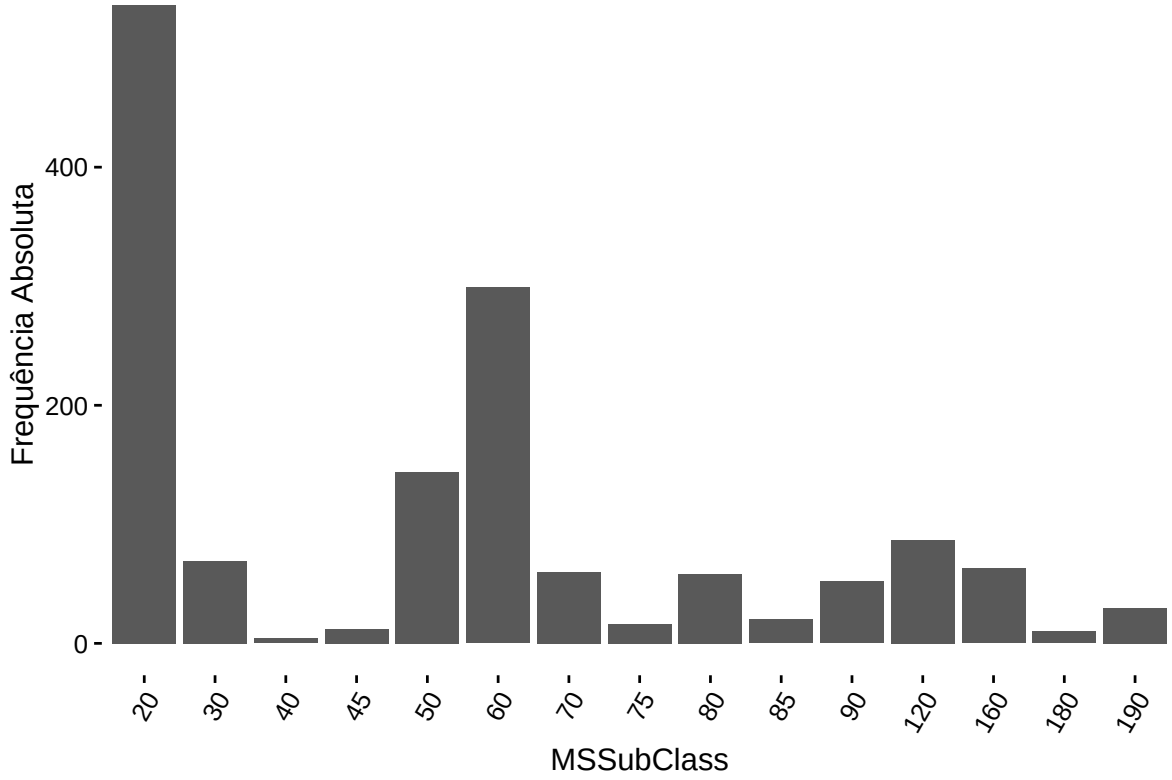
Análise Exploratória

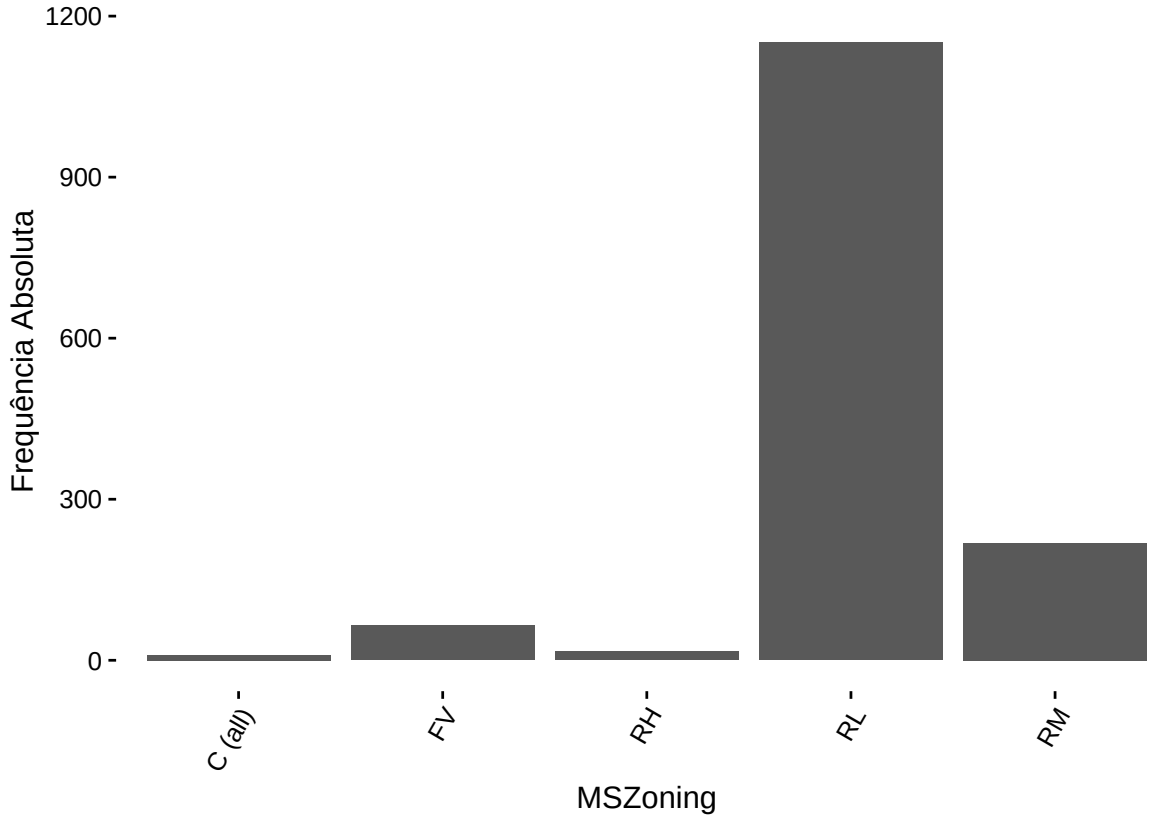
Ossada R. e Hamaguchi L.

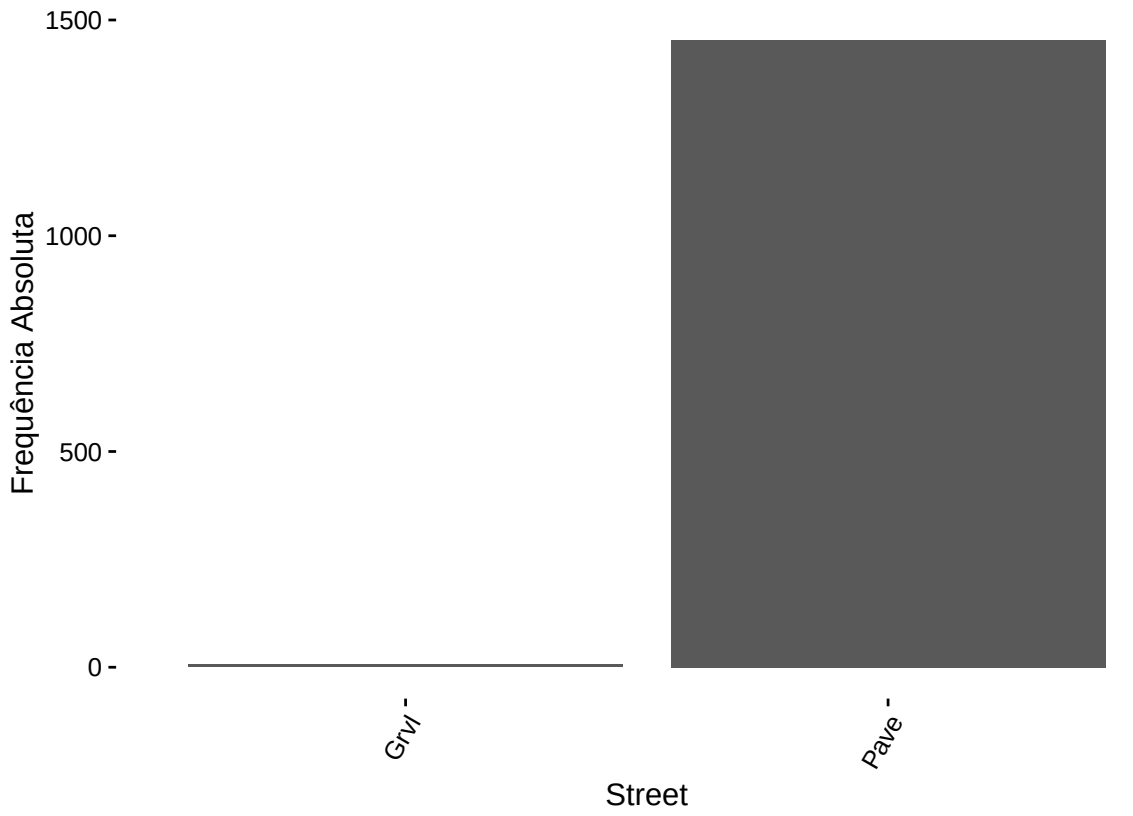
25 de julho de 2017

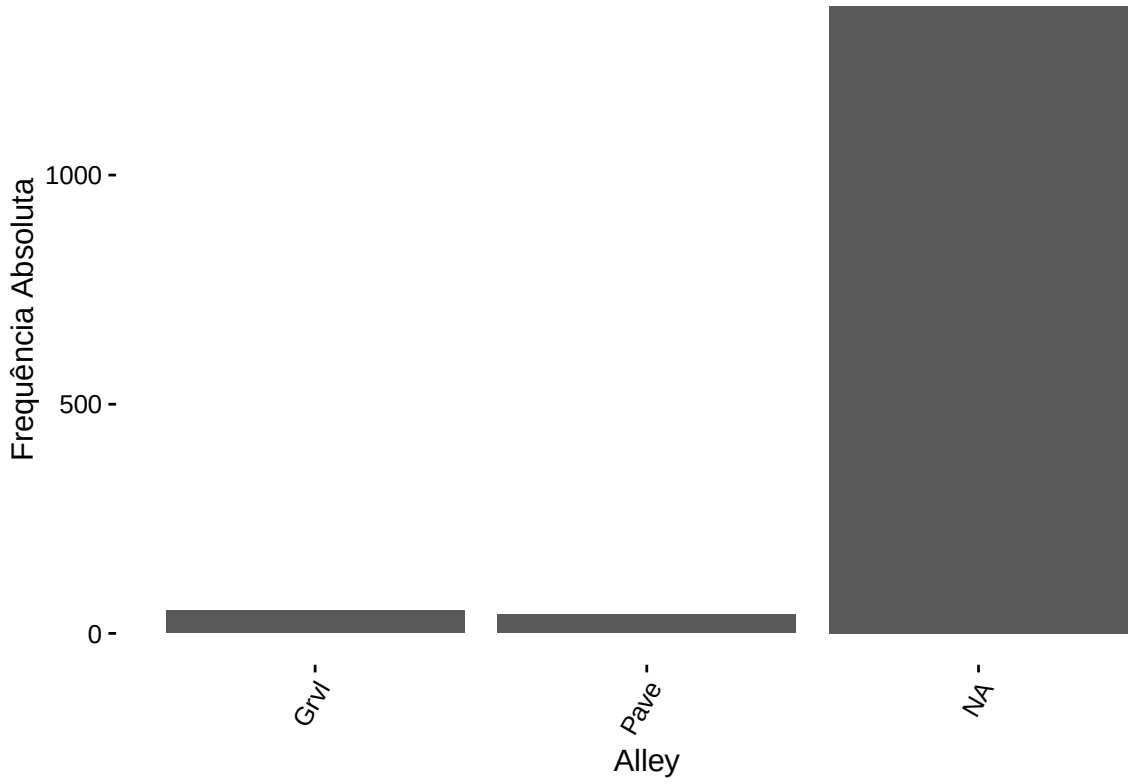
1 ANÁLISE EXPLORATÓRIA DOS DADOS

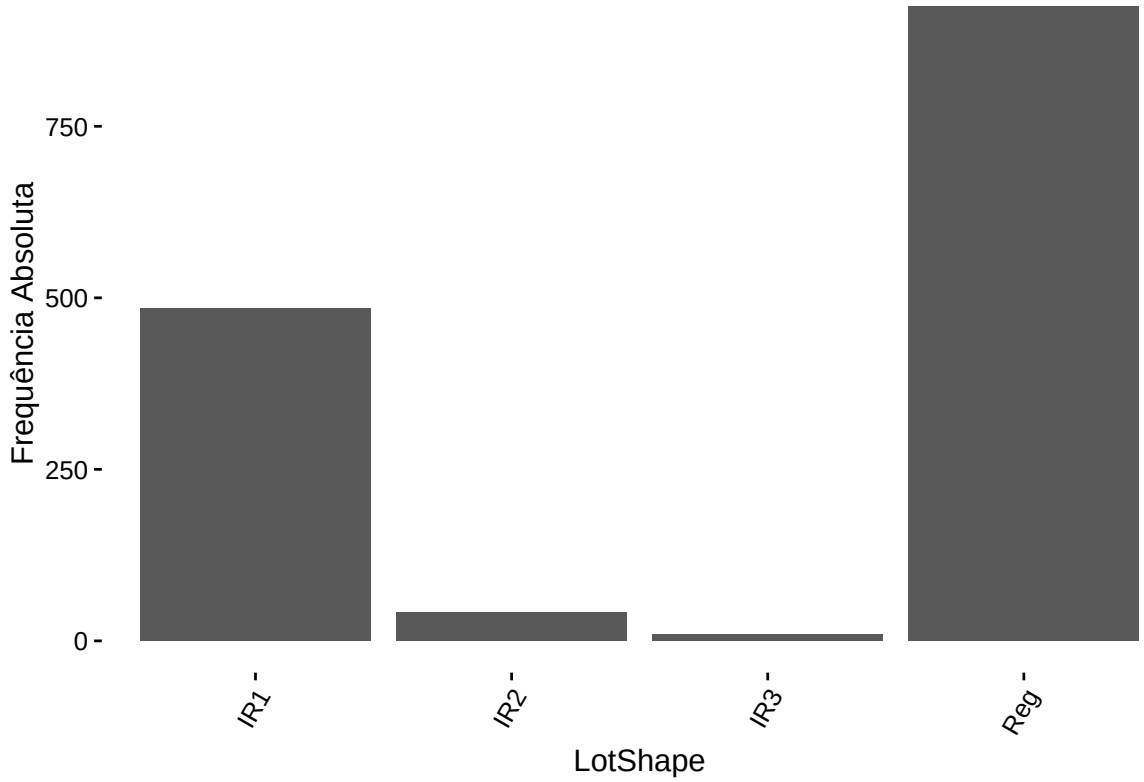
1.1 GRÁFICOS DAS VARIÁVEIS QUALITATIVAS

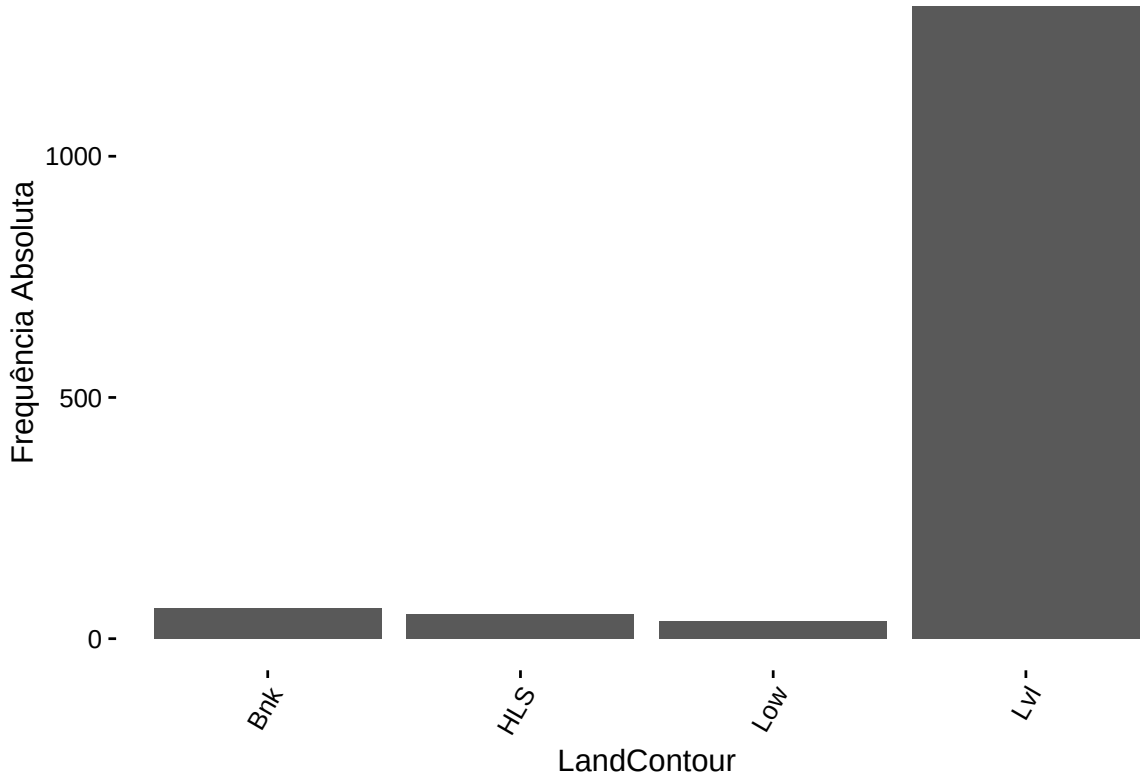


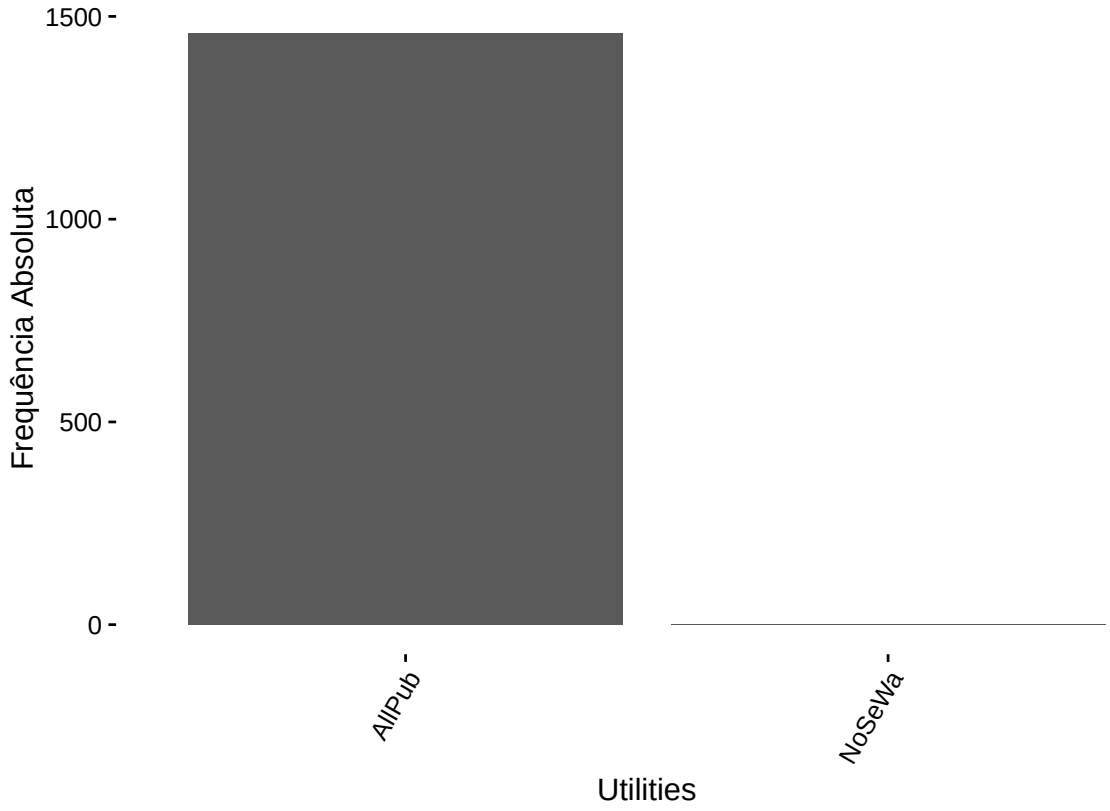


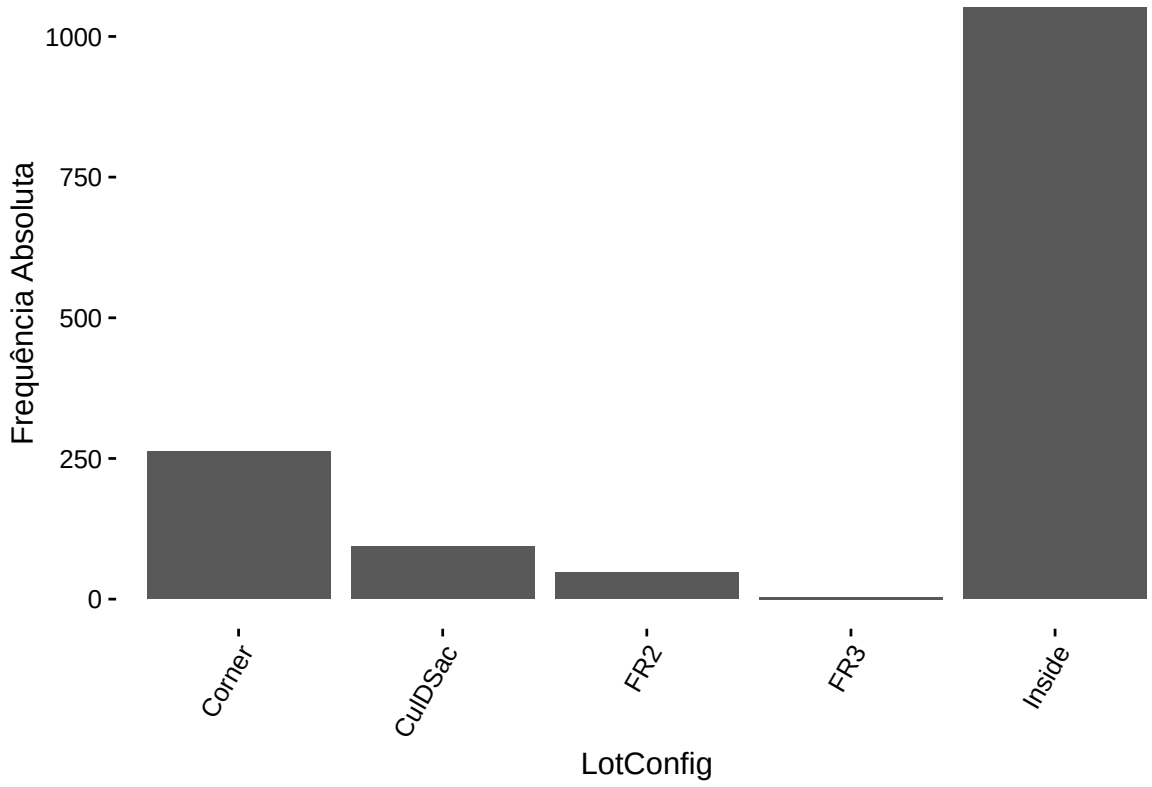

1500
1000
500
0
Frequência Absoluta
Grvl
Pave
Street

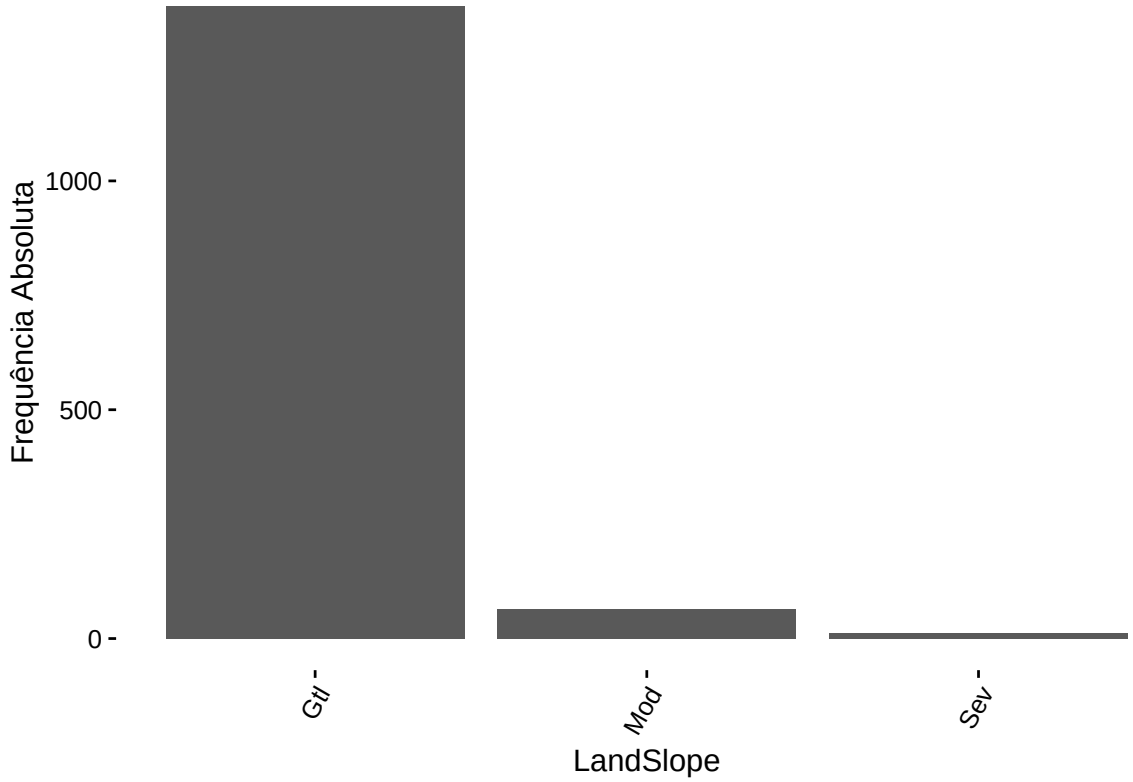

1000
500
0
Frequência Absoluta
Grvl
Pave
NA
Alley

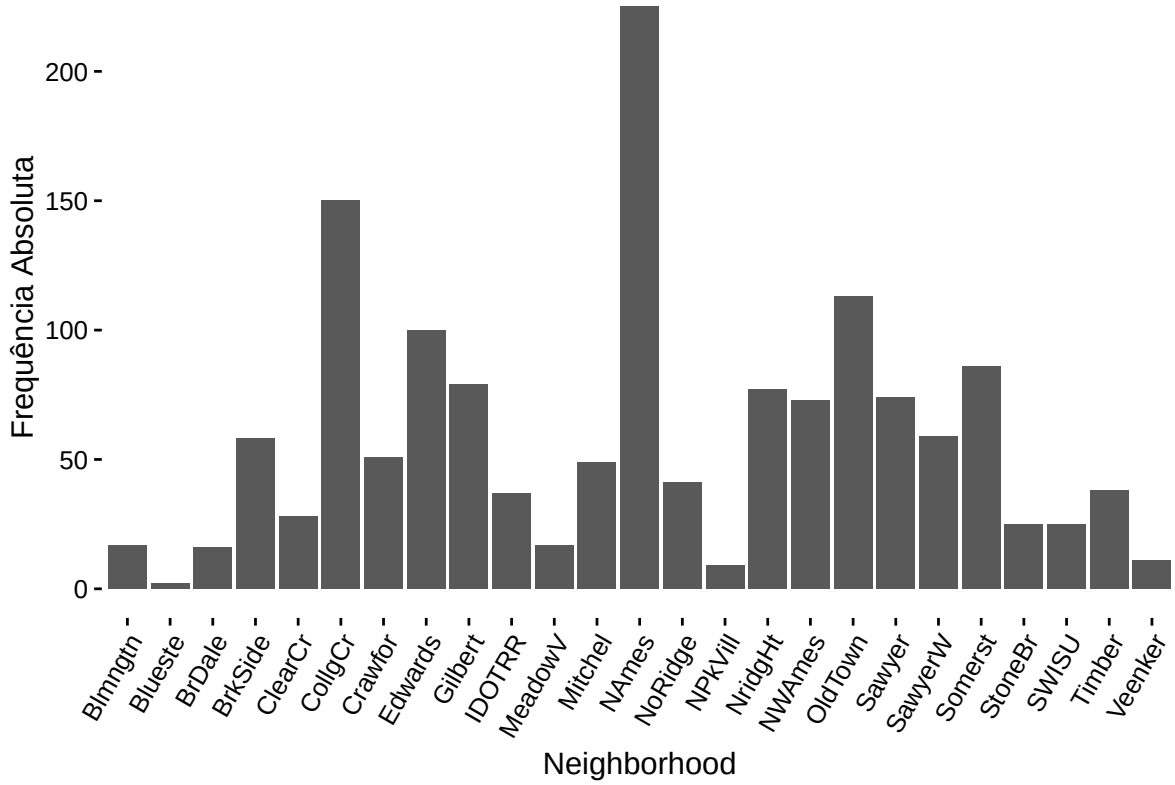

750
500
250
0
Frequência Absoluta
IR1
IR2
IR3
Reg
LotShape

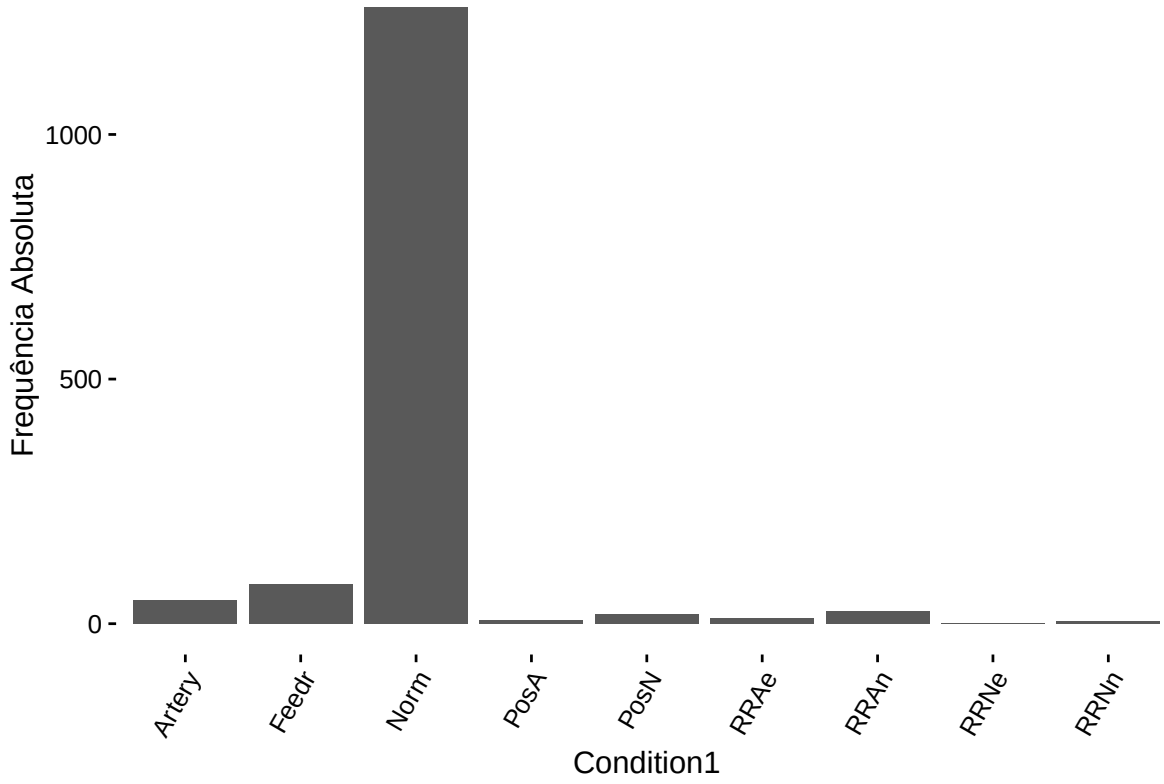

1000
500
0
Frequência Absoluta
Bnk
HLS
Low
Lvl
LandContour

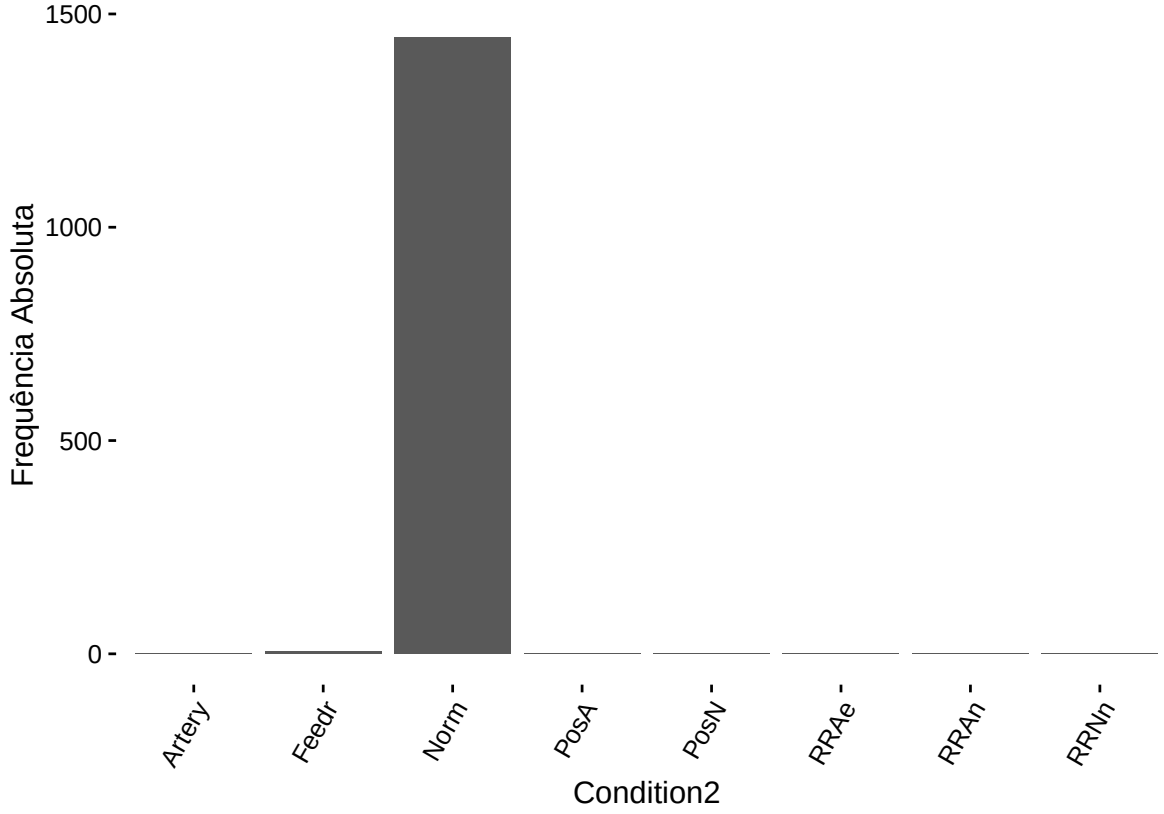

1500
1000
500
0
Frequência Absoluta
AllPub
NoSeWa
Utilities

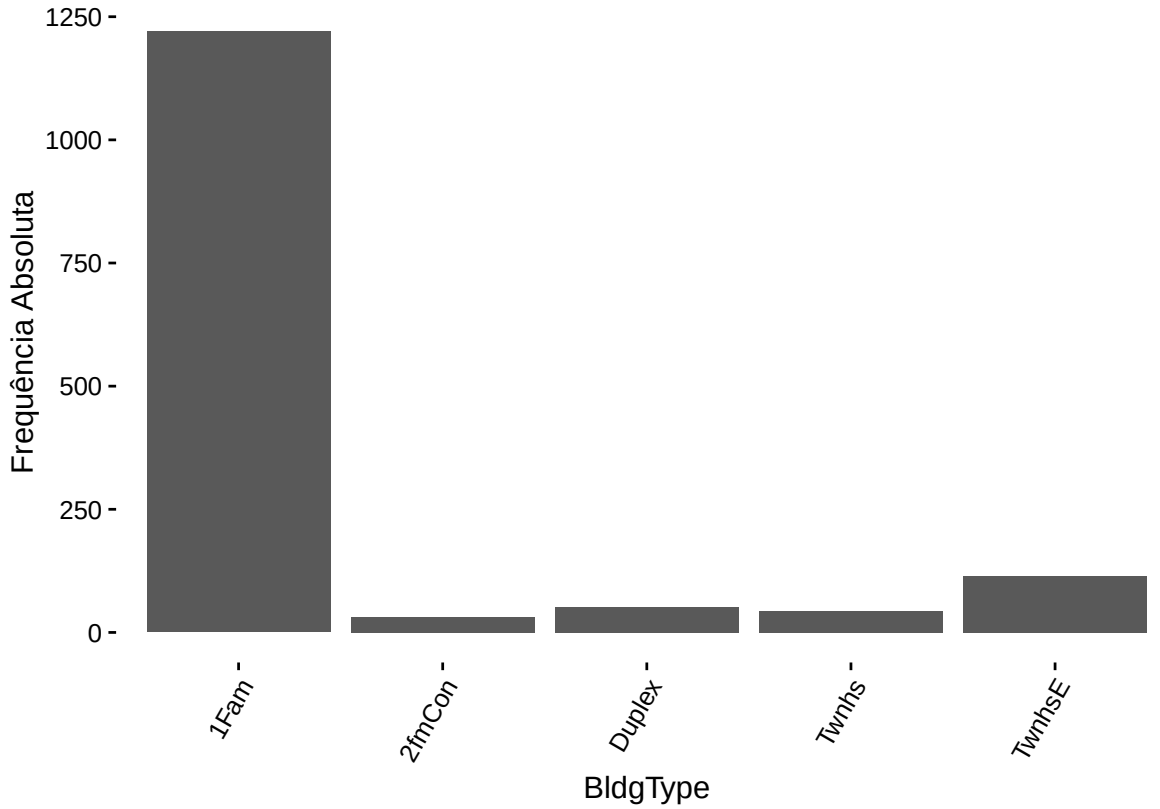

1000
750
500
250
0
Frequência Absoluta
Corner
CulDSac
FR2
FR3
Inside
LotConfig

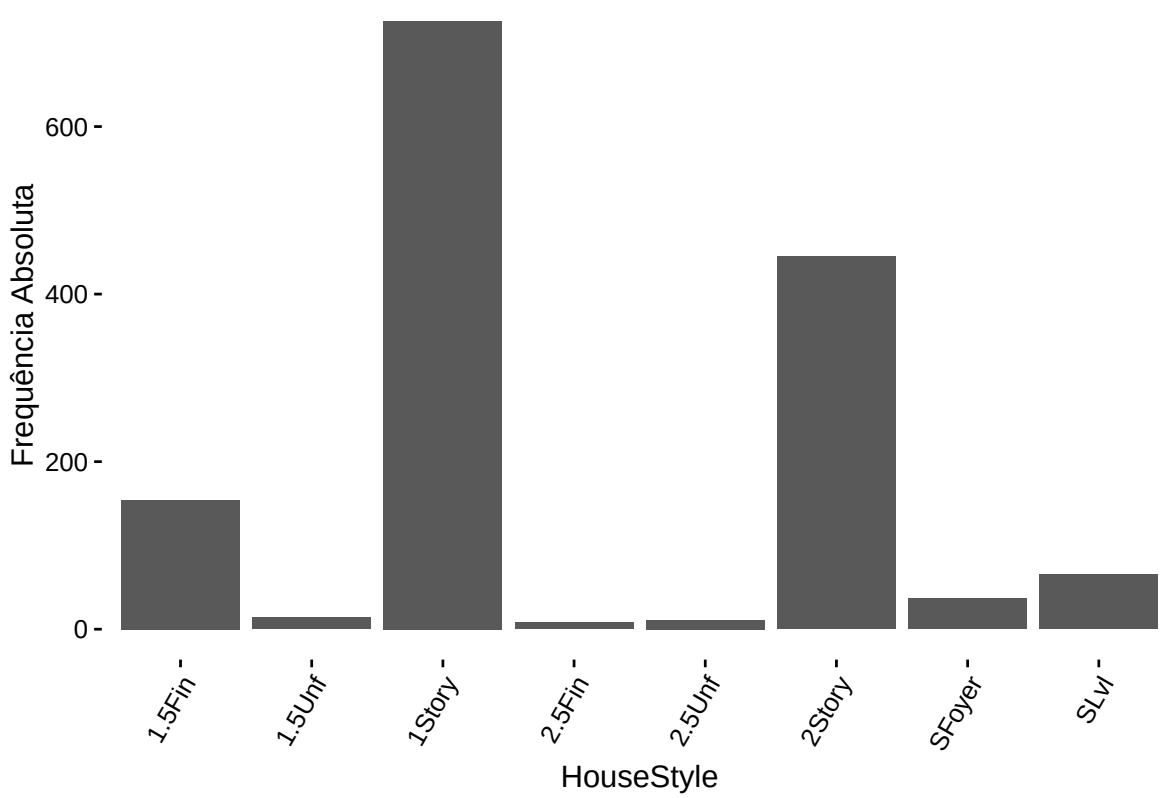

Frequência Absoluta
1000
500
0
Gtl
Mod
Sev
LandSlope

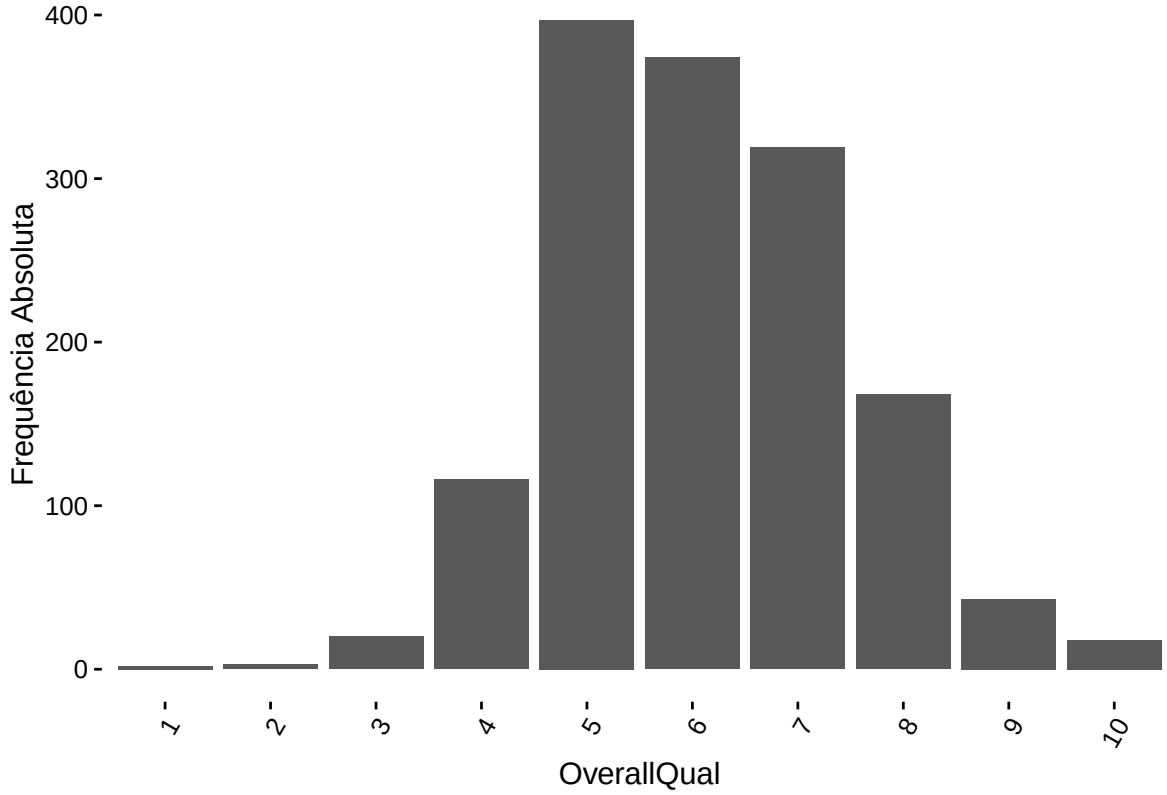

Frequência Absoluta
200
150
100
50
0
Blmngtn
Blueste
BrDale
BrkSide
ClearCr
CollgCr
Crawfor
Edwards
Gilbert
IDOTRR
MeadowV
Mitchel
NAmes
NoRidge
NPkVill
NridgHt
NWAmes
OldTown
Sawyer
SawyerW
Somerst
StoneBr
SWISU
Timber
Veenker
Neighborhood

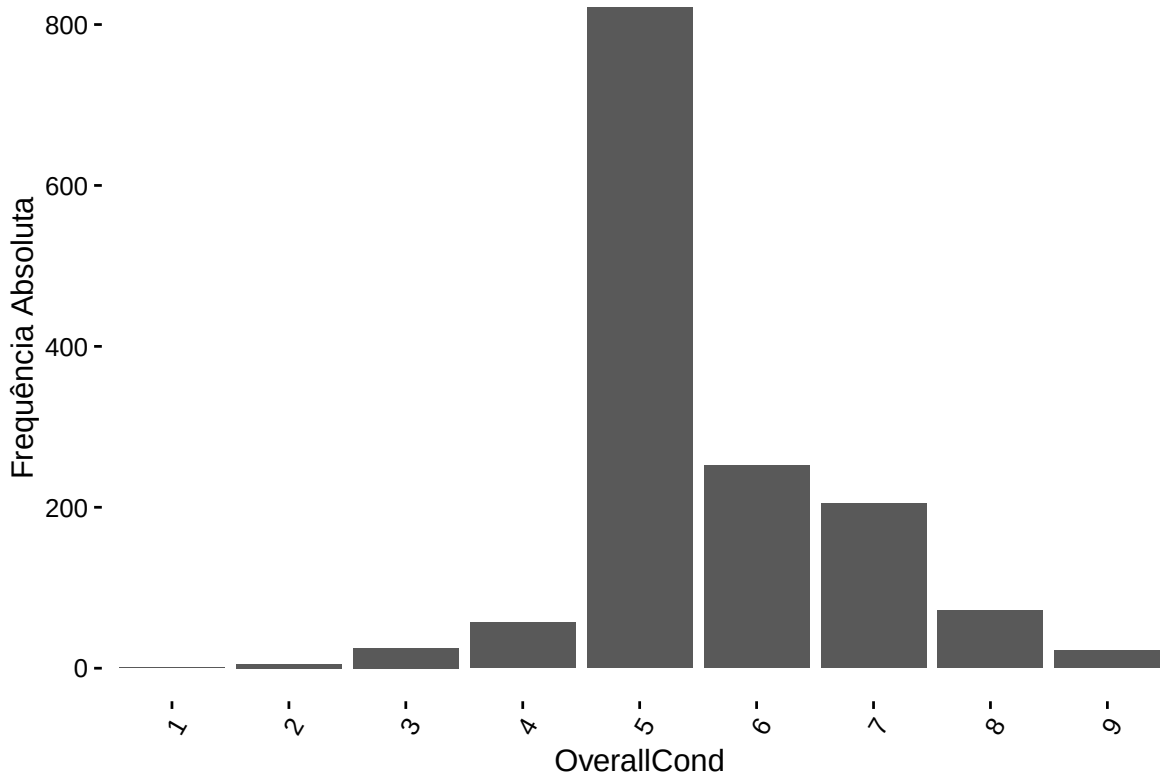

Frequência Absoluta
1000
500
0
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNe
RRNn
Condition1

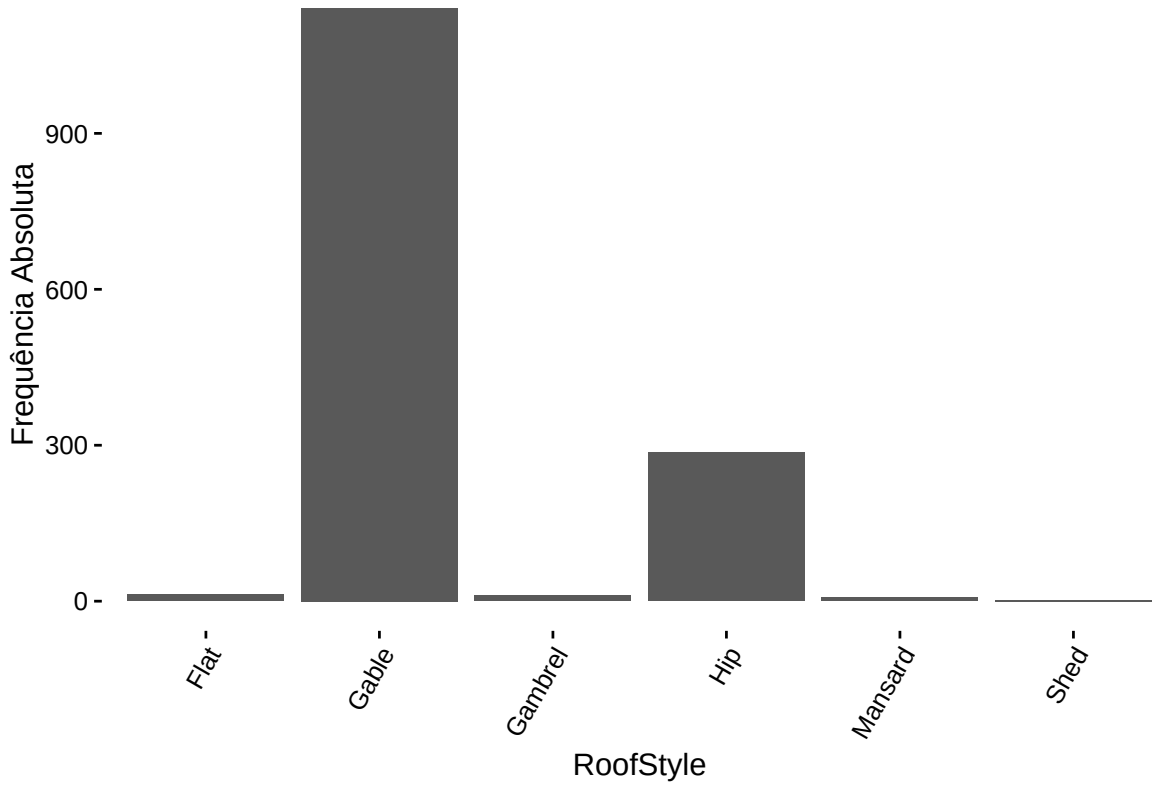

Frequência Absoluta
1500
1000
500
0
Artery
Feedr
Norm
PosA
PosN
RRAe
RRAn
RRNn
Condition2

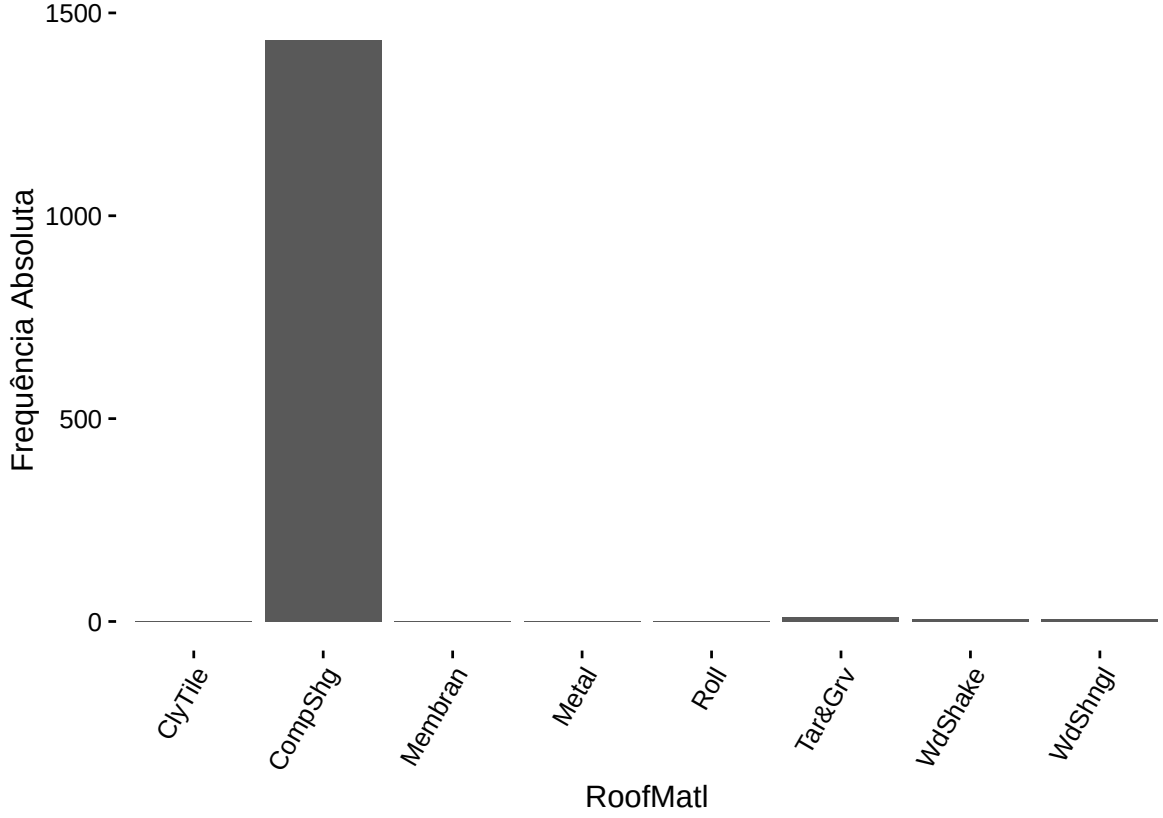

1250
1000
750
500
250
0
Frequência Absoluta
1Fam
2fmCon
Duplex
Twnhs
TwnhsE
BldgType

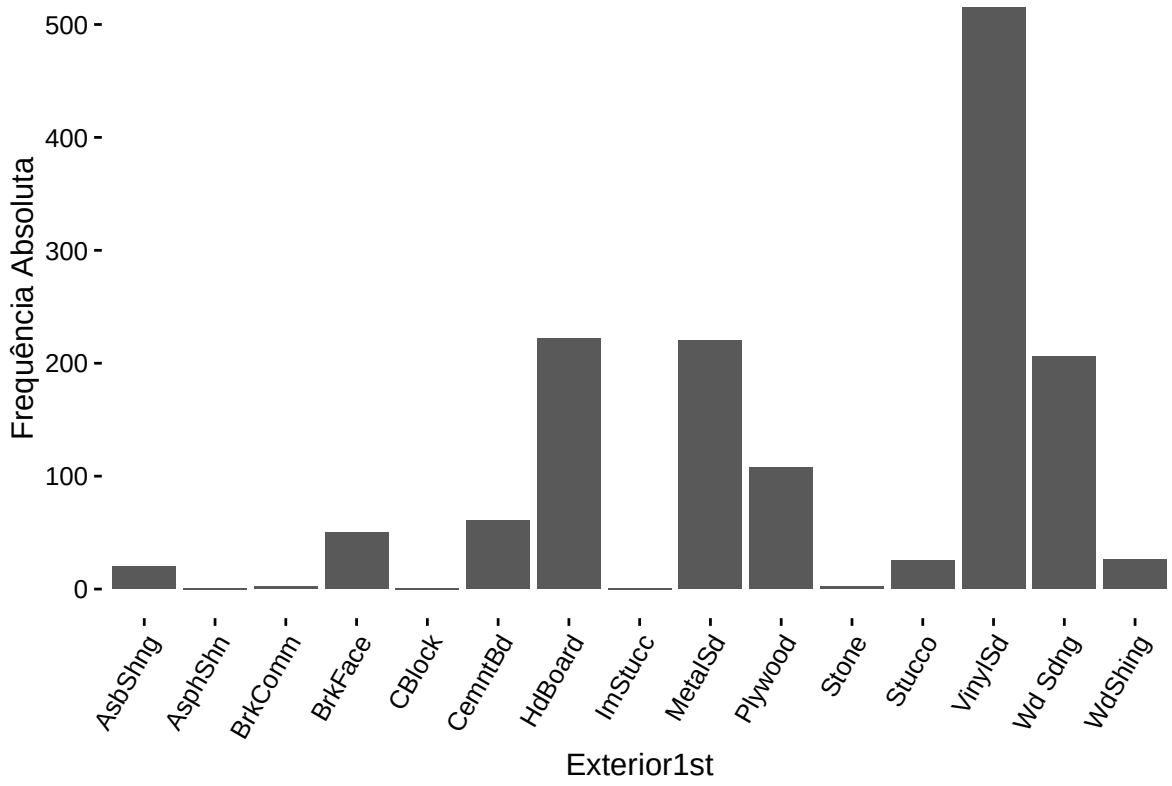

600
400
200
0
Frequência Absoluta
1.5Fin
1.5Unf
1Story
2.5Fin
2.5Unf
2Story
SFoyer
SLvl
HouseStyle

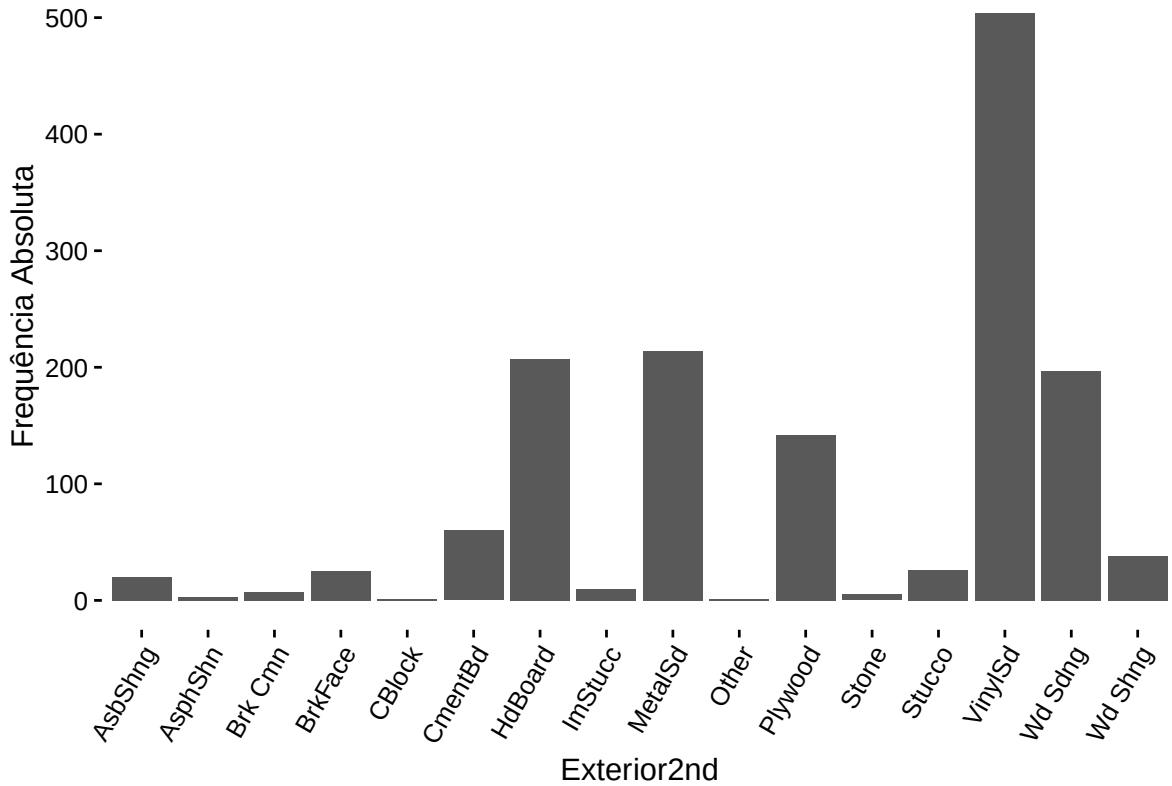

400
300
200
100
0
Frequência Absoluta
1
2
3
4
5
6
7
8
9
10
OverallQual

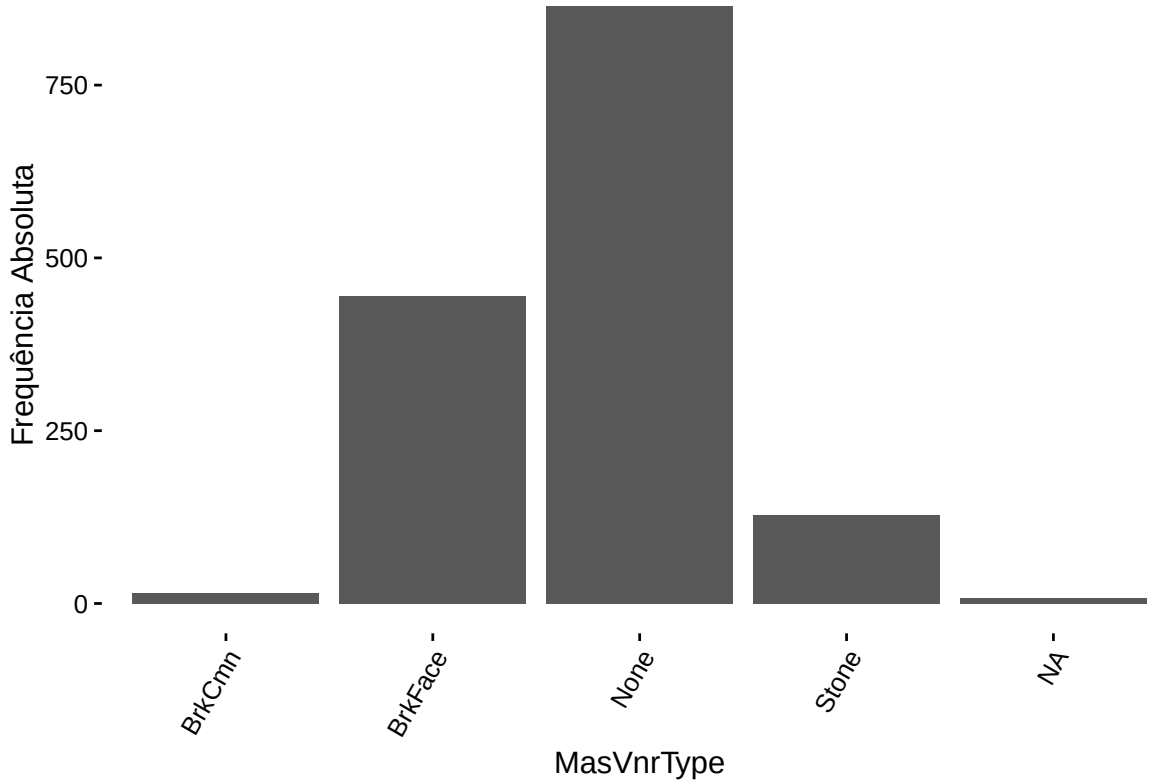

800
600
400
200
0
Frequência Absoluta
1
2
3
4
5
6
7
8
9
OverallCond

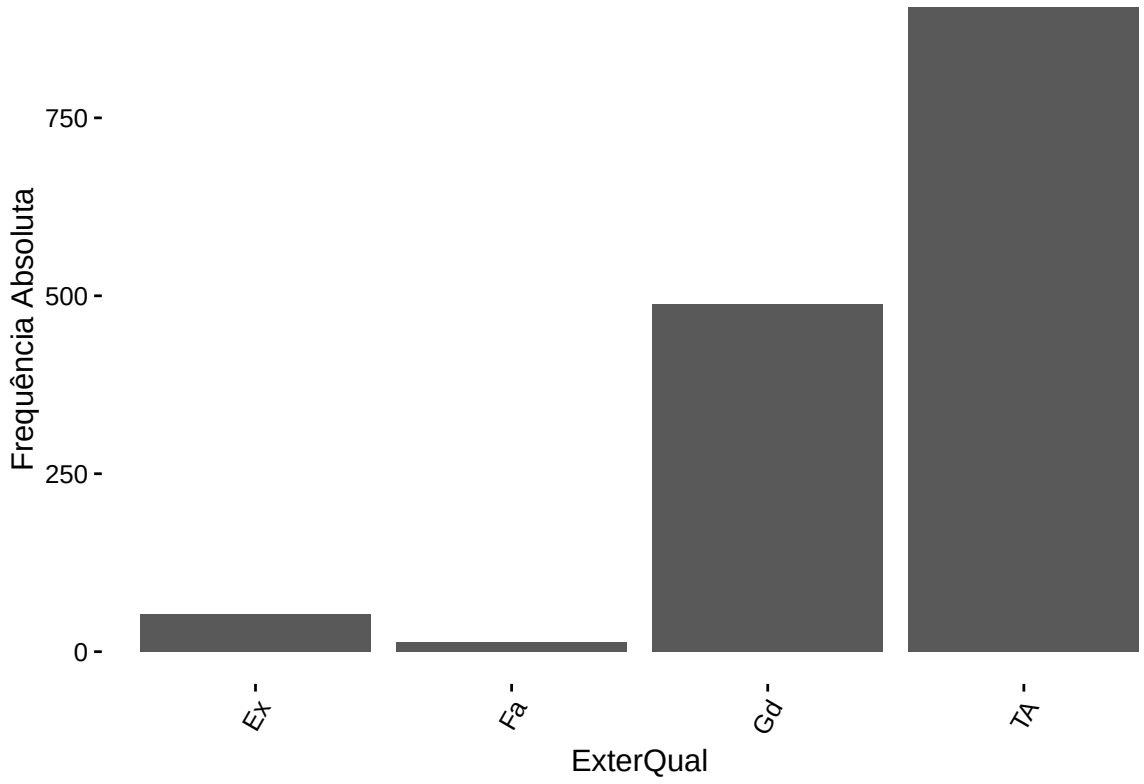

Frequência Absoluta
900
600
300
0
Flat
Gable
Gambrel
Hip
Mansard
Shed
RoofStyle

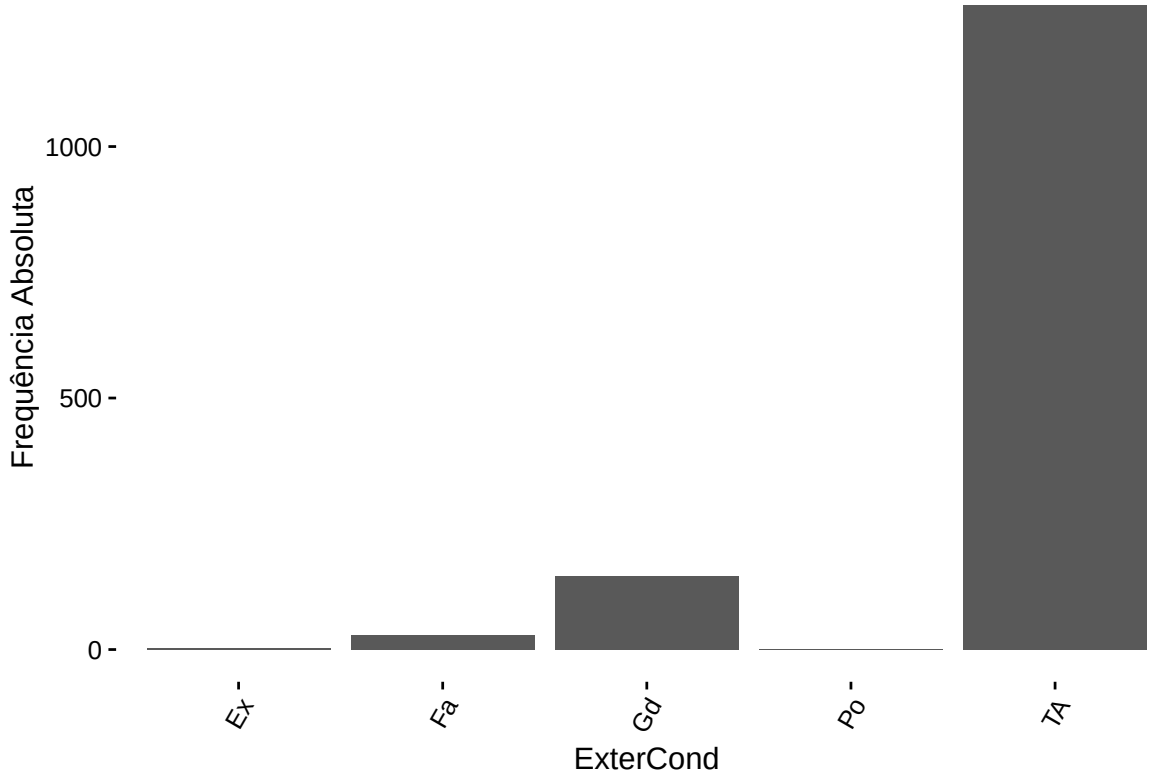

Frequência Absoluta
1500
1000
500
0
ClyTile
CompShg
Membran
Metal
Roll
Tar&Grv
WdShake
WdShngl
RoofMatl

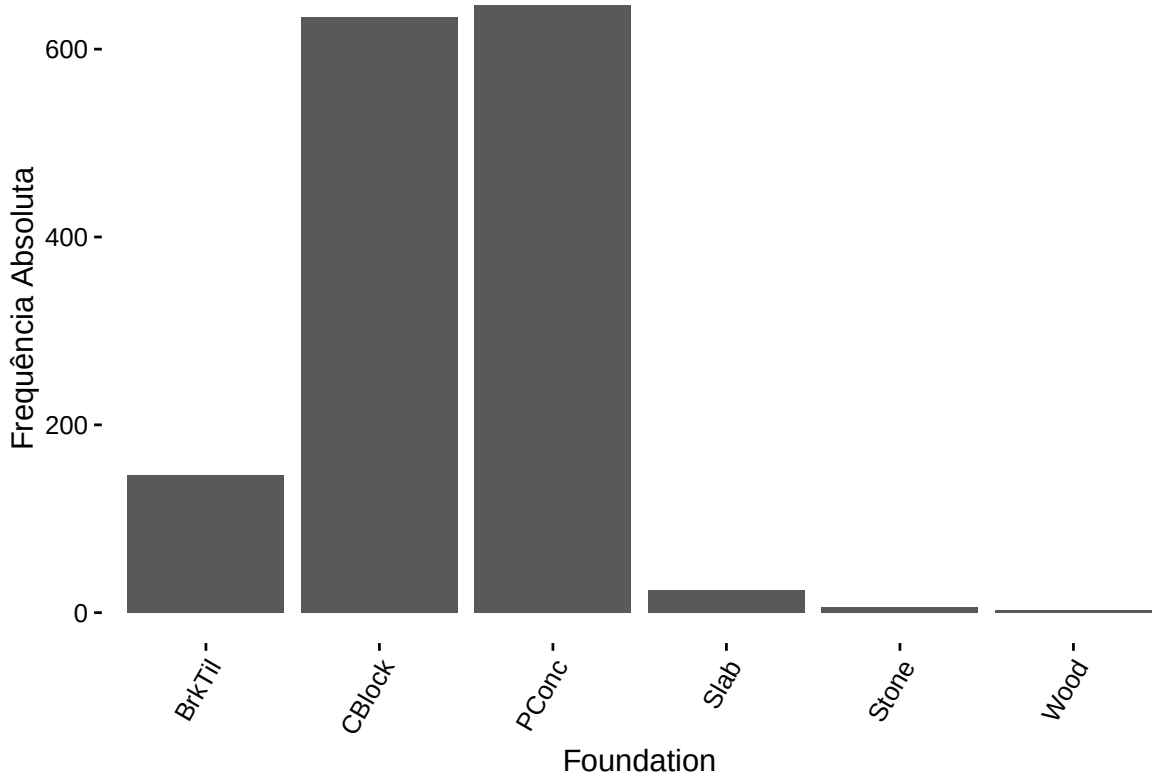

Frequência Absoluta
500
400
300
200
100
0
AsbShng
AsphShn
BrkComm
BrkFace
CBlock
CemntBd
HdBoard
ImStucc
MetalSd
Plywood
Stone
Stucco
VinylSd
Wd Sdng
WdShing
Exterior1st

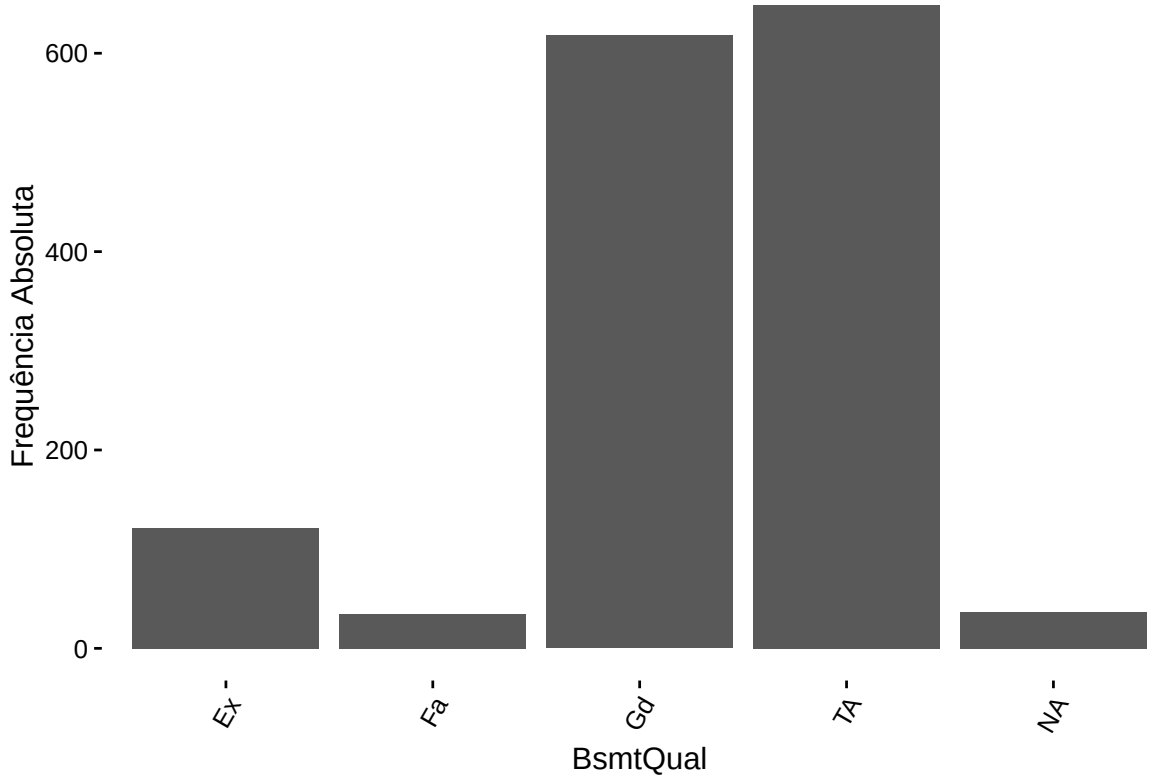

Frequência Absoluta
500
400
300
200
100
0
AsbShng
AsphShn
Brk Cmn
BrkFace
CBlock
CmentBd
HdBoard
ImStucc
MetalSd
Other
Plywood
Stone
Stucco
VinylSd
Wd Sdng
Wd Shng
Exterior2nd

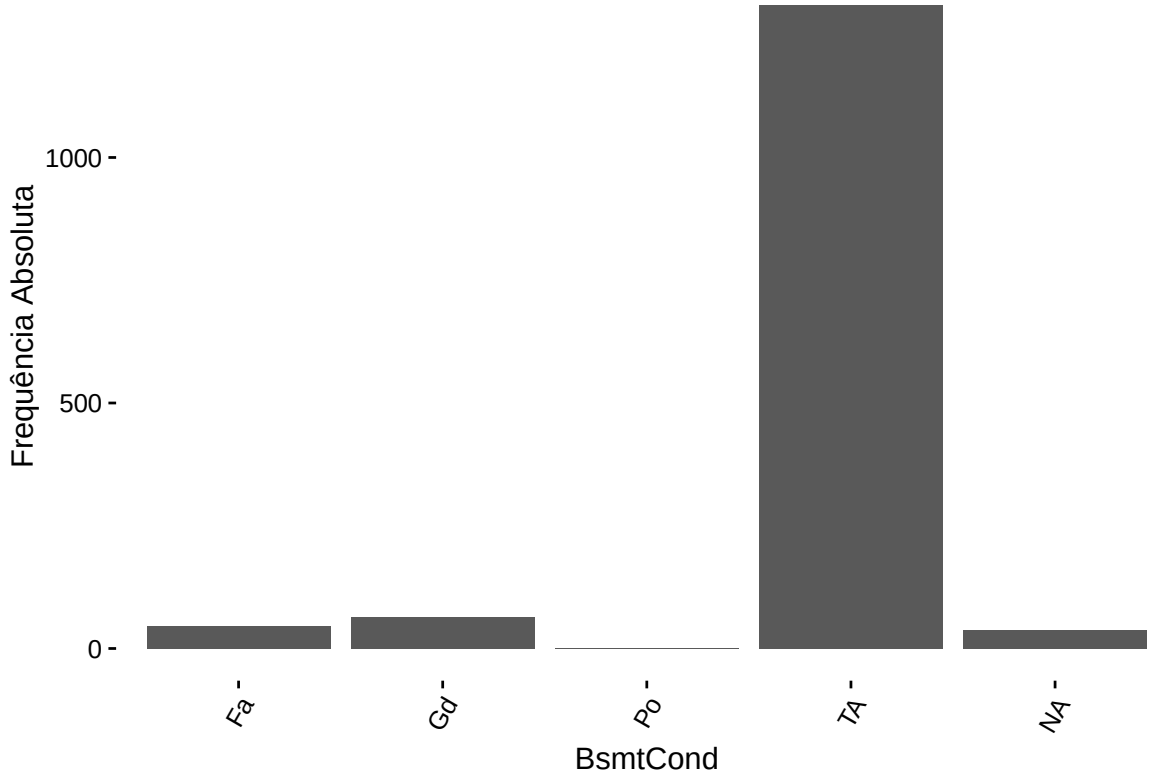

750
500
250
0
Frequência Absoluta
BrkCmn
BrkFace
None
Stone
NA
MasVnrType

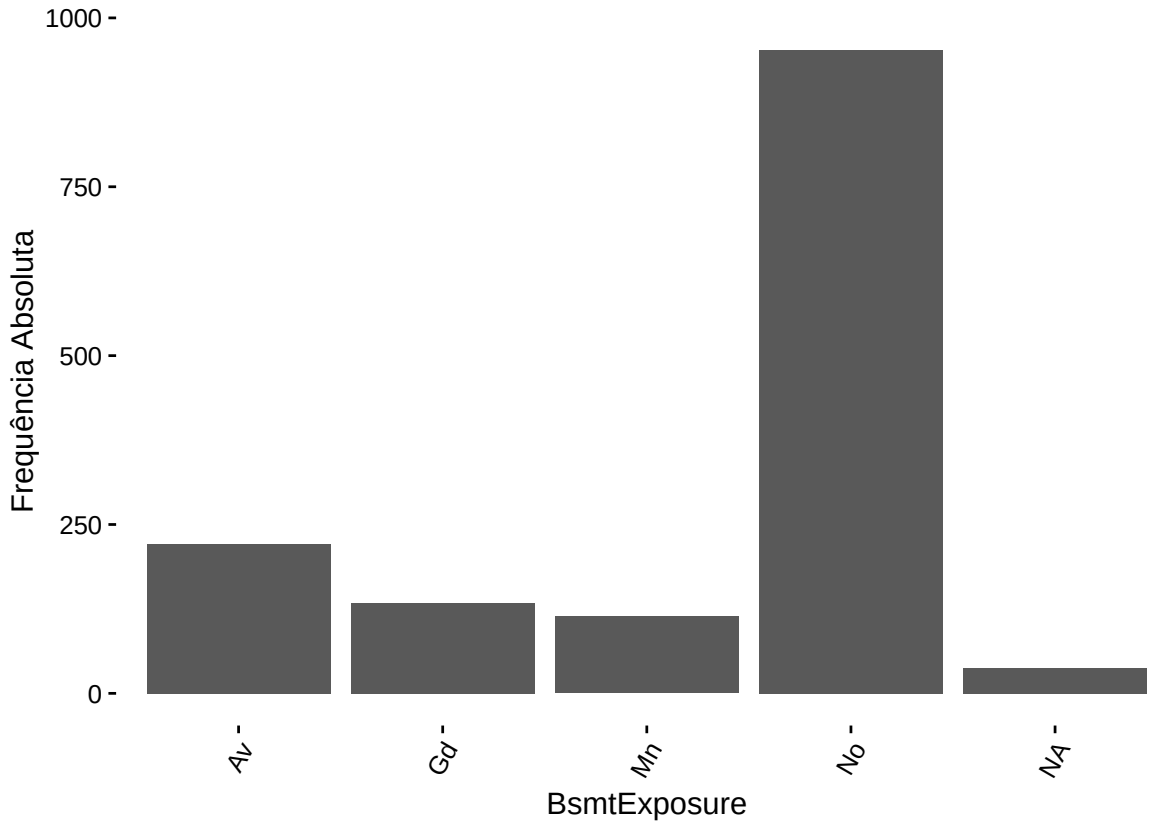

750
500
250
0
Frequência Absoluta
Ex
Fa
Gd
TA
ExterQual

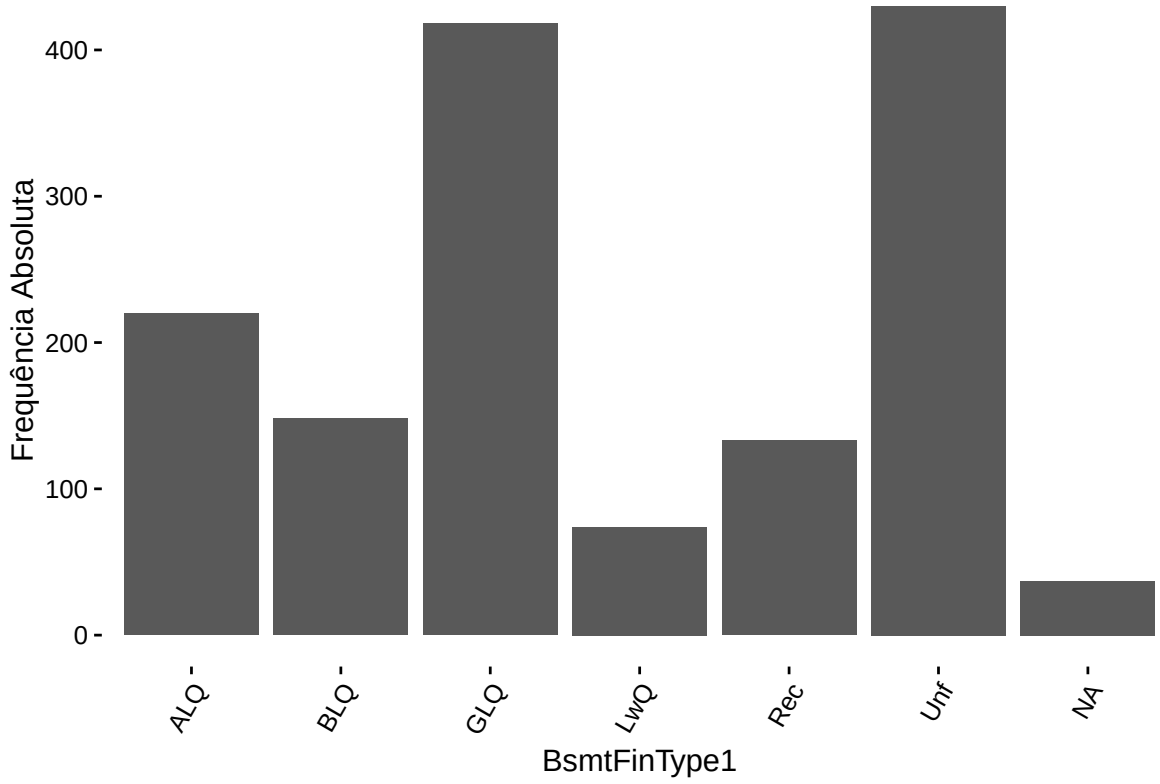

Frequência Absoluta
1000
500
0
Ex
Fa
Gd
Po
TA
ExterCond

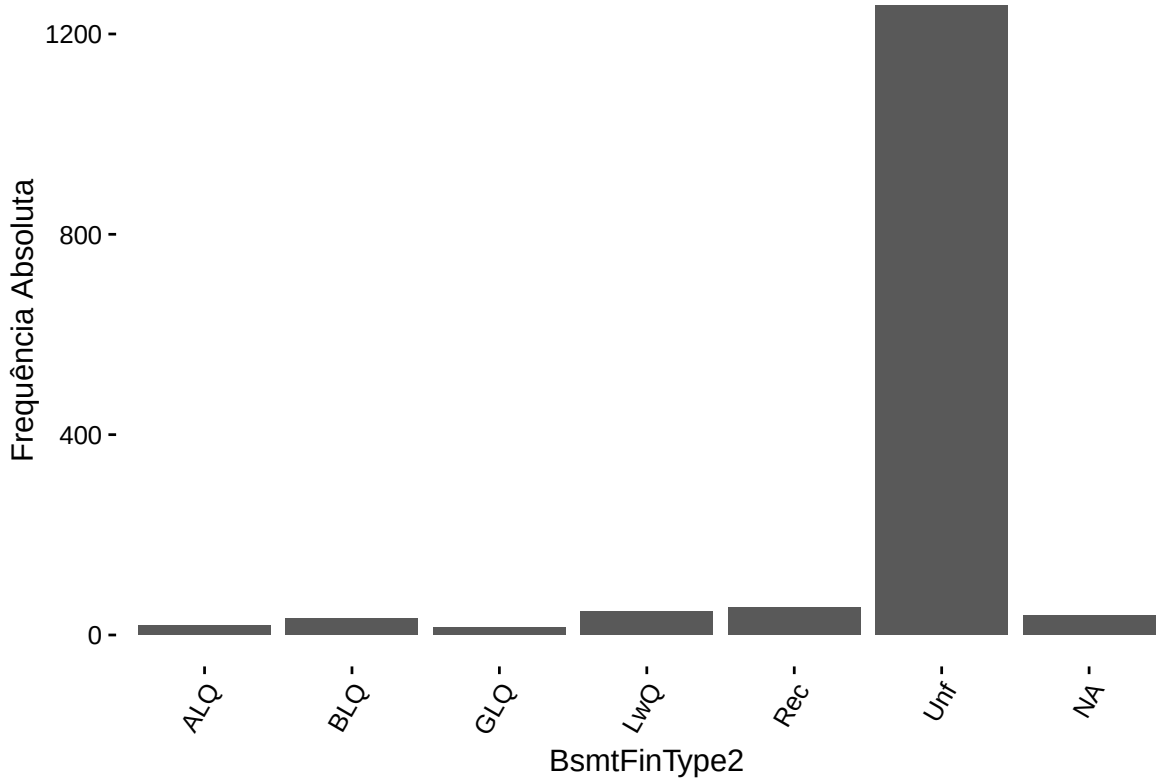

Frequência Absoluta
600
400
200
0
BrkTil
CBlock
PConc
Slab
Stone
Wood
Foundation

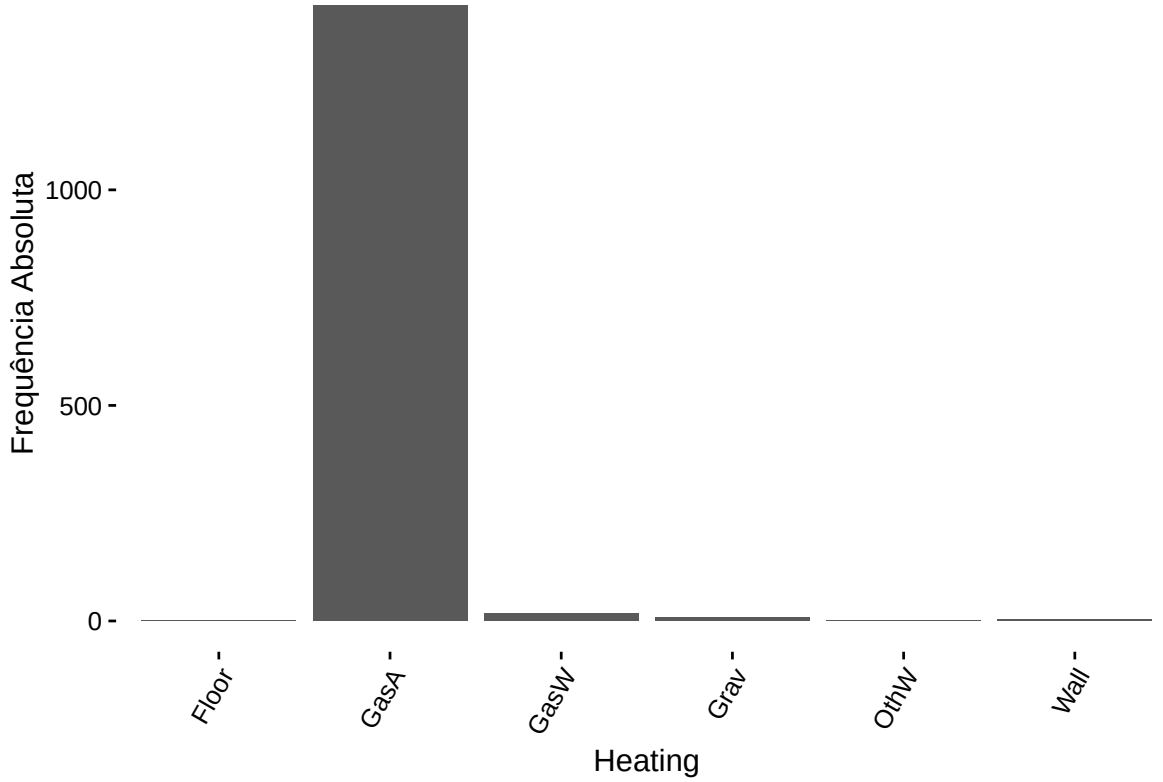

Frequência Absoluta
600
400
200
0
Ex
Fa
Gd
TA
NA
BsmtQual

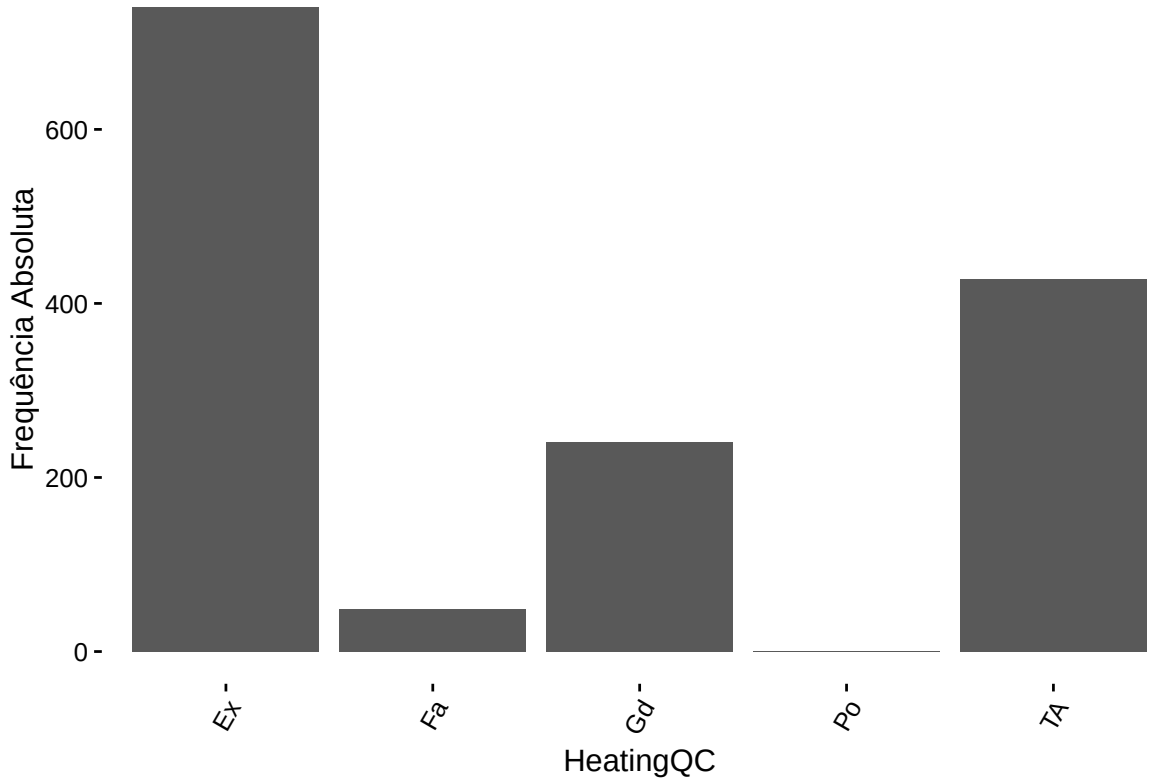

Frequência Absoluta
1000
500
0
Fa
Gd
Po
TA
NA
BsmtCond

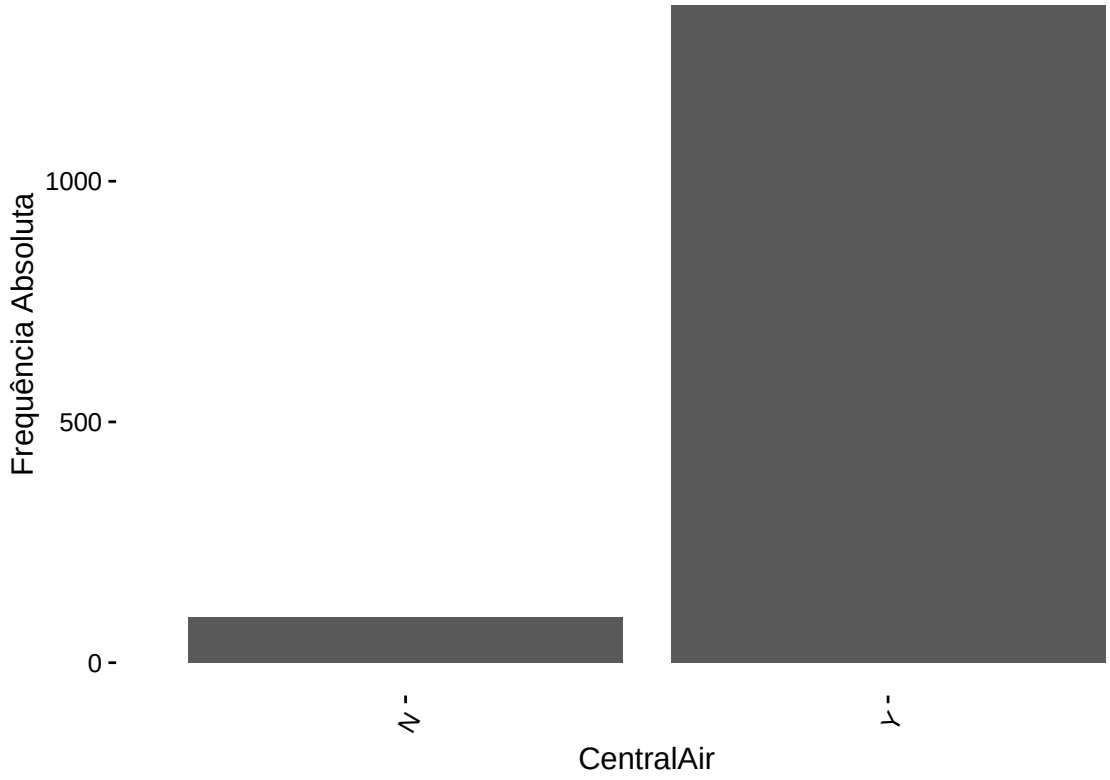

1000
750
500
250
0
Frequência Absoluta
Av
Gd
Mn
No
NA
BsmtExposure

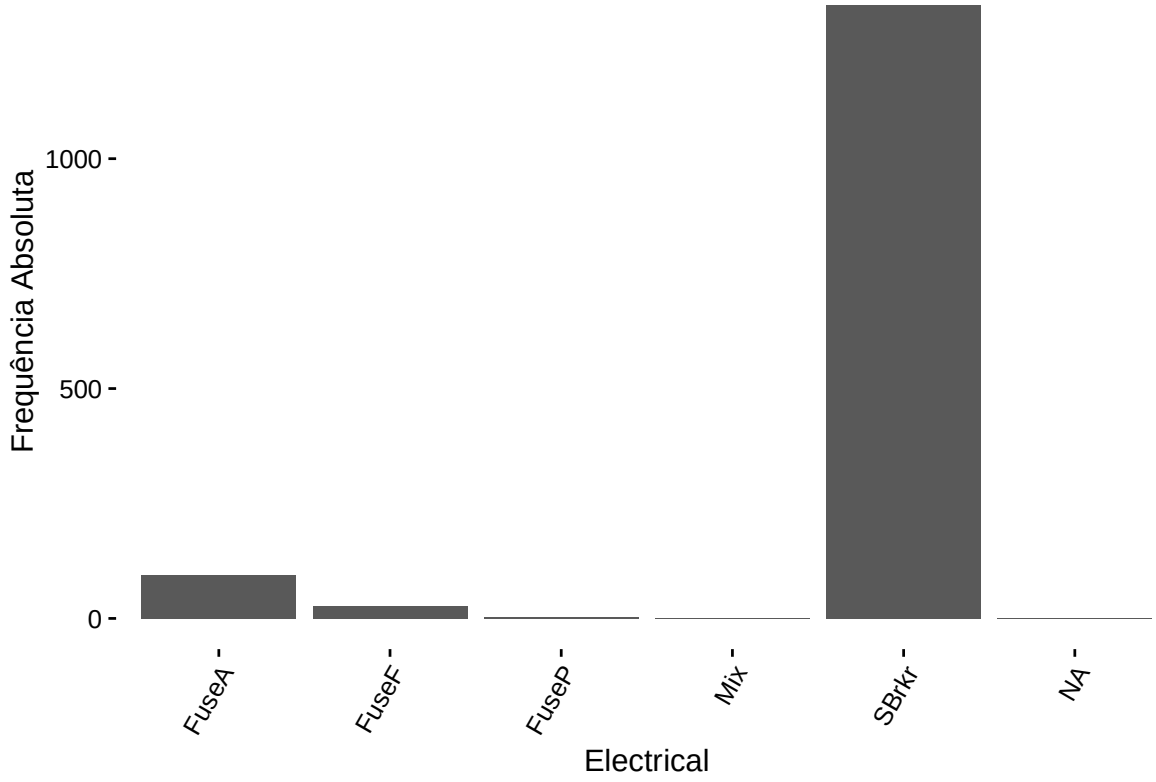

400
300
200
100
0
Frequência Absoluta
ALQ
BLQ
GLQ
LwQ
Rec
Unf
NA
BsmtFinType1

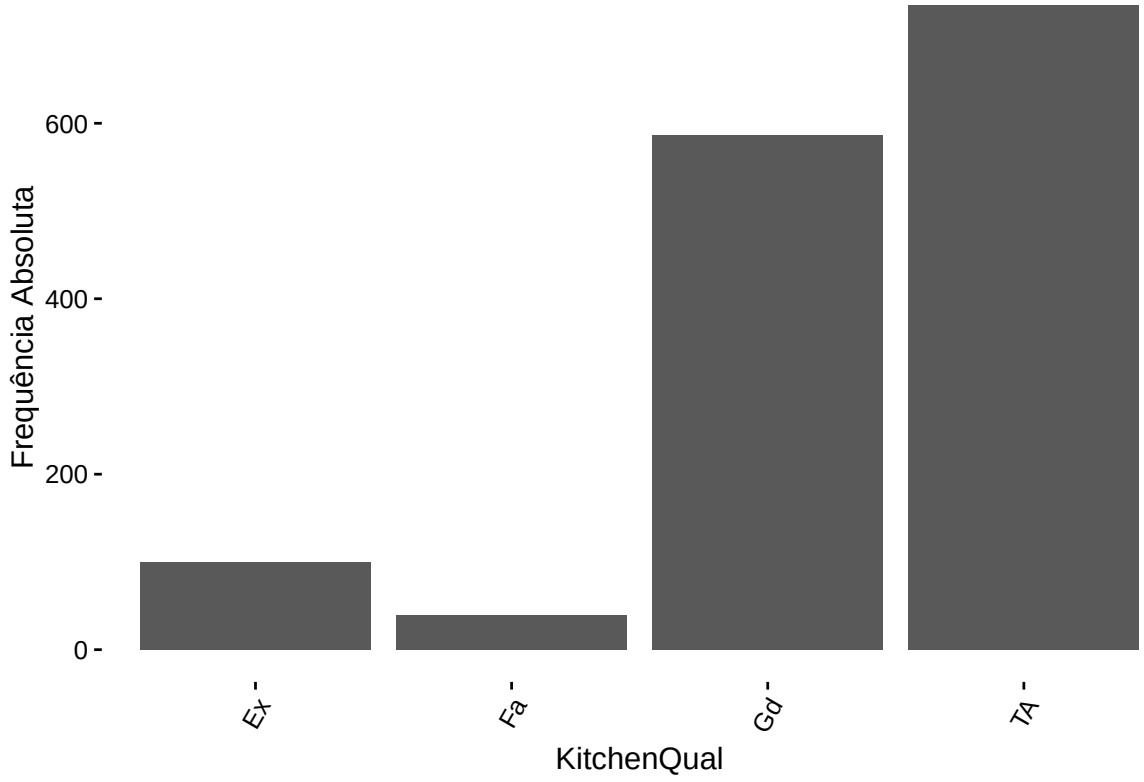

Frequência Absoluta
1200
800
400
0
ALQ
BLQ
GLQ
LwQ
Rec
Unf
NA
BsmtFinType2

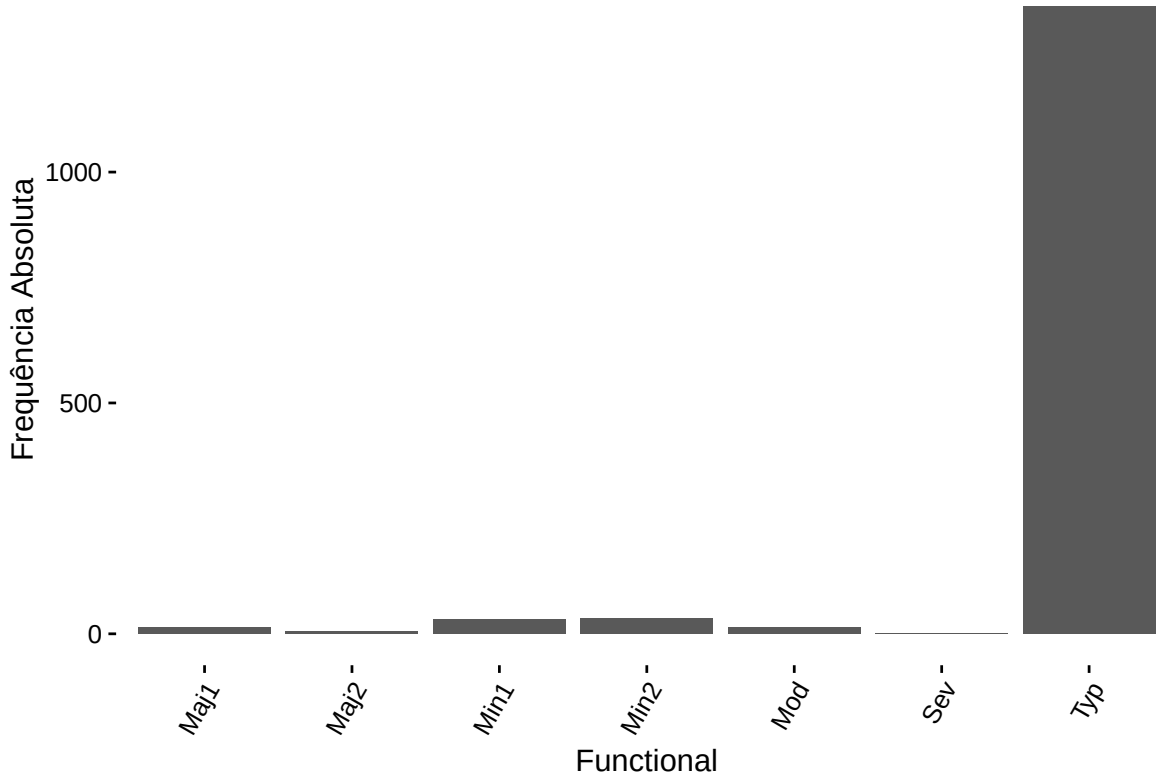

Frequência Absoluta
1000
500
0
Floor
GasA
GasW
Grav
OthW
Wall
Heating

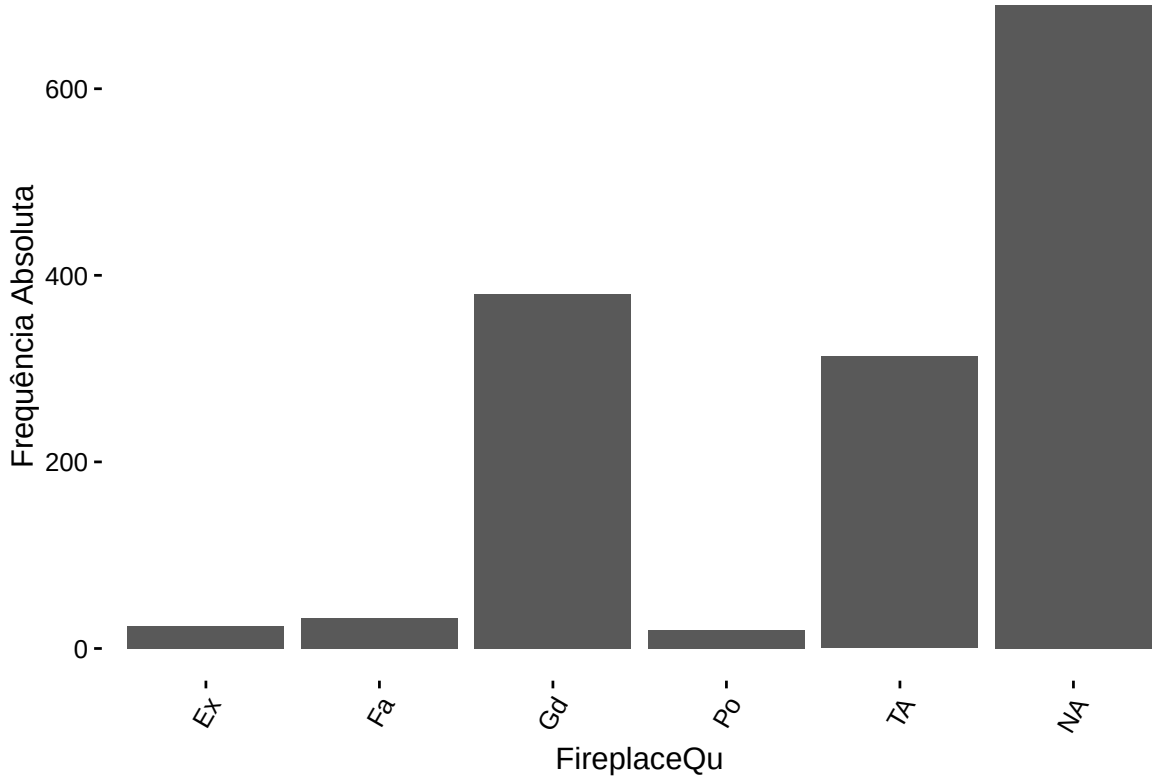

Frequência Absoluta
600
400
200
0
Ex
Fa
Gd
Po
TA
HeatingQC

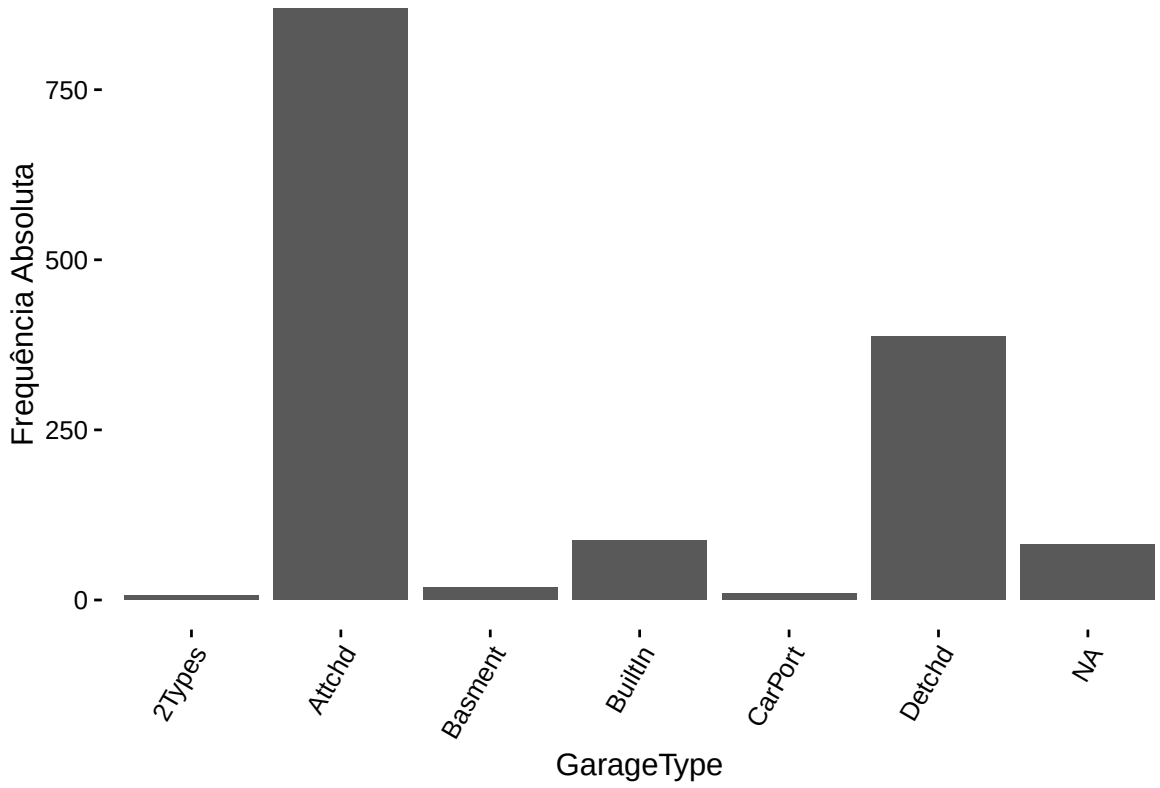

Frequência Absoluta
1000
500
0
N
Y
CentralAir

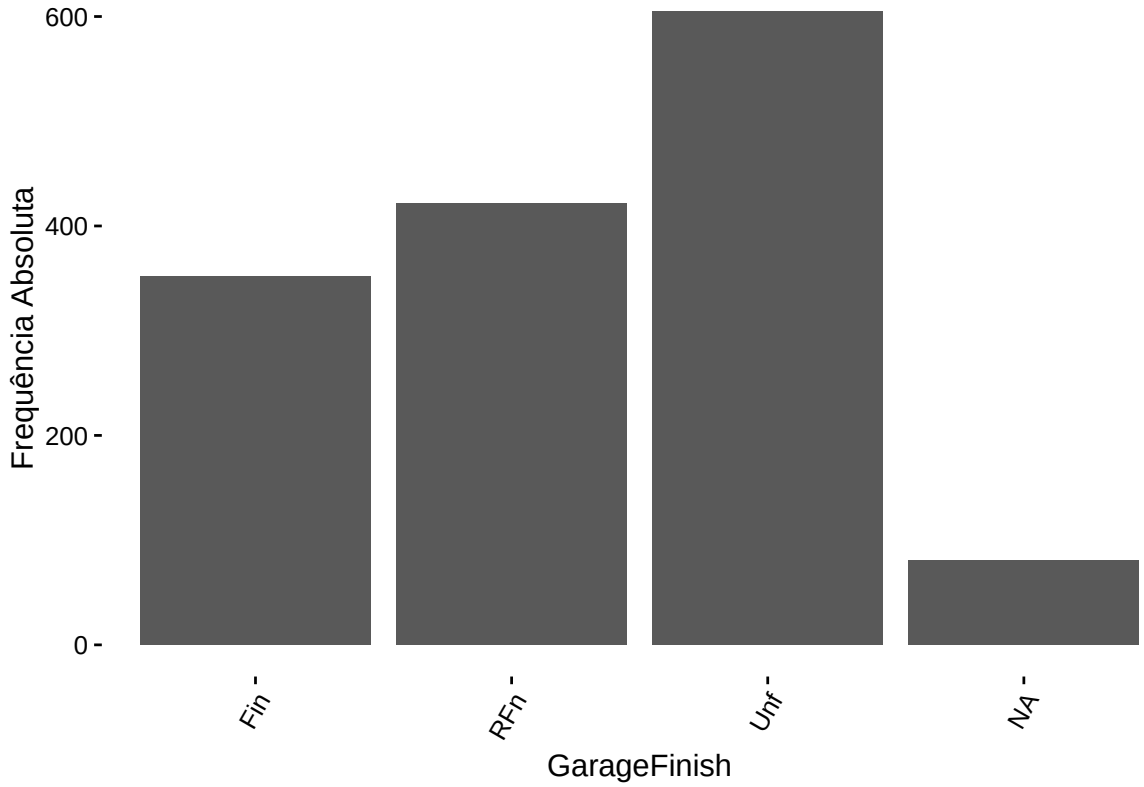

Frequência Absoluta
1000
500
0
FuseA
FuseF
FuseP
Mix
SBrkr
NA
Electrical

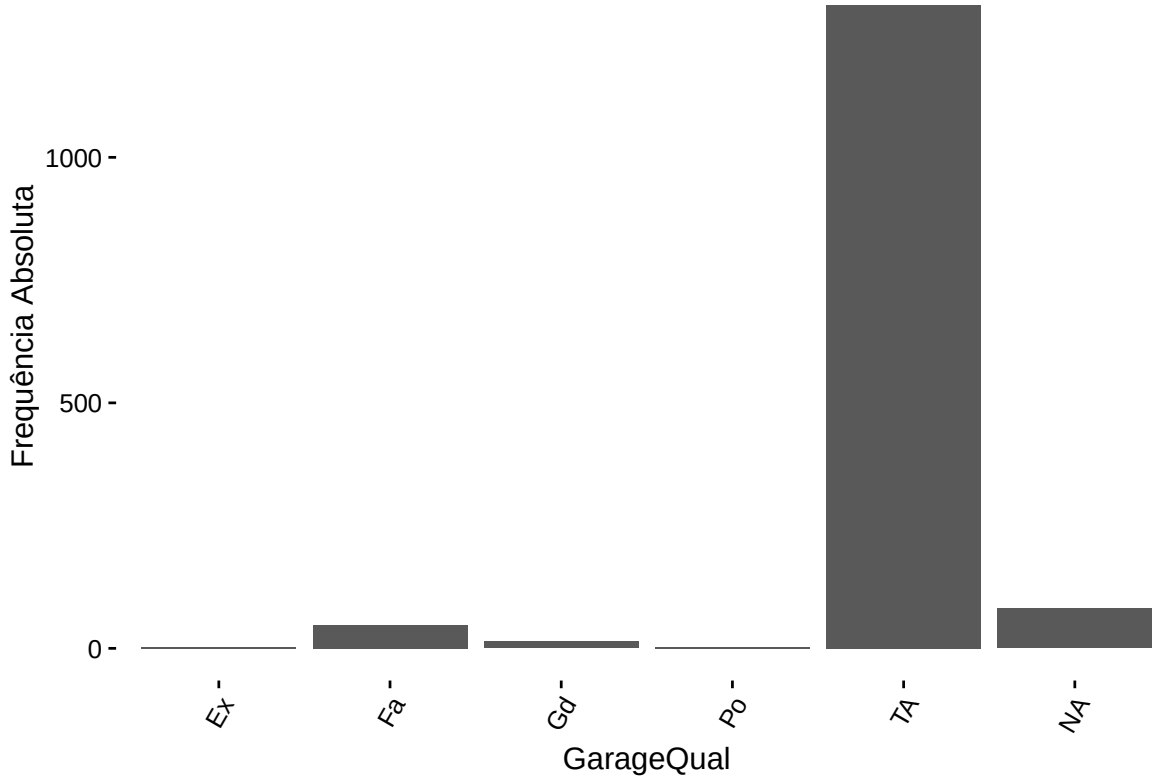

Frequência Absoluta
600
400
200
0
Ex
Fa
Gd
TA
KitchenQual

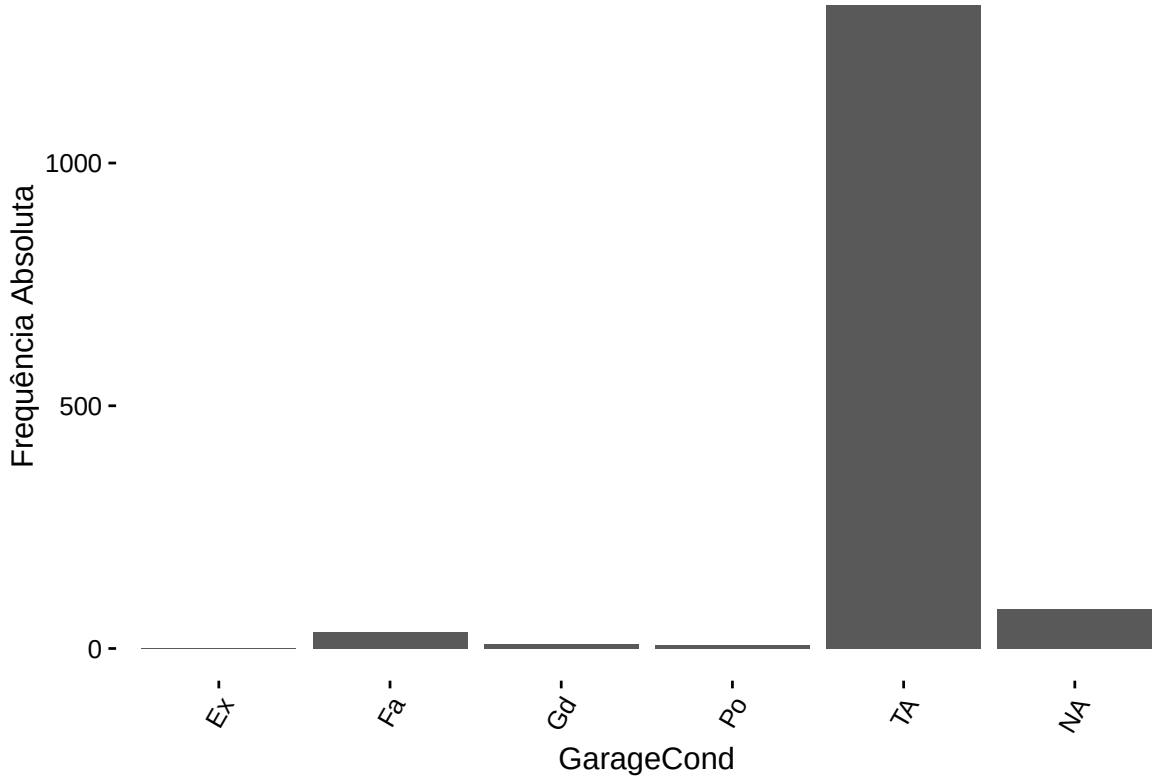

Frequência Absoluta
1000
500
0
Maj1
Maj2
Min1
Min2
Mod
Sev
Typ
Functional

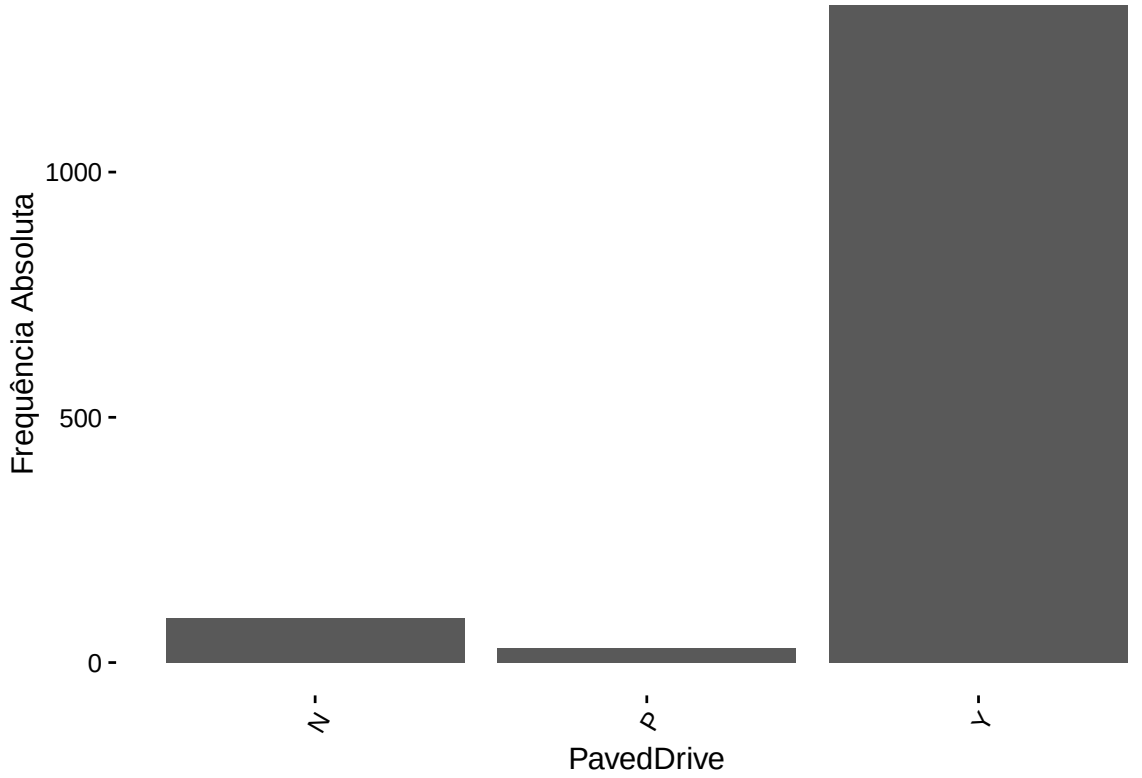

Frequência Absoluta
600
400
200
0
Ex
Fa
Gd
Po
TA
NA
FireplaceQu

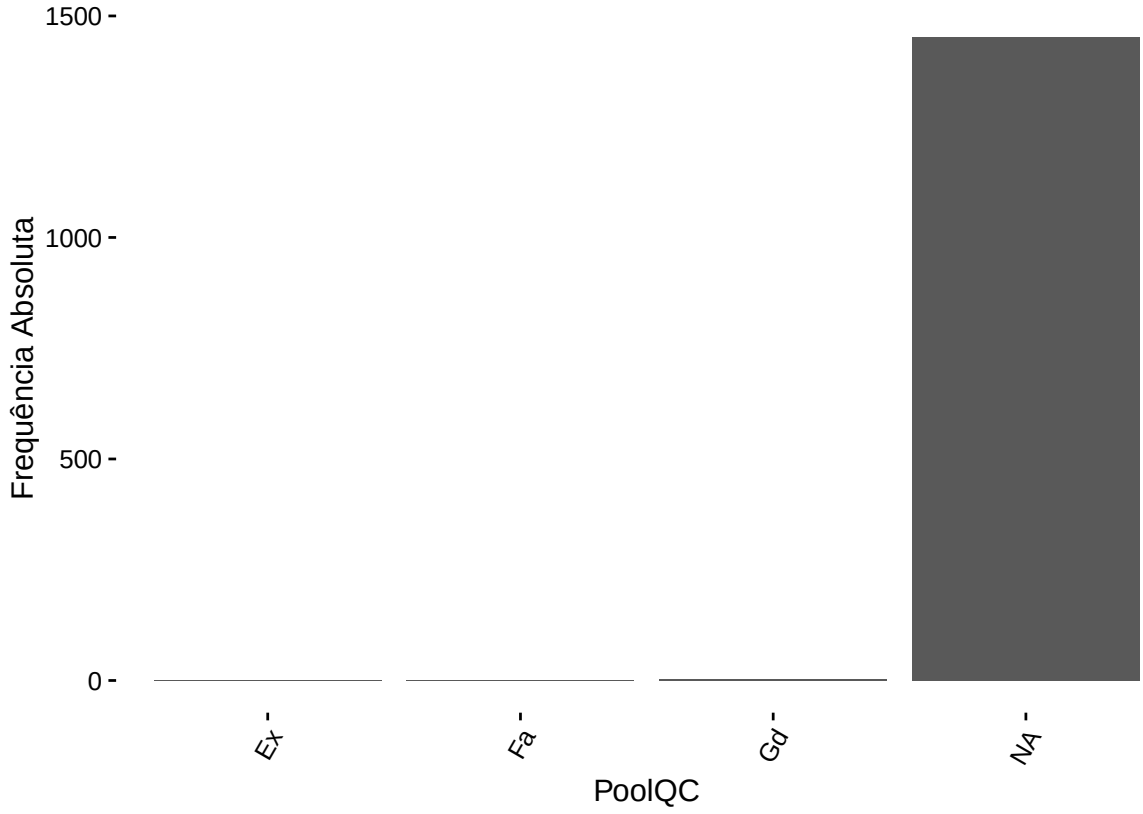

Frequência Absoluta
750
500
250
0
2Types
Attchd
Basment
BuiltIn
CarPort
Detchd
NA
GarageType

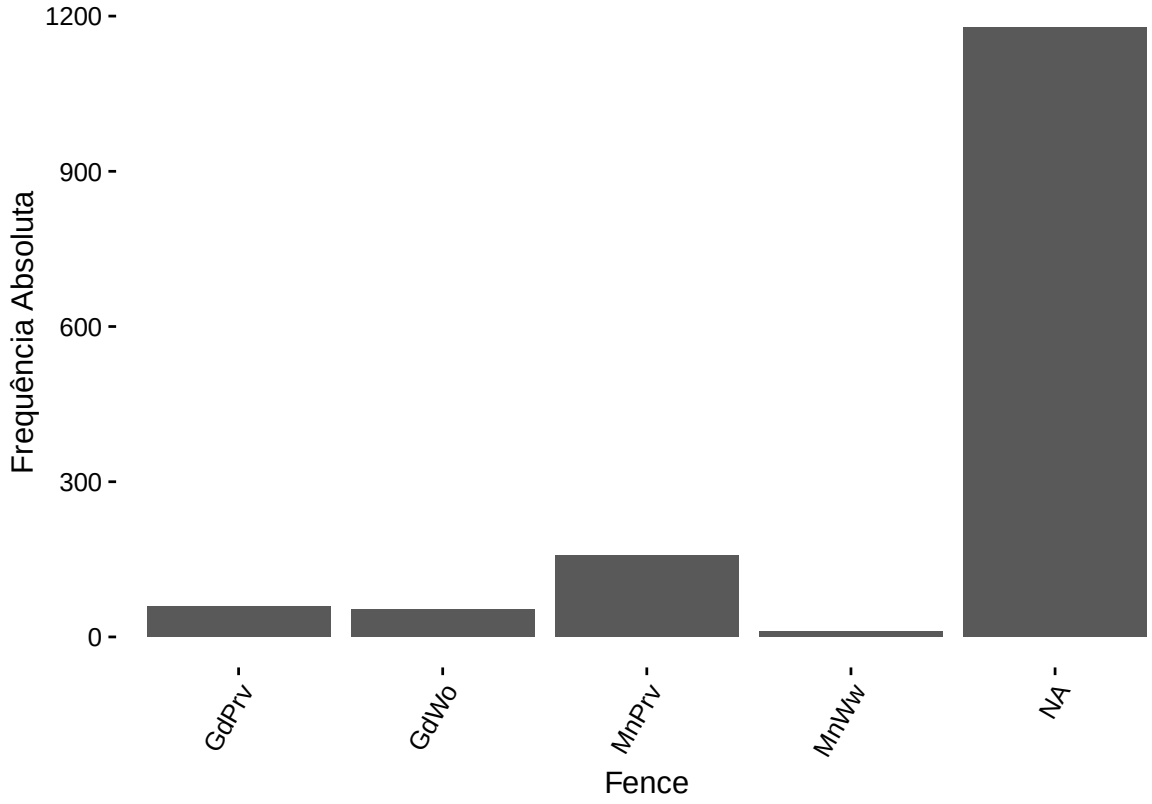

Frequência Absoluta
600
400
200
0
Fin
RFn
Unf
NA
GarageFinish

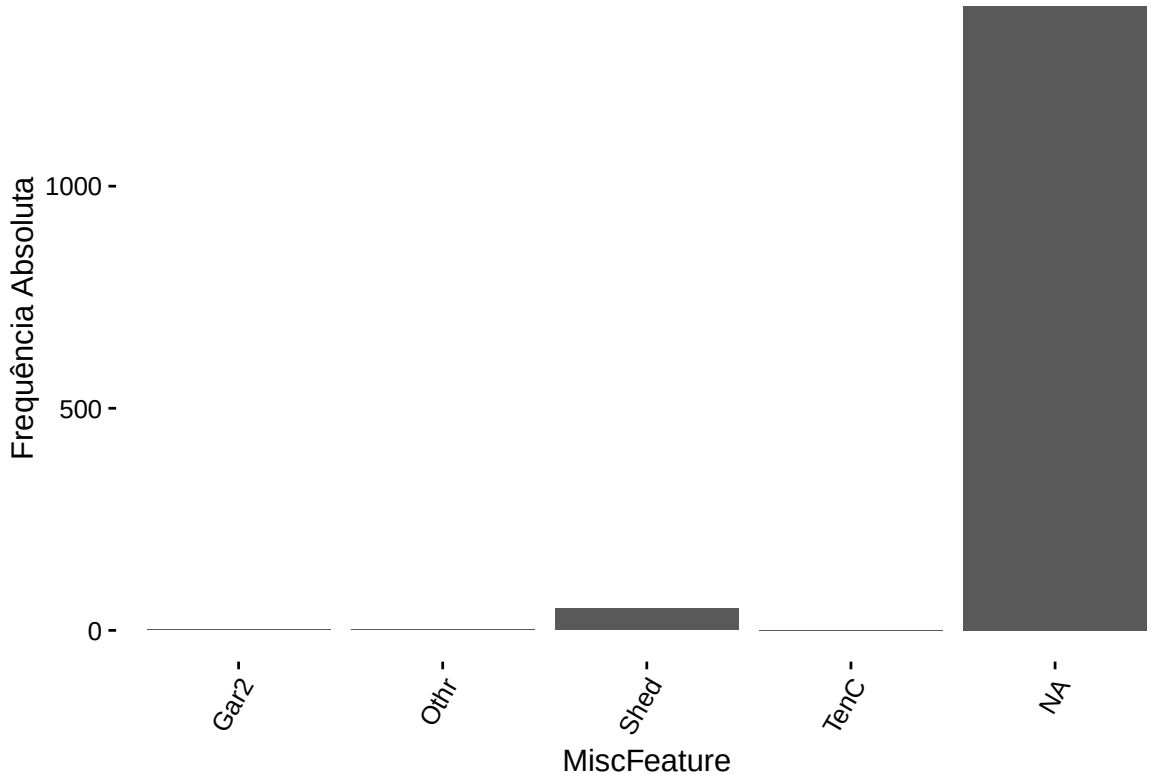

Frequência Absoluta
1000
500
0
Ex
Fa
Gd
Po
TA
NA
GarageQual

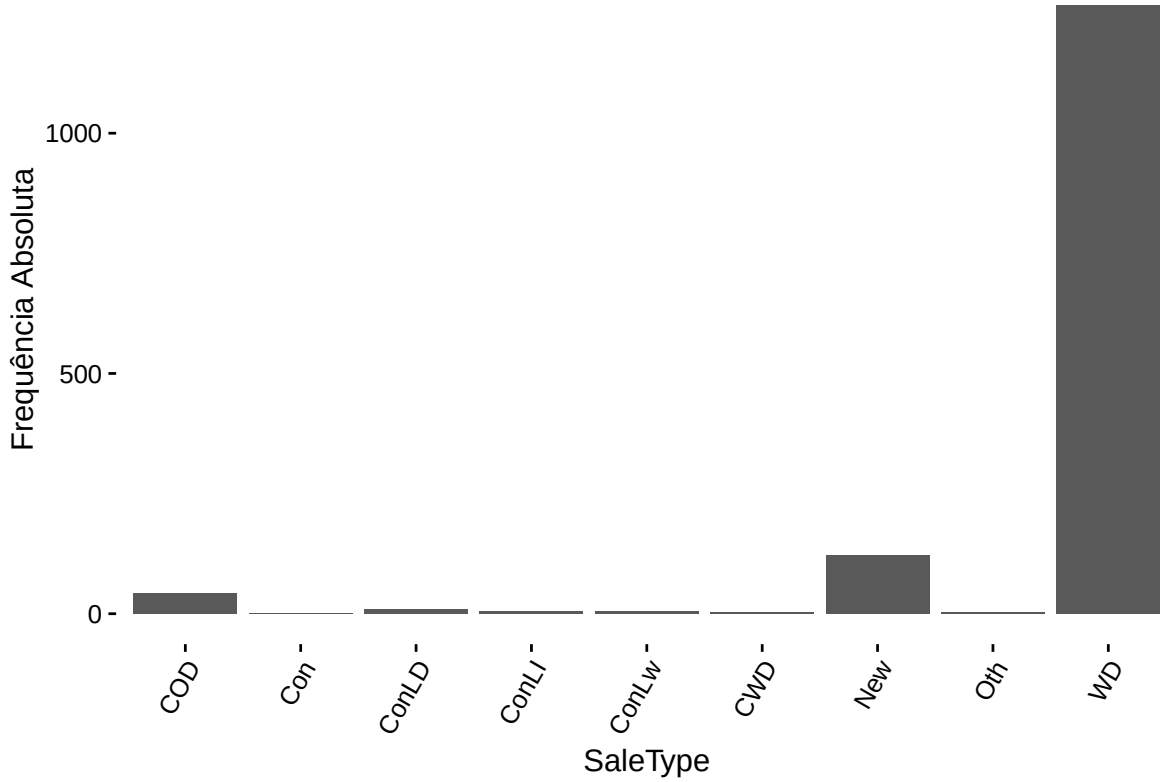

Frequência Absoluta
1000
500
0
Ex
Fa
Gd
Po
TA
NA
GarageCond

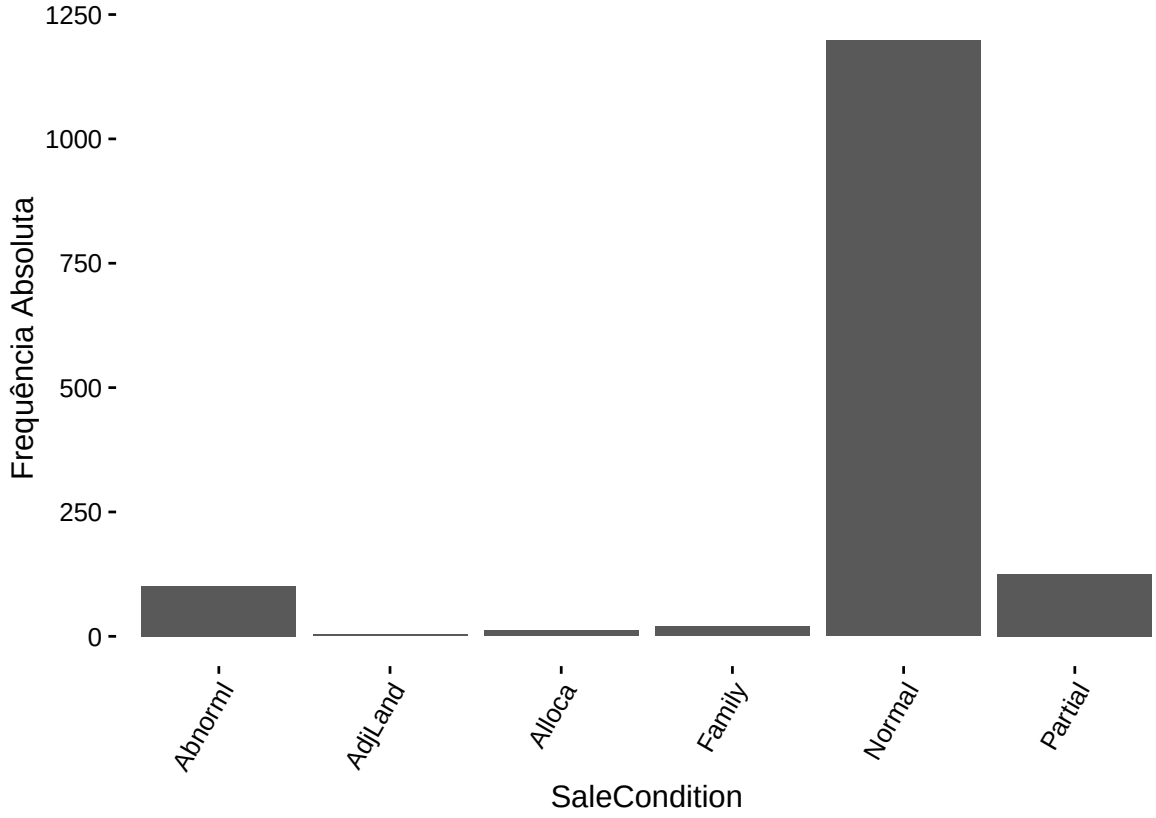

Frequência Absoluta
1000
500
0
N
P
Y
PavedDrive


Frequência Absoluta
1500
1000
500
0
Ex
Fa
Gd
NA
PoolQC


1200
900
600
300
0
Frequência Absoluta
GdPrv
GdWo
MnPrv
MnWw
NA
Fence

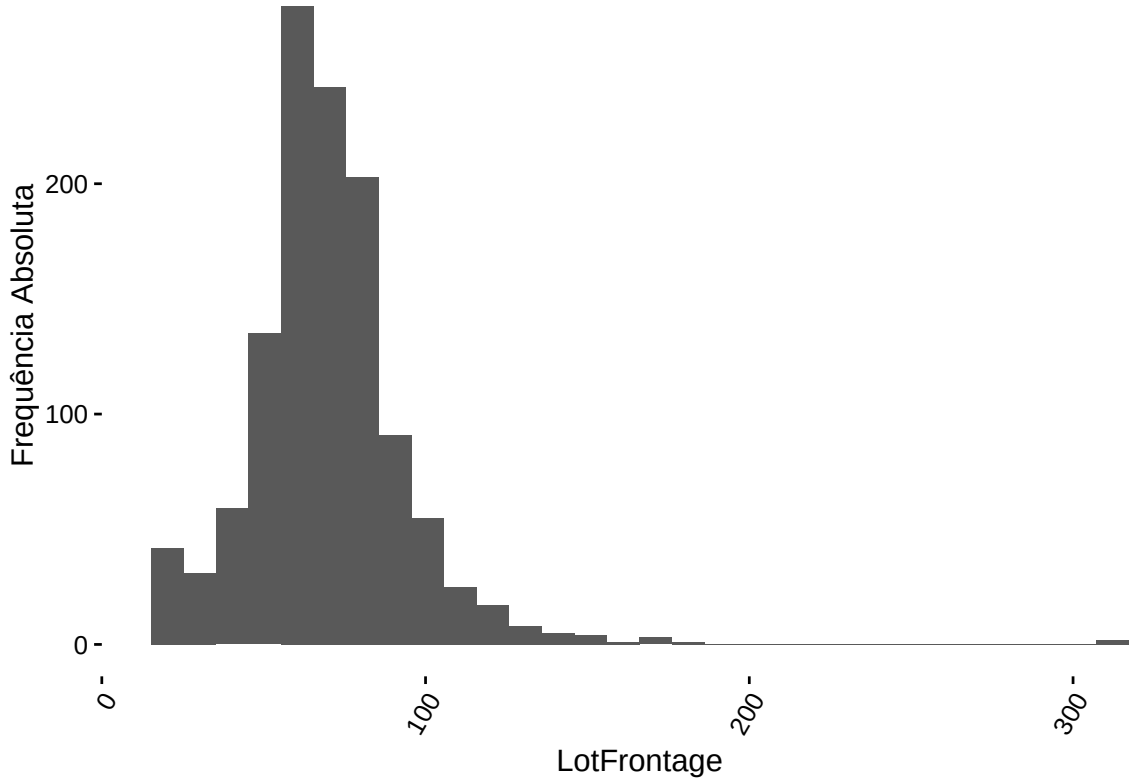

1000
500
0
Frequência Absoluta
Gar2
Othr
Shed
TenC
NA
MiscFeature



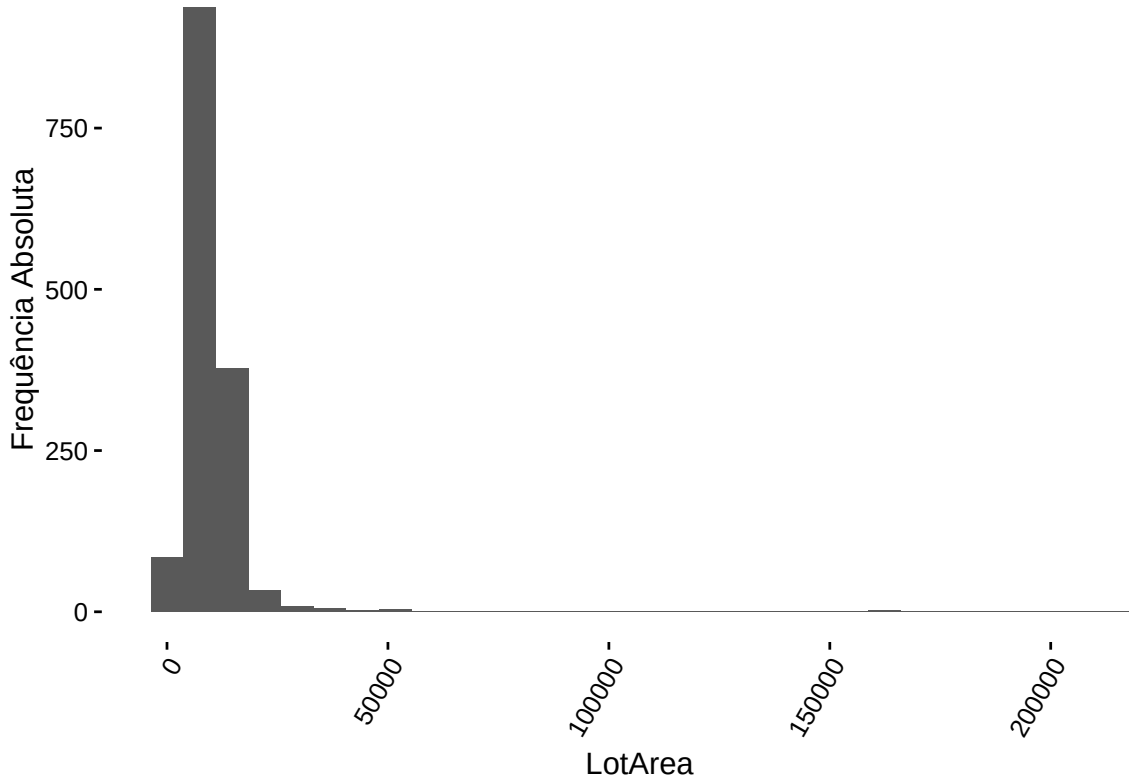


1.2 GRÁFICOS DAS VARIÁVEIS QUANTITATIVAS

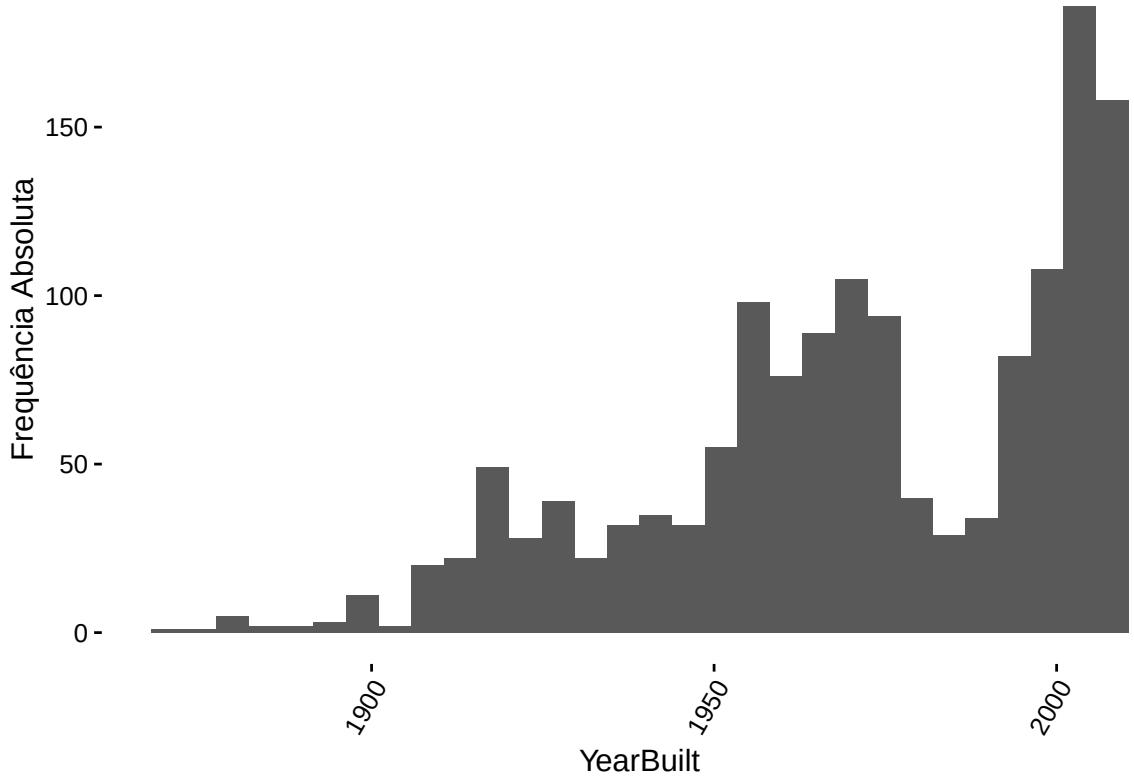
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



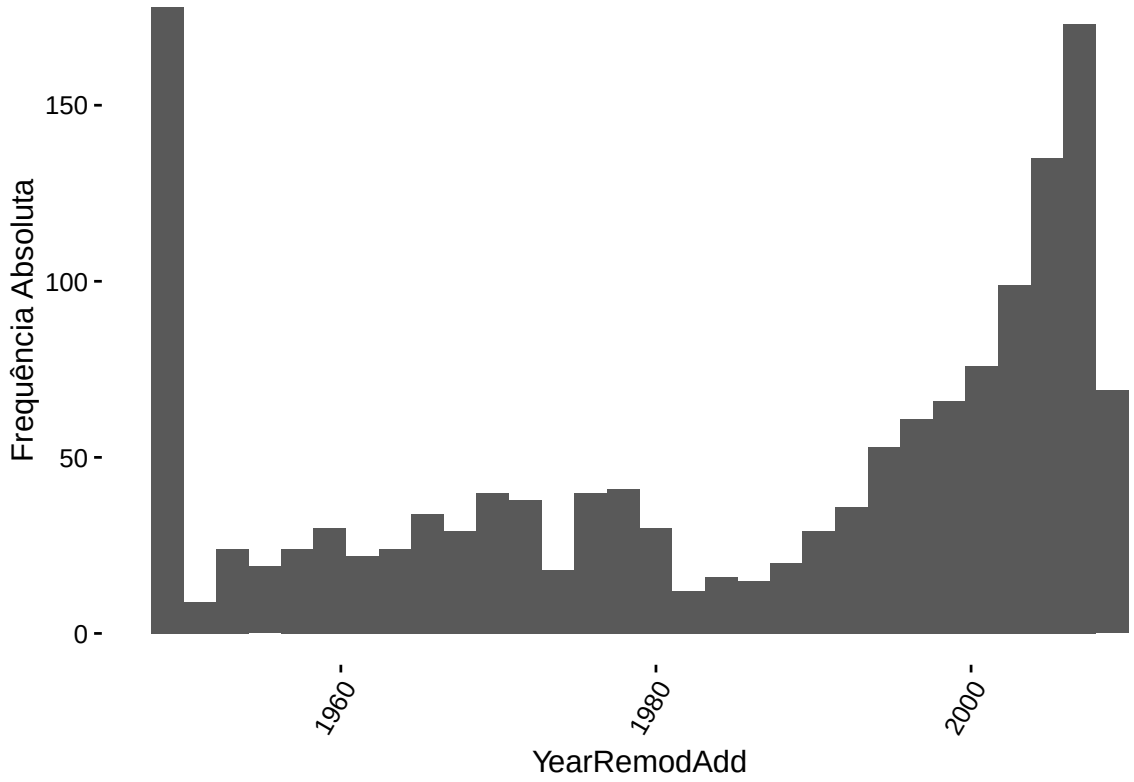

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



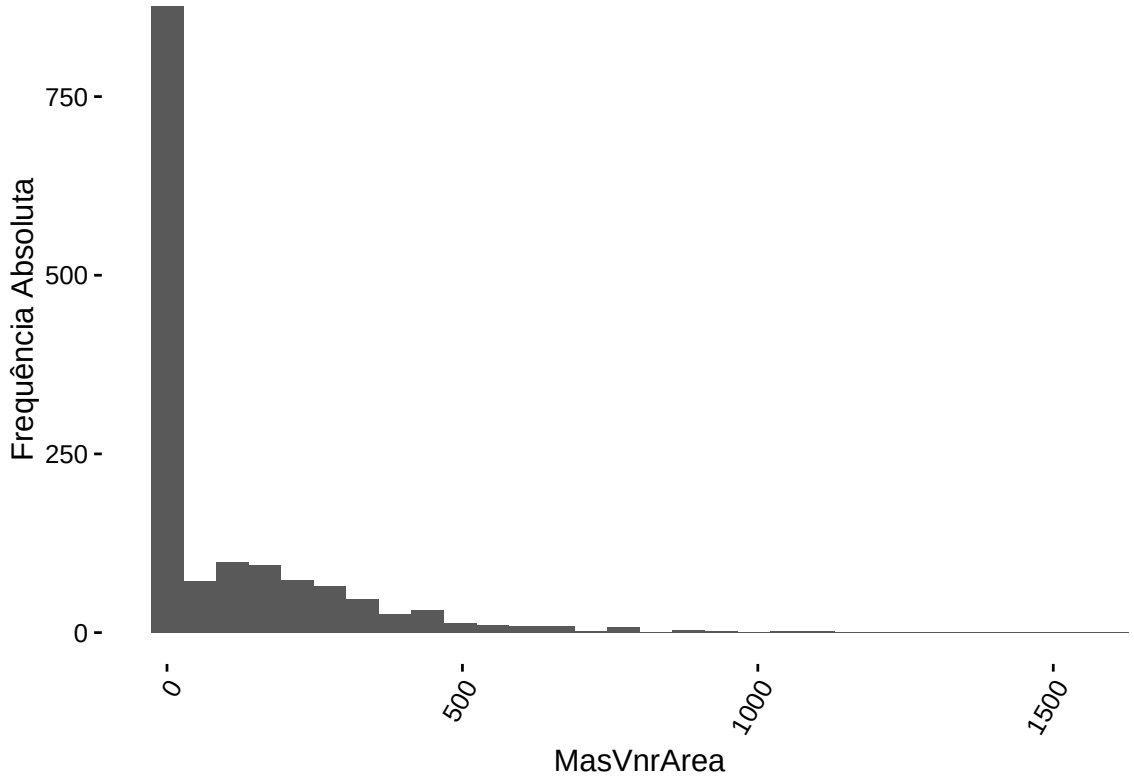

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



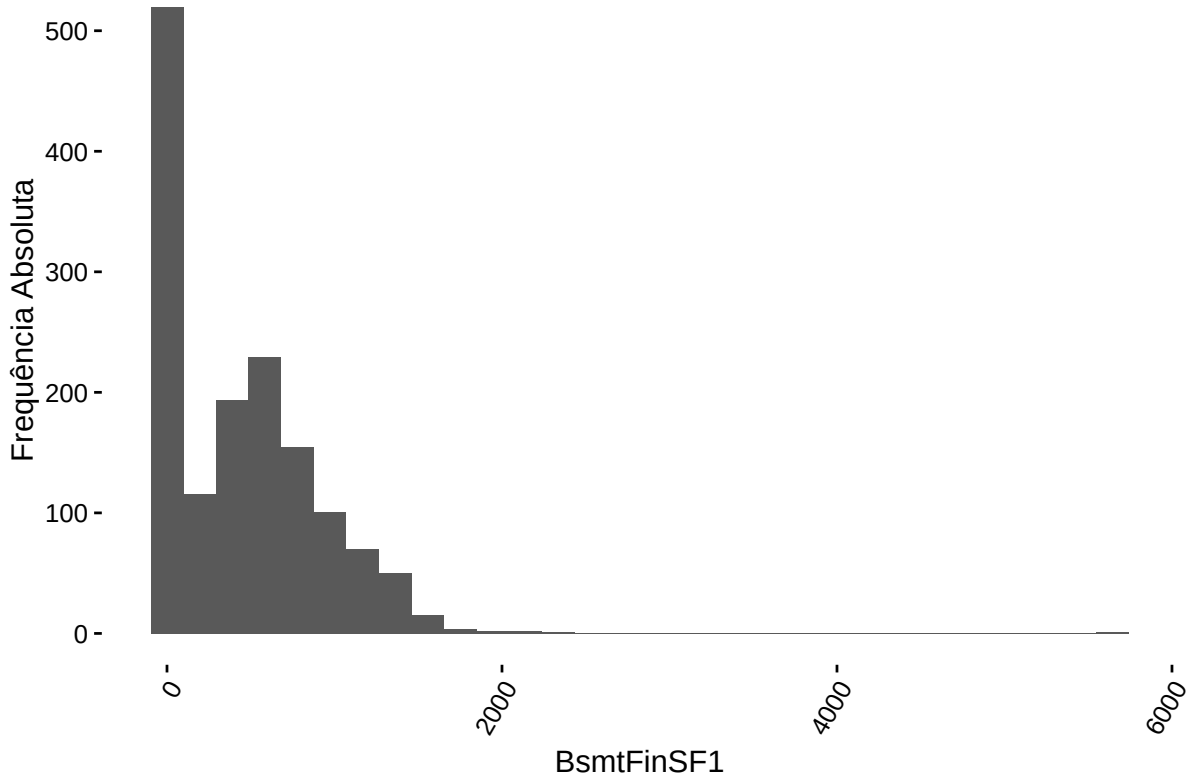

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



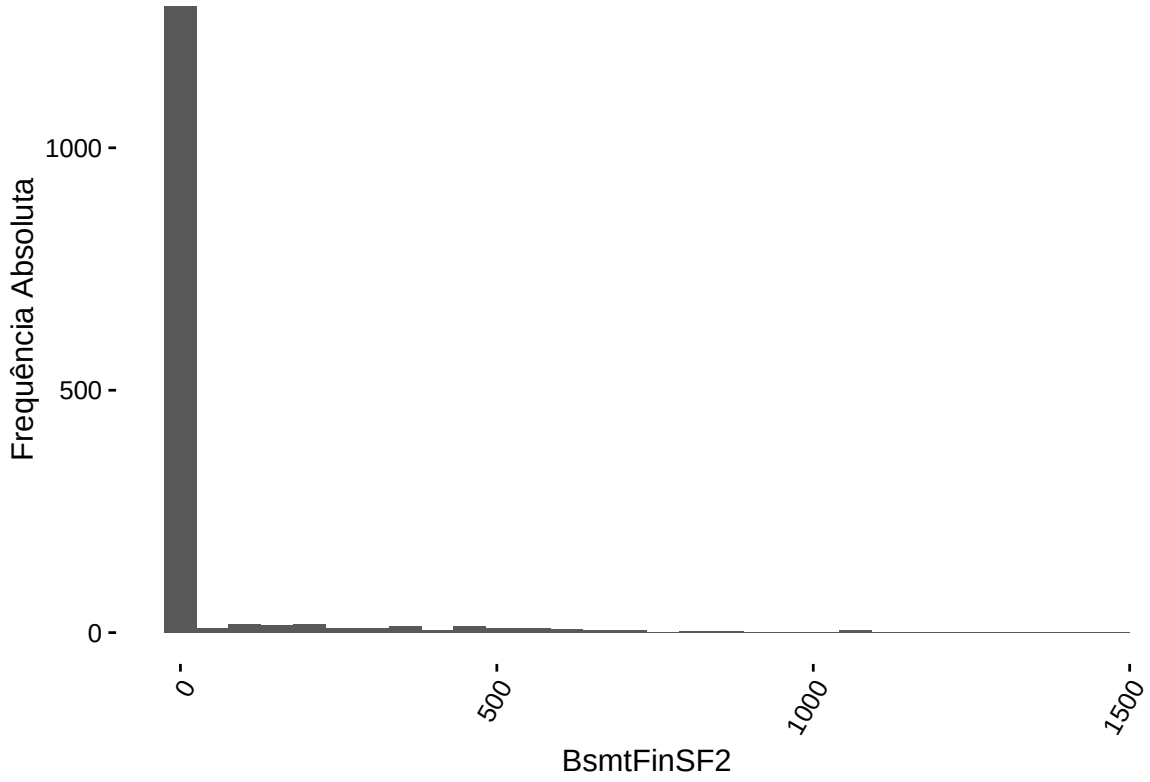

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



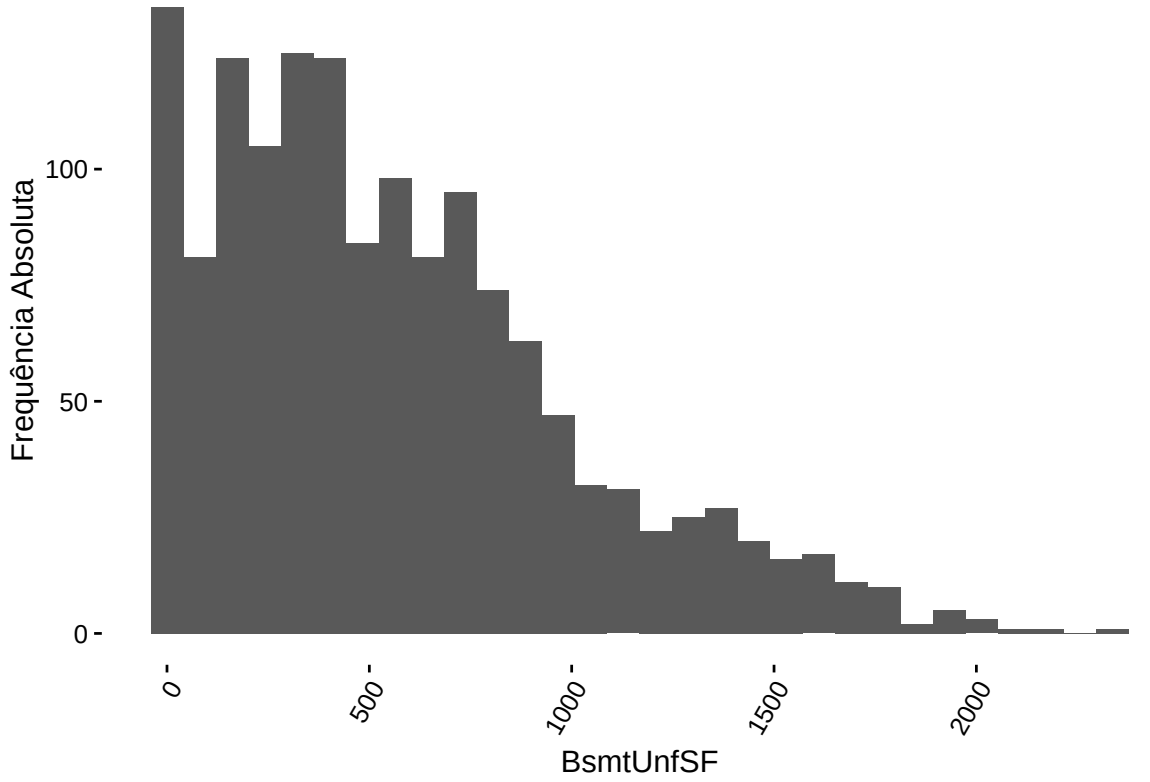

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



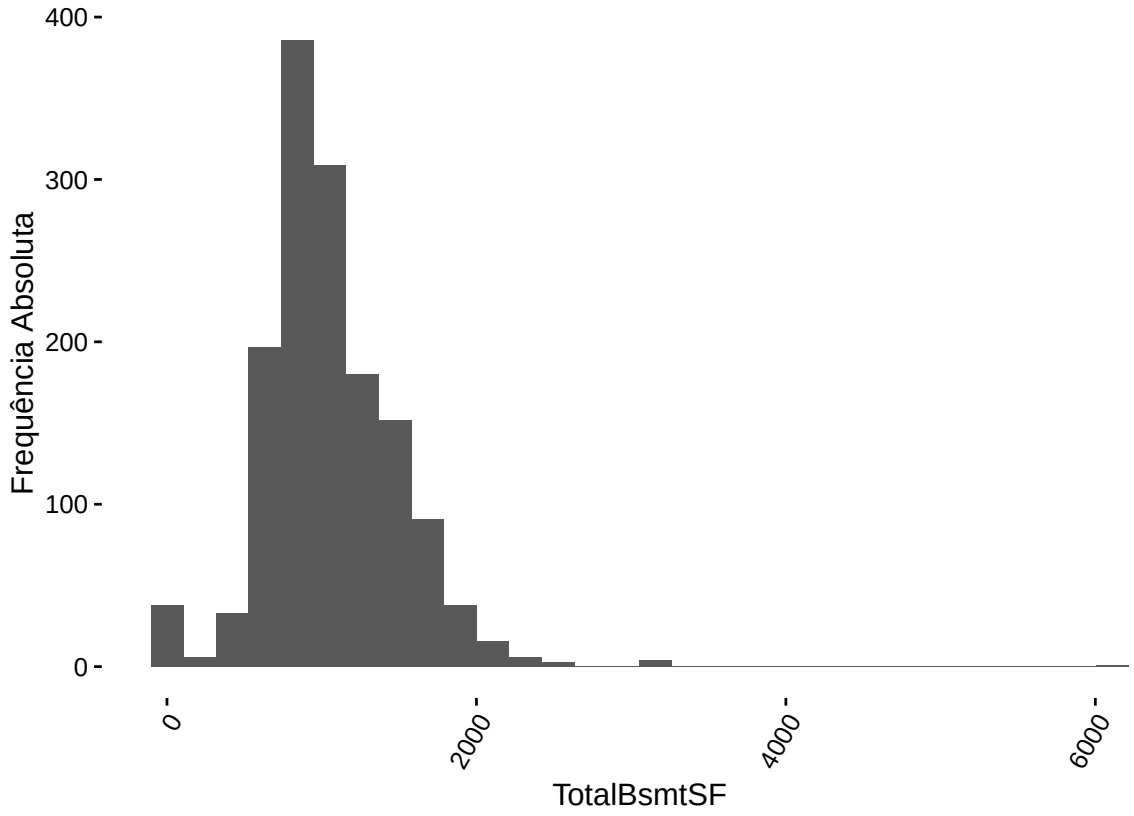

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



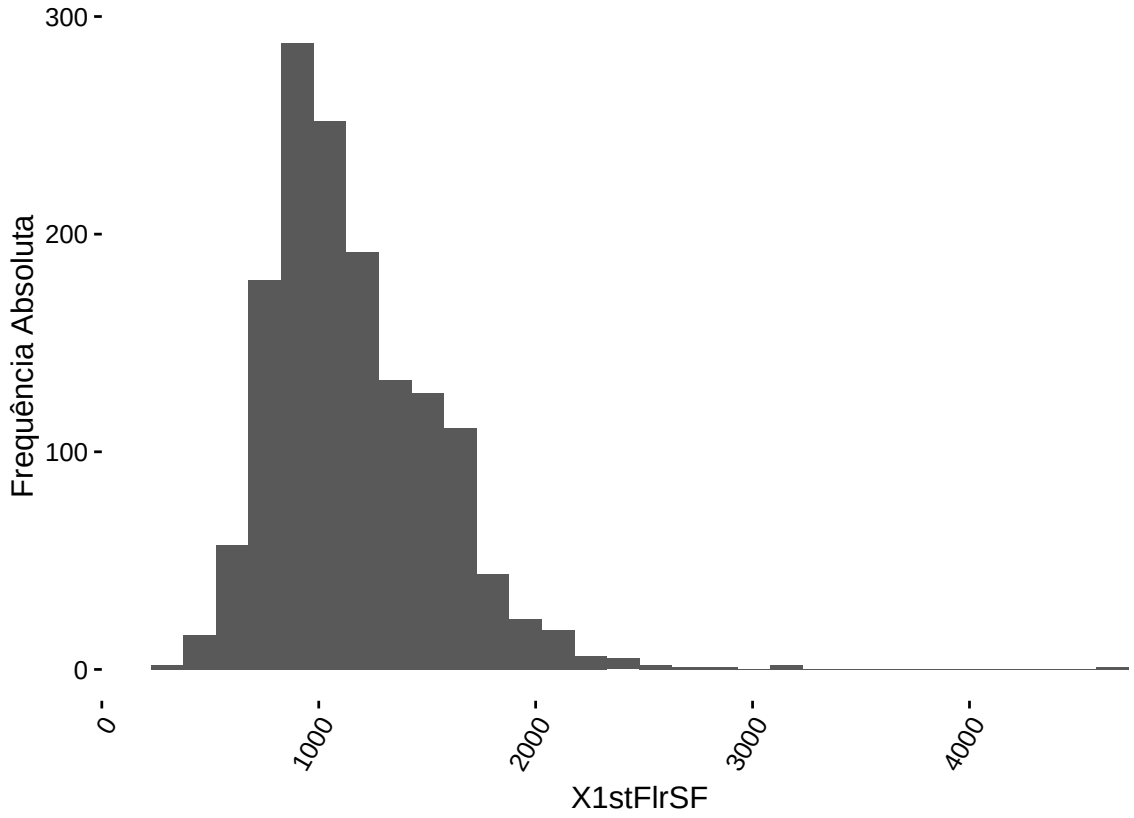

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



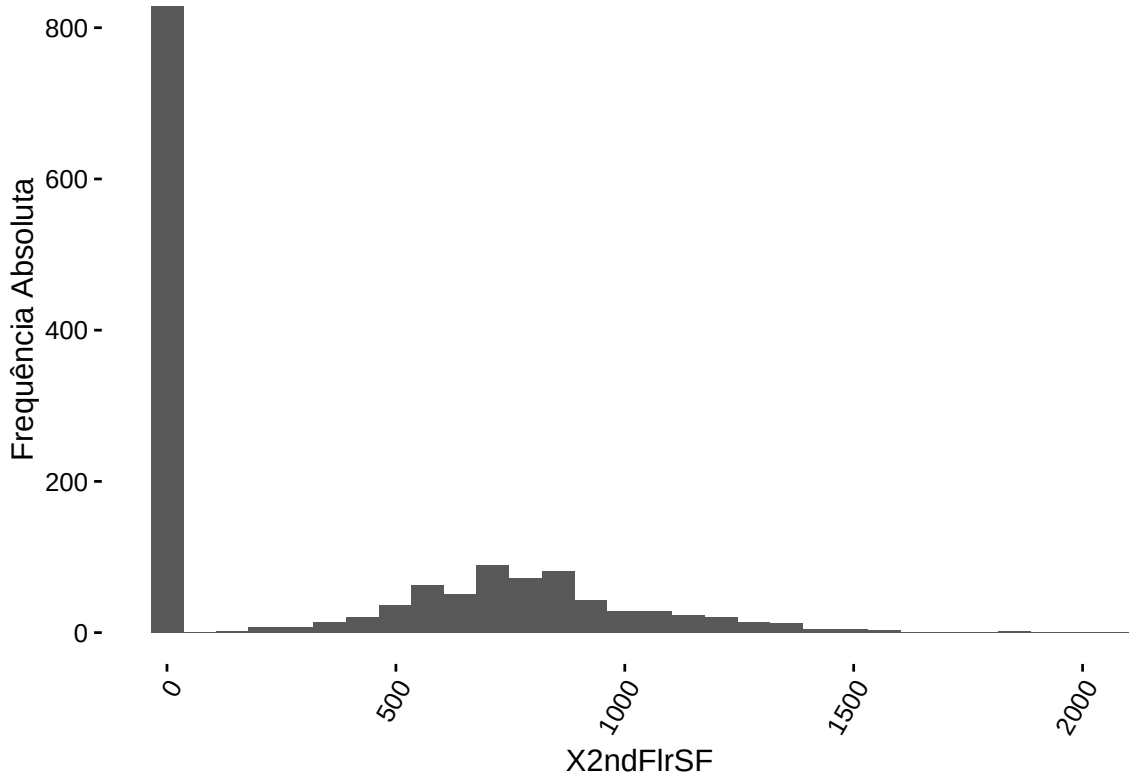

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



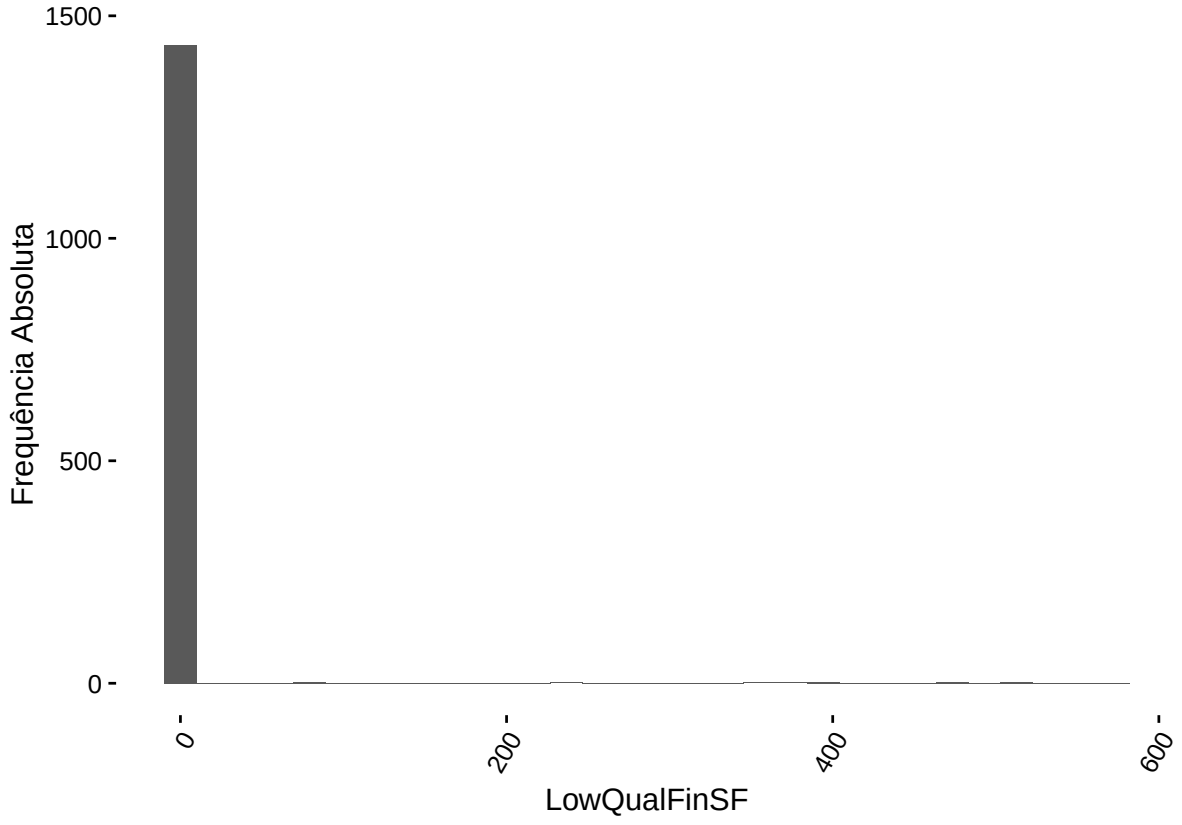

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

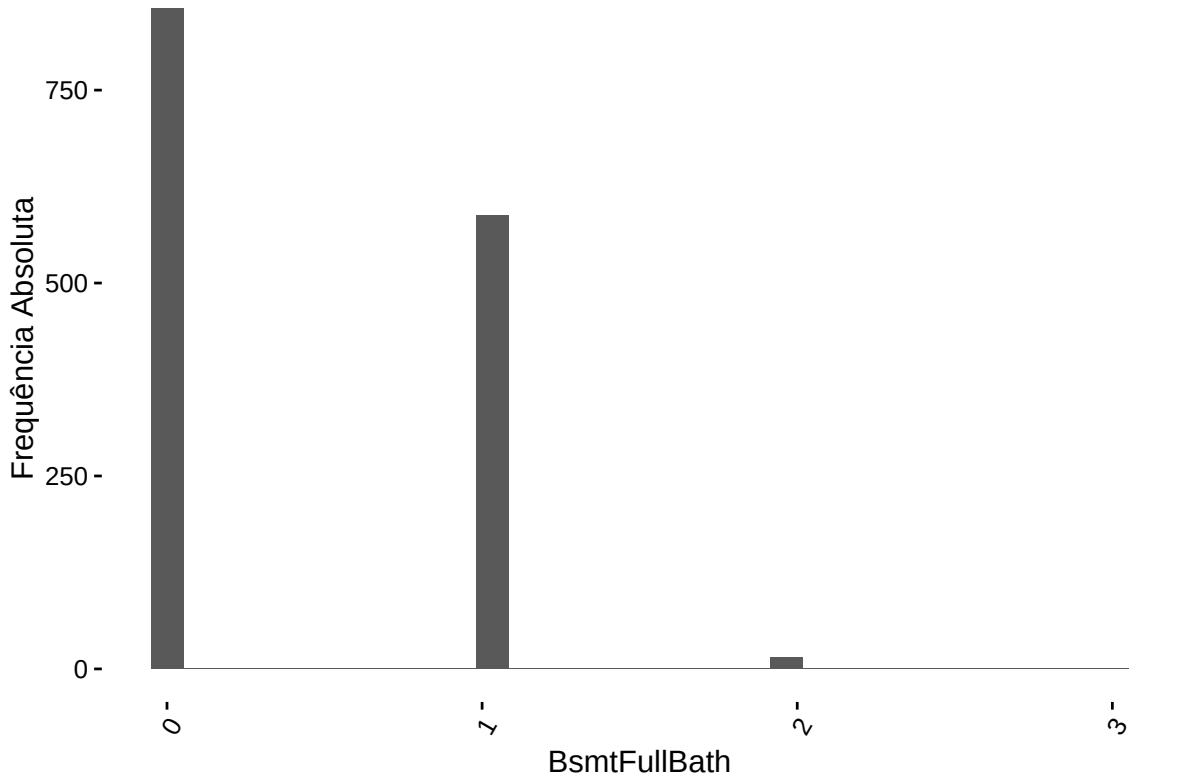



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

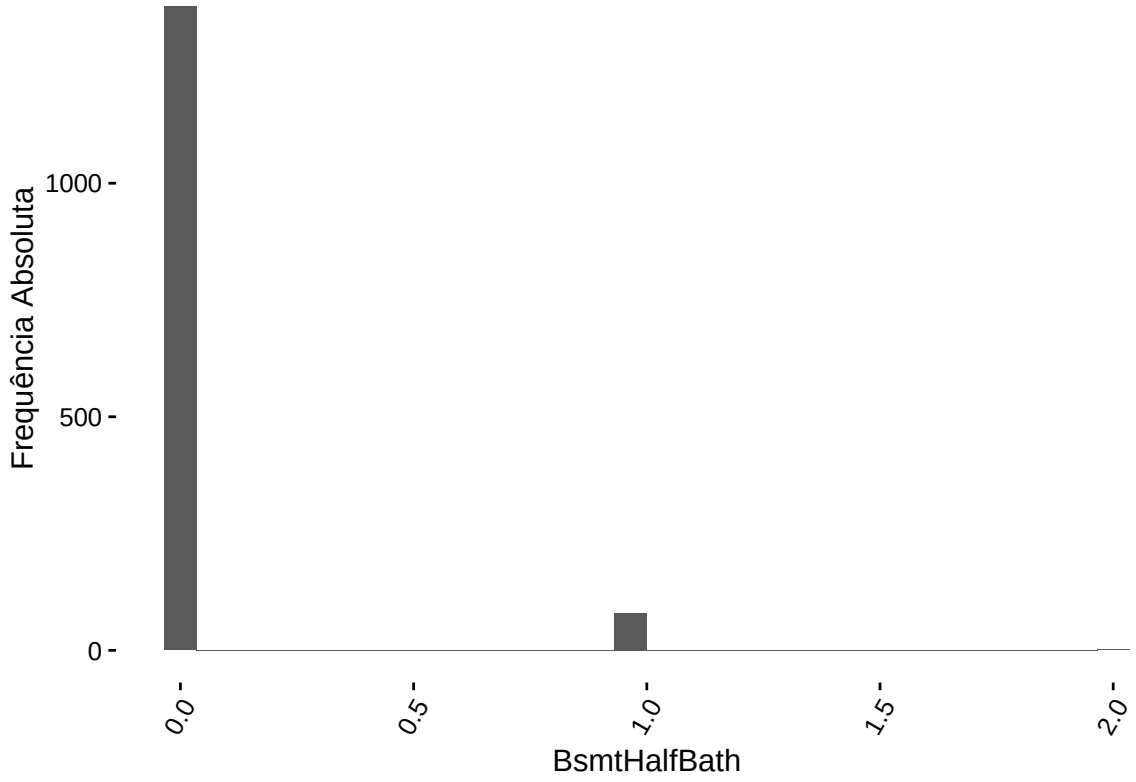



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

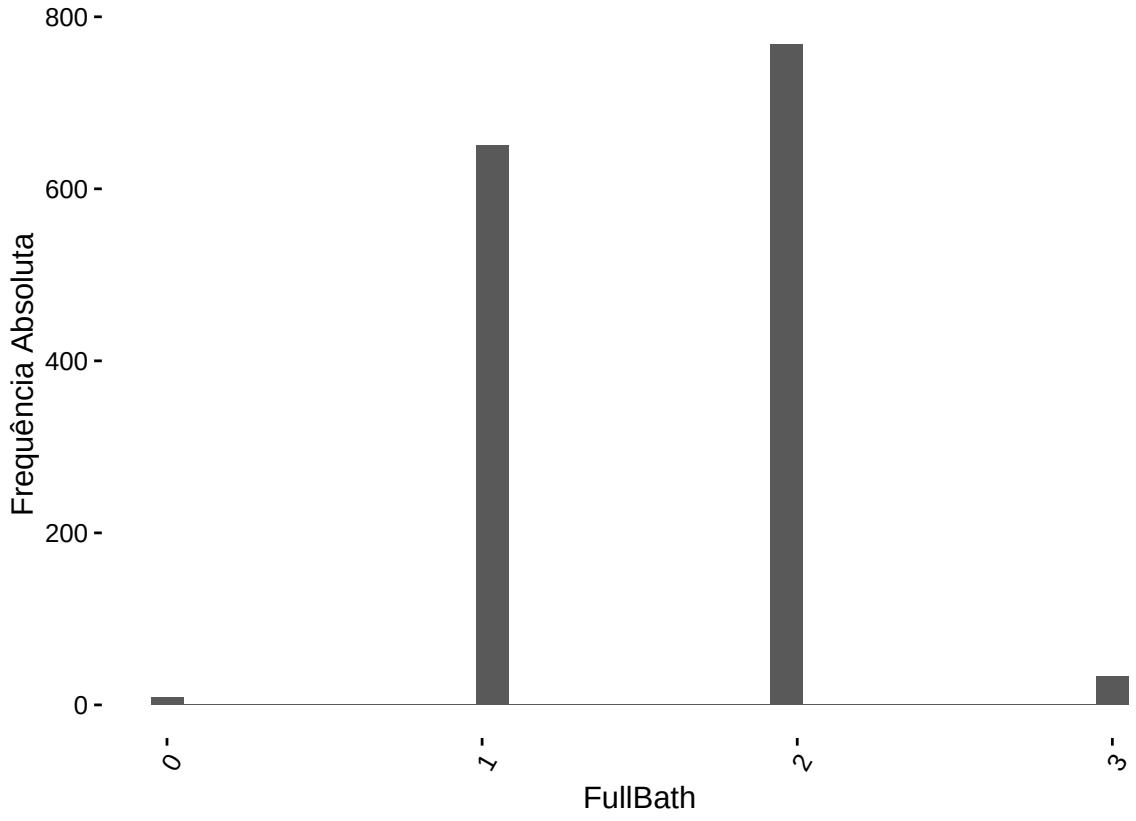

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



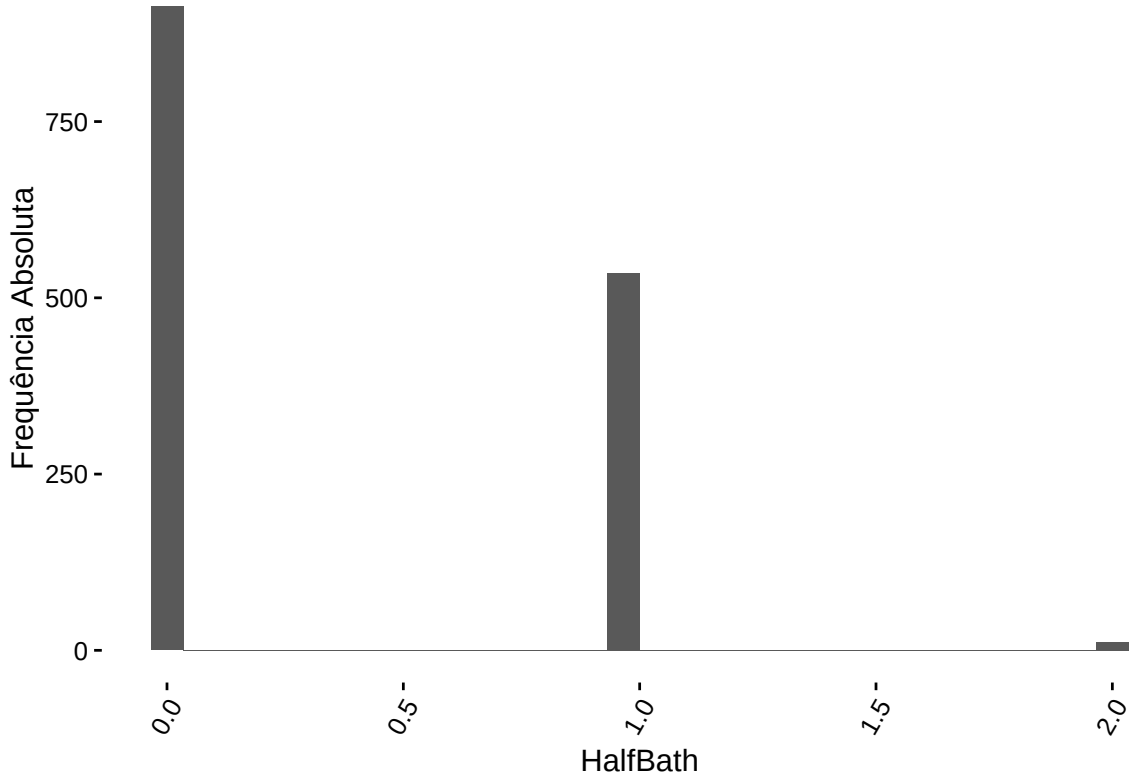

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



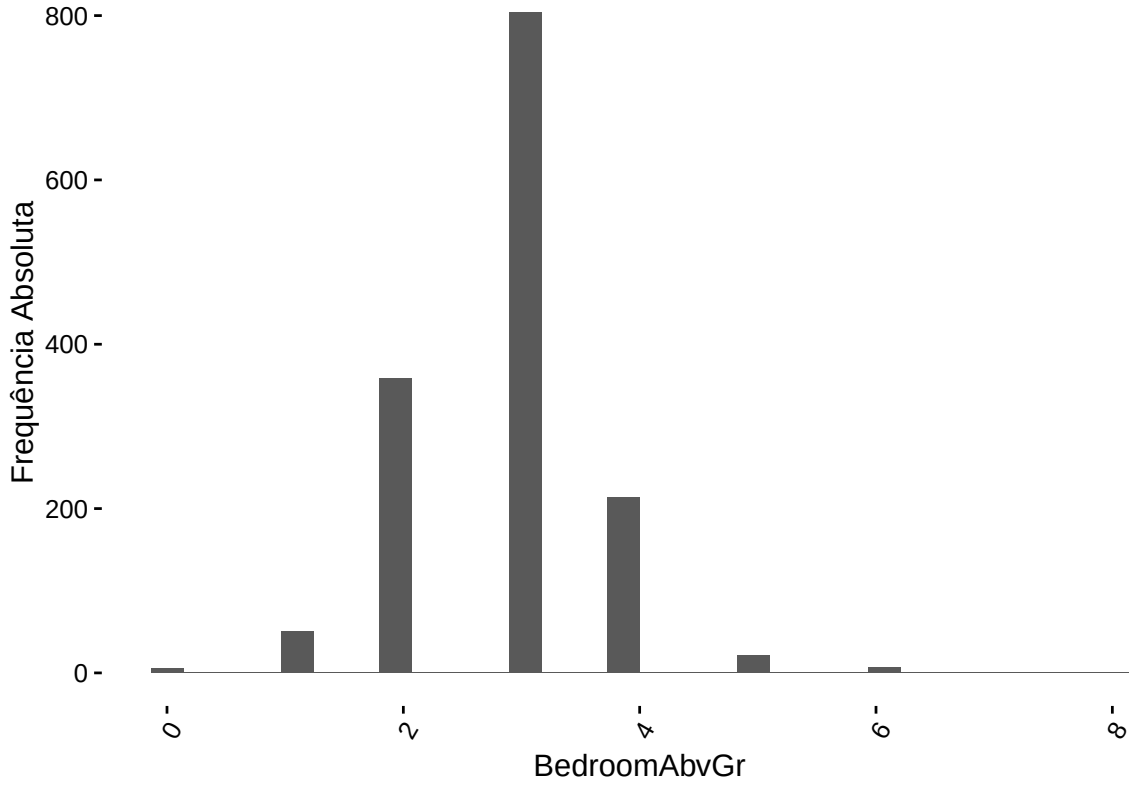

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



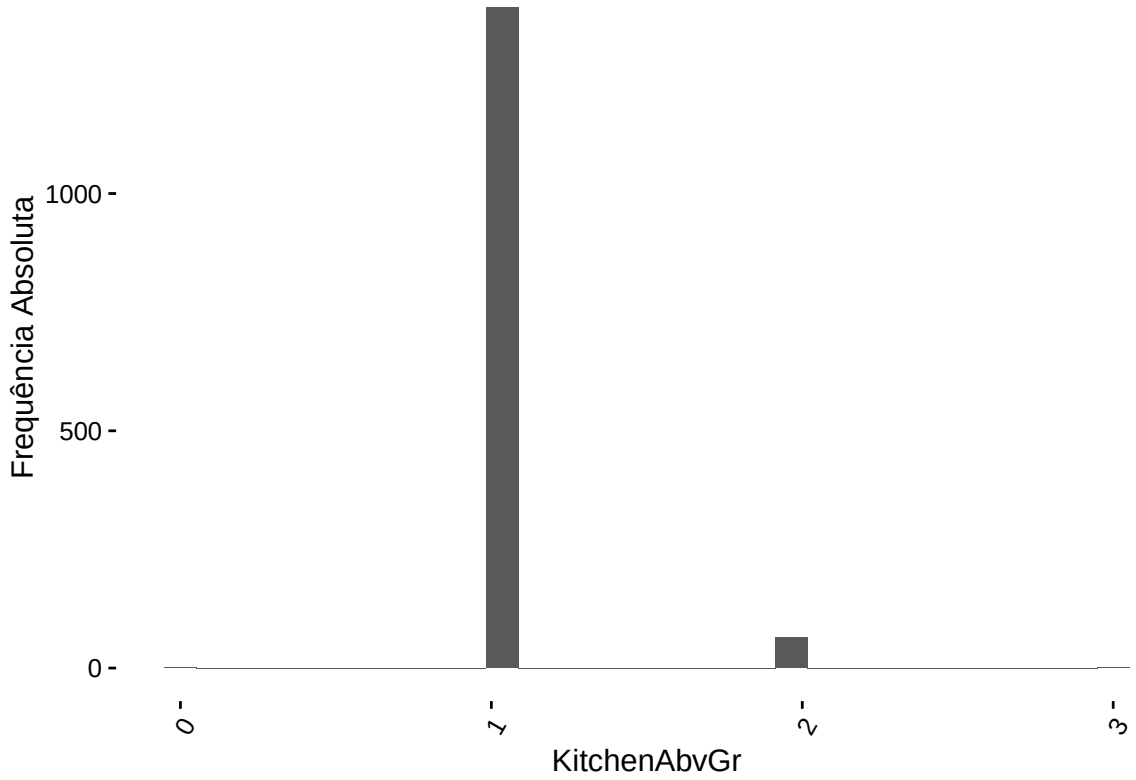

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



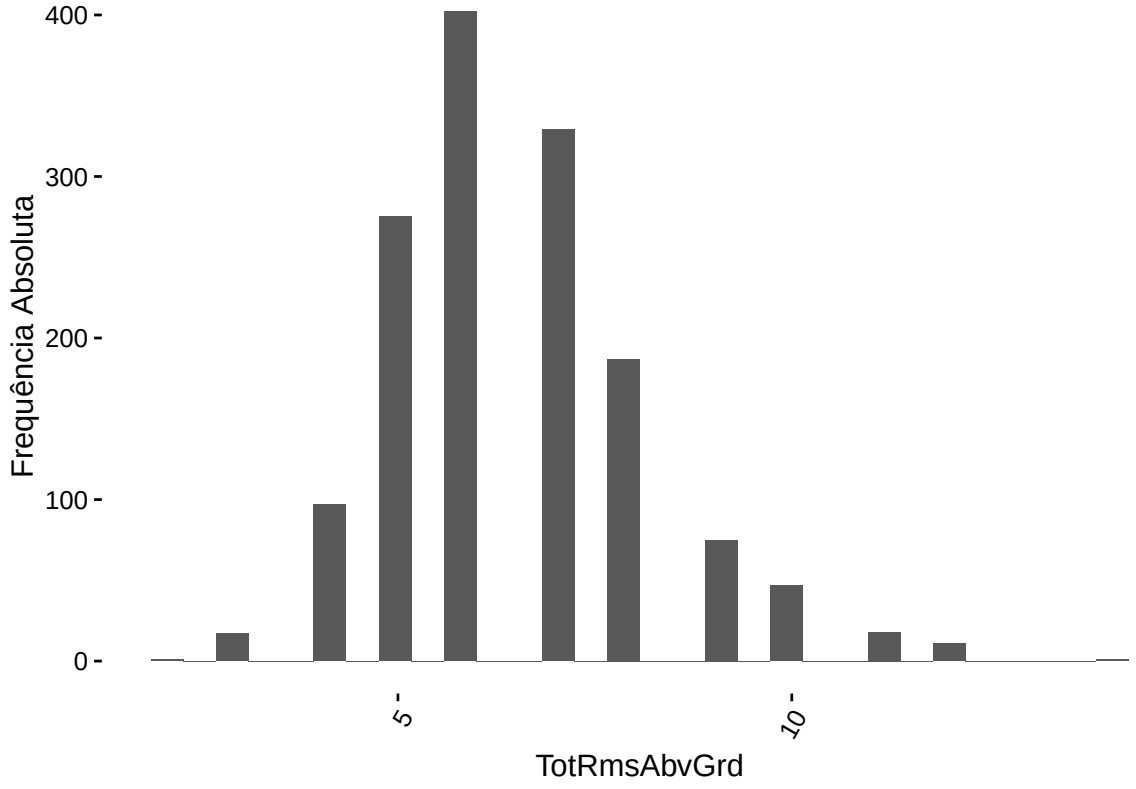

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



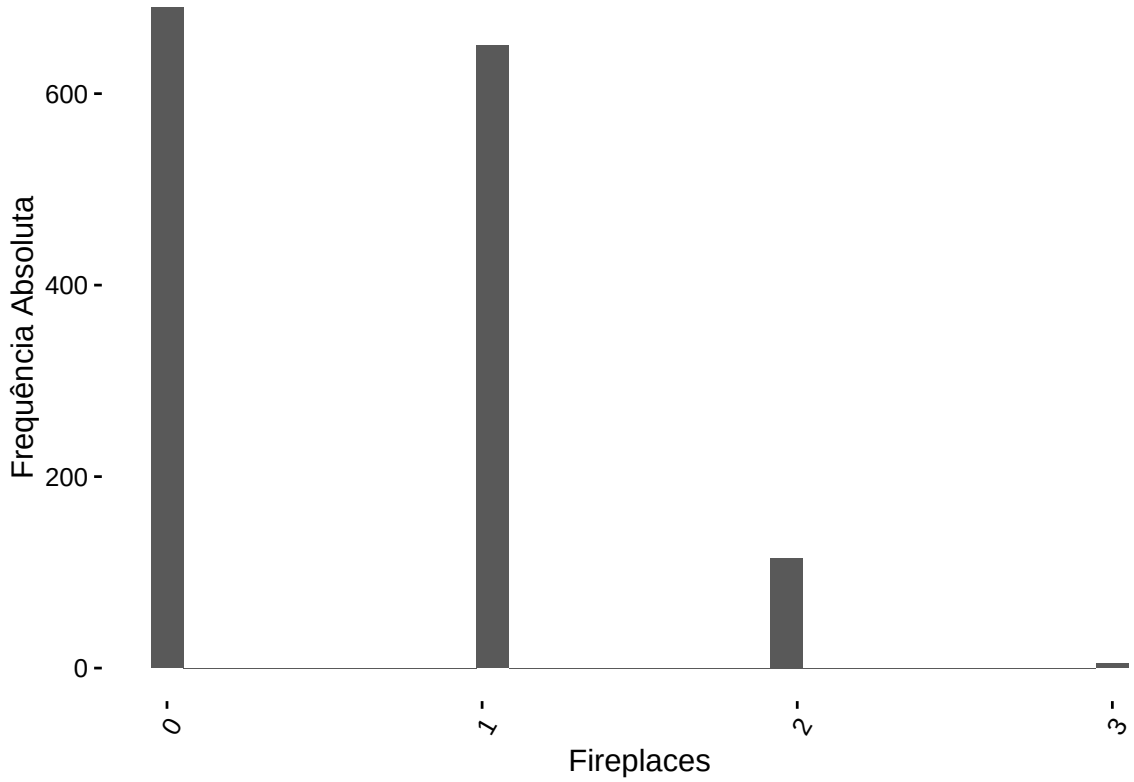

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



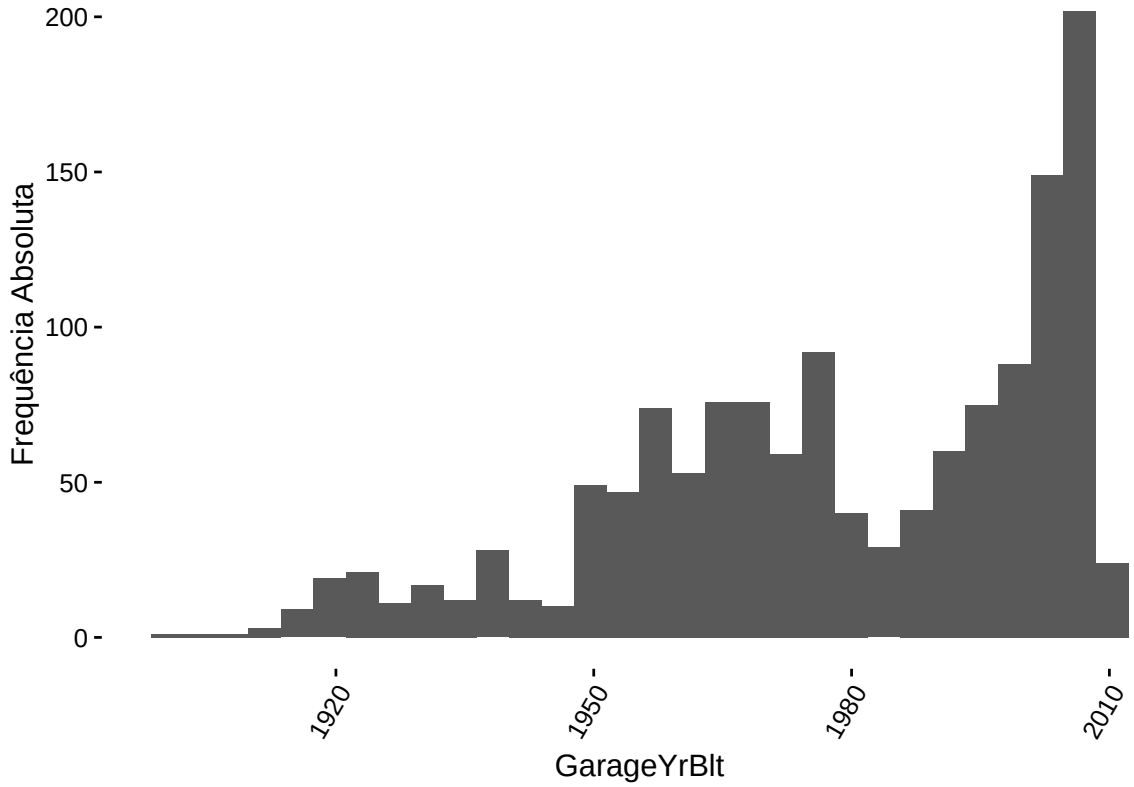

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



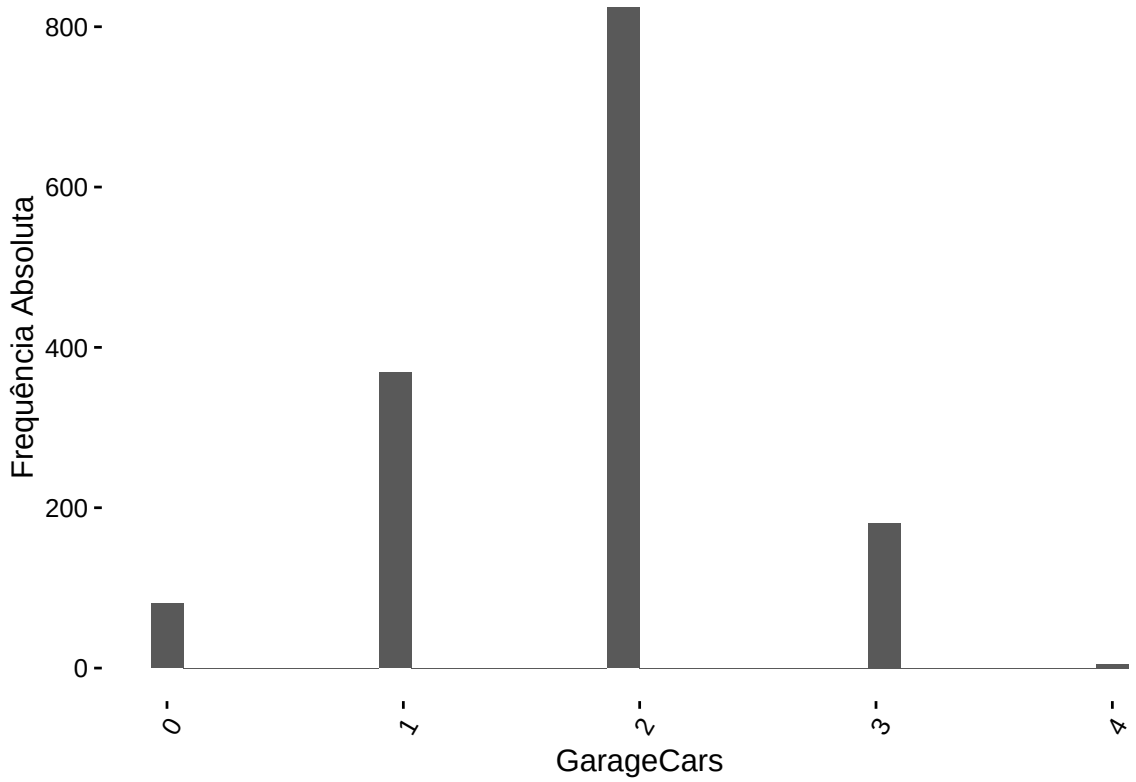

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



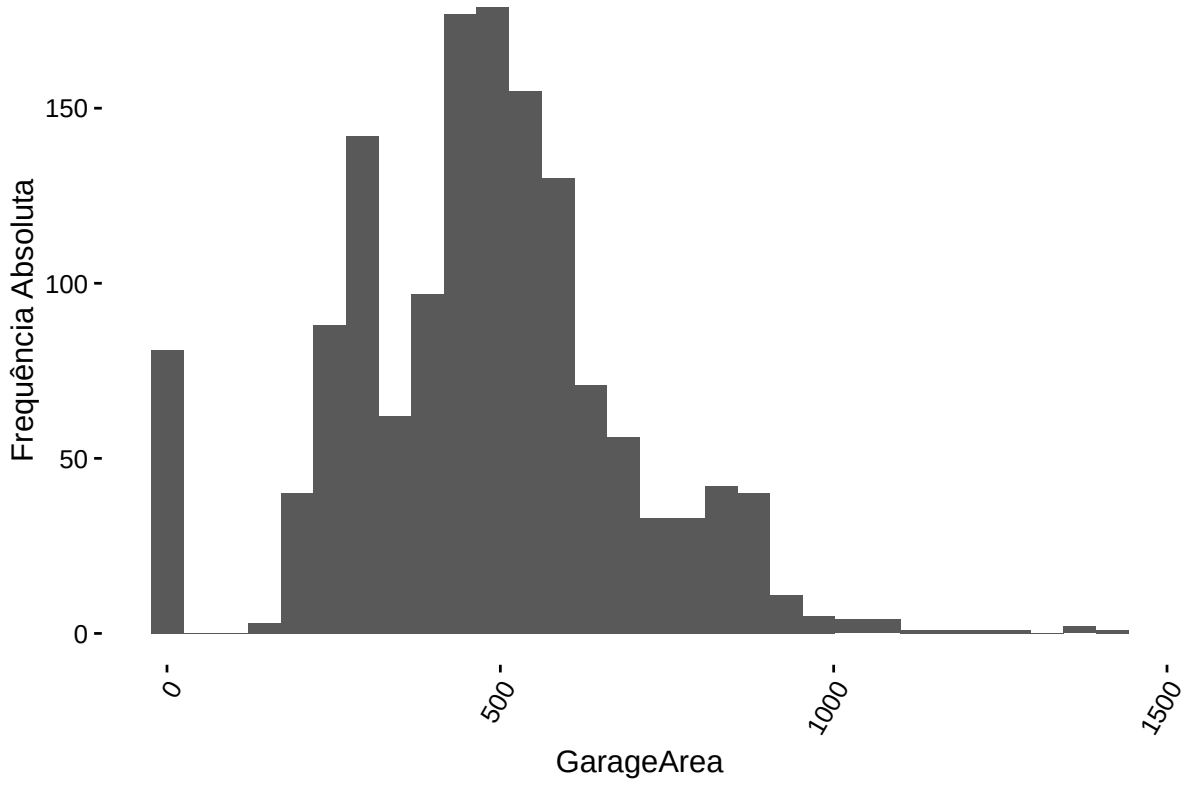

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



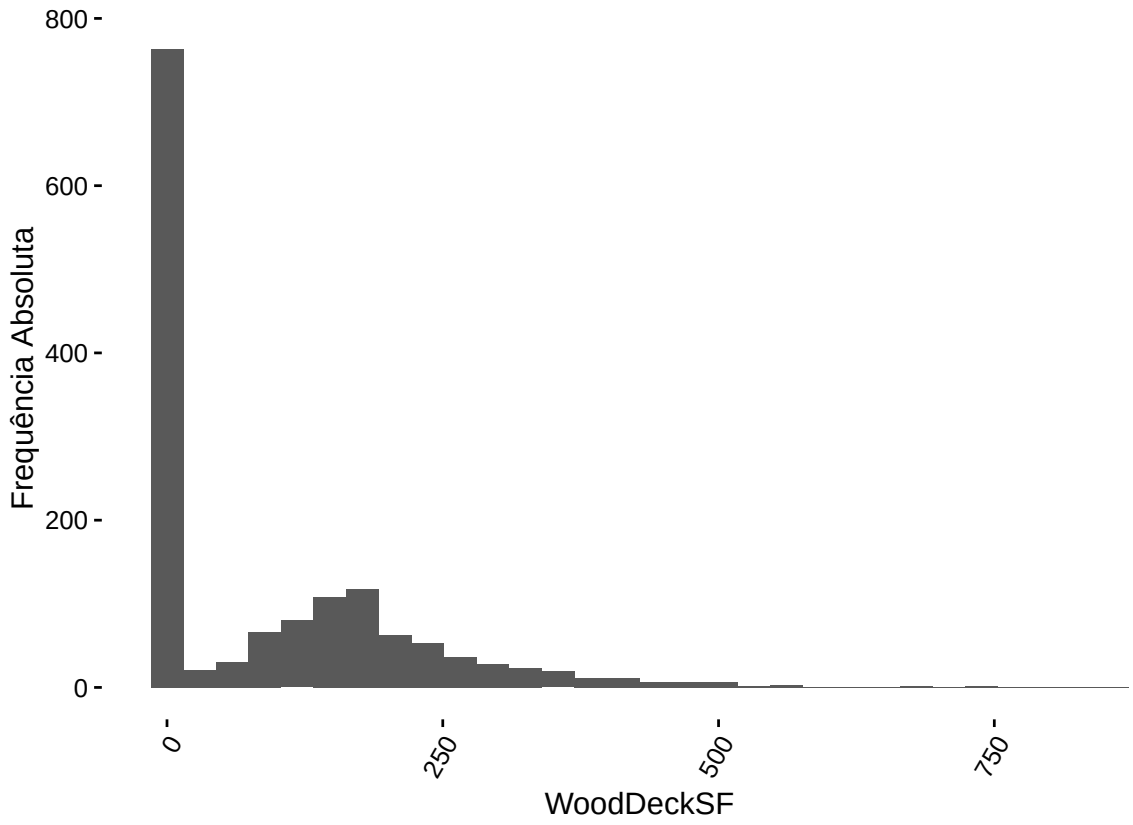

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



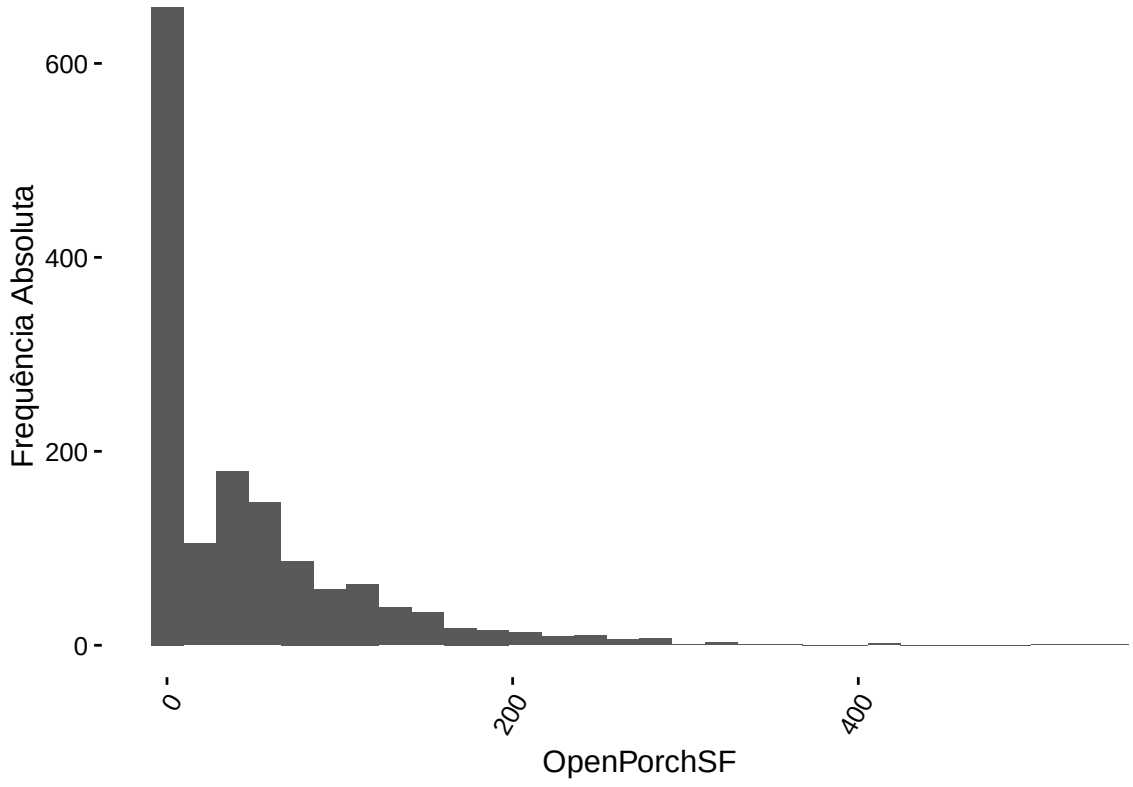

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



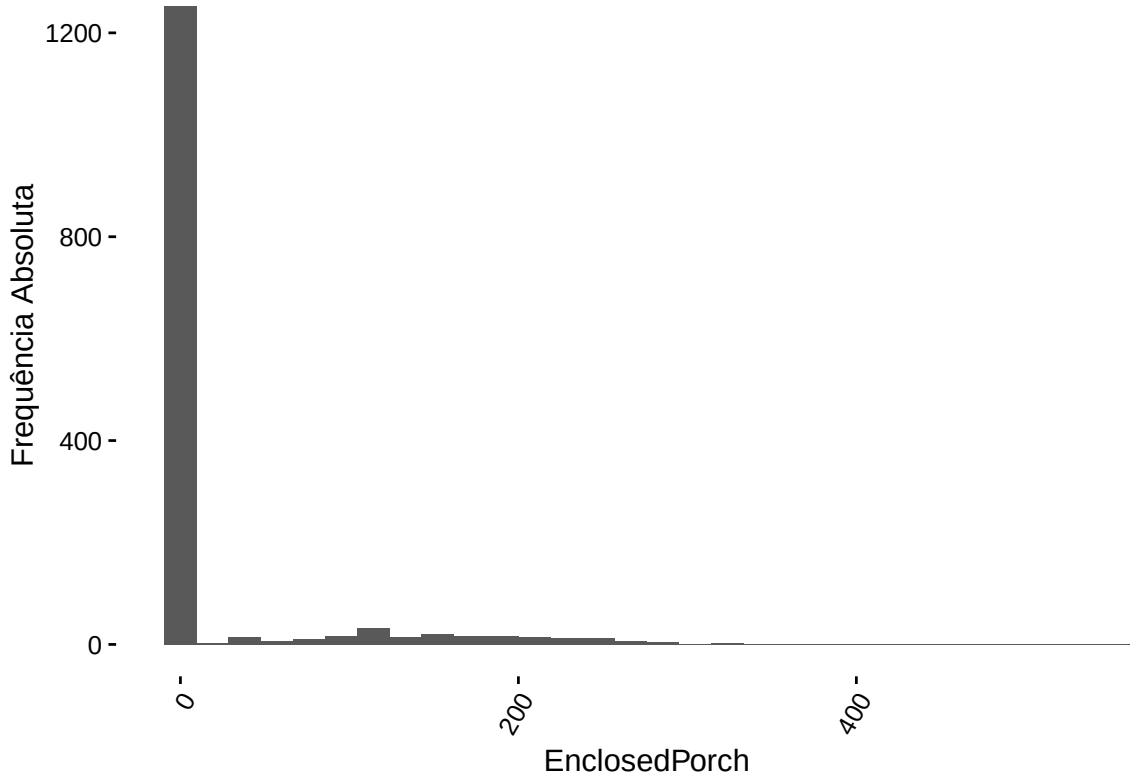

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



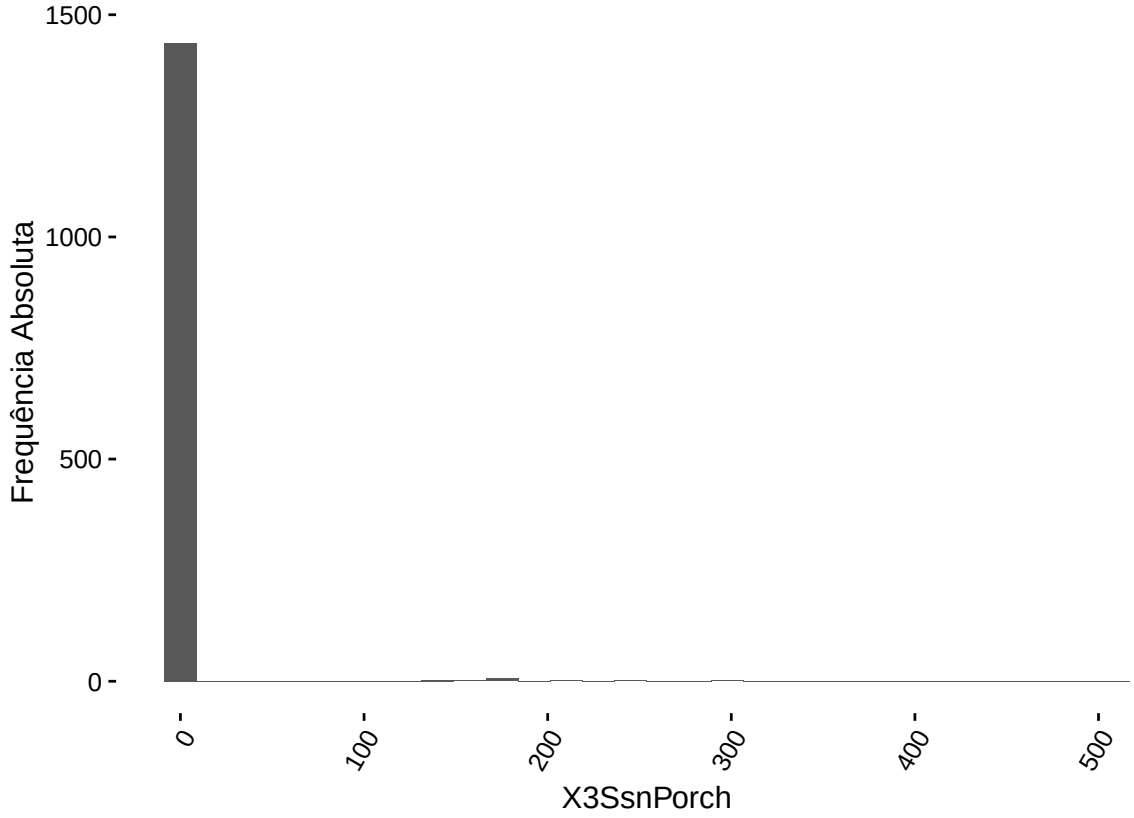

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



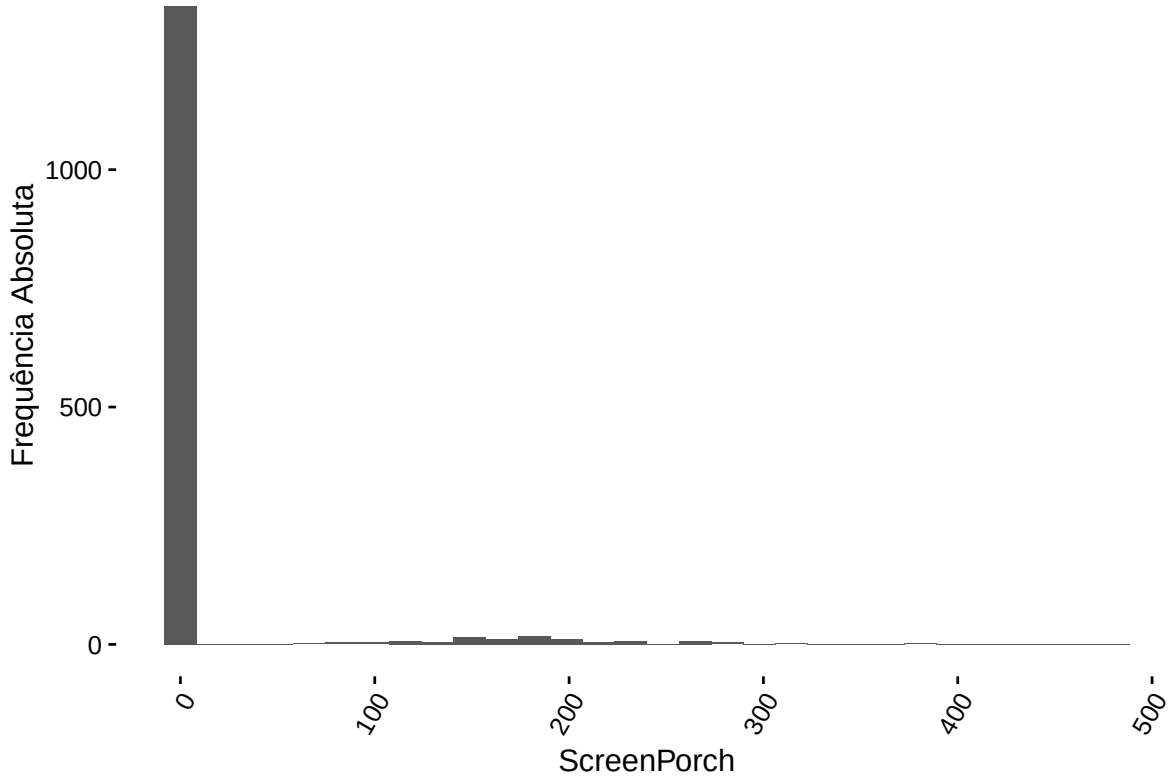

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



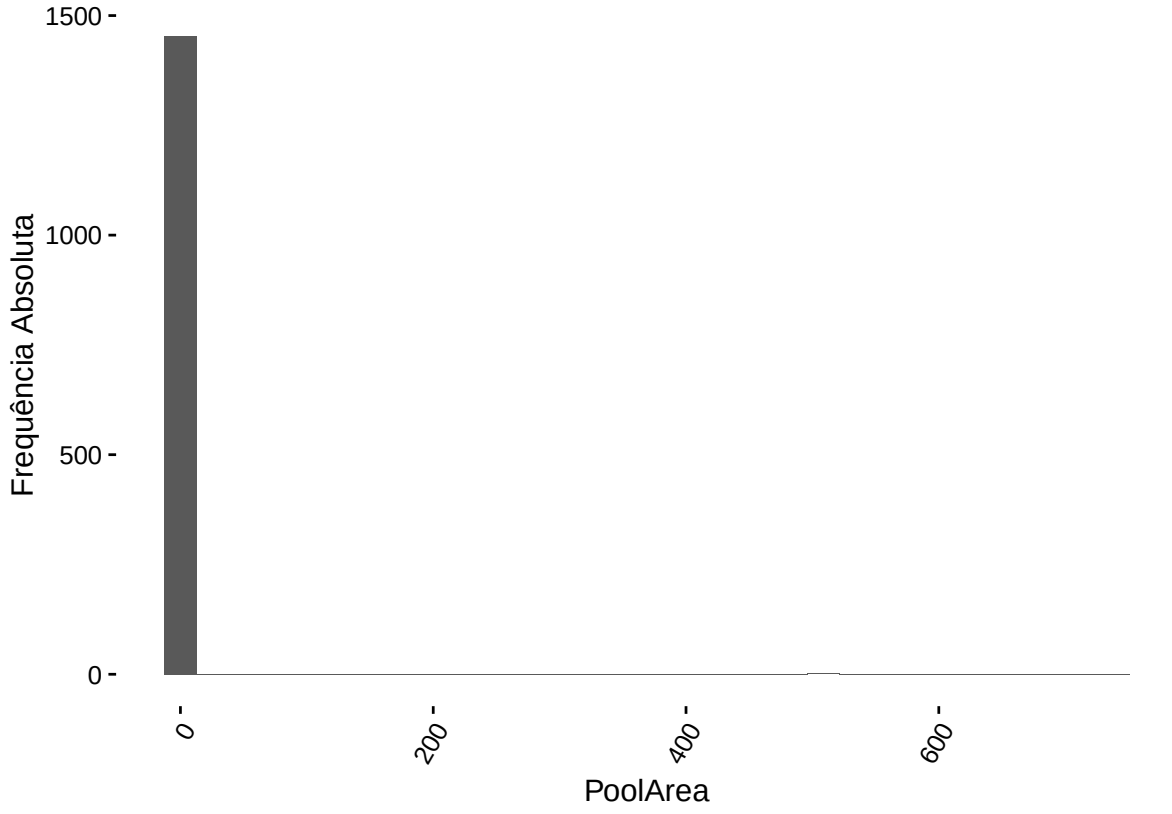

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



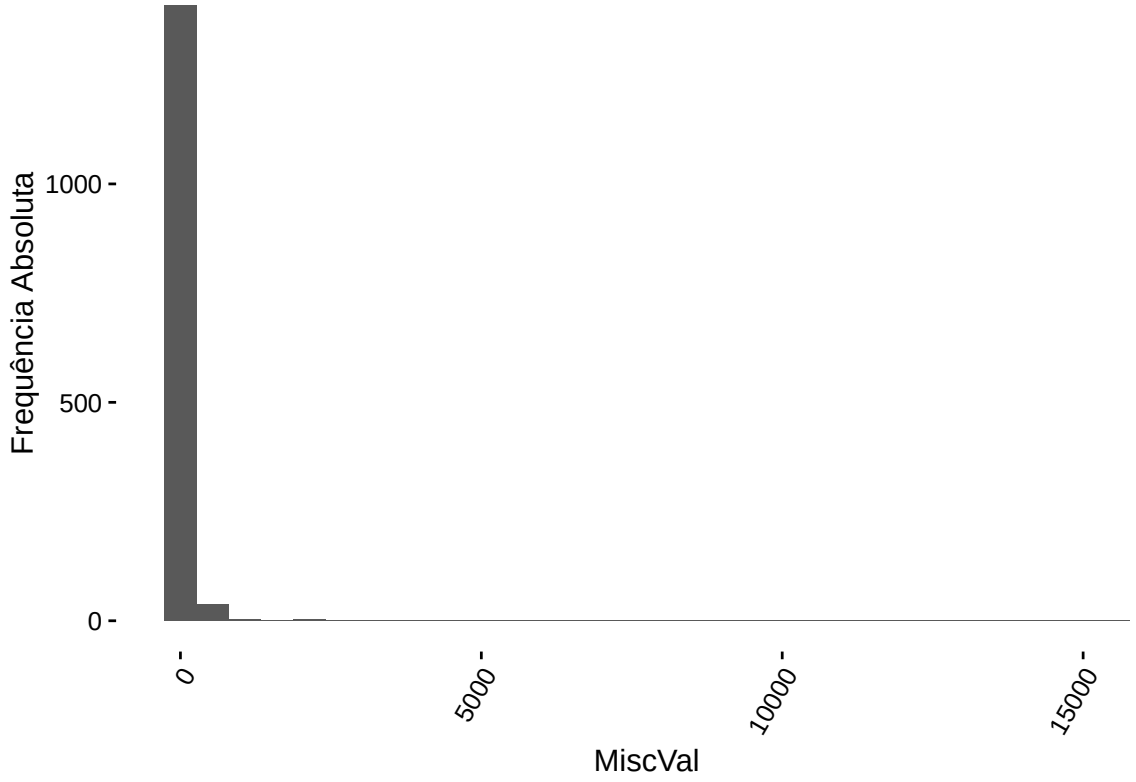

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



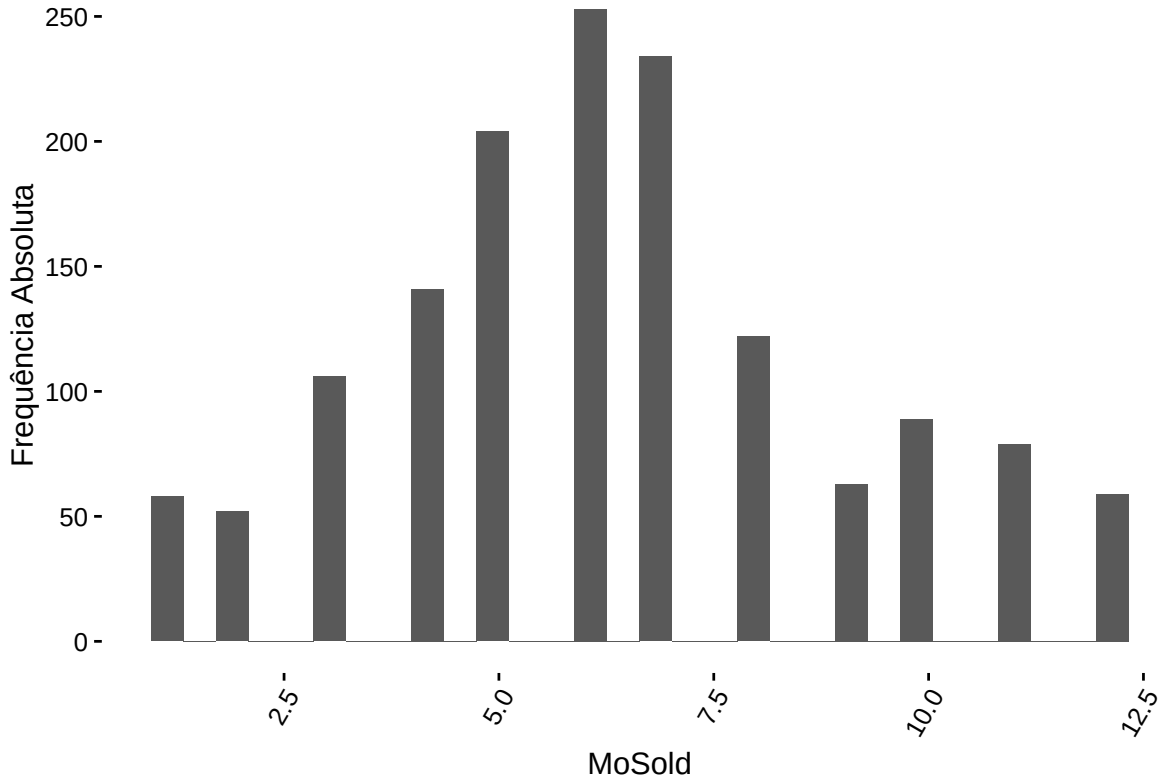

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



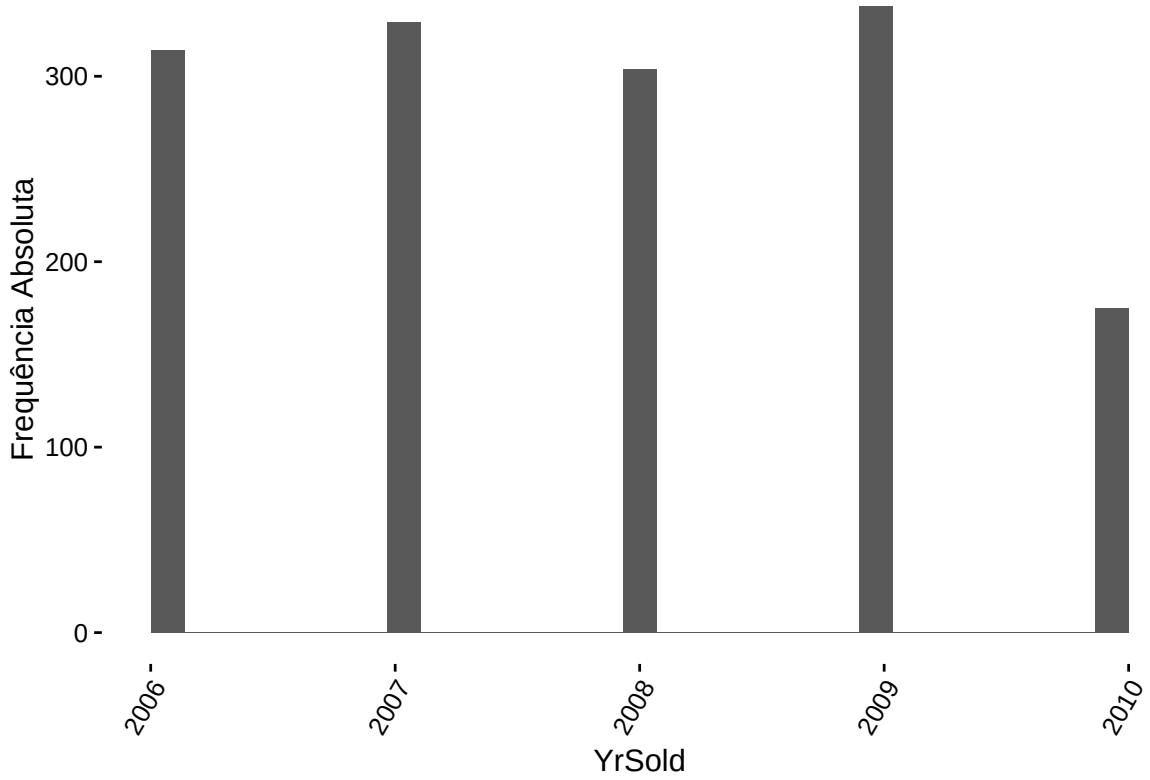

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



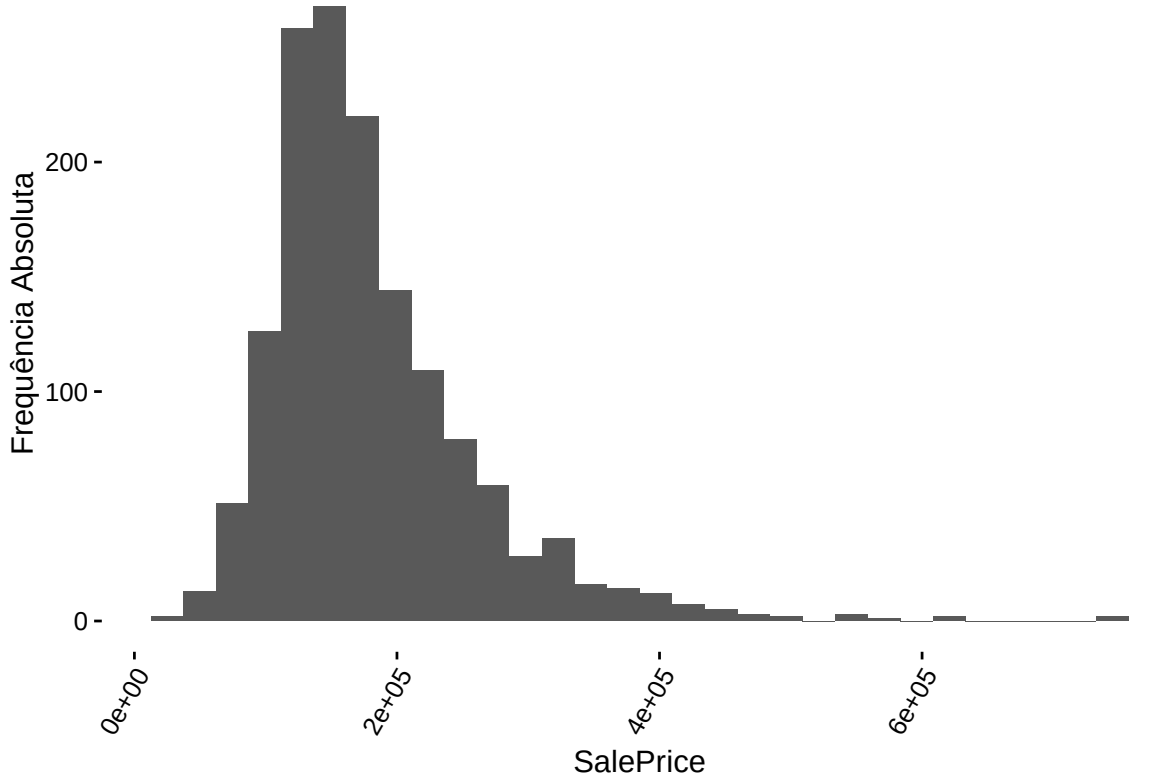

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



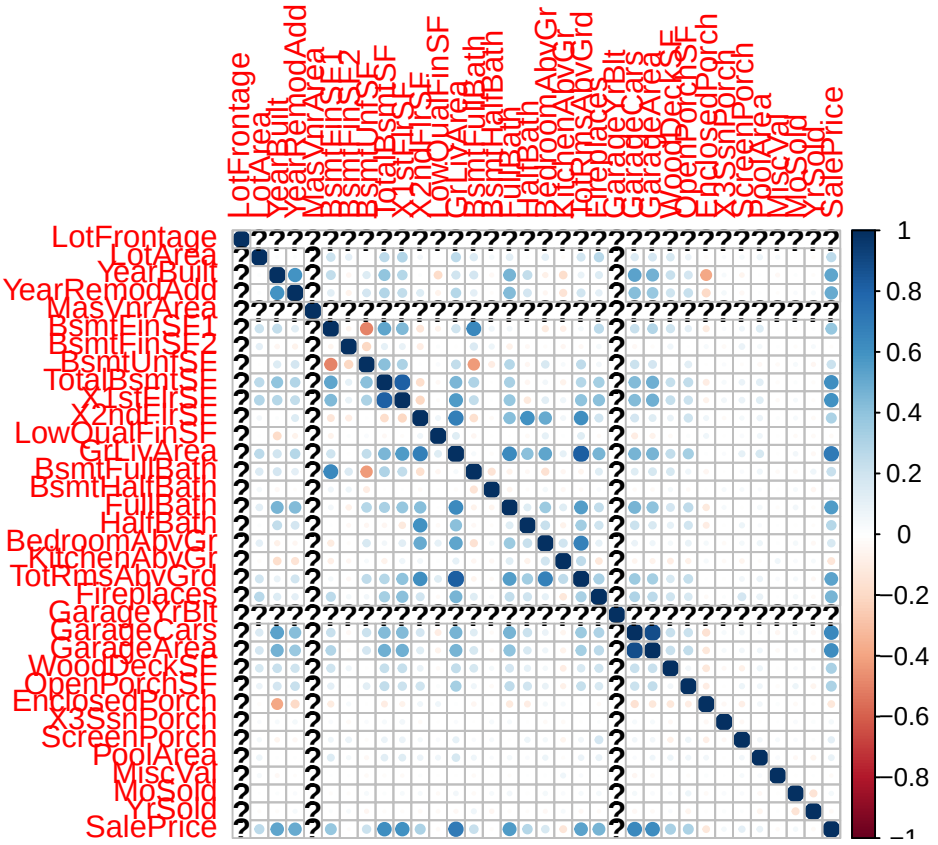

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Matriz de correlação



APÊNDICES

```
#Pacotes que serão utilizados
#ir incluindo em libs novos pacotes
libs = c('ggplot2','corrplot');
for(ii in libs){
  if(!( ii %in% installed.packages())){
    install.packages(ii);
  }
  library(ii, character.only=TRUE );
}
```

```
# Antes de começar dê um setwd na pasta HousePrices
# (onde quer que ela esteja no seu pc)

# setwd("Exploratory/");
```

```
# Importando dados
training_data <- read.csv(file="../Data/train.csv", as.is=FALSE);
remover_colunas <- c("Id"); #Boa Raul!
training_data <- training_data[ , !(names(training_data) %in% remover_colunas ) ];

# Transformando as variáveis qualitativas que aparecem como um código numérico em fatores
training_data$MSSubClass <- as.factor(training_data$MSSubClass);
training_data$OverallQual <- factor(training_data$OverallQual,levels= c(1,2,3,4,5,6,7,8,9,10));
training_data$OverallCond <- factor(training_data$OverallCond,levels= c(1,2,3,4,5,6,7,8,9,10));
```

```r
# Tipos das variáves
vars_tipo <- sapply(X=training_data, FUN=class);
```

```r
qualitativas <- which(vars_tipo=="factor");

for(ii in qualitativas){
  gg <- ggplot(data=training_data, mapping=aes_string(x=colnames(training_data)[ii]) ) +
    geom_bar() +
    scale_y_continuous(name="Frequência Absoluta") +
    theme_classic()+
    theme(axis.text.x = element_text(angle = 60, hjust = 1))
  print(gg);
}
```

```r
quantitativas <- which(vars_tipo=="integer");

for(ii in quantitativas){
  gg <- ggplot(data=training_data,
               mapping=aes_string(x=colnames(training_data)[ii]) ) +
    geom_histogram() +
    scale_y_continuous(name="Frequência Absoluta") +
    theme_classic()+
    theme(axis.text.x = element_text(angle = 60, hjust = 1))
  print(gg);
}
```

```r
#Matriz de Correlação entre as variáveis
var_quant <- training_data[quantitativas] #Seleciona as variáveis quantitativas de training_data
cor_quant <- cor(var_quant) #cria a matriz de correlação

corrplot(cor_quant, method = "circle") #plota a matriz de correlação com círculos
```