605-Final-Project

Michael Y.

December 15, 2019

Contents

FINAL

Your final is due by the end of the last week of class.
You should post your solutions to your GitHub account or RPubs.
You are also expected to make a short presentation via YouTube and post that recording to the board.
This project will show off your ability to understand the elements of the class.

Problem 1.

Using R, generate a random variable X that has 10,000 random **uniform** numbers from 1 to N, where N can be any number of your choosing greater than or equal to 6.

Generate X = U(1,N)

```
# set seed, for reproducibility
set.seed(12344)
### Note: seed =12345 will give the opposite result, causing rejection of the Null Hypothesis

# set maximum value for N
N <- 7

# generate 10,000 random uniform numbers between 1 and N
X <- runif(n = 10000,min = 1,max = N)

# obtain summary statistics on the distribution of X
summary(X)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 1.00071 2.47897 3.97223 3.99107 5.50269 7.00000
```

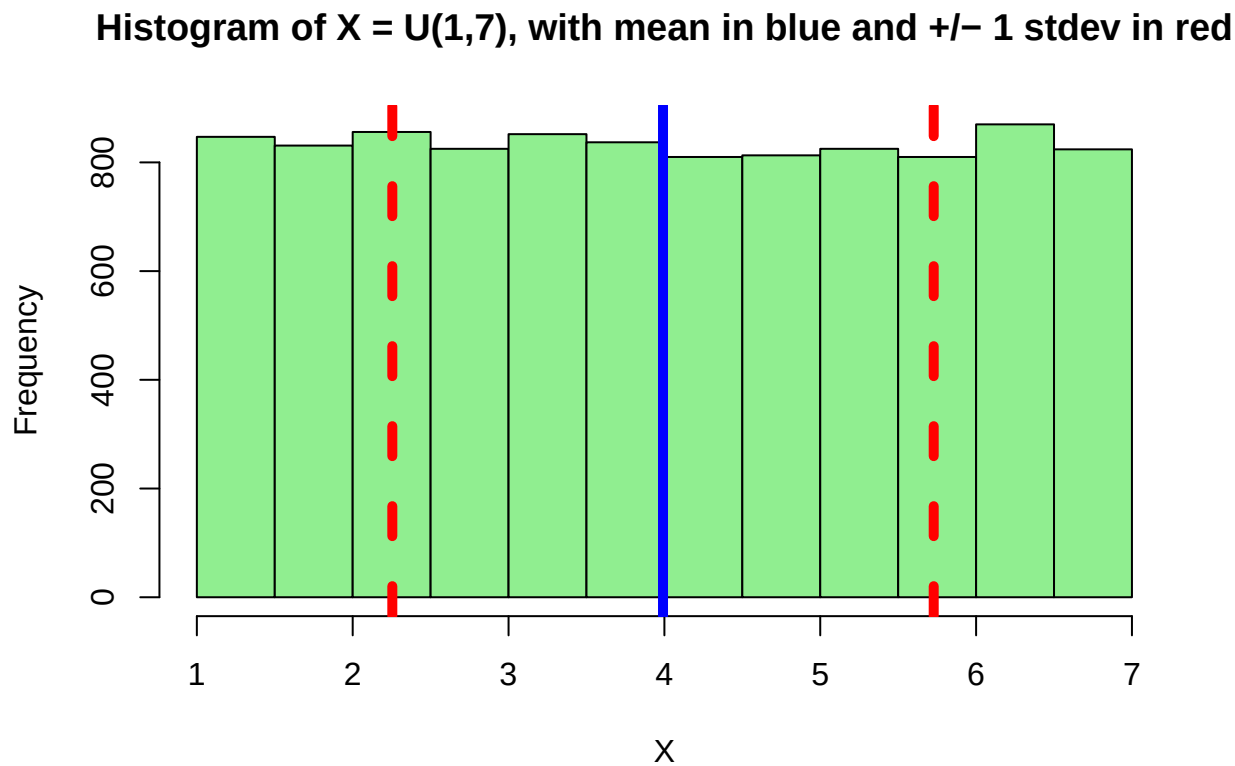
```
Xmean = round(summary(X)["Mean"],4)
Xmedian= round(summary(X)["Median"],4)
Xsd = sd(X)
Xsd_theo = (N-1)/sqrt(12)
print(paste("Actual standard deviation of X     : ",round(Xsd,4)))
```

```
## [1] "Actual standard deviation of X     :  1.7369"
```

```
print(paste("Theoretical standard deviation of X: ",round(Xsd_theo,4)))
```

```
## [1] "Theoretical standard deviation of X:  1.7321"
```

```
# plot a histogram of X, with vertical bars designating the mean and +/- 1 stdev
Xmaintitle = paste0("Histogram of X = U(1,", N,
                    "), with mean in blue and +/- 1 stdev in red")
hist(X,col="lightgreen",main = Xmaintitle)
abline(v=Xmean,col="blue",lwd=5,lty="solid")
abline(v=Xmean-Xsd,col="red",lwd=5,lty="dashed")
abline(v=Xmean+Xsd,col="red",lwd=5,lty="dashed")
```



Then generate a random variable Y that has 10,000 random **normal** numbers with a mean of $\mu = \sigma = \frac{N+1}{2}$.

Generate $Y = N(\mu = \frac{N+1}{2}, \sigma = \frac{N+1}{2})$

```
# generate 10,000 random  numbers with the specified Normal distribution
Y <- rnorm(n = 10000, mean = (N+1)/2, sd = (N+1)/2)
# obtain summary statistics on the distribution of Y
summary(Y)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -11.61246   1.29763   3.94817   3.99160   6.67449  21.22859
```

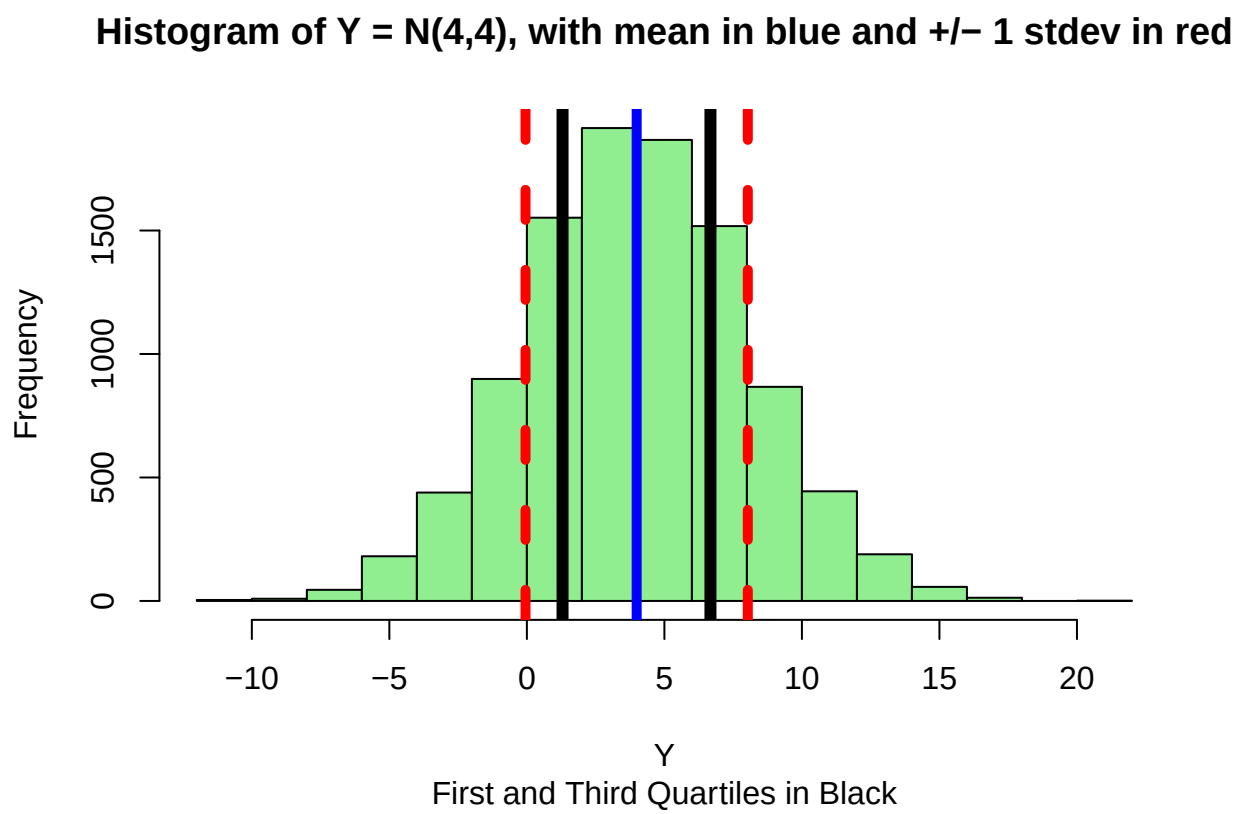
```
Ymean = round(summary(Y)["Mean"],4)
Ymean_theo = (N+1)/2
Ymedian= round(summary(Y)["Median"],4)
Ysd = sd(Y)
Ysd_theo = (N+1)/2
print(paste("Actual standard deviation of X     : ",round(Ysd,4)))
```

```
## [1] "Actual standard deviation of X     :  4.0412"
```

```r
print(paste("Theoretical standard deviation of X: ",round(Ysd_theo,4)))
```

```
## [1] "Theoretical standard deviation of X:  4"
```

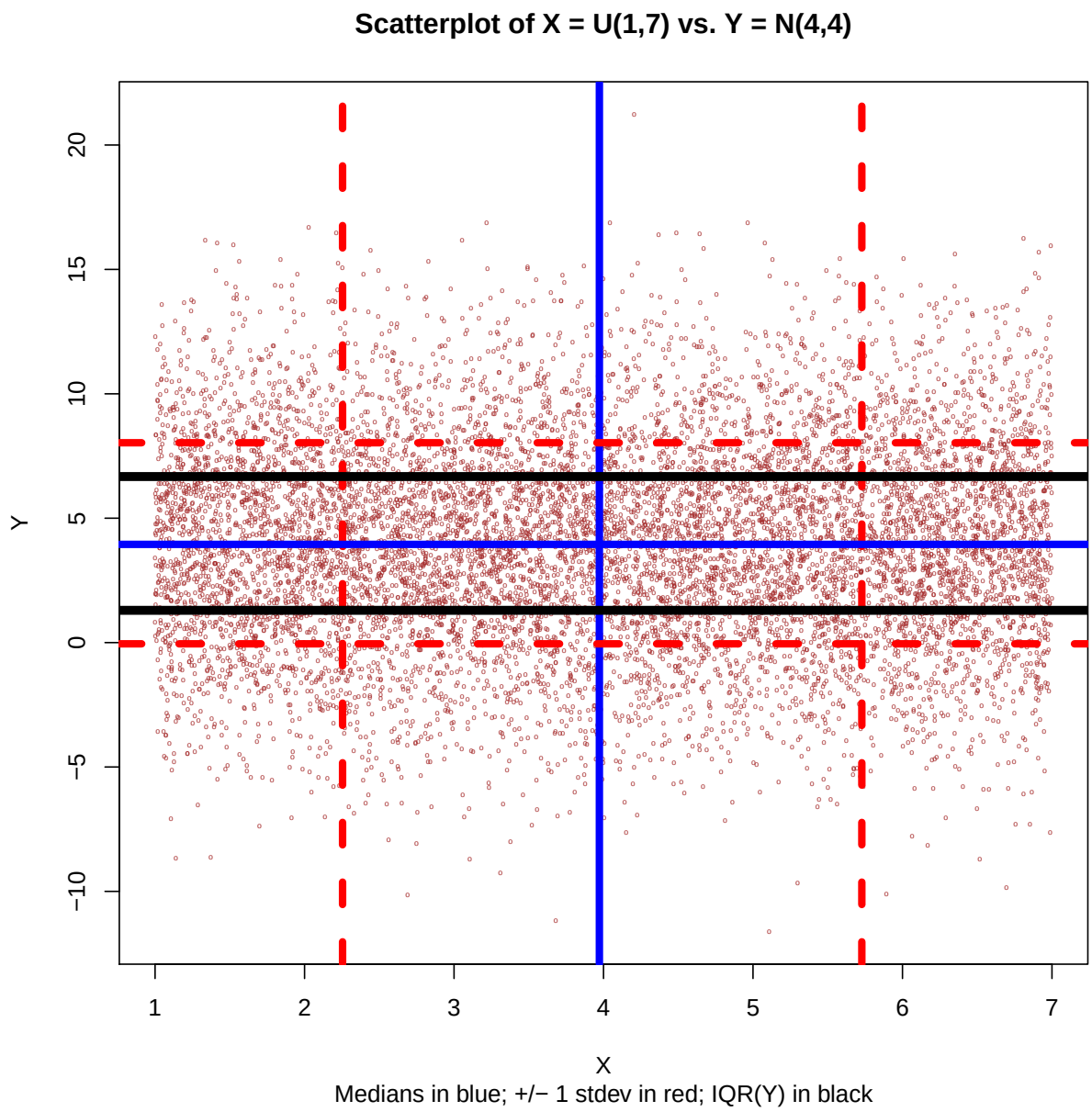
```r
# plot a histogram of Y, with vertical bars designating the median and +/- 1 stdev
Ymaintitle = paste0("Histogram of Y = N(", Ymean_theo, ",", Ysd_theo, "), with mean in blue and +/- 1 s
Ysubtitle = paste0("First and Third Quartiles in Black")
hist(Y,col="lightgreen",breaks = 22, main = Ymaintitle, sub=Ysubtitle)
abline(v=Ymean,col="blue",lwd=5,lty="solid")
abline(v=Ymean-Ysd,col="red",lwd=5,lty="dashed")
abline(v=Ymean+Ysd,col="red",lwd=5,lty="dashed")
abline(v=summary(Y)["1st Qu."], col="black",lwd=6)
abline(v=summary(Y)["3rd Qu."], col="black",lwd=6)
```



Scatterplot of X and Y:

```r
scat_maintitle = paste0("Scatterplot of X = U(1,", N,") vs. Y = N(", Ymean_theo, ",", Ysd_theo, ")")
scat_subtitle = "Medians in blue; +/- 1 stdev in red; IQR(Y) in black"
plot(Y~X, col="brown", pch="o", cex=0.3, main=scat_maintitle, sub=scat_subtitle)
abline(v=Xmedian,col="blue",lwd=5,lty="solid")
abline(v=Xmean-Xsd,col="red",lwd=5,lty="dashed")
abline(v=Xmean+Xsd,col="red",lwd=5,lty="dashed")
```

```r
abline(h=Ymedian,col="blue",lwd=5,lty="solid")
abline(h=Ymean-Ysd,col="red",lwd=5,lty="dashed")
abline(h=Ymean+Ysd,col="red",lwd=5,lty="dashed")
abline(h=summary(Y)["1st Qu."], col="black",lwd=6)
abline(h=summary(Y)["3rd Qu."], col="black",lwd=6)
```



5 points Probability.

Calculate as a minimum the below probabilities a through c.

Assume:

x=Median(X)

- The small letter "x" is estimated as the ***median*** of the X variable, and

```
x <- summary(X)["Median"]
x
```

```
##     Median
## 3.97223332
```

y=First Quartile(Y)

- The small letter "y" is estimated as the ***1st quartile*** of the Y variable.

```
y <- summary(Y)["1st Qu."]
y
```

```
##   1st Qu.
## 1.2976308
```

Interpret the meaning of all probabilities.

a. $P(X > x|X > y)$

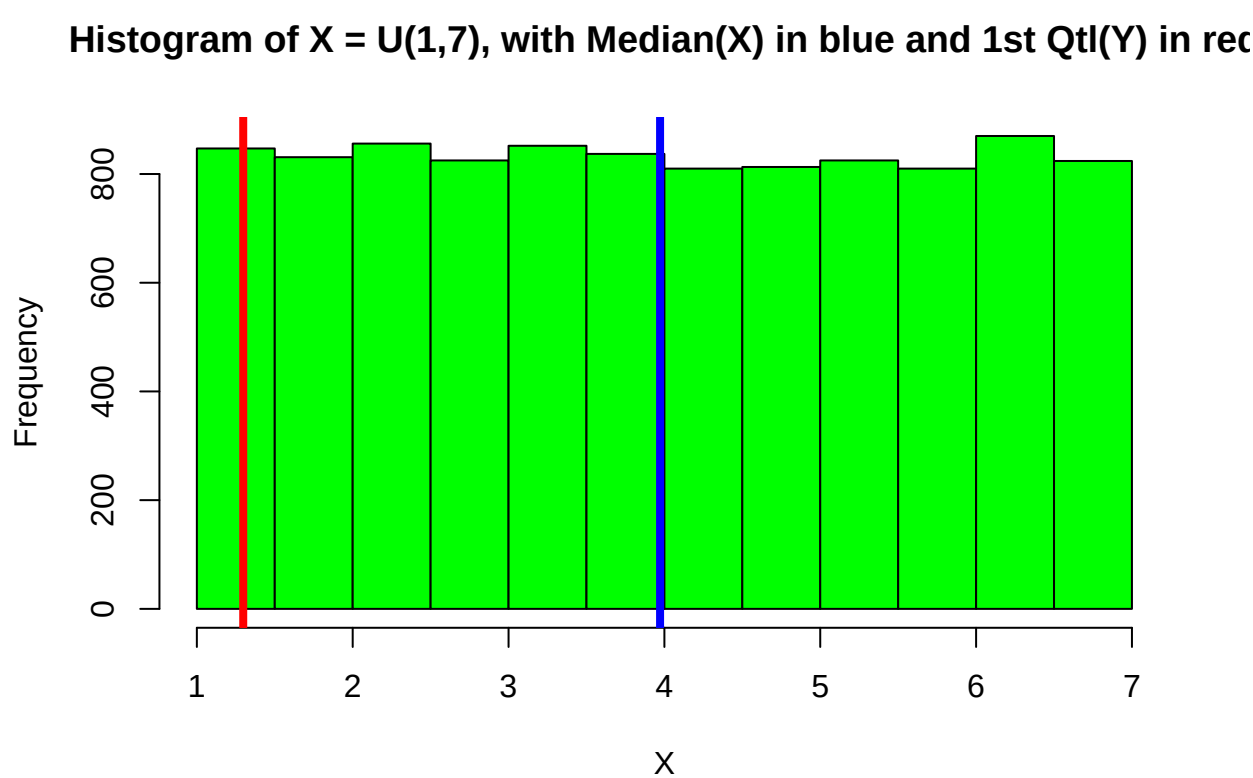
```
### Extract the values of X which are greater than y (the first quartile of Y)
tempa1 <- X[X>y]
### Count them
denom_a <- length(tempa1)
### Extract the values of the above subset which are greater than x (the median of X)
tempa2 <- tempa1[tempa1>x]
### Count them
numer_a <- length(tempa2)
### Compute the probability as the ratio of the above two items
prob_a <- numer_a / denom_a
```

There are 9490 values in X where such value is greater than 1.2976308, the first quartile of Y.

Of these values in X, there are 5000 values where such value is greater than 3.972233315, the median of X.

Therefore, the requested probability $Pr(X > x|X > y) = \frac{5000}{9490} = 0.52687039$.

```
Amaintitle = paste0("Histogram of X = U(1,", N,"), with Median(X) in blue and 1st Qtl(Y) in red")
hist(X,breaks=20, col="green", main=Amaintitle)
abline(v=summary(X)["Median"],col="blue", lwd=4)
abline(v=summary(Y)["1st Qu."], col="red", lwd=4)
```



The statement $Pr(X > x|X > y)$ means:

- "The probability that X is greater than x (i.e., the blue line)
- **given that**
- X is greater than y (i.e., the red line.)"

Of the values of X which are to the right of the red line, the probability is 0.52687039 that they are also to the right of the blue line.

b. $P(X > x, Y > y)$

This is the probability that X is greater than x ***and*** Y is greater than y .

Because X and Y are independently generated, $P(X > x, Y > y) = P(X > x) \cdot P(Y > y)$.

Because x is the median of X, $P(X > x) = 0.5$.
Because y is the first quartile of Y, $P(Y > y) = 0.75$.

Therefore, the answer must be $P(X > x, Y > y) = P(X > x) \cdot P(Y > y) = 0.5 \cdot 0.75 = 0.375$.

Empirically:

```
### Extract those values in X which are greater than the median of X
tempx <- X[X>x]
### Count them  (must be 5000, by definition of median)
length(tempx)
```

```
## [1] 5000
```

```
### Compute the probability (must be half, by definition of median)
probx <- length(tempx)/length(X)

### Extract those values in Y which are greater than the first quartile of Y
tempy <- Y[Y>y]
### Count the (must be 7500, by definition)
length(tempy)
```

```
## [1] 7500
```

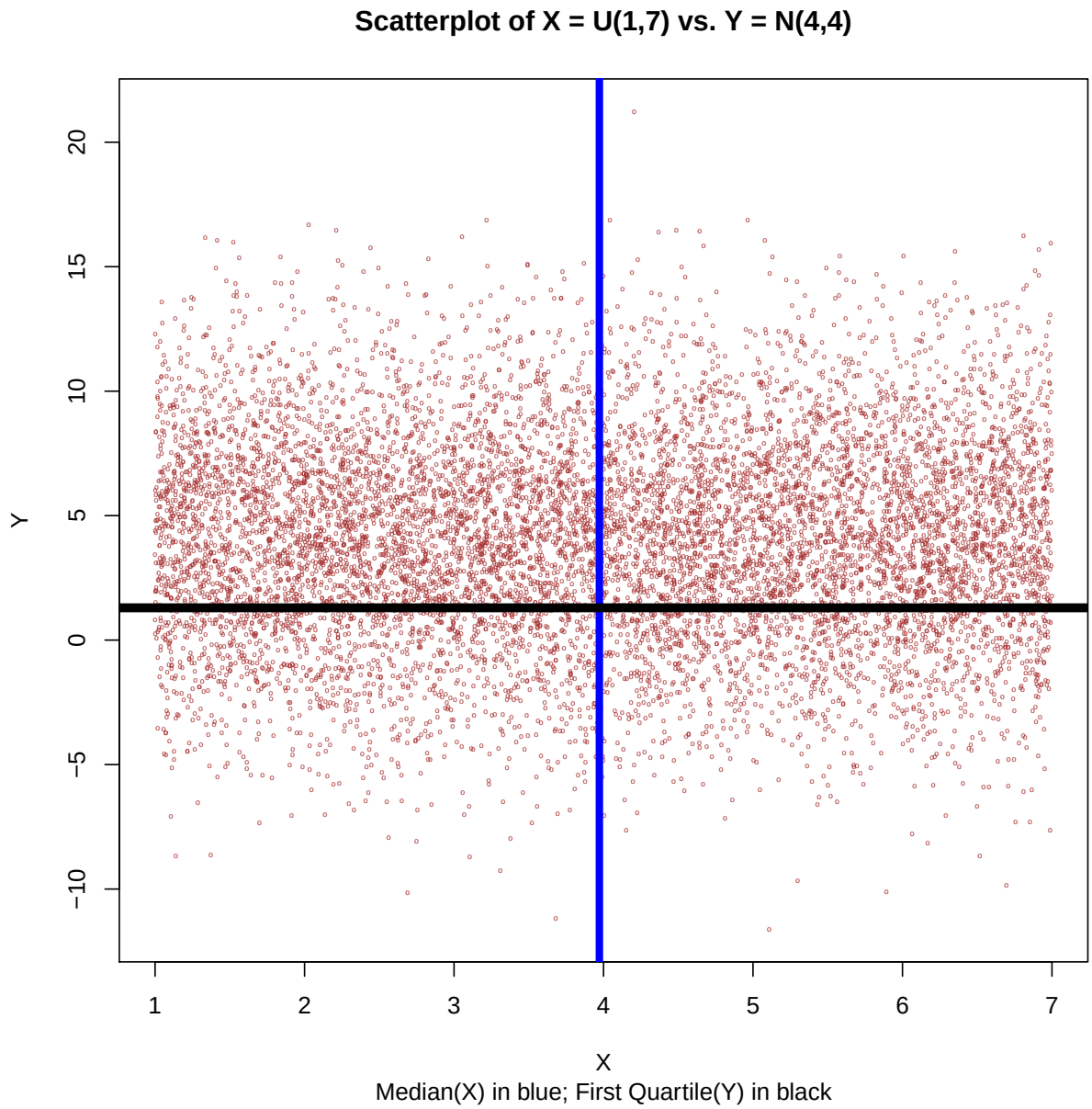
```
### Compute the probability (must be three quarters, by definition of quartiles)
proby <- length(tempy)/length(Y)

### Compute the result, utilize the assumption of independence of X and Y
result <- probx * proby
result
```

```
## [1] 0.375
```

This confirms that the result is 0.375, which equals $\frac{3}{8}$.

Visually:

```
scat_maintitle = paste0("Scatterplot of X = U(1,", N,") vs. Y = N(", Ymean_theo, ",", Ysd_theo, ")")
scat_subtitle = "Median(X) in blue; First Quartile(Y) in black"
plot(Y~X, col="brown", pch="o", cex=0.3, main=scat_maintitle, sub=scat_subtitle)
abline(v=Xmedian,col="blue",lwd=5,lty="solid")
abline(h=summary(Y)["1st Qu."], col="black",lwd=6)
```



Three-quarters of the dots are above the black line (first quartile of Y.)
Of these, half are to the right of the blue line (Median of X).

Therefore, $\frac{3}{8}$ of the dots are in the upper right quadrant in the above scatterplot.

c. $P(X < x|X > y)$

```
### Extract the values of X which are greater than y (the first quartile of Y)
tempc1 <- X[X>y]
### Count them - should be the same as in part (a) above
denom_c <- length(tempc1)
### Extract the values of the above subset which are LESS THAN than x (the median of X)
tempc2 <- tempc1[tempc1<x]
### Count them
numer_c <- length(tempc2)
### Compute the probability as the ratio of the above two items  --
###          should be 1 minus the probability from part (a)
prob_c <- numer_c / denom_c
```

There are 9490 values in X where such value is ***greater*** than 1.2976308, the ***first quartile*** of Y.
Of these values in X, there are 4490 values where such value is ***less*** than 3.972233315, the ***median*** of X.
Therefore, the requested probability is $Pr(X > x|X > y) = \frac{4490}{9490} = 0.47312961$.

It is worth noting that this this result, plus the result from part (a), sum up to 1.

The statement $Pr(X < x|X > y)$ means:

- "The probability that X is ***less*** than x (i.e., the blue line)
- **given that**
- X is greater than y (i.e., the red line.)"

Of the values of X which are to the right of the red line, the probability is 0.52687039 that they are also to the ***left*** of the blue line (i.e., between the red line and the blue line.)

---

5 points. Joint and Marginal

Investigate whether $P(X > x \bigwedge Y > y) = P(X > x) \cdot P(Y > y)$ by building a table and evaluating the marginal and joint probabilities.

```r
### Build a table of the 4 cases
actual <- table(X>x,Y>y,dnn = c("X>x","Y>y"))
actual
```

```
##        Y>y
## X>x     FALSE TRUE
##   FALSE  1235 3765
##   TRUE   1265 3735
```

```r
### Display the above, as probabilities
prop.table(actual)
```

```
##        Y>y
## X>x      FALSE   TRUE
##   FALSE 0.1235 0.3765
##   TRUE  0.1265 0.3735
```

```r
### Display the marginal probabilities, by row
### P(Y>y|X)
prop.table(actual,margin = 1)
```

```
##        Y>y
## X>x     FALSE  TRUE
##   FALSE 0.247 0.753
##   TRUE  0.253 0.747
```

```r
### Display the marginal probabilities, by column
### P(X>x|y)
prop.table(actual,margin = 2)
```

```
##        Y>y
## X>x     FALSE  TRUE
##   FALSE 0.494 0.502
##   TRUE  0.506 0.498
```

```r
### Display the expected results
expected = margin.table(actual, margin=1) %*% t(margin.table(actual, margin=2))/margin.table(actual)
rownames(expected)=colnames(expected)
dimnames(expected)<-dimnames(actual)
expected <- as.table(expected)

### Display the expected results, as probabilities
prop.table(expected)
```

```
##        Y>y
## X>x     FALSE  TRUE
##   FALSE 0.125 0.375
##   TRUE  0.125 0.375
```

```r
### Are Actual and Expected equal?
result <- all.equal(actual,expected)
result
```

```
## [1] "Mean relative difference: 0.006"
```

```r
if(isTRUE(result)) {
  print("Actual and expected are equal")
} else {
  print("Actual and expected are different")
}
```

```
## [1] "Actual and expected are different"
```

---

5 points. Fisher's Exact vs. χ^2 tests

Check to see if independence holds by using ***Fisher's Exact Test*** and the ***Chi Square Test***.

- H_0: Variables X and Y are ***independent***.
- H_a: Variables X and Y are ***not independent***.

Fisher's Exact Test

```
fisher<-fisher.test(actual)
fisher
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  actual
## p-value = 0.503037
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.883753438 1.061400893
## sample estimates:
##  odds ratio
## 0.968497896
```

```
if (fisher$p.value < 0.05) {
  fisher.result <-
    "***reject the null***: Actual differs too much from expected,
    so X and Y are ***not independent***."
} else {
  fisher.result <-
    "***fail to reject*** the null: Actual is close enough to expected,
    so X and Y ***are*** independent."
}
```

The p-value from the Fisher Exact test is 0.503036861 .
This means that we ***fail to reject*** the null: Actual is close enough to expected, so X and Y ***are*** independent.

χ^2 test

```
chi2 <-chisq.test(actual,correct=F)
chi2
```

```
##
##  Pearson's Chi-squared test
##
## data:  actual
## X-squared = 0.48, df = 1, p-value = 0.488422
```

```r
if (chi2$p.value < 0.05) {
  chi2.result <-
    "***reject the null***: Actual differs too much from expected,
    so X and Y are ***not independent***."
} else {
  chi2.result <-
    "***fail to reject*** the null: Actual is close enough to expected,
    so X and Y ***are*** independent."
}
```

The p-value from the χ^2 test test is 0.488422317 .
This means that we ***fail to reject*** the null: Actual is close enough to expected, so X and Y ***are*** independent.

What is the difference between the two?

The χ^2 test requires that the expected number of counts in each cell of the contingency matrix is at least 5, while the Fisher's Exact Test is used when the count data in the respective cells is small (i.e., some cell has fewer than 5 elements.)

Which is most appropriate?

As the counts in each cell are large, that would suggest that the χ^2 test is preferred.

However, because both tests are designed to be used on ***Categorical*** data, **neither is appropriate here** because we are essentially testing the quality of our random number generator, as it is clear that each of the variables X and Y has been generated independently of the other.
In this case, for the above seed, the difference between actual and expected is ***15*** .

I have determined that for these samples of size 10,000, if the counts in the contingency table are ***no more than 43 distant*** from the expected counts, then the tests will pass (more correctly, "fail to reject") at 95% confidence, returning p-values above 0.05.

Whether the tests pass or fail is closely coupled to the selection of the random seed.
For example, running the above with ***seed=12345*** will shift the results sufficiently for **both** tests to ***reject the null*** hypothesis. (For that particular seed, the actual counts differ by **58** from the expected counts, resulting in rejection of the null hypothesis.)

Problem 2.

You are to register for Kaggle.com (free) and compete in the House Prices: Advanced Regression Techniques competition. https://www.kaggle.com/c/house-prices-advanced-regression-techniques .

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------------

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------------------------------------
## x dplyr::filter()     masks stats::filter()
## x dplyr::group_rows() masks kableExtra::group_rows()
## x dplyr::lag()        masks stats::lag()
```

```
library(kableExtra)
na.strings=c("NA","NaN", " ", "")
train.df <- read.csv('train.csv',na.strings=na.strings)     # 1460 obs of 81 variables
test.df <- read.csv('test.csv',na.strings=na.strings)       # 1459 obs of 80 variables
```

Data fields

Let's preview the training dataset: (columns & rows transposed for viewing)

```
## [1] "Number of columns =  81"
```

```
## [1] "Number of rows =  1460"
```

	1	2	3	4	5	6
Id	1	2	3	4	5	6
MSSubClass	60	20	60	70	60	50
MSZoning	RL	RL	RL	RL	RL	RL
LotFrontage	65	80	68	60	84	85
LotArea	8450	9600	11250	9550	14260	14115
Street	Pave	Pave	Pave	Pave	Pave	Pave
Alley	NA	NA	NA	NA	NA	NA
LotShape	Reg	Reg	IR1	IR1	IR1	IR1
LandContour	Lvl	Lvl	Lvl	Lvl	Lvl	Lvl
Utilities	AllPub	AllPub	AllPub	AllPub	AllPub	AllPub
LotConfig	Inside	FR2	Inside	Corner	FR2	Inside
LandSlope	Gtl	Gtl	Gtl	Gtl	Gtl	Gtl
Neighborhood	CollgCr	Veenker	CollgCr	Crawfor	NoRidge	Mitchel
Condition1	Norm	Feedr	Norm	Norm	Norm	Norm
Condition2	Norm	Norm	Norm	Norm	Norm	Norm
BldgType	1Fam	1Fam	1Fam	1Fam	1Fam	1Fam
HouseStyle	2Story	1Story	2Story	2Story	2Story	1.5Fin
OverallQual	7	6	7	7	8	5
OverallCond	5	8	5	5	5	5
YearBuilt	2003	1976	2001	1915	2000	1993
YearRemodAdd	2003	1976	2002	1970	2000	1995
RoofStyle	Gable	Gable	Gable	Gable	Gable	Gable
RoofMatl	CompShg	CompShg	CompShg	CompShg	CompShg	CompShg
Exterior1st	VinylSd	MetalSd	VinylSd	Wd Sdng	VinylSd	VinylSd
Exterior2nd	VinylSd	MetalSd	VinylSd	Wd Shng	VinylSd	VinylSd
MasVnrType	BrkFace	None	BrkFace	None	BrkFace	None
MasVnrArea	196	0	162	0	350	0
ExterQual	Gd	TA	Gd	TA	Gd	TA
ExterCond	TA	TA	TA	TA	TA	TA
Foundation	PConc	CBlock	PConc	BrkTil	PConc	Wood
BsmtQual	Gd	Gd	Gd	TA	Gd	Gd
BsmtCond	TA	TA	TA	Gd	TA	TA
BsmtExposure	No	Gd	Mn	No	Av	No
BsmtFinType1	GLQ	ALQ	GLQ	ALQ	GLQ	GLQ
BsmtFinSF1	706	978	486	216	655	732
BsmtFinType2	Unf	Unf	Unf	Unf	Unf	Unf
BsmtFinSF2	0	0	0	0	0	0
BsmtUnfSF	150	284	434	540	490	64
TotalBsmtSF	856	1262	920	756	1145	796
Heating	GasA	GasA	GasA	GasA	GasA	GasA
HeatingQC	Ex	Ex	Ex	Gd	Ex	Ex

	1	2	3	4	5	6
CentralAir	Y	Y	Y	Y	Y	Y
Electrical	SBrkr	SBrkr	SBrkr	SBrkr	SBrkr	SBrkr
X1stFlrSF	856	1262	920	961	1145	796
X2ndFlrSF	854	0	866	756	1053	566
LowQualFinSF	0	0	0	0	0	0
GrLivArea	1710	1262	1786	1717	2198	1362
BsmtFullBath	1	0	1	1	1	1
BsmtHalfBath	0	1	0	0	0	0
FullBath	2	2	2	1	2	1
HalfBath	1	0	1	0	1	1
BedroomAbvGr	3	3	3	3	4	1
KitchenAbvGr	1	1	1	1	1	1
KitchenQual	Gd	TA	Gd	Gd	Gd	TA
TotRmsAbvGrd	8	6	6	7	9	5
Functional	Typ	Typ	Typ	Typ	Typ	Typ
Fireplaces	0	1	1	1	1	0
FireplaceQu	NA	TA	TA	Gd	TA	NA
GarageType	Attchd	Attchd	Attchd	Detchd	Attchd	Attchd
GarageYrBlt	2003	1976	2001	1998	2000	1993
GarageFinish	RFn	RFn	RFn	Unf	RFn	Unf
GarageCars	2	2	2	3	3	2
GarageArea	548	460	608	642	836	480
GarageQual	TA	TA	TA	TA	TA	TA
GarageCond	TA	TA	TA	TA	TA	TA
PavedDrive	Y	Y	Y	Y	Y	Y
WoodDeckSF	0	298	0	0	192	40
OpenPorchSF	61	0	42	35	84	30
EnclosedPorch	0	0	0	272	0	0
X3SsnPorch	0	0	0	0	0	320
ScreenPorch	0	0	0	0	0	0
PoolArea	0	0	0	0	0	0
PoolQC	NA	NA	NA	NA	NA	NA
Fence	NA	NA	NA	NA	NA	MnPrv
MiscFeature	NA	NA	NA	NA	NA	Shed
MiscVal	0	0	0	0	0	700
MoSold	2	5	9	2	12	10
YrSold	2008	2007	2008	2006	2008	2009
SaleType	WD	WD	WD	WD	WD	WD
SaleCondition	Normal	Normal	Normal	Abnorml	Normal	Normal
SalePrice	208500	181500	223500	140000	250000	143000

Data Description summary

- ***SalePrice* - the property's sale price in dollars. This is the target variable that you're trying to predict.**
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access

- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material

- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade

- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location

- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

Fix categorical (non-numeric) variable MSSubClass

The variable MSSubclass is categorical, but it labels each class using an integer between 20 and 190:

MSSubClass: Identifies the type of dwelling involved in the sale.

```
+---+-------------------------------------------------------+
|Val|  Description                                          |
+===+=======================================================+
|20 |  1-STORY 1946 & NEWER ALL STYLES                      |
+---+-------------------------------------------------------+
|30 |  1-STORY 1945 & OLDER                                 |
+---+-------------------------------------------------------+
|40 |  1-STORY W/FINISHED ATTIC ALL AGES                    |
+---+-------------------------------------------------------+
|45 |  1-1/2 STORY - UNFINISHED ALL AGES                    |
+---+-------------------------------------------------------+
|50 |  1-1/2 STORY FINISHED ALL AGES                        |
+---+-------------------------------------------------------+
|60 |  2-STORY 1946 & NEWER                                 |
+---+-------------------------------------------------------+
|70 |  2-STORY 1945 & OLDER                                 |
+---+-------------------------------------------------------+
|75 |  2-1/2 STORY ALL AGES                                 |
+---+-------------------------------------------------------+
|80 |  SPLIT OR MULTI-LEVEL                                 |
+---+-------------------------------------------------------+
|85 |  SPLIT FOYER                                          |
+---+-------------------------------------------------------+
|90 |  DUPLEX - ALL STYLES AND AGES                         |
+---+-------------------------------------------------------+
|120|  1-STORY PUD (Planned Unit Development) - 1946 & NEWER |
```

```
+---+-------------------------------------------------------+
|150|  1-1/2 STORY PUD - ALL AGES                            |
+---+-------------------------------------------------------+
|160|  2-STORY PUD - 1946 & NEWER                            |
+---+-------------------------------------------------------+
|180|  PUD - MULTILEVEL - INCL SPLIT LEV/FOYER               |
+---+-------------------------------------------------------+
|190|  2 FAMILY CONVERSION - ALL STYLES AND AGES             |
+---+-------------------------------------------------------+
```

Such data was loaded in as numeric, but it should not be treated this was, as there is no ordinal relation between the classes:

```r
train.df$MSSubClass<-as.factor(train.df$MSSubClass)
## implement the same transform for test dataset
test.df$MSSubClass<-as.factor(test.df$MSSubClass)
```

Table 1: Quantity of missing elements in TRAIN

	x
LotFrontage	259
Alley	1369
MasVnrType	8
MasVnrArea	8
BsmtQual	37
BsmtCond	37
BsmtExposure	38
BsmtFinType1	37
BsmtFinType2	38
Electrical	1
FireplaceQu	690
GarageType	81
GarageYrBlt	81
GarageFinish	81
GarageQual	81
GarageCond	81
PoolQC	1453
Fence	1179
MiscFeature	1406

Check missing values

Let's check whether any variables have missing values, e.g., values which are NULL or NA:

check for train

```
miss.cols = apply(train.df, 2, function(x) any(is.na(x)))
miss.cols = miss.cols[miss.cols==T]
num.NAs = apply(train.df, 2, function(x) sum(is.na(x)))
num.NAs = num.NAs[num.NAs>0]
print(paste("Number of TRAIN columns with missing values = ", length(names(miss.cols))))
```

```
## [1] "Number of TRAIN columns with missing values =  19"
```

```
# Number of missing elements in TRAIN:
num.NAs %>%
  kable(caption = "Quantity of missing elements in TRAIN") %>%
  kable_styling(c("striped", "bordered"),full_width = F)
```

check for test

```
test.miss.cols = apply(test.df, 2, function(x) any(is.na(x)))
test.miss.cols = test.miss.cols[test.miss.cols==T]
test.num.NAs = apply(test.df, 2, function(x) sum(is.na(x)))
test.num.NAs = test.num.NAs[test.num.NAs>0]
print(paste("Number of TEST columns with missing values = ", length(names(test.miss.cols))))
```

```
## [1] "Number of TEST columns with missing values =  33"
```

```
# Number of TEST missing elements:
test.num.NAs %>%
  kable(caption = "Quantity of missing elements in TEST") %>%
  kable_styling(c("striped", "bordered"),full_width = F)
```

Table 2: Quantity of missing elements in TEST

	x
MSZoning	4
LotFrontage	227
Alley	1352
Utilities	2
Exterior1st	1
Exterior2nd	1
MasVnrType	16
MasVnrArea	15
BsmtQual	44
BsmtCond	45
BsmtExposure	44
BsmtFinType1	42
BsmtFinSF1	1
BsmtFinType2	42
BsmtFinSF2	1
BsmtUnfSF	1
TotalBsmtSF	1
BsmtFullBath	2
BsmtHalfBath	2
KitchenQual	1
Functional	2
FireplaceQu	730
GarageType	76
GarageYrBlt	78
GarageFinish	78
GarageCars	1
GarageArea	1
GarageQual	78
GarageCond	78
PoolQC	1456
Fence	1169
MiscFeature	1408
SaleType	1

split up TRAINING dataset by numeric|factor ; missing|none missing

Make separate dataframes to split up the variables based on

- variables with some missing elements, vs variables with nothing missing
- numeric vs. categorical (factors)

```
#### Missing vs. none missing
# 19 variables --> 0 after cleaning
train.df.somemissing=train.df[sapply(train.df, function(x) sum(is.na(x))>0)]
# 62 variables --> 81 after cleaning
train.df.nomissing=train.df[sapply(train.df, function(x) sum(is.na(x))==0)]

#### numeric vs. factors
# 38 variables -> 37
train.df.numeric=train.df[sapply(train.df, is.numeric)]
# 43 variables -> 44
train.df.factor=train.df[sapply(train.df, is.factor)]

#### numeric/factor none missing
# 35 variables -> 34 -> 37
train.df.numeric.nomissing = train.df.nomissing[sapply(train.df.nomissing, is.numeric)]
# 27 variables -> 28 -> 44
train.df.factor.nomissing =  train.df.nomissing[sapply(train.df.nomissing, is.factor)]
```

Data cleaning – NAs which mean "Not Applicable" rather than "Unknown" or "Not Available

It appears that there may be two different meanings to the NA values shown.

For some variables, an "NA" may mean "Not Applicable". In such cases, it doesn't make sense to impute various values for such items; rather they should all be set to something like "None" which is treated as a separate factor level, rather than "NA".

This may be the case for the following variables:

Reassign Alley NAs to None

`Alley`: Only some houses are built with an alley behind them, most are not. This is a publically visible feature, the determination of which doesn't require entry to the house. Clearly, the large number of NA values here (1369) indicate that that there is no alley. The non-NA values indicate that an existing alley is "Gravel" or "Paved." We will replace "NA" with "None".

Reassign FireplaceQu NAs to None

`FireplaceQu`: In the train dataset, there are 690 houses with zero fireplaces; for each of these, the rating of Fireplace Quality is "NA". Clearly this is "Not Applicable" (rather than unknown) so we shall change it to "None" rather than "NA", as above.

Reassign Garage NAs to None

In the train dataset, there are 81 houses with no garage; for each of these, the following categorical variables are set to "NA".

- `GarageType`
- `GarageFinish`
- `GarageQual`
- `GarageCond`

Clearly this is "Not Applicable" (rather than unknown) so we shall change it to "None" rather than "NA", as above.

Reassign GarageYrBlt to house's YearBuilt, in case where there is no garage

Additionally, there is a fifth garage variable, `GarageYrBlt`, which is numeric – the year the garage was built. Here it is NA because there is no garage. We will set it to the year in which the house was built:

```
train.df$GarageYrBlt[is.na(train.df$GarageYrBlt)] <- train.df$YearBuilt[is.na(train.df$GarageYrBlt)]
### do the same for the test data:
test.df$GarageYrBlt[is.na(test.df$GarageYrBlt)] <- test.df$YearBuilt[is.na(test.df$GarageYrBlt)]
```

In the test data set, there is one case where the following Numeric variables are set to "NA" :

- `GarageCars`
- `GarageArea`

We will set them to zero, as it appears that the property in question has no garage.

Fix GarageCars and GarageArea

```
summary(test.df$GarageCars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
## 0.00000 1.00000 2.00000 1.76612 2.00000 5.00000       1
```

```
test.df$GarageCars[is.na(test.df$GarageCars)] <- 0
##class(test.df$GarageCars)<-"integer"
summary(test.df$GarageCars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.00000 1.00000 2.00000 1.76491 2.00000 5.00000
```

```
summary(test.df$GarageArea)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    0.000  318.000  480.000  472.769  576.000 1488.000        1
```

```
test.df$GarageArea[is.na(test.df$GarageArea)] <- 0
##class(test.df$GarageArea)<-"integer"
summary(test.df$GarageArea)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000  317.500  480.000  472.445  576.000 1488.000
```

Pools and other items

- Only 7 houses have pools. The other 1453 all have "NA" value for `PoolQC`. We'll change to None.
- Similarly, only 241 houses have a `Fence`, with the other 1179 marked "NA". We'll change to None.
- Finally, only 54 houses have a `MiscFeature` (e.g., a shed or a second garage), with the other 1406 marked "NA". We'll add these items to the list.

No Basements

There are 37 houses with no basements, which is indicated by `TotalBsmtSF`=0 .
For such houses, the following categorical variables are set to NA:

- `BsmtQual`
- `BsmtCond`
- `BsmtExposure`
- `BsmtFinType1`
- `BsmtFinType2`

Additionally, in TEST there is one house without basement, for which the following numerical variables are set to NA (rather than zero):

- BsmtFinSF1
- BsmtFinSF2
- BsmtUnfSF
- TotalBsmtSF
- BsmtFullBath
- BsmtHalfBath

We will set them to zero:

```
basementlist = c("BsmtFinSF1","BsmtFinSF2","BsmtUnfSF","TotalBsmtSF","BsmtFullBath","BsmtHalfBath")
for (g in basementlist) {
  print(g)
  print(summary(test.df[[g]]))
  test.df[[g]][is.na(test.df[[g]])] <- 0
  ##class(test.df[[g]]) <-"integer"
  print(summary(test.df[[g]]))
  print("-----------------------------------------------")
}
```

```
## [1] "BsmtFinSF1"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    0.000    0.000  350.500  439.204  753.500 4010.000        1
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000  350.000  438.903  752.000 4010.000
## [1] "-----------------------------------------------"
## [1] "BsmtFinSF2"
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##    0.0000    0.0000    0.0000   52.6193    0.0000 1526.0000         1
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0000    0.0000    0.0000   52.5833    0.0000 1526.0000
## [1] "-----------------------------------------------"
## [1] "BsmtUnfSF"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    0.000  219.250  460.000  554.295  797.750 2140.000        1
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000  219.000  460.000  553.915  797.500 2140.000
## [1] "-----------------------------------------------"
## [1] "TotalBsmtSF"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00  784.00  988.00 1046.12 1305.00 5095.00       1
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0   784.0   988.0  1045.4  1304.0  5095.0
## [1] "-----------------------------------------------"
## [1] "BsmtFullBath"
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
## 0.000000 0.000000 0.000000 0.434454 1.000000 3.000000        2
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## 0.000000 0.000000 0.000000 0.433859 1.000000 3.000000
## [1] "-----------------------------------------------"
## [1] "BsmtHalfBath"
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
## 0.0000000 0.0000000 0.0000000 0.0652025 0.0000000 2.0000000         2
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
```

```
## 0.0000000 0.0000000 0.0000000 0.0651131 0.0000000 2.0000000
## [1] "----------------------------------------------"
```

Fix Masonry Veneer: MasVnrType (categorical), MasVnrArea (numerical)

There are 8 houses for which these variables are set to NA, suggesting the true value is unknown. Additionally there are about 860 cases where MasVnrType is already set to "None" and MasVnrArea is already set to zero. We attempted using the MICE multiple imputation (below) but it failed. Thus we will set these 8 cases to "None" or 0 as appropriate.

```
summary(train.df$MasVnrType)
```

```
##  BrkCmn BrkFace    None   Stone    NA's
##      15     445     864     128       8
```

```
train.df$MasVnrType[is.na(train.df$MasVnrType)] <- "None"
summary(train.df$MasVnrType)
```

```
##  BrkCmn BrkFace    None   Stone
##      15     445     872     128
```

```
summary(train.df$MasVnrArea)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    0.000    0.000    0.000  103.685  166.000 1600.000        8
```

```
train.df$MasVnrArea[is.na(train.df$MasVnrArea)] <- 0
#class(train.df$MasVnrArea)<-"integer"
summary(train.df$MasVnrArea)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    0.000  103.117  164.250 1600.000
```

```
#### do the same for test
summary(test.df$MasVnrType)
```

```
##  BrkCmn BrkFace    None   Stone    NA's
##      10     434     878     121      16
```

```
test.df$MasVnrType[is.na(test.df$MasVnrType)] <- "None"
summary(test.df$MasVnrType)
```

```
##  BrkCmn BrkFace    None   Stone
##      10     434     894     121
```

```
summary(test.df$MasVnrArea)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##    0.000    0.000    0.000  100.709  164.000 1290.000       15
```

```r
test.df$MasVnrArea[is.na(test.df$MasVnrArea)] <- 0
#class(test.df$MasVnrArea)<-"integer"
summary(test.df$MasVnrArea)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    0.0000    0.0000    0.0000   99.6737  162.0000 1290.0000
```

Electrical

There is a single case where "Electrical" has value NA. Multiple imputation (below) failed, so we add it to this list to be set to None.

On the test data set, multiple imputation (MICE) failed. The following categorical variables have a handful of NA values in test.df, but none in train.df . So, we add them to the list below for replacement with "None".

- MSZoning
- Utilities
- Exterior1st
- Exterior2nd
- KitchenQual
- Functional
- SaleType

List of categorical variables for which "NA" is to be replaced by level "None"

```r
changelist = c("Alley", "FireplaceQu",
               "GarageType", "GarageFinish", "GarageQual", "GarageCond",
               "PoolQC", "Fence", "MiscFeature",
               "BsmtQual","BsmtCond","BsmtExposure","BsmtFinType1","BsmtFinType2",
               "Electrical",
               "MSZoning", "Utilities", "Exterior1st", "Exterior2nd" ,
               "KitchenQual", "Functional", "SaleType" )
```

Loop through above list

```r
for (g in changelist) {

print(paste("TRAIN: CHANGING NAs for ", g))

print(levels(train.df[[g]]))
print(table(train.df[[g]],useNA = "ifany"))

### We need to add one more level on the factor
(levels(train.df[[g]]) <- c(levels(train.df[[g]]),"None"))
print(table(train.df[[g]],useNA = "ifany"))

### Reassign the NAs to None
(train.df[[g]][is.na(train.df[[g]])] <- "None")
```

```r
### See what we have now
print(table(train.df[[g]],useNA = "ifany"))

print("_____________________________________________________________________________________________")

print(paste("***TEST: CHANGING NAs for ", g, "***"))

print(levels(test.df[[g]]))
print(table(test.df[[g]],useNA = "ifany"))

### We need to add one more level on the factor
(levels(test.df[[g]]) <- c(levels(test.df[[g]]),"None"))
print(table(test.df[[g]],useNA = "ifany"))

### Reassign the NAs to None
(test.df[[g]][is.na(test.df[[g]])] <- "None")

### See what we have now
print(table(test.df[[g]],useNA = "ifany"))

print("=============================================================================================")
}
```

```
## [1] "TRAIN: CHANGING NAs for  Alley"
## [1] "Grvl" "Pave"
##
## Grvl Pave <NA>
##   50   41 1369
##
## Grvl Pave None <NA>
##   50   41    0 1369
##
## Grvl Pave None
##   50   41 1369
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  Alley ***"
## [1] "Grvl" "Pave"
##
## Grvl Pave <NA>
##   70   37 1352
##
## Grvl Pave None <NA>
##   70   37    0 1352
##
## Grvl Pave None
##   70   37 1352
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  FireplaceQu"
## [1] "Ex" "Fa" "Gd" "Po" "TA"
##
##   Ex   Fa   Gd   Po   TA <NA>
##   24   33  380   20  313  690
##
```

```
##   Ex   Fa   Gd   Po   TA None <NA>
##   24   33  380   20  313    0  690
##
##   Ex   Fa   Gd   Po   TA None
##   24   33  380   20  313  690
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  FireplaceQu ***"
## [1] "Ex" "Fa" "Gd" "Po" "TA"
##
##   Ex   Fa   Gd   Po   TA <NA>
##   19   41  364   26  279  730
##
##   Ex   Fa   Gd   Po   TA None <NA>
##   19   41  364   26  279    0  730
##
##   Ex   Fa   Gd   Po   TA None
##   19   41  364   26  279  730
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  GarageType"
## [1] "2Types"  "Attchd"  "Basment" "BuiltIn" "CarPort" "Detchd"
##
##  2Types  Attchd Basment BuiltIn CarPort  Detchd    <NA>
##       6     870      19      88       9     387      81
##
##  2Types  Attchd Basment BuiltIn CarPort  Detchd    None    <NA>
##       6     870      19      88       9     387       0      81
##
##  2Types  Attchd Basment BuiltIn CarPort  Detchd    None
##       6     870      19      88       9     387      81
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  GarageType ***"
## [1] "2Types"  "Attchd"  "Basment" "BuiltIn" "CarPort" "Detchd"
##
##  2Types  Attchd Basment BuiltIn CarPort  Detchd    <NA>
##      17     853      17      98       6     392      76
##
##  2Types  Attchd Basment BuiltIn CarPort  Detchd    None    <NA>
##      17     853      17      98       6     392       0      76
##
##  2Types  Attchd Basment BuiltIn CarPort  Detchd    None
##      17     853      17      98       6     392      76
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  GarageFinish"
## [1] "Fin" "RFn" "Unf"
##
##  Fin  RFn  Unf <NA>
##  352  422  605   81
##
##  Fin  RFn  Unf None <NA>
##  352  422  605    0   81
##
##  Fin  RFn  Unf None
##  352  422  605   81
## [1] "_____________________________________________________________________________________________"
```

```
## [1] "***TEST: CHANGING NAs for  GarageFinish ***"
## [1] "Fin" "RFn" "Unf"
##
##  Fin  RFn  Unf <NA>
##  367  389  625   78
##
##  Fin  RFn  Unf None <NA>
##  367  389  625    0   78
##
##  Fin  RFn  Unf None
##  367  389  625   78
## [1] "==========================================================================================="
## [1] "TRAIN: CHANGING NAs for  GarageQual"
## [1] "Ex" "Fa" "Gd" "Po" "TA"
##
##   Ex   Fa   Gd   Po   TA <NA>
##    3   48   14    3 1311   81
##
##   Ex   Fa   Gd   Po   TA None <NA>
##    3   48   14    3 1311    0   81
##
##   Ex   Fa   Gd   Po   TA None
##    3   48   14    3 1311   81
## [1] "-------------------------------------------------------------------------------------------"
## [1] "***TEST: CHANGING NAs for  GarageQual ***"
## [1] "Fa" "Gd" "Po" "TA"
##
##   Fa   Gd   Po   TA <NA>
##   76   10    2 1293   78
##
##   Fa   Gd   Po   TA None <NA>
##   76   10    2 1293    0   78
##
##   Fa   Gd   Po   TA None
##   76   10    2 1293   78
## [1] "==========================================================================================="
## [1] "TRAIN: CHANGING NAs for  GarageCond"
## [1] "Ex" "Fa" "Gd" "Po" "TA"
##
##   Ex   Fa   Gd   Po   TA <NA>
##    2   35    9    7 1326   81
##
##   Ex   Fa   Gd   Po   TA None <NA>
##    2   35    9    7 1326    0   81
##
##   Ex   Fa   Gd   Po   TA None
##    2   35    9    7 1326   81
## [1] "-------------------------------------------------------------------------------------------"
## [1] "***TEST: CHANGING NAs for  GarageCond ***"
## [1] "Ex" "Fa" "Gd" "Po" "TA"
##
##   Ex   Fa   Gd   Po   TA <NA>
##    1   39    6    7 1328   78
##
```

```
##   Ex   Fa   Gd   Po   TA None <NA>
##    1   39    6    7 1328    0   78
##
##   Ex   Fa   Gd   Po   TA None
##    1   39    6    7 1328   78
## [1] "==========================================================================================="
## [1] "TRAIN: CHANGING NAs for  PoolQC"
## [1] "Ex" "Fa" "Gd"
##
##   Ex   Fa   Gd <NA>
##    2    2    3 1453
##
##   Ex   Fa   Gd None <NA>
##    2    2    3    0 1453
##
##   Ex   Fa   Gd None
##    2    2    3 1453
## [1] "___________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  PoolQC ***"
## [1] "Ex" "Gd"
##
##   Ex   Gd <NA>
##    2    1 1456
##
##   Ex   Gd None <NA>
##    2    1    0 1456
##
##   Ex   Gd None
##    2    1 1456
## [1] "==========================================================================================="
## [1] "TRAIN: CHANGING NAs for  Fence"
## [1] "GdPrv" "GdWo"  "MnPrv" "MnWw"
##
## GdPrv  GdWo MnPrv  MnWw  <NA>
##    59    54   157    11  1179
##
## GdPrv  GdWo MnPrv  MnWw  None  <NA>
##    59    54   157    11     0  1179
##
## GdPrv  GdWo MnPrv  MnWw  None
##    59    54   157    11  1179
## [1] "___________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  Fence ***"
## [1] "GdPrv" "GdWo"  "MnPrv" "MnWw"
##
## GdPrv  GdWo MnPrv  MnWw  <NA>
##    59    58   172     1  1169
##
## GdPrv  GdWo MnPrv  MnWw  None  <NA>
##    59    58   172     1     0  1169
##
## GdPrv  GdWo MnPrv  MnWw  None
##    59    58   172     1  1169
## [1] "==========================================================================================="
```

```
## [1] "TRAIN: CHANGING NAs for  MiscFeature"
## [1] "Gar2" "Othr" "Shed" "TenC"
##
## Gar2 Othr Shed TenC <NA>
##    2    2   49    1 1406
##
## Gar2 Othr Shed TenC None <NA>
##    2    2   49    1    0 1406
##
## Gar2 Othr Shed TenC None
##    2    2   49    1 1406
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  MiscFeature ***"
## [1] "Gar2" "Othr" "Shed"
##
## Gar2 Othr Shed <NA>
##    3    2   46 1408
##
## Gar2 Othr Shed None <NA>
##    3    2   46    0 1408
##
## Gar2 Othr Shed None
##    3    2   46 1408
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  BsmtQual"
## [1] "Ex" "Fa" "Gd" "TA"
##
##   Ex   Fa   Gd   TA <NA>
##  121   35  618  649   37
##
##   Ex   Fa   Gd   TA None <NA>
##  121   35  618  649    0   37
##
##   Ex   Fa   Gd   TA None
##  121   35  618  649   37
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  BsmtQual ***"
## [1] "Ex" "Fa" "Gd" "TA"
##
##   Ex   Fa   Gd   TA <NA>
##  137   53  591  634   44
##
##   Ex   Fa   Gd   TA None <NA>
##  137   53  591  634    0   44
##
##   Ex   Fa   Gd   TA None
##  137   53  591  634   44
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  BsmtCond"
## [1] "Fa" "Gd" "Po" "TA"
##
##   Fa   Gd   Po   TA <NA>
##   45   65    2 1311   37
##
```

```
##   Fa   Gd   Po   TA None <NA>
##   45   65    2 1311    0   37
##
##   Fa   Gd   Po   TA None
##   45   65    2 1311   37
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  BsmtCond ***"
## [1] "Fa" "Gd" "Po" "TA"
##
##   Fa   Gd   Po   TA <NA>
##   59   57    3 1295   45
##
##   Fa   Gd   Po   TA None <NA>
##   59   57    3 1295    0   45
##
##   Fa   Gd   Po   TA None
##   59   57    3 1295   45
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  BsmtExposure"
## [1] "Av" "Gd" "Mn" "No"
##
##   Av   Gd   Mn   No <NA>
##  221  134  114  953   38
##
##   Av   Gd   Mn   No None <NA>
##  221  134  114  953    0   38
##
##   Av   Gd   Mn   No None
##  221  134  114  953   38
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  BsmtExposure ***"
## [1] "Av" "Gd" "Mn" "No"
##
##   Av   Gd   Mn   No <NA>
##  197  142  125  951   44
##
##   Av   Gd   Mn   No None <NA>
##  197  142  125  951    0   44
##
##   Av   Gd   Mn   No None
##  197  142  125  951   44
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  BsmtFinType1"
## [1] "ALQ" "BLQ" "GLQ" "LwQ" "Rec" "Unf"
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf <NA>
##  220  148  418   74  133  430   37
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf None <NA>
##  220  148  418   74  133  430    0   37
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf None
##  220  148  418   74  133  430   37
## [1] "_____________________________________________________________________________________________"
```

```
## [1] "***TEST: CHANGING NAs for  BsmtFinType1 ***"
## [1] "ALQ" "BLQ" "GLQ" "LwQ" "Rec" "Unf"
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf <NA>
##  209  121  431   80  155  421   42
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf None <NA>
##  209  121  431   80  155  421    0   42
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf None
##  209  121  431   80  155  421   42
## [1] "==========================================================================================="
## [1] "TRAIN: CHANGING NAs for  BsmtFinType2"
## [1] "ALQ" "BLQ" "GLQ" "LwQ" "Rec" "Unf"
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf <NA>
##   19   33   14   46   54 1256   38
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf None <NA>
##   19   33   14   46   54 1256    0   38
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf None
##   19   33   14   46   54 1256   38
## [1] "___________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  BsmtFinType2 ***"
## [1] "ALQ" "BLQ" "GLQ" "LwQ" "Rec" "Unf"
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf <NA>
##   33   35   20   41   51 1237   42
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf None <NA>
##   33   35   20   41   51 1237    0   42
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf None
##   33   35   20   41   51 1237   42
## [1] "==========================================================================================="
## [1] "TRAIN: CHANGING NAs for  Electrical"
## [1] "FuseA" "FuseF" "FuseP" "Mix"   "SBrkr"
##
## FuseA FuseF FuseP   Mix SBrkr  <NA>
##    94    27     3     1  1334     1
##
## FuseA FuseF FuseP   Mix SBrkr  None  <NA>
##    94    27     3     1  1334     0     1
##
## FuseA FuseF FuseP   Mix SBrkr  None
##    94    27     3     1  1334     1
## [1] "___________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  Electrical ***"
## [1] "FuseA" "FuseF" "FuseP" "SBrkr"
##
## FuseA FuseF FuseP SBrkr
##    94    23     5  1337
##
```

```
## FuseA FuseF FuseP SBrkr  None
##    94    23     5  1337     0
##
## FuseA FuseF FuseP SBrkr  None
##    94    23     5  1337     0
## [1] "==========================================================================================="
## [1] "TRAIN: CHANGING NAs for  MSZoning"
## [1] "C (all)" "FV"      "RH"      "RL"      "RM"
##
## C (all)      FV      RH      RL      RM
##      10      65      16    1151     218
##
## C (all)      FV      RH      RL      RM    None
##      10      65      16    1151     218       0
##
## C (all)      FV      RH      RL      RM    None
##      10      65      16    1151     218       0
## [1] "-------------------------------------------------------------------------------------------"
## [1] "***TEST: CHANGING NAs for  MSZoning ***"
## [1] "C (all)" "FV"      "RH"      "RL"      "RM"
##
## C (all)      FV      RH      RL      RM    <NA>
##      15      74      10    1114     242       4
##
## C (all)      FV      RH      RL      RM    None    <NA>
##      15      74      10    1114     242       0       4
##
## C (all)      FV      RH      RL      RM    None
##      15      74      10    1114     242       4
## [1] "==========================================================================================="
## [1] "TRAIN: CHANGING NAs for  Utilities"
## [1] "AllPub" "NoSeWa"
##
## AllPub NoSeWa
##   1459      1
##
## AllPub NoSeWa   None
##   1459      1      0
##
## AllPub NoSeWa   None
##   1459      1      0
## [1] "-------------------------------------------------------------------------------------------"
## [1] "***TEST: CHANGING NAs for  Utilities ***"
## [1] "AllPub"
##
## AllPub   <NA>
##   1457      2
##
## AllPub   None   <NA>
##   1457      0      2
##
## AllPub   None
##   1457      2
## [1] "==========================================================================================="
```

```
## [1] "TRAIN: CHANGING NAs for  Exterior1st"
##  [1] "AsbShng" "AsphShn" "BrkComm" "BrkFace" "CBlock"  "CemntBd" "HdBoard"
##  [8] "ImStucc" "MetalSd" "Plywood" "Stone"   "Stucco"  "VinylSd" "Wd Sdng"
## [15] "WdShing"
##
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard ImStucc MetalSd Plywood
##      20       1       2      50       1      61     222       1     220     108
##   Stone  Stucco VinylSd Wd Sdng WdShing
##       2      25     515     206      26
##
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard ImStucc MetalSd Plywood
##      20       1       2      50       1      61     222       1     220     108
##   Stone  Stucco VinylSd Wd Sdng WdShing    None
##       2      25     515     206      26       0
##
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard ImStucc MetalSd Plywood
##      20       1       2      50       1      61     222       1     220     108
##   Stone  Stucco VinylSd Wd Sdng WdShing    None
##       2      25     515     206      26       0
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  Exterior1st ***"
##  [1] "AsbShng" "AsphShn" "BrkComm" "BrkFace" "CBlock"  "CemntBd" "HdBoard"
##  [8] "MetalSd" "Plywood" "Stucco"  "VinylSd" "Wd Sdng" "WdShing"
##
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard MetalSd Plywood  Stucco
##      24       1       4      37       1      65     220     230     113      18
## VinylSd Wd Sdng WdShing    <NA>
##     510     205      30       1
##
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard MetalSd Plywood  Stucco
##      24       1       4      37       1      65     220     230     113      18
## VinylSd Wd Sdng WdShing    None    <NA>
##     510     205      30       0       1
##
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard MetalSd Plywood  Stucco
##      24       1       4      37       1      65     220     230     113      18
## VinylSd Wd Sdng WdShing    None
##     510     205      30       1
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  Exterior2nd"
##  [1] "AsbShng" "AsphShn" "Brk Cmn" "BrkFace" "CBlock"  "CmentBd" "HdBoard"
##  [8] "ImStucc" "MetalSd" "Other"   "Plywood" "Stone"   "Stucco"  "VinylSd"
## [15] "Wd Sdng" "Wd Shng"
##
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd   Other
##      20       3       7      25       1      60     207      10     214       1
## Plywood   Stone  Stucco VinylSd Wd Sdng Wd Shng
##     142       5      26     504     197      38
##
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd   Other
##      20       3       7      25       1      60     207      10     214       1
## Plywood   Stone  Stucco VinylSd Wd Sdng Wd Shng    None
##     142       5      26     504     197      38       0
##
```

```
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd   Other
##      20       3       7      25       1      60     207      10     214       1
## Plywood   Stone  Stucco VinylSd Wd Sdng Wd Shng    None
##     142       5      26     504     197      38       0
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  Exterior2nd ***"
##  [1] "AsbShng" "AsphShn" "Brk Cmn" "BrkFace" "CBlock"  "CmentBd" "HdBoard"
##  [8] "ImStucc" "MetalSd" "Plywood" "Stone"   "Stucco"  "VinylSd" "Wd Sdng"
## [15] "Wd Shng"
##
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd Plywood
##      18       1      15      22       2      66     199       5     233     128
##   Stone  Stucco VinylSd Wd Sdng Wd Shng    <NA>
##       1      21     510     194      43       1
##
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd Plywood
##      18       1      15      22       2      66     199       5     233     128
##   Stone  Stucco VinylSd Wd Sdng Wd Shng    None    <NA>
##       1      21     510     194      43       0       1
##
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd Plywood
##      18       1      15      22       2      66     199       5     233     128
##   Stone  Stucco VinylSd Wd Sdng Wd Shng    None
##       1      21     510     194      43       1
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  KitchenQual"
## [1] "Ex" "Fa" "Gd" "TA"
##
##  Ex  Fa  Gd  TA
## 100  39 586 735
##
##   Ex   Fa   Gd   TA None
##  100   39  586  735    0
##
##   Ex   Fa   Gd   TA None
##  100   39  586  735    0
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  KitchenQual ***"
## [1] "Ex" "Fa" "Gd" "TA"
##
##   Ex   Fa   Gd   TA <NA>
##  105   31  565  757    1
##
##   Ex   Fa   Gd   TA None <NA>
##  105   31  565  757    0    1
##
##   Ex   Fa   Gd   TA None
##  105   31  565  757    1
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  Functional"
## [1] "Maj1" "Maj2" "Min1" "Min2" "Mod"  "Sev"  "Typ"
##
## Maj1 Maj2 Min1 Min2  Mod  Sev  Typ
##   14    5   31   34   15    1 1360
```

```
##
## Maj1 Maj2 Min1 Min2  Mod  Sev  Typ None
##   14    5   31   34   15    1 1360    0
##
## Maj1 Maj2 Min1 Min2  Mod  Sev  Typ None
##   14    5   31   34   15    1 1360    0
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  Functional ***"
## [1] "Maj1" "Maj2" "Min1" "Min2" "Mod"  "Sev"  "Typ"
##
## Maj1 Maj2 Min1 Min2  Mod  Sev  Typ <NA>
##    5    4   34   36   20    1 1357    2
##
## Maj1 Maj2 Min1 Min2  Mod  Sev  Typ None <NA>
##    5    4   34   36   20    1 1357    0    2
##
## Maj1 Maj2 Min1 Min2  Mod  Sev  Typ None
##    5    4   34   36   20    1 1357    2
## [1] "=============================================================================================="
## [1] "TRAIN: CHANGING NAs for  SaleType"
## [1] "COD"   "Con"   "ConLD" "ConLI" "ConLw" "CWD"   "New"   "Oth"   "WD"
##
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD
##    43     2     9     5     5     4   122     3  1267
##
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD  None
##    43     2     9     5     5     4   122     3  1267     0
##
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD  None
##    43     2     9     5     5     4   122     3  1267     0
## [1] "_____________________________________________________________________________________________"
## [1] "***TEST: CHANGING NAs for  SaleType ***"
## [1] "COD"   "Con"   "ConLD" "ConLI" "ConLw" "CWD"   "New"   "Oth"   "WD"
##
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD  <NA>
##    44     3    17     4     3     8   117     4  1258     1
##
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD  None  <NA>
##    44     3    17     4     3     8   117     4  1258     0     1
##
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD  None
##    44     3    17     4     3     8   117     4  1258     1
## [1] "=============================================================================================="
```

Impute LotFrontage - set to zero

Numerical variable "LotFrontage" is causing problems with the imputation – it's generating a "computationally singular" problem under multiple imputation. So, I'll manually set its value to zero, because it suggests that the lot on which the property is situated may not have any street frontage.

```
summary(train.df$LotFrontage)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   21.00   59.00   69.00   70.05   80.00  313.00     259
```

```
train.df$LotFrontage[is.na(train.df$LotFrontage)]<-0
#class(train.df$LotFrontage)<-"integer"                  ## otherwise it flipped to numeric
summary(train.df$LotFrontage)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000  42.0000  63.0000  57.6233  79.0000 313.0000
```

```
## do the same for test data
summary(test.df$LotFrontage)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
##  21.0000  58.0000  67.0000  68.5804  80.0000 200.0000      227
```

```
test.df$LotFrontage[is.na(test.df$LotFrontage)]<-0
#class(test.df$LotFrontage)<-"integer"                  ## otherwise it flipped to numeric
summary(test.df$LotFrontage)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   0.0000  44.0000  63.0000  57.9102  78.0000 200.0000
```

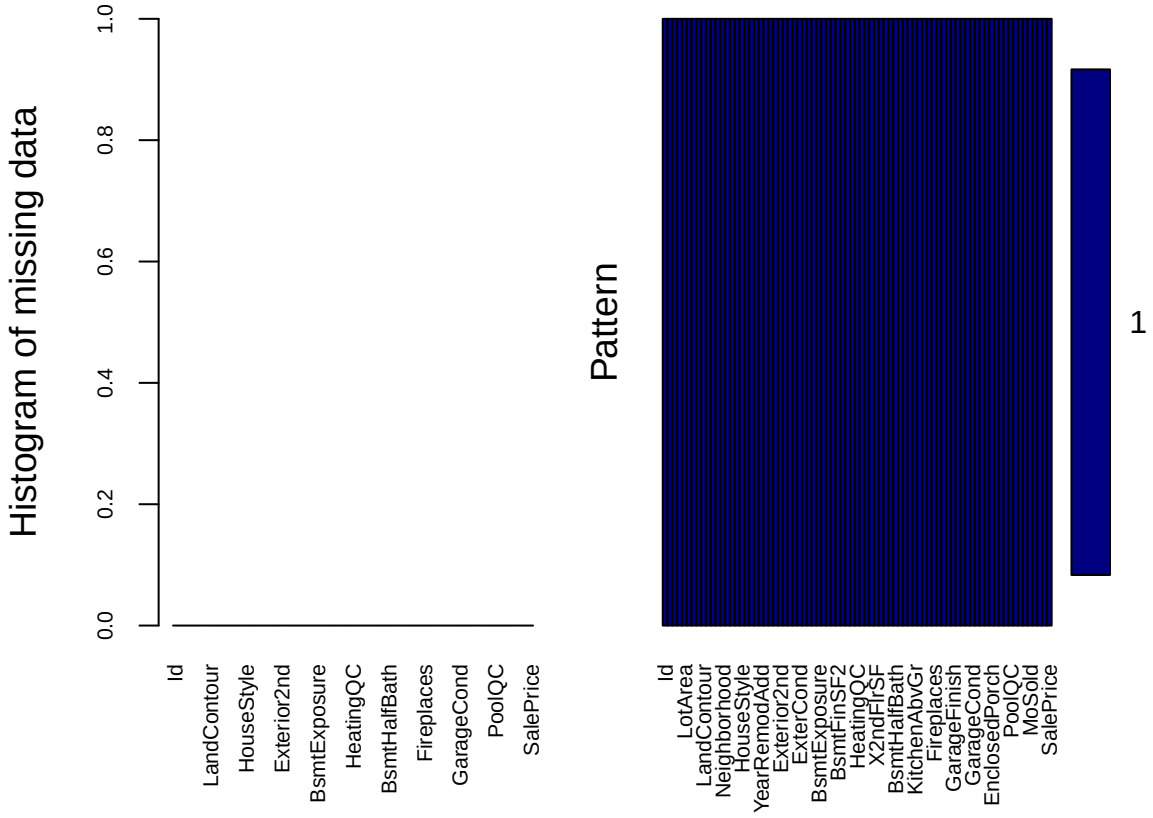
AUTOMATED DATA IMPUTATION

Note: The below usage of MICE did not work – it gave numerous errors.
Therefore, I manually made the above changes to the data as detailed above.
At this point, there are no more NAs in either dataset.

Let's do data imputation for columns with missing values.

Which columns have missing values, and what is a missing pattern?

Let's leverage VIM package to get this information:




```
##
##  Variables sorted by number of missings:
##       Variable Count
##             Id     0
##     MSSubClass     0
##       MSZoning     0
##    LotFrontage     0
##        LotArea     0
##         Street     0
##          Alley     0
##       LotShape     0
##    LandContour     0
##      Utilities     0
```

```
##      LotConfig     0
##      LandSlope     0
##   Neighborhood     0
##     Condition1     0
##     Condition2     0
##       BldgType     0
##     HouseStyle     0
##    OverallQual     0
##    OverallCond     0
##      YearBuilt     0
##   YearRemodAdd     0
##      RoofStyle     0
##       RoofMatl     0
##    Exterior1st     0
##    Exterior2nd     0
##     MasVnrType     0
##     MasVnrArea     0
##      ExterQual     0
##      ExterCond     0
##     Foundation     0
##       BsmtQual     0
##       BsmtCond     0
##   BsmtExposure     0
##   BsmtFinType1     0
##     BsmtFinSF1     0
##   BsmtFinType2     0
##     BsmtFinSF2     0
##      BsmtUnfSF     0
##    TotalBsmtSF     0
##        Heating     0
##      HeatingQC     0
##     CentralAir     0
##     Electrical     0
##      X1stFlrSF     0
##      X2ndFlrSF     0
##   LowQualFinSF     0
##      GrLivArea     0
##   BsmtFullBath     0
##   BsmtHalfBath     0
##       FullBath     0
##       HalfBath     0
##   BedroomAbvGr     0
##   KitchenAbvGr     0
##    KitchenQual     0
##   TotRmsAbvGrd     0
##     Functional     0
##     Fireplaces     0
##    FireplaceQu     0
##     GarageType     0
##    GarageYrBlt     0
##   GarageFinish     0
##     GarageCars     0
##     GarageArea     0
##     GarageQual     0
```

```
##     GarageCond     0
##     PavedDrive     0
##     WoodDeckSF     0
##    OpenPorchSF     0
##  EnclosedPorch     0
##     X3SsnPorch     0
##    ScreenPorch     0
##       PoolArea     0
##         PoolQC     0
##          Fence     0
##    MiscFeature     0
##        MiscVal     0
##         MoSold     0
##         YrSold     0
##       SaleType     0
##  SaleCondition     0
##      SalePrice     0
```

There are now NO columns remaining (in the TRAINING data set) with NA values, as we have manually replaced everything as discussed above. At the same time, we repeated such changes on the TEST data set.

Let's use the mice package to impute missing values

MICE: "Multivariate Imputation by Chained Equations"

(Note: this did not work, when there were still variables to be imputed.)

The **mice** package implements a method to deal with missing data.
The package creates multiple imputations (replacement values) for multivariate missing data.
The method is based on Fully Conditional Specification, where each incomplete variable is imputed by a separate model.
The MICE algorithm can impute mixes of continuous, binary, unordered categorical and ordered categorical data.
In addition, MICE can impute continuous two-level data, and maintain consistency between imputations by means of passive imputation.
Many diagnostic plots are implemented to inspect the quality of the imputations.

run MICE on TRAIN data set

```r
library(mice)
```

```
## Loading required package: lattice
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
##
## Attaching package: 'mice'
```

```
## The following object is masked from 'package:tidyr':
##
##     complete

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```r
save.train.df <- train.df
comp.data <- mice(train.df,m=2,maxit=10,seed=500)
```

```
##
##  iter imp variable
##   1   1
##   1   2
##   2   1
##   2   2
##   3   1
##   3   2
##   4   1
##   4   2
##   5   1
##   5   2
##   6   1
##   6   2
##   7   1
##   7   2
##   8   1
##   8   2
##   9   1
##   9   2
##   10   1
##   10   2
```

```r
train.df = complete(comp.data)
### nothing changed -- because we did all imputations manually, above
all.equal(train.df,save.train.df)
```

```
## [1] TRUE
```

run MICE on TEST data set

Earlier this failed with "too many weights." Thus, each imputation of missing data was performed manually, as discussed above.
Here we confirm that there is nothing further which requires imputation:

```
library(mice)
save.test.df <- test.df
test.comp.data <- mice(test.df,m=2,maxit=10,seed=500)
```

```
##
##  iter imp variable
##   1   1
##   1   2
##   2   1
##   2   2
##   3   1
##   3   2
##   4   1
##   4   2
##   5   1
##   5   2
##   6   1
##   6   2
##   7   1
##   7   2
##   8   1
##   8   2
##   9   1
##   9   2
##   10   1
##   10   2
```

```
test.df = complete(test.comp.data)
### What's changed?
all.equal(train.df,save.train.df)
```

```
## [1] TRUE
```

I want you to do the following:

5 points. Descriptive and Inferential Statistics.

Provide *univariate descriptive statistics* and *appropriate plots* for the training data set.

```r
summary(train.df)
```

```
##        Id            MSSubClass     MSZoning     LotFrontage
##  Min.   :   1.00   20     :536   C (all):  10   Min.   :  0.0000
##  1st Qu.: 365.75   60     :299   FV     :  65   1st Qu.: 42.0000
##  Median : 730.50   50     :144   RH     :  16   Median : 63.0000
##  Mean   : 730.50   120    : 87   RL     :1151   Mean   : 57.6233
##  3rd Qu.:1095.25   30     : 69   RM     : 218   3rd Qu.: 79.0000
##  Max.   :1460.00   160    : 63   None   :   0   Max.   :313.0000
##                    (Other):262
##     LotArea          Street      Alley      LotShape  LandContour  Utilities
##  Min.   :  1300.0   Grvl:   6   Grvl:  50   IR1:484   Bnk:  63    AllPub:1459
##  1st Qu.:  7553.5   Pave:1454   Pave:  41   IR2: 41   HLS:  50    NoSeWa:   1
##  Median :  9478.5               None:1369   IR3: 10   Low:  36    None  :   0
##  Mean   : 10516.8                           Reg:925   Lvl:1311
##  3rd Qu.: 11601.5
##  Max.   :215245.0
##
##    LotConfig    LandSlope   Neighborhood   Condition1     Condition2
##  Corner : 263   Gtl:1382   NAmes  :225   Norm   :1260   Norm   :1445
##  CulDSac:  94   Mod:  65   CollgCr:150   Feedr  :  81   Feedr  :   6
##  FR2    :  47   Sev:  13   OldTown:113   Artery :  48   Artery :   2
##  FR3    :   4              Edwards:100   RRAn   :  26   PosN   :   2
##  Inside :1052              Somerst: 86   PosN   :  19   RRNn   :   2
##                            Gilbert: 79   RRAe   :  11   PosA   :   1
##                            (Other):707   (Other):  15   (Other):   2
##    BldgType      HouseStyle   OverallQual        OverallCond
##  1Fam  :1220   1Story :726   Min.   : 1.00000   Min.   :1.00000
##  2fmCon:  31   2Story :445   1st Qu.: 5.00000   1st Qu.:5.00000
##  Duplex:  52   1.5Fin :154   Median : 6.00000   Median :5.00000
##  Twnhs :  43   SLvl   : 65   Mean   : 6.09932   Mean   :5.57534
##  TwnhsE: 114   SFoyer : 37   3rd Qu.: 7.00000   3rd Qu.:6.00000
##                1.5Unf : 14   Max.   :10.00000   Max.   :9.00000
##                (Other): 19
##    YearBuilt        YearRemodAdd       RoofStyle       RoofMatl
##  Min.   :1872.00   Min.   :1950.00   Flat   :  13   CompShg:1434
##  1st Qu.:1954.00   1st Qu.:1967.00   Gable  :1141   Tar&Grv:  11
##  Median :1973.00   Median :1994.00   Gambrel:  11   WdShngl:   6
##  Mean   :1971.27   Mean   :1984.87   Hip    : 286   WdShake:   5
##  3rd Qu.:2000.00   3rd Qu.:2004.00   Mansard:   7   ClyTile:   1
##  Max.   :2010.00   Max.   :2010.00   Shed   :   2   Membran:   1
##                                                    (Other):   2
##   Exterior1st   Exterior2nd    MasVnrType    MasVnrArea       ExterQual
##  VinylSd:515   VinylSd:504   BrkCmn : 15   Min.   :   0.000   Ex: 52
##  HdBoard:222   MetalSd:214   BrkFace:445   1st Qu.:   0.000   Fa: 14
##  MetalSd:220   HdBoard:207   None   :872   Median :   0.000   Gd:488
##  Wd Sdng:206   Wd Sdng:197   Stone  :128   Mean   : 103.117   TA:906
```

```
##  Plywood:108   Plywood:142                 3rd Qu.: 164.250
##  CemntBd: 61   CmentBd: 60                 Max.   :1600.000
##  (Other):128   (Other):136
##  ExterCond  Foundation  BsmtQual   BsmtCond    BsmtExposure BsmtFinType1
##  Ex:   3   BrkTil:146   Ex  :121   Fa  :  45   Av  :221     ALQ :220
##  Fa:  28   CBlock:634   Fa  : 35   Gd  :  65   Gd  :134     BLQ :148
##  Gd: 146   PConc :647   Gd  :618   Po  :   2   Mn  :114     GLQ :418
##  Po:   1   Slab  : 24   TA  :649   TA  :1311   No  :953     LwQ : 74
##  TA:1282   Stone :  6   None: 37   None:  37   None: 38     Rec :133
##            Wood  :  3                                       Unf :430
##                                                             None: 37
##    BsmtFinSF1      BsmtFinType2   BsmtFinSF2          BsmtUnfSF
##  Min.   :   0.00   ALQ :  19    Min.   :   0.0000   Min.   :   0.00
##  1st Qu.:   0.00   BLQ :  33    1st Qu.:   0.0000   1st Qu.: 223.00
##  Median : 383.50   GLQ :  14    Median :   0.0000   Median : 477.50
##  Mean   : 443.64   LwQ :  46    Mean   :  46.5493   Mean   : 567.24
##  3rd Qu.: 712.25   Rec :  54    3rd Qu.:   0.0000   3rd Qu.: 808.00
##  Max.   :5644.00   Unf :1256    Max.   :1474.0000   Max.   :2336.00
##                    None:  38
##   TotalBsmtSF       Heating     HeatingQC CentralAir Electrical
##  Min.   :   0.00   Floor:   1   Ex:741    N:  95     FuseA:  94
##  1st Qu.: 795.75   GasA :1428   Fa: 49    Y:1365     FuseF:  27
##  Median : 991.50   GasW :  18   Gd:241               FuseP:   3
##  Mean   :1057.43   Grav :   7   Po:  1               Mix  :   1
##  3rd Qu.:1298.25   OthW :   2   TA:428               SBrkr:1334
##  Max.   :6110.00   Wall :   4                        None :   1
##
##    X1stFlrSF         X2ndFlrSF         LowQualFinSF         GrLivArea
##  Min.   : 334.00   Min.   :   0.000   Min.   :  0.00000   Min.   : 334.00
##  1st Qu.: 882.00   1st Qu.:   0.000   1st Qu.:  0.00000   1st Qu.:1129.50
##  Median :1087.00   Median :   0.000   Median :  0.00000   Median :1464.00
##  Mean   :1162.63   Mean   : 346.992   Mean   :  5.84452   Mean   :1515.46
##  3rd Qu.:1391.25   3rd Qu.: 728.000   3rd Qu.:  0.00000   3rd Qu.:1776.75
##  Max.   :4692.00   Max.   :2065.000   Max.   :572.00000   Max.   :5642.00
##
##   BsmtFullBath       BsmtHalfBath          FullBath          HalfBath
##  Min.   :0.000000   Min.   :0.0000000   Min.   :0.00000   Min.   :0.000000
##  1st Qu.:0.000000   1st Qu.:0.0000000   1st Qu.:1.00000   1st Qu.:0.000000
##  Median :0.000000   Median :0.0000000   Median :2.00000   Median :0.000000
##  Mean   :0.425342   Mean   :0.0575342   Mean   :1.56507   Mean   :0.382877
##  3rd Qu.:1.000000   3rd Qu.:0.0000000   3rd Qu.:2.00000   3rd Qu.:1.000000
##  Max.   :3.000000   Max.   :2.0000000   Max.   :3.00000   Max.   :2.000000
##
##   BedroomAbvGr      KitchenAbvGr     KitchenQual  TotRmsAbvGrd
##  Min.   :0.00000   Min.   :0.00000   Ex  :100    Min.   : 2.00000
##  1st Qu.:2.00000   1st Qu.:1.00000   Fa  : 39    1st Qu.: 5.00000
##  Median :3.00000   Median :1.00000   Gd  :586    Median : 6.00000
##  Mean   :2.86644   Mean   :1.04658   TA  :735    Mean   : 6.51781
##  3rd Qu.:3.00000   3rd Qu.:1.00000   None:  0    3rd Qu.: 7.00000
##  Max.   :8.00000   Max.   :3.00000               Max.   :14.00000
##
##    Functional     Fireplaces       FireplaceQu   GarageType   GarageYrBlt
##  Typ    :1360   Min.   :0.000000   Ex  : 24    2Types :  6   Min.   :1872.00
##  Min2   :  34   1st Qu.:0.000000   Fa  : 33    Attchd :870   1st Qu.:1959.00
```

```
##  Min1   :  31   Median :1.000000   Gd  :380    Basment: 19   Median :1978.00
##  Mod    :  15   Mean   :0.613014   Po  : 20    BuiltIn: 88   Mean   :1976.51
##  Maj1   :  14   3rd Qu.:1.000000   TA  :313    CarPort:  9   3rd Qu.:2001.00
##  Maj2   :   5   Max.   :3.000000   None:690    Detchd :387   Max.   :2010.00
##  (Other):   1                                  None   : 81
##  GarageFinish   GarageCars        GarageArea      GarageQual  GarageCond
##  Fin :352     Min.   :0.00000   Min.   :   0.00   Ex  :   3   Ex  :   2
##  RFn :422     1st Qu.:1.00000   1st Qu.: 334.50   Fa  :  48   Fa  :  35
##  Unf :605     Median :2.00000   Median : 480.00   Gd  :  14   Gd  :   9
##  None: 81     Mean   :1.76712   Mean   : 472.98   Po  :   3   Po  :   7
##               3rd Qu.:2.00000   3rd Qu.: 576.00   TA  :1311   TA  :1326
##               Max.   :4.00000   Max.   :1418.00   None:  81   None:  81
##
##  PavedDrive   WoodDeckSF        OpenPorchSF       EnclosedPorch
##  N:  90     Min.   :  0.0000   Min.   :  0.0000   Min.   :  0.0000
##  P:  30     1st Qu.:  0.0000   1st Qu.:  0.0000   1st Qu.:  0.0000
##  Y:1340     Median :  0.0000   Median : 25.0000   Median :  0.0000
##             Mean   : 94.2445   Mean   : 46.6603   Mean   : 21.9541
##             3rd Qu.:168.0000   3rd Qu.: 68.0000   3rd Qu.:  0.0000
##             Max.   :857.0000   Max.   :547.0000   Max.   :552.0000
##
##    X3SsnPorch         ScreenPorch         PoolArea         PoolQC
##  Min.   :  0.00000   Min.   :  0.000   Min.   :  0.0000   Ex  :   2
##  1st Qu.:  0.00000   1st Qu.:  0.000   1st Qu.:  0.0000   Fa  :   2
##  Median :  0.00000   Median :  0.000   Median :  0.0000   Gd  :   3
##  Mean   :  3.40959   Mean   : 15.061   Mean   :  2.7589   None:1453
##  3rd Qu.:  0.00000   3rd Qu.:  0.000   3rd Qu.:  0.0000
##  Max.   :508.00000   Max.   :480.000   Max.   :738.0000
##
##    Fence      MiscFeature    MiscVal              MoSold
##  GdPrv:  59   Gar2:   2   Min.   :    0.000   Min.   : 1.00000
##  GdWo :  54   Othr:   2   1st Qu.:    0.000   1st Qu.: 5.00000
##  MnPrv: 157   Shed:  49   Median :    0.000   Median : 6.00000
##  MnWw :  11   TenC:   1   Mean   :   43.489   Mean   : 6.32192
##  None :1179   None:1406   3rd Qu.:    0.000   3rd Qu.: 8.00000
##                           Max.   :15500.000   Max.   :12.00000
##
##      YrSold           SaleType    SaleCondition    SalePrice
##  Min.   :2006.00   WD     :1267   Abnorml: 101   Min.   : 34900
##  1st Qu.:2007.00   New    : 122   AdjLand:   4   1st Qu.:129975
##  Median :2008.00   COD    :  43   Alloca :  12   Median :163000
##  Mean   :2007.82   ConLD  :   9   Family :  20   Mean   :180921
##  3rd Qu.:2009.00   ConLI  :   5   Normal :1198   3rd Qu.:214000
##  Max.   :2010.00   ConLw  :   5   Partial: 125   Max.   :755000
##                    (Other):   9
```

Plot each variable, along with its summary data

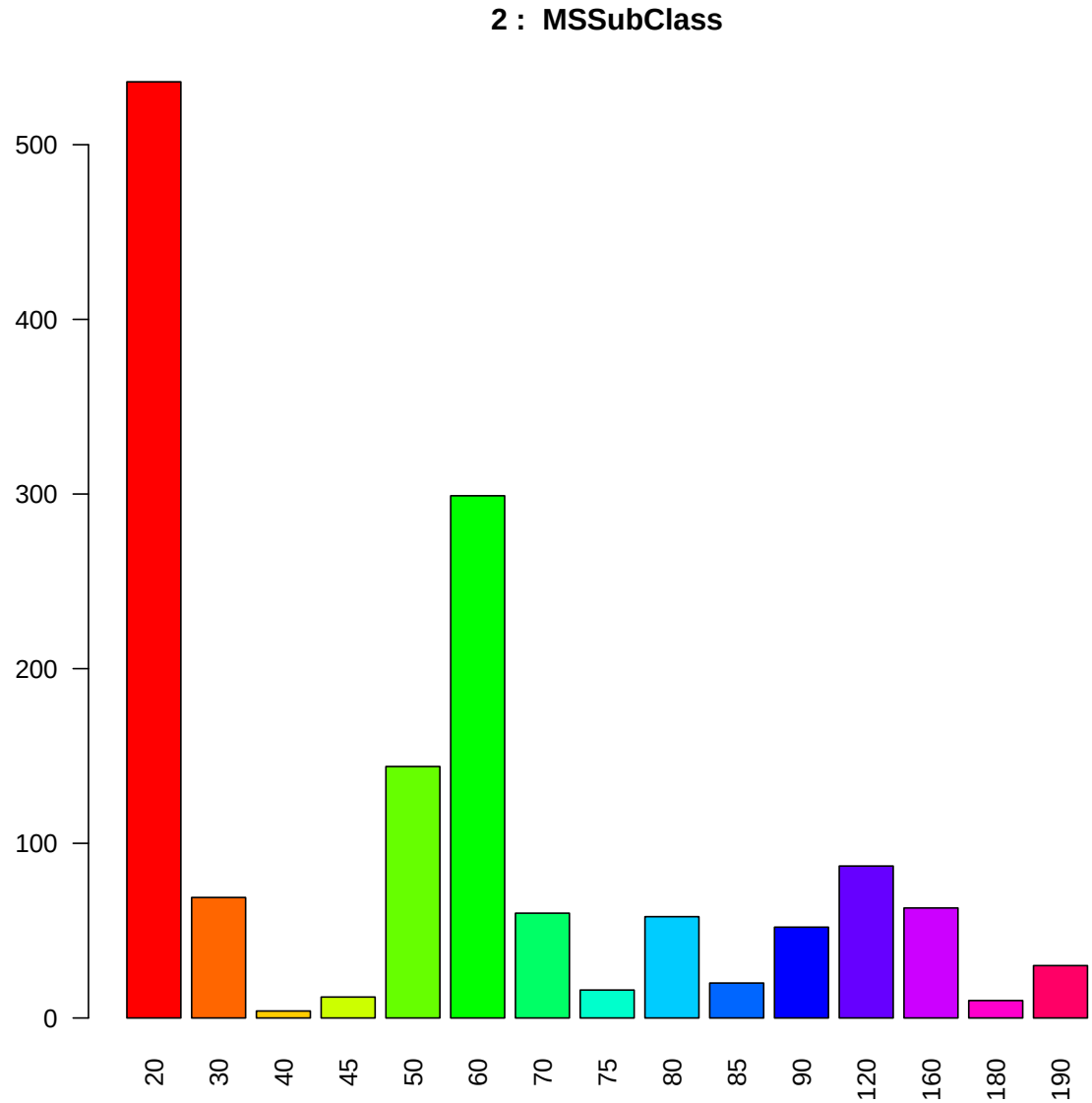
For numeric variables, we will add vertical lines for Mean (black), Median (green) and +/- 1 standard deviation (red-dashed) above and below the mean

Also we will attempt to overlay Normal (blue) and Exponential (purple, dashed) densities on the histogram.

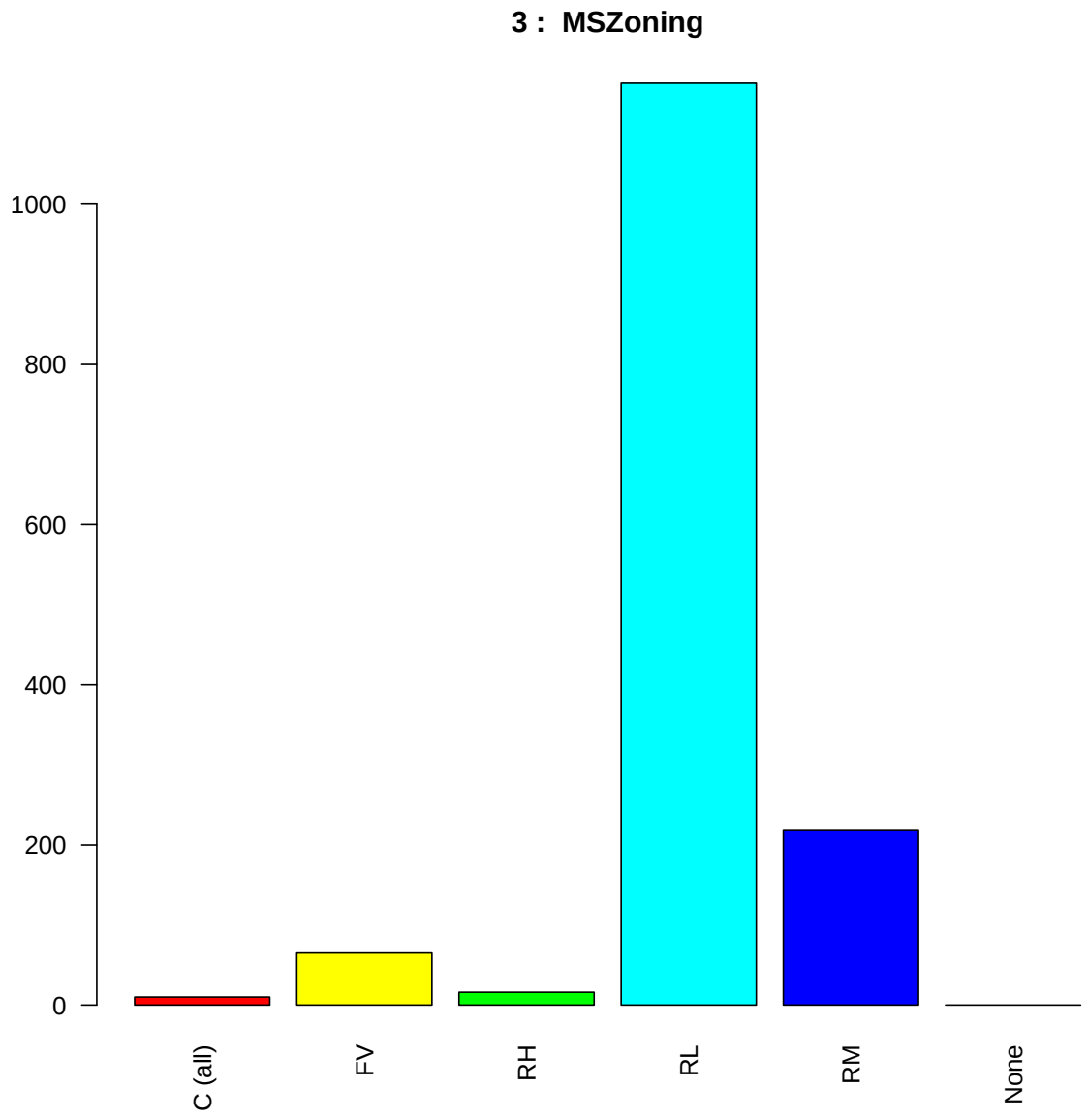
(Depending on the nature of the data, such densities may or may not make sense...)

```
### Don't bother plotting the first variable ("ID") because it's simply an index
for (item in 2:length(train.df) ){ #length(train.df)) {
  thisname = attr(train.df[item],"names")
  # create a title which incorporates the sequence number
  # of the variable along with the name, for guidance
  mainheader = paste(item, ": ", thisname)
  #print(thisname)
  rawitem  = train.df[item]
  thisitem = train.df[[item]]
  thisclass = class(thisitem)
  ### display the factor items
  if(thisclass == "factor"){
    factorresult=table(thisitem,useNA = "ifany",dnn = thisname)
    par(mfrow = c(1, 1))
    (barplot(factorresult,
             col=rainbow(length(factorresult)),
             main=mainheader,
             las=2))
    print(factorresult)

  }
  else if (thisclass=="integer"||thisclass=="numeric") {
    ## Compute the summary statistics, plus the standard deviation
    numresult=c(summary(thisitem),
                STDEV=sd(thisitem,na.rm=T))
    par(mfrow = c(2, 1))
    ## plot the histogram
    hist(thisitem,breaks=30,main = mainheader,xlab=thisname,col="lightblue",probability =  T)

    ## add a normal curve fit
    curve(dnorm(x, mean = mean(thisitem), sd = sd(thisitem)), col="blue", lwd=3 , add=TRUE)

    ## add an exponential curve fit
    curve(dexp(x, rate=1/mean(thisitem)), col="purple", lwd=3, lty="dashed", add=TRUE)

    ## add a vertical line for median
    abline(v=numresult["Median"],col="green", lwd=2)

    ## add a vertical line for mean
    abline(v=numresult["Mean"],col="black",lwd=2)

    ## add vertical lines for down and up one standard deviation
```

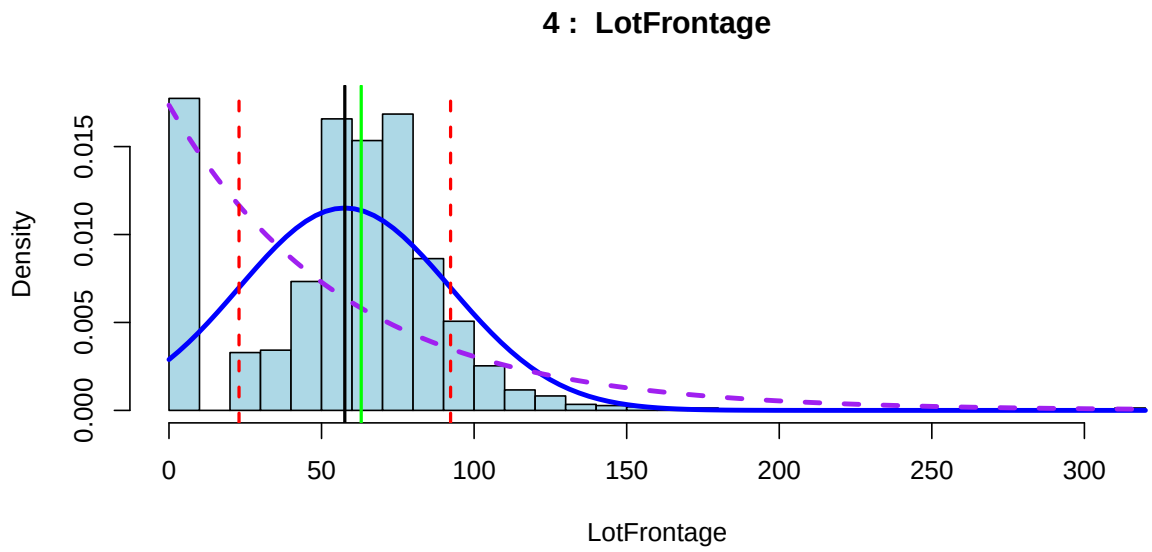
```r
    abline(v=numresult["Mean"]-numresult["STDEV"],col="red",lty="dashed", lwd=2)
    abline(v=numresult["Mean"]+numresult["STDEV"],col="red",lty="dashed", lwd=2)

    ## add a boxplot
    boxplot(thisitem,horizontal = T, col="lightblue", main=mainheader)
    print(numresult)
  }
}
```

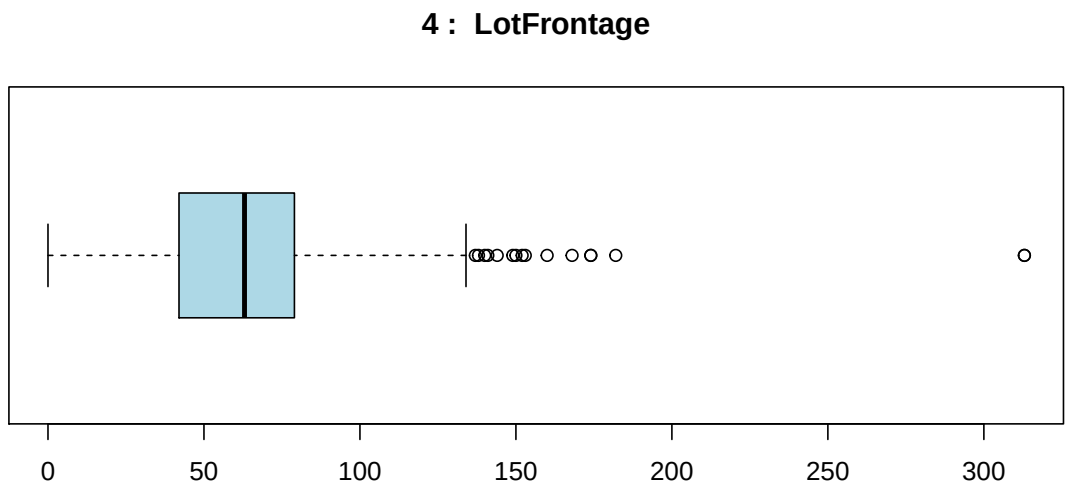



```
## MSSubClass
##  20  30  40  45  50  60  70  75  80  85  90 120 160 180 190
## 536  69   4  12 144 299  60  16  58  20  52  87  63  10  30
```

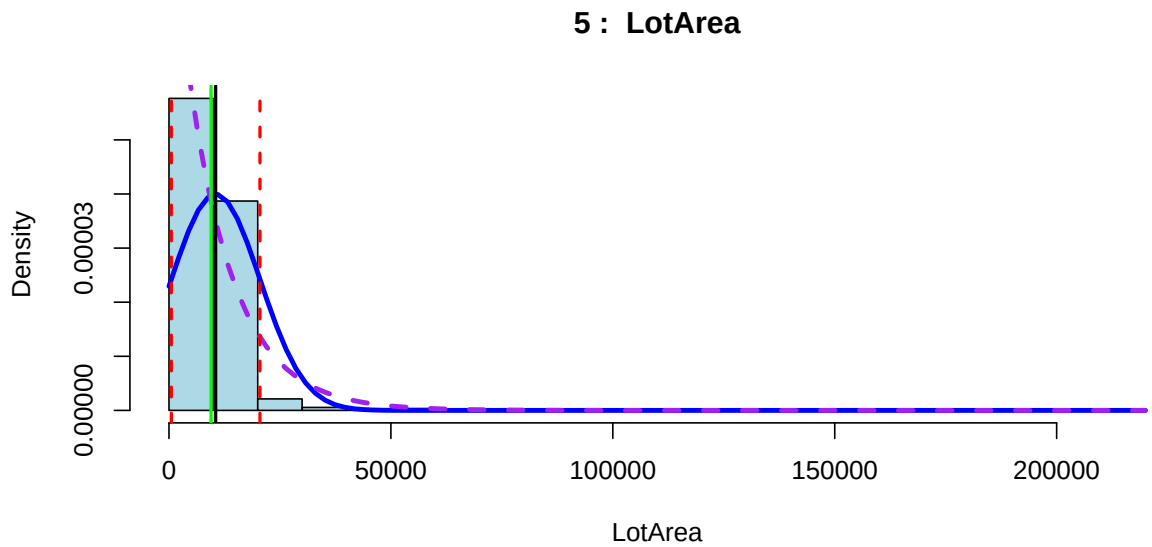


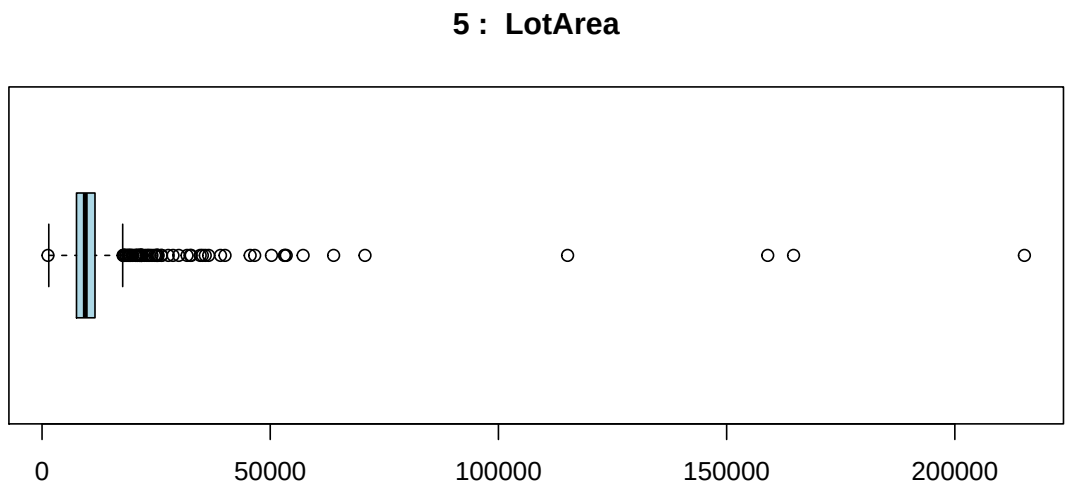

```
## MSZoning
## C (all)      FV      RH      RL      RM    None
##      10      65      16    1151     218       0
```



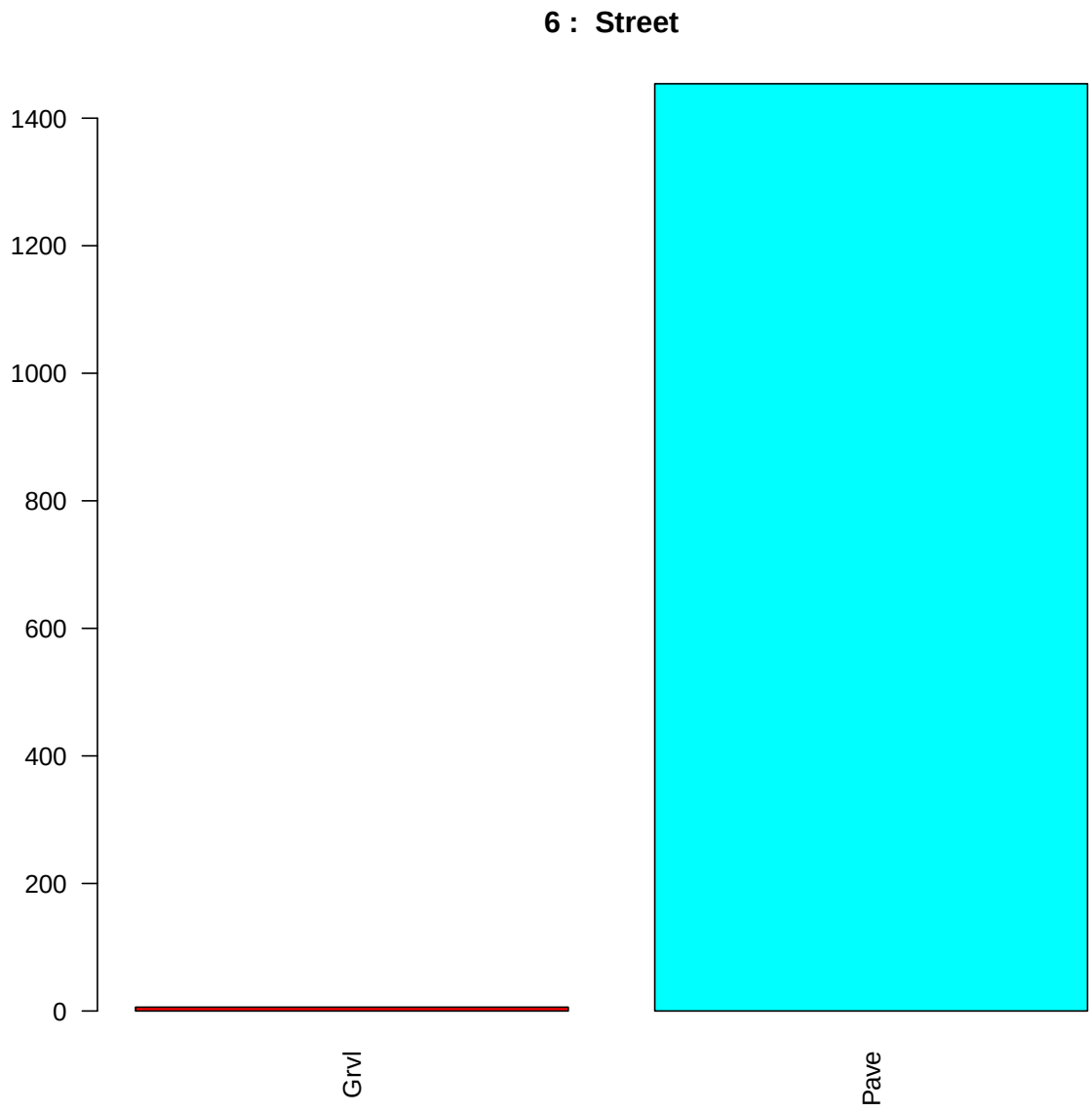



```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##   0.0000000  42.0000000  63.0000000  57.6232877  79.0000000 313.0000000
##       STDEV
##  34.6643042
```

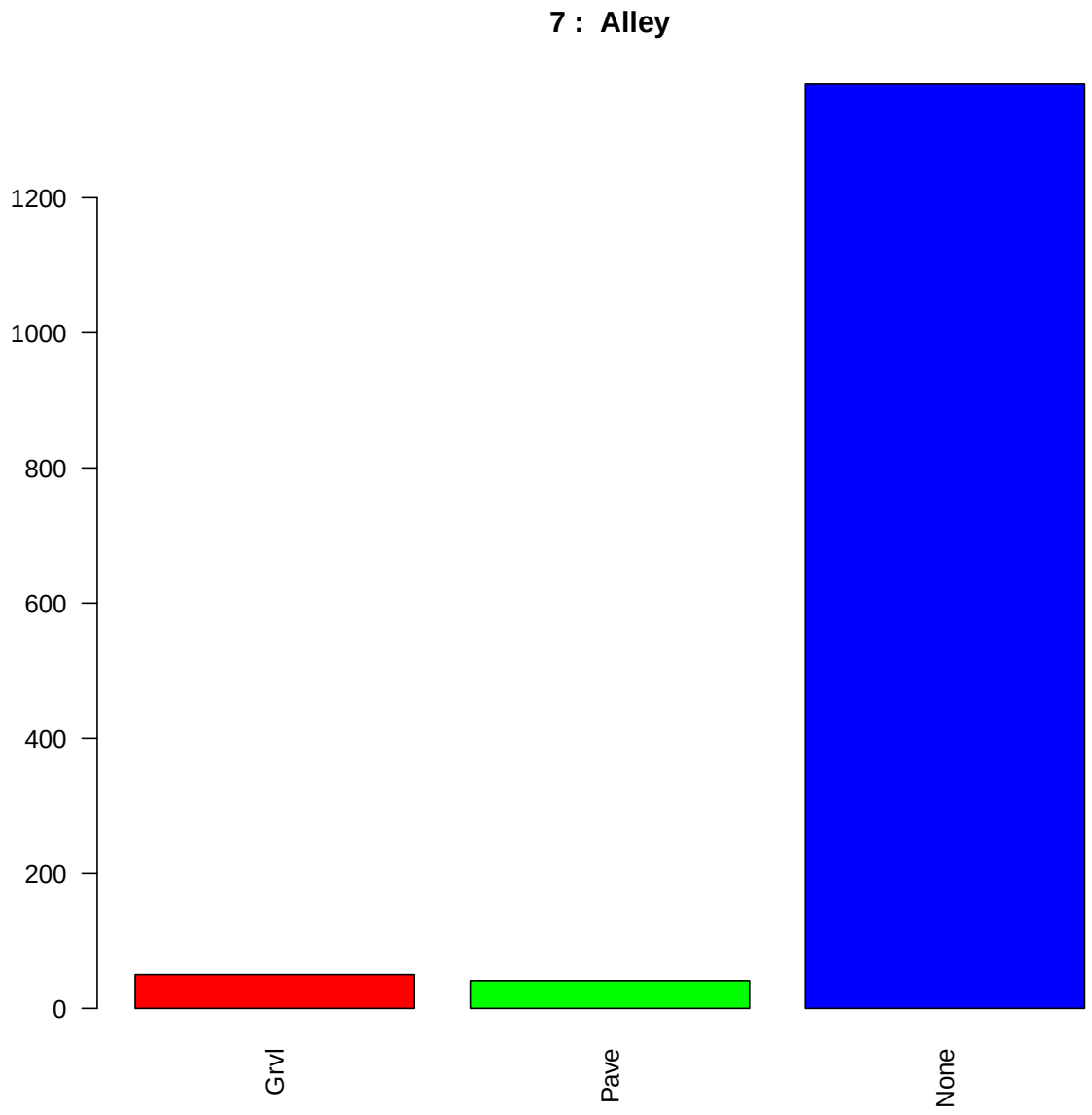




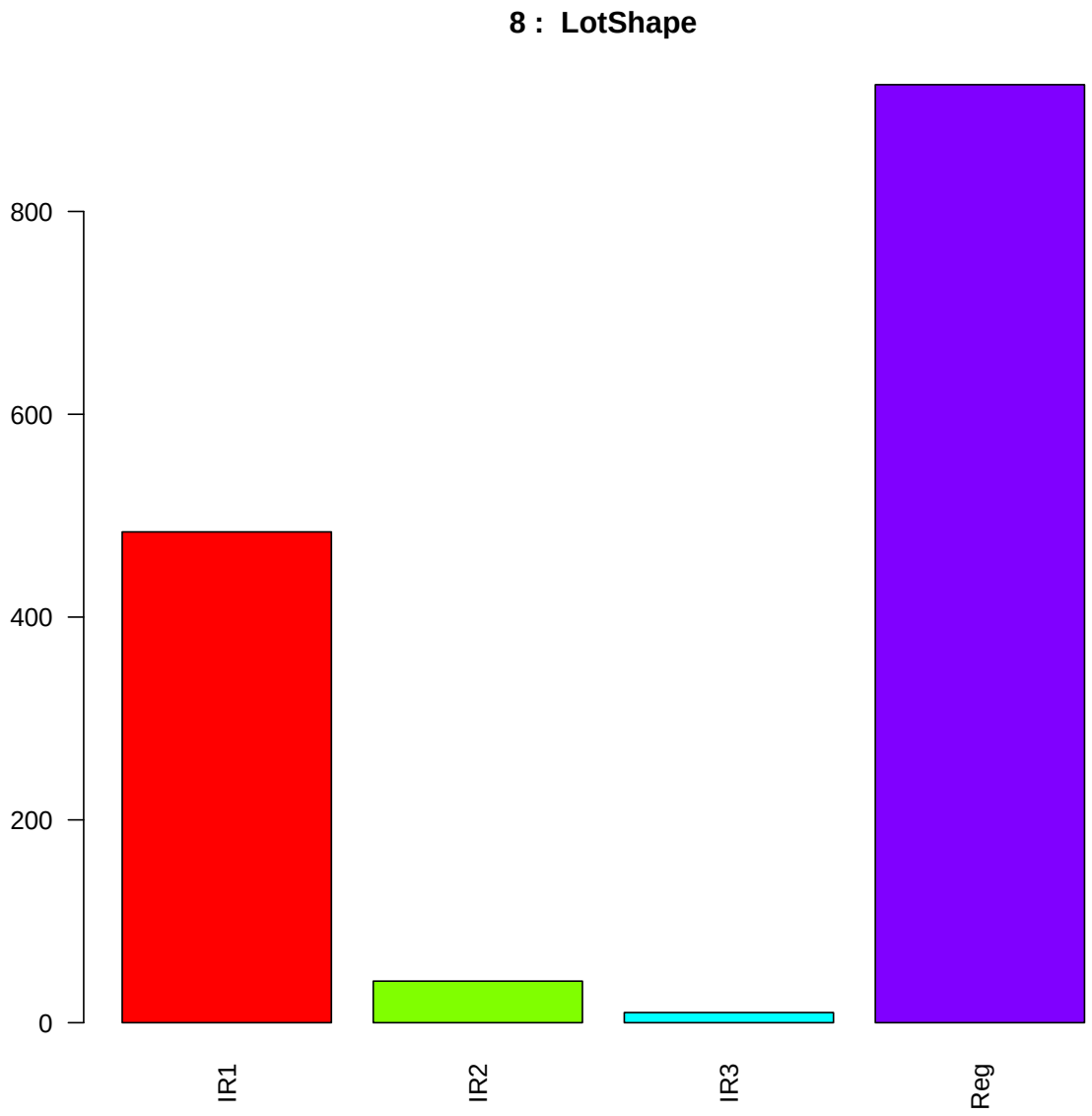

```
##          Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##    1300.00000   7553.50000   9478.50000  10516.82808  11601.50000 215245.00000
##         STDEV
##    9981.26493
```



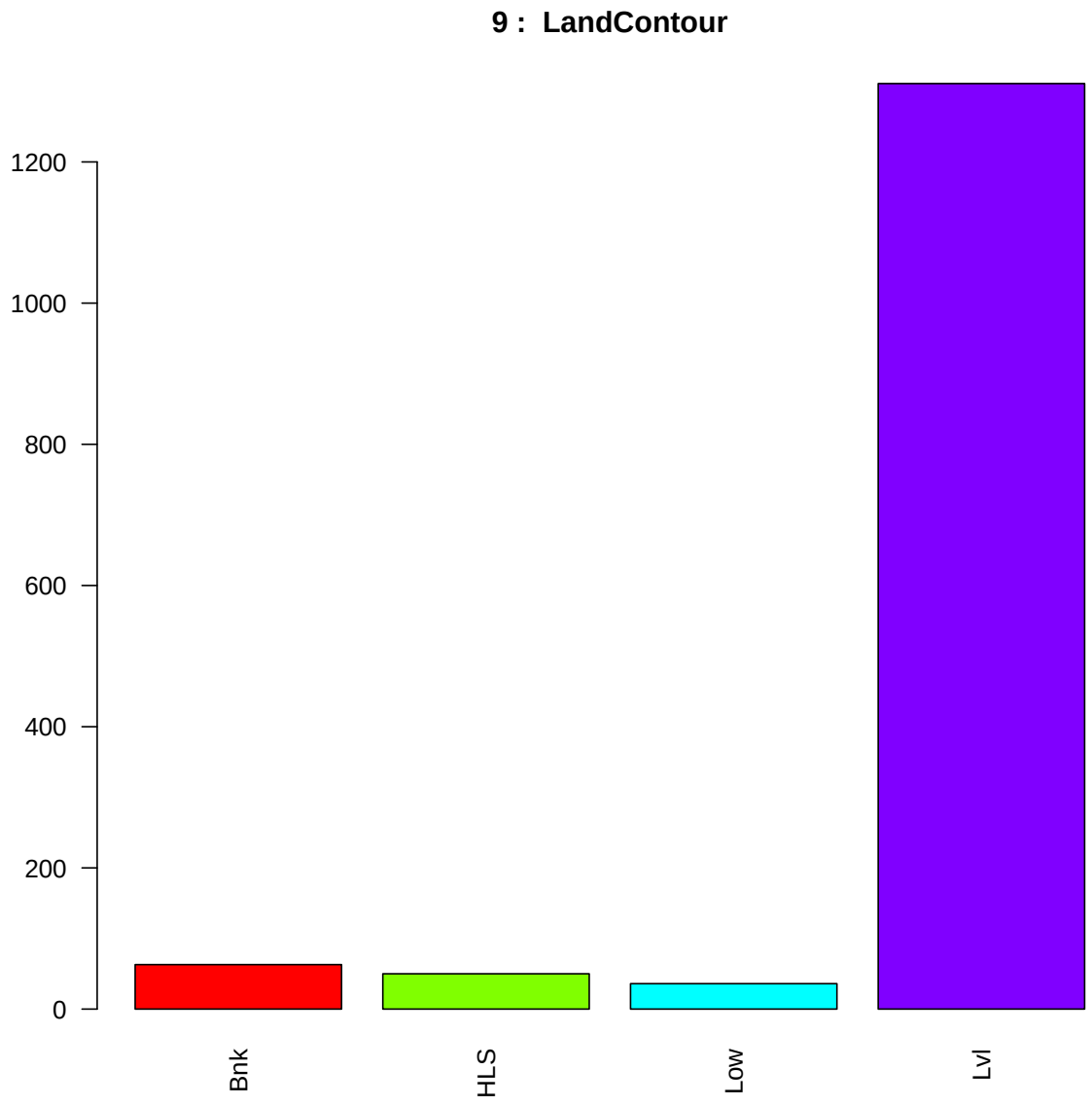

```
## Street
## Grvl Pave
##    6 1454
```



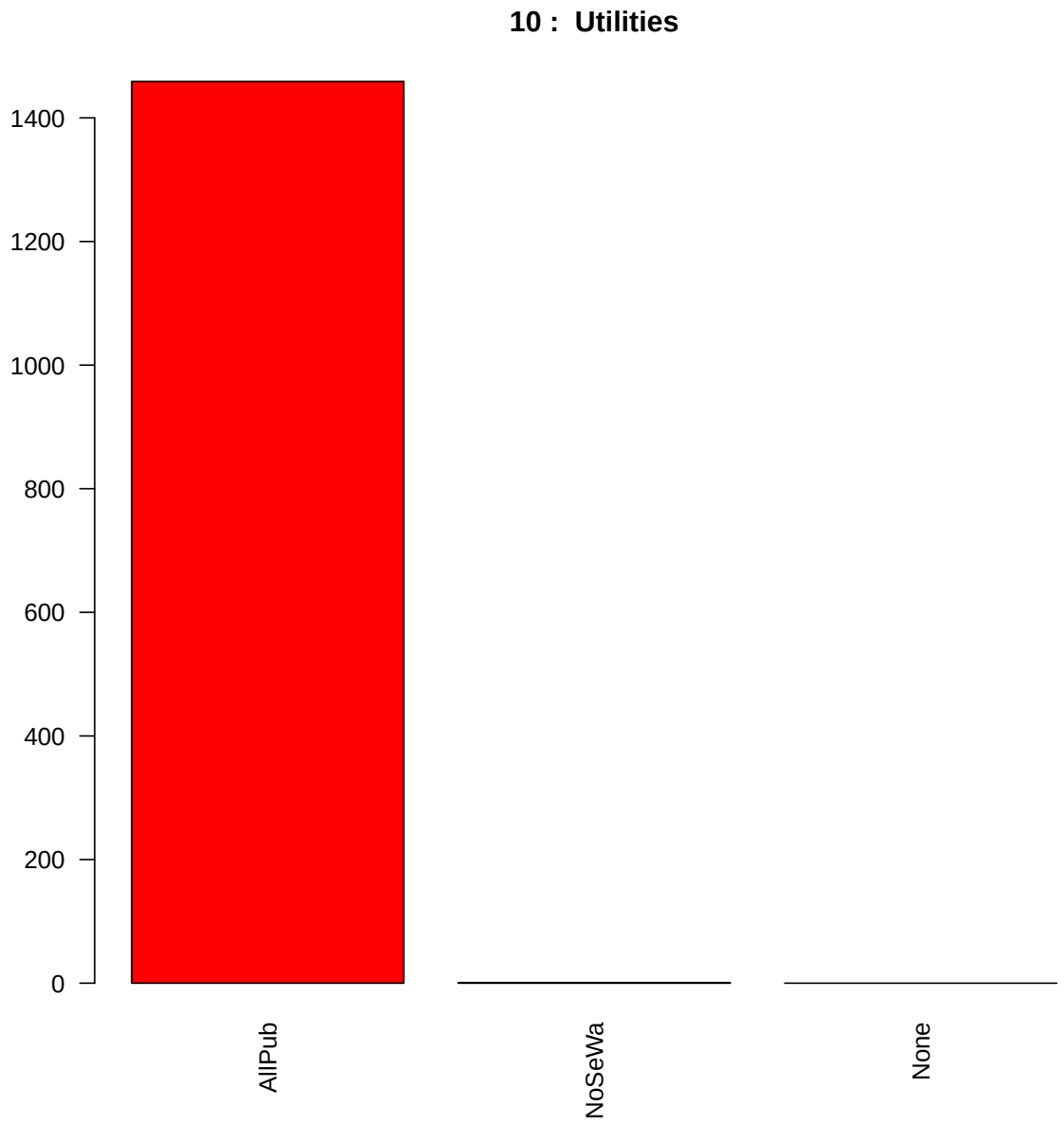

```
## Alley
## Grvl Pave None
##   50   41 1369
```



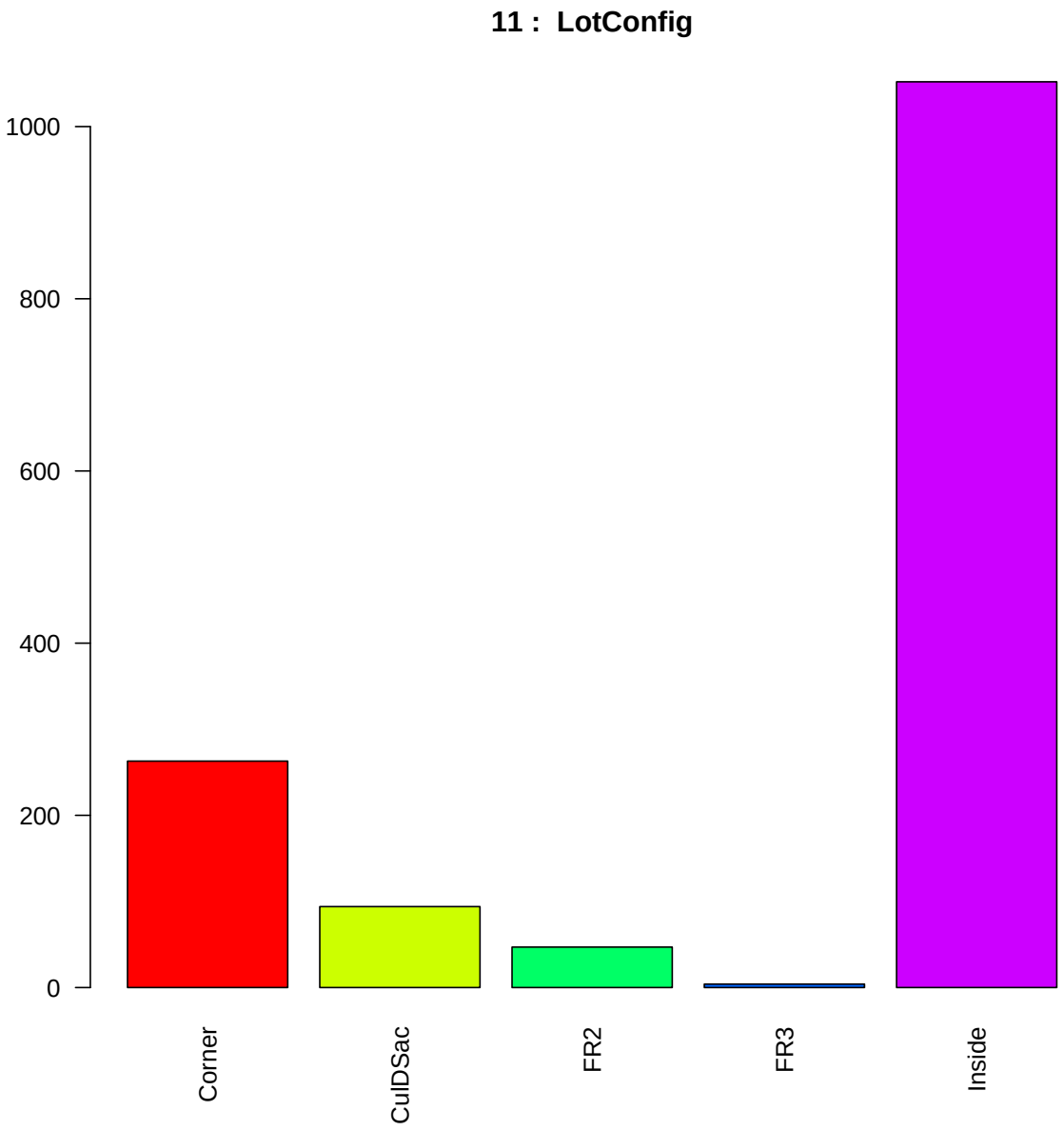

```
## LotShape
## IR1 IR2 IR3 Reg
## 484  41  10 925
```



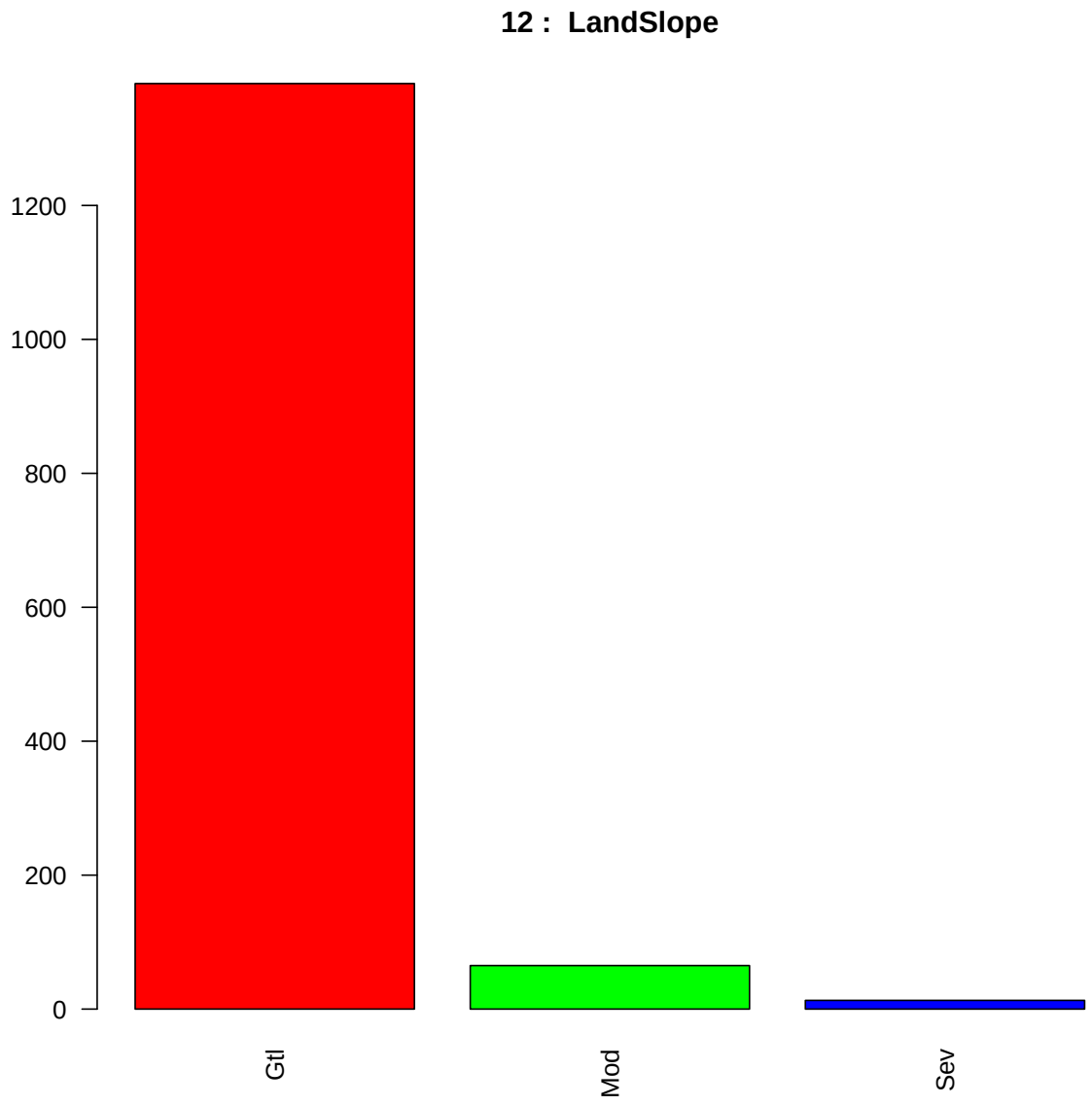

```
## LandContour
##  Bnk  HLS  Low  Lvl
##   63   50   36 1311
```



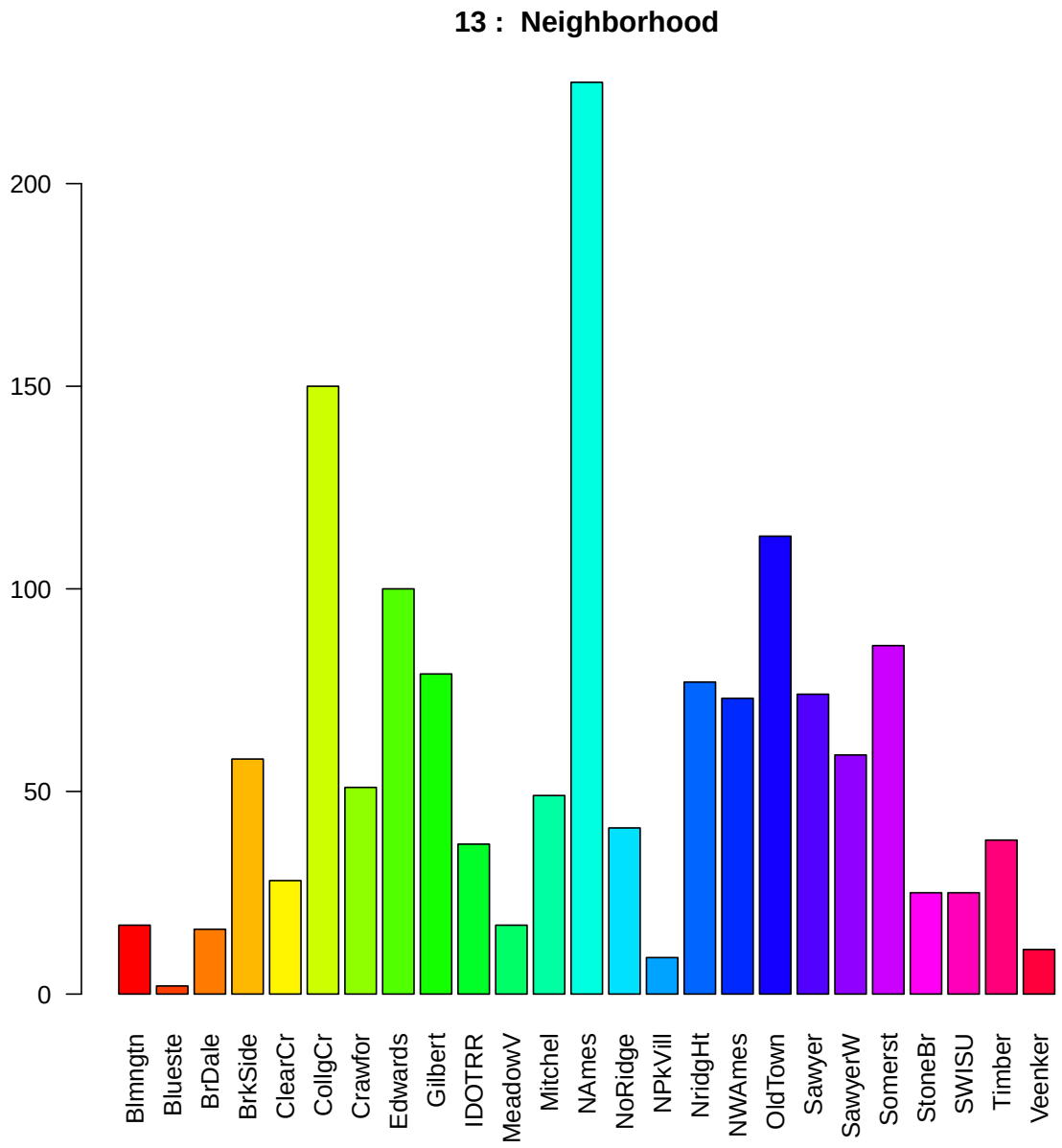

```
## Utilities
## AllPub NoSeWa   None
##   1459      1      0
```



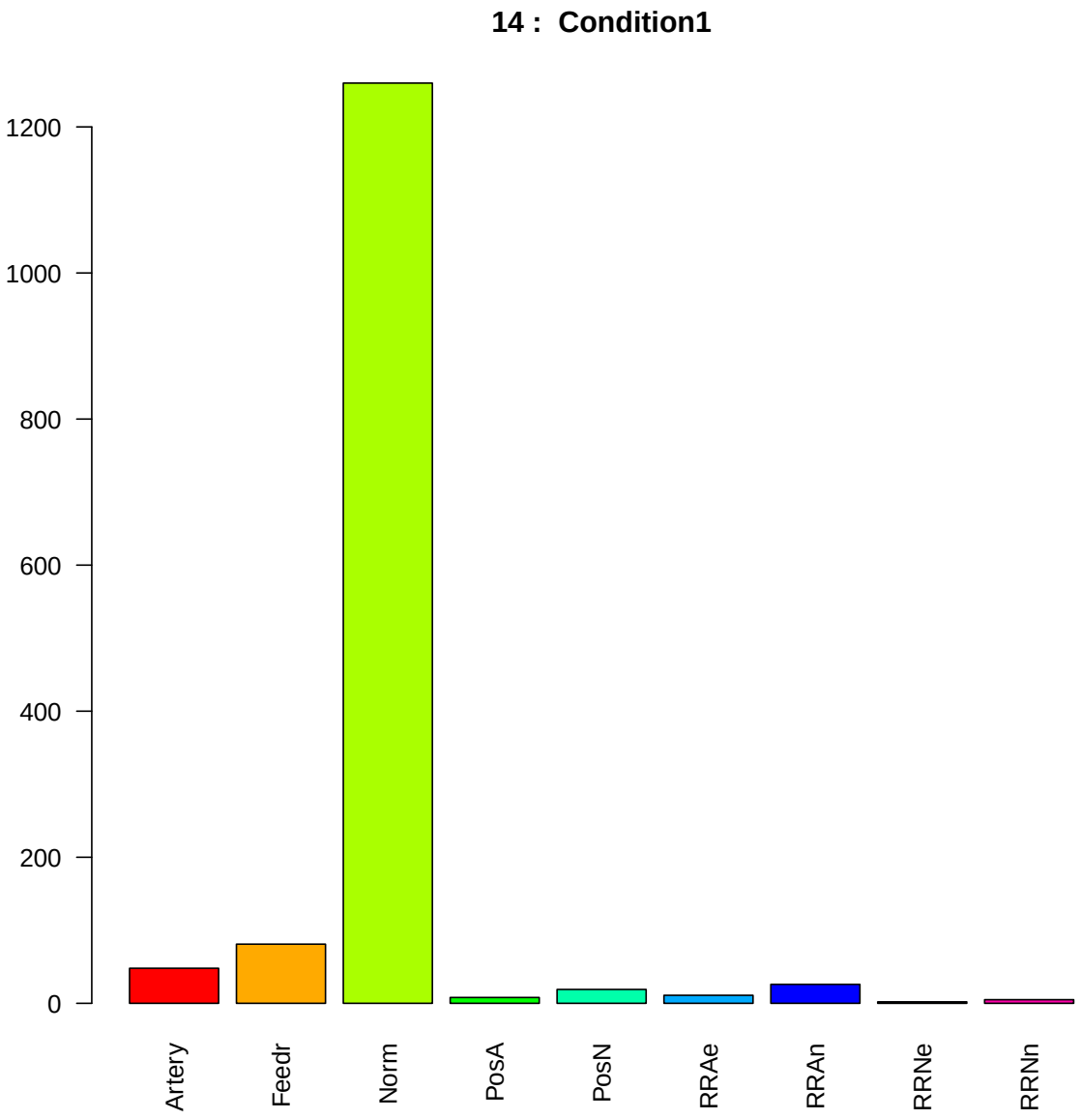

```
## LotConfig
##  Corner CulDSac     FR2     FR3  Inside
##     263      94      47       4    1052
```



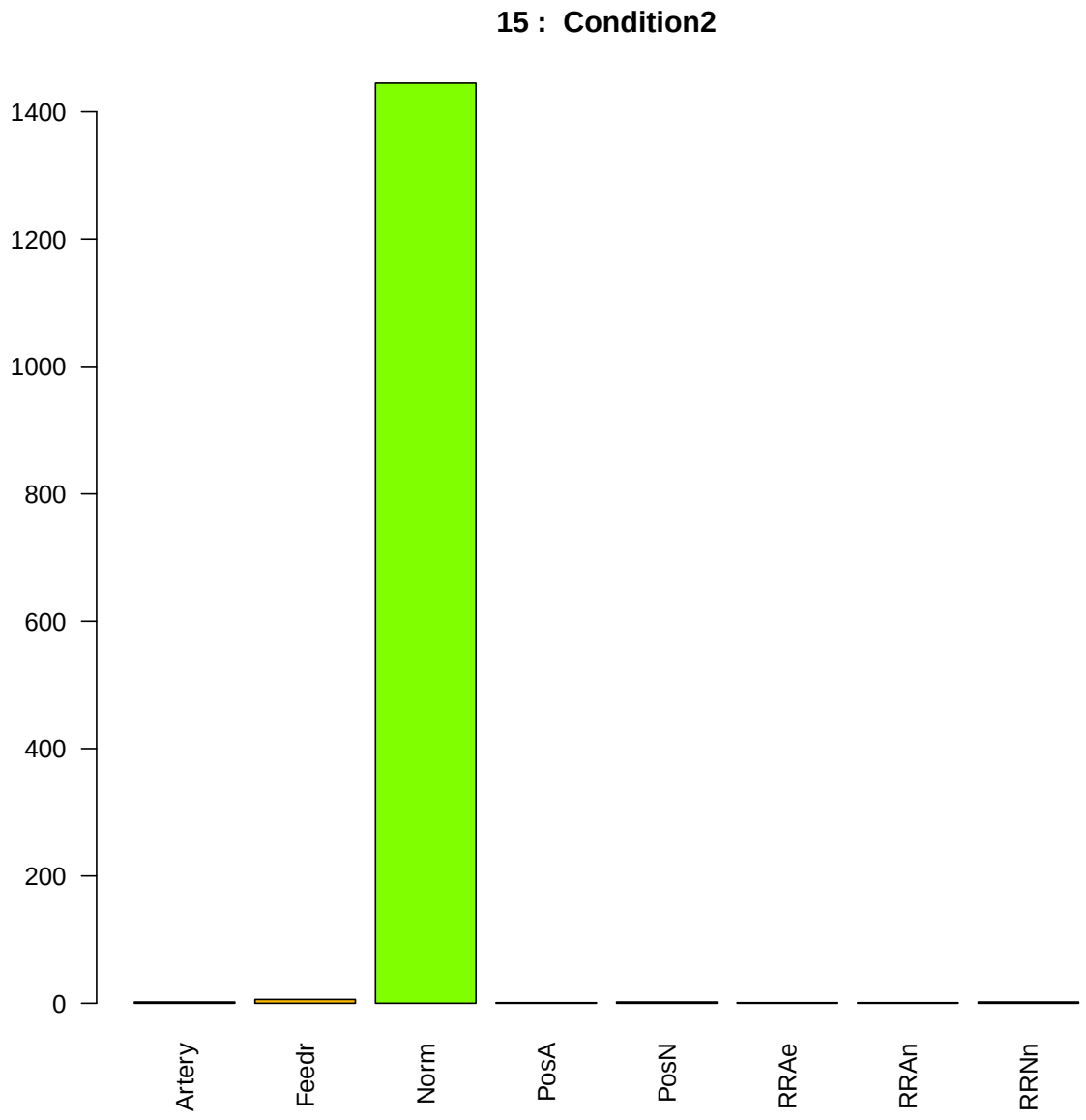

```
## LandSlope
##  Gtl  Mod  Sev
## 1382   65   13
```



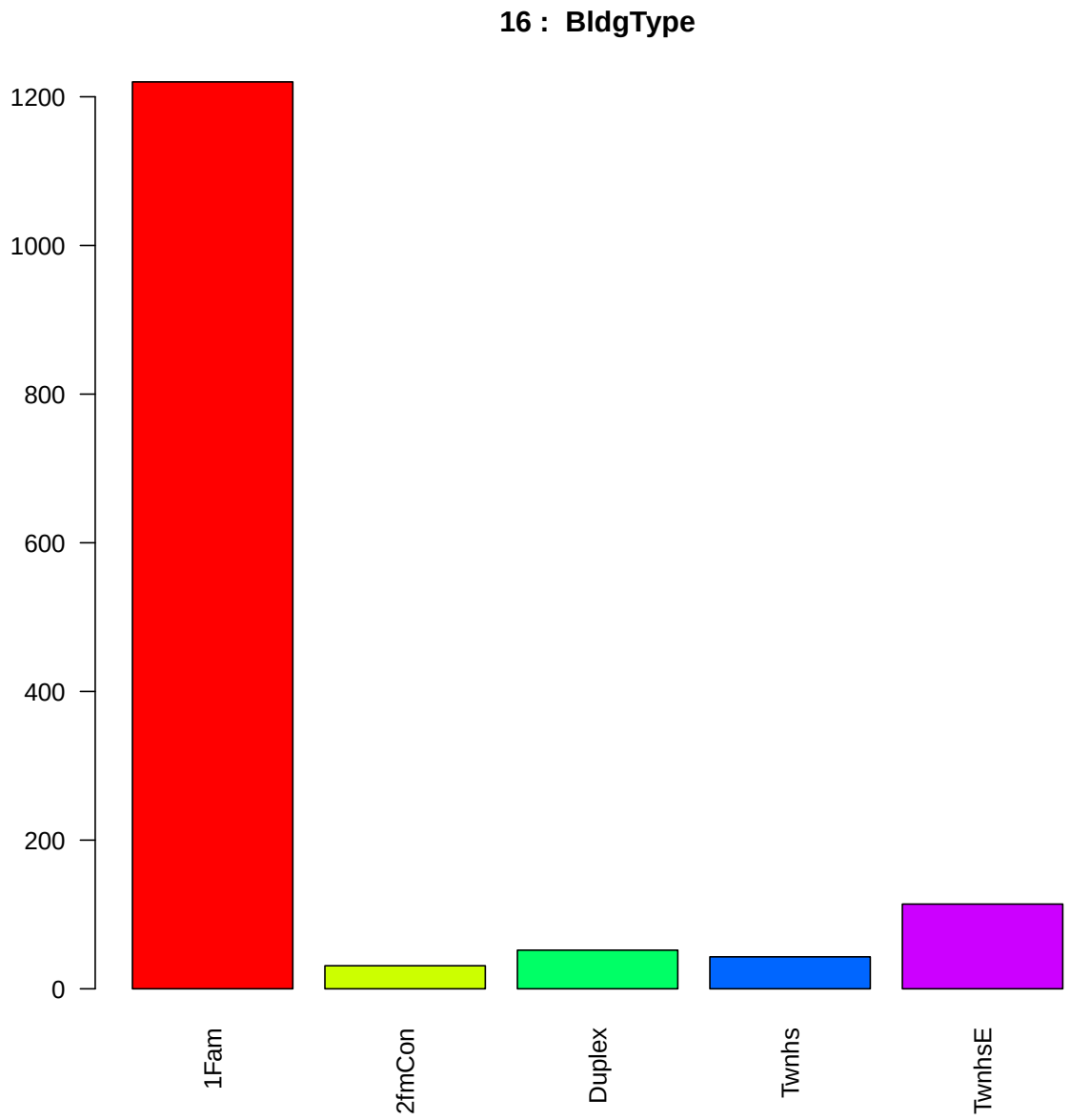

```
## Neighborhood
## Blmngtn Blueste  BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert  IDOTRR
##      17       2      16      58      28     150      51     100      79      37
## MeadowV Mitchel   NAmes NoRidge NPkVill NridgHt  NWAmes OldTown  Sawyer SawyerW
##      17      49     225     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



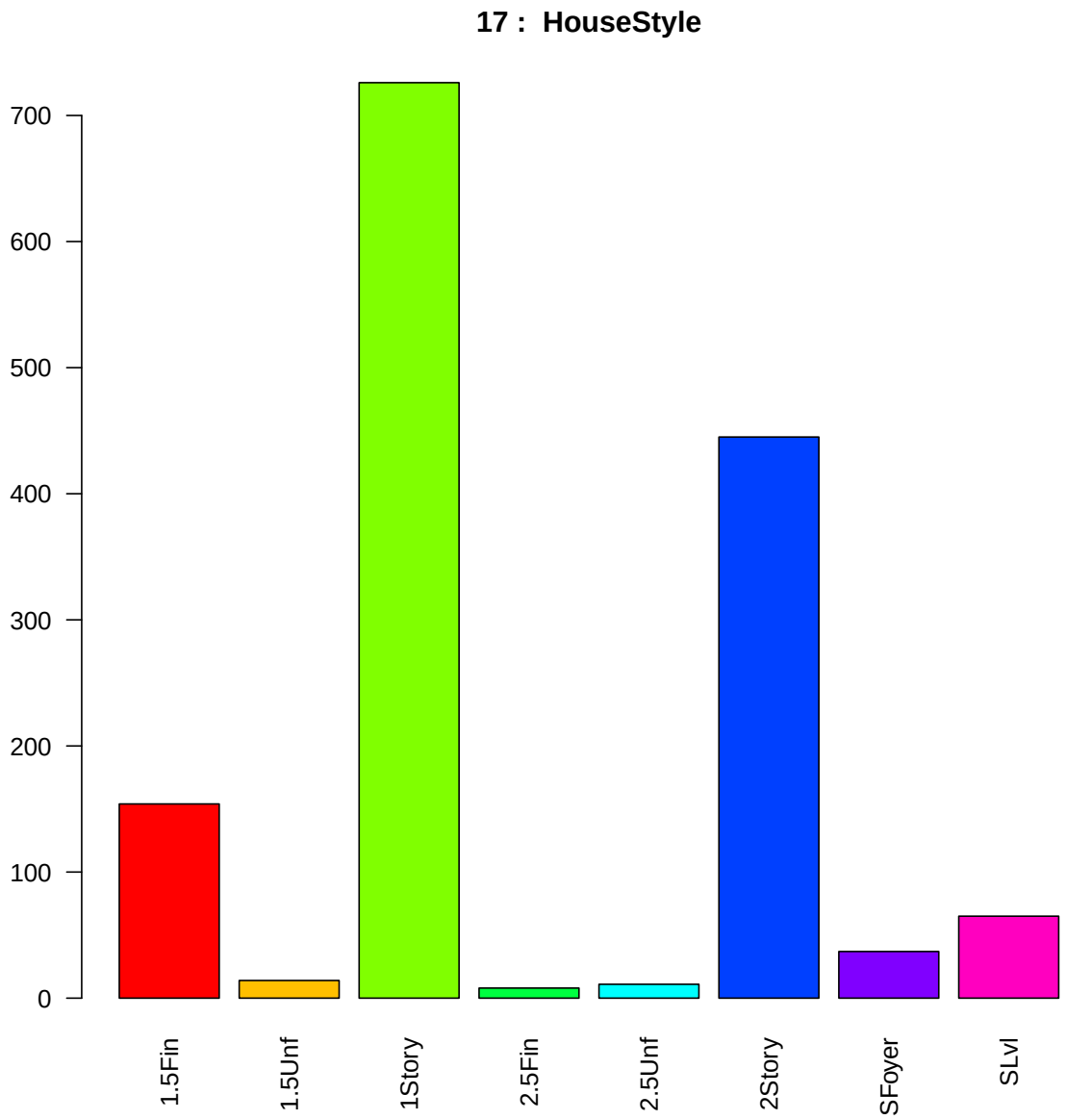

```
## Condition1
## Artery  Feedr   Norm   PosA   PosN   RRAe   RRAn   RRNe   RRNn
##     48     81   1260      8     19     11     26      2      5
```



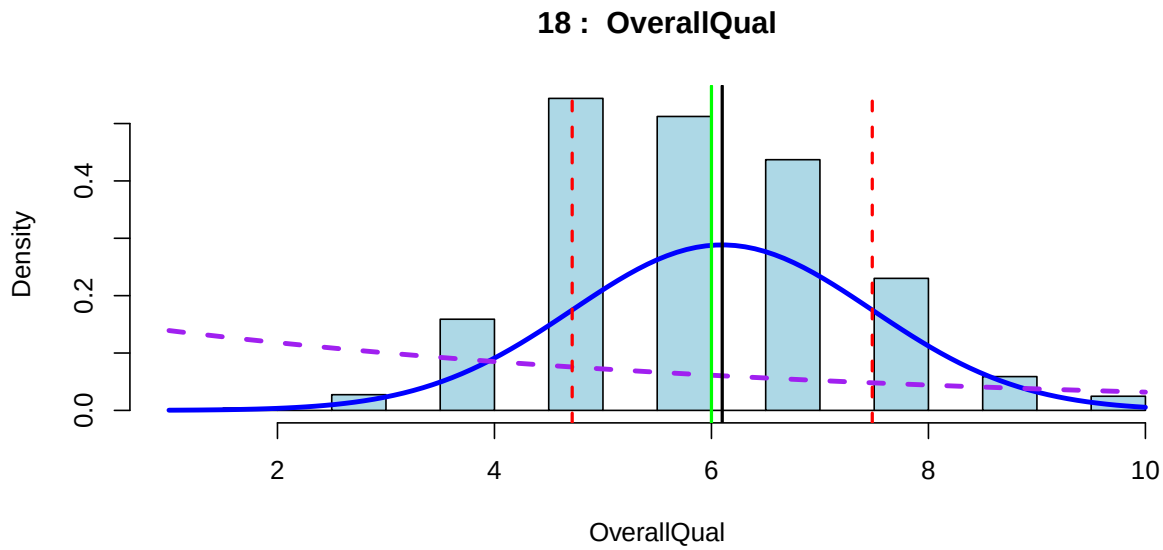

```
## Condition2
## Artery  Feedr   Norm   PosA   PosN   RRAe   RRAn   RRNn
##      2      6   1445      1      2      1      1      2
```

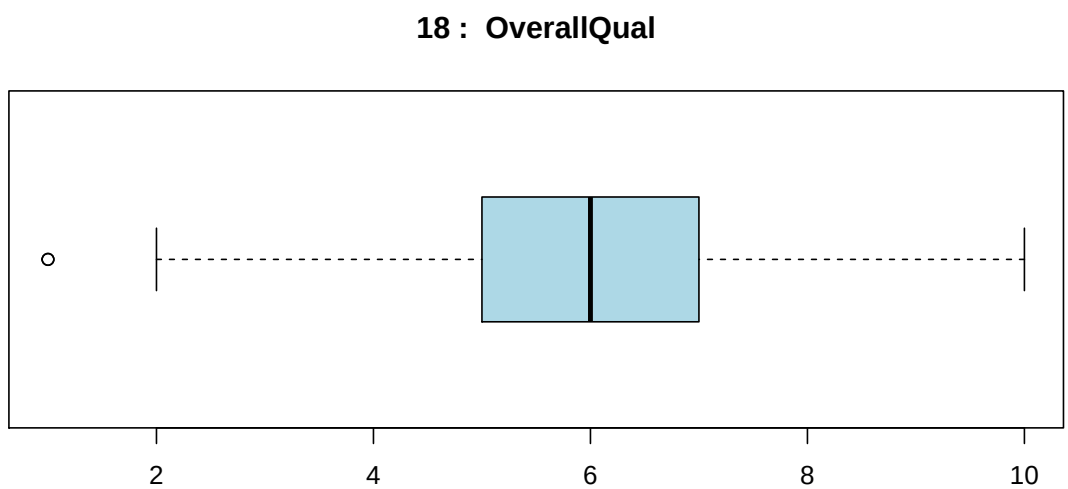



```
## BldgType
##   1Fam 2fmCon Duplex  Twnhs TwnhsE
##   1220     31     52     43    114
```

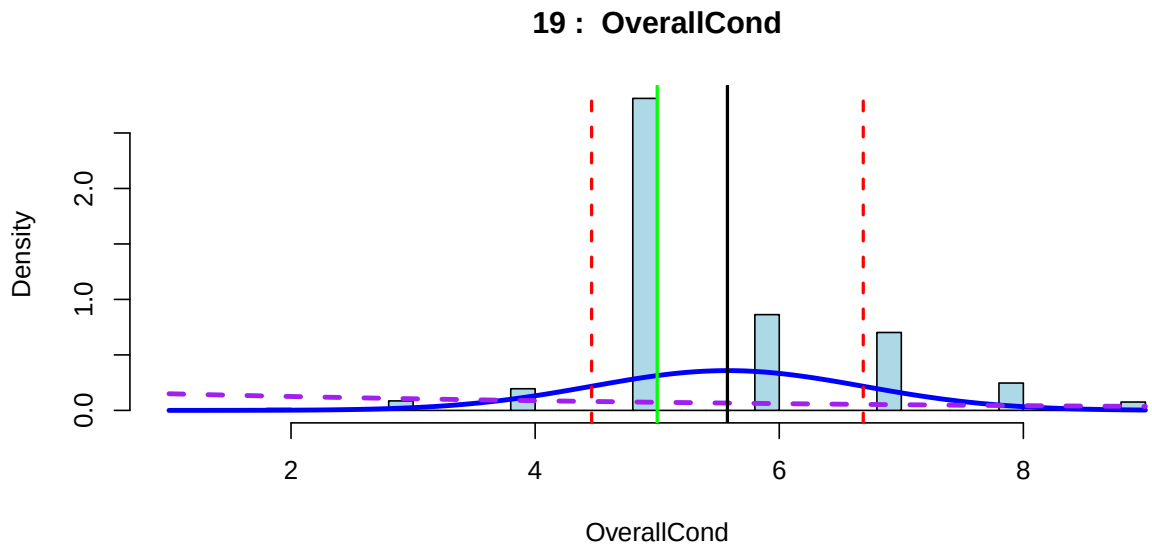


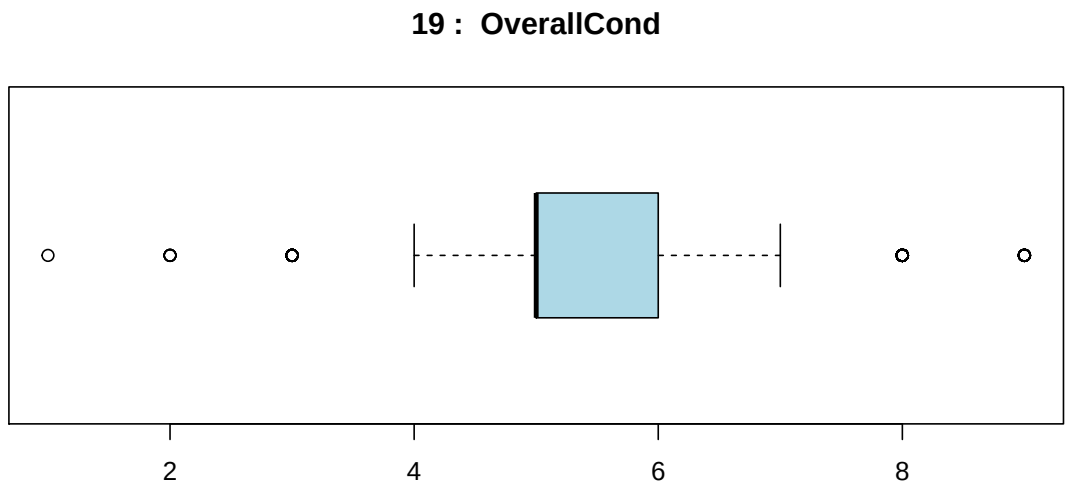

```
## HouseStyle
## 1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SFoyer   SLvl
##    154     14    726      8     11    445     37     65
```



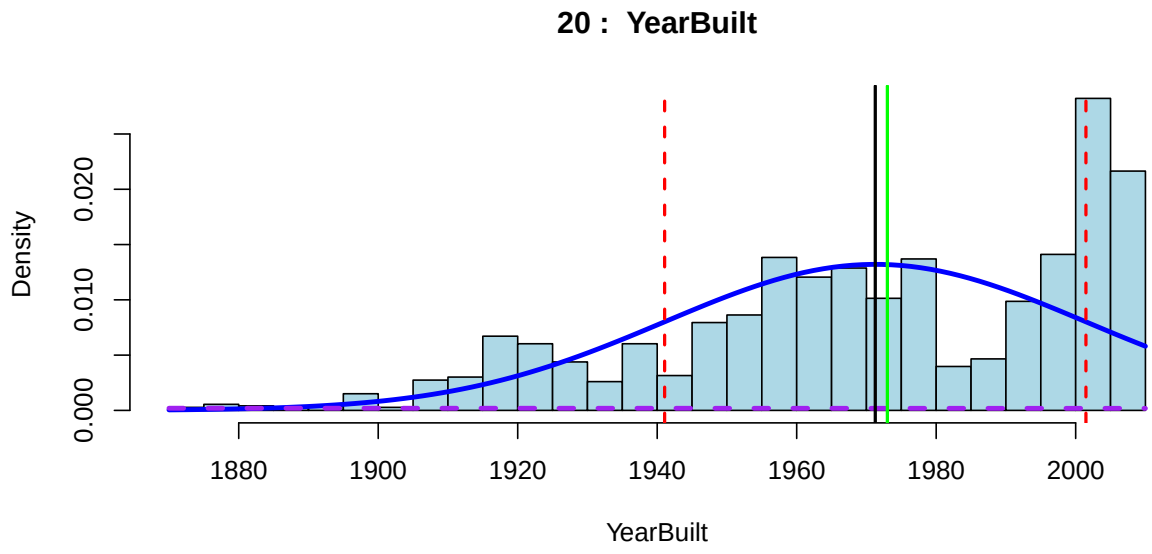


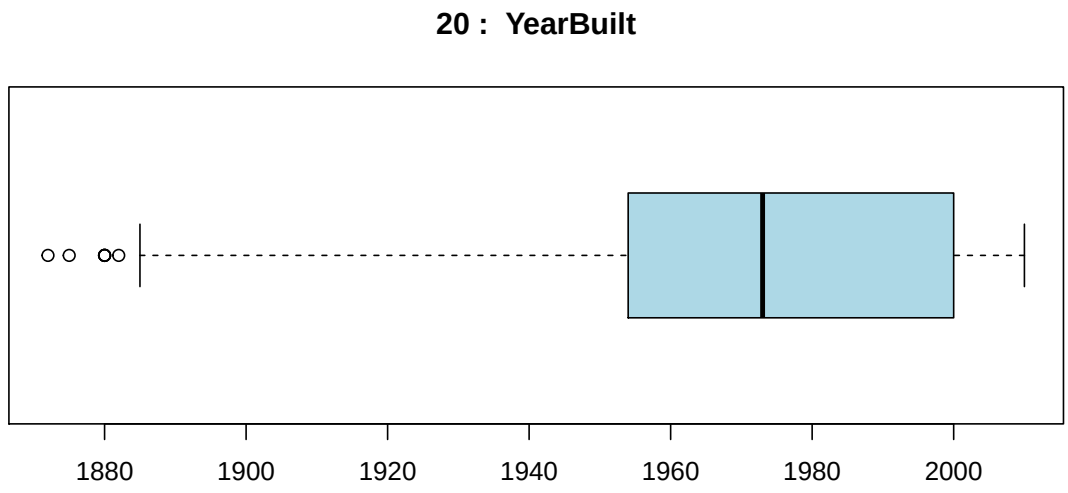

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##  1.00000000  5.00000000  6.00000000  6.09931507  7.00000000 10.00000000
##       STDEV
##  1.38299655
```



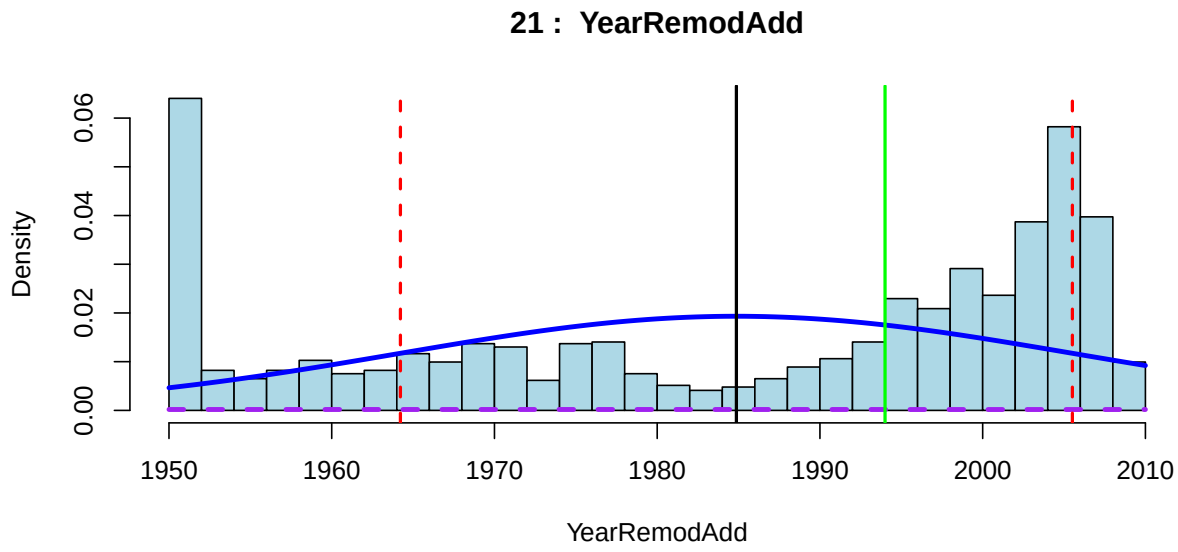


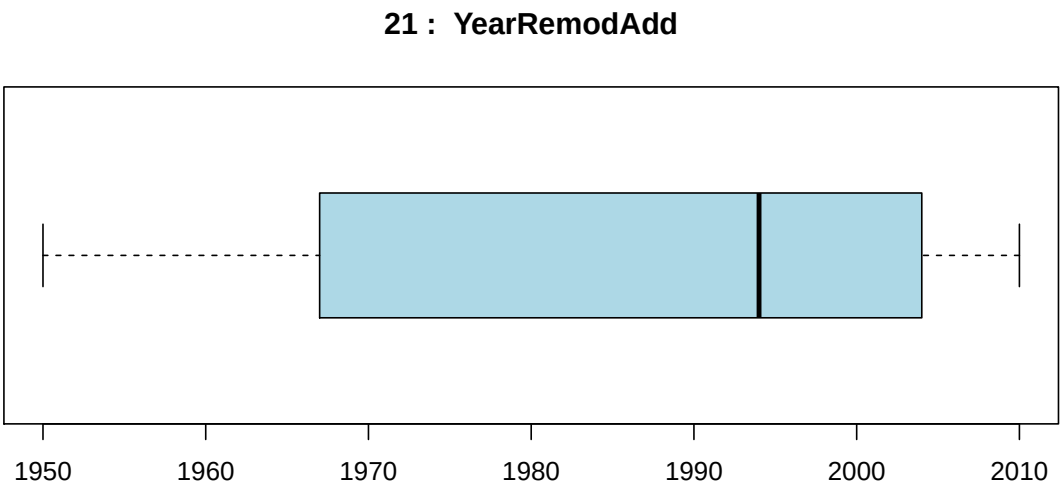

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.      STDEV
## 1.00000000 5.00000000 5.00000000 5.57534247 6.00000000 9.00000000 1.11279934
```



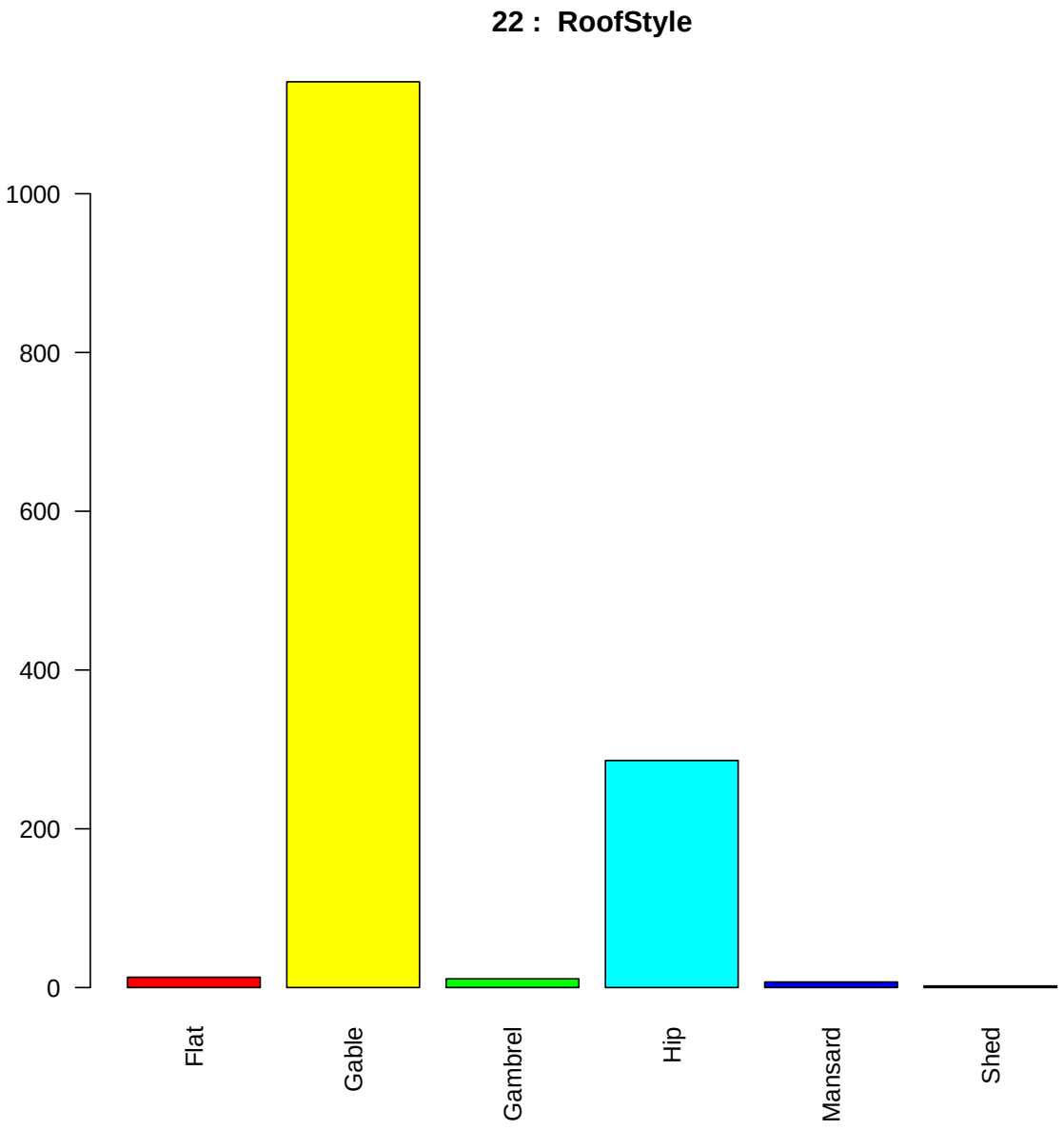



```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## 1872.000000 1954.000000 1973.000000 1971.267808 2000.000000 2010.000000
##       STDEV
##   30.202904
```

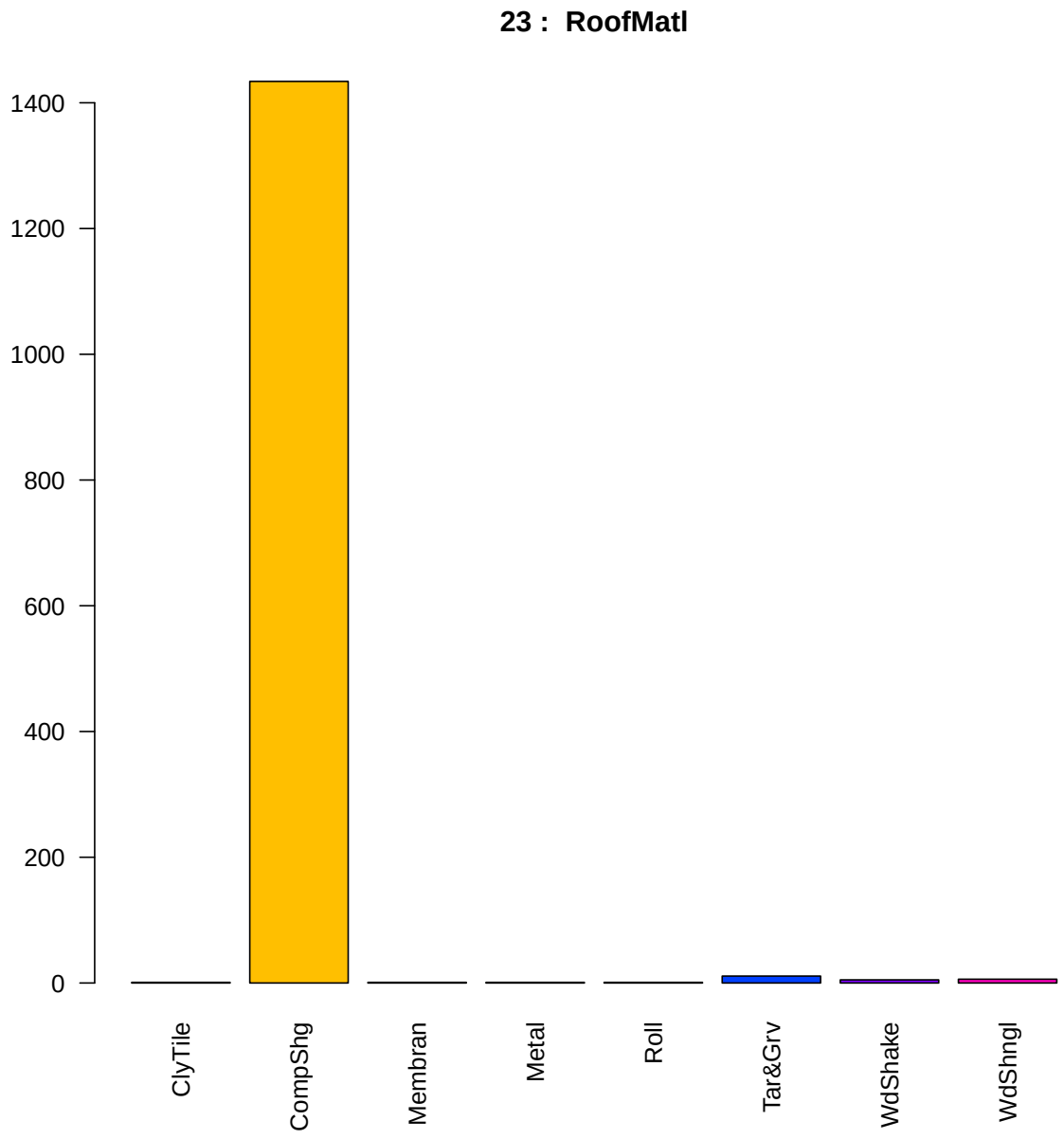




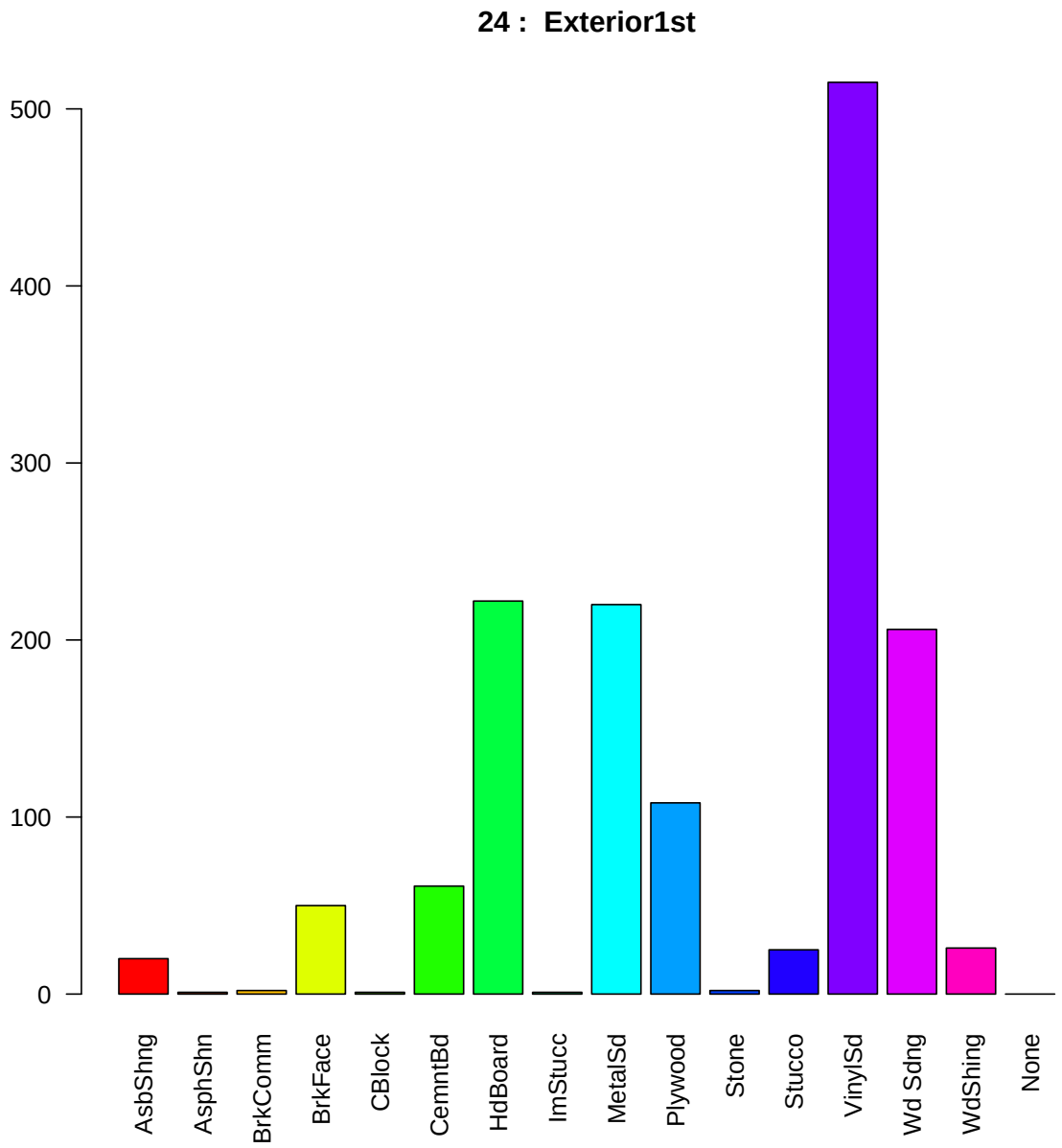

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 1950.0000000 1967.0000000 1994.0000000 1984.8657534 2004.0000000 2010.0000000
##        STDEV
##   20.6454068
```



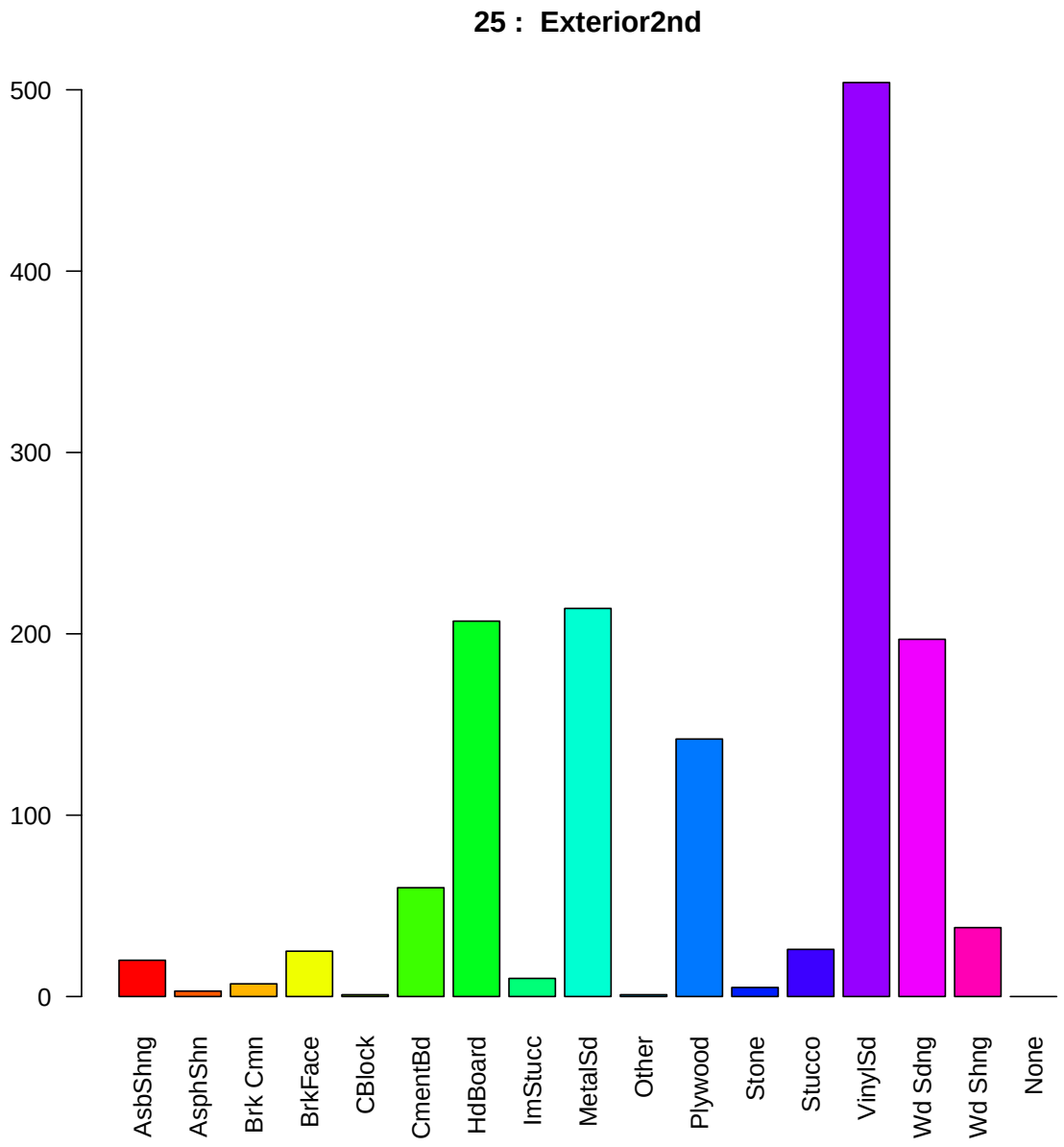

```
## RoofStyle
##    Flat   Gable Gambrel     Hip Mansard    Shed
##      13    1141      11     286       7       2
```



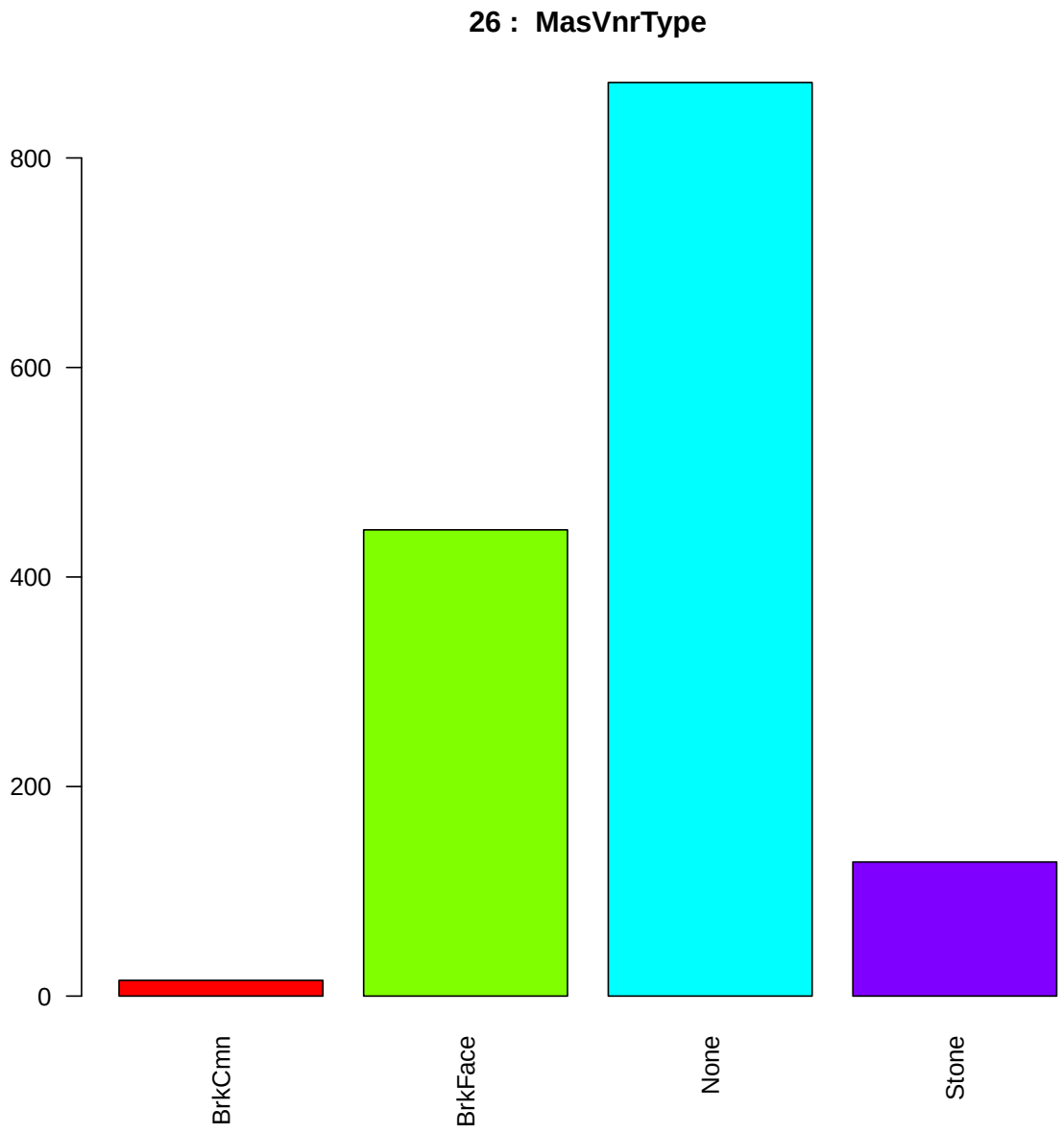

```
## RoofMatl
## ClyTile CompShg Membran   Metal    Roll Tar&Grv WdShake WdShngl
##       1    1434       1       1       1      11       5       6
```



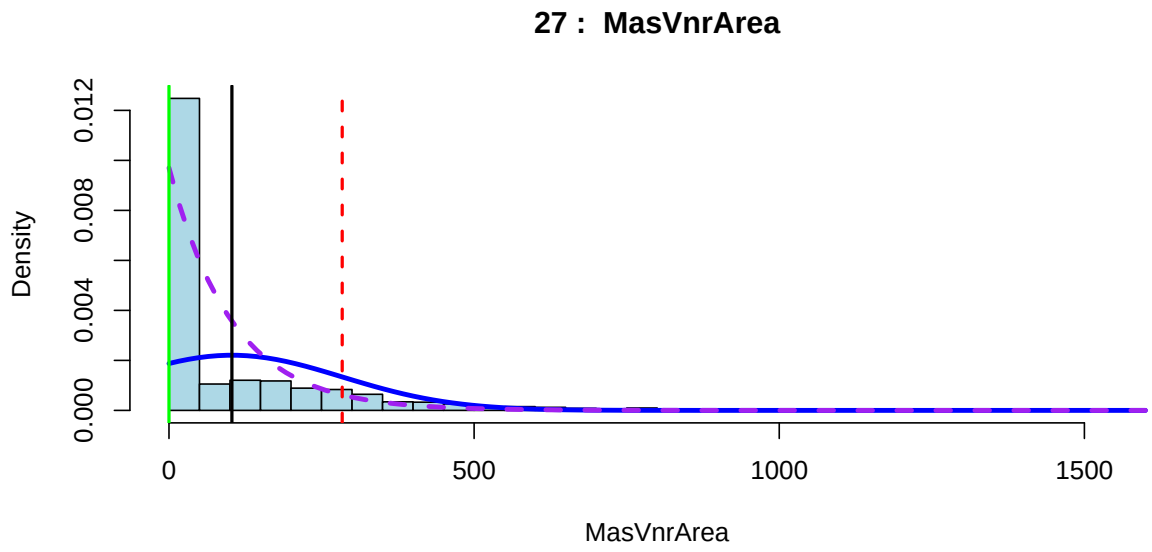

```
## Exterior1st
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard ImStucc MetalSd Plywood
##      20       1       2      50       1      61     222       1     220     108
##   Stone  Stucco VinylSd Wd Sdng WdShing    None
##       2      25     515     206      26       0
```

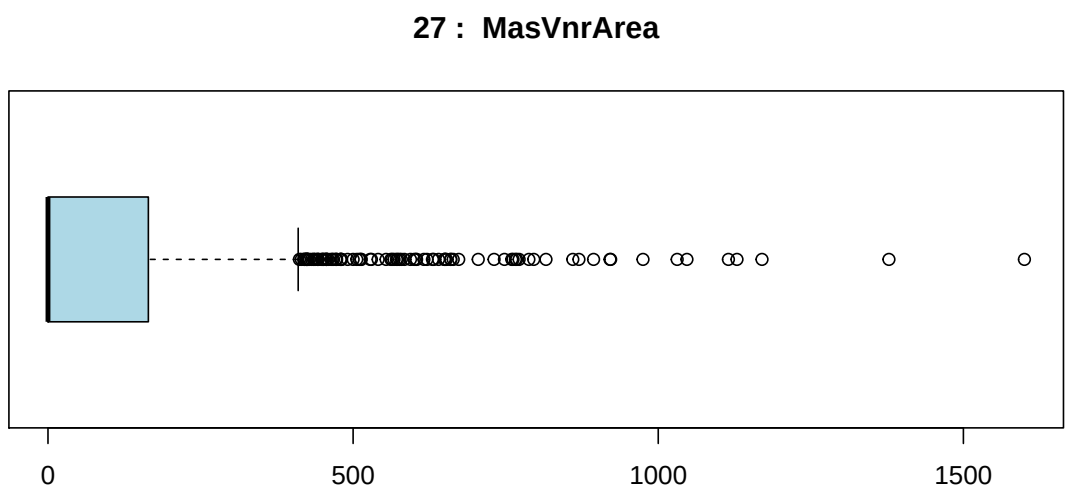



```
## Exterior2nd
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd   Other
##      20       3       7      25       1      60     207      10     214       1
## Plywood   Stone  Stucco VinylSd Wd Sdng Wd Shng    None
##     142       5      26     504     197      38       0
```

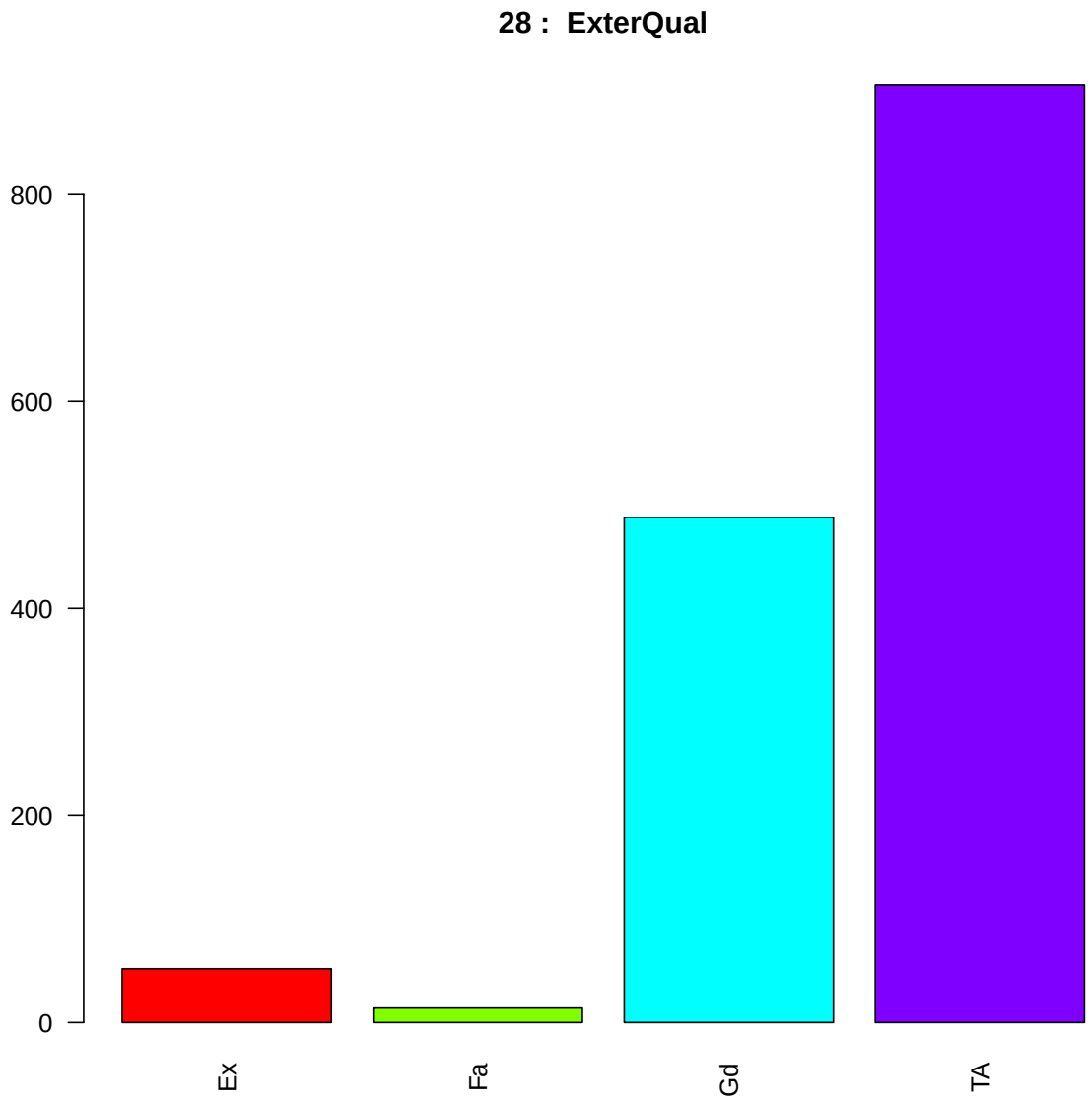



```
## MasVnrType
##  BrkCmn BrkFace    None   Stone
##      15     445     872     128
```

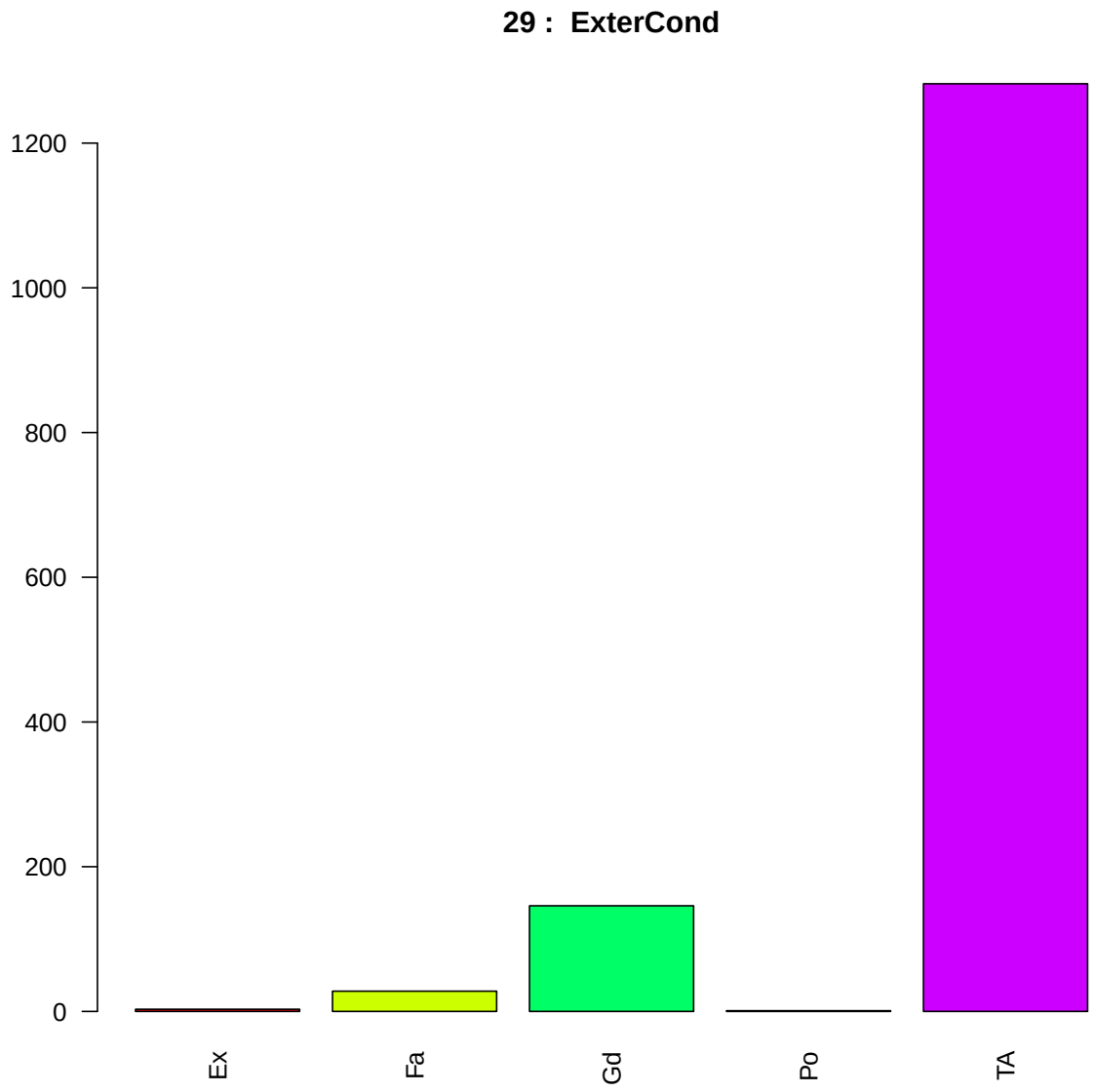




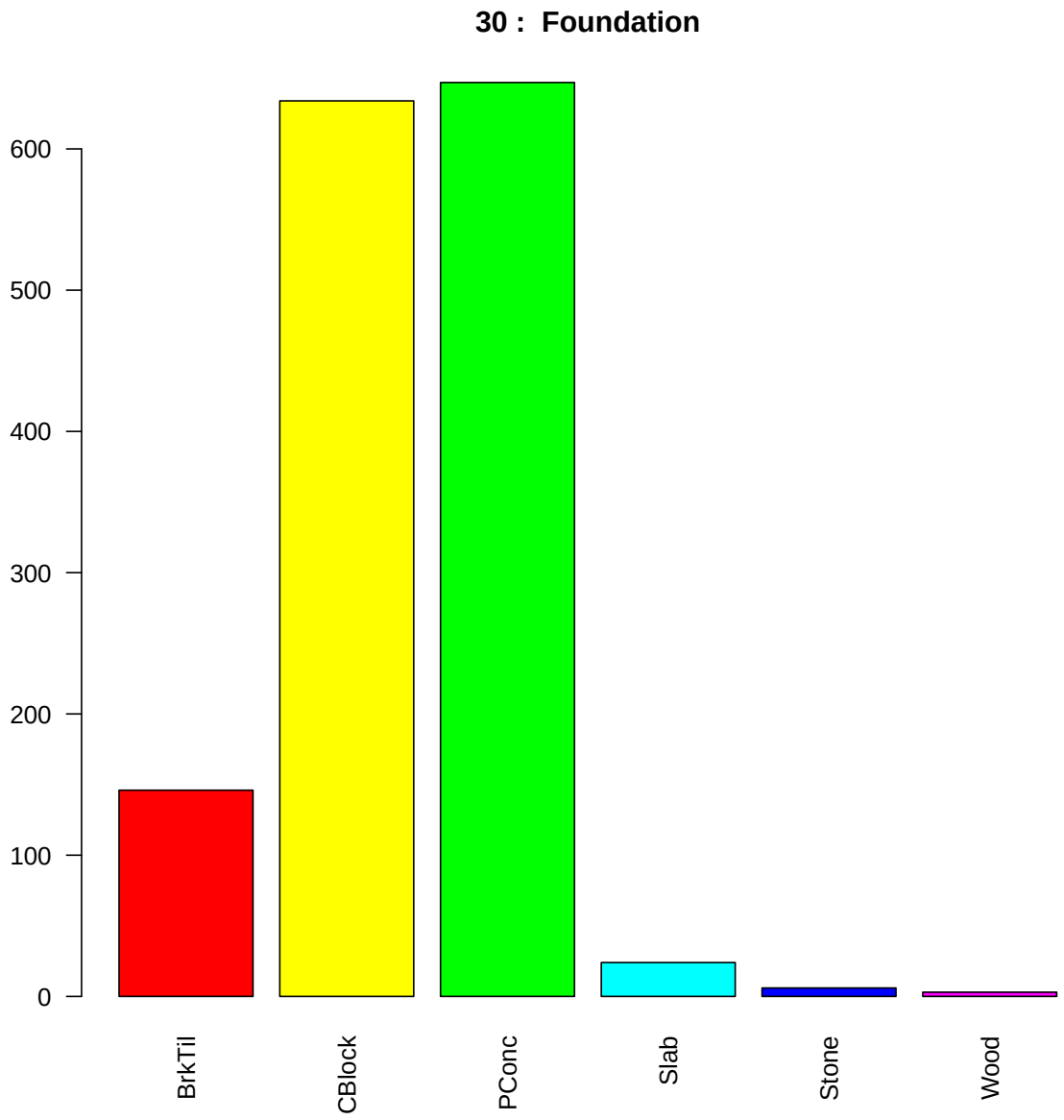

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000    0.000000    0.000000  103.117123  164.250000 1600.000000
##       STDEV
##  180.731373
```



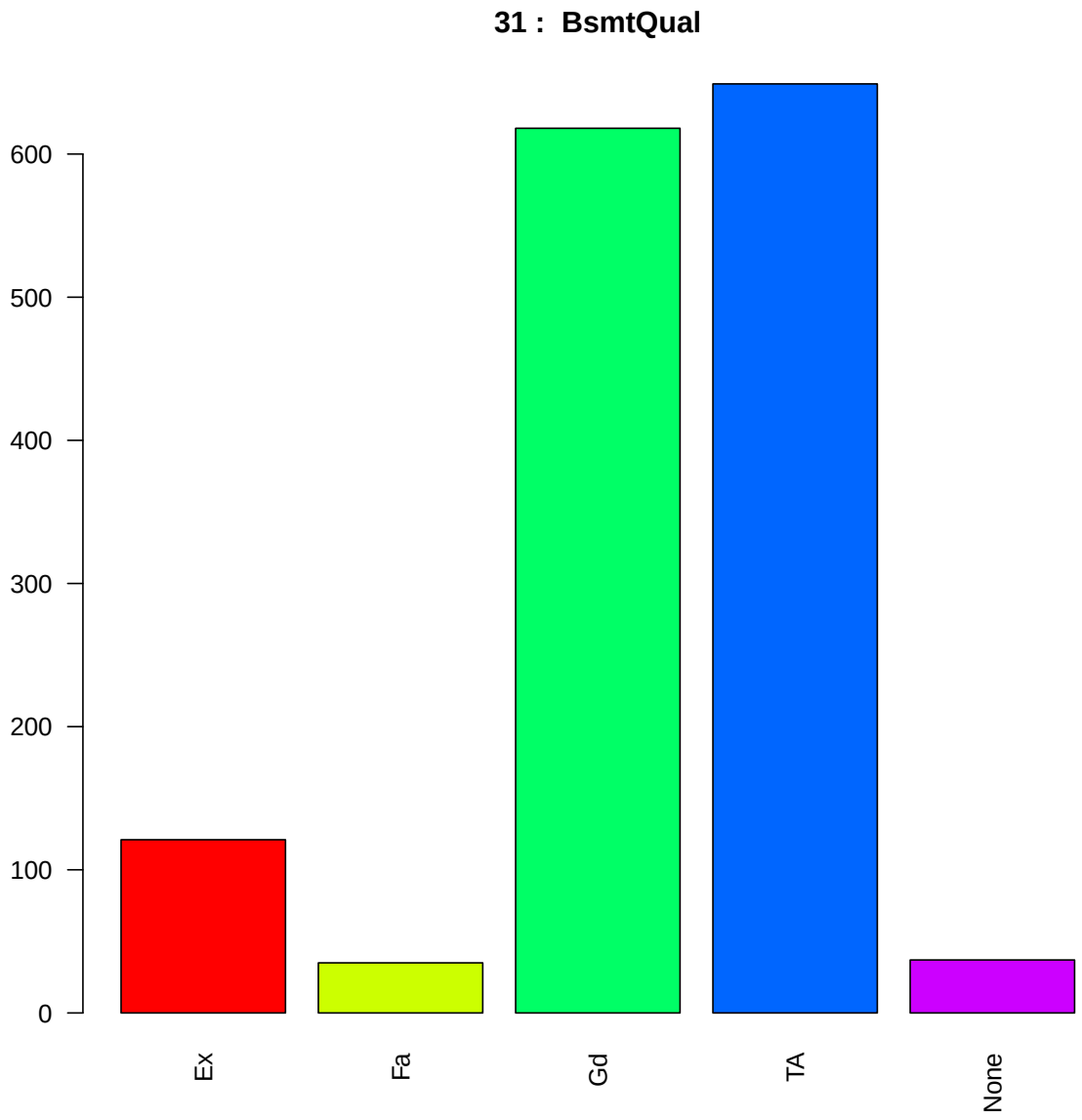

```
## ExterQual
##  Ex  Fa  Gd  TA
##  52  14 488 906
```



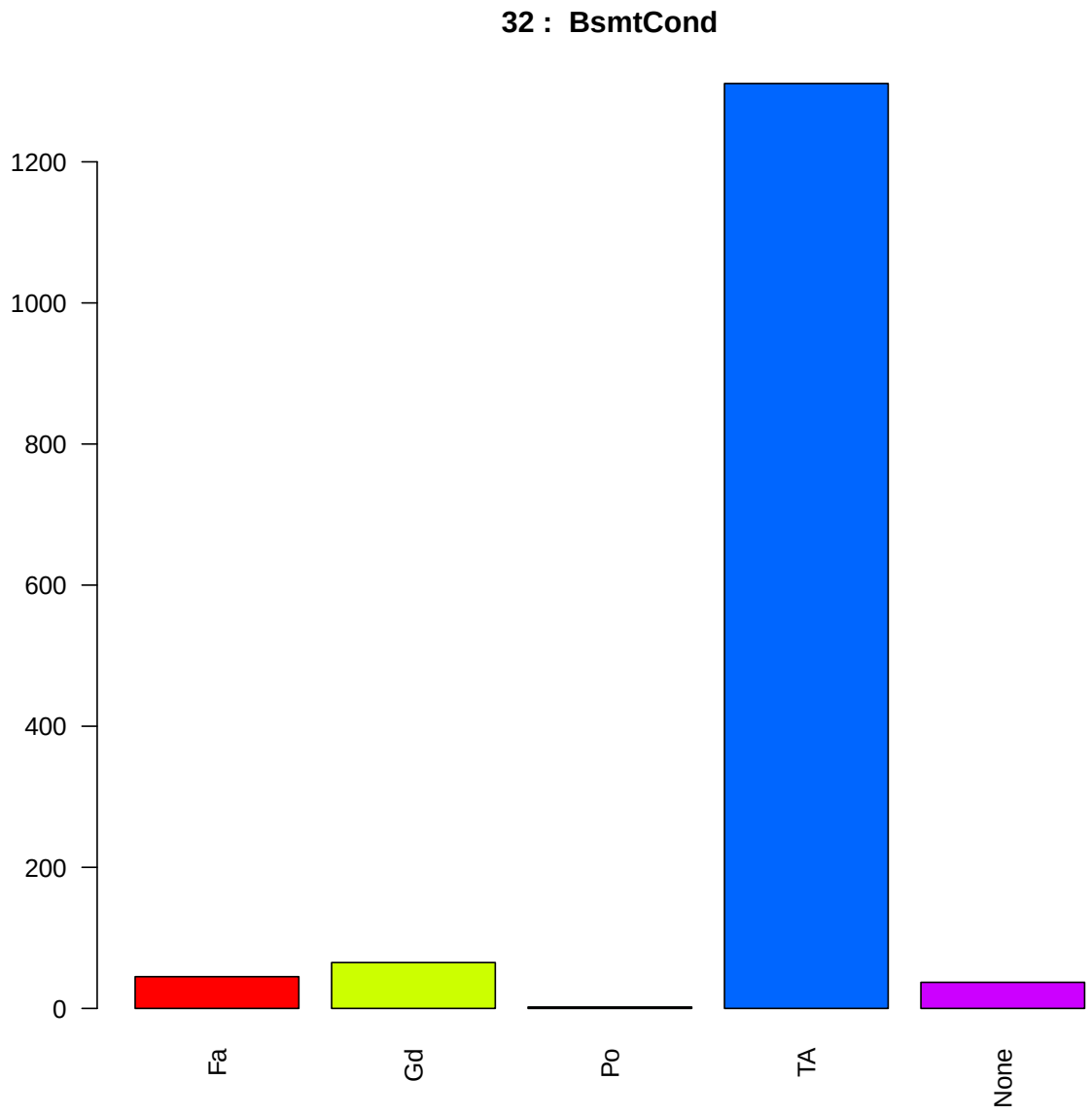

```
## ExterCond
##   Ex   Fa   Gd   Po   TA
##    3   28  146    1 1282
```



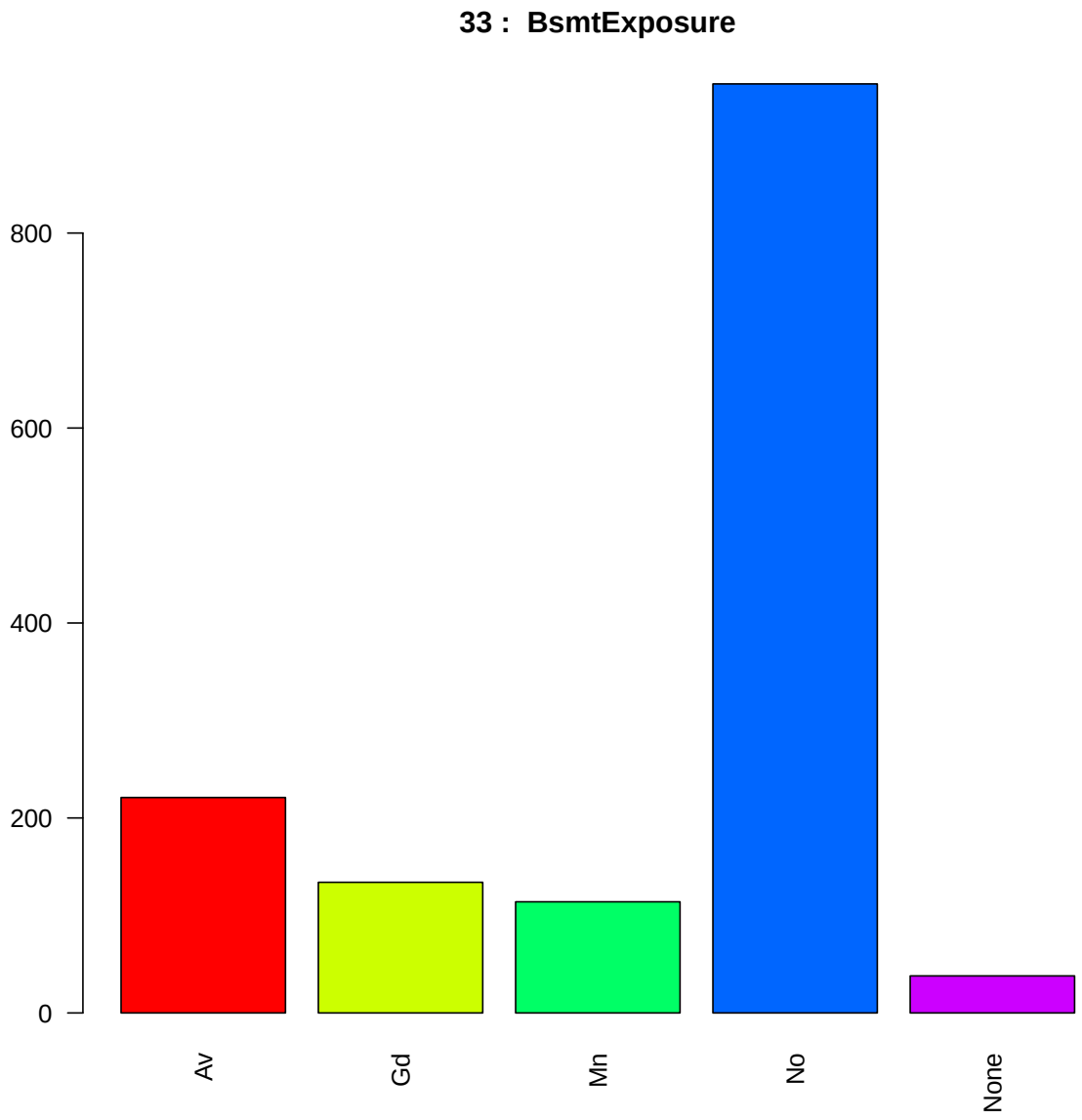

```
## Foundation
## BrkTil CBlock  PConc   Slab  Stone   Wood
##    146    634    647     24      6      3
```



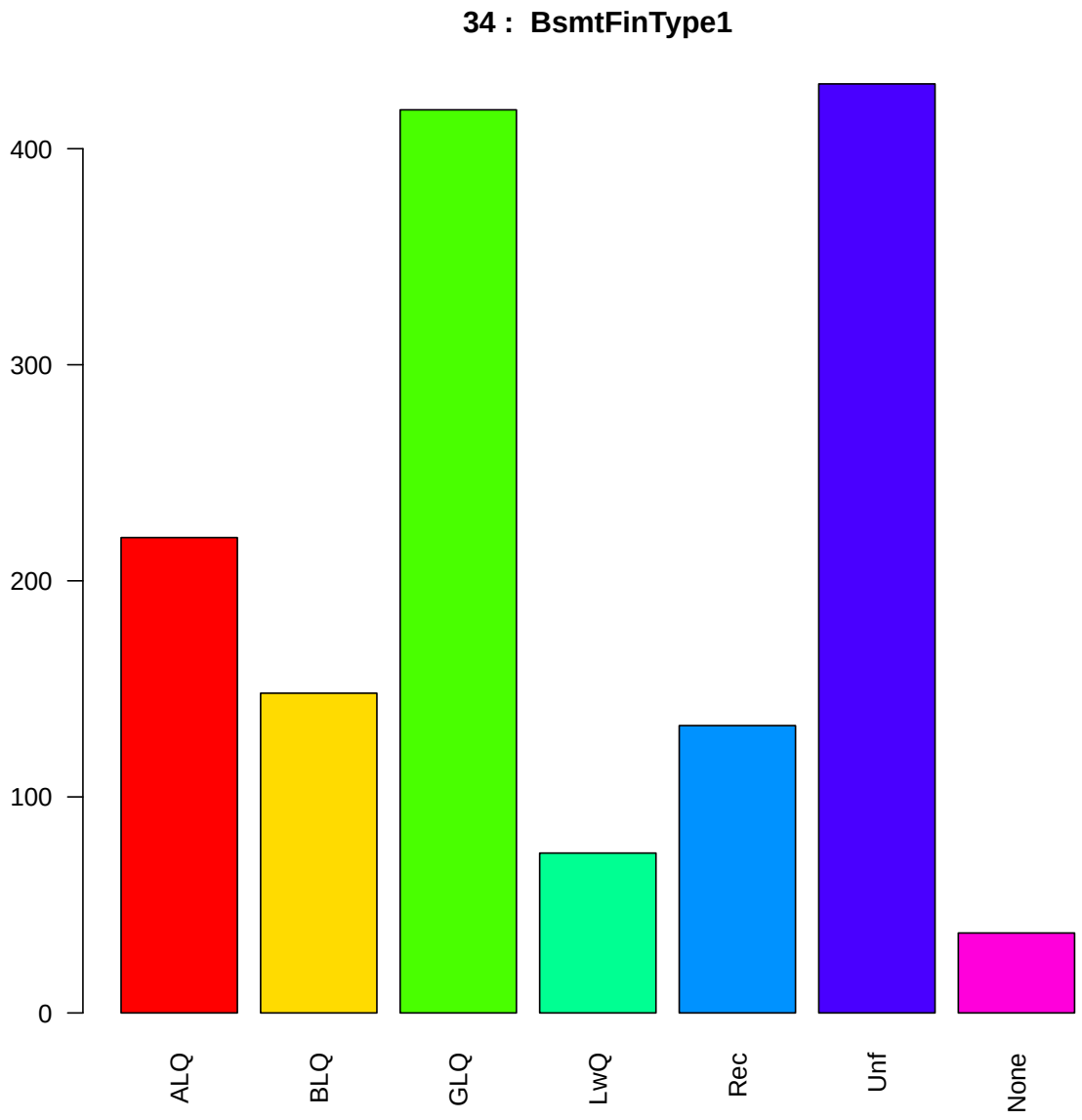

```
## BsmtQual
##   Ex   Fa   Gd   TA None
##  121   35  618  649   37
```



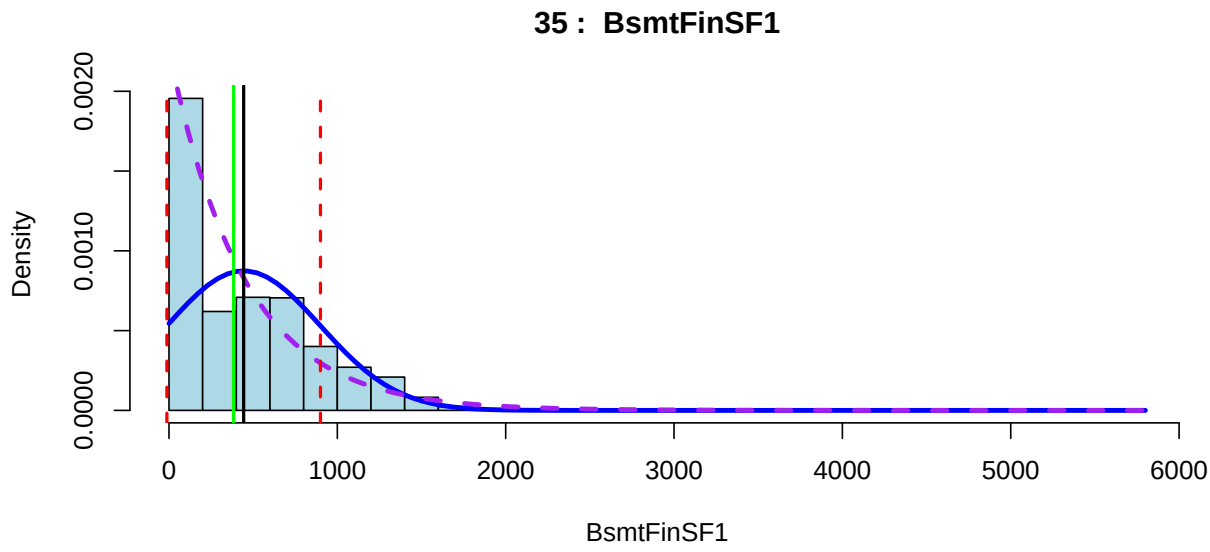

```
## BsmtCond
##   Fa   Gd   Po   TA None
##   45   65    2 1311   37
```

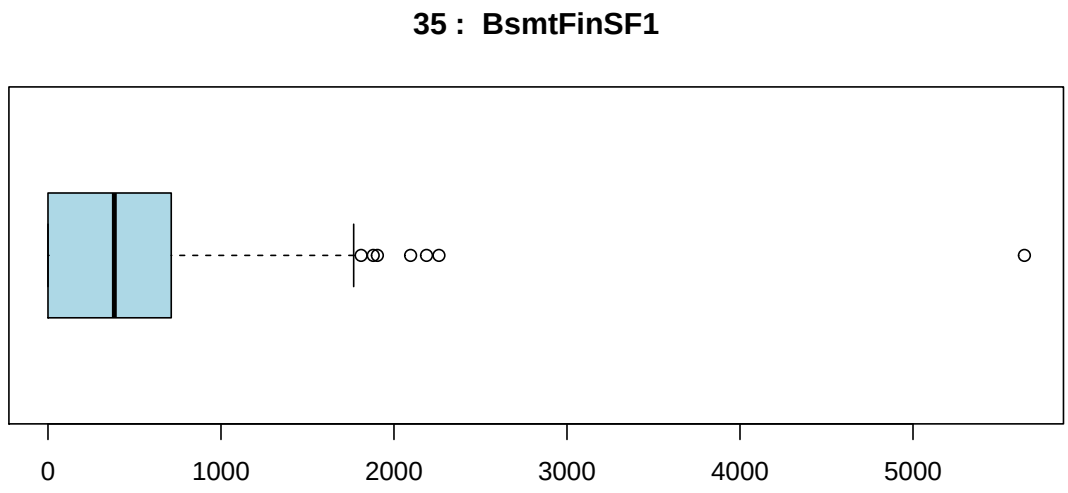



```
## BsmtExposure
##   Av   Gd   Mn   No None
##  221  134  114  953   38
```

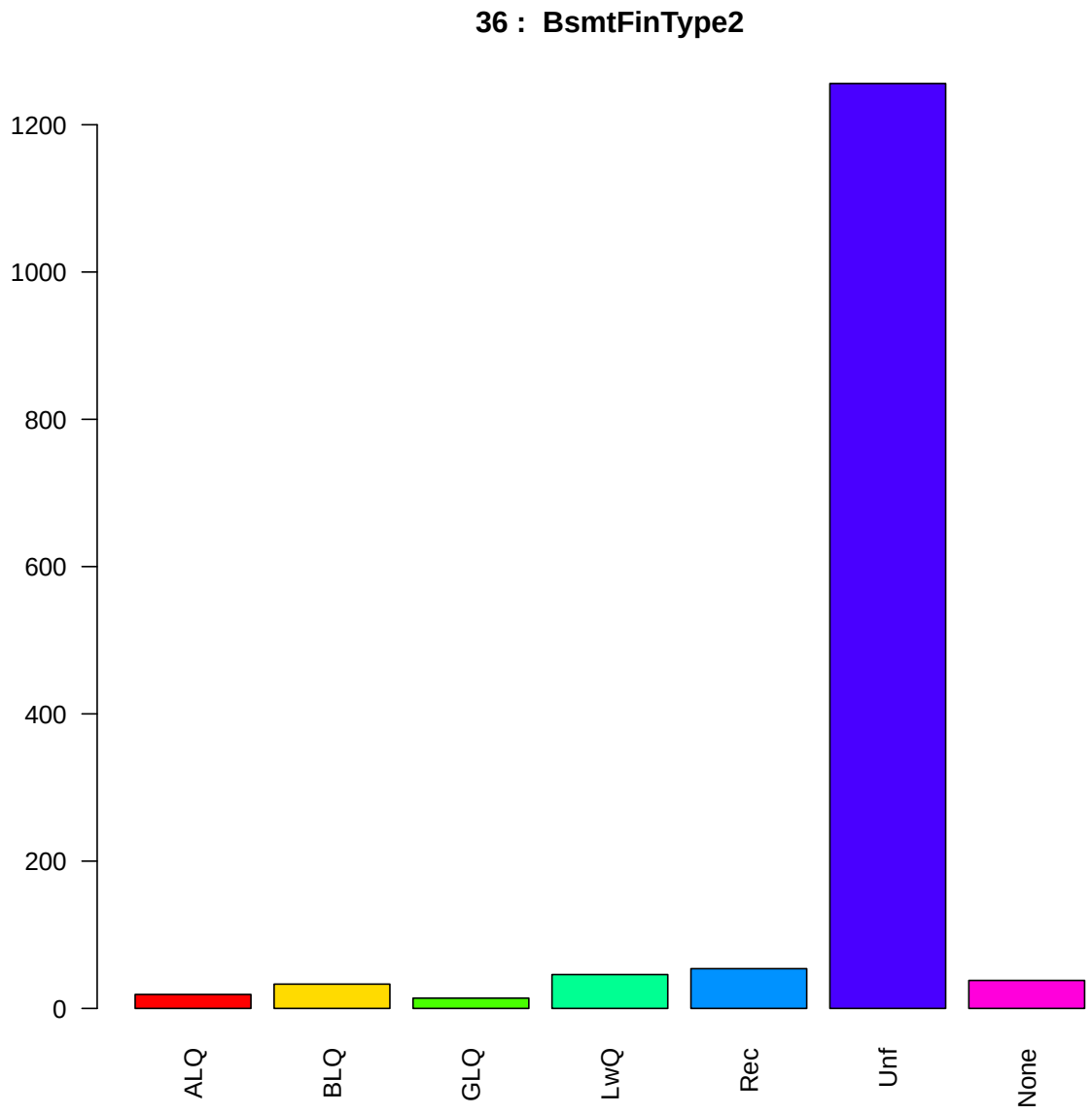



```
## BsmtFinType1
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf None
##  220  148  418   74  133  430   37
```

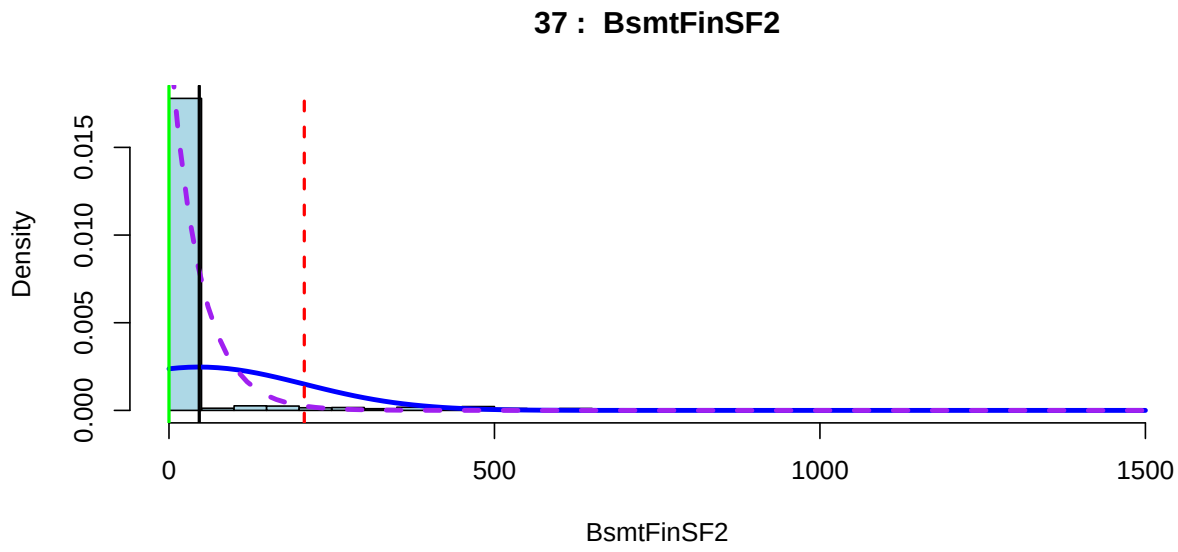


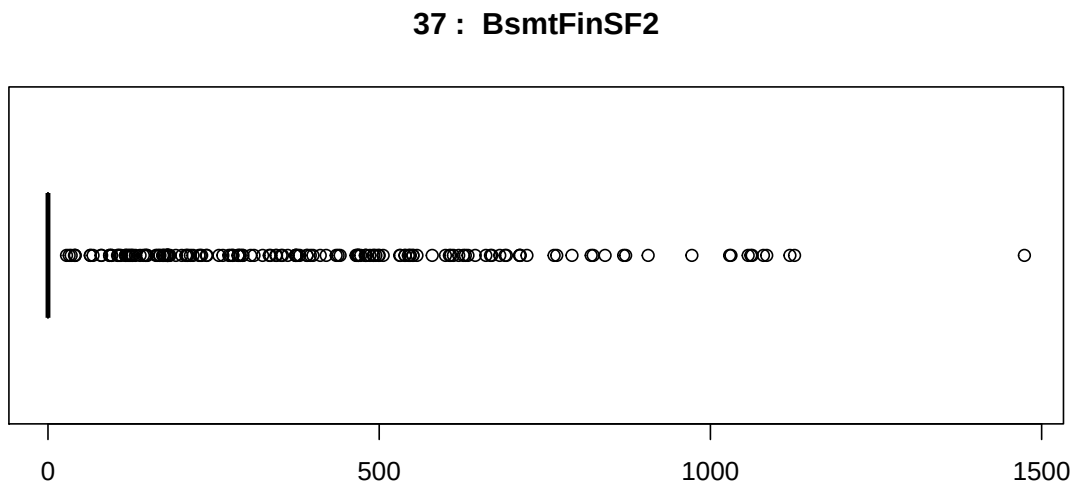



```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000    0.000000  383.500000  443.639726  712.250000 5644.000000
##       STDEV
##  456.098091
```

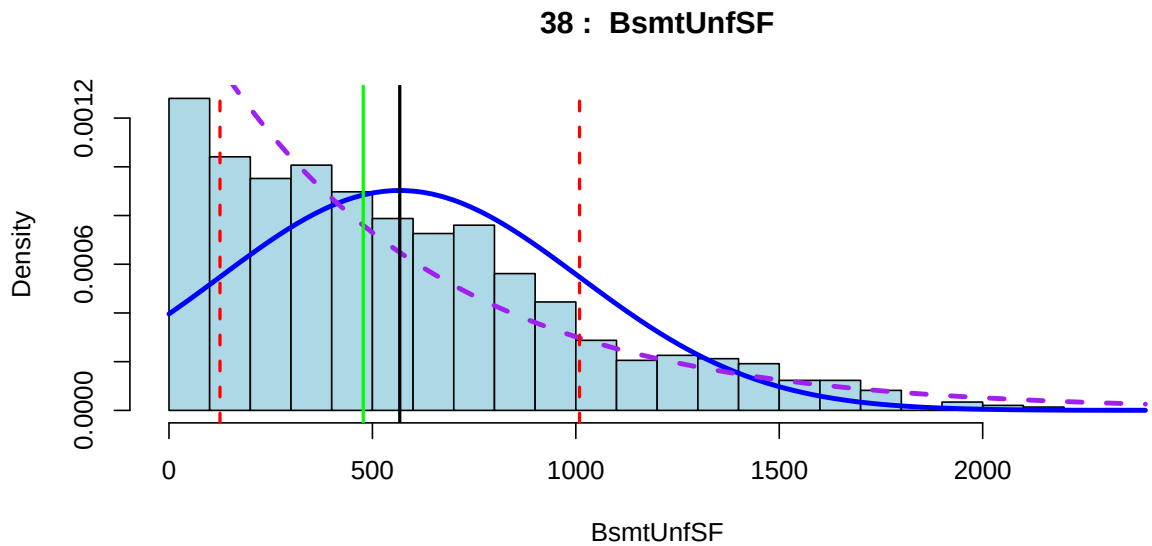


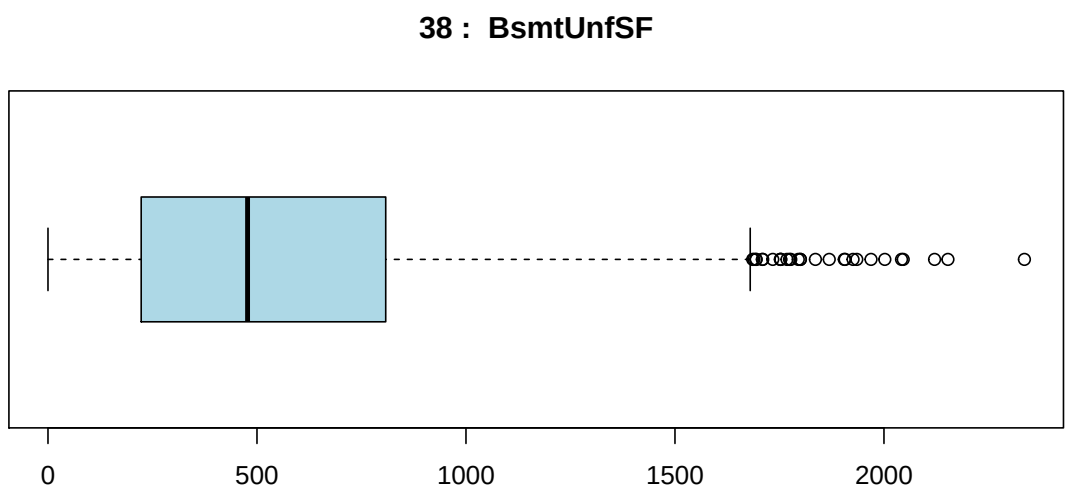

```
## BsmtFinType2
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf None
##   19   33   14   46   54 1256   38
```



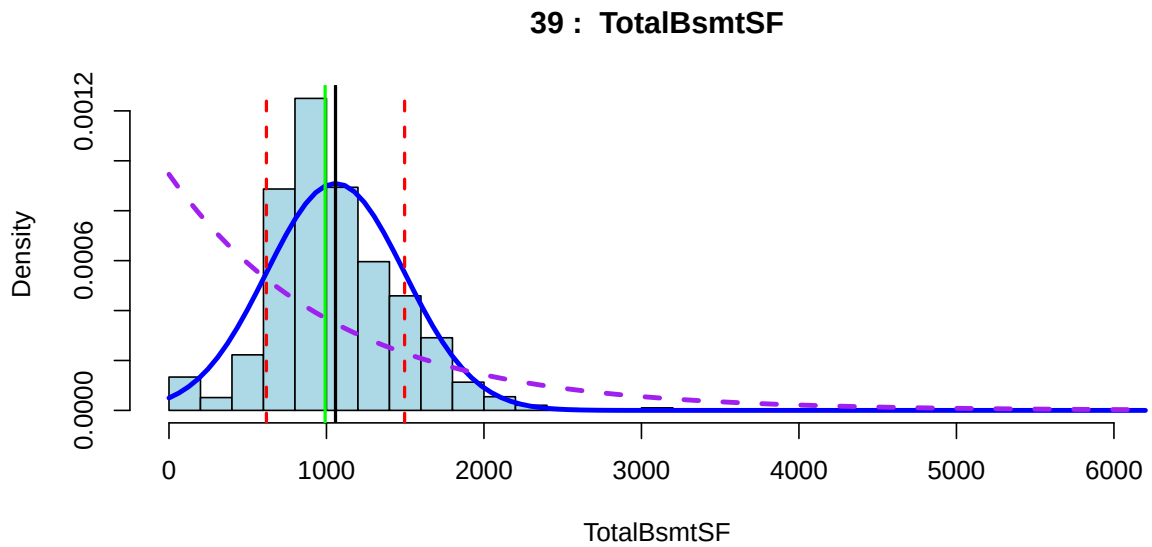


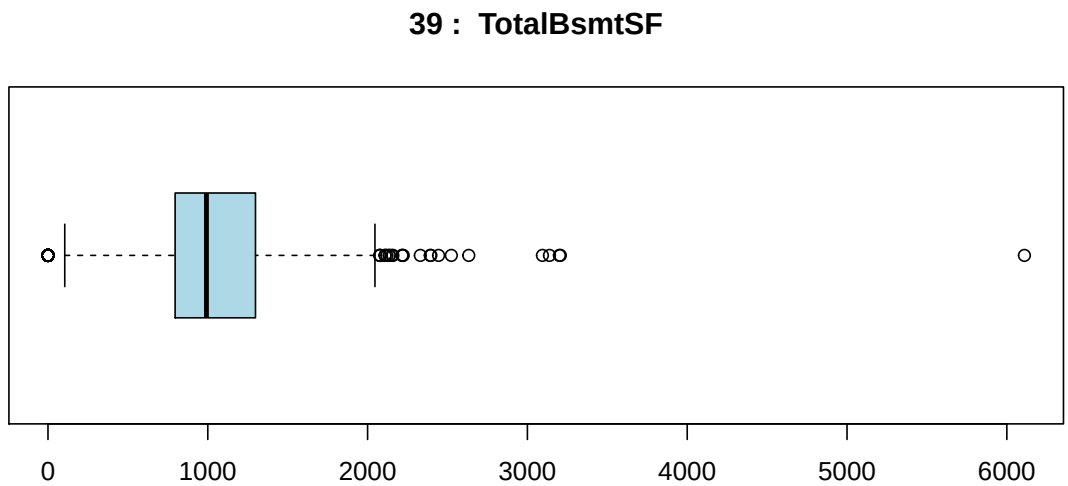

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##    0.0000000    0.0000000    0.0000000   46.5493151    0.0000000 1474.0000000
##        STDEV
##  161.3192728
```



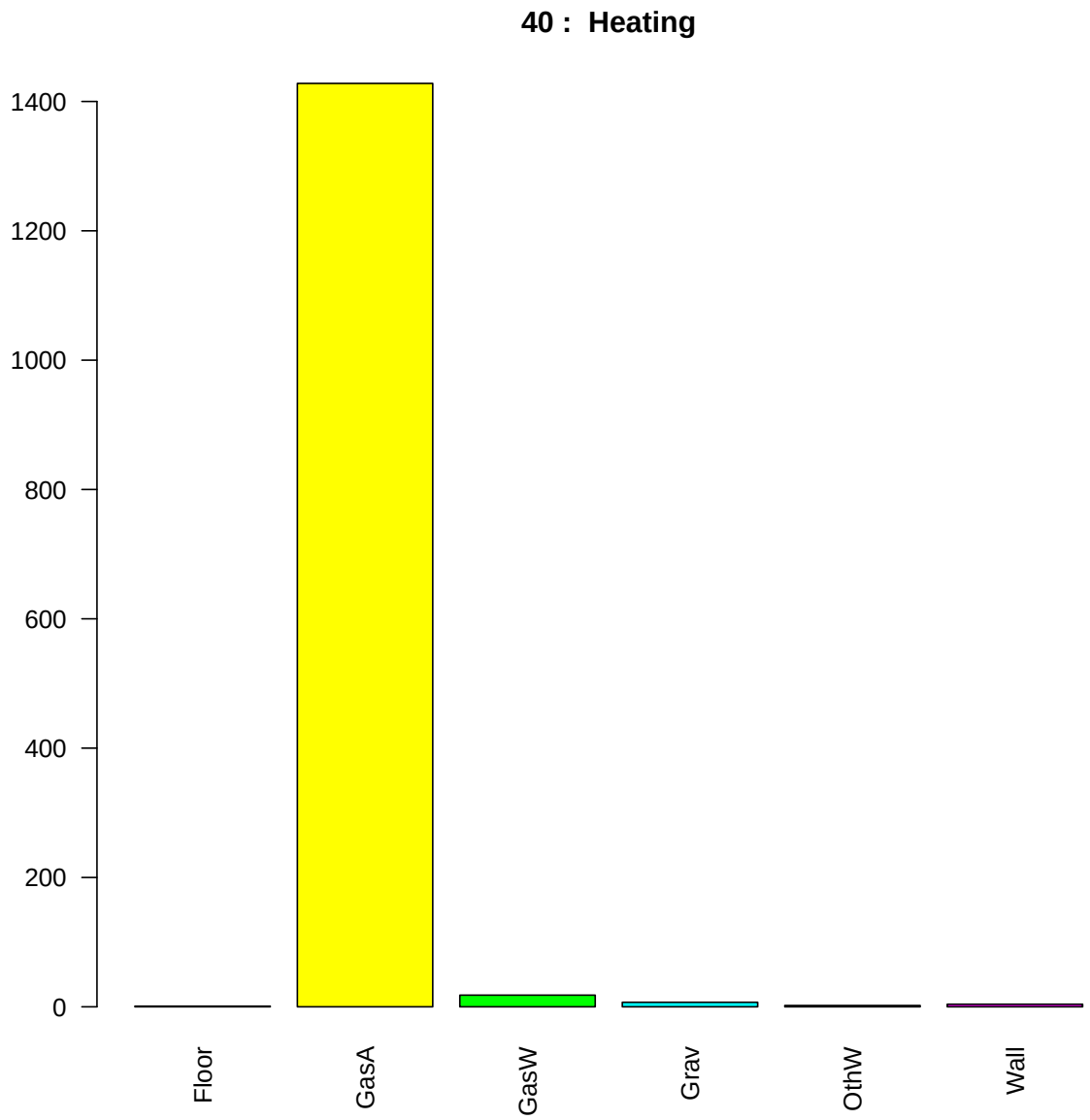



```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000  223.000000  477.500000  567.240411  808.000000 2336.000000
##       STDEV
##  441.866955
```

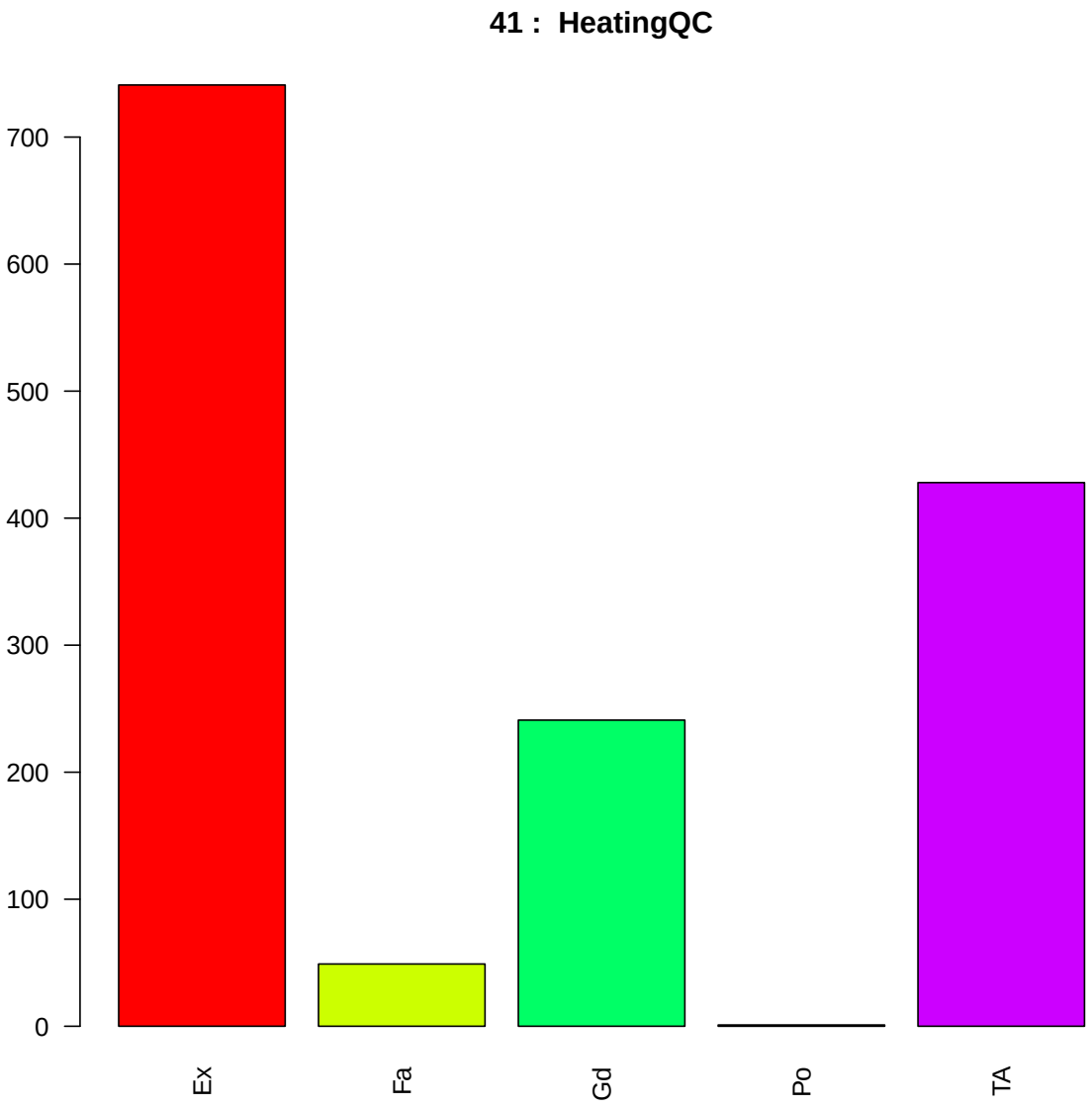




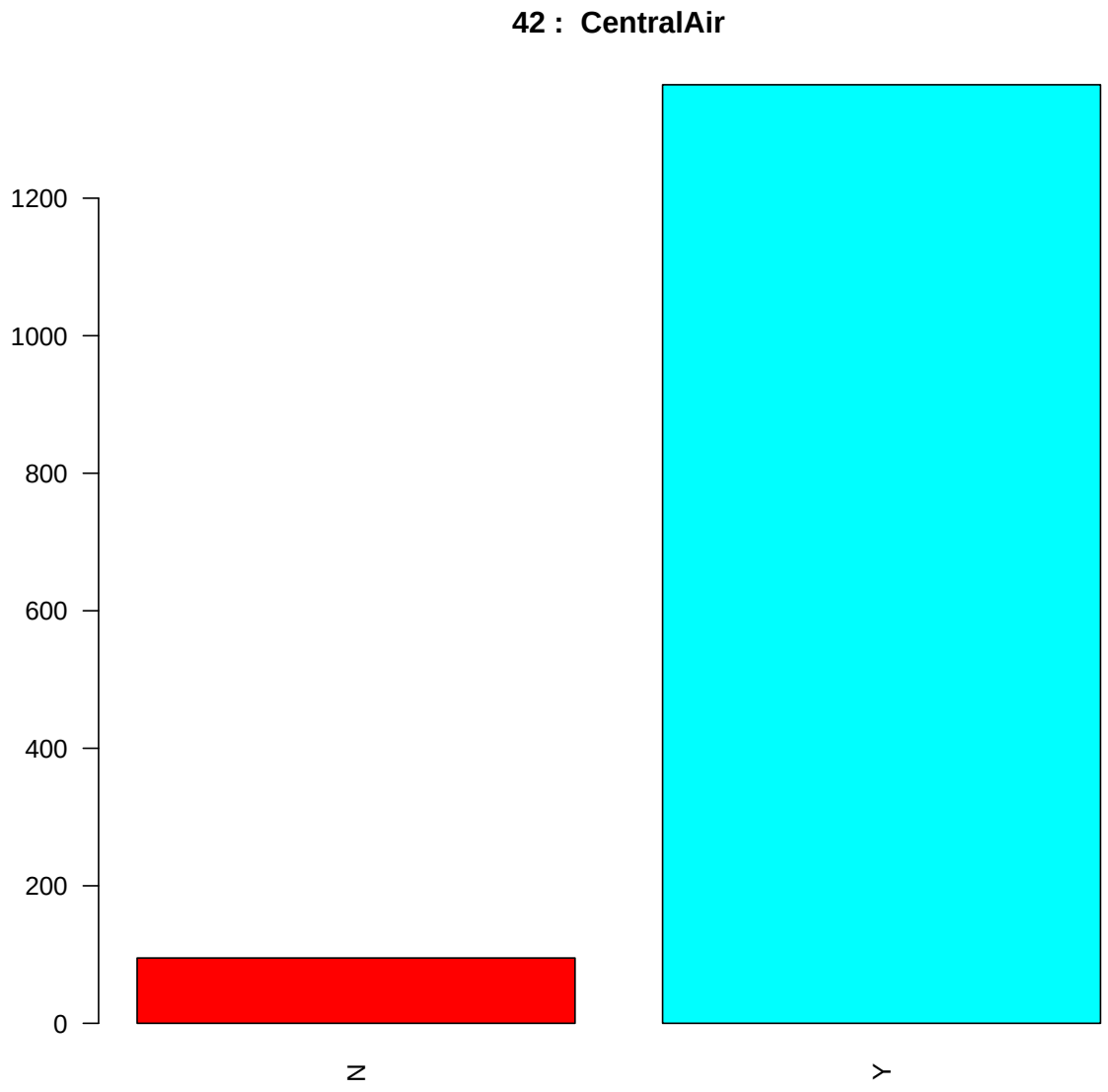

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000  795.750000  991.500000 1057.429452 1298.250000 6110.000000
##       STDEV
##  438.705324
```



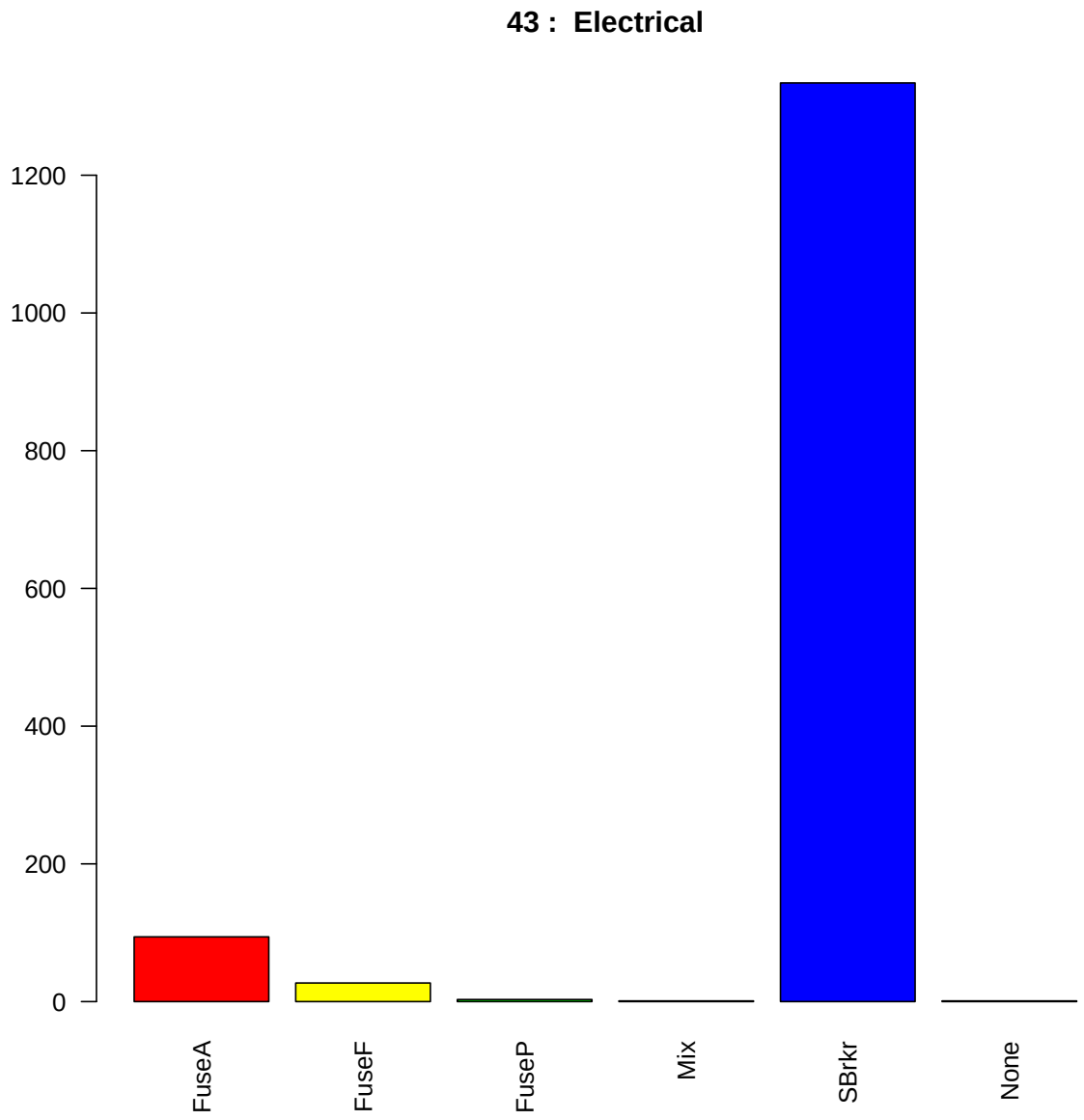

```
## Heating
## Floor  GasA  GasW  Grav  OthW  Wall
##     1  1428    18     7     2     4
```



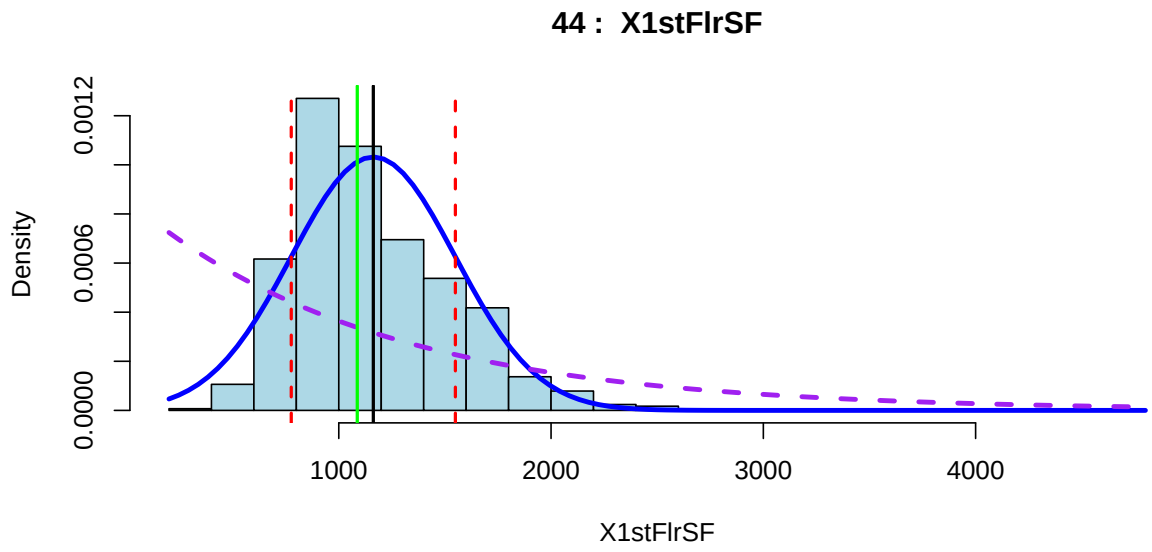

```
## HeatingQC
##  Ex  Fa  Gd  Po  TA
## 741  49 241   1 428
```

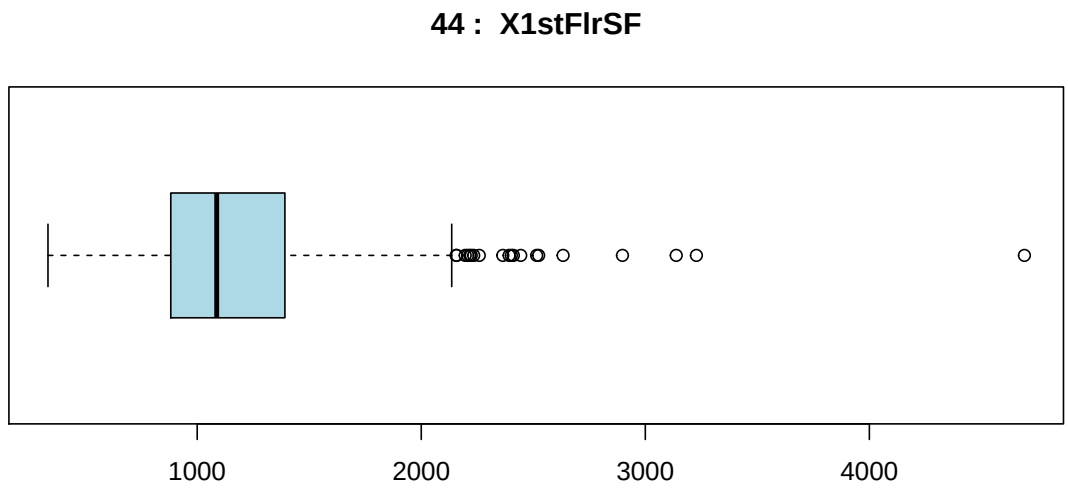



```
## CentralAir
##    N    Y
##   95 1365
```

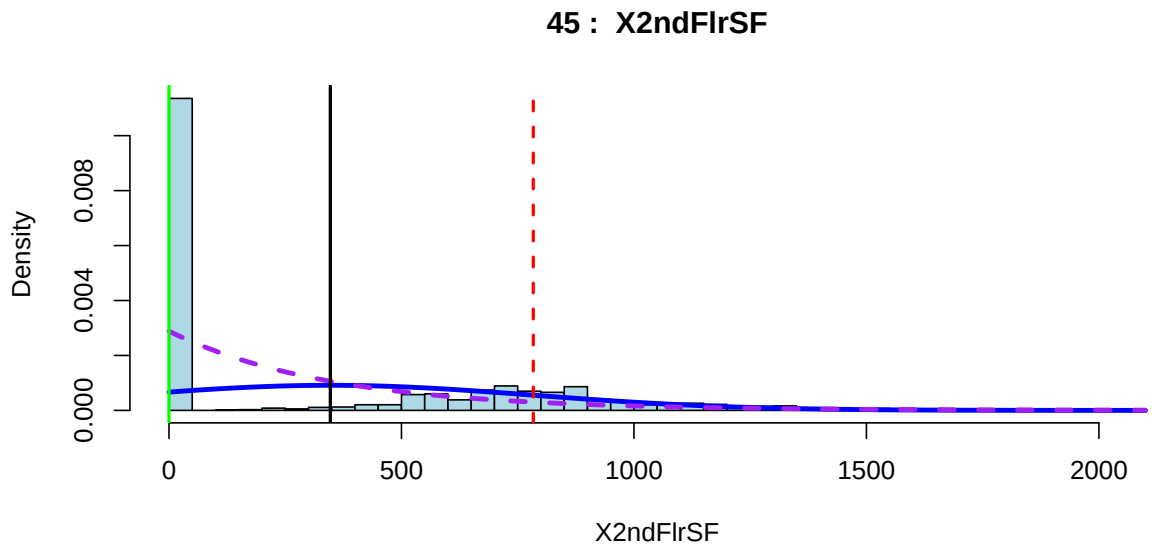


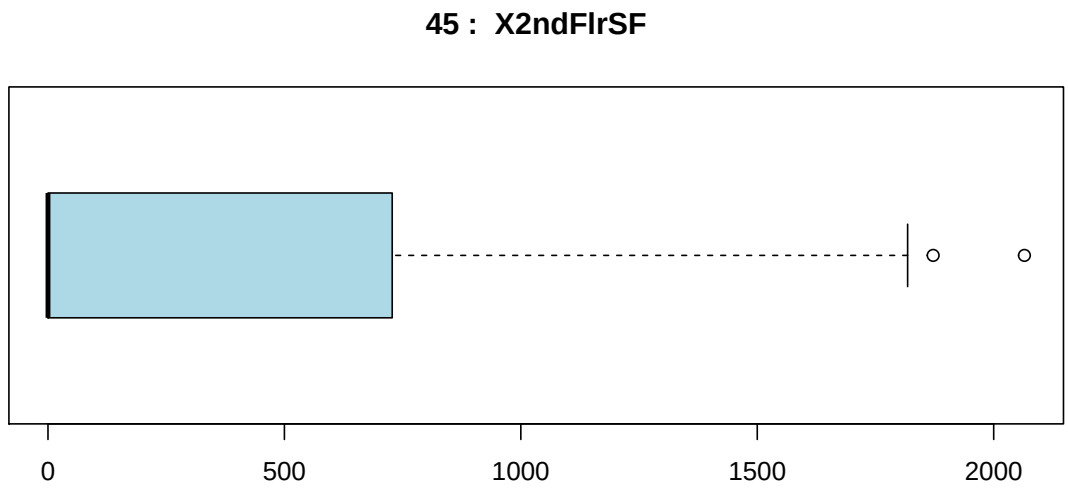

```
## Electrical
## FuseA FuseF FuseP   Mix SBrkr  None
##    94    27     3     1  1334     1
```



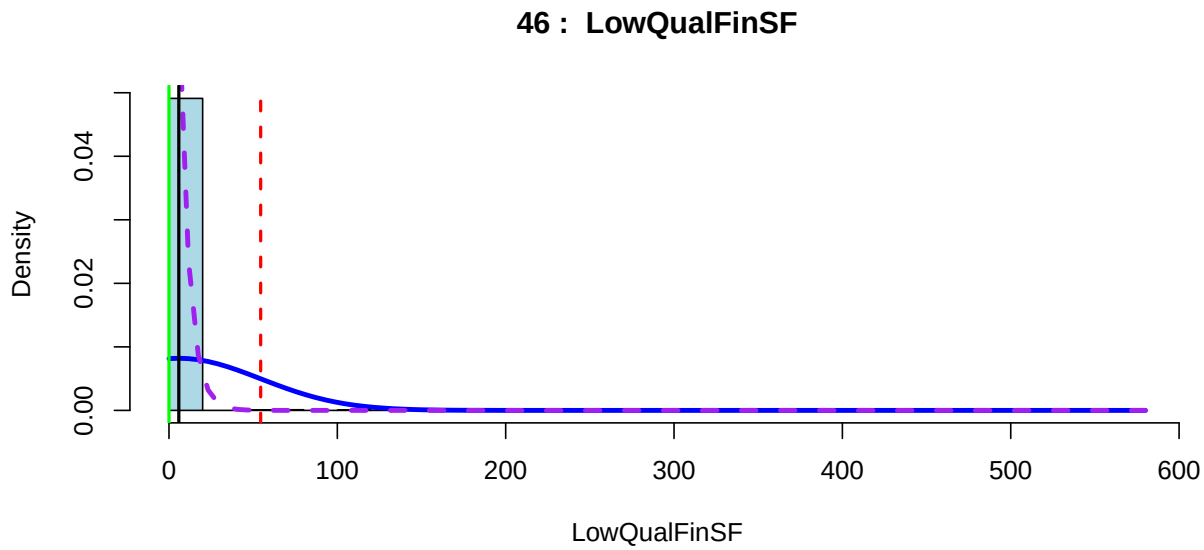


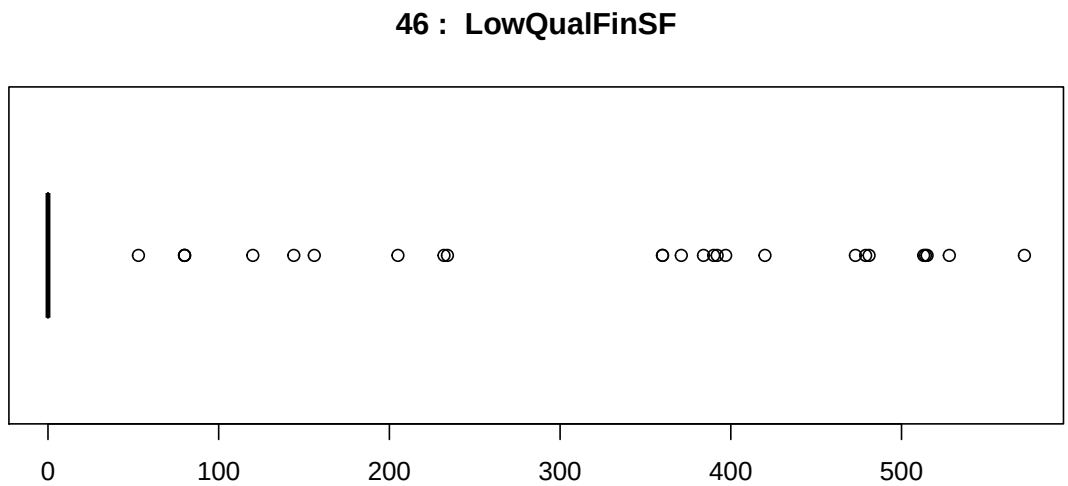

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##  334.000000  882.000000 1087.000000 1162.626712 1391.250000 4692.000000
##       STDEV
##  386.587738
```



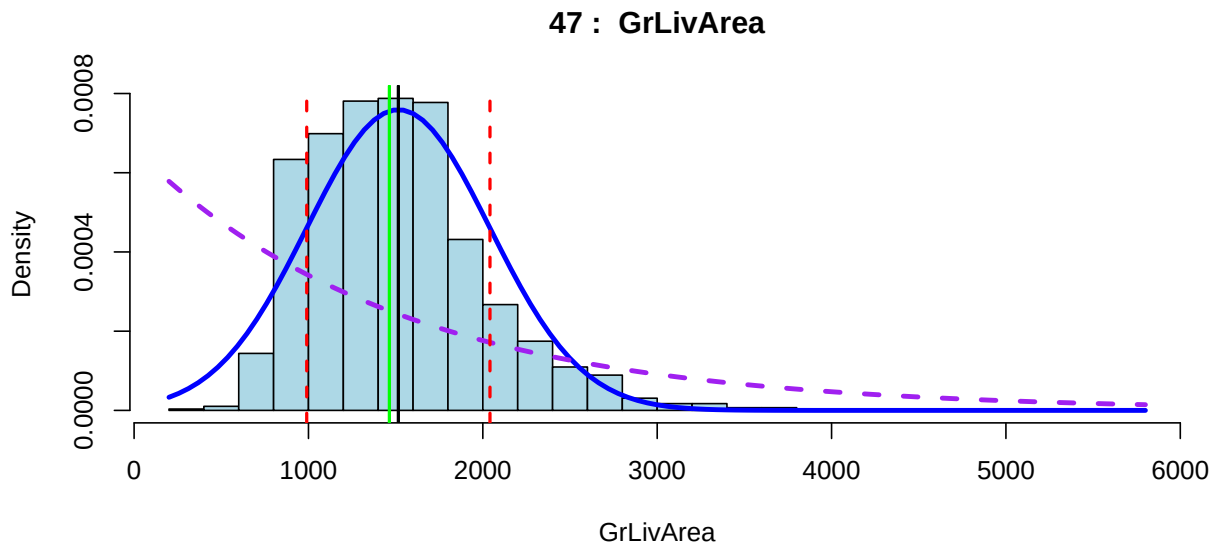


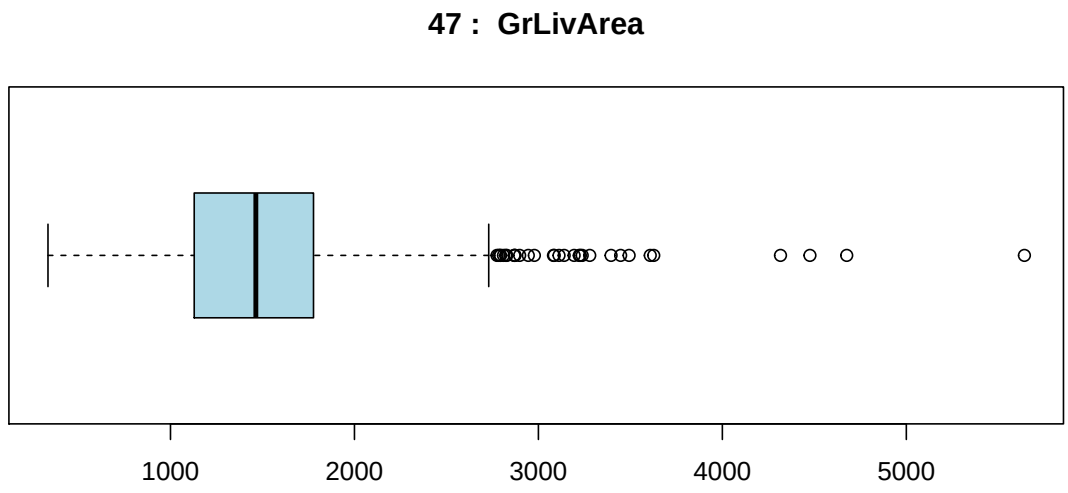

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000    0.000000    0.000000  346.992466  728.000000 2065.000000
##       STDEV
##  436.528436
```



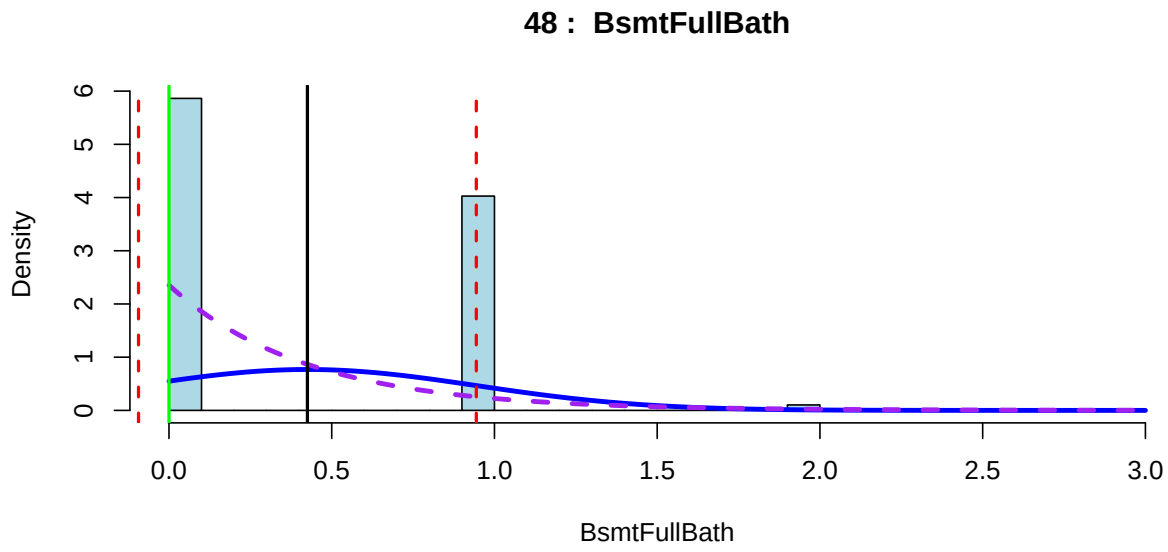


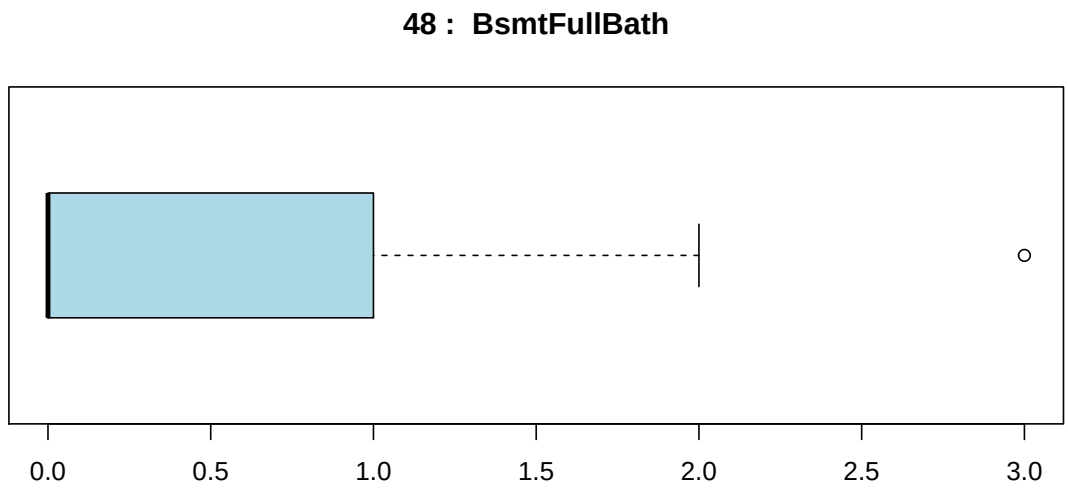

```
##          Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##   0.00000000   0.00000000   0.00000000   5.84452055   0.00000000 572.00000000
##         STDEV
##  48.62308143
```



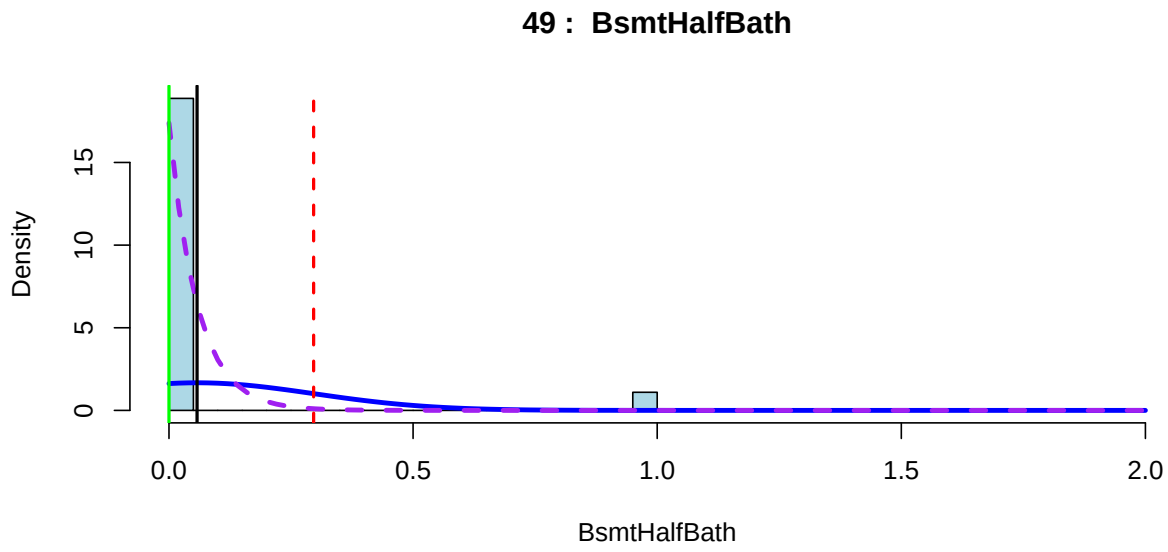


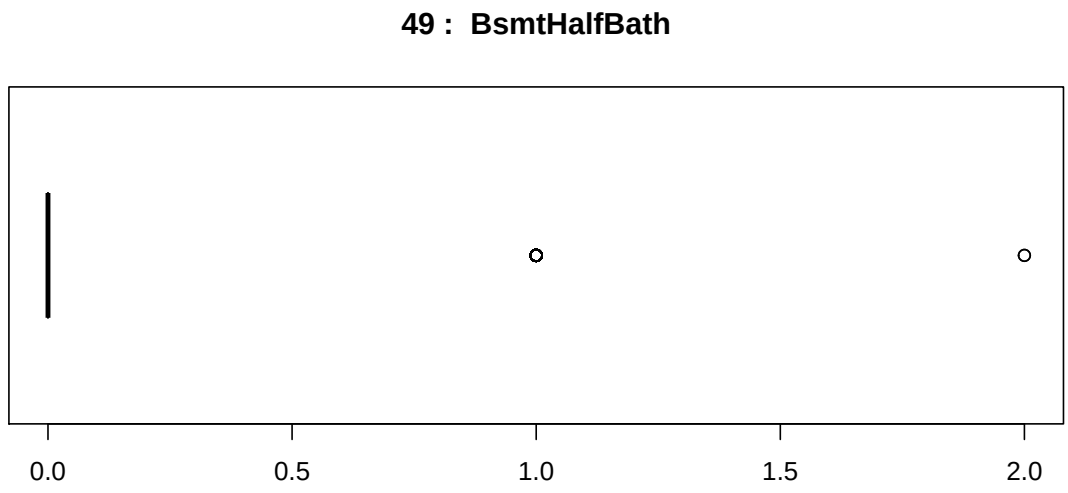

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##  334.000000 1129.500000 1464.000000 1515.463699 1776.750000 5642.000000
##       STDEV
##  525.480383
```



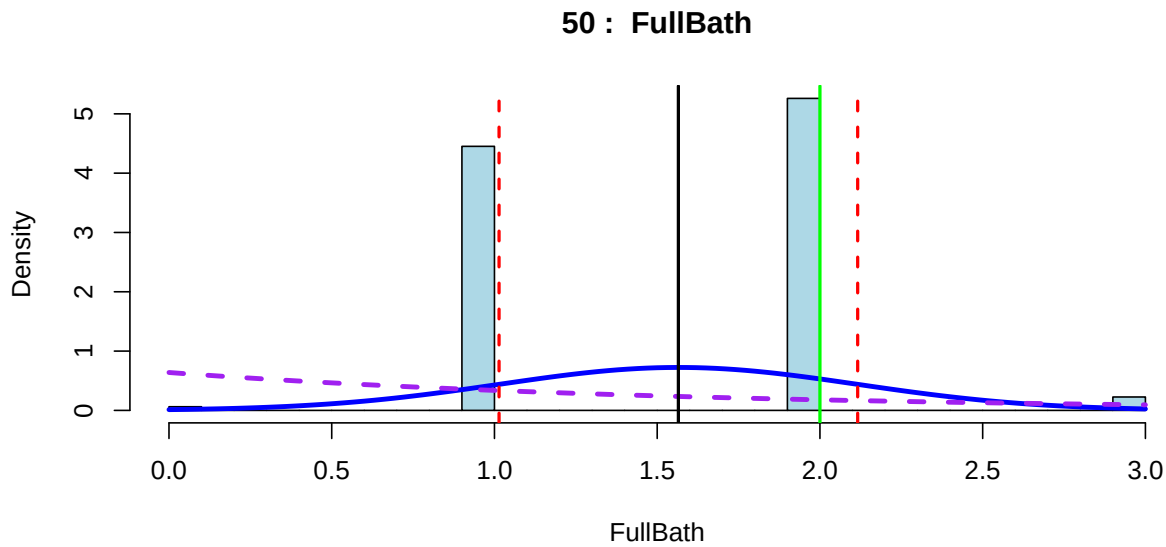


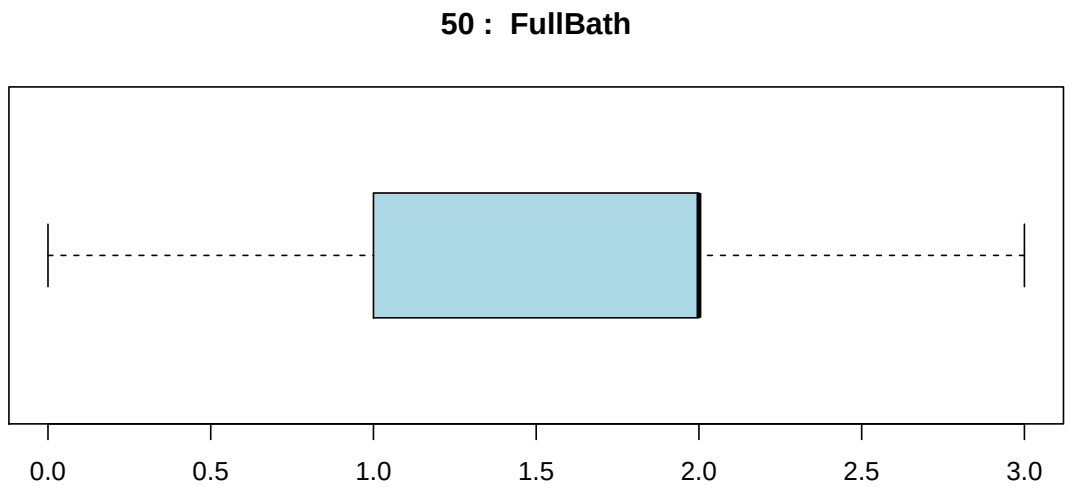

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## 0.000000000 0.000000000 0.000000000 0.425342466 1.000000000 3.000000000
##       STDEV
## 0.518910606
```



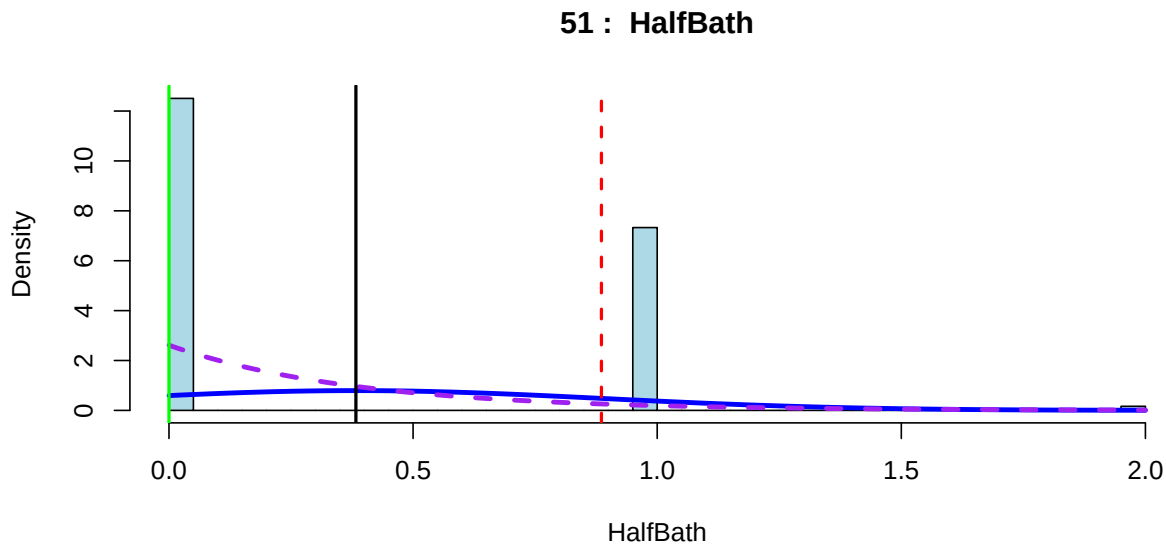


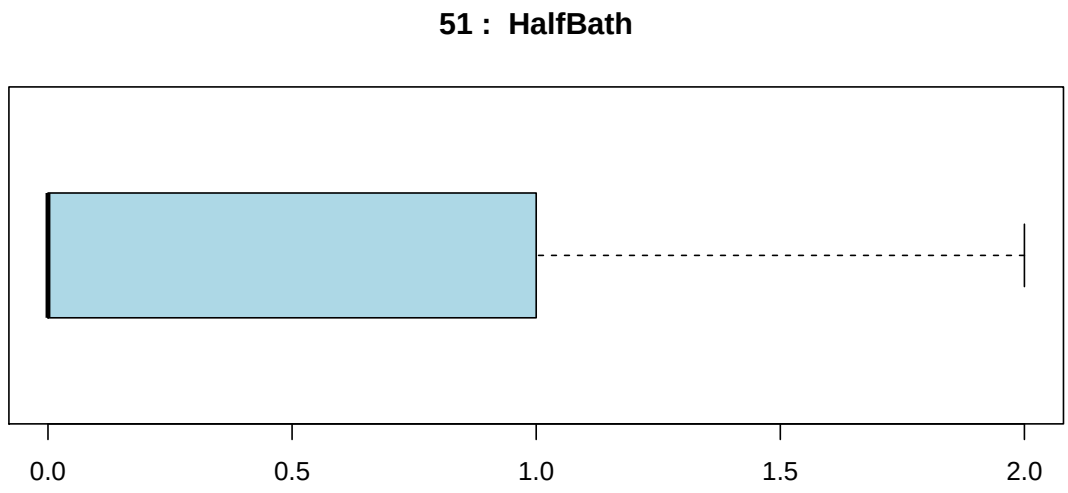

```
##          Min.       1st Qu.        Median          Mean       3rd Qu.          Max.
## 0.0000000000 0.0000000000 0.0000000000 0.0575342466 0.0000000000 2.0000000000
##         STDEV
## 0.2387526463
```



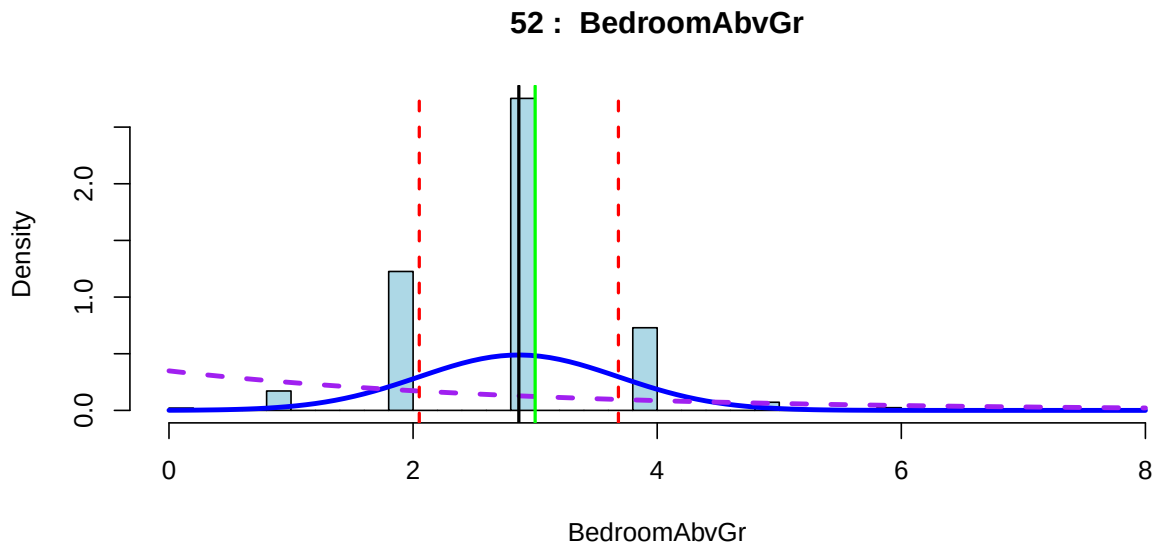


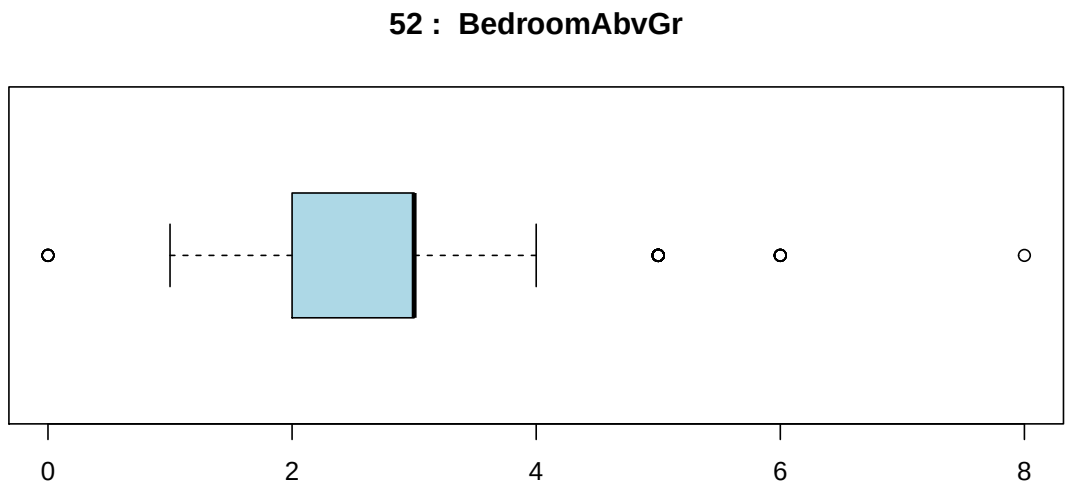

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## 0.000000000 1.000000000 2.000000000 1.565068493 2.000000000 3.000000000
##       STDEV
## 0.550915801
```



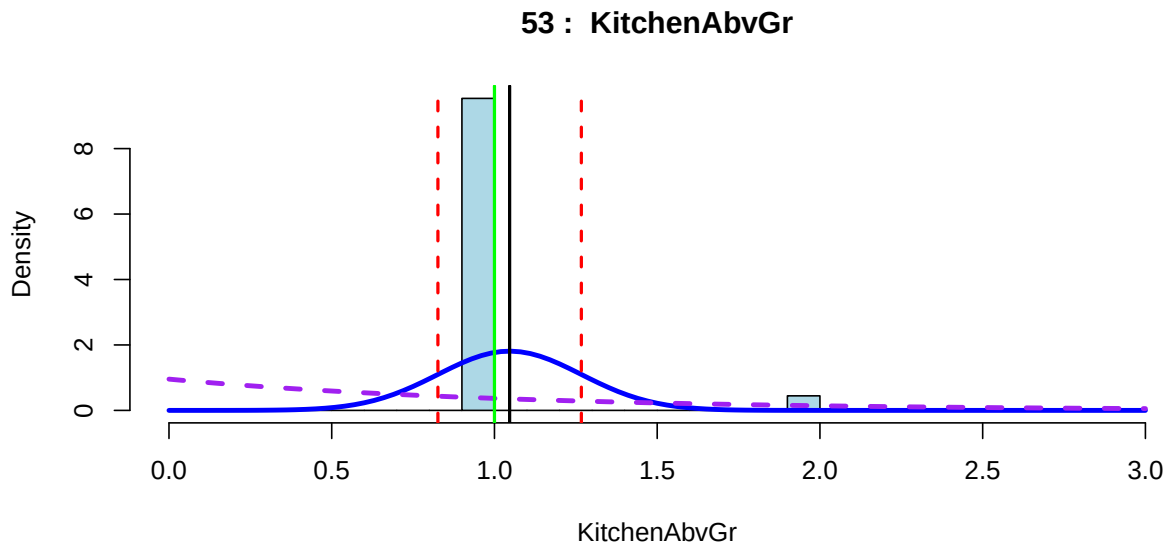


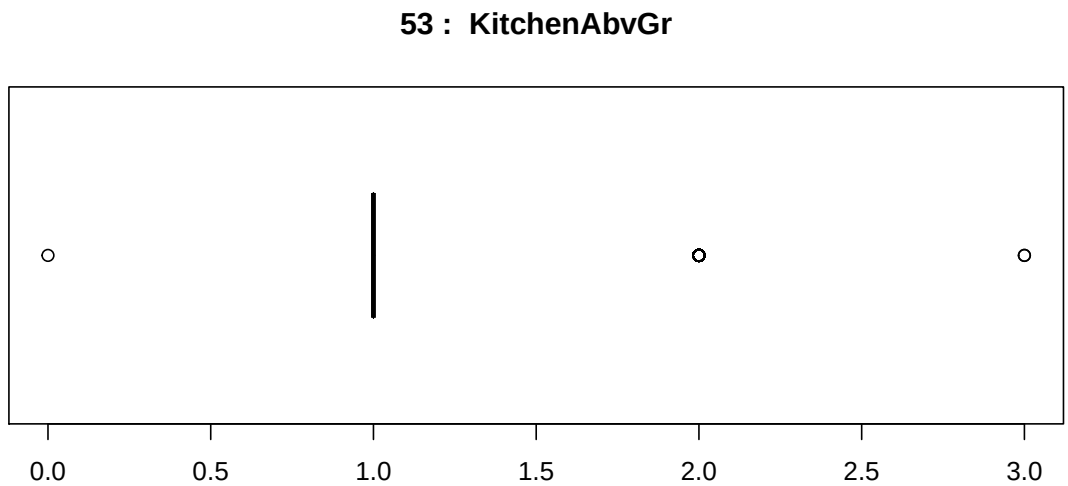

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## 0.000000000 0.000000000 0.000000000 0.382876712 1.000000000 2.000000000
##       STDEV
## 0.502885381
```



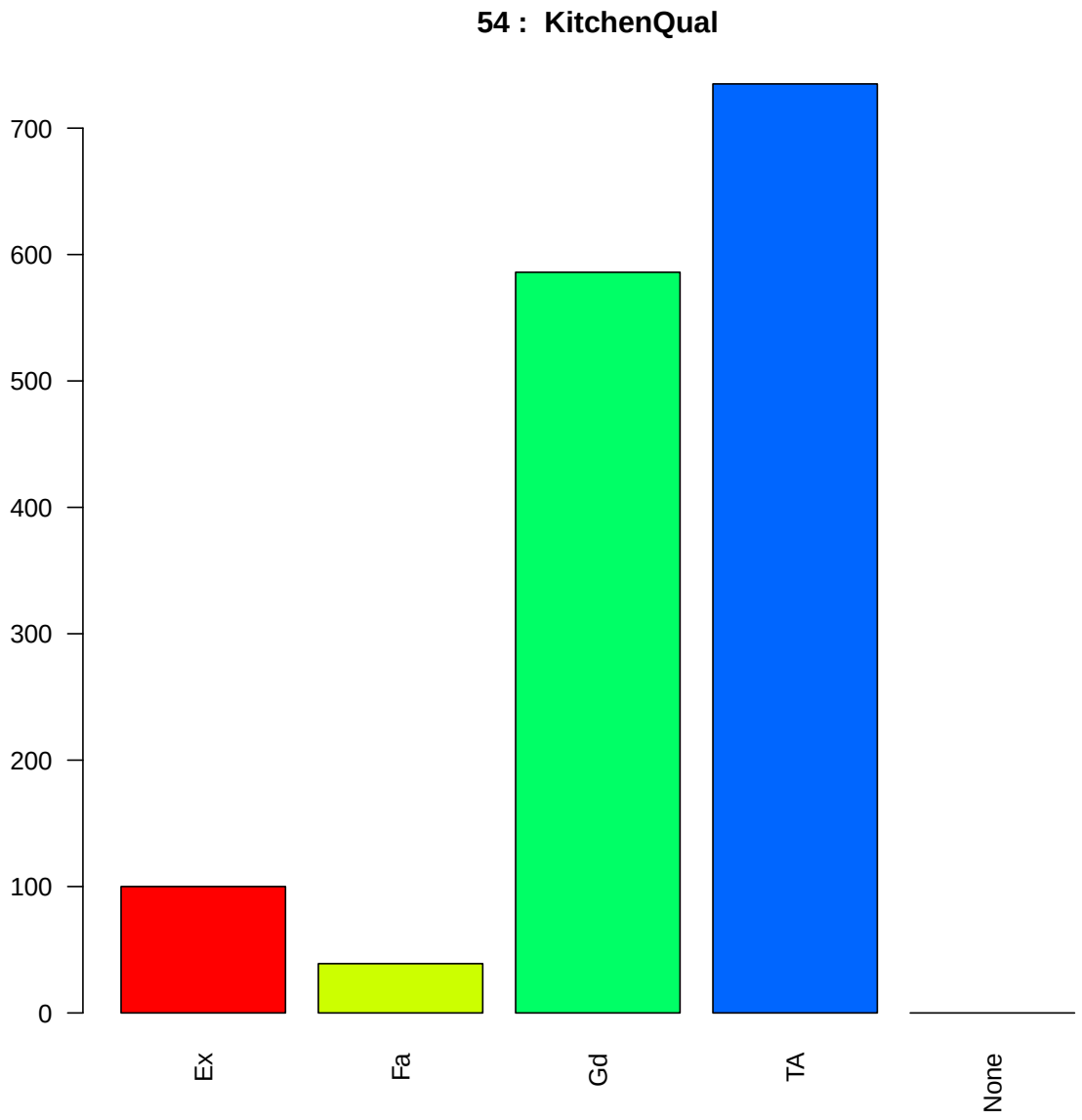



```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## 0.000000000 2.000000000 3.000000000 2.866438356 3.000000000 8.000000000
##       STDEV
## 0.815778044
```

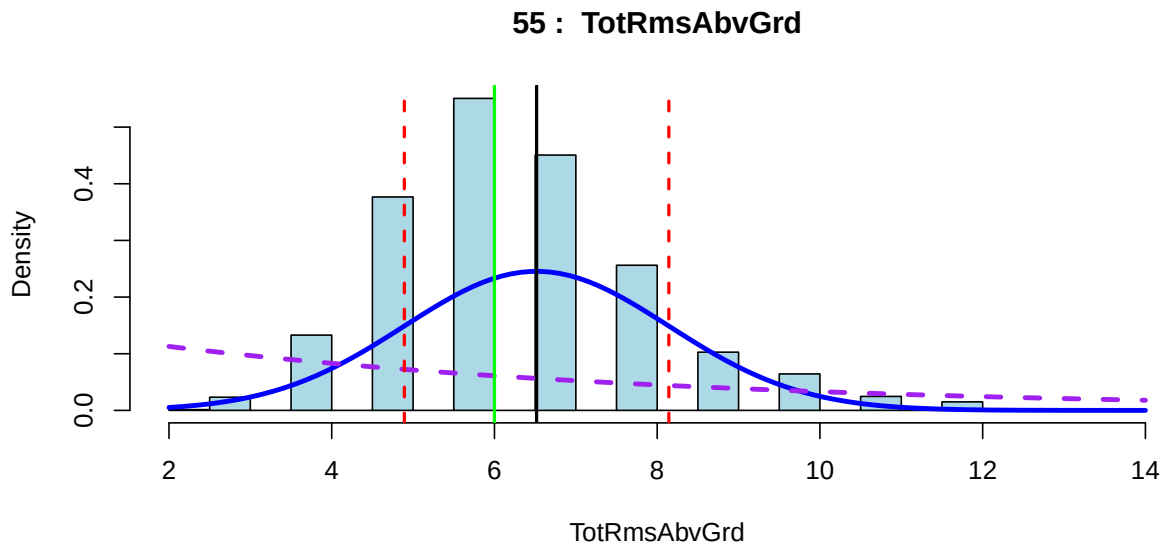


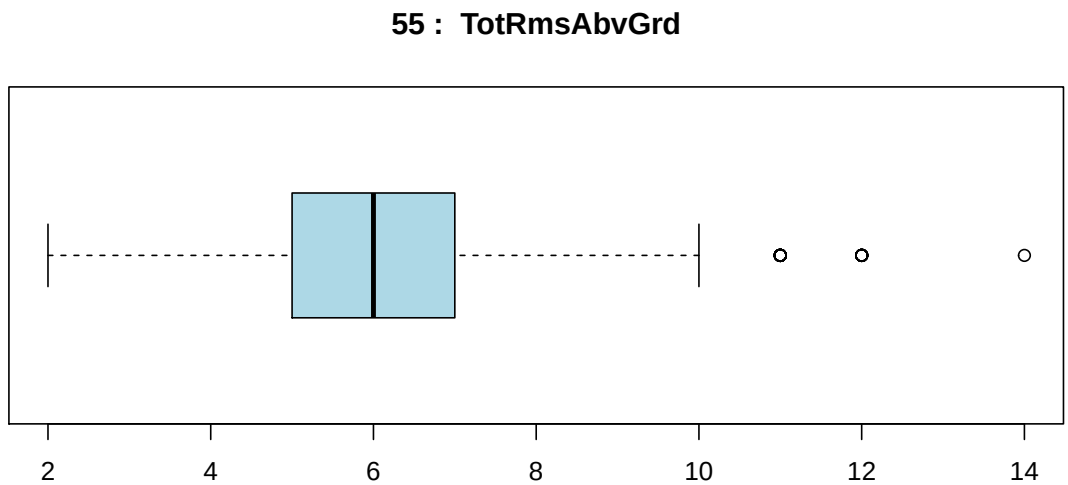



```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## 0.000000000 1.000000000 1.000000000 1.046575342 1.000000000 3.000000000
##       STDEV
## 0.220338198
```

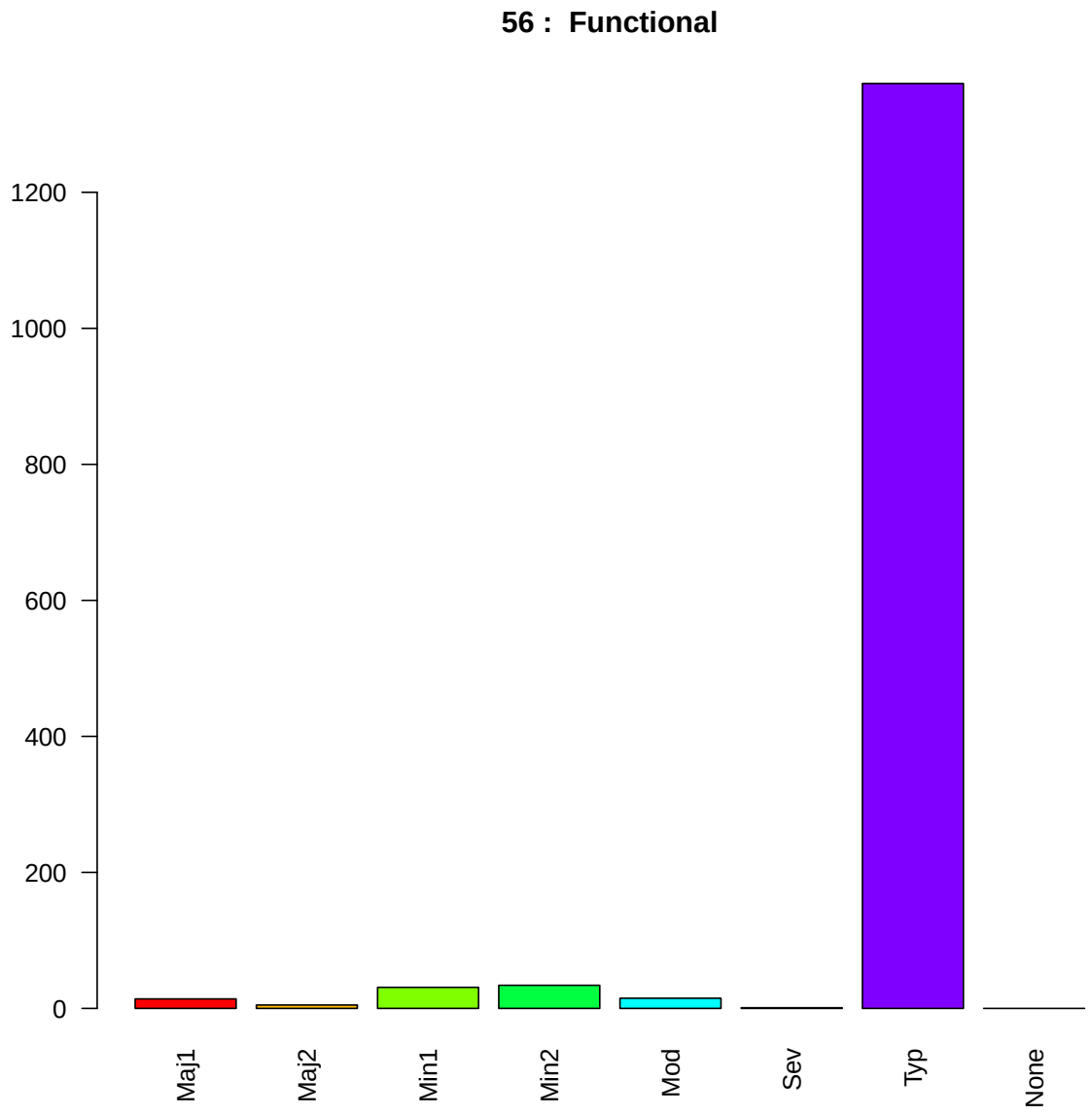



```
## KitchenQual
##   Ex   Fa   Gd   TA None
##  100   39  586  735    0
```

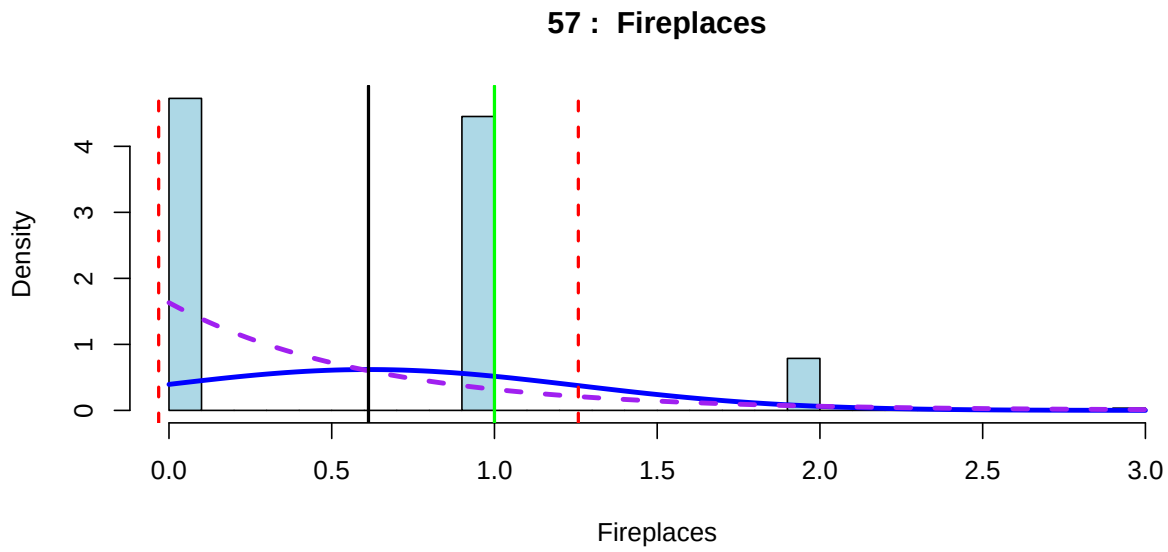


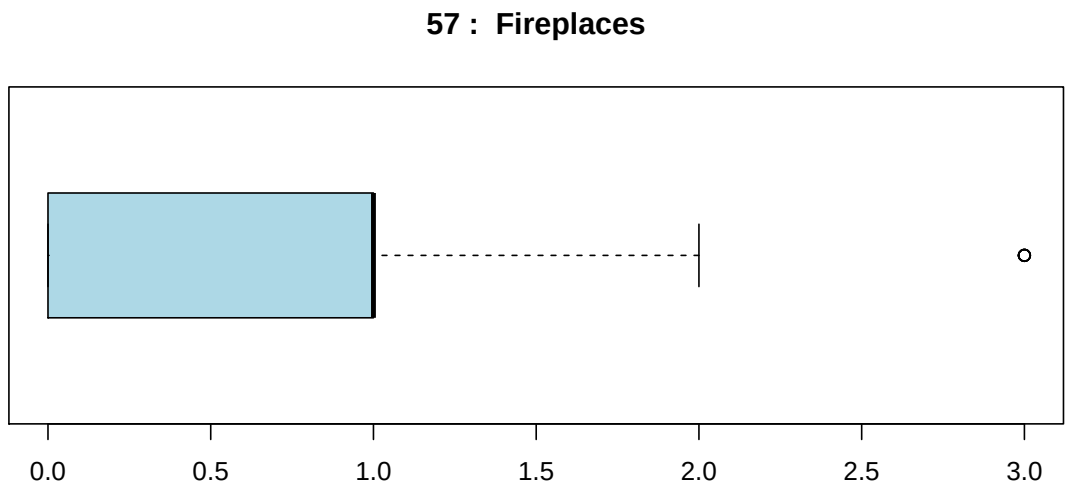



```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##  2.00000000  5.00000000  6.00000000  6.51780822  7.00000000 14.00000000
##       STDEV
##  1.62539329
```

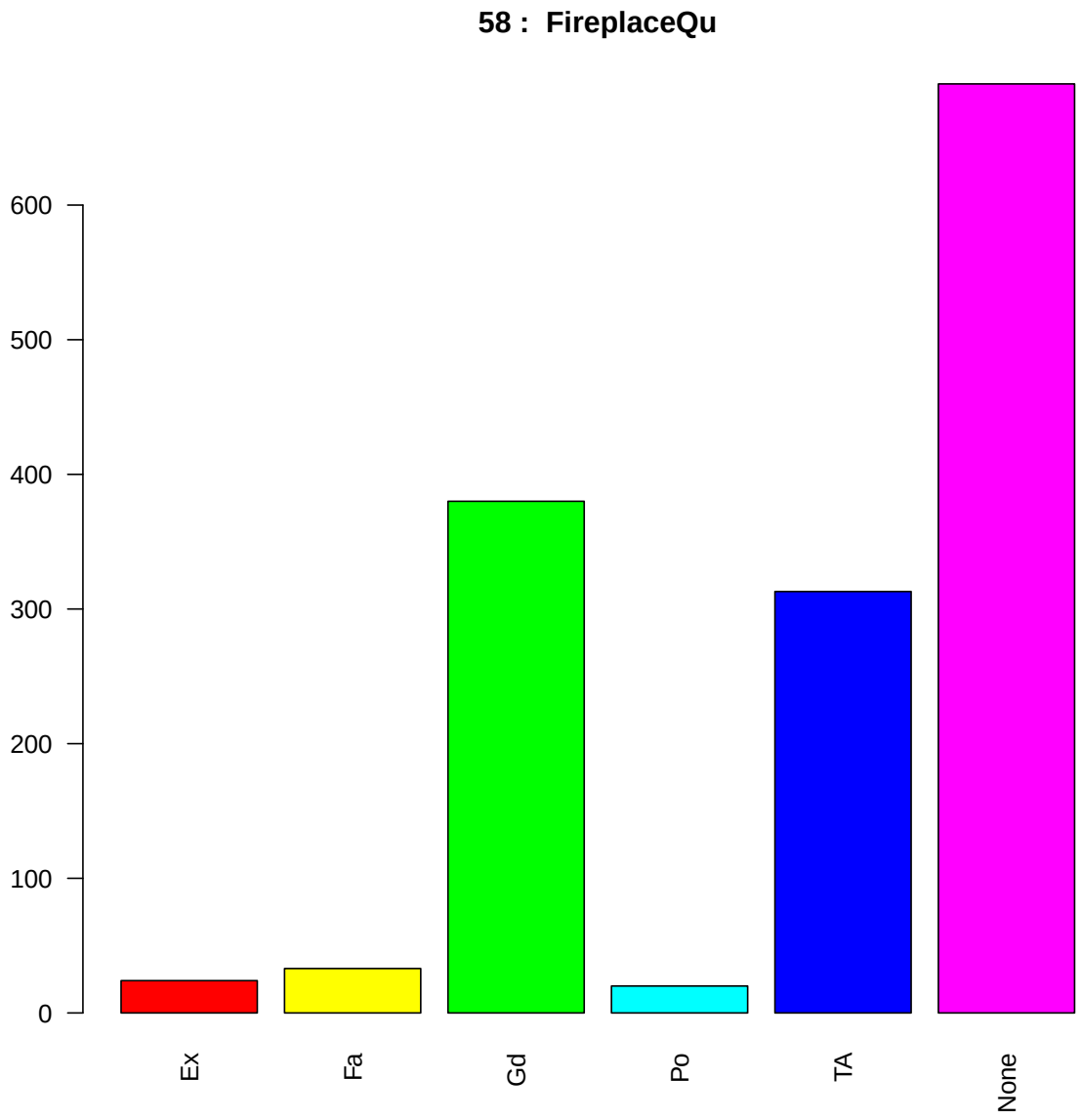



```
## Functional
## Maj1 Maj2 Min1 Min2  Mod  Sev  Typ None
##   14    5   31   34   15    1 1360    0
```

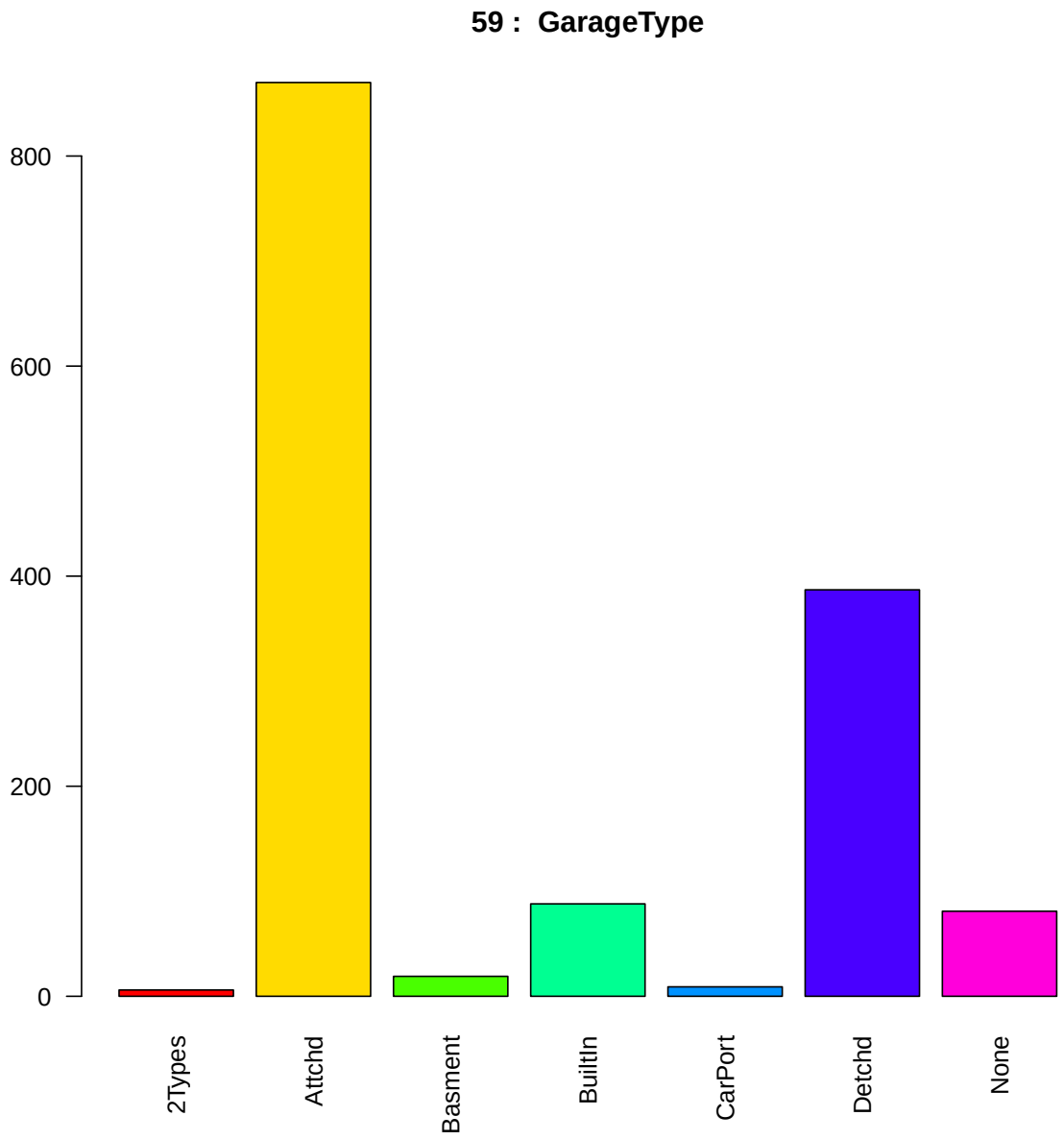




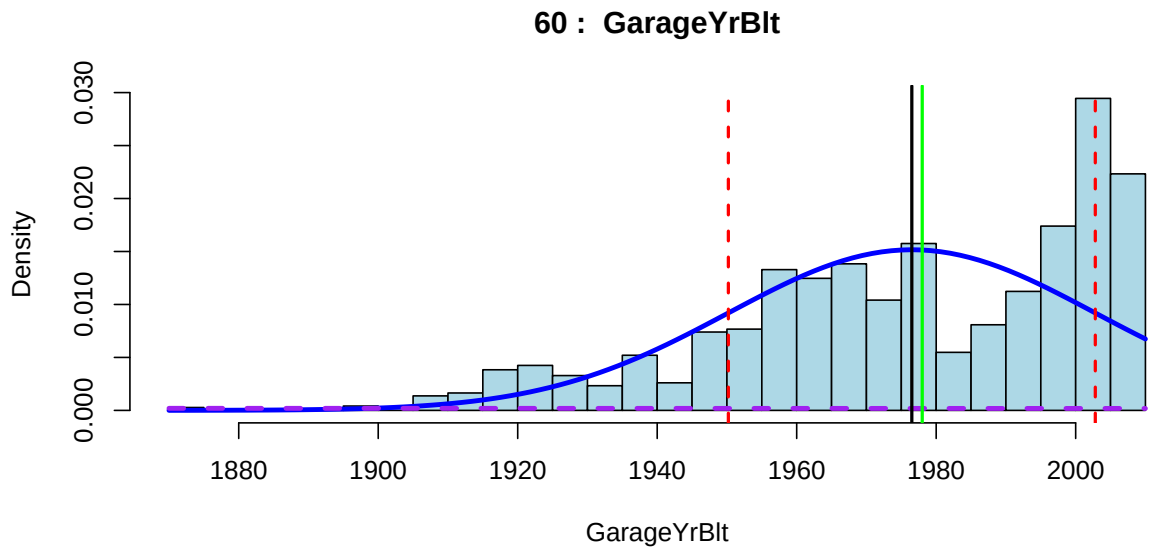

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
## 0.000000000 0.000000000 1.000000000 0.613013699 1.000000000 3.000000000
##       STDEV
## 0.644666386
```

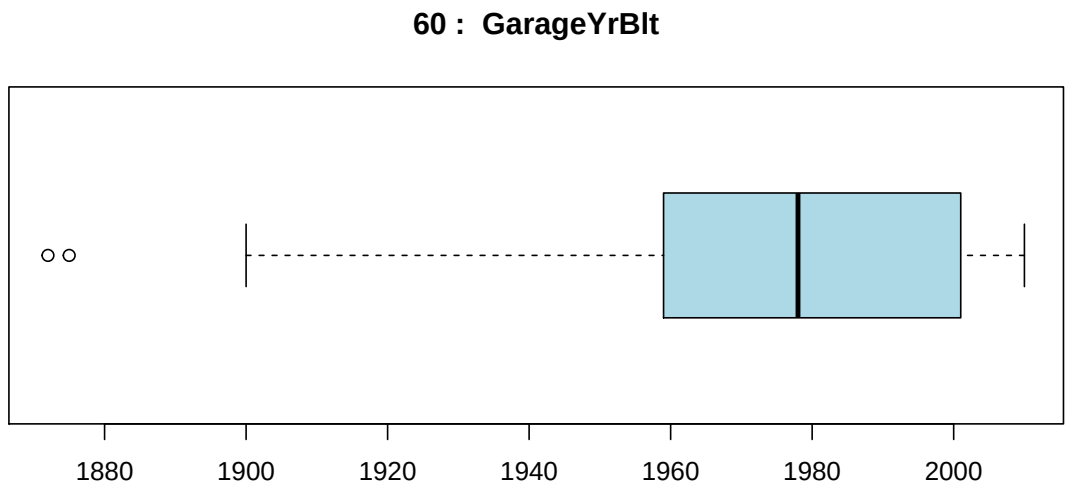



```
## FireplaceQu
##   Ex   Fa   Gd   Po   TA None
##   24   33  380   20  313  690
```

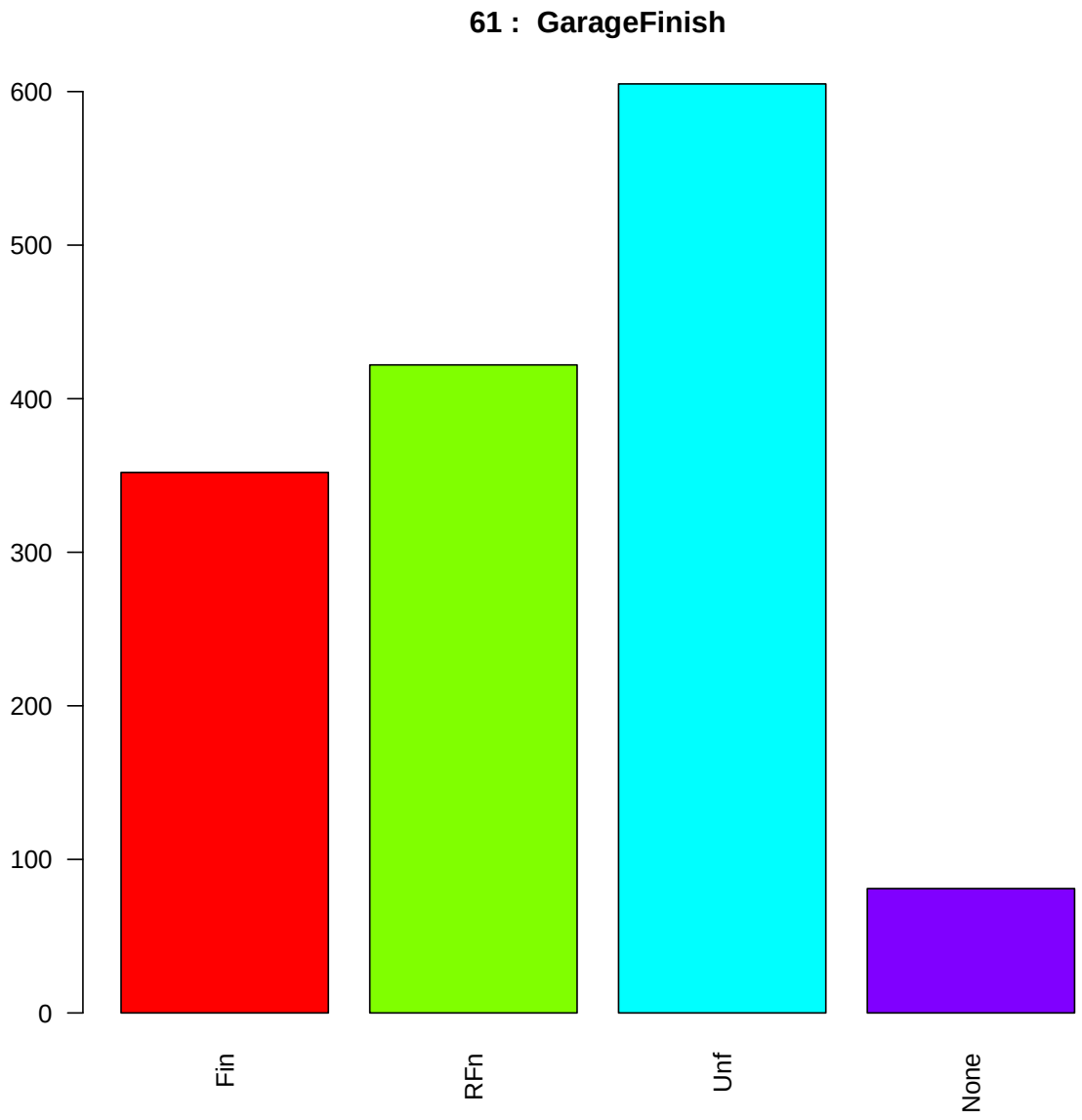



```
## GarageType
##  2Types  Attchd Basment BuiltIn CarPort  Detchd    None
##       6     870      19      88       9     387      81
```

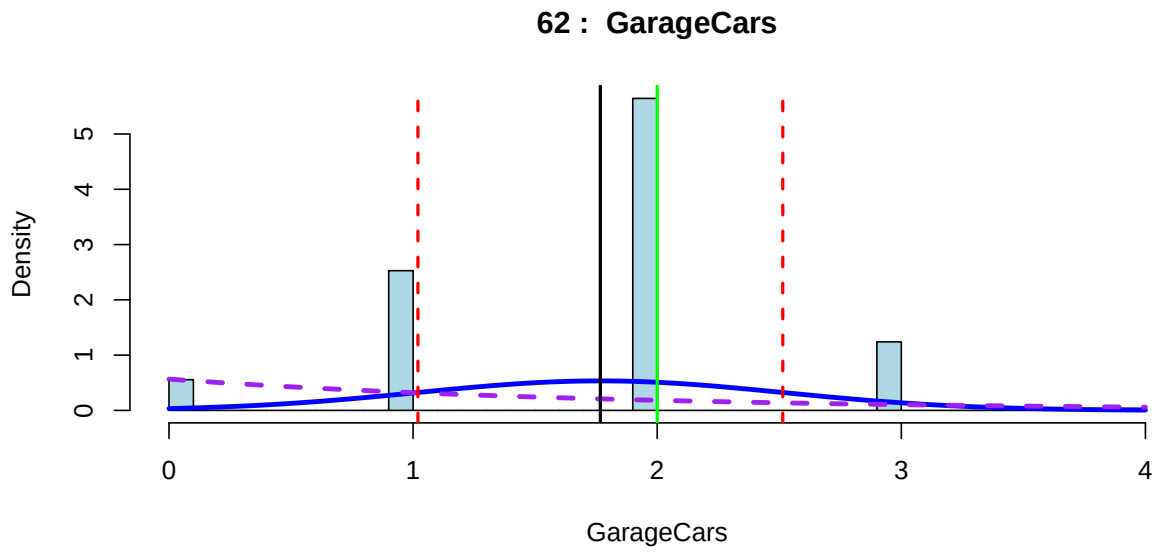


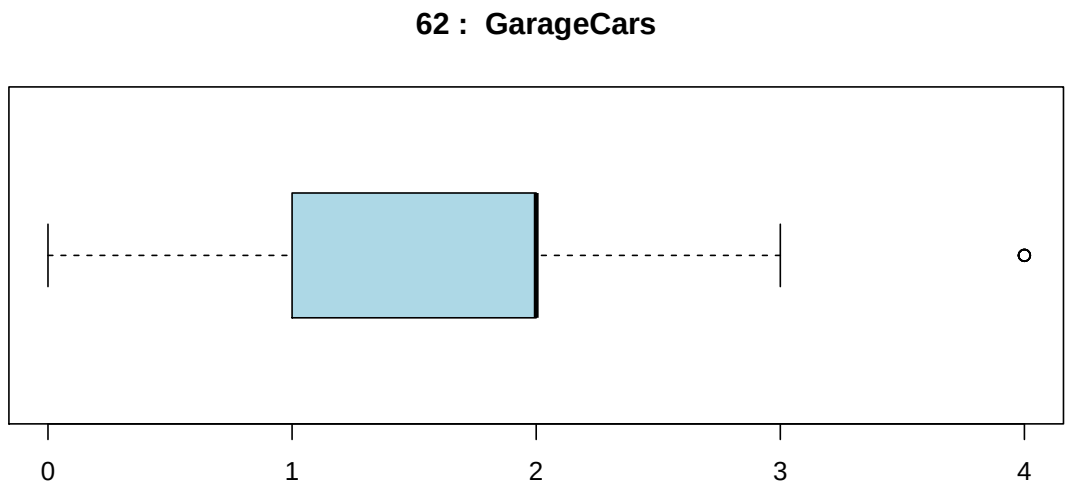



```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## 1872.0000000 1959.0000000 1978.0000000 1976.5075342 2001.0000000 2010.0000000
##        STDEV
##   26.3067386
```

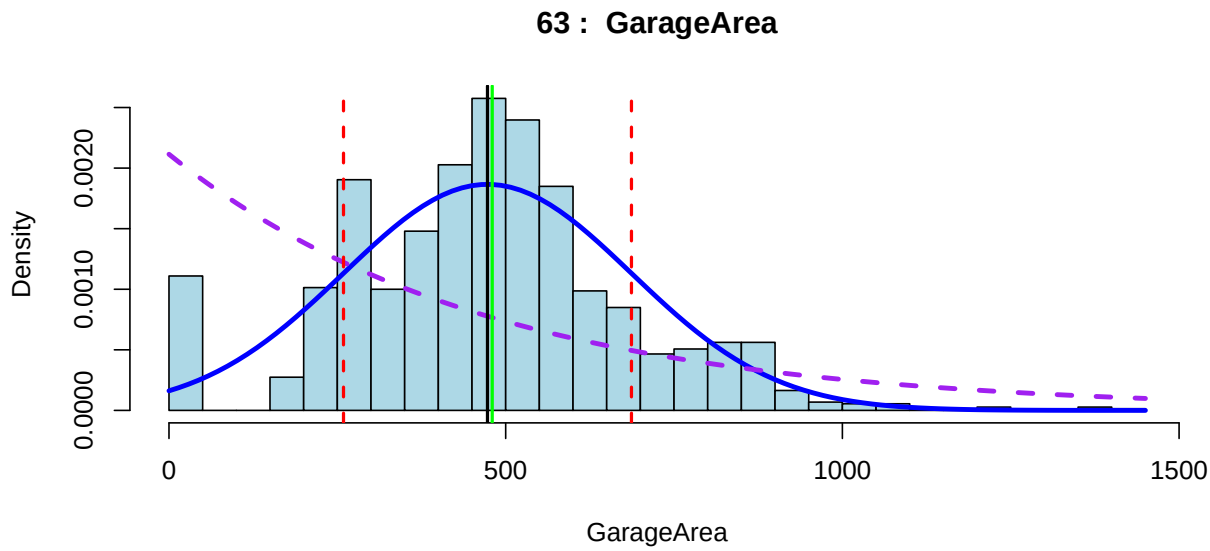


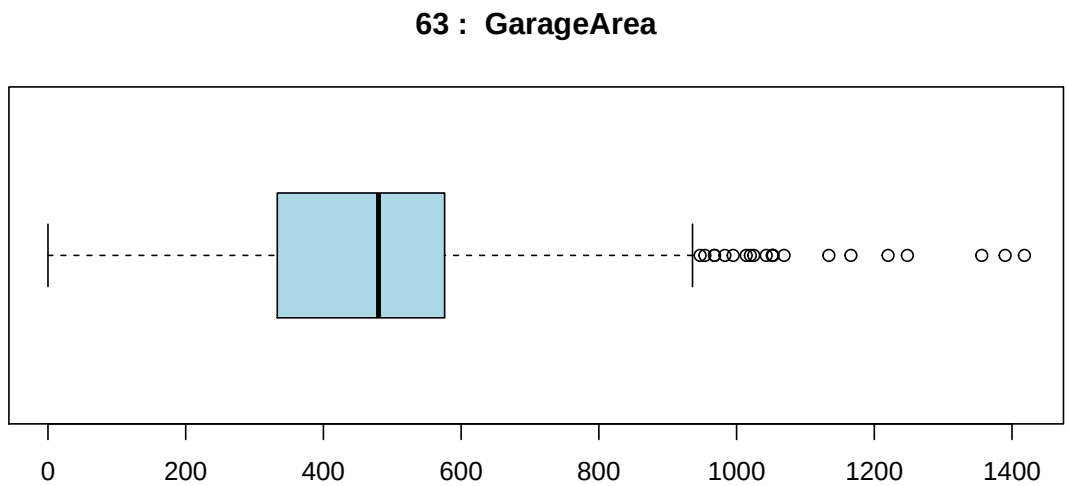

```
## GarageFinish
##  Fin  RFn  Unf None
##  352  422  605   81
```



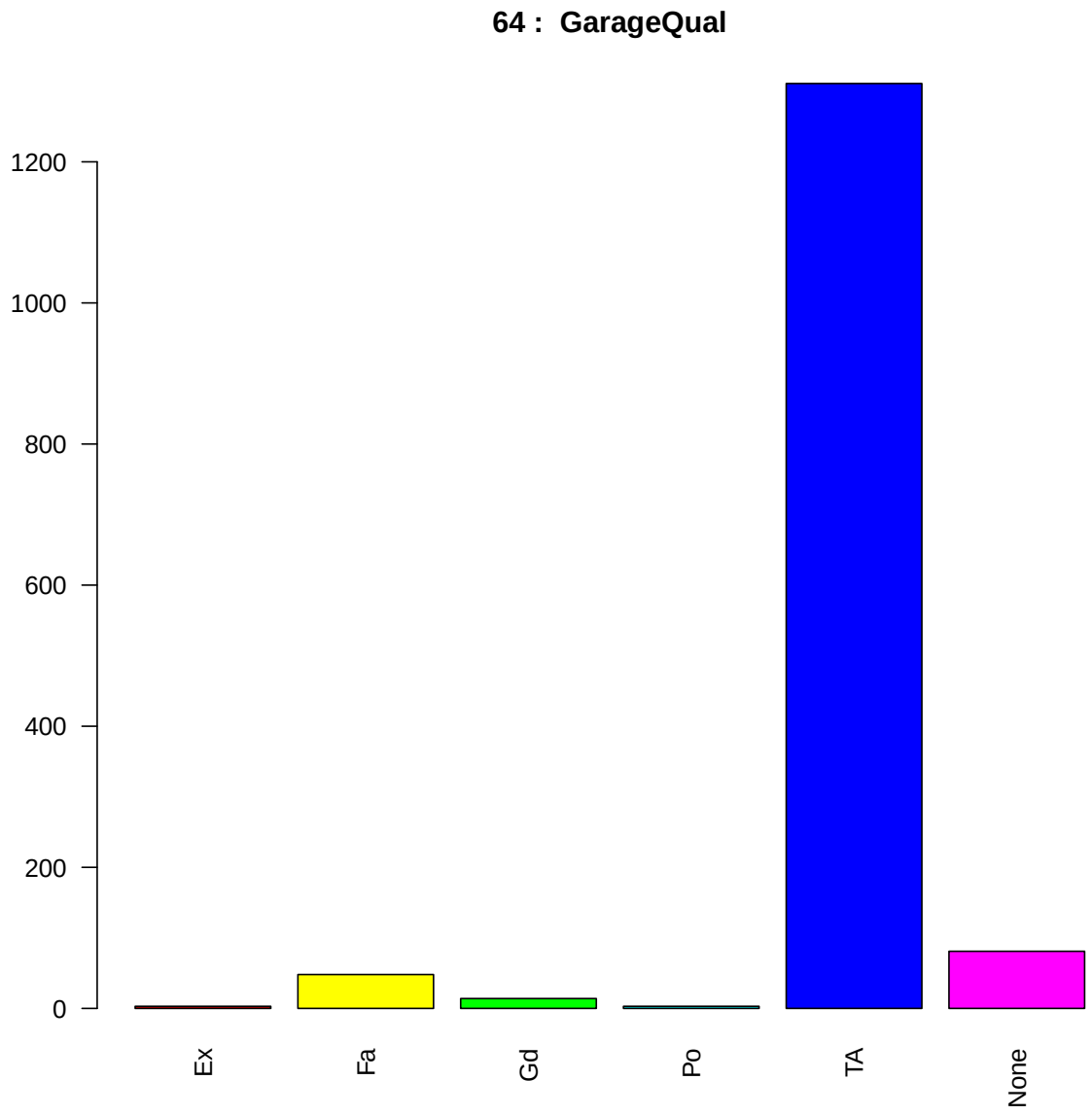



```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.      STDEV
## 0.00000000 1.00000000 2.00000000 1.76712329 2.00000000 4.00000000 0.74731501
```

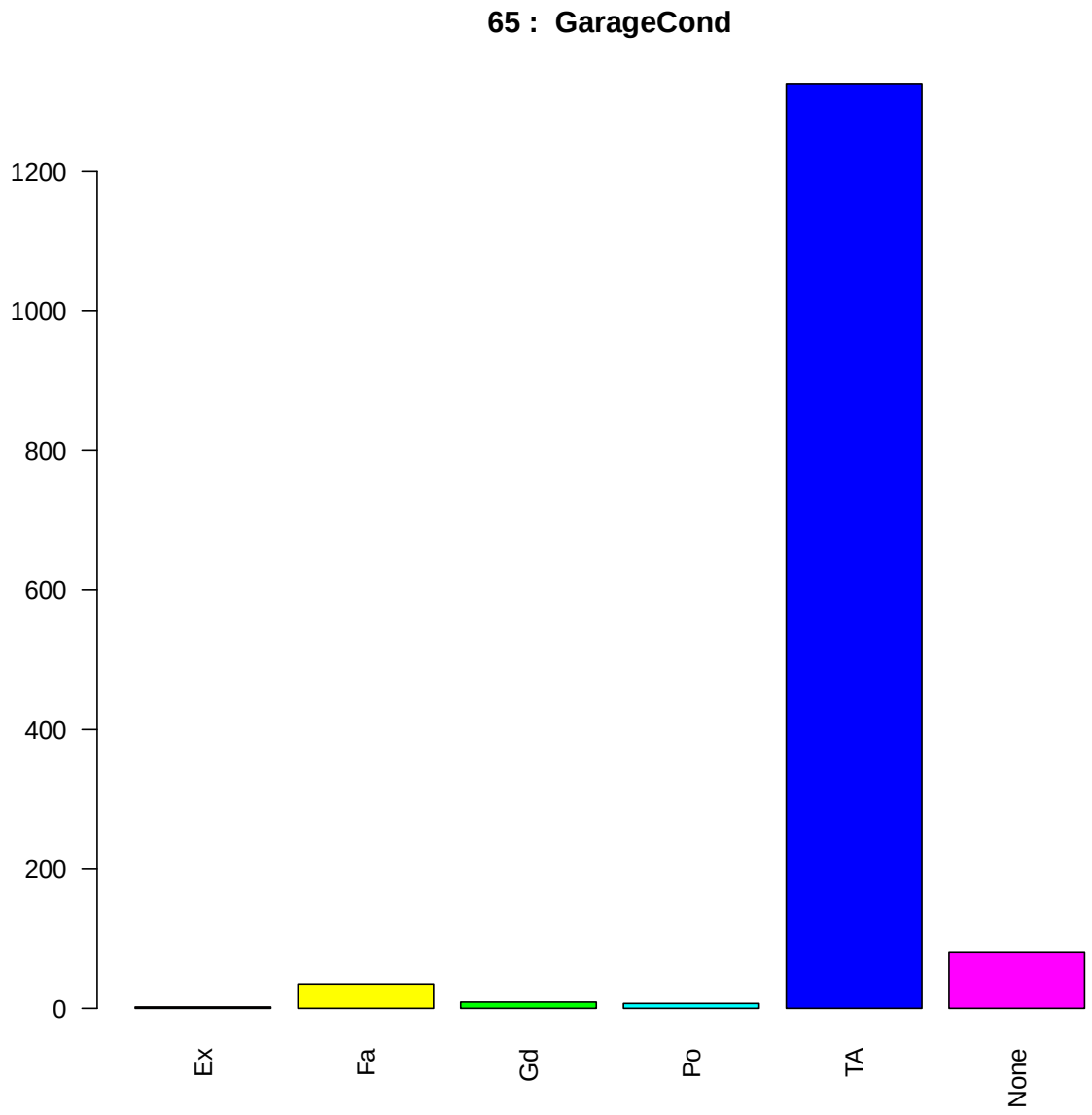




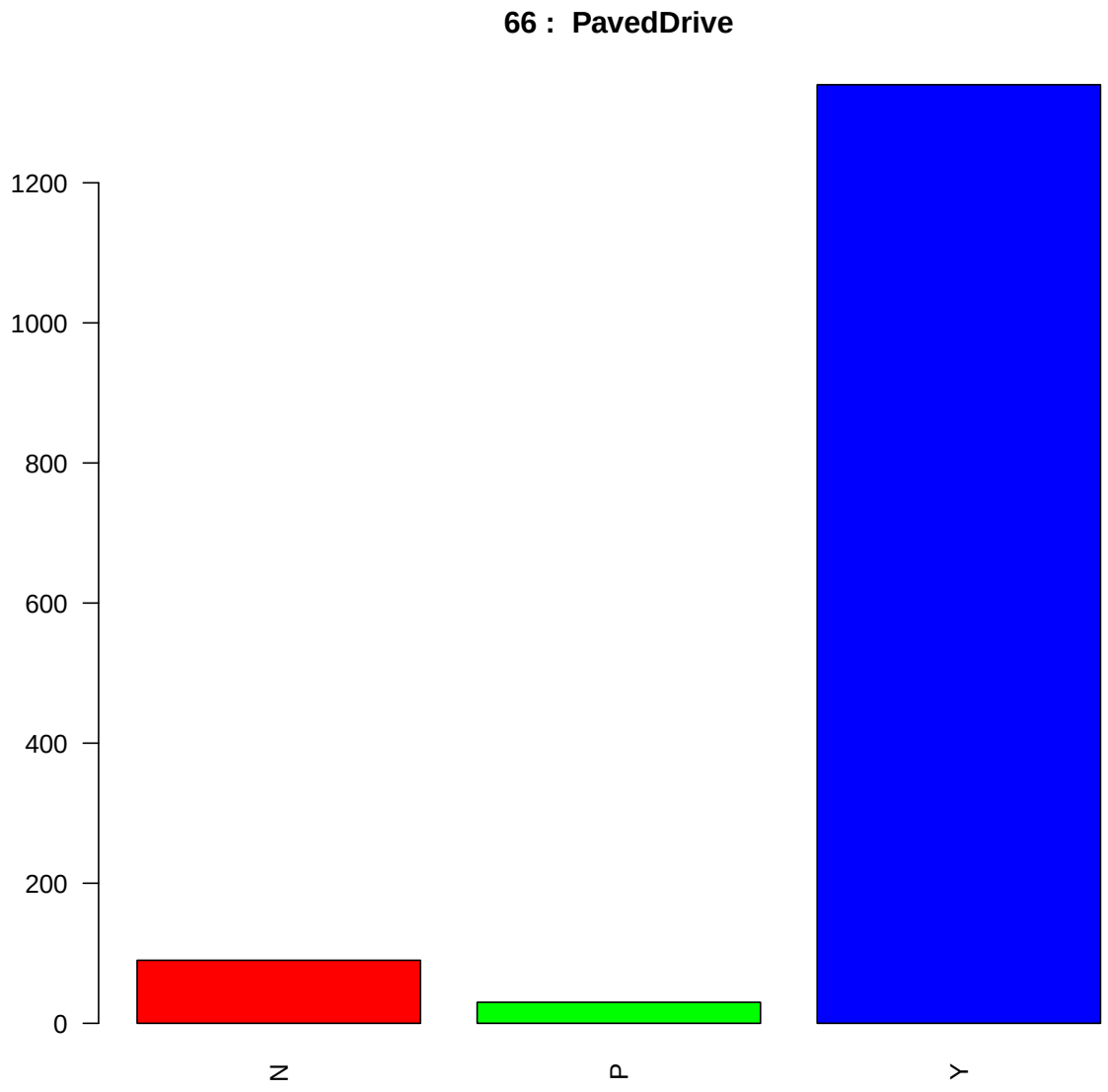

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000  334.500000  480.000000  472.980137  576.000000 1418.000000
##       STDEV
##  213.804841
```



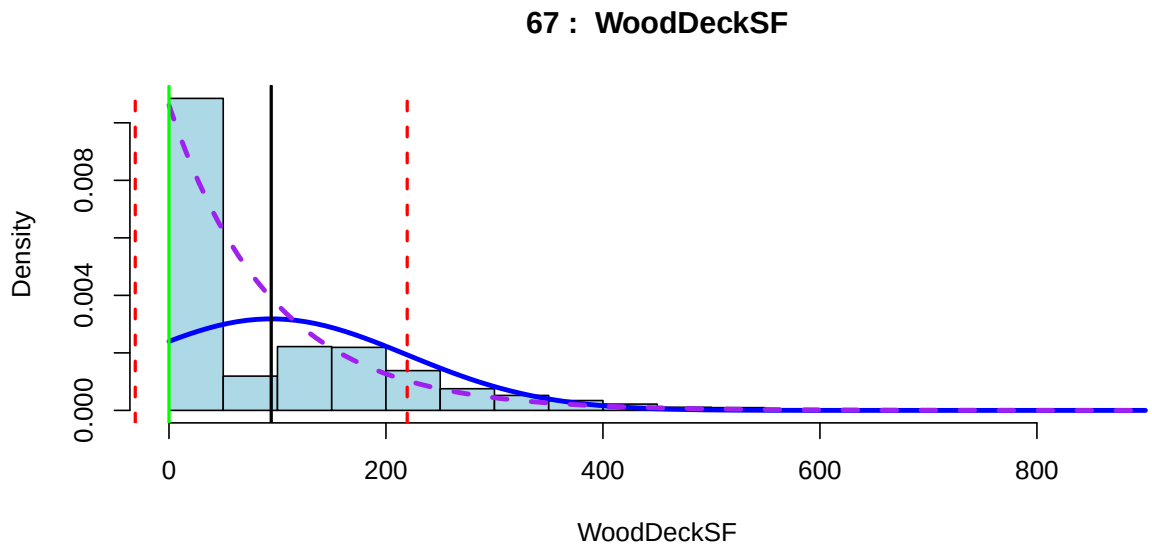

```
## GarageQual
##   Ex   Fa   Gd   Po   TA None
##    3   48   14    3 1311   81
```

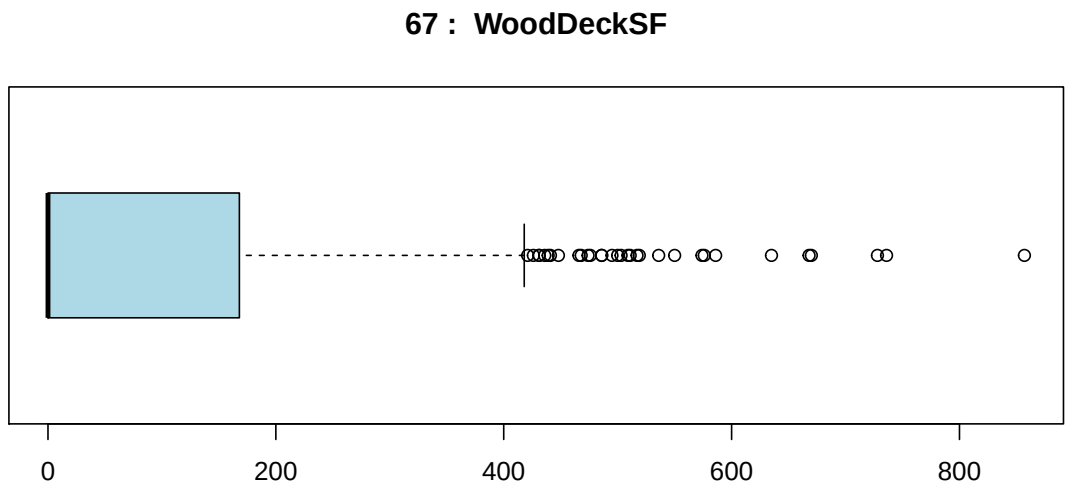



```
## GarageCond
##   Ex   Fa   Gd   Po   TA None
##    2   35    9    7 1326   81
```

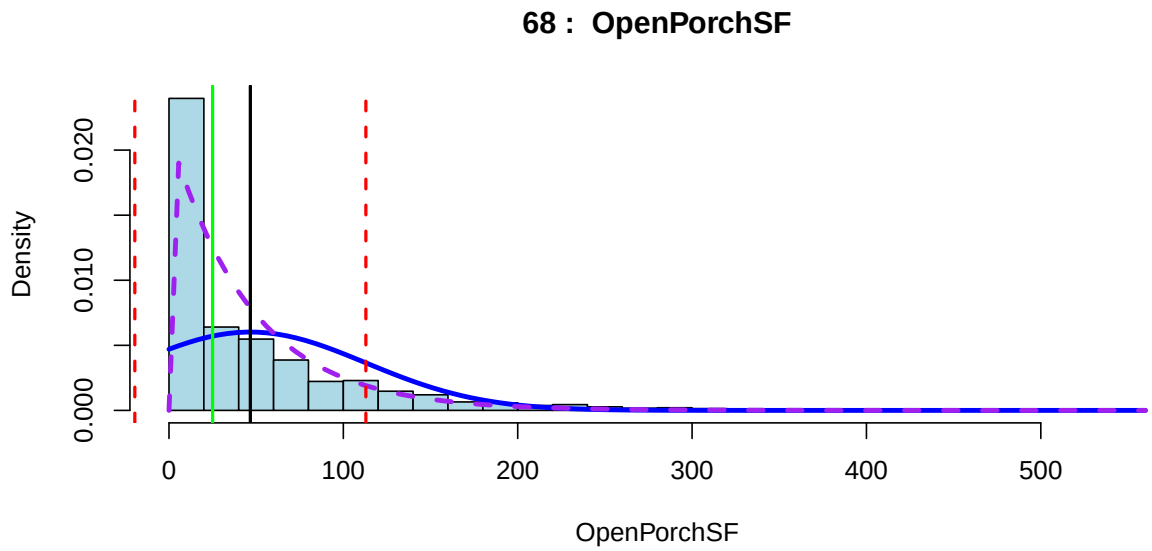


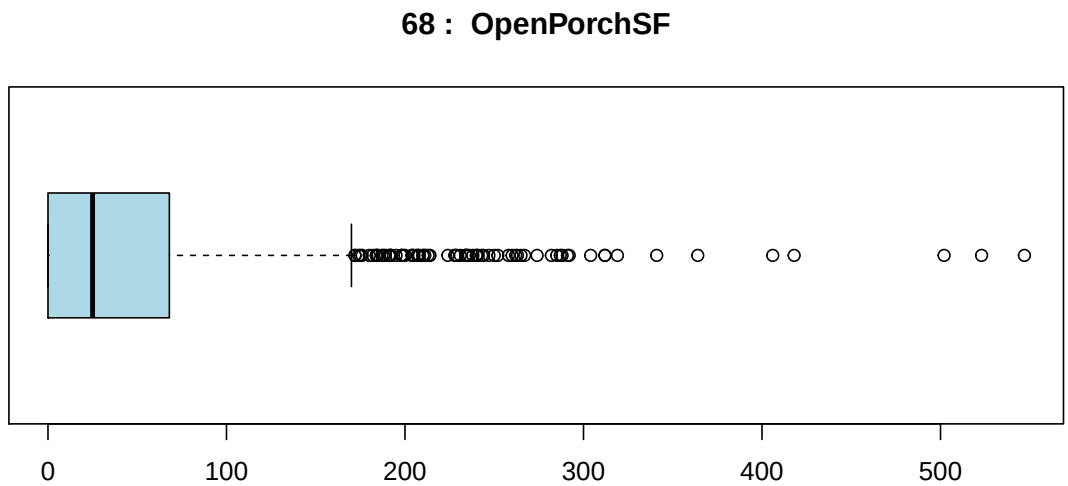

```
## PavedDrive
##    N    P    Y
##   90   30 1340
```



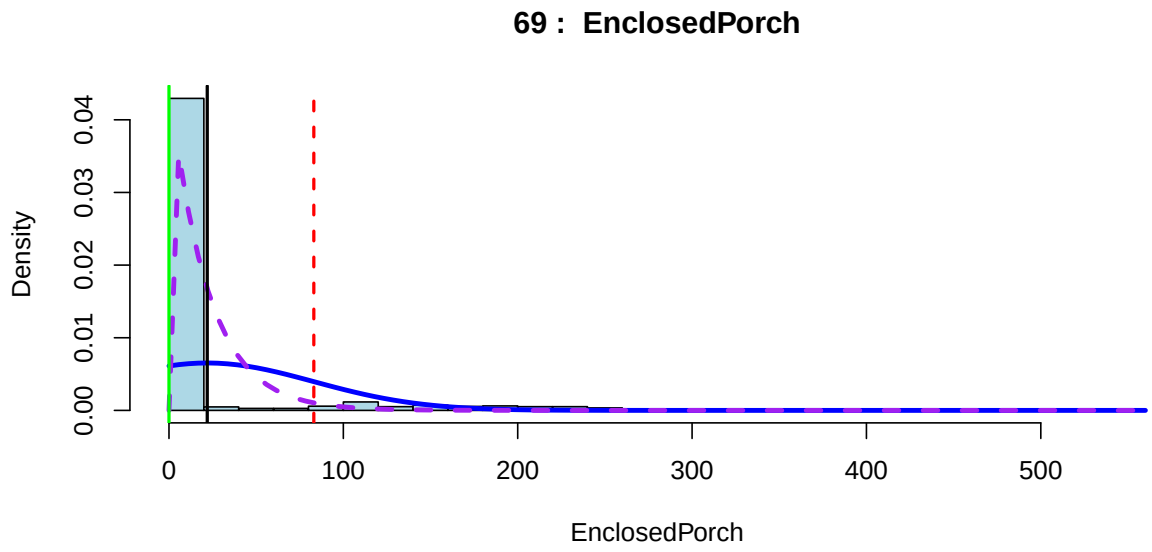


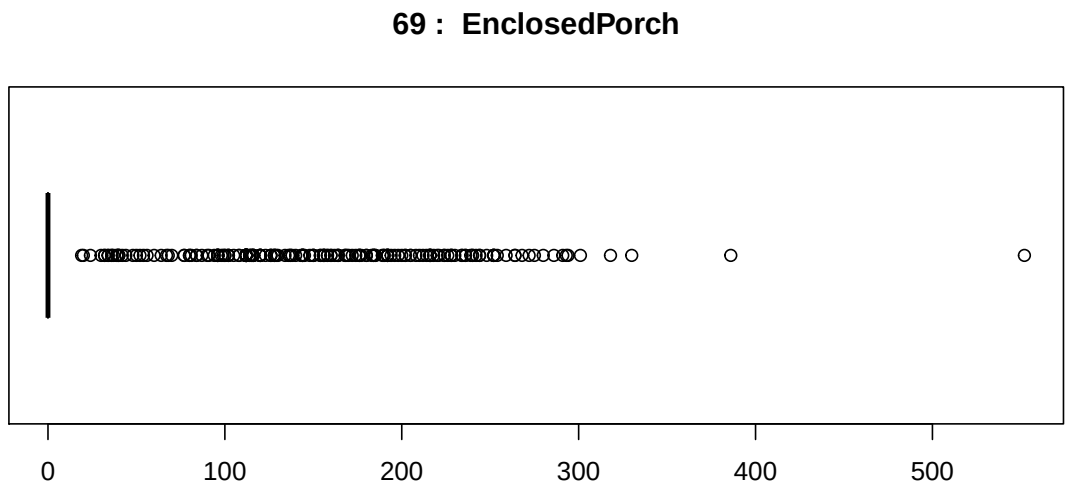

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##   0.0000000   0.0000000   0.0000000  94.2445205 168.0000000 857.0000000
##       STDEV
## 125.3387944
```



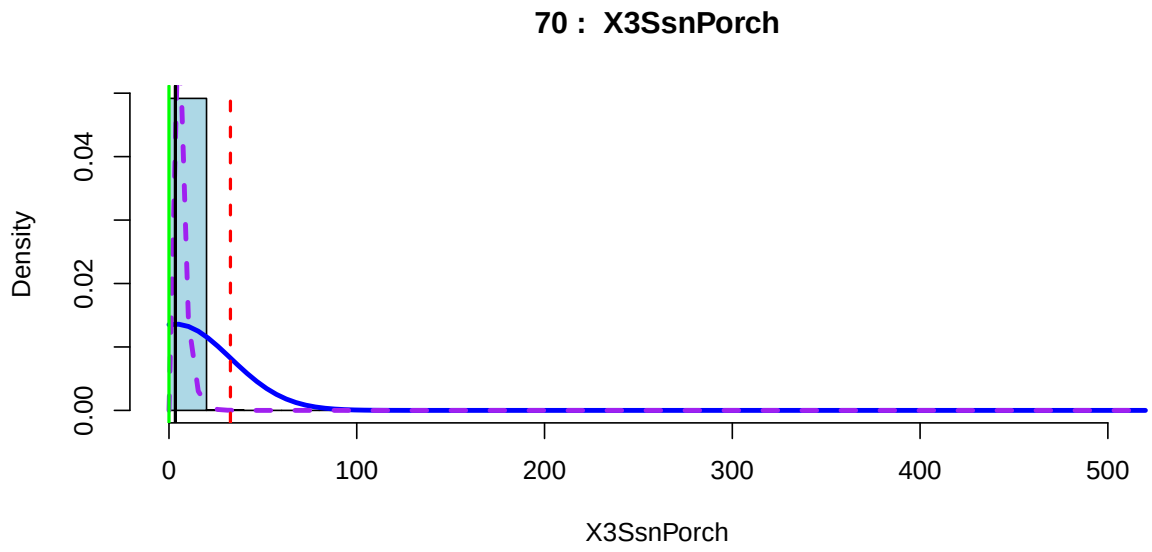


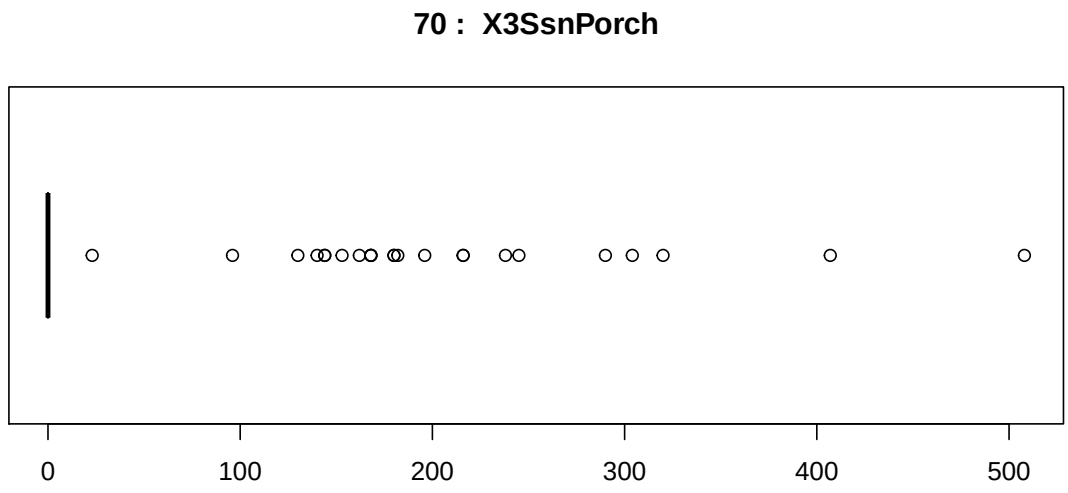

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##   0.0000000   0.0000000  25.0000000  46.6602740  68.0000000 547.0000000
##       STDEV
##  66.2560277
```



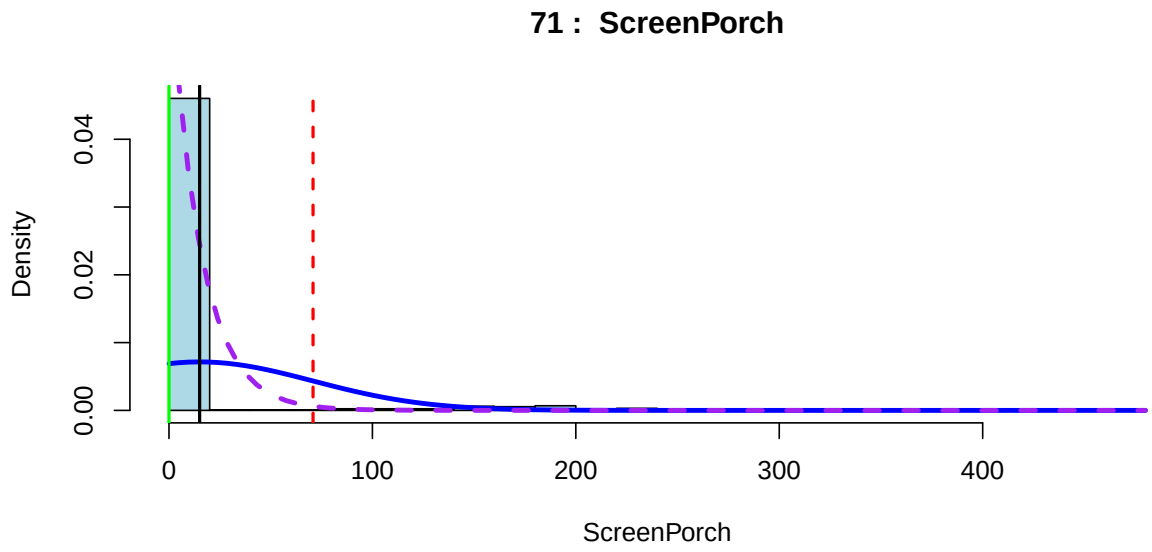


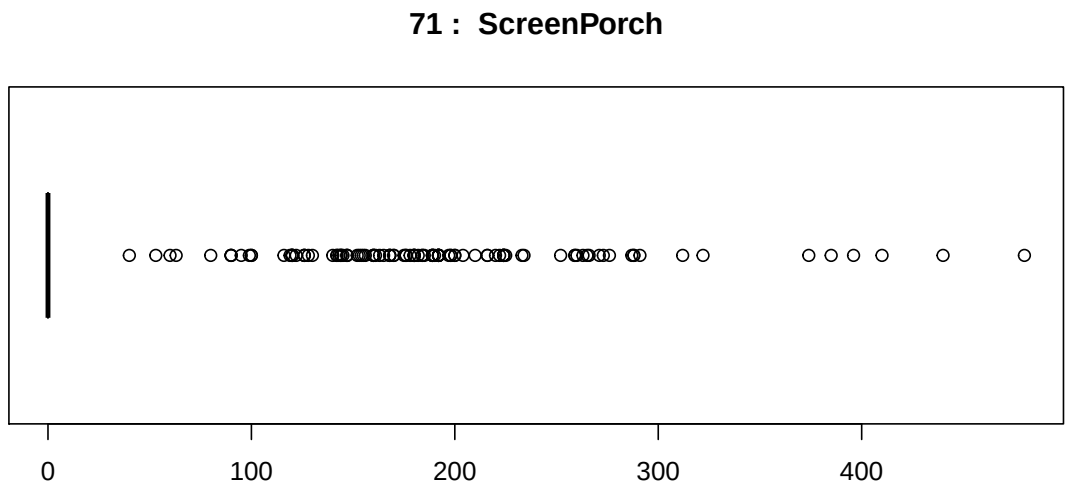

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##   0.0000000   0.0000000   0.0000000  21.9541096   0.0000000 552.0000000
##       STDEV
##  61.1191486
```



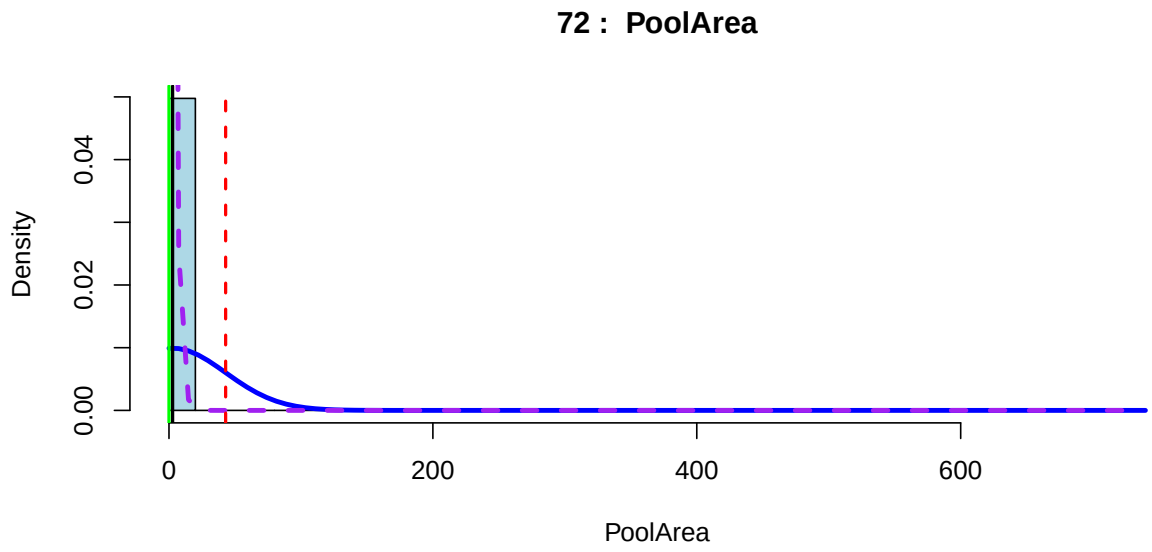


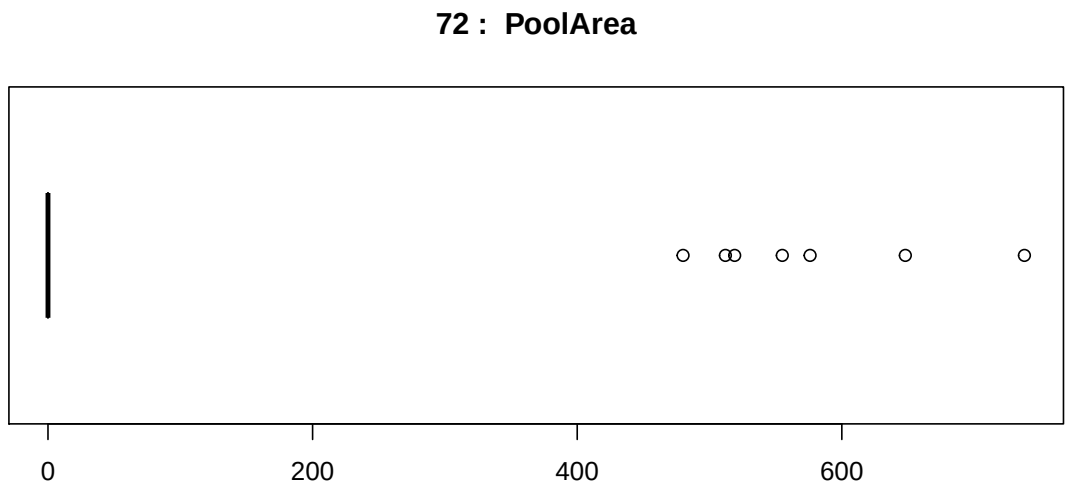

```
##          Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##   0.00000000   0.00000000   0.00000000   3.40958904   0.00000000 508.00000000
##         STDEV
##   29.31733056
```



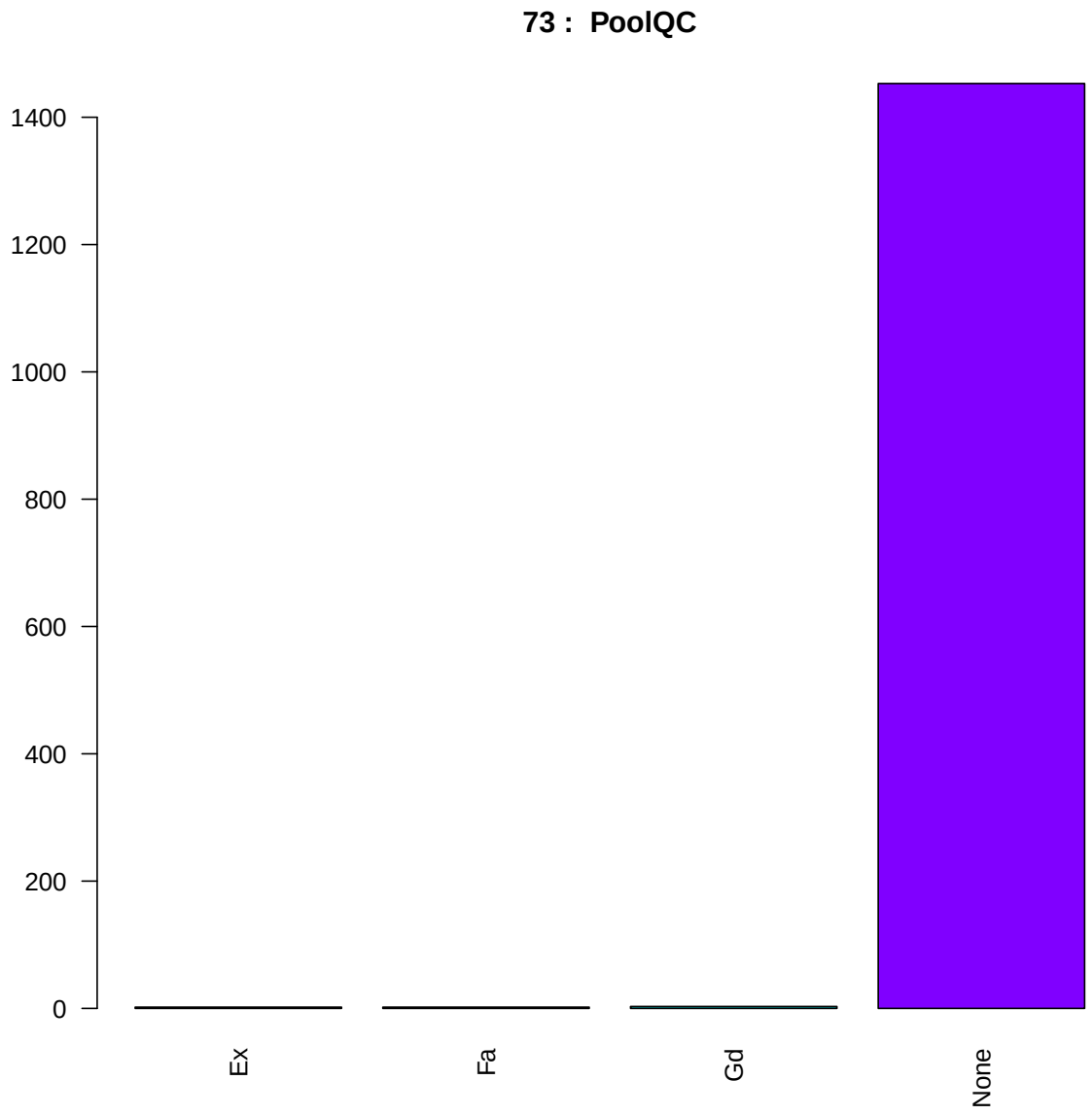



```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##   0.0000000   0.0000000   0.0000000  15.0609589   0.0000000 480.0000000
##       STDEV
##  55.7574153
```

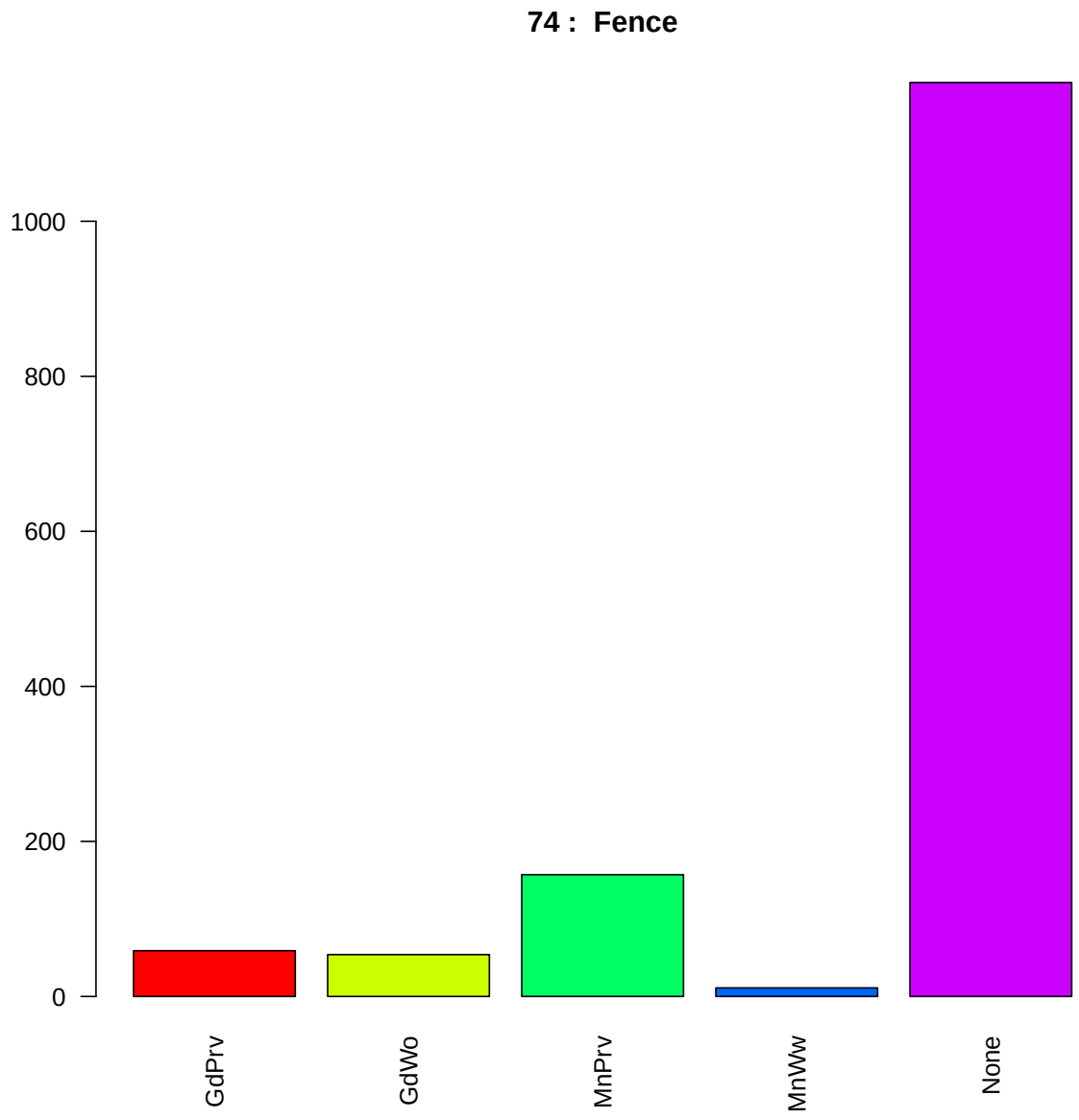




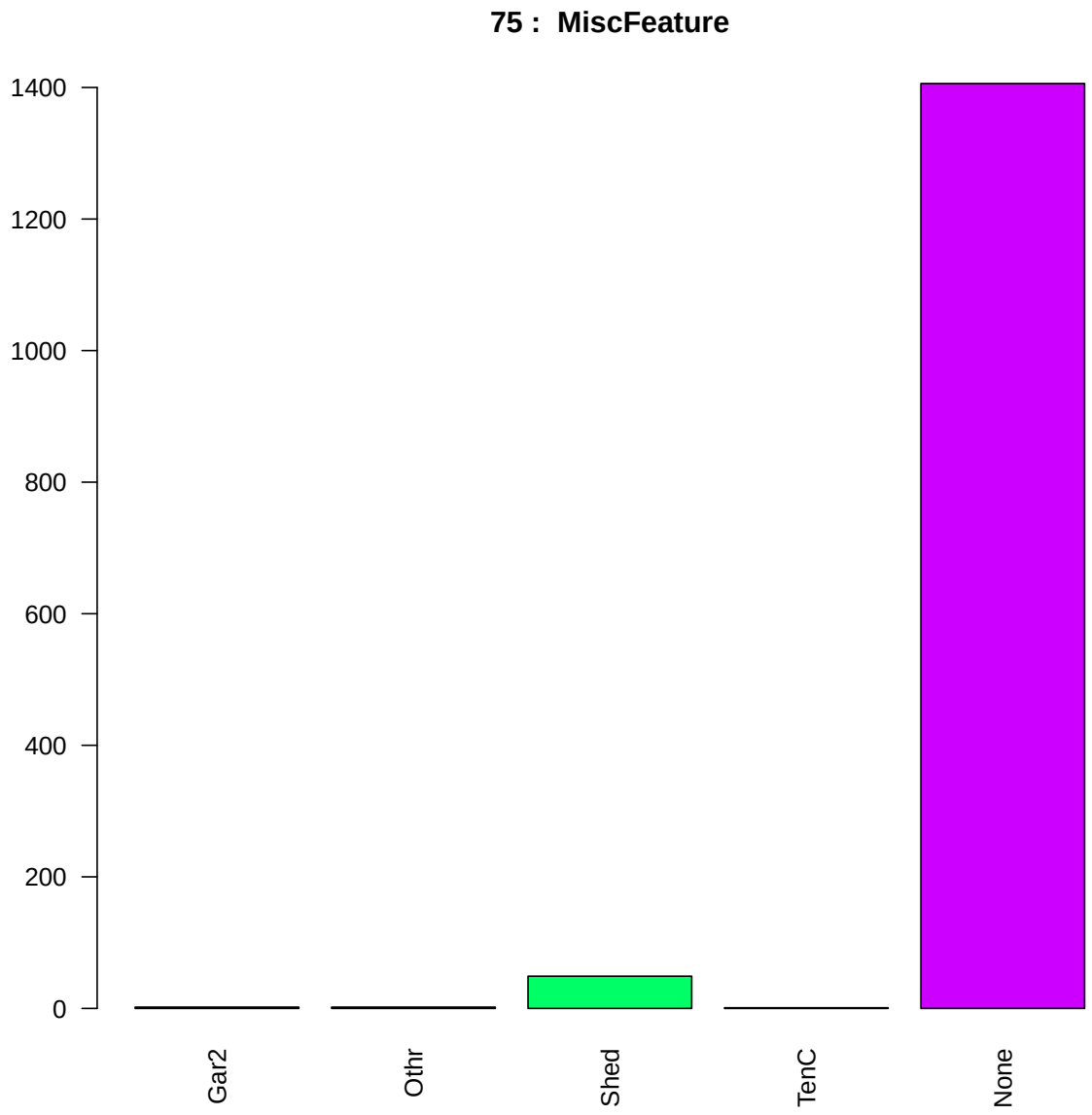

```
##          Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
##   0.00000000   0.00000000   0.00000000   2.75890411   0.00000000 738.00000000
##         STDEV
##  40.17730694
```



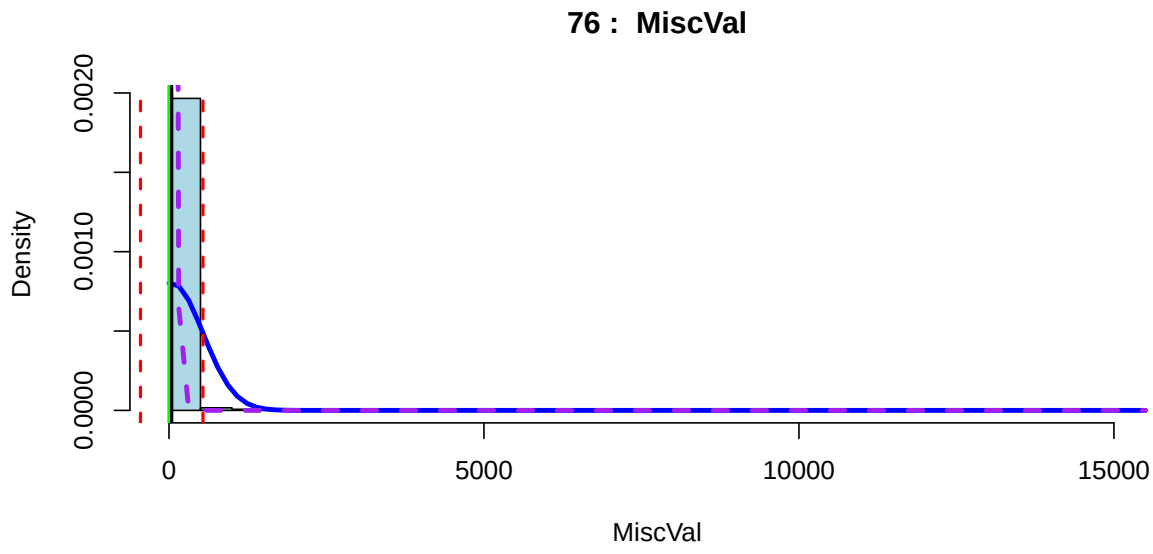

```
## PoolQC
##   Ex   Fa   Gd None
##    2    2    3 1453
```

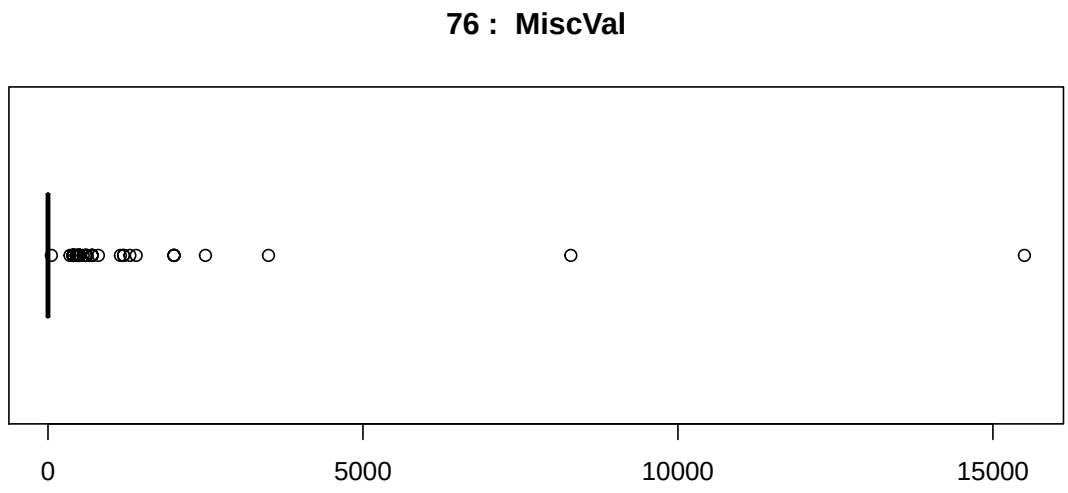



```
## Fence
## GdPrv  GdWo MnPrv  MnWw  None
##    59    54   157    11  1179
```

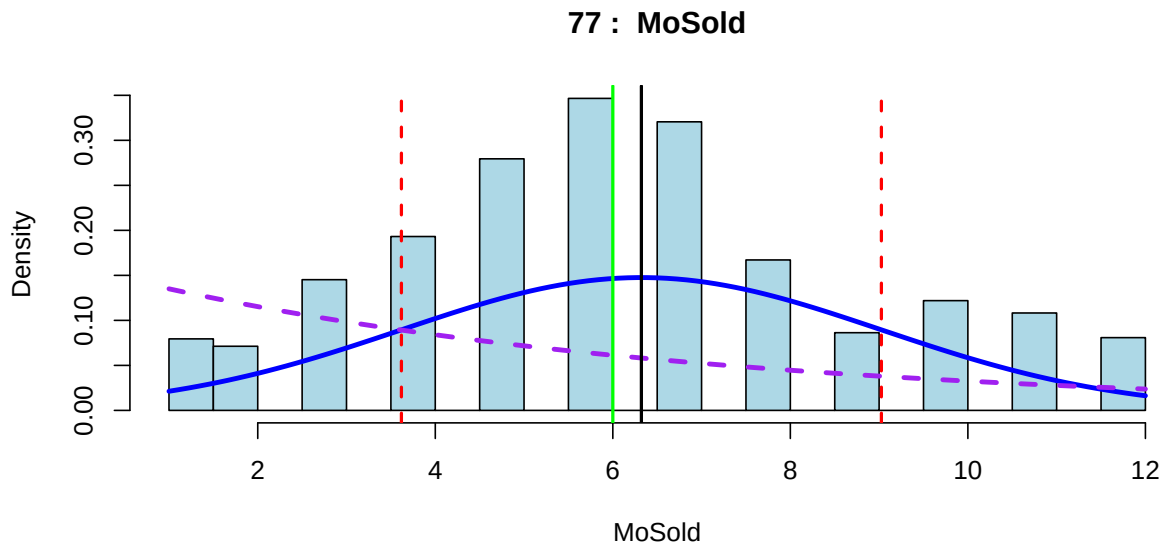


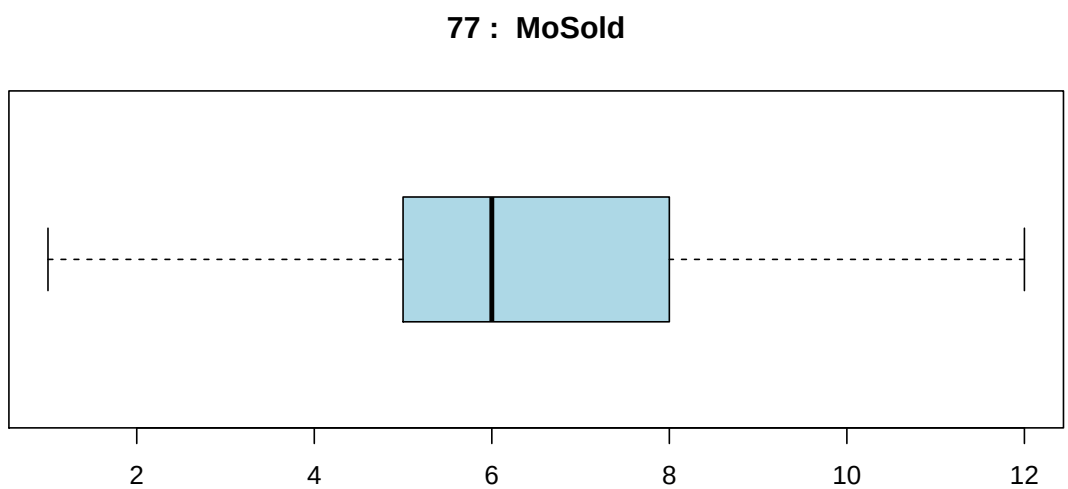

```
## MiscFeature
## Gar2 Othr Shed TenC None
##    2    2   49    1 1406
```



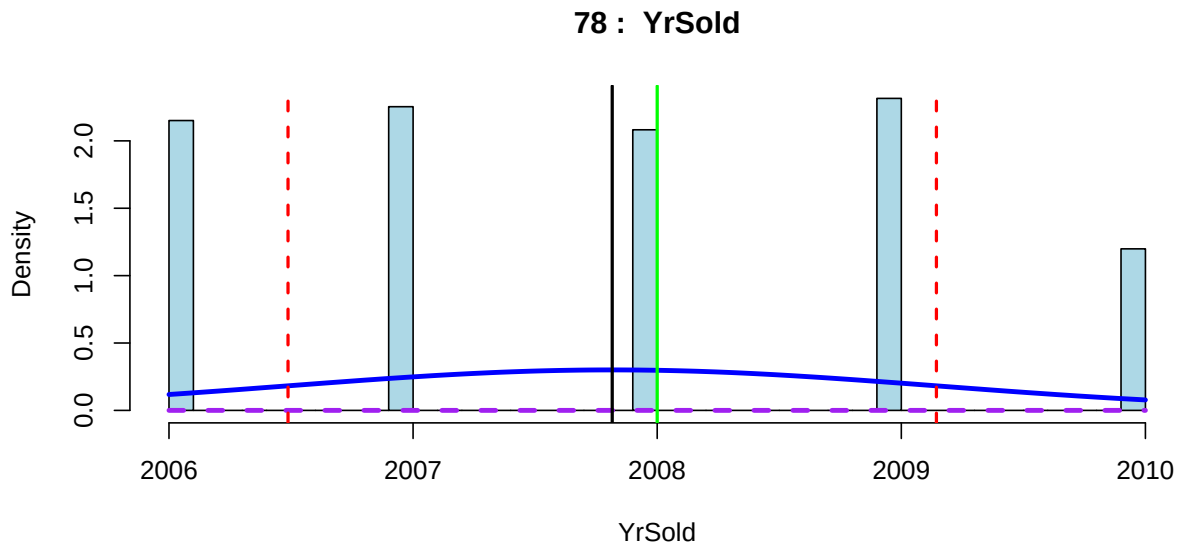


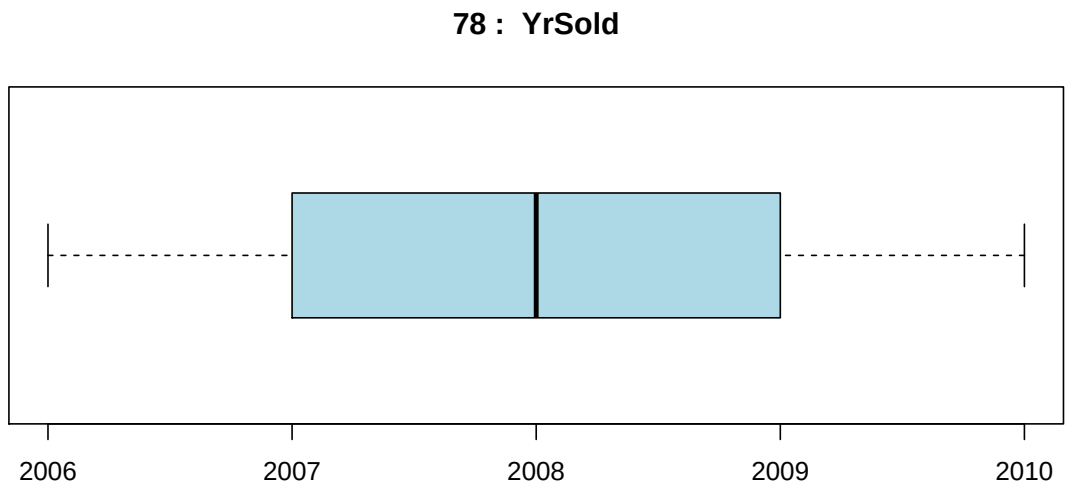

```
##          Min.       1st Qu.        Median          Mean       3rd Qu.
##     0.0000000     0.0000000     0.0000000    43.4890411     0.0000000
##          Max.         STDEV
## 15500.0000000   496.1230245
```



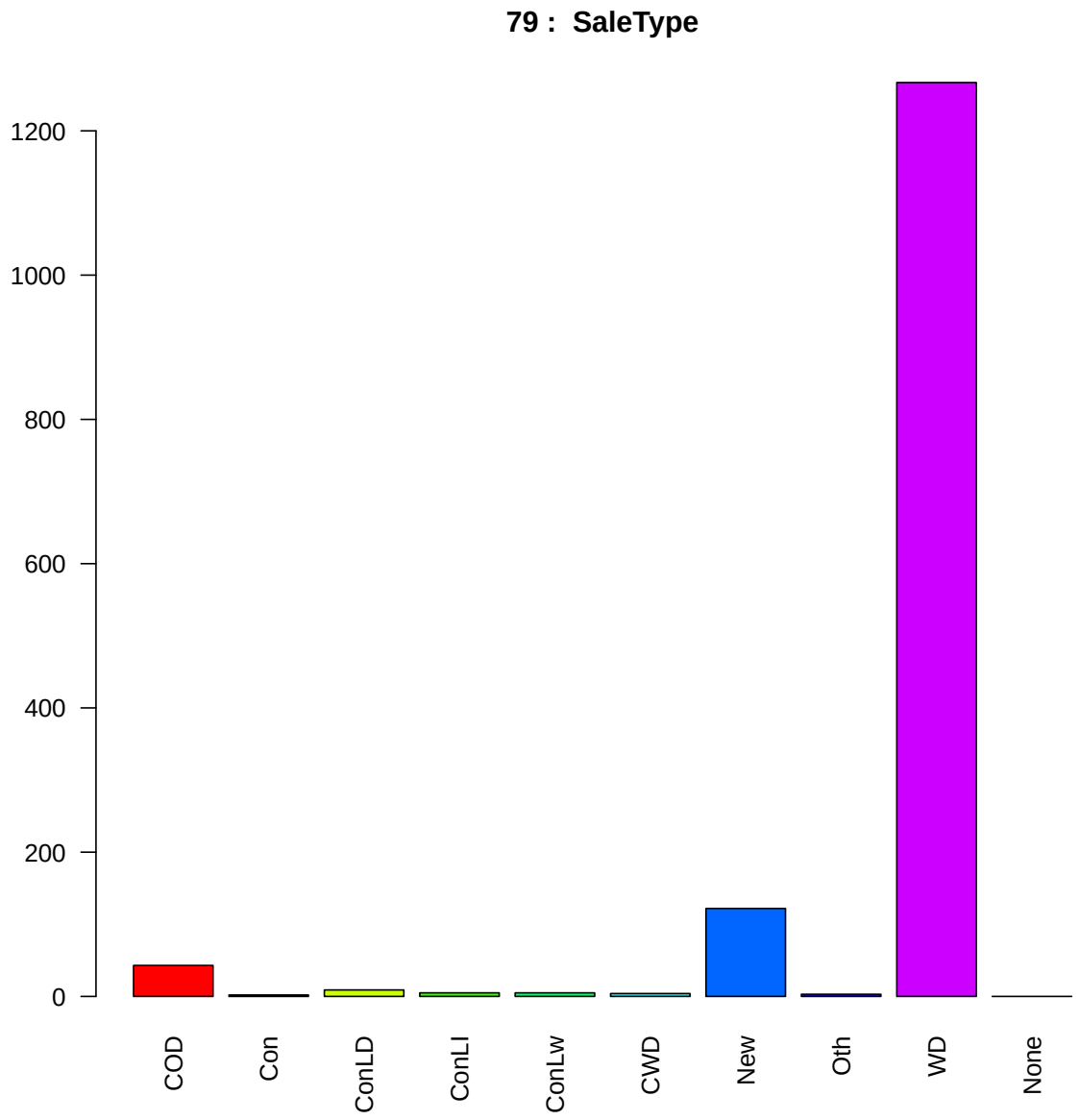



```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##  1.00000000  5.00000000  6.00000000  6.32191781  8.00000000 12.00000000
##       STDEV
##  2.70362621
```

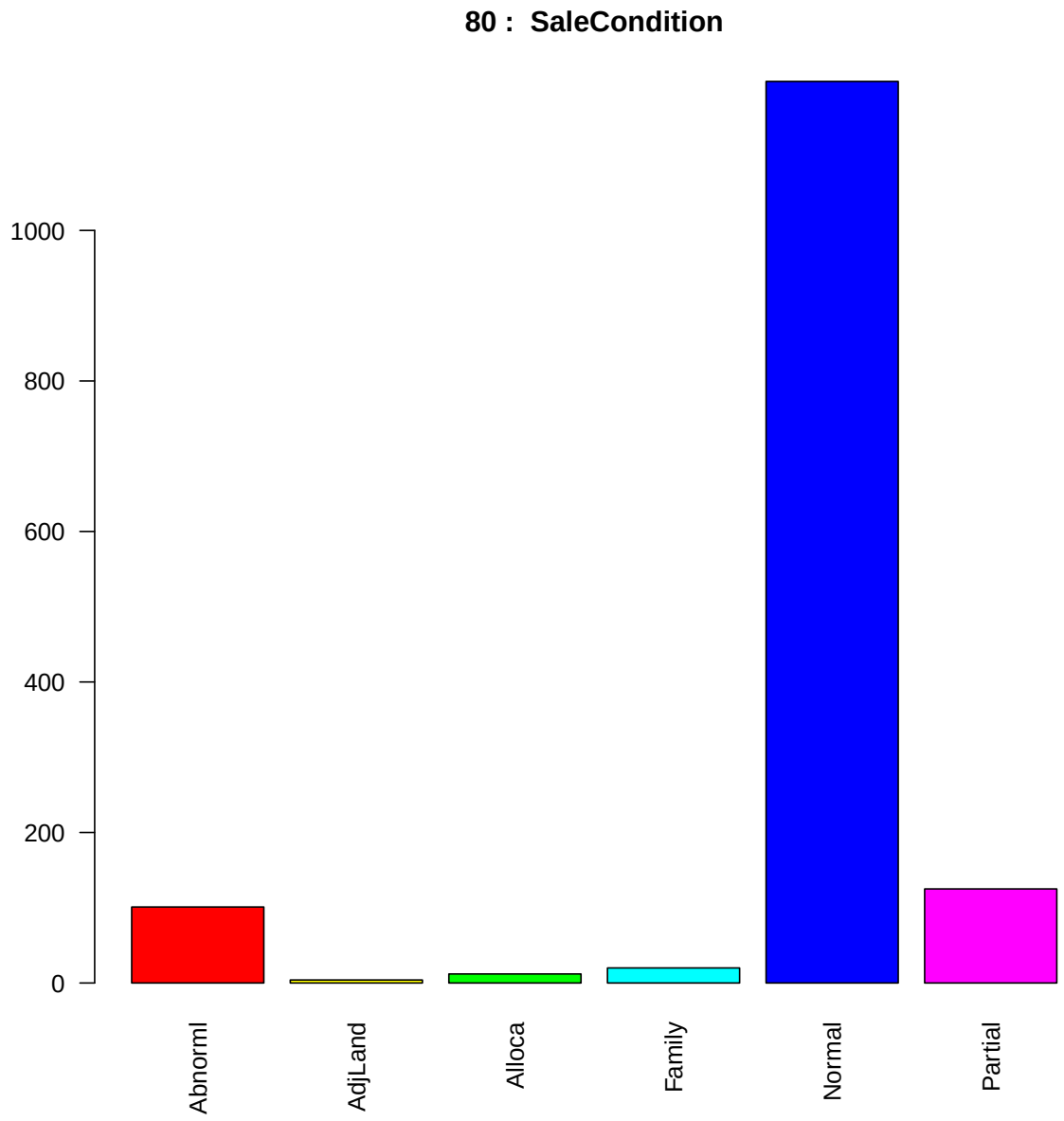




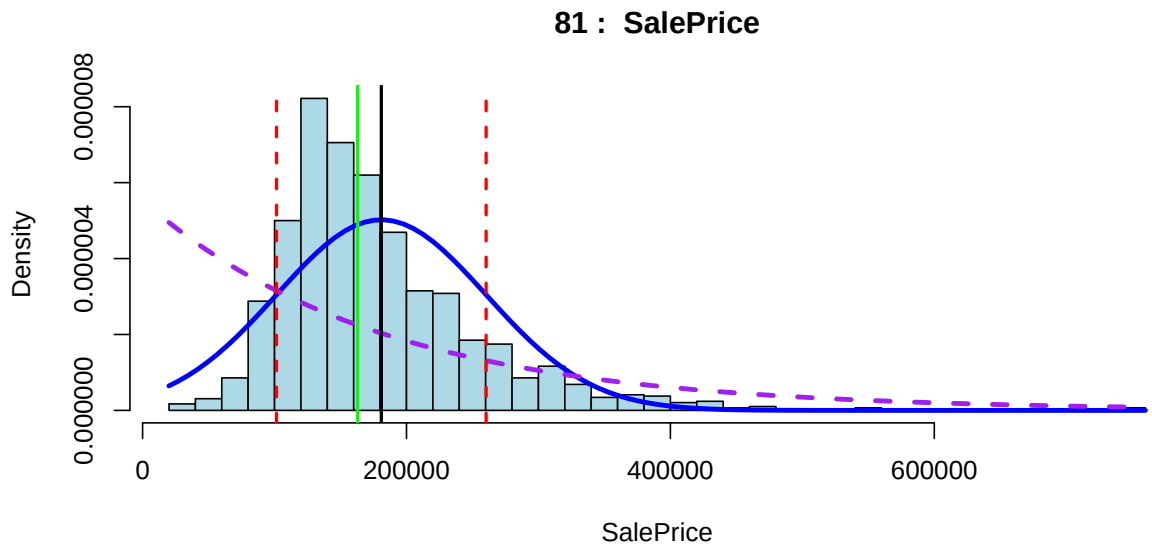

```
##          Min.       1st Qu.        Median          Mean       3rd Qu.
## 2006.00000000 2007.00000000 2008.00000000 2007.81575342 2009.00000000
##          Max.          STDEV
## 2010.00000000    1.32809512
```

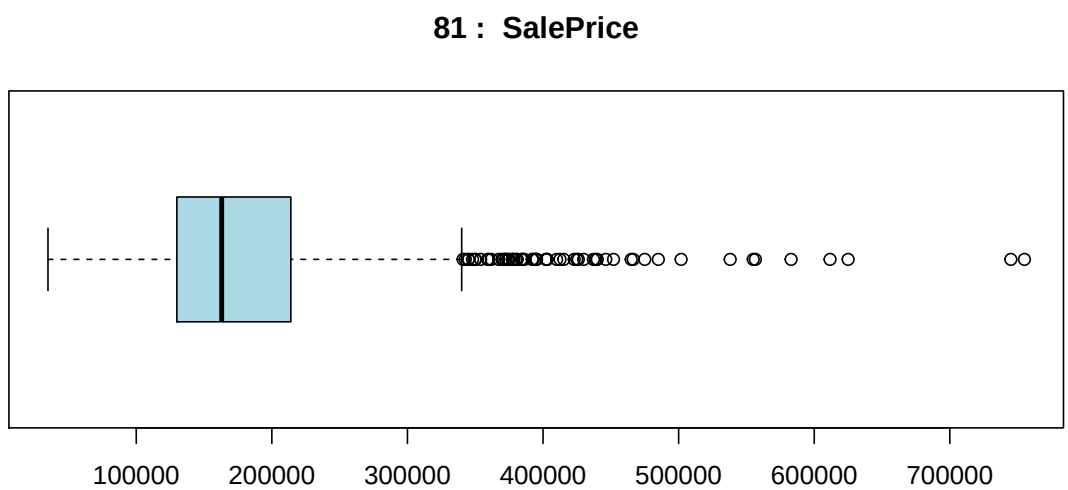



```
## SaleType
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD  None
##    43     2     9     5     5     4   122     3  1267     0
```




```
## SaleCondition
## Abnorml AdjLand  Alloca  Family  Normal Partial
##     101       4      12      20    1198     125
```






```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##  34900.0000 129975.0000 163000.0000 180921.1959 214000.0000 755000.0000
##       STDEV
##  79442.5029
```

Transform SalePrice by taking log

If we try to predict SalePrice ~ [some list of variables] there is a chance that the result could be negative. This would not make sense.
We want to ensure that the model only returns positive values for SalePrice. One way to do this is to fit log(SalePrice) rather than Saleprice.

```r
summary(train.df$SalePrice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34900  129975  163000  180921  214000  755000
```

```r
logtrain.df <- train.df %>% mutate(SalePrice = log(SalePrice))
summary(logtrain.df$SalePrice)
```

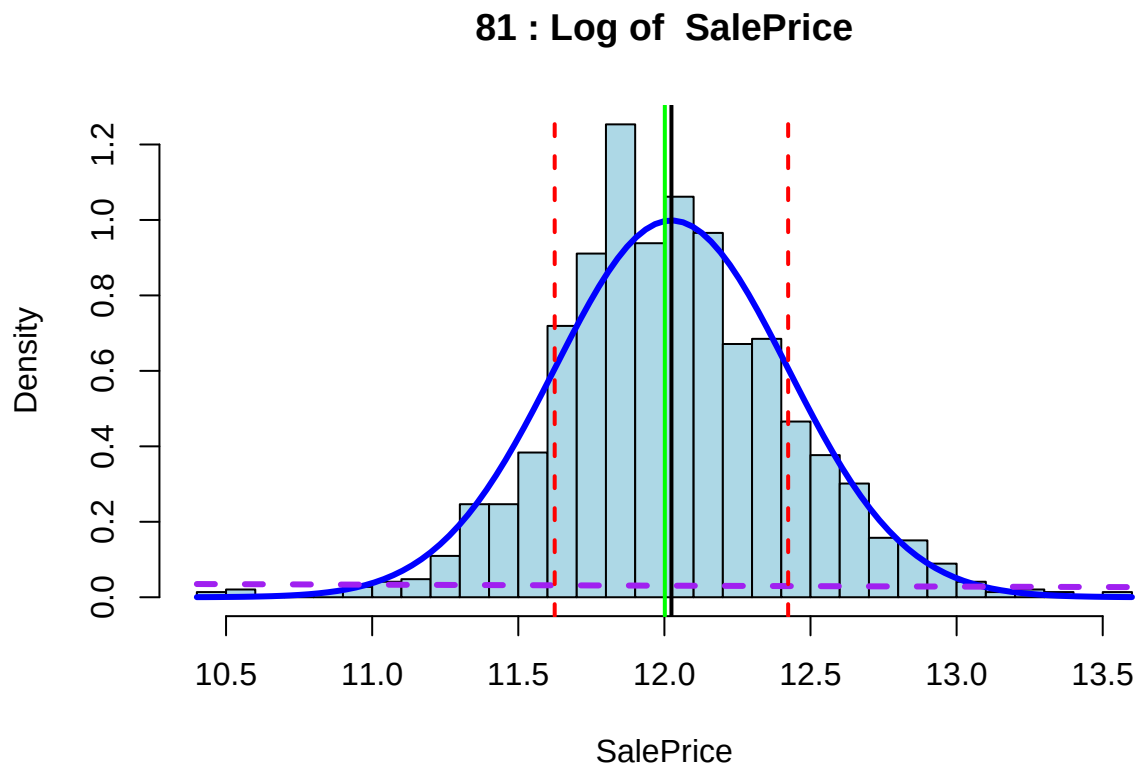
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 10.4602 11.7751 12.0015 12.0241 12.2737 13.5345
```

Note that we have not renamed the variable (e.g., to LogSalePrice.)

Rather, we need to remember that it is still named "SalePrice" in the new dataframe "logtrain.df",

but the values now represent the logarithm of the SalePrice.

```r
item=81
  thisname = attr(logtrain.df[item],"names")
  # create a title which incorporates the sequence number
  # of the variable along with the name, for guidance
  mainheader = paste(item, ": Log of ", thisname)
  #print(thisname)
  rawitem  = logtrain.df[item]
  thisitem = logtrain.df[[item]]
  thisclass = class(thisitem)
    ## Compute the summary statistics, plus the standard deviation
    numresult=c(summary(thisitem),
                STDEV=sd(thisitem,na.rm=T))

    ## plot the histogram
    hist(thisitem,breaks=30,main = mainheader,xlab=thisname,col="lightblue",probability =  T)

    ## add a normal curve fit
    curve(dnorm(x, mean = mean(thisitem), sd = sd(thisitem)), col="blue", lwd=3 , add=TRUE)

    ## add an exponential curve fit
    curve(dexp(x, rate=1/mean(thisitem)), col="purple", lwd=3, lty="dashed", add=TRUE)

    ## add a vertical line for median
    abline(v=numresult["Median"],col="green", lwd=2)

    ## add a vertical line for mean
    abline(v=numresult["Mean"],col="black",lwd=2)

    ## add vertical lines for down and up one standard deviation
    abline(v=numresult["Mean"]-numresult["STDEV"],col="red",lty="dashed", lwd=2)
    abline(v=numresult["Mean"]+numresult["STDEV"],col="red",lty="dashed", lwd=2)
```




```
    print(numresult)
```

```
##          Min.       1st Qu.        Median          Mean       3rd Qu.          Max.
## 10.460242108 11.775097348 12.001505480 12.024050901 12.273731294 13.534473028
##         STDEV
##   0.399451868
```

The transformed SalePrice now resembles a Normal distribution.

split up LOGtrain dataset - post cleaning

Note that we are now using the dataframe where SalePrice has been replaced by its logarithm...

```
# 19 variables --> 0 after cleaning
train.df.somemissing=logtrain.df[sapply(logtrain.df, function(x) sum(is.na(x))>0)]
# 62 variables --> 81 after cleaning
train.df.nomissing=logtrain.df[sapply(logtrain.df, function(x) sum(is.na(x))==0)]

#### Note that SalePrice is now "numeric" rather than "integer" because of logarithm
# 38 variables -> 37
train.df.numeric=logtrain.df[sapply(logtrain.df, is.numeric)]
# 43 variables -> 44
train.df.factor=logtrain.df[sapply(logtrain.df, is.factor)]

# 35 variables -> 37
```

```r
train.df.numeric.nomissing = train.df.nomissing[sapply(train.df.nomissing, is.numeric)]
# 27 fariables -> 44
train.df.factor.nomissing =  train.df.nomissing[sapply(train.df.nomissing, is.factor)]
```

Boruta: Variable Importance

The importance of each original variable is ranked using the `Boruta` function.

Boruta is a feature selection algorithm, using Random Forest to select "important" variables.

It classifies the feature variables into 3 levels based on the p-value specified: Confirmed, Tentative, or Rejected.

Use the Boruta algorithm to determine the "importance" of the various variables in predicting SalePrice.

We run this search on just those variables which have no "missing" values - which should now cover everything...

```r
library(Boruta)
```

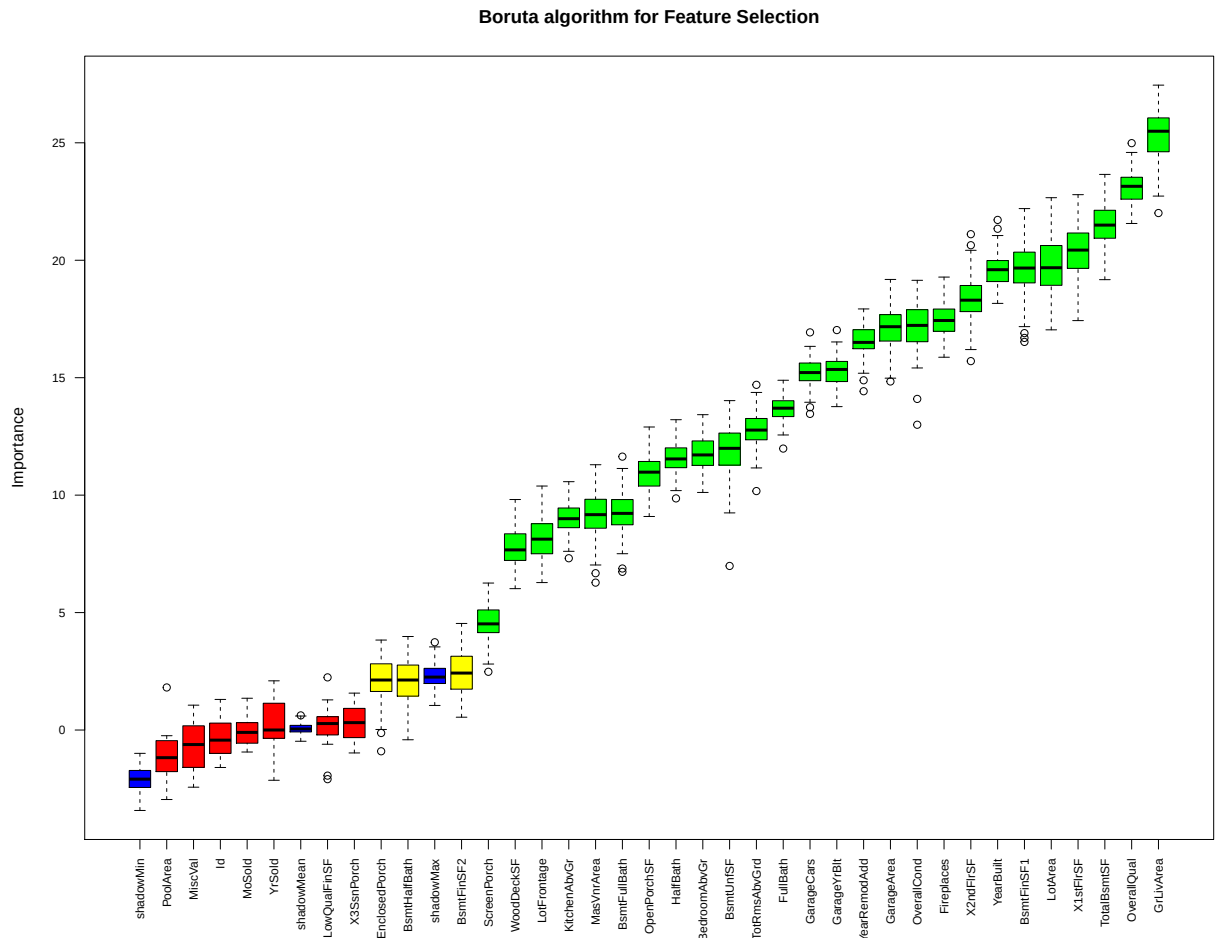
```
## Loading required package: ranger
```

```r
set.seed(1)

# Use a version of the dataframe which incorporates just those variables
# which have no "Missing" (NA) values.
# Consider all such variables (SalePrice ~ .)
# to assess "importance" in prediction of SalePrice
#
# I have found a problem when including categorical (factor) variables.
# The issue occurs when a factor variable contains a level which is
#     present in the test set,
#     but is not present in the training set.
# The model can be built and optimized,
# but when it comes time to use the "predict()" function at the end,
# it will fail if there is any new level in a factor in the test data
# which was not present in the training data.
#
# For this reason, I will now try the below on numerical values only.
# Boruta(SalePrice  ~ .  , data=train.df.nomissing)->Bor.hvo

Boruta(SalePrice  ~ .  , data=train.df.numeric.nomissing)->Bor.hvo
### Note that in the above database, SalePrice has been replaced by its logarithm.
### Thus we are measuring importance relative to fitting log(SalePrice)

print(Bor.hvo)
```

```
## Boruta performed 99 iterations in 6.35424332 mins.
##  26 attributes confirmed important: BedroomAbvGr, BsmtFinSF1,
## BsmtFullBath, BsmtUnfSF, Fireplaces and 21 more;
##  7 attributes confirmed unimportant: Id, LowQualFinSF, MiscVal, MoSold,
## PoolArea and 2 more;
##  3 tentative attributes left: BsmtFinSF2, BsmtHalfBath, EnclosedPorch;
```

Plot Borura results

```
plot(Bor.hvo,  cex.axis=0.75, las=2, main="Boruta algorithm for Feature Selection", xlab="")
```



The above graph can be interpreted as follows:

- The 26 variables on the right (in Green) are ***Confirmed*** to be "important" .
- The most "important" variables are `GrLivArea` and `OverallQual`.
- The 7 variables on the left (in Red) are ***Rejected*** as "not important" .
- The 3 variables in the middle (in Yellow) are marked as ***Tentative*** .
- A few variables are generated by the algorithm and given names like "shadowMin", "shadowMean", and "shadowMax" are shown above in Blue – these can be ignored.

Tabular listing of Boruta results:

```
BorutaFinal       <- Bor.hvo$finalDecision

# Here is the alphabetized list of Boruta decision:
BorutaFinalAlpha  <- BorutaFinal[order(names(BorutaFinal))] %>% t %>% t

# Extract the numerical median results from the Boruta algorithm
BorutaMedian      <- apply(X = Bor.hvo$ImpHistory, MARGIN = 2, FUN = median)
# drop the three "shadow" variables from the list
BorutaMedian      <- BorutaMedian[BorutaMedian %>% names %>% grep("shadow",.,invert=T)]
# alphabetize the list
BorutaMedianAlpha <- BorutaMedian[order(names(BorutaMedian))]
BorutaMedianAlphaNum <- as.numeric(BorutaMedianAlpha)

BorutaMedianAlpha <- BorutaMedian[order(names(BorutaMedian))] %>% t %>% t

BorutaJoinedAlpha <- cbind(BorutaFinalAlpha,BorutaMedianAlpha)

BorutaFinalAlphaResults <- as.character(BorutaFinalAlpha)
BorutaFinalAlphaNames <- BorutaFinal[names(BorutaFinal) %>% order] %>% names()

# Here's the alphabetical list of the Boruta results:
BorutaByAlpha <- cbind(BorutaFinalAlphaNames,BorutaFinalAlphaResults,BorutaMedianAlphaNum)
BorutaByAlpha
```

```
##       BorutaFinalAlphaNames BorutaFinalAlphaResults BorutaMedianAlphaNum
##  [1,] "BedroomAbvGr"        "Confirmed"             "11.7115631649261"
##  [2,] "BsmtFinSF1"          "Confirmed"             "19.668547222657"
##  [3,] "BsmtFinSF2"          "Tentative"             "2.42498455658155"
##  [4,] "BsmtFullBath"        "Confirmed"             "9.22319328637244"
##  [5,] "BsmtHalfBath"        "Tentative"             "2.1295463138818"
##  [6,] "BsmtUnfSF"           "Confirmed"             "11.9925035688715"
##  [7,] "EnclosedPorch"       "Tentative"             "2.12930375713697"
##  [8,] "Fireplaces"          "Confirmed"             "17.4344983179933"
##  [9,] "FullBath"            "Confirmed"             "13.7016169122141"
## [10,] "GarageArea"          "Confirmed"             "17.1704912492497"
## [11,] "GarageCars"          "Confirmed"             "15.2188326902613"
## [12,] "GarageYrBlt"         "Confirmed"             "15.3506342409513"
## [13,] "GrLivArea"           "Confirmed"             "25.4930927972869"
## [14,] "HalfBath"            "Confirmed"             "11.5415123599231"
## [15,] "Id"                  "Rejected"              "-Inf"
## [16,] "KitchenAbvGr"        "Confirmed"             "8.99698259543654"
## [17,] "LotArea"             "Confirmed"             "19.6812800913697"
## [18,] "LotFrontage"         "Confirmed"             "8.12636235442263"
## [19,] "LowQualFinSF"        "Rejected"              "-Inf"
## [20,] "MasVnrArea"          "Confirmed"             "9.16946036047692"
## [21,] "MiscVal"             "Rejected"              "-Inf"
## [22,] "MoSold"              "Rejected"              "-Inf"
## [23,] "OpenPorchSF"         "Confirmed"             "10.9784229705144"
## [24,] "OverallCond"         "Confirmed"             "17.2263281722049"
## [25,] "OverallQual"         "Confirmed"             "23.1480970224686"
## [26,] "PoolArea"            "Rejected"              "-Inf"
## [27,] "ScreenPorch"         "Confirmed"             "4.51936048129632"
```

```
## [28,] "TotalBsmtSF"         "Confirmed"             "21.4959257874995"
## [29,] "TotRmsAbvGrd"        "Confirmed"             "12.7699265542664"
## [30,] "WoodDeckSF"          "Confirmed"             "7.66919114098259"
## [31,] "X1stFlrSF"           "Confirmed"             "20.4328426049174"
## [32,] "X2ndFlrSF"           "Confirmed"             "18.3004906803124"
## [33,] "X3SsnPorch"          "Rejected"              "-Inf"
## [34,] "YearBuilt"           "Confirmed"             "19.5982630334252"
## [35,] "YearRemodAdd"        "Confirmed"             "16.5023514299386"
## [36,] "YrSold"              "Rejected"              "-Inf"
```

```r
# Here's the numerical list of the Boruta results:
BorutaByNum <- BorutaByAlpha[order(BorutaMedianAlphaNum),]
BorutaByNum %>% kable() %>% kable_styling(c("bordered","striped"),full_width = F)
```

BorutaFinalAlphaNames	BorutaFinalAlphaResults	BorutaMedianAlphaNum
Id	Rejected	-Inf
LowQualFinSF	Rejected	-Inf
MiscVal	Rejected	-Inf
MoSold	Rejected	-Inf
PoolArea	Rejected	-Inf
X3SsnPorch	Rejected	-Inf
YrSold	Rejected	-Inf
EnclosedPorch	Tentative	2.12930375713697
BsmtHalfBath	Tentative	2.1295463138818
BsmtFinSF2	Tentative	2.42498455658155
ScreenPorch	Confirmed	4.51936048129632
WoodDeckSF	Confirmed	7.66919114098259
LotFrontage	Confirmed	8.12636235442263
KitchenAbvGr	Confirmed	8.99698259543654
MasVnrArea	Confirmed	9.16946036047692
BsmtFullBath	Confirmed	9.22319328637244
OpenPorchSF	Confirmed	10.9784229705144
HalfBath	Confirmed	11.5415123599231
BedroomAbvGr	Confirmed	11.7115631649261
BsmtUnfSF	Confirmed	11.9925035688715
TotRmsAbvGrd	Confirmed	12.7699265542664
FullBath	Confirmed	13.7016169122141
GarageCars	Confirmed	15.2188326902613
GarageYrBlt	Confirmed	15.3506342409513
YearRemodAdd	Confirmed	16.5023514299386
GarageArea	Confirmed	17.1704912492497
OverallCond	Confirmed	17.2263281722049
Fireplaces	Confirmed	17.4344983179933
X2ndFlrSF	Confirmed	18.3004906803124
YearBuilt	Confirmed	19.5982630334252
BsmtFinSF1	Confirmed	19.668547222657
LotArea	Confirmed	19.6812800913697
X1stFlrSF	Confirmed	20.4328426049174
TotalBsmtSF	Confirmed	21.4959257874995
OverallQual	Confirmed	23.1480970224686
GrLivArea	Confirmed	25.4930927972869

```
BorutaConfirmed <- BorutaByNum[BorutaByNum[,"BorutaFinalAlphaResults"]=="Confirmed",]
BorutaConfirmedFeatures<-BorutaConfirmed[,1]

# Include the name of the target variable on the list
BorutaPlusTarget<-c("SalePrice",BorutaConfirmedFeatures)

library(tidyselect)
# make a dataframe which includes just these "confirmed important" variables
train.df.Boruta <- logtrain.df %>% dplyr::select(vars_select(.vars = names(logtrain.df),BorutaPlusTarge
```

According to Boruta, the two most "important" variables for predicting SalePrice are:

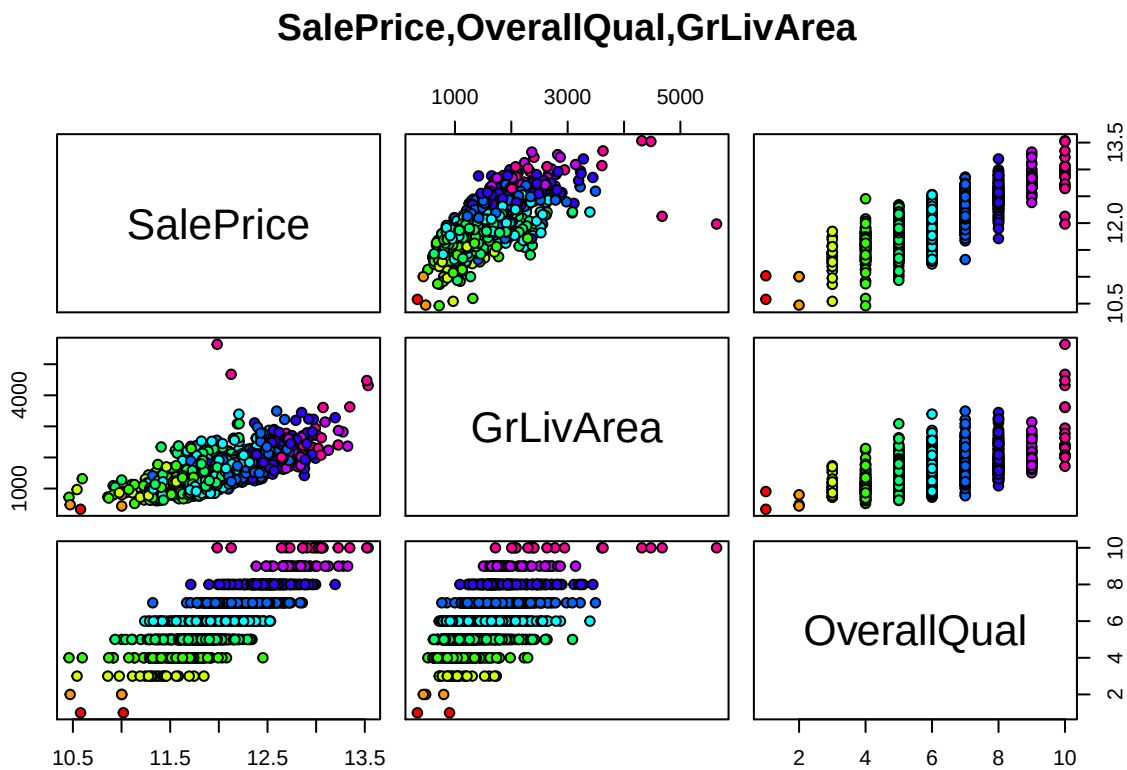
- GrLivArea (continuous)
- OverallQual (numeric, integer rating from 1 through 10)

Pull just these three variables into their own data frame

```
my.train.df.correl <- logtrain.df %>% dplyr::select(vars_select(.vars = names(logtrain.df),
                                                    c("SalePrice","GrLivArea","OverallQual")))
```

Provide a *scatterplot matrix* for at least two of the independent variables and the dependent variable.

```
pairs(my.train.df.correl,
      main = "SalePrice,OverallQual,GrLivArea",
      pch = 21,
      bg = rainbow(10)[as.factor(my.train.df.correl$OverallQual)])
```



Derive a *correlation matrix* for any three quantitative variables in the dataset.

```
### Here are the correlations of the three variables selected above:

cor(logtrain.df$SalePrice,logtrain.df$OverallQual)
```

```
## [1] 0.817184418
```

```
cor(logtrain.df$SalePrice,logtrain.df$GrLivArea)
```

```
## [1] 0.700926653
```

```
cor(logtrain.df$OverallQual,logtrain.df$GrLivArea)
```

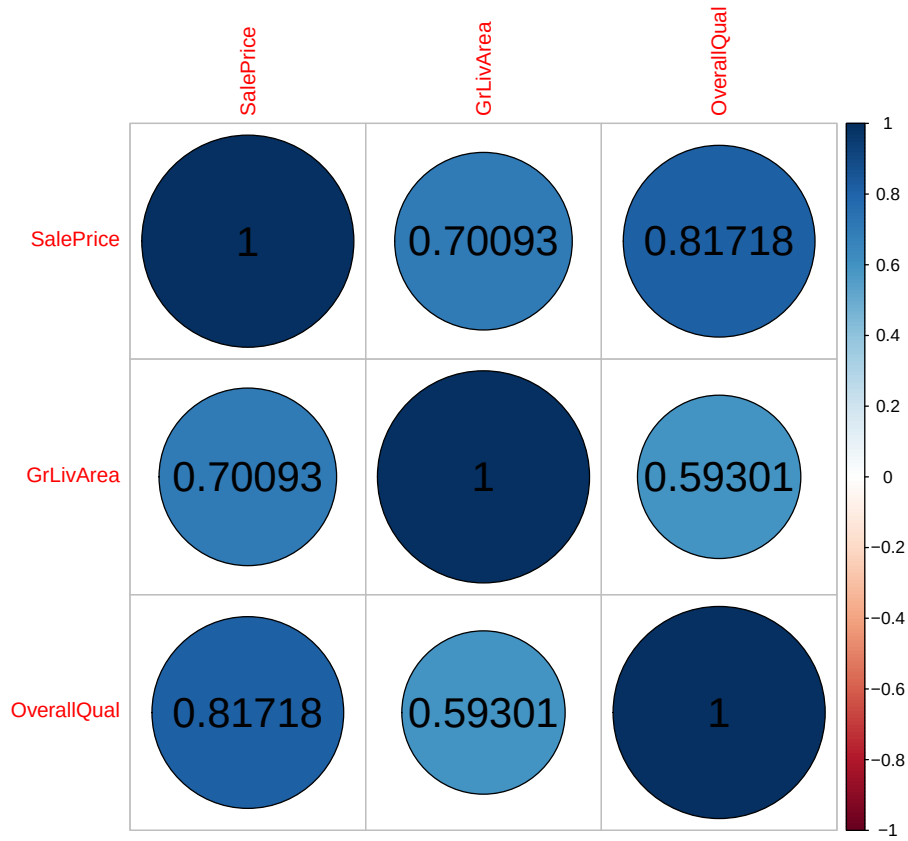
```
## [1] 0.59300743
```

```
### Make a correlation matrix
mycor3 <- cor(my.train.df.correl)
mycor3 %>% kable() %>% kable_styling(c("striped", "bordered"),full_width = F)
```

	SalePrice	GrLivArea	OverallQual
SalePrice	1.000000000	0.700926653	0.817184418
GrLivArea	0.700926653	1.000000000	0.593007430
OverallQual	0.817184418	0.593007430	1.000000000

Colorful plot of correlations

```
## corrplot 0.84 loaded
```



Check variable correlation more widely

```
library(Hmisc)
```

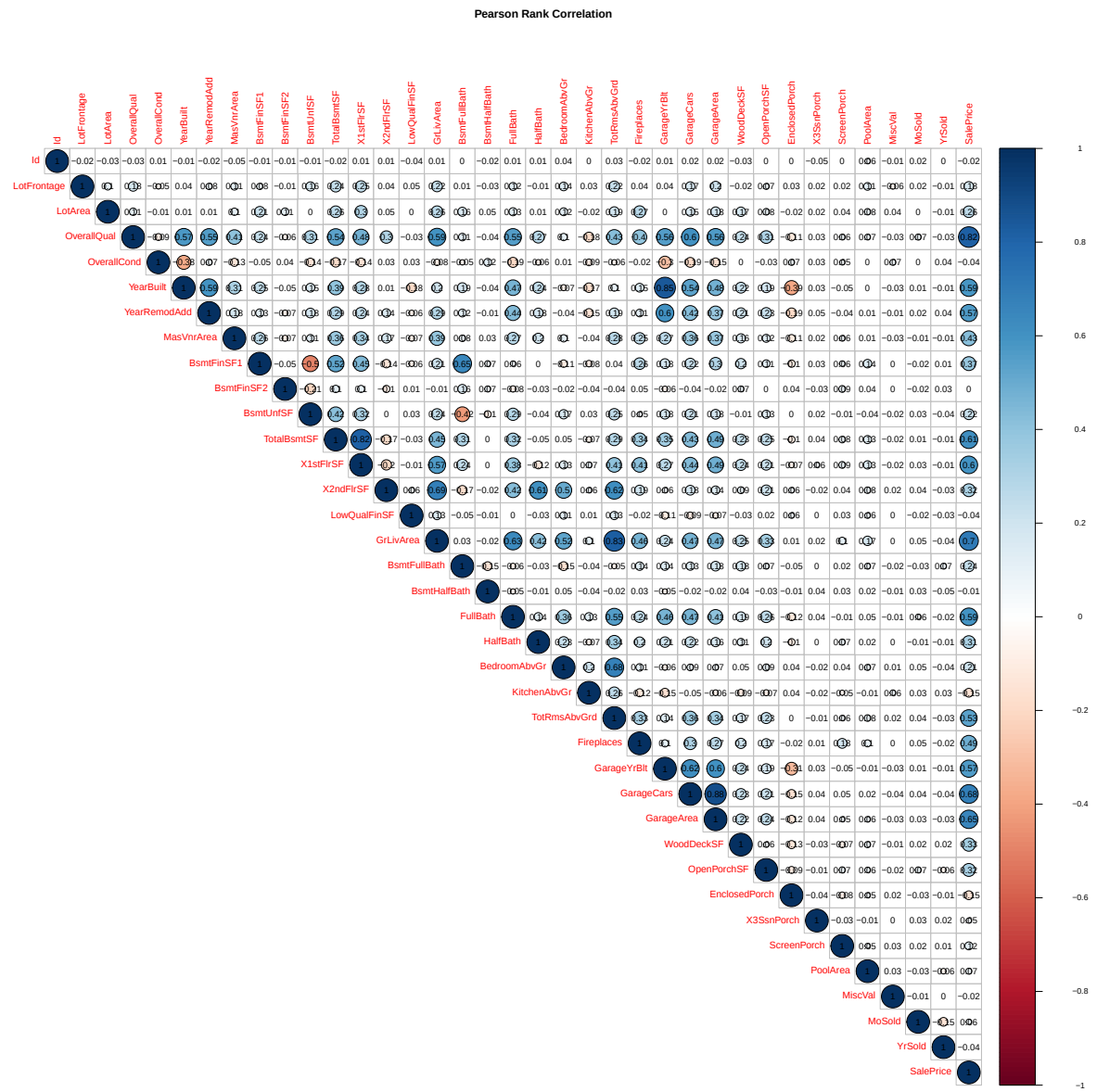
```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':
##
##     src, summarize

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
res2<-rcorr(as.matrix(train.df.numeric))
respearson=rcorr(as.matrix(train.df.numeric),type = "pearson")
resspearman=rcorr(as.matrix(train.df.numeric),type = "spearman")
res3 <- cor(as.matrix(train.df.numeric))
```

Pearson rank correlation

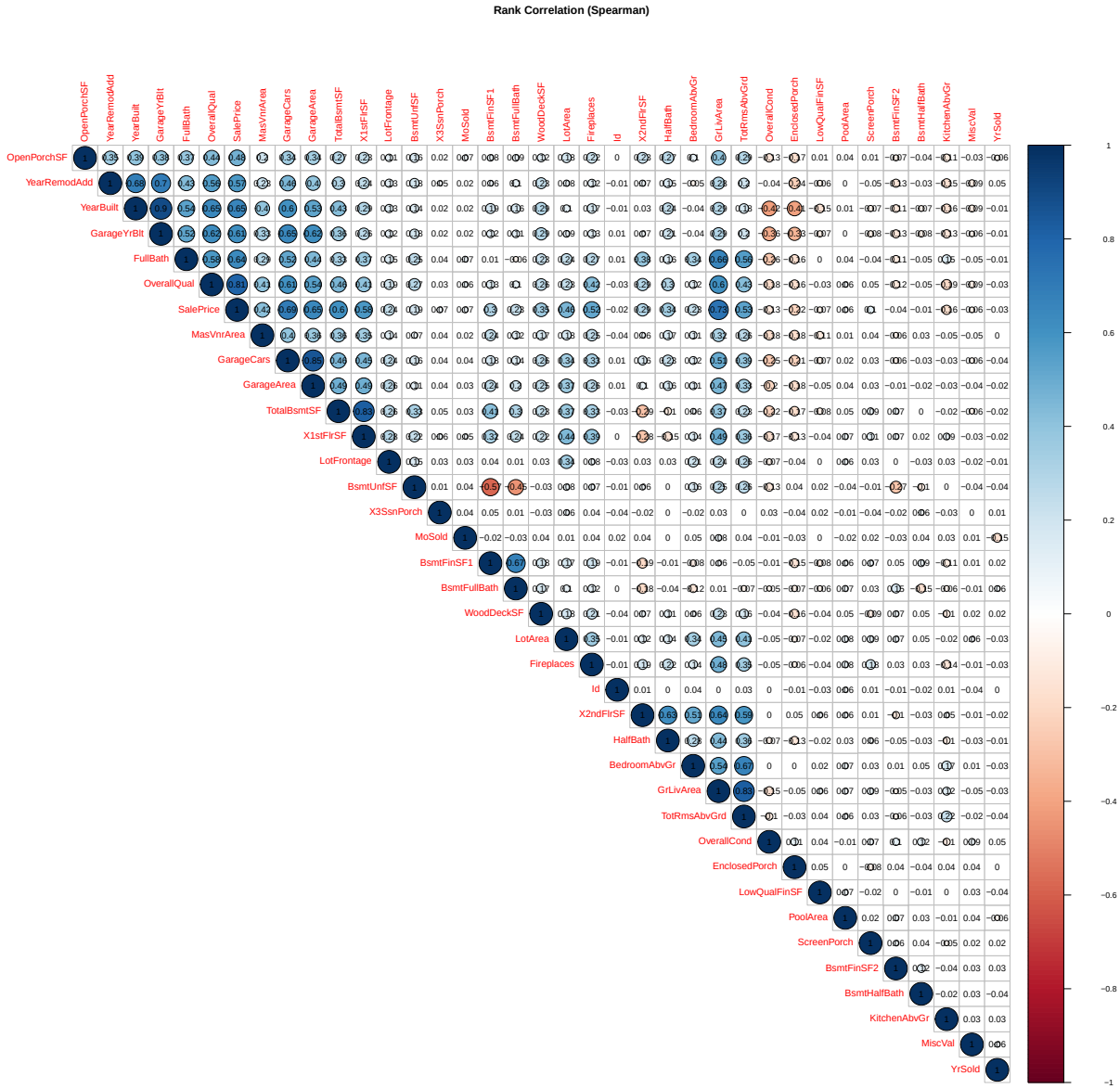


The above shows the target variable, `SalePrice` in the rightmost column, along with the pairwise correlations of tall the variables considered (for technical reasons, we have only looked at those variables without any "missing" values.)

The independent variables which display a high correlation are reflected in larger circles, with a darker blue color.

Spearman rank correlation

This is quite similar to that above, except we change the sequence in which the variables are listed. Now they are clustered together based upon similarity.



Test the hypotheses **that the correlations between each pairwise set of variables is 0 and provide an 80% confidence interval.**

SalePrice + GrLivArea

```
cor.test(~ SalePrice + GrLivArea, data=my.train.df.correl,method="pearson",conf.level = 0.8)
```

```
##
```

```
##  Pearson's product-moment correlation
##
## data:  SalePrice and GrLivArea
## t = 37.52491, df = 1458, p-value < 0.000000000000000222
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.683442483 0.717607124
## sample estimates:
##         cor
## 0.700926653
```

The p-value is zero and the 80-percent confidence interval does not include zero, so we reject H_0: true correlation equals zero, in favor of H_A .

SalePrice + OverallQual

```r
cor.test(~ SalePrice + OverallQual, data=my.train.df.correl,method="pearson",conf.level = 0.8)
```

```
##
##  Pearson's product-moment correlation
##
## data:  SalePrice and OverallQual
## t = 54.13682, df = 1458, p-value < 0.000000000000000222
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.805720508 0.828036294
## sample estimates:
##         cor
## 0.817184418
```

The p-value is zero and the 80-percent confidence interval does not include zero, so we reject H_0: true correlation equals zero, in favor of H_A .

GrLivArea + OverallQual

```r
cor.test(~ GrLivArea + OverallQual, data=my.train.df.correl,method="pearson",conf.level = 0.8)
```

```
##
##  Pearson's product-moment correlation
##
## data:  GrLivArea and OverallQual
## t = 28.12139, df = 1458, p-value < 0.000000000000000222
## alternative hypothesis: true correlation is not equal to 0
## 80 percent confidence interval:
##  0.570806147 0.614342245
## sample estimates:
##        cor
## 0.59300743
```

The p-value is zero and the 80-percent confidence interval does not include zero, so we reject H_0: true correlation equals zero, in favor of H_A .

Discuss the meaning of your analysis.

As these variables are highly correlated with each other, we can hope that the independent variables will help explain the dependent variable (SalesPrice) when we use them in the regression model.

Of course, correlation does not imply causation, but intuitively it does make sense that houses which are larger, and which are rated as being of high quality, would likely sell for higher prices.

Would you be worried about *familywise error*?

Definition: The ***familywise error rate*** (FWE or FWER) is the probability of a coming to ***at least one false conclusion*** in a series of hypothesis tests.
In other words, it's the probability of making at least one Type I Error.
(The FWER is also called ***alpha inflation*** or ***cumulative Type I error.***)

The formula to estimate the familywise error rate is: $FWE \leq 1\breve{}(1\breve{}\alpha_i)^c$, where:

- α_i = alpha level for an individual test (here, .20 for an 80% confidence interval), and
- c = Number of comparisons.

So here, $FWER \leq= 1\breve{}(1\breve{}0.2)^3 = 1 - 0.8^3 = 1 - 0.512 = 0.488$.

This means that we have nearly a 50 percent chance of coming to a false conclusion across three hypothesis tests.

Why or why not?

I'm not concerned, because we could implement the Bonferroni correction, reducing the alpha on each of the three tests, and still would come up with an acceptable result.

---

5 points. Linear Algebra and Correlation.

Invert your correlation matrix from above.

(This is known as the ***precision matrix*** and contains ***variance inflation factors*** on the diagonal.)

```
### My correlation matrix
mycor3
```

```
##               SalePrice   GrLivArea OverallQual
## SalePrice   1.000000000 0.700926653 0.817184418
## GrLivArea   0.700926653 1.000000000 0.593007430
## OverallQual 0.817184418 0.593007430 1.000000000
```

```
### Inverted
precision_matrix <- solve(mycor3)
precision_matrix
```

```
##               SalePrice    GrLivArea  OverallQual
## SalePrice    3.84574814 -1.283198226 -2.381739374
## GrLivArea   -1.28319823  1.970555950 -0.119944724
## OverallQual -2.38173937 -0.119944724  3.017448416
```

Multiply the correlation matrix by the precision matrix, and then

```
result1 <- precision_matrix %*% mycor3
result1
```

```
##                              SalePrice                  GrLivArea
## SalePrice   1.000000000000000222044605 0.000000000000000222044605
## GrLivArea   0.000000000000000111022302 1.000000000000000222044605
## OverallQual 0.000000000000000000000000 0.000000000000000222044605
##                            OverallQual
## SalePrice   0.000000000000000000000000
## GrLivArea   0.000000000000000194289029
## OverallQual 1.000000000000000000000000
```

```
round(result1,digits=8)
```

```
##             SalePrice GrLivArea OverallQual
## SalePrice           1         0           0
## GrLivArea           0         1           0
## OverallQual         0         0           1
```

multiply the precision matrix by the correlation matrix.

```
result2 <-  mycor3 %*% precision_matrix
result2
```

```
##                             SalePrice                   GrLivArea
## SalePrice   1.000000000000000022204460 -0.000000000000000001110223025
## GrLivArea   0.000000000000000044408921  1.000000000000000000000000
## OverallQual 0.000000000000000000000000 -0.000000000000000000277555756
##                            OverallQual
## SalePrice   0.000000000000000000000000
## GrLivArea   0.000000000000000222044605
## OverallQual 1.000000000000000000000000
```

```
round(result2,digits=8)
```

```
##             SalePrice GrLivArea OverallQual
## SalePrice           1         0           0
## GrLivArea           0         1           0
## OverallQual         0         0           1
```

Conduct *LU decomposition* on the matrix.

```
library(matrixcalc)
myLU <- lu.decomposition(mycor3)
myLU
```

```
## $L
##             [,1]        [,2] [,3]
## [1,] 1.000000000 0.000000000    0
## [2,] 0.700926653 1.000000000    0
## [3,] 0.817184418 0.039750381    1
##
## $U
##      [,1]        [,2]         [,3]
## [1,]    1 0.700926653 0.8171844179
## [2,]    0 0.508701828 0.0202210915
## [3,]    0 0.000000000 0.3314058310
```

```
### check the results
checkLU <- myLU$L %*% myLU$U
checkLU
```

```
##             [,1]        [,2]        [,3]
## [1,] 1.000000000 0.700926653 0.817184418
## [2,] 0.700926653 1.000000000 0.593007430
## [3,] 0.817184418 0.593007430 1.000000000
```

```
### check the difference
checkLU - mycor3
```

```
##             SalePrice GrLivArea OverallQual
## SalePrice           0         0           0
## GrLivArea           0         0           0
## OverallQual         0         0           0
```

```r
all.equal(checkLU , mycor3)
```

```
## [1] "Attributes: < Length mismatch: comparison on first 1 components >"
```

```r
### difference is due to dimnames, so make them same
dimnames(checkLU) <- dimnames(mycor3)

### again, check for equality
all.equal(checkLU , mycor3)
```

```
## [1] TRUE
```

---

5 points. Calculus-Based Probability & Statistics.

Many times, it makes sense to ***fit a closed form distribution to data.***

Select a variable in the Kaggle.com training dataset that is ***skewed to the right***, shift it so that the minimum value is absolutely above zero if necessary.

Here we will compute the skewness for each quantitative value in the dataset:

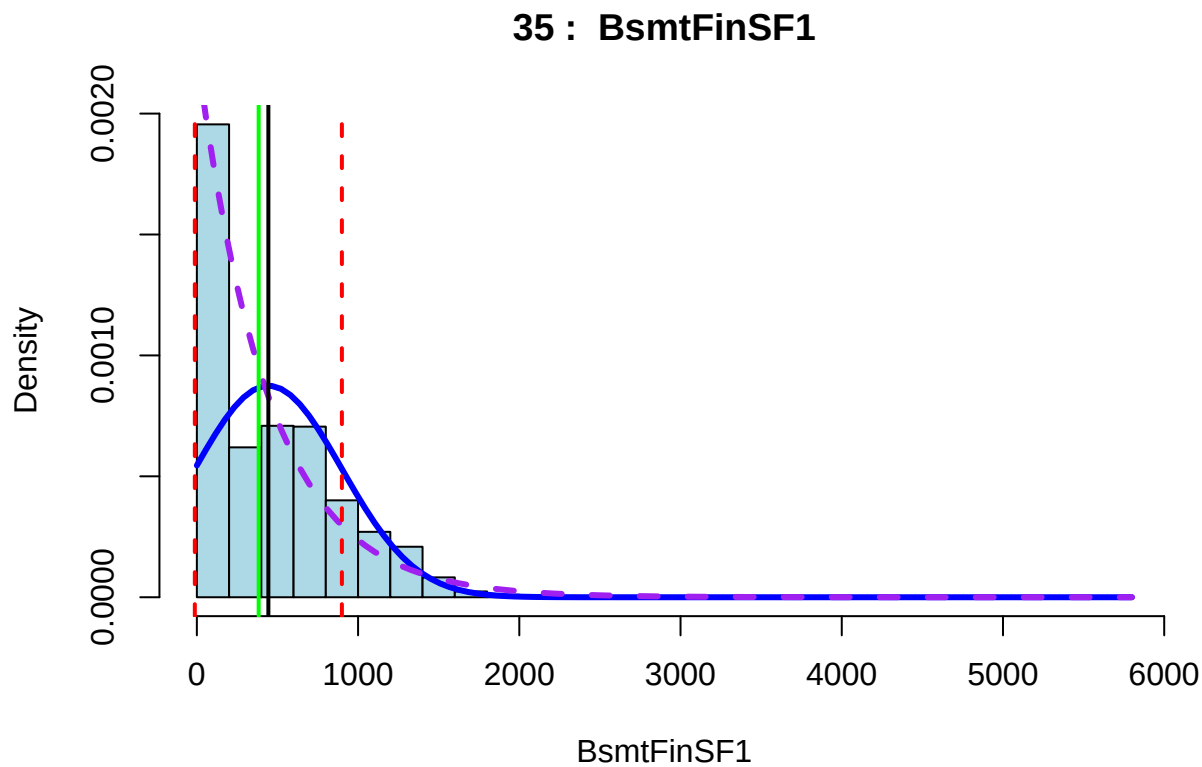
```
library(moments)
### Compute the skewness for each numeric variable (n=37) in the dataset
skewlist <- sapply(X=train.df.numeric, FUN = skewness) %>% sort%>% t %>% t
skewlist
```

```
##                        [,1]
## GarageYrBlt   -0.6936154112
## YearBuilt     -0.6128307242
## YearRemodAdd  -0.5030444968
## GarageCars    -0.3421968954
## Id             0.0000000000
## FullBath       0.0365239844
## YrSold         0.0961695796
## SalePrice      0.1212103673
## GarageArea     0.1797959421
## BedroomAbvGr   0.2115724416
## MoSold         0.2118350602
## OverallQual    0.2167209765
## LotFrontage    0.2675471493
## BsmtFullBath   0.5954540376
## Fireplaces     0.6488976310
## HalfBath       0.6752028348
## TotRmsAbvGrd   0.6756457673
## OverallCond    0.6923552136
## X2ndFlrSF      0.8121942732
## BsmtUnfSF      0.9193227016
## GrLivArea      1.3651559548
## X1stFlrSF      1.3753417422
## TotalBsmtSF    1.5226880870
## WoodDeckSF     1.5397916998
## BsmtFinSF1     1.6837708962
## OpenPorchSF    2.3619119286
## MasVnrArea     2.6748646898
## EnclosedPorch  3.0866964714
## BsmtHalfBath   4.0991856695
## ScreenPorch    4.1179773828
## BsmtFinSF2     4.2508880171
## KitchenAbvGr   4.4837840939
## LowQualFinSF   9.0020804177
## X3SsnPorch    10.2937523572
## LotArea       12.1951421251
## PoolArea      14.8131346604
## MiscVal       24.4516396173
```

The items at the bottom of the list are most heavily skewed to the right.
Some of these have a large number of zero values, for example since there are only 7 homes with pools, the other 1453 homes have a zero for `PoolArea` .
We note that Variable number 35: `BsmtFinSF1` has a smaller number of zero entries (467) and a reasonable right skewness (1.68377) .

Its plot looks somewhat exponential:

```
for (item in 35 ){
  thisname = attr(logtrain.df[item],"names")
  # create a title which incorporates the sequence number of the variable along with the name, for guid
  mainheader = paste(item, ": ", thisname)
  #print(thisname)
  rawitem  = logtrain.df[item]
  thisitem = logtrain.df[[item]]
  thisclass = class(thisitem)

  if (thisclass=="integer"||thisclass=="numeric") {
    numresult=c(summary(thisitem),
                STDEV=sd(thisitem,na.rm=T))
    hist(thisitem,breaks=30,main = mainheader,xlab=thisname,col="lightblue",probability =  T)
    curve(dnorm(x, mean = mean(thisitem), sd = sd(thisitem)), col="blue", lwd=3 , add=TRUE)
    curve(dexp(x, rate=1/mean(thisitem)), col="purple", lwd=3, lty="dashed", add=TRUE)
    abline(v=numresult["Median"],col="green", lwd=2)
    abline(v=numresult["Mean"],col="black",lwd=2)
    abline(v=numresult["Mean"]-numresult["STDEV"],col="red",lty="dashed", lwd=2)
    abline(v=numresult["Mean"]+numresult["STDEV"],col="red",lty="dashed", lwd=2)
    print(numresult)
  }
}
```



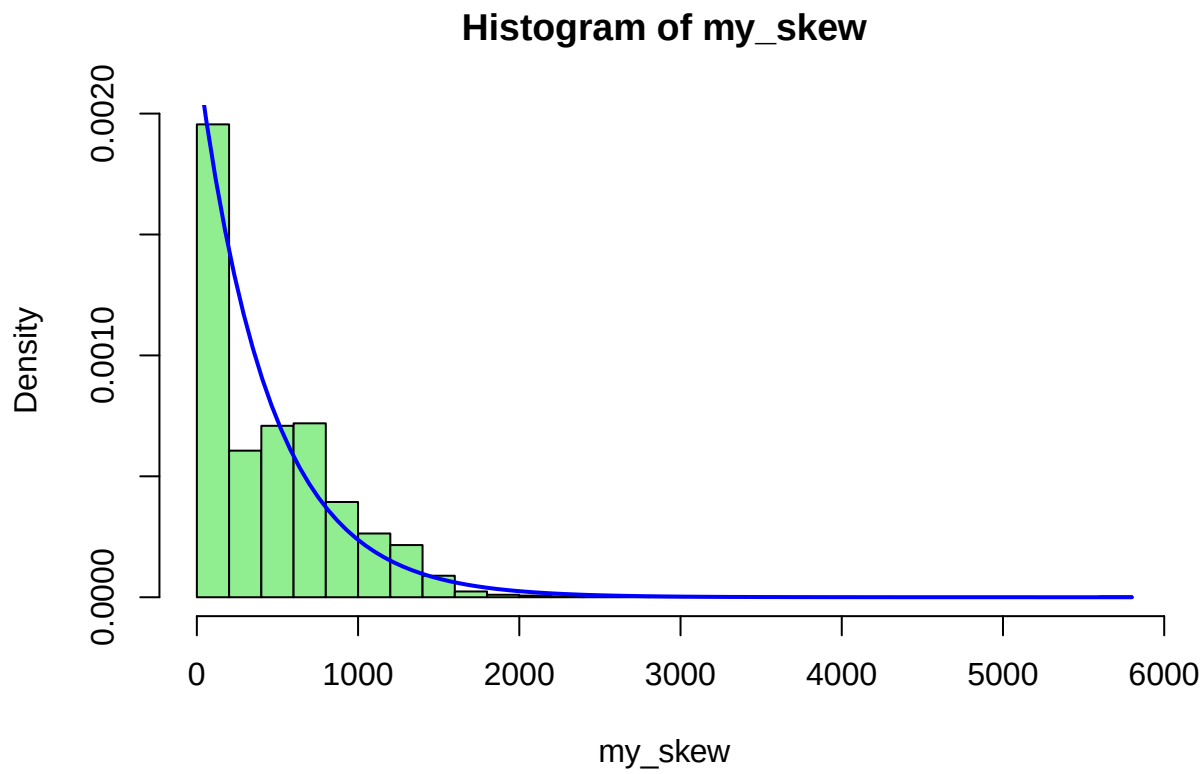

```
##        Min.     1st Qu.      Median        Mean     3rd Qu.        Max.
##    0.000000    0.000000  383.500000  443.639726  712.250000 5644.000000
##       STDEV
##  456.098091
```

(Here, the purple curve represents an exponential fit.)

select skewed variable > 0

Shift the distribution by adding 1

```
### add one to the variable to move the zero entries to 1
my_skew = train.df$BsmtFinSF1 + 1

hh=hist(my_skew,breaks=30,probability = T,col="lightgreen") ;
curve(dexp(x,rate=1/mean(my_skew)),col="blue",lwd=2,add=T)
```



Then load the MASS package and run `fitdistr` to fit an ***exponential*** probability density function. (See https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html).

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
expfit <- fitdistr(my_skew , densfun = "exponential")
expfit
```

```
##         rate
##   0.0022490118212
##  (0.0000588593295)
```

```
# Check that it is the reciprocal of the mean
expmean <- mean(my_skew)
expmean
```

```
## [1] 444.639726
```

```r
1/expmean
```

```
## [1] 0.00224901182
```

```r
# are they equal?
1/expmean - expfit$estimate
```

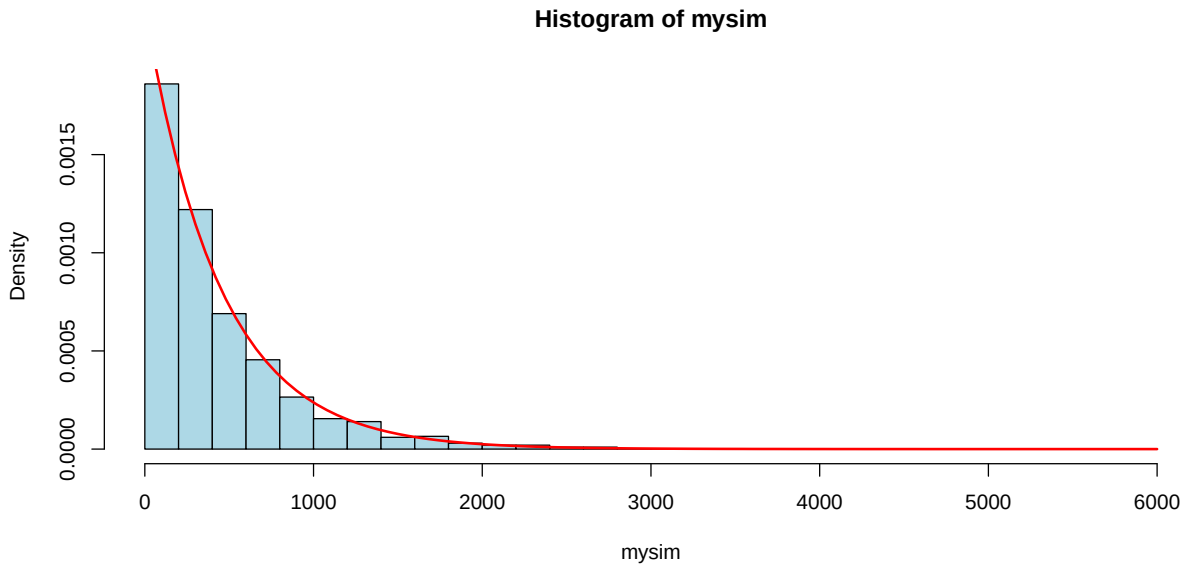
```
## rate
##    0
```

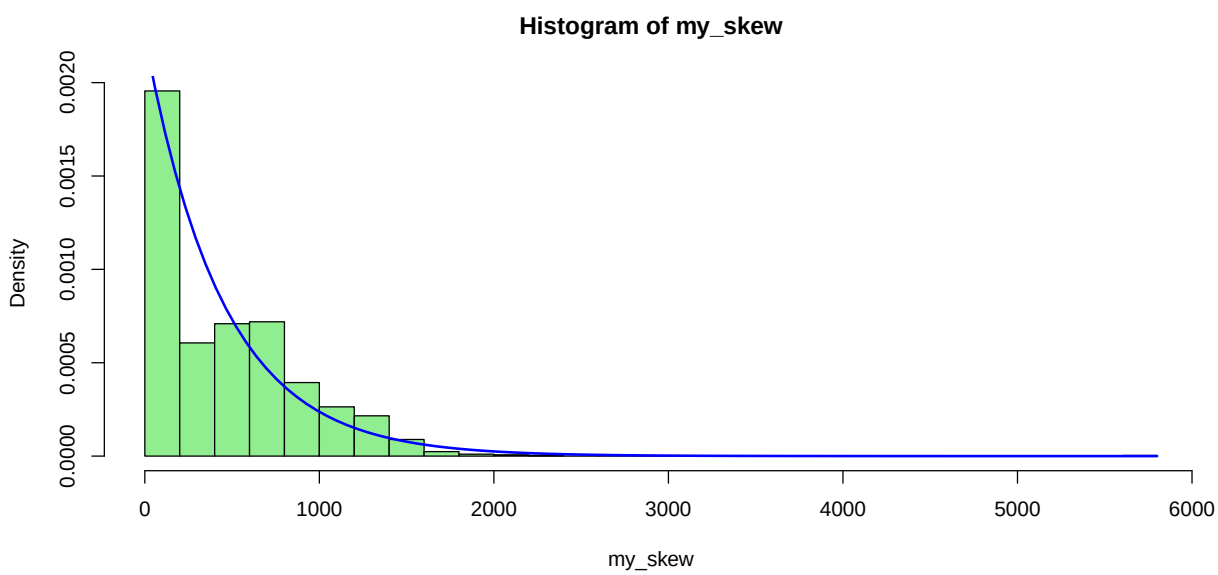
Find the optimal value of λ for this distribution,
and then ***take 1000 samples*** from this exponential distribution using this value (e.g., rexp(1000, λ)).

```r
lamda = expfit$estimate
set.seed(12344)
mysim <- rexp(1000,lamda)
```

Plot a histogram and compare it with a histogram of your original variable.

```r
par(mfrow = c(2, 1))


ss=hist(mysim,breaks=hh$breaks,probability = T, xlim=c(0,6000), col="lightblue")
curve(dexp(x,rate=lamda),col="red",lwd=2,add=T)

hist(my_skew,breaks=30,probability = T,col="lightgreen") ;
curve(dexp(x,rate=1/mean(my_skew)),col="blue",lwd=2,add=T)
```





The plots do look rather similar, excepting for a smaller density in the second bucket of the empirical distribution.

Using the exponential pdf, *find the 5th and 95th percentiles* using the cumulative distribution function (CDF).

```
qexp(c(.05, .95), rate = lamda)
```

```
## [1]   22.8070364 1332.0215774
```

The 5th percentile of the exponential distribution, at this lambda, is 22.807, while the 95th percentile is 1332.

Also *generate a 95% confidence interval* from the empirical data, assuming normality.

Assuming that the original data is normally distributed (which it clearly is not), to generate a 95 percent confidence interval we look to the tails of 0.025 and 0.975.

```
my_mean = mean(my_skew)
my_mean
```

```
## [1] 444.639726
```

```
my_sd = sd(my_skew)
my_sd
```

```
## [1] 456.098091
```

```
qnorm(c(.025, .975), mean=my_mean, sd=my_sd)
```

```
## [1] -449.296105 1338.575557
```

This indicates that the 95% confidence interval for the data (**not** for the **mean**) is (-.449.496, 1338.576) . Of course, it doesn't make sense to have negative values, so the left tail would have to be truncated at zero. This indicates that a Normal distribution is not appropriate for this data.

Finally, provide the *empirical 5th percentile and 95th percentile* of the data.

```
quantile(x = my_skew,probs = c(0.05,0.95))
```

```
##   5%  95%
##    1 1275
```

Because we'ce shifted the data up by 1, the empirical quantiles are (1,1275).
(We would subtract 1 to reflect the original data.)

Discuss.

Because there are so many zero values in the dataset (e.g., some houses do not have a basement; others may have a basement but it may be "unfinshed", i.e, not built up like a living space), the square footage measured for such houses is zero.

This would suggest that a more appropriate model could be something which handles "Zero-Inflated" cases, e.g., Zero-Inflated Poisson; Zero-Inflated Negative Binomial, etc.

---

10 points. Modeling.

Build some type of multiple regression model and submit your model to the competition board.

We will implement forward and backward stepwise regression to select features to determine the "best" model, where the criteria used here is minimizing the Akaike Information Criterion (AIC). (Because the stepwise algorithm is "greedy", it is possible that the forward and backward algorithms may not converge onto the same model, especially in the case of a large number of potential features, as we have here.)

Create full and null models, for starting points of the stepwise regressions

Null model (intercept only)

```
library(MASS)
lm_Null    <- lm(formula = SalePrice ~ 1, data = train.df.Boruta)
lm_Null_sum <- summary(lm_Null)
lm_Null_sum
```

```
##
## Call:
## lm(formula = SalePrice ~ 1, data = train.df.Boruta)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.5638088 -0.2489536 -0.0225454  0.2496804  1.5104221
##
## Coefficients:
##               Estimate Std. Error t value               Pr(>|t|)
## (Intercept) 12.0240509  0.0104541 1150.17 < 0.000000000000000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.399452 on 1459 degrees of freedom
```

The above model has only an intercept; it does not yield sufficient diagnostics to compute R^2.
The standard error of the residuals is $\sigma = 0.399451868$, which is quite high.

Full model (all variables)

```
lm_Boruta1 <- lm(formula = SalePrice ~ ., data = train.df.Boruta)
lm_Boruta1_sum <- summary(lm_Boruta1)
lm_Boruta1_sum
```

```
##
## Call:
## lm(formula = SalePrice ~ ., data = train.df.Boruta)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.1308273 -0.0677725  0.0065547  0.0794669  0.4669958
##
```

```
## Coefficients:
##                     Estimate      Std. Error  t value               Pr(>|t|)
## (Intercept)   3.282063268629  0.587086263881  5.59043        0.0000000270807
## ScreenPorch   0.000321375962  0.000073015506  4.40148        0.0000115484354
## WoodDeckSF    0.000105599969  0.000034173322  3.09013             0.00203921
## LotFrontage  -0.000055217608  0.000120652332 -0.45766             0.64726693
## KitchenAbvGr -0.100113598526  0.020802898882 -4.81248        0.0000016482116
## MasVnrArea   -0.000002027470  0.000025420689 -0.07976             0.93644191
## BsmtFullBath  0.055912650430  0.010646733627  5.25162        0.0000001734915
## OpenPorchSF  -0.000037275769  0.000064752796 -0.57566             0.56493357
## HalfBath      0.023469431708  0.011419622896  2.05518             0.04004210
## BedroomAbvGr  0.003350907359  0.007184106113  0.46643             0.64097617
## BsmtUnfSF    -0.000010435339  0.000026363139 -0.39583             0.69228886
## TotRmsAbvGrd  0.019327963741  0.005255515593  3.67765             0.00024409
## FullBath      0.037031594460  0.012054124411  3.07211             0.00216556
## GarageCars    0.070371127364  0.012225088179  5.75629        0.0000000105013
## GarageYrBlt  -0.000106801000  0.000327452891 -0.32616             0.74435332
## YearRemodAdd  0.001120324351  0.000287820538  3.89244             0.00010380
## GarageArea    0.000041716049  0.000043301033  0.96340             0.33551129
## OverallCond   0.049055904838  0.004351551820 11.27320 < 0.000000000000000222
## Fireplaces    0.044845376983  0.007640816403  5.86919        0.0000000054325
## X2ndFlrSF     0.000030556294  0.000085385108  0.35786             0.72049746
## YearBuilt     0.002675457665  0.000306540473  8.72791 < 0.000000000000000222
## BsmtFinSF1    0.000009743230  0.000025995329  0.37481             0.70785950
## LotArea       0.000002034571  0.000000430716  4.72369        0.0000025431632
## X1stFlrSF     0.000088541513  0.000086460381  1.02407             0.30597489
## TotalBsmtSF   0.000072081653  0.000030077177  2.39656             0.01667744
## OverallQual   0.083185689249  0.005043727398 16.49290 < 0.000000000000000222
## GrLivArea     0.000106984560  0.000084431427  1.26712             0.20531908
##
## (Intercept)  ***
## ScreenPorch  ***
## WoodDeckSF   **
## LotFrontage
## KitchenAbvGr ***
## MasVnrArea
## BsmtFullBath ***
## OpenPorchSF
## HalfBath     *
## BedroomAbvGr
## BsmtUnfSF
## TotRmsAbvGrd ***
## FullBath     **
## GarageCars   ***
## GarageYrBlt
## YearRemodAdd ***
## GarageArea
## OverallCond  ***
## Fireplaces   ***
## X2ndFlrSF
## YearBuilt    ***
## BsmtFinSF1
## LotArea      ***
## X1stFlrSF
```

```
## TotalBsmtSF  *
## OverallQual  ***
## GrLivArea
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.149421 on 1433 degrees of freedom
## Multiple R-squared:  0.862568,   Adjusted R-squared:  0.860075
## F-statistic: 345.923 on 26 and 1433 DF,  p-value: < 0.000000000000000222
```

The full model gives $R^2 = 0.862568245$ and adj-$R^2 = 0.860074717$, which seem rather good.
The standard error of the residuals is $\sigma = 0.149421315$, which is considerably better.
The model incorporates a large number of variables (25).

Interestingly, with all the other variables included, the "GrLivingArea" variable (which was measured as "most important" by the Boruta algorithm) is not significant here, as the p-value is quite large (see the final entry in the above table.)

Forward Stepwise

Forward stepwise regression starts from an empty model (here, just the intercept) and adds variables one-by-one when doing so improves the AIC.
The algorithm ends when no more variables can be added which would improve the AIC.

```
stepforward <- stepAIC(lm_Null,
                       direction="both",
                       scope=list(upper=lm_Boruta1,
                                  lower=lm_Null))
```

```
## Start:  AIC=-2678.57
## SalePrice ~ 1
##
##                 Df Sum of Sq       RSS       AIC
## + OverallQual   1 155.46204  77.33862 -4285.477
## + GrLivArea     1 114.37454 118.42612 -3663.378
## + GarageCars    1 107.84494 124.95572 -3585.019
## + GarageArea    1  98.62707 134.17359 -3481.104
## + TotalBsmtSF   1  87.23227 145.56839 -3362.097
## + X1stFlrSF     1  82.96698 149.83368 -3319.932
## + FullBath      1  82.35370 150.44696 -3313.968
## + YearBuilt     1  80.09848 152.70218 -3292.245
## + GarageYrBlt   1  75.29120 157.50946 -3246.991
## + YearRemodAdd  1  74.47578 158.32488 -3239.452
## + TotRmsAbvGrd  1  66.48952 166.31114 -3167.604
## + Fireplaces    1  55.76990 177.03076 -3076.408
## + MasVnrArea    1  42.40162 190.39904 -2970.122
## + BsmtFinSF1    1  32.21988 200.58078 -2894.063
## + WoodDeckSF    1  25.99132 206.80934 -2849.416
## + OpenPorchSF   1  23.99593 208.80473 -2835.397
## + X2ndFlrSF     1  23.73460 209.06605 -2833.571
## + HalfBath      1  22.95062 209.85004 -2828.106
## + LotArea       1  15.41455 217.38611 -2776.594
## + BsmtFullBath  1  12.99070 219.80996 -2760.405
## + BsmtUnfSF     1  11.47180 221.32886 -2750.351
## + BedroomAbvGr  1  10.17322 222.62744 -2741.810
## + LotFrontage   1   7.48443 225.31623 -2724.283
## + KitchenAbvGr  1   5.06817 227.73249 -2708.709
## + ScreenPorch   1   3.42014 229.38052 -2698.182
## <none>                      232.80066 -2678.573
## + OverallCond   1   0.31643 232.48423 -2678.559
##
## Step:  AIC=-4285.48
## SalePrice ~ OverallQual
##
##                 Df Sum of Sq       RSS       AIC
## + GrLivArea     1  16.80406  60.53456 -4641.147
## + GarageCars    1  13.11560  64.22302 -4554.792
## + X1stFlrSF     1  13.00331  64.33531 -4552.241
## + GarageArea    1  12.49369  64.84493 -4540.722
## + TotalBsmtSF   1   9.76268  67.57594 -4480.492
## + TotRmsAbvGrd  1   9.76098  67.57764 -4480.455
## + BsmtFinSF1    1   7.66563  69.67299 -4435.873
```

```
## + Fireplaces    1   7.54216  69.79646 -4433.288
## + FullBath      1   7.00749  70.33113 -4422.147
## + LotArea       1   6.87289  70.46573 -4419.355
## + BsmtFullBath  1   4.98570  72.35292 -4380.768
## + YearBuilt     1   4.89240  72.44622 -4378.887
## + WoodDeckSF    1   4.76274  72.57588 -4376.276
## + YearRemodAdd  1   4.46484  72.87378 -4370.296
## + GarageYrBlt   1   4.46032  72.87830 -4370.205
## + BedroomAbvGr  1   3.73190  73.60672 -4355.685
## + MasVnrArea    1   2.46472  74.87390 -4330.764
## + HalfBath      1   2.06155  75.27707 -4322.924
## + X2ndFlrSF     1   1.54502  75.79360 -4312.940
## + OpenPorchSF   1   1.21426  76.12436 -4306.582
## + ScreenPorch   1   1.08686  76.25176 -4304.141
## + OverallCond   1   0.34364  76.99498 -4289.979
## + LotFrontage   1   0.29469  77.04393 -4289.051
## + BsmtUnfSF     1   0.22901  77.10961 -4287.807
## <none>                       77.33862 -4285.477
## + KitchenAbvGr  1   0.00178  77.33684 -4283.511
## - OverallQual   1 155.46204 232.80066 -2678.573
##
## Step:  AIC=-4641.15
## SalePrice ~ OverallQual + GrLivArea
##
##                Df Sum of Sq       RSS       AIC
## + YearBuilt     1   9.95757  50.57699 -4901.535
## + GarageCars    1   8.75442  51.78014 -4867.210
## + GarageArea    1   7.60192  52.93264 -4835.070
## + GarageYrBlt   1   7.35704  53.17752 -4828.332
## + BsmtFinSF1    1   5.90994  54.62462 -4789.132
## + BsmtFullBath  1   5.72985  54.80471 -4784.327
## + YearRemodAdd  1   5.55041  54.98415 -4779.554
## + TotalBsmtSF   1   5.52831  55.00624 -4778.968
## + X1stFlrSF     1   4.59425  55.94030 -4754.384
## + X2ndFlrSF     1   3.97270  56.56186 -4738.251
## + LotArea       1   2.71658  57.81797 -4706.182
## + WoodDeckSF    1   2.69957  57.83498 -4705.753
## + Fireplaces    1   2.45213  58.08243 -4699.520
## + KitchenAbvGr  1   1.16467  59.36989 -4667.511
## + FullBath      1   0.79627  59.73828 -4658.479
## + BsmtUnfSF     1   0.62184  59.91272 -4654.222
## + MasVnrArea    1   0.59053  59.94402 -4653.460
## + ScreenPorch   1   0.52299  60.01157 -4651.815
## + OverallCond   1   0.51158  60.02298 -4651.538
## + BedroomAbvGr  1   0.27292  60.26163 -4645.744
## + OpenPorchSF   1   0.10277  60.43179 -4641.628
## <none>                       60.53456 -4641.147
## + TotRmsAbvGrd  1   0.02458  60.50998 -4639.740
## + HalfBath      1   0.00980  60.52476 -4639.383
## + LotFrontage   1   0.00312  60.53143 -4639.222
## - GrLivArea     1  16.80406  77.33862 -4285.477
## - OverallQual   1  57.89156 118.42612 -3663.378
##
## Step:  AIC=-4901.53
```

```
## SalePrice ~ OverallQual + GrLivArea + YearBuilt
##
##                 Df Sum of Sq      RSS       AIC
## + OverallCond   1  4.898304 45.67869 -5048.257
## + GarageCars    1  3.963084 46.61391 -5018.667
## + BsmtFullBath  1  3.801978 46.77501 -5013.630
## + BsmtFinSF1    1  3.769772 46.80722 -5012.625
## + GarageArea    1  3.728421 46.84857 -5011.336
## + TotalBsmtSF   1  3.382309 47.19468 -5000.589
## + X1stFlrSF     1  3.269107 47.30788 -4997.092
## + X2ndFlrSF     1  3.008985 47.56801 -4989.086
## + Fireplaces    1  2.904549 47.67244 -4985.884
## + LotArea       1  2.756435 47.82056 -4981.355
## + WoodDeckSF    1  1.442559 49.13443 -4941.782
## + YearRemodAdd  1  1.402293 49.17470 -4940.586
## + ScreenPorch   1  1.040243 49.53675 -4929.876
## + KitchenAbvGr  1  0.958092 49.61890 -4927.457
## + BsmtUnfSF     1  0.528484 50.04851 -4914.871
## + HalfBath      1  0.272638 50.30435 -4907.426
## + GarageYrBlt   1  0.201654 50.37534 -4905.367
## + BedroomAbvGr  1  0.146136 50.43085 -4903.759
## + MasVnrArea    1  0.082222 50.49477 -4901.910
## + FullBath      1  0.081061 50.49593 -4901.877
## <none>                      50.57699 -4901.535
## + OpenPorchSF   1  0.018509 50.55848 -4900.069
## + LotFrontage   1  0.010526 50.56646 -4899.839
## + TotRmsAbvGrd  1  0.000430 50.57656 -4899.547
## - YearBuilt     1  9.957566 60.53456 -4641.147
## - OverallQual   1 19.398118 69.97511 -4429.556
## - GrLivArea     1 21.869229 72.44622 -4378.887
##
## Step:  AIC=-5048.26
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond
##
##                 Df Sum of Sq      RSS       AIC
## + GarageCars    1  4.158290 41.52040 -5185.610
## + TotalBsmtSF   1  4.083610 41.59508 -5182.986
## + X1stFlrSF     1  3.901811 41.77688 -5176.619
## + GarageArea    1  3.716021 41.96267 -5170.141
## + X2ndFlrSF     1  3.677462 42.00123 -5168.800
## + BsmtFullBath  1  3.667137 42.01155 -5168.441
## + BsmtFinSF1    1  3.361125 42.31756 -5157.845
## + Fireplaces    1  2.848699 42.82999 -5140.272
## + LotArea       1  2.670263 43.00842 -5134.202
## + WoodDeckSF    1  1.002513 44.67617 -5078.657
## + ScreenPorch   1  0.917219 44.76147 -5075.872
## + KitchenAbvGr  1  0.497952 45.18074 -5062.261
## + HalfBath      1  0.423023 45.25566 -5059.841
## + BedroomAbvGr  1  0.210400 45.46829 -5052.998
## + BsmtUnfSF     1  0.191599 45.48709 -5052.394
## + GarageYrBlt   1  0.161348 45.51734 -5051.424
## + YearRemodAdd  1  0.142536 45.53615 -5050.820
## + MasVnrArea    1  0.128861 45.54983 -5050.382
## <none>                      45.67869 -5048.257
```

```
## + LotFrontage   1  0.055465 45.62322 -5048.031
## + FullBath      1  0.046469 45.63222 -5047.743
## + OpenPorchSF   1  0.005383 45.67330 -5046.430
## + TotRmsAbvGrd  1  0.003605 45.67508 -5046.373
## - OverallCond   1  4.898304 50.57699 -4901.535
## - YearBuilt     1 14.344293 60.02298 -4651.538
## - OverallQual   1 14.852571 60.53126 -4639.226
## - GrLivArea     1 24.232020 69.91071 -4428.900
##
## Step:  AIC=-5185.61
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars
##
##                Df Sum of Sq      RSS       AIC
## + TotalBsmtSF   1  3.463060 38.05734 -5310.763
## + BsmtFullBath  1  3.348454 38.17194 -5306.373
## + BsmtFinSF1    1  3.089798 38.43060 -5296.513
## + X1stFlrSF     1  2.943442 38.57696 -5290.963
## + X2ndFlrSF     1  2.823431 38.69697 -5286.429
## + Fireplaces    1  2.503370 39.01703 -5274.403
## + LotArea       1  2.167420 39.35298 -5261.885
## + WoodDeckSF    1  0.845160 40.67524 -5213.635
## + ScreenPorch   1  0.783371 40.73703 -5211.419
## + KitchenAbvGr  1  0.632839 40.88756 -5206.034
## + HalfBath      1  0.292416 41.22798 -5193.929
## + BsmtUnfSF     1  0.253353 41.26704 -5192.546
## + GarageArea    1  0.244450 41.27595 -5192.231
## + YearRemodAdd  1  0.115384 41.40501 -5187.673
## + BedroomAbvGr  1  0.096924 41.42347 -5187.022
## + GarageYrBlt   1  0.089378 41.43102 -5186.756
## + FullBath      1  0.069962 41.45044 -5186.072
## <none>                     41.52040 -5185.610
## + MasVnrArea    1  0.036120 41.48428 -5184.881
## + OpenPorchSF   1  0.011597 41.50880 -5184.018
## + LotFrontage   1  0.007858 41.51254 -5183.886
## + TotRmsAbvGrd  1  0.000357 41.52004 -5183.623
## - GarageCars    1  4.158290 45.67869 -5048.257
## - OverallCond   1  5.093509 46.61391 -5018.667
## - YearBuilt     1  8.801675 50.32207 -4906.912
## - OverallQual   1 10.844211 52.36461 -4848.823
## - GrLivArea     1 18.179797 59.70019 -4657.410
##
## Step:  AIC=-5310.76
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF
##
##                Df Sum of Sq      RSS       AIC
## + Fireplaces    1  1.894910 36.16243 -5383.330
## + BsmtFullBath  1  1.809178 36.24816 -5379.873
## + BsmtUnfSF     1  1.253284 36.80405 -5357.652
## + LotArea       1  1.246988 36.81035 -5357.403
## + BsmtFinSF1    1  1.016557 37.04078 -5348.291
## + ScreenPorch   1  0.597412 37.45993 -5331.863
## + WoodDeckSF    1  0.584247 37.47309 -5331.350
```

```
## + KitchenAbvGr  1  0.557704 37.49963 -5330.316
## + X1stFlrSF     1  0.232047 37.82529 -5317.692
## + X2ndFlrSF     1  0.194805 37.86253 -5316.255
## + YearRemodAdd  1  0.182271 37.87507 -5315.772
## + TotRmsAbvGrd  1  0.056977 38.00036 -5310.950
## <none>                      38.05734 -5310.763
## + LotFrontage   1  0.037263 38.02007 -5310.193
## + HalfBath      1  0.026680 38.03066 -5309.787
## + GarageYrBlt   1  0.021221 38.03612 -5309.577
## + GarageArea    1  0.008172 38.04916 -5309.076
## + BedroomAbvGr  1  0.003133 38.05420 -5308.883
## + MasVnrArea    1  0.000907 38.05643 -5308.798
## + FullBath      1  0.000070 38.05727 -5308.766
## + OpenPorchSF   1  0.000019 38.05732 -5308.764
## - TotalBsmtSF   1  3.463060 41.52040 -5185.610
## - GarageCars    1  3.537740 41.59508 -5182.986
## - OverallCond   1  5.734460 43.79180 -5107.848
## - YearBuilt     1  7.719887 45.77722 -5043.111
## - OverallQual   1  7.895792 45.95313 -5037.512
## - GrLivArea     1 14.533262 52.59060 -4840.535
##
## Step:  AIC=-5383.33
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces
##
##                 Df Sum of Sq      RSS       AIC
## + BsmtFullBath  1  1.479699 34.68273 -5442.327
## + BsmtUnfSF     1  0.839812 35.32261 -5415.636
## + LotArea       1  0.808783 35.35364 -5414.354
## + BsmtFinSF1    1  0.681044 35.48138 -5409.088
## + WoodDeckSF    1  0.435629 35.72680 -5399.024
## + YearRemodAdd  1  0.371269 35.79116 -5396.397
## + ScreenPorch   1  0.333329 35.82910 -5394.850
## + KitchenAbvGr  1  0.295425 35.86700 -5393.306
## + TotRmsAbvGrd  1  0.121173 36.04125 -5386.230
## + X1stFlrSF     1  0.088460 36.07397 -5384.906
## + X2ndFlrSF     1  0.079617 36.08281 -5384.548
## <none>                      36.16243 -5383.330
## + GarageArea    1  0.040196 36.12223 -5382.954
## + FullBath      1  0.018170 36.14426 -5382.064
## + BedroomAbvGr  1  0.011221 36.15121 -5381.783
## + MasVnrArea    1  0.005934 36.15649 -5381.569
## + GarageYrBlt   1  0.005225 36.15720 -5381.541
## + HalfBath      1  0.004925 36.15750 -5381.529
## + LotFrontage   1  0.002590 36.15984 -5381.434
## + OpenPorchSF   1  0.000071 36.16236 -5381.333
## - Fireplaces    1  1.894910 38.05734 -5310.763
## - TotalBsmtSF   1  2.854601 39.01703 -5274.403
## - GarageCars    1  3.308949 39.47138 -5257.499
## - OverallCond   1  5.616727 41.77915 -5174.540
## - OverallQual   1  6.903187 43.06561 -5130.262
## - YearBuilt     1  8.171308 44.33373 -5087.891
## - GrLivArea     1 11.247458 47.40988 -4989.947
##
```

```
## Step:  AIC=-5442.33
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath
##
##                 Df Sum of Sq      RSS       AIC
## + LotArea        1  0.618779 34.06395 -5466.610
## + ScreenPorch    1  0.333410 34.34932 -5454.430
## + YearRemodAdd   1  0.327650 34.35508 -5454.185
## + KitchenAbvGr   1  0.316786 34.36594 -5453.724
## + WoodDeckSF     1  0.274410 34.40832 -5451.924
## + TotRmsAbvGrd   1  0.205075 34.47765 -5448.985
## + FullBath       1  0.133789 34.54894 -5445.970
## + X1stFlrSF      1  0.071452 34.61128 -5443.338
## + BedroomAbvGr   1  0.066731 34.61600 -5443.139
## + X2ndFlrSF      1  0.062734 34.61999 -5442.970
## <none>                       34.68273 -5442.327
## + BsmtUnfSF      1  0.037815 34.64491 -5441.920
## + BsmtFinSF1     1  0.016554 34.66617 -5441.024
## + GarageYrBlt    1  0.009977 34.67275 -5440.747
## + GarageArea     1  0.008701 34.67403 -5440.693
## + MasVnrArea     1  0.003761 34.67897 -5440.485
## + HalfBath       1  0.002022 34.68071 -5440.412
## + OpenPorchSF    1  0.000239 34.68249 -5440.337
## + LotFrontage    1  0.000072 34.68266 -5440.330
## - BsmtFullBath   1  1.479699 36.16243 -5383.330
## - Fireplaces     1  1.565431 36.24816 -5379.873
## - TotalBsmtSF    1  1.635093 36.31782 -5377.069
## - GarageCars     1  3.240795 37.92352 -5313.905
## - OverallCond    1  5.381631 40.06436 -5233.729
## - YearBuilt      1  7.377411 42.06014 -5162.753
## - OverallQual    1  7.395073 42.07780 -5162.140
## - GrLivArea      1 11.946087 46.62881 -5012.201
##
## Step:  AIC=-5466.61
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath + LotArea
##
##                 Df Sum of Sq      RSS       AIC
## + ScreenPorch    1  0.355295 33.70865 -5479.918
## + YearRemodAdd   1  0.351570 33.71238 -5479.757
## + KitchenAbvGr   1  0.278300 33.78565 -5476.587
## + TotRmsAbvGrd   1  0.216240 33.84771 -5473.908
## + WoodDeckSF     1  0.210132 33.85382 -5473.644
## + FullBath       1  0.120400 33.94355 -5469.780
## + BedroomAbvGr   1  0.058191 34.00576 -5467.106
## + X1stFlrSF      1  0.050857 34.01309 -5466.791
## <none>                       34.06395 -5466.610
## + X2ndFlrSF      1  0.045248 34.01870 -5466.551
## + BsmtUnfSF      1  0.021616 34.04233 -5465.537
## + GarageYrBlt    1  0.016613 34.04734 -5465.322
## + BsmtFinSF1     1  0.011532 34.05242 -5465.104
## + HalfBath       1  0.006596 34.05735 -5464.893
## + GarageArea     1  0.004655 34.05929 -5464.810
## + MasVnrArea     1  0.002269 34.06168 -5464.707
```

```
## + LotFrontage   1  0.001340 34.06261 -5464.668
## + OpenPorchSF   1  0.000151 34.06380 -5464.617
## - LotArea       1  0.618779 34.68273 -5442.327
## - Fireplaces    1  1.234603 35.29855 -5416.631
## - BsmtFullBath  1  1.289696 35.35364 -5414.354
## - TotalBsmtSF   1  1.304406 35.36835 -5413.746
## - GarageCars    1  3.065448 37.12940 -5342.803
## - OverallCond   1  5.292031 39.35598 -5257.774
## - YearBuilt     1  7.552467 41.61641 -5176.238
## - OverallQual   1  7.823525 41.88747 -5166.759
## - GrLivArea     1 10.921064 44.98501 -5062.599
##
## Step:  AIC=-5479.92
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath + LotArea +
##     ScreenPorch
##
##                 Df Sum of Sq      RSS       AIC
## + YearRemodAdd  1  0.388031 33.32062 -5494.822
## + WoodDeckSF    1  0.281373 33.42728 -5490.156
## + KitchenAbvGr  1  0.251885 33.45677 -5488.869
## + TotRmsAbvGrd  1  0.234012 33.47464 -5488.089
## + FullBath      1  0.141632 33.56702 -5484.066
## + X1stFlrSF     1  0.055951 33.65270 -5480.344
## + BedroomAbvGr  1  0.053969 33.65468 -5480.258
## + X2ndFlrSF     1  0.049047 33.65961 -5480.044
## <none>                      33.70865 -5479.918
## + GarageYrBlt   1  0.019965 33.68869 -5478.783
## + BsmtUnfSF     1  0.013383 33.69527 -5478.498
## + BsmtFinSF1    1  0.009888 33.69877 -5478.347
## + GarageArea    1  0.004550 33.70410 -5478.115
## + MasVnrArea    1  0.003928 33.70473 -5478.088
## + OpenPorchSF   1  0.001625 33.70703 -5477.989
## + HalfBath      1  0.001490 33.70716 -5477.983
## + LotFrontage   1  0.001399 33.70725 -5477.979
## - ScreenPorch   1  0.355295 34.06395 -5466.610
## - LotArea       1  0.640665 34.34932 -5454.430
## - Fireplaces    1  1.021814 34.73047 -5438.319
## - TotalBsmtSF   1  1.240402 34.94906 -5429.158
## - BsmtFullBath  1  1.286726 34.99538 -5427.224
## - GarageCars    1  3.009216 36.71787 -5357.075
## - OverallCond   1  5.199065 38.90772 -5272.499
## - YearBuilt     1  7.763780 41.47243 -5179.298
## - OverallQual   1  7.799240 41.50789 -5178.050
## - GrLivArea     1 10.909369 44.61802 -5072.559
##
## Step:  AIC=-5494.82
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath + LotArea +
##     ScreenPorch + YearRemodAdd
##
##                 Df Sum of Sq      RSS       AIC
## + WoodDeckSF    1  0.268370 33.05225 -5504.629
## + KitchenAbvGr  1  0.244229 33.07639 -5503.563
```

```
## + TotRmsAbvGrd  1  0.237447 33.08317 -5503.264
## + BedroomAbvGr  1  0.100956 33.21967 -5497.253
## + FullBath      1  0.088770 33.23185 -5496.717
## + X1stFlrSF     1  0.050462 33.27016 -5495.035
## <none>                      33.32062 -5494.822
## + X2ndFlrSF     1  0.041743 33.27888 -5494.652
## + BsmtUnfSF     1  0.040036 33.28059 -5494.578
## + BsmtFinSF1    1  0.028534 33.29209 -5494.073
## + GarageArea    1  0.010231 33.31039 -5493.271
## + OpenPorchSF   1  0.006453 33.31417 -5493.105
## + HalfBath      1  0.005093 33.31553 -5493.045
## + LotFrontage   1  0.002934 33.31769 -5492.951
## + GarageYrBlt   1  0.000446 33.32018 -5492.842
## + MasVnrArea    1  0.000097 33.32052 -5492.826
## - YearRemodAdd  1  0.388031 33.70865 -5479.918
## - ScreenPorch   1  0.391757 33.71238 -5479.757
## - LotArea       1  0.667413 33.98803 -5467.867
## - Fireplaces    1  1.159795 34.48042 -5446.868
## - BsmtFullBath  1  1.239066 34.55969 -5443.516
## - TotalBsmtSF   1  1.284458 34.60508 -5441.599
## - GarageCars    1  2.940884 36.26151 -5373.335
## - OverallCond   1  3.510397 36.83102 -5350.583
## - YearBuilt     1  4.575318 37.89594 -5308.968
## - OverallQual   1  6.881567 40.20219 -5222.715
## - GrLivArea     1 10.261025 43.58165 -5104.871
##
## Step:  AIC=-5504.63
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath + LotArea +
##     ScreenPorch + YearRemodAdd + WoodDeckSF
##
##                 Df Sum of Sq      RSS       AIC
## + TotRmsAbvGrd  1  0.249451 32.80280 -5513.690
## + KitchenAbvGr  1  0.212197 32.84005 -5512.032
## + BedroomAbvGr  1  0.110093 32.94216 -5507.500
## + FullBath      1  0.093831 32.95842 -5506.780
## + X1stFlrSF     1  0.047366 33.00489 -5504.723
## <none>                      33.05225 -5504.629
## + X2ndFlrSF     1  0.039191 33.01306 -5504.361
## + BsmtUnfSF     1  0.030622 33.02163 -5503.982
## + BsmtFinSF1    1  0.025224 33.02703 -5503.744
## + GarageArea    1  0.009848 33.04240 -5503.064
## + HalfBath      1  0.004608 33.04764 -5502.833
## + OpenPorchSF   1  0.002497 33.04976 -5502.739
## + GarageYrBlt   1  0.000319 33.05193 -5502.643
## + LotFrontage   1  0.000131 33.05212 -5502.635
## + MasVnrArea    1  0.000000 33.05225 -5502.629
## - WoodDeckSF    1  0.268370 33.32062 -5494.822
## - YearRemodAdd  1  0.375029 33.42728 -5490.156
## - ScreenPorch   1  0.463531 33.51578 -5486.296
## - LotArea       1  0.594028 33.64628 -5480.622
## - Fireplaces    1  1.076554 34.12881 -5459.833
## - BsmtFullBath  1  1.105366 34.15762 -5458.601
## - TotalBsmtSF   1  1.244391 34.29664 -5452.671
```

```
## - GarageCars    1  2.892235 35.94449 -5384.155
## - OverallCond   1  3.328716 36.38097 -5366.533
## - YearBuilt     1  4.298219 37.35047 -5328.135
## - OverallQual   1  6.941529 39.99378 -5228.303
## - GrLivArea     1  9.754316 42.80657 -5129.070
##
## Step:  AIC=-5513.69
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath + LotArea +
##     ScreenPorch + YearRemodAdd + WoodDeckSF + TotRmsAbvGrd
##
##                Df Sum of Sq      RSS       AIC
## + KitchenAbvGr  1  0.384391 32.41841 -5528.899
## + BsmtUnfSF     1  0.062807 32.73999 -5514.488
## + FullBath      1  0.061936 32.74086 -5514.449
## + BsmtFinSF1    1  0.052107 32.75069 -5514.011
## <none>                      32.80280 -5513.690
## + X1stFlrSF     1  0.043651 32.75915 -5513.634
## + X2ndFlrSF     1  0.035815 32.76699 -5513.285
## + GarageArea    1  0.019681 32.78312 -5512.566
## + BedroomAbvGr  1  0.010219 32.79258 -5512.145
## + HalfBath      1  0.006531 32.79627 -5511.980
## + LotFrontage   1  0.003011 32.79979 -5511.824
## + OpenPorchSF   1  0.000900 32.80190 -5511.730
## + MasVnrArea    1  0.000167 32.80263 -5511.697
## + GarageYrBlt   1  0.000016 32.80278 -5511.690
## - TotRmsAbvGrd  1  0.249451 33.05225 -5504.629
## - WoodDeckSF    1  0.280374 33.08317 -5503.264
## - YearRemodAdd  1  0.378207 33.18101 -5498.953
## - ScreenPorch   1  0.486208 33.28901 -5494.208
## - LotArea       1  0.604949 33.40775 -5489.010
## - Fireplaces    1  1.135904 33.93870 -5465.988
## - BsmtFullBath  1  1.183435 33.98624 -5463.945
## - TotalBsmtSF   1  1.339949 34.14275 -5457.237
## - GarageCars    1  2.785800 35.58860 -5396.683
## - GrLivArea     1  2.826453 35.62925 -5395.016
## - OverallCond   1  3.357194 36.15999 -5373.428
## - YearBuilt     1  4.377788 37.18059 -5332.791
## - OverallQual   1  7.052452 39.85525 -5231.369
##
## Step:  AIC=-5528.9
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath + LotArea +
##     ScreenPorch + YearRemodAdd + WoodDeckSF + TotRmsAbvGrd +
##     KitchenAbvGr
##
##                Df Sum of Sq      RSS       AIC
## + FullBath      1  0.139000 32.27941 -5533.173
## + X1stFlrSF     1  0.102808 32.31560 -5531.537
## + X2ndFlrSF     1  0.082261 32.33615 -5530.609
## + BsmtUnfSF     1  0.054056 32.36435 -5529.336
## + BsmtFinSF1    1  0.049686 32.36872 -5529.139
## <none>                      32.41841 -5528.899
## + GarageArea    1  0.013086 32.40532 -5527.489
```

```
## + LotFrontage   1  0.004396 32.41401 -5527.097
## + OpenPorchSF   1  0.004318 32.41409 -5527.094
## + BedroomAbvGr  1  0.003802 32.41461 -5527.071
## + GarageYrBlt   1  0.000485 32.41792 -5526.921
## + MasVnrArea    1  0.000161 32.41825 -5526.907
## + HalfBath      1  0.000145 32.41826 -5526.906
## - WoodDeckSF    1  0.239632 32.65804 -5520.147
## - YearRemodAdd  1  0.370547 32.78896 -5514.306
## - KitchenAbvGr  1  0.384391 32.80280 -5513.690
## - TotRmsAbvGrd  1  0.421645 32.84005 -5512.032
## - ScreenPorch   1  0.447733 32.86614 -5510.873
## - LotArea       1  0.567137 32.98555 -5505.578
## - Fireplaces    1  0.966552 33.38496 -5488.006
## - BsmtFullBath  1  1.250327 33.66874 -5475.648
## - TotalBsmtSF   1  1.374884 33.79329 -5470.257
## - GrLivArea     1  2.768677 35.18709 -5411.248
## - GarageCars    1  2.873315 35.29172 -5406.913
## - OverallCond   1  3.089673 35.50808 -5397.990
## - YearBuilt     1  4.177219 36.59563 -5353.944
## - OverallQual   1  6.262381 38.68079 -5273.039
##
## Step:  AIC=-5533.17
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath + LotArea +
##     ScreenPorch + YearRemodAdd + WoodDeckSF + TotRmsAbvGrd +
##     KitchenAbvGr + FullBath
##
##                 Df Sum of Sq      RSS       AIC
## + X1stFlrSF      1  0.096853 32.18256 -5535.560
## + X2ndFlrSF      1  0.077794 32.20162 -5534.696
## + BsmtUnfSF      1  0.063465 32.21594 -5534.046
## + BsmtFinSF1     1  0.057777 32.22163 -5533.788
## <none>                       32.27941 -5533.173
## + HalfBath       1  0.023394 32.25602 -5532.231
## + GarageArea     1  0.020442 32.25897 -5532.098
## + OpenPorchSF    1  0.006052 32.27336 -5531.446
## + LotFrontage    1  0.002605 32.27680 -5531.291
## + MasVnrArea     1  0.001387 32.27802 -5531.235
## + GarageYrBlt    1  0.001213 32.27820 -5531.228
## + BedroomAbvGr   1  0.000172 32.27924 -5531.181
## - FullBath       1  0.139000 32.41841 -5528.899
## - WoodDeckSF     1  0.240200 32.51961 -5524.349
## - YearRemodAdd   1  0.305120 32.58453 -5521.437
## - TotRmsAbvGrd   1  0.389336 32.66875 -5517.668
## - KitchenAbvGr   1  0.461455 32.74086 -5514.449
## - ScreenPorch    1  0.463597 32.74301 -5514.353
## - LotArea        1  0.545335 32.82474 -5510.713
## - Fireplaces     1  0.991622 33.27103 -5490.997
## - BsmtFullBath   1  1.366856 33.64627 -5474.623
## - TotalBsmtSF    1  1.424184 33.70359 -5472.137
## - GrLivArea      1  2.202611 34.48202 -5438.800
## - GarageCars     1  2.839141 35.11855 -5412.095
## - OverallCond    1  3.155466 35.43488 -5399.003
## - YearBuilt      1  3.515172 35.79458 -5384.257
```

```
## - OverallQual   1  6.084449 38.36386 -5283.051
##
## Step:  AIC=-5535.56
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath + LotArea +
##     ScreenPorch + YearRemodAdd + WoodDeckSF + TotRmsAbvGrd +
##     KitchenAbvGr + FullBath + X1stFlrSF
##
##                Df Sum of Sq      RSS       AIC
## + HalfBath      1  0.097233 32.08532 -5537.978
## + BsmtUnfSF     1  0.055638 32.12692 -5536.086
## + BsmtFinSF1    1  0.052676 32.12988 -5535.952
## <none>                      32.18256 -5535.560
## + GarageArea    1  0.013691 32.16886 -5534.181
## + X2ndFlrSF     1  0.010569 32.17199 -5534.040
## + LotFrontage   1  0.004587 32.17797 -5533.768
## + OpenPorchSF   1  0.003417 32.17914 -5533.715
## + BedroomAbvGr  1  0.002213 32.18034 -5533.660
## + MasVnrArea    1  0.001458 32.18110 -5533.626
## + GarageYrBlt   1  0.001087 32.18147 -5533.609
## - X1stFlrSF     1  0.096853 32.27941 -5533.173
## - FullBath      1  0.133044 32.31560 -5531.537
## - WoodDeckSF    1  0.232434 32.41499 -5527.053
## - YearRemodAdd  1  0.299083 32.48164 -5524.054
## - TotalBsmtSF   1  0.350892 32.53345 -5521.728
## - TotRmsAbvGrd  1  0.401503 32.58406 -5519.458
## - ScreenPorch   1  0.466483 32.64904 -5516.549
## - LotArea       1  0.516653 32.69921 -5514.308
## - KitchenAbvGr  1  0.520289 32.70285 -5514.145
## - Fireplaces    1  0.881426 33.06398 -5498.111
## - BsmtFullBath  1  1.355761 33.53832 -5477.315
## - GrLivArea     1  1.975338 34.15789 -5450.589
## - GarageCars    1  2.714619 34.89717 -5419.327
## - OverallCond   1  3.166269 35.34883 -5400.553
## - YearBuilt     1  3.570599 35.75316 -5383.948
## - OverallQual   1  6.170681 38.35324 -5281.455
##
## Step:  AIC=-5537.98
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath + LotArea +
##     ScreenPorch + YearRemodAdd + WoodDeckSF + TotRmsAbvGrd +
##     KitchenAbvGr + FullBath + X1stFlrSF + HalfBath
##
##                Df Sum of Sq      RSS       AIC
## + BsmtUnfSF     1  0.050084 32.03524 -5538.259
## + BsmtFinSF1    1  0.049918 32.03540 -5538.251
## <none>                      32.08532 -5537.978
## + GarageArea    1  0.018946 32.06638 -5536.840
## + OpenPorchSF   1  0.006877 32.07845 -5536.291
## + X2ndFlrSF     1  0.004420 32.08090 -5536.179
## + LotFrontage   1  0.002880 32.08244 -5536.109
## + BedroomAbvGr  1  0.002708 32.08262 -5536.101
## + MasVnrArea    1  0.000766 32.08456 -5536.013
## + GarageYrBlt   1  0.000469 32.08485 -5535.999
```

```
## - HalfBath      1  0.097233 32.18256 -5535.560
## - X1stFlrSF     1  0.170693 32.25602 -5532.231
## - FullBath      1  0.206768 32.29209 -5530.599
## - WoodDeckSF    1  0.229434 32.31476 -5529.575
## - YearRemodAdd  1  0.298719 32.38404 -5526.448
## - TotalBsmtSF   1  0.349952 32.43527 -5524.140
## - TotRmsAbvGrd  1  0.394222 32.47955 -5522.149
## - ScreenPorch   1  0.440258 32.52558 -5520.081
## - LotArea       1  0.524299 32.60962 -5516.313
## - KitchenAbvGr  1  0.531608 32.61693 -5515.986
## - Fireplaces    1  0.816821 32.90214 -5503.275
## - GrLivArea     1  1.151876 33.23720 -5488.482
## - BsmtFullBath  1  1.382078 33.46740 -5478.405
## - GarageCars    1  2.680635 34.76596 -5422.827
## - YearBuilt     1  2.781148 34.86647 -5418.612
## - OverallCond   1  3.156496 35.24182 -5402.979
## - OverallQual   1  6.211230 38.29655 -5281.614
##
## Step:  AIC=-5538.26
## SalePrice ~ OverallQual + GrLivArea + YearBuilt + OverallCond +
##     GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath + LotArea +
##     ScreenPorch + YearRemodAdd + WoodDeckSF + TotRmsAbvGrd +
##     KitchenAbvGr + FullBath + X1stFlrSF + HalfBath + BsmtUnfSF
##
##                 Df Sum of Sq      RSS       AIC
## <none>                       32.03524 -5538.259
## - BsmtUnfSF     1  0.050084 32.08532 -5537.978
## + GarageArea    1  0.015238 32.02000 -5536.953
## + OpenPorchSF   1  0.005929 32.02931 -5536.529
## + BedroomAbvGr  1  0.004496 32.03074 -5536.463
## + X2ndFlrSF     1  0.003796 32.03144 -5536.432
## + BsmtFinSF1    1  0.003561 32.03168 -5536.421
## + LotFrontage   1  0.002771 32.03247 -5536.385
## + GarageYrBlt   1  0.000273 32.03497 -5536.271
## + MasVnrArea    1  0.000107 32.03513 -5536.263
## - HalfBath      1  0.091680 32.12692 -5536.086
## - X1stFlrSF     1  0.158345 32.19358 -5533.060
## - FullBath      1  0.213547 32.24879 -5530.558
## - WoodDeckSF    1  0.220034 32.25527 -5530.265
## - YearRemodAdd  1  0.325975 32.36121 -5525.477
## - TotalBsmtSF   1  0.395384 32.43062 -5522.349
## - TotRmsAbvGrd  1  0.424413 32.45965 -5521.043
## - ScreenPorch   1  0.426873 32.46211 -5520.932
## - LotArea       1  0.506351 32.54159 -5517.362
## - KitchenAbvGr  1  0.521862 32.55710 -5516.666
## - BsmtFullBath  1  0.625287 32.66053 -5512.036
## - Fireplaces    1  0.791233 32.82647 -5504.636
## - GrLivArea     1  1.089460 33.12470 -5491.432
## - YearBuilt     1  2.662371 34.69761 -5423.700
## - GarageCars    1  2.699709 34.73495 -5422.130
## - OverallCond   1  2.984066 35.01931 -5410.227
## - OverallQual   1  6.260774 38.29601 -5279.635
```

```r
stepforward_sum <- summary(stepforward)
stepforward_sum
```

```
##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea + YearBuilt +
##     OverallCond + GarageCars + TotalBsmtSF + Fireplaces + BsmtFullBath +
##     LotArea + ScreenPorch + YearRemodAdd + WoodDeckSF + TotRmsAbvGrd +
##     KitchenAbvGr + FullBath + X1stFlrSF + HalfBath + BsmtUnfSF,
##     data = train.df.Boruta)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.1228365 -0.0675088  0.0052979  0.0804568  0.4638405
##
## Coefficients:
##                     Estimate      Std. Error  t value               Pr(>|t|)
## (Intercept)   3.242967967491  0.551178309080  5.88370     0.0000000049810693
## OverallQual   0.082635011540  0.004924159248 16.78155 < 0.000000000000000222
## GrLivArea     0.000138689940  0.000019811681  7.00041     0.0000000000038998
## YearBuilt     0.002653533255  0.000242477931 10.94340 < 0.000000000000000222
## OverallCond   0.049703505611  0.004290074352 11.58570 < 0.000000000000000222
## GarageCars    0.079146922430  0.007182200493 11.01987 < 0.000000000000000222
## TotalBsmtSF   0.000081259449  0.000019268425  4.21723     0.0000262727716261
## Fireplaces    0.044529438548  0.007464091604  5.96582     0.0000000030582129
## BsmtFullBath  0.056050340454  0.010568675260  5.30344     0.0000001313750844
## LotArea       0.000002042779  0.000000428033  4.77248     0.0000020046455322
## ScreenPorch   0.000318029066  0.000072577112  4.38195     0.0000126139384168
## YearRemodAdd  0.001056153428  0.000275814246  3.82922             0.00013406
## WoodDeckSF    0.000106292100  0.000033786122  3.14603             0.00168889
## TotRmsAbvGrd  0.020111839040  0.004602985393  4.36930     0.0000133565886396
## KitchenAbvGr -0.099780697131  0.020594473679 -4.84502     0.0000014026515103
## FullBath      0.036608474616  0.011811808680  3.09931             0.00197725
## X1stFlrSF     0.000058463863  0.000021906188  2.66883             0.00769706
## HalfBath      0.022809299275  0.011232015913  2.03074             0.04246471
## BsmtUnfSF    -0.000019830857  0.000013212217 -1.50095             0.13358796
##
## (Intercept)  ***
## OverallQual  ***
## GrLivArea    ***
## YearBuilt    ***
## OverallCond  ***
## GarageCars   ***
## TotalBsmtSF  ***
## Fireplaces   ***
## BsmtFullBath ***
## LotArea      ***
## ScreenPorch  ***
## YearRemodAdd ***
## WoodDeckSF   **
## TotRmsAbvGrd ***
## KitchenAbvGr ***
## FullBath     **
```

```
## X1stFlrSF    **
## HalfBath     *
## BsmtUnfSF
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.149101 on 1441 degrees of freedom
## Multiple R-squared:  0.862392,   Adjusted R-squared:  0.860673
## F-statistic:  501.71 on 18 and 1441 DF,  p-value: < 0.000000000000000222
```

The ***forward stepwise*** algorithm starts from an empty model with an AIC of -2678.57 and, at each step, adds the locally "best" unused variable into the model.

The first such entry is `OverallQual`, which is not surprising. It reduces the AIC to -4285.48 .

The next variable to enter is `GrLivArea`, which is also not surprising, as these two variables were confirmed as "most important" by Boruta. Adding this variable reduces the AIC to -4641.15 .

Successively adding a total of 18 variables reduces the AIC to -5538.26 , yielding an $R^2 =$ '$rstepforward_sum$r.squared'$ and an adj-$R^2 = 0.860673066$.

The standard error of the residuals has been reduced to $\sigma = 0.149101494$.

The variables selected under forward stepwise include:

1. OverallQual
2. GrLivArea
3. YearBuilt
4. OverallCond
5. GarageCars
6. TotalBsmtSF
7. Fireplaces
8. BsmtFullBath
9. LotArea
10. ScreenPorch
11. YearRemodAdd
12. WoodDeckSF
13. TotRmsAbvGrd
14. KitchenAbvGr
15. FullBath
16. X1stFlrSF
17. HalfBath
18. BsmtUnfSF

Backward Stepwise

Backward stepwise regression starts from a full model and deletes variables when doing so improves the AIC. The algorithm ends when no more variables can be deleted which would improve the AIC.

```
stepbackward <- stepAIC(lm_Boruta1,
                        direction="both",
                        scope=list(upper=lm_Boruta1,
                                   lower=lm_Null))
```

```
## Start:  AIC=-5524.13
## SalePrice ~ ScreenPorch + WoodDeckSF + LotFrontage + KitchenAbvGr +
##     MasVnrArea + BsmtFullBath + OpenPorchSF + HalfBath + BedroomAbvGr +
##     BsmtUnfSF + TotRmsAbvGrd + FullBath + GarageCars + GarageYrBlt +
##     YearRemodAdd + GarageArea + OverallCond + Fireplaces + X2ndFlrSF +
##     YearBuilt + BsmtFinSF1 + LotArea + X1stFlrSF + TotalBsmtSF +
##     OverallQual + GrLivArea
##
##                Df Sum of Sq      RSS       AIC
## - MasVnrArea    1  0.000142 31.99435 -5526.124
## - GarageYrBlt   1  0.002375 31.99658 -5526.022
## - X2ndFlrSF     1  0.002859 31.99706 -5526.000
## - BsmtFinSF1    1  0.003136 31.99734 -5525.987
## - BsmtUnfSF     1  0.003498 31.99770 -5525.970
## - LotFrontage   1  0.004676 31.99888 -5525.917
## - BedroomAbvGr  1  0.004857 31.99906 -5525.908
## - OpenPorchSF   1  0.007399 32.00160 -5525.792
## - GarageArea    1  0.020722 32.01493 -5525.185
## - X1stFlrSF     1  0.023414 32.01762 -5525.062
## - GrLivArea     1  0.035848 32.03005 -5524.495
## <none>                      31.99420 -5524.130
## - HalfBath      1  0.094303 32.08851 -5521.833
## - TotalBsmtSF   1  0.128233 32.12244 -5520.290
## - FullBath      1  0.210717 32.20492 -5516.546
## - WoodDeckSF    1  0.213196 32.20740 -5516.433
## - TotRmsAbvGrd  1  0.301972 32.29618 -5512.415
## - YearRemodAdd  1  0.338274 32.33248 -5510.774
## - ScreenPorch   1  0.432535 32.42674 -5506.524
## - LotArea       1  0.498182 32.49239 -5503.571
## - KitchenAbvGr  1  0.517087 32.51129 -5502.722
## - BsmtFullBath  1  0.615761 32.60996 -5498.298
## - GarageCars    1  0.739793 32.73400 -5492.755
## - Fireplaces    1  0.769097 32.76330 -5491.449
## - YearBuilt     1  1.700770 33.69497 -5450.511
## - OverallCond   1  2.837392 34.83160 -5402.074
## - OverallQual   1  6.073222 38.06742 -5272.376
##
## Step:  AIC=-5526.12
## SalePrice ~ ScreenPorch + WoodDeckSF + LotFrontage + KitchenAbvGr +
##     BsmtFullBath + OpenPorchSF + HalfBath + BedroomAbvGr + BsmtUnfSF +
##     TotRmsAbvGrd + FullBath + GarageCars + GarageYrBlt + YearRemodAdd +
##     GarageArea + OverallCond + Fireplaces + X2ndFlrSF + YearBuilt +
##     BsmtFinSF1 + LotArea + X1stFlrSF + TotalBsmtSF + OverallQual +
##     GrLivArea
```

```
##
##                 Df Sum of Sq      RSS       AIC
## - GarageYrBlt   1  0.002313 31.99666 -5528.018
## - X2ndFlrSF     1  0.002797 31.99714 -5527.996
## - BsmtFinSF1    1  0.003030 31.99738 -5527.985
## - BsmtUnfSF     1  0.003592 31.99794 -5527.960
## - LotFrontage   1  0.004665 31.99901 -5527.911
## - BedroomAbvGr  1  0.004900 31.99925 -5527.900
## - OpenPorchSF   1  0.007306 32.00165 -5527.790
## - GarageArea    1  0.020587 32.01493 -5527.184
## - X1stFlrSF     1  0.023285 32.01763 -5527.061
## - GrLivArea     1  0.036025 32.03037 -5526.480
## <none>                      31.99435 -5526.124
## + MasVnrArea    1  0.000142 31.99420 -5524.130
## - HalfBath      1  0.094164 32.08851 -5523.833
## - TotalBsmtSF   1  0.128436 32.12278 -5522.274
## - FullBath      1  0.211341 32.20569 -5518.511
## - WoodDeckSF    1  0.213061 32.20741 -5518.433
## - TotRmsAbvGrd  1  0.301912 32.29626 -5514.411
## - YearRemodAdd  1  0.341622 32.33597 -5512.617
## - ScreenPorch   1  0.432414 32.42676 -5508.523
## - LotArea       1  0.498776 32.49312 -5505.538
## - KitchenAbvGr  1  0.517119 32.51146 -5504.714
## - BsmtFullBath  1  0.620076 32.61442 -5500.098
## - GarageCars    1  0.739732 32.73408 -5494.752
## - Fireplaces    1  0.769139 32.76348 -5493.441
## - YearBuilt     1  1.721413 33.71576 -5451.611
## - OverallCond   1  2.837384 34.83173 -5404.068
## - OverallQual   1  6.112453 38.10680 -5272.867
##
## Step:  AIC=-5528.02
## SalePrice ~ ScreenPorch + WoodDeckSF + LotFrontage + KitchenAbvGr +
##     BsmtFullBath + OpenPorchSF + HalfBath + BedroomAbvGr + BsmtUnfSF +
##     TotRmsAbvGrd + FullBath + GarageCars + YearRemodAdd + GarageArea +
##     OverallCond + Fireplaces + X2ndFlrSF + YearBuilt + BsmtFinSF1 +
##     LotArea + X1stFlrSF + TotalBsmtSF + OverallQual + GrLivArea
##
##                 Df Sum of Sq      RSS       AIC
## - BsmtFinSF1    1  0.003086 31.99974 -5529.877
## - X2ndFlrSF     1  0.003290 31.99995 -5529.868
## - BsmtUnfSF     1  0.003671 32.00033 -5529.850
## - LotFrontage   1  0.004285 32.00094 -5529.822
## - BedroomAbvGr  1  0.005235 32.00189 -5529.779
## - OpenPorchSF   1  0.007173 32.00383 -5529.691
## - GarageArea    1  0.018355 32.01501 -5529.181
## - X1stFlrSF     1  0.024921 32.02158 -5528.881
## - GrLivArea     1  0.034759 32.03142 -5528.433
## <none>                      31.99666 -5528.018
## + GarageYrBlt   1  0.002313 31.99435 -5526.124
## + MasVnrArea    1  0.000080 31.99658 -5526.022
## - HalfBath      1  0.094448 32.09111 -5525.715
## - TotalBsmtSF   1  0.129460 32.12612 -5524.123
## - FullBath      1  0.209591 32.20625 -5520.486
## - WoodDeckSF    1  0.210972 32.20763 -5520.423
```

```
## - TotRmsAbvGrd  1  0.301134 32.29779 -5516.342
## - YearRemodAdd  1  0.343353 32.34001 -5514.434
## - ScreenPorch   1  0.432397 32.42906 -5510.420
## - LotArea       1  0.502633 32.49929 -5507.261
## - KitchenAbvGr  1  0.515787 32.51245 -5506.670
## - BsmtFullBath  1  0.621224 32.61788 -5501.943
## - GarageCars    1  0.739388 32.73605 -5496.664
## - Fireplaces    1  0.784905 32.78156 -5494.635
## - YearBuilt     1  2.530866 34.52752 -5418.875
## - OverallCond   1  2.855914 34.85257 -5405.195
## - OverallQual   1  6.112199 38.10886 -5274.788
##
## Step:  AIC=-5529.88
## SalePrice ~ ScreenPorch + WoodDeckSF + LotFrontage + KitchenAbvGr +
##     BsmtFullBath + OpenPorchSF + HalfBath + BedroomAbvGr + BsmtUnfSF +
##     TotRmsAbvGrd + FullBath + GarageCars + YearRemodAdd + GarageArea +
##     OverallCond + Fireplaces + X2ndFlrSF + YearBuilt + LotArea +
##     X1stFlrSF + TotalBsmtSF + OverallQual + GrLivArea
##
##                Df Sum of Sq      RSS       AIC
## - X2ndFlrSF     1  0.003462 32.00321 -5531.719
## - LotFrontage   1  0.004188 32.00393 -5531.686
## - BedroomAbvGr  1  0.004999 32.00474 -5531.649
## - OpenPorchSF   1  0.007304 32.00705 -5531.544
## - GarageArea    1  0.018900 32.01865 -5531.015
## - X1stFlrSF     1  0.025214 32.02496 -5530.727
## - GrLivArea     1  0.034444 32.03419 -5530.306
## <none>                      31.99974 -5529.877
## - BsmtUnfSF     1  0.046058 32.04580 -5529.777
## + BsmtFinSF1    1  0.003086 31.99666 -5528.018
## + GarageYrBlt   1  0.002369 31.99738 -5527.985
## + MasVnrArea    1  0.000009 31.99974 -5527.878
## - HalfBath      1  0.093581 32.09333 -5527.614
## - WoodDeckSF    1  0.209161 32.20891 -5522.365
## - FullBath      1  0.209268 32.20901 -5522.360
## - TotRmsAbvGrd  1  0.300738 32.30048 -5518.220
## - YearRemodAdd  1  0.344056 32.34380 -5516.263
## - TotalBsmtSF   1  0.386454 32.38620 -5514.351
## - ScreenPorch   1  0.429470 32.42921 -5512.413
## - LotArea       1  0.500014 32.49976 -5509.240
## - KitchenAbvGr  1  0.513680 32.51343 -5508.626
## - BsmtFullBath  1  0.627423 32.62717 -5503.528
## - GarageCars    1  0.737982 32.73773 -5498.589
## - Fireplaces    1  0.787174 32.78692 -5496.397
## - YearBuilt     1  2.540064 34.53981 -5420.356
## - OverallCond   1  2.854517 34.85426 -5407.124
## - OverallQual   1  6.132096 38.13184 -5275.907
##
## Step:  AIC=-5531.72
## SalePrice ~ ScreenPorch + WoodDeckSF + LotFrontage + KitchenAbvGr +
##     BsmtFullBath + OpenPorchSF + HalfBath + BedroomAbvGr + BsmtUnfSF +
##     TotRmsAbvGrd + FullBath + GarageCars + YearRemodAdd + GarageArea +
##     OverallCond + Fireplaces + YearBuilt + LotArea + X1stFlrSF +
##     TotalBsmtSF + OverallQual + GrLivArea
```

```
##
##                 Df Sum of Sq      RSS       AIC
## - LotFrontage   1  0.004322 32.00753 -5533.522
## - BedroomAbvGr  1  0.005294 32.00850 -5533.478
## - OpenPorchSF   1  0.007311 32.01052 -5533.386
## - GarageArea    1  0.018842 32.02205 -5532.860
## <none>                      32.00321 -5531.719
## - BsmtUnfSF     1  0.046698 32.04990 -5531.590
## + X2ndFlrSF     1  0.003462 31.99974 -5529.877
## + BsmtFinSF1    1  0.003258 31.99995 -5529.868
## + GarageYrBlt   1  0.002883 32.00032 -5529.851
## + MasVnrArea    1  0.000001 32.00321 -5529.719
## - HalfBath      1  0.099130 32.10234 -5529.204
## - X1stFlrSF     1  0.157024 32.16023 -5526.573
## - WoodDeckSF    1  0.209468 32.21267 -5524.194
## - FullBath      1  0.212732 32.21594 -5524.046
## - TotRmsAbvGrd  1  0.298900 32.30211 -5520.147
## - YearRemodAdd  1  0.342279 32.34549 -5518.187
## - TotalBsmtSF   1  0.386343 32.38955 -5516.200
## - ScreenPorch   1  0.428020 32.43123 -5514.322
## - LotArea       1  0.501705 32.50491 -5511.009
## - KitchenAbvGr  1  0.510355 32.51356 -5510.620
## - BsmtFullBath  1  0.625992 32.62920 -5505.437
## - GarageCars    1  0.743877 32.74708 -5500.172
## - Fireplaces    1  0.795879 32.79909 -5497.855
## - GrLivArea     1  1.022327 33.02553 -5487.810
## - YearBuilt     1  2.574222 34.57743 -5420.766
## - OverallCond   1  2.876066 34.87927 -5408.077
## - OverallQual   1  6.136385 38.13959 -5277.611
##
## Step:  AIC=-5533.52
## SalePrice ~ ScreenPorch + WoodDeckSF + KitchenAbvGr + BsmtFullBath +
##     OpenPorchSF + HalfBath + BedroomAbvGr + BsmtUnfSF + TotRmsAbvGrd +
##     FullBath + GarageCars + YearRemodAdd + GarageArea + OverallCond +
##     Fireplaces + YearBuilt + LotArea + X1stFlrSF + TotalBsmtSF +
##     OverallQual + GrLivArea
##
##                 Df Sum of Sq      RSS       AIC
## - BedroomAbvGr  1  0.004881 32.01241 -5535.299
## - OpenPorchSF   1  0.007131 32.01466 -5535.197
## - GarageArea    1  0.017695 32.02522 -5534.715
## <none>                      32.00753 -5533.522
## - BsmtUnfSF     1  0.046886 32.05441 -5533.385
## + LotFrontage   1  0.004322 32.00321 -5531.719
## + X2ndFlrSF     1  0.003596 32.00393 -5531.686
## + BsmtFinSF1    1  0.003161 32.00437 -5531.666
## + GarageYrBlt   1  0.002465 32.00506 -5531.634
## + MasVnrArea    1  0.000001 32.00753 -5531.522
## - HalfBath      1  0.100872 32.10840 -5530.928
## - X1stFlrSF     1  0.155574 32.16310 -5528.443
## - WoodDeckSF    1  0.215955 32.22348 -5525.704
## - FullBath      1  0.216617 32.22415 -5525.674
## - TotRmsAbvGrd  1  0.295573 32.30310 -5522.102
## - YearRemodAdd  1  0.339586 32.34711 -5520.114
```

```
## - TotalBsmtSF   1  0.382783 32.39031 -5518.165
## - ScreenPorch   1  0.428292 32.43582 -5516.115
## - LotArea       1  0.498834 32.50636 -5512.944
## - KitchenAbvGr  1  0.509218 32.51675 -5512.477
## - BsmtFullBath  1  0.628642 32.63617 -5507.125
## - GarageCars    1  0.746400 32.75393 -5501.866
## - Fireplaces    1  0.810706 32.81823 -5499.003
## - GrLivArea     1  1.021281 33.02881 -5489.665
## - YearBuilt     1  2.597691 34.60522 -5421.593
## - OverallCond   1  2.896411 34.90394 -5409.044
## - OverallQual   1  6.136808 38.14434 -5279.429
##
## Step:  AIC=-5535.3
## SalePrice ~ ScreenPorch + WoodDeckSF + KitchenAbvGr + BsmtFullBath +
##     OpenPorchSF + HalfBath + BsmtUnfSF + TotRmsAbvGrd + FullBath +
##     GarageCars + YearRemodAdd + GarageArea + OverallCond + Fireplaces +
##     YearBuilt + LotArea + X1stFlrSF + TotalBsmtSF + OverallQual +
##     GrLivArea
##
##                Df Sum of Sq      RSS       AIC
## - OpenPorchSF   1  0.007592 32.02000 -5536.953
## - GarageArea    1  0.016901 32.02931 -5536.529
## <none>                      32.01241 -5535.299
## - BsmtUnfSF     1  0.045143 32.05755 -5535.242
## + BedroomAbvGr  1  0.004881 32.00753 -5533.522
## + LotFrontage   1  0.003909 32.00850 -5533.478
## + X2ndFlrSF     1  0.003878 32.00853 -5533.476
## + BsmtFinSF1    1  0.002936 32.00947 -5533.433
## + GarageYrBlt   1  0.002835 32.00957 -5533.429
## + MasVnrArea    1  0.000000 32.01241 -5533.299
## - HalfBath      1  0.100295 32.11270 -5532.732
## - X1stFlrSF     1  0.151638 32.16405 -5530.400
## - WoodDeckSF    1  0.214809 32.22722 -5527.535
## - FullBath      1  0.227051 32.23946 -5526.981
## - YearRemodAdd  1  0.335655 32.34806 -5522.071
## - TotalBsmtSF   1  0.384601 32.39701 -5519.863
## - TotRmsAbvGrd  1  0.427901 32.44031 -5517.913
## - ScreenPorch   1  0.431547 32.44396 -5517.749
## - LotArea       1  0.503003 32.51541 -5514.537
## - KitchenAbvGr  1  0.519593 32.53200 -5513.792
## - BsmtFullBath  1  0.628033 32.64044 -5508.934
## - GarageCars    1  0.744812 32.75722 -5503.720
## - Fireplaces    1  0.806014 32.81842 -5500.994
## - GrLivArea     1  1.045117 33.05753 -5490.396
## - YearBuilt     1  2.632864 34.64527 -5421.904
## - OverallCond   1  2.959538 34.97195 -5408.202
## - OverallQual   1  6.266391 38.27880 -5276.291
##
## Step:  AIC=-5536.95
## SalePrice ~ ScreenPorch + WoodDeckSF + KitchenAbvGr + BsmtFullBath +
##     HalfBath + BsmtUnfSF + TotRmsAbvGrd + FullBath + GarageCars +
##     YearRemodAdd + GarageArea + OverallCond + Fireplaces + YearBuilt +
##     LotArea + X1stFlrSF + TotalBsmtSF + OverallQual + GrLivArea
##
```

```
##                Df Sum of Sq      RSS       AIC
## - GarageArea    1  0.015238 32.03524 -5538.259
## <none>                     32.02000 -5536.953
## - BsmtUnfSF     1  0.046376 32.06638 -5536.840
## + OpenPorchSF   1  0.007592 32.01241 -5535.299
## + BedroomAbvGr  1  0.005341 32.01466 -5535.197
## + X2ndFlrSF     1  0.003896 32.01610 -5535.131
## + LotFrontage   1  0.003714 32.01629 -5535.123
## + BsmtFinSF1    1  0.003058 32.01694 -5535.093
## + GarageYrBlt   1  0.002713 32.01729 -5535.077
## + MasVnrArea    1  0.000019 32.01998 -5534.954
## - HalfBath      1  0.096388 32.11639 -5534.565
## - X1stFlrSF     1  0.153541 32.17354 -5531.969
## - WoodDeckSF    1  0.220630 32.24063 -5528.928
## - FullBath      1  0.222517 32.24252 -5528.842
## - YearRemodAdd  1  0.330804 32.35081 -5523.947
## - TotalBsmtSF   1  0.379812 32.39981 -5521.737
## - ScreenPorch   1  0.427886 32.44789 -5519.572
## - TotRmsAbvGrd  1  0.431409 32.45141 -5519.414
## - LotArea       1  0.502458 32.52246 -5516.221
## - KitchenAbvGr  1  0.514505 32.53451 -5515.680
## - BsmtFullBath  1  0.623331 32.64333 -5510.805
## - GarageCars    1  0.760134 32.78013 -5504.699
## - Fireplaces    1  0.806153 32.82615 -5502.651
## - GrLivArea     1  1.037591 33.05759 -5492.393
## - YearBuilt     1  2.638575 34.65858 -5423.344
## - OverallCond   1  2.958683 34.97868 -5409.921
## - OverallQual   1  6.258819 38.27882 -5278.291
##
## Step:  AIC=-5538.26
## SalePrice ~ ScreenPorch + WoodDeckSF + KitchenAbvGr + BsmtFullBath +
##     HalfBath + BsmtUnfSF + TotRmsAbvGrd + FullBath + GarageCars +
##     YearRemodAdd + OverallCond + Fireplaces + YearBuilt + LotArea +
##     X1stFlrSF + TotalBsmtSF + OverallQual + GrLivArea
##
##                Df Sum of Sq      RSS       AIC
## <none>                     32.03524 -5538.259
## - BsmtUnfSF     1  0.050084 32.08532 -5537.978
## + GarageArea    1  0.015238 32.02000 -5536.953
## + OpenPorchSF   1  0.005929 32.02931 -5536.529
## + BedroomAbvGr  1  0.004496 32.03074 -5536.463
## + X2ndFlrSF     1  0.003796 32.03144 -5536.432
## + BsmtFinSF1    1  0.003561 32.03168 -5536.421
## + LotFrontage   1  0.002771 32.03247 -5536.385
## + GarageYrBlt   1  0.000273 32.03497 -5536.271
## + MasVnrArea    1  0.000107 32.03513 -5536.263
## - HalfBath      1  0.091680 32.12692 -5536.086
## - X1stFlrSF     1  0.158345 32.19358 -5533.060
## - FullBath      1  0.213547 32.24879 -5530.558
## - WoodDeckSF    1  0.220034 32.25527 -5530.265
## - YearRemodAdd  1  0.325975 32.36121 -5525.477
## - TotalBsmtSF   1  0.395384 32.43062 -5522.349
## - TotRmsAbvGrd  1  0.424413 32.45965 -5521.043
## - ScreenPorch   1  0.426873 32.46211 -5520.932
```

```
## - LotArea       1  0.506351 32.54159 -5517.362
## - KitchenAbvGr  1  0.521862 32.55710 -5516.666
## - BsmtFullBath  1  0.625287 32.66053 -5512.036
## - Fireplaces    1  0.791233 32.82647 -5504.636
## - GrLivArea     1  1.089460 33.12470 -5491.432
## - YearBuilt     1  2.662371 34.69761 -5423.700
## - GarageCars    1  2.699709 34.73495 -5422.130
## - OverallCond   1  2.984066 35.01931 -5410.227
## - OverallQual   1  6.260774 38.29601 -5279.635
```

```r
stepbackward_sum <- summary(stepbackward)
stepbackward_sum
```

```
##
## Call:
## lm(formula = SalePrice ~ ScreenPorch + WoodDeckSF + KitchenAbvGr +
##     BsmtFullBath + HalfBath + BsmtUnfSF + TotRmsAbvGrd + FullBath +
##     GarageCars + YearRemodAdd + OverallCond + Fireplaces + YearBuilt +
##     LotArea + X1stFlrSF + TotalBsmtSF + OverallQual + GrLivArea,
##     data = train.df.Boruta)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.1228365 -0.0675088  0.0052979  0.0804568  0.4638405
##
## Coefficients:
##                     Estimate      Std. Error  t value               Pr(>|t|)
## (Intercept)   3.242967967491  0.551178309080  5.88370     0.0000000049810693
## ScreenPorch   0.000318029066  0.000072577112  4.38195     0.0000126139384168
## WoodDeckSF    0.000106292100  0.000033786122  3.14603             0.00168889
## KitchenAbvGr -0.099780697131  0.020594473679 -4.84502     0.0000014026515103
## BsmtFullBath  0.056050340454  0.010568675260  5.30344     0.0000001313750844
## HalfBath      0.022809299275  0.011232015913  2.03074             0.04246471
## BsmtUnfSF    -0.000019830857  0.000013212217 -1.50095             0.13358796
## TotRmsAbvGrd  0.020111839040  0.004602985393  4.36930     0.0000133565886396
## FullBath      0.036608474616  0.011811808680  3.09931             0.00197725
## GarageCars    0.079146922430  0.007182200493 11.01987 < 0.000000000000000222
## YearRemodAdd  0.001056153428  0.000275814246  3.82922             0.00013406
## OverallCond   0.049703505611  0.004290074352 11.58570 < 0.000000000000000222
## Fireplaces    0.044529438548  0.007464091604  5.96582     0.0000000030582129
## YearBuilt     0.002653533255  0.000242477931 10.94340 < 0.000000000000000222
## LotArea       0.000002042779  0.000000428033  4.77248     0.0000020046455322
## X1stFlrSF     0.000058463863  0.000021906188  2.66883             0.00769706
## TotalBsmtSF   0.000081259449  0.000019268425  4.21723     0.0000262727716261
## OverallQual   0.082635011540  0.004924159248 16.78155 < 0.000000000000000222
## GrLivArea     0.000138689940  0.000019811681  7.00041     0.0000000000038998
##
## (Intercept)  ***
## ScreenPorch  ***
## WoodDeckSF   **
## KitchenAbvGr ***
## BsmtFullBath ***
## HalfBath     *
## BsmtUnfSF
```

```
## TotRmsAbvGrd ***
## FullBath     **
## GarageCars   ***
## YearRemodAdd ***
## OverallCond  ***
## Fireplaces   ***
## YearBuilt    ***
## LotArea      ***
## X1stFlrSF    **
## TotalBsmtSF  ***
## OverallQual  ***
## GrLivArea    ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.149101 on 1441 degrees of freedom
## Multiple R-squared:  0.862392,   Adjusted R-squared:  0.860673
## F-statistic:  501.71 on 18 and 1441 DF,  p-value: < 0.000000000000000222
```

The ***backwards stepwise*** algorithm starts from a full model, containing 26 variables, with an AIC of -5524.13. At each step, it eliminates the locally "worst" variable from the model, causing the AIC to improve (i.e., become more negative) until there are no longer any variables which will further improve the AIC when dropped.

Successively removing 8 variables reduces the AIC to -5538.26 , yielding an $R^2 = 0.862391973$ and an adj-$R^2 =$ '$tstepforward_sum$adj.r.squared'$.

The standard error of the residuals has been reduced to $\sigma = 0.149101494$.

The variables selected under backward stepwise include:

1. ScreenPorch
2. WoodDeckSF
3. KitchenAbvGr
4. BsmtFullBath
5. HalfBath
6. BsmtUnfSF
7. TotRmsAbvGrd
8. FullBath
9. GarageCars
10. YearRemodAdd
11. OverallCond
12. Fireplaces
13. YearBuilt
14. LotArea
15. X1stFlrSF
16. TotalBsmtSF
17. OverallQual
18. GrLivArea

These variables are the same as those obtained from the forward stepwise algorithm, which means that the two methods have converged (this does not always occur.)

Provide your complete model summary and results with analysis.

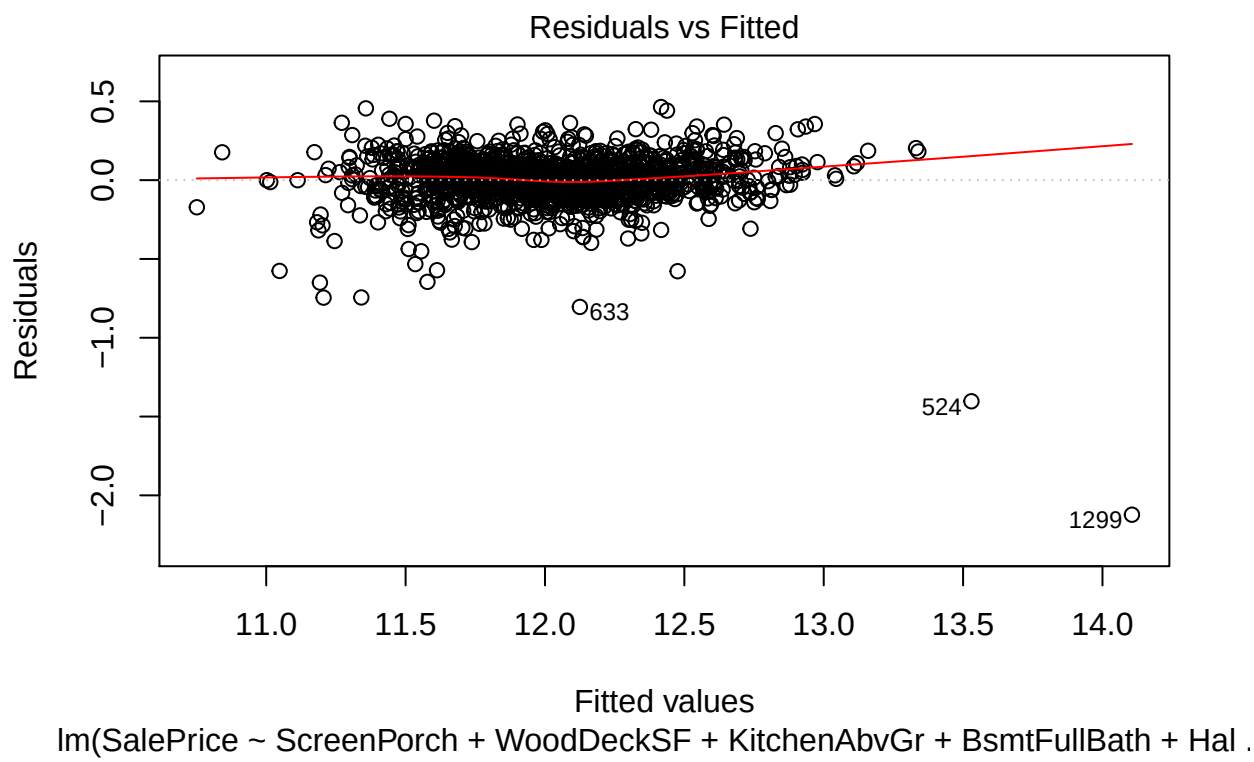
The model is

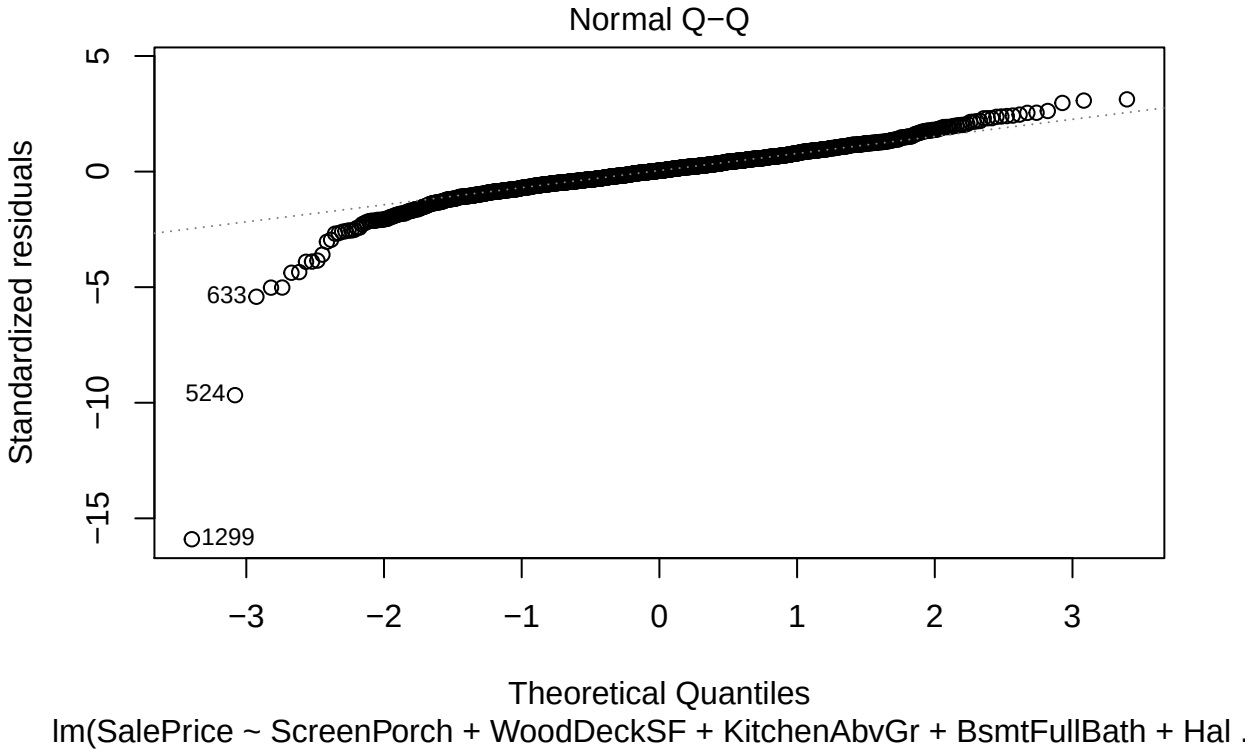
$$
\begin{aligned}
log(SalePrice) = 3.242967967491 &+ 0.082635011 \cdot OverallQual && + 0.000138689 \cdot GrLivArea \\
&+ 0.002653533 \cdot YearBuilt && + 0.049703505 \cdot OverallCond \\
&+ 0.079146922 \cdot GarageCars && + 0.000081259 \cdot TotalBsmtSF \\
&+ 0.044529438 \cdot Fireplaces && + 0.056050340 \cdot BsmtFullBath \\
&+ 0.000002042 \cdot LotArea && + 0.000318029 \cdot ScreenPorch \\
&+ 0.001056153 \cdot YearRemodAdd && + 0.000106292 \cdot WoodDeckSF \\
&+ 0.020111839 \cdot TotRmsAbvGrd && - 0.099780697 \cdot KitchenAbvGr \\
&+ 0.036608474 \cdot FullBath && + 0.000058463 \cdot X1stFlrSF \\
&+ 0.022809299 \cdot HalfBath && - 0.000019830 \cdot BsmtUnfSF
\end{aligned}
$$

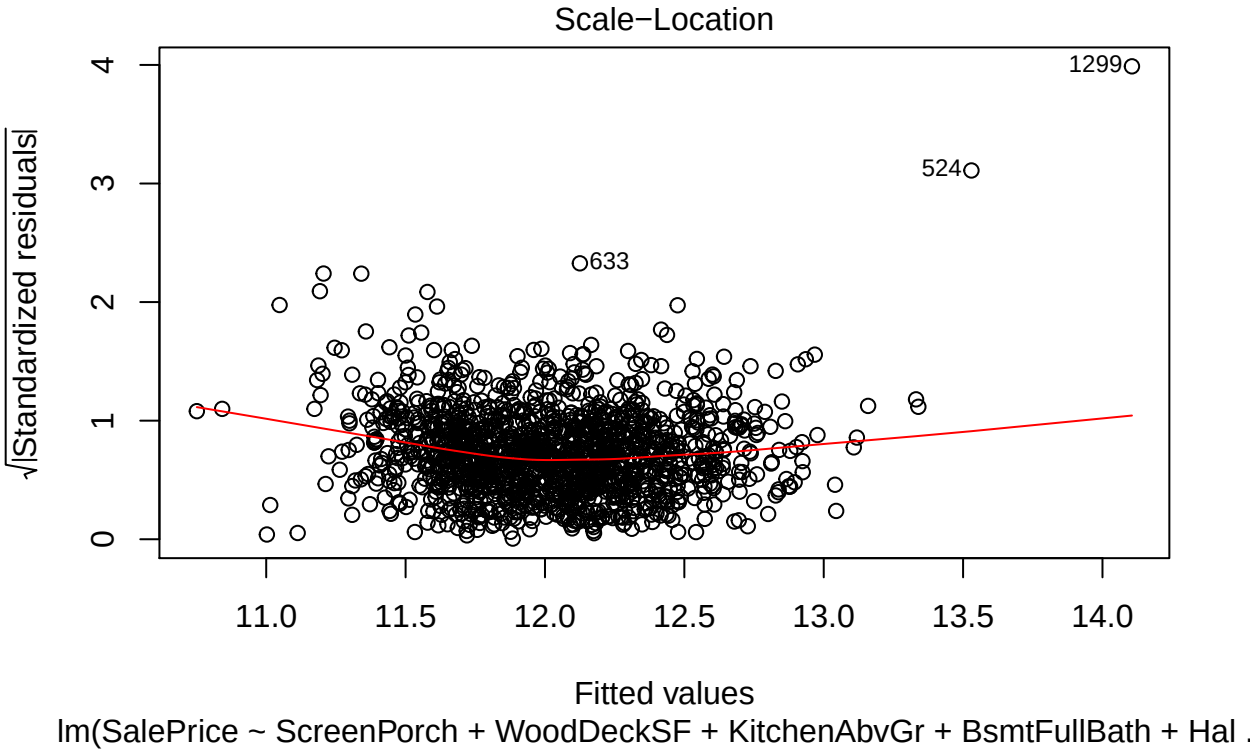
Each of the variables is significant, except for the final one.

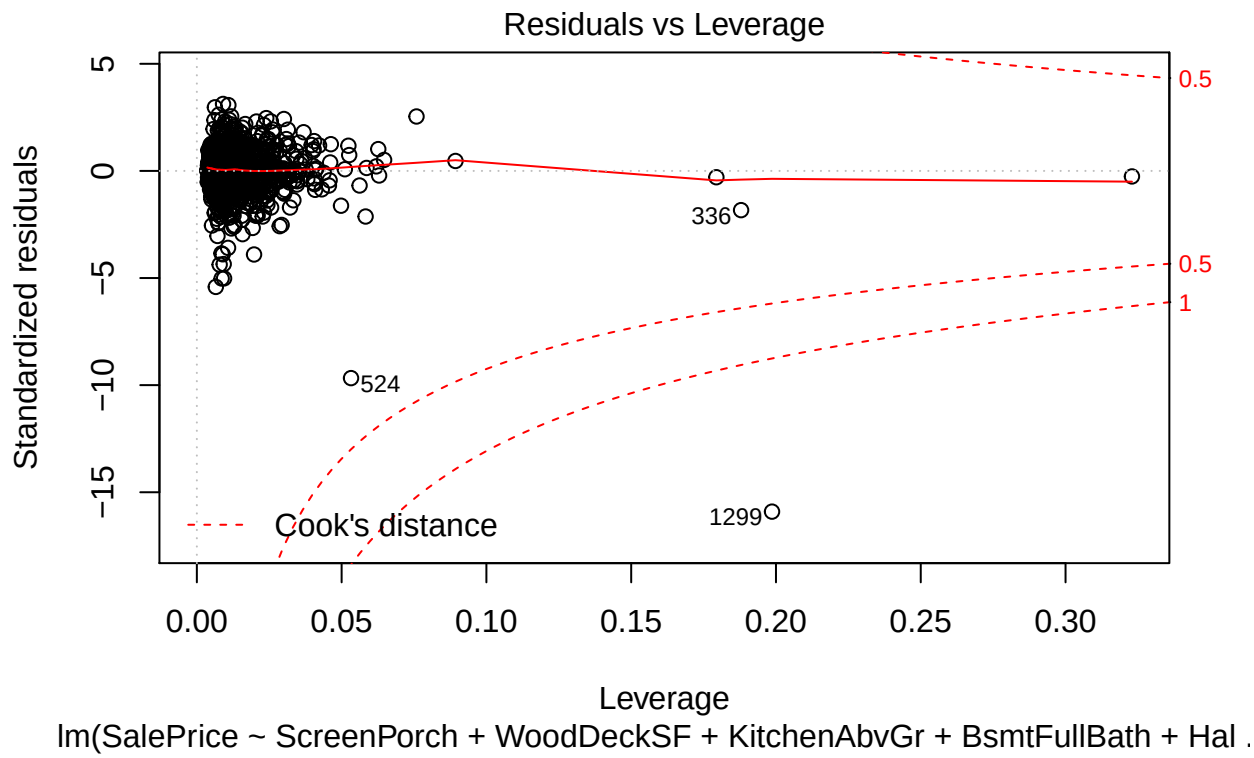
Diagnostics

```
plot(stepbackward)
```

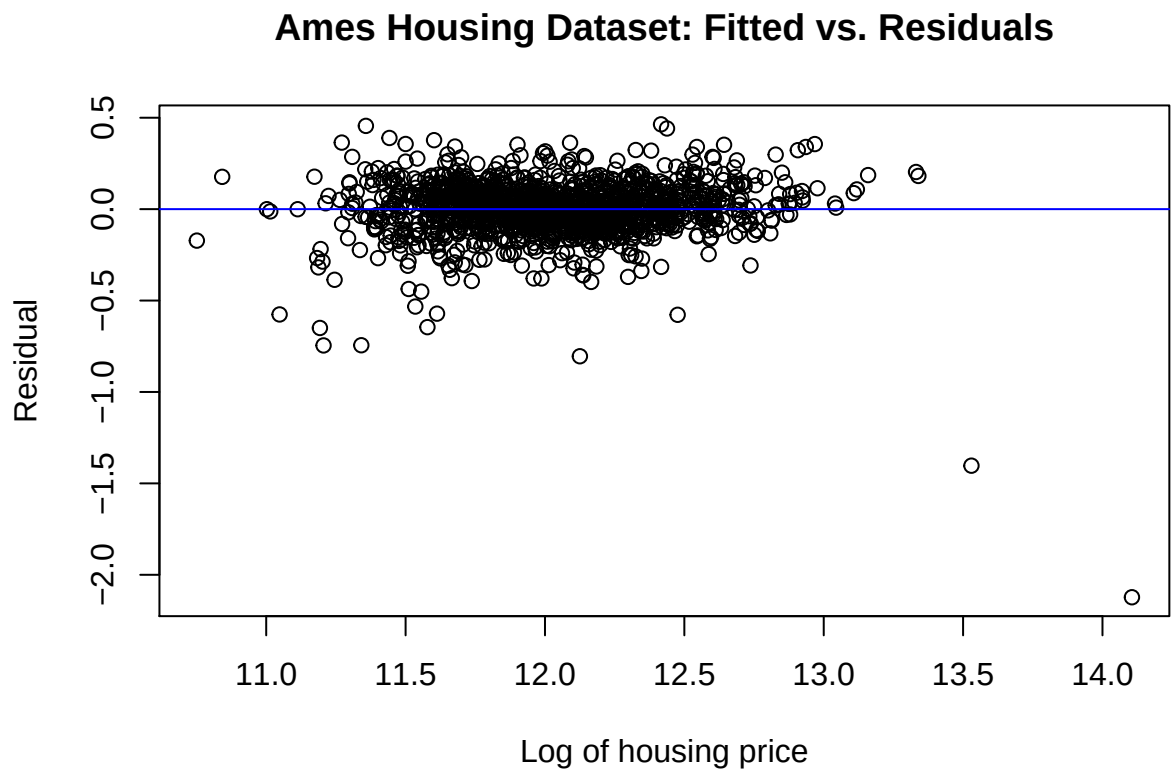



Normal Q−Q
Standardized residuals
5
0
−5
−10
−15
633
524
1299
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(SalePrice ~ ScreenPorch + WoodDeckSF + KitchenAbvGr + BsmtFullBath + Hal .

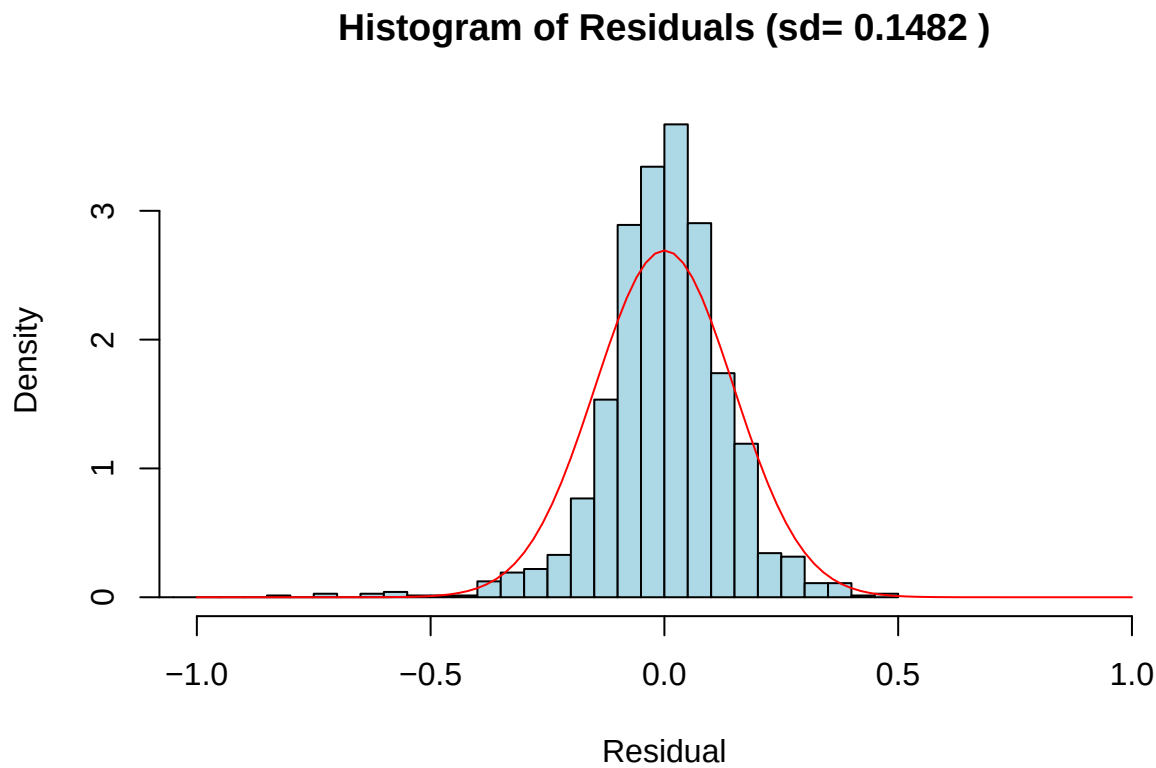

Scale−Location
√|Standardized residuals|
0
1
2
3
4
1299
524
633
11.0
11.5
12.0
12.5
13.0
13.5
14.0
Fitted values
lm(SalePrice ~ ScreenPorch + WoodDeckSF + KitchenAbvGr + BsmtFullBath + Hal .




```
#### Fitted vs. Residuals
Residual = resid(stepbackward)
Fitted = fitted(stepbackward)
plot(Fitted,Residual,
     main="Ames Housing Dataset: Fitted vs. Residuals",
     xlab="Log of housing price")
abline(h=0, col="blue")
```




```
#### Histogram of residuals
titl = paste("Histogram of Residuals (sd=",round(sd(Residual),4),")" )
hist(Residual, main = titl, ylab = "Density",
     #ylim = c(0, 0.35),
     xlim = c(-1,1),
     prob = TRUE,breaks=40, col="lightblue")
curve(dnorm(x, mean = mean(Residual), sd = sd(Residual)), col="red", add=TRUE)
```




```r
#### Tests for normality
library(olsrr)
ols_test_normality(stepbackward)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.8424         0.0000
## Kolmogorov-Smirnov        0.0869         0.0000
## Cramer-von Mises        375.1229         0.0886
## Anderson-Darling         22.3934         0.0000
## -----------------------------------------------
```

Only the Cramer-von Mises test passes – the other three tests fail.

Homogeneity of residuals

```r
library(lmSupport)
modelAssumptions(stepbackward,"NORMAL")
```

```
## Descriptive Statistics for Studentized Residuals
##
## Call:
## lm(formula = SalePrice ~ ScreenPorch + WoodDeckSF + KitchenAbvGr +
```

```
##     BsmtFullBath + HalfBath + BsmtUnfSF + TotRmsAbvGrd + FullBath +
##     GarageCars + YearRemodAdd + OverallCond + Fireplaces + YearBuilt +
##     LotArea + X1stFlrSF + TotalBsmtSF + OverallQual + GrLivArea,
##     data = train.df.Boruta)
##
## Coefficients:
##    (Intercept)     ScreenPorch      WoodDeckSF    KitchenAbvGr    BsmtFullBath
##  3.24296796749   0.00031802907   0.00010629210  -0.09978069713   0.05605034045
##       HalfBath       BsmtUnfSF    TotRmsAbvGrd        FullBath      GarageCars
##  0.02280929927  -0.00001983086   0.02011183904   0.03660847462   0.07914692243
##   YearRemodAdd     OverallCond      Fireplaces       YearBuilt         LotArea
##  0.00105615343   0.04970350561   0.04452943855   0.00265353326   0.00000204278
##      X1stFlrSF     TotalBsmtSF     OverallQual       GrLivArea
##  0.00005846386   0.00008125945   0.08263501154   0.00013868994
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = Model)
##
##                         Value     p-value                   Decision
## Global Stat        80510.9151 0.000000000 Assumptions NOT satisfied!
## Skewness            2338.9716 0.000000000 Assumptions NOT satisfied!
## Kurtosis           78087.1383 0.000000000 Assumptions NOT satisfied!
## Link Function         73.6082 0.000000000 Assumptions NOT satisfied!
## Heteroscedasticity    11.1970 0.000819316 Assumptions NOT satisfied!
```

All tests fail.

Despite the transformation, the model does not satisfy the conditions required for multiple linear regression.

Additional transformations are needed in order to improve the model to satisfy the conditions.

Compute the RMSE (on the training data)

```r
### this result is the (log(saleprice)) - - needed for RMSE calculation
log_res_train <- predict(object=stepbackward,newdata=logtrain.df)
summary(log_res_train)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 10.7511 11.7482 12.0162 12.0241 12.2667 14.1058
```

```r
###

### need to exponentiate  to get res_train - the predictions, in dollars:
res_train = exp(log_res_train)
summary(res_train)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   46681.5  126528.3  165417.7  179206.7  212491.5 1336768.4
```

```r
### obtain the test predictions (log_SalePrice) - used (within kaggle) for RMSE score
log_res_test <- predict(object=stepbackward,newdata=test.df)
summary(log_res_test)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 10.9195 11.7498 11.9792 12.0109 12.2532 13.9837
```

```r
### need to exponential to get predictions on dollars (for submission to Kaggle)
res_test = exp(log_res_test)
summary(res_test)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##   55241.3  126732.7  159397.0  176983.7  209660.1 1183173.8
```

```r
### Compute the RMSE on the training dataset
library(Metrics)
log_sale_price = logtrain.df$SalePrice     # log of actual sales prices, for train
### Here is the RMSE of the log of the actual sale price vs. the log of the predicted price:
rmse(log_res_train,log_sale_price)
```

```
## [1] 0.148128136
```

Report your Kaggle.com user name and score.

```
### res_test are the exponentiated results, in dollars, which is the format required for Kaggle submiss
kaggle_sub <- cbind(Id=test.df$Id,SalePrice=res_test)
head(kaggle_sub)
```

```
##     Id  SalePrice
## 1 1461 116464.922
## 2 1462 146224.390
## 3 1463 168950.113
## 4 1464 197376.831
## 5 1465 192292.327
## 6 1466 176263.344
```

```
### check for any NA values in the submission -- this would cause a problem
#### Number of NAs ?
sum(is.na(kaggle_sub[,2]))
```

```
## [1] 0
```

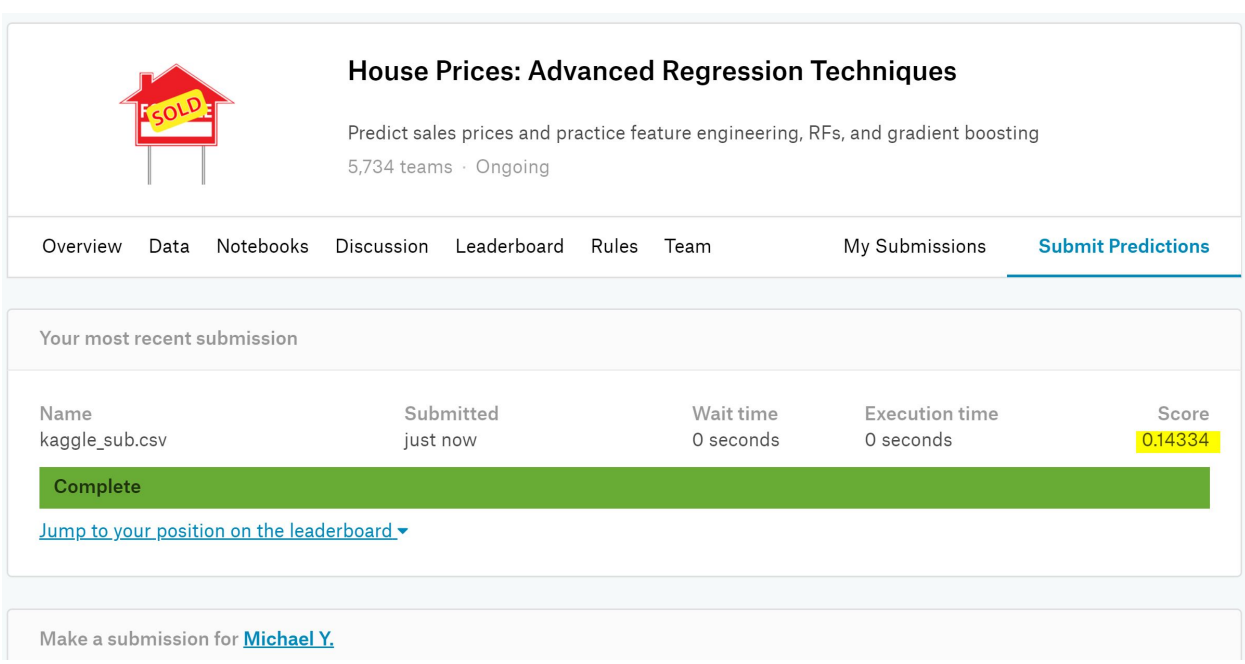
```
#### Listing of rows for which the model generated NA predictions:
kaggle_sub[is.na(kaggle_sub[,2]),]
```

```
##      Id SalePrice
```

```
### Create a csv file in the format required for submission to Kaggle
write_csv(data.frame(kaggle_sub),"kaggle_sub.csv")
```

Kaggle results

```
knitr::include_graphics("Kaggle_results.JPG")
```



```r
knitr::include_graphics("Kaggle_results_listing.JPG")
```

TeamId	TeamName	SubmissionDate	Score	
4033161	Michael Y.	12/16/2019 8:45	0.4089	testing sample submission file
4033161	Michael Y.	12/16/2019 20:58	0.14334	stepwise regression predictions

---

The results from the Kaggle submission for this model: $rmse = 0.14334$