Homework 2

Urbani Ludovico, Mohamed Fadhla Mutua, Florin Razvan Burlacu

2024-02-01

# Impostazioni progetto

library(ggplot2)  
  
plot\_theme = theme\_minimal() +  
 theme(  
 plot.title = element\_text(face = "bold", hjust = 0.5),  
 plot.subtitle = element\_text(hjust = 0.5),  
 axis.line = element\_line(color = "darkgrey", size = 0.4),  
 axis.ticks = element\_line(color = "black"))  
theme\_set(plot\_theme)

# House Price - Univariata Quantitativa

## Lettura dati

HP <- read.csv2(  
 "./house\_price.csv",   
 sep = ",",   
 stringsAsFactors = T,  
 row.names=1)

## Scelta delle colonne

column\_selection <- function(data){  
 HP\_numeric\_bool = sapply(data, is.numeric)  
 HP\_numeric\_fcolnames = names(data[HP\_numeric\_bool])  
 # Rimozione delle colonne  
 HP\_numeric\_fcolnames = HP\_numeric\_fcolnames[c(-1,-4,-5)]  
 print("Colonne scelte: ")  
 print(HP\_numeric\_fcolnames)  
 return(HP\_numeric\_fcolnames)  
}  
HP\_numeric\_colnames = column\_selection(HP)

## [1] "Colonne scelte: "  
## [1] "LotFrontage" "LotArea" "YearBuilt" "YearRemodAdd"   
## [5] "MasVnrArea" "BsmtFinSF1" "BsmtFinSF2" "BsmtUnfSF"   
## [9] "TotalBsmtSF" "X1stFlrSF" "X2ndFlrSF" "LowQualFinSF"   
## [13] "GrLivArea" "BsmtFullBath" "BsmtHalfBath" "FullBath"   
## [17] "HalfBath" "BedroomAbvGr" "KitchenAbvGr" "TotRmsAbvGrd"   
## [21] "Fireplaces" "GarageYrBlt" "GarageCars" "GarageArea"   
## [25] "WoodDeckSF" "OpenPorchSF" "EnclosedPorch" "X3SsnPorch"   
## [29] "ScreenPorch" "PoolArea" "MiscVal" "MoSold"   
## [33] "YrSold" "SalePrice"

## Analisi Primaria

Analisi\_Primaria <- function(data){  
 print(paste("Numeri di osservazione: ",ncol(data)))  
 print(paste("Numeri di variabili: ",nrow(data)))  
}  
Analisi\_Primaria(HP[ ,HP\_numeric\_colnames])

## [1] "Numeri di osservazione: 34"  
## [1] "Numeri di variabili: 1460"

str(HP[,HP\_numeric\_colnames])

## 'data.frame': 1460 obs. of 34 variables:  
## $ LotFrontage : int 65 80 68 60 84 85 75 NA 51 50 ...  
## $ LotArea : int 8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...  
## $ YearBuilt : int 2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...  
## $ YearRemodAdd : int 2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...  
## $ MasVnrArea : int 196 0 162 0 350 0 186 240 0 0 ...  
## $ BsmtFinSF1 : int 706 978 486 216 655 732 1369 859 0 851 ...  
## $ BsmtFinSF2 : int 0 0 0 0 0 0 0 32 0 0 ...  
## $ BsmtUnfSF : int 150 284 434 540 490 64 317 216 952 140 ...  
## $ TotalBsmtSF : int 856 1262 920 756 1145 796 1686 1107 952 991 ...  
## $ X1stFlrSF : int 856 1262 920 961 1145 796 1694 1107 1022 1077 ...  
## $ X2ndFlrSF : int 854 0 866 756 1053 566 0 983 752 0 ...  
## $ LowQualFinSF : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ GrLivArea : int 1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...  
## $ BsmtFullBath : int 1 0 1 1 1 1 1 1 0 1 ...  
## $ BsmtHalfBath : int 0 1 0 0 0 0 0 0 0 0 ...  
## $ FullBath : int 2 2 2 1 2 1 2 2 2 1 ...  
## $ HalfBath : int 1 0 1 0 1 1 0 1 0 0 ...  
## $ BedroomAbvGr : int 3 3 3 3 4 1 3 3 2 2 ...  
## $ KitchenAbvGr : int 1 1 1 1 1 1 1 1 2 2 ...  
## $ TotRmsAbvGrd : int 8 6 6 7 9 5 7 7 8 5 ...  
## $ Fireplaces : int 0 1 1 1 1 0 1 2 2 2 ...  
## $ GarageYrBlt : int 2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...  
## $ GarageCars : int 2 2 2 3 3 2 2 2 2 1 ...  
## $ GarageArea : int 548 460 608 642 836 480 636 484 468 205 ...  
## $ WoodDeckSF : int 0 298 0 0 192 40 255 235 90 0 ...  
## $ OpenPorchSF : int 61 0 42 35 84 30 57 204 0 4 ...  
## $ EnclosedPorch: int 0 0 0 272 0 0 0 228 205 0 ...  
## $ X3SsnPorch : int 0 0 0 0 0 320 0 0 0 0 ...  
## $ ScreenPorch : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ PoolArea : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ MiscVal : int 0 0 0 0 0 700 0 350 0 0 ...  
## $ MoSold : int 2 5 9 2 12 10 8 11 4 1 ...  
## $ YrSold : int 2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...  
## $ SalePrice : int 208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...

## Analisi Univariata

univariate\_analysis <- function(data, colname) {  
 result = list()  
   
 result[[colname]] = list(  
 # Summary statistics including min, 1st and 3rd quartile, mean, median, max  
 Summary = summary(data),  
 # Variance  
 Variance = round(var(data), 2),  
 # Standard Deviation  
 Standard\_Deviation = round(sd(data), 2),  
 # Mode calculation  
 Moda = Mode(data)  
 )  
 return(result)  
}  
  
Mode <- function(data) {  
 uniq\_vals = unique(data)  
 tab = tabulate(match(data, uniq\_vals))  
 modes = uniq\_vals[tab == max(tab)]  
 list(Mode = modes, Frequency = max(tab))  
}

## Analisi non modificato

sum\_info\_h = lapply(HP\_numeric\_colnames, function(ele) {  
 colummn = na.omit(HP[, ele])  
 univariate\_analysis(colummn, ele)  
})

## Funzione di Visualizazione continuo

visualization\_continuos <- function(data,sumary,colname){  
 par(mfrow = c(2,2))  
 print(sumary)  
 boxplot(data, horizontal = T, main = paste("Boxplot of ", colname))  
 hist(data, probability = T, col = 2, main = paste("Hist of ", colname))  
 curve(dnorm(x, mean = mean(data), sd = sd(data)),  
 add=T, col = 4)  
 print(density(data))  
 plot(density(data), main = paste("Desnity Distribution of ", colname))  
 qqnorm(data, col = 3, main = paste("QQnorm of ", colname))  
 qqline(data, col = 5, pch = 10)  
}

## Funzione di visualizazione discreta

visualization\_discrete <- function(data,sumary,colname){  
 par(mfrow = c(1,1))  
 print(sumary)  
 print("freq: ")  
 cat("\n")  
 print(table(data))  
 cat("\n")  
 print("percintile: ")  
 cat("\n")  
 print(prop.table(table(data))\*100)  
 cat("\n")  
 barplot(table(data), col = 2, main = paste("Barplot of ", colname))  
}

## Rimozione degli zero

rm\_0 <- function(data){  
 imp\_data = data  
 imp\_data[imp\_data == 0] = NA  
 imp\_data = na.omit(imp\_data)  
 return(imp\_data)  
}

## Rimozione degli outlier attraverso intervallo di confidenza

rm\_oultiers <- function(data, CI){  
 mean\_data = mean(data)  
 sd\_data = sd(data)  
 confidence\_level = CI  
 z\_score = qnorm((1 + confidence\_level) / 2)  
   
 lower\_bound = mean\_data - z\_score \* sd\_data  
 upper\_bound = mean\_data + z\_score \* sd\_data  
   
 data\_in\_ci = data[data >= lower\_bound & data <= upper\_bound]  
 return(data\_in\_ci)  
}

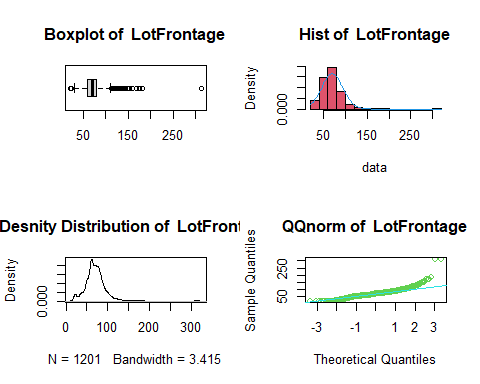
## Analisi

### LotFrontage

visualization\_continuos(na.omit(HP[,"LotFrontage"]), sum\_info\_h[1],"LotFrontage")

## [[1]]  
## [[1]]$LotFrontage  
## [[1]]$LotFrontage$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 21.00 59.00 69.00 70.05 80.00 313.00   
##   
## [[1]]$LotFrontage$Variance  
## [1] 589.75  
##   
## [[1]]$LotFrontage$Standard\_Deviation  
## [1] 24.28  
##   
## [[1]]$LotFrontage$Moda  
## [[1]]$LotFrontage$Moda$Mode  
## [1] 60  
##   
## [[1]]$LotFrontage$Moda$Frequency  
## [1] 143

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1201 obs.); Bandwidth 'bw' = 3.415  
##   
## x y   
## Min. : 10.75 Min. :0.0000000   
## 1st Qu.: 88.88 1st Qu.:0.0000000   
## Median :167.00 Median :0.0001649   
## Mean :167.00 Mean :0.0031969   
## 3rd Qu.:245.12 3rd Qu.:0.0032738   
## Max. :323.25 Max. :0.0238537



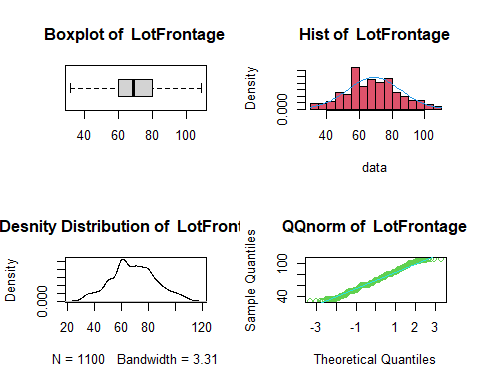
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli lotti ha un percorso verso la strada compreso tra 59 e 80 piedi, essendo rispettivamente il 1° e il 3° quantile e maggiormente alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(na.omit(HP[,"LotFrontage"]),.90)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"LotFrontage"),"LotFrontage")

## $LotFrontage  
## $LotFrontage$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 32.00 60.00 69.00 68.91 80.00 109.00   
##   
## $LotFrontage$Variance  
## [1] 261.63  
##   
## $LotFrontage$Standard\_Deviation  
## [1] 16.17  
##   
## $LotFrontage$Moda  
## $LotFrontage$Moda$Mode  
## [1] 60  
##   
## $LotFrontage$Moda$Frequency  
## [1] 143

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1100 obs.); Bandwidth 'bw' = 3.31  
##   
## x y   
## Min. : 22.07 Min. :5.250e-06   
## 1st Qu.: 46.28 1st Qu.:3.296e-03   
## Median : 70.50 Median :7.436e-03   
## Mean : 70.50 Mean :1.031e-02   
## 3rd Qu.: 94.72 3rd Qu.:1.884e-02   
## Max. :118.93 Max. :2.641e-02



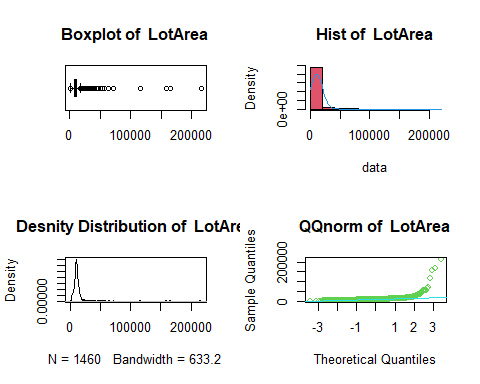
Possiamo osservare che la distribuzione segue per lo piu quella gaussiana e che la maggior parte degli lotti ha un percorso verso la strada compreso tra 60 e 80 piedi, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### LotArea

visualization\_continuos(na.omit(HP[,"LotArea"]), sum\_info\_h[2],"LotArea")

## [[1]]  
## [[1]]$LotArea  
## [[1]]$LotArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1300 7554 9478 10517 11602 215245   
##   
## [[1]]$LotArea$Variance  
## [1] 99625650  
##   
## [[1]]$LotArea$Standard\_Deviation  
## [1] 9981.26  
##   
## [[1]]$LotArea$Moda  
## [[1]]$LotArea$Moda$Mode  
## [1] 7200  
##   
## [[1]]$LotArea$Moda$Frequency  
## [1] 25

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 633.2  
##   
## x y   
## Min. : -599.5 Min. :0.000e+00   
## 1st Qu.: 53836.5 1st Qu.:0.000e+00   
## Median :108272.5 Median :0.000e+00   
## Mean :108272.5 Mean :4.587e-06   
## 3rd Qu.:162708.5 3rd Qu.:2.697e-07   
## Max. :217144.5 Max. :1.404e-04



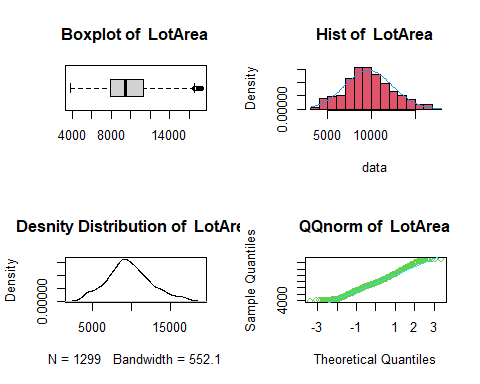
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli lotti ha un area compreso tra 7554 e 11602 piedi^2, essendo rispettivamente il 1° e il 3° quantile e e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(na.omit(HP[,"LotArea"]),.50)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"LotArea"),"LotArea")

## $LotArea  
## $LotArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3842 7877 9505 9666 11325 17217   
##   
## $LotArea$Variance  
## [1] 7453350  
##   
## $LotArea$Standard\_Deviation  
## [1] 2730.08  
##   
## $LotArea$Moda  
## $LotArea$Moda$Mode  
## [1] 7200  
##   
## $LotArea$Moda$Frequency  
## [1] 25

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1299 obs.); Bandwidth 'bw' = 552.1  
##   
## x y   
## Min. : 2186 Min. :2.276e-08   
## 1st Qu.: 6358 1st Qu.:1.739e-05   
## Median :10530 Median :4.514e-05   
## Mean :10530 Mean :5.987e-05   
## 3rd Qu.:14701 3rd Qu.:1.008e-04   
## Max. :18873 Max. :1.621e-04



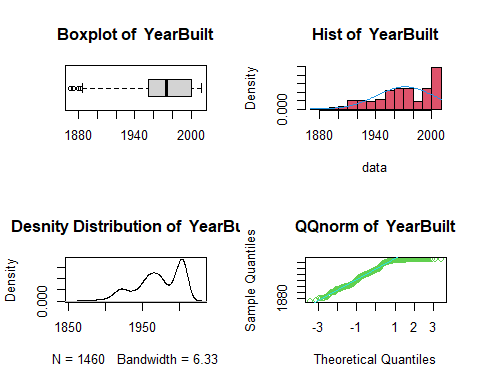
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli lotti ha un area compreso tra 7500 e 11362 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Da notare che si butano il 50% dei dat

### YearBuilt

visualization\_continuos(na.omit(HP[,"YearBuilt"]), sum\_info\_h[3],"YearBuilt")

## [[1]]  
## [[1]]$YearBuilt  
## [[1]]$YearBuilt$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1872 1954 1973 1971 2000 2010   
##   
## [[1]]$YearBuilt$Variance  
## [1] 912.22  
##   
## [[1]]$YearBuilt$Standard\_Deviation  
## [1] 30.2  
##   
## [[1]]$YearBuilt$Moda  
## [[1]]$YearBuilt$Moda$Mode  
## [1] 2006  
##   
## [[1]]$YearBuilt$Moda$Frequency  
## [1] 67

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 6.33  
##   
## x y   
## Min. :1853 Min. :6.120e-07   
## 1st Qu.:1897 1st Qu.:5.052e-04   
## Median :1941 Median :4.774e-03   
## Mean :1941 Mean :5.677e-03   
## 3rd Qu.:1985 3rd Qu.:9.652e-03   
## Max. :2029 Max. :1.871e-02



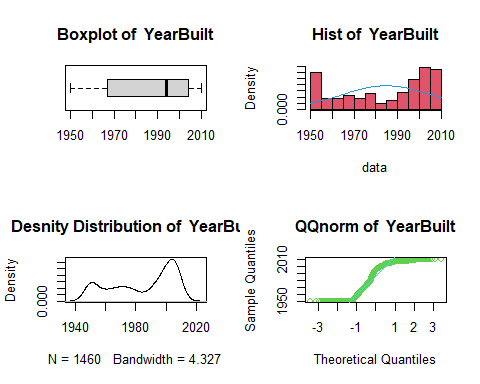
Possiamo dedurre che la distribuzione è negativamente asimetrica e quindi il numero di edifici costruiti crescono ogni anno La maggior parte degli edifici erano costruiti fra 1954 e 2000, essendo rispettivamente il 1° e il 3° quantile e l’anno con è il massimo costruitto è nel 2006 con 67 edifici

### YearRemodAd

visualization\_continuos(na.omit(HP[,"YearRemodAdd"]), sum\_info\_h[4], "YearBuilt")

## [[1]]  
## [[1]]$YearRemodAdd  
## [[1]]$YearRemodAdd$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1950 1967 1994 1985 2004 2010   
##   
## [[1]]$YearRemodAdd$Variance  
## [1] 426.23  
##   
## [[1]]$YearRemodAdd$Standard\_Deviation  
## [1] 20.65  
##   
## [[1]]$YearRemodAdd$Moda  
## [[1]]$YearRemodAdd$Moda$Mode  
## [1] 1950  
##   
## [[1]]$YearRemodAdd$Moda$Frequency  
## [1] 178

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 4.327  
##   
## x y   
## Min. :1937 Min. :2.807e-05   
## 1st Qu.:1959 1st Qu.:7.214e-03   
## Median :1980 Median :1.023e-02   
## Mean :1980 Mean :1.162e-02   
## 3rd Qu.:2001 3rd Qu.:1.389e-02   
## Max. :2023 Max. :3.250e-02



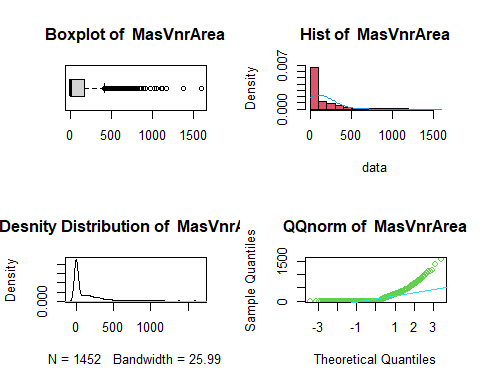
Possiamo notare che la distribuzione ha un pico negli anni 50 con il massimo di rimodellazione nel 1950 con 178 edifici. Dopodichè, si ha un calo quasi immediato e i numeri di edifici rimodellati resta quasi costante fino agli anni 80 dove fra il 1980 al 1985 si nota il minimo numero di rimodellazione. Dopodiche, la distribuzione inizia a crescere quasi costantemente con pico negli anni fra 2000 e 2005

### MasVnrArea

visualization\_continuos(na.omit(HP[,"MasVnrArea"]), sum\_info\_h[5],"MasVnrArea")

## [[1]]  
## [[1]]$MasVnrArea  
## [[1]]$MasVnrArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 0.0 103.7 166.0 1600.0   
##   
## [[1]]$MasVnrArea$Variance  
## [1] 32784.97  
##   
## [[1]]$MasVnrArea$Standard\_Deviation  
## [1] 181.07  
##   
## [[1]]$MasVnrArea$Moda  
## [[1]]$MasVnrArea$Moda$Mode  
## [1] 0  
##   
## [[1]]$MasVnrArea$Moda$Frequency  
## [1] 861

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1452 obs.); Bandwidth 'bw' = 25.99  
##   
## x y   
## Min. : -77.98 Min. :3.000e-09   
## 1st Qu.: 361.01 1st Qu.:9.362e-06   
## Median : 800.00 Median :4.165e-05   
## Mean : 800.00 Mean :5.685e-04   
## 3rd Qu.:1238.99 3rd Qu.:3.960e-04   
## Max. :1677.98 Max. :9.300e-03



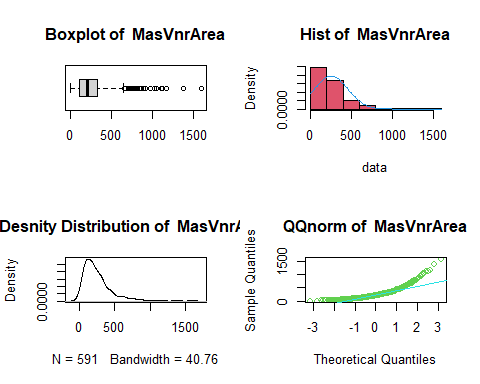
Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area di MasVnrArea è prelevantemete 0 qundi non applicato. Per fare un analisi sugli edifici che hanno il Masonry veneer, serve ignorare quelli senza

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(HP[,"MasVnrArea"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"MasVnrArea"),"MasVnrArea")

## $MasVnrArea  
## $MasVnrArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 113.0 203.0 254.7 330.5 1600.0   
##   
## $MasVnrArea$Variance  
## [1] 42084.13  
##   
## $MasVnrArea$Standard\_Deviation  
## [1] 205.14  
##   
## $MasVnrArea$Moda  
## $MasVnrArea$Moda$Mode  
## [1] 180 108 72  
##   
## $MasVnrArea$Moda$Frequency  
## [1] 8

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (591 obs.); Bandwidth 'bw' = 40.76  
##   
## x y   
## Min. :-121.3 Min. :1.888e-07   
## 1st Qu.: 339.6 1st Qu.:1.440e-05   
## Median : 800.5 Median :7.853e-05   
## Mean : 800.5 Mean :5.419e-04   
## 3rd Qu.:1261.4 3rd Qu.:6.832e-04   
## Max. :1722.3 Max. :2.868e-03



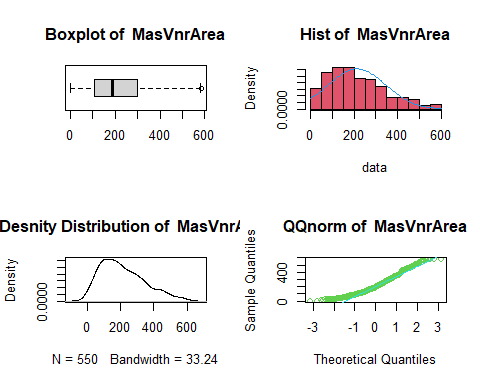
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli edifici ha un area coperto da Masonry veneer tra 113.0 e 330.5 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di oultier

rm\_outlier\_data = rm\_oultiers(rm\_0\_data,.90)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"MasVnrArea"),"MasVnrArea")

## $MasVnrArea  
## $MasVnrArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.0 108.0 186.0 213.5 297.8 584.0   
##   
## $MasVnrArea$Variance  
## [1] 17023.88  
##   
## $MasVnrArea$Standard\_Deviation  
## [1] 130.48  
##   
## $MasVnrArea$Moda  
## $MasVnrArea$Moda$Mode  
## [1] 180 108 72  
##   
## $MasVnrArea$Moda$Frequency  
## [1] 8

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (550 obs.); Bandwidth 'bw' = 33.24  
##   
## x y   
## Min. :-98.73 Min. :7.966e-07   
## 1st Qu.: 96.89 1st Qu.:2.708e-04   
## Median :292.50 Median :9.103e-04   
## Mean :292.50 Mean :1.277e-03   
## 3rd Qu.:488.11 3rd Qu.:2.248e-03   
## Max. :683.73 Max. :3.137e-03



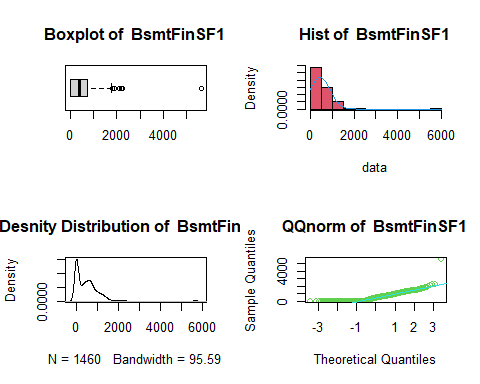
Possiamo dedurre nella distribuzione che la maggior parte degli edifici ha un area coperto da Masonry veneer tra 108.5 e 304.0 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### BsmtFinSF1

visualization\_continuos(na.omit(HP[,"BsmtFinSF1"]), sum\_info\_h[6],"BsmtFinSF1")

## [[1]]  
## [[1]]$BsmtFinSF1  
## [[1]]$BsmtFinSF1$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 0.0 383.5 443.6 712.2 5644.0   
##   
## [[1]]$BsmtFinSF1$Variance  
## [1] 208025.5  
##   
## [[1]]$BsmtFinSF1$Standard\_Deviation  
## [1] 456.1  
##   
## [[1]]$BsmtFinSF1$Moda  
## [[1]]$BsmtFinSF1$Moda$Mode  
## [1] 0  
##   
## [[1]]$BsmtFinSF1$Moda$Frequency  
## [1] 467

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 95.59  
##   
## x y   
## Min. :-286.8 Min. :0.000e+00   
## 1st Qu.:1267.6 1st Qu.:0.000e+00   
## Median :2822.0 Median :4.431e-07   
## Mean :2822.0 Mean :1.606e-04   
## 3rd Qu.:4376.4 3rd Qu.:1.658e-04   
## Max. :5930.8 Max. :1.525e-03



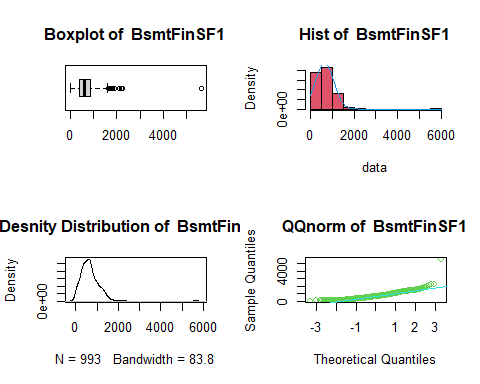
Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area finito di semiterrato finito di tipo 1 è prelevantemete 0 qundi non costruitto. Per fare un analisi sugli edifici che hanno il semiterrato finito di tipo 1, serve ignorare quelli senza

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(HP[,"BsmtFinSF1"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"BsmtFinSF1"),"BsmtFinSF1")

## $BsmtFinSF1  
## $BsmtFinSF1$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.0 371.0 604.0 652.3 867.0 5644.0   
##   
## $BsmtFinSF1$Variance  
## [1] 169727.8  
##   
## $BsmtFinSF1$Standard\_Deviation  
## [1] 411.98  
##   
## $BsmtFinSF1$Moda  
## $BsmtFinSF1$Moda$Mode  
## [1] 24  
##   
## $BsmtFinSF1$Moda$Frequency  
## [1] 12

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (993 obs.); Bandwidth 'bw' = 83.8  
##   
## x y   
## Min. :-249.4 Min. :0.000e+00   
## 1st Qu.:1286.8 1st Qu.:0.000e+00   
## Median :2823.0 Median :4.138e-07   
## Mean :2823.0 Mean :1.626e-04   
## 3rd Qu.:4359.2 3rd Qu.:1.431e-04   
## Max. :5895.4 Max. :1.121e-03



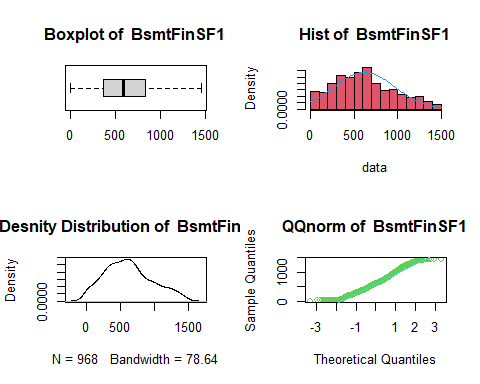
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli edifici ha un semiterrato finito di tipo 1 di area compresa tra 371.0 e 867.0 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(rm\_0\_data,.95)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"BsmtFinSF1"),"BsmtFinSF1")

## $BsmtFinSF1  
## $BsmtFinSF1$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.0 364.8 594.5 621.4 833.8 1456.0   
##   
## $BsmtFinSF1$Variance  
## [1] 119445.6  
##   
## $BsmtFinSF1$Standard\_Deviation  
## [1] 345.61  
##   
## $BsmtFinSF1$Moda  
## $BsmtFinSF1$Moda$Mode  
## [1] 24  
##   
## $BsmtFinSF1$Moda$Frequency  
## [1] 12

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (968 obs.); Bandwidth 'bw' = 78.64  
##   
## x y   
## Min. :-233.9 Min. :4.036e-07   
## 1st Qu.: 247.5 1st Qu.:1.898e-04   
## Median : 729.0 Median :4.584e-04   
## Mean : 729.0 Mean :5.187e-04   
## 3rd Qu.:1210.5 3rd Qu.:8.656e-04   
## Max. :1691.9 Max. :1.159e-03



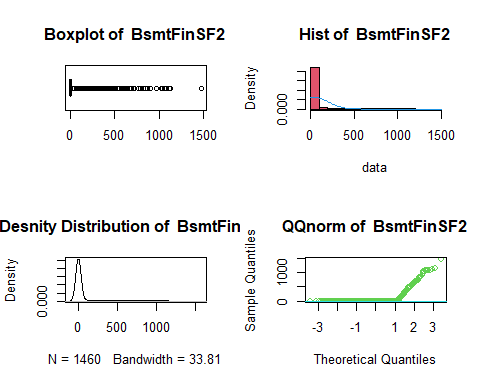
Possiamo dedurre nella distribuzione semi-normato che la maggior parte degli edifici ha un semiterrato finito di tipo 1 di area compresa tra 364.8 e 833.8 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### BsmtFinSF2

visualization\_continuos(na.omit(HP[,"BsmtFinSF2"]), sum\_info\_h[7],"BsmtFinSF2")

## [[1]]  
## [[1]]$BsmtFinSF2  
## [[1]]$BsmtFinSF2$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 46.55 0.00 1474.00   
##   
## [[1]]$BsmtFinSF2$Variance  
## [1] 26023.91  
##   
## [[1]]$BsmtFinSF2$Standard\_Deviation  
## [1] 161.32  
##   
## [[1]]$BsmtFinSF2$Moda  
## [[1]]$BsmtFinSF2$Moda$Mode  
## [1] 0  
##   
## [[1]]$BsmtFinSF2$Moda$Frequency  
## [1] 1293

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 33.81  
##   
## x y   
## Min. :-101.4 Min. :0.000e+00   
## 1st Qu.: 317.8 1st Qu.:1.196e-05   
## Median : 737.0 Median :5.220e-05   
## Mean : 737.0 Mean :5.951e-04   
## 3rd Qu.:1156.2 3rd Qu.:1.424e-04   
## Max. :1575.4 Max. :1.047e-02



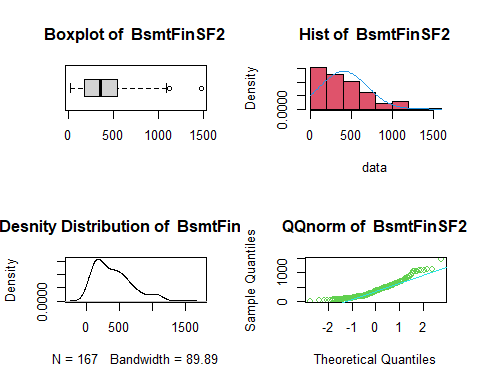
Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area finito di semiterrato finito di tipo 2 è prelevantemete 0 qundi non costruitto. Per fare un analisi sugli edifici che hanno il semiterrato finito di tipo 2, serve ignorare quelli senza

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(HP[,"BsmtFinSF2"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"BsmtFinSF2"),"BsmtFinSF2")

## $BsmtFinSF2  
## $BsmtFinSF2$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 28.0 178.5 354.0 407.0 551.0 1474.0   
##   
## $BsmtFinSF2$Variance  
## [1] 81173.38  
##   
## $BsmtFinSF2$Standard\_Deviation  
## [1] 284.91  
##   
## $BsmtFinSF2$Moda  
## $BsmtFinSF2$Moda$Mode  
## [1] 180  
##   
## $BsmtFinSF2$Moda$Frequency  
## [1] 5

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (167 obs.); Bandwidth 'bw' = 89.89  
##   
## x y   
## Min. :-241.7 Min. :2.985e-07   
## 1st Qu.: 254.7 1st Qu.:2.557e-05   
## Median : 751.0 Median :2.538e-04   
## Mean : 751.0 Mean :5.032e-04   
## 3rd Qu.:1247.3 3rd Qu.:1.001e-03   
## Max. :1743.7 Max. :1.592e-03



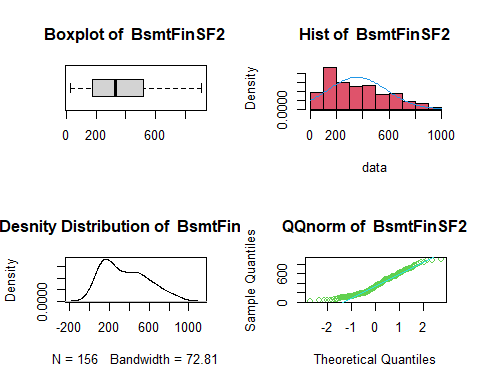
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli edifici ha un semiterrato finito di tipo 2 di area compresa tra 178.5 e 551.0 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(rm\_0\_data,.95)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"BsmtFinSF2"),"BsmtFinSF2")

## $BsmtFinSF2  
## $BsmtFinSF2$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 28.0 173.8 329.0 358.1 512.2 906.0   
##   
## $BsmtFinSF2$Variance  
## [1] 49340.55  
##   
## $BsmtFinSF2$Standard\_Deviation  
## [1] 222.13  
##   
## $BsmtFinSF2$Moda  
## $BsmtFinSF2$Moda$Mode  
## [1] 180  
##   
## $BsmtFinSF2$Moda$Frequency  
## [1] 5

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (156 obs.); Bandwidth 'bw' = 72.81  
##   
## x y   
## Min. :-190.4 Min. :5.917e-07   
## 1st Qu.: 138.3 1st Qu.:1.521e-04   
## Median : 467.0 Median :7.180e-04   
## Mean : 467.0 Mean :7.597e-04   
## 3rd Qu.: 795.7 3rd Qu.:1.262e-03   
## Max. :1124.4 Max. :1.840e-03



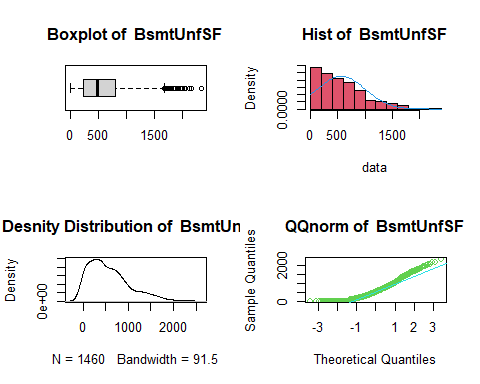
Possiamo dedurre nella distribuzione semi-normato che la maggior parte degli edifici ha un semiterrato finito di tipo 2 di area compresa tra 173.8 e 512.2 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### BsmtUnfSF

visualization\_continuos(na.omit(HP[,"BsmtUnfSF"]), sum\_info\_h[8],"BsmtUnfSF")

## [[1]]  
## [[1]]$BsmtUnfSF  
## [[1]]$BsmtUnfSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 223.0 477.5 567.2 808.0 2336.0   
##   
## [[1]]$BsmtUnfSF$Variance  
## [1] 195246.4  
##   
## [[1]]$BsmtUnfSF$Standard\_Deviation  
## [1] 441.87  
##   
## [[1]]$BsmtUnfSF$Moda  
## [[1]]$BsmtUnfSF$Moda$Mode  
## [1] 0  
##   
## [[1]]$BsmtUnfSF$Moda$Frequency  
## [1] 118

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 91.5  
##   
## x y   
## Min. :-274.5 Min. :3.380e-08   
## 1st Qu.: 446.7 1st Qu.:2.885e-05   
## Median :1168.0 Median :2.109e-04   
## Mean :1168.0 Mean :3.462e-04   
## 3rd Qu.:1889.3 3rd Qu.:7.062e-04   
## Max. :2610.5 Max. :9.942e-04



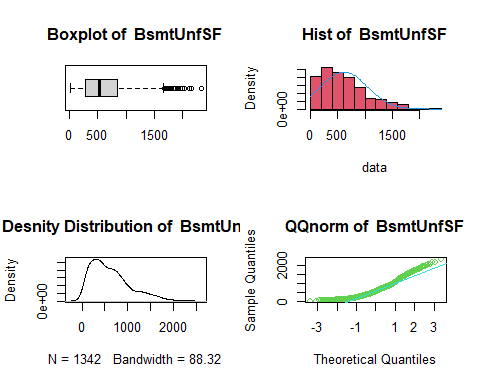
Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area di semiterrato non finito è prelevantemete 0 qundi non costruitto. Per fare un analisi sugli edifici che hanno il semiterrato non finito, serve ignorare quelli con

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(HP[,"BsmtUnfSF"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"BsmtUnfSF"),"BsmtUnfSF")

## $BsmtUnfSF  
## $BsmtUnfSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 14.0 288.0 536.0 617.1 843.2 2336.0   
##   
## $BsmtUnfSF$Variance  
## [1] 181624.3  
##   
## $BsmtUnfSF$Standard\_Deviation  
## [1] 426.17  
##   
## $BsmtUnfSF$Moda  
## $BsmtUnfSF$Moda$Mode  
## [1] 728  
##   
## $BsmtUnfSF$Moda$Frequency  
## [1] 9

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1342 obs.); Bandwidth 'bw' = 88.32  
##   
## x y   
## Min. :-251 Min. :3.810e-08   
## 1st Qu.: 462 1st Qu.:2.536e-05   
## Median :1175 Median :2.098e-04   
## Mean :1175 Mean :3.503e-04   
## 3rd Qu.:1888 3rd Qu.:7.008e-04   
## Max. :2601 Max. :1.084e-03



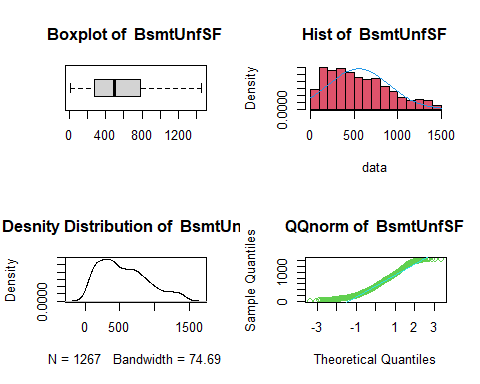
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli edifici ha un semiterrato non finito di area compresa tra 288.0 e 843.2 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(rm\_0\_data,.95)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"BsmtUnfSF"),"BsmtUnfSF")

## $BsmtUnfSF  
## $BsmtUnfSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 14.0 278.5 495.0 554.4 780.0 1450.0   
##   
## $BsmtUnfSF$Variance  
## [1] 120004.2  
##   
## $BsmtUnfSF$Standard\_Deviation  
## [1] 346.42  
##   
## $BsmtUnfSF$Moda  
## $BsmtUnfSF$Moda$Mode  
## [1] 728  
##   
## $BsmtUnfSF$Moda$Frequency  
## [1] 9

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1267 obs.); Bandwidth 'bw' = 74.69  
##   
## x y   
## Min. :-210.1 Min. :2.806e-07   
## 1st Qu.: 261.0 1st Qu.:1.939e-04   
## Median : 732.0 Median :4.555e-04   
## Mean : 732.0 Mean :5.302e-04   
## 3rd Qu.:1203.0 3rd Qu.:8.862e-04   
## Max. :1674.1 Max. :1.164e-03



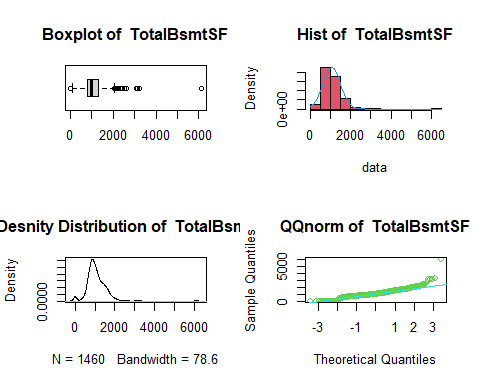
Possiamo dedurre nella distribuzione semi-normato che la maggior parte degli edifici ha un semiterrato non finito di area compresa tra 278.5 e 780.0 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### TotalBsmtSF

visualization\_continuos(na.omit(HP[,"TotalBsmtSF"]), sum\_info\_h[9],"TotalBsmtSF")

## [[1]]  
## [[1]]$TotalBsmtSF  
## [[1]]$TotalBsmtSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 795.8 991.5 1057.4 1298.2 6110.0   
##   
## [[1]]$TotalBsmtSF$Variance  
## [1] 192462.4  
##   
## [[1]]$TotalBsmtSF$Standard\_Deviation  
## [1] 438.71  
##   
## [[1]]$TotalBsmtSF$Moda  
## [[1]]$TotalBsmtSF$Moda$Mode  
## [1] 0  
##   
## [[1]]$TotalBsmtSF$Moda$Frequency  
## [1] 37

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 78.6  
##   
## x y   
## Min. :-235.8 Min. :0.000e+00   
## 1st Qu.:1409.6 1st Qu.:0.000e+00   
## Median :3055.0 Median :3.073e-06   
## Mean :3055.0 Mean :1.518e-04   
## 3rd Qu.:4700.4 3rd Qu.:1.024e-04   
## Max. :6345.8 Max. :1.254e-03



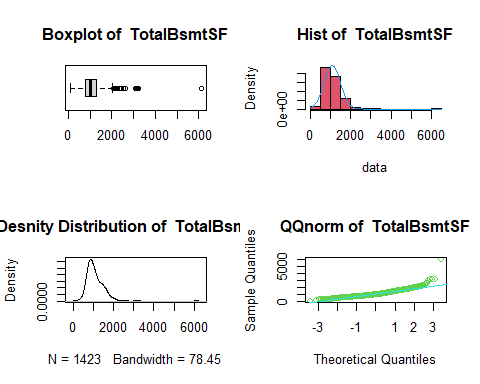
Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area totale semiterrato è prelevantemete 0 qundi non costruitto. Per fare un analisi sugli edifici che hanno il semiterrato non finito, serve ignorare quelli senza

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(HP[,"TotalBsmtSF"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"TotalBsmtSF"),"TotalBsmtSF")

## $TotalBsmtSF  
## $TotalBsmtSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 105.0 810.5 1004.0 1084.9 1309.5 6110.0   
##   
## $TotalBsmtSF$Variance  
## [1] 167619.6  
##   
## $TotalBsmtSF$Standard\_Deviation  
## [1] 409.41  
##   
## $TotalBsmtSF$Moda  
## $TotalBsmtSF$Moda$Mode  
## [1] 864  
##   
## $TotalBsmtSF$Moda$Frequency  
## [1] 35

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1423 obs.); Bandwidth 'bw' = 78.45  
##   
## x y   
## Min. :-130.4 Min. :0.000e+00   
## 1st Qu.:1488.6 1st Qu.:0.000e+00   
## Median :3107.5 Median :2.040e-06   
## Mean :3107.5 Mean :1.543e-04   
## 3rd Qu.:4726.4 3rd Qu.:8.769e-05   
## Max. :6345.4 Max. :1.287e-03



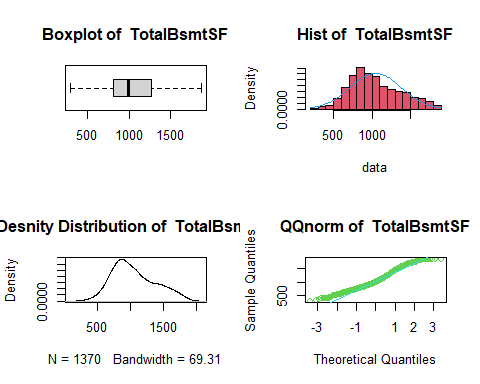
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli edifici ha un semiterrato di area compresa tra 810.5 e 1309.5 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(rm\_0\_data,.95)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"TotalBsmtSF"),"TotalBsmtSF")

## $TotalBsmtSF  
## $TotalBsmtSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 290.0 806.0 990.5 1048.2 1264.0 1869.0   
##   
## $TotalBsmtSF$Variance  
## [1] 106622.3  
##   
## $TotalBsmtSF$Standard\_Deviation  
## [1] 326.53  
##   
## $TotalBsmtSF$Moda  
## $TotalBsmtSF$Moda$Mode  
## [1] 864  
##   
## $TotalBsmtSF$Moda$Frequency  
## [1] 35

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1370 obs.); Bandwidth 'bw' = 69.31  
##   
## x y   
## Min. : 82.06 Min. :6.440e-08   
## 1st Qu.: 580.78 1st Qu.:8.634e-05   
## Median :1079.50 Median :4.377e-04   
## Mean :1079.50 Mean :5.008e-04   
## 3rd Qu.:1578.22 3rd Qu.:8.098e-04   
## Max. :2076.94 Max. :1.364e-03



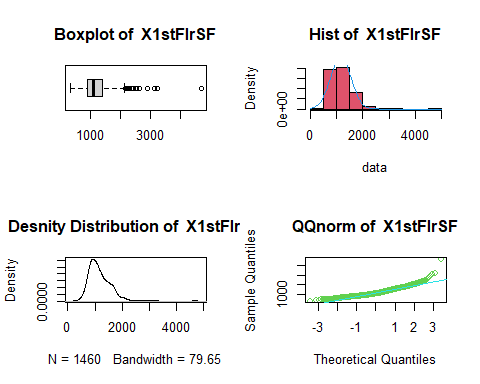
Possiamo notare che la maggior parte degli edifici ha un semiterrato di area compresa tra 804.0 e 1266.8 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### X1stFlrSF

visualization\_continuos(na.omit(HP[,"X1stFlrSF"]), sum\_info\_h[10],"X1stFlrSF")

## [[1]]  
## [[1]]$X1stFlrSF  
## [[1]]$X1stFlrSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 882 1087 1163 1391 4692   
##   
## [[1]]$X1stFlrSF$Variance  
## [1] 149450.1  
##   
## [[1]]$X1stFlrSF$Standard\_Deviation  
## [1] 386.59  
##   
## [[1]]$X1stFlrSF$Moda  
## [[1]]$X1stFlrSF$Moda$Mode  
## [1] 864  
##   
## [[1]]$X1stFlrSF$Moda$Frequency  
## [1] 25

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 79.65  
##   
## x y   
## Min. : 95.04 Min. :0.000e+00   
## 1st Qu.:1304.02 1st Qu.:2.549e-07   
## Median :2513.00 Median :5.676e-06   
## Mean :2513.00 Mean :2.066e-04   
## 3rd Qu.:3721.98 3rd Qu.:2.578e-04   
## Max. :4930.96 Max. :1.237e-03



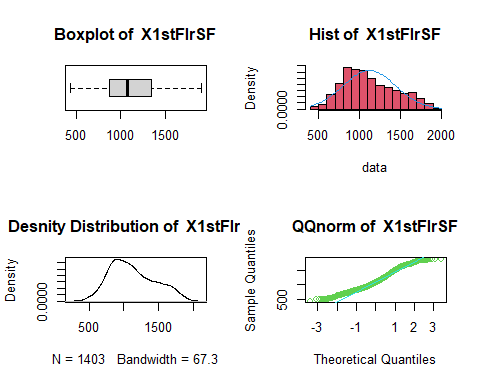
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli edifici ha il primo piano di area compresa tra 882 e 1391 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(HP[,"X1stFlrSF"],.95)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"X1stFlrSF"),"X1stFlrSF")

## $X1stFlrSF  
## $X1stFlrSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 438 874 1072 1122 1344 1905   
##   
## $X1stFlrSF$Variance  
## [1] 101485.5  
##   
## $X1stFlrSF$Standard\_Deviation  
## [1] 318.57  
##   
## $X1stFlrSF$Moda  
## $X1stFlrSF$Moda$Mode  
## [1] 864  
##   
## $X1stFlrSF$Moda$Frequency  
## [1] 25

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1403 obs.); Bandwidth 'bw' = 67.3  
##   
## x y   
## Min. : 236.1 Min. :9.690e-08   
## 1st Qu.: 703.8 1st Qu.:1.096e-04   
## Median :1171.5 Median :5.082e-04   
## Mean :1171.5 Mean :5.340e-04   
## 3rd Qu.:1639.2 3rd Qu.:8.647e-04   
## Max. :2106.9 Max. :1.313e-03



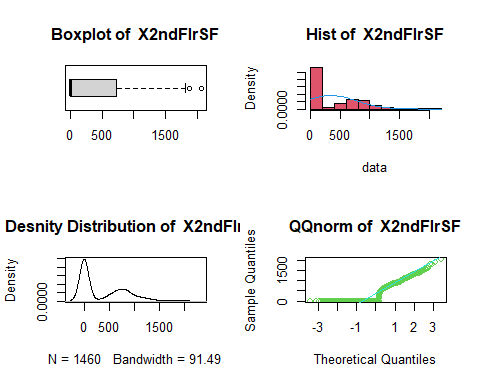
Possiamo notare che la maggior parte degli edifici ha il primo piano di area compresa tra 874 e 1344 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### X2ndFlrSF

visualization\_continuos(na.omit(HP[,"X2ndFlrSF"]), sum\_info\_h[11],"X2ndFlrSF")

## [[1]]  
## [[1]]$X2ndFlrSF  
## [[1]]$X2ndFlrSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 0 0 347 728 2065   
##   
## [[1]]$X2ndFlrSF$Variance  
## [1] 190557.1  
##   
## [[1]]$X2ndFlrSF$Standard\_Deviation  
## [1] 436.53  
##   
## [[1]]$X2ndFlrSF$Moda  
## [[1]]$X2ndFlrSF$Moda$Mode  
## [1] 0  
##   
## [[1]]$X2ndFlrSF$Moda$Frequency  
## [1] 829

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 91.49  
##   
## x y   
## Min. :-274.5 Min. :3.360e-08   
## 1st Qu.: 379.0 1st Qu.:1.064e-05   
## Median :1032.5 Median :1.700e-04   
## Mean :1032.5 Mean :3.819e-04   
## 3rd Qu.:1686.0 3rd Qu.:5.216e-04   
## Max. :2339.5 Max. :2.478e-03



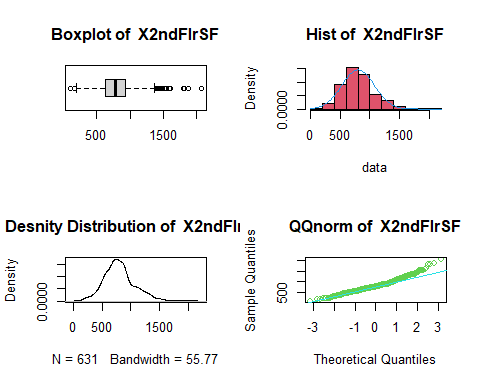
Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area del secondo piano è prelevantemete 0 qundi non costruitto. Per fare un analisi sugli edifici che hanno il secondo piano, serve ignorare quelli senza

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(HP[,"X2ndFlrSF"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"X2ndFlrSF"),"X2ndFlrSF")

## $X2ndFlrSF  
## $X2ndFlrSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 110.0 625.0 776.0 802.9 926.5 2065.0   
##   
## $X2ndFlrSF$Variance  
## [1] 74718.56  
##   
## $X2ndFlrSF$Standard\_Deviation  
## [1] 273.35  
##   
## $X2ndFlrSF$Moda  
## $X2ndFlrSF$Moda$Mode  
## [1] 728  
##   
## $X2ndFlrSF$Moda$Frequency  
## [1] 10

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (631 obs.); Bandwidth 'bw' = 55.77  
##   
## x y   
## Min. : -57.32 Min. :1.284e-07   
## 1st Qu.: 515.09 1st Qu.:1.705e-05   
## Median :1087.50 Median :1.611e-04   
## Mean :1087.50 Mean :4.363e-04   
## 3rd Qu.:1659.91 3rd Qu.:6.302e-04   
## Max. :2232.32 Max. :1.738e-03



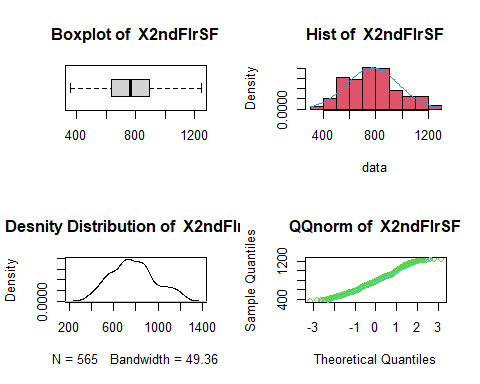
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli edifici ha il secondo piano di area compresa tra 625.0 e 926.5 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(rm\_0\_data,.90)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"X2ndFlrSF"),"X2ndFlrSF")

## $X2ndFlrSF  
## $X2ndFlrSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 358 634 762 776 895 1243   
##   
## $X2ndFlrSF$Variance  
## [1] 39522.8  
##   
## $X2ndFlrSF$Standard\_Deviation  
## [1] 198.8  
##   
## $X2ndFlrSF$Moda  
## $X2ndFlrSF$Moda$Mode  
## [1] 728  
##   
## $X2ndFlrSF$Moda$Frequency  
## [1] 10

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (565 obs.); Bandwidth 'bw' = 49.36  
##   
## x y   
## Min. : 209.9 Min. :5.443e-07   
## 1st Qu.: 505.2 1st Qu.:2.248e-04   
## Median : 800.5 Median :6.737e-04   
## Mean : 800.5 Mean :8.458e-04   
## 3rd Qu.:1095.8 3rd Qu.:1.437e-03   
## Max. :1391.1 Max. :1.975e-03



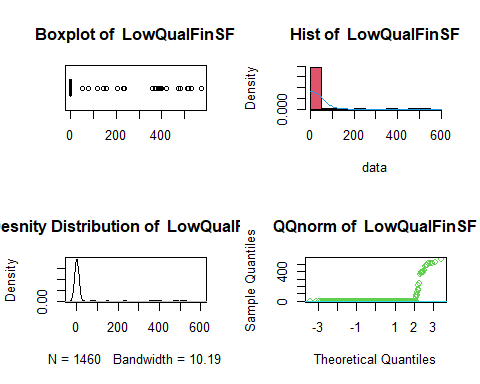
Possiamo notare che la maggior parte degli edifici ha il secondo piano di area compresa tra 634 e 895 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### LowQualFinSF

visualization\_continuos(na.omit(HP[,"LowQualFinSF"]), sum\_info\_h[12],"LowQualFinSF")

## [[1]]  
## [[1]]$LowQualFinSF  
## [[1]]$LowQualFinSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 5.845 0.000 572.000   
##   
## [[1]]$LowQualFinSF$Variance  
## [1] 2364.2  
##   
## [[1]]$LowQualFinSF$Standard\_Deviation  
## [1] 48.62  
##   
## [[1]]$LowQualFinSF$Moda  
## [[1]]$LowQualFinSF$Moda$Mode  
## [1] 0  
##   
## [[1]]$LowQualFinSF$Moda$Frequency  
## [1] 1434

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 10.19  
##   
## x y   
## Min. :-30.57 Min. :0.000e+00   
## 1st Qu.:127.71 1st Qu.:9.700e-06   
## Median :286.00 Median :2.900e-05   
## Mean :286.00 Mean :1.576e-03   
## 3rd Qu.:444.29 3rd Qu.:6.513e-05   
## Max. :602.57 Max. :3.834e-02



Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area dei piani fatti male è prelevantemete 0 Per fare un analisi sugli edifici che hanno i piani fatti male, serve ignorare quelli senza

#### Rimozione degli zero

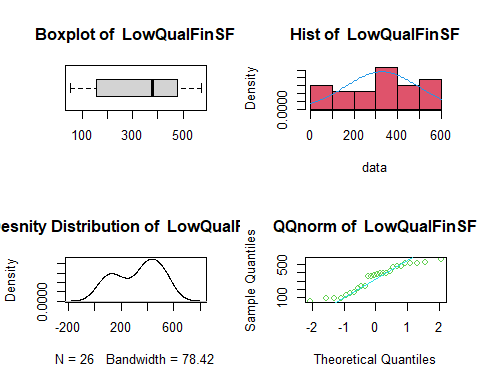
rm\_0\_data = rm\_0(na.omit(HP[,"LowQualFinSF"]))  
print(paste("Ratio: ",length(rm\_0\_data)/length(na.omit(HP[,"LowQualFinSF"]))))

## [1] "Ratio: 0.0178082191780822"

visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"LowQualFinSF"),"LowQualFinSF")

## $LowQualFinSF  
## $LowQualFinSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 53.0 168.2 377.5 328.2 477.5 572.0   
##   
## $LowQualFinSF$Variance  
## [1] 27951.2  
##   
## $LowQualFinSF$Standard\_Deviation  
## [1] 167.19  
##   
## $LowQualFinSF$Moda  
## $LowQualFinSF$Moda$Mode  
## [1] 80  
##   
## $LowQualFinSF$Moda$Frequency  
## [1] 3

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (26 obs.); Bandwidth 'bw' = 78.42  
##   
## x y   
## Min. :-182.27 Min. :3.188e-06   
## 1st Qu.: 65.11 1st Qu.:1.957e-04   
## Median : 312.50 Median :1.231e-03   
## Mean : 312.50 Mean :1.009e-03   
## 3rd Qu.: 559.89 3rd Qu.:1.471e-03   
## Max. : 807.27 Max. :2.264e-03



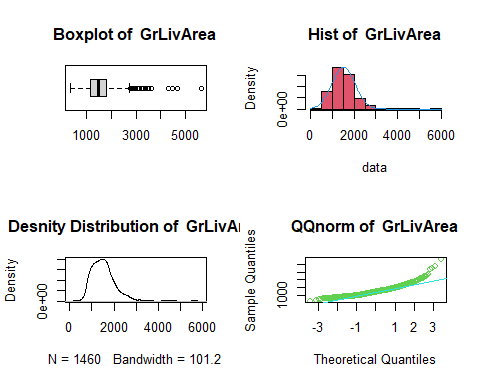
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli edifici ha i piani fatti male di area compresa tra 168.2 e 477.5 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Gli edifici che hanno i piani fatti male sono cosi piccoli (rispetto al totale ~=1,8%) che si puo trascurare

### GrLivArea

visualization\_continuos(na.omit(HP[,"GrLivArea"]), sum\_info\_h[13],"GrLivArea")

## [[1]]  
## [[1]]$GrLivArea  
## [[1]]$GrLivArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 334 1130 1464 1515 1777 5642   
##   
## [[1]]$GrLivArea$Variance  
## [1] 276129.6  
##   
## [[1]]$GrLivArea$Standard\_Deviation  
## [1] 525.48  
##   
## [[1]]$GrLivArea$Moda  
## [[1]]$GrLivArea$Moda$Mode  
## [1] 864  
##   
## [[1]]$GrLivArea$Moda$Frequency  
## [1] 22

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 101.2  
##   
## x y   
## Min. : 30.29 Min. :1.000e-10   
## 1st Qu.:1509.14 1st Qu.:1.295e-06   
## Median :2988.00 Median :1.176e-05   
## Mean :2988.00 Mean :1.689e-04   
## 3rd Qu.:4466.85 3rd Qu.:2.169e-04   
## Max. :5945.71 Max. :8.045e-04



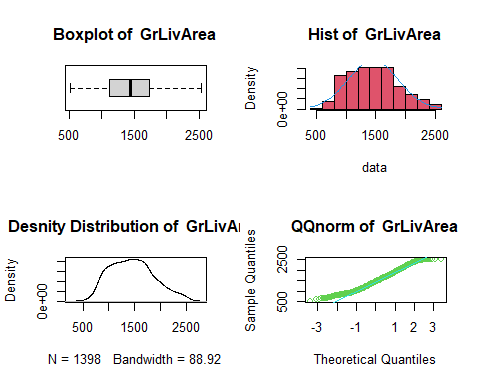
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli edifici che sono sopra il seminterrato sono di area compresa tra 1130 e 1777 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(na.omit(HP[,"GrLivArea"]),.95)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"GrLivArea"),"GrLivArea")

## $GrLivArea  
## $GrLivArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 520 1121 1440 1456 1728 2531   
##   
## $GrLivArea$Variance  
## [1] 176895.8  
##   
## $GrLivArea$Standard\_Deviation  
## [1] 420.59  
##   
## $GrLivArea$Moda  
## $GrLivArea$Moda$Mode  
## [1] 864  
##   
## $GrLivArea$Moda$Frequency  
## [1] 22

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1398 obs.); Bandwidth 'bw' = 88.92  
##   
## x y   
## Min. : 253.2 Min. :4.100e-08   
## 1st Qu.: 889.4 1st Qu.:7.707e-05   
## Median :1525.5 Median :3.389e-04   
## Mean :1525.5 Mean :3.926e-04   
## 3rd Qu.:2161.6 3rd Qu.:7.425e-04   
## Max. :2797.8 Max. :8.456e-04



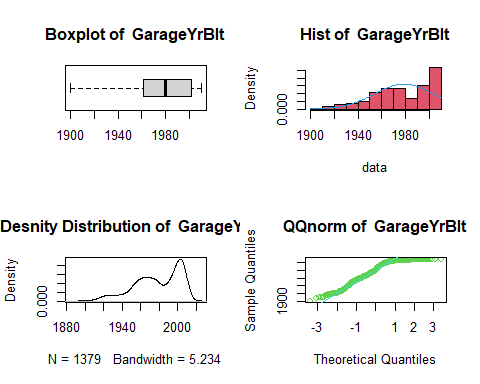
Possiamo notare che la maggior parte degli edifici che sono sopra il seminterrato sono compresi tra 1121 e 1728 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### GarageYrBlt

visualization\_continuos(na.omit(HP[,"GarageYrBlt"]), sum\_info\_h[22],"GarageYrBlt")

## [[1]]  
## [[1]]$GarageYrBlt  
## [[1]]$GarageYrBlt$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1900 1961 1980 1979 2002 2010   
##   
## [[1]]$GarageYrBlt$Variance  
## [1] 609.58  
##   
## [[1]]$GarageYrBlt$Standard\_Deviation  
## [1] 24.69  
##   
## [[1]]$GarageYrBlt$Moda  
## [[1]]$GarageYrBlt$Moda$Mode  
## [1] 2005  
##   
## [[1]]$GarageYrBlt$Moda$Frequency  
## [1] 65

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1379 obs.); Bandwidth 'bw' = 5.234  
##   
## x y   
## Min. :1884 Min. :6.370e-07   
## 1st Qu.:1920 1st Qu.:1.156e-03   
## Median :1955 Median :4.492e-03   
## Mean :1955 Mean :7.065e-03   
## 3rd Qu.:1990 3rd Qu.:1.213e-02   
## Max. :2026 Max. :2.341e-02



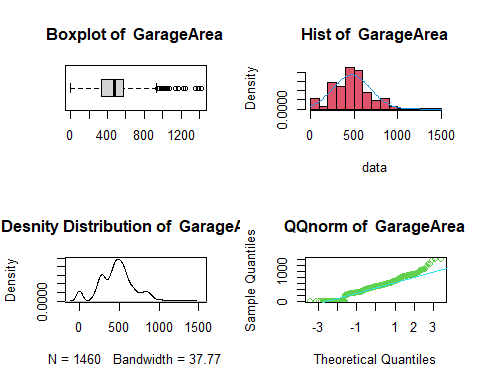
Possiamo dedurre che la distribuzione è positivamente asimetrica e quindi il numero di garage costruiti crescono ogni anno La maggior parte degli garage erano costruiti fra 1961 e 2002, essendo rispettivamente il 1° e il 3° quantile e l’anno con è il massimo costruitto è nel 2005 con 65 garage

### GarageArea

visualization\_continuos(na.omit(HP[,"GarageArea"]), sum\_info\_h[24],"GarageArea")

## [[1]]  
## [[1]]$GarageArea  
## [[1]]$GarageArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0 334.5 480.0 473.0 576.0 1418.0   
##   
## [[1]]$GarageArea$Variance  
## [1] 45712.51  
##   
## [[1]]$GarageArea$Standard\_Deviation  
## [1] 213.8  
##   
## [[1]]$GarageArea$Moda  
## [[1]]$GarageArea$Moda$Mode  
## [1] 0  
##   
## [[1]]$GarageArea$Moda$Frequency  
## [1] 81

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 37.77  
##   
## x y   
## Min. :-113.3 Min. :8.960e-08   
## 1st Qu.: 297.8 1st Qu.:1.752e-05   
## Median : 709.0 Median :3.136e-04   
## Mean : 709.0 Mean :6.074e-04   
## 3rd Qu.:1120.2 3rd Qu.:1.028e-03   
## Max. :1531.3 Max. :2.465e-03



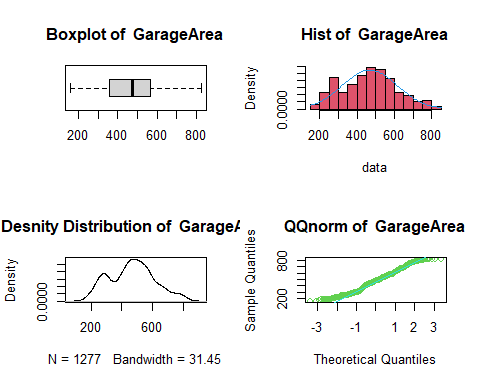
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte degli garage sono di area compresa tra 334.5 e 576.0 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(na.omit(HP[,"GarageArea"]),.90)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"GarageArea"),"GarageArea")

## $GarageArea  
## $GarageArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 160.0 358.0 476.0 467.7 566.0 824.0   
##   
## $GarageArea$Variance  
## [1] 21343.39  
##   
## $GarageArea$Standard\_Deviation  
## [1] 146.09  
##   
## $GarageArea$Moda  
## $GarageArea$Moda$Mode  
## [1] 440  
##   
## $GarageArea$Moda$Frequency  
## [1] 49

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1277 obs.); Bandwidth 'bw' = 31.45  
##   
## x y   
## Min. : 65.65 Min. :4.482e-07   
## 1st Qu.:278.82 1st Qu.:3.274e-04   
## Median :492.00 Median :1.095e-03   
## Mean :492.00 Mean :1.172e-03   
## 3rd Qu.:705.18 3rd Qu.:1.846e-03   
## Max. :918.35 Max. :2.875e-03



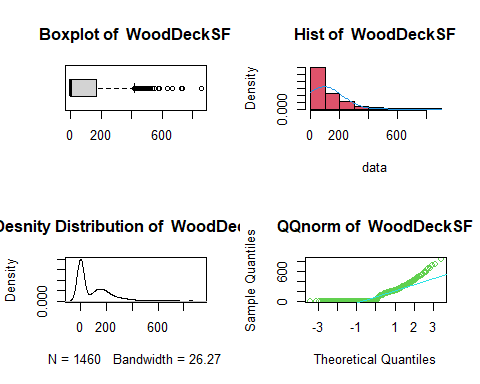
Possiamo notare che la maggior parte degli garage sono compresi tra 358.0 e 566.0 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### WoodDeckSF

visualization\_continuos(na.omit(HP[,"WoodDeckSF"]), sum\_info\_h[25],"WoodDeckSF")

## [[1]]  
## [[1]]$WoodDeckSF  
## [[1]]$WoodDeckSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 94.24 168.00 857.00   
##   
## [[1]]$WoodDeckSF$Variance  
## [1] 15709.81  
##   
## [[1]]$WoodDeckSF$Standard\_Deviation  
## [1] 125.34  
##   
## [[1]]$WoodDeckSF$Moda  
## [[1]]$WoodDeckSF$Moda$Mode  
## [1] 0  
##   
## [[1]]$WoodDeckSF$Moda$Frequency  
## [1] 761

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 26.27  
##   
## x y   
## Min. :-78.81 Min. :1.180e-07   
## 1st Qu.:174.85 1st Qu.:2.217e-05   
## Median :428.50 Median :1.736e-04   
## Mean :428.50 Mean :9.840e-04   
## 3rd Qu.:682.15 3rd Qu.:1.377e-03   
## Max. :935.81 Max. :8.060e-03



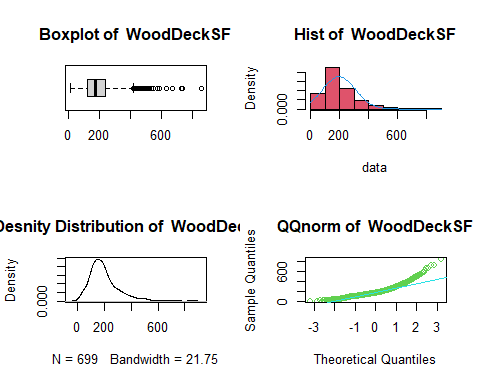
Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area del ponte di legno è prelevantemete 0 con 761 occorrenze qundi non costruitto. Per fare un analisi sui ponti di legno costruiti, serve ignorare quelli non

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(HP[,"WoodDeckSF"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"WoodDeckSF"),"WoodDeckSF")

## $WoodDeckSF  
## $WoodDeckSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 12.0 120.0 171.0 196.8 240.0 857.0   
##   
## $WoodDeckSF$Variance  
## [1] 12611.23  
##   
## $WoodDeckSF$Standard\_Deviation  
## [1] 112.3  
##   
## $WoodDeckSF$Moda  
## $WoodDeckSF$Moda$Mode  
## [1] 192  
##   
## $WoodDeckSF$Moda$Frequency  
## [1] 38

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (699 obs.); Bandwidth 'bw' = 21.75  
##   
## x y   
## Min. :-53.24 Min. :3.000e-07   
## 1st Qu.:190.63 1st Qu.:4.159e-05   
## Median :434.50 Median :2.577e-04   
## Mean :434.50 Mean :1.024e-03   
## 3rd Qu.:678.37 3rd Qu.:1.424e-03   
## Max. :922.24 Max. :4.913e-03



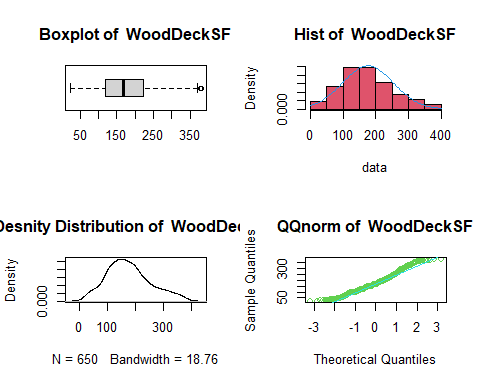
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte dei ponte di legno sono di area compresa tra 120.0 e 240.0 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(rm\_0\_data,.90)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"WoodDeckSF"),"WoodDeckSF")

## $WoodDeckSF  
## $WoodDeckSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 24.0 120.0 168.0 176.3 222.0 380.0   
##   
## $WoodDeckSF$Variance  
## [1] 6040.51  
##   
## $WoodDeckSF$Standard\_Deviation  
## [1] 77.72  
##   
## $WoodDeckSF$Moda  
## $WoodDeckSF$Moda$Mode  
## [1] 192  
##   
## $WoodDeckSF$Moda$Frequency  
## [1] 38

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (650 obs.); Bandwidth 'bw' = 18.76  
##   
## x y   
## Min. :-32.27 Min. :1.060e-06   
## 1st Qu.: 84.86 1st Qu.:6.233e-04   
## Median :202.00 Median :1.584e-03   
## Mean :202.00 Mean :2.132e-03   
## 3rd Qu.:319.13 3rd Qu.:3.725e-03   
## Max. :436.27 Max. :5.362e-03



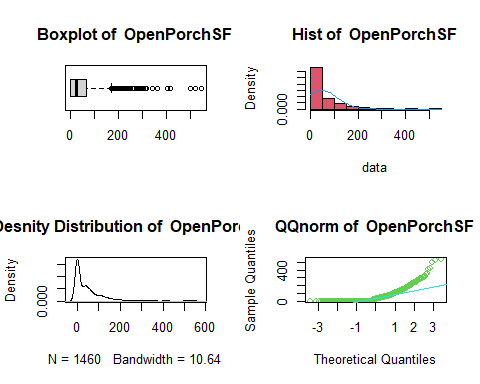
Possiamo notare che la maggior parte dei ponte di legno sono di area compresa tra 120.0 e 220.5 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### OpenPorchSF

visualization\_continuos(na.omit(HP[,"OpenPorchSF"]), sum\_info\_h[26],"OpenPorchSF")

## [[1]]  
## [[1]]$OpenPorchSF  
## [[1]]$OpenPorchSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 25.00 46.66 68.00 547.00   
##   
## [[1]]$OpenPorchSF$Variance  
## [1] 4389.86  
##   
## [[1]]$OpenPorchSF$Standard\_Deviation  
## [1] 66.26  
##   
## [[1]]$OpenPorchSF$Moda  
## [[1]]$OpenPorchSF$Moda$Mode  
## [1] 0  
##   
## [[1]]$OpenPorchSF$Moda$Frequency  
## [1] 656

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 10.64  
##   
## x y   
## Min. :-31.91 Min. :2.300e-08   
## 1st Qu.:120.80 1st Qu.:2.790e-05   
## Median :273.50 Median :1.925e-04   
## Mean :273.50 Mean :1.635e-03   
## 3rd Qu.:426.20 3rd Qu.:1.603e-03   
## Max. :578.91 Max. :1.730e-02



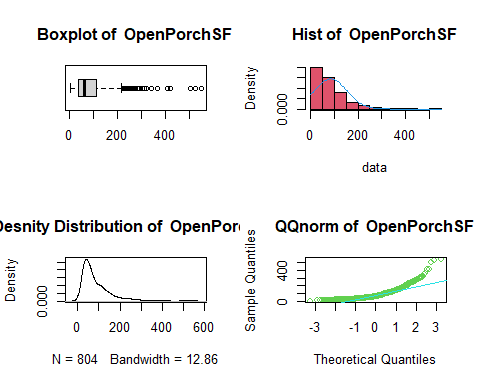
Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area del portico aperto è prelevantemete 0 con 656 occorrenze qundi non costruitto. Per fare un analisi sui portico aperti, serve ignorare quelli non

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(HP[,"OpenPorchSF"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"OpenPorchSF"),"OpenPorchSF")

## $OpenPorchSF  
## $OpenPorchSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.00 39.00 63.00 84.73 112.00 547.00   
##   
## $OpenPorchSF$Variance  
## [1] 4746.27  
##   
## $OpenPorchSF$Standard\_Deviation  
## [1] 68.89  
##   
## $OpenPorchSF$Moda  
## $OpenPorchSF$Moda$Mode  
## [1] 36  
##   
## $OpenPorchSF$Moda$Frequency  
## [1] 29

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (804 obs.); Bandwidth 'bw' = 12.86  
##   
## x y   
## Min. :-34.59 Min. :3.960e-07   
## 1st Qu.:120.45 1st Qu.:4.309e-05   
## Median :275.50 Median :2.546e-04   
## Mean :275.50 Mean :1.611e-03   
## 3rd Qu.:430.55 3rd Qu.:1.808e-03   
## Max. :585.60 Max. :1.078e-02



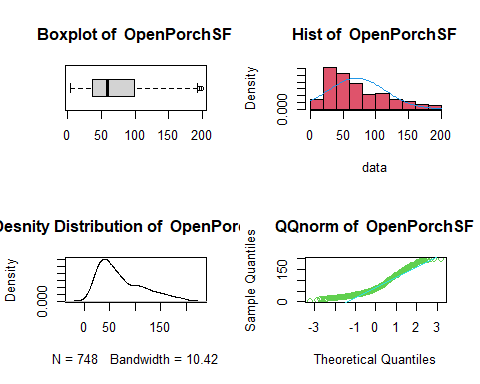
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte dei portico aperti sono di area compresa tra 39.00 e 112.00 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(rm\_0\_data,.90)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"OpenPorchSF"),"OpenPorchSF")

## $OpenPorchSF  
## $OpenPorchSF$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4.00 36.00 59.00 70.86 98.25 198.00   
##   
## $OpenPorchSF$Variance  
## [1] 1891.16  
##   
## $OpenPorchSF$Standard\_Deviation  
## [1] 43.49  
##   
## $OpenPorchSF$Moda  
## $OpenPorchSF$Moda$Mode  
## [1] 36  
##   
## $OpenPorchSF$Moda$Frequency  
## [1] 29

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (748 obs.); Bandwidth 'bw' = 10.42  
##   
## x y   
## Min. :-27.26 Min. :1.210e-06   
## 1st Qu.: 36.87 1st Qu.:7.911e-04   
## Median :101.00 Median :2.782e-03   
## Mean :101.00 Mean :3.895e-03   
## 3rd Qu.:165.13 3rd Qu.:5.686e-03   
## Max. :229.26 Max. :1.222e-02



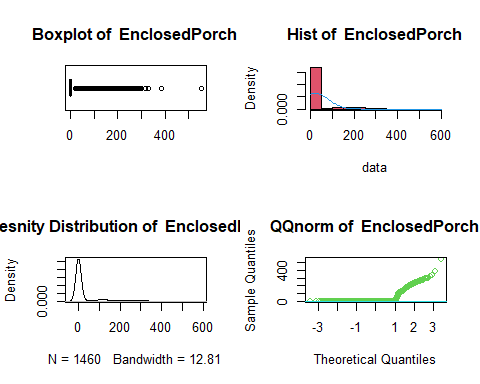
Possiamo notare che la maggior parte dei portico aperti sono di area compresa tra 120.0 e 220.5 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### EnclosedPorch

visualization\_continuos(na.omit(HP[,"EnclosedPorch"]), sum\_info\_h[27],"EnclosedPorch")

## [[1]]  
## [[1]]$EnclosedPorch  
## [[1]]$EnclosedPorch$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 21.95 0.00 552.00   
##   
## [[1]]$EnclosedPorch$Variance  
## [1] 3735.55  
##   
## [[1]]$EnclosedPorch$Standard\_Deviation  
## [1] 61.12  
##   
## [[1]]$EnclosedPorch$Moda  
## [[1]]$EnclosedPorch$Moda$Mode  
## [1] 0  
##   
## [[1]]$EnclosedPorch$Moda$Frequency  
## [1] 1252

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 12.81  
##   
## x y   
## Min. :-38.43 Min. :0.000e+00   
## 1st Qu.:118.78 1st Qu.:6.346e-06   
## Median :276.00 Median :1.886e-04   
## Mean :276.00 Mean :1.587e-03   
## 3rd Qu.:433.21 3rd Qu.:6.071e-04   
## Max. :590.43 Max. :2.664e-02



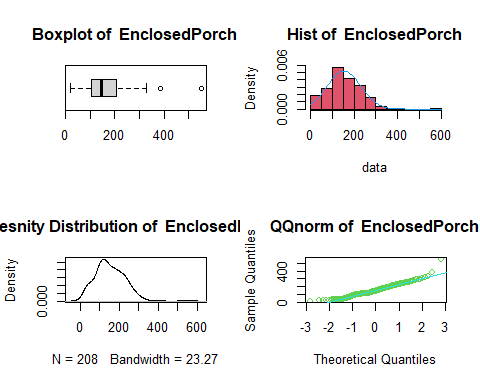
Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area del portico chiuso è prelevantemete 0 con 1252 occorrenze qundi non costruitto. Per fare un analisi sui portico chiusi, serve ignorare quelli non

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(HP[,"EnclosedPorch"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"EnclosedPorch"),"EnclosedPorch")

## $EnclosedPorch  
## $EnclosedPorch$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.0 104.2 144.5 154.1 205.0 552.0   
##   
## $EnclosedPorch$Variance  
## [1] 5866.98  
##   
## $EnclosedPorch$Standard\_Deviation  
## [1] 76.6  
##   
## $EnclosedPorch$Moda  
## $EnclosedPorch$Moda$Mode  
## [1] 112  
##   
## $EnclosedPorch$Moda$Frequency  
## [1] 15

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (208 obs.); Bandwidth 'bw' = 23.27  
##   
## x y   
## Min. :-50.81 Min. :2.910e-07   
## 1st Qu.:117.35 1st Qu.:3.794e-05   
## Median :285.50 Median :3.211e-04   
## Mean :285.50 Mean :1.485e-03   
## 3rd Qu.:453.65 3rd Qu.:2.944e-03   
## Max. :621.81 Max. :5.414e-03



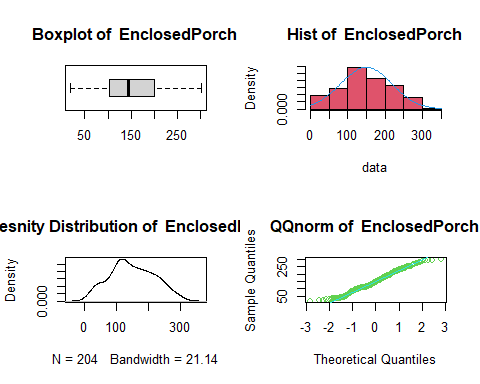
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte dei portico chiusi sono di area compresa tra 104.2 e 205.0 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(rm\_0\_data,.95)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"EnclosedPorch"),"EnclosedPorch")

## $EnclosedPorch  
## $EnclosedPorch$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 19.0 102.0 144.0 149.3 200.5 301.0   
##   
## $EnclosedPorch$Variance  
## [1] 4630.31  
##   
## $EnclosedPorch$Standard\_Deviation  
## [1] 68.05  
##   
## $EnclosedPorch$Moda  
## $EnclosedPorch$Moda$Mode  
## [1] 112  
##   
## $EnclosedPorch$Moda$Frequency  
## [1] 15

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (204 obs.); Bandwidth 'bw' = 21.14  
##   
## x y   
## Min. :-44.42 Min. :2.099e-06   
## 1st Qu.: 57.79 1st Qu.:5.220e-04   
## Median :160.00 Median :2.447e-03   
## Mean :160.00 Mean :2.443e-03   
## 3rd Qu.:262.21 3rd Qu.:4.105e-03   
## Max. :364.42 Max. :5.684e-03



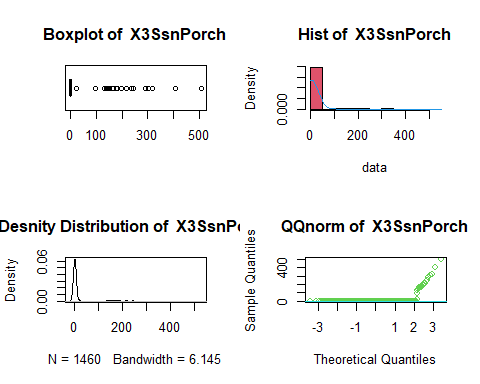
Possiamo notare che la maggior parte dei portico chiusi sono di area compresa tra 102.0 e 200.5 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### X3SsnPorch

visualization\_continuos(na.omit(HP[,"X3SsnPorch"]), sum\_info\_h[28],"X3SsnPorch")

## [[1]]  
## [[1]]$X3SsnPorch  
## [[1]]$X3SsnPorch$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 3.41 0.00 508.00   
##   
## [[1]]$X3SsnPorch$Variance  
## [1] 859.51  
##   
## [[1]]$X3SsnPorch$Standard\_Deviation  
## [1] 29.32  
##   
## [[1]]$X3SsnPorch$Moda  
## [[1]]$X3SsnPorch$Moda$Mode  
## [1] 0  
##   
## [[1]]$X3SsnPorch$Moda$Frequency  
## [1] 1436

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 6.145  
##   
## x y   
## Min. :-18.43 Min. :0.000e+00   
## 1st Qu.:117.78 1st Qu.:1.300e-07   
## Median :254.00 Median :2.019e-05   
## Mean :254.00 Mean :1.832e-03   
## 3rd Qu.:390.22 3rd Qu.:5.386e-05   
## Max. :526.43 Max. :6.356e-02



Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area del Portico per tre stagioni è prelevantemete 0 con 1252 occorrenze qundi non costruitto. Per fare un analisi sui Portico per tre stagioni, serve ignorare quelli non

#### Rimozione degli zero

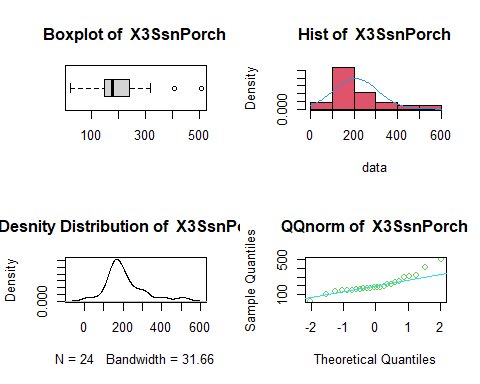
rm\_0\_data = rm\_0(na.omit(HP[,"X3SsnPorch"]))  
print(paste("Ratio: ",length(rm\_0\_data)/length(na.omit(HP[,"X3SsnPorch"]))))

## [1] "Ratio: 0.0164383561643836"

visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"X3SsnPorch"),"X3SsnPorch")

## $X3SsnPorch  
## $X3SsnPorch$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 23.0 150.8 180.0 207.4 239.8 508.0   
##   
## $X3SsnPorch$Variance  
## [1] 10368.34  
##   
## $X3SsnPorch$Standard\_Deviation  
## [1] 101.83  
##   
## $X3SsnPorch$Moda  
## $X3SsnPorch$Moda$Mode  
## [1] 168  
##   
## $X3SsnPorch$Moda$Frequency  
## [1] 3

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (24 obs.); Bandwidth 'bw' = 31.66  
##   
## x y   
## Min. :-71.97 Min. :0.0000059   
## 1st Qu.: 96.76 1st Qu.:0.0003565   
## Median :265.50 Median :0.0005635   
## Mean :265.50 Mean :0.0014800   
## 3rd Qu.:434.24 3rd Qu.:0.0019201   
## Max. :602.97 Max. :0.0062139



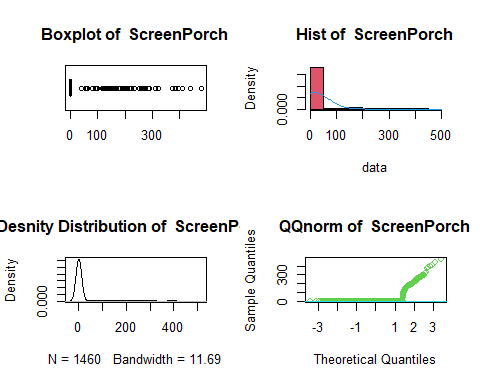
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte dei Portico per tre stagioni sono di area compresa tra 150.8 e 239.8 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Gli Portico per tre stagioni costruiti sono cosi piccoli (rispetto al totale ~=1,6%) che si puo trascurare

### ScreenPorch

visualization\_continuos(na.omit(HP[,"ScreenPorch"]), sum\_info\_h[29],"ScreenPorch")

## [[1]]  
## [[1]]$ScreenPorch  
## [[1]]$ScreenPorch$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 15.06 0.00 480.00   
##   
## [[1]]$ScreenPorch$Variance  
## [1] 3108.89  
##   
## [[1]]$ScreenPorch$Standard\_Deviation  
## [1] 55.76  
##   
## [[1]]$ScreenPorch$Moda  
## [[1]]$ScreenPorch$Moda$Mode  
## [1] 0  
##   
## [[1]]$ScreenPorch$Moda$Frequency  
## [1] 1344

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 11.69  
##   
## x y   
## Min. :-35.06 Min. :2.640e-07   
## 1st Qu.:102.47 1st Qu.:2.890e-05   
## Median :240.00 Median :1.415e-04   
## Mean :240.00 Mean :1.814e-03   
## 3rd Qu.:377.53 3rd Qu.:4.037e-04   
## Max. :515.06 Max. :3.135e-02



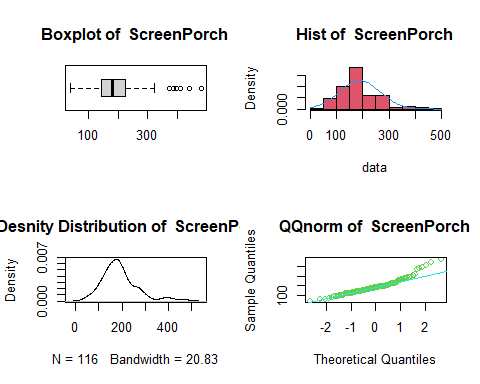
Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area del Portico dello schermo è prelevantemete 0 con 1344 occorrenze qundi non costruitto. Per fare un analisi sui Portico dello schermo, serve ignorare quelli non

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(HP[,"ScreenPorch"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"ScreenPorch"),"ScreenPorch")

## $ScreenPorch  
## $ScreenPorch$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 40.0 143.8 180.0 189.6 224.0 480.0   
##   
## $ScreenPorch$Variance  
## [1] 6076.54  
##   
## $ScreenPorch$Standard\_Deviation  
## [1] 77.95  
##   
## $ScreenPorch$Moda  
## $ScreenPorch$Moda$Mode  
## [1] 192  
##   
## $ScreenPorch$Moda$Frequency  
## [1] 6

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (116 obs.); Bandwidth 'bw' = 20.83  
##   
## x y   
## Min. :-22.49 Min. :1.862e-06   
## 1st Qu.:118.75 1st Qu.:2.460e-04   
## Median :260.00 Median :5.576e-04   
## Mean :260.00 Mean :1.768e-03   
## 3rd Qu.:401.25 3rd Qu.:2.672e-03   
## Max. :542.49 Max. :6.750e-03



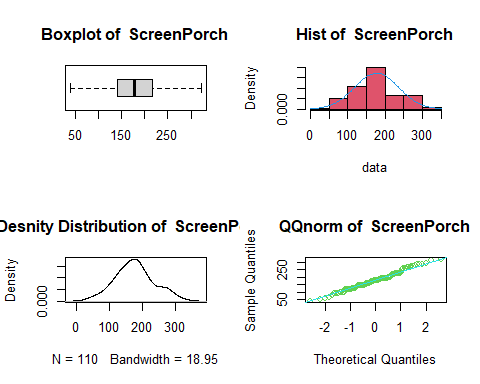
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte dei Portico dello schermo stagioni sono di area compresa tra 143.8 e 224.0 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(rm\_0\_data,.95)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"ScreenPorch"),"ScreenPorch")

## $ScreenPorch  
## $ScreenPorch$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 40.0 142.2 177.0 177.3 214.5 322.0   
##   
## $ScreenPorch$Variance  
## [1] 3410.93  
##   
## $ScreenPorch$Standard\_Deviation  
## [1] 58.4  
##   
## $ScreenPorch$Moda  
## $ScreenPorch$Moda$Mode  
## [1] 192  
##   
## $ScreenPorch$Moda$Frequency  
## [1] 6

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (110 obs.); Bandwidth 'bw' = 18.95  
##   
## x y   
## Min. :-16.86 Min. :2.451e-06   
## 1st Qu.: 82.07 1st Qu.:4.027e-04   
## Median :181.00 Median :1.985e-03   
## Mean :181.00 Mean :2.524e-03   
## 3rd Qu.:279.93 3rd Qu.:4.225e-03   
## Max. :378.86 Max. :7.267e-03



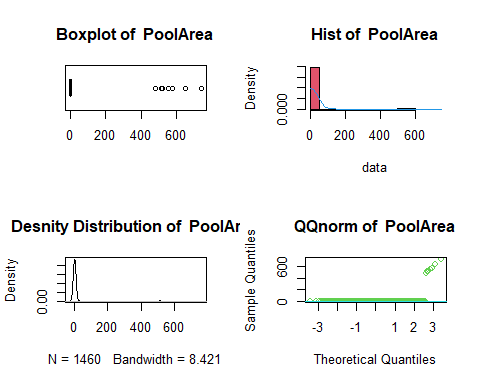
Possiamo notare che la maggior parte dei Portico dello schermo sono di area compresa tra 142.2 e 214.5 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

### PoolArea

visualization\_continuos(na.omit(HP[,"PoolArea"]), sum\_info\_h[30],"PoolArea")

## [[1]]  
## [[1]]$PoolArea  
## [[1]]$PoolArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 0.000 0.000 2.759 0.000 738.000   
##   
## [[1]]$PoolArea$Variance  
## [1] 1614.22  
##   
## [[1]]$PoolArea$Standard\_Deviation  
## [1] 40.18  
##   
## [[1]]$PoolArea$Moda  
## [[1]]$PoolArea$Moda$Mode  
## [1] 0  
##   
## [[1]]$PoolArea$Moda$Frequency  
## [1] 1453

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 8.421  
##   
## x y   
## Min. :-25.26 Min. :0.0000000   
## 1st Qu.:171.87 1st Qu.:0.0000000   
## Median :369.00 Median :0.0000000   
## Mean :369.00 Mean :0.0012658   
## 3rd Qu.:566.13 3rd Qu.:0.0000153   
## Max. :763.26 Max. :0.0467836



Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che l’area della piscina è prelevantemete 0 con 1453 occorrenze qundi non costruitto. Per fare un analisi sulle piscine, serve ignorare quelli non

#### Rimozione degli zero

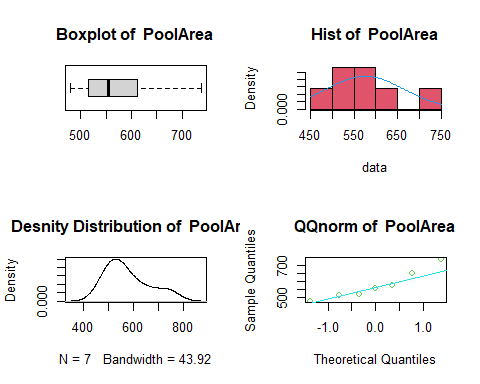
rm\_0\_data = rm\_0(na.omit(HP[,"PoolArea"]))  
print(paste("Ration: ", length(rm\_0\_data)/length(na.omit(HP[,"PoolArea"]))))

## [1] "Ration: 0.00479452054794521"

visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"PoolArea"),"PoolArea")

## $PoolArea  
## $PoolArea$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 480.0 515.5 555.0 575.4 612.0 738.0   
##   
## $PoolArea$Variance  
## [1] 8071.29  
##   
## $PoolArea$Standard\_Deviation  
## [1] 89.84  
##   
## $PoolArea$Moda  
## $PoolArea$Moda$Mode  
## [1] 512 648 576 555 480 519 738  
##   
## $PoolArea$Moda$Frequency  
## [1] 1

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (7 obs.); Bandwidth 'bw' = 43.92  
##   
## x y   
## Min. :348.2 Min. :1.456e-05   
## 1st Qu.:478.6 1st Qu.:4.850e-04   
## Median :609.0 Median :1.564e-03   
## Mean :609.0 Mean :1.915e-03   
## 3rd Qu.:739.4 3rd Qu.:3.040e-03   
## Max. :869.8 Max. :5.021e-03



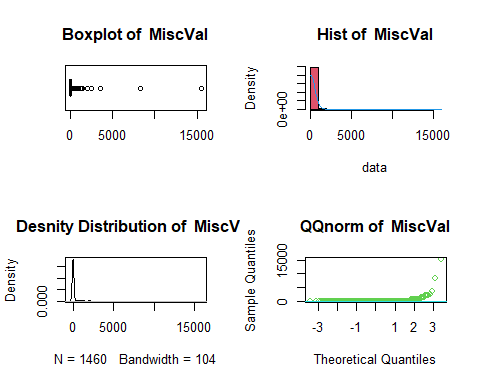
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte delle piscine sono di area compresa tra 515.5 e 612.0 piedi^2, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Gli Portico per tre stagioni costruiti sono cosi piccoli (rispetto al totale <1%) che si puo trascurare

### MiscVal

visualization\_continuos(na.omit(HP[,"MiscVal"]), sum\_info\_h[31],"MiscVal")

## [[1]]  
## [[1]]$MiscVal  
## [[1]]$MiscVal$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 0.00 0.00 43.49 0.00 15500.00   
##   
## [[1]]$MiscVal$Variance  
## [1] 246138.1  
##   
## [[1]]$MiscVal$Standard\_Deviation  
## [1] 496.12  
##   
## [[1]]$MiscVal$Moda  
## [[1]]$MiscVal$Moda$Mode  
## [1] 0  
##   
## [[1]]$MiscVal$Moda$Frequency  
## [1] 1408

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 104  
##   
## x y   
## Min. : -311.9 Min. :0.000e+00   
## 1st Qu.: 3719.0 1st Qu.:0.000e+00   
## Median : 7750.0 Median :0.000e+00   
## Mean : 7750.0 Mean :6.210e-05   
## 3rd Qu.:11781.0 3rd Qu.:5.010e-07   
## Max. :15811.9 Max. :3.682e-03



Possiamo dedurre che la distribuzione è positivamente asimetrica. Si puo osservare che il costo delle funzionalità varie è prelevantemete 0 con 1453 occorrenze qundi non costruitto. Per fare un analisi suli costi delle funzionalità varie, serve ignorare quelli non

#### Rimozione degli zero

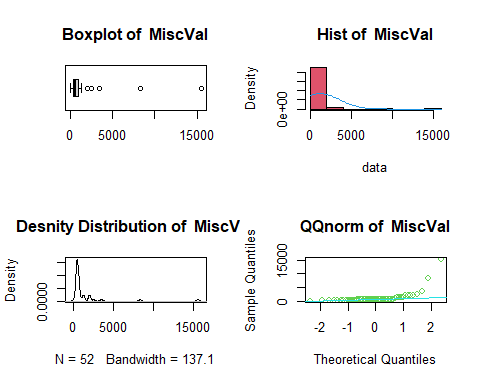
rm\_0\_data = rm\_0(na.omit(HP[,"MiscVal"]))  
print(paste("Ratio: ", length(rm\_0\_data)/length(na.omit(HP[,"MiscVal"]))))

## [1] "Ratio: 0.0356164383561644"

visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"MiscVal"),"MiscVal")

## $MiscVal  
## $MiscVal$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 54.0 437.5 500.0 1221.0 887.5 15500.0   
##   
## $MiscVal$Variance  
## [1] 5575453  
##   
## $MiscVal$Standard\_Deviation  
## [1] 2361.24  
##   
## $MiscVal$Moda  
## $MiscVal$Moda$Mode  
## [1] 400  
##   
## $MiscVal$Moda$Frequency  
## [1] 11

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (52 obs.); Bandwidth 'bw' = 137.1  
##   
## x y   
## Min. : -357.4 Min. :0.000e+00   
## 1st Qu.: 3709.8 1st Qu.:0.000e+00   
## Median : 7777.0 Median :0.000e+00   
## Mean : 7777.0 Mean :6.137e-05   
## 3rd Qu.:11844.2 3rd Qu.:1.989e-05   
## Max. :15911.4 Max. :1.626e-03



print(paste("Not in IQR: ",rm\_0\_data[rm\_0\_data<437.5]+rm\_0\_data[rm\_0\_data>887.5]))

## [1] "Not in IQR: 15850" "Not in IQR: 1600" "Not in IQR: 2400"   
## [4] "Not in IQR: 2400" "Not in IQR: 3900" "Not in IQR: 1700"   
## [7] "Not in IQR: 1600" "Not in IQR: 2400" "Not in IQR: 1454"   
## [10] "Not in IQR: 8700" "Not in IQR: 1550" "Not in IQR: 2400"   
## [13] "Not in IQR: 2900"

print(paste("In IQR: ",rm\_0\_data[rm\_0\_data>=437.5 & rm\_0\_data<=887.5]))

## [1] "In IQR: 700" "In IQR: 700" "In IQR: 500" "In IQR: 700" "In IQR: 480"  
## [6] "In IQR: 450" "In IQR: 450" "In IQR: 500" "In IQR: 450" "In IQR: 700"  
## [11] "In IQR: 800" "In IQR: 480" "In IQR: 600" "In IQR: 500" "In IQR: 600"  
## [16] "In IQR: 600" "In IQR: 500" "In IQR: 450" "In IQR: 500" "In IQR: 500"  
## [21] "In IQR: 500" "In IQR: 620" "In IQR: 560" "In IQR: 500" "In IQR: 700"  
## [26] "In IQR: 600"

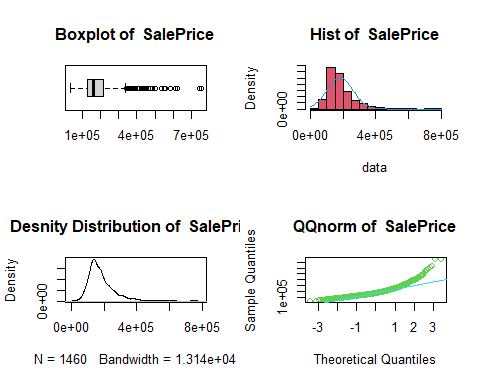
Possiamo notare che la maggior parte dei costi delle funzionalità varie sono comprese tra 437.5 e 887.5 , essendo rispettivamente il 1° e il 3° quantile. Dato la variabilita dei dati e la mancanza di coerenza, dovuta alla mancanza di dati (~=3.56%) dei dati totali, allora si potrebbe trascurare

### SalePrice

visualization\_continuos(na.omit(HP[,"SalePrice"]), sum\_info\_h[34],"SalePrice")

## [[1]]  
## [[1]]$SalePrice  
## [[1]]$SalePrice$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 34900 129975 163000 180921 214000 755000   
##   
## [[1]]$SalePrice$Variance  
## [1] 6311111264  
##   
## [[1]]$SalePrice$Standard\_Deviation  
## [1] 79442.5  
##   
## [[1]]$SalePrice$Moda  
## [[1]]$SalePrice$Moda$Mode  
## [1] 140000  
##   
## [[1]]$SalePrice$Moda$Frequency  
## [1] 20

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1460 obs.); Bandwidth 'bw' = 1.314e+04  
##   
## x y   
## Min. : -4527 Min. :1.000e-12   
## 1st Qu.:195211 1st Qu.:2.700e-08   
## Median :394950 Median :1.687e-07   
## Mean :394950 Mean :1.250e-06   
## 3rd Qu.:594689 3rd Qu.:1.535e-06   
## Max. :794427 Max. :7.682e-06



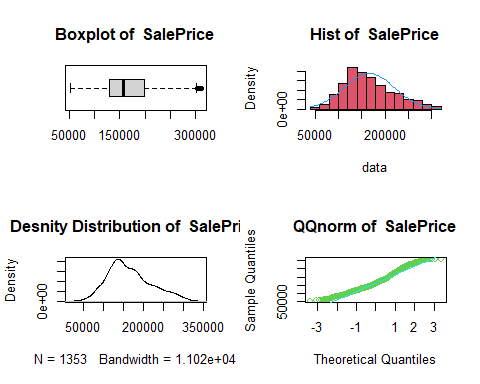
Possiamo dedurre che la distribuzione è positivamente asimetrica e che la maggior parte del prezzo di vendita è compresa tra 129975 e 214000, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Rimangono ancora degli outlier

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(na.omit(HP[,"SalePrice"]),.90)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"SalePrice"),"SalePrice")

## $SalePrice  
## $SalePrice$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 52000 128500 157500 166074 197900 311500   
##   
## $SalePrice$Variance  
## [1] 2827579437  
##   
## $SalePrice$Standard\_Deviation  
## [1] 53174.99  
##   
## $SalePrice$Moda  
## $SalePrice$Moda$Mode  
## [1] 140000  
##   
## $SalePrice$Moda$Frequency  
## [1] 20

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (1353 obs.); Bandwidth 'bw' = 1.102e+04  
##   
## x y   
## Min. : 18936 Min. :1.089e-09   
## 1st Qu.:100343 1st Qu.:6.209e-07   
## Median :181750 Median :2.388e-06   
## Mean :181750 Mean :3.068e-06   
## 3rd Qu.:263157 3rd Qu.:5.234e-06   
## Max. :344564 Max. :8.585e-06



Possiamo notare che la maggior parte del prezzo di vendita sono compresa tra 128988 e 196500, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma.

# Credit Cards - Univariata Quantitativa

## Lettura dati

credit\_card = read.csv2("./credit\_card.csv", sep = ",", stringsAsFactors=T, row.names=1)

## Trasformazione in numerica

str(credit\_card)

## 'data.frame': 10127 obs. of 22 variables:  
## $ Attrition\_Flag : Factor w/ 2 levels "Attrited Customer",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Customer\_Age : int 45 49 51 40 40 44 51 32 37 48 ...  
## $ Gender : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 2 2 2 2 ...  
## $ Dependent\_count : int 3 5 3 4 3 2 4 0 3 2 ...  
## $ Education\_Level : Factor w/ 7 levels "College","Doctorate",..: 4 3 3 4 6 3 7 4 6 3 ...  
## $ Marital\_Status : Factor w/ 4 levels "Divorced","Married",..: 2 3 2 4 2 2 2 4 3 3 ...  
## $ Income\_Category : Factor w/ 6 levels "$120K +","$40K - $60K",..: 3 5 4 5 3 2 1 3 3 4 ...  
## $ Card\_Category : Factor w/ 4 levels "Blue","Gold",..: 1 1 1 1 1 1 2 4 1 1 ...  
## $ Months\_on\_book : int 39 44 36 34 21 36 46 27 36 36 ...  
## $ Total\_Relationship\_Count : int 5 6 4 3 5 3 6 2 5 6 ...  
## $ Months\_Inactive\_12\_mon : int 1 1 1 4 1 1 1 2 2 3 ...  
## $ Contacts\_Count\_12\_mon : int 3 2 0 1 0 2 3 2 0 3 ...  
## $ Credit\_Limit : Factor w/ 6205 levels "10001","10007",..: 519 5724 3719 3613 4464 4081 3748 3136 2207 333 ...  
## $ Total\_Revolving\_Bal : int 777 864 0 2517 0 1247 2264 1396 2517 1677 ...  
## $ Avg\_Open\_To\_Buy : Factor w/ 6813 levels "10","100","1000",..: 525 5881 3964 6124 4699 3018 3501 3023 2055 6808 ...  
## $ Total\_Amt\_Chng\_Q4\_Q1 : Factor w/ 1158 levels "0","0.01","0.018",..: 988 1074 1155 1021 1146 1010 1137 1148 1157 1070 ...  
## $ Total\_Trans\_Amt : int 1144 1291 1887 1171 816 1088 1330 1538 1350 1441 ...  
## $ Total\_Trans\_Ct : int 42 33 20 20 28 24 31 36 24 32 ...  
## $ Total\_Ct\_Chng\_Q4\_Q1 : Factor w/ 830 levels "0","0.028","0.029",..: 792 830 819 819 822 508 400 393 711 539 ...  
## $ Avg\_Utilization\_Ratio : Factor w/ 964 levels "0","0.004","0.005",..: 59 103 1 758 1 309 64 46 111 142 ...  
## $ Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1: Factor w/ 1704 levels "0.00010018","0.00010043",..: 1667 1466 1303 149 1305 1453 113 1636 1403 547 ...  
## $ Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2: Factor w/ 640 levels "0.00041998","0.00042446",..: 632 635 639 628 639 635 629 632 637 611 ...

Come si vede, alcuni dei variabili quantitativi sono fattori

is\_numeric\_factor <- function(factor\_column) {  
 all(!is.na(suppressWarnings(as.numeric(as.character(factor\_column)))))  
}  
  
# Fuzione che converte fact a num  
convert\_factors\_to\_numeric <- function(data) {  
 for (col in names(data)) {  
 if (is.factor(data[[col]])) {  
 if (is\_numeric\_factor(data[[col]])) {  
 data[[col]] = as.numeric(as.character(data[[col]]))  
 }  
 }  
 }  
 return(data)  
}  
  
credit\_w\_numeric = convert\_factors\_to\_numeric(credit\_card)  
str(credit\_w\_numeric)

## 'data.frame': 10127 obs. of 22 variables:  
## $ Attrition\_Flag : Factor w/ 2 levels "Attrited Customer",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Customer\_Age : int 45 49 51 40 40 44 51 32 37 48 ...  
## $ Gender : Factor w/ 2 levels "F","M": 2 1 2 1 2 2 2 2 2 2 ...  
## $ Dependent\_count : int 3 5 3 4 3 2 4 0 3 2 ...  
## $ Education\_Level : Factor w/ 7 levels "College","Doctorate",..: 4 3 3 4 6 3 7 4 6 3 ...  
## $ Marital\_Status : Factor w/ 4 levels "Divorced","Married",..: 2 3 2 4 2 2 2 4 3 3 ...  
## $ Income\_Category : Factor w/ 6 levels "$120K +","$40K - $60K",..: 3 5 4 5 3 2 1 3 3 4 ...  
## $ Card\_Category : Factor w/ 4 levels "Blue","Gold",..: 1 1 1 1 1 1 2 4 1 1 ...  
## $ Months\_on\_book : int 39 44 36 34 21 36 46 27 36 36 ...  
## $ Total\_Relationship\_Count : int 5 6 4 3 5 3 6 2 5 6 ...  
## $ Months\_Inactive\_12\_mon : int 1 1 1 4 1 1 1 2 2 3 ...  
## $ Contacts\_Count\_12\_mon : int 3 2 0 1 0 2 3 2 0 3 ...  
## $ Credit\_Limit : num 12691 8256 3418 3313 4716 ...  
## $ Total\_Revolving\_Bal : int 777 864 0 2517 0 1247 2264 1396 2517 1677 ...  
## $ Avg\_Open\_To\_Buy : num 11914 7392 3418 796 4716 ...  
## $ Total\_Amt\_Chng\_Q4\_Q1 : num 1.33 1.54 2.59 1.4 2.17 ...  
## $ Total\_Trans\_Amt : int 1144 1291 1887 1171 816 1088 1330 1538 1350 1441 ...  
## $ Total\_Trans\_Ct : int 42 33 20 20 28 24 31 36 24 32 ...  
## $ Total\_Ct\_Chng\_Q4\_Q1 : num 1.62 3.71 2.33 2.33 2.5 ...  
## $ Avg\_Utilization\_Ratio : num 0.061 0.105 0 0.76 0 0.311 0.066 0.048 0.113 0.144 ...  
## $ Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_1: num 9.34e-05 5.69e-05 2.11e-05 1.34e-04 2.17e-05 ...  
## $ Naive\_Bayes\_Classifier\_Attrition\_Flag\_Card\_Category\_Contacts\_Count\_12\_mon\_Dependent\_count\_Education\_Level\_Months\_Inactive\_12\_mon\_2: num 1 1 1 1 1 ...

Adesso tutti i variabili quantitaivi sono numerici

## Scelta delle colonne

credit\_card\_numeric\_bool = sapply(credit\_w\_numeric,is.numeric)  
credit\_card\_numeric\_names = names(credit\_card[credit\_card\_numeric\_bool])  
# Rimozione dell'ultimo e penultimo in quanto usato per machine learning  
# e del primo in quanto ID delle persone  
credit\_card\_numeric\_names = credit\_card\_numeric\_names[c(-15,-16)]  
print("Colonne scelte: ")

## [1] "Colonne scelte: "

print(credit\_card\_numeric\_names)

## [1] "Customer\_Age" "Dependent\_count"   
## [3] "Months\_on\_book" "Total\_Relationship\_Count"  
## [5] "Months\_Inactive\_12\_mon" "Contacts\_Count\_12\_mon"   
## [7] "Credit\_Limit" "Total\_Revolving\_Bal"   
## [9] "Avg\_Open\_To\_Buy" "Total\_Amt\_Chng\_Q4\_Q1"   
## [11] "Total\_Trans\_Amt" "Total\_Trans\_Ct"   
## [13] "Total\_Ct\_Chng\_Q4\_Q1" "Avg\_Utilization\_Ratio"

## Analisi Primaria

Analisi\_Primaria <- function(data){  
 print(paste("Numeri di osservazione: ",ncol(data)))  
 print(paste("Numeri di variabili: ",nrow(data)))  
}  
Analisi\_Primaria(credit\_w\_numeric[ ,credit\_card\_numeric\_names])

## [1] "Numeri di osservazione: 14"  
## [1] "Numeri di variabili: 10127"

## Analisi non modificato

sum\_info\_c = lapply(credit\_card\_numeric\_names, function(ele) {  
 colummn = na.omit(credit\_w\_numeric[, ele])  
 univariate\_analysis(colummn, ele)  
})

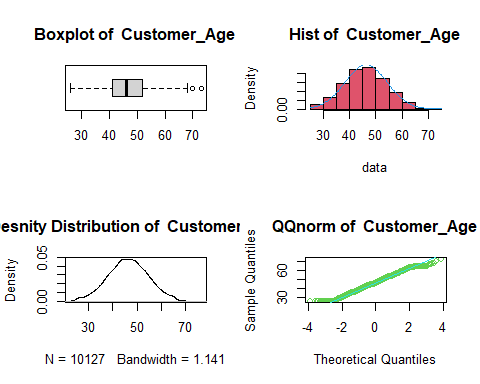
## Analisi

### Customer\_Age

visualization\_continuos(na.omit(credit\_w\_numeric[,"Customer\_Age"]), sum\_info\_c[1],"Customer\_Age")

## [[1]]  
## [[1]]$Customer\_Age  
## [[1]]$Customer\_Age$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 26.00 41.00 46.00 46.33 52.00 73.00   
##   
## [[1]]$Customer\_Age$Variance  
## [1] 64.27  
##   
## [[1]]$Customer\_Age$Standard\_Deviation  
## [1] 8.02  
##   
## [[1]]$Customer\_Age$Moda  
## [[1]]$Customer\_Age$Moda$Mode  
## [1] 44  
##   
## [[1]]$Customer\_Age$Moda$Frequency  
## [1] 500

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (10127 obs.); Bandwidth 'bw' = 1.141  
##   
## x y   
## Min. :22.58 Min. :3.900e-07   
## 1st Qu.:36.04 1st Qu.:3.047e-03   
## Median :49.50 Median :1.262e-02   
## Mean :49.50 Mean :1.855e-02   
## 3rd Qu.:62.96 3rd Qu.:3.461e-02   
## Max. :76.42 Max. :4.824e-02



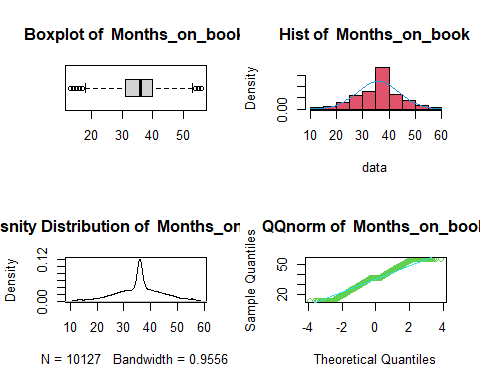
Possiamo dedurre che la distribuzione è per lo più una gaussiana e che la maggior parte delle persone compreso tra 41.00 e 52.00 anni, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma La piu grande occoranze è 500 a 44 anni alineati alla retta qq della norma. Gli outlier invece si trovano ad eta di 73 e 70

### Months\_on\_book

visualization\_continuos(na.omit(credit\_w\_numeric[,"Months\_on\_book"]), sum\_info\_c[3],"Months\_on\_book")

## [[1]]  
## [[1]]$Months\_on\_book  
## [[1]]$Months\_on\_book$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 13.00 31.00 36.00 35.93 40.00 56.00   
##   
## [[1]]$Months\_on\_book$Variance  
## [1] 63.78  
##   
## [[1]]$Months\_on\_book$Standard\_Deviation  
## [1] 7.99  
##   
## [[1]]$Months\_on\_book$Moda  
## [[1]]$Months\_on\_book$Moda$Mode  
## [1] 36  
##   
## [[1]]$Months\_on\_book$Moda$Frequency  
## [1] 2463

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (10127 obs.); Bandwidth 'bw' = 0.9556  
##   
## x y   
## Min. :10.13 Min. :3.333e-05   
## 1st Qu.:22.32 1st Qu.:5.783e-03   
## Median :34.50 Median :1.472e-02   
## Mean :34.50 Mean :2.050e-02   
## 3rd Qu.:46.68 3rd Qu.:2.774e-02   
## Max. :58.87 Max. :1.209e-01

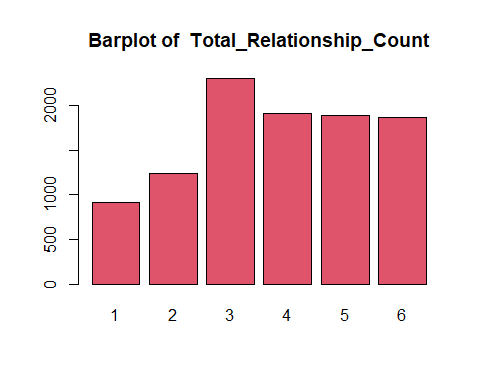


Possiamo dedurre che la distribuzione è per lo più una gaussiana e che la maggior parte delle persone sono registrati con la banca fra 31.00 e 40.00 mesi, La piu grande occoranze è 2463 a 36 mesi

### Total\_Relationship\_Count

visualization\_discrete(na.omit(  
 credit\_w\_numeric[,"Total\_Relationship\_Count"]), sum\_info\_c[4],"Total\_Relationship\_Count")

## [[1]]  
## [[1]]$Total\_Relationship\_Count  
## [[1]]$Total\_Relationship\_Count$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 3.000 4.000 3.813 5.000 6.000   
##   
## [[1]]$Total\_Relationship\_Count$Variance  
## [1] 2.42  
##   
## [[1]]$Total\_Relationship\_Count$Standard\_Deviation  
## [1] 1.55  
##   
## [[1]]$Total\_Relationship\_Count$Moda  
## [[1]]$Total\_Relationship\_Count$Moda$Mode  
## [1] 3  
##   
## [[1]]$Total\_Relationship\_Count$Moda$Frequency  
## [1] 2305  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 1 2 3 4 5 6   
## 910 1243 2305 1912 1891 1866   
##   
## [1] "percintile: "  
##   
## data  
## 1 2 3 4 5 6   
## 8.985879 12.274119 22.760936 18.880221 18.672855 18.425990

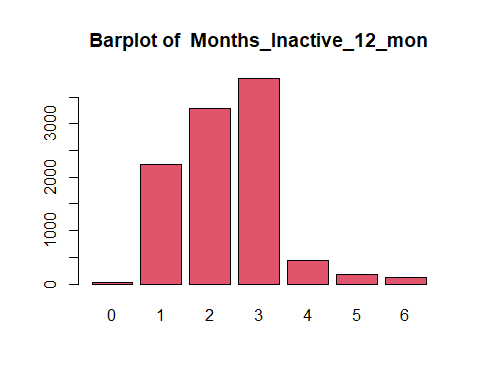


si osservare che la maggior parte delle persone hanno tra 3 e 5 prodotti coprendo il ~=60.31%, con la mggior parte avendo 3 prodotti a 2305 occoranze (~=22.76%) Si osserva che la differenza tra 4,5,6 prodotti è <1% e quindi si puo considerare di uguale possesso

### Months\_Inactive\_12\_mon

visualization\_discrete(na.omit(  
 credit\_w\_numeric[,"Months\_Inactive\_12\_mon",]), sum\_info\_c[5],"Months\_Inactive\_12\_mon")

## [[1]]  
## [[1]]$Months\_Inactive\_12\_mon  
## [[1]]$Months\_Inactive\_12\_mon$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 2.000 2.000 2.341 3.000 6.000   
##   
## [[1]]$Months\_Inactive\_12\_mon$Variance  
## [1] 1.02  
##   
## [[1]]$Months\_Inactive\_12\_mon$Standard\_Deviation  
## [1] 1.01  
##   
## [[1]]$Months\_Inactive\_12\_mon$Moda  
## [[1]]$Months\_Inactive\_12\_mon$Moda$Mode  
## [1] 3  
##   
## [[1]]$Months\_Inactive\_12\_mon$Moda$Frequency  
## [1] 3846  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 0 1 2 3 4 5 6   
## 29 2233 3282 3846 435 178 124   
##   
## [1] "percintile: "  
##   
## data  
## 0 1 2 3 4 5 6   
## 0.2863632 22.0499654 32.4084132 37.9776834 4.2954478 1.7576775 1.2244495

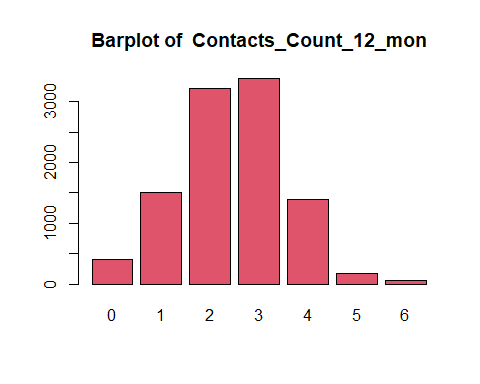


Possiamo osservare che la maggior parte delle persone rimangono innativi tra 1 e 3 mesi, con la maggior parte stando per 3 mesi con 3846 occoranze Si osserva che la maggior parte non vanno piu di 3 mesi prima di riattivare ma quasi sicuramente 1 mese e piu in quanto <1% rimangono attivi costantemente (0 mesi di innativita)

### Contacts\_Count\_12\_mon

visualization\_discrete(na.omit(  
 credit\_w\_numeric[,"Contacts\_Count\_12\_mon"]), sum\_info\_c[6],"Contacts\_Count\_12\_mon")

## [[1]]  
## [[1]]$Contacts\_Count\_12\_mon  
## [[1]]$Contacts\_Count\_12\_mon$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 2.000 2.000 2.455 3.000 6.000   
##   
## [[1]]$Contacts\_Count\_12\_mon$Variance  
## [1] 1.22  
##   
## [[1]]$Contacts\_Count\_12\_mon$Standard\_Deviation  
## [1] 1.11  
##   
## [[1]]$Contacts\_Count\_12\_mon$Moda  
## [[1]]$Contacts\_Count\_12\_mon$Moda$Mode  
## [1] 3  
##   
## [[1]]$Contacts\_Count\_12\_mon$Moda$Frequency  
## [1] 3380  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 0 1 2 3 4 5 6   
## 399 1499 3227 3380 1392 176 54   
##   
## [1] "percintile: "  
##   
## data  
## 0 1 2 3 4 5 6   
## 3.939962 14.802014 31.865311 33.376123 13.745433 1.737928 0.533228



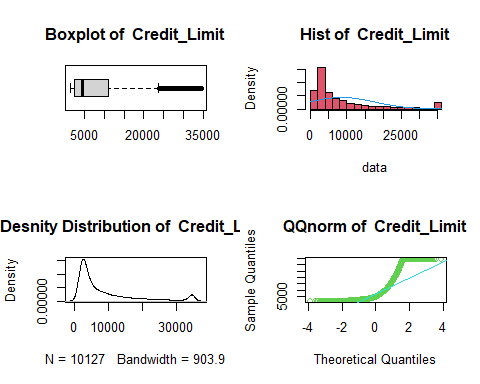
Possiamo osservare che la maggior parte delle persone hanno tra 2 e 3 contanti/12mesi, con la maggior parte avendo 3 contanti con 3380 occoranze

### Credit\_Limit

visualization\_continuos(na.omit(credit\_w\_numeric[,"Credit\_Limit"]), sum\_info\_c[7],"Credit\_Limit")

## [[1]]  
## [[1]]$Credit\_Limit  
## [[1]]$Credit\_Limit$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1438 2555 4549 8632 11068 34516   
##   
## [[1]]$Credit\_Limit$Variance  
## [1] 82605861  
##   
## [[1]]$Credit\_Limit$Standard\_Deviation  
## [1] 9088.78  
##   
## [[1]]$Credit\_Limit$Moda  
## [[1]]$Credit\_Limit$Moda$Mode  
## [1] 34516  
##   
## [[1]]$Credit\_Limit$Moda$Frequency  
## [1] 508

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (10127 obs.); Bandwidth 'bw' = 903.9  
##   
## x y   
## Min. :-1273 Min. :2.553e-07   
## 1st Qu.: 8352 1st Qu.:5.913e-06   
## Median :17977 Median :1.145e-05   
## Mean :17977 Mean :2.594e-05   
## 3rd Qu.:27602 3rd Qu.:2.660e-05   
## Max. :37228 Max. :1.574e-04



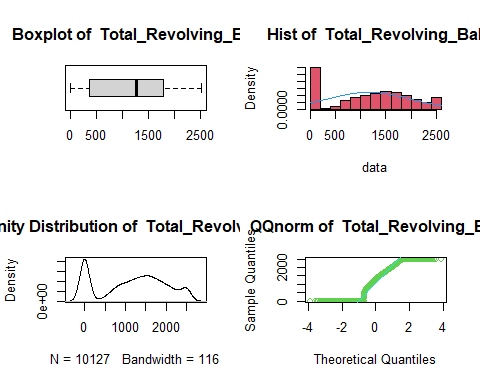
Possiamo osservare che la maggior parte delle persone hanno un limite di credito tra 2555 e 11068, ed esiste un anomalia al limite superiore dove non rispetta la descrescita delle occoranze del limite di credito con 34516 avendo 508 occoranze ed essendo il max pur non facendo parte del IQR

### Total\_Revolving\_Bal

visualization\_continuos(na.omit(  
 credit\_w\_numeric[,"Total\_Revolving\_Bal"]), sum\_info\_c[8],"Total\_Revolving\_Bal")

## [[1]]  
## [[1]]$Total\_Revolving\_Bal  
## [[1]]$Total\_Revolving\_Bal$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 359 1276 1163 1784 2517   
##   
## [[1]]$Total\_Revolving\_Bal$Variance  
## [1] 664204.4  
##   
## [[1]]$Total\_Revolving\_Bal$Standard\_Deviation  
## [1] 814.99  
##   
## [[1]]$Total\_Revolving\_Bal$Moda  
## [[1]]$Total\_Revolving\_Bal$Moda$Mode  
## [1] 0  
##   
## [[1]]$Total\_Revolving\_Bal$Moda$Frequency  
## [1] 2470

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (10127 obs.); Bandwidth 'bw' = 116  
##   
## x y   
## Min. :-347.9 Min. :2.141e-06   
## 1st Qu.: 455.3 1st Qu.:1.690e-04   
## Median :1258.5 Median :3.021e-04   
## Mean :1258.5 Mean :3.108e-04   
## 3rd Qu.:2061.7 3rd Qu.:4.447e-04   
## Max. :2864.9 Max. :8.415e-04



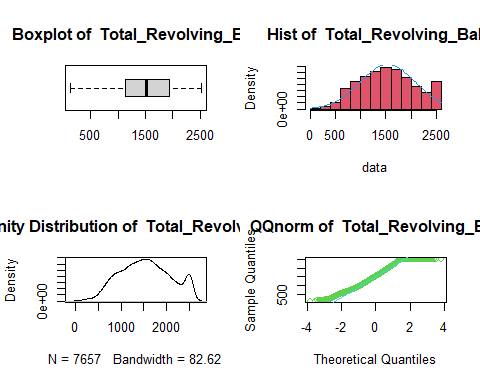
Possiamo notare che la maggior parte delle persone hanno Prestito a saldo rotativo compresa tra 359 e 1784, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Possiamo notare che la maggior parte delle persone hanno un Prestito a saldo rotativo a 0, 2470 in totale Per analizare quelli che hanno serve ignorare quelli senza

#### Rimozione degli zero

rm\_0\_data = rm\_0(na.omit(  
 credit\_w\_numeric[,"Total\_Revolving\_Bal"]))  
visualization\_continuos(rm\_0\_data, univariate\_analysis(rm\_0\_data,"Total\_Revolving\_Bal"),"Total\_Revolving\_Bal")

## $Total\_Revolving\_Bal  
## $Total\_Revolving\_Bal$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 132 1132 1524 1538 1935 2517   
##   
## $Total\_Revolving\_Bal$Variance  
## [1] 301542.4  
##   
## $Total\_Revolving\_Bal$Standard\_Deviation  
## [1] 549.13  
##   
## $Total\_Revolving\_Bal$Moda  
## $Total\_Revolving\_Bal$Moda$Mode  
## [1] 2517  
##   
## $Total\_Revolving\_Bal$Moda$Frequency  
## [1] 508

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (7657 obs.); Bandwidth 'bw' = 82.62  
##   
## x y   
## Min. :-115.9 Min. :4.030e-08   
## 1st Qu.: 604.3 1st Qu.:1.053e-04   
## Median :1324.5 Median :3.810e-04   
## Mean :1324.5 Mean :3.468e-04   
## 3rd Qu.:2044.7 3rd Qu.:5.380e-04   
## Max. :2764.9 Max. :6.810e-04



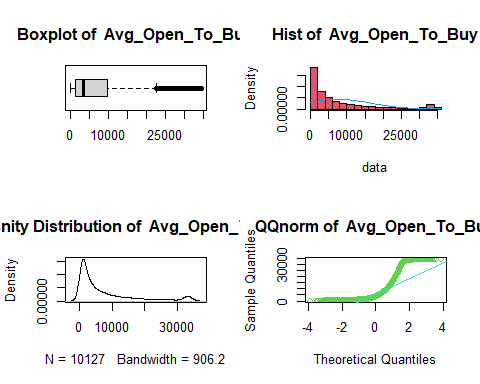
Possiamo notare che la maggior parte delle persone hanno Prestito a saldo rotativo compresa tra 1132 e 1935, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Possiamo notare che la maggior parte delle persone hanno un Prestito a saldo rotativo a 2517 per un totale di 508 persone esendo fuori dal IQR

### Avg\_Open\_To\_Buy

visualization\_continuos(na.omit(  
 credit\_w\_numeric[,"Avg\_Open\_To\_Buy"]), sum\_info\_c[9],"Avg\_Open\_To\_Buy")

## [[1]]  
## [[1]]$Avg\_Open\_To\_Buy  
## [[1]]$Avg\_Open\_To\_Buy$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3 1324 3474 7469 9859 34516   
##   
## [[1]]$Avg\_Open\_To\_Buy$Variance  
## [1] 82640560  
##   
## [[1]]$Avg\_Open\_To\_Buy$Standard\_Deviation  
## [1] 9090.69  
##   
## [[1]]$Avg\_Open\_To\_Buy$Moda  
## [[1]]$Avg\_Open\_To\_Buy$Moda$Mode  
## [1] 1438.3  
##   
## [[1]]$Avg\_Open\_To\_Buy$Moda$Frequency  
## [1] 324

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (10127 obs.); Bandwidth 'bw' = 906.2  
##   
## x y   
## Min. :-2716 Min. :5.349e-08   
## 1st Qu.: 7272 1st Qu.:5.472e-06   
## Median :17260 Median :1.138e-05   
## Mean :17260 Mean :2.501e-05   
## 3rd Qu.:27247 3rd Qu.:2.493e-05   
## Max. :37235 Max. :1.607e-04



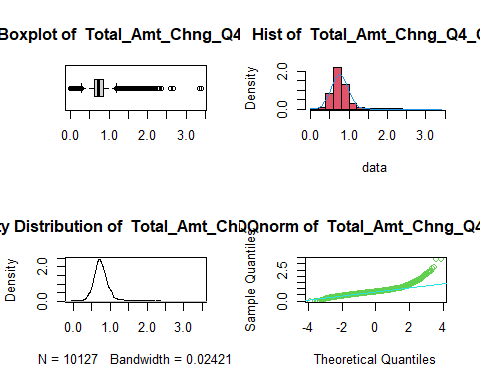
Possiamo notare che la maggior parte delle persone disponibili per acquistare una linea di credito sono compresa tra 1324 e 9859, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Possiamo notare che la maggior parte delle persone hanno un Prestito a saldo rotativo a 2517 per un totale di 508 persone esendo fuori dal IQR

### Total\_Amt\_Chng\_Q4\_Q1

visualization\_continuos(na.omit(  
 credit\_w\_numeric[,"Total\_Amt\_Chng\_Q4\_Q1"]), sum\_info\_c[10],"Total\_Amt\_Chng\_Q4\_Q1")

## [[1]]  
## [[1]]$Total\_Amt\_Chng\_Q4\_Q1  
## [[1]]$Total\_Amt\_Chng\_Q4\_Q1$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.6310 0.7360 0.7599 0.8590 3.3970   
##   
## [[1]]$Total\_Amt\_Chng\_Q4\_Q1$Variance  
## [1] 0.05  
##   
## [[1]]$Total\_Amt\_Chng\_Q4\_Q1$Standard\_Deviation  
## [1] 0.22  
##   
## [[1]]$Total\_Amt\_Chng\_Q4\_Q1$Moda  
## [[1]]$Total\_Amt\_Chng\_Q4\_Q1$Moda$Mode  
## [1] 0.791  
##   
## [[1]]$Total\_Amt\_Chng\_Q4\_Q1$Moda$Frequency  
## [1] 36

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (10127 obs.); Bandwidth 'bw' = 0.02421  
##   
## x y   
## Min. :-0.07263 Min. :0.0000000   
## 1st Qu.: 0.81294 1st Qu.:0.0009398   
## Median : 1.69850 Median :0.0068703   
## Mean : 1.69850 Mean :0.2820299   
## 3rd Qu.: 2.58406 3rd Qu.:0.1170530   
## Max. : 3.46963 Max. :2.4673806



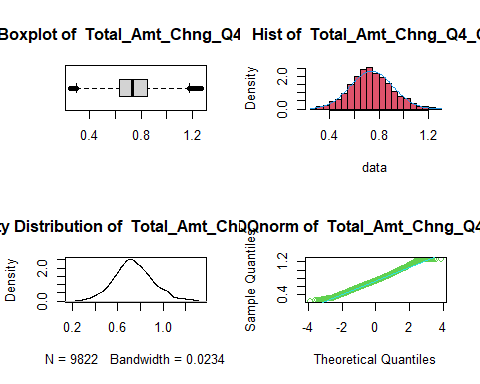
Possiamo notare che la maggior parte delle Variazione dell’importo della transazione (Q4 rispetto a Q1) è compresa tra 0.6310 e 0.8590, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Possiamo notare che la maggior parte delle Variazione dell’importo della transazione (Q4 rispetto a Q1) sono a 0.791 con 36 Variazione

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(na.omit(credit\_w\_numeric[,"Total\_Amt\_Chng\_Q4\_Q1"]),.98)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"Total\_Amt\_Chng\_Q4\_Q1"),"Total\_Amt\_Chng\_Q4\_Q1")

## $Total\_Amt\_Chng\_Q4\_Q1  
## $Total\_Amt\_Chng\_Q4\_Q1$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.2510 0.6290 0.7320 0.7411 0.8480 1.2690   
##   
## $Total\_Amt\_Chng\_Q4\_Q1$Variance  
## [1] 0.03  
##   
## $Total\_Amt\_Chng\_Q4\_Q1$Standard\_Deviation  
## [1] 0.17  
##   
## $Total\_Amt\_Chng\_Q4\_Q1$Moda  
## $Total\_Amt\_Chng\_Q4\_Q1$Moda$Mode  
## [1] 0.791  
##   
## $Total\_Amt\_Chng\_Q4\_Q1$Moda$Frequency  
## [1] 36

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (9822 obs.); Bandwidth 'bw' = 0.0234  
##   
## x y   
## Min. :0.1808 Min. :0.0000433   
## 1st Qu.:0.4704 1st Qu.:0.1271346   
## Median :0.7600 Median :0.5689233   
## Mean :0.7600 Mean :0.8624335   
## 3rd Qu.:1.0496 3rd Qu.:1.5598759   
## Max. :1.3392 Max. :2.5517658



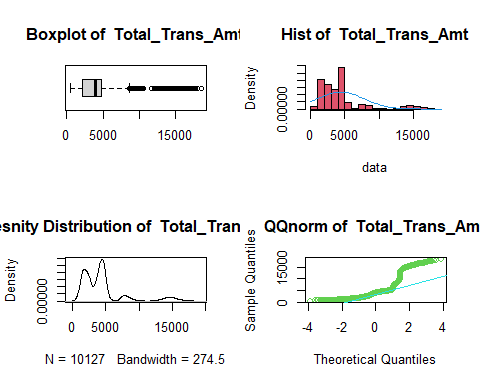
Possiamo dedurre che la distribuzione è semi-normato e che la maggior parte delle Variazione dell’importo della transazione (Q4 rispetto a Q1) sono compresa tra 0.6290 e 0.8480, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Possiamo notare che la maggior parte delle Variazione delle Variazione dell’importo della transazione (Q4 rispetto a Q1) sono a 0.791 con 36 Variazione

### Total\_Trans\_Amt

visualization\_continuos(na.omit(  
 credit\_w\_numeric[,"Total\_Trans\_Amt"]), sum\_info\_c[11],"Total\_Trans\_Amt")

## [[1]]  
## [[1]]$Total\_Trans\_Amt  
## [[1]]$Total\_Trans\_Amt$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 510 2156 3899 4404 4741 18484   
##   
## [[1]]$Total\_Trans\_Amt$Variance  
## [1] 11540487  
##   
## [[1]]$Total\_Trans\_Amt$Standard\_Deviation  
## [1] 3397.13  
##   
## [[1]]$Total\_Trans\_Amt$Moda  
## [[1]]$Total\_Trans\_Amt$Moda$Mode  
## [1] 4253 4509  
##   
## [[1]]$Total\_Trans\_Amt$Moda$Frequency  
## [1] 11

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (10127 obs.); Bandwidth 'bw' = 274.5  
##   
## x y   
## Min. : -313.6 Min. :1.640e-09   
## 1st Qu.: 4591.7 1st Qu.:2.146e-06   
## Median : 9497.0 Median :1.323e-05   
## Mean : 9497.0 Mean :5.092e-05   
## 3rd Qu.:14402.3 3rd Qu.:3.987e-05   
## Max. :19307.6 Max. :2.976e-04



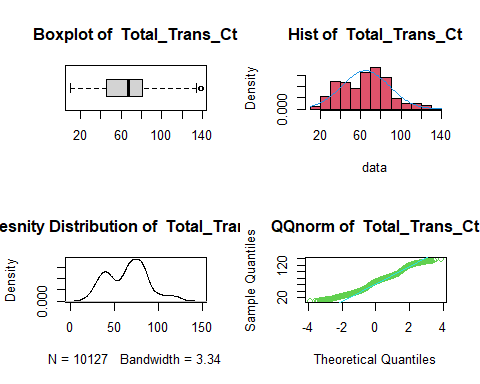
Possiamo notare che la maggior parte delle Importe totali delle transazioni sono compresa tra 2156 e 4741, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Possiamo notare che la maggior parte delle Importe totali delle transazioni sono a 4253 e 4509 con 11 transazioni

### Total\_Trans\_Ct

visualization\_continuos(na.omit(  
 credit\_w\_numeric[,"Total\_Trans\_Ct"]), sum\_info\_c[12],"Total\_Trans\_Ct")

## [[1]]  
## [[1]]$Total\_Trans\_Ct  
## [[1]]$Total\_Trans\_Ct$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 10.00 45.00 67.00 64.86 81.00 139.00   
##   
## [[1]]$Total\_Trans\_Ct$Variance  
## [1] 550.96  
##   
## [[1]]$Total\_Trans\_Ct$Standard\_Deviation  
## [1] 23.47  
##   
## [[1]]$Total\_Trans\_Ct$Moda  
## [[1]]$Total\_Trans\_Ct$Moda$Mode  
## [1] 81  
##   
## [[1]]$Total\_Trans\_Ct$Moda$Frequency  
## [1] 208

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (10127 obs.); Bandwidth 'bw' = 3.34  
##   
## x y   
## Min. : -0.01909 Min. :1.860e-07   
## 1st Qu.: 37.24046 1st Qu.:1.141e-03   
## Median : 74.50000 Median :4.058e-03   
## Mean : 74.50000 Mean :6.703e-03   
## 3rd Qu.:111.75954 3rd Qu.:1.182e-02   
## Max. :149.01909 Max. :1.885e-02



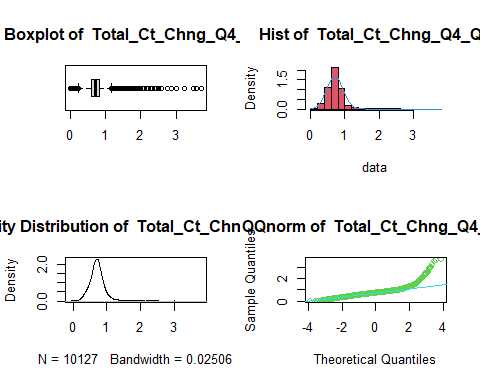
Possiamo dedurre che la distribuzione è semi-normato positivamente assimetrica e che la maggior parte delle Conteggio totale delle transazioni sono compresa tra 45.00 e 81.00, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Possiamo notare che la maggior parte delle Variazione delle Conteggio totale delle transazioni sono a 81 con 208 occoranze

### Total\_Ct\_Chng\_Q4\_Q1

visualization\_continuos(na.omit(  
 credit\_w\_numeric[,"Total\_Ct\_Chng\_Q4\_Q1"]), sum\_info\_c[13],"Total\_Ct\_Chng\_Q4\_Q1")

## [[1]]  
## [[1]]$Total\_Ct\_Chng\_Q4\_Q1  
## [[1]]$Total\_Ct\_Chng\_Q4\_Q1$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.5820 0.7020 0.7122 0.8180 3.7140   
##   
## [[1]]$Total\_Ct\_Chng\_Q4\_Q1$Variance  
## [1] 0.06  
##   
## [[1]]$Total\_Ct\_Chng\_Q4\_Q1$Standard\_Deviation  
## [1] 0.24  
##   
## [[1]]$Total\_Ct\_Chng\_Q4\_Q1$Moda  
## [[1]]$Total\_Ct\_Chng\_Q4\_Q1$Moda$Mode  
## [1] 0.667  
##   
## [[1]]$Total\_Ct\_Chng\_Q4\_Q1$Moda$Frequency  
## [1] 171

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (10127 obs.); Bandwidth 'bw' = 0.02506  
##   
## x y   
## Min. :-0.07518 Min. :0.000000   
## 1st Qu.: 0.89091 1st Qu.:0.001310   
## Median : 1.85700 Median :0.007164   
## Mean : 1.85700 Mean :0.258526   
## 3rd Qu.: 2.82309 3rd Qu.:0.120710   
## Max. : 3.78918 Max. :2.289734



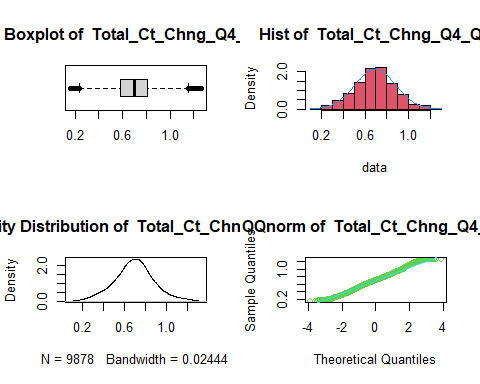
Possiamo notare che la maggior parte delle Variazione delle Variazione nel conteggio delle transazioni (Q4 rispetto a Q1) è compresa tra 0.5820 e 0.8180, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Possiamo notare che la maggior parte delle Variazione dell’importo della transazione (Q4 rispetto a Q1) sono a 0.667 con 171 Variazione

#### Rimozione di outlier

rm\_outlier\_data = rm\_oultiers(na.omit(credit\_w\_numeric[,"Total\_Ct\_Chng\_Q4\_Q1"]),.98)  
visualization\_continuos(rm\_outlier\_data, univariate\_analysis(rm\_outlier\_data,"Total\_Ct\_Chng\_Q4\_Q1"),"Total\_Ct\_Chng\_Q4\_Q1")

## $Total\_Ct\_Chng\_Q4\_Q1  
## $Total\_Ct\_Chng\_Q4\_Q1$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.1610 0.5810 0.6980 0.6959 0.8100 1.2630   
##   
## $Total\_Ct\_Chng\_Q4\_Q1$Variance  
## [1] 0.03  
##   
## $Total\_Ct\_Chng\_Q4\_Q1$Standard\_Deviation  
## [1] 0.19  
##   
## $Total\_Ct\_Chng\_Q4\_Q1$Moda  
## $Total\_Ct\_Chng\_Q4\_Q1$Moda$Mode  
## [1] 0.667  
##   
## $Total\_Ct\_Chng\_Q4\_Q1$Moda$Frequency  
## [1] 171

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (9878 obs.); Bandwidth 'bw' = 0.02444  
##   
## x y   
## Min. :0.08769 Min. :0.0001072   
## 1st Qu.:0.39985 1st Qu.:0.1444697   
## Median :0.71200 Median :0.5058148   
## Mean :0.71200 Mean :0.8001016   
## 3rd Qu.:1.02415 3rd Qu.:1.3695197   
## Max. :1.33631 Max. :2.3510407



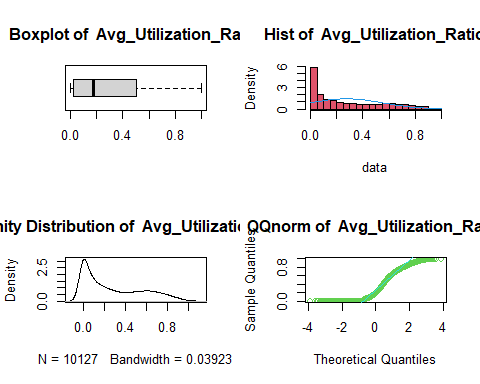
Possiamo dedurre che la distribuzione è semi-normato e che la maggior parte delle Variazione delle Variazione nel conteggio delle transazioni (Q4 rispetto a Q1) è compresa tra 0.5810 e 0.8100, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Possiamo notare che la maggior parte delle Variazione elle Variazione delle Variazione nel conteggio delle transazioni (Q4 rispetto a Q1) sono a 0.667 con 171 Variazione

### Avg\_Utilization\_Ratio

visualization\_continuos(na.omit(  
 credit\_w\_numeric[,"Avg\_Utilization\_Ratio"]), sum\_info\_c[14],"Avg\_Utilization\_Ratio")

## [[1]]  
## [[1]]$Avg\_Utilization\_Ratio  
## [[1]]$Avg\_Utilization\_Ratio$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0230 0.1760 0.2749 0.5030 0.9990   
##   
## [[1]]$Avg\_Utilization\_Ratio$Variance  
## [1] 0.08  
##   
## [[1]]$Avg\_Utilization\_Ratio$Standard\_Deviation  
## [1] 0.28  
##   
## [[1]]$Avg\_Utilization\_Ratio$Moda  
## [[1]]$Avg\_Utilization\_Ratio$Moda$Mode  
## [1] 0  
##   
## [[1]]$Avg\_Utilization\_Ratio$Moda$Frequency  
## [1] 2470

##   
## Call:  
## density.default(x = data)  
##   
## Data: data (10127 obs.); Bandwidth 'bw' = 0.03923  
##   
## x y   
## Min. :-0.1177 Min. :0.0000855   
## 1st Qu.: 0.1909 1st Qu.:0.3492169   
## Median : 0.4995 Median :0.6801743   
## Mean : 0.4995 Mean :0.8090999   
## 3rd Qu.: 0.8081 3rd Qu.:0.9021584   
## Max. : 1.1167 Max. :3.1604425



Possiamo dedurre che la distribuzione è semi-normato e che la maggior parte delle Rapporto medio di utilizzo della carta è compresa tra 0.0230 e 0.5030, essendo rispettivamente il 1° e il 3° quantile e con maggior densita alineati alla retta qq della norma. Possiamo notare che la maggior parte delle Variazione elle Variazione delle Variazione nel conteggio delle transazioni (Q4 rispetto a Q1) sono a 0 con 2470 Variazione

#House price Univariate Qualitative

## Pulizia dei dati

HP$MSZoning <- factor(  
 HP$MSZoning,  
 levels = c("RL", "RM", "RH", "FV", "C (all)"),  
 labels = c("Residential Low Density", "Residential Medium Density", "Residential High Density", "Floating Village Residential", "Commercial"))  
  
HP$MSSubClass <- factor(  
 HP$MSSubClass,  
 levels= c("20","30","40","45","50","60","70","75","80","85","90","120","150","160","180","190"),  
 labels = c("1-STORY 1946 & NEWER ALL STYLES","1-STORY 1945 & OLDER","1-STORY FINISHED ATTIC ALL AGES" ,"1-1/2 STORY UNFINISHED ALL AGES", "1-1/2 STORY FINISHED ALL AGES", "2-STORY 1946 & NEWER", "2-STORY 1945 & OLDER", "2-1/2 STORY ALL AGES", "SPLIT OR MULTI-LEVEL", "SPLIT FOYER", "DUPLEX - ALL STYLES AND AGES", "1-STORY PUD 1946 & NEWER", "1-1/2 STORY PUD ALL AGES", "2-STORY PUD 1946 & NEWER", "PUD MULTILEVEL - INCL SPLIT LEV/FOYER", "FAMILY CONVERSION ALL STYLES AND AGES"),)  
  
  
HP$LotShape <- factor(  
 HP$LotShape,  
 levels = c("Reg", "IR1", "IR2", "IR3"),  
 labels = c("Regular", "Slightly irregular", "Moderately irregular", "Irregular"))  
  
HP$LandContour <- factor(  
 HP$LandContour,  
 levels = c("Lvl", "Bnk", "HLS", "Low"),  
 labels = c("Near Flat/Level", "Banked", "Hillside", "Depression"))  
  
HP$Neighborhood <- factor(  
 HP$Neighborhood,  
 levels = c("Blmngtn", "Blueste", "BrDale", "BrkSide", "ClearCr", "CollgCr", "Crawfor", "Edwards", "Gilbert", "IDOTRR", "MeadowV", "Mitchel", "NAmes", "NoRidge", "NPkVill", "NridgHt", "NWAmes", "OldTown", "SWISU", "Sawyer", "SawyerW", "Somerst", "StoneBr", "Timber", "Veenker"),  
 labels = c("Bloomington Heights", "Bluestem", "Briardale", "Brookside", "Clear Creek", "College Creek", "Crawford", "Edwards", "Gilbert", "Iowa DOT and Rail Road", "Meadow Village", "Mitchell", "North Ames", "Northridge", "Northpark Villa", "Northridge Heights", "Northwest Ames", "Old Town", "South & West of Iowa State University", "Sawyer", "Sawyer West", "Somerset", "Stone Brook", "Timberland", "Veenker"))  
  
HP$BldgType <- factor(  
 HP$BldgType,  
 levels = c("1Fam", "2fmCon", "Duplex", "TwnhsE", "Twnhs"),  
 labels = c("Single-family Detached", "Two-family Conversion", "Duplex", "Townhouse End Unit", "Townhouse Inside Unit"))  
  
HP$MasVnrArea[HP$MasVnrArea == 0 | is.na(HP$MasVnrType)] <- NA  
  
for (f in c("ExterQual", "ExterCond", "BsmtQual", "BsmtCond", "HeatingQC", "KitchenQual", "FireplaceQu", "GarageQual", "GarageCond")) {  
 HP[, f] <- factor(  
 HP[, f],  
 levels = c("Po", "Fa", "TA", "Gd", "Ex"),  
 labels = c("Poor", "Fair", "Average", "Good", "Excellent"))  
}  
  
levels(HP$CentralAir) <- c("No", "Yes")  
  
HP$BsmtExposure <- factor(  
 HP$BsmtExposure,  
 levels = c("No", "Mn", "Av", "Gd"),  
 labels = c("No", "Minimum", "Average", "Good"))  
  
HP$GarageFinish <- factor(  
 HP$GarageFinish,  
 levels = c("Unf", "RFn", "Fin"),  
 labels = c("Unfinished", "Rough finished", "Finished"))  
  
HP$PavedDrive <- factor(  
 HP$PavedDrive,  
 levels = c("N", "P", "Y"),  
 labels = c("Dirt/Graver", "Partial paved", "Paved"))  
  
HP$Fence <- factor(  
 HP$Fence,  
 levels = c("MnWw", "GdWo", "MnPrv", "GdPrv"),  
 labels = c("Minimum Wood/Wire", "Good Wood", "Minimum Privacy", "Good Privacy"))  
  
HP$SaleType <- factor(  
 HP$SaleType,  
 levels = c("WD", "CWD", "VWD", "New", "COD", "Con", "ConLw", "ConLI", "ConLD", "Oth"),  
 labels = c("Warranty Deed - Conventional", "Warranty Deed - Cash", "Warranty Deed - VA Loan", "Home just constructed and sold", "Court Officer", "Contract 15% Down payment", "Contract Low Down Payment", "Contract Low Interest", "Contract Low Down", "Other"))  
  
HP$SaleCondition <- factor(  
 HP$SaleCondition,  
 levels = c("Normal", "Abnorml", "AdjLand", "Alloca", "Family", "Partial"),  
 labels = c("Normal Sale", "Abnormal Sale", "Adjoining Land Purchase", "Allocation", "Sale between family members", "Home not completed"))  
  
  
HP$OverallQual <- factor(  
 HP$OverallQual,  
 levels=c("1","2","3","4","5","6","7","8","9","10"))  
  
HP$OverallCond <- factor(  
 HP$OverallCond,  
 levels=c("1","2","3","4","5","6","7","8","9","10"))   
  
for (f in c("BsmtFullBath", "BsmtHalfBath", "FullBath", "HalfBath", "BedroomAbvGr", "KitchenAbvGr", "TotRmsAbvGrd", "Fireplaces","GarageCars","MoSold","YrSold")) {  
 HP[, f] <- factor(HP[, f])  
}  
  
HP$MiscFeature <- factor(  
 HP$MiscFeature,  
 levels =c("Elev","Gar2","Othr","Shed","TenC",NA) ,  
 labels=c("Elevator","2nd Garage","Other","Shed","Tennis Court","No Feature"),   
 exclude = NULL)  
attach(HP)  
attach(credit\_card)

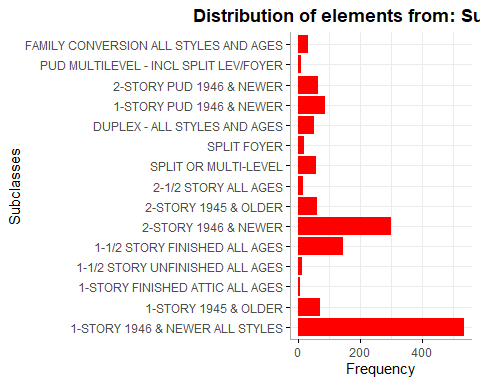
##Funzione di analisi

Univar\_Quality <- function(df,clmn, label=NULL, title=NULL) {  
   
 Entries<-clmn  
 tab<-sort(table(Entries), decreasing = T)  
 freqtab<-round(prop.table(tab)\*100,2)  
  
 print(tab)  
 print(freqtab)  
   
 if(length(levels(clmn))<4){  
 bar\_plot <- ggplot(df,aes(x=clmn)) +  
 geom\_bar(fill = "red") +  
 labs(title = paste("Distribution of elements from:", title),  
 x = label,  
 y = "Frequency")  
 print(bar\_plot)  
 } else{  
 bar\_plot <- ggplot(df, aes(x=clmn)) +  
 geom\_bar(fill = "red") +  
 coord\_flip()+  
 labs(title = paste("Distribution of elements from:", title),  
 x = label,  
 y = "Frequency")  
 print(bar\_plot)  
 }  
}

### MSSubClass

Univar\_Quality(  
 HP, MSSubClass,  
 label="Subclasses",  
 title="Sub Classes")

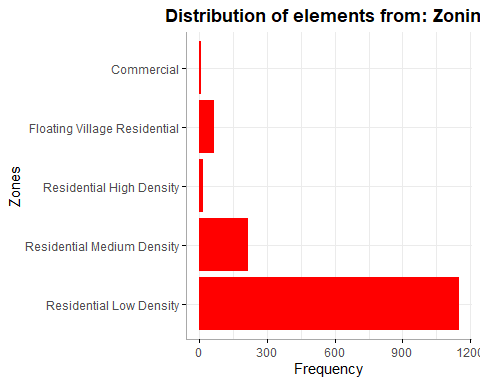
## Entries  
## 1-STORY 1946 & NEWER ALL STYLES 2-STORY 1946 & NEWER   
## 536 299   
## 1-1/2 STORY FINISHED ALL AGES 1-STORY PUD 1946 & NEWER   
## 144 87   
## 1-STORY 1945 & OLDER 2-STORY PUD 1946 & NEWER   
## 69 63   
## 2-STORY 1945 & OLDER SPLIT OR MULTI-LEVEL   
## 60 58   
## DUPLEX - ALL STYLES AND AGES FAMILY CONVERSION ALL STYLES AND AGES   
## 52 30   
## SPLIT FOYER 2-1/2 STORY ALL AGES   
## 20 16   
## 1-1/2 STORY UNFINISHED ALL AGES PUD MULTILEVEL - INCL SPLIT LEV/FOYER   
## 12 10   
## 1-STORY FINISHED ATTIC ALL AGES 1-1/2 STORY PUD ALL AGES   
## 4 0   
## Entries  
## 1-STORY 1946 & NEWER ALL STYLES 2-STORY 1946 & NEWER   
## 36.71 20.48   
## 1-1/2 STORY FINISHED ALL AGES 1-STORY PUD 1946 & NEWER   
## 9.86 5.96   
## 1-STORY 1945 & OLDER 2-STORY PUD 1946 & NEWER   
## 4.73 4.32   
## 2-STORY 1945 & OLDER SPLIT OR MULTI-LEVEL   
## 4.11 3.97   
## DUPLEX - ALL STYLES AND AGES FAMILY CONVERSION ALL STYLES AND AGES   
## 3.56 2.05   
## SPLIT FOYER 2-1/2 STORY ALL AGES   
## 1.37 1.10   
## 1-1/2 STORY UNFINISHED ALL AGES PUD MULTILEVEL - INCL SPLIT LEV/FOYER   
## 0.82 0.68   
## 1-STORY FINISHED ATTIC ALL AGES 1-1/2 STORY PUD ALL AGES   
## 0.27 0.00

 Possiamo notare che delle varie tipologie di abitazione che stiamo analizzando, le più comuni, con il 36.7% e il 20.5%, sono quelle a un piano e quelle a due piani successive al 1946 di tutti i stili. Inoltre le case con elementi ancora in costruzione come le “PUD” e le “Unfinished” hanno meno successo di quelle finite o recenti. Quattro elementi hanno valori trascurabili in quanto inferiori all ~1%, con “1-1/2 STORY PUD - ALL AGES” che non compare in nessun caso.

### Zoning

Univar\_Quality(  
 HP,MSZoning,  
 label="Zones",  
 title="Zoning")

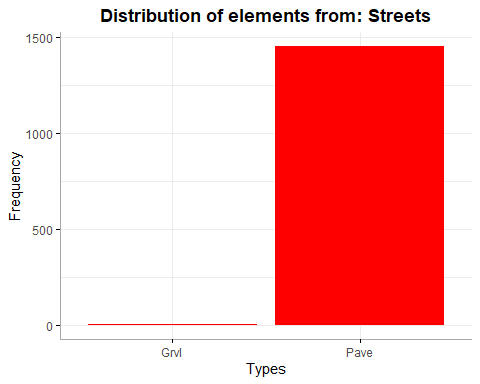
## Entries  
## Residential Low Density Residential Medium Density   
## 1151 218   
## Floating Village Residential Residential High Density   
## 65 16   
## Commercial   
## 10   
## Entries  
## Residential Low Density Residential Medium Density   
## 78.84 14.93   
## Floating Village Residential Residential High Density   
## 4.45 1.10   
## Commercial   
## 0.68

 La maggior parte della popolazione (~79%) risiede in zone residenziali a bassa densità. Le zone commerciali rappresentano meno di ~1%, quindi appaiono trascurabili. Noto un diminuire della popolazione nelle aree “Residential” in base all’aumento della densità.

### Street

Univar\_Quality(  
 HP,Street,  
 label="Types",  
 title="Streets")

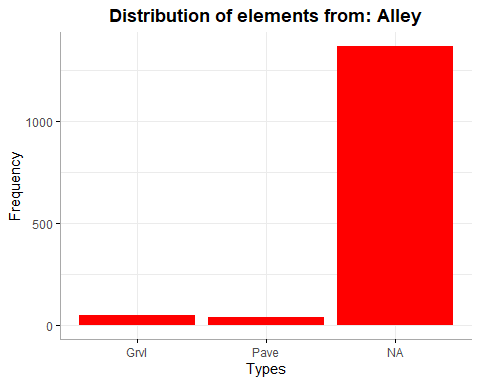
## Entries  
## Pave Grvl   
## 1454 6   
## Entries  
## Pave Grvl   
## 99.59 0.41

 Notiamo uno squilibrio dei dati (1454 entrate) verso “Pave”, cioè marciapiede. Le strade di ghiaia rappresentano meno dell’~1% dei dati (6 entrate), quindi deduco siano trascurabili.

### Alley

Univar\_Quality(  
 HP,Alley,  
 label="Types",  
 title="Alley")

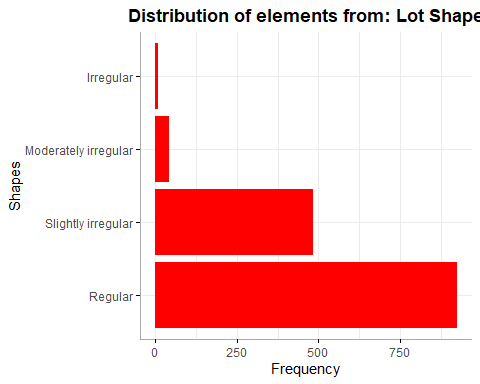
## Entries  
## Grvl Pave   
## 50 41   
## Entries  
## Grvl Pave   
## 54.95 45.05

 Noto che la maggior parte delle case non ha strade interne di accesso, ma tra quelle che lo hanno (91 elementi) la distribuzione tra ghiaia e marciapiedi è bilanciata.

### LotShape

Univar\_Quality(  
 HP,LotShape,  
 label="Shapes",  
 title="Lot Shape")

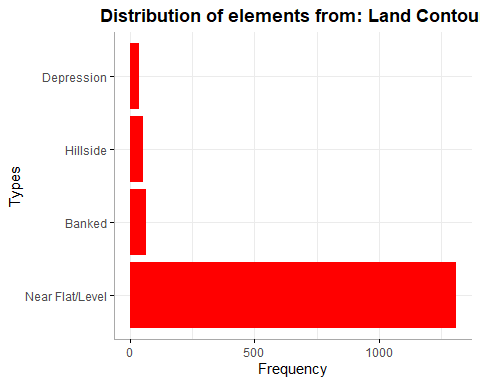
## Entries  
## Regular Slightly irregular Moderately irregular   
## 925 484 41   
## Irregular   
## 10   
## Entries  
## Regular Slightly irregular Moderately irregular   
## 63.36 33.15 2.81   
## Irregular   
## 0.68

 Notiamo che 925 (il ~63%) delle abitazioni hanno forma regolare e il 484 (~33%) ha piccole irregolarità. Le case più irregolari in cambio rappresentano meno dell’ ~1% delle frequenze, mostrando come tendenzialmente esse siano meno comprate/vendute per questa qualità.

### LandContour

Univar\_Quality(  
 HP,LandContour,  
 label="Types",  
 title="Land Contour")

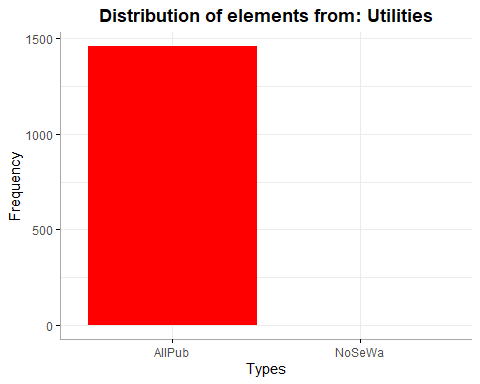
## Entries  
## Near Flat/Level Banked Hillside Depression   
## 1311 63 50 36   
## Entries  
## Near Flat/Level Banked Hillside Depression   
## 89.79 4.32 3.42 2.47

 Notiamo che tendenzialmente il ~90% circa delle case prese in analisi sono in posizioni quasi completamente piane. Case situate in zone con depressioni rappresentano la minoranza con solo 36 entrate in confronto alle 1311 della pianura.

### Utilities

Univar\_Quality(  
 HP,Utilities,  
 label="Types",  
 title="Utilities")

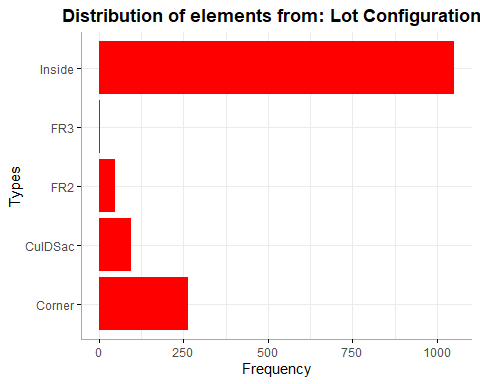
## Entries  
## AllPub NoSeWa   
## 1459 1   
## Entries  
## AllPub NoSeWa   
## 99.93 0.07

 Notiamo che il resto delle categorie di Utilities è completamente trascurabile, visto che solo una casa appare sotto la categoria “Senza collegamento alle fogne”, mentre tutte le altre possiedono tutte i principali servizi pubblici.

### LotConfig

Univar\_Quality(  
 HP,LotConfig,  
 label="Types",  
 title="Lot Configurations")

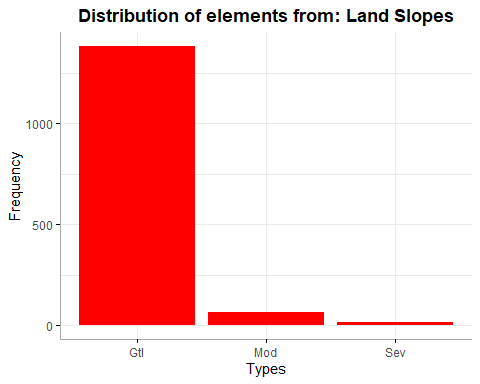
## Entries  
## Inside Corner CulDSac FR2 FR3   
## 1052 263 94 47 4   
## Entries  
## Inside Corner CulDSac FR2 FR3   
## 72.05 18.01 6.44 3.22 0.27

 Notiamo che la configurazione prevalente del lotto è quella interna con una frequenza di 1052, seguita poi da quella ad angolo di frequenza 263. Invece quelle con tre facciate sull’immobile sono solo 4, quindi trascurabili.

### LandSlope

Univar\_Quality(  
 HP,LandSlope,  
 label="Types",  
 title="Land Slopes")

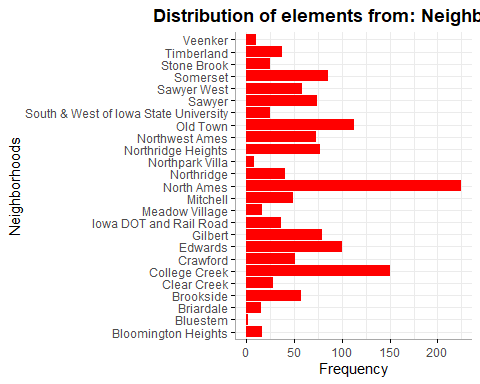
## Entries  
## Gtl Mod Sev   
## 1382 65 13   
## Entries  
## Gtl Mod Sev   
## 94.66 4.45 0.89

 Notiamo che la maggior parte delle proprietà (~94%) è inclinata solo leggermente, dato coerente con LotContour che indicava una prevaelenza di case in pianura. Le abitazioni maggiormente inclinate hanno frequenza inferiore all’ ~1%, quindi sono trascurabili.

### Neighborhood

Univar\_Quality(  
 HP,Neighborhood,   
 label="Neighborhoods",  
 title="Neighborhoods")

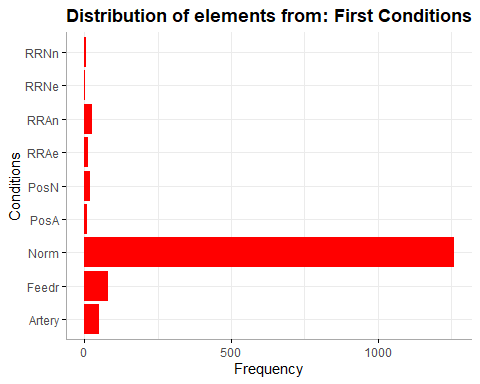
## Entries  
## North Ames College Creek   
## 225 150   
## Old Town Edwards   
## 113 100   
## Somerset Gilbert   
## 86 79   
## Northridge Heights Sawyer   
## 77 74   
## Northwest Ames Sawyer West   
## 73 59   
## Brookside Crawford   
## 58 51   
## Mitchell Northridge   
## 49 41   
## Timberland Iowa DOT and Rail Road   
## 38 37   
## Clear Creek South & West of Iowa State University   
## 28 25   
## Stone Brook Bloomington Heights   
## 25 17   
## Meadow Village Briardale   
## 17 16   
## Veenker Northpark Villa   
## 11 9   
## Bluestem   
## 2   
## Entries  
## North Ames College Creek   
## 15.41 10.27   
## Old Town Edwards   
## 7.74 6.85   
## Somerset Gilbert   
## 5.89 5.41   
## Northridge Heights Sawyer   
## 5.27 5.07   
## Northwest Ames Sawyer West   
## 5.00 4.04   
## Brookside Crawford   
## 3.97 3.49   
## Mitchell Northridge   
## 3.36 2.81   
## Timberland Iowa DOT and Rail Road   
## 2.60 2.53   
## Clear Creek South & West of Iowa State University   
## 1.92 1.71   
## Stone Brook Bloomington Heights   
## 1.71 1.16   
## Meadow Village Briardale   
## 1.16 1.10   
## Veenker Northpark Villa   
## 0.75 0.62   
## Bluestem   
## 0.14

 Notiamo che le frequenze maggiori si verificano per North Ames (225), seguita da College Creek (150), Old town e Edwards che sono vicini (113 e 100). I Vicinati meno significativi a livello di frequenza sono “Veenker”, “North Park Villa” e “Bluestem”, tutti sotto l’ ~1%.

### Condition1

Univar\_Quality(  
 HP,Condition1,  
 label="Conditions",  
 title="First Conditions")

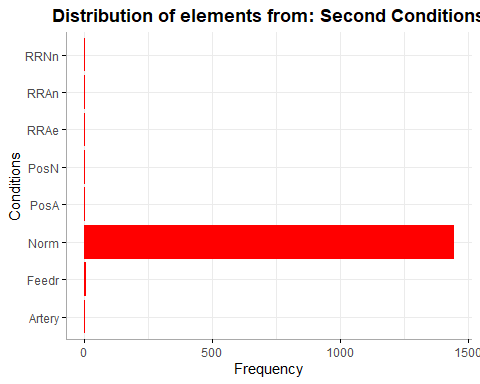
## Entries  
## Norm Feedr Artery RRAn PosN RRAe PosA RRNn RRNe   
## 1260 81 48 26 19 11 8 5 2   
## Entries  
## Norm Feedr Artery RRAn PosN RRAe PosA RRNn RRNe   
## 86.30 5.55 3.29 1.78 1.30 0.75 0.55 0.34 0.14

 Notiamo che la maggior parte della popolazione analizzata risiede in luoghi senza vicinanze degne di nota. Le più diffuse tra queste sono quella a una strada di svincolo o a una strada arteria. Molto meno diffuse invece sono le case in prossimità della ferrovia, le quali in questo caso risultano trascurabili.

### Condition2

Univar\_Quality(  
 HP,Condition2,  
 label="Conditions",  
 title="Second Conditions")

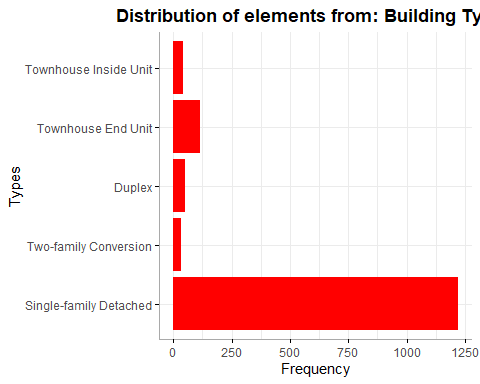
## Entries  
## Norm Feedr Artery PosN RRNn PosA RRAe RRAn   
## 1445 6 2 2 2 1 1 1   
## Entries  
## Norm Feedr Artery PosN RRNn PosA RRAe RRAn   
## 98.97 0.41 0.14 0.14 0.14 0.07 0.07 0.07

 Notiamo che i dati sono molto sbilanciati, con 1445 (o il 98%) case senza alcuna vicinanza degna di nota. Tutte le altre condizioni hanno frequenza inferiore all’ ~1%.

### BldgType

Univar\_Quality(  
 HP,BldgType,   
 label="Types",  
 title="Building Types")

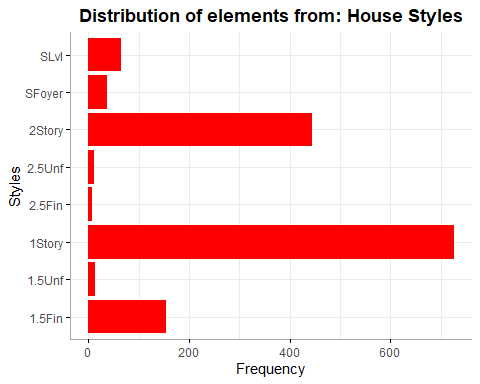
## Entries  
## Single-family Detached Townhouse End Unit Duplex   
## 1220 114 52   
## Townhouse Inside Unit Two-family Conversion   
## 43 31   
## Entries  
## Single-family Detached Townhouse End Unit Duplex   
## 83.56 7.81 3.56   
## Townhouse Inside Unit Two-family Conversion   
## 2.95 2.12

 Notiamo che il tipo di abitazione più diffuso è quello staccato per una famiglia con l’ ~84% di frequenze. Le unità terminali di residenze a schiera hanno frequenza 114, mentre le altre tipologie hanno frequenze decrescenti e molto basse, ma non inferiori all’~1%.

### HouseStyle

Univar\_Quality(  
 HP,HouseStyle,  
 label="Styles",  
 title="House Styles")

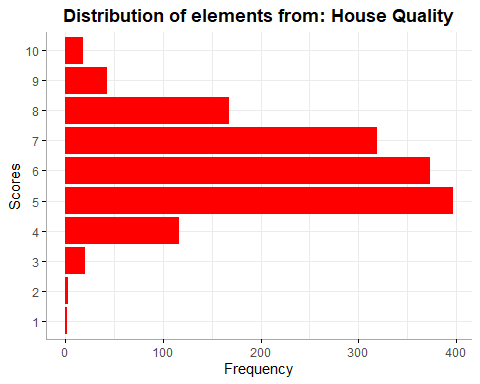
## Entries  
## 1Story 2Story 1.5Fin SLvl SFoyer 1.5Unf 2.5Unf 2.5Fin   
## 726 445 154 65 37 14 11 8   
## Entries  
## 1Story 2Story 1.5Fin SLvl SFoyer 1.5Unf 2.5Unf 2.5Fin   
## 49.73 30.48 10.55 4.45 2.53 0.96 0.75 0.55

 Notiamo che gli stili più diffusi sono quello a un piano e quello a due piani che insieme compongono l’\80% degli elementi. Le abitazioni in stato non finito invece rappresentano la minoranza in questo grafico, risultando ognuna sotto l’\1%.

### OverallQual

Univar\_Quality(  
 HP,OverallQual,  
 label="Scores",  
 title="House Quality")

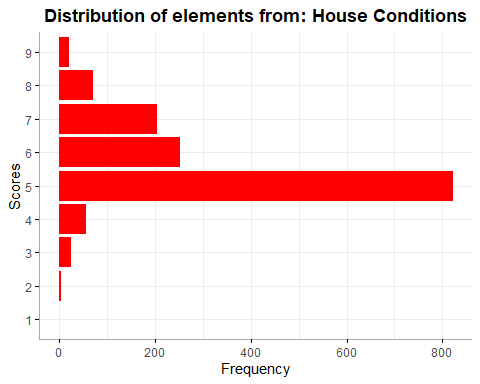
## Entries  
## 5 6 7 8 4 9 3 10 2 1   
## 397 374 319 168 116 43 20 18 3 2   
## Entries  
## 5 6 7 8 4 9 3 10 2 1   
## 27.19 25.62 21.85 11.51 7.95 2.95 1.37 1.23 0.21 0.14

 Notiamo che la maggioranza delle abitazioni considerate vengono considerate a livello di qualità attorno o superiore alla media, con frequenza di 397 per “nella media”, 374 per “sopra la media” e 319 per “buono”. Tra i voti negativi il più degno di nota è il 4, con 116 entrate. Il numero complessivo di case valutate sotto il 3 è di 5, quindi sono frequenze trascurabili in questo caso.

### OverallCond

Univar\_Quality(  
 HP,OverallCond,  
 label="Scores",  
 title="House Conditions")

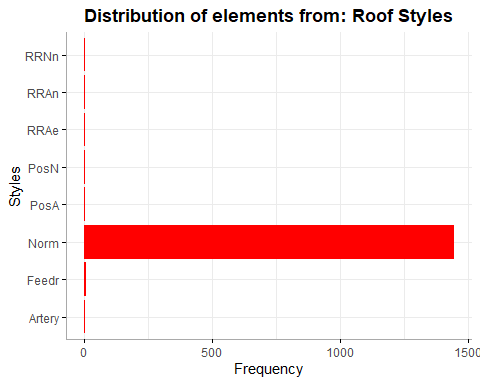
## Entries  
## 5 6 7 8 4 3 9 2 1 10   
## 821 252 205 72 57 25 22 5 1 0   
## Entries  
## 5 6 7 8 4 3 9 2 1 10   
## 56.23 17.26 14.04 4.93 3.90 1.71 1.51 0.34 0.07 0.00

 Troviamo la maggior parte delle frequenze sotto la valutazione “nella media”, con una prevalenza verso le valutazioni positive. Notiamo che nessuna casa ha valutazione 10, e le case valutate sotto il 3 sono in totale 6, un numero trascurabile rispetto alle 821 valutate 5.

### RoofStyle

Univar\_Quality(  
 HP,Condition2,  
 label="Styles",  
 title="Roof Styles")

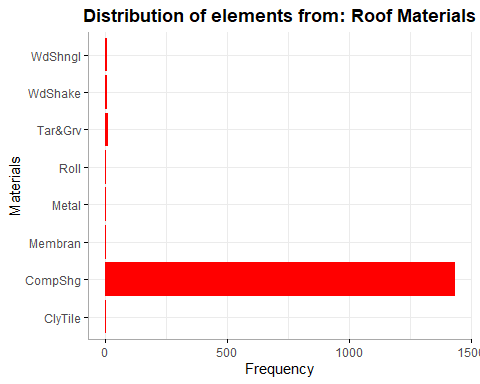
## Entries  
## Norm Feedr Artery PosN RRNn PosA RRAe RRAn   
## 1445 6 2 2 2 1 1 1   
## Entries  
## Norm Feedr Artery PosN RRNn PosA RRAe RRAn   
## 98.97 0.41 0.14 0.14 0.14 0.07 0.07 0.07

 I dati sono sbilanciati completamente verso lo stile normale (~99% circa), rendendo trascurabili tutte le altre opzioni aventi tutte meno del 0.5% frequenza.

### RoofMatl

Univar\_Quality(  
 HP,RoofMatl,   
 label="Materials",  
 title="Roof Materials")

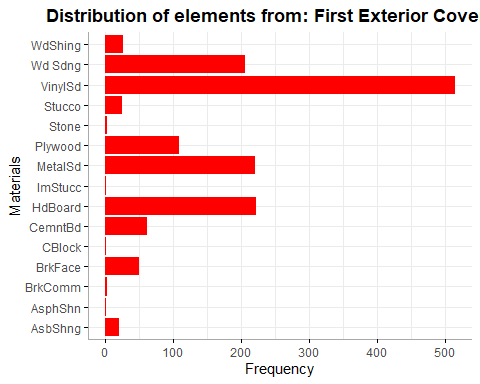
## Entries  
## CompShg Tar&Grv WdShngl WdShake ClyTile Membran Metal Roll   
## 1434 11 6 5 1 1 1 1   
## Entries  
## CompShg Tar&Grv WdShngl WdShake ClyTile Membran Metal Roll   
## 98.22 0.75 0.41 0.34 0.07 0.07 0.07 0.07

 Notiamo che i dati sono sbilanciati (~98%) verso l’utilizzo di una tipologia normale di materiale per il tetto, mentre tutte le altre sono estremamente basse .

### Exterior1st

Univar\_Quality(  
 HP,Exterior1st,  
 label="Materials",  
 title="First Exterior Covering")

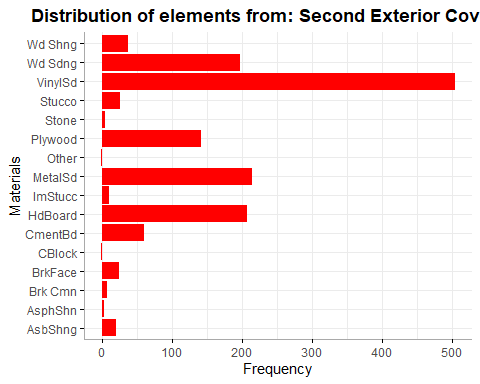
## Entries  
## VinylSd HdBoard MetalSd Wd Sdng Plywood CemntBd BrkFace WdShing Stucco AsbShng   
## 515 222 220 206 108 61 50 26 25 20   
## BrkComm Stone AsphShn CBlock ImStucc   
## 2 2 1 1 1   
## Entries  
## VinylSd HdBoard MetalSd Wd Sdng Plywood CemntBd BrkFace WdShing Stucco AsbShng   
## 35.27 15.21 15.07 14.11 7.40 4.18 3.42 1.78 1.71 1.37   
## BrkComm Stone AsphShn CBlock ImStucc   
## 0.14 0.14 0.07 0.07 0.07

 Notiamo che le coperture in vinile sono le più diffuse con una frequenza del ~35%, mentre quelle in tavola dura, quelle in metallo e quelle in legno sono tutte diffuse in modo bilanciato, avendo a ~15% di frequenza ognuno. Più elementi, come le coperture in pietra, asfalto, mattoni e stucco sono presenti in quantità trascurabili.

### Exterior2nd

Univar\_Quality(  
 HP,Exterior2nd,  
 label="Materials",  
 title="Second Exterior Covering")

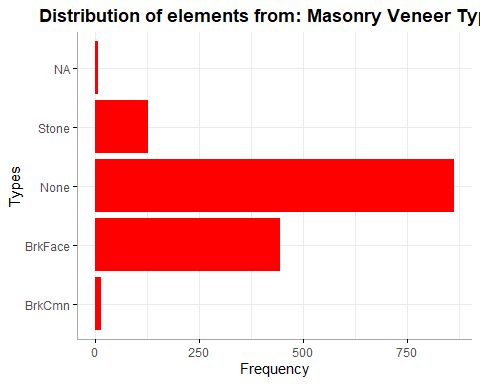
## Entries  
## VinylSd MetalSd HdBoard Wd Sdng Plywood CmentBd Wd Shng Stucco BrkFace AsbShng   
## 504 214 207 197 142 60 38 26 25 20   
## ImStucc Brk Cmn Stone AsphShn CBlock Other   
## 10 7 5 3 1 1   
## Entries  
## VinylSd MetalSd HdBoard Wd Sdng Plywood CmentBd Wd Shng Stucco BrkFace AsbShng   
## 34.52 14.66 14.18 13.49 9.73 4.11 2.60 1.78 1.71 1.37   
## ImStucc Brk Cmn Stone AsphShn CBlock Other   
## 0.68 0.48 0.34 0.21 0.07 0.07

 Notiamo che tra i secondi rivestimenti, il più diffuso rimane quello in vinile con 504 di frequenza, con quello di metallo, quello di tavola dura e quelli in legno distribuiti equamente ognuno con all’incirca 220 di frequenza. I meno presenti invece risultano essere quelli in imitazione di stucco, in asfato, in pietra,in mattoni e in calcestruzzo.

### MasVnrType

Univar\_Quality(  
 HP,MasVnrType,  
 label="Types",  
 title="Masonry Veneer Types")

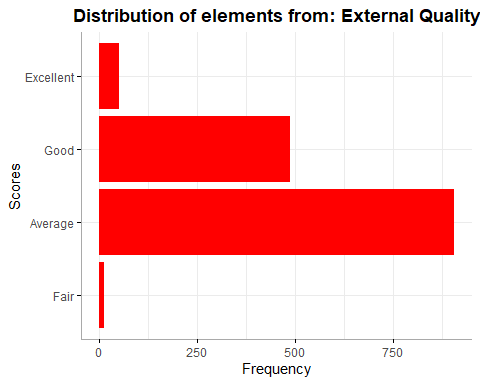
## Entries  
## None BrkFace Stone BrkCmn   
## 864 445 128 15   
## Entries  
## None BrkFace Stone BrkCmn   
## 59.50 30.65 8.82 1.03

 Notiamo che la cosa più diffusa è l’assenza di rivestimenti in muratura (~\60%). Tra gli elementi restanti, il rivestimento più diffuso è quello in mattoni (~30%), seguito da quello in pietra (~9%).

### ExterQual

Univar\_Quality(  
 HP,ExterQual,  
 label="Scores",  
 title="External Quality")

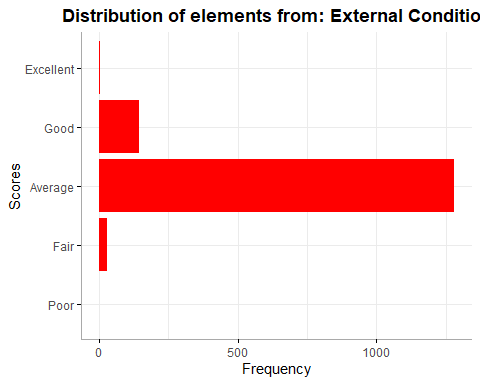
## Entries  
## Average Good Excellent Fair Poor   
## 906 488 52 14 0   
## Entries  
## Average Good Excellent Fair Poor   
## 62.05 33.42 3.56 0.96 0.00

 Notiamo che la maggior parte delle valutazioni cade sotto “Nella media” con frequenza di 906. Le valutazioni “buone” rappresentano il ~33% di quelle analizzate, mentre quelle negative risultano trascurabili.

### ExterCond

Univar\_Quality(  
 HP,ExterCond,  
 label="Scores",  
 title="External Condition")

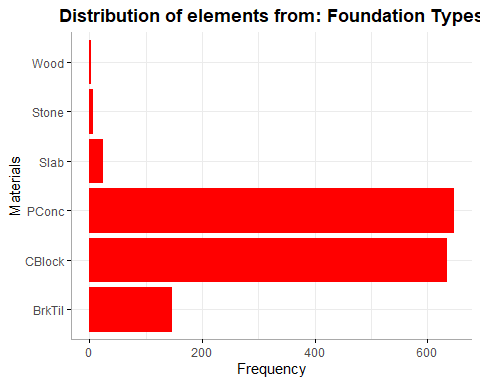
## Entries  
## Average Good Fair Excellent Poor   
## 1282 146 28 3 1   
## Entries  
## Average Good Fair Excellent Poor   
## 87.81 10.00 1.92 0.21 0.07

 Notiamo che tra le valutazioni delle condizioni esterne, la più diffusa è “nella media” rappresentando l’ ~88% delle frequenze. Le restanti percentuali sono principalmente trovate in “buono” (~10%) e “accettabile” (~2%). Le restanti valutazioni risultano trascurabili.

### Foundation

Univar\_Quality(  
 HP,Foundation,  
 label="Materials",  
 title="Foundation Types")

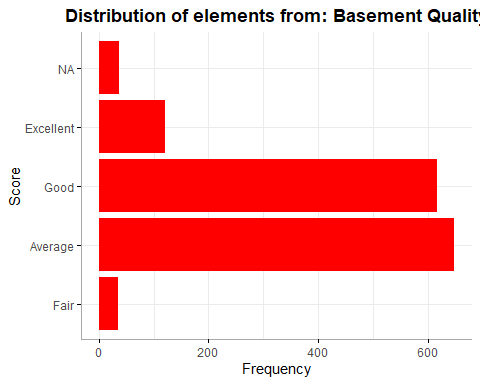
## Entries  
## PConc CBlock BrkTil Slab Stone Wood   
## 647 634 146 24 6 3   
## Entries  
## PConc CBlock BrkTil Slab Stone Wood   
## 44.32 43.42 10.00 1.64 0.41 0.21

 Tra le tipologie di fondamenta notiamo una equa distibuzione della maggior parte delle frequenze tra cemento (~44%) e calcestruzzo (~43%). Tra le restanti tipologie, quella in mattoni rappresenta il ~10% con 146 di frequenza, mentre quelle in legno e pietra appaiono trascurabili.

### BsmtQual

Univar\_Quality(  
 HP,BsmtQual,  
 label="Score",  
 title="Basement Quality")

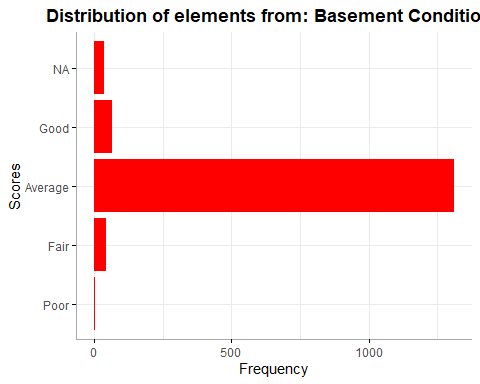
## Entries  
## Average Good Excellent Fair Poor   
## 649 618 121 35 0   
## Entries  
## Average Good Excellent Fair Poor   
## 45.61 43.43 8.50 2.46 0.00

 Notiamo che tra gli immobili che hanno il seminterrato, la maggior parte di questi si presentano di qualità medio/buona con frequenze di rispettivamente 649 e 618. Notiamo anche che 36 (~2,5%) abitazioni non possiedono il seminterrato.

### BsmtCond

Univar\_Quality(  
 HP,BsmtCond,  
 label="Scores",  
 title="Basement Conditions")

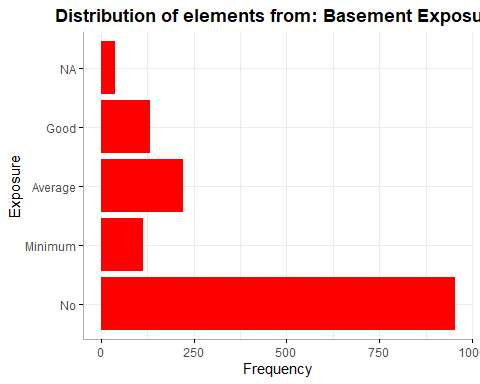
## Entries  
## Average Good Fair Poor Excellent   
## 1311 65 45 2 0   
## Entries  
## Average Good Fair Poor Excellent   
## 92.13 4.57 3.16 0.14 0.00

 Notiamo che 1311 valutazioni “nella media” compongono il ~92% delle frequenze, con “buono” e “accettabile” appena a ~4.6% e ~3.2%. Nessun seminterrato è stato valutato “eccellente” e “povero” risulta trascurabile in quanto applicata solo a 2 elementi. Notiamo anche che 36 (~2,5%) abitazioni non possiedono il seminterrato.

### BsmtExposure

Univar\_Quality(  
 HP,BsmtExposure,  
 label="Exposure",  
 title="Basement Exposure")

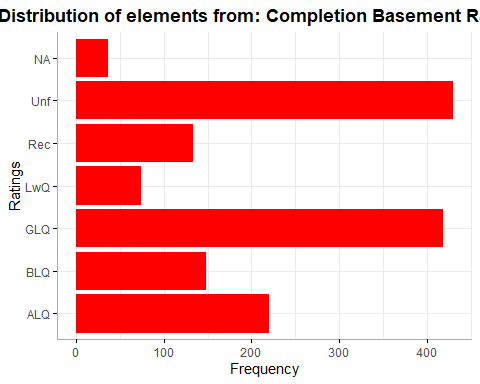
## Entries  
## No Average Good Minimum   
## 953 221 134 114   
## Entries  
## No Average Good Minimum   
## 67.02 15.54 9.42 8.02

 Notiamo che nella maggior parte delle abitazioni con seminterrati essi non sono esposti affatto (frequenza di 953), mentre i restanti sono divisi in “mediamente esposti” con 221 e quasi equamente tra “esposizione buona” 134 e “esposizione minima” 114. Notiamo anche che 36 (~2,5%) abitazioni non possiedono il seminterrato.

### BsmtFinType1

Univar\_Quality(  
 HP,BsmtFinType1,  
 label="Ratings",  
 title="Completion Basement Ratings")

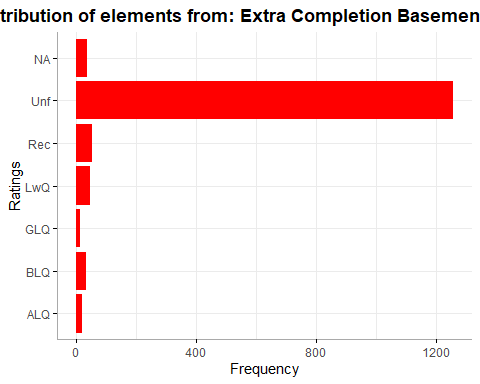
## Entries  
## Unf GLQ ALQ BLQ Rec LwQ   
## 430 418 220 148 133 74   
## Entries  
## Unf GLQ ALQ BLQ Rec LwQ   
## 30.22 29.37 15.46 10.40 9.35 5.20

 Analizzando i livelli di completamento notiamo che abbiamo delle distribuzioni simili tra i “non finiti” e quelli in “buone condizioni, entrambi attorno al ~30%. Non ci sono informazioni trascurabili.

### BsmtFinType2

Univar\_Quality(  
 HP,BsmtFinType2,  
 label="Ratings",  
 title="Extra Completion Basement Ratings")

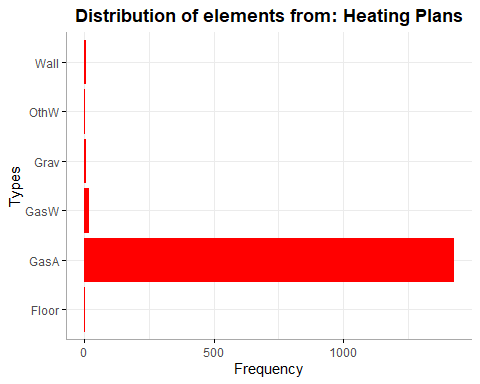
## Entries  
## Unf Rec LwQ BLQ ALQ GLQ   
## 1256 54 46 33 19 14   
## Entries  
## Unf Rec LwQ BLQ ALQ GLQ   
## 88.33 3.80 3.23 2.32 1.34 0.98

 Notiamo che nel caso in cui i seminterrati sono composti da più camere, quelle non principali risultano nel nostro caso essere prevalentemente “non finite” con una frequenza di 1256.

### Heating

Univar\_Quality(  
 HP,Heating,  
 label="Types",  
 title="Heating Plans")

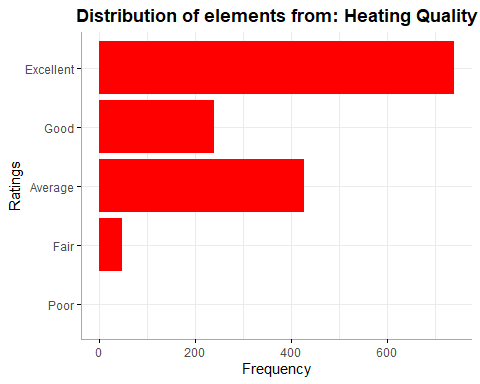
## Entries  
## GasA GasW Grav Wall OthW Floor   
## 1428 18 7 4 2 1   
## Entries  
## GasA GasW Grav Wall OthW Floor   
## 97.81 1.23 0.48 0.27 0.14 0.07

 Notiamo che il \98% circa della case analizzate hanno tutte il riscaldamento basato su fornaci ad aria calda. Delle restanti tipologie, quelle ad acqua calda sono le più prevalenti, anche se sono rappresentate tutte da basse frequenze.

### HeatingQC

Univar\_Quality(  
 HP,HeatingQC,  
 label="Ratings"  
 ,title="Heating Quality")

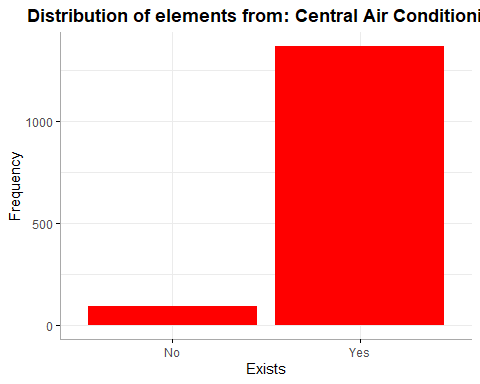
## Entries  
## Excellent Average Good Fair Poor   
## 741 428 241 49 1   
## Entries  
## Excellent Average Good Fair Poor   
## 50.75 29.32 16.51 3.36 0.07

 Notiamo che la maggior parte dei voti risulta positiva con frequenze 741 per “eccellente”, 428 per “nella media” e 241 per “buono”. Con frequenza 49, “accettabile” appare l’unico voto negativo degno di nota.

### CentralAir

Univar\_Quality(  
 HP,CentralAir,  
 label="Exists",  
 title="Central Air Conditioning")

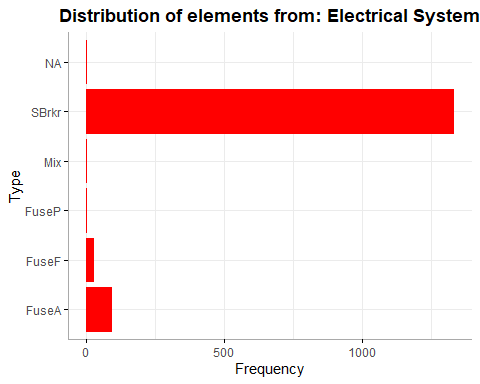
## Entries  
## Yes No   
## 1365 95   
## Entries  
## Yes No   
## 93.49 6.51

 Notiamo che con frequenza 1365, la maggior parte delle abitazioni ha installato un sistema di condizionatori dell’aria centralizzato.

### Electrical

Univar\_Quality(  
 HP,Electrical,  
 label="Type",  
 title="Electrical System")

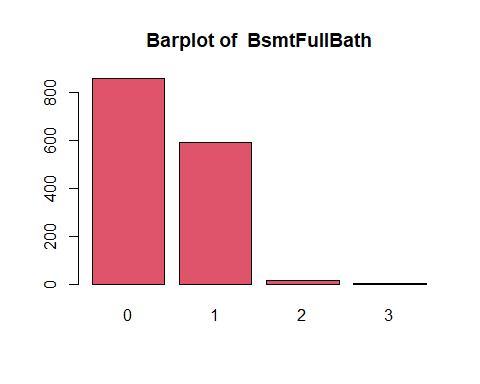
## Entries  
## SBrkr FuseA FuseF FuseP Mix   
## 1334 94 27 3 1   
## Entries  
## SBrkr FuseA FuseF FuseP Mix   
## 91.43 6.44 1.85 0.21 0.07

 Notiamo che la maggior parte dei sistemi elettrici installati (\91%) impiegano il salvavita, mentre il resto delle popolazioni significanti sono quelle del fusibile A e del fusibile F.

### BsmtFullBath

visualization\_discrete(  
 na.omit(HP[,"BsmtFullBath"]),  
 sum\_info\_h[14],  
 "BsmtFullBath")

## [[1]]  
## [[1]]$BsmtFullBath  
## [[1]]$BsmtFullBath$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.4253 1.0000 3.0000   
##   
## [[1]]$BsmtFullBath$Variance  
## [1] 0.27  
##   
## [[1]]$BsmtFullBath$Standard\_Deviation  
## [1] 0.52  
##   
## [[1]]$BsmtFullBath$Moda  
## [[1]]$BsmtFullBath$Moda$Mode  
## [1] 0  
##   
## [[1]]$BsmtFullBath$Moda$Frequency  
## [1] 856  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 0 1 2 3   
## 856 588 15 1   
##   
## [1] "percintile: "  
##   
## data  
## 0 1 2 3   
## 58.63013699 40.27397260 1.02739726 0.06849315

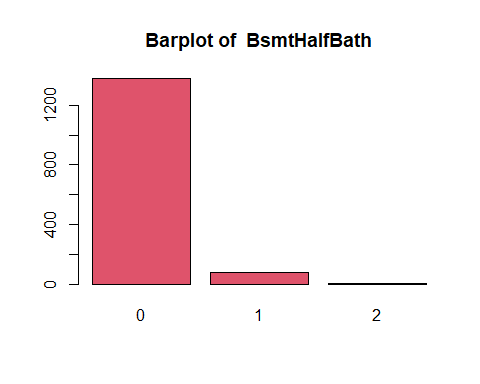


Si osserva che la maggior parte dei semiterrati non hanno un bagno completo essendo il ~=58.6% con frequenza di 856 Si osserva che la minima freq dei numeri di bagno completi e 1 per 3 bangi e quindi si puo trascurare essendo <1% Si osserva che la freq dei numeri di bagni completi per 2 bagni e 15 che è ~= 1.03% e quindi anch’esso trascurabile Infine si osserva che la maggior parte dei semiterrati che hanno 1 bagno completo è ~=40.3% con frequenza di 588

### BsmtHalfBath

visualization\_discrete(  
 na.omit(HP[,"BsmtHalfBath"]),  
 sum\_info\_h[15],  
 "BsmtHalfBath")

## [[1]]  
## [[1]]$BsmtHalfBath  
## [[1]]$BsmtHalfBath$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00000 0.00000 0.00000 0.05753 0.00000 2.00000   
##   
## [[1]]$BsmtHalfBath$Variance  
## [1] 0.06  
##   
## [[1]]$BsmtHalfBath$Standard\_Deviation  
## [1] 0.24  
##   
## [[1]]$BsmtHalfBath$Moda  
## [[1]]$BsmtHalfBath$Moda$Mode  
## [1] 0  
##   
## [[1]]$BsmtHalfBath$Moda$Frequency  
## [1] 1378  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 0 1 2   
## 1378 80 2   
##   
## [1] "percintile: "  
##   
## data  
## 0 1 2   
## 94.3835616 5.4794521 0.1369863

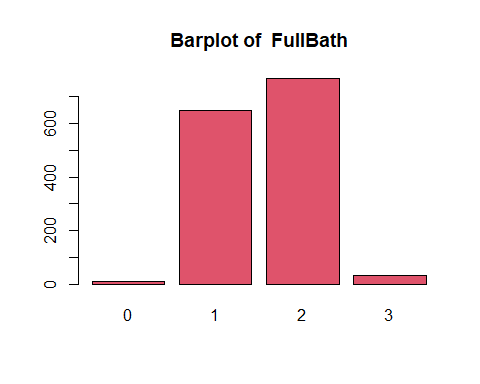


Si osserva che la maggior parte dei semiterrati non hanno un bagno mezzo completo essendo il ~94% con frequenza di 1378 Si osserva che la freq dei numeri di bagni mezzi completi per 2 bagni e 2 che è ~= 1% e quindi anch’esso trascurabile Infine si osserva che la maggior parte dei semiterrati che hanno 1 bagno mezzo completo è ~5% con frequenza di 80

### FullBath

visualization\_discrete(  
 na.omit(HP[,"FullBath"]),  
 sum\_info\_h[16],"FullBath")

## [[1]]  
## [[1]]$FullBath  
## [[1]]$FullBath$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 2.000 1.565 2.000 3.000   
##   
## [[1]]$FullBath$Variance  
## [1] 0.3  
##   
## [[1]]$FullBath$Standard\_Deviation  
## [1] 0.55  
##   
## [[1]]$FullBath$Moda  
## [[1]]$FullBath$Moda$Mode  
## [1] 2  
##   
## [[1]]$FullBath$Moda$Frequency  
## [1] 768  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 0 1 2 3   
## 9 650 768 33   
##   
## [1] "percintile: "  
##   
## data  
## 0 1 2 3   
## 0.6164384 44.5205479 52.6027397 2.2602740

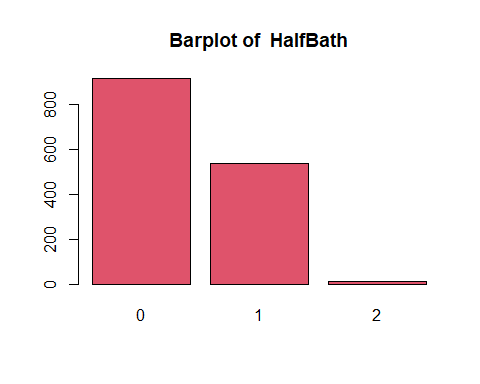


Si osserva che la maggior parte degli edifici che sono sopra il seminterrato hanno 2 bagni completi essendo il ~=52.6% con frequenza di 768 Si osserva che la minima freq dei numeri di bagno completi e 9 per nessun bango e quindi si puo trascurare essendo <1% Si osserva che la freq dei numeri di bango completi per 3 bangi e 33 che è ~= 2.2% e quindi anch’esso trascurabile Infine si osserva che gli edifici che sono sopra il seminterrato che hanno 1 bagno completo sono il ~=44.5% con frequenza di 650

### HalfBath

visualization\_discrete(  
 na.omit(HP[,"HalfBath"]),  
 sum\_info\_h[17],  
 "HalfBath")

## [[1]]  
## [[1]]$HalfBath  
## [[1]]$HalfBath$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0000 0.0000 0.0000 0.3829 1.0000 2.0000   
##   
## [[1]]$HalfBath$Variance  
## [1] 0.25  
##   
## [[1]]$HalfBath$Standard\_Deviation  
## [1] 0.5  
##   
## [[1]]$HalfBath$Moda  
## [[1]]$HalfBath$Moda$Mode  
## [1] 0  
##   
## [[1]]$HalfBath$Moda$Frequency  
## [1] 913  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 0 1 2   
## 913 535 12   
##   
## [1] "percintile: "  
##   
## data  
## 0 1 2   
## 62.5342466 36.6438356 0.8219178

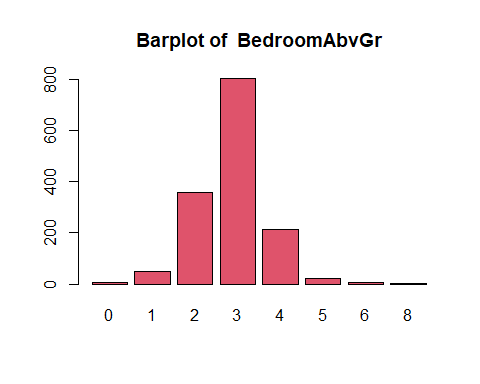


Si osserva che la maggior parte degli edifici che sono sopra il seminterrato hanno 0 bagni mezzo completi essendo il ~=62.5% con frequenza di 913 Si osserva che la minima freq dei numeri di bagno completi e 12 per 2 bango e quindi si puo trascurare essendo <1% Infine si osserva che gli edifici che sono sopra il seminterrato che hanno 1 bagno completo sono il ~=36.6% con frequenza di 535

### BedroomAbvGr

visualization\_discrete(  
 na.omit(HP[,"BedroomAbvGr"]),  
 sum\_info\_h[18],  
 "BedroomAbvGr")

## [[1]]  
## [[1]]$BedroomAbvGr  
## [[1]]$BedroomAbvGr$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 2.000 3.000 2.866 3.000 8.000   
##   
## [[1]]$BedroomAbvGr$Variance  
## [1] 0.67  
##   
## [[1]]$BedroomAbvGr$Standard\_Deviation  
## [1] 0.82  
##   
## [[1]]$BedroomAbvGr$Moda  
## [[1]]$BedroomAbvGr$Moda$Mode  
## [1] 3  
##   
## [[1]]$BedroomAbvGr$Moda$Frequency  
## [1] 804  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 0 1 2 3 4 5 6 8   
## 6 50 358 804 213 21 7 1   
##   
## [1] "percintile: "  
##   
## data  
## 0 1 2 3 4 5   
## 0.41095890 3.42465753 24.52054795 55.06849315 14.58904110 1.43835616   
## 6 8   
## 0.47945205 0.06849315

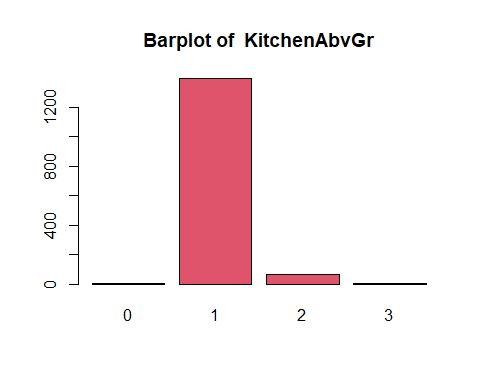


Si osserva che la maggior parte degli edifici che sono sopra il seminterrato hanno 3 camere essendo il ~=55.07% con frequenza di 804 seguito da 2, 4 e 1 camere essendo ~=24.5%, ~=14.6% e ~=3.4% con frequenze 358, 213 e 50 rispettivamente. Si osserva che la minima freq dei numeri di camere e 1 per 8 camere seguito da 5, 6 e 0 camere con frequenze 21, 7 e 6 rispettivamente e quindi si puo trascurare essendo tutti <1.5%

### KitchenAbvGr

visualization\_discrete(  
 na.omit(HP[,"KitchenAbvGr"]),  
 sum\_info\_h[19],  
 "KitchenAbvGr")

## [[1]]  
## [[1]]$KitchenAbvGr  
## [[1]]$KitchenAbvGr$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 1.000 1.047 1.000 3.000   
##   
## [[1]]$KitchenAbvGr$Variance  
## [1] 0.05  
##   
## [[1]]$KitchenAbvGr$Standard\_Deviation  
## [1] 0.22  
##   
## [[1]]$KitchenAbvGr$Moda  
## [[1]]$KitchenAbvGr$Moda$Mode  
## [1] 1  
##   
## [[1]]$KitchenAbvGr$Moda$Frequency  
## [1] 1392  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 0 1 2 3   
## 1 1392 65 2   
##   
## [1] "percintile: "  
##   
## data  
## 0 1 2 3   
## 0.06849315 95.34246575 4.45205479 0.13698630

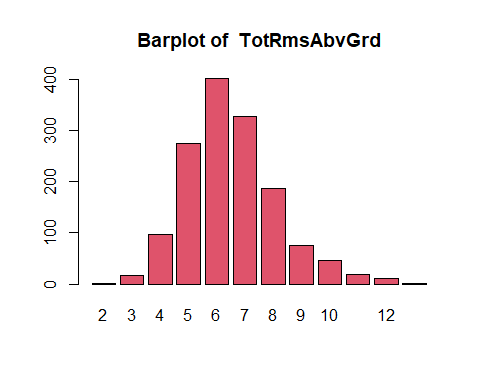


Si osserva che la maggior parte degli edifici che sono sopra il seminterrato hanno 1 cucina essendo il ~=95.34% con frequenza di 1392 seguito da 2 cucine essendo ~=4.45% con frequenze 65. Si osserva che la minima freq dei numeri di camere e 1 per 0 cucine seguito da 3 cucine con frequenze 2 e quindi si puo trascurare essendo tutti <1%

### TotRmsAbvGrd

visualization\_discrete(  
 na.omit(HP[,"TotRmsAbvGrd"]),  
 sum\_info\_h[20],  
 "TotRmsAbvGrd")

## [[1]]  
## [[1]]$TotRmsAbvGrd  
## [[1]]$TotRmsAbvGrd$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.000 5.000 6.000 6.518 7.000 14.000   
##   
## [[1]]$TotRmsAbvGrd$Variance  
## [1] 2.64  
##   
## [[1]]$TotRmsAbvGrd$Standard\_Deviation  
## [1] 1.63  
##   
## [[1]]$TotRmsAbvGrd$Moda  
## [[1]]$TotRmsAbvGrd$Moda$Mode  
## [1] 6  
##   
## [[1]]$TotRmsAbvGrd$Moda$Frequency  
## [1] 402  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 2 3 4 5 6 7 8 9 10 11 12 14   
## 1 17 97 275 402 329 187 75 47 18 11 1   
##   
## [1] "percintile: "  
##   
## data  
## 2 3 4 5 6 7   
## 0.06849315 1.16438356 6.64383562 18.83561644 27.53424658 22.53424658   
## 8 9 10 11 12 14   
## 12.80821918 5.13698630 3.21917808 1.23287671 0.75342466 0.06849315

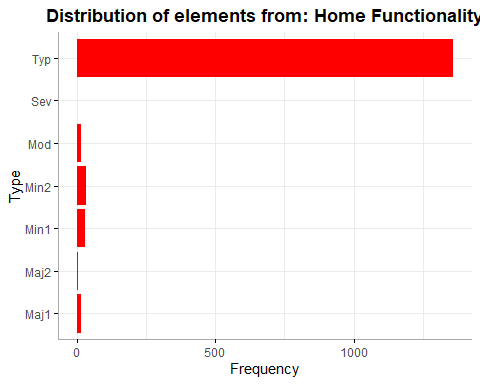


Si osserva che la maggior parte degli edifici che sono sopra il seminterrato sono nel intervallo fra 5 e 7 stanze (non considerando i bagni) essendo il ~=68.8% con il piu grande di occorenze 6 stanze (~=27.5%) con frequenza di 402. Si osserva che la minima freq dei numeri di stanze e 1 per 2 e 14 stanze seguito da 12, 3 e 11 stanze con frequenze 11, 17 e 18 rispettivamente e quindi si puo trascurare essendo tutti <1.5%

### Functional

Univar\_Quality(  
 HP,Functional,  
 label="Type",  
 title="Home Functionality")

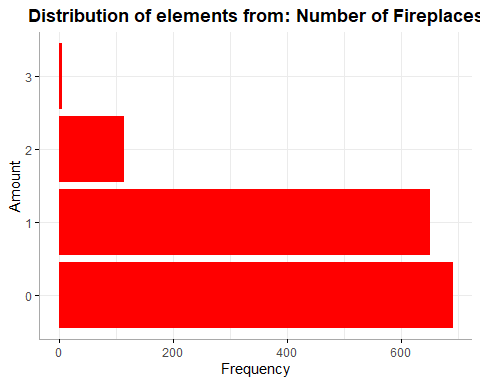
## Entries  
## Typ Min2 Min1 Mod Maj1 Maj2 Sev   
## 1360 34 31 15 14 5 1   
## Entries  
## Typ Min2 Min1 Mod Maj1 Maj2 Sev   
## 93.15 2.33 2.12 1.03 0.96 0.34 0.07

 I dati sono fortemente sbilanciati verso una funzionalità tipica della casa, rappresentando il \93% delle frequenze. Le abitazioni danneggiate rappresentano una minoranza in questi dati, risultando trascurabili.

### Fireplaces

Univar\_Quality(  
 HP,Fireplaces,  
 label="Amount",  
 title="Number of Fireplaces")

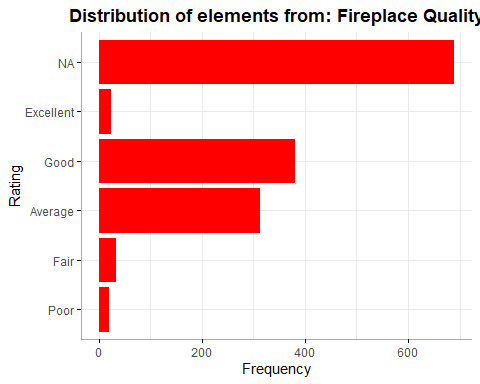
## Entries  
## 0 1 2 3   
## 690 650 115 5   
## Entries  
## 0 1 2 3   
## 47.26 44.52 7.88 0.34

 Notiamo che nel ~47% degli immobili analizzati essi risultano senza stufe a legna. Di quelli che ne posseggono, osserviamo che il ~45% contiene una sola stufa, l’~8% due e meno dell’~1% tre.

### FireplaceQu

Univar\_Quality(  
 HP,FireplaceQu,  
 label="Rating",  
 title="Fireplace Quality")

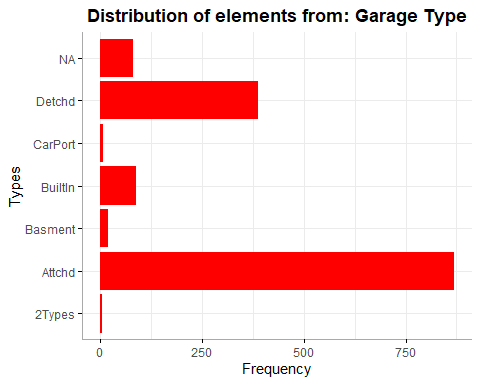
## Entries  
## Good Average Fair Excellent Poor   
## 380 313 33 24 20   
## Entries  
## Good Average Fair Excellent Poor   
## 49.35 40.65 4.29 3.12 2.60

 Notiamo che tra le abitazioni dotate di camini, 380 (~49%) di esse lo hanno valutato come “buono”, 313 (~40%) come “nella media”. Gli “eccellenti” e i “poveri” risultano avere numeri paragonabili, con 24 (~3.1%) e 20 di frequenza.(~2.6%)

### GarageType

Univar\_Quality(  
 HP,GarageType,  
 label="Types",  
 title="Garage Type")

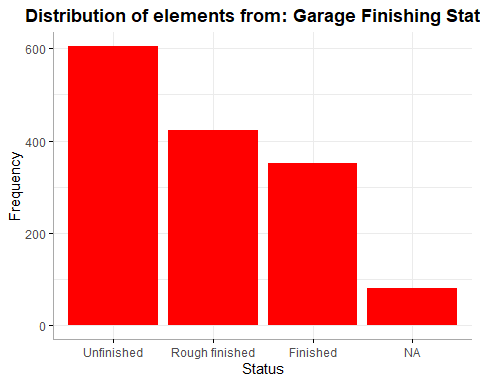
## Entries  
## Attchd Detchd BuiltIn Basment CarPort 2Types   
## 870 387 88 19 9 6   
## Entries  
## Attchd Detchd BuiltIn Basment CarPort 2Types   
## 63.09 28.06 6.38 1.38 0.65 0.44

 Notiamo che tra i tipi di garage associati alle varie abitazioni controllate risultano 870 (~63%) di essi attaccati alla stessa, 387 (~28%) staccati e 88 (~6%) costruiti dentro l’edificio. Osserviamo anche che le abitazioni con due garage di tipi diversi o l’entrata per la macchina risultano nel complessivo trascurabili.

### GarageFinish

Univar\_Quality(  
 HP,GarageFinish,  
 label="Status",  
 title="Garage Finishing Status")

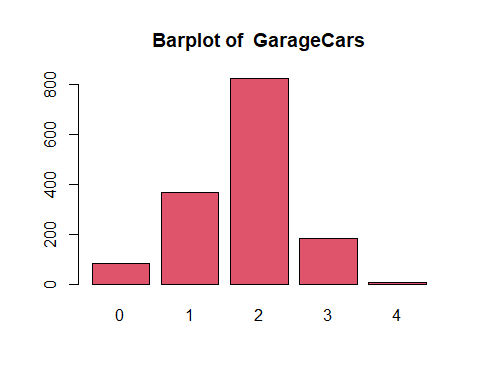
## Entries  
## Unfinished Rough finished Finished   
## 605 422 352   
## Entries  
## Unfinished Rough finished Finished   
## 43.87 30.60 25.53

 Tra le case dotate di garage, osserviamo che 605 (~44%) di essi risulta non completato, mentre quelli finiti e quelli quasi completi sono 422 (~31%) e 352 (~26%). Quindi nel complessivo il numero di garage generalmente quasi completi è superiore a quello dei garage incompleti. Osserviamo anche che 81 immobili non hanno il garage.

### GarageCars

visualization\_discrete(  
 na.omit(HP[,"GarageCars"]),  
 sum\_info\_h[23],  
 "GarageCars")

## [[1]]  
## [[1]]$GarageCars  
## [[1]]$GarageCars$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.000 1.000 2.000 1.767 2.000 4.000   
##   
## [[1]]$GarageCars$Variance  
## [1] 0.56  
##   
## [[1]]$GarageCars$Standard\_Deviation  
## [1] 0.75  
##   
## [[1]]$GarageCars$Moda  
## [[1]]$GarageCars$Moda$Mode  
## [1] 2  
##   
## [[1]]$GarageCars$Moda$Frequency  
## [1] 824  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 0 1 2 3 4   
## 81 369 824 181 5   
##   
## [1] "percintile: "  
##   
## data  
## 0 1 2 3 4   
## 5.5479452 25.2739726 56.4383562 12.3972603 0.3424658

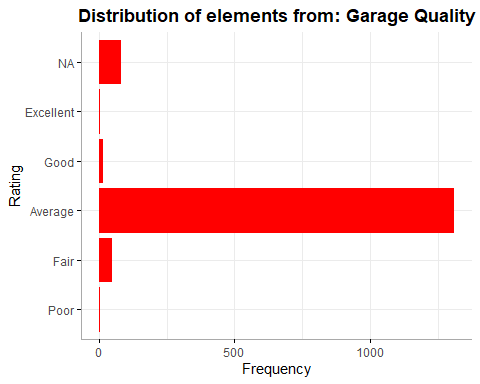


Si osserva che la maggior parte degli Garage possono avere 2 auto essendo il ~=56.44% con frequenza di 824 seguito da 1, 3 e 0 auto essendo ~=25.27% con frequenze 369, ~=12.4% con frequenze 181 e ~=5.55% con frequenze 81 rispettivamente. Notiamo che la minima freq dei numeri di auto e 5 per 4 auto e quindi si puo trascurare essendo <1%

### GarageQual

Univar\_Quality(HP,GarageQual, label="Rating",title="Garage Quality")

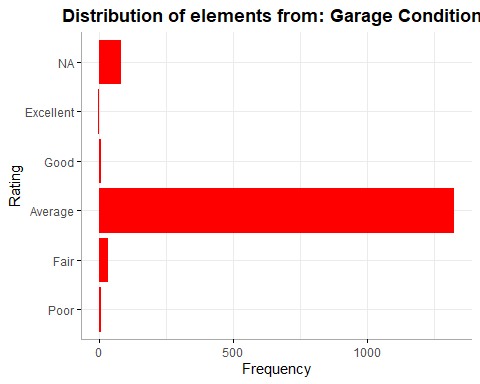
## Entries  
## Average Fair Good Poor Excellent   
## 1311 48 14 3 3   
## Entries  
## Average Fair Good Poor Excellent   
## 95.07 3.48 1.02 0.22 0.22

 Notiamo che i dati sono principalmente concentrati su “nella media”, con una frequenza di 1311 (~95%). I voti “eccellente” e “povero” risultano applicati a una popolazione molto piccola dalla frequenza di 3 ognuno, quindi appaiono trascurabili.

### GarageCond

Univar\_Quality(  
 HP,GarageCond,  
 label="Rating",  
 title="Garage Conditions")

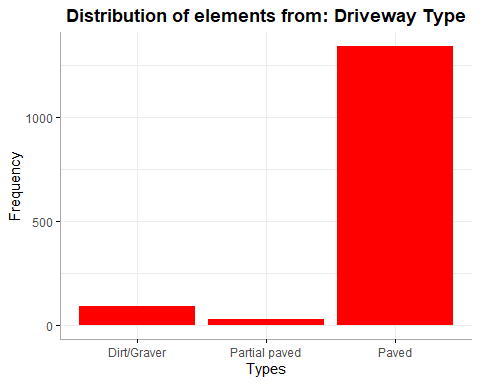
## Entries  
## Average Fair Good Poor Excellent   
## 1326 35 9 7 2   
## Entries  
## Average Fair Good Poor Excellent   
## 96.16 2.54 0.65 0.51 0.15

 Le distribuzioni riguardanti le condizioni sembrano rispecchiare quelli della qualità, questa volta con 1326 (~96%) di frequenza.

### PavedDrive

Univar\_Quality(  
 HP,PavedDrive,  
 label="Types",  
 title="Driveway Type")

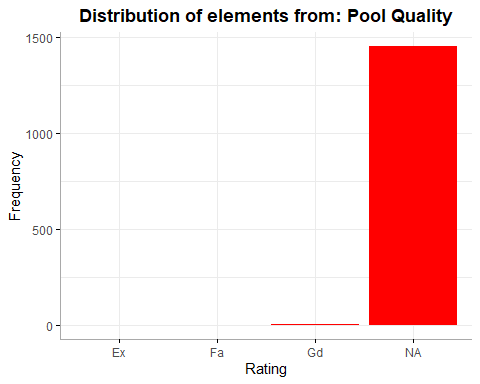
## Entries  
## Paved Dirt/Graver Partial paved   
## 1340 90 30   
## Entries  
## Paved Dirt/Graver Partial paved   
## 91.78 6.16 2.05

 Notiamo che le frequenze riguardanti il tipo di viale si concentrano principalmente su “pavimentato” con frequenza del 1340 (~92%).

### PoolQC

Univar\_Quality(  
 HP,PoolQC,  
 label="Rating",  
 title="Pool Quality")

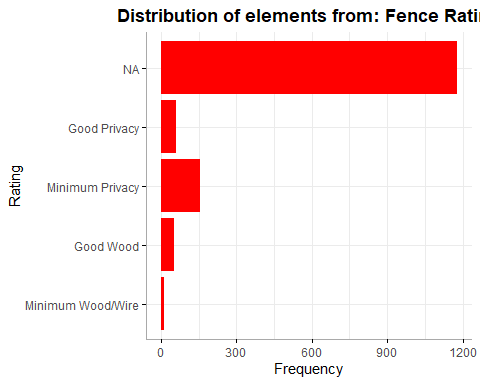
## Entries  
## Gd Ex Fa   
## 3 2 2   
## Entries  
## Gd Ex Fa   
## 42.86 28.57 28.57

 Osserviamo che 1453 delle abitazioni considerate non posseggono la piscina, con i restanti 7 quasi equamente distribuiti sulle altre valutazioni. Per questo motivo il dato risulta quasi interamente trascurabile.

### Fence

Univar\_Quality(  
 HP,Fence,  
 label="Rating",  
 title="Fence Rating")

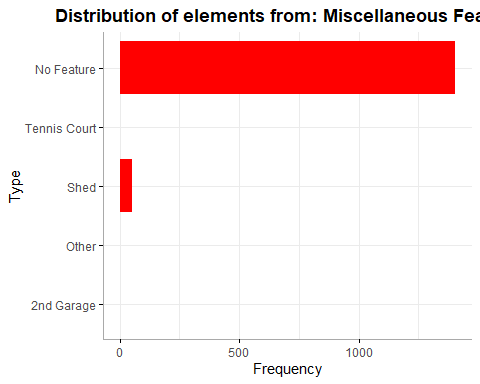
## Entries  
## Minimum Privacy Good Privacy Good Wood Minimum Wood/Wire   
## 157 59 54 11   
## Entries  
## Minimum Privacy Good Privacy Good Wood Minimum Wood/Wire   
## 55.87 21.00 19.22 3.91

 Notiamo che le case senza alcun recinto hanno una frequenza di 1179, quindi sono la maggioranza. Delle restanti 281, 157 hanno privacy minima (~55%), 59 hanno buona privacy (~21%) e 54 (~20%) hanno buon legno.

### MiscFeature

Univar\_Quality(  
 HP,MiscFeature,  
 label="Type",  
 title="Miscellaneous Feature")

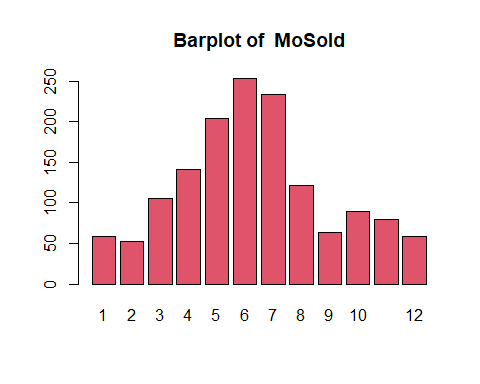
## Entries  
## No Feature Shed 2nd Garage Other Tennis Court Elevator   
## 1406 49 2 2 1 0   
## Entries  
## No Feature Shed 2nd Garage Other Tennis Court Elevator   
## 96.30 3.36 0.14 0.14 0.07 0.00

 Notiamo che 1406 immobili risultano non avere alcun elemento aggiuntivo, mentre tra quelle che lo hanno, le uniche degne di nota sembrano essere quelle con un capanno, le quali compongono il 3% (frequenza 49) delle case complessive, mentre le altre risultano trascurabili.

### MoSold

visualization\_discrete(na.omit(HP[,"MoSold"]), sum\_info\_h[32],"MoSold")

## [[1]]  
## [[1]]$MoSold  
## [[1]]$MoSold$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 5.000 6.000 6.322 8.000 12.000   
##   
## [[1]]$MoSold$Variance  
## [1] 7.31  
##   
## [[1]]$MoSold$Standard\_Deviation  
## [1] 2.7  
##   
## [[1]]$MoSold$Moda  
## [[1]]$MoSold$Moda$Mode  
## [1] 6  
##   
## [[1]]$MoSold$Moda$Frequency  
## [1] 253  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 1 2 3 4 5 6 7 8 9 10 11 12   
## 58 52 106 141 204 253 234 122 63 89 79 59   
##   
## [1] "percintile: "  
##   
## data  
## 1 2 3 4 5 6 7 8   
## 3.972603 3.561644 7.260274 9.657534 13.972603 17.328767 16.027397 8.356164   
## 9 10 11 12   
## 4.315068 6.095890 5.410959 4.041096

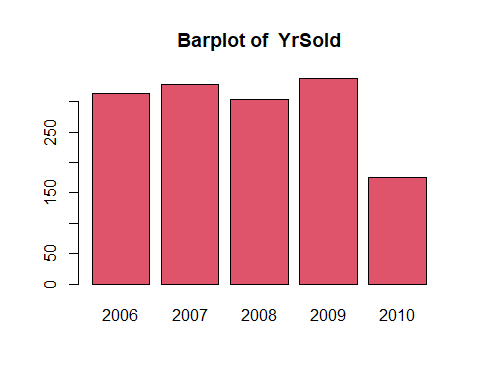


Possiamo notare che i mesi con le moggior vendute sono fra il 5 e 8 mese con8 ~=55.69% Dal 1 si ha una crescita di vendute fino al 6 mese dove si il massimo di vendite con 253 vendite dopodicche la vendita decresce

### YrSold

visualization\_discrete(  
 na.omit(HP[,"YrSold"]),  
 sum\_info\_h[33],  
 "YrSold")

## [[1]]  
## [[1]]$YrSold  
## [[1]]$YrSold$Summary  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2006 2007 2008 2008 2009 2010   
##   
## [[1]]$YrSold$Variance  
## [1] 1.76  
##   
## [[1]]$YrSold$Standard\_Deviation  
## [1] 1.33  
##   
## [[1]]$YrSold$Moda  
## [[1]]$YrSold$Moda$Mode  
## [1] 2009  
##   
## [[1]]$YrSold$Moda$Frequency  
## [1] 338  
##   
##   
##   
##   
## [1] "freq: "  
##   
## data  
## 2006 2007 2008 2009 2010   
## 314 329 304 338 175   
##   
## [1] "percintile: "  
##   
## data  
## 2006 2007 2008 2009 2010   
## 21.50685 22.53425 20.82192 23.15068 11.98630

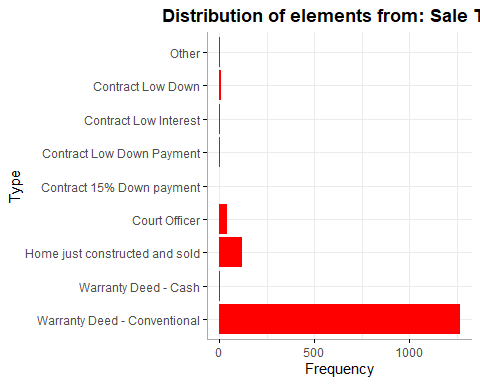


Si puo notare che la vendita per anno dal 2006 a 2009 e per lo piu uniforme con l’anno di maggior vendita nel 2009 con 338 venidite Nel 2010 invece si ha il minimo di vendite con 175 vendite

### SaleType

Univar\_Quality(  
 HP,SaleType,  
 label="Type",  
 title="Sale Type")

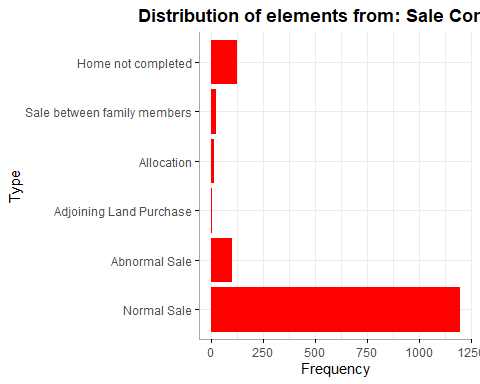
## Entries  
## Warranty Deed - Conventional Home just constructed and sold   
## 1267 122   
## Court Officer Contract Low Down   
## 43 9   
## Contract Low Down Payment Contract Low Interest   
## 5 5   
## Warranty Deed - Cash Other   
## 4 3   
## Contract 15% Down payment Warranty Deed - VA Loan   
## 2 0   
## Entries  
## Warranty Deed - Conventional Home just constructed and sold   
## 86.78 8.36   
## Court Officer Contract Low Down   
## 2.95 0.62   
## Contract Low Down Payment Contract Low Interest   
## 0.34 0.34   
## Warranty Deed - Cash Other   
## 0.27 0.21   
## Contract 15% Down payment Warranty Deed - VA Loan   
## 0.14 0.00

 Osserviamo che le distribuzioni di questa tabella sono sbilanciate verso “certificato di garanzia”, avendo frequenza di 1267 (~87%). Tra le altre tipologie di pagamento noto “casa appena costruita e venduta” e “tramite ufficiale del tribunale”, che rappresentano l’ ~8% e il ~3% delle frequenze. Le restanti categorie risultano trascurabili.

### SaleCondition

Univar\_Quality(  
 HP,SaleCondition,  
 label="Type",  
 title="Sale Condition")

## Entries  
## Normal Sale Home not completed   
## 1198 125   
## Abnormal Sale Sale between family members   
## 101 20   
## Allocation Adjoining Land Purchase   
## 12 4   
## Entries  
## Normal Sale Home not completed   
## 82.05 8.56   
## Abnormal Sale Sale between family members   
## 6.92 1.37   
## Allocation Adjoining Land Purchase   
## 0.82 0.27

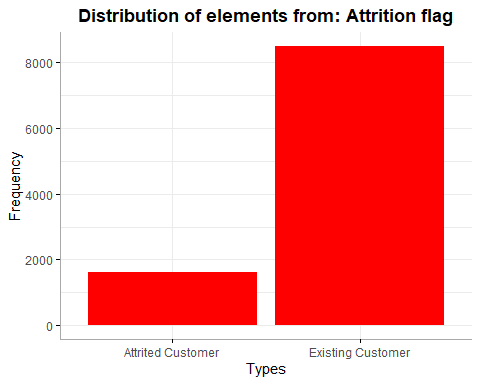
 Osserviamo che la maggior parte degli acquisti si è svolta in maniera normale, data la frequenza di 1198 (~86.78%), mentre il resto si è prevalentemente eseguito ad abitazione non completata (~9%), in maniera anomala (~7%) o in alcuni casi tra parenti (~1%). Le restanti modalità si presentano meno dell’~1% delle volte

# Credit Card univariata qualitativa

### Attrition\_Flag

Univar\_Quality(credit\_card,Attrition\_Flag,  
 label="Types",  
 title="Attrition flag")

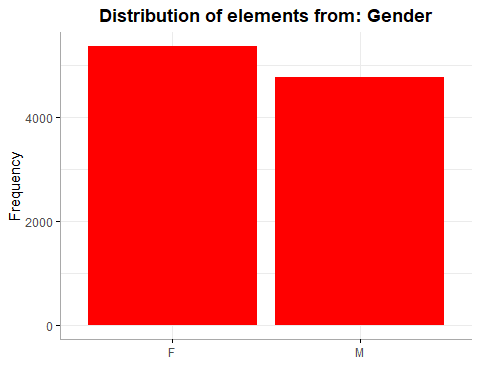
## Entries  
## Existing Customer Attrited Customer   
## 8500 1627   
## Entries  
## Existing Customer Attrited Customer   
## 83.93 16.07

 Osserviamo che dei 10127 clienti, 8500 di essi (~84%) sono utenti esistenti e attivi, mentre i restanti 1627 (~16%) risultano essere utenti inattivi o dal conto bloccato.

### Gender

Univar\_Quality(  
 credit\_card,Gender,  
 label="",  
 title="Gender")

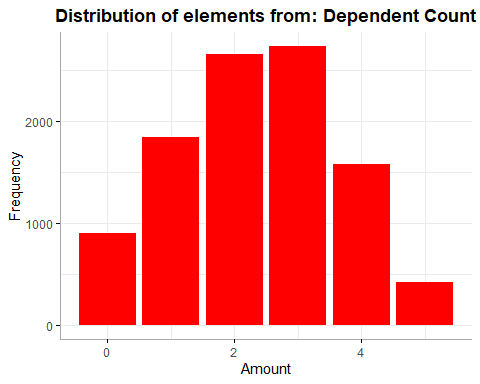
## Entries  
## F M   
## 5358 4769   
## Entries  
## F M   
## 52.91 47.09

 Notiamo che la distribuzione dei sessi dei clienti risulta quasi bilanciata, con un lieve prevalere di quello femminile, dalla frequenza di 5358 (~52%), su quello maschile, di frequenza 4769 (~47%).

### Dependent\_count

Univar\_Quality(  
 credit\_card,Dependent\_count,  
 label="Amount",  
 title="Dependent Count")

## Entries  
## 3 2 1 4 0 5   
## 2732 2655 1838 1574 904 424   
## Entries  
## 3 2 1 4 0 5   
## 26.98 26.22 18.15 15.54 8.93 4.19

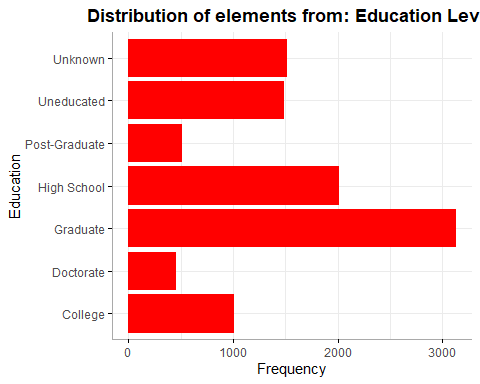


Si può osservare che la maggior parte delle persone hanno tra 1 e 3 dipendenti coprendo il ~=71.35%, con la maggior parte avendo 3 dipendenti a 2732 occorrenze (~=26.98%)

### Education\_Level

Univar\_Quality(  
 credit\_card,Education\_Level,  
 label="Education",  
 title="Education Level")

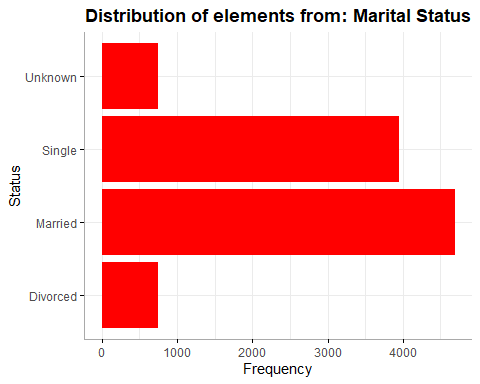
## Entries  
## Graduate High School Unknown Uneducated College   
## 3128 2013 1519 1487 1013   
## Post-Graduate Doctorate   
## 516 451   
## Entries  
## Graduate High School Unknown Uneducated College   
## 30.89 19.88 15.00 14.68 10.00   
## Post-Graduate Doctorate   
## 5.10 4.45

 Notiamo che tra i membri di questa banca risultano esserci una prevalenza di laureati, avendo frequenza 3128, ovvero il ~=31% dei clienti complessivi, seguito da quelli che hanno superato gli studi in una scuola superiore (2013) che sono il ~=20%. Osserviamo che le frequenze di “sconosciuto” e “senza studi” sono paragonabili , avendo entrambi quasi il ~=15% ognuno del complessivo. Sono vicini anche il numero di “post-laureati” e “Dottorandi”, attorno al 5% delle frequenze totali ognuno.

### Marital\_Status

Univar\_Quality(  
 credit\_card,Marital\_Status,  
 label="Status",  
 title="Marital Status")

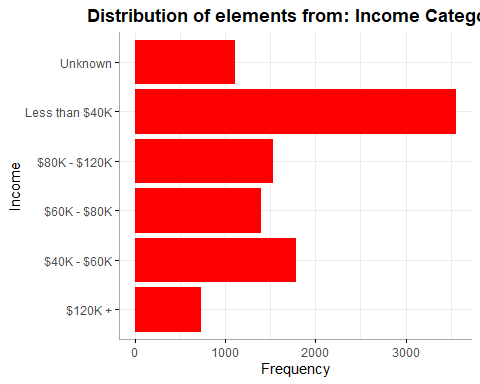
## Entries  
## Married Single Unknown Divorced   
## 4687 3943 749 748   
## Entries  
## Married Single Unknown Divorced   
## 46.28 38.94 7.40 7.39

 La distribuzione di questa variabile è relativamente bilanciata tra “Single” (single+divorziati) e “Sposati” con ~=46% ognuno, quindi accorperei single e divorziati in un unico livello per un miglior bilanciamento. Inoltre abbiamo 749 utenti dei quali non conosciamo lo status, quindi probabilmente non risulteranno utili in un modello predittivo.

### Income\_Category

Univar\_Quality(  
 credit\_card,Income\_Category,  
 label="Income",  
 title="Income Category")

## Entries  
## Less than $40K $40K - $60K $80K - $120K $60K - $80K Unknown   
## 3561 1790 1535 1402 1112   
## $120K +   
## 727   
## Entries  
## Less than $40K $40K - $60K $80K - $120K $60K - $80K Unknown   
## 35.16 17.68 15.16 13.84 10.98   
## $120K +   
## 7.18

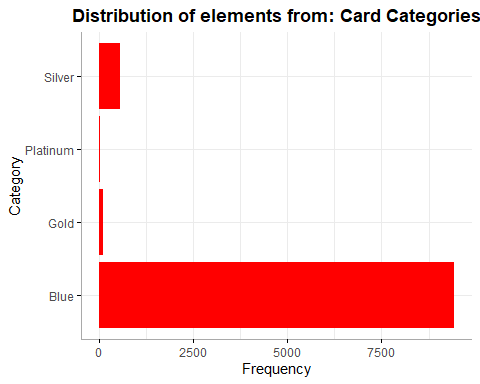


Osservo che il ~=52% degli utenti complessivi appartiene alle fasce inferiori ai $60K, con un maggior numero di utenti sotto i $40K (3561) rispetto a quelli tra $40k e $60k. Tra i restanti, 1112 di essi non sappiamo in che categoria ricadano, quindi è probabilmente un dato trascurabile o da chiarificare; e gli utenti con più di $60K rappresentano il ~=36% degli iscritti complessivi.

### Card\_Category

Univar\_Quality(  
 credit\_card,Card\_Category,  
 label="Category",  
 title="Card Categories")

## Entries  
## Blue Silver Gold Platinum   
## 9436 555 116 20   
## Entries  
## Blue Silver Gold Platinum   
## 93.18 5.48 1.15 0.20



Il ~=93% degli utenti possiede una carta di tipo “Blue”, seguiti dagli utenti “Silver” (~=5%) e “Gold” (~=1%). Gli utenti di categoria “Platinum” rappresentano meno dell’ ~=1%, con una frequenza di 20.

detach(HP)  
detach(credit\_card)

# House Price - Bivariata

## Lettura dati

HP <- read.csv2(  
 "./house\_price.csv",   
 sep = ",",   
 stringsAsFactors = T)

## Pulizia dei dati

HP$MSSubClass <- factor(  
 HP$MSSubClass,  
 levels = c(20, 30, 40, 45, 50, 60, 70, 75, 80, 85, 90, 120, 150, 160, 180, 190),  
 labels = c("1-STORY 1946 & NEWER ALL STYLES", "1-STORY 1945 & OLDER", "1-STORY W/FINISHED ATTIC ALL AGES", "1-1/2 STORY - UNFINISHED ALL AGES", "1-1/2 STORY FINISHED ALL AGES", "2-STORY 1946 & NEWER", "2-STORY 1945 & OLDER", "2-1/2 STORY ALL AGES", "SPLIT OR MULTI-LEVEL", "SPLIT FOYER", "DUPLEX - ALL STYLES AND AGES", "1-STORY PUD - 1946 & NEWER", "1-1/2 STORY PUD - ALL AGES", "2-STORY PUD - 1946 & NEWER", "PUD - MULTILEVEL - INCL SPLIT LEV/FOYER", "2 FAMILY CONVERSION - ALL STYLES AND AGES"))  
  
HP$MSZoning <- factor(  
 HP$MSZoning,  
 levels = c("RL", "RM", "RH", "FV", "C (all)"),  
 labels = c("Residential Low Density", "Residential Medium Density", "Residential High Density", "Floating Village Residential", "Commercial"))  
  
HP$LotShape <- factor(  
 HP$LotShape,  
 levels = c("Reg", "IR1", "IR2", "IR3"),  
 labels = c("Regular", "Slightly irregular", "Moderately irregular", "Irregular"))  
  
HP$LandContour <- factor(  
 HP$LandContour,  
 levels = c("Lvl", "Bnk", "HLS", "Low"),  
 labels = c("Near Flat/Level", "Banked", "Hillside", "Depression"))  
  
HP$Neighborhood <- factor(  
 HP$Neighborhood,  
 levels = c("Blmngtn", "Blueste", "BrDale", "BrkSide", "ClearCr", "CollgCr", "Crawfor", "Edwards", "Gilbert", "IDOTRR", "MeadowV", "Mitchel", "Names", "NoRidge", "NPkVill", "NridgHt", "NWAmes", "OldTown", "SWISU", "Sawyer", "SawyerW", "Somerst", "StoneBr", "Timber", "Veenker"),  
 labels = c("Bloomington Heights", "Bluestem", "Briardale", "Brookside", "Clear Creek", "College Creek", "Crawford", "Edwards", "Gilbert", "Iowa DOT and Rail Road", "Meadow Village", "Mitchell", "North Ames", "Northridge", "Northpark Villa", "Northridge Heights", "Northwest Ames", "Old Town", "South & West of Iowa State University", "Sawyer", "Sawyer West", "Somerset", "Stone Brook", "Timberland", "Veenker"))  
  
HP$BldgType <- factor(  
 HP$BldgType,  
 levels = c("1Fam", "2fmCon", "Duplex", "TwnhsE", "Twnhs"),  
 labels = c("Single-family Detached", "Two-family Conversion", "Duplex", "Townhouse End Unit", "Townhouse Inside Unit"))  
  
HP$MasVnrArea[HP$MasVnrArea == 0 | is.na(HP$MasVnrType)] <- NA  
  
for (f in c("ExterQual", "ExterCond", "BsmtQual", "BsmtCond", "HeatingQC", "KitchenQual", "FireplaceQu", "GarageQual", "GarageCond")) {  
 HP[, f] <- factor(  
 HP[, f],  
 levels = c("Po", "Fa", "TA", "Gd", "Ex"),  
 labels = c("Poor", "Fair", "Average", "Good", "Excellent"))  
}  
  
levels(HP$CentralAir) <- c("No", "Yes")  
  
HP$BsmtExposure <- factor(  
 HP$BsmtExposure,  
 levels = c("No", "Mn", "Av", "Gd"),  
 labels = c("No", "Minimu", "Average", "Good"))  
  
HP$GarageFinish <- factor(  
 HP$GarageFinish,  
 levels = c("Unf", "RFn", "Fin"),  
 labels = c("Unfinished", "Rough finished", "Finished"))  
  
HP$PavedDrive <- factor(  
 HP$PavedDrive,  
 levels = c("N", "P", "Y"),  
 labels = c("Dirt/Graver", "Partial paved", "Paved"))  
  
HP$Fence <- factor(  
 HP$Fence,  
 levels = c("MnWw", "GdWo", "MnPrv", "GdPrv"),  
 labels = c("Minimum Wood/Wire", "Good Wood", "Minimum Privacy", "Good Privacy"))  
  
HP$SaleType <- factor(  
 HP$SaleType,  
 levels = c("WD", "CWD", "VWD", "New", "COD", "Con", "ConLw", "ConLI", "ConLD", "Oth"),  
 labels = c("Warranty Deed - Conventional", "Warranty Deed - Cash", "Warranty Deed - VA Loan", "Home just constructed and sold", "Court Officer", "Contract 15% Down payment", "Contract Low Down Payment", "Contract Low Interest", "Contract Low Down", "Other"))  
  
HP$SaleCondition <- factor(  
 HP$SaleCondition,  
 levels = c("Normal", "Abnorml", "AdjLand", "Alloca", "Family", "Partial"),  
 labels = c("Normal Sale", "Abnormal Sale", "Adjoining Land Purchase", "Allocation", "Sale between family members", "Home not completed"))  
  
attach(HP)

## Funzioni di analisi

Seconda variabile qualitativa:

analysis.qualitative <- function(data, horizontal = F, lab = NULL, title = NULL) {  
 if (missing(data) | !is.factor(data)) {  
 stop("Missing data or data is not a factor.")  
 }  
 if (!is.logical(horizontal)) {  
 stop("Horizontal must be boolean")  
 }  
   
 df <- subset(data.frame(SalePrice, data), !is.na(data))  
   
 test <- aov(SalePrice ~ data)  
 p <- summary(test)[[1]][["Pr(>F)"]][1]  
   
 plt <- ggplot(  
 df,  
 aes(  
 x = if (horizontal) SalePrice else data,   
 y = if (horizontal) data else SalePrice)) +  
 geom\_violin(  
 trim = F,   
 draw\_quantiles = c(0.25, 0.5, 0.75)) +  
 stat\_summary(  
 fun = "mean",  
 geom = "point",  
 color = "red",  
 fill = "red",  
 shape = 20,  
 size = 2) +  
 labs(  
 title = if (is.null(title)) NULL else paste("Price by", title, sep = " "),  
 subtitle = paste("Anova p-value = ", signif(p, 3), sep = ""),  
 x = if (horizontal) "price" else lab,  
 y = if (horizontal) lab else "price")  
   
 plot(plt)  
}

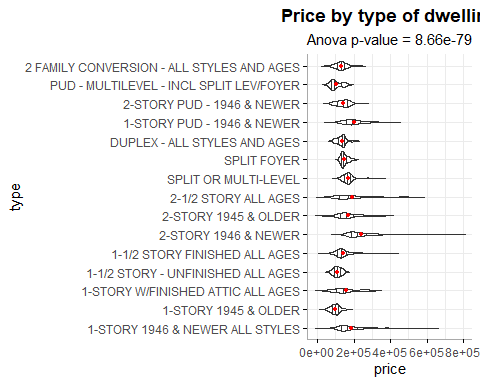
Seconda variabile quantitativa:

analysis.quantitative <- function(data, horizontal = F, lab = NULL, title = NULL) {  
 if (missing(data)) {  
 stop("Missing data.")  
 }  
 if (!is.logical(horizontal)) {  
 stop("Horizontal must be boolean")  
 }  
   
 sp <- log(SalePrice)  
 dt <- log(data)  
   
 mod <- lm(sp ~ dt)  
 p <- summary(mod)$coefficients[2,4]  
   
 plt <- ggplot(  
 data.frame(sp, dt),  
 aes(  
 x = if (horizontal) sp else dt,   
 y = if (horizontal) dt else sp)) +  
 geom\_point(size = 1) +  
 geom\_abline(  
 intercept = coef(mod)[1],   
 slope = coef(mod)[2],  
 color = "red") +  
 labs(  
 title = if (is.null(title)) NULL else paste("Price by", title, sep = " "),  
 subtitle = paste("Linear model p-value = ", signif(p, 3), sep = ""),  
 x = if (horizontal) "log price" else paste("log", lab, sep = " "),  
 y = if (horizontal) paste("log", lab, sep = " ") else "log price")  
   
 plot(plt)  
}

## Analisi

### MSSubClass

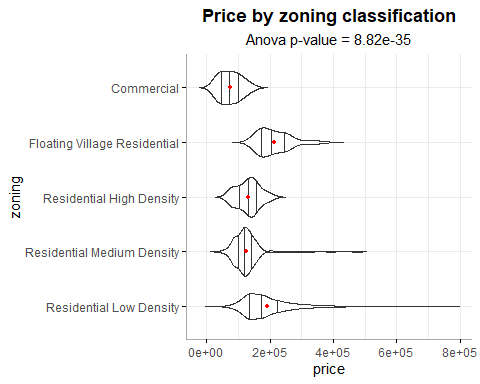
analysis.qualitative(  
 MSSubClass,  
 horizontal = T,  
 lab = "type",  
 title = "type of dwelling")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di ^{-79} non c’è supporto a questa ipotesi. Quindi c’è una correlazione tra il prezzo di vendità e il tipo della casa.

### Zoning

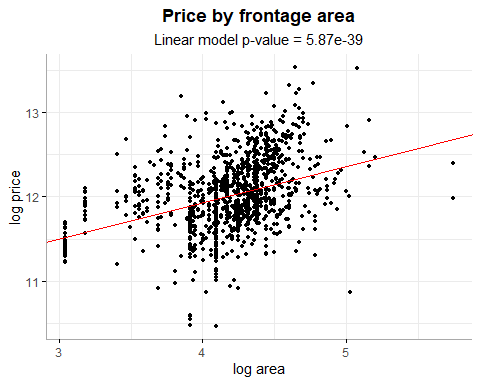
analysis.qualitative(  
 MSZoning,  
 horizontal = T,  
 lab = "zoning",  
 title = "zoning classification")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-79} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il tipo della casa.

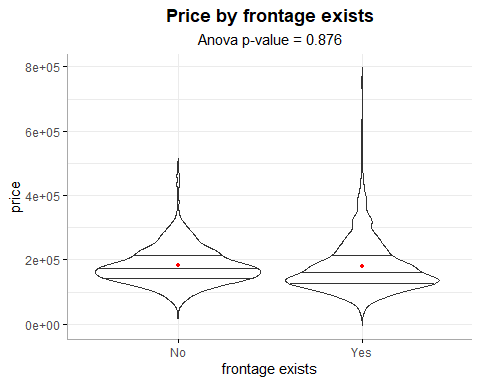
### LotFrontage

analysis.quantitative(  
 LotFrontage,  
 lab = "area",  
 title = "frontage area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-39} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la distanza dalla strada, che è approssiabile da un modello lineare con pendenza positiva.

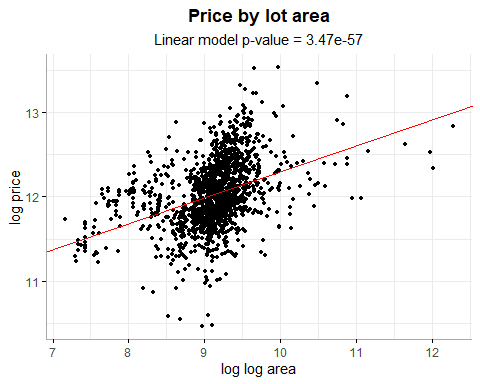
LotFrontage.exists <- factor(  
 !is.na(LotFrontage),   
 levels = c(F, T),   
 labels = c("No", "Yes"))  
  
analysis.qualitative(  
 LotFrontage.exists,  
 lab = "frontage exists",  
 title = "frontage exists")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ $ c’è supporto all’ipotesi nulla. Quindi non c’è una correlazione tra il prezzo di vendità e l’esistenza di spazio tra casa e strada.

### LotArea

analysis.quantitative(  
 LotArea,  
 lab = "log area",  
 title = "lot area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-59} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la dimensione del lotto, che è approssiabile da un modello lineare con pendenza positiva.

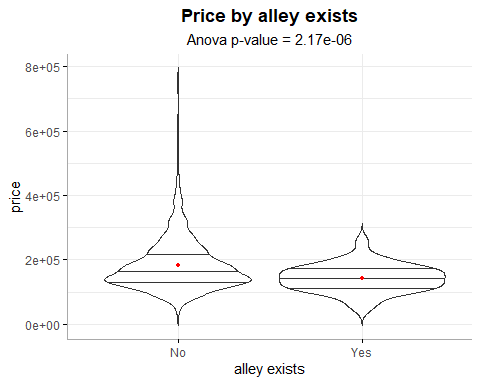
### Street

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## Street  
## Pave Grvl   
## 99.59 0.41

### Alley

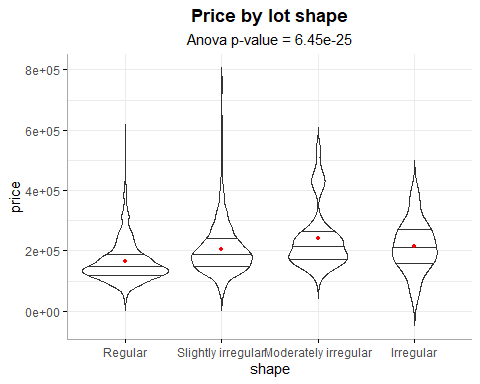
Alley.exists <- factor(  
 !is.na(Alley),   
 levels = c(F, T),   
 labels = c("No", "Yes"))  
  
analysis.qualitative(  
 Alley.exists,  
 lab = "alley exists",  
 title = "alley exists")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-6} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’esistenza del vialetto.

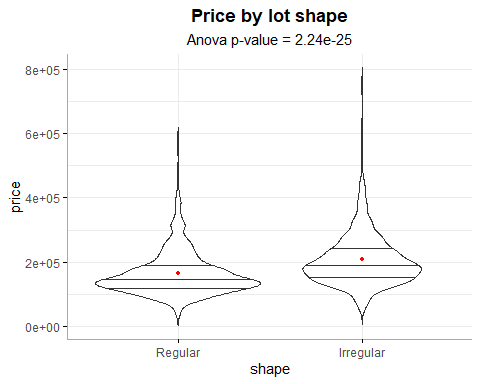
### LotShape

analysis.qualitative(  
 LotShape,  
 lab = "shape",  
 title = "lot shape")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-25} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la forma del lotto.

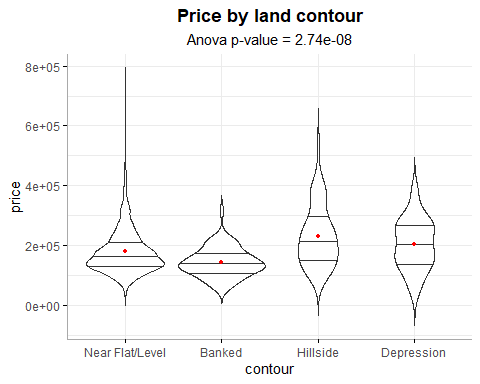
LotShape.cut <- factor(  
 LotShape == "Regular",   
 levels = c(T, F),   
 labels = c("Regular", "Irregular"))  
  
analysis.qualitative(  
 LotShape.cut,  
 lab = "shape",  
 title = "lot shape")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-25} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la forma regolare o irregolare del lotto.

### LandContour

analysis.qualitative(  
 LandContour,  
 lab = "contour",  
 title = "land contour")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-8} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la zona in cui è situata l’abitazione.

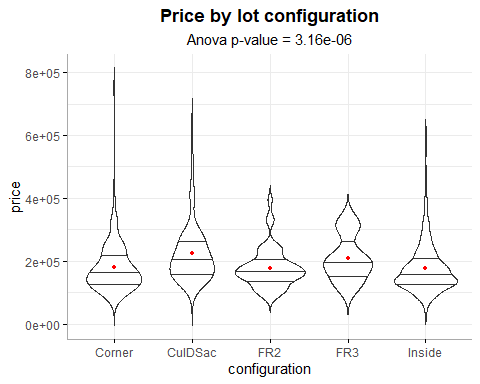
### Utilities

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## Utilities  
## AllPub NoSeWa   
## 99.93 0.07

### LotConfig

analysis.qualitative(  
 LotConfig,  
 lab = "configuration",  
 title = "lot configuration")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-6} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la configurazione del lotto.

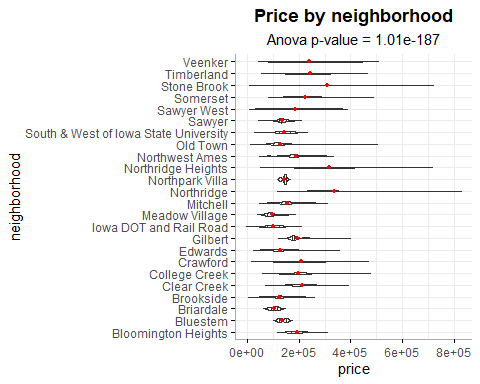
### LandSlope

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## LandSlope  
## Gtl Mod Sev   
## 94.66 4.45 0.89

### Neighborhood

analysis.qualitative(  
 Neighborhood,  
 horizontal = T,  
 lab = "neighborhood",  
 title = "neighborhood")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-187} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il vicinato della casa.

### Condition1

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## Condition1  
## Norm Feedr Artery RRAn PosN RRAe PosA RRNn RRNe   
## 86.30 5.55 3.29 1.78 1.30 0.75 0.55 0.34 0.14

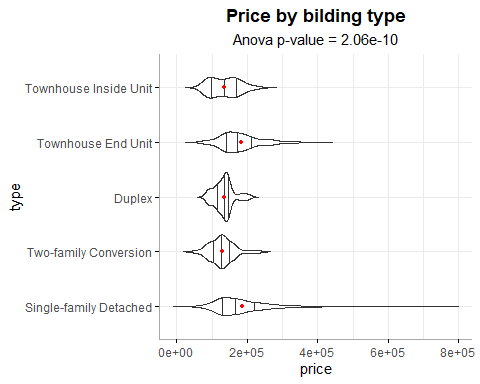
### Condition2

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## Condition2  
## Norm Feedr Artery PosN RRNn PosA RRAe RRAn   
## 98.97 0.41 0.14 0.14 0.14 0.07 0.07 0.07

### BldgType

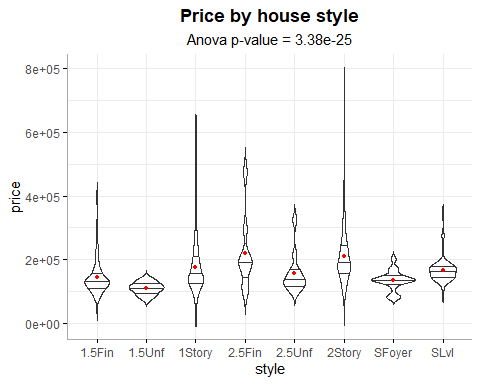
analysis.qualitative(  
 BldgType,  
 horizontal = T,  
 lab = "type",  
 title = "bilding type")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-10} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il tipo di abitazione.

### HouseStyle

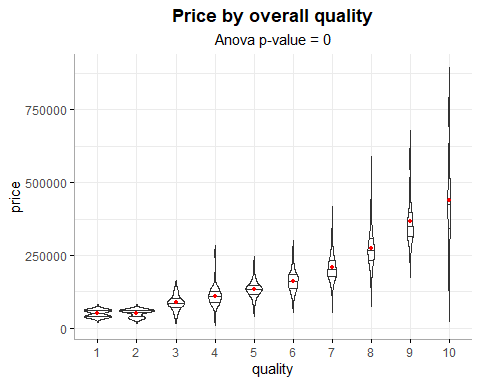
analysis.qualitative(  
 HouseStyle,  
 lab = "style",  
 title = "house style")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-25} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e lo stile della casa.

### OverallQual

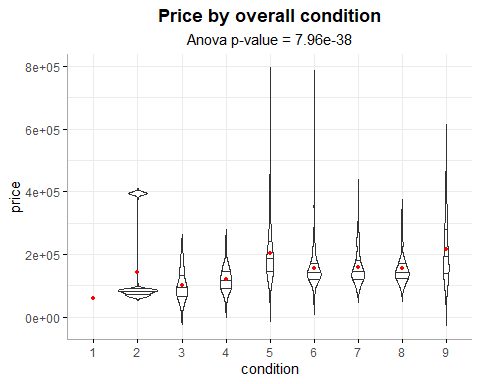
OverallQual.fact <- factor(  
 OverallQual,  
 levels = 1:10)  
  
analysis.qualitative(  
 OverallQual.fact,  
 lab = "quality",  
 title = "overall quality")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la qualità dei materiali.

### OverallCond

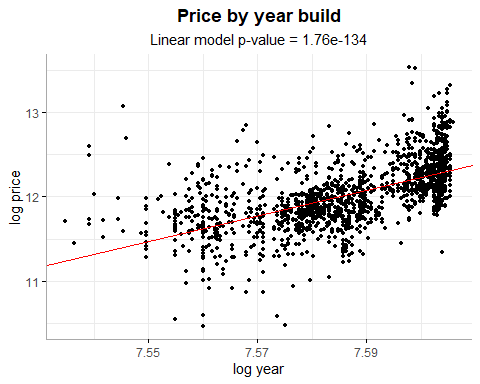
OverallCond.fact <- factor(  
 OverallCond,  
 levels = 1:10)  
  
analysis.qualitative(  
 OverallCond.fact,  
 lab = "condition",  
 title = "overall condition")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-38} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e le condizioni attuali dei materiali.

### YearBuilt

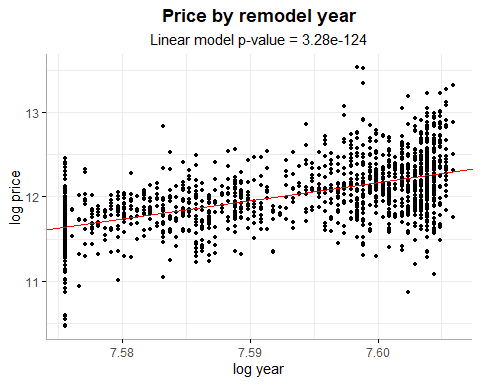
analysis.quantitative(  
 YearBuilt,  
 lab = "year",  
 title = "year build")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-134} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’anno di costruzione, che è approssiabile da un modello lineare con pendenza positiva.

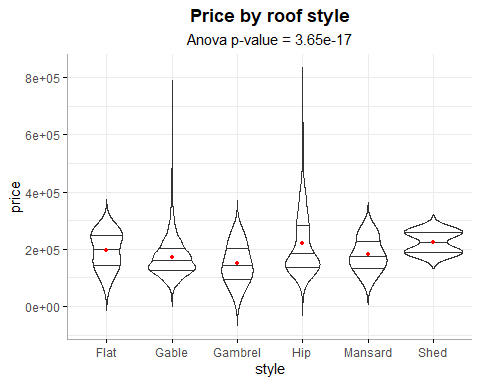
### YearRemodAdd

analysis.quantitative(  
 YearRemodAdd,  
 lab = "year",  
 title = "remodel year")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-124} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’anno di ristrutturazione, che è approssiabile da un modello lineare con pendenza positiva. ### RoofStyle

analysis.qualitative(  
 RoofStyle,  
 lab = "style",  
 title = "roof style")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-17} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e lo stile del tetto.

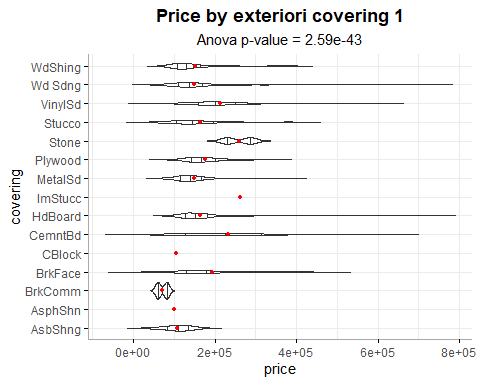
### RoofMatl

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## RoofMatl  
## CompShg Tar&Grv WdShngl WdShake ClyTile Membran Metal Roll   
## 98.22 0.75 0.41 0.34 0.07 0.07 0.07 0.07

### Exterior1st

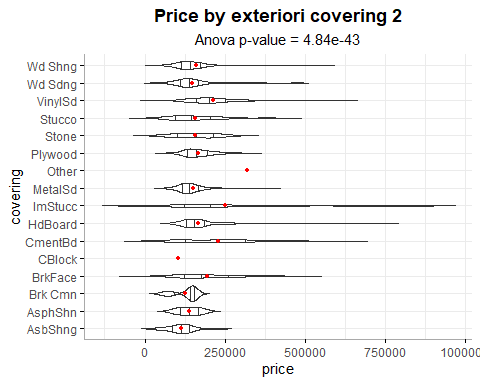
analysis.qualitative(  
 Exterior1st,  
 horizontal = T,  
 lab = "covering",  
 title = "exteriori covering 1")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-43} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il rivestimento esterno.

### Exterior2nd

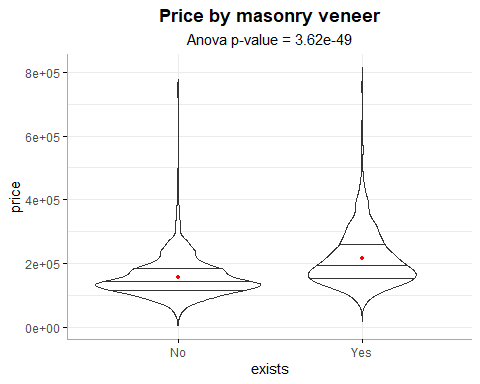
analysis.qualitative(  
 Exterior2nd,  
 horizontal = T,  
 lab = "covering",  
 title = "exteriori covering 2")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-43} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il rivestimento esterno.

### MasVnrType

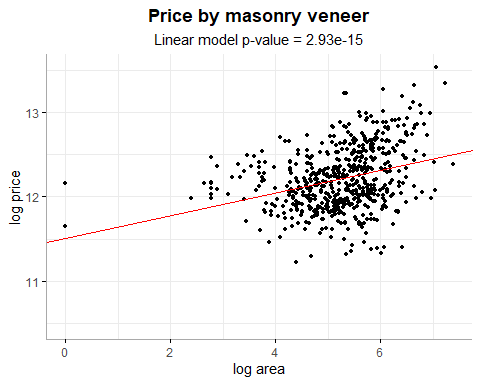
MasVnrType.exists <- factor(  
 MasVnrType != "None",  
 levels = c(F, T),   
 labels = c("No", "Yes"))  
  
analysis.qualitative(  
 MasVnrType.exists,  
 lab = "exists",  
 title = "masonry veneer")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-49} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la presenza dell’impiallacciatura in muratura.

### MasVnrArea

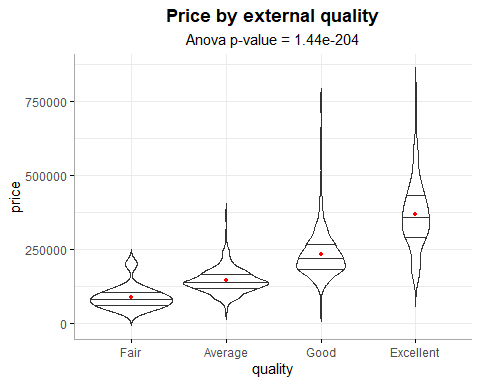
analysis.quantitative(  
 MasVnrArea,  
 lab = "area",  
 title = "masonry veneer")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-15} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’area di impiallacciatura in muratura, che è approssiabile da un modello lineare con pendenza positiva.

### ExterQual

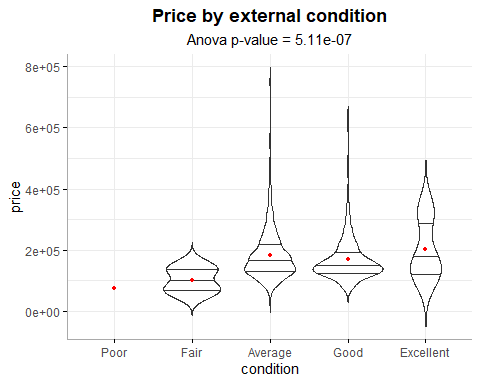
analysis.qualitative(  
 ExterQual,  
 lab = "quality",  
 title = "external quality")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-204} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la qualità degli esterni.

### ExterCond

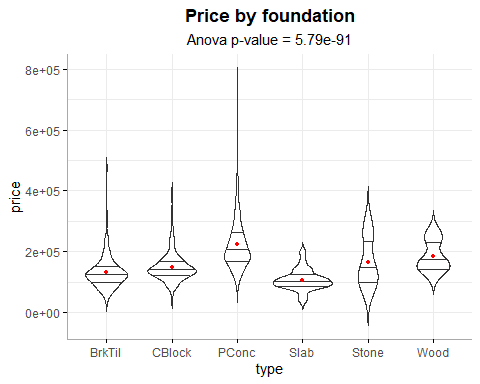
analysis.qualitative(  
 ExterCond,  
 lab = "condition",  
 title = "external condition")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-7} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e le attuali condizioni degli esterni.

### Foundation

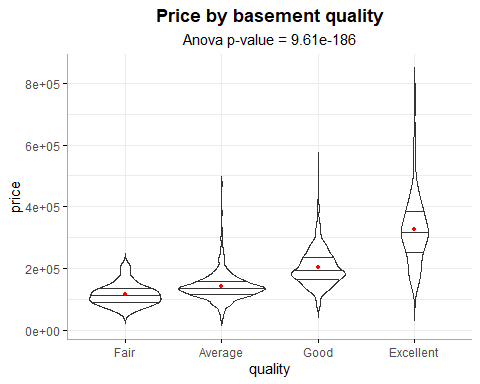
analysis.qualitative(  
 Foundation,  
 lab = "type",  
 title = "foundation")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-91} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il tipo di fondamenta.

### BsmtQual

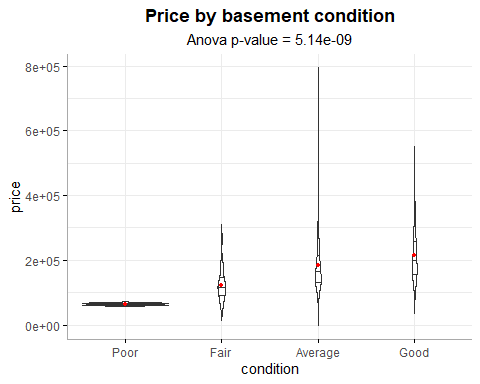
analysis.qualitative(  
 BsmtQual,  
 lab = "quality",  
 title = "basement quality")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-186} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la qualità del seminterrato.

### BsmtCond

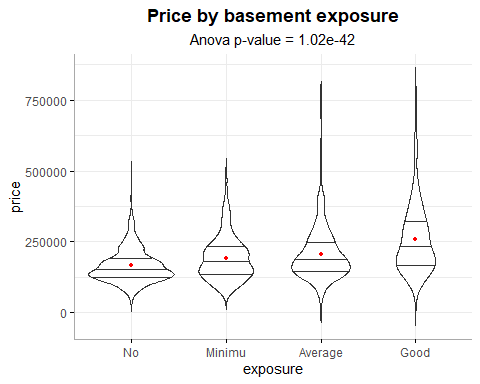
analysis.qualitative(  
 BsmtCond,  
 lab = "condition",  
 title = "basement condition")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-9} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e le condizioni attuali del meninterrato.

### BsmtExposure

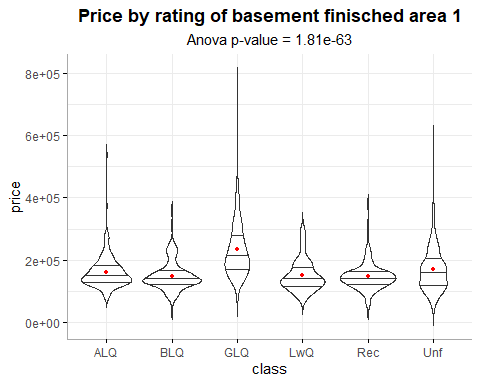
analysis.qualitative(  
 BsmtExposure,  
 lab = "exposure",  
 title = "basement exposure")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-6} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’esposizione del seminterrato.

### BsmtFinType1

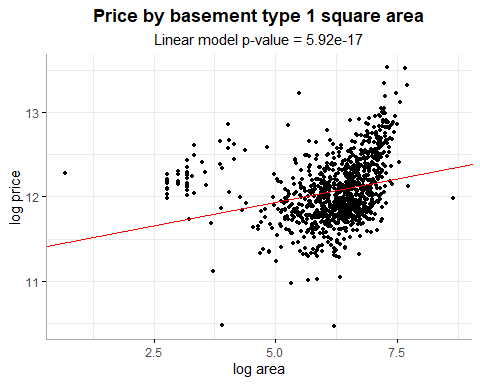
analysis.qualitative(  
 BsmtFinType1,  
 lab = "class",  
 title = "rating of basement finisched area 1")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-63} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’area finita di seminterrato.

### BsmtFinSF1

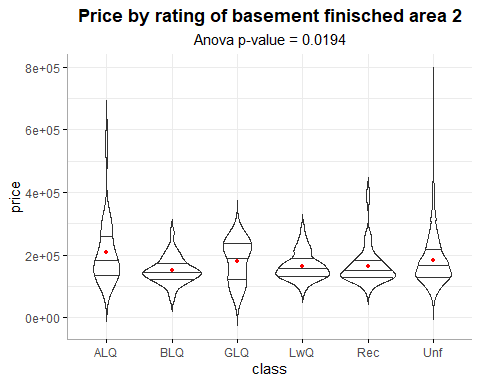
BsmtFinSF1.positive <- BsmtFinSF1  
BsmtFinSF1.positive[BsmtFinSF1 == 0] <- NA  
  
analysis.quantitative(  
 BsmtFinSF1.positive,  
 lab = "area",  
 title = "basement type 1 square area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-17} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’area del seminterrato, che è approssiabile da un modello lineare con pendenza positiva.

### BsmtFinType2

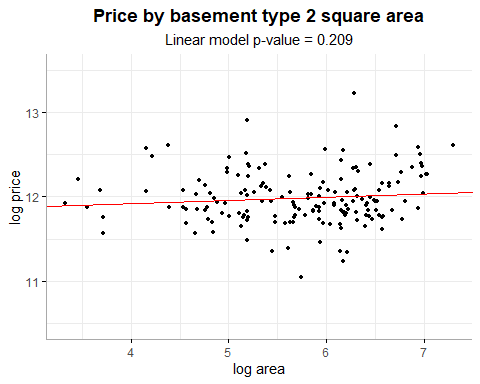
analysis.qualitative(  
 BsmtFinType2,  
 lab = "class",  
 title = "rating of basement finisched area 2")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ $ c’è un debole supporto all’ipotesi nulla. Quindi c’è una debole correlazione tra il prezzo di vendità e la seconda area finita del seminterrato.

### BsmtFinSF2

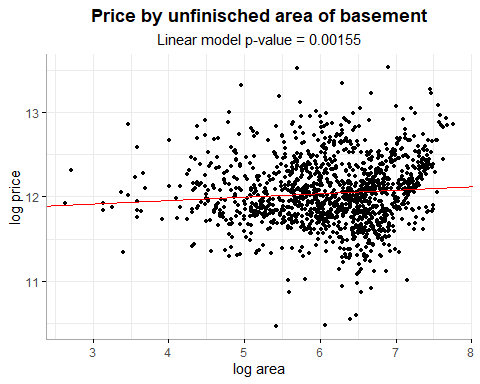
BsmtFinSF2.positive <- BsmtFinSF2  
BsmtFinSF2.positive[BsmtFinSF2 == 0] <- NA  
  
analysis.quantitative(  
 BsmtFinSF2.positive,  
 lab = "area",  
 title = "basement type 2 square area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ $ c’è supporto all’ipotesi nulla. Quindi non c’è una correlazione tra il prezzo di vendità e l’area del secondo seminterrato. Questo probabilmente è dovuto alla scaristà nel numero dei dati.

### BsmtUnfSF

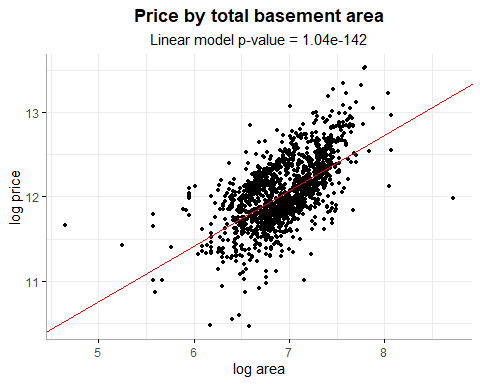
BsmtUnfSF.positive <- BsmtUnfSF  
BsmtUnfSF.positive[BsmtUnfSF == 0] <- NA  
  
analysis.quantitative(  
 BsmtUnfSF.positive,  
 lab = "area",  
 title = "unfinisched area of basement")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ $ c’è un debole supporto all’ipotesi nulla. Quindi c’è una correlazione abbastanza forte tra il prezzo di vendità e l’area non finita di seminterrato, che è approssiabile da un modello lineare con pendenza positiva.

### TotalBsmtSF

TotalBsmtSF.positive <- TotalBsmtSF  
TotalBsmtSF.positive[TotalBsmtSF == 0] <- NA  
  
analysis.quantitative(  
 TotalBsmtSF.positive,  
 lab = "area",  
 title = "total basement area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-142} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’area del seminterrato, che è approssiabile da un modello lineare con pendenza positiva.

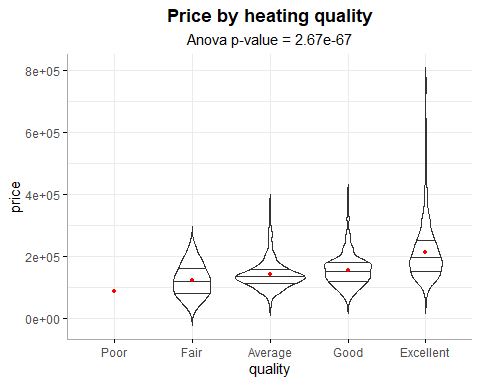
### Heating

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## Heating  
## GasA GasW Grav Wall OthW Floor   
## 97.81 1.23 0.48 0.27 0.14 0.07

### HeatingQC

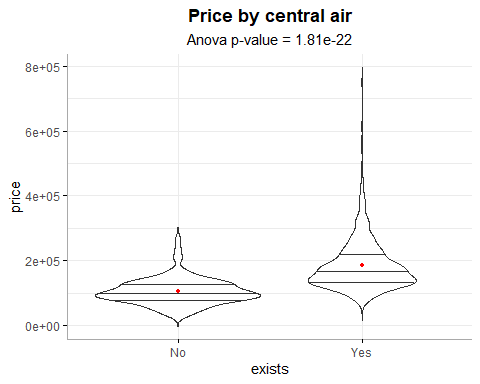
analysis.qualitative(  
 HeatingQC,  
 lab = "quality",  
 title = "heating quality")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-67} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la qualità del riscaldamento.

### CentralAir

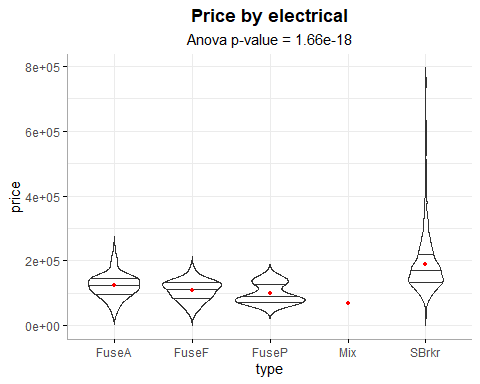
analysis.qualitative(  
 CentralAir,  
 lab = "exists",  
 title = "central air")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-22} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’esistenza dell’aria condizionata centralizzata.

### Electrical

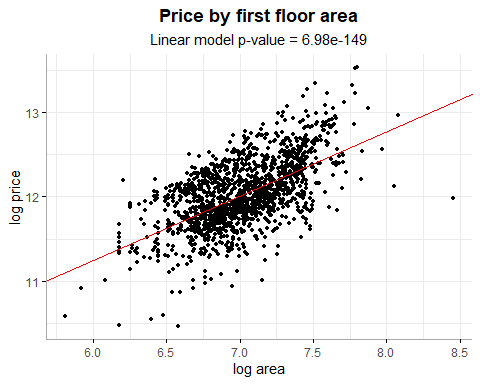
analysis.qualitative(  
 Electrical,  
 lab = "type",  
 title = "electrical")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-18} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il tipo di impianto elettrico.

### X1stFlrSF

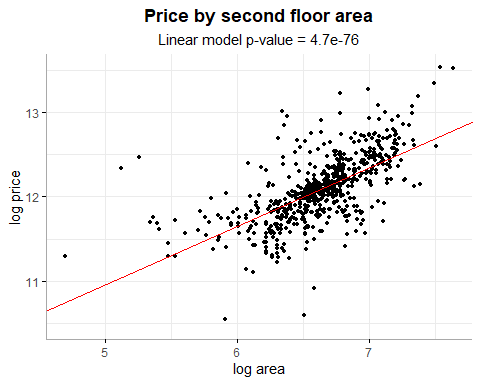
analysis.quantitative(  
 X1stFlrSF,  
 lab = "area",  
 title = "first floor area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-149} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’area calpestabile del piano terra, che è approssiabile da un modello lineare con pendenza positiva.

### X2ndFlrSF

X2ndFlrSF.positive <- X2ndFlrSF  
X2ndFlrSF.positive[X2ndFlrSF == 0] <- NA  
  
analysis.quantitative(  
 X2ndFlrSF.positive,  
 lab = "area",  
 title = "second floor area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-76} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’area calpestabile del secondo piano, che è approssiabile da un modello lineare con pendenza positiva.

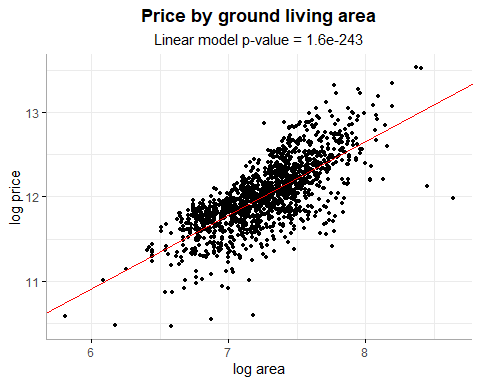
### LowQualFinSF

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## LowQualFinSF  
## 0 80 360 53 120 144 156 205 232 234 371 384 390   
## 98.22 0.21 0.14 0.07 0.07 0.07 0.07 0.07 0.07 0.07 0.07 0.07 0.07   
## 392 397 420 473 479 481 513 514 515 528 572   
## 0.07 0.07 0.07 0.07 0.07 0.07 0.07 0.07 0.07 0.07 0.07

### GrLivArea

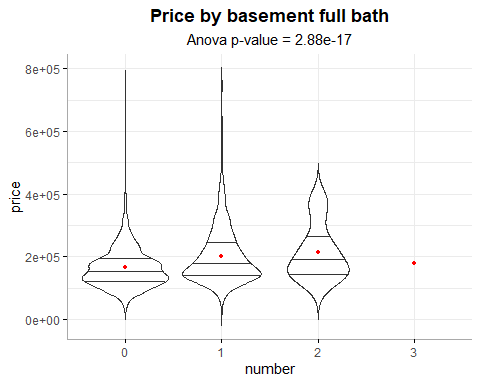
analysis.quantitative(  
 GrLivArea,  
 lab = "area",  
 title = "ground living area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-243} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’area calpestabile sopra il piano strada, che è approssiabile da un modello lineare con pendenza positiva.

### BsmtFullBath

BsmtFullBath.fact <- factor(  
 BsmtFullBath,  
 levels = 0:max(BsmtFullBath))  
  
analysis.qualitative(  
 BsmtFullBath.fact,  
 lab = "number",  
 title = "basement full bath")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-17} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il numero di bagni completi.

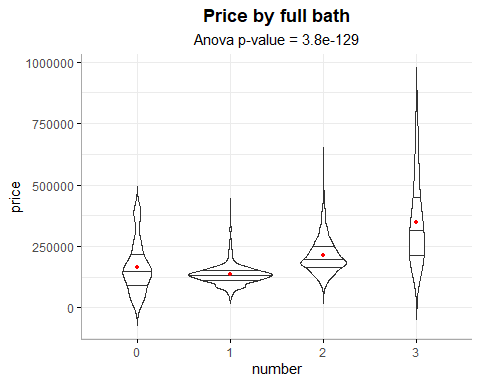
### BsmtHalfBath

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## BsmtHalfBath  
## 0 1 2   
## 94.38 5.48 0.14

### FullBath

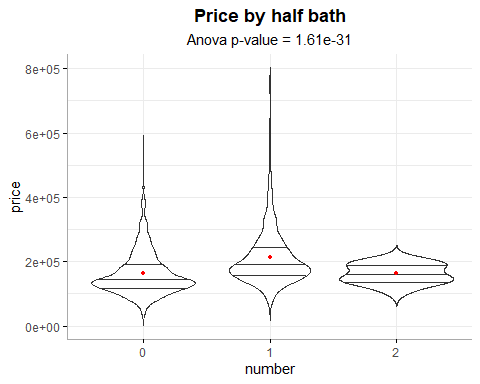
FullBath.fact <- factor(  
 FullBath,  
 levels = 0:max(FullBath))  
  
analysis.qualitative(  
 FullBath.fact,  
 lab = "number",  
 title = "full bath")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-129} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il numero di bagni completi.

### HalfBath

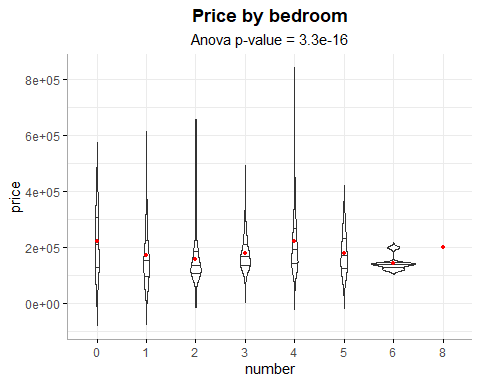
HalfBath.fact <- factor(  
 HalfBath,  
 levels = 0:max(HalfBath))  
  
analysis.qualitative(  
 HalfBath.fact,  
 lab = "number",  
 title = "half bath")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-31} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il numero di bagni di servizio.

### BedroomAbvGr

BedroomAbvGr.fact <- factor(  
 BedroomAbvGr,  
 levels = 0:max(BedroomAbvGr))  
  
analysis.qualitative(  
 BedroomAbvGr.fact,  
 lab = "number",  
 title = "bedroom")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-16} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il numero di camere da letto.

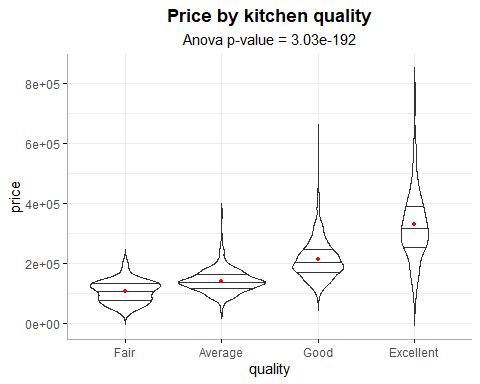
### KitchenAbvGr

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## KitchenAbvGr  
## 1 2 3 0   
## 95.34 4.45 0.14 0.07

### KitchenQual

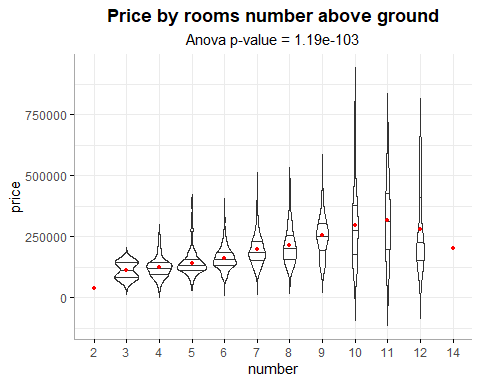
analysis.qualitative(  
 KitchenQual,  
 lab = "quality",  
 title = "kitchen quality")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-192} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la qualità della cucina.

### TotRmsAbvGrd

TotRmsAbvGrd.fact <- factor(  
 TotRmsAbvGrd,  
 levels = 0:max(TotRmsAbvGrd))  
  
analysis.qualitative(  
 TotRmsAbvGrd.fact,  
 lab = "number",  
 title = "rooms number above ground")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-103} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il numero di stanze sopra il livello del terreno.

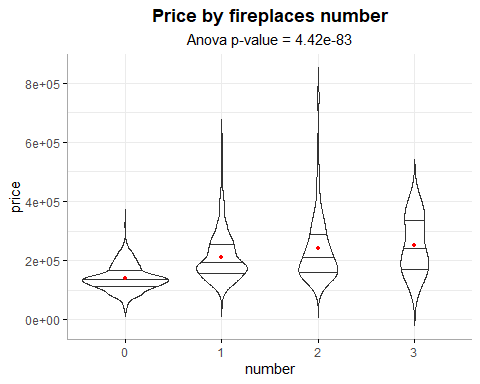
### Functional

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## Functional  
## Typ Min2 Min1 Mod Maj1 Maj2 Sev   
## 93.15 2.33 2.12 1.03 0.96 0.34 0.07

### Fireplaces

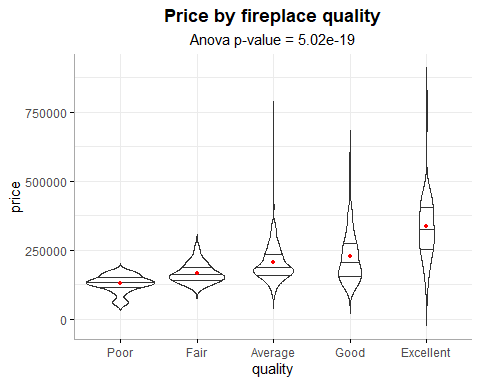
Fireplaces.fact <- factor(  
 Fireplaces,  
 levels = 0:max(Fireplaces))  
  
analysis.qualitative(  
 Fireplaces.fact,  
 lab = "number",  
 title = "fireplaces number")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-83} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il numero di caminetti.

### FireplaceQu

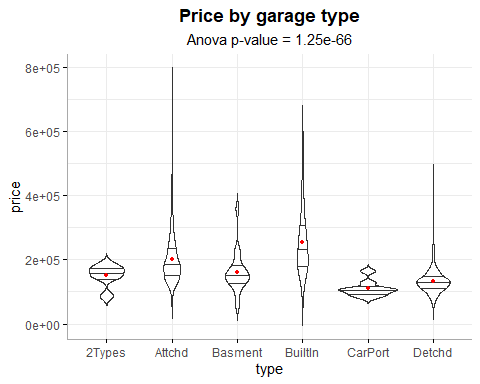
analysis.qualitative(  
 FireplaceQu,  
 lab = "quality",  
 title = "fireplace quality")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-19} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la qualità dei caminetti.

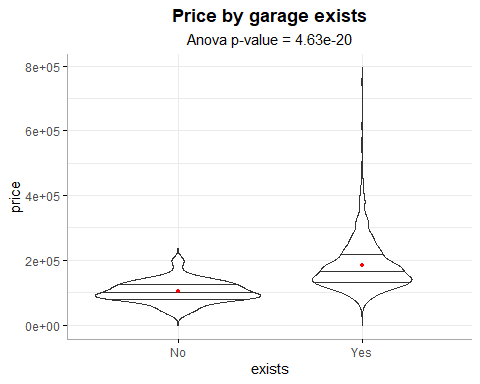
### GarageType

analysis.qualitative(  
 GarageType,  
 lab = "type",  
 title = "garage type")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-66} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il tipo di garage.

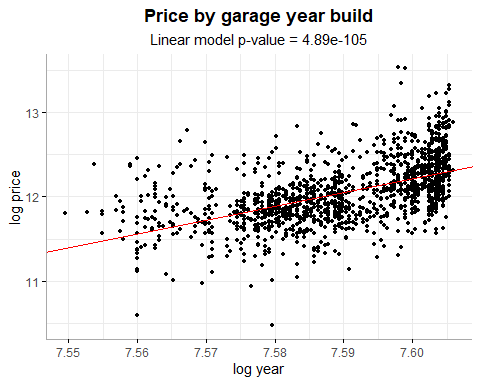
Garage.exists <- factor(  
 !is.na(GarageType),  
 levels = c(F, T),   
 labels = c("No", "Yes"))  
  
analysis.qualitative(  
 Garage.exists,  
 lab = "exists",  
 title = "garage exists")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-20} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’esistenza del garage.

### GarageYrBlt

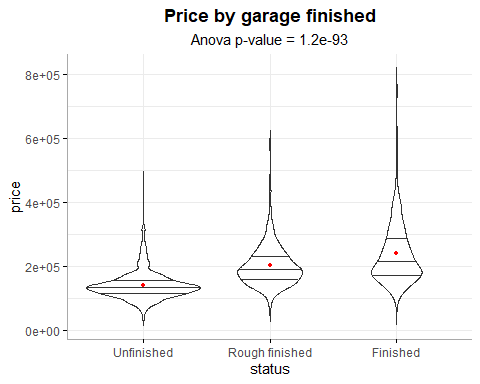
analysis.quantitative(  
 GarageYrBlt,  
 lab = "year",  
 title = "garage year build")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-105} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’anno di costruzione del garage, che è approssiabile da un modello lineare con pendenza positiva.

### GarageFinish

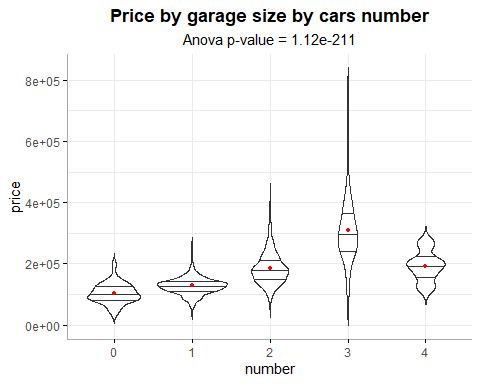
analysis.qualitative(  
 GarageFinish,  
 lab = "status",  
 title = "garage finished")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-93} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e lo stato di costruzione del garage.

### GarageCars

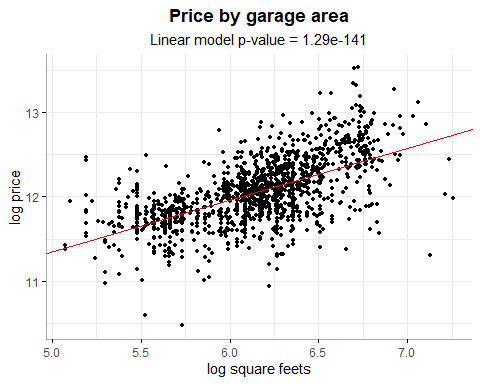
GarageCars.fact <- factor(  
 GarageCars,  
 levels = 0:max(GarageCars))  
  
analysis.qualitative(  
 GarageCars.fact,  
 lab = "number",  
 title = "garage size by cars number")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-6} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la dimensione del garage misurata in numero di auto contenute.

### GarageArea

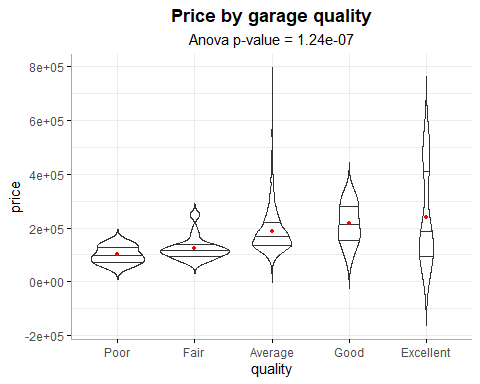
GarageArea.positive <- GarageArea  
GarageArea.positive[GarageArea == 0] <- NA  
  
analysis.quantitative(  
 GarageArea.positive,  
 lab = "square feets",  
 title = "garage area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-141} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’area del garage, che è approssiabile da un modello lineare con pendenza positiva.

### GarageQual

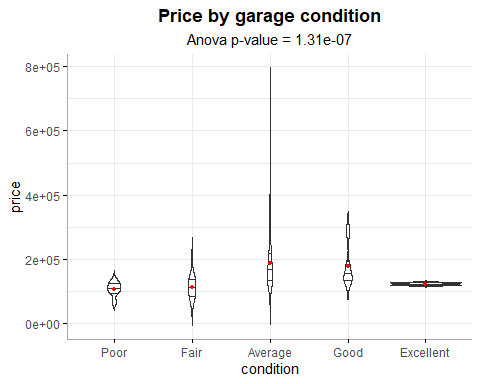
analysis.qualitative(  
 GarageQual,  
 lab = "quality",  
 title = "garage quality")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-7} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la qualità del garage.

### GarageCond

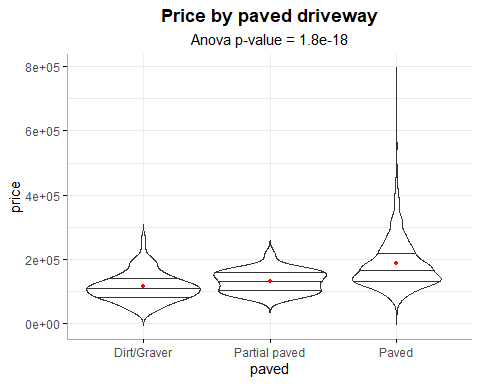
analysis.qualitative(  
 GarageCond,  
 lab = "condition",  
 title = "garage condition")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-7} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e le condizioni del garage.

### PavedDrive

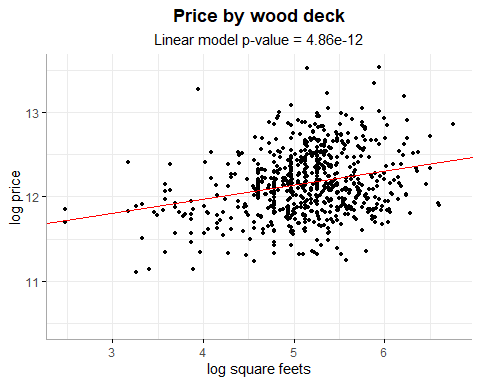
analysis.qualitative(  
 PavedDrive,  
 lab = "paved",  
 title = "paved driveway")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-18} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la pavimentazione del vialetto.

### WoodDeckSF

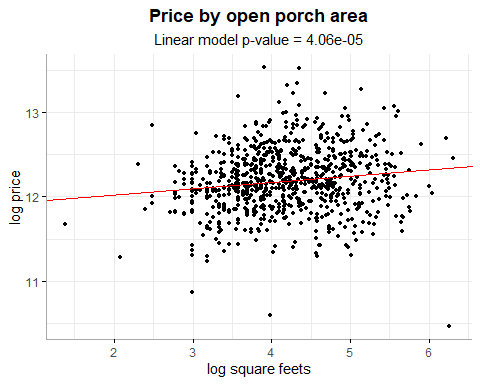
WoodDeckSF.positive <- WoodDeckSF  
WoodDeckSF.positive[WoodDeckSF == 0] <- NA  
  
analysis.quantitative(  
 WoodDeckSF.positive,  
 lab = "square feets",  
 title = "wood deck")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-12} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e la dimensione del portico in legno, che è approssiabile da un modello lineare con pendenza positiva.

### OpenPorchSF

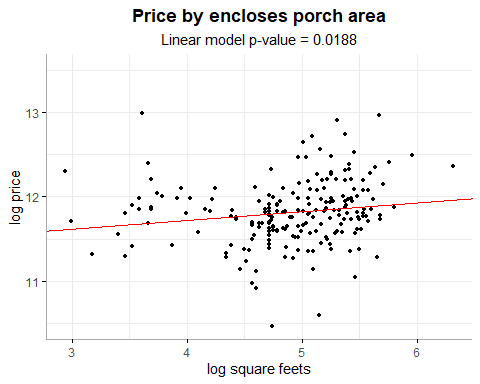
OpenPorchSF.positive <- OpenPorchSF  
OpenPorchSF.positive[OpenPorchSF == 0] <- NA  
  
analysis.quantitative(  
 OpenPorchSF.positive,  
 lab = "square feets",  
 title = "open porch area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ ^{-5} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’area del porticato, che è approssiabile da un modello lineare con pendenza positiva.

### EnclosedPorch

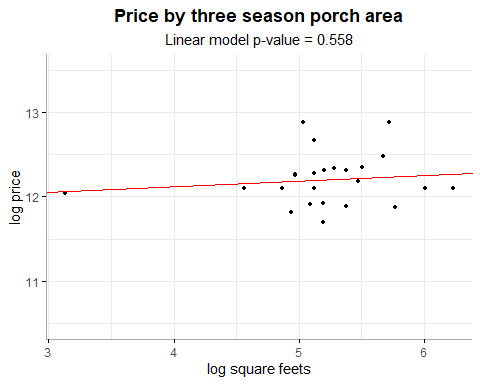
EnclosedPorch.positive <- EnclosedPorch  
EnclosedPorch.positive[EnclosedPorch == 0] <- NA  
  
analysis.quantitative(  
 EnclosedPorch.positive,  
 lab = "square feets",  
 title = "encloses porch area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ $ c’è un debole supporto all’ipotesi nulla. Quindi non c’è una forte correlazione tra il prezzo di vendità e l’area chiusa del portico.

### X3SsnPorch

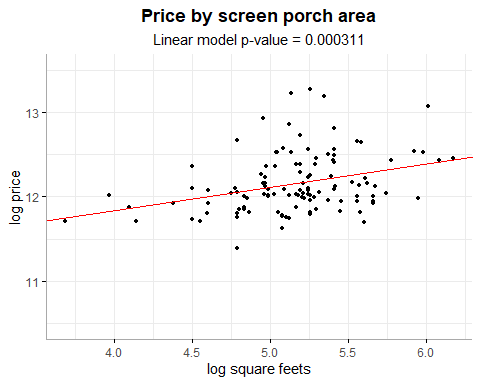
X3SsnPorch.positive <- X3SsnPorch  
X3SsnPorch.positive[X3SsnPorch == 0] <- NA  
  
analysis.quantitative(  
 X3SsnPorch.positive,  
 lab = "square feets",  
 title = "three season porch area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ $ c’è supporto all’ipotesi nulla. Quindi non c’è una correlazione tra il prezzo di vendità e l’area del portico tre stagioni.

### ScreenPorch

ScreenPorch.positive <- ScreenPorch  
ScreenPorch.positive[ScreenPorch == 0] <- NA  
  
analysis.quantitative(  
 ScreenPorch.positive,  
 lab = "square feets",  
 title = "screen porch area")



Usando un modello di regressione lineare si può approssimare il rapporto tra le due variabili. Data come ipotesi nulla la non correttezza del modello, osserviamo che con un p-value di $ $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il portico con zanzariera, che è approssiabile da un modello lineare con pendenza positiva.

### PoolArea

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati.

## PoolArea.exists  
## FALSE TRUE   
## 0.995 0.005

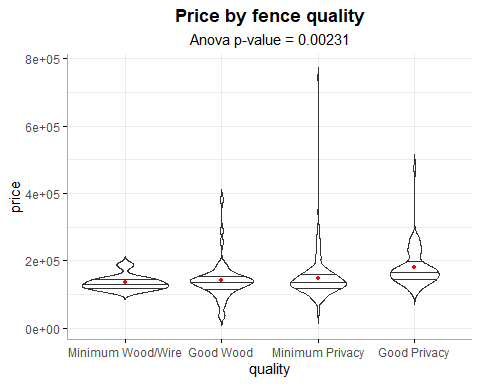
### PoolQC

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta in una sola categoria.

## PoolQC.na  
## <NA> Gd Ex Fa   
## 99.52 0.21 0.14 0.14

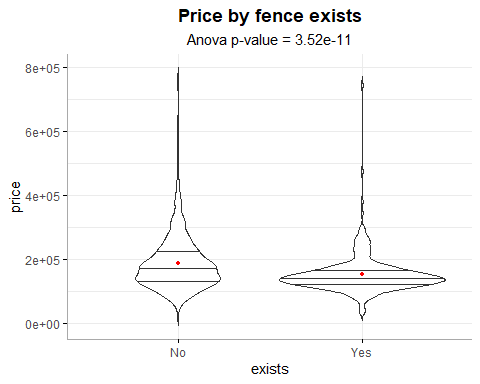
### Fence

analysis.qualitative(  
 Fence,  
 lab = "quality",  
 title = "fence quality")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ $ c’è un debolissimo supporto all’ipotesi nulla. Quindi c’è una correlazione abbastanza forte tra il prezzo di vendità e la qualità della recinzione.

Fence.exists <- factor(  
 !is.na(Fence),  
 levels = c(F, T),   
 labels = c("No", "Yes"))  
  
analysis.qualitative(  
 Fence.exists,  
 lab = "exists",  
 title = "fence exists")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-11} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e l’esistenza della recinzione

### MiscFeature

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta nella categoria NA.

## MiscFeature.na  
## <NA> Shed Gar2 Othr TenC   
## 96.30 3.36 0.14 0.14 0.07

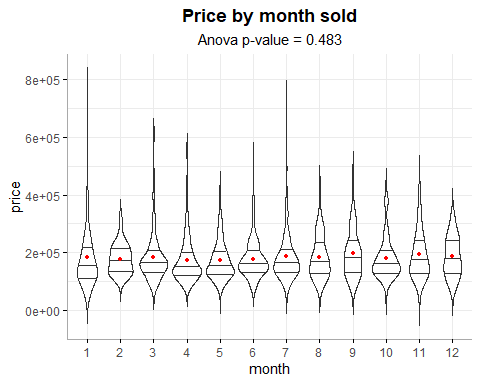
### MiscVal

Questa variabile non viene messa in correlazione perché i dati sono troppo sbilanciati. Infatti la maggiorparte dei dati sta nella categoria 0 (non ci sono costi).

##   
## 0 > 0   
## 96.44 3.56

### MoSold

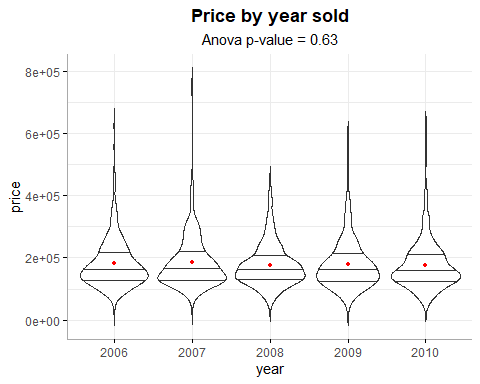
MoSold.fact <- factor(  
 MoSold,  
 levels = 0:max(MoSold))  
  
analysis.qualitative(  
 MoSold.fact,  
 lab = "month",  
 title = "month sold")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ $ c’è supporto all’ipotesi nulla. Quindi non c’è una correlazione tra il prezzo di vendità e il mese di vendità.

### YrSold

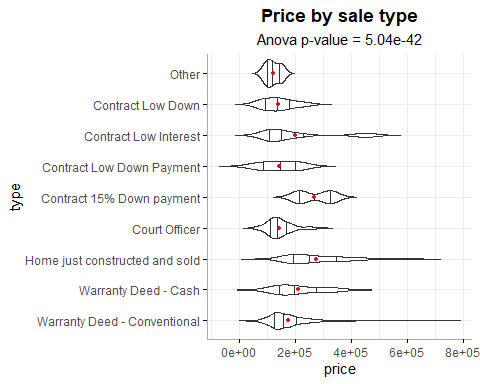
YrSold.fact <- factor(  
 YrSold,  
 levels = 0:max(YrSold))  
  
analysis.qualitative(  
 YrSold.fact,  
 lab = "year",  
 title = "year sold")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ $ c’è supporto all’ipotesi nulla. Quindi non c’è una correlazione tra il prezzo di vendità e l’anno di vendita.

### SaleType

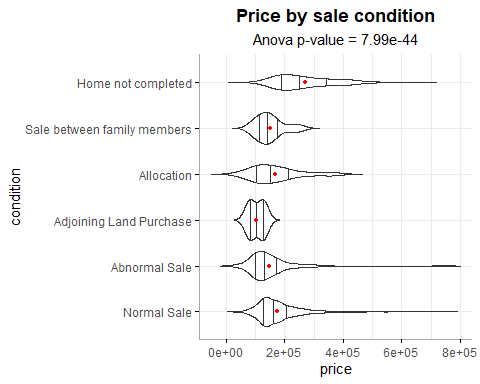
analysis.qualitative(  
 SaleType,  
 horizontal = T,  
 lab = "type",  
 title = "sale type")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-42} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e il tipo di contratto di vendita.

### SaleCondition

analysis.qualitative(  
 SaleCondition,  
 horizontal = T,  
 lab = "condition",  
 title = "sale condition")



Data l’uguaglianza delle medie come ipotesi nulla, osserviamo che con un p-value di $ ^{-44} $ non c’è supporto all’ipotesi nulla. Quindi c’è una correlazione tra il prezzo di vendità e le condizioni di vendita.