devianza_tra_gruppi
## [1] 9.904e+11
##
## $devianza_entro_gruppi
## [1] 8.217511e+12
##
## $eta2
## [1] 0.1075597
```

```
ggplot(case, aes(x = factor(MSZoning), y = SalePrice, fill = factor(MSZoning))) +
geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) +
stat_summary(fun = mean, geom = "point", shape = 18, size = 1, color = "red") +
geom_hline(yintercept = mean(case$SalePrice), linetype = "dashed", color = "blue")
+ labs(x = "MSZoning")
```



In relazione a “SalePrice” abbiamo quasi esclusivamente devianza “Entro”, Eta Quadro è infatti basso. Questo perché le medie dei gruppi sono tutte vicine a quella generale e per via del fatto che i valori estremi di “RL” incidono molto sul calcolo della devianza essendo esso il gruppo più comune. Vediamo che il gruppo “C” è quello con la media dei prezzi più bassa e che nessunodi essi supera la media generale.

Variabile LotFrontage

```
cor(case$LotFrontage, case$SalePrice, use="complete.obs")
```

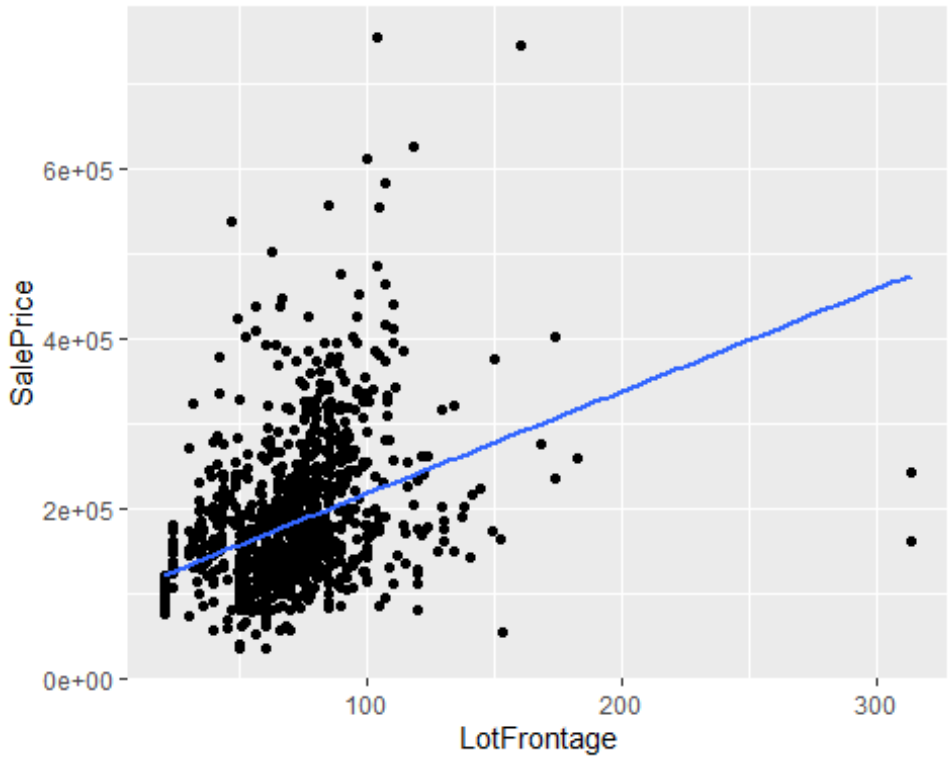
```
## [1] 0.3517991
```

```r
model <- lm(SalePrice ~ LotFrontage, data = case)
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotFrontage, data = case)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -314258  -48878  -19402   33290  533217
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 96149.04    6881.97   13.97   <2e-16 ***
## LotFrontage  1208.02      92.83   13.01   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 78090 on 1199 degrees of freedom
##   (259 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.1238, Adjusted R-squared:  0.123
## F-statistic: 169.4 on 1 and 1199 DF,  p-value: < 2.2e-16
```

```r
ggplot(case, aes(x = LotFrontage, y = SalePrice)) + geom_point(na.rm = T) +
geom_smooth(method = "lm", se = FALSE, na.rm = T)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



```r
#Senza i valori estremi
Frontage <- case[case$LotFrontage < 200,]
cor(Frontage$LotFrontage, Frontage$SalePrice, use="complete.obs")
```

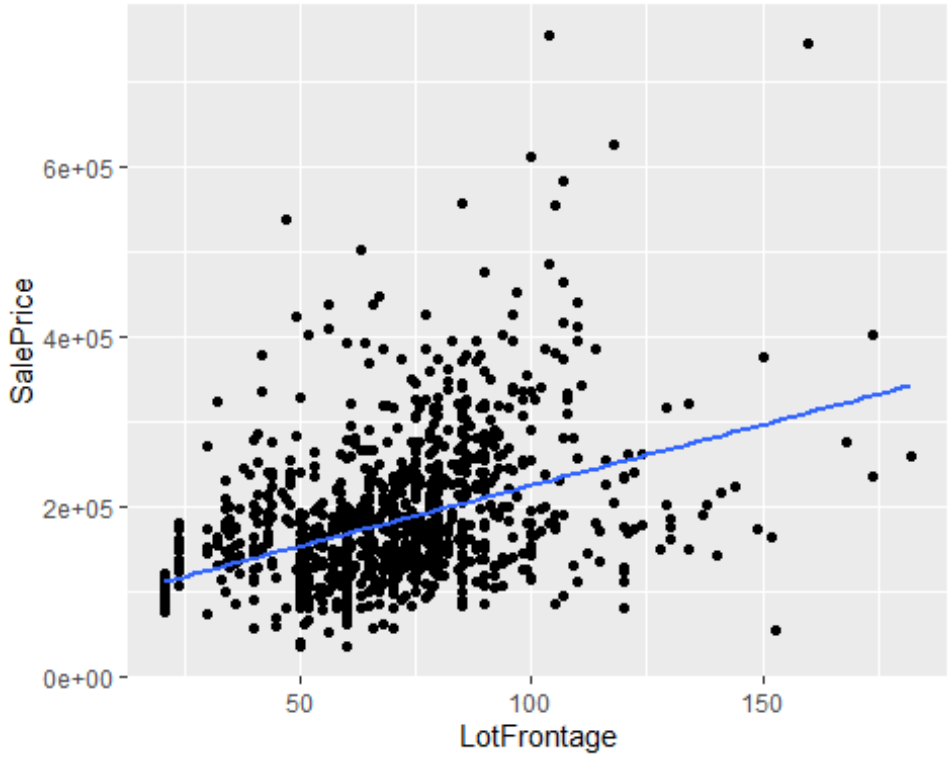
```
## [1] 0.3811296
```

```r
model <- lm(SalePrice ~ LotFrontage, data = Frontage)
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotFrontage, data = Frontage)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -247739  -49060  -18047   33626  525010
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  80891.0     7346.7   11.01   <2e-16 ***
## LotFrontage   1433.6      100.5   14.26   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 77170 on 1197 degrees of freedom
##   (259 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.1453, Adjusted R-squared:  0.1445
## F-statistic: 203.4 on 1 and 1197 DF,  p-value: < 2.2e-16
```

```r
ggplot(Frontage, aes(x = LotFrontage, y = SalePrice)) + geom_point(na.rm = T) +
geom_smooth(method = "lm", se = FALSE, na.rm = T)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Con target la variabile “SalePrice” vediamo che il coefficiente di correlazione è 0.3517991, vediamo quindi una correlazione bassa, ma comunque presente, tra le due variabili. R Quadro è invece molto basso e infatti vediamo una dispersione molto ampia dei punti. Anche rimuovendo i valori più estremi la correlazione cambia poco.

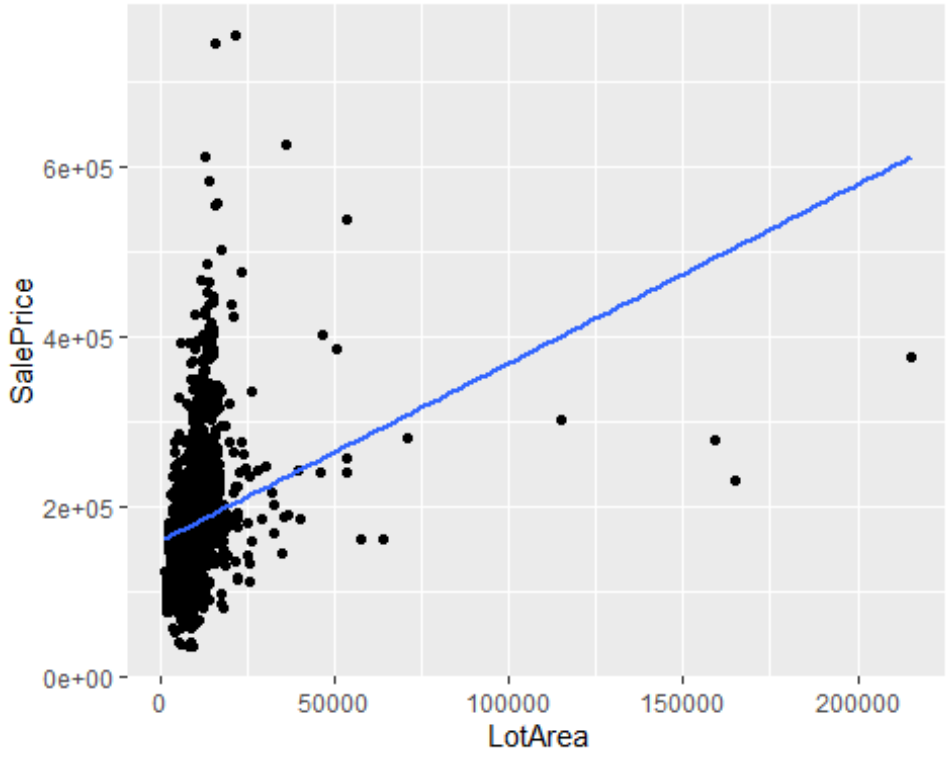
Variabile LotArea

```
cor(case$LotArea, case$SalePrice, use="complete.obs")

## [1] 0.2638434

model <- lm(SalePrice ~ LotArea, data = case)
summary(model)

##
## Call:
## lm(formula = SalePrice ~ LotArea, data = case)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -275668  -48169  -17725   31248  553356
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.588e+05  2.915e+03   54.49   <2e-16 ***
## LotArea     2.100e+00  2.011e-01   10.45   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 76650 on 1458 degrees of freedom
## Multiple R-squared:  0.06961,    Adjusted R-squared:  0.06898
## F-statistic: 109.1 on 1 and 1458 DF,  p-value: < 2.2e-16
```

```
ggplot(case, aes(x = LotArea, y = SalePrice)) + geom_point() + geom_smooth(method
= "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
#LotArea Senza Valori Estremi:
Area <- case[case$LotArea < 100000,]
cor(Area$LotArea, Area$SalePrice, use="complete.obs")
```

```
## [1] 0.3544944
```

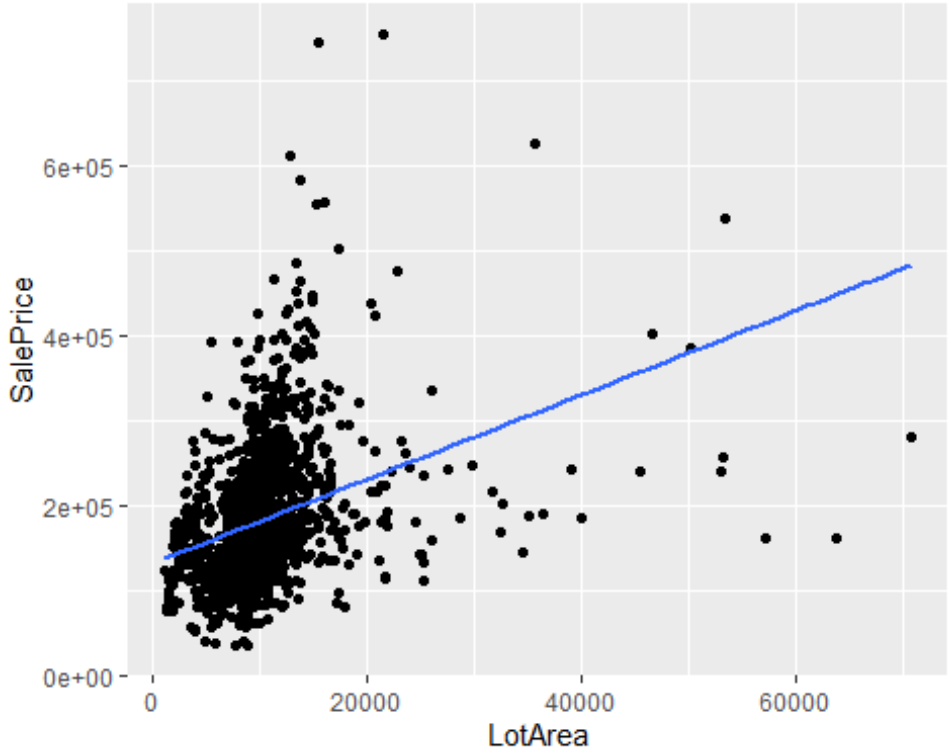
```
model <- lm(SalePrice ~ LotArea, data = Area)
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ LotArea, data = Area)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -288209  -46423  -16475   31845  536900
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.304e+05  3.981e+03   32.75   <2e-16 ***
```

```
## LotArea     4.975e+00  3.441e-01   14.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 74150 on 1454 degrees of freedom
## Multiple R-squared:  0.1257, Adjusted R-squared:  0.1251
## F-statistic:   209 on 1 and 1454 DF,  p-value: < 2.2e-16
```

```
ggplot(Area, aes(x = LotArea, y = SalePrice)) + geom_point() + geom_smooth(method
= "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



La correlazione lineare con "SalePrice" è positiva ma bassa ed R Quadro pure ha un valore molto vicino allo zero. Rimuovendo i valori di "LotArea" maggiori di 100000 vediamo che la correlazione sale di quasi 0.10 . C'è quindi una correlazione lineare molto più marcata una volta rimossi i 4 dati più estremi. Questa è un osservazione importante perché, anche se il numero di queste osservazioni non è elevato, ci si aspetterebbe che abitazioni con Aree sopra i 100000 metri quadrati fossero estremamente più costose rispetto alle altre.

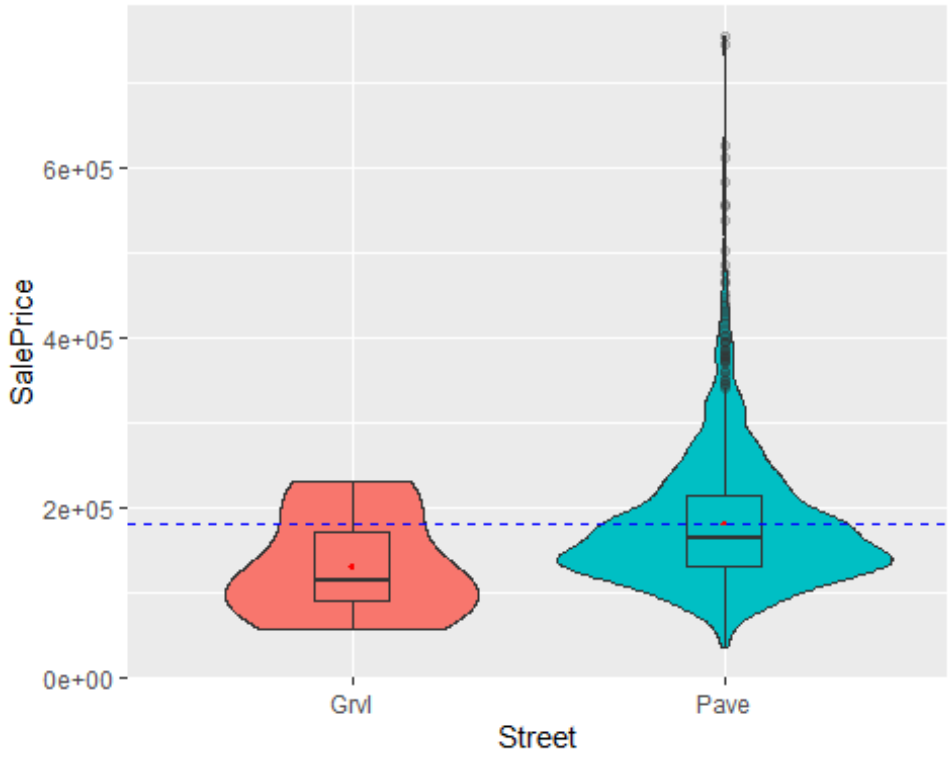
Variabile Street

```
calcola_devianza(case$SalePrice, factor(case$Street))
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 15505341615
```

```
##
## $devianza_entro_gruppi
## [1] 9.192406e+12
##
## $eta2
## [1] 0.001683915
```

```
ggplot(case, aes(x = factor(Street), y = SalePrice, fill = factor(Street))) +
geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) +
stat_summary(fun = mean, geom = "point", shape = 18, size = 1, color = "red") +
geom_hline(yintercept = mean(case$SalePrice), linetype = "dashed", color = "blue")
+ labs(x = "Street")
```



Visto che il tipo "Grvl" corrisponde a sole 6 osservazioni, questo pesa molto poco sulla devianza che infatti è quasi solamente composta di devianza "Entro". L'indice Eta Quadro è infatti uguale a 0.001683915 (la devianza "tra" è inferiore allo 0.2% di quella totale).

Variabile Alley

```
Alley_F <- factor(replace(case$Alley, is.na(case$Alley), "Non Presente"))
```

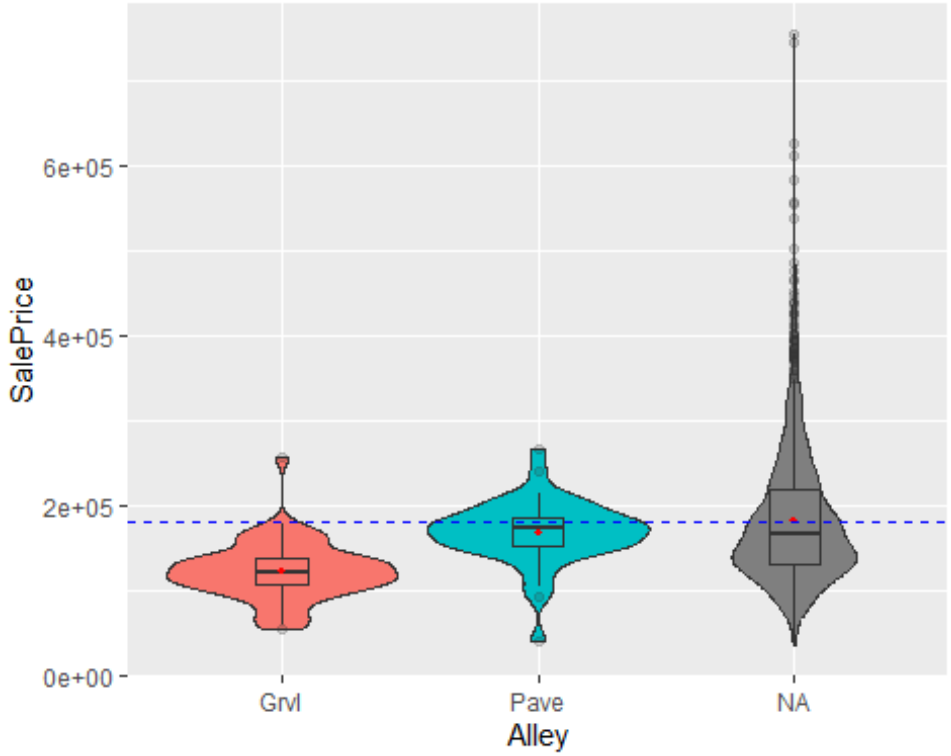
```
## Warning in `[<-.factor`(`*tmp*`, list, value = "Non Presente"): livello factore
## non valido, generato NA
```

```
calcola_devianza(case$SalePrice, Alley_F)
```

```
## $devianza_totale
## [1] 165383231761
##
## $devianza_tra_gruppi
```

```
## [1] 47216371196
##
## $devianza_entro_gruppi
## [1] 118166860566
##
## $eta2
## [1] 0.2854967
```

```
ggplot(case, aes(x = Alley_F, y = SalePrice, fill = Alley_F)) + geom_violin() +
geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) + stat_summary(fun =
mean, geom = "point", shape = 18, size = 1, color = "red") +
geom_hline(yintercept = mean(case$SalePrice), linetype = "dashed", color = "blue")
+ labs(x = "Alley")
```



Come per la variabile precedente, abbiamo un gruppo drasticamente più numeroso rispetto gli altri. La devianza è quindi principalmente composta dalla devianza "Entro" ed Eta Quadro è infatti 0.02040754. Si nota che tutti i valori più elevati di "SalePrice" si trovano all'interno del gruppo "Non Presente".

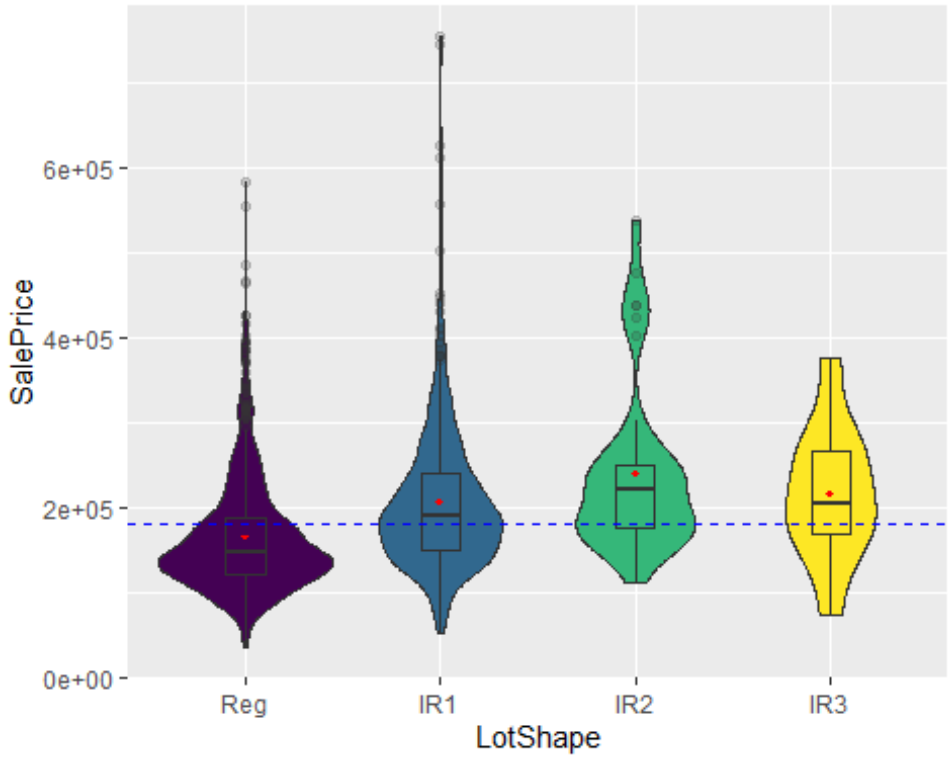
Variabile LotShape

```
calcola_devianza(case$SalePrice, factor(case$LotShape))
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 703260743057
##
```

```
## $devianza_entro_gruppi
## [1] 8.504651e+12
##
## $eta2
## [1] 0.07637571
```

```
ggplot(case, aes(x = ordered(factor(LotShape),levels =
c("Reg","IR1","IR2","IR3")), y = SalePrice, fill = factor(LotShape))) +
geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) +
stat_summary(fun = mean, geom = "point", shape = 18, size = 1, color = "red") +
geom_hline(yintercept = mean(case$SalePrice), linetype = "dashed", color = "blue")
+ labs(x = "LotShape")
```



Anche in questo caso la devianza "Tra" è decisamente inferiore rispetto alla "Entro". Eta quadro e infatti 0.07637571. In questo caso il valore basso della deviazione "Tra" è dovuto al fatto che i 4 gruppi possiedono tutti medie molto vicine a quella generale.

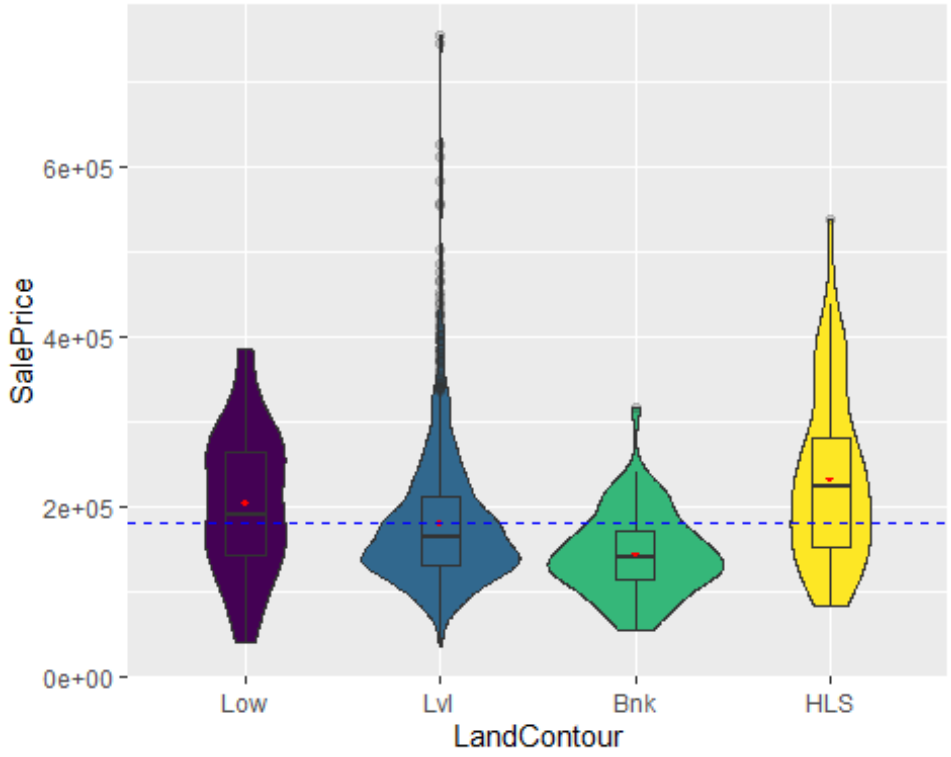
Variabile LandContour

```
calcola_devianza(case$SalePrice, factor(case$LandContour))
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 237509651885
##
## $devianza_entro_gruppi
## [1] 8.970402e+12
```

```
##
## $eta2
## [1] 0.02579409
```

```
ggplot(case, aes(x = ordered(factor(LandContour),levels =
c("Low","Lvl","Bnk","HLS")), y = SalePrice, fill = factor(LandContour))) +
geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) +
stat_summary(fun = mean, geom = "point", shape = 18, size = 1, color = "red") +
geom_hline(yintercept = mean(case$SalePrice), linetype = "dashed", color = "blue")
+ labs(x = "LandContour")
```



La devianza "Tra" è ancora molto bassa con Eta quadro uguale 0.02579409. Anche se il gruppo "Lvl" non comprendesse 90% delle osservazioni, tutte le medie dei gruppi sono molto vicine a quella generale. I gruppi hanno range di valori simili fatta eccezione di "Lvl" che possiede anche quelli più estremi.

Variabile Utilities

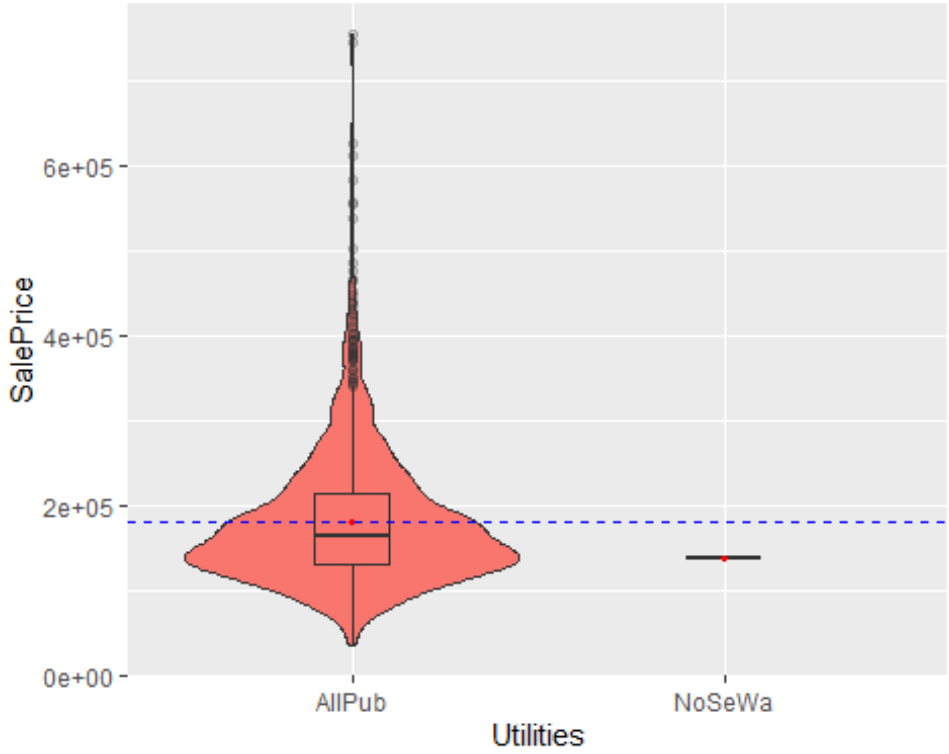
```
calcola_devianza(case$SalePrice, factor(case$Utilities))
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 1886692508
##
## $devianza_entro_gruppi
## [1] 9.206025e+12
##
```

```
## $eta2
## [1] 0.0002048991
```

```
ggplot(case, aes(x = factor(Utilities), y = SalePrice, fill = factor(Utilities)))
+ geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) +
stat_summary(fun = mean, geom = "point", shape = 18, size = 1, color = "red") +
geom_hline(yintercept = mean(case$SalePrice), linetype = "dashed", color = "blue")
+ labs(x = "Utilities")
```

```
## Warning: Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
```



Il valore della devianza "Tra" è estremamente minore di quella "Entro" ed infatti Eta Quadro è uguale a solo 0.0002048991. L'unico valore del gruppo NoSeWa è inoltre molto vicino alla media generale.

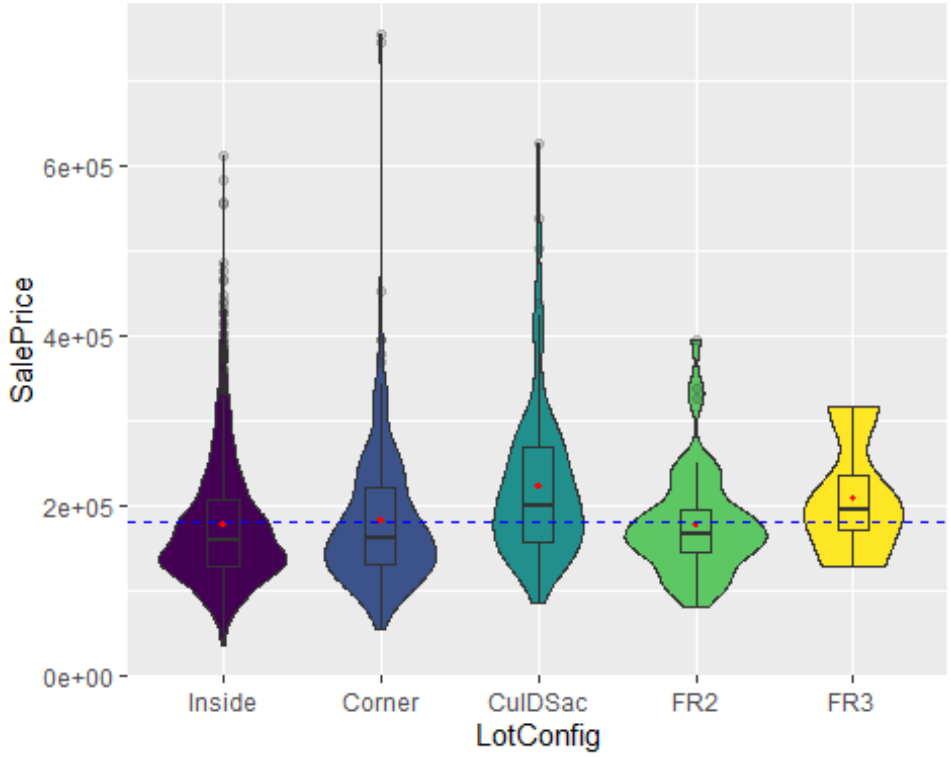
Variabile LotConfig

```
calcola_devianza(case$SalePrice, factor(case$LotConfig))
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 193544444976
##
## $devianza_entro_gruppi
## [1] 9.014367e+12
##
```

```
## $eta2
## [1] 0.02101936
```

```
ggplot(case, aes(x = ordered(factor(LotConfig), levels =
c("Inside","Corner","CulDSac","FR2","FR3")), y = SalePrice, fill =
factor(LotConfig))) + geom_violin() + geom_boxplot(width=0.2, alpha=1/5) +
guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape = 18, size =
1, color = "red") +  geom_hline(yintercept = mean(case$SalePrice), linetype =
"dashed", color = "blue") + labs(x = "LotConfig")
```



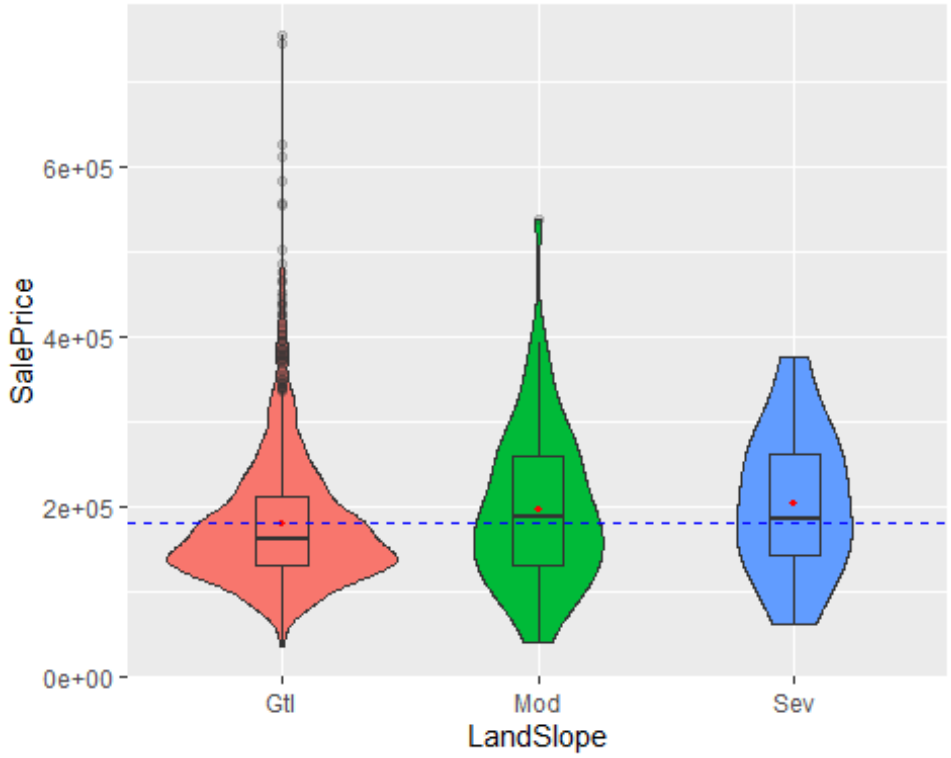
Un alto caso in cui il valore di Eta Quadro è estremamente basso. I gruppi hanno infatti medie molto vicine a quella generale e anche range simili.

Variabile LandSlope

```
calcola_devianza(case$SalePrice, factor(case$LandSlope))
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 24692170428
##
## $devianza_entro_gruppi
## [1] 9.183219e+12
##
## $eta2
## [1] 0.002681626
```

```r
ggplot(case, aes(x = factor(LandSlope), y = SalePrice, fill = factor(LandSlope)))
+ geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) +
stat_summary(fun = mean, geom = "point", shape = 18, size = 1, color = "red") +
geom_hline(yintercept = mean(case$SalePrice), linetype = "dashed", color = "blue")
+ labs(x = "LandSlope")
```



Come in molti casi precedenti la devianza "Entro" è estremamente alta essendo il gruppo "Gtl" disproporzionalmente numeroso. Inoltre, anche le medie di "Mod" e "Sev" sono vicine alla media generale. I valori più alti della variabile prezzo sono presenti nel gruppo "Gtl"

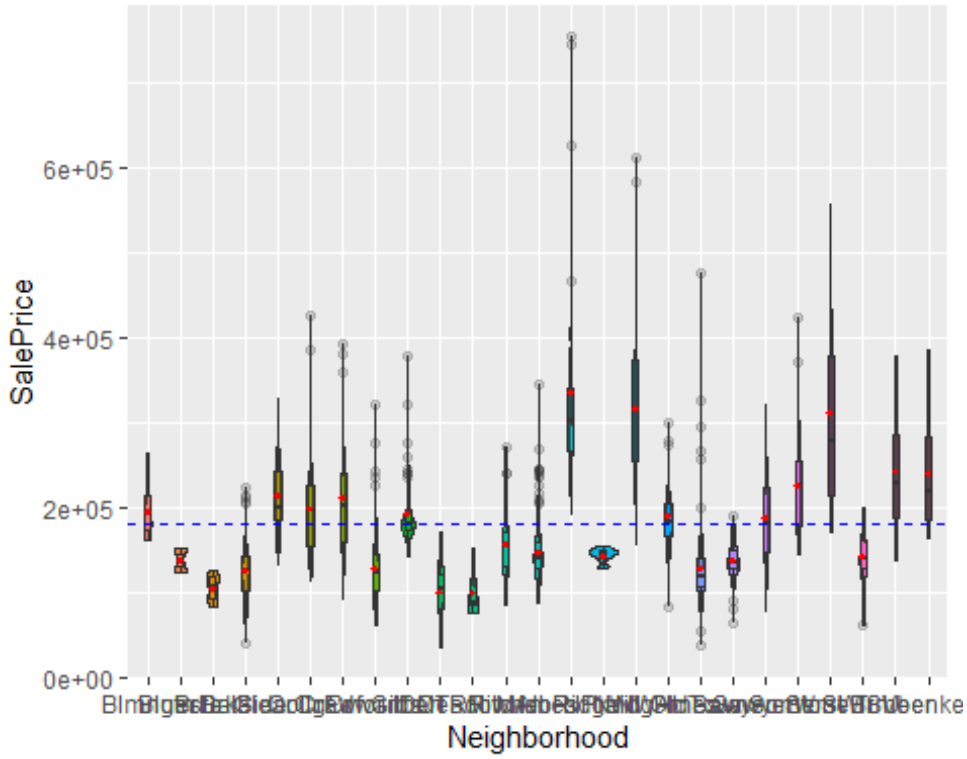
Variabile Neighborhood

```r
calcola_devianza(case$SalePrice, factor(case$Neighborhood))
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 5.023606e+12
##
## $devianza_entro_gruppi
## [1] 4.184305e+12
##
## $eta2
## [1] 0.545575
```

```r
ggplot(case, aes(x = factor(Neighborhood), y = SalePrice, fill =
factor(Neighborhood))) + geom_violin() + geom_boxplot(width=0.2, alpha=1/5) +
guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape = 18, size =
```

```
1, color = "red") +  geom_hline(yintercept = mean(case$SalePrice), linetype =
"dashed", color = "blue") + labs(x = "Neighborhood")
```

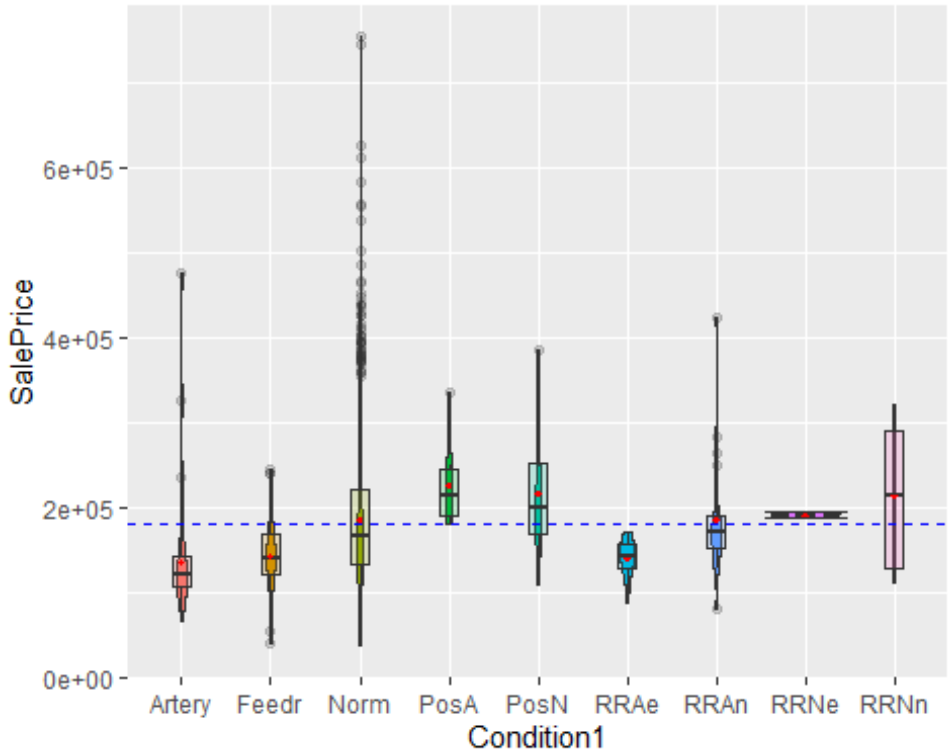


In questo caso Eta Quadro è uguale a 0.545575 e infatti la devianza di tipo "Tra"è quella prevalente. Si può vedere, infatti, che ci sono molti gruppi la cui media è decisamente distante da quella generale. Con questa variabile in particolare, il sapere a che gruppo appartiene una certa abitazione può aiutarci a prevedere quale sarà il suo prezzo. Proprietà situate in NoRidge, NridgHt o StoneBr sono, come si può vedere dal grafico, molto più care rispetto la media generale mentre quelle situate in Blueste, MeadowV, BrDale, NPkVill e IDOTRR sono sotto di essa. Bisogna però fare attenzione perché queste 5 hanno un sample size molto basso.

Variabile Condition1

```
calcola_devianza(case$SalePrice, factor(case$Condition1))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 3.0046e+11
##
## $devianza_entro_gruppi
## [1] 8.907451e+12
##
## $eta2
## [1] 0.03263064

ggplot(case, aes(x = factor(Condition1), y = SalePrice, fill =
factor(Condition1))) + geom_violin() + geom_boxplot(width=0.2, alpha=1/5) +
```

```
guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape = 18, size =
1, color = "red") +  geom_hline(yintercept = mean(case$SalePrice), linetype =
"dashed", color = "blue") + labs(x = "Condition1")
```

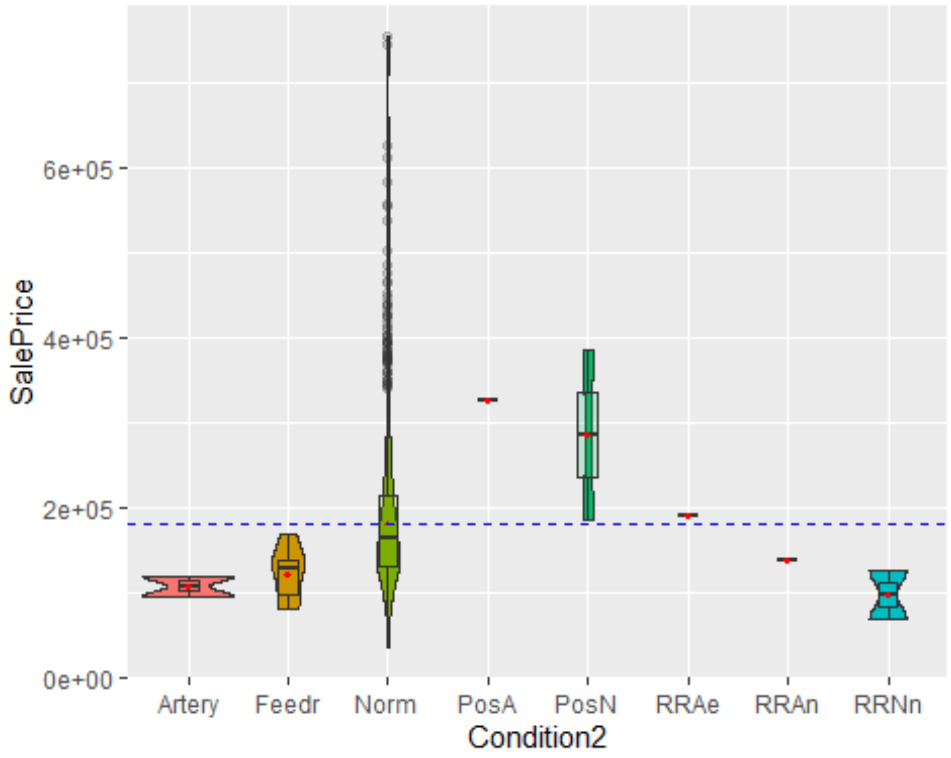


Eta Quadro è molto basso in questo caso, infatti il gruppo "Norm" influisce sproporzionalmente di più sulla devianza rispetto gli altri. Tutti i gruppi hanno comunque una media molto vicina a quella generale, il che contribuisce anche questo a tenere bassa da devianza "Tra" e quindi anche Eta Quadro.

Variabile Condition2

```
calcola_devianza(case$SalePrice, factor(case$Condition2))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 91150594164
##
## $devianza_entro_gruppi
## [1] 9.116761e+12
##
## $eta2
## [1] 0.009899161

ggplot(case, aes(x = factor(Condition2), y = SalePrice, fill =
factor(Condition2))) + geom_violin() + geom_boxplot(width=0.2, alpha=1/5) +
guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape = 18, size =
```

```
1, color = "red") +  geom_hline(yintercept = mean(case$SalePrice), linetype =
"dashed", color = "blue") + labs(x = "Condition2")

## Warning: Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
## Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
## Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
```

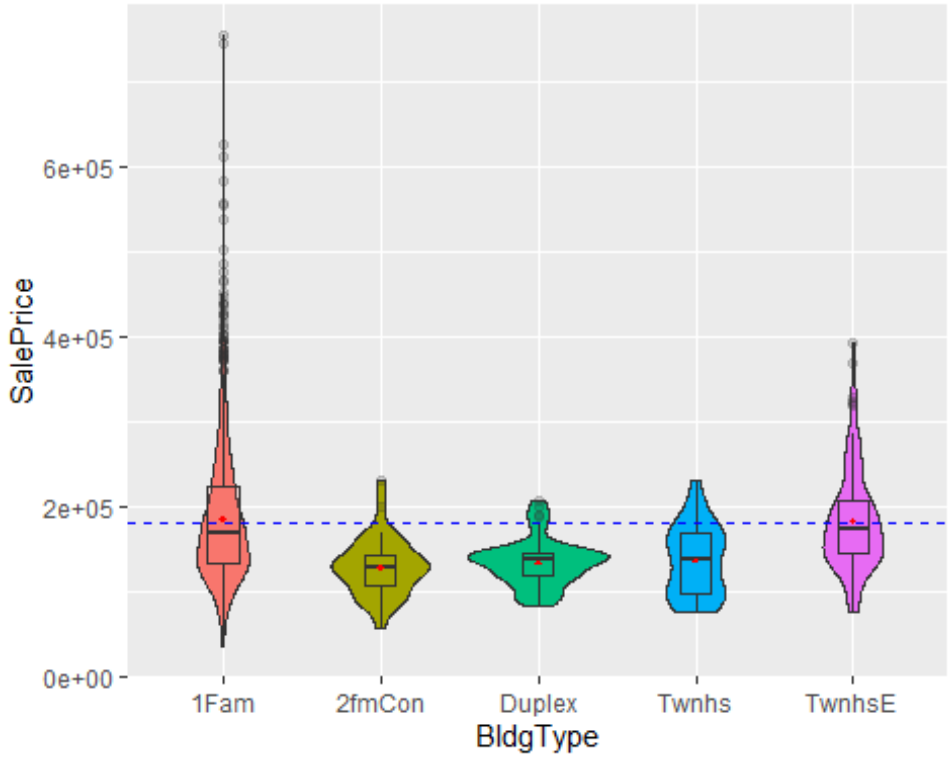


Eta Quadro è anche qui molto basso ed essendo tutti i gruppi diversi da "Norm" formati da al 6 proprietà, non abbiamo una sample size abbastanza grande per fare previsioni sul comportamento della variabile "SalePrice" al variare di "Condition2".

Variabile BldgType

```
calcola_devianza(case$SalePrice, factor(case$BldgType))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 317986257619
##
## $devianza_entro_gruppi
## [1] 8.889925e+12
##
## $eta2
## [1] 0.03453403
```

```
ggplot(case, aes(x = factor(BldgType), y = SalePrice, fill = factor(BldgType))) +
geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) +
stat_summary(fun = mean, geom = "point", shape = 18, size = 1, color = "red") +
geom_hline(yintercept = mean(case$SalePrice), linetype = "dashed", color = "blue")
+ labs(x = "BldgType")
```

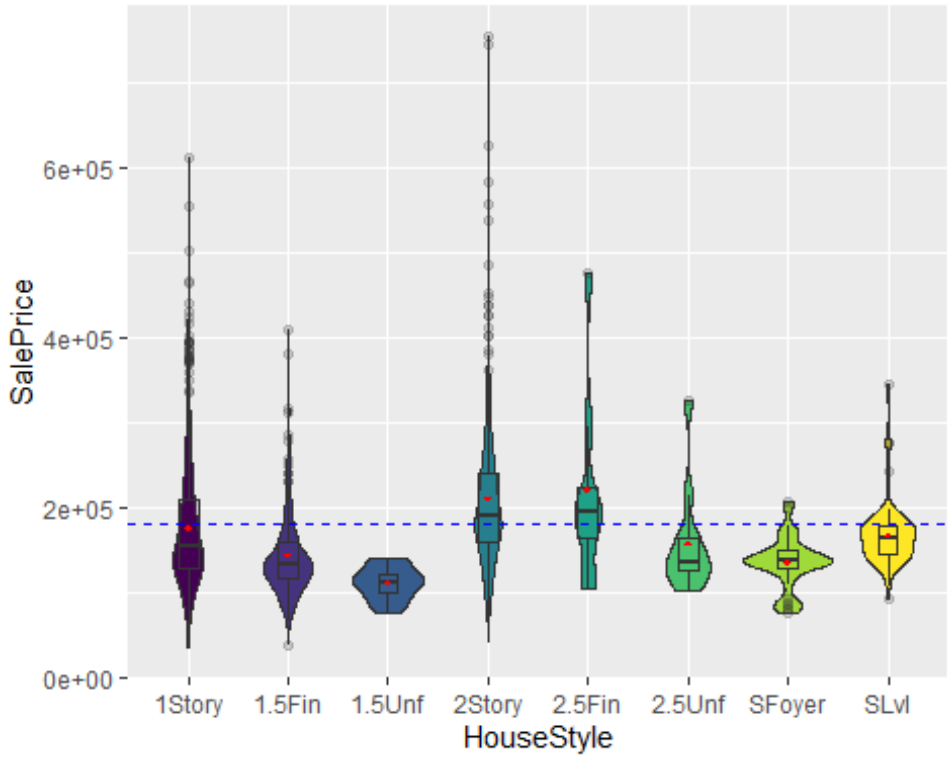


Eta Quadro è 0.03453403, infatti la devianza "Entro" è quella prevalente. Vediamo che le medie di tutti i gruppi sono vicine a quella generale e che l'unico gruppo con valori che si discostano tanto dalla media è "1Fam".

Variabile HouseStyle

```
calcola_devianza(case$SalePrice, factor(case$HouseStyle))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 7.94759e+11
##
## $devianza_entro_gruppi
## [1] 8.413152e+12
##
## $eta2
## [1] 0.08631263

ggplot(case, aes(x = ordered(factor(HouseStyle), levels =
c("1Story","1.5Fin","1.5Unf","2Story","2.5Fin","2.5Unf","SFoyer","SLvl")), y =
SalePrice, fill = factor(HouseStyle))) + geom_violin() + geom_boxplot(width=0.2,
```

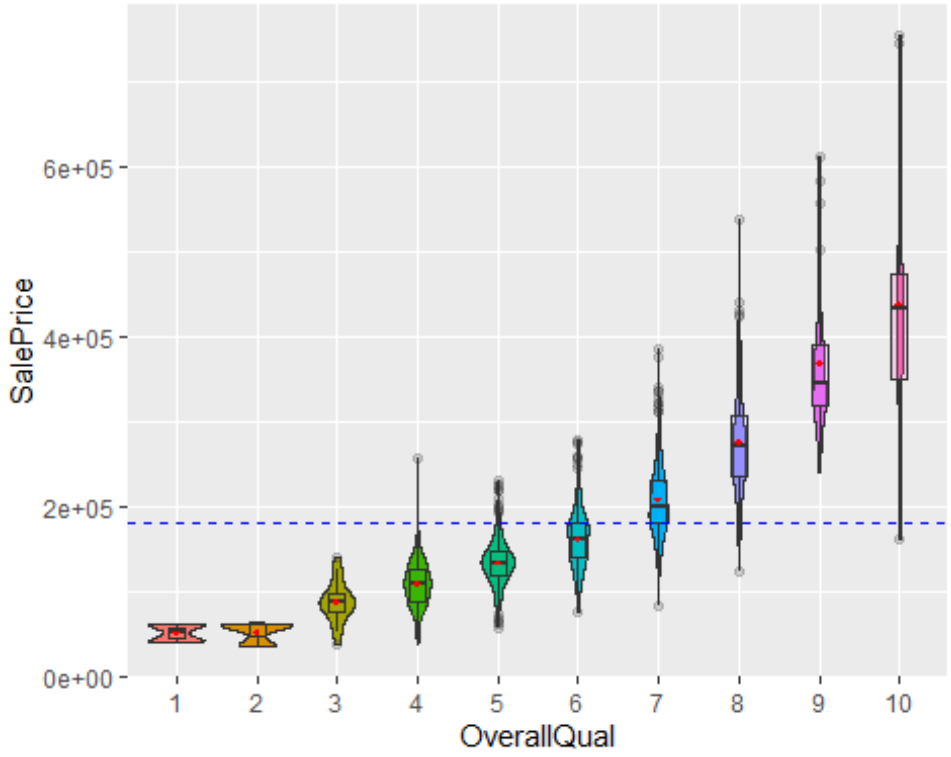
```
alpha=1/5) + guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape
= 18, size = 1, color = "red") +  geom_hline(yintercept = mean(case$SalePrice),
linetype = "dashed", color = "blue") + labs(x = "HouseStyle")
```



Eta Quadro, uguale a 0.08631263, è molto basso. Tutte le medie sono infatti molto vicine a quella generale anche se la presenza di ulteriori piani ad un edificio farebbe pensare ad una correlazione stretta con l'aumento di "SalePrice"

Variabile OverallQual

```
calcola_devianza(case$SalePrice, factor(case$OverallQual))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 6.299881e+12
##
## $devianza_entro_gruppi
## [1] 2.908031e+12
##
## $eta2
## [1] 0.6841813

ggplot(case, aes(x = factor(OverallQual), y = SalePrice, fill =
factor(OverallQual))) + geom_violin() + geom_boxplot(width=0.2, alpha=1/5) +
guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape = 18, size =
1, color = "red") +  geom_hline(yintercept = mean(case$SalePrice), linetype =
"dashed", color = "blue") + labs(x = "OverallQual")
```



Per l'analisi bivariata la consideriamo come una Variabile Qualitativa. Abbiamo Eta Quadro uguale a 0.6841813, il che denota una prevalenza di devianza "Tra". Si vede infatti come all'aumentare del valore di "OverallQual" la variabile "SalePrice" salga. Volendo comunque usare la correlazione lineare con questa variabile, vediamo che essa arriva a 0.7909816, ovvero è presente una forte correlazione tra le due variabili.

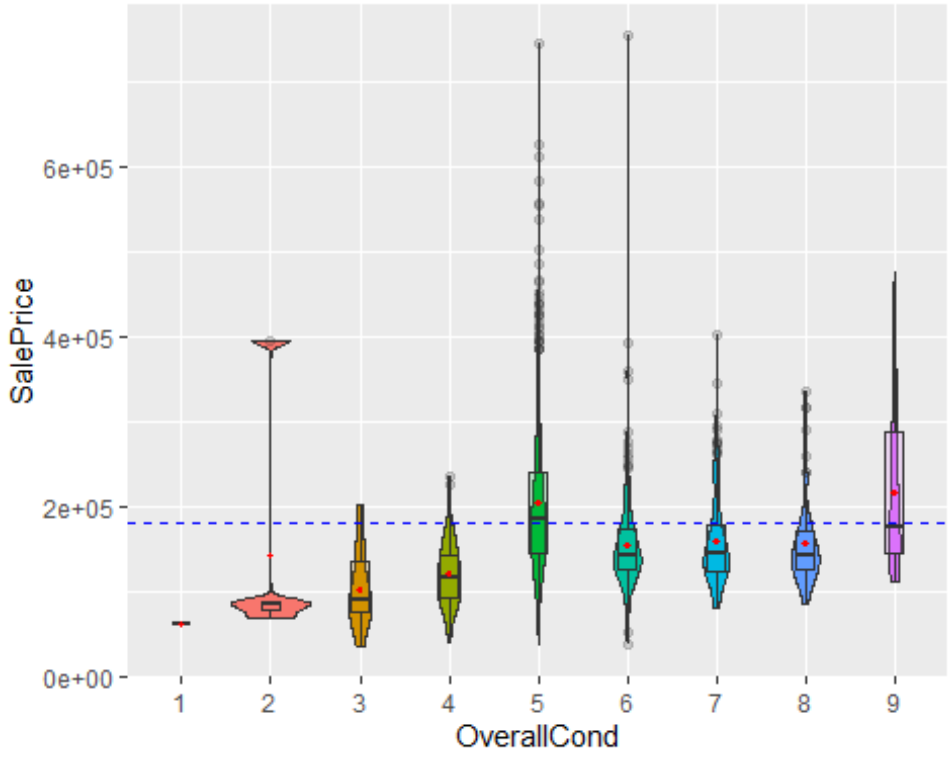
Variabile OverallCond

```
calcola_devianza(case$SalePrice, factor(case$OverallCond))
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 1.154581e+12
##
## $devianza_entro_gruppi
## [1] 8.05333e+12
##
## $eta2
## [1] 0.1253901
```

```
ggplot(case, aes(x = factor(OverallCond), y = SalePrice, fill =
factor(OverallCond))) + geom_violin() + geom_boxplot(width=0.2, alpha=1/5) +
guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape = 18, size =
1, color = "red") +  geom_hline(yintercept = mean(case$SalePrice), linetype =
"dashed", color = "blue") + labs(x = "OverallCond")
```

```
## Warning: Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
```



Dalla Devianza vediamo che quella principale è quella di tipo "Entro", Eta Quadro è infatti basso a 0.1253901. Si vede che le medie dei gruppi sono tutte vicine a quella generale con i gruppi "5" e "6" quelli con i valori più estremi.

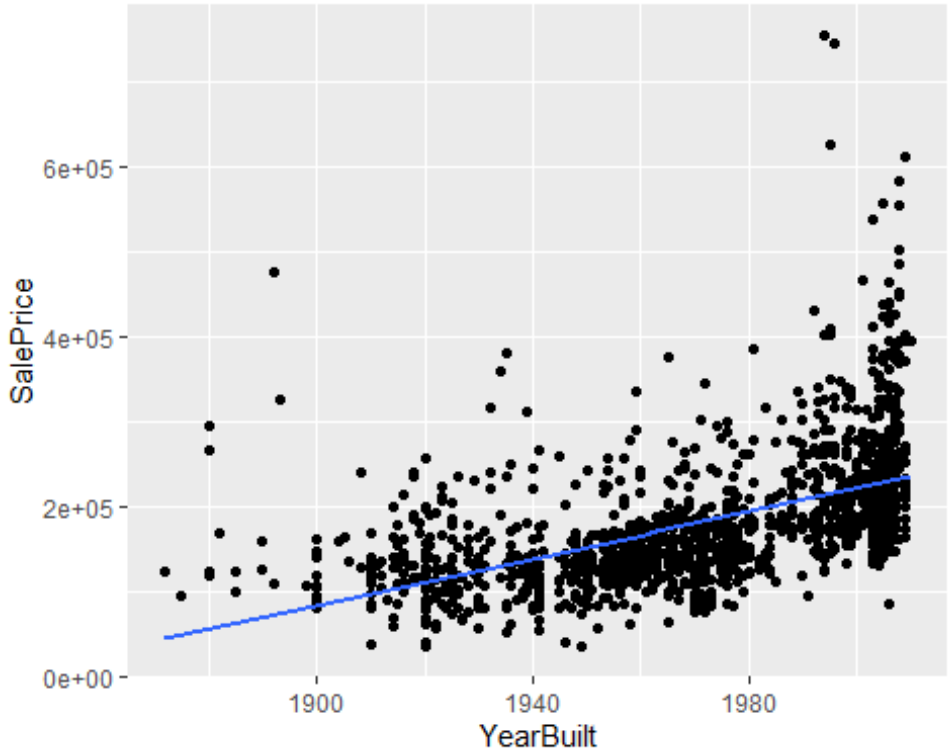
Variabile YearBuilt

```
cor(case$YearBuilt, case$SalePrice, use="complete.obs")

## [1] 0.5228973

model <- lm(SalePrice ~ YearBuilt, data = case)
summary(model)

##
## Call:
## lm(formula = SalePrice ~ YearBuilt, data = case)
##
## 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 ***
## 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
```

```
ggplot(case, aes(x = YearBuilt, y = SalePrice)) + geom_point() +
geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Il coefficiente di correlazione in questo caso è uguale a 0.5228973, il che denota una discreta correlazione tra le due variabili. R Quadro ha un valore di 0.2729 che mi indica una dispersione dei punti ampia intorno alla mia retta di regressione.

Variabile YearRemodAdd

```
calcolo_cov_cor(case$YearRemodAdd)
```

```
##          cov          cor
## 8.317079e+05 5.071010e-01
```

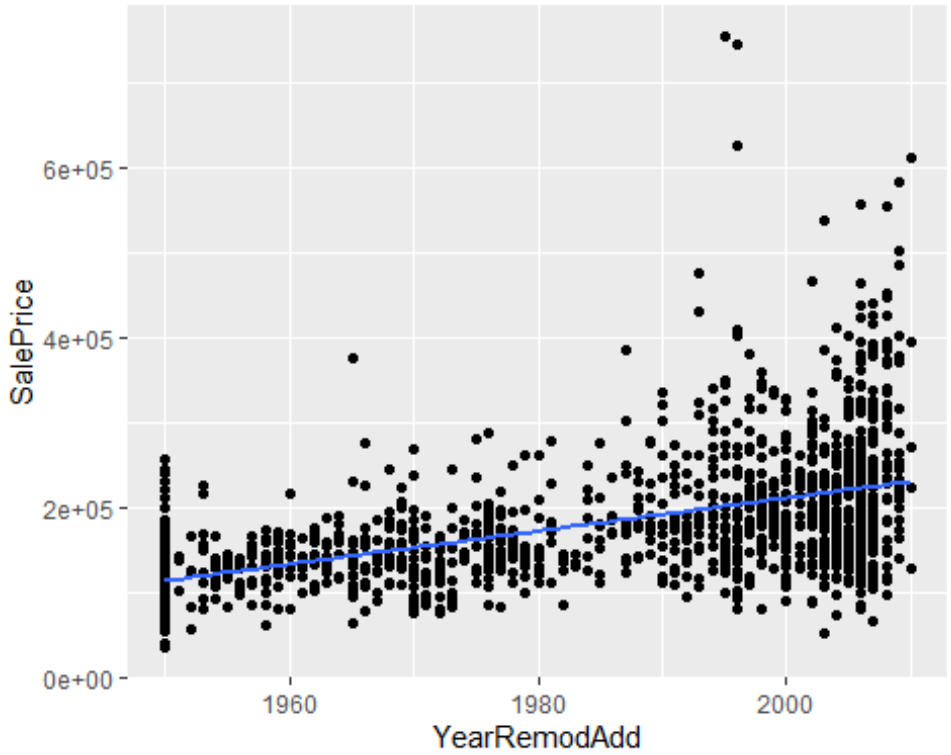
```
model <- lm(SalePrice ~ YearRemodAdd, data = case)
summary(model)
```

```
##
## Call:
## lm(formula = SalePrice ~ YearRemodAdd, data = case)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -164307  -39541   -8159   24603  554304
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.692e+06  1.724e+05  -21.41   <2e-16 ***
## YearRemodAdd  1.951e+03  8.686e+01   22.47   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 68490 on 1458 degrees of freedom
## Multiple R-squared:  0.2572, Adjusted R-squared:  0.2566
## F-statistic: 504.7 on 1 and 1458 DF,  p-value: < 2.2e-16
```

```
ggplot(case, aes(x = YearRemodAdd, y = SalePrice)) + geom_point() +
geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



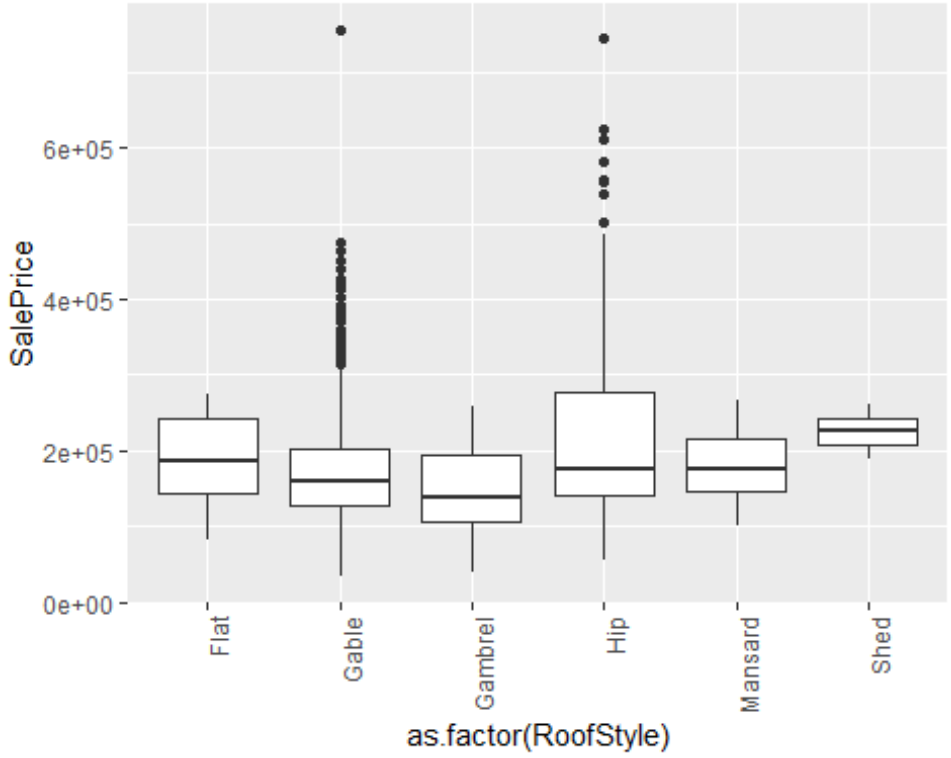
Il coefficiente di correlazione lineare è circa 0.5, il che denota una buona correlazione tra le 2 variabili. Dal modello di regressione lineare osserviamo un valore di R^2 pari a circa 0.26 che mi indica una dispersione ampia dei valori intorno alla retta di regressione.

Variabile RoofStyle

```
case$RoofStyle <- factor(case$RoofStyle)
calcola_devianza(case$SalePrice, case$RoofStyle)
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 531265454526
##
## $devianza_entro_gruppi
## [1] 8.676646e+12
##
## $eta2
## [1] 0.05769663
```

```
ggplot(case, aes(x = as.factor(RoofStyle), y = SalePrice)) + geom_boxplot() +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Si nota che la devianza within ha un valore molto elevato, dovuta soprattuto alle classi Gabel e Hip che hanno dei valori outlier. Le case con i prezzi più alti appartengono alle due classi sopracitate, ma la variabile RoofStyle comunque non sembra influenzare il prezzo delle case in quanto la media dei prezzi di ogni classe è comunque allineata.

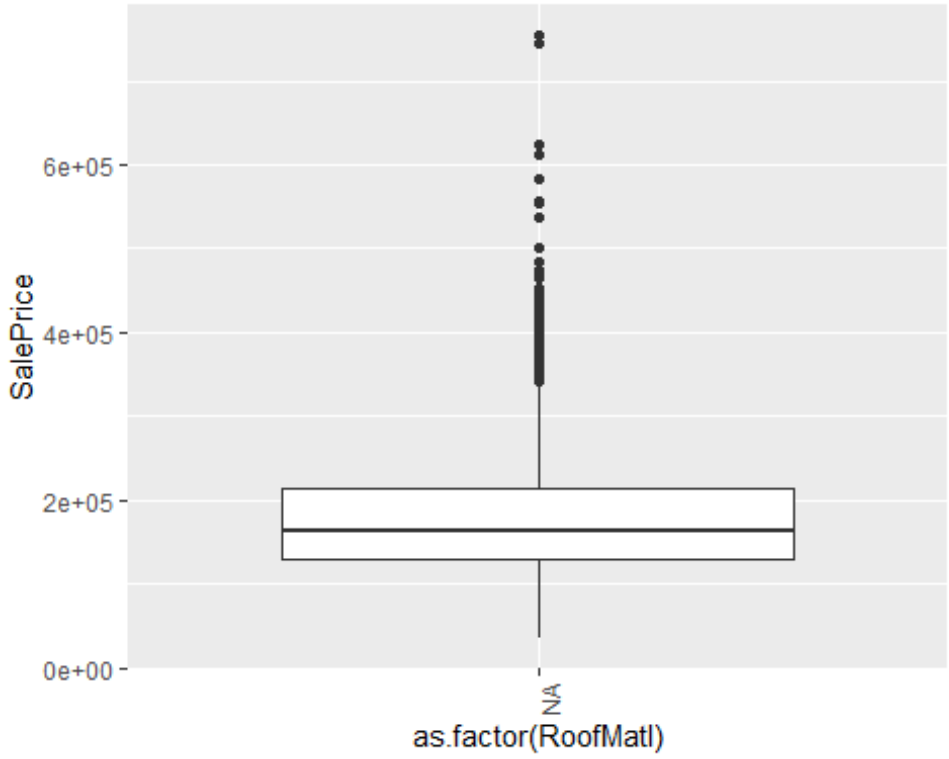
Variabile RoofMatl

```
case$RoofMatl <- factor(case$RoofMatl, levels = c("Roll", "ClyTile", "CompShg",
"Tar&Grv", "Metal", "Membran", "WdShake", "WdShngl"))
calcola_devianza(case$SalePrice, case$RoofMatl)
```

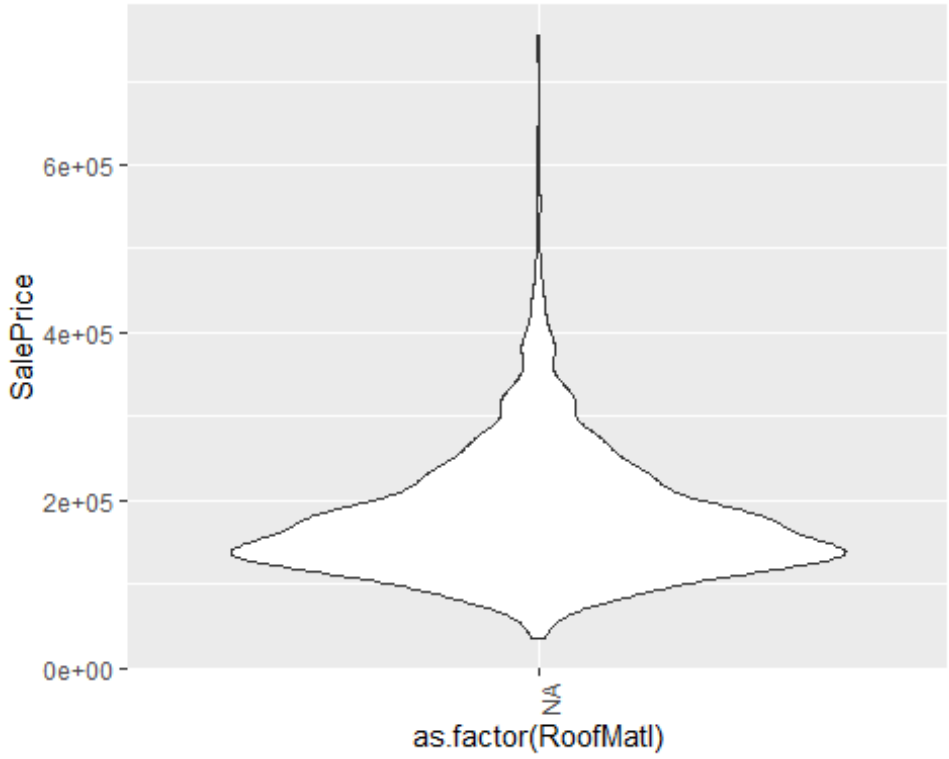
```
## $devianza_totale
## [1] 0
##
```

```
## $devianza_tra_gruppi
## [1] NaN
##
## $devianza_entro_gruppi
## [1] 0
##
## $eta2
## [1] NaN
```

```
ggplot(case, aes(x = as.factor(RoofMatl), y = SalePrice)) + geom_boxplot() +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



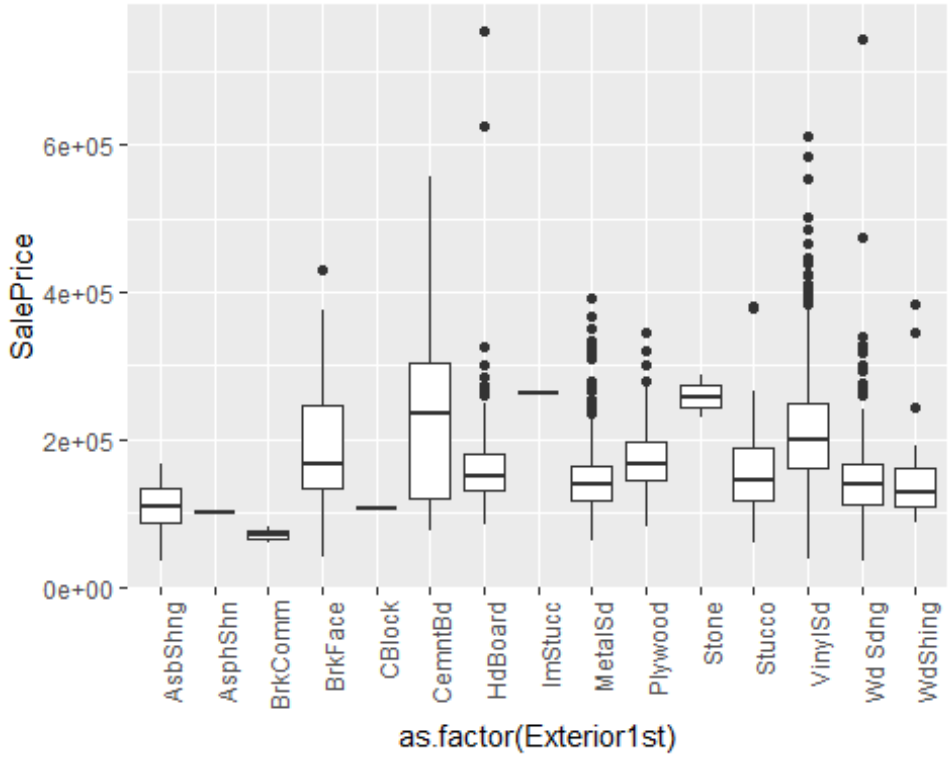

```
ggplot(case, aes(x = as.factor(RoofMatl), y = SalePrice)) + geom_violin() +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Si nota che la devianza within ha un valore molto elevato, dovuta soprattuto alla classa CompShg che ha numerosi valori outlier. L'andamento dei prezzi medi delle case sembrano seguire l'ordine dei factor, tuttavia, come si evince dal grafico a violino, si hanno pochi dati per alcune classe di variabili. In particolare per Roll, ClyTile, Metal e Membran la variabile RoofMatl non sembra influenzare il prezzo delle case in quanto il valore di ETA^2 è comunque basso

Variabile Exterior1st

```
case$Exterior1st <- factor(case$Exterior1st)
calcola_devianza(case$SalePrice, case$Exterior1st)

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 1.406721e+12
##
## $devianza_entro_gruppi
## [1] 7.80119e+12
##
## $eta2
## [1] 0.1527731

ggplot(case, aes(x = as.factor(Exterior1st), y = SalePrice)) + geom_boxplot() +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



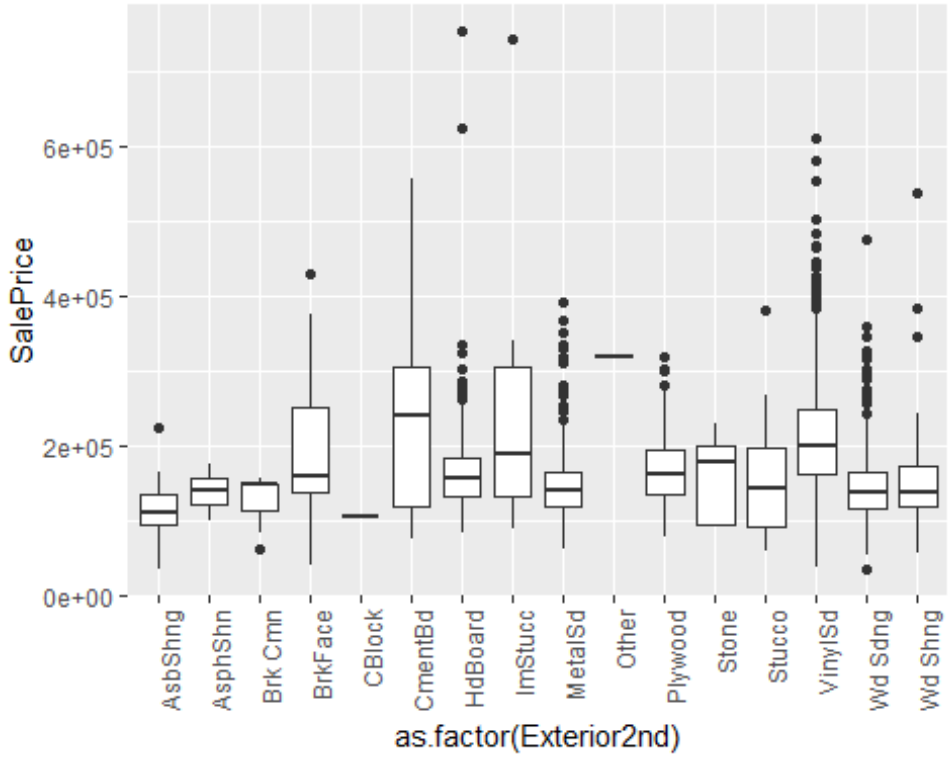
Le varianze within e between hanno valori elevati, e la ETA^2 ha un valore di circa 0.15. Di conseguenza c'è una bassa correlazione tra la variabile Exterior1st e il prezzo.

Variabile Exterior2nd

```
case$Exterior2nd <- factor(case$Exterior2nd)
calcola_devianza(case$SalePrice, case$Exterior2nd)
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 1.416452e+12
##
## $devianza_entro_gruppi
## [1] 7.79146e+12
##
## $eta2
## [1] 0.1538299
```

```
ggplot(case, aes(x = as.factor(Exterior2nd), y = SalePrice)) + geom_boxplot() +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



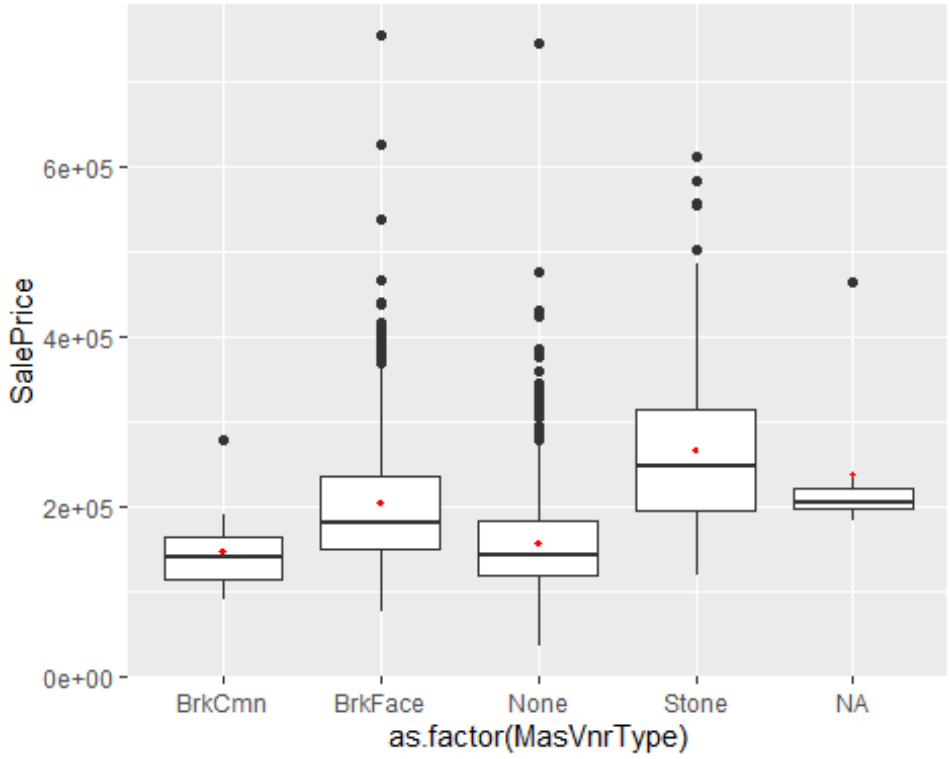
Le varianze within e between hanno valori elevati, e la ETA^2 ha un valore di circa 0.15 . Di conseguenza, c'è una bassa correlazione tra la variabile Exterior1st e il prezzo.

Variabile MasVnrType

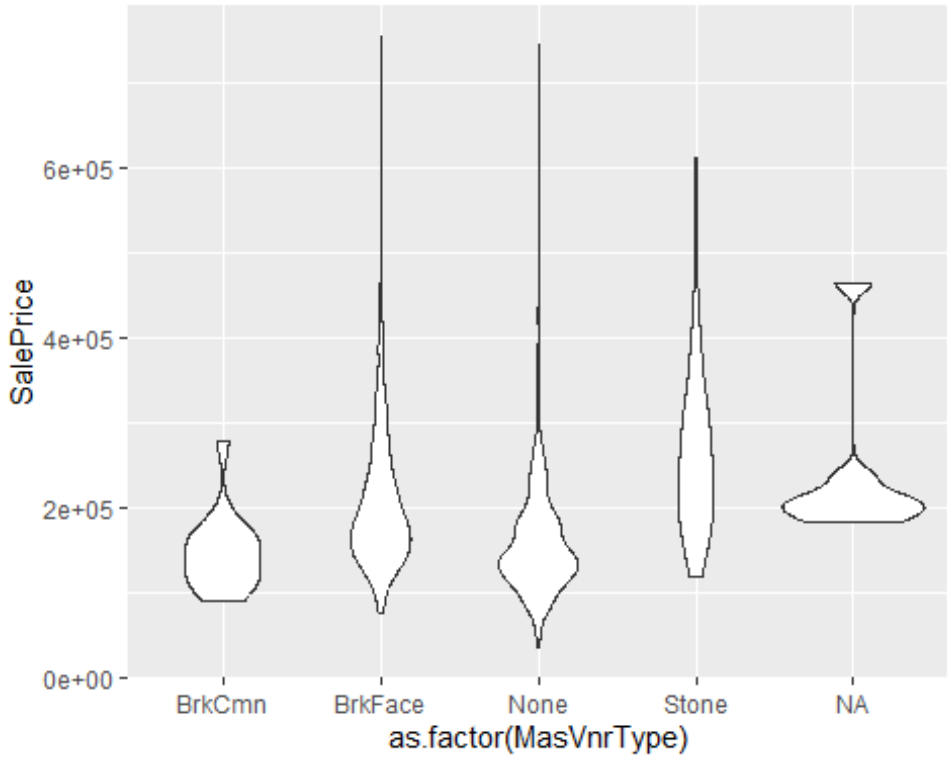
```
case$MasVnrType <- factor(case$MasVnrType)
calcola_devianza(case$SalePrice, case$MasVnrType)
```

```
## $devianza_totale
## [1] 9.121272e+12
##
## $devianza_tra_gruppi
## [1] 1.713827e+12
##
## $devianza_entro_gruppi
## [1] 7.407445e+12
##
## $eta2
## [1] 0.1878934
```

```
ggplot(case, aes(x = as.factor(MasVnrType), y = SalePrice)) + geom_boxplot() +
stat_summary(fun = mean, geom = "point", shape = 18, size = 1, color = "red")
```



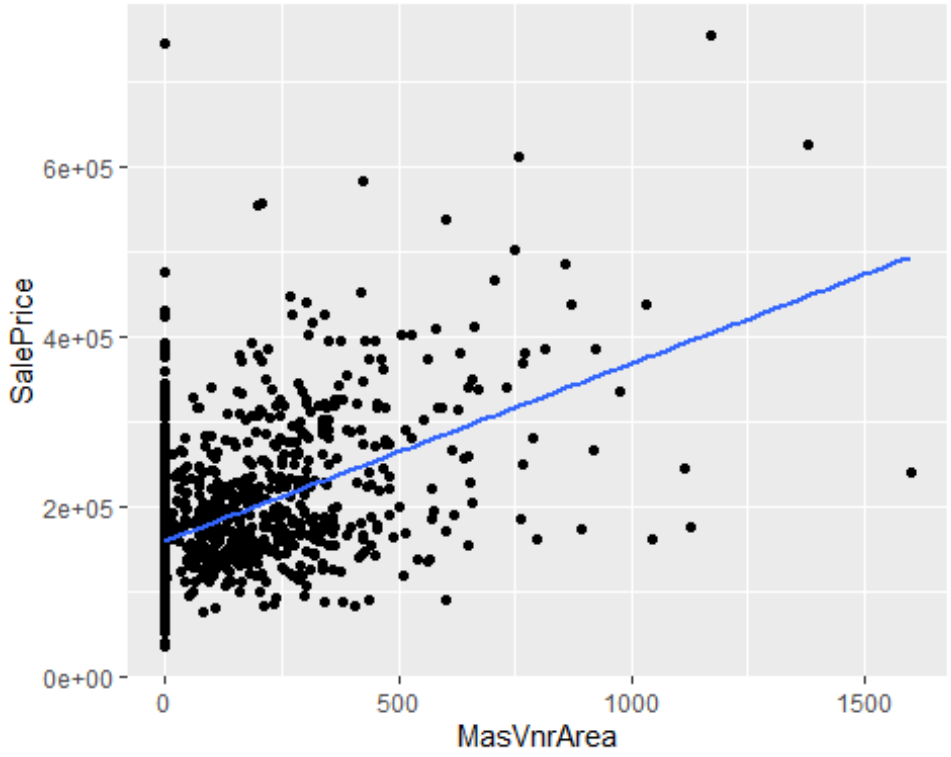

```
ggplot(case, aes(x = as.factor(MasVnrType), y = SalePrice)) + geom_violin()
```



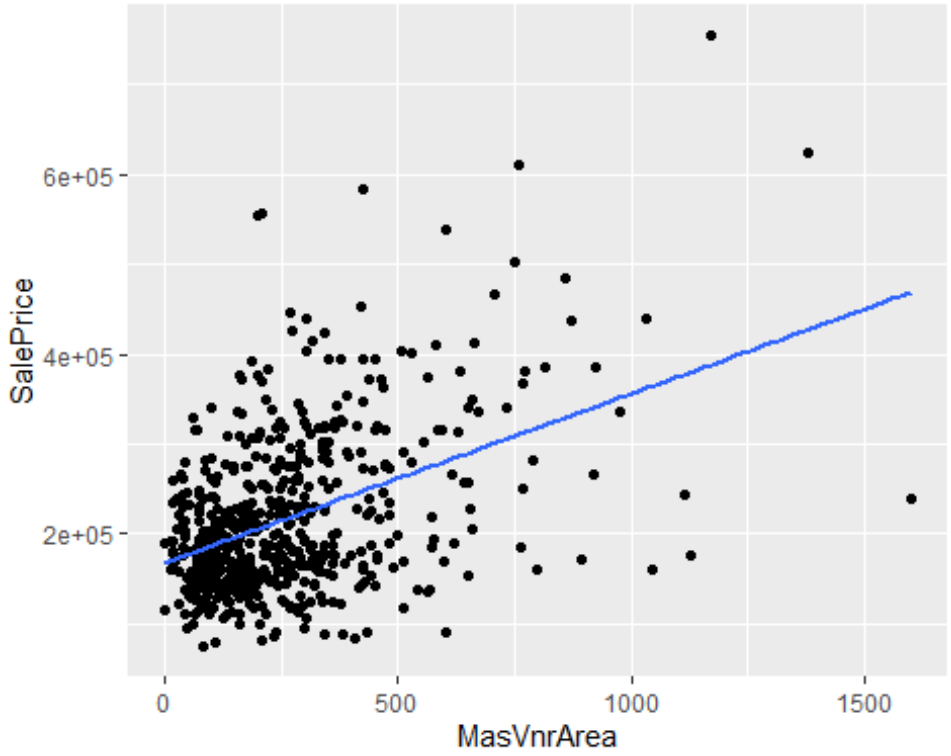
Il valore di ETA^2 è di circa 0.19 che mostra una leggera correlazione tra il prezzo e MasVnrType. In particolare, in media le case rifinite in Stone e in BrickFace hanno un prezzo più alto.

Variabile MasVnrArea

```
cor(case$MasVnrArea, case$SalePrice, use = "complete.obs")

## [1] 0.477493

model <- lm(SalePrice ~ MasVnrArea, data = subset(case, MasVnrArea != 0),
na.action = "na.omit")
summary(model)

##
## Call:
## lm(formula = SalePrice ~ MasVnrArea, data = subset(case, MasVnrArea !=
##     0), na.action = "na.omit")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -229678  -54764   -9620   43111  367196
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 167751.58    5258.52   31.90   <2e-16 ***
## MasVnrArea     188.08      16.08   11.69   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80140 on 589 degrees of freedom
## Multiple R-squared:  0.1884, Adjusted R-squared:  0.1871
## F-statistic: 136.8 on 1 and 589 DF,  p-value: < 2.2e-16

ggplot(case, aes(x = MasVnrArea, y = SalePrice)) + geom_point() +
geom_smooth(method = "lm", se = FALSE)

## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 8 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 8 rows containing missing values or values outside the scale
range
## (`geom_point()`).
```




```
ggplot(data = subset(case, MasVnrArea != 0), aes(x = MasVnrArea, y = SalePrice)) +
geom_point() + geom_smooth(method = "lm", se = FALSE)

## `geom_smooth()` using formula = 'y ~ x'
```



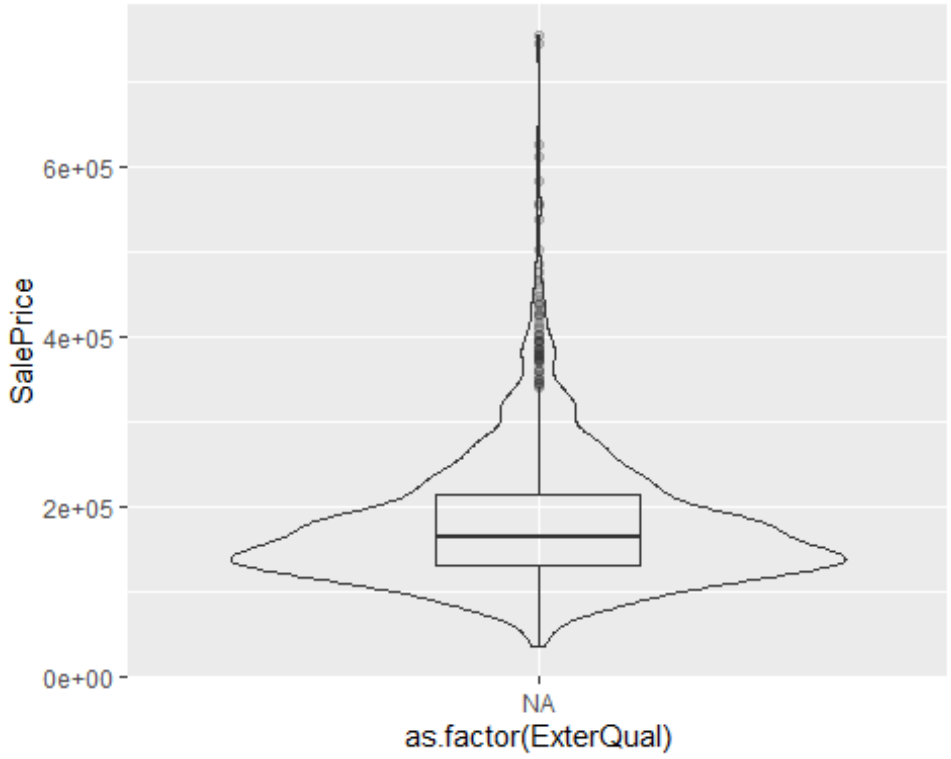
Si nota che la correlazione è di circa 0.48. La variabile MasVnrArea quindi influenza debolmente il prezzo delle case, come si può vedere dal grafico.

Variabile ExterQual

```
case$ExterQual <- factor(case$ExterQual, levels = c("Fa","TA","Gd","Ex"))
calcola_devianza(case$SalePrice, case$ExterQual)
```

```
## $devianza_totale
## [1] 0
##
## $devianza_tra_gruppi
## [1] NaN
##
## $devianza_entro_gruppi
## [1] 0
##
## $eta2
## [1] NaN
```

```
ggplot(case, aes(x = as.factor(ExterQual), y = SalePrice, fill =
as.factor(ExterQual))) + geom_violin() + geom_boxplot(width=0.3, alpha=1/5) +
scale_fill_brewer(palette = "RdYlGn") + guides(fill = FALSE)
```



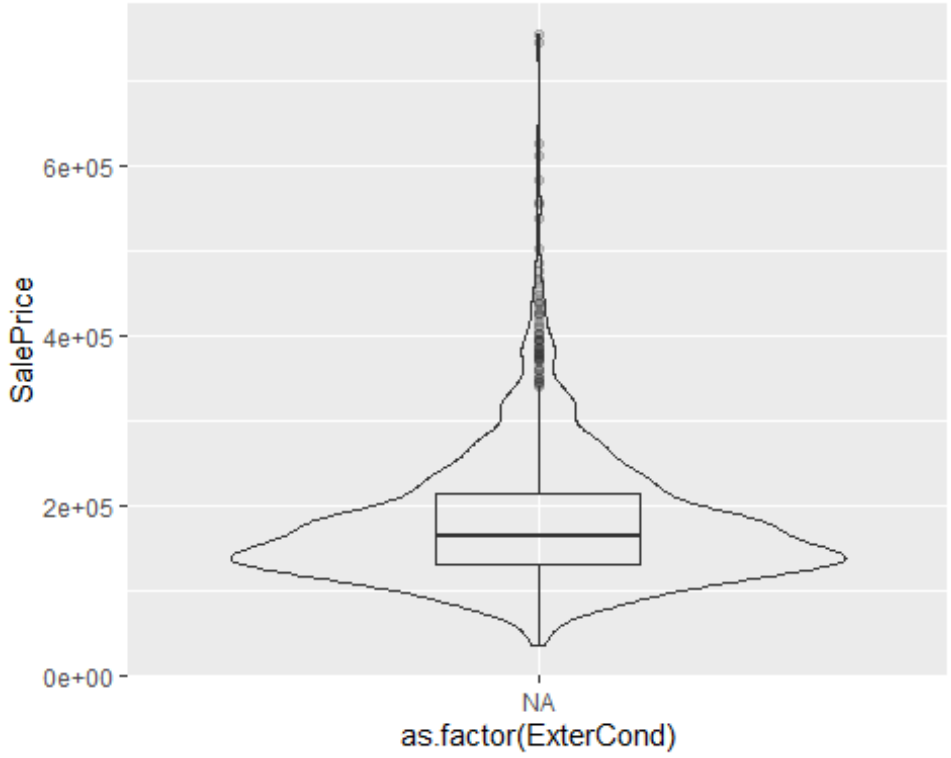
La variabile ExterQual influenza fortemente il prezzo di una casa. Si può infatti notare che il rapporto tra la la devianza between e la devianza totale è di 0.48 . Come si vede anche dal grafico più la finitura esterna è di qualità e più il prezzo della casa sale.

Variabile ExterCond

```
case$ExterCond <- factor(case$ExterCond, levels = c("Po","Fa","TA","Gd","Ex"))
calcola_devianza(case$SalePrice, case$ExterCond)
```

```
## $devianza_totale
## [1] 0
##
## $devianza_tra_gruppi
## [1] NaN
##
## $devianza_entro_gruppi
## [1] 0
##
## $eta2
## [1] NaN
```

```
ggplot(case, aes(x = as.factor(ExterCond), y = SalePrice, fill =
as.factor(ExterCond))) + geom_violin() + geom_boxplot(width=0.3, alpha=1/5)  +
scale_fill_brewer(palette = "RdYlGn") + guides(fill = FALSE)
```



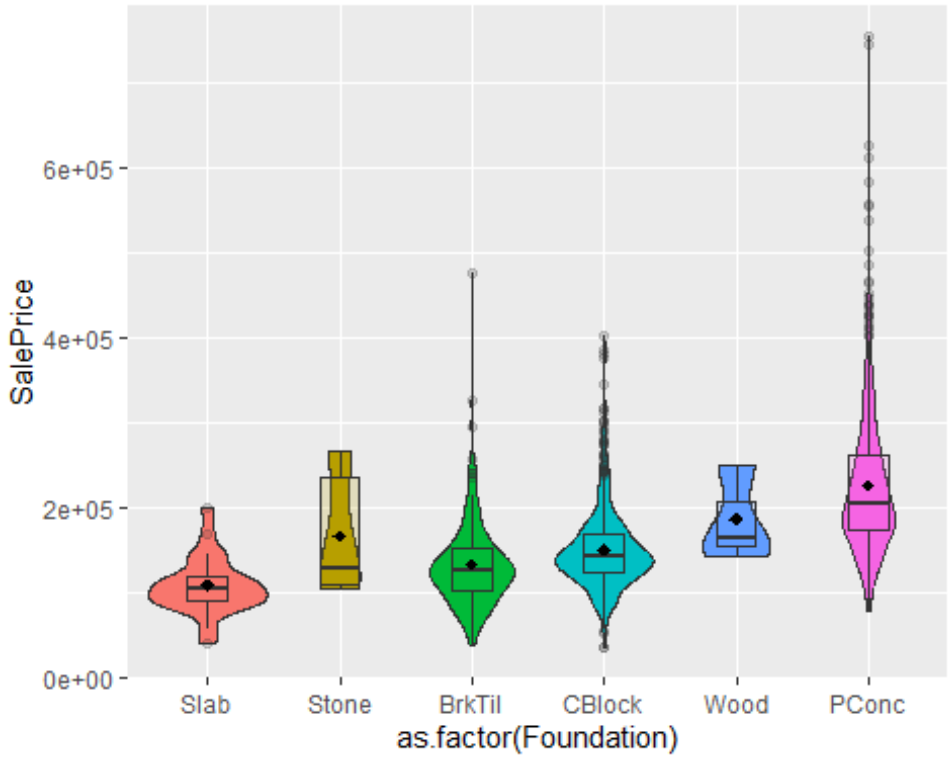
La variabile ExterCond influenza debolmente il prezzo. Il valore di ETA^2 è di 0.02 infatti dal grafico si evince che le classi Average/typical e Good contengono numerosi valori outlier, con un prezzo più alto della media.

Variabile Foundation

```
case$Foundation <- factor(case$Foundation, levels =
c("Slab","Stone","BrkTil","CBlock","Wood","PConc"))
calcola_devianza(case$SalePrice, case$Foundation)
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 2.360618e+12
##
## $devianza_entro_gruppi
## [1] 6.847294e+12
##
## $eta2
## [1] 0.2563684
```

```
ggplot(case, aes(x = as.factor(Foundation), y = SalePrice, fill =
as.factor(Foundation))) + geom_violin() + geom_boxplot(width=0.3, alpha=1/5) +
guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape = 18, size =
2, color = "black")
```



Il tipo di fondamenta influenza moderatamente il prezzo delle case. Infatti, le case con le fondamenta in Poured Contrete hanno in media il prezzo più alto e, anche in assoluto, le case con il prezzo più alto hanno le fondamenta in Poured Contrete. Tuttavia BrkTil e CBlock contengono alcuni valori outlier di conseguenza il valore di ETA^2 non è elevatissimo.

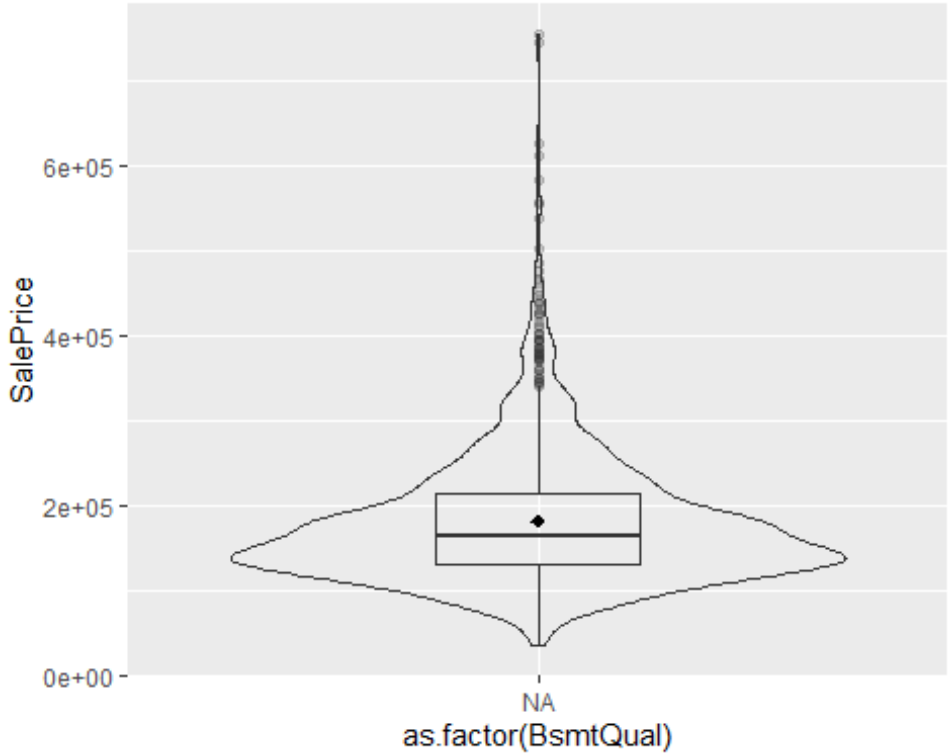
Variabile BsmtQual

```
case$BsmtQual <- factor(case$BsmtQual, levels = c("Fa","TA","Gd","Ex"))
calcola_devianza(case$SalePrice, case$BsmtQual)
```

```
## $devianza_totale
## [1] 0
```

```
##
## $devianza_tra_gruppi
## [1] NaN
##
## $devianza_entro_gruppi
## [1] 0
##
## $eta2
## [1] NaN
```

```
ggplot(case, aes(x = as.factor(BsmtQual), y = SalePrice, fill =
as.factor(BsmtQual))) + geom_violin() + geom_boxplot(width=0.3, alpha=1/5) +
scale_fill_brewer(palette = "RdYlGn") + guides(fill = FALSE) + stat_summary(fun =
mean, geom = "point", shape = 18, size = 2, color = "black")
```



Si nota una forte correlazione tra la qualità del seminterrato e il prezzo delle case. Il rapporto tra la devianza tra i gruppi e la Devianza Totale è vicina al 45% che è dimostrato dal grafico in cui si vede che la media dei prezzi delle case con la qualità del seminterrato Fair è la più bassa mentre la media dei prezzi delle case con la qualità del seminterrato Excellent è la più alta.

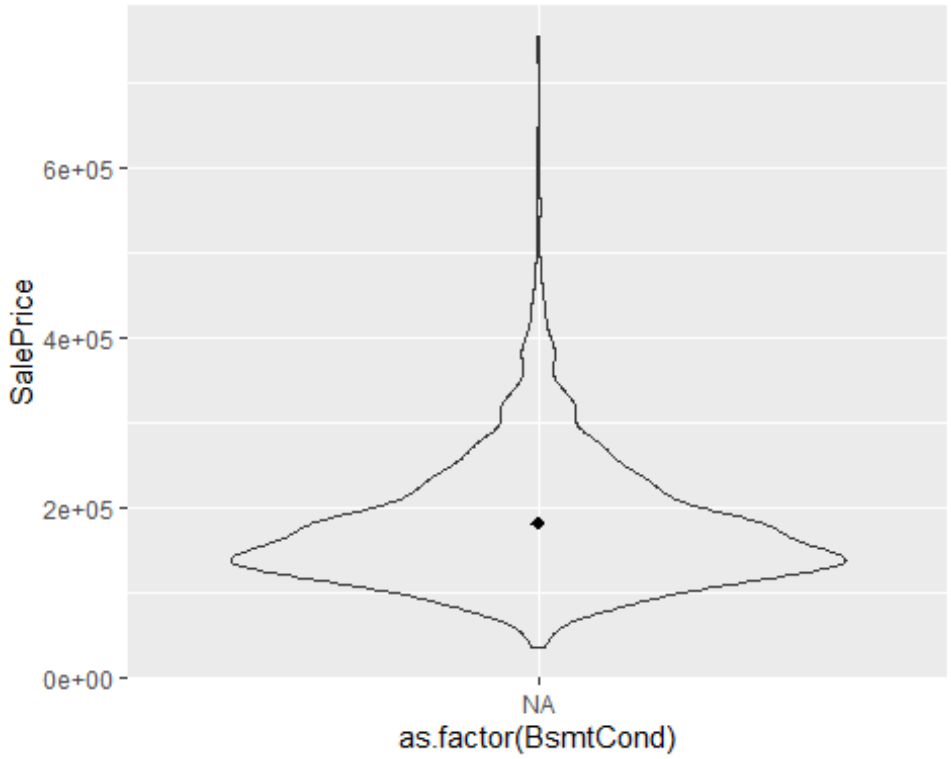
Variabile BsmtCond

```
case$BsmtCond <- factor(case$BsmtCond, levels = c("Po","Fa","TA","Gd"))
calcola_devianza(case$SalePrice, case$BsmtCond)
```

```
## $devianza_totale
## [1] 0
##
## $devianza_tra_gruppi
```

```
## [1] NaN
##
## $devianza_entro_gruppi
## [1] 0
##
## $eta2
## [1] NaN
```

```
ggplot(case, aes(x = as.factor(BsmtCond), y = SalePrice, fill =
as.factor(BsmtCond))) + geom_violin() + scale_fill_brewer(palette = "RdYlGn") +
guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape = 18, size =
2, color = "black")
```



La correlazione tra la condizione del seminterrato e la variabile "SalePrice", a casusa di numerosi valori outlier presenti nella classe Average/Typical e per il fatto che la devianza within è molto elevata, è bassa. Infatti, la devianza between è molto minore rispetto alla devianza totale.

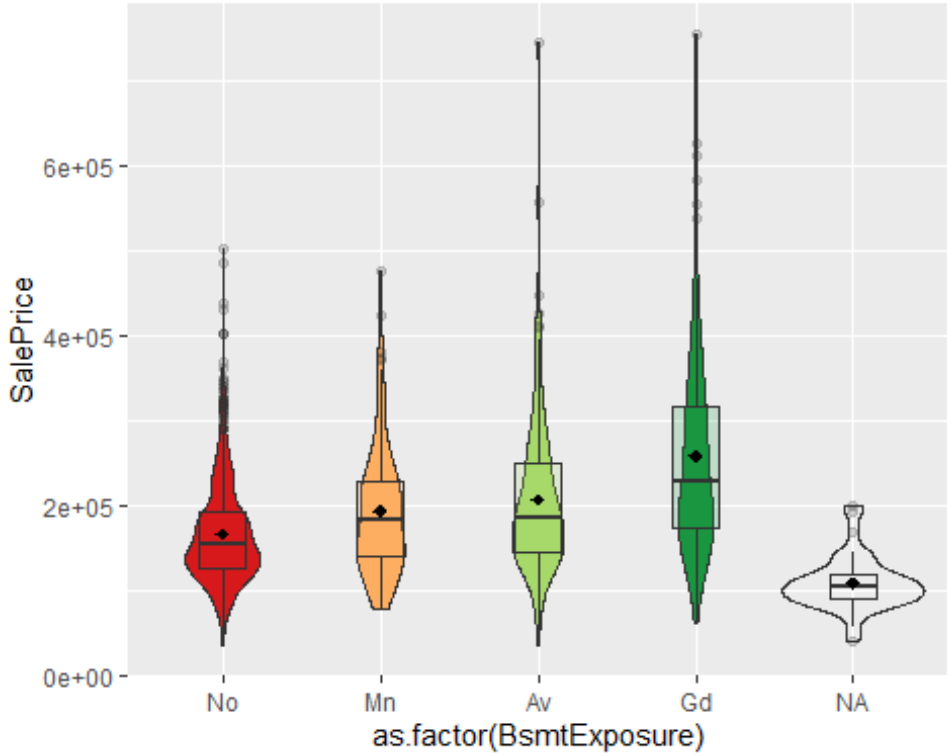
Variabile BsmtExposure

```
case$BsmtExposure <- factor(case$BsmtExposure, levels = c("No", "Mn", "Av", "Gd"))
calcola_devianza(case$SalePrice, case$BsmtExposure)
```

```
## $devianza_totale
## [1] 8.961891e+12
##
## $devianza_tra_gruppi
## [1] 1.16877e+12
##
## $devianza_entro_gruppi
```

```
## [1] 7.793121e+12
##
## $eta2
## [1] 0.1304156
```

```
ggplot(case, aes(x = as.factor(BsmtExposure), y = SalePrice, fill =
as.factor(BsmtExposure))) + geom_violin() + geom_boxplot(width=0.3, alpha=1/5) +
scale_fill_brewer(palette = "RdYlGn") + guides(fill = FALSE) + stat_summary(fun =
mean, geom = "point", shape = 18, size = 2, color = "black")
```



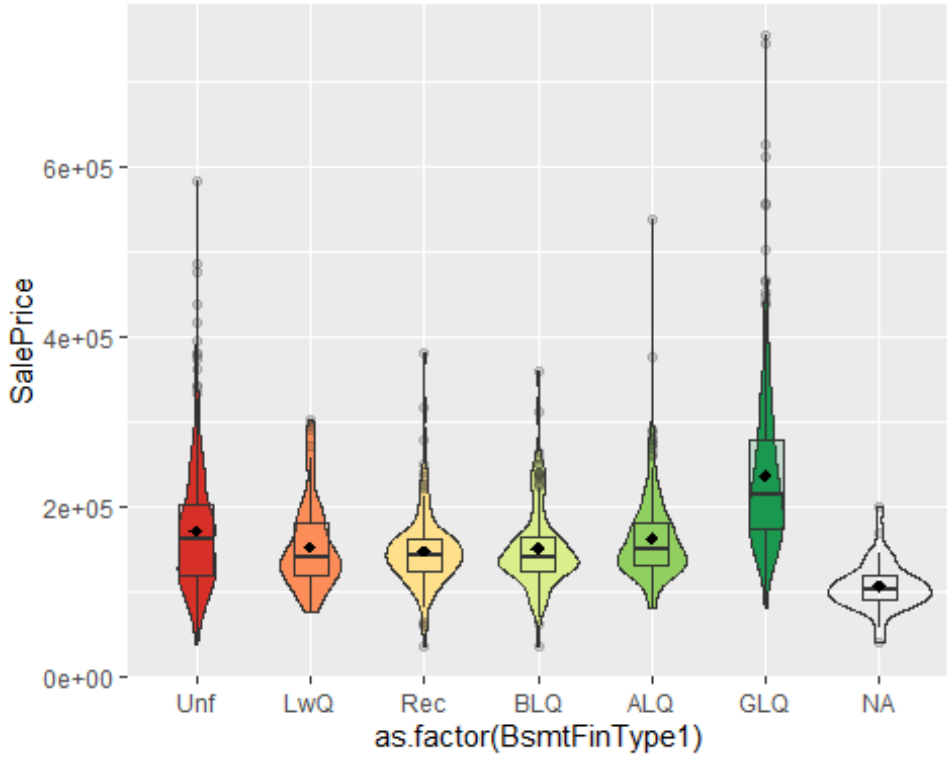
L'influenza con la variabile BsmtExposure è bassa, infatti le medie dei vari gruppi sono piuttosto vicine tra loro. Tuttavia, in parte si osserva che le case con una buona esposizione hanno un prezzo medio leggermente maggiore rispetto alle altre.

Variabile BsmtFinType1

```
case$BsmtFinType1 <- factor(case$BsmtFinType1, levels = c("Unf", "LwQ", "Rec",
"BLQ", "ALQ", "GLQ"))
calcola_devianza(case$SalePrice, case$BsmtFinType1)
```

```
## $devianza_totale
## [1] 8.961984e+12
##
## $devianza_tra_gruppi
## [1] 1.726056e+12
##
## $devianza_entro_gruppi
## [1] 7.235927e+12
##
```

```
## $eta2
## [1] 0.1925976

ggplot(case, aes(x = as.factor(BsmtFinType1), y = SalePrice, fill =
as.factor(BsmtFinType1))) + geom_violin() + geom_boxplot(width=0.3, alpha=1/5) +
scale_fill_brewer(palette = "RdYlGn") + guides(fill = FALSE) + stat_summary(fun =
mean, geom = "point", shape = 18, size = 2, color = "black")
```



Dal grafico si nota che le medie e le mediane di ogni classe sono tutte sullo stesso livello. Questo indica un basso valore di varianza between, infatti il rapporto con la varianza totale è di 0.1925976 . Questo a sua volta indica una bassa influenza della variabile BsmtFinType1 sui prezzi delle case.

Variabile BsmtFinSF1

```
cor(case$BsmtFinSF1, case$SalePrice, use = "complete.obs")

## [1] 0.3864198

model <- lm(SalePrice ~ BsmtFinSF1, data = case, na.action = "na.omit")
model <- lm(SalePrice ~ BsmtFinSF1, data = subset(case, BsmtFinSF1 != 0),
na.action = "na.omit")
summary(model)

##
## Call:
## lm(formula = SalePrice ~ BsmtFinSF1, data = subset(case, BsmtFinSF1 !=
##     0), na.action = "na.omit")
##
## Residuals:
```

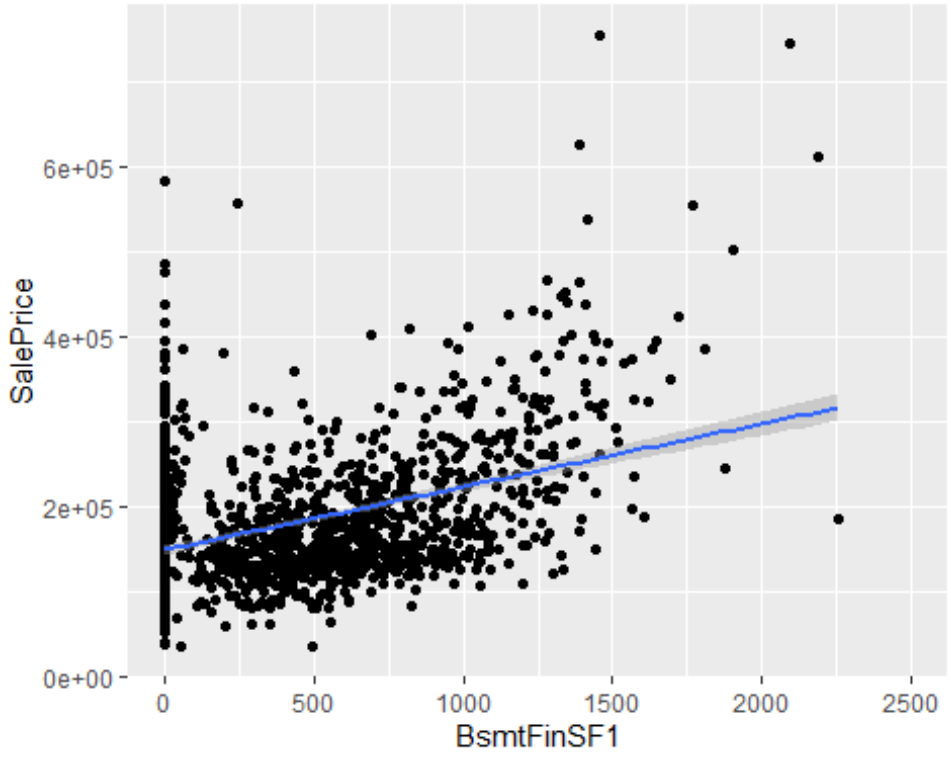
```
##     Min      1Q  Median      3Q     Max
## -494705  -47157  -14425   33288  491811
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 127200.60    4281.20   29.71   <2e-16 ***
## BsmtFinSF1      93.46       5.55   16.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 72020 on 991 degrees of freedom
## Multiple R-squared:  0.2225, Adjusted R-squared:  0.2217
## F-statistic: 283.6 on 1 and 991 DF,  p-value: < 2.2e-16
```

```
ggplot(case, aes(x = BsmtFinSF1, y = SalePrice), ) + geom_point() +
geom_smooth(method = "lm") + xlim(0,2500)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

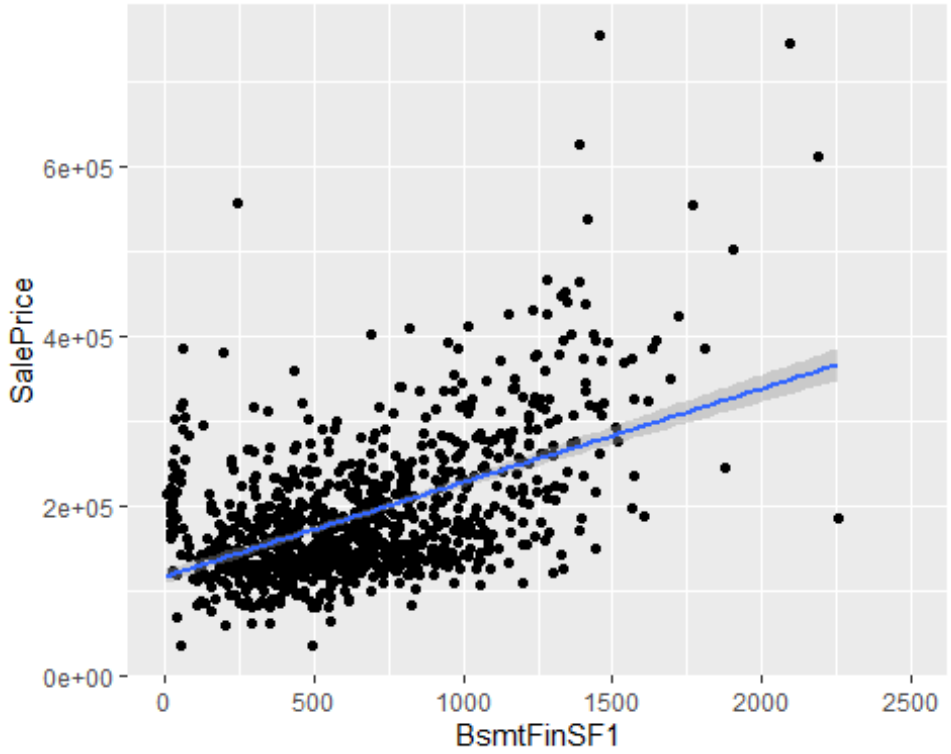
```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 1 row containing missing values or values outside the scale
range
## (`geom_point()`).
```




```
ggplot(data = subset(case, BsmtFinSF1 != 0), aes(x = BsmtFinSF1, y = SalePrice), )
+ geom_point() + geom_smooth(method = "lm") + xlim(0,2500)
```

```
## `geom_smooth()` using formula = 'y ~ x'

## Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).
## Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



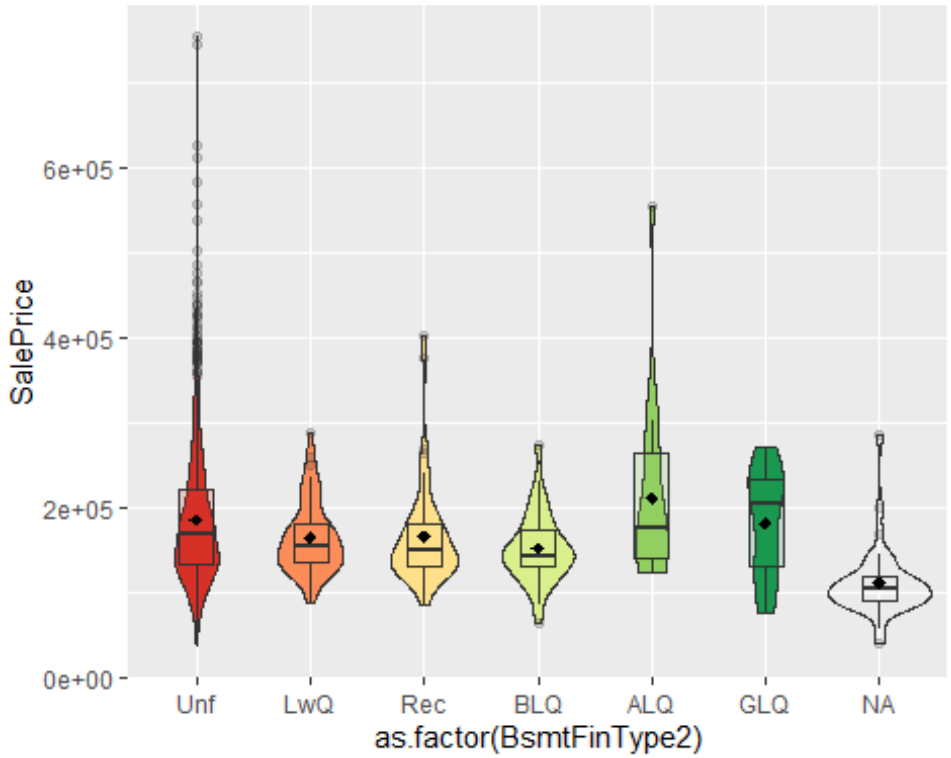
La correlazione delle due variabili è di 0.39 e come si nota dal grafico la il modello non ha una un'ottima accuratezza. La variabile BsmtFinSF1 quindi, non influenza pesantemente il prezzo delle case.

Variabile BsmtFinType2

```
case$BsmtFinType2 <- factor(case$BsmtFinType2, levels = c("Unf", "LwQ", "Rec",
"BLQ", "ALQ", "GLQ"))
calcola_devianza(case$SalePrice, case$BsmtFinType2)
```

```
## $devianza_totale
## [1] 8.951751e+12
##
## $devianza_tra_gruppi
## [1] 84615086215
##
## $devianza_entro_gruppi
## [1] 8.867136e+12
##
## $eta2
## [1] 0.009452351
```

```
ggplot(case, aes(x = as.factor(BsmtFinType2), y = SalePrice, fill =
as.factor(BsmtFinType2))) + geom_violin() + geom_boxplot(width=0.3, alpha=1/5) +
scale_fill_brewer(palette = "RdYlGn") + guides(fill = FALSE) + stat_summary(fun =
mean, geom = "point", shape = 18, size = 2, color = "black")
```



Le due variabili non sono correlate infatti un grande numero di case, anche con prezzi elevati, hanno il seminterrato incompleto. Questo si può notare anche dal valore di ETA^2 che è molto basso (0.009452351) dovuto alla codevianza between molto bassa.

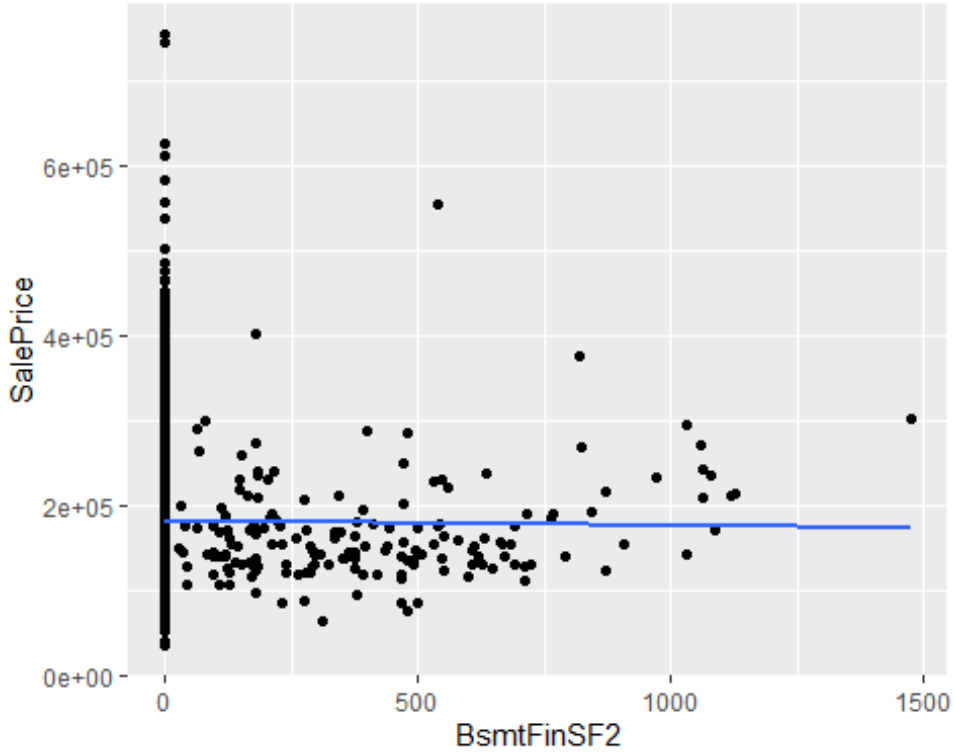
Variabile BsmtFinSF2

```
cor(case$BsmtFinSF2, case$SalePrice, use = "complete.obs")

## [1] -0.01137812

model <- lm(SalePrice ~ BsmtFinSF2, data = case, na.action = "na.omit")
model <- lm(SalePrice ~ BsmtFinSF2, data = subset(case, BsmtFinSF2 != 0),
na.action = "na.omit")
summary(model)

##
## Call:
## lm(formula = SalePrice ~ BsmtFinSF2, data = subset(case, BsmtFinSF2 !=
##     0), na.action = "na.omit")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -102685  -36647  -15770   20085  380073
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 151619.17    8229.57  18.424  < 2e-16 ***
## BsmtFinSF2      43.24      16.58   2.608  0.00995 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60870 on 165 degrees of freedom
## Multiple R-squared:  0.03958,    Adjusted R-squared:  0.03376
## F-statistic:   6.8 on 1 and 165 DF,  p-value: 0.009949
```

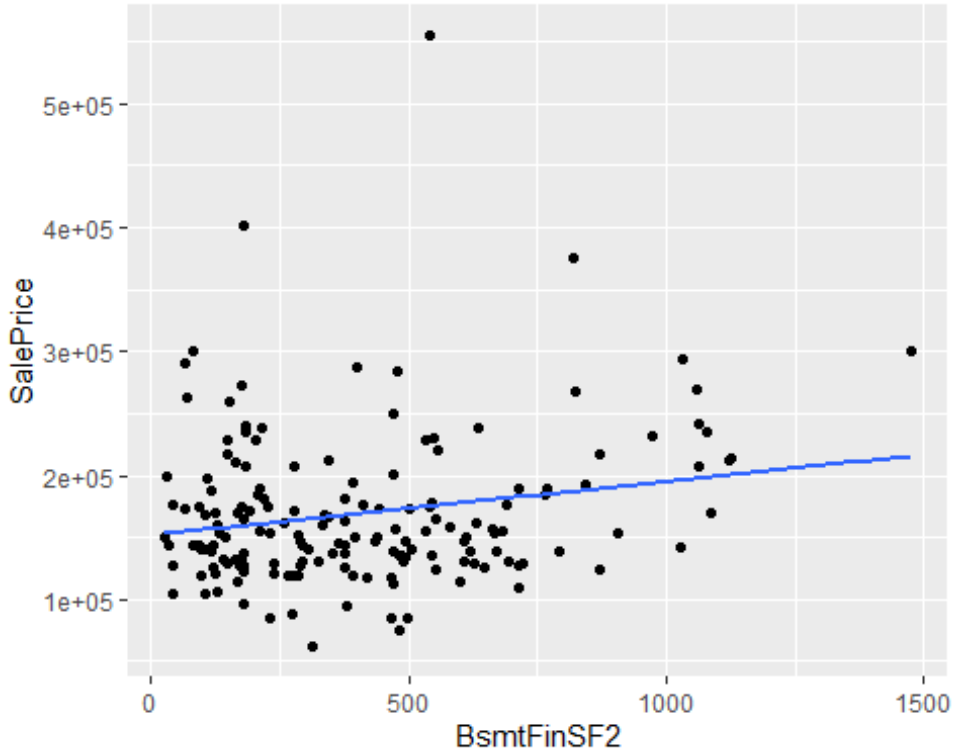
```
ggplot(case, aes(x = BsmtFinSF2, y = SalePrice), ) + geom_point() +
geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(data = subset(case, BsmtFinSF2 != 0), aes(x = BsmtFinSF2, y = SalePrice), )
+ geom_point() + geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Essendo questa variabile la superficie della finitura descritta alla variabile precedente, anche qui non c'è nessuna relazione tra le due variabili, il valore della correlazione è di -0.11% : molto prossima allo zero.

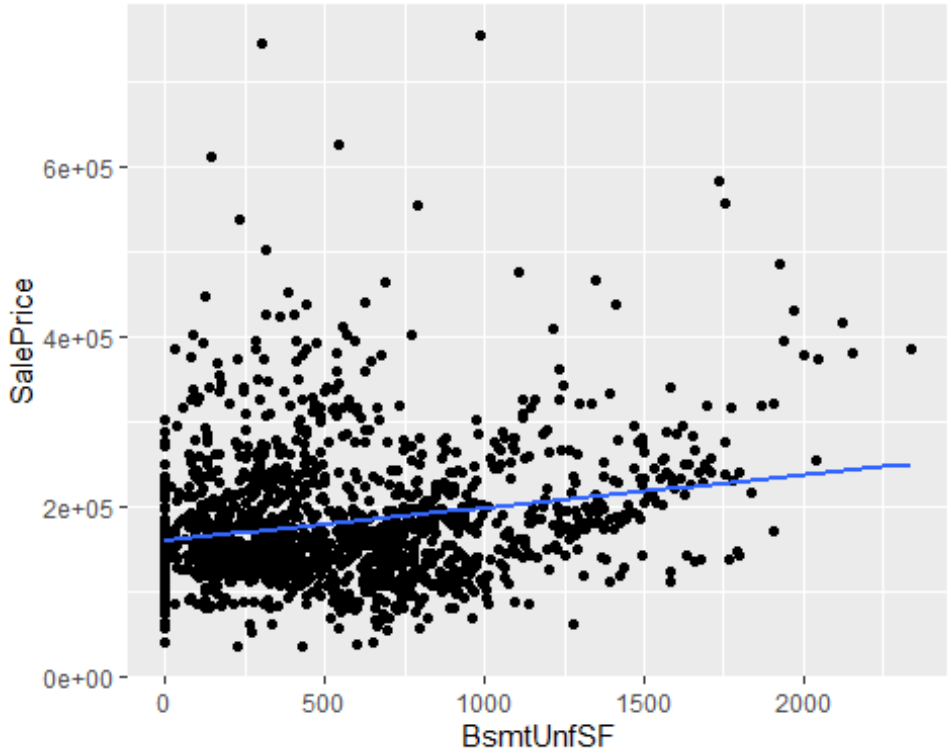
Variabile BsmtUnfSF

```
cor(case$BsmtUnfSF, case$SalePrice, use = "complete.obs")

## [1] 0.2144791

model <- lm(SalePrice ~ BsmtUnfSF, data = case, na.action = "na.omit")
model <- lm(SalePrice ~ BsmtUnfSF, data = subset(case, BsmtUnfSF != 0), na.action
= "na.omit")
summary(model)

##
## Call:
## lm(formula = SalePrice ~ BsmtUnfSF, data = subset(case, BsmtUnfSF !=
##     0), na.action = "na.omit")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -146344  -49762  -16828   29291  570339
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.651e+05  3.810e+03  43.326  < 2e-16 ***
## BsmtUnfSF   3.194e+01  5.081e+00   6.287 4.38e-10 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 79300 on 1340 degrees of freedom
## Multiple R-squared:  0.02865,    Adjusted R-squared:  0.02792
## F-statistic: 39.52 on 1 and 1340 DF,  p-value: 4.379e-10
```

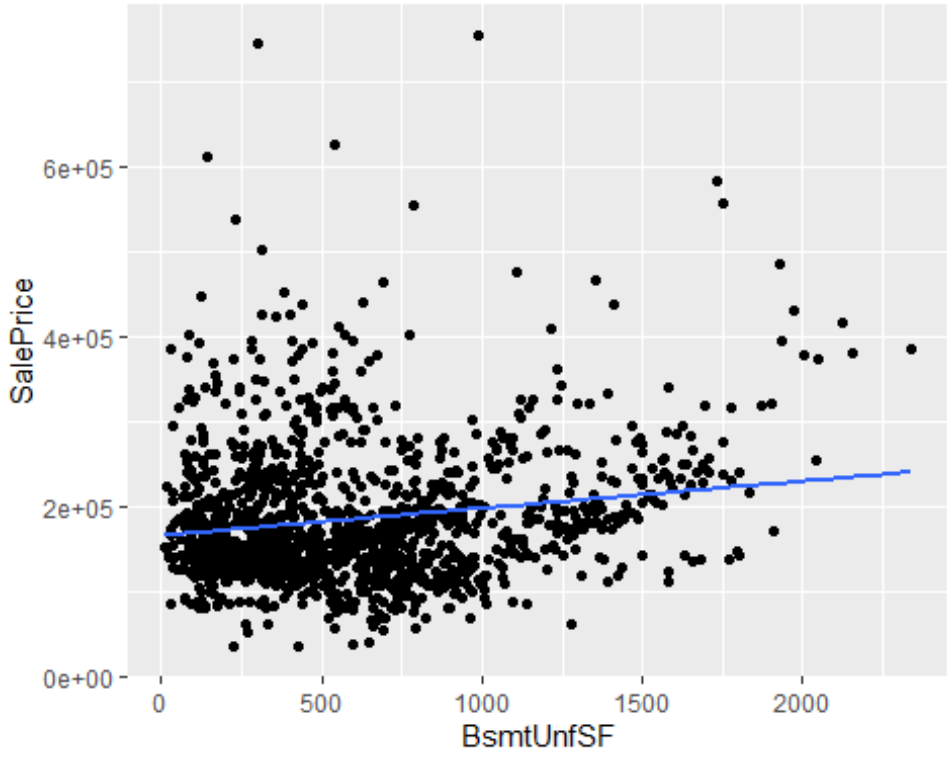
```
ggplot(case, aes(x = BsmtUnfSF, y = SalePrice), ) + geom_point() +
geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
ggplot(data = subset(case, BsmtUnfSF != 0), aes(x = BsmtUnfSF, y = SalePrice), ) +
geom_point() + geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Non si nota una forte correlazione tra la variabile Superficie di seminterrato incompleto e il prezzo dell'abitazione. La correlazione è del 0.21448 e il valore di R^2 è 0.045 . Anche escludendo tutti i valori che hanno superficie di seminterrato incompleto pari a zero si nota che il valore di R^2 rimane basso.

Variabile TotalBsmtSF

```
cor(case$TotalBsmtSF, case$SalePrice, use = "complete.obs")

## [1] 0.6135806

model <- lm(SalePrice ~ TotalBsmtSF, data = case, na.action = "na.omit")
model <- lm(SalePrice ~ TotalBsmtSF, data = subset(case, TotalBsmtSF != 0),
na.action = "na.omit")
summary(model)

##
## Call:
## lm(formula = SalePrice ~ TotalBsmtSF, data = subset(case, TotalBsmtSF !=
##     0), na.action = "na.omit")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -616945  -39053  -14330   34192  411451
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54618.116   4727.844   11.55   <2e-16 ***
## TotalBsmtSF   118.220      4.077   29.00   <2e-16 ***
```

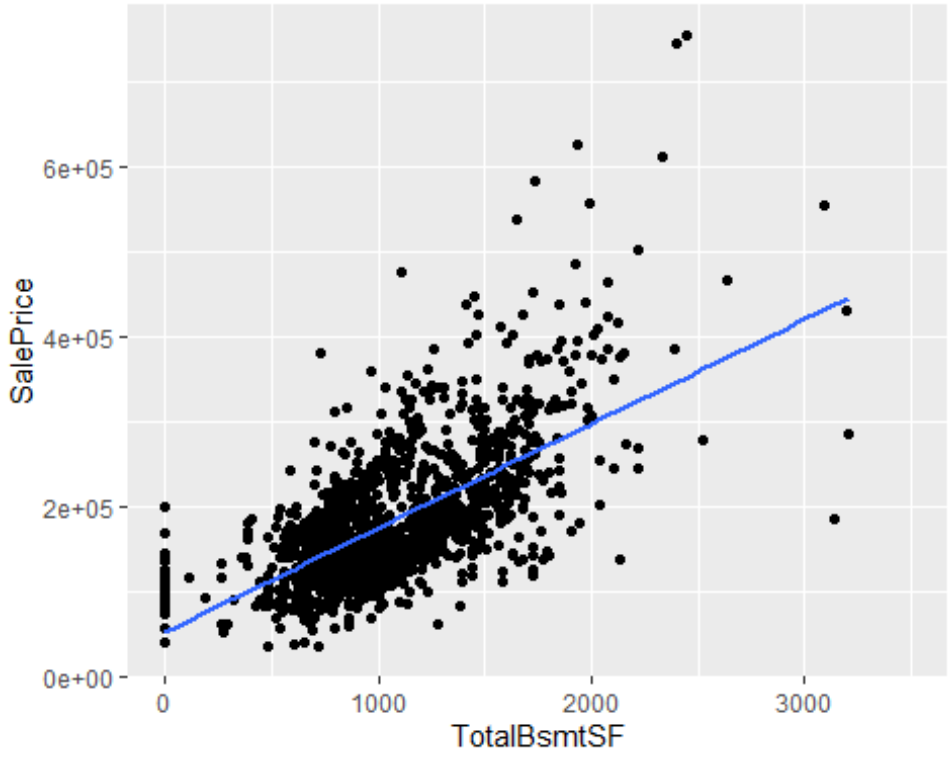
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62950 on 1421 degrees of freedom
## Multiple R-squared:  0.3717, Adjusted R-squared:  0.3713
## F-statistic: 840.7 on 1 and 1421 DF,  p-value: < 2.2e-16
```

```
ggplot(case, aes(x = TotalBsmtSF, y = SalePrice), ) + geom_point() +
geom_smooth(method = "lm", se = FALSE) + xlim(0,3500)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
## (`stat_smooth()`).
```

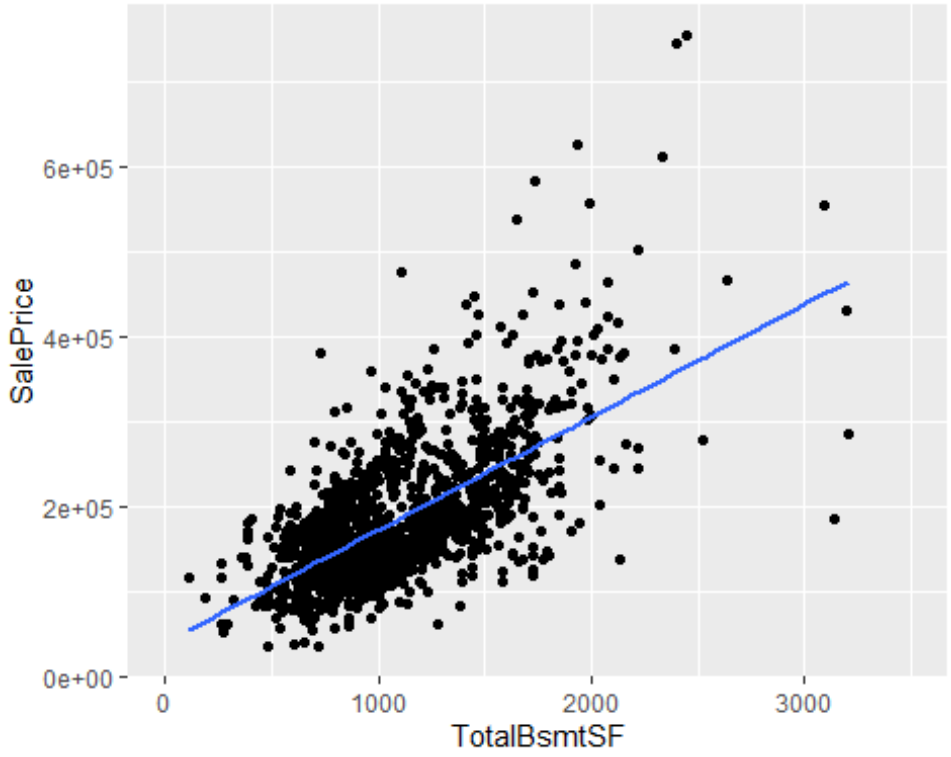
```
## Warning: Removed 1 row containing missing values or values outside the scale
range
## (`geom_point()`).
```




```
ggplot(data = subset(case, TotalBsmtSF != 0), aes(x = TotalBsmtSF, y = SalePrice),
) + geom_point() + geom_smooth(method = "lm", se = FALSE) + xlim(0,3500)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 1 row containing non-finite outside the scale range
(`stat_smooth()`).
## Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```



La dipendenza tra la variabile Superficie del seminterrato e la variabile prezzo è evidente. Il valore della correlazione è di 0.6135806 e dal grafico si nota come all'aumentare della superficie il valore del prezzo tende ad essere più alto.

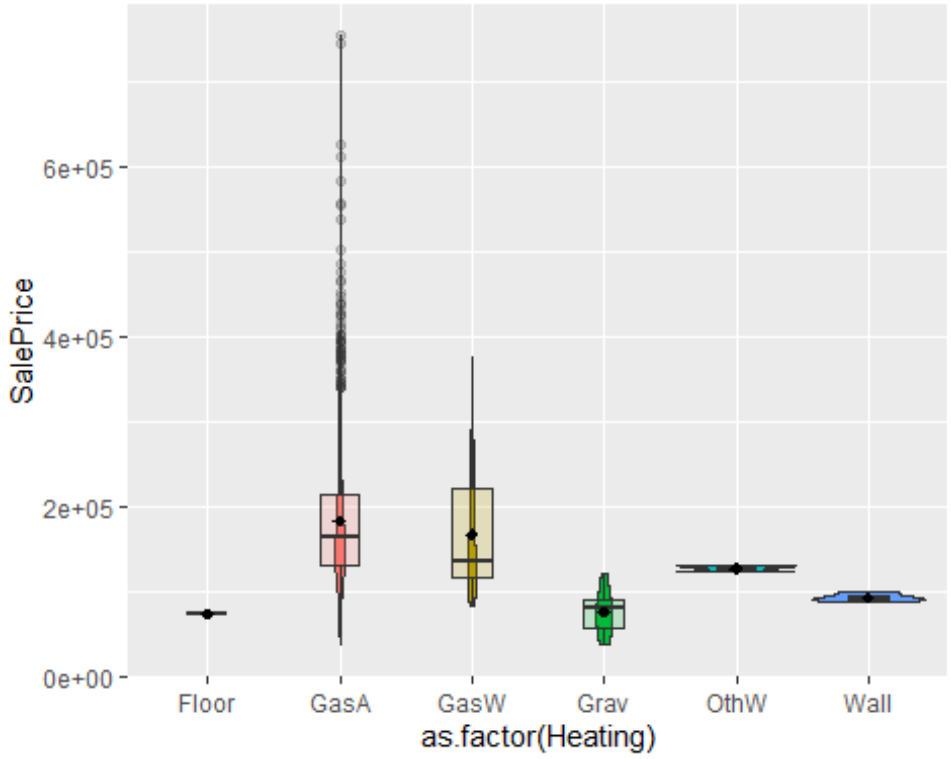
Variabile Heating

```
case$Heating <- factor(case$Heating)
calcola_devianza(case$SalePrice, case$Heating)
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 132935862931
##
## $devianza_entro_gruppi
## [1] 9.074975e+12
##
## $eta2
## [1] 0.01443714
```

```
ggplot(case, aes(x = as.factor(Heating), y = SalePrice, fill =
as.factor(Heating))) + geom_violin() + geom_boxplot(width=0.3, alpha=1/5) +
guides(fill = FALSE) + stat_summary(fun = mean, geom = "point", shape = 18, size =
2, color = "black")
```

```
## Warning: Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
```



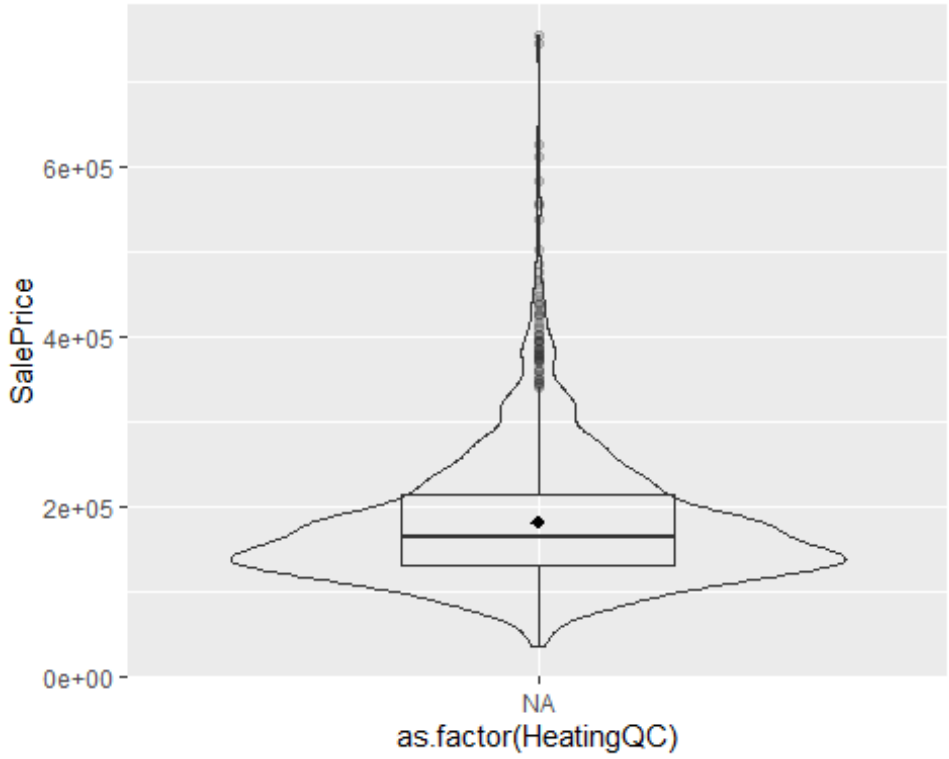
La variabile Heating ha una bassa influenza sulla variabile SalePrice. La devianza between ha un valore piuttosto basso, infatti le medie dei vari gruppo sono all'incirca alla stessa altezza. Si nota poi la presenza di numerosi valori outlier nella categoria GasA.

Variabile HeatingQC

```
case$HeatingQC <- factor(case$HeatingQC, levels = c("Po","Fa","TA","Gd","Ex"))
calcola_devianza(case$SalePrice, case$HeatingQC)
```

```
## $devianza_totale
## [1] 0
##
## $devianza_tra_gruppi
## [1] NaN
##
## $devianza_entro_gruppi
## [1] 0
##
## $eta2
## [1] NaN
```

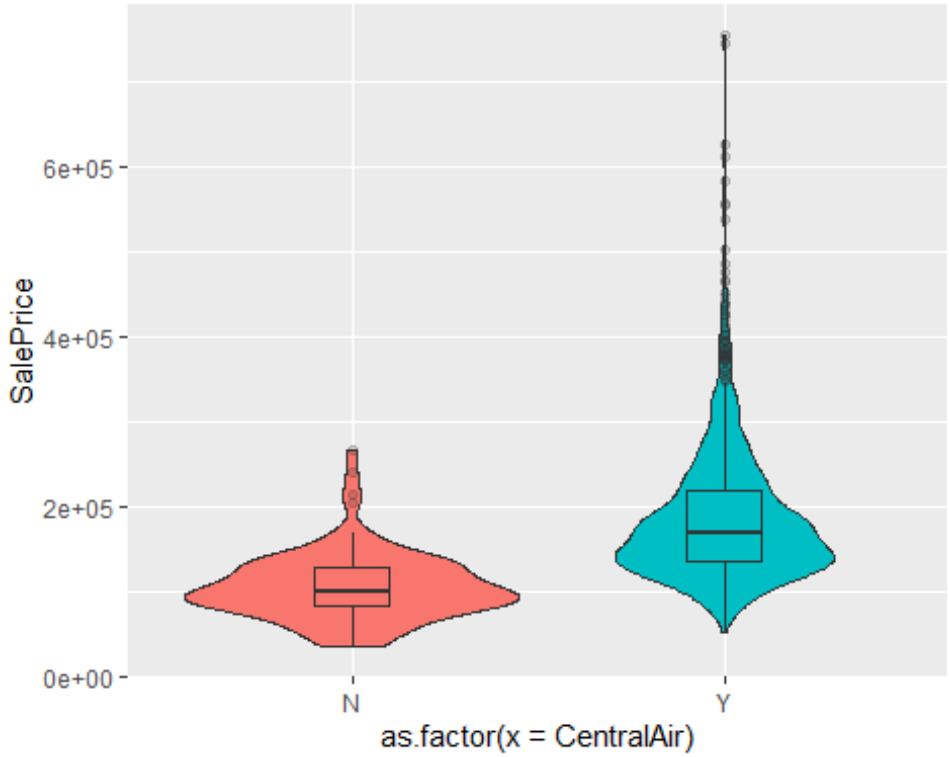
```
ggplot(case, aes(x = as.factor(HeatingQC), y = SalePrice, fill =
as.factor(HeatingQC))) + geom_violin() + geom_boxplot(width=0.4, alpha=1/5) +
scale_fill_brewer(palette = "RdYlGn") + guides(fill = FALSE) + stat_summary(fun =
mean, geom = "point", shape = 18, size = 2, color = "black")
```



La qualità dell'impianto di riscaldamento è una variabile che influenza il prezzo di vendita dell'abitazione. Si nota che la categoria Eccellente contiene le case con il prezzo più alto.

Variabile CentralAir

```
case$CentralAir <- factor(case$CentralAir)
ggplot(case, aes(x = as.factor(x = CentralAir), y = SalePrice, fill = as.factor(x
=CentralAir))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)
```




```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
case$CentralAir)
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 581625508784
##
## $devianza_entro_gruppi
## [1] 8.626286e+12
##
## $eta2
## [1] 0.06316585
```

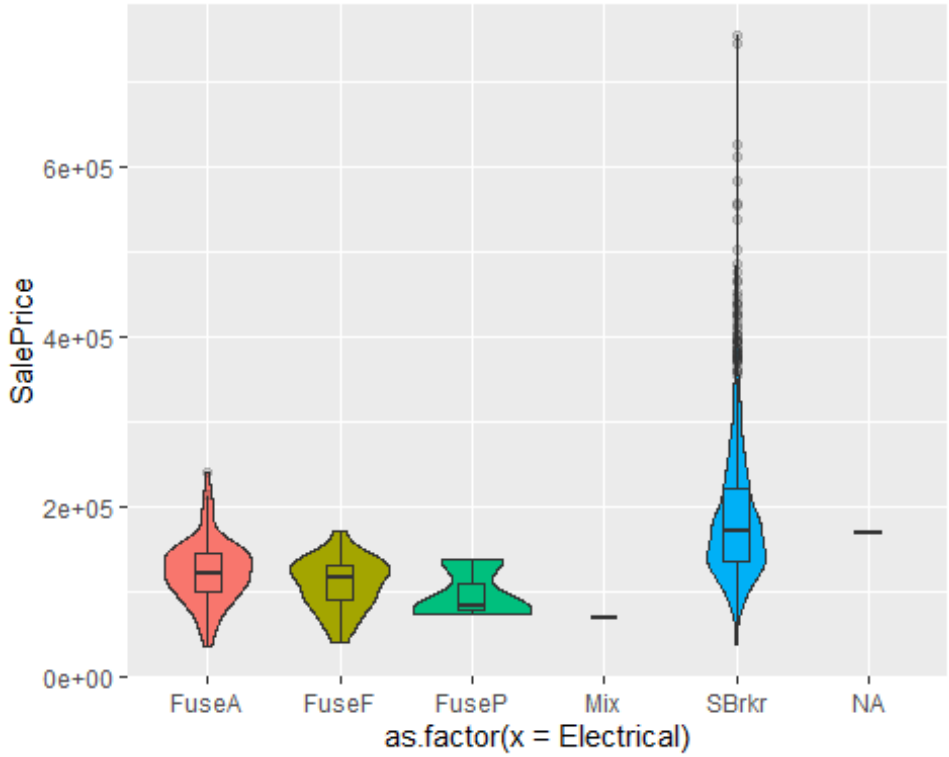
Il sistema di aria centralizzato non sembra avere una grande influenza sul prezzo, si osserva che le mediane fra i due gruppi sono simili. Si osserva inoltre una maggiore dispersione nelle case del gruppo yes.

Variabile Electrical

```
case$Electrical <- factor(case$Electrical)
ggplot(case, aes(x = as.factor(x = Electrical), y = SalePrice, fill = as.factor(x
= Electrical))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)
```

```
## Warning: Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
```

```
## Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
```



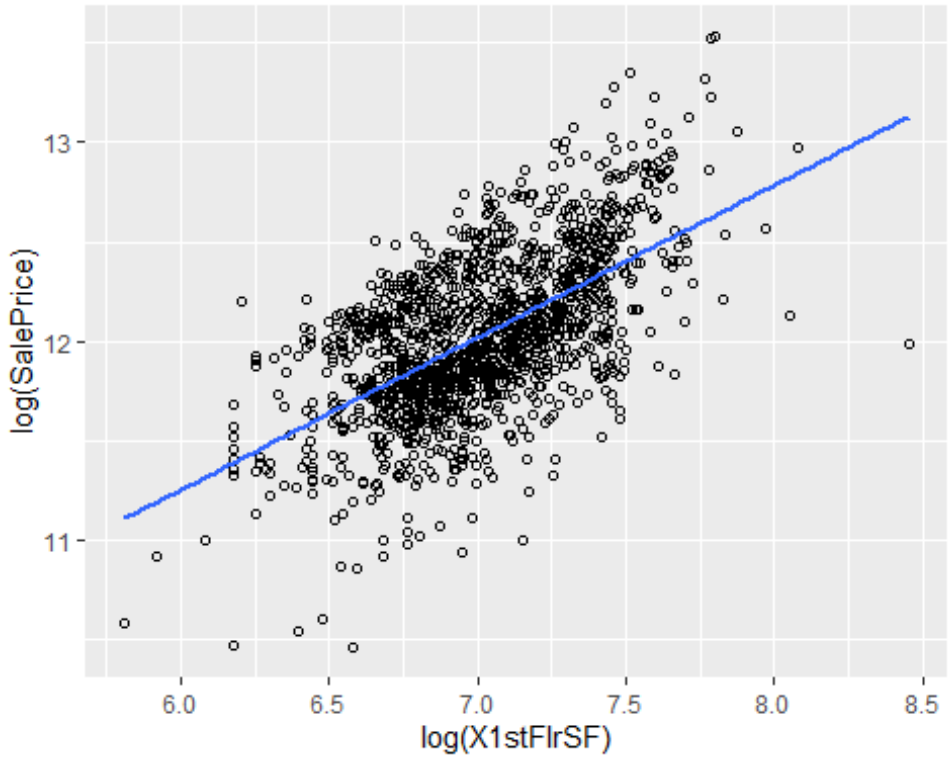

```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$Electrical))

## $devianza_totale
## [1] 9.207731e+12
##
## $devianza_tra_gruppi
## [1] 549453419750
##
## $devianza_entro_gruppi
## [1] 8.658278e+12
##
## $eta2
## [1] 0.05967305
```

Il tipo di sistema elettrico non sembra avere influenza sul prezzo, le mediane dei gruppi sono simili, il gruppo SBrkr cioè il sistema elettrico standard, ha la varianza maggiore, bisogna però notare che è il gruppo con il la maggiore numerosità.

Variabile X1stFlrSF

```
calcolo_cov_cor(case$X1stFlrSF)

##          cov          cor
## 1.860663e+07 6.058522e-01
```

```
lmodel <- lm(data = case, formula = (X1stFlrSF~SalePrice))
summary(lmodel)

##
## Call:
## lm(formula = (X1stFlrSF ~ SalePrice), data = case)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -719.5 -212.0  -18.7  172.6 3591.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.292e+02  2.003e+01   31.41   <2e-16 ***
## SalePrice   2.948e-03  1.014e-04   29.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307.7 on 1458 degrees of freedom
## Multiple R-squared:  0.3671, Adjusted R-squared:  0.3666
## F-statistic: 845.5 on 1 and 1458 DF,  p-value: < 2.2e-16

ggplot(data = case , aes(x=log(X1stFlrSF), y=log(SalePrice))) +
  geom_point(shape=1) + geom_smooth(method = 'lm', se = F)

## `geom_smooth()` using formula = 'y ~ x'
```



abbiamo un coefficiente di correlazione lineare di circa 0.6, dal modello lineare osserviamo un valore di R^2 di circa 0.366.

Variabile X2ndFlrSF

```
calcolo_cov_cor(case$X2ndFlrSF)
```

```
##          cov          cor
## 1.107415e+07 3.193338e-01
```

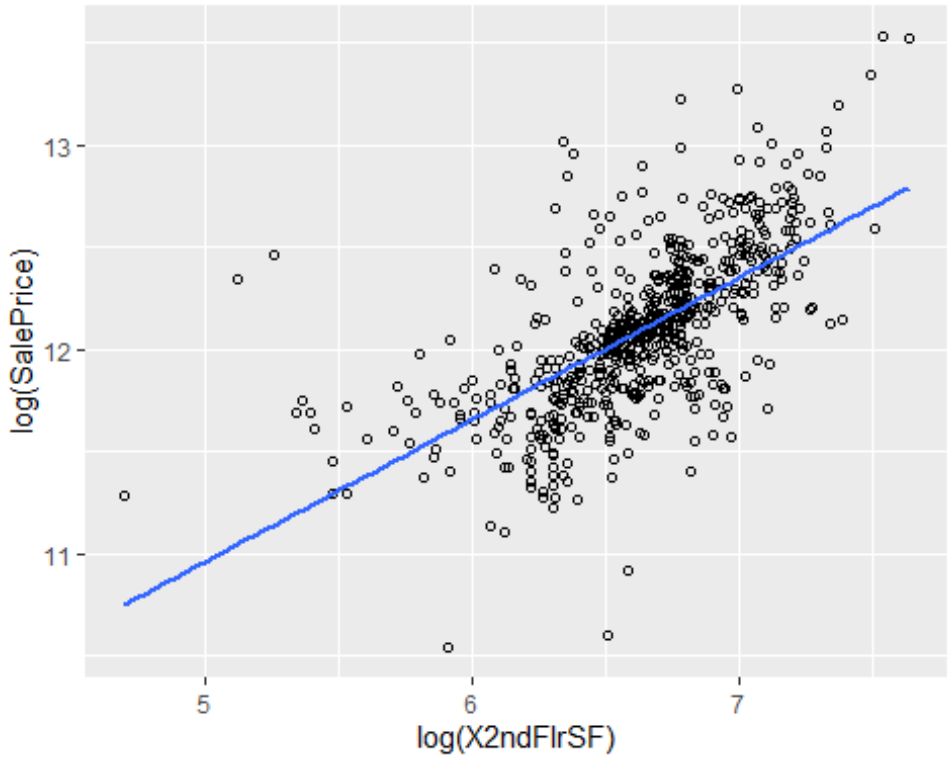
```
lmodel <- lm(data = case, formula = (X1stFlrSF~SalePrice))
summary(lmodel)
```

```
##
## Call:
## lm(formula = (X1stFlrSF ~ SalePrice), data = case)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -719.5 -212.0  -18.7  172.6 3591.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.292e+02  2.003e+01   31.41   <2e-16 ***
## SalePrice   2.948e-03  1.014e-04   29.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307.7 on 1458 degrees of freedom
## Multiple R-squared:  0.3671, Adjusted R-squared:  0.3666
## F-statistic: 845.5 on 1 and 1458 DF,  p-value: < 2.2e-16
```

```
lmodel_case_con_2_piano <- lm(data = subset(case, X2ndFlrSF != 0), formula =
(X2ndFlrSF~SalePrice))
summary(lmodel_case_con_2_piano)
```

```
##
## Call:
## lm(formula = (X2ndFlrSF ~ SalePrice), data = subset(case, X2ndFlrSF !=
##     0))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -797.23  -93.58  -10.77  100.69  817.82
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.820e+02  2.011e+01   19.00   <2e-16 ***
## SalePrice   2.175e-03  9.525e-05   22.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 202.3 on 629 degrees of freedom
## Multiple R-squared:  0.4533, Adjusted R-squared:  0.4525
## F-statistic: 521.6 on 1 and 629 DF,  p-value: < 2.2e-16
```

```
ggplot(data = subset(case, X2ndFlrSF != 0) , aes(x=log(X2ndFlrSF),
y=log(SalePrice))) +
  geom_point(shape=1) + geom_smooth(method = 'lm', se = F)

## `geom_smooth()` using formula = 'y ~ x'
```

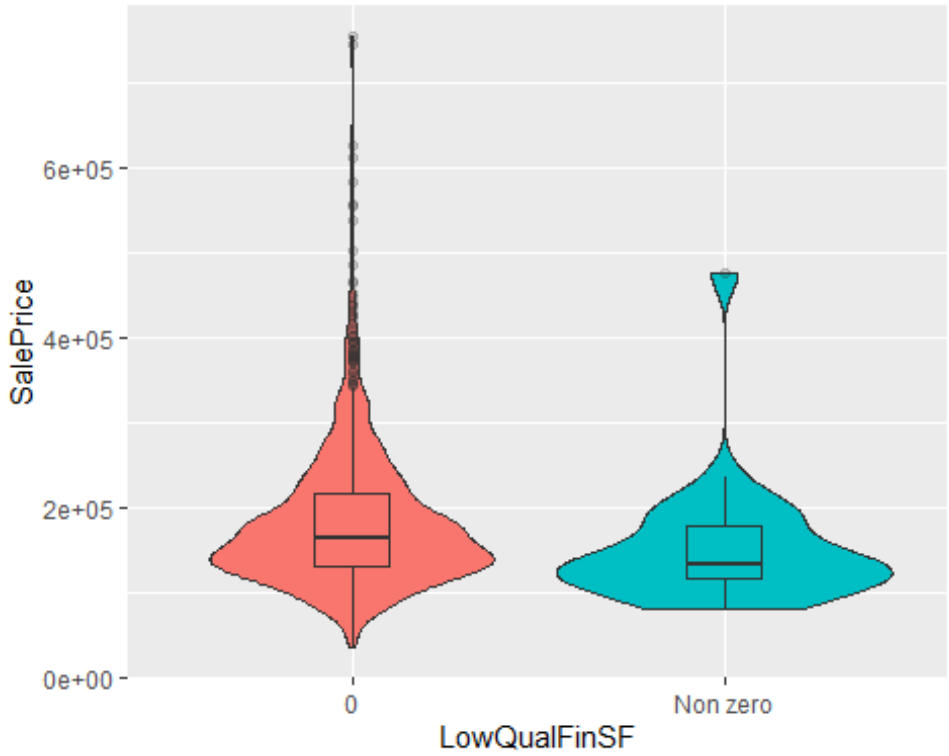


Osserviamo che vi è un discreto numero di case che non ha un secondo piano, plottando i punti vi è una concertazione di punti nell'ascissa 0. Facendo il modello di regressione lineare sia sul set completo che sul subset senza gli zeri si nota che in quest'ultimo il valore di R^2 aumenta quasi del 10%.

Variabile lowqualityfinsf

```
case$LowQualFinSF<- factor(replace(case$LowQualFinSF, (case$LowQualFinSF) > 0,
"Non zero"))
calcola_devianza(case$SalePrice, case$LowQualFinSF)

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 21154558355
##
## $devianza_entro_gruppi
## [1] 9.186757e+12
##
## $eta2
## [1] 0.002297433
```

```
ggplot(case, aes(x = LowQualFinSF, y = SalePrice, fill = as.factor(x =
LowQualFinSF))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = F)
```



La maggior parte dei valori è 0, divido in due gruppi uno in cui il valore è diverso da zero e uno in cui è uguale. Dal grafico e dell'analisi della devianza si osserva che le mediane dei due gruppi sono molto vivine tra loro. IL coefficiente eta^2 è quasi zero.

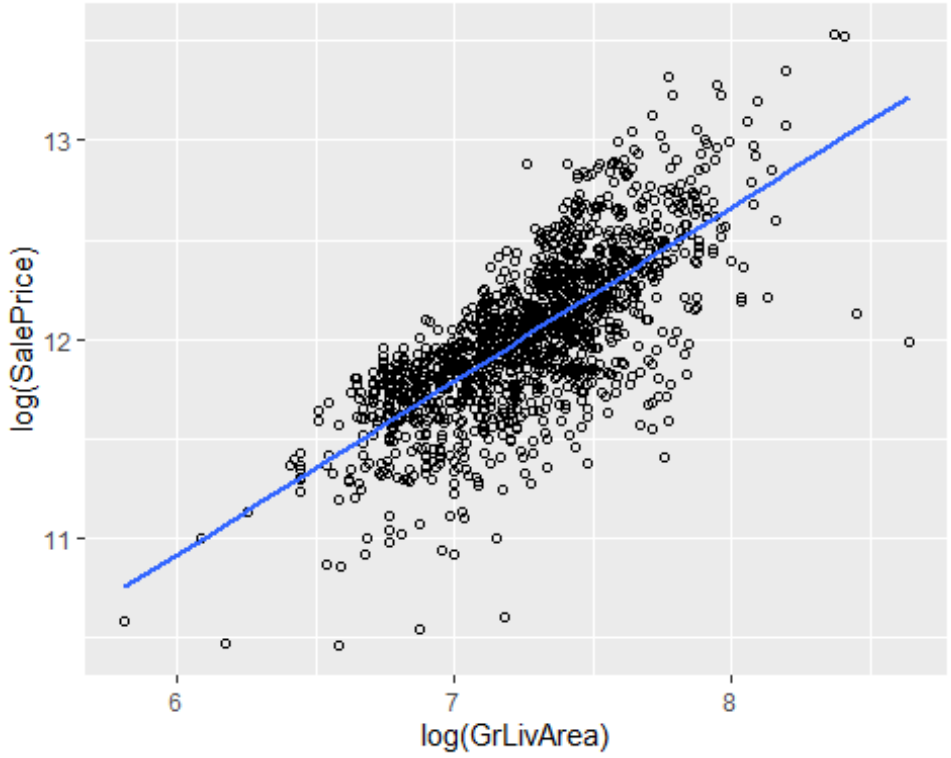
Variabile GrLivArea

```
calcolo_cov_cor(case$GrLivArea)

##          cov          cor
## 2.958187e+07 7.086245e-01

lmodel <- lm(data = case, formula = (GrLivArea~SalePrice))
summary(lmodel)

##
## Call:
## lm(formula = (GrLivArea ~ SalePrice), data = case)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1170.4  -255.2   -52.7   187.5  4224.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.674e+02  2.415e+01   27.64   <2e-16 ***
```

```
## SalePrice   4.687e-03  1.222e-04   38.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 370.9 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
```

```
ggplot(data = case , aes(x=log(GrLivArea), y=log(SalePrice))) +
  geom_point(shape=1) + geom_smooth(method = 'lm', se = F)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

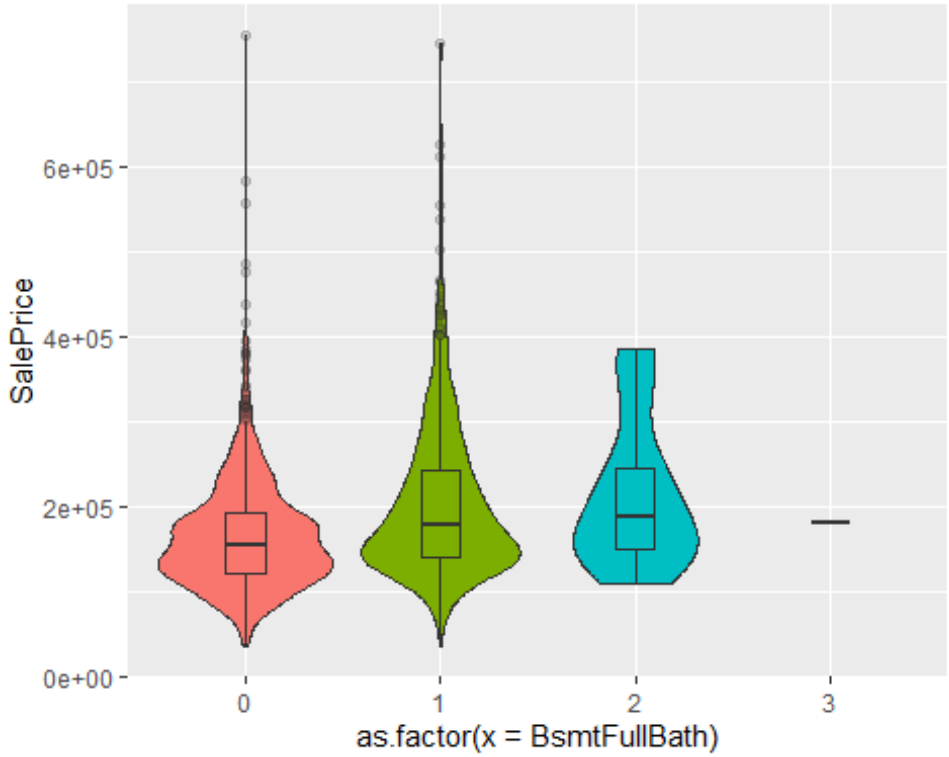


R^2 vale 0.50 e il coefficiente di correlazione lineare 0.70 dunque vi è una forte correlazione lineare tra le variabili.

Variabile BsmtFullBath

```
ggplot(case, aes(x = as.factor(x = BsmtFullBath), y = SalePrice, fill =
as.factor(x = BsmtFullBath))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)
```

```
## Warning: Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
```



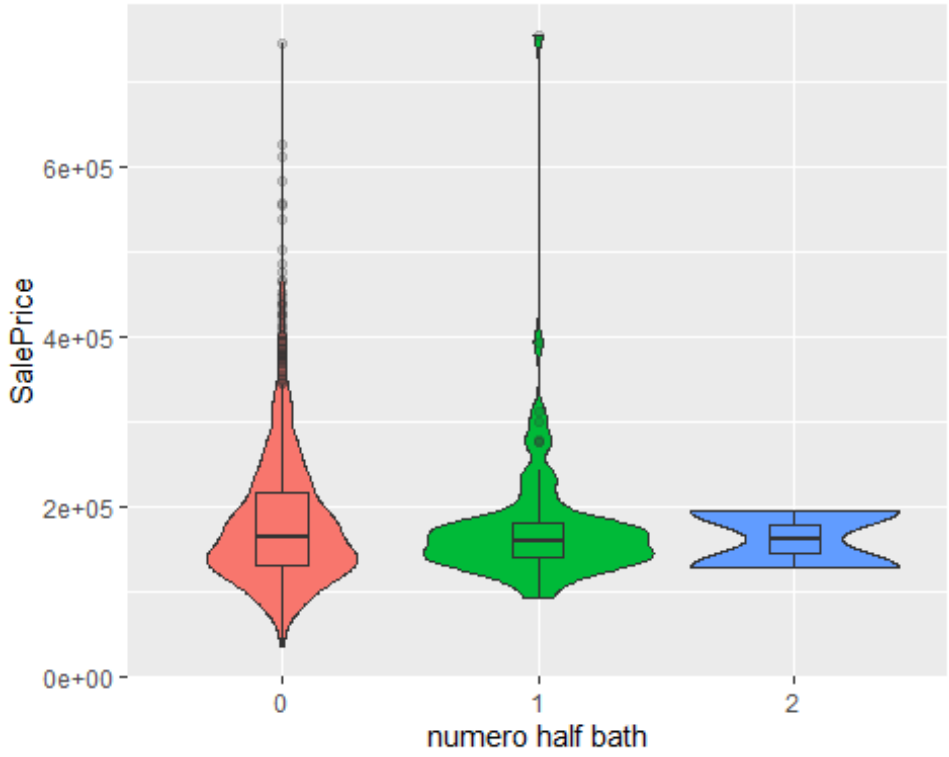

```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$BsmtFullBath))
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 492878470644
##
## $devianza_entro_gruppi
## [1] 8.715033e+12
##
## $eta2
## [1] 0.05352772
```

La variabile rappresenta i Basement full Bathrooms. La devianza tra i gruppi è bassa, il coefficiente eta^2 vale circa 0.053. Le mediane dei gruppi sono simili. Dal grafico si nota che le case con 0 e 1 bagno hanno una varianza maggiore.

Variabile BsmtHalfBath

```
ggplot(case, aes(x = as.factor(x = BsmtHalfBath), y = SalePrice, fill =
as.factor(x = BsmtHalfBath))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) +
labs(x = 'numero half bath')
```

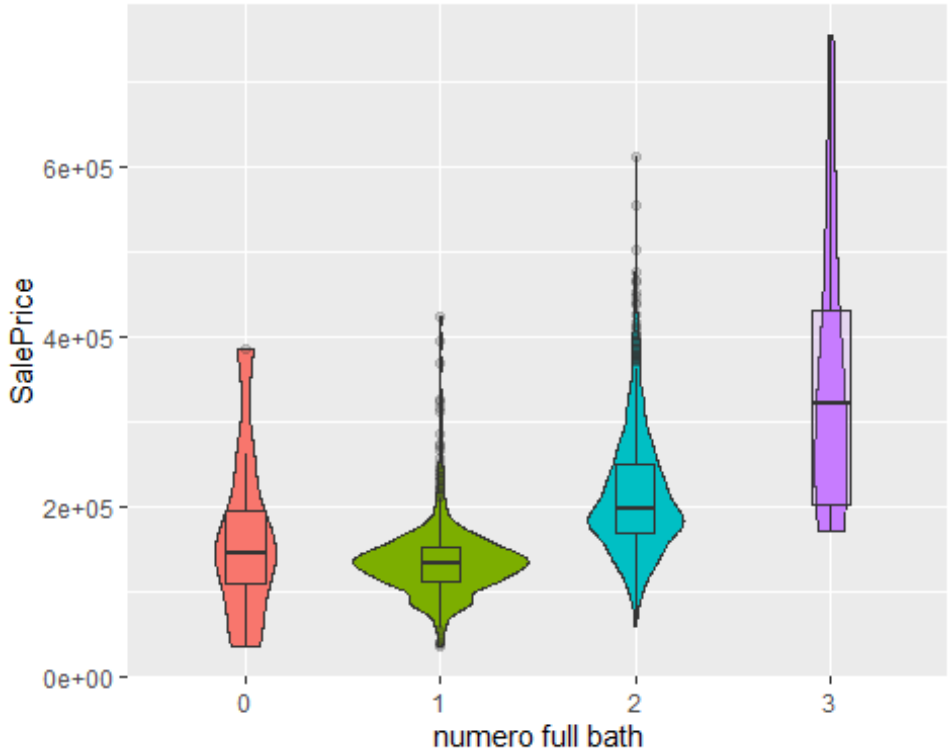



```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$BsmtHalfBath))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 2798313904
##
## $devianza_entro_gruppi
## [1] 9.205113e+12
##
## $eta2
## [1] 0.0003039032
```

La variabile rappresenta i Basement half Bathrooms. La devianza tra i gruppi è bassa, eta^2 è quasi 0. Le mediane dei gruppi sono simili. Dal grafico si nota che le case con 0 e 1 bagno hanno una varianza maggiore.

Variabile FullBath

```
ggplot(case, aes(x = as.factor(x = FullBath), y = SalePrice, fill = as.factor(x =
FullBath))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) +
labs(x = 'numero full bath')
```

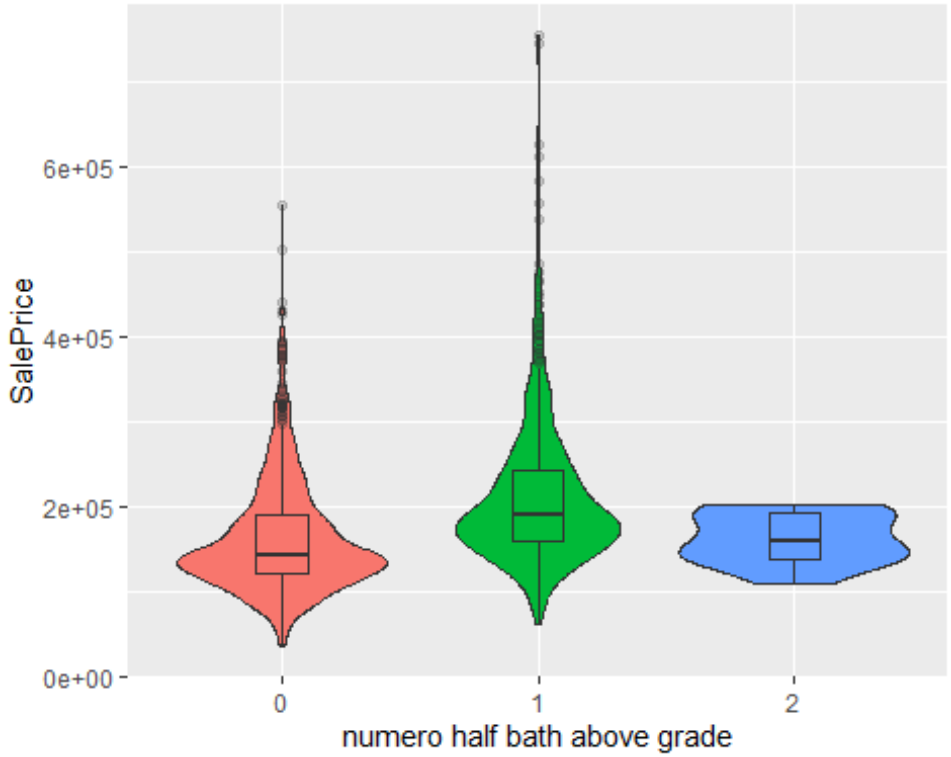



```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$FullBath))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 3.097843e+12
##
## $devianza_entro_gruppi
## [1] 6.110069e+12
##
## $eta2
## [1] 0.3364327
```

La variabile rappresenta i Full bathrooms above grade. La devianza tra i gruppi è alta, eta^2 vale circa 0.34. Dal grafico si nota che il prezzo è influenzato dal numero di full bath above grade, anche la devianza entro i gruppi è alta.

Variabile HalfBath

```
ggplot(case, aes(x = as.factor(x = HalfBath), y = SalePrice, fill = as.factor(x =
HalfBath))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE) +
labs(x = 'numero half bath above grade')
```

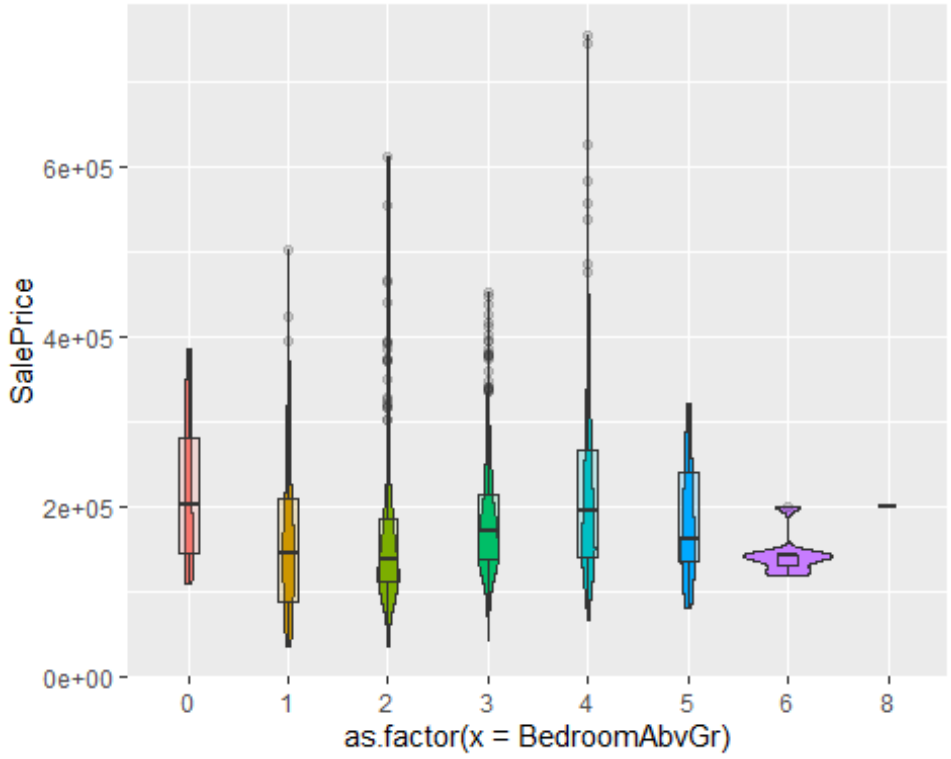



```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$HalfBath))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 8.53968e+11
##
## $devianza_entro_gruppi
## [1] 8.353943e+12
##
## $eta2
## [1] 0.09274285
```

La variabile rappresenta gli half Bathrooms above grade. La devianza entro i gruppi è maggiore della devianza tra i gruppi, il coefficiente eta^2 è circa 0.0927. Le case nel gruppo 1 sono quelle che presentano una maggiore varianza.

Variabile BedroomAbvGr

```
ggplot(case, aes(x = as.factor(x = BedroomAbvGr), y = SalePrice, fill =
as.factor(x = BedroomAbvGr))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)

## Warning: Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
```

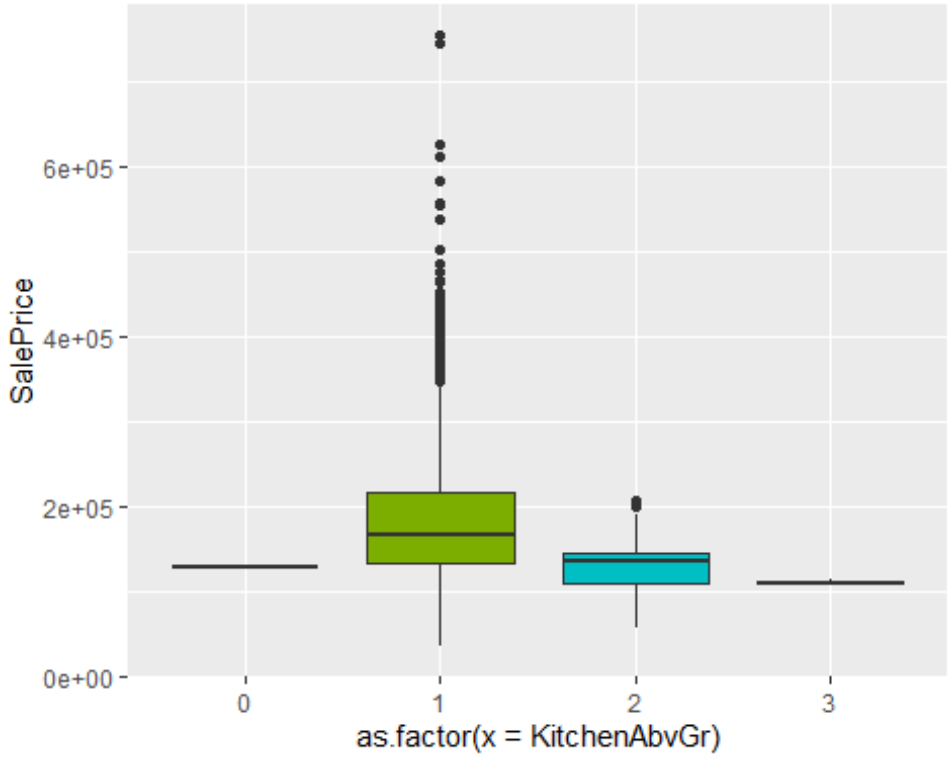



```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$BedroomAbvGr))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 5.40113e+11
##
## $devianza_entro_gruppi
## [1] 8.667798e+12
##
## $eta2
## [1] 0.0586575
```

Le case con 4 camere da letto sono quelle che presentano la maggiore varianza. La devianza entro i gruppi è maggiore di quella tra i gruppi. il coefficienta eta^2 è circa 0.5866.

Variabile KitchenAbvGr

```
ggplot(case, aes(x = as.factor(x = KitchenAbvGr), y = SalePrice, fill =
as.factor(x = KitchenAbvGr))) +
  geom_boxplot() + guides(fill = FALSE)
```

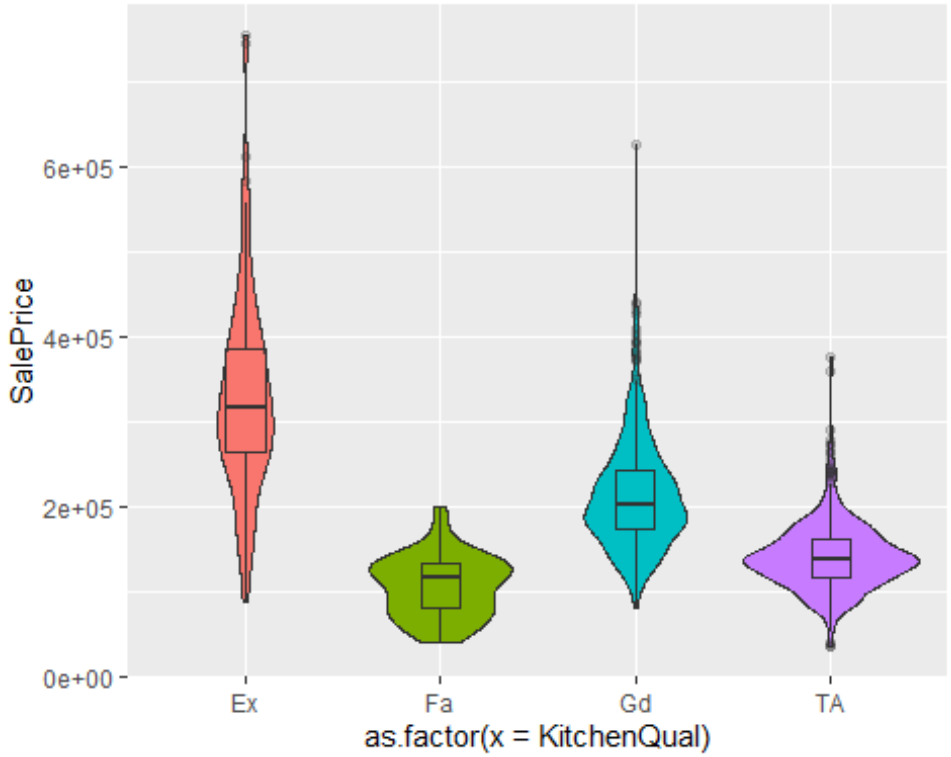



```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$KitchenAbvGr))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 182896479951
##
## $devianza_entro_gruppi
## [1] 9.025015e+12
##
## $eta2
## [1] 0.01986297
```

la devianza tra i gruppi è bassa, il coefficiente eta^2 vale circa 0.0199. La correlazione è bassa.

Variabile KitchenQual

```
ggplot(case, aes(x = as.factor(x = KitchenQual), y = SalePrice, fill = as.factor(x
= KitchenQual))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)
```




```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$KitchenQual))
```

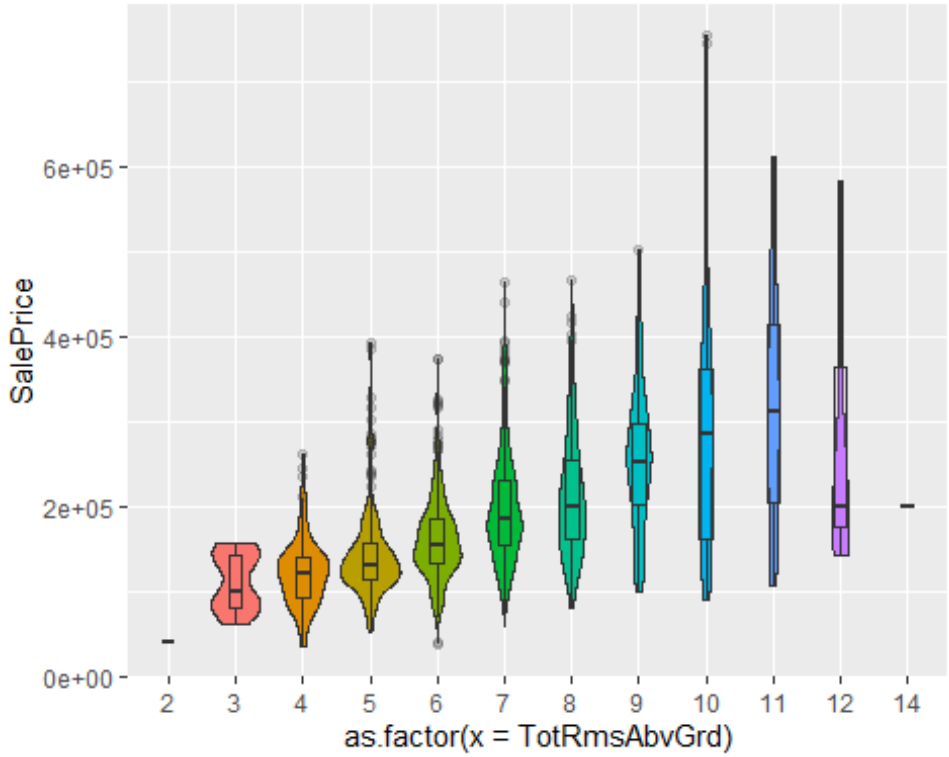
```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] NaN
##
## $devianza_entro_gruppi
## [1] 5.003592e+12
##
## $eta2
## [1] NaN
```

Il coefficiente eta^2 vale circa 0.457 e dal grafico si nota una correlazione tra il prezzo della casa e la qualità della cucina. Notiamo una maggiore varianza nelle case che hanno una cucina eccellente.

Variabile TotRmsAbvGrd

```
ggplot(case, aes(x = as.factor(x = TotRmsAbvGrd), y = SalePrice, fill =
as.factor(x = TotRmsAbvGrd))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)
```

```
## Warning: Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
## Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
```




```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$TotRmsAbvGrd))
```

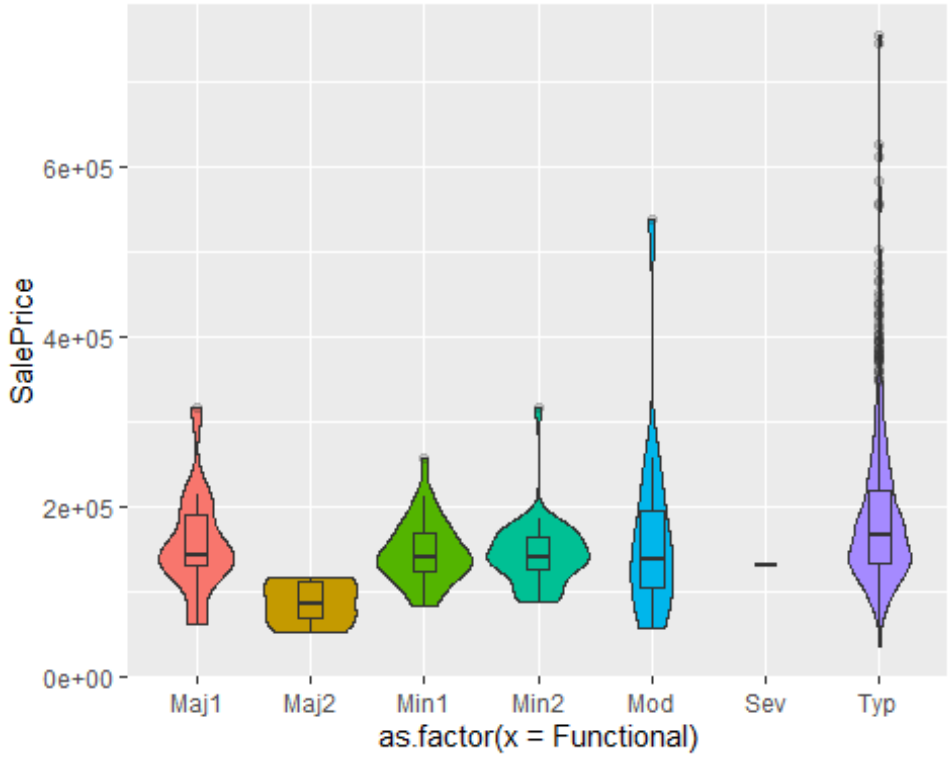
```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 2.753747e+12
##
## $devianza_entro_gruppi
## [1] 6.454164e+12
##
## $eta2
## [1] 0.2990631
```

il coefficiente eta^2 vale 0.2991 graficamente osserviamo una chiara correlazione tra il numero delle stanza above grade e il prezzo.

Variabile Functional

```
ggplot(case, aes(x = as.factor(x = Functional), y = SalePrice, fill = as.factor(x
= Functional))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)
```

```
## Warning: Groups with fewer than two datapoints have been dropped.
## ℹ Set `drop = FALSE` to consider such groups for position adjustment purposes.
```

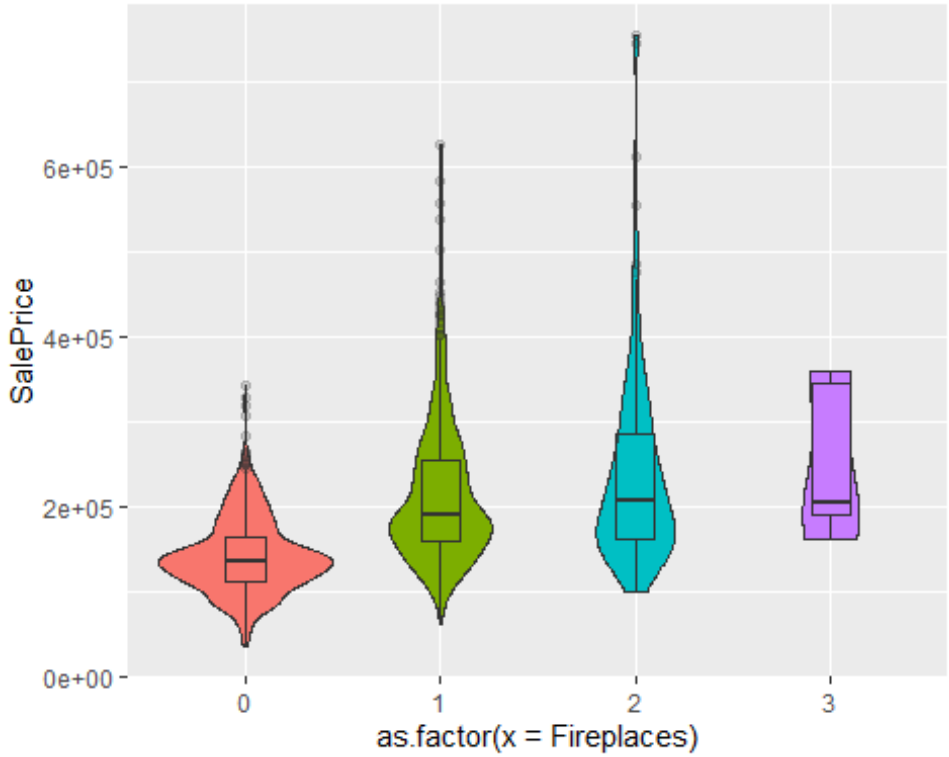



```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$Functional))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 151749930483
##
## $devianza_entro_gruppi
## [1] 9.056161e+12
##
## $eta2
## [1] 0.01648039
```

Il coefficiente eta^2 vale circa 0.165 e la devianza tra i gruppi e molto più bassa della devianza entro i gruppi. Non sembra esserci una correlazione tra l'appartenere ad un gruppo e il prezzo.

Variabile Fireplaces

```
ggplot(case, aes(x = as.factor(x = Fireplaces), y = SalePrice, fill = as.factor(x
= Fireplaces))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)
```

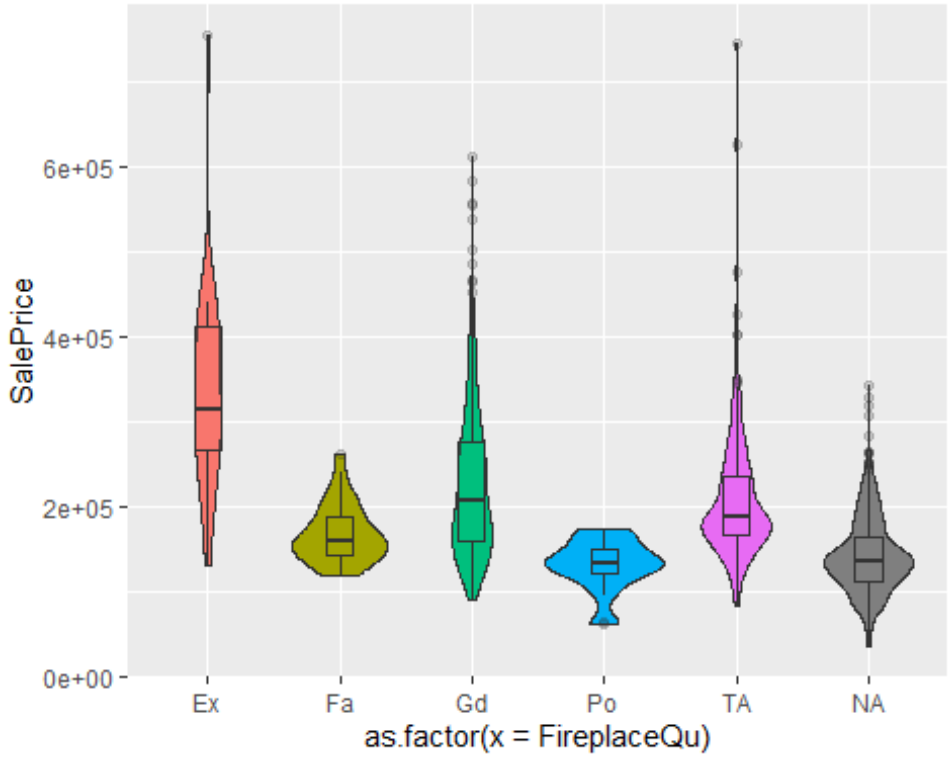



```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$Fireplaces))

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 2.137691e+12
##
## $devianza_entro_gruppi
## [1] 7.070221e+12
##
## $eta2
## [1] 0.232158
```

Il coefficiente eta^2 vale 0.2322 la devianza entro i gruppi e la devianza tra i gruppi hhanno lo stesso ordine di grandezza 10^12, sembra esserci correlazione tra le variabili.

Variabile FireplaceQu

```
ggplot(case, aes(x = as.factor(x = FireplaceQu), y = SalePrice, fill = as.factor(x
= FireplaceQu))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)
```

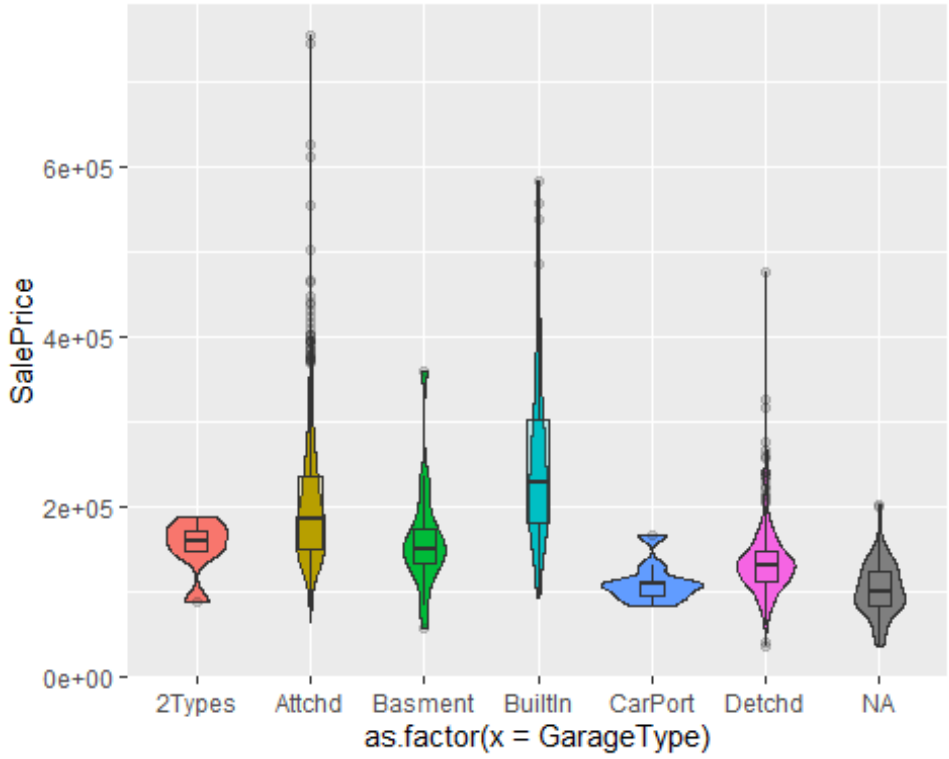



```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$FireplaceQu))

## $devianza_totale
## [1] 5.799693e+12
##
## $devianza_tra_gruppi
## [1] 656188256681
##
## $devianza_entro_gruppi
## [1] 5.143505e+12
##
## $eta2
## [1] 0.1131419
```

Il coefficiente eta^2 è circa 0.113, e la devianza tra i gruppi è minore di quella entro i gruppi.

Variabile GarageType

```
ggplot(case, aes(x = as.factor(x = GarageType), y = SalePrice, fill = as.factor(x
= GarageType))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)
```




```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$GarageType))
```

```
## $devianza_totale
## [1] 8.6053e+12
##
## $devianza_tra_gruppi
## [1] 1.778186e+12
##
## $devianza_entro_gruppi
## [1] 6.827115e+12
##
## $eta2
## [1] 0.2066384
```

Il coefficiente eta^2 è circa 0.2066, sembra esserci correlazione tra le due variabili. IL gruppo che presenta la maggiore varianza è il gruppo delle case con il garage di tipo Attached , si ricordi che questas il gruppo con la numerosità maggiore.

Variabile GarageYrBlt

```
cor(case$GarageYrBlt, case$SalePrice, use="complete.obs")
```

```
## [1] 0.4863617
```

```
lmodel <- lm(data = case, formula = (GarageYrBlt~SalePrice))
summary(lmodel)
```

```
##
## Call:
```

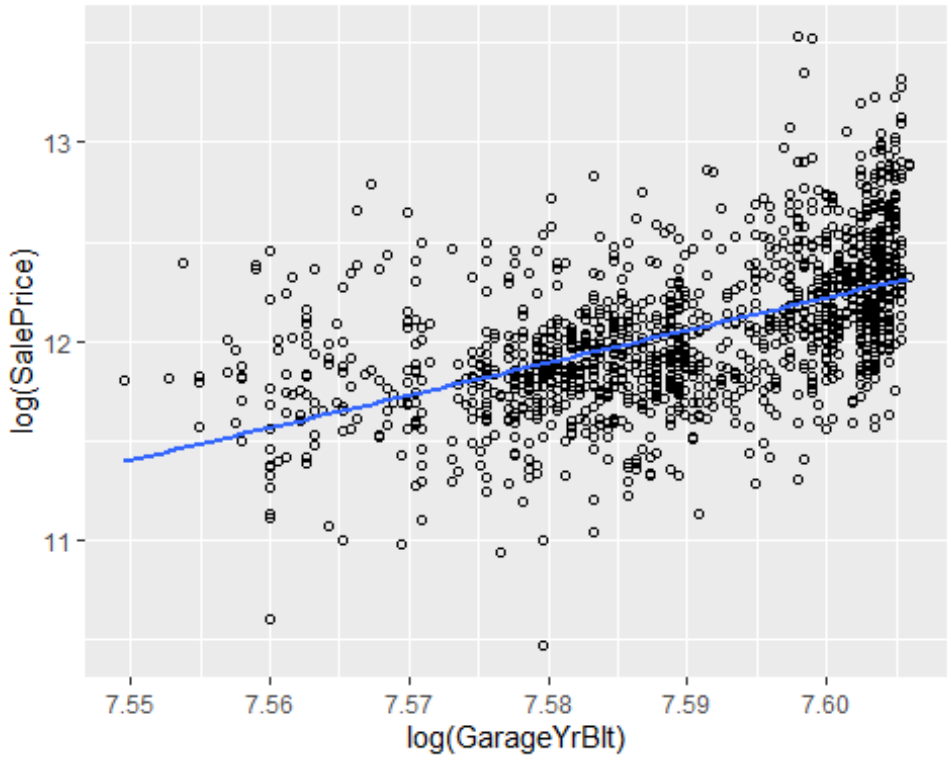
```
## lm(formula = (GarageYrBlt ~ SalePrice), data = case)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -78.791 -12.802   2.164  16.943  39.659
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.950e+03  1.483e+00 1315.02   <2e-16 ***
## SalePrice   1.520e-04  7.357e-06   20.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.58 on 1377 degrees of freedom
##   (81 osservazioni eliminate a causa di valori mancanti)
## Multiple R-squared:  0.2365, Adjusted R-squared:  0.236
## F-statistic: 426.6 on 1 and 1377 DF,  p-value: < 2.2e-16
```

```
ggplot(data = case , aes(x=log(GarageYrBlt), y=log(SalePrice))) +
  geom_point(shape=1) + geom_smooth(method = 'lm', se = F)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 81 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

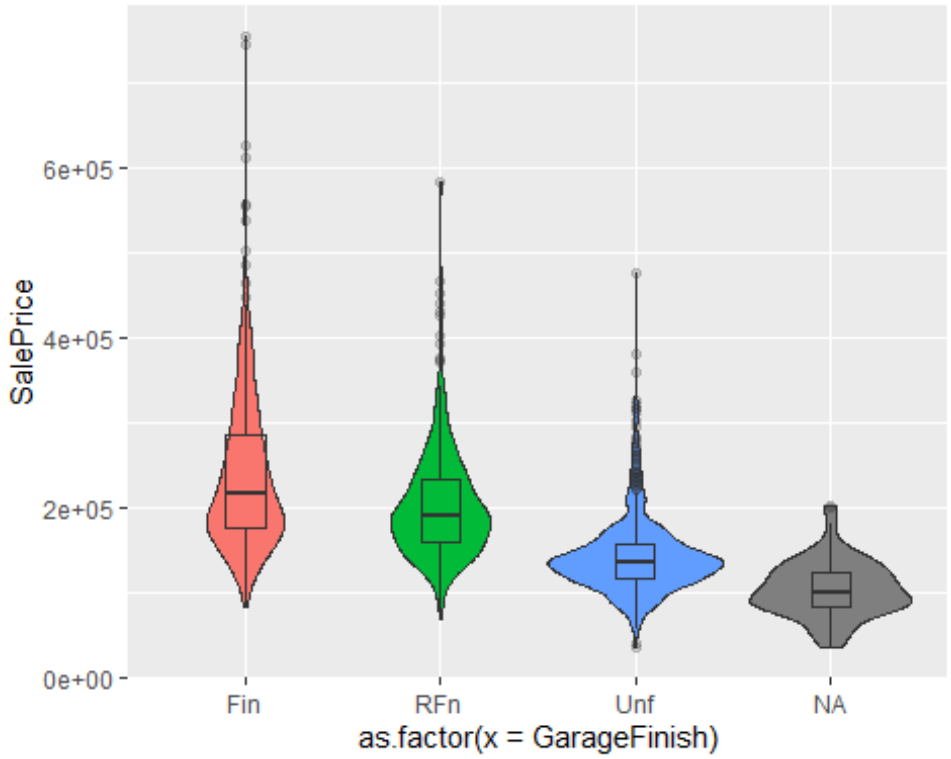
```
## Warning: Removed 81 rows containing missing values or values outside the scale
range
## (`geom_point()`).
```



il coefficiente di correlazione lineare vale 0.5, dal modello di regressione lineare osserviamo un valore di R^2 pari a 0.23.

Variabile GarageFinish

```
ggplot(case, aes(x = as.factor(x = GarageFinish), y = SalePrice, fill =
as.factor(x = GarageFinish))) +
  geom_violin() + geom_boxplot(width=0.2, alpha=1/5) + guides(fill = FALSE)
```




```
calcola_devianza(numerical_var = case$SalePrice, categorical_var =
as.factor(case$GarageFinish))
```

```
## $devianza_totale
## [1] 8.6053e+12
##
## $devianza_tra_gruppi
## [1] NaN
##
## $devianza_entro_gruppi
## [1] 6.305307e+12
##
## $eta2
## [1] NaN
```

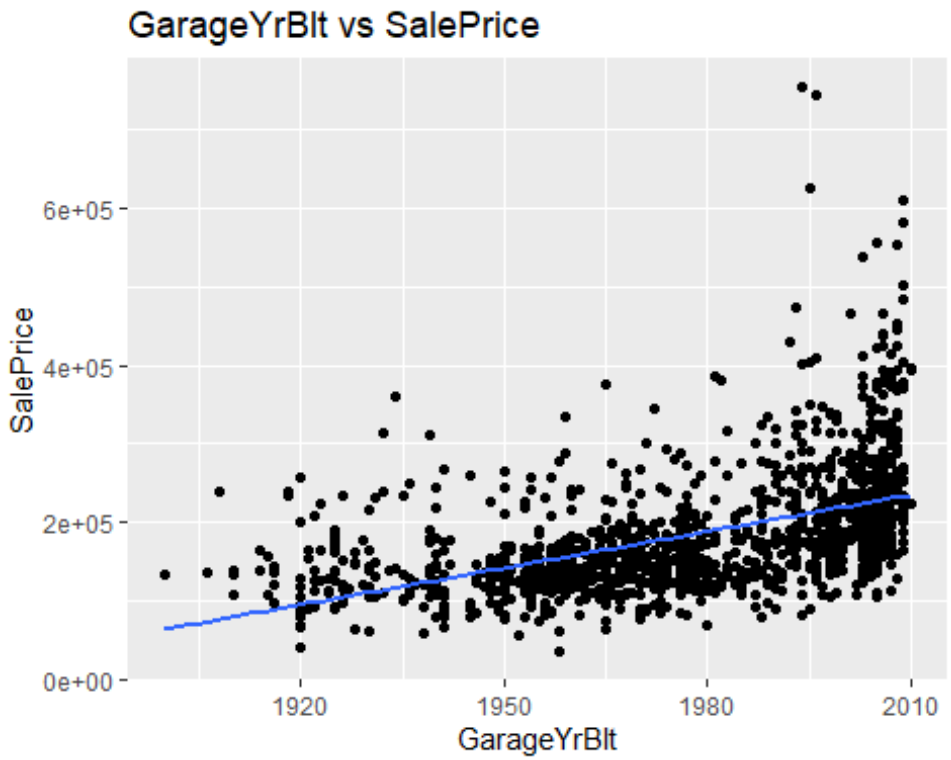
il valore di eta^2 è circa 0.3, sembra esserci correlazione tra le variabili. Nel gruppo di case che hanno il garage ultimato si riscontra lamaggiore varianza.

Variabile 'GarageYrBlt'

```
risultati_cov_cor_GarageYrBlt <- calcolo_cov_cor(case$GarageYrBlt)
risultati_cov_cor_GarageYrBlt
```

```
##          cov          cor
## 9.489296e+05 4.863617e-01
```

```
ggplot(case, aes(x = GarageYrBlt, y = SalePrice)) + geom_point() +
geom_smooth(method = "lm", se = FALSE) + ggtitle("GarageYrBlt vs SalePrice")
```



La variabile GarageYrBlt ha una covarianza di 948929.6 e una correlazione di 0.4864 con SalePrice. Questo indica una relazione positiva tra GarageYrBlt e SalePrice. Questo potrebbe suggerire che i garage costruiti più recentemente tendono ad essere associati a case con un prezzo di vendita più alto, forse perché indicano una manutenzione più recente e tecnologie più moderne.

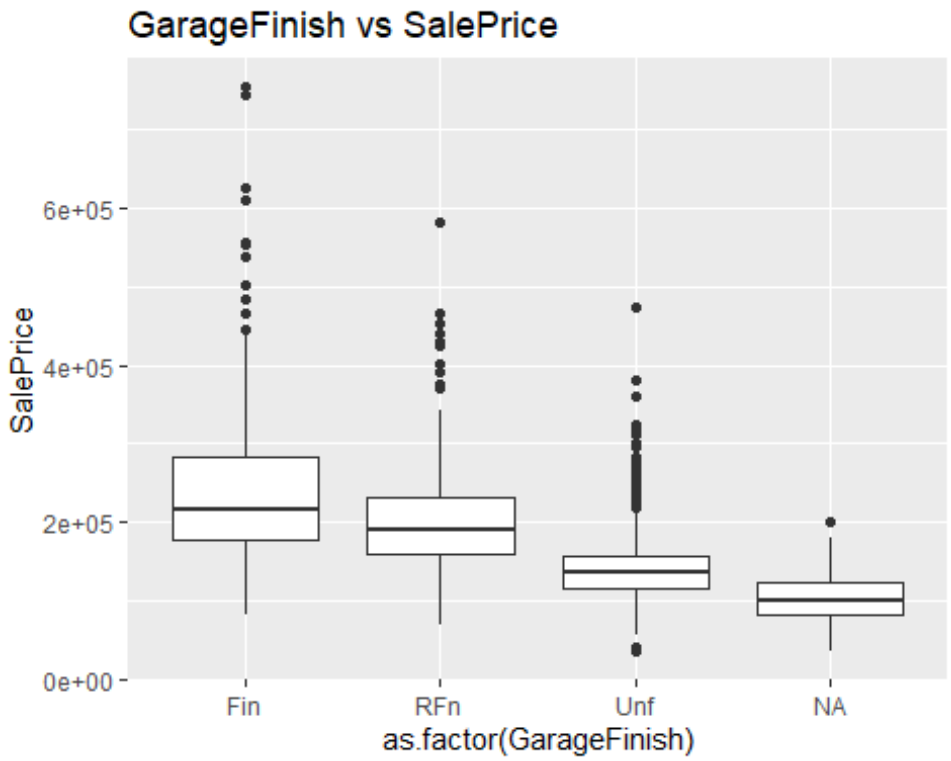
Variabile 'GarageFinish'

```
risultati_devianza_GarageFinish <- calcola_devianza(case$SalePrice,
case$GarageFinish)
print(risultati_devianza_GarageFinish)
```

```
## $devianza_totale
## [1] 8.6053e+12
##
## $devianza_tra_gruppi
## [1] 2.299993e+12
## $devianza_entro_gruppi
```

```
## [1] 6.305307e+12
##
## $eta2
## [1] 0.2672764
```

```
ggplot(case, aes(x = as.factor(GarageFinish), y = SalePrice)) + geom_boxplot() +
ggtitle("GarageFinish vs SalePrice")
```



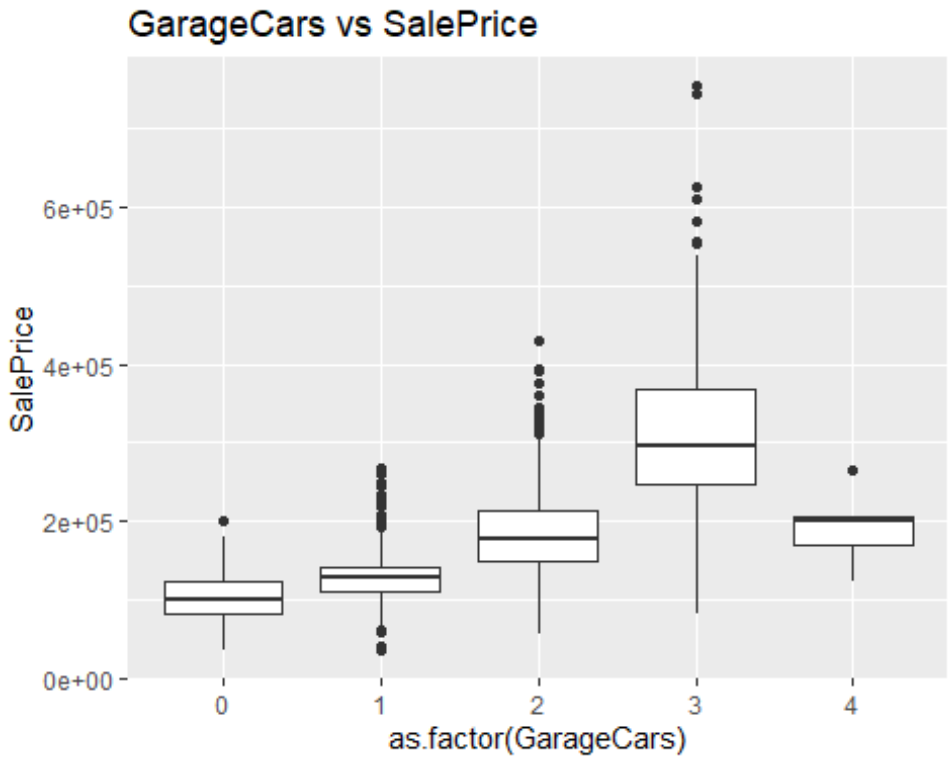
La variabile GarageFinish ha una devianza totale di 8.6053e+12 , con una devianza tra gruppi di 2.299993e+12 e una devianza entro gruppi di 6.305307e+12 . Eta² è 0.2672764 indicando che il 0.2672764 % della varianza di SalePrice è spiegata da GarageFinish. Un garage finito (o meglio rifinito) potrebbe essere percepito come un segno di qualità e cura della proprietà, contribuendo quindi ad un prezzo di vendita più alto.

Variabile 'GarageCars'

```
risultati_cov_cor_GarageCars <- calcolo_cov_cor(case$GarageCars)
print(risultati_cov_cor_GarageCars)
```

```
##          cov          cor
## 3.802018e+04 6.404092e-01
```

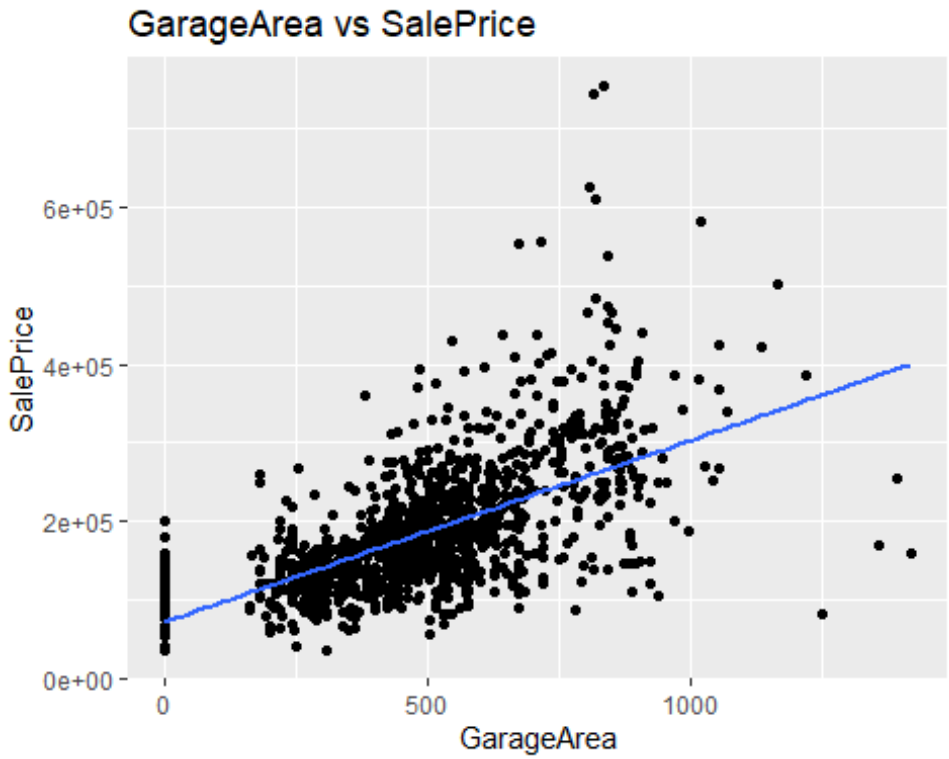
```
# Plot
ggplot(case, aes(x = as.factor(GarageCars), y = SalePrice)) + geom_boxplot() +
ggtitle("GarageCars vs SalePrice")
```



La variabile GarageCars ha una covarianza di 38020.18 e una correlazione di 0.6404 con SalePrice. Questo indica una relazione positiva tra GarageCars e SalePrice. Questo potrebbe suggerire che case con un maggior numero di garage tendono ad avere un prezzo di vendita più alto, forse perché indicano una maggiore capacità di parcheggio e spazio di stoccaggio.

Variabile 'GarageArea'

```
risultati_cov_cor_GarageArea <- calcolo_cov_cor(case$GarageArea)
print(risultati_cov_cor_GarageArea)

##          cov          cor
## 1.058910e+07 6.234314e-01

# Plot
ggplot(case, aes(x = GarageArea, y = SalePrice)) + geom_point() +
geom_smooth(method = "lm", se = FALSE) + ggtitle("GarageArea vs SalePrice")

## `geom_smooth()` using formula = 'y ~ x'
```



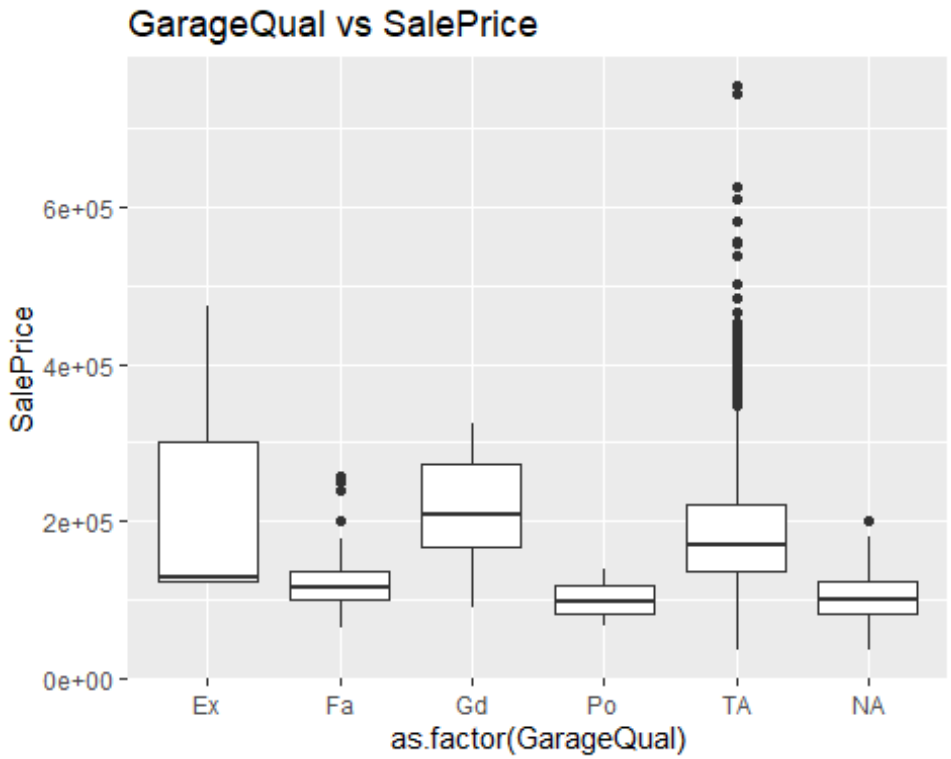
La variabile GarageArea ha una covarianza di 10589103 e una correlazione di 0.6234 con SalePrice. Questo indica una relazione positiva tra GarageArea e SalePrice. Questo potrebbe suggerire che case con un'area garage più grande tendono ad avere un prezzo di vendita più alto, forse perché indicano una maggiore capacità di parcheggio e spazio di stoccaggio.

Variabile 'GarageQual'

```
risultati_devianza_GarageQual <- calcola_devianza(case$SalePrice, case$GarageQual)
print(risultati_devianza_GarageQual)
```

```
## $devianza_totale
## [1] 8.6053e+12
##
## $devianza_tra_gruppi
## [1] 233256815766
##
## $devianza_entro_gruppi
## [1] 8.372044e+12
##
## $eta2
## [1] 0.02710618
```

```
# Plot
ggplot(case, aes(x = as.factor(GarageQual), y = SalePrice)) + geom_boxplot() +
ggtitle("GarageQual vs SalePrice")
```



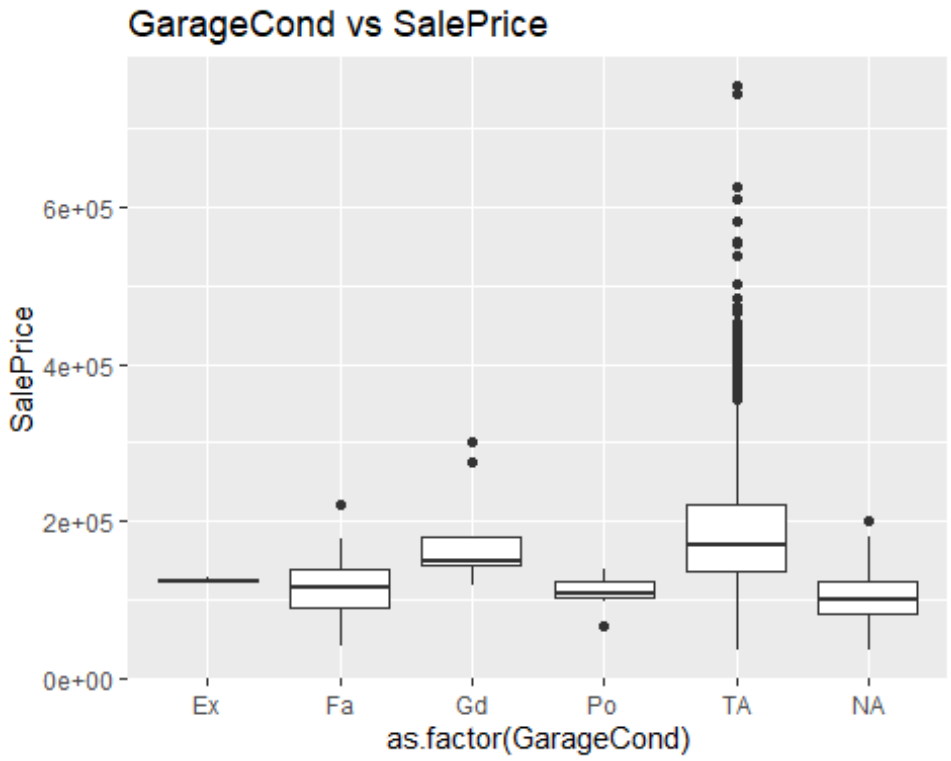
La variabile GarageQual ha una devianza totale di 8.6053e+12 , con una devianza tra gruppi di 233256815766 e una devianza entro gruppi di 8.372044e+12 . Eta² è 0.0271 indicando che il 2.71 % della varianza di SalePrice è spiegata da GarageQual. Un garage di qualità superiore potrebbe essere percepito come un segno di qualità e cura della proprietà, contribuendo quindi ad un prezzo di vendita più alto.

Variabile 'GarageCond'

```
risultati_devianza_GarageCond <- calcola_devianza(case$SalePrice, case$GarageCond)
print(risultati_devianza_GarageCond)
```

```
## $devianza_totale
## [1] 8.6053e+12
##
## $devianza_tra_gruppi
## [1] 232563691396
##
## $devianza_entro_gruppi
## [1] 8.372737e+12
##
## $eta2
## [1] 0.02702563
```

```
# Plot
ggplot(case, aes(x = as.factor(GarageCond), y = SalePrice)) + geom_boxplot() +
ggtitle("GarageCond vs SalePrice")
```



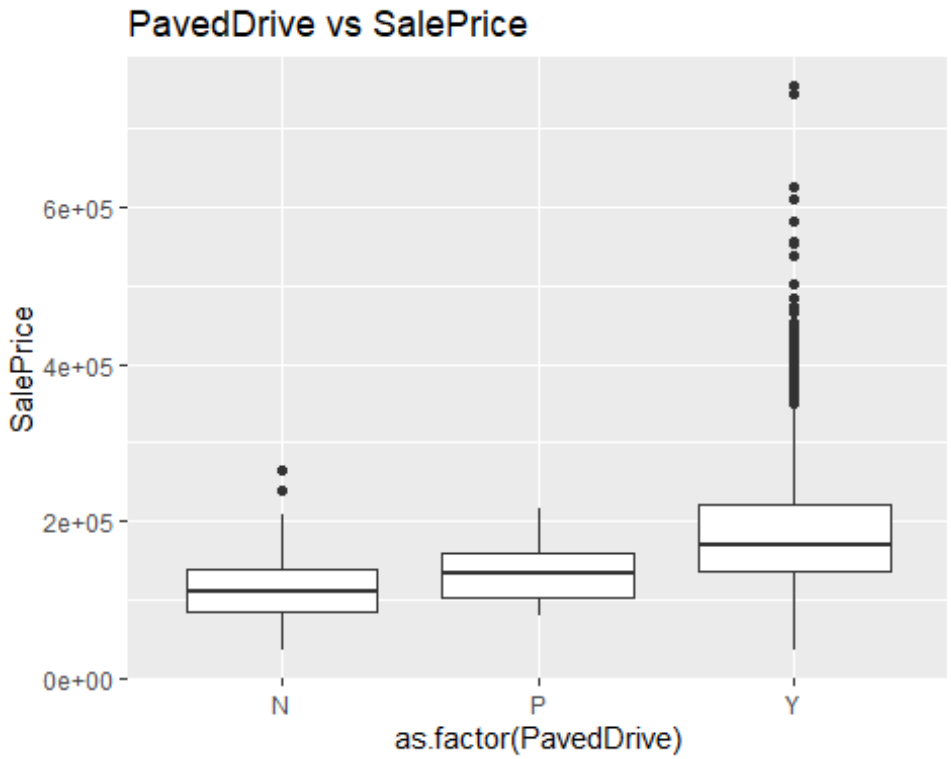
La variabile GarageCond ha una devianza totale di 8.6053e+12 , con una devianza tra gruppi di 232563691396 e una devianza entro gruppi di 8.372737e+12 . Eta² è 0.027 indicando che il 2.7 % della varianza di SalePrice è spiegata da GarageCond. Una buona condizione del garage è cruciale per la funzionalità e l'estetica della proprietà, influenzando positivamente il prezzo di vendita.

Variabile 'PavedDrive'

```
risultati_devianza_PavedDrive <- calcola_devianza(case$SalePrice, case$PavedDrive)
print(risultati_devianza_PavedDrive)
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 5.02197e+11
##
## $devianza_entro_gruppi
## [1] 8.705714e+12
##
## $eta2
## [1] 0.05453973
```

```
# Plot
ggplot(case, aes(x = as.factor(PavedDrive), y = SalePrice)) + geom_boxplot() +
ggtitle("PavedDrive vs SalePrice")
```



La variabile PavedDrive ha una devianza totale di 9.207911e+12 , con una devianza tra gruppi di 5.02197e+11 e una devianza entro gruppi di 8.705714e+12 . Eta² è 0.0545 indicando che il 5.45 % della varianza di SalePrice è spiegata da PavedDrive. Un vialetto pavimentato può migliorare l'aspetto estetico della casa e la comodità, influenzando così positivamente il prezzo di vendita.

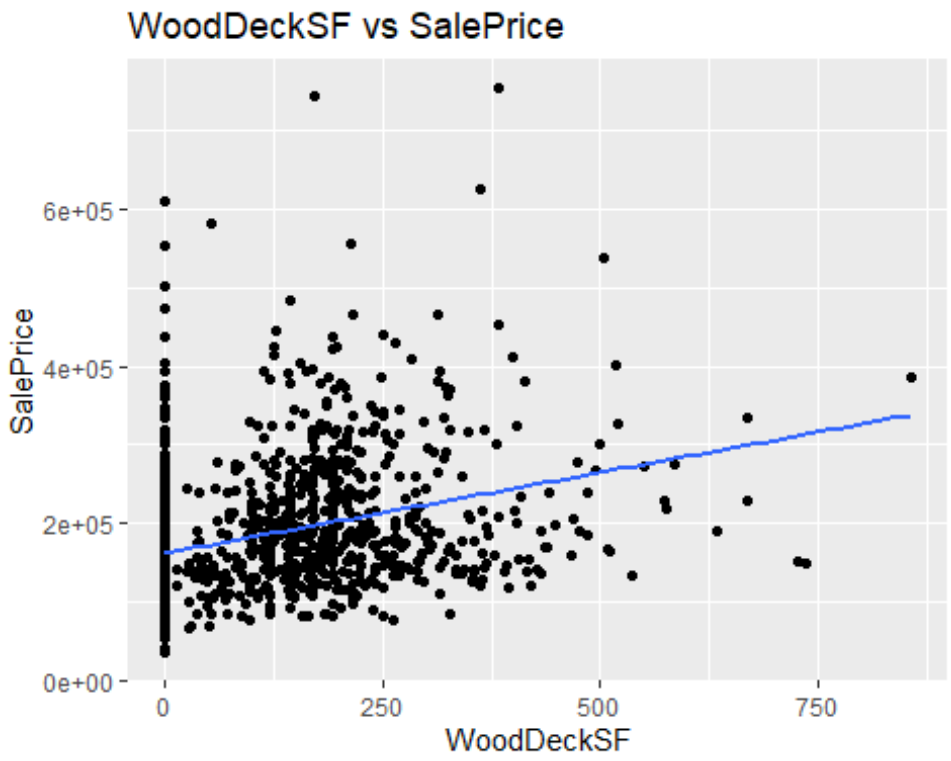
Variabile 'WoodDeckSF'

```
risultati_cov_cor_WoodDeckSF <- calcolo_cov_cor(case$WoodDeckSF)
print(risultati_cov_cor_WoodDeckSF)
```

```
##          cov          cor
## 3.230258e+06 3.244134e-01
```

```
# Plot
ggplot(case, aes(x = WoodDeckSF, y = SalePrice)) + geom_point() +
geom_smooth(method = "lm", se = FALSE) + ggtitle("WoodDeckSF vs SalePrice")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



La variabile WoodDeckSF ha una covarianza di 3230258 e una correlazione di 0.3244 con SalePrice. Questo indica una relazione positiva tra WoodDeckSF e SalePrice. La presenza di un'ampia terrazza in legno può essere un fattore attrattivo per gli acquirenti, offrendo spazio all'aperto per il relax e l'intrattenimento, e quindi aumentando il valore della casa.

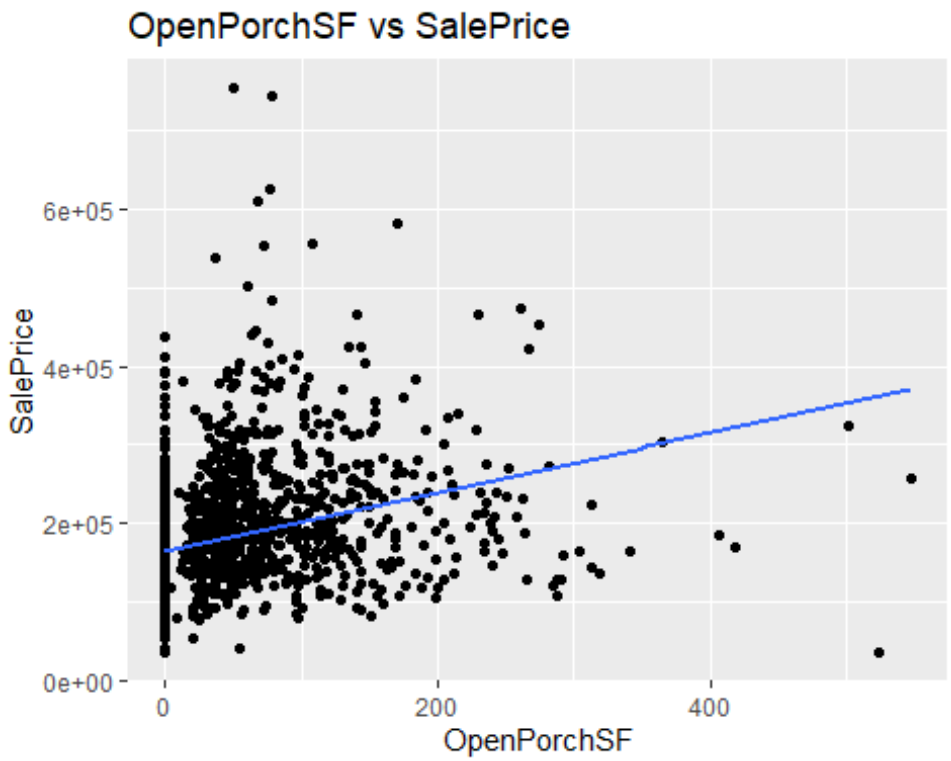
Variabile 'OpenPorchSF'

```
risultati_cov_cor_OpenPorchSF <- calcolo_cov_cor(case$OpenPorchSF)
print(risultati_cov_cor_OpenPorchSF)
```

```
##          cov          cor
## 1.662523e+06 3.158562e-01
```

```
# Plot
ggplot(case, aes(x = OpenPorchSF, y = SalePrice)) + geom_point() +
geom_smooth(method = "lm", se = FALSE) + ggtitle("OpenPorchSF vs SalePrice")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



La variabile OpenPorchSF ha una covarianza di 1662523 e una correlazione di 0.3159 con SalePrice. Questo indica una relazione positiva tra OpenPorchSF e SalePrice. Gli spazi aperti come le verande possono migliorare la qualità della vita e l'attrattiva estetica della proprietà, contribuendo a un prezzo di vendita più alto.

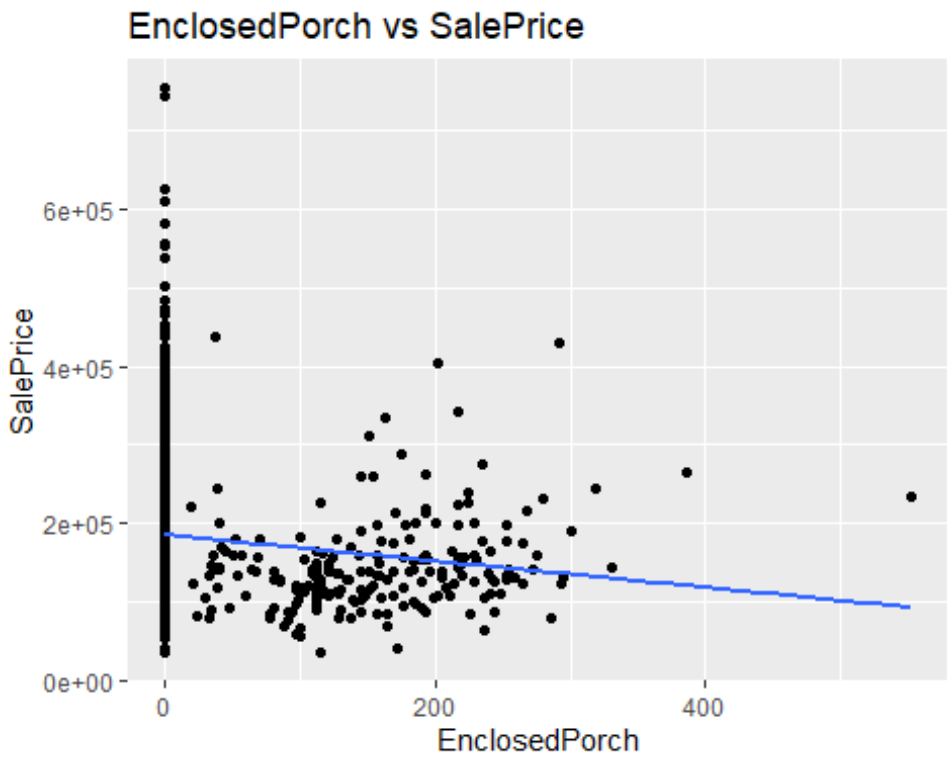
Variabile 'EnclosedPorch'

```
risultati_cov_cor_EnclosedPorch <- calcolo_cov_cor(case$EnclosedPorch)
print(risultati_cov_cor_EnclosedPorch)
```

```
##           cov           cor
## -6.243049e+05 -1.285780e-01
```

```
# Plot
ggplot(case, aes(x = EnclosedPorch, y = SalePrice)) + geom_point() +
geom_smooth(method = "lm", se = FALSE) + ggtitle("EnclosedPorch vs SalePrice")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



La variabile EnclosedPorch ha una covarianza di -624304.9 e una correlazione di -0.1286 con SalePrice. Questo indica una relazione negativa tra EnclosedPorch e SalePrice. Una veranda chiusa può aggiungere spazio utilizzabile alla casa e offrire protezione dagli elementi atmosferici, rendendo la proprietà più preziosa.

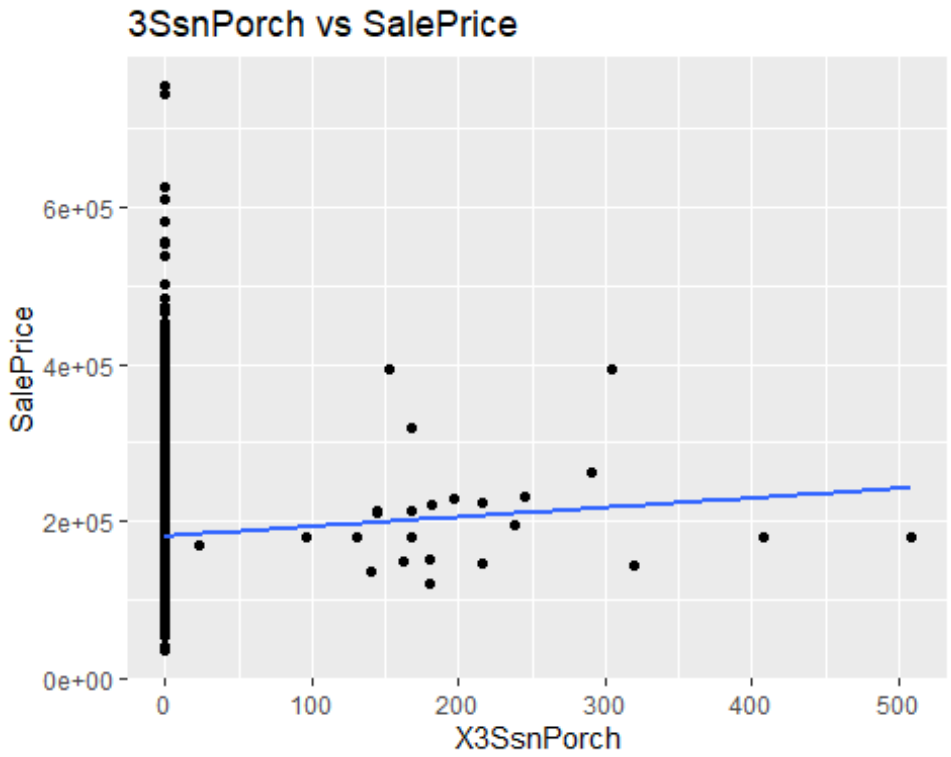
Variabile 'X3SsnPorch'

```
risultati_cov_cor_X3SsnPorch <- calcolo_cov_cor(case$`X3SsnPorch`)
print(risultati_cov_cor_X3SsnPorch)
```

```
##          cov          cor
## 1.038372e+05 4.458367e-02
```

```
# Plot
ggplot(case, aes(x = `X3SsnPorch`, y = SalePrice)) + geom_point() +
geom_smooth(method = "lm", se = FALSE) + ggtitle("3SsnPorch vs SalePrice")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



La variabile 3SsnPorch ha una covarianza di 103837.2 e una correlazione di 0.0446 con SalePrice. Questo indica una relazione positiva tra 3SsnPorch e SalePrice. Una veranda a tre stagioni può essere vista come un valore aggiunto, offrendo un'area utilizzabile per gran parte dell'anno e aumentando così il valore della proprietà.

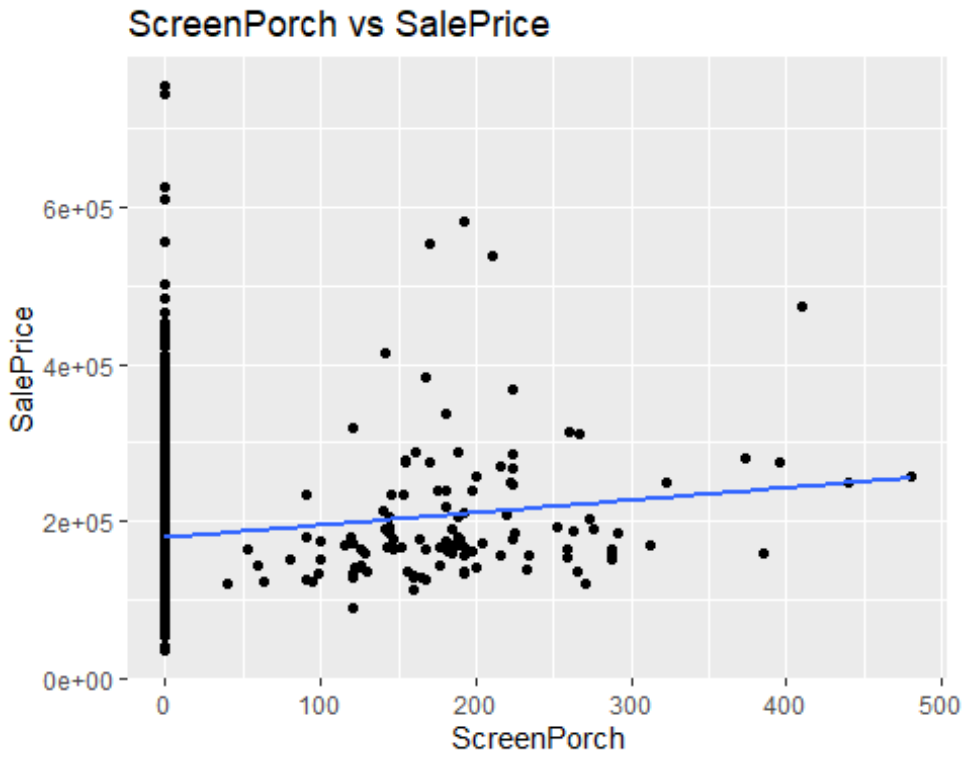
Variabile 'ScreenPorch'

```
risultati_cov_cor_ScreenPorch <- calcolo_cov_cor(case$ScreenPorch)
print(risultati_cov_cor_ScreenPorch)
```

```
##          cov          cor
## 4.936535e+05 1.114466e-01
```

```
# Plot
ggplot(case, aes(x = ScreenPorch, y = SalePrice)) + geom_point() +
geom_smooth(method = "lm", se = FALSE) + ggtitle("ScreenPorch vs SalePrice")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



La variabile ScreenPorch ha una covarianza di 493653.5 e una correlazione di 0.1114 con SalePrice. Questo indica una relazione positiva tra ScreenPorch e SalePrice. Una veranda schermata può fornire uno spazio esterno, migliorando la qualità della vita e il valore della casa.

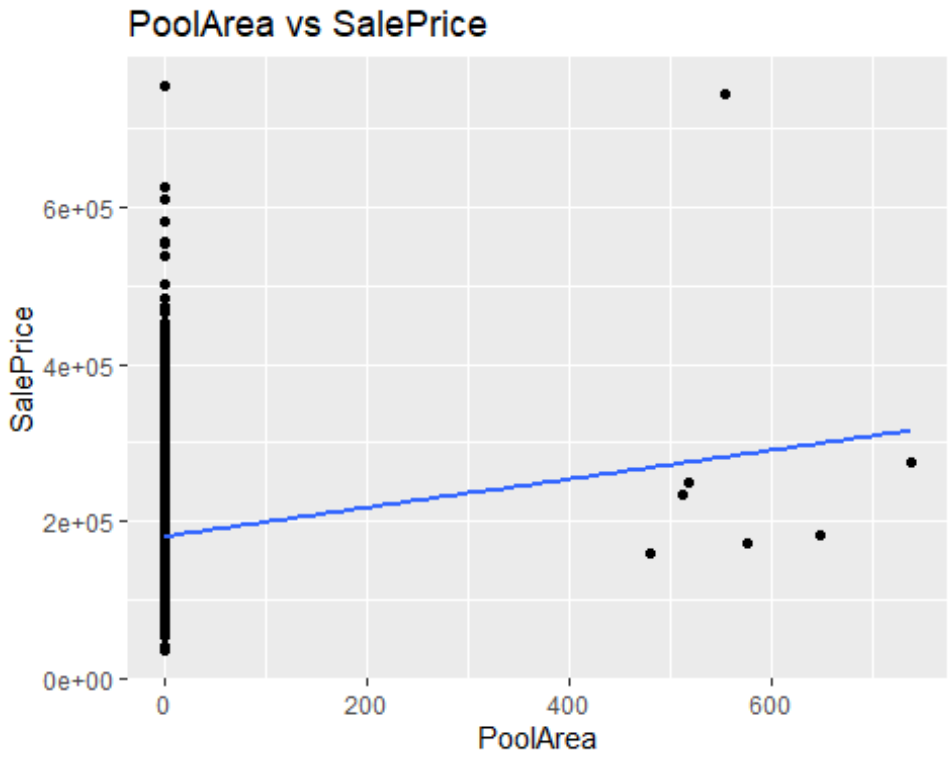
Variabile 'PoolArea'

```
risultati_cov_cor_PoolArea <- calcolo_cov_cor(case$PoolArea)
print(risultati_cov_cor_PoolArea)
```

```
##          cov          cor
## 2.949323e+05 9.240355e-02
```

```
# Plot
ggplot(case, aes(x = PoolArea, y = SalePrice)) + geom_point() + geom_smooth(method
= "lm", se = FALSE) + ggtitle("PoolArea vs SalePrice")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



La variabile PoolArea ha una covarianza di 294932.3 e una correlazione di 0.0924 con SalePrice. Questo indica una relazione positiva tra PoolArea e SalePrice. Una piscina può essere un lusso desiderabile in una proprietà, contribuendo significativamente al prezzo di vendita, anche se questo può variare a seconda della località.

Variabile 'MiscVal'

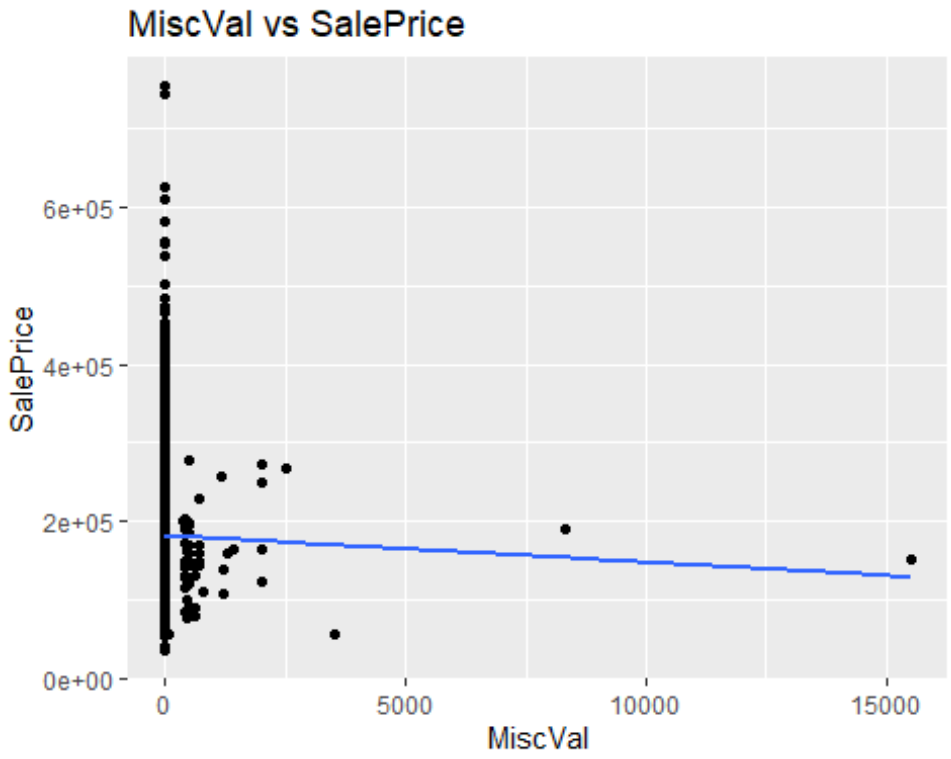
```
risultati_cov_cor_MiscVal <- calcolo_cov_cor(case$MiscVal)
print(risultati_cov_cor_MiscVal)
```

```
##           cov           cor
## -8.351503e+05 -2.118958e-02
```

```
cat("Commento: La variabile MiscVal ha una covarianza di",
round(risultati_cov_cor_MiscVal["cov"], 2),
    "e una correlazione di", round(risultati_cov_cor_MiscVal["cor"], 4),
    "con SalePrice. Questo indica una relazione",
ifelse(risultati_cov_cor_MiscVal["cor"] > 0, "positiva", "negativa"),
    "tra MiscVal e SalePrice. I valori vari possono rappresentare miglioramenti o
caratteristiche aggiuntive della proprietà che non rientrano nelle categorie
standard, influenzando così il prezzo di vendita.\n")
```

```
# Plot
ggplot(case, aes(x = MiscVal, y = SalePrice)) + geom_point() + geom_smooth(method
= "lm", se = FALSE) + ggtitle("MiscVal vs SalePrice")
```

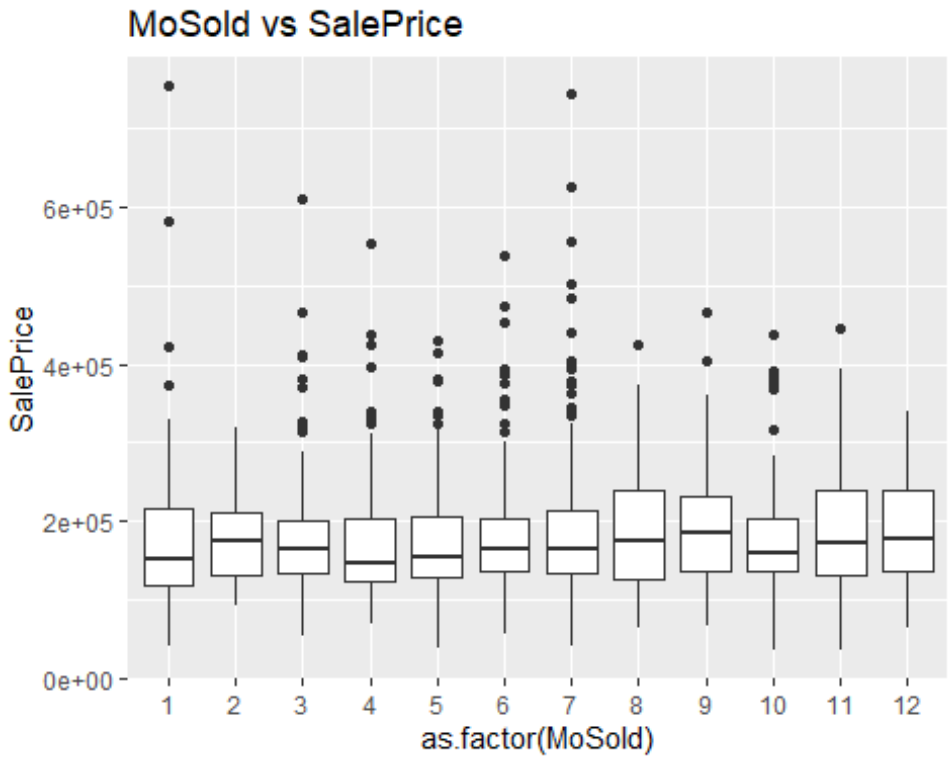
```
## `geom_smooth()` using formula = 'y ~ x'
```



La variabile MiscVal ha una covarianza di -835150.3 e una correlazione di -0.0212 con SalePrice. Questo indica una relazione negativa tra MiscVal e SalePrice. I valori vari possono rappresentare miglioramenti o caratteristiche aggiuntive della proprietà che non rientrano nelle categorie standard, influenzando così il prezzo di vendita.

Variabile 'MoSold'

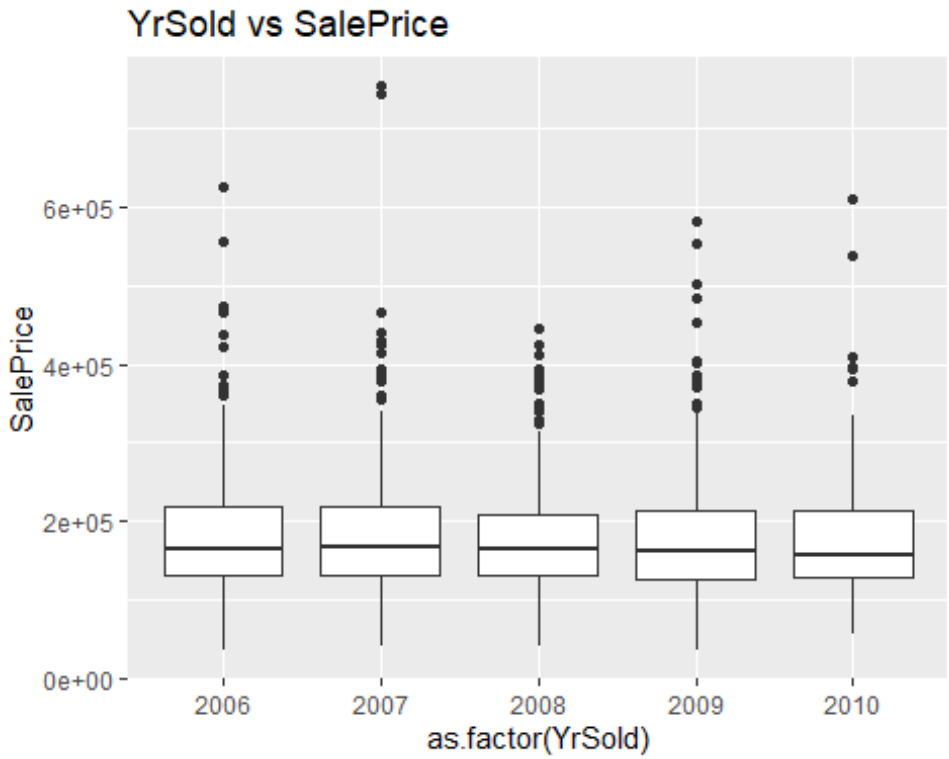
```
risultati_cov_cor_MoSold <- calcolo_cov_cor(case$MoSold)
print(risultati_cov_cor_MoSold)

##          cov          cor
## 9.972849e+03 4.643225e-02

# Plot
ggplot(case, aes(x = as.factor(MoSold), y = SalePrice)) + geom_boxplot() +
ggtitle("MoSold vs SalePrice")
```



La variabile MoSold ha una covarianza di 9972.85 e una correlazione di 0.0464 con SalePrice. Questo indica una relazione positiva tra MoSold e SalePrice. Il mese di vendita può influenzare il prezzo di vendita a causa della stagionalità del mercato immobiliare, con certi mesi che potrebbero avere una domanda più alta e quindi prezzi più alti.

Variabile 'YrSold'

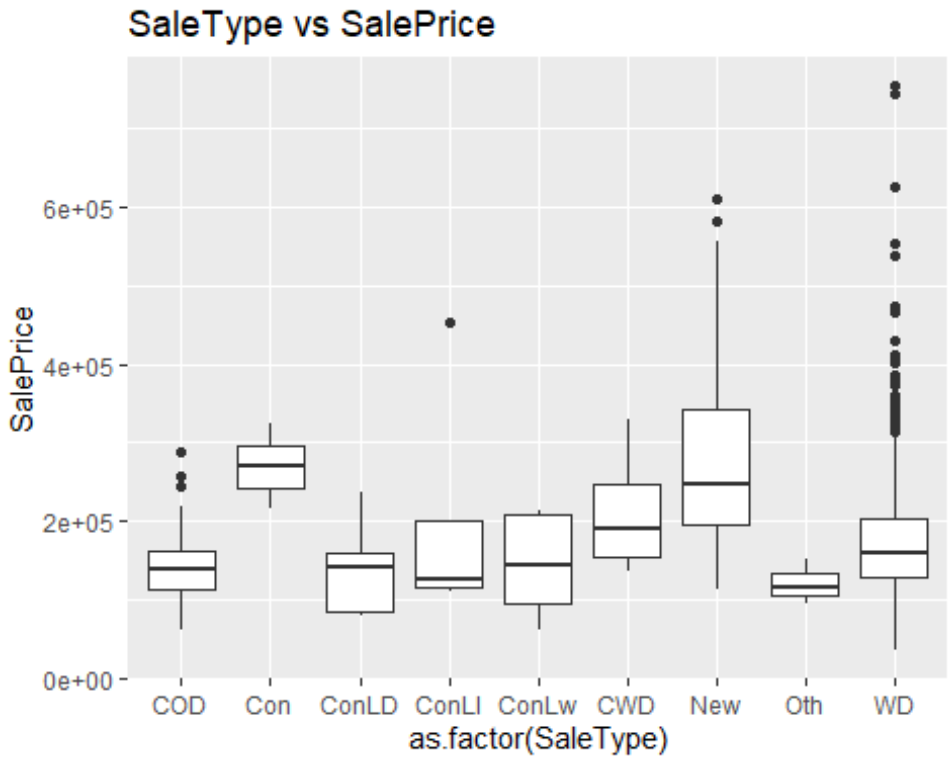
```
risultati_cov_cor_YrSold <- calcolo_cov_cor(case$YrSold)
print(risultati_cov_cor_YrSold)

##           cov           cor
## -3.051541e+03 -2.892259e-02

# Plot
ggplot(case, aes(x = as.factor(YrSold), y = SalePrice)) + geom_boxplot() +
ggtitle("YrSold vs SalePrice")
```



La variabile YrSold ha una covarianza di -3051.54 e una correlazione di -0.0289 con SalePrice. Questo indica una relazione negativa tra YrSold e SalePrice. L'anno di vendita può riflettere le condizioni economiche generali e le tendenze del mercato immobiliare, influenzando i prezzi di vendita.

Variabile 'SaleType'

```
risultati_devianza_SaleType <- calcola_devianza(case$SalePrice, case$SaleType)
print(risultati_devianza_SaleType)

## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 1.264131e+12
##
## $devianza_entro_gruppi
## [1] 7.94378e+12
##
## $eta2
## [1] 0.1372875

# Plot
ggplot(case, aes(x = as.factor(SaleType), y = SalePrice)) + geom_boxplot() +
ggtitle("SaleType vs SalePrice")
```



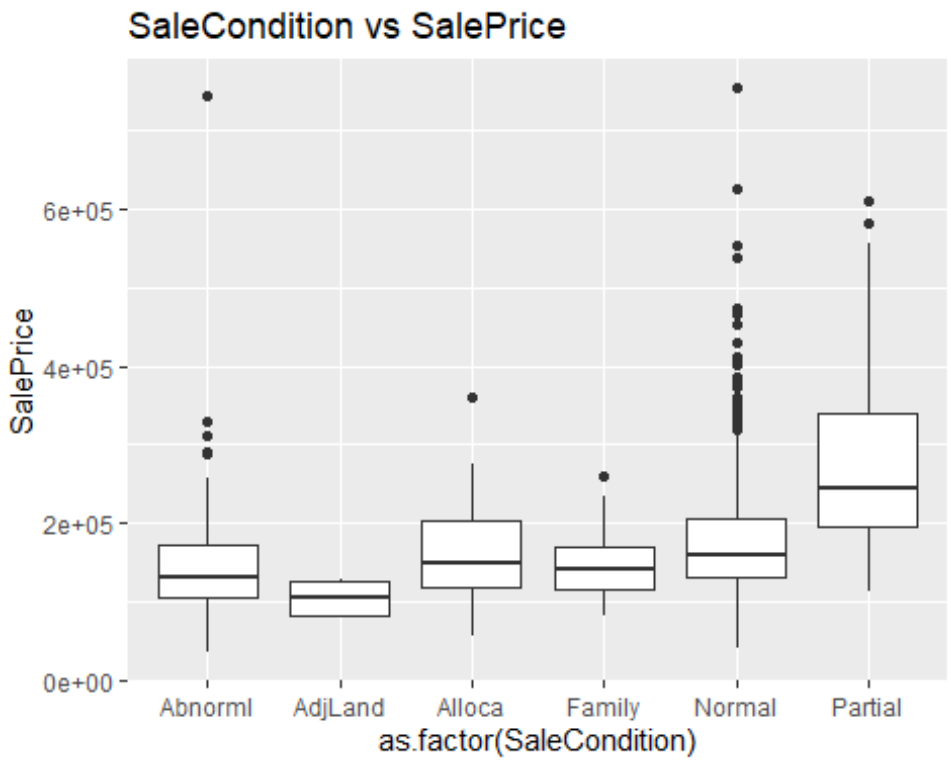
La variabile SaleType ha una devianza totale di 9.207911e+12 , con una devianza tra gruppi di 1.264131e+12 e una devianza entro gruppi di 7.94378e+12 . Eta² è 0.1373 indicando che il 13.73 % della varianza di SalePrice è spiegata da SaleType. Il tipo di vendita può influenzare il prezzo di vendita, con vendite all'asta o forzate che potrebbero portare a prezzi più bassi rispetto a vendite tradizionali.

Variabile 'SaleCondition'

```
risultati_devianza_SaleCondition <- calcola_devianza(case$SalePrice,
case$SaleCondition)
print(risultati_devianza_SaleCondition)
```

```
## $devianza_totale
## [1] 9.207911e+12
##
## $devianza_tra_gruppi
## [1] 1.247649e+12
##
## $devianza_entro_gruppi
## [1] 7.960263e+12
##
## $eta2
## [1] 0.1354975
```

```
# Plot
ggplot(case, aes(x = as.factor(SaleCondition), y = SalePrice)) + geom_boxplot() +
ggtitle("SaleCondition vs SalePrice")
```



La variabile SaleCondition ha una devianza totale di 9.207911e+12 , con una devianza tra gruppi di 1.247649e+12 e una devianza entro gruppi di 7.960263e+12 . Eta² è 0.1355 indicando che il 13.55 % della varianza di SalePrice è spiegata da SaleCondition. La condizione di vendita può influenzare il prezzo di vendita, con vendite come le foreclosures che potrebbero portare a prezzi più bassi rispetto a vendite tradizionali.