House price

ForozanHeidaryan

2022-12-03

This data set is about the sale price of buildings in the city of Ame Iva which has been Collected by statistics named Dean de cock

Understanding the Business Question

Price Recommendation for house Data inspection

Dataset Description

File descriptions

data_description.txt - full description of each column, originally prepared by Dean De Cock

but lightly edited to match the column names used here

sample_submission.csv - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

Data fields

Here's a brief version of what you'll find in the data description file.

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

read data from file

```
data <- read.csv("train.csv" , header = TRUE )
```

summary of the data

```
dim(data)
```

```
## [1] 1460   81
```

```
length(unique(data$Id))
```

```
## [1] 1460
```

```
summary(data)
```

```
##        Id           MSSubClass      MSZoning          LotFrontage
##  Min.   :   1.0   Min.   : 20.0   Length:1460        Min.   : 21.00
##  1st Qu.: 365.8   1st Qu.: 20.0   Class :character   1st Qu.: 59.00
##  Median : 730.5   Median : 50.0   Mode  :character   Median : 69.00
##  Mean   : 730.5   Mean   : 56.9                      Mean   : 70.05
##  3rd Qu.:1095.2   3rd Qu.: 70.0                      3rd Qu.: 80.00
##  Max.   :1460.0   Max.   :190.0                      Max.   :313.00
##                                                      NA's   :259
##     LotArea          Street             Alley             LotShape
##  Min.   :  1300   Length:1460        Length:1460        Length:1460
##  1st Qu.:  7554   Class :character   Class :character   Class :character
##  Median :  9478   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 10517
##  3rd Qu.: 11602
##  Max.   :215245
##
##  LandContour         Utilities          LotConfig          LandSlope
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  Neighborhood        Condition1         Condition2          BldgType
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   HouseStyle         OverallQual      OverallCond      YearBuilt
##  Length:1460        Min.   : 1.000   Min.   :1.000   Min.   :1872
##  Class :character   1st Qu.: 5.000   1st Qu.:5.000   1st Qu.:1954
##  Mode  :character   Median : 6.000   Median :5.000   Median :1973
##                     Mean   : 6.099   Mean   :5.575   Mean   :1971
##                     3rd Qu.: 7.000   3rd Qu.:6.000   3rd Qu.:2000
##                     Max.   :10.000   Max.   :9.000   Max.   :2010
##
##   YearRemodAdd   RoofStyle           RoofMatl         Exterior1st
##  Min.   :1950   Length:1460        Length:1460        Length:1460
##  1st Qu.:1967   Class :character   Class :character   Class :character
##  Median :1994   Mode  :character   Mode  :character   Mode  :character
##  Mean   :1985
##  3rd Qu.:2004
##  Max.   :2010
##
```

```
##  Exterior2nd         MasVnrType          MasVnrArea      ExterQual
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median :   0.0   Mode  :character
##                                        Mean   : 103.7
##                                        3rd Qu.: 166.0
##                                        Max.   :1600.0
##                                        NA's   :8
##   ExterCond          Foundation          BsmtQual           BsmtCond
##  Length:1460        Length:1460        Length:1460        Length:1460
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  BsmtExposure       BsmtFinType1         BsmtFinSF1     BsmtFinType2
##  Length:1460        Length:1460        Min.   :   0.0   Length:1460
##  Class :character   Class :character   1st Qu.:   0.0   Class :character
##  Mode  :character   Mode  :character   Median : 383.5   Mode  :character
##                                        Mean   : 443.6
##                                        3rd Qu.: 712.2
##                                        Max.   :5644.0
##
##    BsmtFinSF2        BsmtUnfSF       TotalBsmtSF       Heating
##  Min.   :   0.00   Min.   :   0.0   Min.   :   0.0   Length:1460
##  1st Qu.:   0.00   1st Qu.: 223.0   1st Qu.: 795.8   Class :character
##  Median :   0.00   Median : 477.5   Median : 991.5   Mode  :character
##  Mean   :  46.55   Mean   : 567.2   Mean   :1057.4
##  3rd Qu.:   0.00   3rd Qu.: 808.0   3rd Qu.:1298.2
##  Max.   :1474.00   Max.   :2336.0   Max.   :6110.0
##
##   HeatingQC          CentralAir         Electrical          X1stFlrSF
##  Length:1460        Length:1460        Length:1460        Min.   : 334
##  Class :character   Class :character   Class :character   1st Qu.: 882
##  Mode  :character   Mode  :character   Mode  :character   Median :1087
##                                                           Mean   :1163
##                                                           3rd Qu.:1391
##                                                           Max.   :4692
##
##    X2ndFlrSF     LowQualFinSF       GrLivArea     BsmtFullBath
##  Min.   :   0   Min.   :  0.000   Min.   : 334   Min.   :0.0000
##  1st Qu.:   0   1st Qu.:  0.000   1st Qu.:1130   1st Qu.:0.0000
##  Median :   0   Median :  0.000   Median :1464   Median :0.0000
##  Mean   : 347   Mean   :  5.845   Mean   :1515   Mean   :0.4253
##  3rd Qu.: 728   3rd Qu.:  0.000   3rd Qu.:1777   3rd Qu.:1.0000
##  Max.   :2065   Max.   :572.000   Max.   :5642   Max.   :3.0000
##
##   BsmtHalfBath        FullBath        HalfBath       BedroomAbvGr
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000   Min.   :0.000
##  1st Qu.:0.00000   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
##  Median :0.00000   Median :2.000   Median :0.0000   Median :3.000
##  Mean   :0.05753   Mean   :1.565   Mean   :0.3829   Mean   :2.866
##  3rd Qu.:0.00000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000
```

```
##  Max.   :2.00000   Max.   :3.000   Max.   :2.0000   Max.   :8.000
##
##   KitchenAbvGr   KitchenQual         TotRmsAbvGrd     Functional
##  Min.   :0.000   Length:1460        Min.   : 2.000   Length:1460
##  1st Qu.:1.000   Class :character   1st Qu.: 5.000   Class :character
##  Median :1.000   Mode  :character   Median : 6.000   Mode  :character
##  Mean   :1.047                      Mean   : 6.518
##  3rd Qu.:1.000                      3rd Qu.: 7.000
##  Max.   :3.000                      Max.   :14.000
##
##    Fireplaces    FireplaceQu         GarageType         GarageYrBlt
##  Min.   :0.000   Length:1460        Length:1460        Min.   :1900
##  1st Qu.:0.000   Class :character   Class :character   1st Qu.:1961
##  Median :1.000   Mode  :character   Mode  :character   Median :1980
##  Mean   :0.613                                         Mean   :1979
##  3rd Qu.:1.000                                         3rd Qu.:2002
##  Max.   :3.000                                         Max.   :2010
##                                                        NA's   :81
##  GarageFinish         GarageCars      GarageArea      GarageQual
##  Length:1460        Min.   :0.000   Min.   :   0.0   Length:1460
##  Class :character   1st Qu.:1.000   1st Qu.: 334.5   Class :character
##  Mode  :character   Median :2.000   Median : 480.0   Mode  :character
##                     Mean   :1.767   Mean   : 473.0
##                     3rd Qu.:2.000   3rd Qu.: 576.0
##                     Max.   :4.000   Max.   :1418.0
##
##   GarageCond         PavedDrive          WoodDeckSF      OpenPorchSF
##  Length:1460        Length:1460        Min.   :  0.00   Min.   :  0.00
##  Class :character   Class :character   1st Qu.:  0.00   1st Qu.:  0.00
##  Mode  :character   Mode  :character   Median :  0.00   Median : 25.00
##                                        Mean   : 94.24   Mean   : 46.66
##                                        3rd Qu.:168.00   3rd Qu.: 68.00
##                                        Max.   :857.00   Max.   :547.00
##
##  EnclosedPorch      X3SsnPorch      ScreenPorch        PoolArea
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.000
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.000
##  Median :  0.00   Median :  0.00   Median :  0.00   Median :  0.000
##  Mean   : 21.95   Mean   :  3.41   Mean   : 15.06   Mean   :  2.759
##  3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.00   3rd Qu.:  0.000
##  Max.   :552.00   Max.   :508.00   Max.   :480.00   Max.   :738.000
##
##     PoolQC             Fence           MiscFeature          MiscVal
##  Length:1460        Length:1460        Length:1460        Min.   :    0.00
##  Class :character   Class :character   Class :character   1st Qu.:    0.00
##  Mode  :character   Mode  :character   Mode  :character   Median :    0.00
##                                                           Mean   :   43.49
##                                                           3rd Qu.:    0.00
##                                                           Max.   :15500.00
##
##      MoSold           YrSold       SaleType         SaleCondition
##  Min.   : 1.000   Min.   :2006   Length:1460        Length:1460
##  1st Qu.: 5.000   1st Qu.:2007   Class :character   Class :character
##  Median : 6.000   Median :2008   Mode  :character   Mode  :character
```

```
##  Mean   : 6.322   Mean   :2008
##  3rd Qu.: 8.000   3rd Qu.:2009
##  Max.   :12.000   Max.   :2010
##
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
##
```

Convert categorical variables to factor

```
cat_var <- c("MSZoning" , "Street" , "Alley" ,"LotShape" , "LandContour", "Utilities" , "LotConfig" , "I
             "Neighborhood" , "Condition1" , "Condition2" , "BldgType" ,
             "HouseStyle", "OverallQual", "OverallCond" , "RoofStyle" , "Exterior1st" , "Exterior2nd" ,
             "ExterQual" ,"ExterCond" , "Foundation" , "BsmtQual" ,"BsmtCond" ,
             "BsmtExposure" ,"BsmtFinType1" ,"BsmtFinType2" , "Heating" , "HeatingQC" , "CentralAir" ,
             "KitchenQual" , "Functional" , "Fireplaces" , "FireplaceQu" ,
             "GarageType" , "GarageFinish" , "GarageCars" , "GarageQual" ,"GarageCond" , "PavedDrive" ,
             "Fence","RoofMatl" ,
             "MiscFeature" , "SaleType" , "SaleCondition")
```

```
data[,cat_var]  <-  lapply(data[,cat_var] , factor)
knitr::kable(summary(data))
```

Id	[illegible]
[illegible]	[illegible]

```
knitr::kable(str(data))
```

```
## 'data.frame':    1460 obs. of  81 variables:
```

```
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
##  $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...
##  $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...
##  $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...
##  $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...
##  $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...
##  $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...
##  $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...
##  $ OverallQual  : Factor w/ 10 levels "1","2","3","4",..: 7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : Factor w/ 9 levels "1","2","3","4",..: 5 8 5 5 5 5 5 6 5 6 ...
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
##  $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...
##  $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...
##  $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
##  $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...
##  $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...
##  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...
##  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...
##  $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
##  $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
##  $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
##  $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...
```

```
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...
##  $ Fireplaces   : Factor w/ 4 levels "0","1","2","3": 1 2 2 2 2 1 2 3 3 3 ...
##  $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...
##  $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
##  $ GarageCars   : Factor w/ 5 levels "0","1","2","3",..: 3 3 3 4 4 3 3 3 3 2 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...
##  $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
##  $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...
##  $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
##  $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...
##  $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

Identification missing values

```
mv_summary2        <-  data.frame('variables name' = colnames(data))
mv_summary2$freq   <- apply(data , 2 , function(x)  sum(is.na(x)))
mv_summary2$pers   <- round(mv_summary2$freq / nrow(data) , 3) * 100
mv_summary2_1      <- as.data.frame(mv_summary2[mv_summary2$pers > 0 & mv_summary2$pers <10, ])
mv_summary2_2      <- as.data.frame(mv_summary2[mv_summary2$pers > 10 , ])
knitr::kable(mv_summary2_1)
```

	variables.name	freq	pers
26	MasVnrType	8	0.5
27	MasVnrArea	8	0.5
31	BsmtQual	37	2.5
32	BsmtCond	37	2.5
33	BsmtExposure	38	2.6
34	BsmtFinType1	37	2.5
36	BsmtFinType2	38	2.6
43	Electrical	1	0.1
59	GarageType	81	5.5
60	GarageYrBlt	81	5.5
61	GarageFinish	81	5.5
64	GarageQual	81	5.5
65	GarageCond	81	5.5

```
knitr::kable(mv_summary2_2)
```

	variables.name	freq	pers
4	LotFrontage	259	17.7
7	Alley	1369	93.8
58	FireplaceQu	690	47.3
73	PoolQC	1453	99.5
74	Fence	1179	80.8
75	MiscFeature	1406	96.3

Removing columns that have more than 10% missing value In my opinion, these columns have high missing values and cause problems in over modelling

```
t1 <- data[,c("LotFrontage","Alley","FireplaceQu" ,"Fence" ,
              "MiscFeature" , "PoolQC")]
```

```
data1 <- data[,-which(data %in% t1)]
dim(data1)
```

```
## [1] 1460   75
```

I also removed all rows that contained missing values. this is the easiest category to get rid of missing values

```
data2 <- data1[apply(data1,1,function(x) any(is.na(x))) == F,]
```

For convenience, I divided the data into continuous and discrete parts

```
cat <- data2[,c("MSZoning" , "Street" ,"LotShape" , "LandContour", "Utilities" , "LotConfig" , "LandSlo
                "Neighborhood" , "Condition1" , "Condition2" , "BldgType" ,
                "HouseStyle", "OverallQual", "OverallCond" , "RoofStyle" , "Exterior1st" , "Exterior2n
                "MasVnrType"    ,
                "ExterQual" ,"ExterCond" , "Foundation" , "BsmtQual" ,"BsmtCond" ,
                "BsmtExposure" ,"BsmtFinType1" ,"BsmtFinType2" , "Heating" , "HeatingQC" , "CentralAir
                "Electrical"  ,
                "KitchenQual" , "Functional" , "GarageType" , "GarageFinish" , "GarageCars" , "GarageQ
                "GarageCond" , "PavedDrive" ,
                "RoofMatl" ,
                "SaleType" , "SaleCondition" ,"Fireplaces" )]
```

I prefer to convert the columns that contain the date field to age, this helps me more easily determine the relationship between the price and the life of the house.

```
data3 <- data2[,-c(18,19,57,72)]
today <- as.Date("2022"  , format = "%Y")
```

```
data3$ageBuilt <- as.Date(as.character(data2$YearBuilt) , format = "%Y")

data3$ageBuilt   <- as.numeric(today - data3$ageBuilt)

data3$ageBuilt<- round(data3$ageBuilt/365 )
summary(data$ageBuilt)
```

```
## Length  Class   Mode
##      0   NULL   NULL
```

```
data3$ageRemodAdd  <- as.Date(as.character(data2$YearRemodAdd), format = "%Y")

data3$ageRemodAdd <- as.numeric(today - data3$ageRemodAdd)

data3$ageRemodAdd <- round(data3$ageRemodAdd / 365 )

summary(data3$ageRemodAdd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   18.00   27.50   36.33   54.00   72.00
```

```
data3$GarageageBlt   <- as.Date(as.character(data2$GarageYrBlt),format = "%Y")
data3$GarageageBlt   <- as.numeric(today - data3$GarageageBlt)
data3$GarageageBlt   <-round(data3$GarageageBlt / 365 )

summary(data3$GarageageBlt)
```

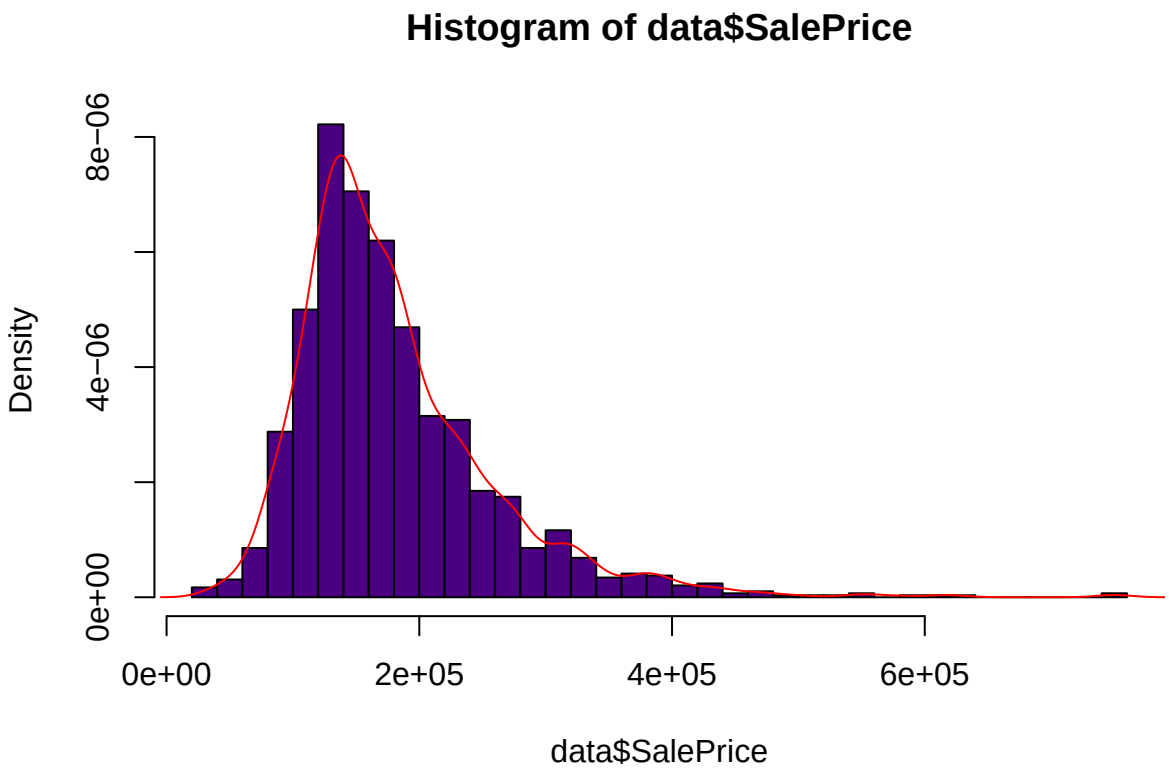
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    12.0    20.0    42.0    43.4    60.0   122.0
```

```
#Regarding the sale date, it also helps me to better recognize the rise and fall of prices
data3$ageSold  <- as.Date(as.character(data2$YrSold) , format = "%Y")
data3$ageSold  <- as.numeric(today - data3$ageSold)
data3$ageSold  <- round(data3$ageSold / 365 )
summary(data3$ageSold)
```

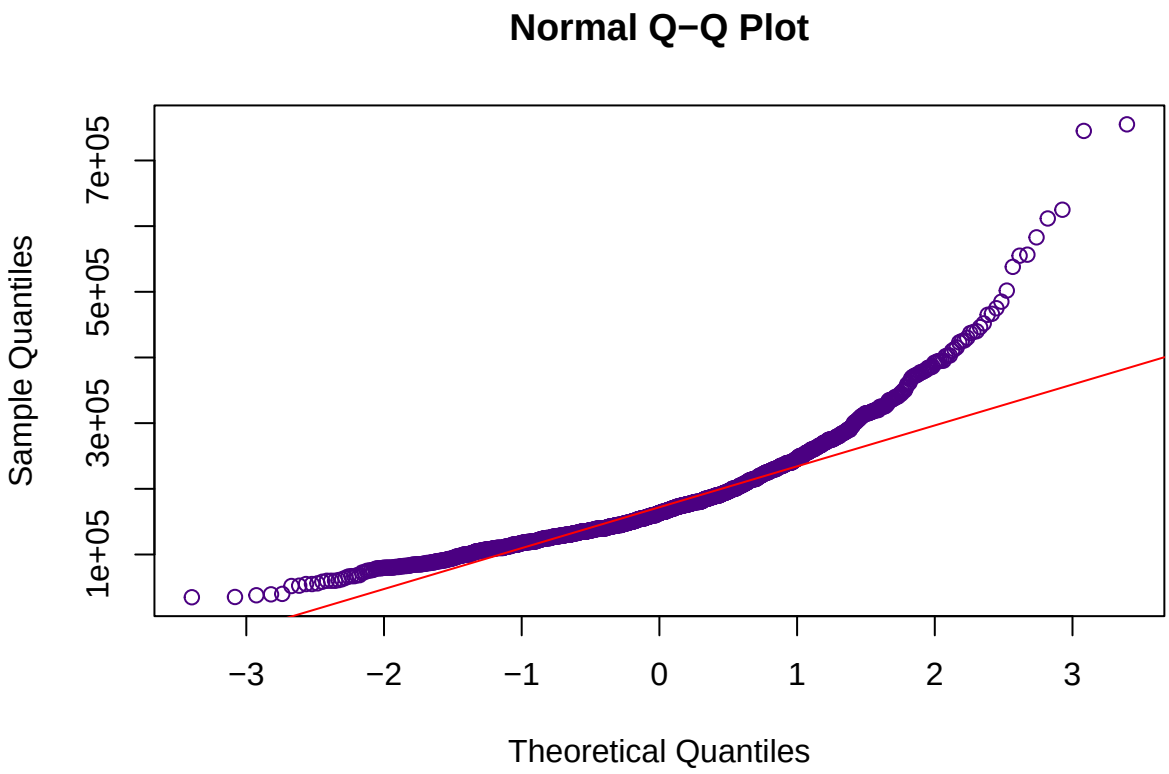
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.00   13.00   14.00   14.19   15.00   16.00
```

Examining the distribution of the response variable

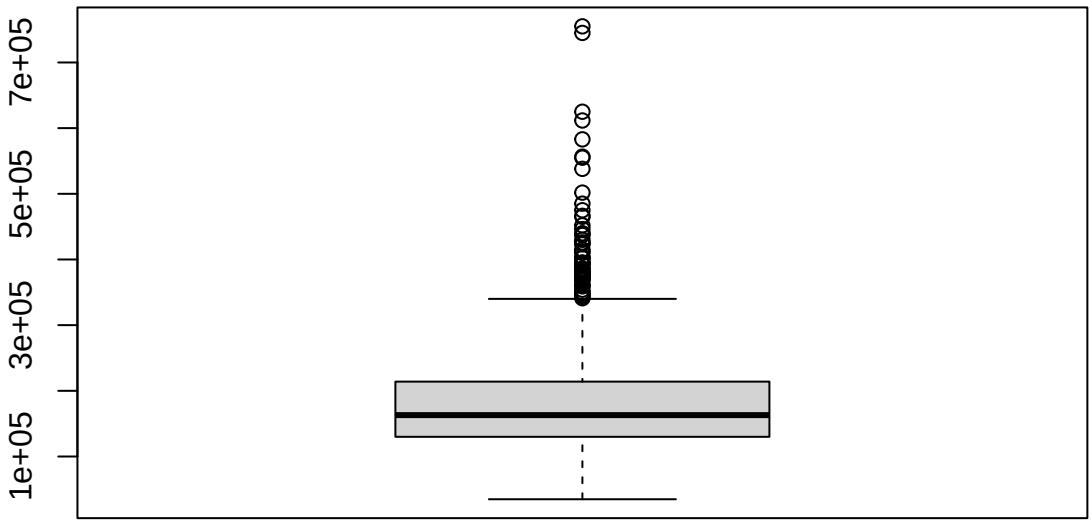
```
par(mfrow = c(1,1))
hist(data$SalePrice , breaks = 50,probability = TRUE , col = "#4B0082")
lines(density(data$SalePrice) , col = "red")
```




```
qqnorm(data$SalePrice , col = "#4B0082")
qqline(data$SalePrice , col = "red")
```




```
par(mfrow = c(1,1))
boxplot(data$SalePrice)
```




```
library("moments")

jarque.test(data$SalePrice)
```

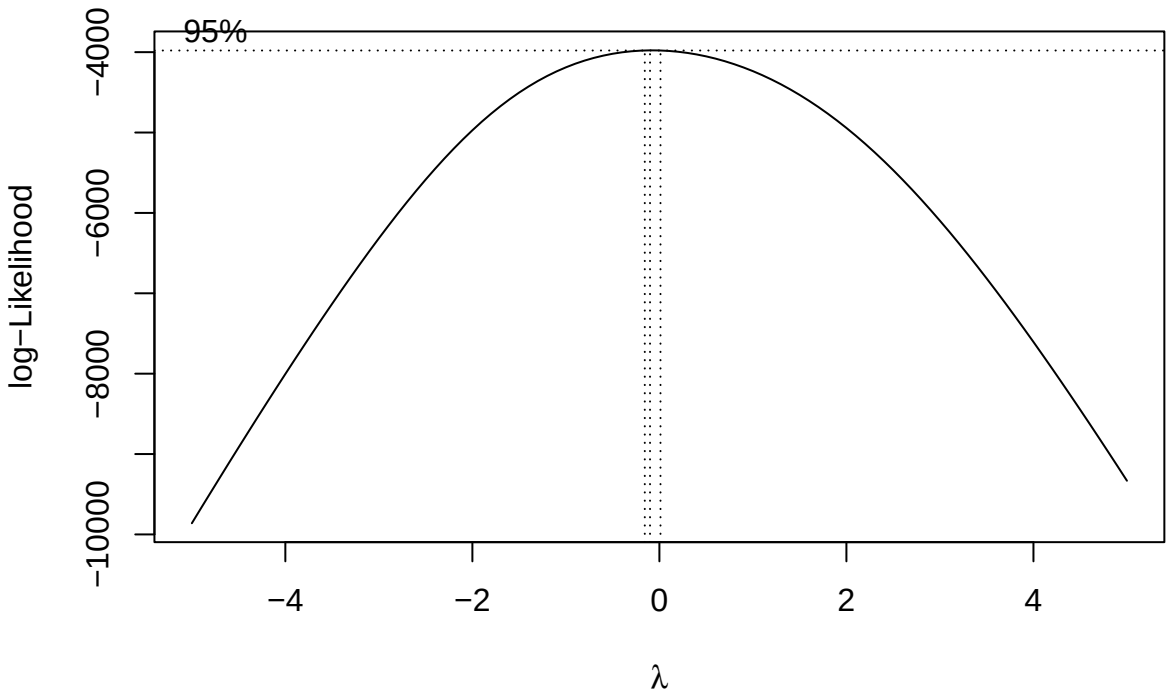
```
##
##  Jarque-Bera Normality Test
##
## data:  data$SalePrice
## JB = 3438.9, p-value < 2.2e-16
## alternative hypothesis: greater
```

```
#pvalue < 0   -> h0 regect
```

Graphs and statical tests indicate that the response variable does not follow A normal distribution which is quite reasonable for house prices

I would like to make the response variable as close to a normal distribution As possible this may help me in the modelling for this I use the box cox transformation

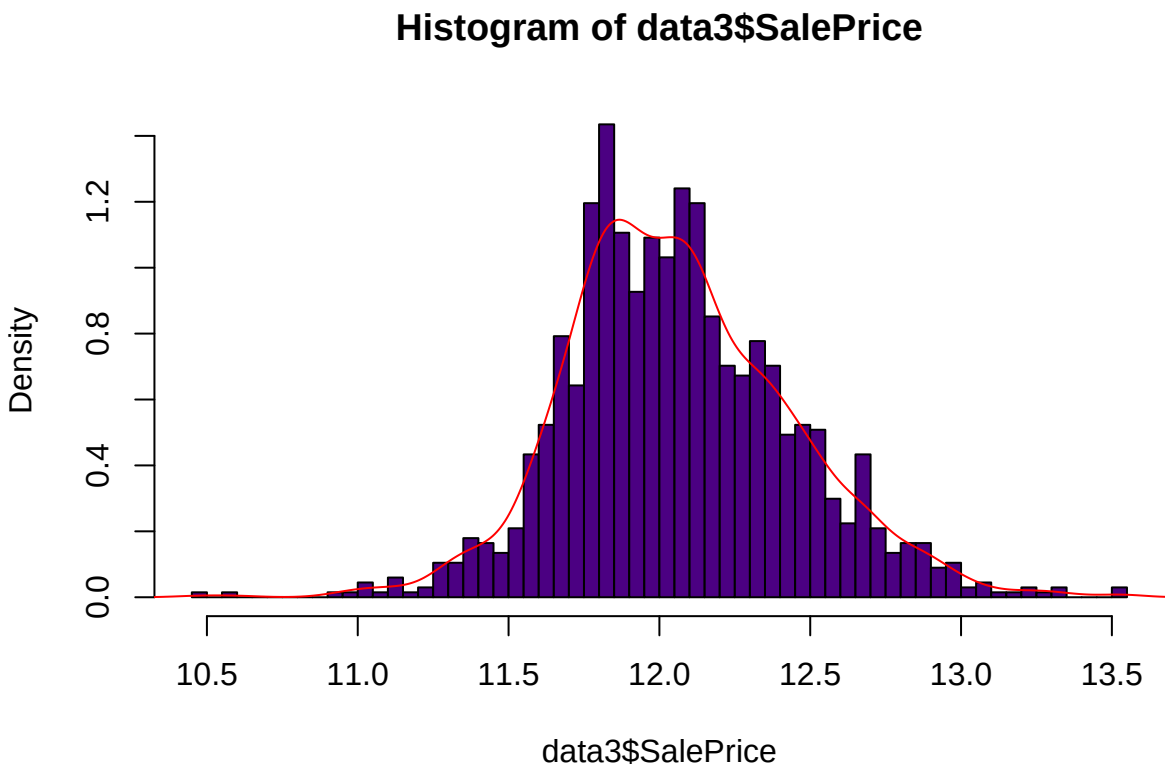
```
library("MASS")
box_result <- boxcox(data$SalePrice ~ 1 , lambda = seq(-5 , 5 , 0.1))
```



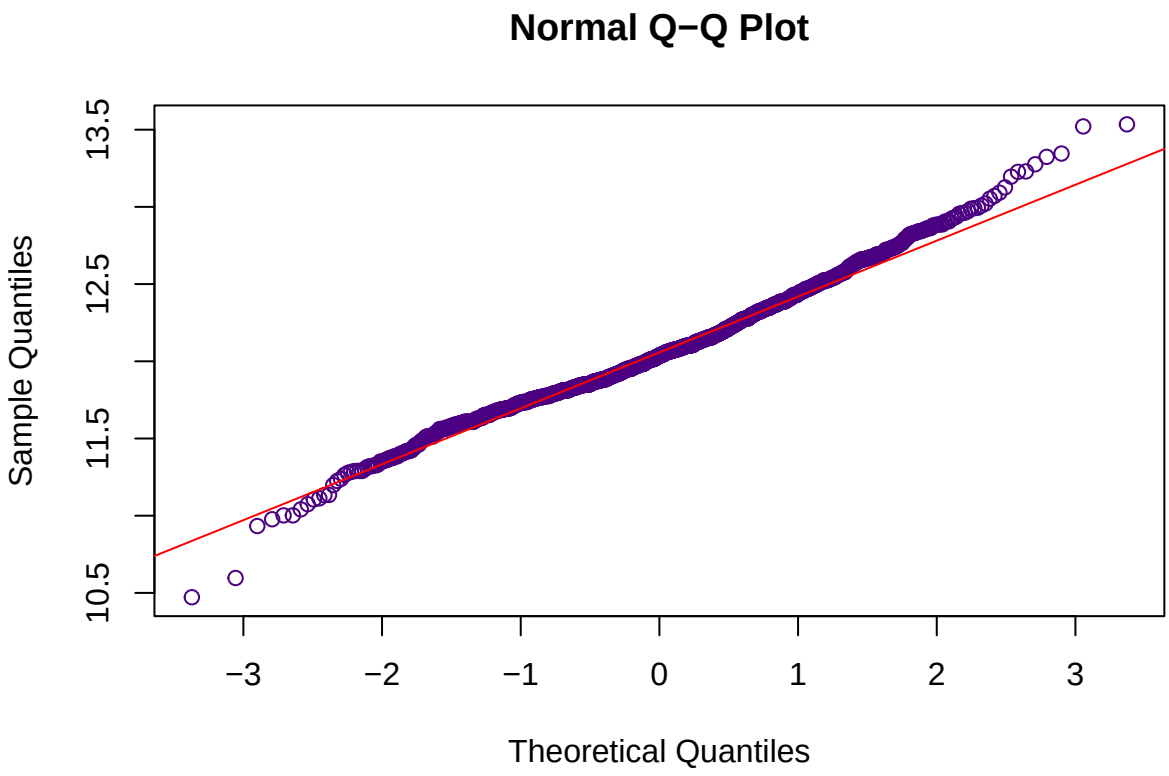

```
box_result <- data.frame(box_result)
lambda <- box_result[which(box_result$y == max(box_result$y)),]
#It observes that zero is inside the confidence interval, so I use logarithm variable change
data3$SalePrice <- log(data3$SalePrice)
```

It is clear that zero is inside this confidence interval Therefore, I use logarithmic transformation

```
hist(data3$SalePrice , breaks = 50 , probability = TRUE , col = "#4B0082")
lines(density(data3$SalePrice) , col = "red")
```




```
qqnorm(data3$SalePrice , col = "#4B0082")
qqline(data3$SalePrice , col = "red")
```



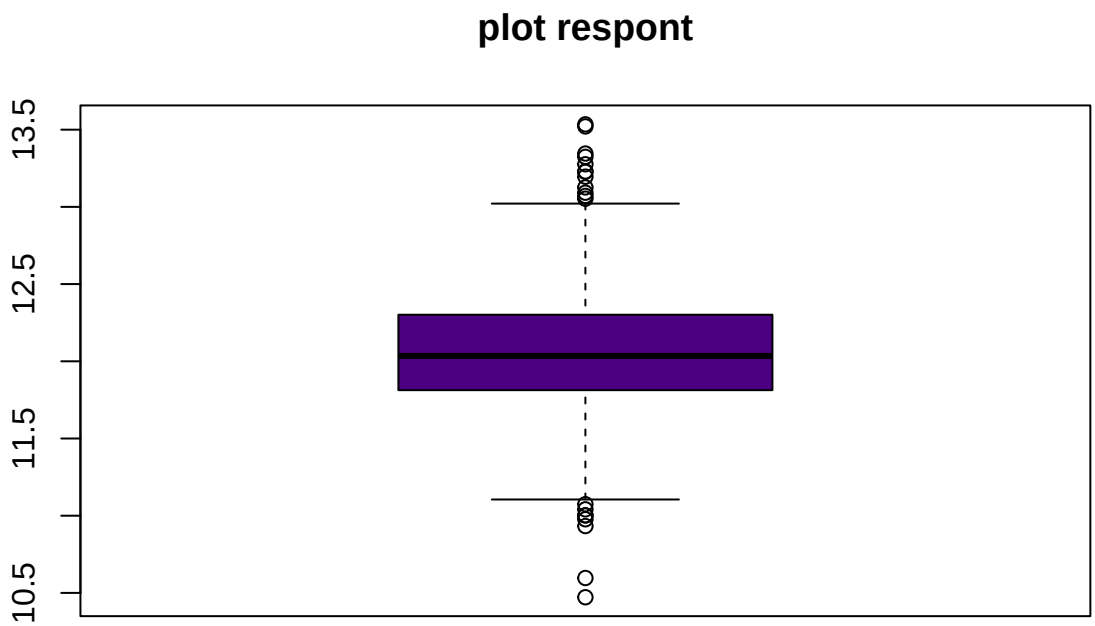

```
jarque.test(data3$SalePrice)
```

```
##
##  Jarque-Bera Normality Test
##
## data:  data3$SalePrice
## JB = 50.086, p-value = 1.33e-11
## alternative hypothesis: greater
```

```
anscombe.test(data3$SalePrice)
```

```
##
##  Anscombe-Glynn kurtosis test
##
## data:  data3$SalePrice
## kurt = 3.750, z = 4.278, p-value = 1.886e-05
## alternative hypothesis: kurtosis is not equal to 3
```

```
par(mfrow = c(1,1))
boxplot(data3$SalePrice ,main = "plot respont", col = "#4B0082")
```



It seems that the data is far from the normal distribution

```
tukey_u <- quantile(data3$SalePrice , probs = 0.75) + 1.5 * IQR(data3$SalePrice)

sum(data3$SalePrice > tukey_u)
```

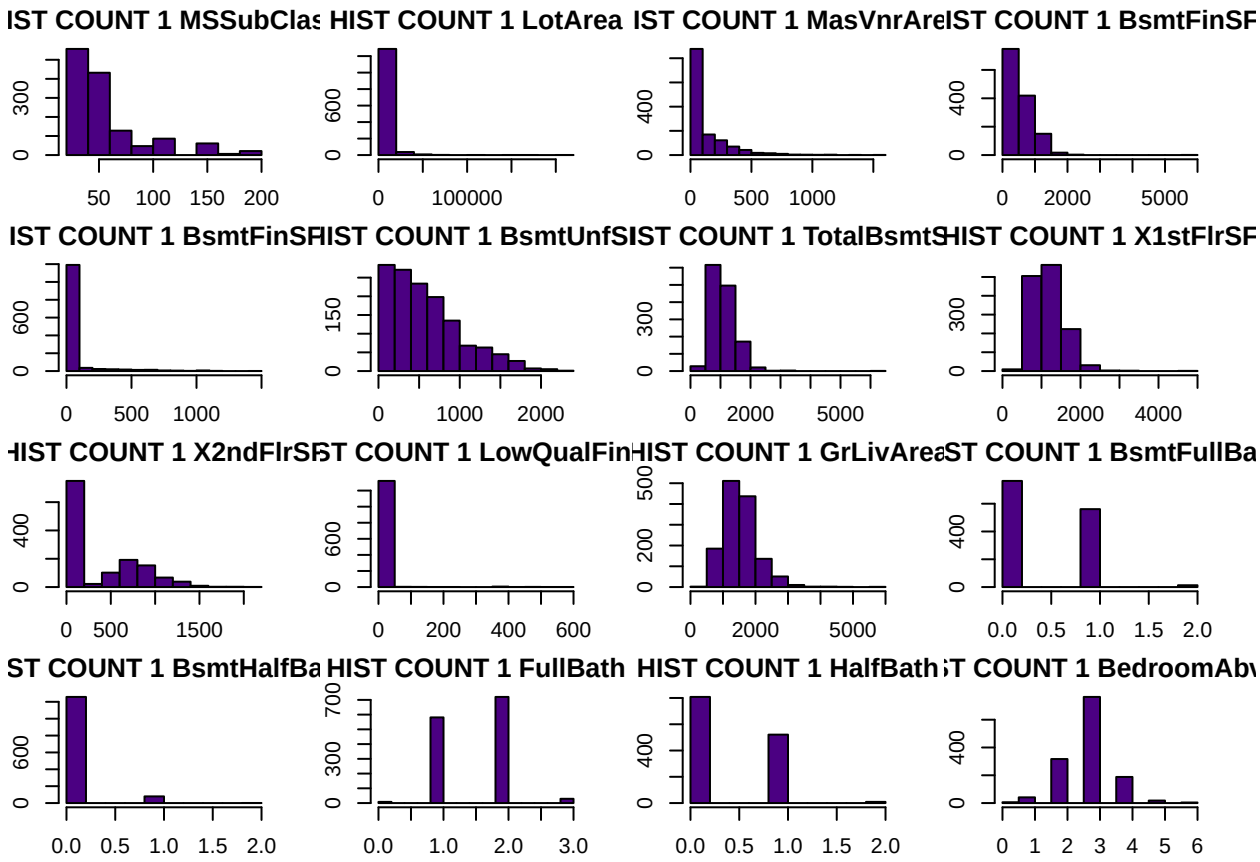
```
## [1] 12
```

For convenience, I divided the data into continuous and discrete parts

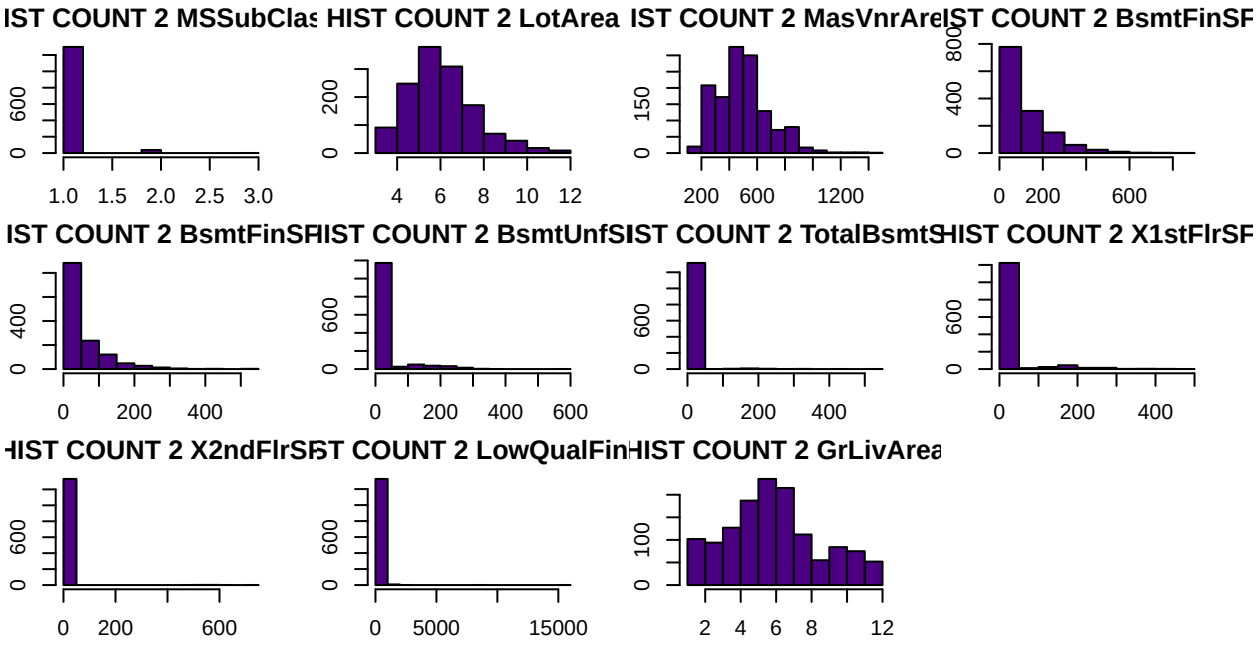
```
count <- data3[,-which(data3 %in% cat)]
cunt1 <- count[,c(1:17)]
cunt2 <- count[,c(18:29)]

cunt1 <- cunt1[,-1]
cunt1$SalePrice  <- cunt2$SalePrice
```

Histograms relate to continuous variables

```
par(mar = c(2,2,2,2))
par(mfrow = c(4,4))
for (i in 1:16 ) {
  hist(cunt1[,i] , xlab =   "" ,col = "#4B0082", main = paste("HIST COUNT 1" ,
                                                              names(cunt1[i])))
}
```




```
par(mar = c(2,2,2,2))
par(mfrow = c(4,4))
for (i in 1:11) {
  hist(cunt2[,i] , xlab = "" , col = "#4B0082" , main = paste("HIST COUNT 2" ,
                                                              names(cunt1[i])))
}
```

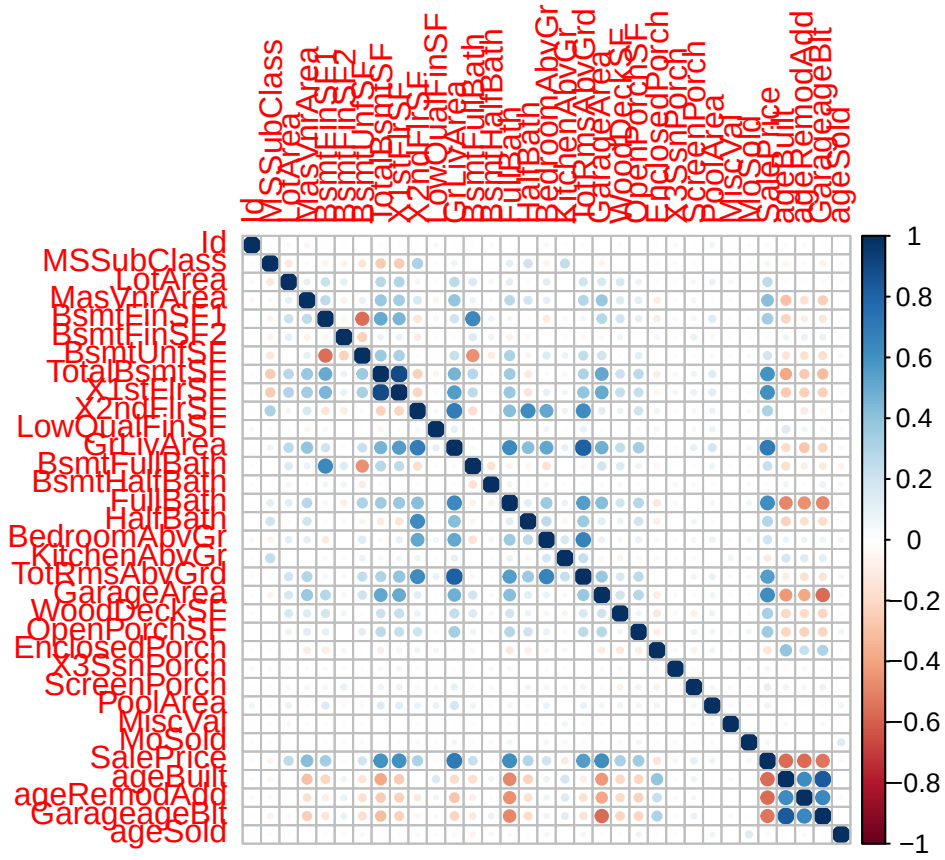


Correlation of continuous variables versus response variable

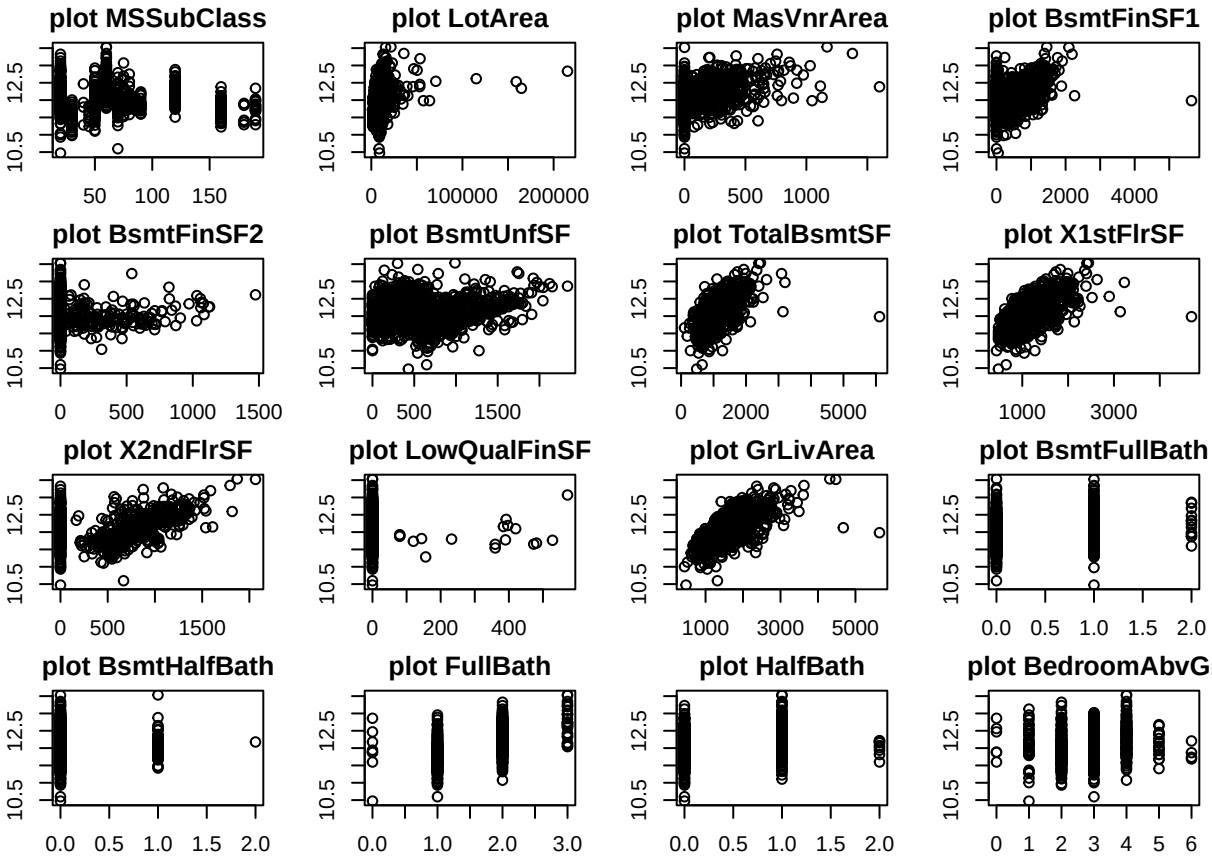
```
cros_tab <- round(cor(count),2)

library("corrplot")
```

```
## corrplot 0.92 loaded
```

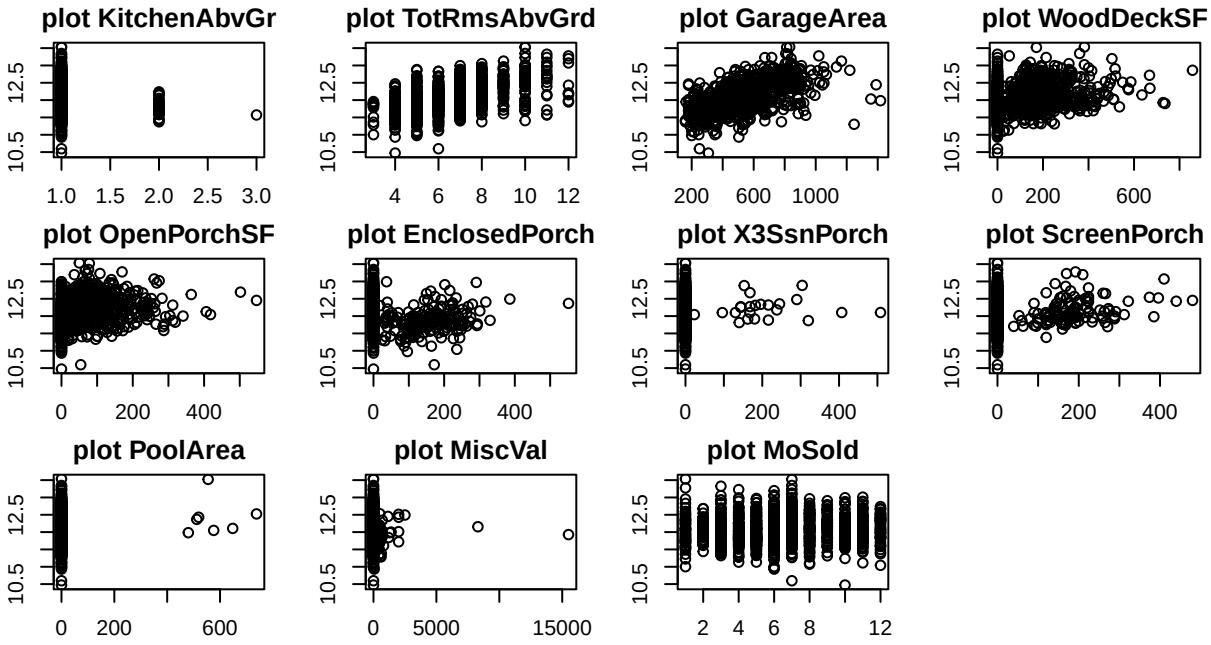
```
par(mfrow = c(1,1))
corrplot(cros_tab)
```




```
par(mar=c(2,2,2,2))
par(mfrow = c(4,4))
for (i in 1:16) {
  plot(cunt1[,i] , cunt2$SalePrice,xlab = "" , main = paste("plot" ,names(cunt1)[i] ))
}
```




```
par(mar=c(2,2,2,2))
par(mfrow = c(4,4))
for (i in 1:11) {
  plot(cunt2[,i] , cunt2$SalePrice,xlab = "" , main = paste("plot" ,names(cunt2)[i] ))
}
```



We go to the distribution of discrete variables

```
categori <- data3[,which( data3 %in% cat)]
colnames(categori)
```

```
##  [1] "MSZoning"      "Street"        "LotShape"      "LandContour"
##  [5] "Utilities"     "LotConfig"     "LandSlope"     "Neighborhood"
##  [9] "Condition1"    "Condition2"    "BldgType"      "HouseStyle"
## [13] "OverallQual"   "OverallCond"   "RoofStyle"     "RoofMatl"
## [17] "Exterior1st"   "Exterior2nd"   "MasVnrType"    "ExterQual"
## [21] "ExterCond"     "Foundation"    "BsmtQual"      "BsmtCond"
## [25] "BsmtExposure"  "BsmtFinType1"  "BsmtFinType2"  "Heating"
## [29] "HeatingQC"     "CentralAir"    "Electrical"    "KitchenQual"
## [33] "Functional"    "Fireplaces"    "GarageType"    "GarageFinish"
## [37] "GarageCars"    "GarageQual"    "GarageCond"    "PavedDrive"
## [41] "SaleType"      "SaleCondition"
```

```
dim(categori)
```

```
## [1] 1338   42
```

Examining the distribution of discrete variables

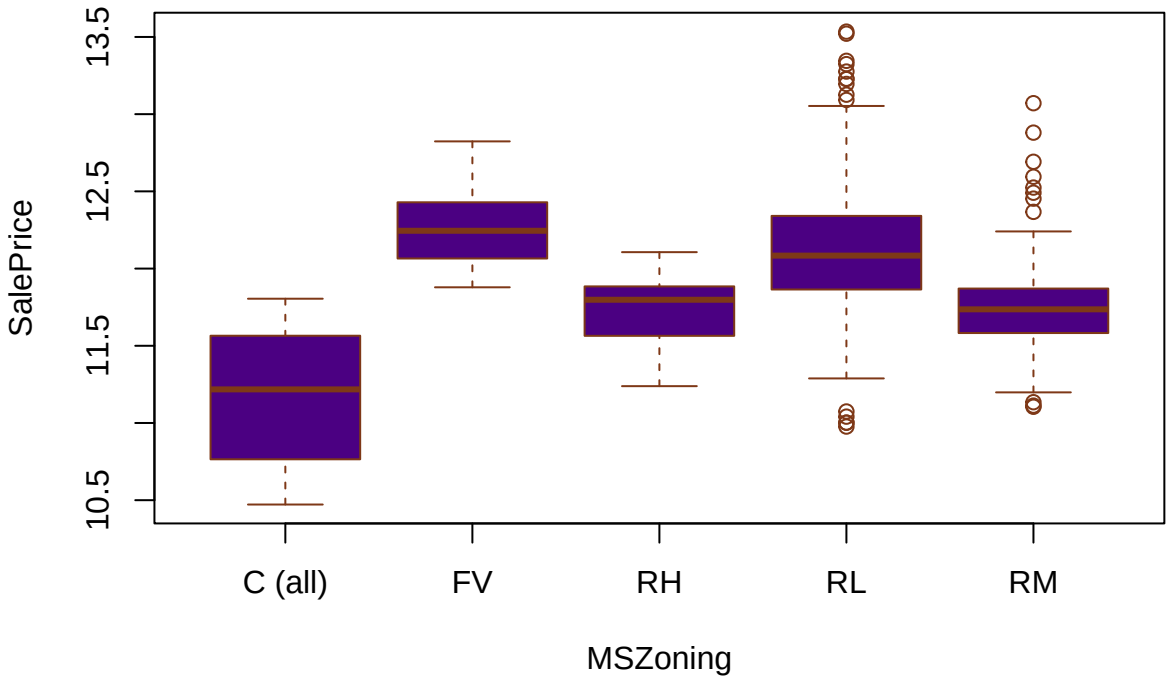
It is observed that residential low-density and residential following villages have the highest average price, of course, it is obvious that these two types have a higher frequency than the other

```
table(data3$MSZoning)
```

```
##
```

```
## C (all)      FV      RH      RL      RM
##       8      62      11    1066     191
```

```
boxplot(SalePrice ~ MSZoning , data = data3 , col = "#4B0082" , border = "#7E3817")
```

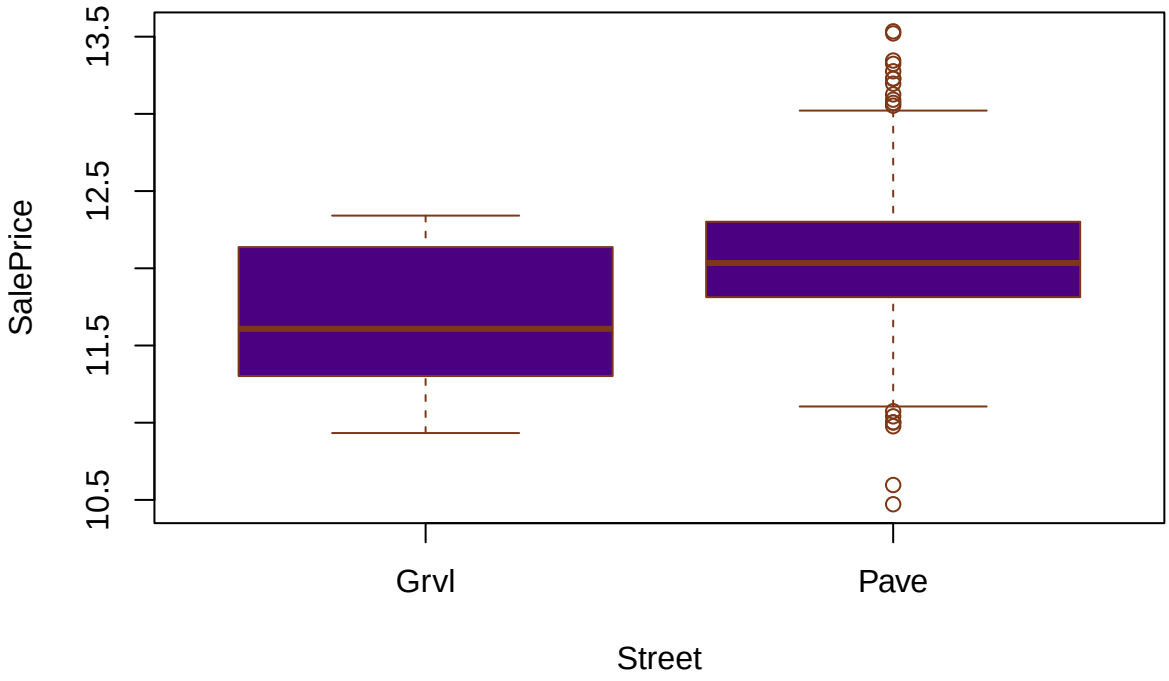


It is observed that residential low-density and residential following villages have the highest average price, of course, it is obvious that these two types have a higher frequency than the other

```
table(data3$Street)
```

```
##
## Grvl Pave
##    5 1333
```

```
boxplot(SalePrice ~  Street , data = data3 , col = "#4B0082" , border = "#7E3817")
```

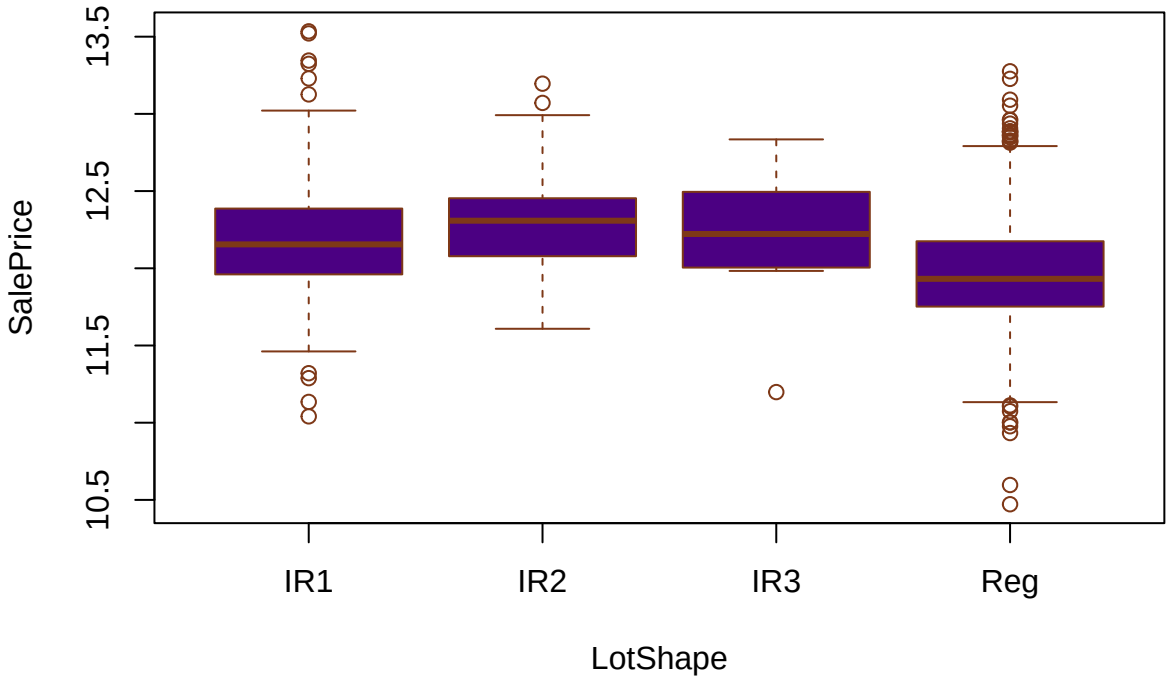


Buildings that are located on paved streets have a higher average price,although according to the table it is evident that they are mor expensive than other type, so no decision can be made about this

```
table(data3$LotShape)
```

```
##
## IR1 IR2 IR3 Reg
## 459  40  10 829
```

```
boxplot(SalePrice ~ LotShape , data = data3 , col = "#4B0082" , border = "#7E3817")
```

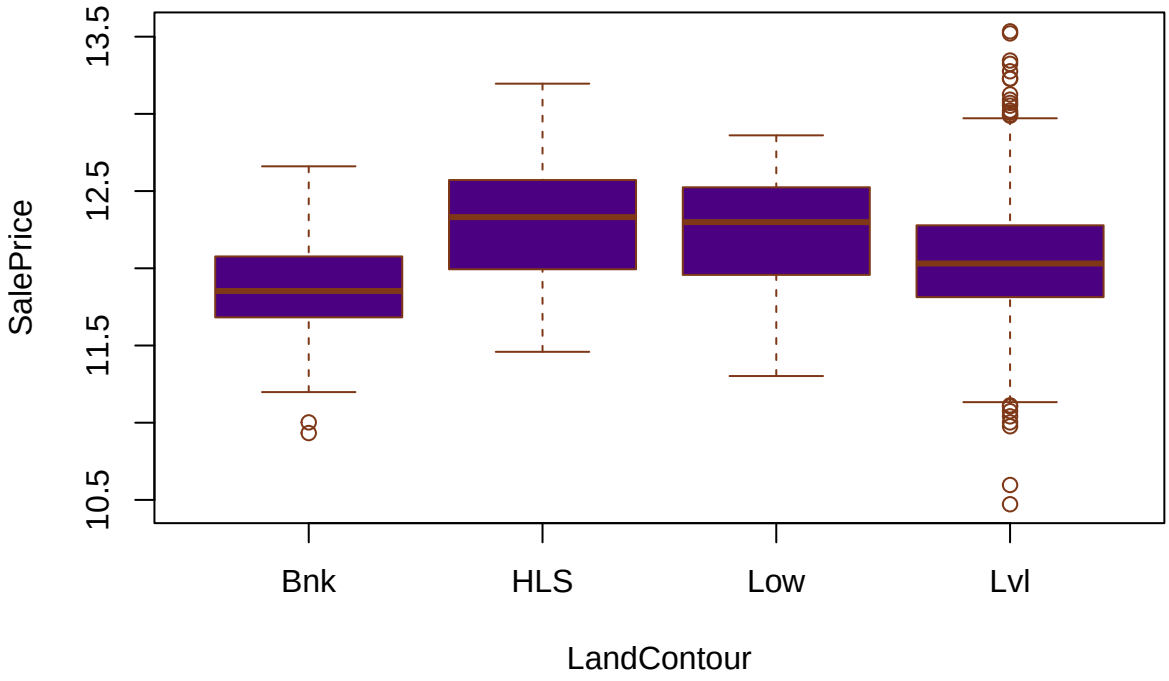


if the shape of the building more regular, it will have higher average

```
table(data3$LandContour)
```

```
##
##  Bnk  HLS  Low  Lvl
##   52   48   32 1206
```

```
boxplot(SalePrice ~ LandContour , data = data3 , col = "#4B0082" , border = "#7E3817")
```

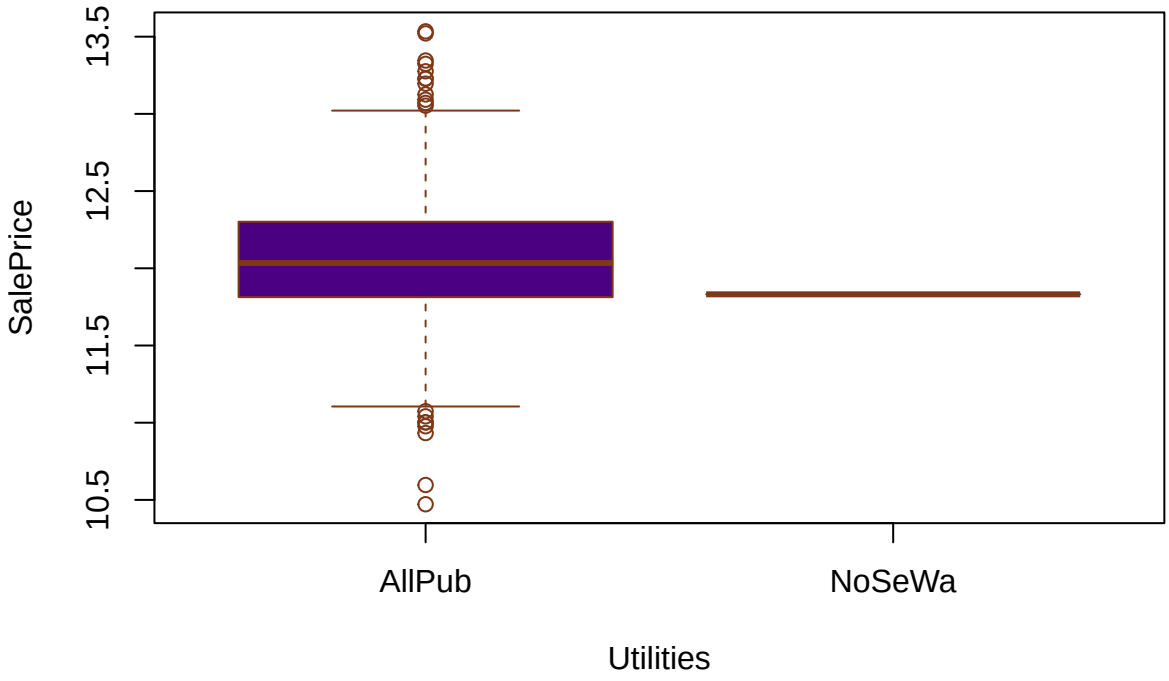


buildings that are higer than the grouand level (compared to the street) they supposed to have higher average price

```
table(data3$Utilities)
```

```
##
## AllPub NoSeWa
##   1337      1
```

```
boxplot(SalePrice ~  Utilities , data = data3 , col = "#4B0082" , border = "#7E3817")
```

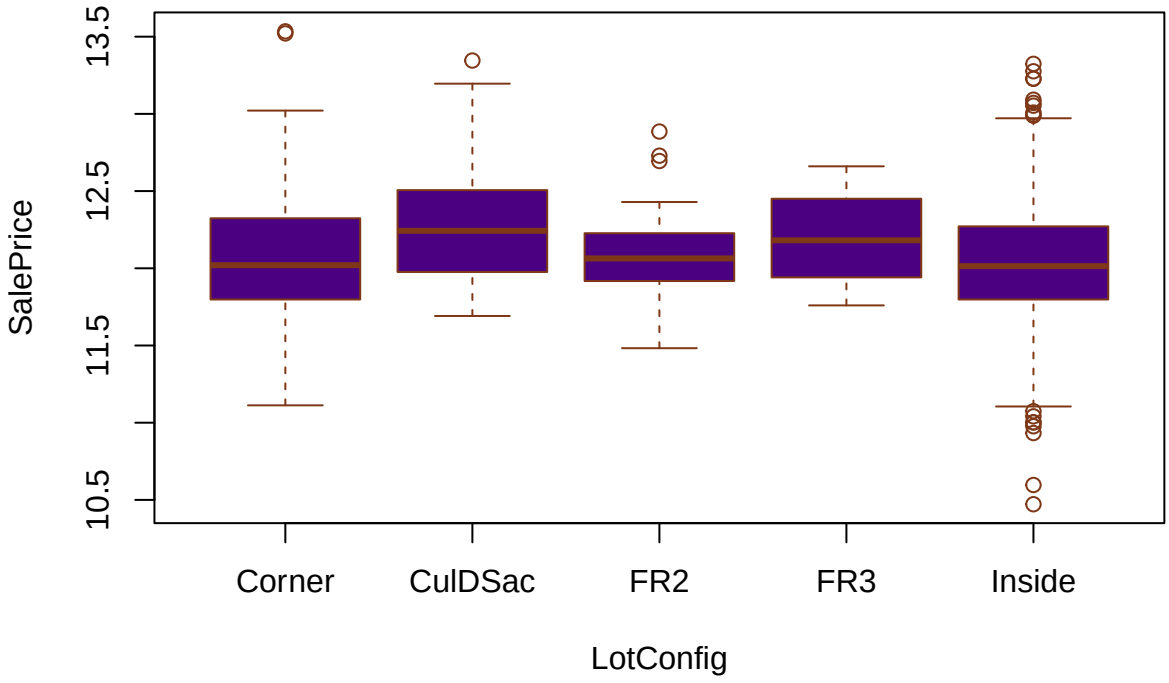


I cant make any decisions about this because I only have one sample of the nosewat type _ I'll probablity drop this variable altogether because it cant be a good explanation

```
table(data3$LotConfig)
```

```
##
##  Corner CulDSac     FR2     FR3  Inside
##     244      90      43       4     957
```

```
boxplot(SalePrice ~  LotConfig , data = data3 , col = "#4B0082" , border = "#7E3817")
```

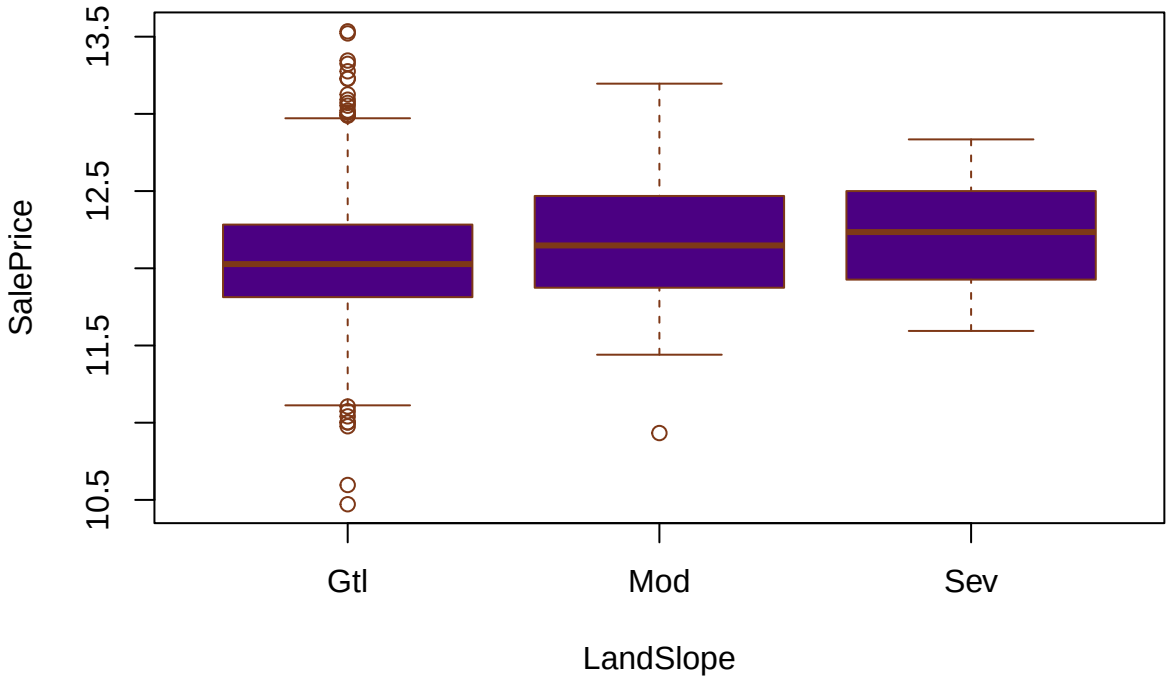


this variable vague for me, may be I will catch some thing in the moddeling process. Ather wise I'll delete irt

```
table(data3$LandSlope)
```

```
##
##  Gtl  Mod  Sev
## 1265   61   12
```

```
boxplot(SalePrice ~  LandSlope , data = data3 , col = "#4B0082" , border = "#7E3817")
```

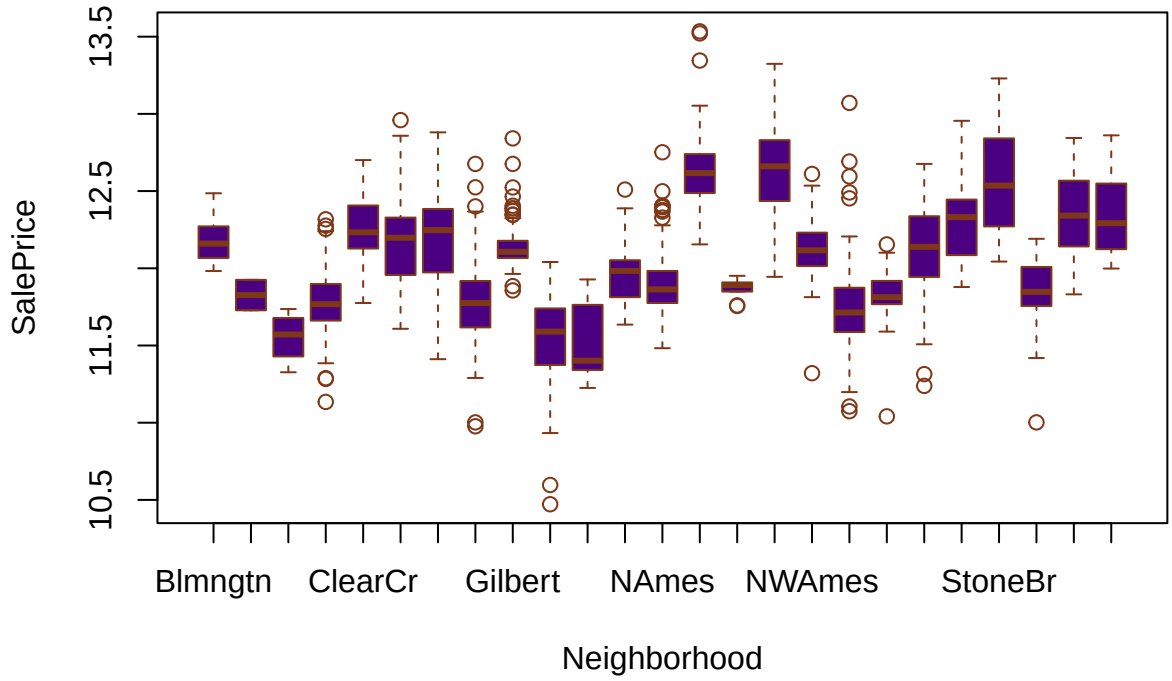


the avrage of the variable is close to each other, and it seems that the presence of outlier variable in the gentle slope is due to the high frequency of this type, according to observing image of Ames, we found that this city is flat - in the case of colinearity I will delete it

```
table(data3$Neighborhood)
```

```
##
## Blmngtn Blueste  BrDale BrkSide ClearCr CollgCr Crawfor Edwards Gilbert  IDOTRR
##      17       2      15      47      26     146      50      70      77      29
## MeadowV Mitchel   NAmes NoRidge NPkVill NridgHt  NWAmes OldTown  Sawyer SawyerW
##      12      42     209      41       9      75      73     100      69      53
## Somerst StoneBr   SWISU  Timber Veenker
##      83      25      20      37      11
```

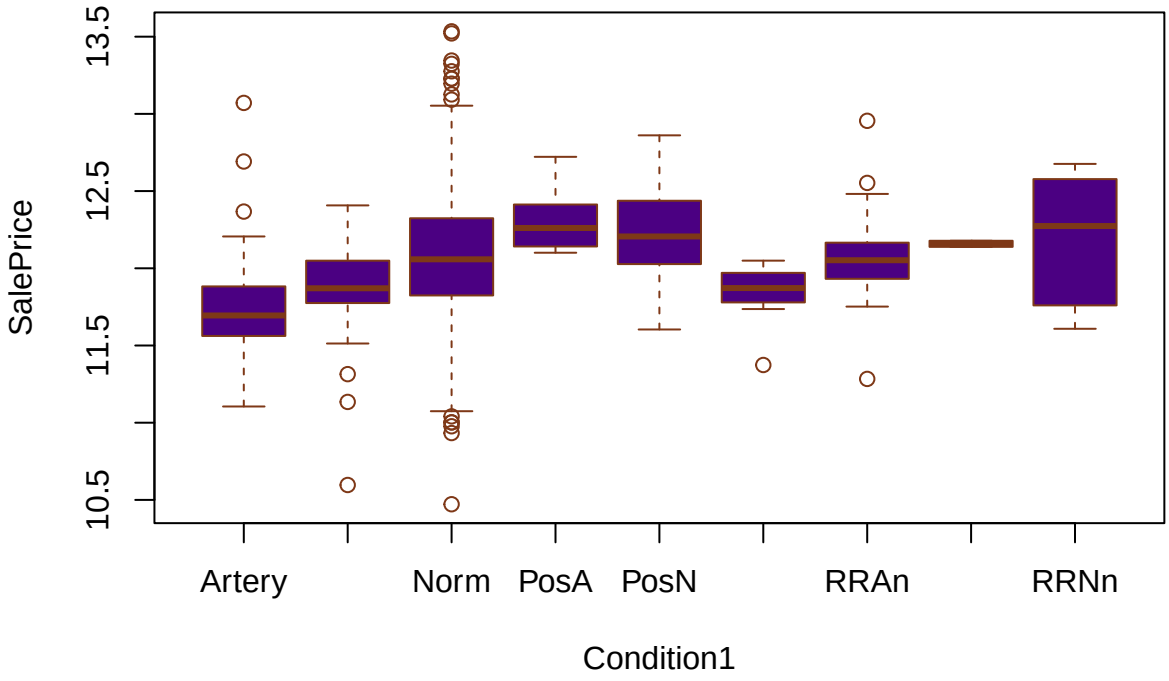
```
boxplot(SalePrice ~ Neighborhood , data = data3 , col = "#4B0082" , border = "#7E3817")
```




```
table(data3$Condition1)
```

```
##
## Artery  Feedr   Norm   PosA   PosN   RRAe   RRAn   RRNe   RRNn
##     43     63   1162      8     19     10     26      2      5
```

```
boxplot(SalePrice ~ Condition1 , data = data3 , col = "#4B0082" , border = "#7E3817")
```

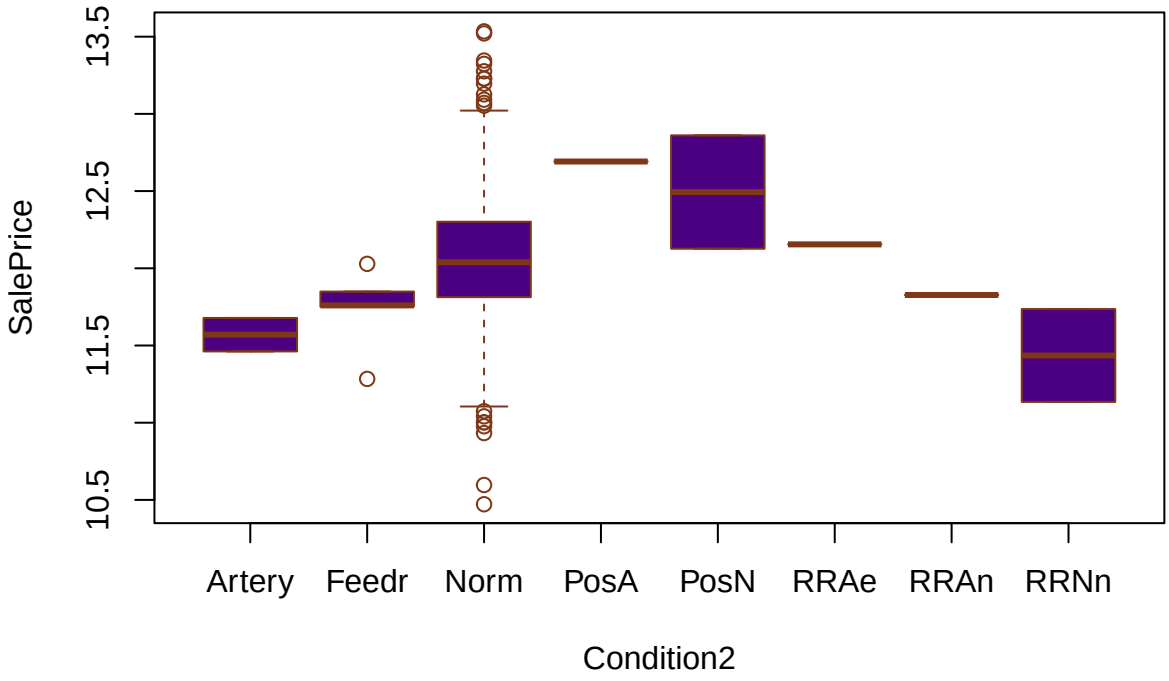


most of the building are in normal location. In the case of(arteria street) I suspect that the noise of the cars is involved in lowering the price

```
table(data3$Condition2)
```

```
##
## Artery  Feedr   Norm   PosA   PosN   RRAe   RRAn   RRNn
##      2      5   1324      1      2      1      1      2
```

```
boxplot(SalePrice ~ Condition2 , data = data3 , col = "#4B0082" , border = "#7E3817")
```

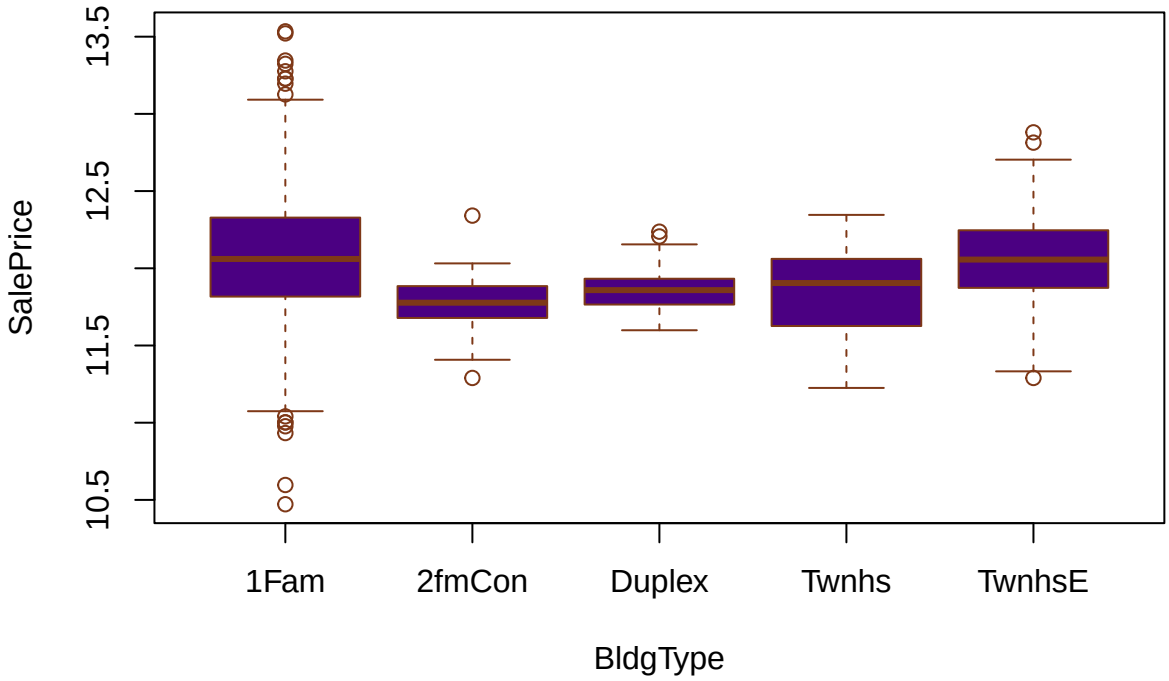


conditions1 and 2 are very similar and I have the same opinion about these two conditions

```
table(data3$BldgType)
```

```
##
##   1Fam 2fmCon Duplex  Twnhs TwnhsE
##   1138     22     28     38    112
```

```
boxplot(SalePrice ~ BldgType , data = data3 , col = "#4B0082" , border = "#7E3817")
```

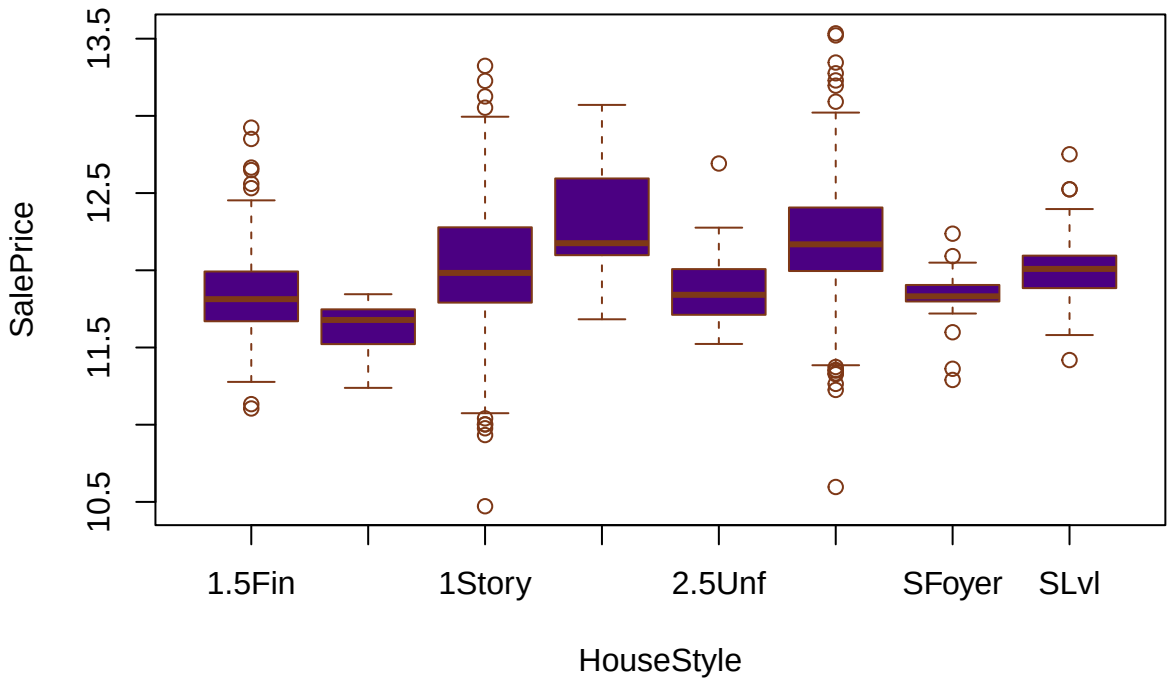


single_family detached house and town houses inside units have a higher average price than others.and the houses that have been converted for these two families_ have the Lowest price

```
table(data3$HouseStyle)
```

```
##
## 1.5Fin 1.5Unf 1Story 2.5Fin 2.5Unf 2Story SFoyer   SLvl
##    134     11    657      6     10    426     30     64
```

```
boxplot(SalePrice ~ HouseStyle , data = data3 , col = "#4B0082" , border = "#7E3817")
```

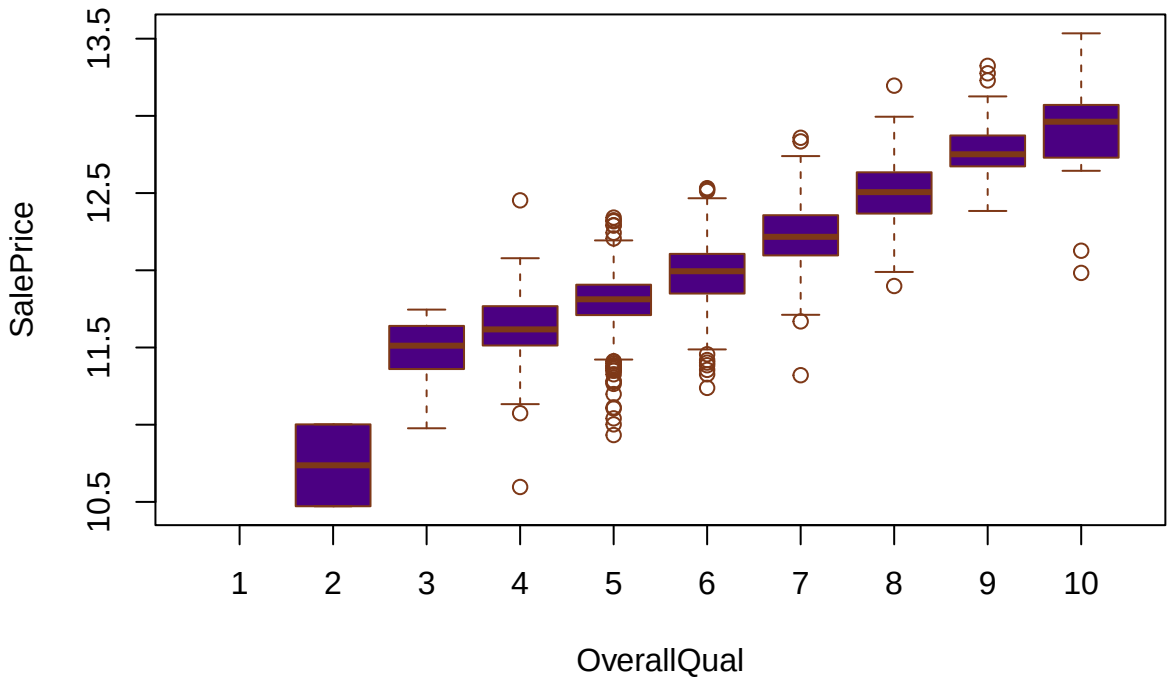


house style raises all issues related to the high square footage and floors of the average house price

```
table(data3$OverallQual)
```

```
##
##   1   2   3   4   5   6   7   8   9  10
##   0   2   8  81 351 359 312 165  43  17
```

```
boxplot(SalePrice ~ OverallQual , data = data3 , col = "#4B0082" , border = "#7E3817")
```

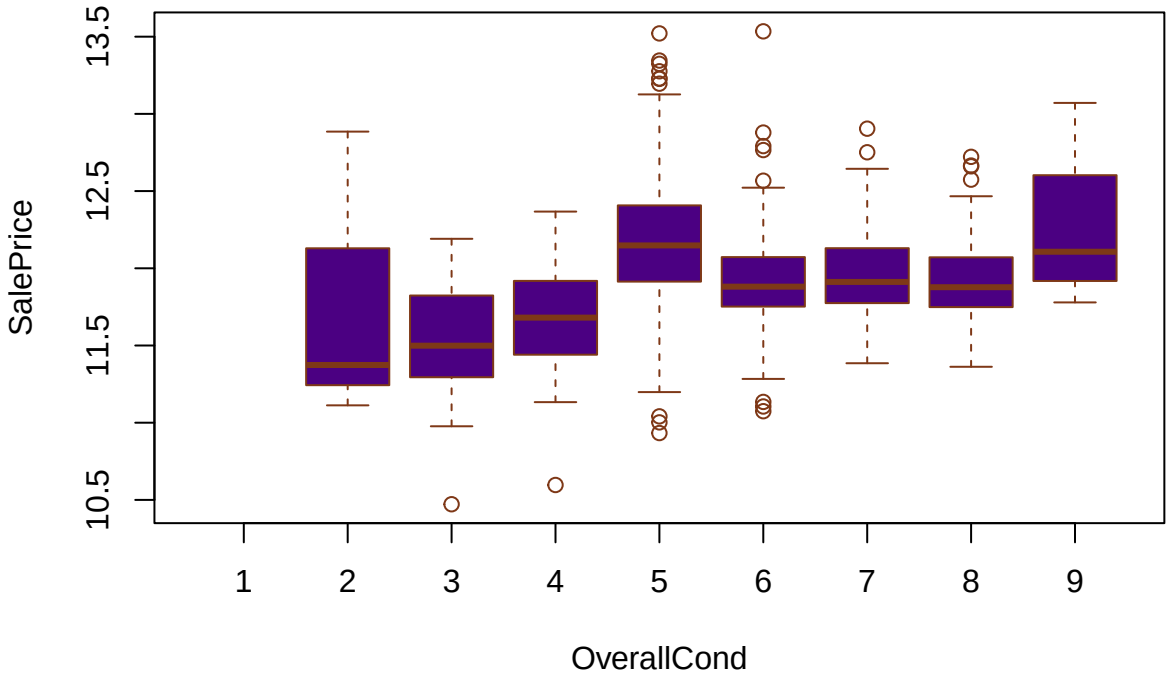


the overall quality of the building raises the average price and standard deviation

```
table(data3$OverallCond)
```

```
##
##   1   2   3   4   5   6   7   8   9
##   0   3  15  46 770 233 183  68  20
```

```
boxplot(SalePrice ~ OverallCond , data = data3 , col = "#4B0082" , border = "#7E3817")
```

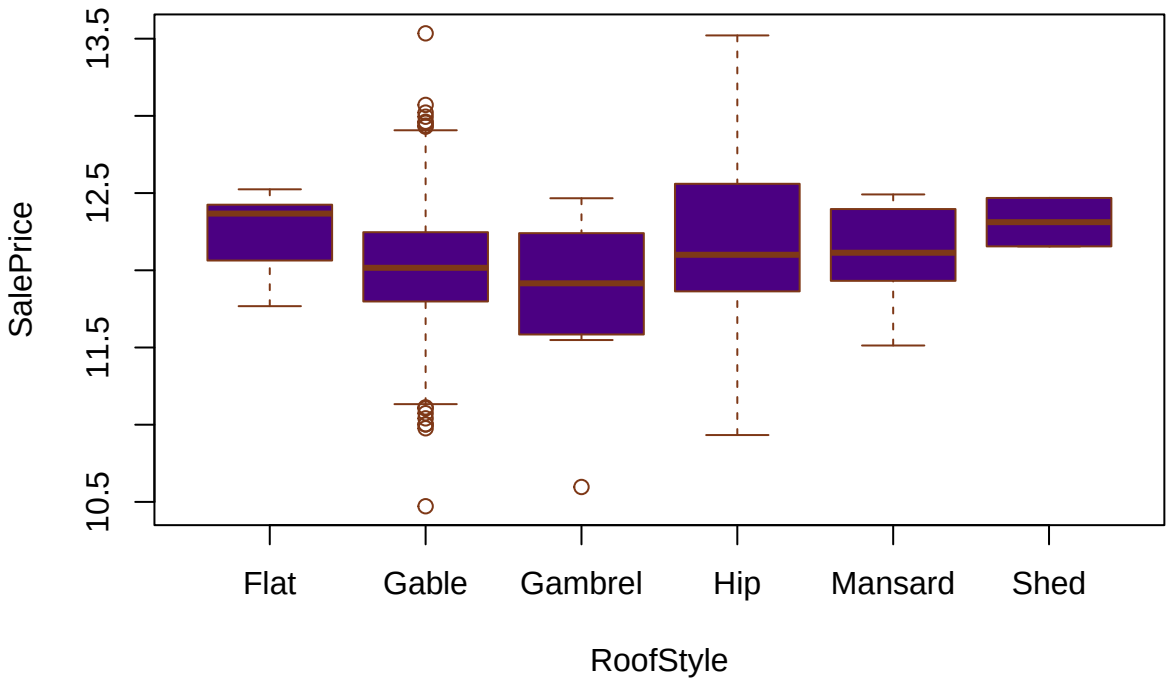


of course the better the overall condition of the house, the higher the average price should be, but what is evident here is that houses with the average condition have a higher average price, maybe it is because of the abundance of this group

```
table(data3$RoofStyle)
```

```
##
##    Flat   Gable Gambrel     Hip Mansard    Shed
##      11    1037      10     272       6       2
```

```
boxplot(SalePrice ~ RoofStyle , data = data3 , col = "#4B0082"  , border = "#7E3817")
```

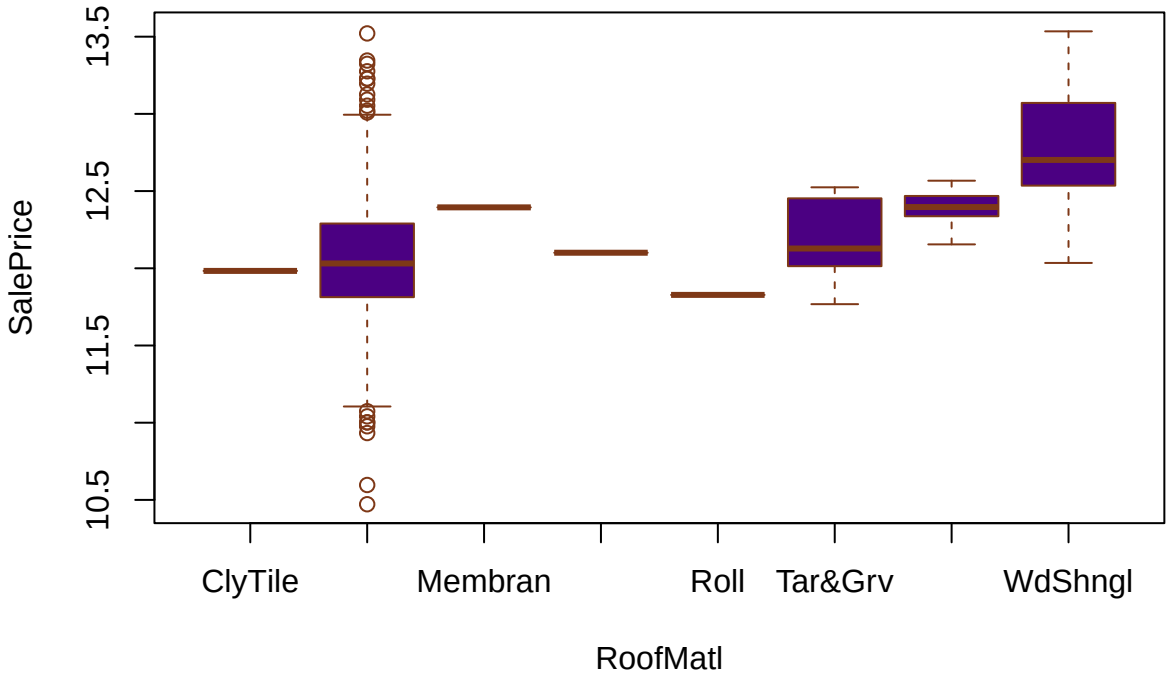


the average is close to each other

```
table(data3$RoofMatl)
```

```
##
## ClyTile CompShg Membran   Metal    Roll Tar&Grv WdShake WdShngl
##       1    1314       1       1       1       9       5       6
```

```
boxplot(SalePrice ~  RoofMatl , data = data3 , col = "#4B0082" , border = "#7E3817")
```

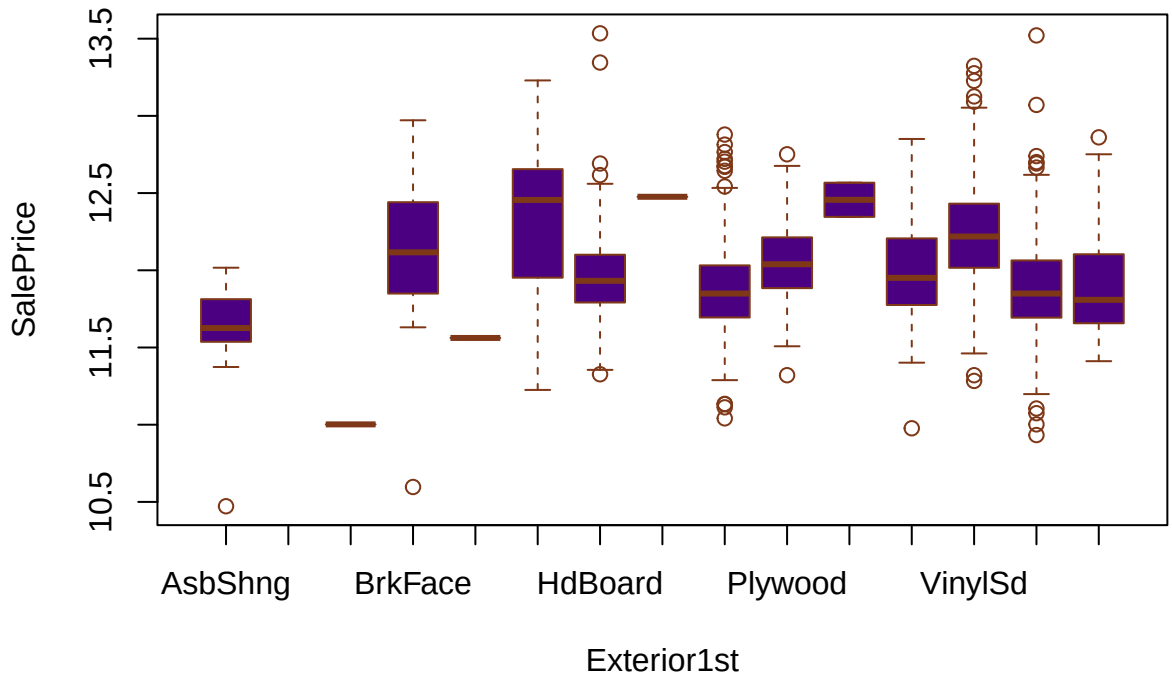


almost all types of standard (composite) single has been formed. I don't this variable is a good explanation think

```
table(data3$Exterior1st)
```

```
##
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard ImStucc MetalSd Plywood
##      15       0       1      44       1      52     211       1     201     100
##   Stone  Stucco VinylSd Wd Sdng WdShing
##       2      21     486     183      20
```

```
boxplot(SalePrice ~ Exterior1st , data = data3 , col = "#4B0082" , border = "#7E3817")
```

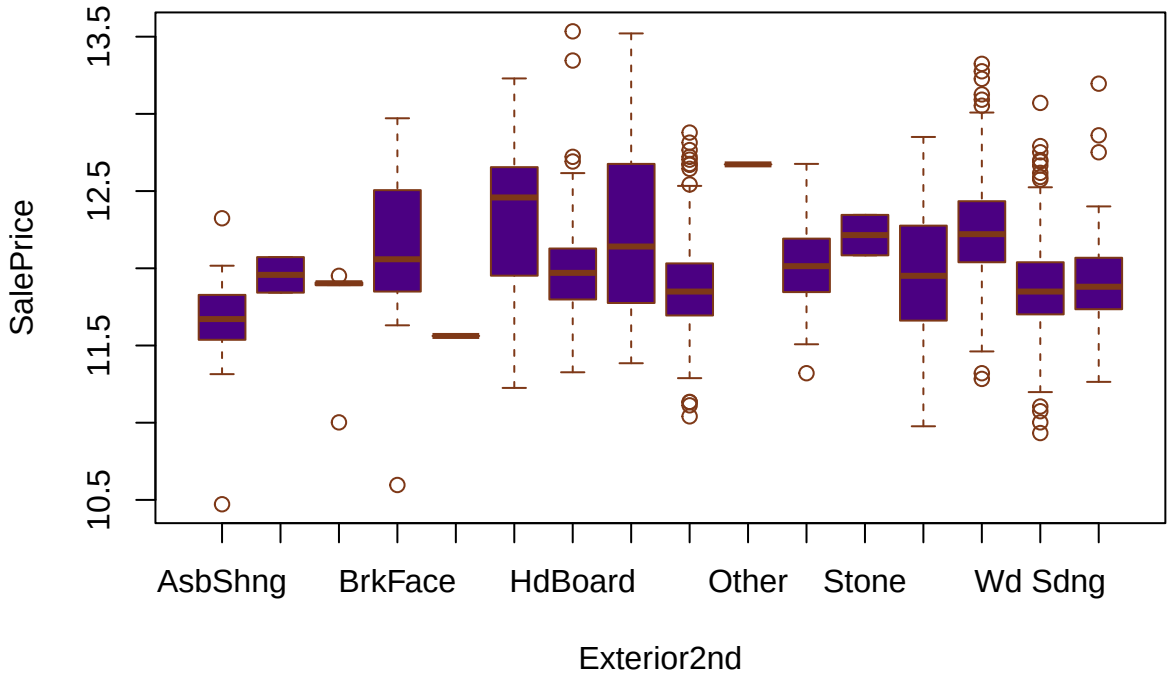


Houses covered with stone and concrete have a relatively higher average price than house covered with wood

```
table(data3$Exterior2nd)
```

```
##
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd   Other
##      16       2       6      22       1      51     197      10     197       1
## Plywood   Stone  Stucco VinylSd Wd Sdng Wd Shng
##     127       2      23     475     176      32
```

```
boxplot(SalePrice ~ Exterior2nd , data = data3 , col = "#4B0082" , border = "#7E3817")
```

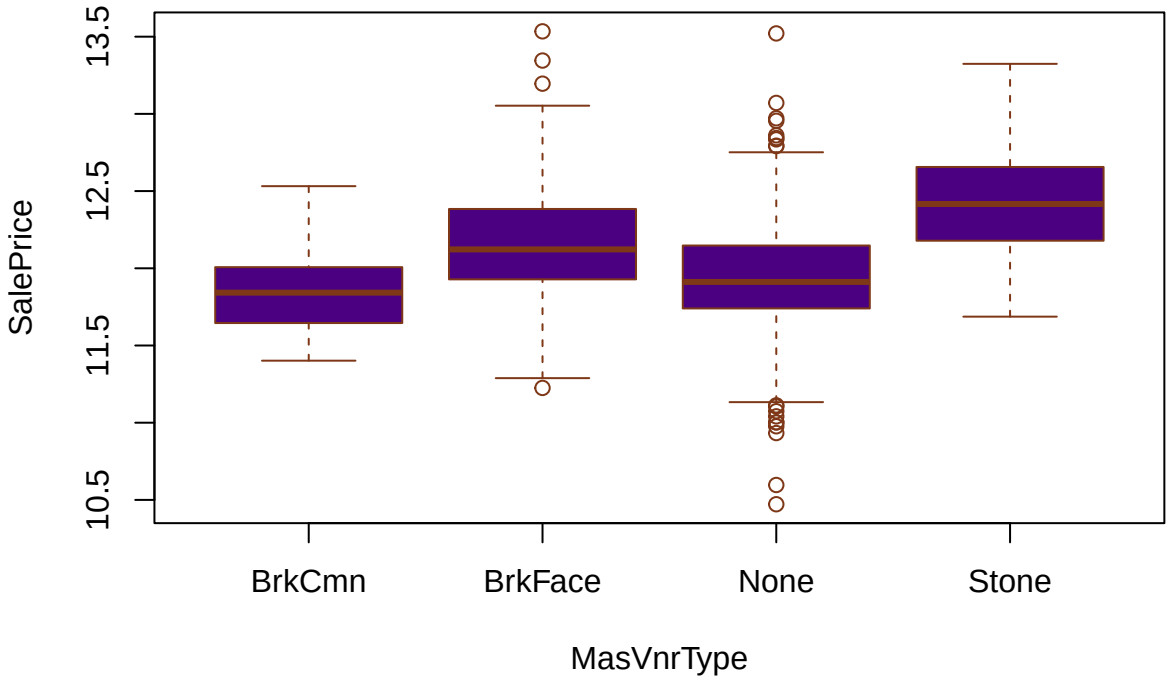


Vinyl siding covers most of the house, the rest of the covers are around this average

```
table(data3$MasVnrType)
```

```
##
##  BrkCmn BrkFace    None   Stone
##      15     432     763     128
```

```
boxplot(SalePrice ~ MasVnrType , data = data3 , col = "#4B0082" , border = "#7E3817")
```

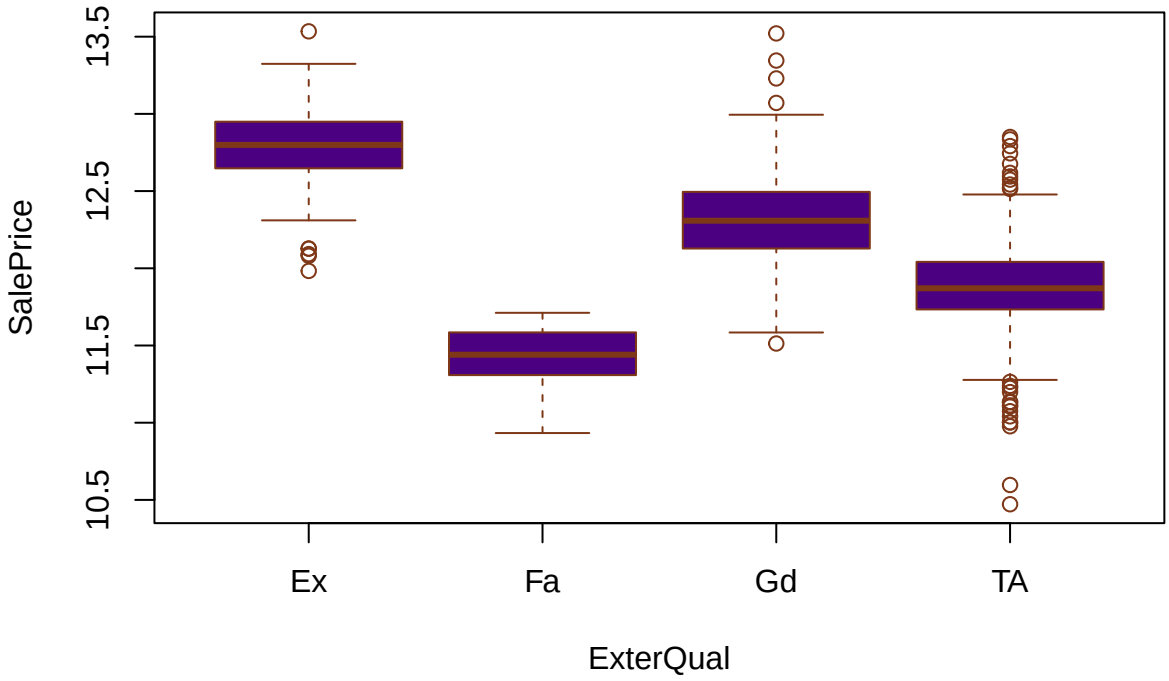


stone cladding has the highest average price and brick common has the lowest price

```
table(data3$ExterQual)
```

```
##
##  Ex  Fa  Gd  TA
##  51   7 477 803
```

```
boxplot(SalePrice ~ ExterQual ,data = data3 , col = "#4B0082" , border = "#7E3817")
```

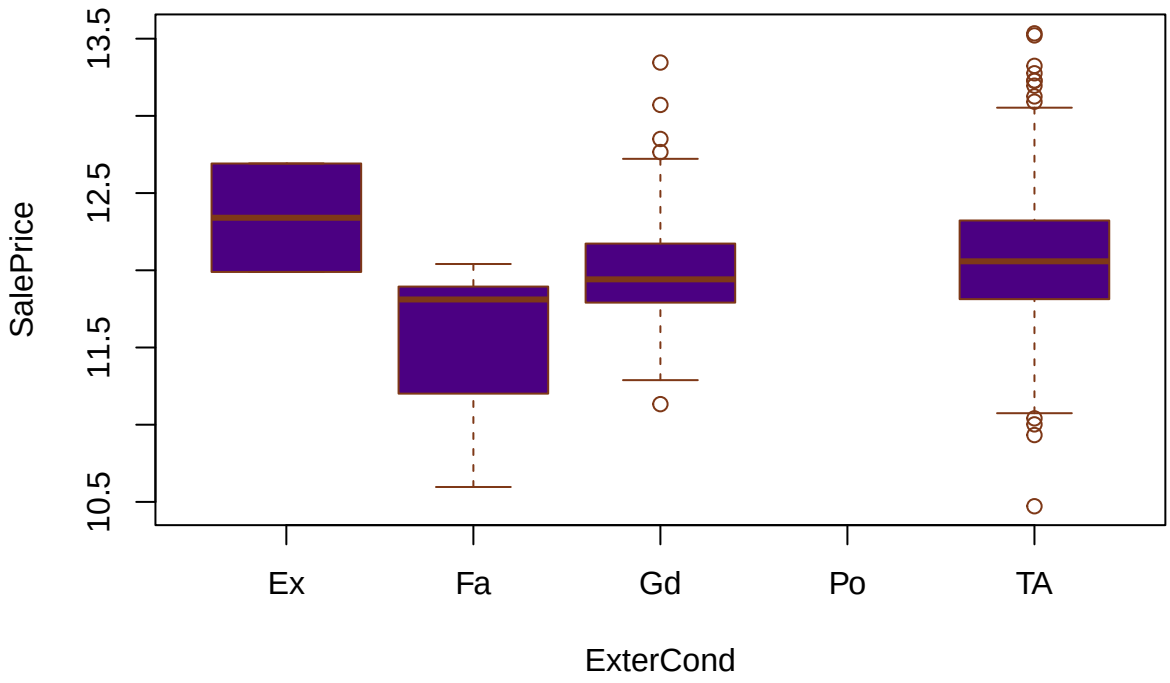


the excellent quality of the material, for the exterior of the building, has the highest average price, and the lower the quality of the materials, the lower the average price

```
table(data3$ExterCond)
```

```
##
##   Ex   Fa   Gd   Po   TA
##    2   16  137    0 1183
```

```
boxplot(SalePrice ~ ExterCond , data = data3 , col = "#4B0082" , border = "#7E3817")
```

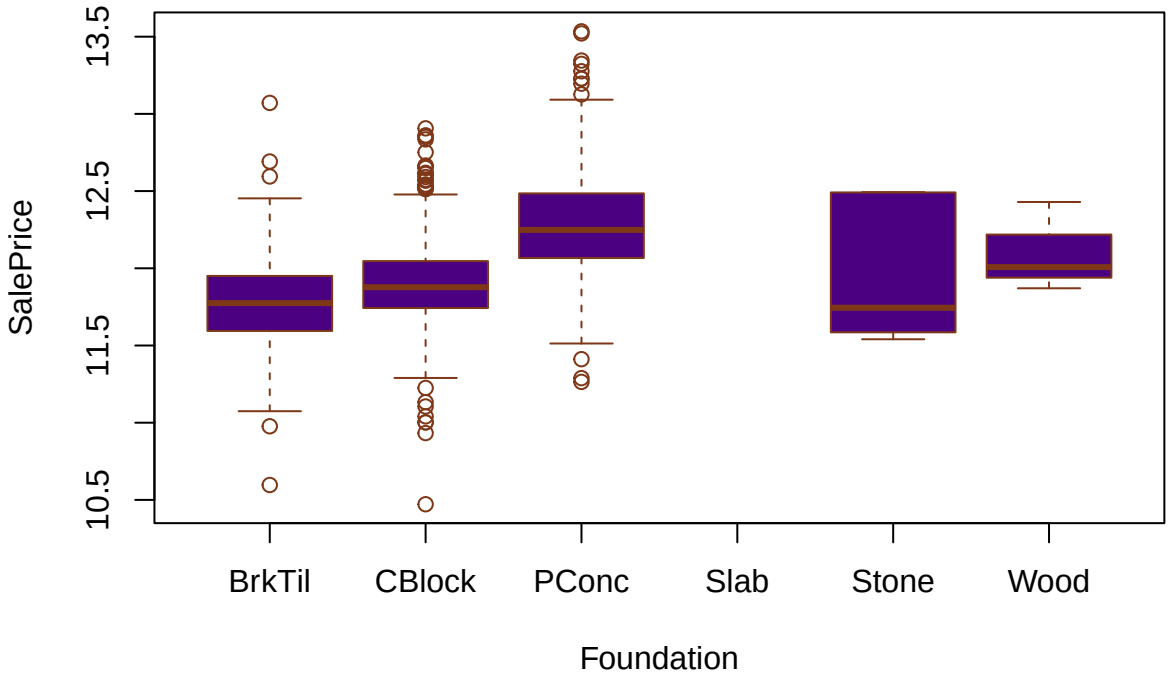


The variabels distribution is almost similar to the extent qual

```
table(data3$Foundation)
```

```
##
## BrkTil CBlock  PConc   Slab  Stone   Wood
##    129    580    620      0      6      3
```

```
boxplot(SalePrice ~ Foundation , data = data3 , col = "#4B0082" , border = "#7E3817")
```

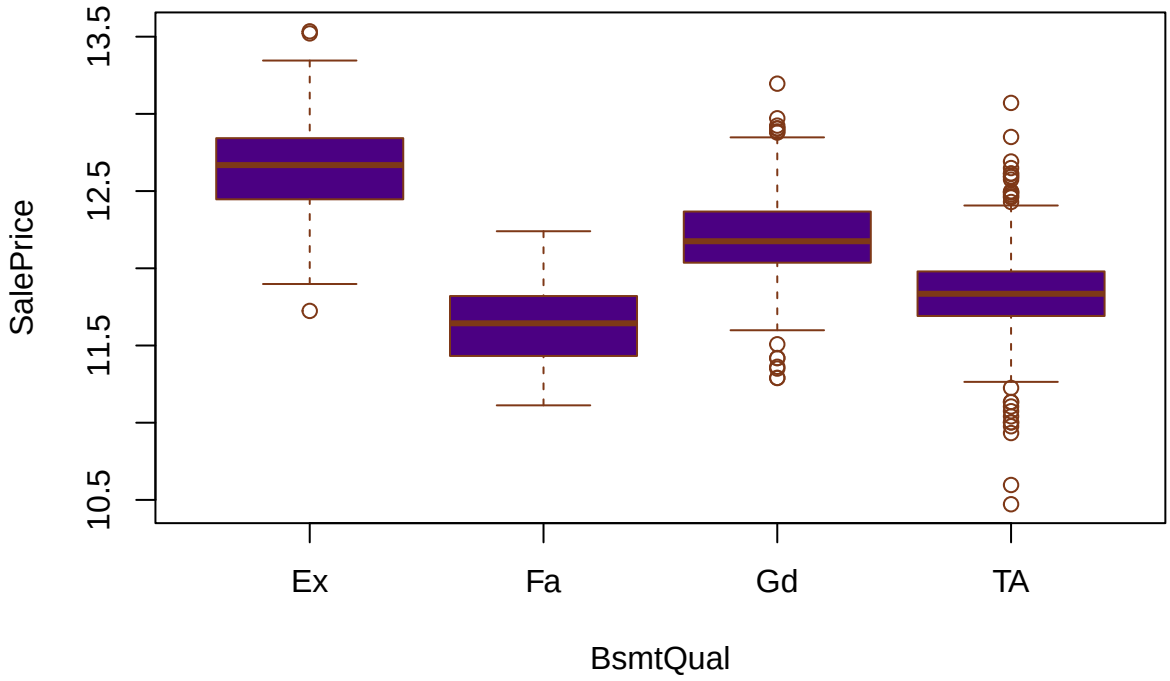


poured concrete and finder block from the most frequent and also have the highest average price

```
table(data3$BsmtQual)
```

```
##
##  Ex  Fa  Gd  TA
## 120  32 592 594
```

```
boxplot(SalePrice ~ BsmtQual , data = data3 , col = "#4B0082" , border = "#7E3817")
```

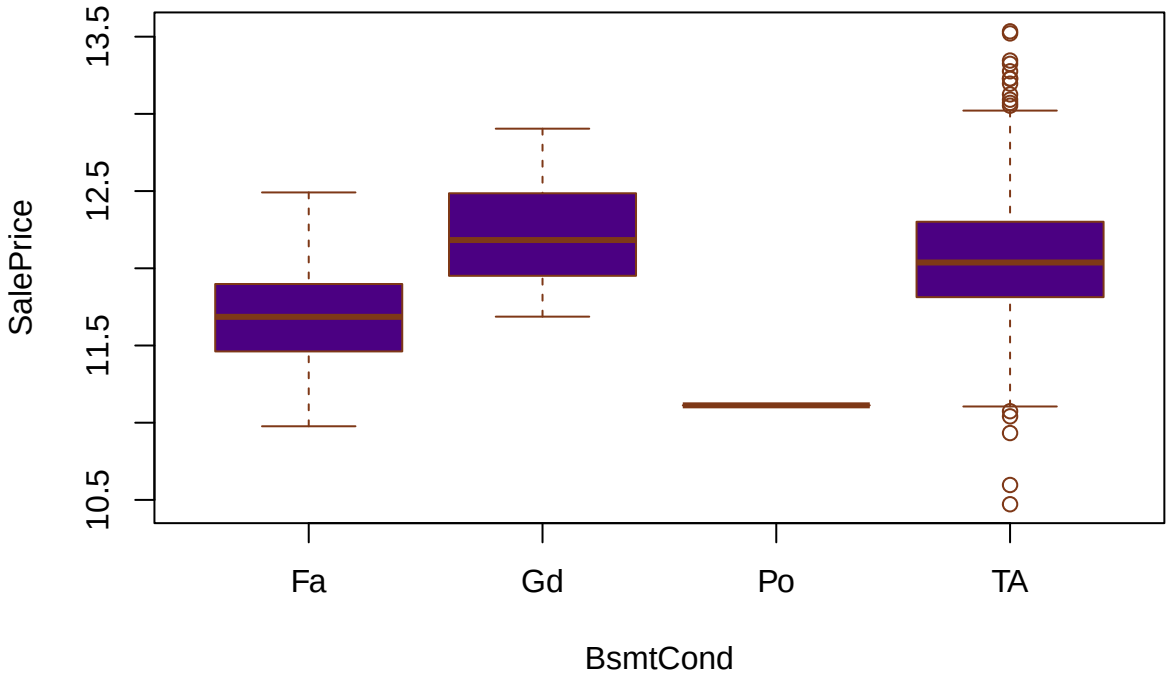


what so ever the height of the basement is high, the average price rise

```
table(data3$BsmtCond)
```

```
##
##   Fa   Gd   Po   TA
##   38   62    1 1237
```

```
boxplot(SalePrice ~ BsmtCond , data = data3 , col = "#4B0082" , border = "#7E3817")
```

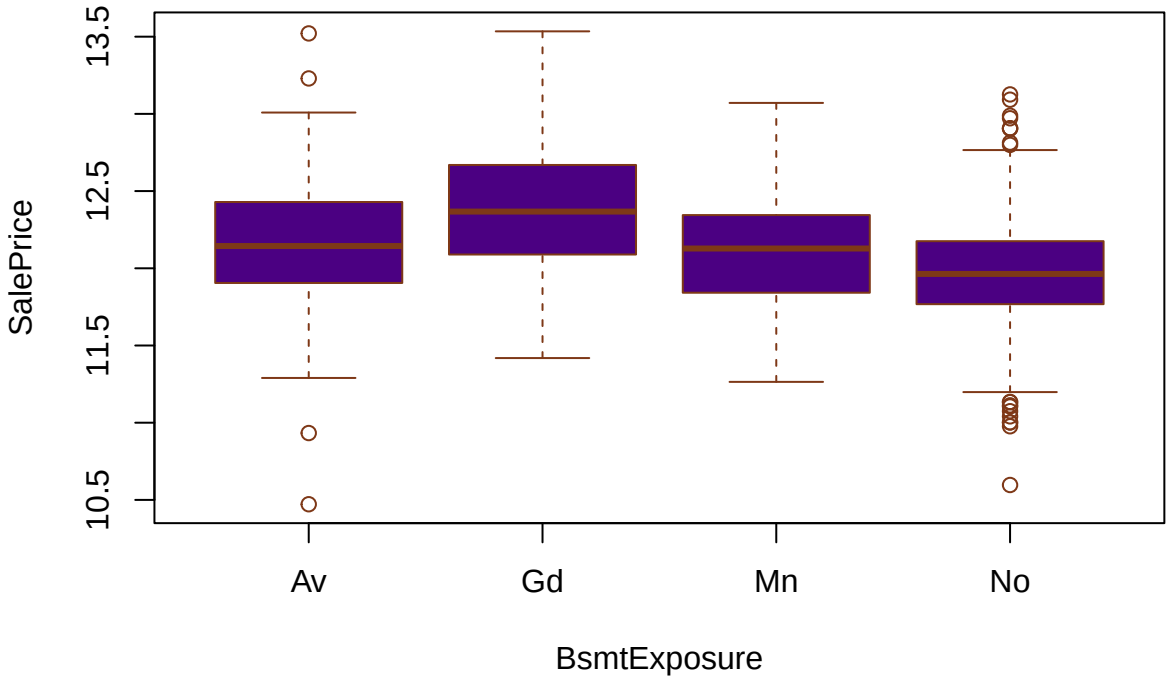


the overall quality of the basement raises the average price

```
table(data3$BsmtExposure)
```

```
##
##  Av  Gd  Mn  No
## 213 127 111 887
```

```
boxplot(SalePrice ~ BsmtExposure , data = data3 , col = "#4B0082" , border = "#7E3817")
```

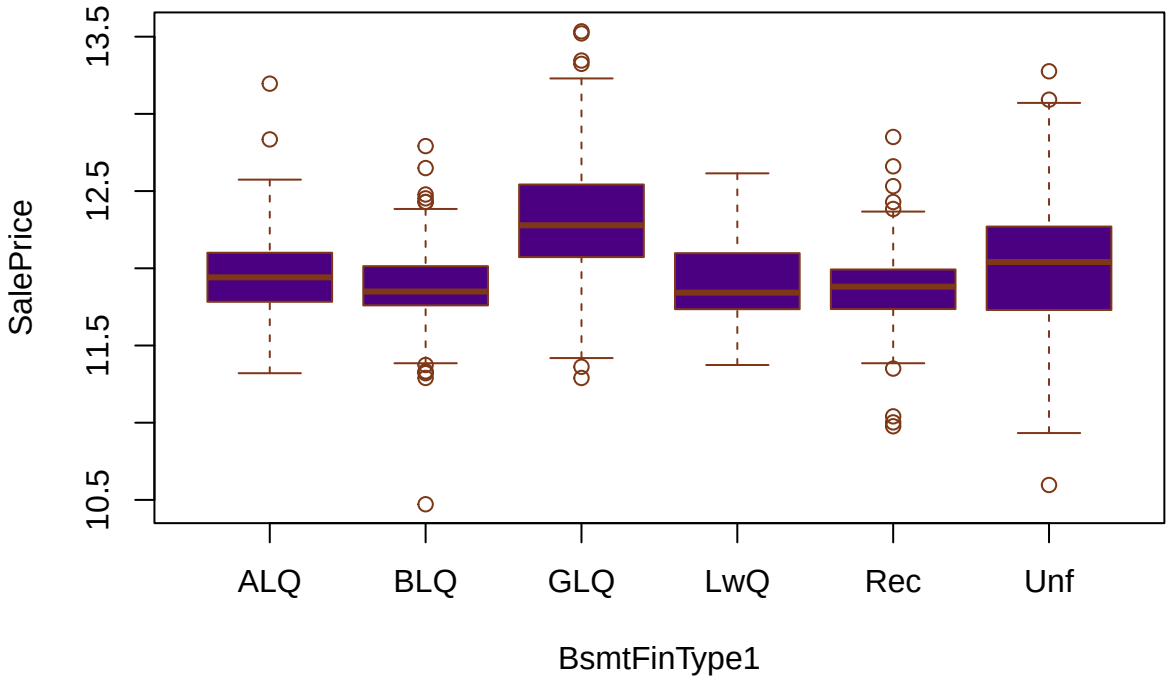


the quality of the pavement be better the average price will raise

```
table(data3$BsmtFinType1)
```

```
##
## ALQ BLQ GLQ LwQ Rec Unf
## 209 141 402  69 125 392
```

```
boxplot(SalePrice ~ BsmtFinType1 , data = data3 , col = "#4B0082" , border = "#7E3817")
```

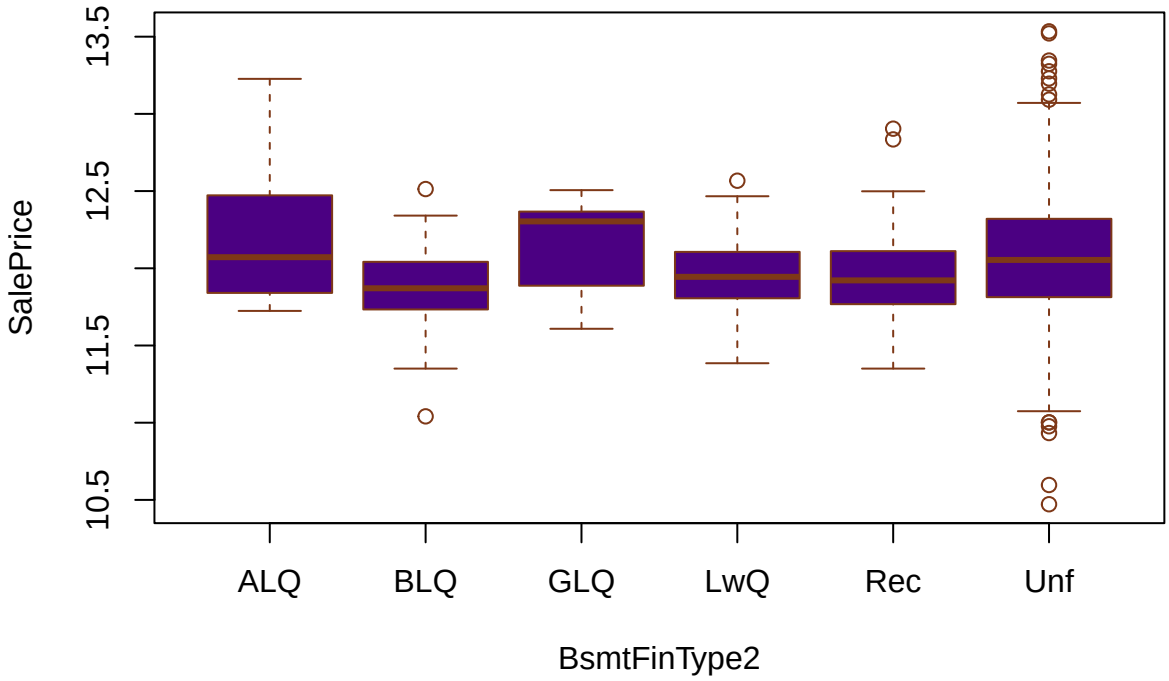


As the area of the land increases, the average price increases, of course It can be in line with the total area of the building

```
table(data3$BsmtFinType2)
```

```
##
##  ALQ  BLQ  GLQ  LwQ  Rec  Unf
##   19   32   12   46   53 1176
```

```
boxplot(SalePrice ~ BsmtFinType2 , data = data3 , col = "#4B0082" , border = "#7E3817")
```

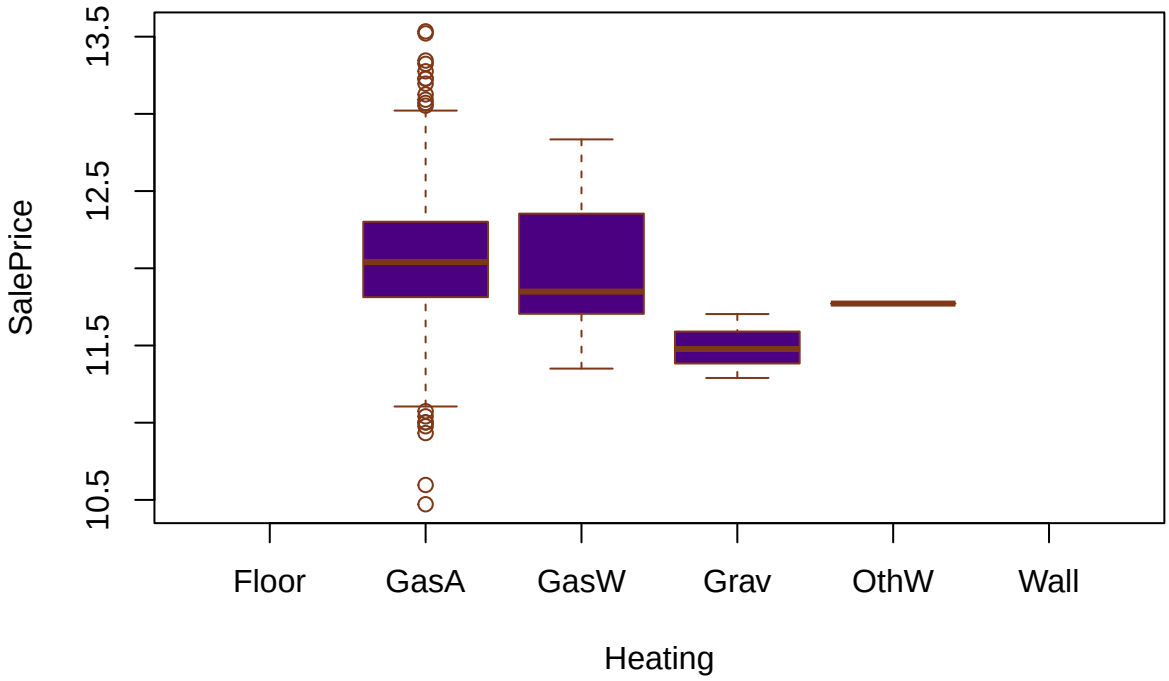


the quality of the land has a direct relationship with the price, but apparently it has formed the majority of the land (probability of multi-collinearity and lack of explicit explanation)

```
table(data3$Heating)
```

```
##
## Floor  GasA  GasW  Grav  OthW  Wall
##     0  1318    16     3     1     0
```

```
boxplot(SalePrice ~ Heating ,data = data3 , col = "#4B0082" , border = "#7E3817")
```

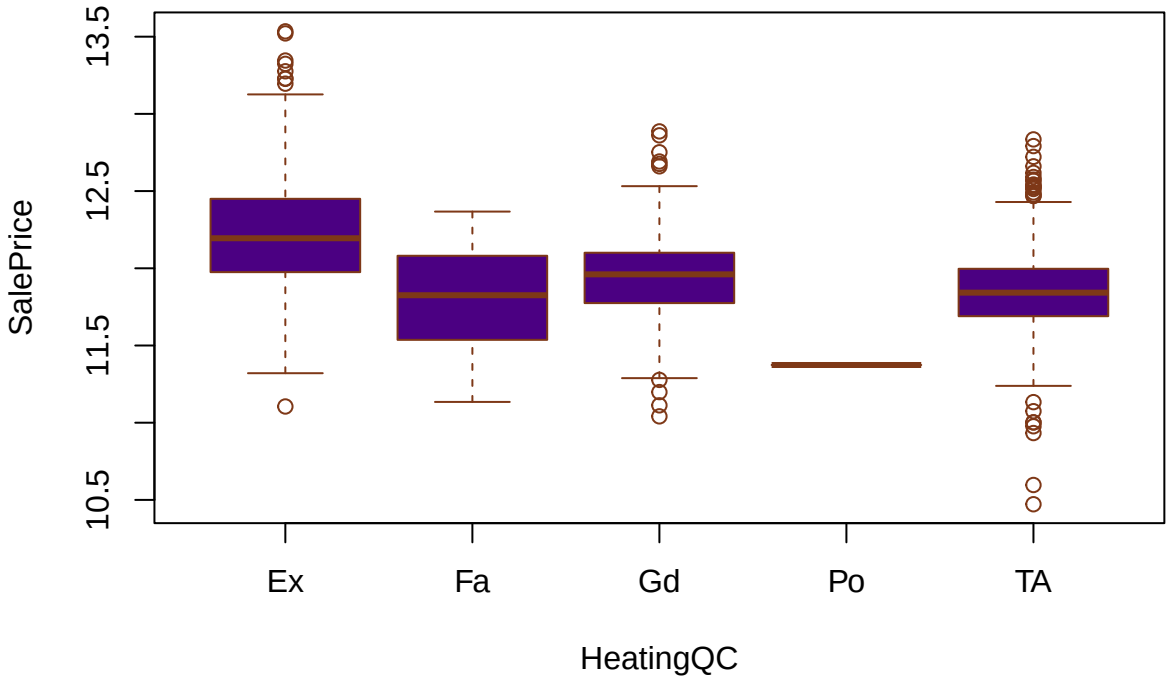


Houses with gas have the most abundance and the highest price

```
table(data3$HeatingQC)
```

```
##
##  Ex  Fa  Gd  Po  TA
## 704  36 217   1 380
```

```
boxplot(SalePrice ~ HeatingQC , data = data3 , col = "#4B0082" , border = "#7E3817")
```

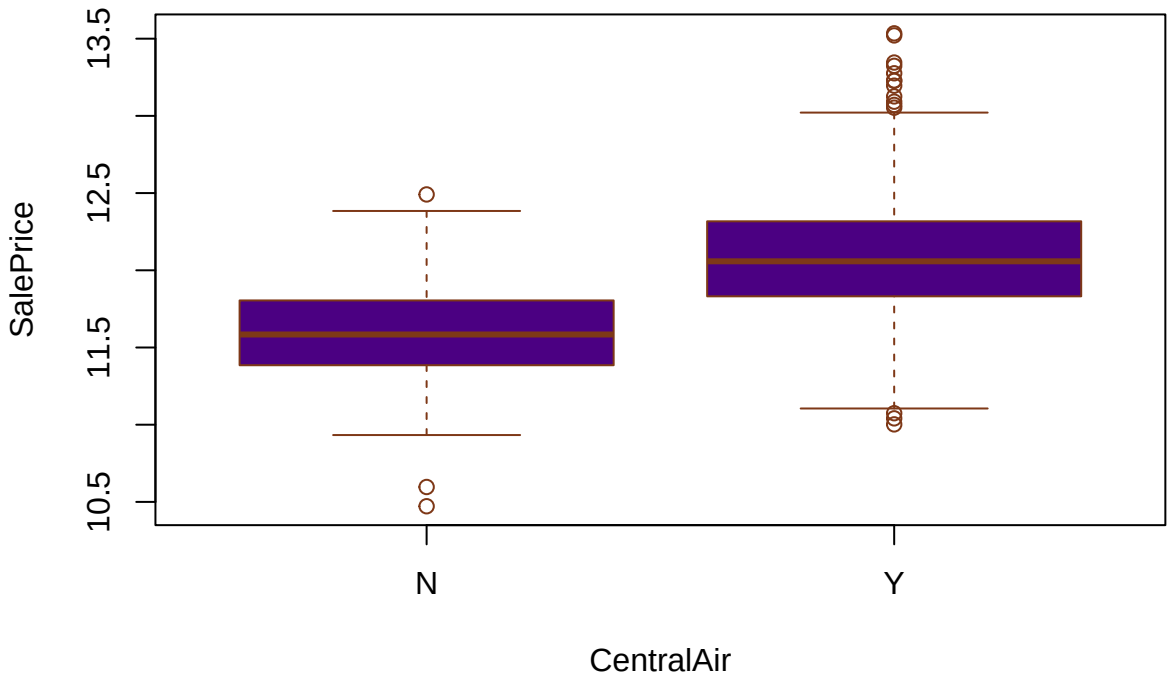


the quality of heating has a direct relationship with the price

```
table(data3$CentralAir)
```

```
##
##    N    Y
##   61 1277
```

```
boxplot(SalePrice ~ CentralAir ,data = data3 , col = "#4B0082" , border = "#7E3817")
```

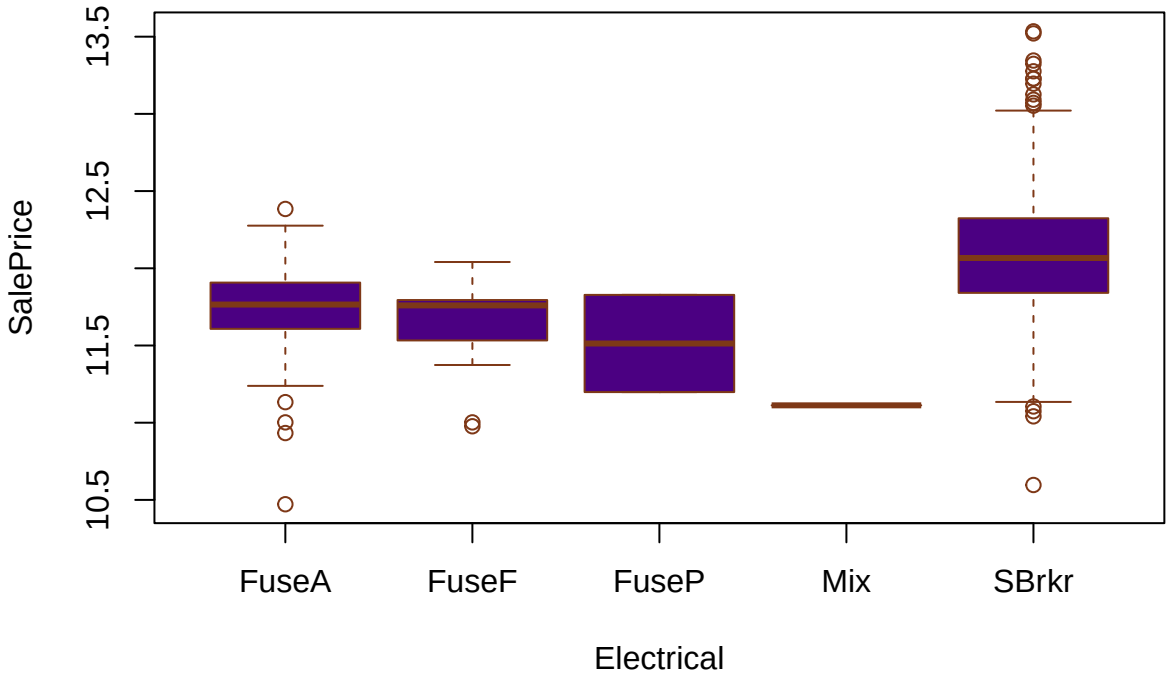


Having central air conditioning has a direct relationship with the average price

```
table(data3$Electrical)
```

```
##
## FuseA FuseF FuseP   Mix SBrkr
##    76    17     2     1  1242
```

```
boxplot(SalePrice ~ Electrical , data = data3 , col = "#4B0082" , border = "#7E3817")
```

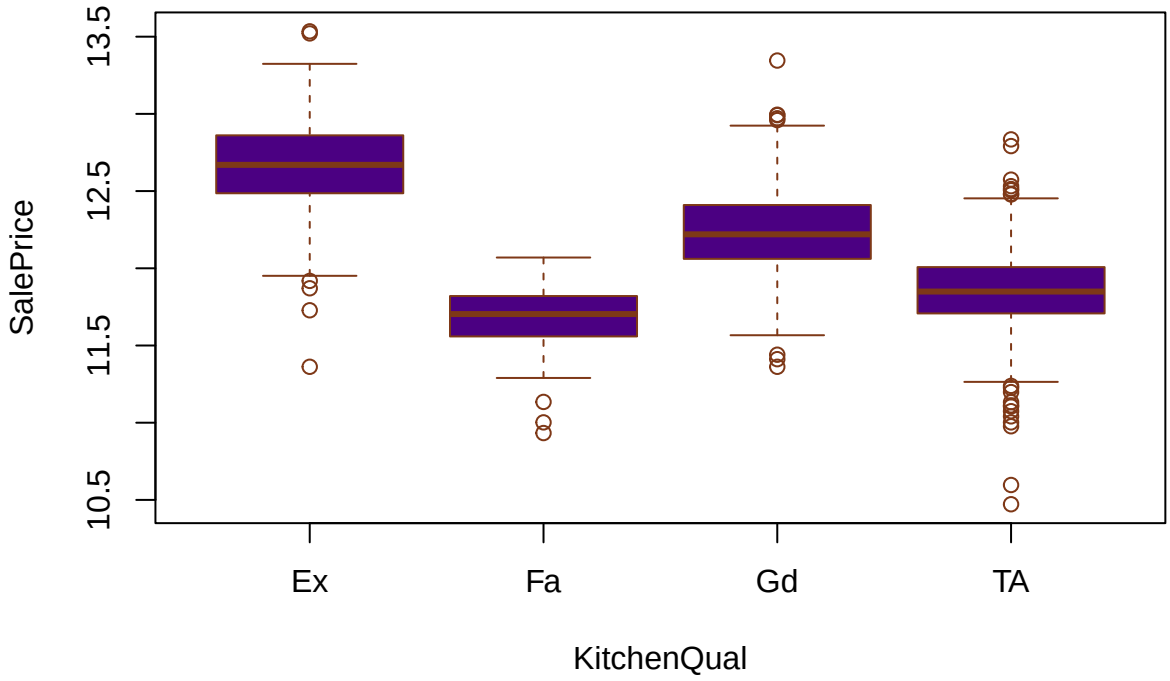


Most houses with SBrkr have a higher average price

```
table(data3$KitchenQual)
```

```
##
##  Ex  Fa  Gd  TA
##  97  23 568 650
```

```
boxplot(SalePrice ~ KitchenQual , data = data3, col = "#4B0082" , border = "#7E3817")
```

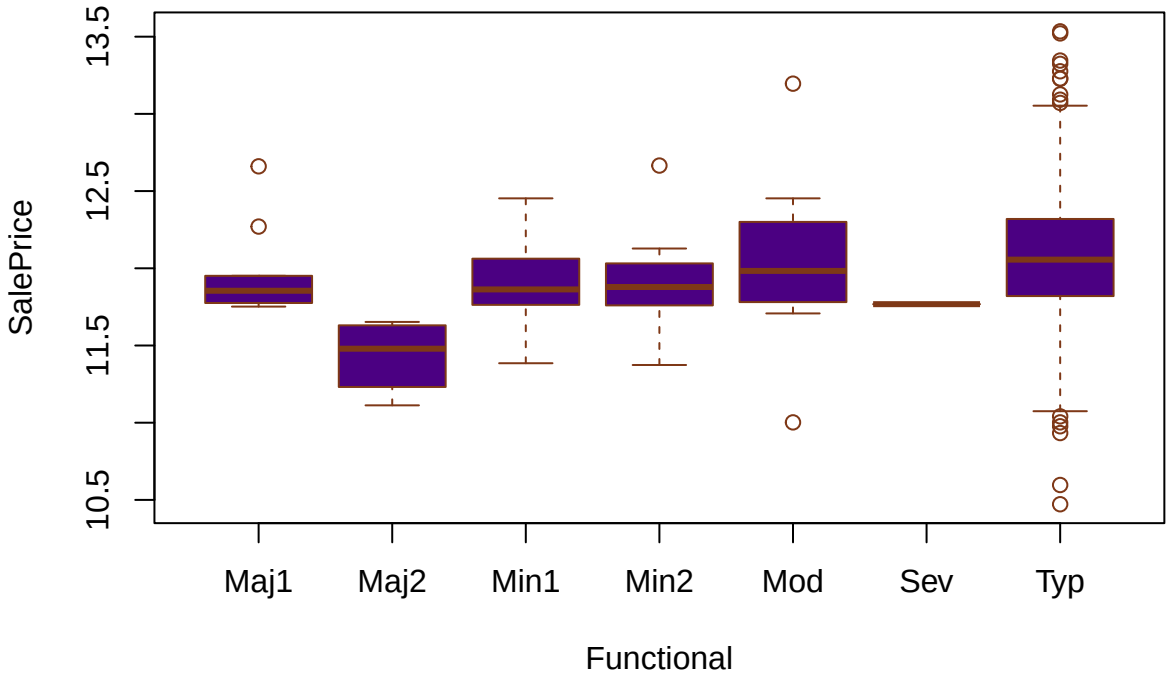


the quality of kitchens has a direct relationship with the price

```
table(data3$Functional)
```

```
##
## Maj1 Maj2 Min1 Min2  Mod  Sev  Typ
##   10    4   28   30   11    1 1254
```

```
boxplot(SalePrice ~ Functional, data = data3 , col = "#4B0082" , border = "#7E3817")
```

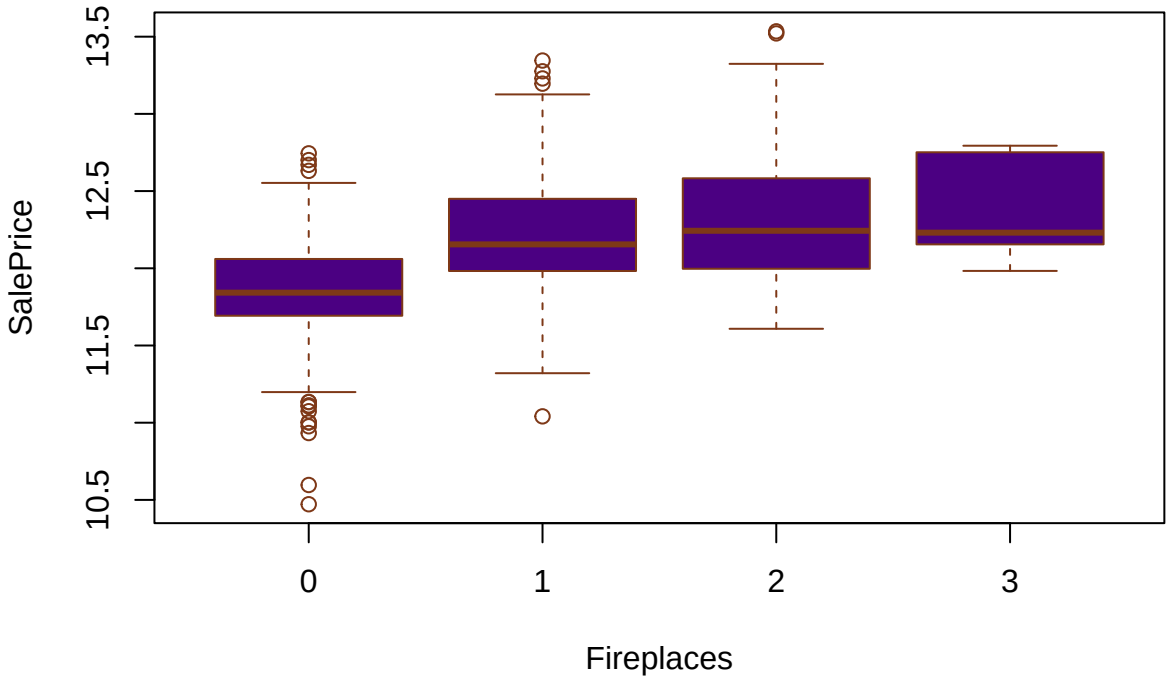


this category is vague to me maybe I didn't comment on it and maybe I catch sight in the modelling process

```
table(data3$Fireplaces)
```

```
##
##   0   1   2   3
## 591 631 111   5
```

```
boxplot(SalePrice ~ Fireplaces , data = data3 , col = "#4B0082" , border = "#7E3817")
```

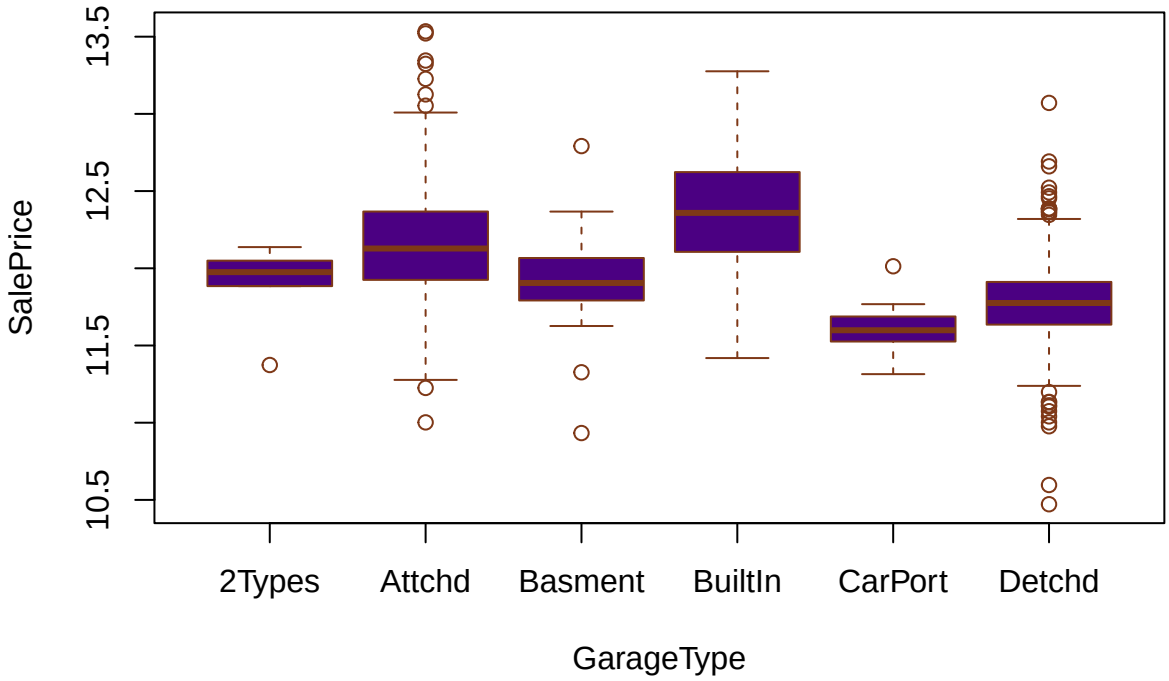


the number of fireplaces has a direct relationship with the price of the house maybe it is in line with the square footage

```
table(data3$GarageType)
```

```
##
##  2Types  Attchd Basment BuiltIn CarPort  Detchd
##       6     852      19      85       7     369
```

```
boxplot(SalePrice ~ GarageType , data = data3 , col = "#4B0082" , border = "#7E3817")
```

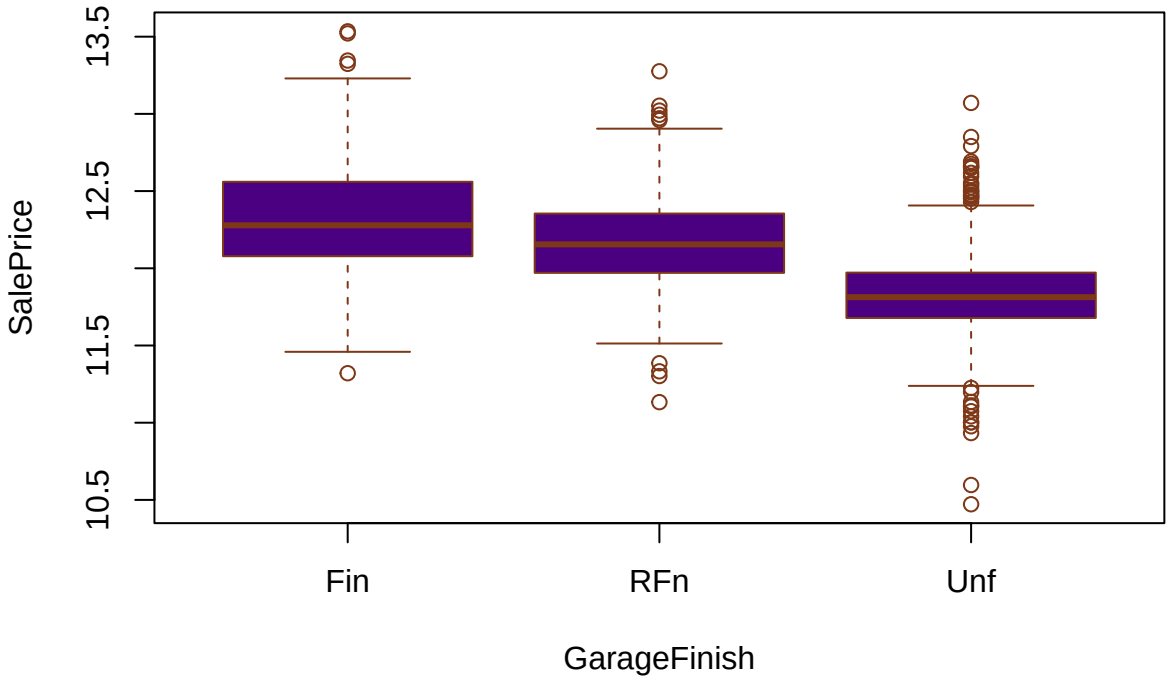


Garages inside the house and attached to the house have a higher price and houses with A separate garage and without a garage have a lower average price

```
table(data3$GarageFinish)
```

```
##
## Fin RFn Unf
## 345 413 580
```

```
boxplot(SalePrice ~ GarageFinish , data = data3 , col = "#4B0082" , border = "#7E3817")
```

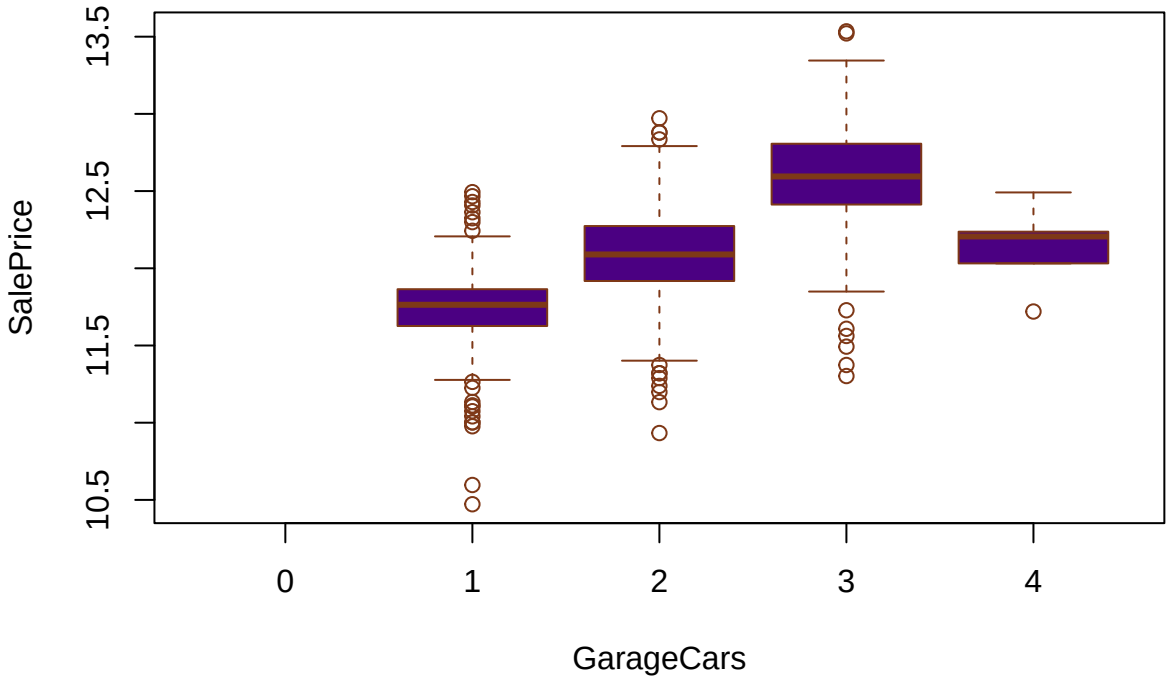


complete garages have a higher average price than unfinished garages

```
table(data3$GarageCars)
```

```
##
##   0   1   2   3   4
##   0 361 793 179   5
```

```
boxplot(SalePrice ~ GarageCars , data = data3 , col = "#4B0082" , border = "#7E3817")
```

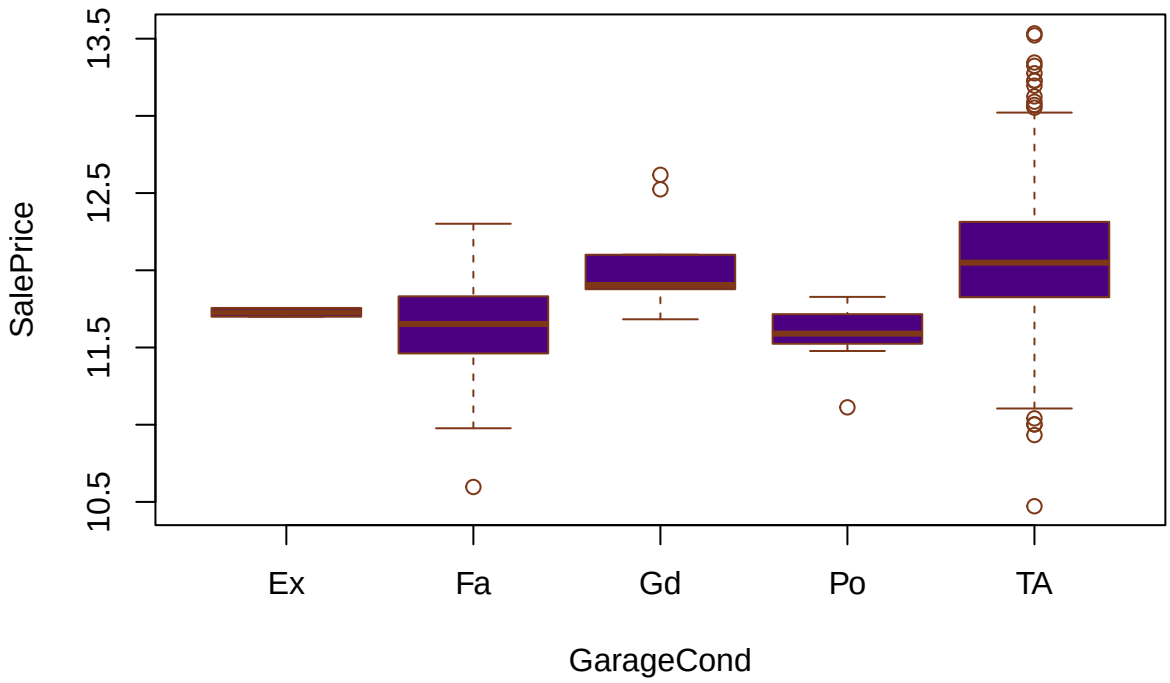


the number of garages has a direct relationship with the price but in the case of houses with 4 garages, a price drop is observed

```
table(data3$GarageCond)
```

```
##
##   Ex   Fa   Gd   Po   TA
##    2   33    9    7 1287
```

```
boxplot(SalePrice ~ GarageCond , data = data3 , col = "#4B0082" , border = "#7E3817")
```

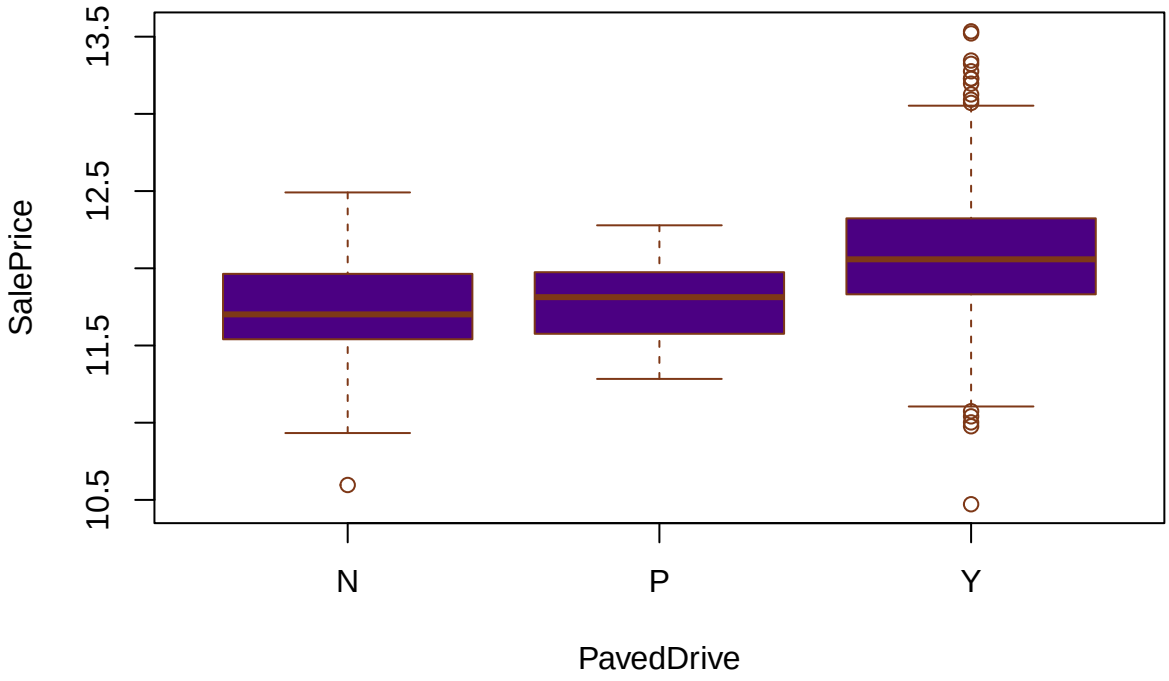


paved and equipped sidewalks have a direct relationship with the price

```
table(data3$PavedDrive)
```

```
##
##    N    P    Y
##   54   27 1257
```

```
boxplot(SalePrice ~ PavedDrive , data = data3 , col = "#4B0082" , border = "#7E3817")
```

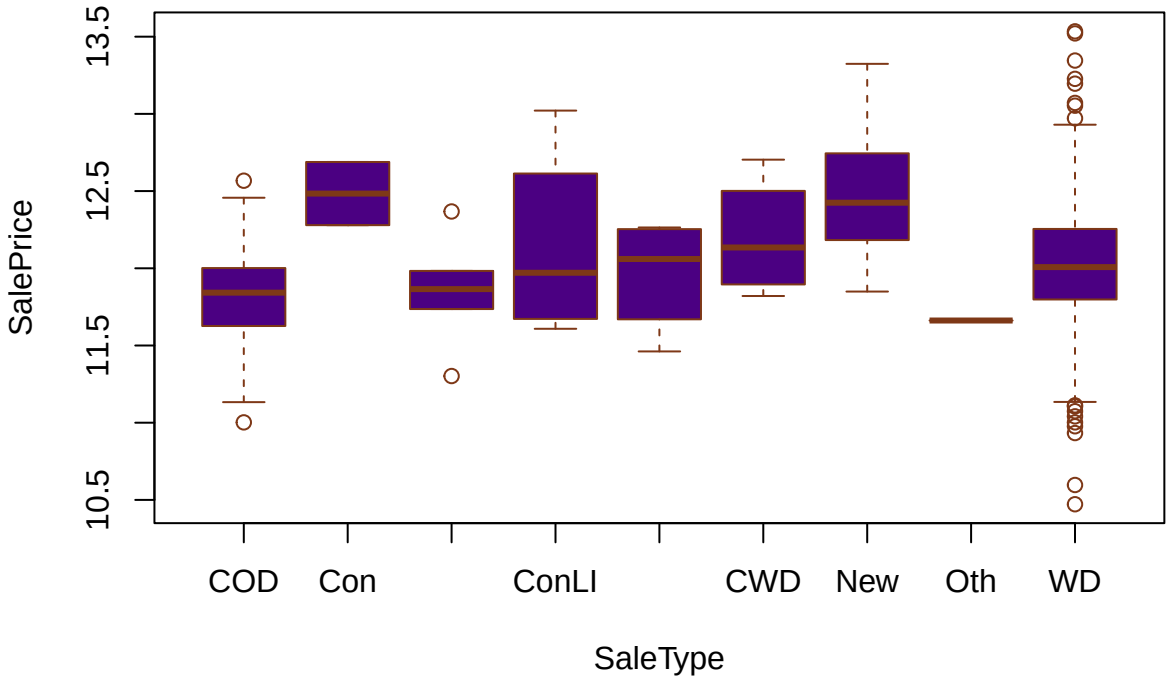


Normal and conditional sales have the most frequent and average prices and are higher in the sale type, but sales with advance payment and also newly built houses have a higher price

```
table(data3$SaleType)
```

```
##
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD
##    42     2     6     4     4     4   117     1  1158
```

```
boxplot(SalePrice ~ SaleType , data = data3 , col = "#4B0082" , border = "#7E3817")
```

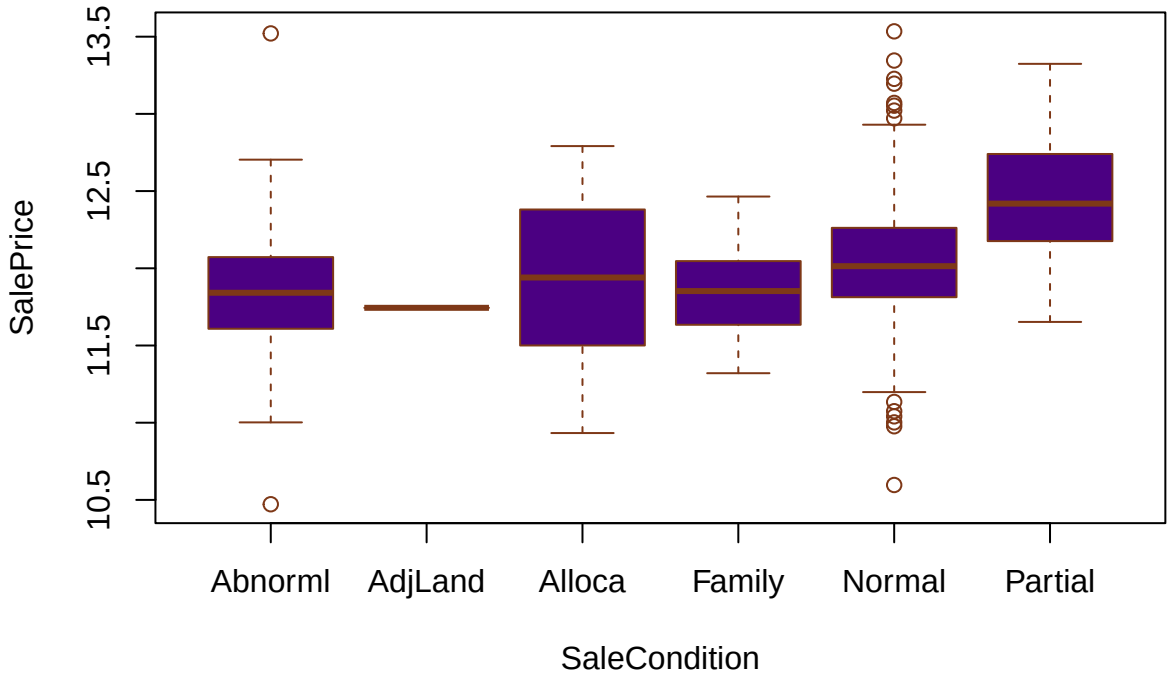


the average of all prices are close to each other, except partial, which may be due to home renovation

```
table(data3$SaleCondition)
```

```
##
## Abnorml AdjLand  Alloca  Family  Normal Partial
##      86       1       7      20    1104     120
```

```
boxplot(SalePrice ~ SaleCondition , data = data3 , col = "#4B0082" , border = "#7E3817")
```



Go to modelling