Cleaning data

Arjuna Anilkumar, A20446963

11/8/2020

Introduction

This project aims to predict the final price of houses using the Ames housing dataset.

Data description

The Ames Housing dataset was compiled by Dean De Cock for use in data science education. It's an alternative to the Boston Housing dataset and is for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset.

The Ames housing data contains With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa.

Data Processing

Install packages

```
#install.packages(c("Amelia","purrr","tidyr","ggplot2","rpart","plyr"))
```

Load data

```
df <- read.table("../data/raw/train.csv", sep = ",",header = T)
head(df)
```

```
##   Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape LandContour
## 1  1         60       RL          65    8450   Pave  <NA>      Reg         Lvl
## 2  2         20       RL          80    9600   Pave  <NA>      Reg         Lvl
## 3  3         60       RL          68   11250   Pave  <NA>      IR1         Lvl
## 4  4         70       RL          60    9550   Pave  <NA>      IR1         Lvl
## 5  5         60       RL          84   14260   Pave  <NA>      IR1         Lvl
## 6  6         50       RL          85   14115   Pave  <NA>      IR1         Lvl
##   Utilities LotConfig LandSlope Neighborhood Condition1 Condition2 BldgType
## 1    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 2    AllPub       FR2       Gtl      Veenker      Feedr       Norm     1Fam
## 3    AllPub    Inside       Gtl      CollgCr       Norm       Norm     1Fam
## 4    AllPub    Corner       Gtl      Crawfor       Norm       Norm     1Fam
## 5    AllPub       FR2       Gtl      NoRidge       Norm       Norm     1Fam
## 6    AllPub    Inside       Gtl      Mitchel       Norm       Norm     1Fam
##   HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle RoofMatl
## 1     2Story           7           5      2003         2003     Gable  CompShg
## 2     1Story           6           8      1976         1976     Gable  CompShg
## 3     2Story           7           5      2001         2002     Gable  CompShg
## 4     2Story           7           5      1915         1970     Gable  CompShg
## 5     2Story           8           5      2000         2000     Gable  CompShg
## 6     1.5Fin           5           5      1993         1995     Gable  CompShg
##   Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond Foundation
## 1     VinylSd     VinylSd    BrkFace        196        Gd        TA      PConc
## 2     MetalSd     MetalSd       None          0        TA        TA     CBlock
## 3     VinylSd     VinylSd    BrkFace        162        Gd        TA      PConc
## 4     Wd Sdng     Wd Shng       None          0        TA        TA     BrkTil
## 5     VinylSd     VinylSd    BrkFace        350        Gd        TA      PConc
## 6     VinylSd     VinylSd       None          0        TA        TA       Wood
##   BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2
## 1       Gd       TA           No          GLQ        706          Unf
## 2       Gd       TA           Gd          ALQ        978          Unf
## 3       Gd       TA           Mn          GLQ        486          Unf
## 4       TA       Gd           No          ALQ        216          Unf
## 5       Gd       TA           Av          GLQ        655          Unf
## 6       Gd       TA           No          GLQ        732          Unf
##   BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical
## 1          0       150         856    GasA        Ex          Y      SBrkr
## 2          0       284        1262    GasA        Ex          Y      SBrkr
## 3          0       434         920    GasA        Ex          Y      SBrkr
## 4          0       540         756    GasA        Gd          Y      SBrkr
## 5          0       490        1145    GasA        Ex          Y      SBrkr
## 6          0        64         796    GasA        Ex          Y      SBrkr
##   X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath FullBath
## 1       856       854            0      1710            1            0        2
## 2      1262         0            0      1262            0            1        2
## 3       920       866            0      1786            1            0        2
## 4       961       756            0      1717            1            0        1
## 5      1145      1053            0      2198            1            0        2
## 6       796       566            0      1362            1            0        1
##   HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd Functional
## 1        1            3            1          Gd            8        Typ
## 2        0            3            1          TA            6        Typ
## 3        1            3            1          Gd            6        Typ
## 4        0            3            1          Gd            7        Typ
## 5        1            4            1          Gd            9        Typ
## 6        1            1            1          TA            5        Typ
##   Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish GarageCars
## 1          0        <NA>     Attchd        2003          RFn          2
## 2          1          TA     Attchd        1976          RFn          2
## 3          1          TA     Attchd        2001          RFn          2
## 4          1          Gd     Detchd        1998          Unf          3
## 5          1          TA     Attchd        2000          RFn          3
## 6          0        <NA>     Attchd        1993          Unf          2
##   GarageArea GarageQual GarageCond PavedDrive WoodDeckSF OpenPorchSF
## 1        548         TA         TA          Y          0          61
## 2        460         TA         TA          Y        298           0
## 3        608         TA         TA          Y          0          42
## 4        642         TA         TA          Y          0          35
## 5        836         TA         TA          Y        192          84
## 6        480         TA         TA          Y         40          30
##   EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC Fence MiscFeature
## 1             0          0           0        0   <NA>  <NA>        <NA>
## 2             0          0           0        0   <NA>  <NA>        <NA>
## 3             0          0           0        0   <NA>  <NA>        <NA>
## 4           272          0           0        0   <NA>  <NA>        <NA>
## 5             0          0           0        0   <NA>  <NA>        <NA>
## 6             0        320           0        0   <NA> MnPrv        Shed
##   MiscVal MoSold YrSold SaleType SaleCondition SalePrice
## 1       0      2   2008       WD        Normal    208500
## 2       0      5   2007       WD        Normal    181500
## 3       0      9   2008       WD        Normal    223500
## 4       0      2   2006       WD       Abnorml    140000
## 5       0     12   2008       WD        Normal    250000
## 6     700     10   2009       WD        Normal    143000
```

```r
df2 <- read.table("../data/raw/test.csv", sep = ",",header = T)
head(df2)
```

```
##     Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape
## 1 1461         20       RH          80   11622   Pave  <NA>      Reg
## 2 1462         20       RL          81   14267   Pave  <NA>      IR1
## 3 1463         60       RL          74   13830   Pave  <NA>      IR1
## 4 1464         60       RL          78    9978   Pave  <NA>      IR1
## 5 1465        120       RL          43    5005   Pave  <NA>      IR1
## 6 1466         60       RL          75   10000   Pave  <NA>      IR1
##   LandContour Utilities LotConfig LandSlope Neighborhood Condition1 Condition2
## 1         Lvl    AllPub    Inside       Gtl        NAmes      Feedr       Norm
## 2         Lvl    AllPub    Corner       Gtl        NAmes       Norm       Norm
## 3         Lvl    AllPub    Inside       Gtl      Gilbert       Norm       Norm
## 4         Lvl    AllPub    Inside       Gtl      Gilbert       Norm       Norm
## 5         HLS    AllPub    Inside       Gtl      StoneBr       Norm       Norm
## 6         Lvl    AllPub    Corner       Gtl      Gilbert       Norm       Norm
##   BldgType HouseStyle OverallQual OverallCond YearBuilt YearRemodAdd RoofStyle
## 1     1Fam     1Story           5           6      1961         1961     Gable
## 2     1Fam     1Story           6           6      1958         1958       Hip
## 3     1Fam     2Story           5           5      1997         1998     Gable
## 4     1Fam     2Story           6           6      1998         1998     Gable
## 5   TwnhsE     1Story           8           5      1992         1992     Gable
## 6     1Fam     2Story           6           5      1993         1994     Gable
##   RoofMatl Exterior1st Exterior2nd MasVnrType MasVnrArea ExterQual ExterCond
## 1  CompShg     VinylSd     VinylSd       None          0        TA        TA
## 2  CompShg     Wd Sdng     Wd Sdng    BrkFace        108        TA        TA
## 3  CompShg     VinylSd     VinylSd       None          0        TA        TA
## 4  CompShg     VinylSd     VinylSd    BrkFace         20        TA        TA
## 5  CompShg     HdBoard     HdBoard       None          0        Gd        TA
## 6  CompShg     HdBoard     HdBoard       None          0        TA        TA
##   Foundation BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1
## 1     CBlock       TA       TA           No          Rec        468
## 2     CBlock       TA       TA           No          ALQ        923
## 3      PConc       Gd       TA           No          GLQ        791
## 4      PConc       TA       TA           No          GLQ        602
## 5      PConc       Gd       TA           No          ALQ        263
## 6      PConc       Gd       TA           No          Unf          0
##   BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir
## 1          LwQ        144       270         882    GasA        TA          Y
## 2          Unf          0       406        1329    GasA        TA          Y
## 3          Unf          0       137         928    GasA        Gd          Y
## 4          Unf          0       324         926    GasA        Ex          Y
## 5          Unf          0      1017        1280    GasA        Ex          Y
## 6          Unf          0       763         763    GasA        Gd          Y
##   Electrical X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath
## 1      SBrkr       896         0            0       896            0
## 2      SBrkr      1329         0            0      1329            0
## 3      SBrkr       928       701            0      1629            0
## 4      SBrkr       926       678            0      1604            0
## 5      SBrkr      1280         0            0      1280            0
## 6      SBrkr       763       892            0      1655            0
##   BsmtHalfBath FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual
## 1            0        1        0            2            1          TA
## 2            0        1        1            3            1          Gd
## 3            0        2        1            3            1          TA
## 4            0        2        1            3            1          Gd
## 5            0        2        0            2            1          Gd
## 6            0        2        1            3            1          TA
##   TotRmsAbvGrd Functional Fireplaces FireplaceQu GarageType GarageYrBlt
## 1            5        Typ          0        <NA>     Attchd        1961
## 2            6        Typ          0        <NA>     Attchd        1958
## 3            6        Typ          1          TA     Attchd        1997
## 4            7        Typ          1          Gd     Attchd        1998
## 5            5        Typ          0        <NA>     Attchd        1992
## 6            7        Typ          1          TA     Attchd        1993
##   GarageFinish GarageCars GarageArea GarageQual GarageCond PavedDrive
## 1          Unf          1        730         TA         TA          Y
## 2          Unf          1        312         TA         TA          Y
## 3          Fin          2        482         TA         TA          Y
## 4          Fin          2        470         TA         TA          Y
## 5          RFn          2        506         TA         TA          Y
## 6          Fin          2        440         TA         TA          Y
##   WoodDeckSF OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC
## 1        140           0             0          0         120        0   <NA>
## 2        393          36             0          0           0        0   <NA>
## 3        212          34             0          0           0        0   <NA>
## 4        360          36             0          0           0        0   <NA>
## 5          0          82             0          0         144        0   <NA>
## 6        157          84             0          0           0        0   <NA>
##   Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition
## 1 MnPrv        <NA>       0      6   2010       WD        Normal
## 2  <NA>        Gar2   12500      6   2010       WD        Normal
## 3 MnPrv        <NA>       0      3   2010       WD        Normal
## 4  <NA>        <NA>       0      6   2010       WD        Normal
## 5  <NA>        <NA>       0      1   2010       WD        Normal
## 6  <NA>        <NA>       0      4   2010       WD        Normal
```

```
train <- df
test <- df2
str(train)
```

```
## ’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     : chr  "RL" "RL" "RL" "RL" ...
##  $ 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       : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr  NA NA NA NA ...
##  $ LotShape     : chr  "Reg" "Reg" "IR1" "IR1" ...
##  $ LandContour  : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : chr  "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ LotConfig    : chr  "Inside" "FR2" "Inside" "Corner" ...
##  $ LandSlope    : chr  "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood : chr  "CollgCr" "Veenker" "CollgCr" "Crawfor" ...
##  $ Condition1   : chr  "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition2   : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType     : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr  "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  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    : chr  "Gable" "Gable" "Gable" "Gable" ...
##  $ RoofMatl     : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...
##  $ Exterior2nd  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...
##  $ MasVnrType   : chr  "BrkFace" "None" "BrkFace" "None" ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : chr  "Gd" "TA" "Gd" "TA" ...
##  $ ExterCond    : chr  "TA" "TA" "TA" "TA" ...
##  $ Foundation   : chr  "PConc" "CBlock" "PConc" "BrkTil" ...
##  $ BsmtQual     : chr  "Gd" "Gd" "Gd" "TA" ...
##  $ BsmtCond     : chr  "TA" "TA" "TA" "Gd" ...
##  $ BsmtExposure : chr  "No" "Gd" "Mn" "No" ...
##  $ BsmtFinType1 : chr  "GLQ" "ALQ" "GLQ" "ALQ" ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : chr  "Unf" "Unf" "Unf" "Unf" ...
##  $ 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      : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC    : chr  "Ex" "Ex" "Ex" "Gd" ...
##  $ CentralAir   : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical   : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ 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  : chr  "Gd" "TA" "Gd" "Gd" ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : chr  NA "TA" "TA" "Gd" ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : chr  "RFn" "RFn" "RFn" "Unf" ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : chr  "TA" "TA" "TA" "TA" ...
##  $ GarageCond   : chr  "TA" "TA" "TA" "TA" ...
##  $ PavedDrive   : chr  "Y" "Y" "Y" "Y" ...
##  $ 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       : chr  NA NA NA NA ...
##  $ Fence        : chr  NA NA NA NA ...
##  $ MiscFeature  : chr  NA NA 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     : chr  "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr  "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

```r
str(test)
```

```
## ’data.frame':    1459 obs. of  80 variables:
##  $ Id          : int  1461 1462 1463 1464 1465 1466 1467 1468 1469 1470 ...
##  $ MSSubClass  : int  20 20 60 60 120 60 20 60 20 20 ...
##  $ MSZoning    : chr  "RH" "RL" "RL" "RL" ...
##  $ LotFrontage : int  80 81 74 78 43 75 NA 63 85 70 ...
##  $ LotArea     : int  11622 14267 13830 9978 5005 10000 7980 8402 10176 8400 ...
##  $ Street      : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley       : chr  NA NA NA NA ...
##  $ LotShape    : chr  "Reg" "IR1" "IR1" "IR1" ...
##  $ LandContour : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities   : chr  "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ LotConfig   : chr  "Inside" "Corner" "Inside" "Inside" ...
##  $ LandSlope   : chr  "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood: chr  "NAmes" "NAmes" "Gilbert" "Gilbert" ...
##  $ Condition1  : chr  "Feedr" "Norm" "Norm" "Norm" ...
##  $ Condition2  : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType    : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle  : chr  "1Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual : int  5 6 5 6 8 6 6 6 7 4 ...
##  $ OverallCond : int  6 6 5 6 5 5 7 5 5 5 ...
##  $ YearBuilt   : int  1961 1958 1997 1998 1992 1993 1992 1998 1990 1970 ...
##  $ YearRemodAdd: int  1961 1958 1998 1998 1992 1994 2007 1998 1990 1970 ...
##  $ RoofStyle   : chr  "Gable" "Hip" "Gable" "Gable" ...
##  $ RoofMatl    : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st : chr  "VinylSd" "Wd Sdng" "VinylSd" "VinylSd" ...
##  $ Exterior2nd : chr  "VinylSd" "Wd Sdng" "VinylSd" "VinylSd" ...
##  $ MasVnrType  : chr  "None" "BrkFace" "None" "BrkFace" ...
##  $ MasVnrArea  : int  0 108 0 20 0 0 0 0 0 0 ...
##  $ ExterQual   : chr  "TA" "TA" "TA" "TA" ...
##  $ ExterCond   : chr  "TA" "TA" "TA" "TA" ...
##  $ Foundation  : chr  "CBlock" "CBlock" "PConc" "PConc" ...
##  $ BsmtQual    : chr  "TA" "TA" "Gd" "TA" ...
##  $ BsmtCond    : chr  "TA" "TA" "TA" "TA" ...
##  $ BsmtExposure: chr  "No" "No" "No" "No" ...
##  $ BsmtFinType1: chr  "Rec" "ALQ" "GLQ" "GLQ" ...
##  $ BsmtFinSF1  : int  468 923 791 602 263 0 935 0 637 804 ...
##  $ BsmtFinType2: chr  "LwQ" "Unf" "Unf" "Unf" ...
##  $ BsmtFinSF2  : int  144 0 0 0 0 0 0 0 0 78 ...
##  $ BsmtUnfSF   : int  270 406 137 324 1017 763 233 789 663 0 ...
##  $ TotalBsmtSF : int  882 1329 928 926 1280 763 1168 789 1300 882 ...
##  $ Heating     : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC   : chr  "TA" "TA" "Gd" "Ex" ...
##  $ CentralAir  : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical  : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ X1stFlrSF   : int  896 1329 928 926 1280 763 1187 789 1341 882 ...
##  $ X2ndFlrSF   : int  0 0 701 678 0 892 0 676 0 0 ...
##  $ LowQualFinSF: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea   : int  896 1329 1629 1604 1280 1655 1187 1465 1341 882 ...
##  $ BsmtFullBath: int  0 0 0 0 0 0 1 0 1 1 ...
##  $ BsmtHalfBath: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ FullBath    : int  1 1 2 2 2 2 2 2 1 1 ...
##  $ HalfBath    : int  0 1 1 1 0 1 0 1 1 0 ...
##  $ BedroomAbvGr: int  2 3 3 3 2 3 3 3 2 2 ...
##  $ KitchenAbvGr: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ KitchenQual : chr  "TA" "Gd" "TA" "Gd" ...
##  $ TotRmsAbvGrd: int  5 6 6 7 5 7 6 7 5 4 ...
##  $ Functional  : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces  : int  0 0 1 1 0 1 0 1 1 0 ...
##  $ FireplaceQu : chr  NA NA "TA" "Gd" ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Attchd" ...
##  $ GarageYrBlt  : int  1961 1958 1997 1998 1992 1993 1992 1998 1990 1970 ...
##  $ GarageFinish : chr  "Unf" "Unf" "Fin" "Fin" ...
##  $ GarageCars   : int  1 1 2 2 2 2 2 2 2 2 ...
##  $ GarageArea   : int  730 312 482 470 506 440 420 393 506 525 ...
##  $ GarageQual   : chr  "TA" "TA" "TA" "TA" ...
##  $ GarageCond   : chr  "TA" "TA" "TA" "TA" ...
##  $ PavedDrive   : chr  "Y" "Y" "Y" "Y" ...
##  $ WoodDeckSF   : int  140 393 212 360 0 157 483 0 192 240 ...
##  $ OpenPorchSF  : int  0 36 34 36 82 84 21 75 0 0 ...
##  $ EnclosedPorch: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ScreenPorch  : int  120 0 0 0 144 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : chr  NA NA NA NA ...
##  $ Fence        : chr  "MnPrv" NA "MnPrv" NA ...
##  $ MiscFeature  : chr  NA "Gar2" NA NA ...
##  $ MiscVal      : int  0 12500 0 0 0 0 500 0 0 0 ...
##  $ MoSold       : int  6 6 3 6 1 4 3 5 2 4 ...
##  $ YrSold       : int  2010 2010 2010 2010 2010 2010 2010 2010 2010 2010 ...
##  $ SaleType     : chr  "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr  "Normal" "Normal" "Normal" "Normal" ...
```

Combine data

Combining data so we can clean and analyze the entire dataset simultaneously.

```
test$SalePrice <- rep(NA, 1459) # adding NA’s to test data sales price so we can join train and test data into
combined <- rbind(train,test)
str(combined)
```

```
## ’data.frame':    2919 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     : chr  "RL" "RL" "RL" "RL" ...
##  $ 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       : chr  "Pave" "Pave" "Pave" "Pave" ...
##  $ Alley        : chr  NA NA NA NA ...
##  $ LotShape     : chr  "Reg" "Reg" "IR1" "IR1" ...
##  $ LandContour  : chr  "Lvl" "Lvl" "Lvl" "Lvl" ...
##  $ Utilities    : chr  "AllPub" "AllPub" "AllPub" "AllPub" ...
##  $ LotConfig    : chr  "Inside" "FR2" "Inside" "Corner" ...
##  $ LandSlope    : chr  "Gtl" "Gtl" "Gtl" "Gtl" ...
##  $ Neighborhood : chr  "CollgCr" "Veenker" "CollgCr" "Crawfor" ...
##  $ Condition1   : chr  "Norm" "Feedr" "Norm" "Norm" ...
##  $ Condition2   : chr  "Norm" "Norm" "Norm" "Norm" ...
##  $ BldgType     : chr  "1Fam" "1Fam" "1Fam" "1Fam" ...
##  $ HouseStyle   : chr  "2Story" "1Story" "2Story" "2Story" ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  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    : chr  "Gable" "Gable" "Gable" "Gable" ...
##  $ RoofMatl     : chr  "CompShg" "CompShg" "CompShg" "CompShg" ...
##  $ Exterior1st  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Sdng" ...
##  $ Exterior2nd  : chr  "VinylSd" "MetalSd" "VinylSd" "Wd Shng" ...
##  $ MasVnrType   : chr  "BrkFace" "None" "BrkFace" "None" ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : chr  "Gd" "TA" "Gd" "TA" ...
##  $ ExterCond    : chr  "TA" "TA" "TA" "TA" ...
##  $ Foundation   : chr  "PConc" "CBlock" "PConc" "BrkTil" ...
##  $ BsmtQual     : chr  "Gd" "Gd" "Gd" "TA" ...
##  $ BsmtCond     : chr  "TA" "TA" "TA" "Gd" ...
##  $ BsmtExposure : chr  "No" "Gd" "Mn" "No" ...
##  $ BsmtFinType1 : chr  "GLQ" "ALQ" "GLQ" "ALQ" ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : chr  "Unf" "Unf" "Unf" "Unf" ...
##  $ 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      : chr  "GasA" "GasA" "GasA" "GasA" ...
##  $ HeatingQC    : chr  "Ex" "Ex" "Ex" "Gd" ...
##  $ CentralAir   : chr  "Y" "Y" "Y" "Y" ...
##  $ Electrical   : chr  "SBrkr" "SBrkr" "SBrkr" "SBrkr" ...
##  $ 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  : chr  "Gd" "TA" "Gd" "Gd" ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : chr  "Typ" "Typ" "Typ" "Typ" ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : chr  NA "TA" "TA" "Gd" ...
##  $ GarageType   : chr  "Attchd" "Attchd" "Attchd" "Detchd" ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
##  $ GarageFinish : chr  "RFn" "RFn" "RFn" "Unf" ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : chr  "TA" "TA" "TA" "TA" ...
##  $ GarageCond   : chr  "TA" "TA" "TA" "TA" ...
##  $ PavedDrive   : chr  "Y" "Y" "Y" "Y" ...
##  $ 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       : chr  NA NA NA NA ...
##  $ Fence        : chr  NA NA NA NA ...
##  $ MiscFeature  : chr  NA NA 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     : chr  "WD" "WD" "WD" "WD" ...
##  $ SaleCondition: chr  "Normal" "Normal" "Normal" "Abnorml" ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000 200000 129900 118000 ...
```

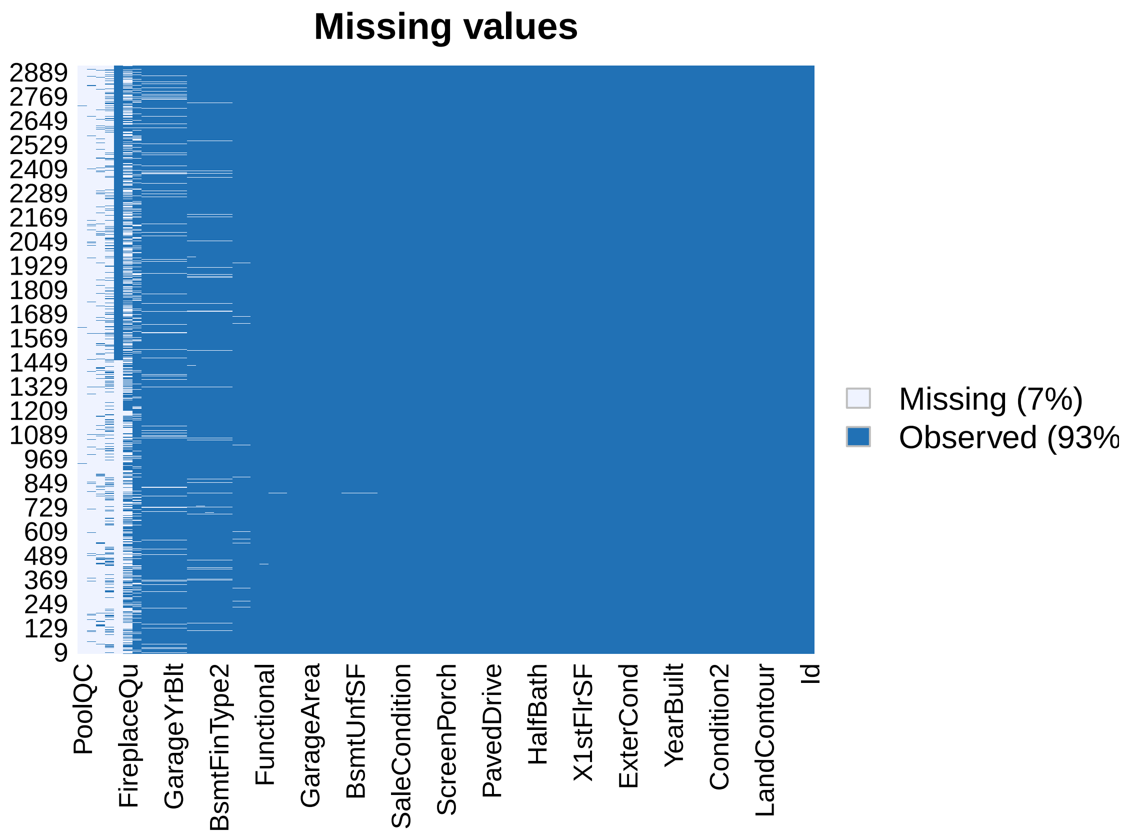
Missing values and label encoding

```r
library(Amelia)
```

```
## Warning: package 'Amelia' was built under R version 4.0.3
## Loading required package: Rcpp

## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.6, built: 2019-11-24)
## ## Copyright (C) 2005-2020 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
misscounts <- sapply(combined,function(x) sum(is.na(x)))
missmap(combined, main = "Missing values")
```



```
sort(colSums(sapply(combined, is.na)), decreasing = T)
```

```
##        PoolQC   MiscFeature         Alley         Fence     SalePrice
##          2909          2814          2721          2348          1459
##   FireplaceQu   LotFrontage   GarageYrBlt  GarageFinish    GarageQual
##          1420           486           159           159           159
##    GarageCond    GarageType      BsmtCond  BsmtExposure      BsmtQual
##           159           157            82            82            81
##  BsmtFinType2  BsmtFinType1    MasVnrType    MasVnrArea      MSZoning
##            80            79            24            23             4
##     Utilities  BsmtFullBath  BsmtHalfBath    Functional   Exterior1st
##             2             2             2             2             1
##   Exterior2nd    BsmtFinSF1    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF
##             1             1             1             1             1
##    Electrical   KitchenQual    GarageCars    GarageArea      SaleType
##             1             1             1             1             1
##            Id    MSSubClass       LotArea        Street      LotShape
##             0             0             0             0             0
##   LandContour     LotConfig     LandSlope  Neighborhood    Condition1
##             0             0             0             0             0
##    Condition2      BldgType    HouseStyle   OverallQual   OverallCond 
##             0             0             0             0             0 
##     YearBuilt  YearRemodAdd     RoofStyle      RoofMatl     ExterQual 
##             0             0             0             0             0 
##     ExterCond    Foundation       Heating     HeatingQC    CentralAir 
##             0             0             0             0             0 
##     X1stFlrSF     X2ndFlrSF  LowQualFinSF     GrLivArea      FullBath 
##             0             0             0             0             0 
##      HalfBath  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd    Fireplaces 
##             0             0             0             0             0 
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch 
##             0             0             0             0             0 
##   ScreenPorch      PoolArea       MiscVal        MoSold        YrSold 
##             0             0             0             0             0 
## SaleCondition 
##             0
```

pool variables

The PoolQC has the most missing values. Pool area does not have missing values but it is related to PoolQC as it does not make sense to have a pool quality data when there is zero pool area or no pool. Its description from the data description document is.

PoolQC: Pool quality

```
Ex   Excellent
Gd   Good
TA   Average/Typical
Fa   Fair
NA   No Pool
```

Since a house with no pool has NA they are not really missing values. we can check with other pool related variables to see if there are any actual missing values in our data.

```
table(is.na(combined$PoolQC))
```

```
## 
## FALSE  TRUE 
##    10  2909
```

```
table(combined$PoolArea, combined$PoolQC, useNA = 'ifany')
```

```
##       
##          Ex   Fa   Gd <NA>
##   0       0    0    0 2906
##   144     1    0    0    0
##   228     1    0    0    0
##   368     0    0    0    1
##   444     0    0    0    1
##   480     0    0    1    0
##   512     1    0    0    0
##   519     0    1    0    0
##   555     1    0    0    0
##   561     0    0    0    1
##   576     0    0    1    0
##   648     0    1    0    0
##   738     0    0    1    0
##   800     0    0    1    0
```

Here we have some actual missing values. We have 13 houses with pool area data but we have only 10 PoolQC data available.

```
library(plyr)
```

```
## Warning: package ’plyr’ was built under R version 4.0.3
```

```
combined[combined$PoolArea==0,]$PoolQC <- "None"

# convert all NA’s in PoolQC to none except for the 3 actual missing values.

combined[is.na(combined$PoolQC),c("OverallQual","PoolArea")]
```

```
##      OverallQual PoolArea
## 2421           4      368
## 2504           6      444
## 2600           3      561
```

```
# imputing the values of poolQC according to overall quality and pool area.

combined[is.na(combined$PoolQC),"PoolQC"] <- c("TA","Gd","TA")

# label encoding as the values are ordinal.

encoding_levels <- c(’None’ = 0, ’Po’ = 1, ’Fa’ = 2, ’TA’ = 3, ’Gd’ = 4, ’Ex’ = 5)

combined$PoolQC <- as.integer(plyr::revalue(combined$PoolQC,encoding_levels))
```

```
## The following ‘from‘ values were not present in ‘x‘: Po
```

```
table(combined$PoolQC)
```

```
##
##    0    2    3    4    5
## 2906    2    2    5    4
```

```
str(as.factor(combined$PoolQC))
```

```
##  Factor w/ 5 levels "0","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...
```

MiscFeature variable

```
table(combined$MiscFeature, useNA = "ifany")
```

```
##
## Gar2 Othr Shed TenC <NA>
##    5    4   95    1 2814
```

In MiscFeature variable, there are 2814 missing values that have to be replaced by none.

```
library(ggplot2)
```

```
## Warning: package ’ggplot2’ was built under R version 4.0.3
```

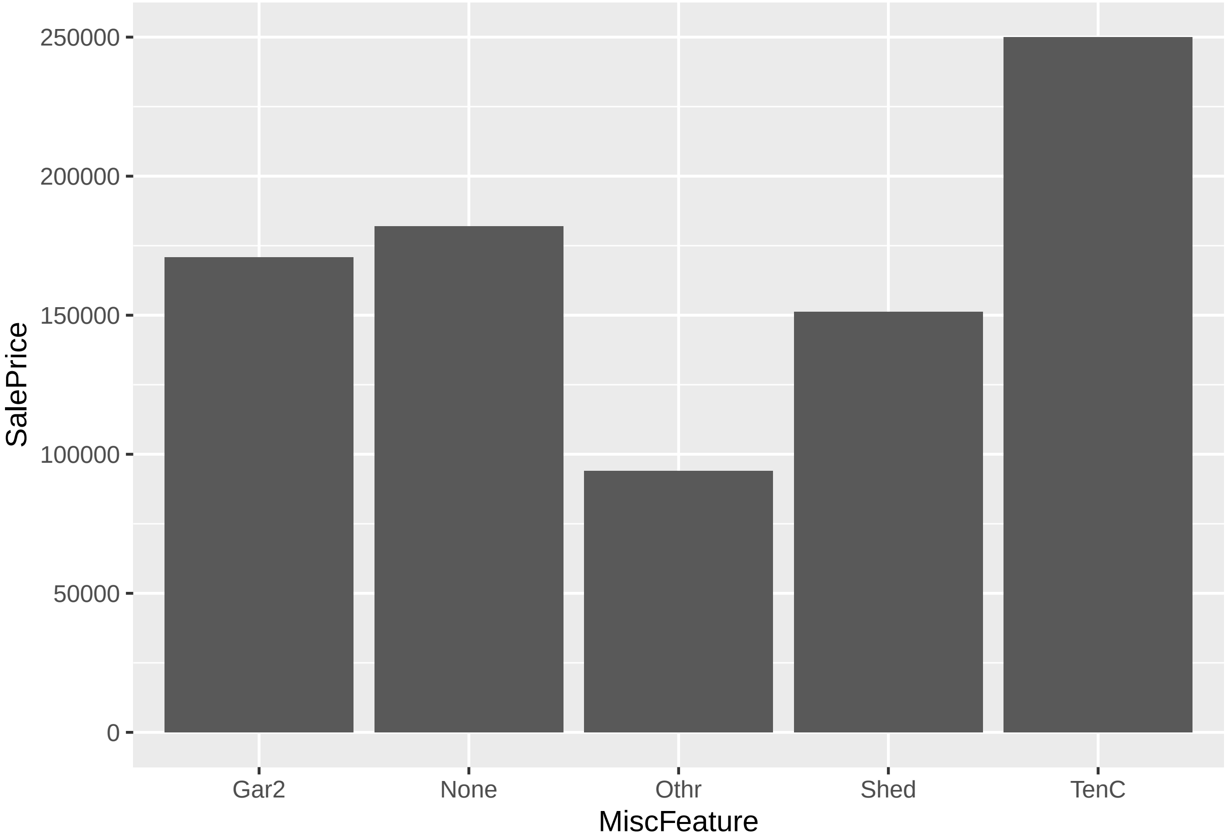
```
# convert all NA's in MiscFeature to none.
combined[is.na(combined$MiscFeature),"MiscFeature"] <- "None"

# convert to factor
combined$MiscFeature <- as.factor(combined$MiscFeature)

ggplot(combined, aes(x=MiscFeature, y = SalePrice)) + geom_bar(stat = 'summary')
```

```
## Warning: Removed 1459 rows containing non-finite values (stat_summary).
```

```
## No summary function supplied, defaulting to `mean_se()`
```



Alley Predictor

```
table(combined$Alley, useNA = "ifany")
```

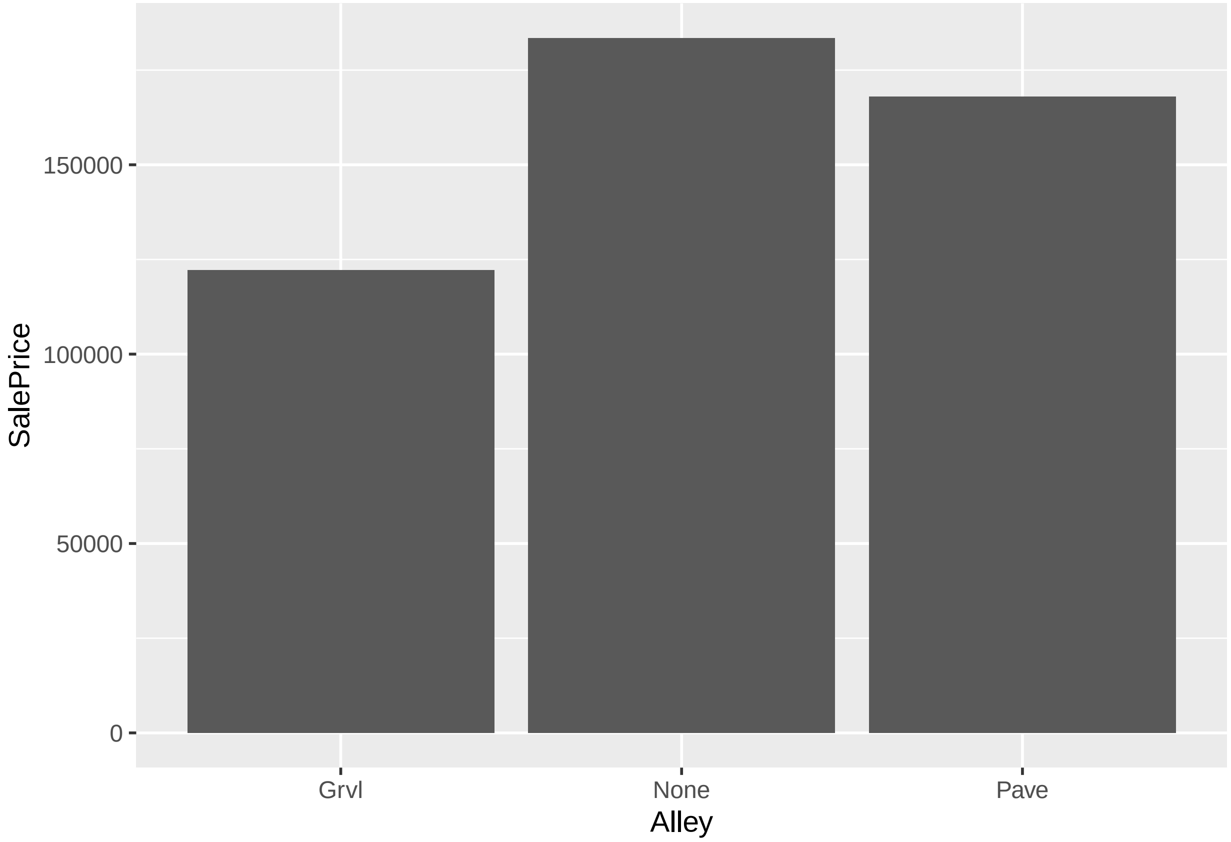
```
##
## Grvl Pave <NA>
##  120   78 2721
```

```
# convert all NA's in Alley to none.
combined[is.na(combined$Alley),"Alley"] <- "None"

# convert to factor
combined$Alley <- as.factor(combined$Alley)

ggplot(combined, aes(x=Alley, y = SalePrice)) + geom_bar(stat = 'summary')
```

```
## Warning: Removed 1459 rows containing non-finite values (stat_summary).
## No summary function supplied, defaulting to ‘mean_se()‘
```



Fence predictor

```
table(combined$Fence, useNA = "ifany")
```

```
##
## GdPrv  GdWo MnPrv  MnWw  <NA>
##   118   112   329    12  2348
```

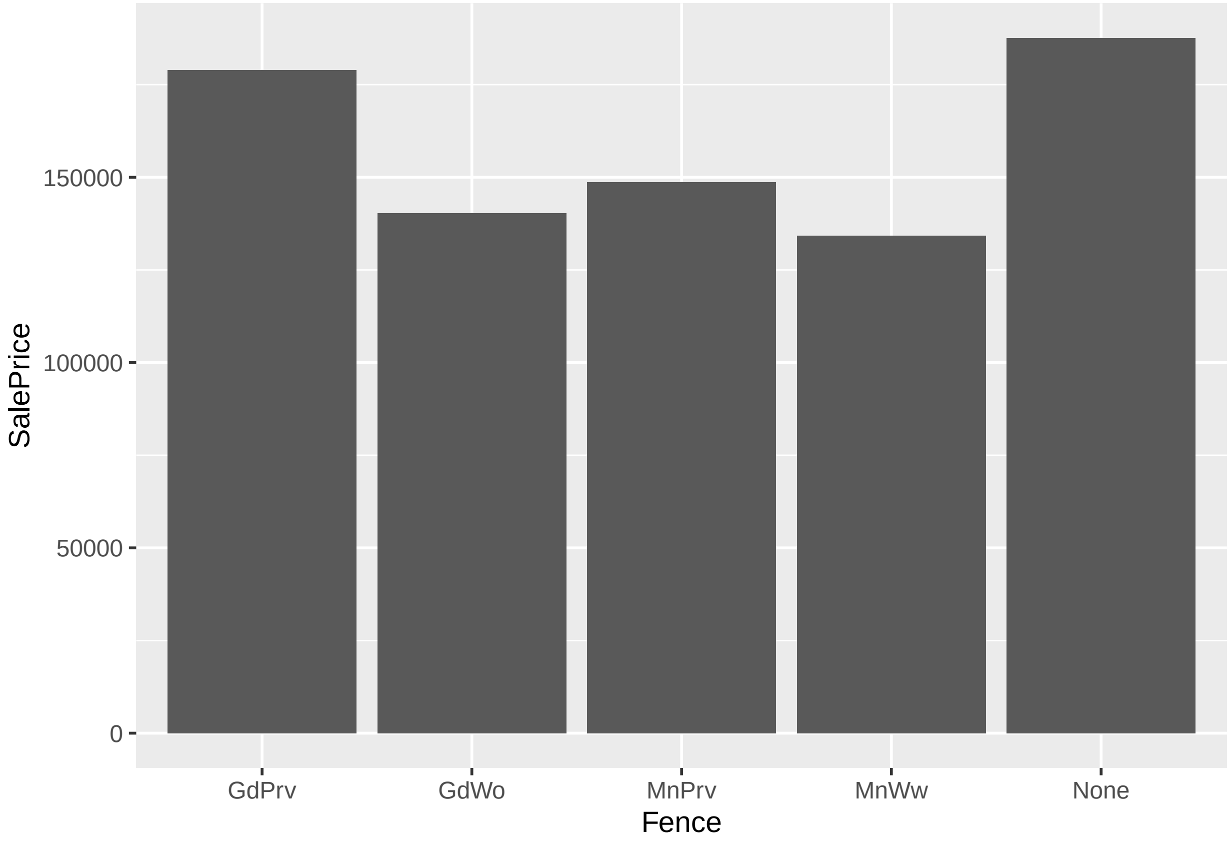
```
# convert all NA’s in Fence to none.
combined[is.na(combined$Fence),"Fence"] <- "None"

# convert to factor
combined$Fence <- as.factor(combined$Fence)

ggplot(combined, aes(x=Fence, y = SalePrice)) + geom_bar(stat = ’summary’)
```

```
## Warning: Removed 1459 rows containing non-finite values (stat_summary).
```

```
## No summary function supplied, defaulting to ‘mean_se()‘
```



Fireplace variables

Fireplace quality

```
table(combined$FireplaceQu, useNA = "ifany")
```

```
##
##   Ex   Fa   Gd   Po   TA <NA>
##   43   74  744   46  592 1420
```

```
# convert all NA's in FireplaceQu to none.

combined[is.na(combined$FireplaceQu),"FireplaceQu"] <- "None"

# Changing and converting to factor levels from character.

combined$FireplaceQu <- as.integer(plyr::revalue(combined$FireplaceQu,encoding_levels))

table(combined$FireplaceQu)
```

```
##
##    0    1    2    3    4    5
## 1420   46   74  592  744   43
```

```
str(combined$FireplaceQu)
```

```
##  int [1:2919] 0 3 3 4 3 0 4 3 3 3 ...
```

```
anyNA(combined$FireplaceQu)
```

```
## [1] FALSE
```

Lot variables

LotFrontage LotShape LotConfig LotArea

```
table(is.na(combined$LotFrontage))
```

```
##
## FALSE  TRUE
##  2433   486
```

Here we have 486 missing values which cannot be replaced by none as it is a numerical variable. So we predict using rpart.

http://r-statistics.co/Missing-Value-Treatment-With-R.html

```
# predictors that lotfrontage variable might depend on.
predictors <- c("MSSubClass", "MSZoning", "LotFrontage", "LotArea", "Street", "Alley", "LotShape", "LandContour"
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 4.0.3
```

```
mod <- rpart(LotFrontage~., data = combined[!is.na(combined$LotFrontage),predictors], method = "anova", na.actio

pred <- predict(mod, combined[is.na(combined$LotFrontage),predictors])
pred <- round(pred)
combined$LotFrontage[is.na(combined$LotFrontage)] <- pred
anyNA(combined$LotFrontage)
```

```
## [1] FALSE
```

```
table(combined$LotShape)
```

```
##
##  IR1  IR2  IR3  Reg
##  968   76   16 1859
```

```
combined$LotShape <- as.integer(plyr::revalue(combined$LotShape,c("Reg" = 3, "IR1" = 2, "IR2" = 1, "IR3" = 0 )))
table(combined$LotConfig)
```

```
##
##  Corner CulDSac     FR2     FR3  Inside
##     511     176      85      14    2133
```

```
combined$LotConfig <- as.factor(combined$LotConfig)
```

```
sort(colSums(sapply(combined, is.na)), decreasing = T)
```

```
##     SalePrice   GarageYrBlt  GarageFinish    GarageQual    GarageCond
##          1459           159           159           159           159
##    GarageType      BsmtCond  BsmtExposure      BsmtQual  BsmtFinType2
##           157            82            82            81            80
##  BsmtFinType1    MasVnrType    MasVnrArea      MSZoning     Utilities
##            79            24            23             4             2
##  BsmtFullBath  BsmtHalfBath    Functional   Exterior1st   Exterior2nd
##             2             2             2             1             1
##    BsmtFinSF1    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF    Electrical
##             1             1             1             1             1
##   KitchenQual    GarageCars    GarageArea      SaleType            Id
##             1             1             1             1             0
##    MSSubClass   LotFrontage       LotArea        Street         Alley
##             0             0             0             0             0
##      LotShape   LandContour     LotConfig     LandSlope  Neighborhood
##             0             0             0             0             0
##    Condition1    Condition2      BldgType    HouseStyle   OverallQual
##             0             0             0             0             0
##   OverallCond     YearBuilt  YearRemodAdd     RoofStyle      RoofMatl
##             0             0             0             0             0
##     ExterQual     ExterCond    Foundation       Heating     HeatingQC
##             0             0             0             0             0
##    CentralAir     X1stFlrSF     X2ndFlrSF  LowQualFinSF     GrLivArea
##             0             0             0             0             0
##      FullBath      HalfBath  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd
##             0             0             0             0             0
##    Fireplaces   FireplaceQu    PavedDrive    WoodDeckSF   OpenPorchSF
##             0             0             0             0             0
## EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea        PoolQC
##             0             0             0             0             0
##         Fence   MiscFeature       MiscVal        MoSold        YrSold
##             0             0             0             0             0
## SaleCondition
##             0
```

Garage variables

GarageYrBlt GarageType GarageFinish, GarageQual, GarageCond, GarageCars, GarageArea

```
garage <- c("GarageYrBlt","GarageType","GarageFinish","GarageQual","GarageCond","GarageCars","GarageArea")
sort(colSums(sapply(combined[,garage], is.na)), decreasing = T)
```

```
##  GarageYrBlt GarageFinish   GarageQual   GarageCond   GarageType   GarageCars
##          159          159          159          159          157            1
##   GarageArea
##            1
```

```
combined$GarageYrBlt[is.na(combined$GarageYrBlt)] <- combined$YearBuilt[is.na(combined$GarageYrBlt)]

which(!is.na(combined$GarageType) & is.na(combined$GarageFinish) & is.na(combined$GarageCond) & is.na(combined$G
```

```
## [1] 2127 2577
```

```
combined[c(2127,2577),c("GarageType","GarageFinish","GarageCond","GarageQual","GarageCars","GarageArea")]
```

```
##      GarageType GarageFinish GarageCond GarageQual GarageCars GarageArea
## 2127     Detchd         <NA>       <NA>       <NA>          1        360
## 2577     Detchd         <NA>       <NA>       <NA>         NA         NA
```

```
# impute mode
combined[c(2127),"GarageFinish"] <- names(sort(-table(combined$GarageFinish)))[1]
combined[c(2127),"GarageCond"] <- names(sort(-table(combined$GarageCond)))[1]
combined[c(2127),"GarageQual"] <- names(sort(-table(combined$GarageQual)))[1]
```

```
combined[c(2577),"GarageFinish"] <- "None"
combined[c(2577),"GarageCond"] <- "None"
combined[c(2577),"GarageQual"] <- "None"
combined[c(2577),"GarageType"] <- "None"
combined[c(2577),"GarageCars"] <- 0
combined[c(2577),"GarageArea"] <- 0

which(!is.na(combined$GarageType) & is.na(combined$GarageFinish) & is.na(combined$GarageCond) & is.na(combined$G
```

```
## integer(0)
```

```
combined$GarageType[is.na(combined$GarageType)] <- "None"
combined$GarageFinish[is.na(combined$GarageFinish)] <- "None"
combined$GarageCond[is.na(combined$GarageCond)] <- "None"
combined$GarageQual[is.na(combined$GarageQual)] <- "None"
sort(colSums(sapply(combined[,garage], is.na)), decreasing = T)
```

```
##  GarageYrBlt   GarageType GarageFinish   GarageQual   GarageCond   GarageCars
##            0            0            0            0            0            0
##   GarageArea
##            0
```

```
# convert into factor
combined$GarageType <- as.factor(combined$GarageType)
table(combined$GarageType)
```

```
##
##  2Types  Attchd Basment BuiltIn CarPort  Detchd    None
##      23    1723      36     186      15     778     158
```

```
# convert into ordinal
Finish <- c(’None’=0, ’Unf’=1, ’RFn’=2, ’Fin’=3)
combined$GarageFinish<-as.integer(revalue(combined$GarageFinish, Finish))
table(combined$GarageFinish)
```

```
##
##    0    1    2    3
##  158 1231  811  719
```

```
combined$GarageCond<-as.integer(revalue(combined$GarageCond, encoding_levels))
table(combined$GarageCond)
```

```
##
##    0    1    2    3    4    5
##  158   14   74 2655   15    3
```

```
combined$GarageQual<-as.integer(revalue(combined$GarageQual, encoding_levels))
table(combined$GarageQual)
```

```
##
##    0    1    2    3    4    5
##  158    5  124 2605   24    3
```

Basement variables

there are 11 basement variables

BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2, BsmtFullBath, BsmtHalfBath, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF,

```
basement <- c("BsmtQual","BsmtCond","BsmtExposure","BsmtFinType1","BsmtFinType2","BsmtFullBath","BsmtHalfBath","
sort(colSums(sapply(combined[,basement], is.na)), decreasing = T)
```

```
##     BsmtCond BsmtExposure     BsmtQual BsmtFinType2 BsmtFinType1 BsmtFullBath
##           82           82           81           80           79            2
## BsmtHalfBath   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF
##            2            1            1            1            1
```

```
x <- which(!is.na(combined$BsmtFinType1) & (is.na(combined$BsmtCond)|is.na(combined$BsmtExposure)|is.na(combined

combined[x,basement]
```

```
##      BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtFullBath
## 333        Gd       TA           No          GLQ         <NA>            1
## 949        Gd       TA         <NA>          Unf          Unf            0
## 1488       Gd       TA         <NA>          Unf          Unf            0
## 2041       Gd     <NA>           Mn          GLQ          Rec            1
## 2186       TA     <NA>           No          BLQ          Unf            0
## 2218     <NA>       Fa           No          Unf          Unf            0
## 2219     <NA>       TA           No          Unf          Unf            0
## 2349       Gd       TA         <NA>          Unf          Unf            0
## 2525       TA     <NA>           Av          ALQ          Unf            0
##      BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 333             0       1124        479      1603        3206
## 949             0          0          0       936         936
## 1488            0          0          0      1595        1595
## 2041            0       1044        382         0        1426
## 2186            1       1033          0        94        1127
## 2218            0          0          0       173         173
## 2219            0          0          0       356         356
## 2349            0          0          0       725         725
## 2525            0        755          0       240         995
```

```
# impute mode
combined[c(2218,2219),"BsmtQual"] <- names(sort(-table(combined$BsmtQual)))[1]
combined[c(2041,2186,2525),"BsmtCond"] <- names(sort(-table(combined$BsmtCond)))[1]
combined[c(949,1488,2349),"BsmtExposure"] <- names(sort(-table(combined$BsmtExposure)))[1]
combined[c(333),"BsmtFinType2"] <- names(sort(-table(combined$BsmtFinType2)))[1]
combined[x,basement]
```

```
##      BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtFullBath
## 333        Gd       TA           No          GLQ          Unf            1
## 949        Gd       TA           No          Unf          Unf            0
## 1488       Gd       TA           No          Unf          Unf            0
## 2041       Gd       TA           Mn          GLQ          Rec            1
## 2186       TA       TA           No          BLQ          Unf            0
## 2218       TA       Fa           No          Unf          Unf            0
## 2219       TA       TA           No          Unf          Unf            0
## 2349       Gd       TA           No          Unf          Unf            0
## 2525       TA       TA           Av          ALQ          Unf            0
##      BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 333             0       1124        479      1603        3206
## 949             0          0          0       936         936
## 1488            0          0          0       1595        1595
## 2041            0       1044        382          0        1426
## 2186            1       1033          0         94        1127
## 2218            0          0          0        173         173
## 2219            0          0          0        356         356
## 2349            0          0          0        725         725
## 2525            0        755          0        240         995
```

```
anyNA(combined[x,basement])
```

```
## [1] FALSE
```

```
sort(colSums(sapply(combined[,basement], is.na)), decreasing = T)
```

```
##     BsmtQual     BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtFullBath
##           79           79           79           79           79            2
## BsmtHalfBath   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF
##            2            1            1            1            1
```

```
combined[is.na(combined[,"TotalBsmtSF"]),basement]
```

```
##      BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtFullBath
## 2121     <NA>     <NA>         <NA>         <NA>         <NA>           NA
##      BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 2121           NA         NA         NA        NA          NA
```

```
combined[2121,"BsmtQual"]  <- "None"
combined[2121,"BsmtCond"]  <- "None"
combined[2121,"BsmtExposure"]  <- "None"
combined[2121,"BsmtFinType1"]  <- "None"
combined[2121,"BsmtFinType2"]  <- "None"
combined[2121,"BsmtFullBath"]  <- 0
combined[2121,"BsmtHalfBath"]  <- 0
combined[2121,"BsmtFinSF1"]  <- 0
combined[2121,"BsmtFinSF2"]  <- 0
combined[2121,"BsmtUnfSF"]  <- 0
combined[2121,"TotalBsmtSF"]  <- 0

sort(colSums(sapply(combined[,basement], is.na)), decreasing = T)
```

```
##     BsmtQual     BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtFullBath
##           78           78           78           78           78            1
## BsmtHalfBath   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF
##            1            0            0            0            0
```

```
combined[is.na(combined[,"BsmtHalfBath"]),basement]
```

```
##      BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtFullBath
## 2189     <NA>     <NA>         <NA>         <NA>         <NA>           NA
##      BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 2189           NA          0          0         0           0
```

```
combined$BsmtQual[is.na(combined$BsmtQual)] <- "None"
combined$BsmtCond[is.na(combined$BsmtCond)] <- "None"
combined$BsmtExposure[is.na(combined$BsmtExposure)] <- "None"
combined$BsmtFinType1[is.na(combined$BsmtFinType1)] <- "None"
combined$BsmtFinType2[is.na(combined$BsmtFinType2)] <- "None"
```

```
combined$BsmtFullBath[is.na(combined$BsmtFullBath)] <- 0
combined$BsmtHalfBath[is.na(combined$BsmtHalfBath)] <- 0

sort(colSums(sapply(combined[,basement], is.na)), decreasing = T)
```

```
##     BsmtQual     BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtFullBath
##            0            0            0            0            0            0
## BsmtHalfBath   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF
##            0            0            0            0            0
```

```
# convert to ordinal

combined$BsmtQual<-as.integer(revalue(combined$BsmtQual, encoding_levels))
```

```
## The following ‘from‘ values were not present in ‘x‘: Po
```

```
table(combined$BsmtQual)
```

```
##
##    0    2    3    4    5
##   79   88 1285 1209  258
```

```
combined$BsmtCond<-as.integer(revalue(combined$BsmtCond, encoding_levels))
```

```
## The following ‘from‘ values were not present in ‘x‘: Ex
```

```
table(combined$BsmtCond)
```

```
##
##    0    1    2    3    4
##   79    5  104 2609  122
```

```
exposure <- c(’Gd’ = 4,’Av’ = 3,’Mn’ = 2,’No’ = 1,’None’ = 0)
combined$BsmtExposure<-as.integer(revalue(combined$BsmtExposure, exposure))
table(combined$BsmtExposure)
```

```
##
##    0    1    2    3    4
##   79 1907  239  418  276
```

```
rating <- c(’GLQ’ = 6,’ALQ’ = 5,’BLQ’ = 4,’Rec’ = 3,’LwQ’ = 2,’Unf’ = 1,’None’ = 0)
combined$BsmtFinType1<-as.integer(revalue(combined$BsmtFinType1, rating))
table(combined$BsmtFinType1)
```

```
##
##   0   1   2   3   4   5   6
##  79 851 154 288 269 429 849
```

```
combined$BsmtFinType2<-as.integer(revalue(combined$BsmtFinType2, rating))
table(combined$BsmtFinType2)
```

```
##
##    0    1    2    3    4    5    6
##   79 2494   87  105   68   52   34
```

masonry variables

```r
sort(colSums(sapply(combined, is.na)), decreasing = T)
```

```
##     SalePrice    MasVnrType    MasVnrArea      MSZoning     Utilities
##          1459            24            23             4             2
##    Functional   Exterior1st   Exterior2nd    Electrical   KitchenQual
##             2             1             1             1             1
##      SaleType            Id    MSSubClass   LotFrontage       LotArea
##             1             0             0             0             0
##        Street         Alley      LotShape   LandContour     LotConfig
##             0             0             0             0             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl     ExterQual     ExterCond    Foundation
##             0             0             0             0             0
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##             0             0             0             0             0
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##             0             0             0             0             0
##     HeatingQC    CentralAir     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             0             0             0             0
##  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd    Fireplaces   FireplaceQu
##             0             0             0             0             0
##    GarageType   GarageYrBlt  GarageFinish    GarageCars    GarageArea
##             0             0             0             0             0
##    GarageQual    GarageCond    PavedDrive    WoodDeckSF   OpenPorchSF
##             0             0             0             0             0
## EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea        PoolQC
##             0             0             0             0             0
##         Fence   MiscFeature       MiscVal        MoSold        YrSold
##             0             0             0             0             0
## SaleCondition
##             0
```

```r
x <- which(!is.na(combined$MasVnrArea) & is.na(combined$MasVnrType) )
combined[x,c("MasVnrArea","MasVnrType")]
```

```
##      MasVnrArea MasVnrType
## 2611        198       <NA>
```

```r
combined[2611,"MasVnrType"] <- names(sort(-table(combined$MasVnrType)))[1]
combined$MasVnrType[is.na(combined$MasVnrType)] <- "None"
combined$MasVnrArea[is.na(combined$MasVnrArea)] <- 0
combined$MasVnrType <- as.factor(combined$MasVnrType)
table(combined$MasVnrType)
```

```
##
##  BrkCmn BrkFace    None   Stone
##      25     879    1766     249
```

catogorical variables

Below are categorical variables identified from data description:

GarageType MSZoning, Exterior1st, Exterior2nd, Electrical, SaleType, SaleCondition, Foundation, Heating, CentralAir, RoofStyle, RoofMatl, LandContour, BldgType, HouseStyle, Neighborhood, Condition1, Condition2, Street, MSSubClass, MoSold, YrSold

```
sort(colSums(sapply(combined, is.na)), decreasing = T)
```

```
##     SalePrice      MSZoning     Utilities    Functional   Exterior1st
##          1459             4             2             2             1
##   Exterior2nd    Electrical   KitchenQual      SaleType            Id
##             1             1             1             1             0
##    MSSubClass   LotFrontage       LotArea        Street         Alley
##             0             0             0             0             0
##      LotShape   LandContour     LotConfig     LandSlope  Neighborhood
##             0             0             0             0             0
##    Condition1    Condition2      BldgType    HouseStyle   OverallQual
##             0             0             0             0             0
##   OverallCond     YearBuilt  YearRemodAdd     RoofStyle      RoofMatl
##             0             0             0             0             0
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##             0             0             0             0             0
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##             0             0             0             0             0
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##             0             0             0             0             0
##     HeatingQC    CentralAir     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             0             0             0             0
##  BedroomAbvGr  KitchenAbvGr  TotRmsAbvGrd    Fireplaces   FireplaceQu
##             0             0             0             0             0
##    GarageType   GarageYrBlt  GarageFinish    GarageCars    GarageArea
##             0             0             0             0             0
##    GarageQual    GarageCond    PavedDrive    WoodDeckSF   OpenPorchSF
##             0             0             0             0             0
## EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea        PoolQC
##             0             0             0             0             0
##         Fence   MiscFeature       MiscVal        MoSold        YrSold
##             0             0             0             0             0
## SaleCondition
##             0
```

```
categorical_variables <- c(’GarageType’,"MSZoning","Utilities","Exterior1st","Exterior2nd","Electrical","SaleTyp

table(combined$MSZoning, useNA = "ifany")
```

```
##
## C (all)      FV      RH      RL      RM    <NA>
##      25     139      26    2265     460       4
```

```
combined$MSZoning[is.na(combined$MSZoning)] <- names(sort(-table(combined$MSZoning)))[1]
combined$MSZoning <- as.factor(combined$MSZoning)

table(combined$Utilities, useNA = "ifany")
```

```
##
## AllPub NoSeWa   <NA>
##   2916      1      2
```

```
combined$Utilities[is.na(combined$Utilities)] <- names(sort(-table(combined$Utilities)))[1]
combined$Utilities <- as.factor(combined$Utilities)

table(combined$Exterior1st, useNA = "ifany")
```

```
##
## AsbShng AsphShn BrkComm BrkFace  CBlock CemntBd HdBoard ImStucc MetalSd Plywood
##      44       2       6      87       2     126     442       1     450     221
##   Stone  Stucco VinylSd Wd Sdng WdShing    <NA>
##       2      43    1025     411      56       1
```

```
combined$Exterior1st[is.na(combined$Exterior1st)] <-
names(sort(-table(combined$Exterior1st)))[1]
combined$Exterior1st <- as.factor(combined$Exterior1st)

table(combined$Exterior2nd, useNA = "ifany")
```

```
##
## AsbShng AsphShn Brk Cmn BrkFace  CBlock CmentBd HdBoard ImStucc MetalSd   Other
##      38       4      22      47       3     126     406      15     447       1
## Plywood   Stone  Stucco VinylSd Wd Sdng Wd Shng    <NA>
##     270       6      47    1014     391      81       1
```

```
combined$Exterior2nd[is.na(combined$Exterior2nd)] <-
names(sort(-table(combined$Exterior2nd)))[1]
combined$Exterior2nd <- as.factor(combined$Exterior2nd)

table(combined$Electrical, useNA = "ifany")
```

```
##
## FuseA FuseF FuseP   Mix SBrkr  <NA>
##   188    50     8     1  2671     1
```

```
combined$Electrical[is.na(combined$Electrical)] <-
names(sort(-table(combined$Electrical)))[1]
combined$Electrical <- as.factor(combined$Electrical)

table(combined$SaleType, useNA = "ifany")
```

```
##
##   COD   Con ConLD ConLI ConLw   CWD   New   Oth    WD  <NA>
##    87     5    26     9     8    12   239     7  2525     1
```

```
combined$SaleType[is.na(combined$SaleType)] <-
names(sort(-table(combined$SaleType)))[1]
combined$SaleType <- as.factor(combined$SaleType)

x <- sort(colSums(sapply(combined[,categorical_variables], is.na)), decreasing = T)
x
```

```
##    GarageType      MSZoning     Utilities   Exterior1st   Exterior2nd
##             0             0             0             0             0
##    Electrical      SaleType SaleCondition    Foundation       Heating
##             0             0             0             0             0
##    CentralAir     RoofStyle      RoofMatl   LandContour      BldgType
##             0             0             0             0             0
##    HouseStyle  Neighborhood    Condition1    Condition2        Street
##             0             0             0             0             0
##    MSSubClass        MoSold        YrSold
##             0             0             0
```

```r
for(i in 1:length(names(x)))
{
        combined[,names(x)[i]] <- as.factor(combined[,names(x)[i]])
}

str(combined[,categorical_variables])
```

```
## ’data.frame':    2919 obs. of  23 variables:
##  $ GarageType   : Factor w/ 7 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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 ...
##  $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
##  $ 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 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ 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 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ 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 ...
##  $ 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 ...
##  $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
##  $ MSSubClass   : Factor w/ 16 levels "20","30","40",..: 6 1 6 7 6 5 1 6 5 16 ...
##  $ MoSold       : Factor w/ 12 levels "1","2","3","4",..: 2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : Factor w/ 5 levels "2006","2007",..: 3 2 3 1 3 4 2 4 3 3 ...
```

```r
sort(colSums(sapply(combined, is.na)), decreasing = T)
```

```
##     SalePrice    Functional   KitchenQual            Id    MSSubClass
##          1459             2             1             0             0
##      MSZoning   LotFrontage       LotArea        Street         Alley
##             0             0             0             0             0
##      LotShape   LandContour     Utilities     LotConfig     LandSlope
##             0             0             0             0             0
##  Neighborhood    Condition1    Condition2      BldgType    HouseStyle
##             0             0             0             0             0
##   OverallQual   OverallCond     YearBuilt  YearRemodAdd     RoofStyle
##             0             0             0             0             0
##      RoofMatl   Exterior1st   Exterior2nd    MasVnrType    MasVnrArea
##             0             0             0             0             0
##     ExterQual     ExterCond    Foundation      BsmtQual      BsmtCond
##             0             0             0             0             0
##  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2    BsmtFinSF2
##             0             0             0             0             0
##     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC    CentralAir
##             0             0             0             0             0
##    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF     GrLivArea
##             0             0             0             0             0
##  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath  BedroomAbvGr
##             0             0             0             0             0
##  KitchenAbvGr  TotRmsAbvGrd    Fireplaces   FireplaceQu    GarageType
##             0             0             0             0             0
##   GarageYrBlt  GarageFinish    GarageCars    GarageArea    GarageQual
##             0             0             0             0             0
##    GarageCond    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch
##             0             0             0             0             0
##    X3SsnPorch   ScreenPorch      PoolArea        PoolQC         Fence
##             0             0             0             0             0
##   MiscFeature       MiscVal        MoSold        YrSold      SaleType
##             0             0             0             0             0
## SaleCondition
##             0
```

Ordinal variables

Below are ordinal variables identified from data description:

```
combined[is.na(combined$Functional),"Functional"] <- names(sort(-table(combined$Functional)))[1]
functionality <- c(’Sal’=0, ’Sev’=1, ’Maj2’=2, ’Maj1’=3, ’Mod’=4, ’Min2’=5, ’Min1’=6, ’Typ’=7)
combined$Functional <- as.integer(revalue(combined$Functional,functionality))
```

```
## The following ‘from‘ values were not present in ‘x‘: Sal
```

```
combined[is.na(combined$KitchenQual),"KitchenQual"] <- names(sort(-table(combined$KitchenQual)))[1]
combined$KitchenQual <- as.integer(revalue(combined$KitchenQual,encoding_levels))
```

```
## The following ‘from‘ values were not present in ‘x‘: None, Po
```

```
sort(colSums(sapply(combined, is.na)), decreasing = T)
```

```
##     SalePrice            Id    MSSubClass      MSZoning   LotFrontage
##          1459             0             0             0             0
##       LotArea        Street         Alley      LotShape   LandContour
##             0             0             0             0             0
##     Utilities     LotConfig     LandSlope  Neighborhood    Condition1
##             0             0             0             0             0
##    Condition2      BldgType    HouseStyle   OverallQual   OverallCond
##             0             0             0             0             0
##     YearBuilt  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st
##             0             0             0             0             0
##   Exterior2nd    MasVnrType    MasVnrArea     ExterQual     ExterCond
##             0             0             0             0             0
##    Foundation      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1
##             0             0             0             0             0
##    BsmtFinSF1  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF
##             0             0             0             0             0
##       Heating     HeatingQC    CentralAir    Electrical     X1stFlrSF
##             0             0             0             0             0
##     X2ndFlrSF  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath
##             0             0             0             0             0
##      FullBath      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual
##             0             0             0             0             0
##  TotRmsAbvGrd    Functional    Fireplaces   FireplaceQu    GarageType
##             0             0             0             0             0
##   GarageYrBlt  GarageFinish    GarageCars    GarageArea    GarageQual
##             0             0             0             0             0
##    GarageCond    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch
##             0             0             0             0             0
##    X3SsnPorch   ScreenPorch      PoolArea        PoolQC         Fence
##             0             0             0             0             0
##   MiscFeature       MiscVal        MoSold        YrSold      SaleType
##             0             0             0             0             0
## SaleCondition
##             0
```

```
char_columns <- names(combined[,sapply(combined, is.character)])
char_columns
```

```
## [1] "LandSlope"  "ExterQual"  "ExterCond"  "HeatingQC"  "PavedDrive"
```

```
#label encode remaining char variables
combined$LandSlope <- as.integer(revalue(combined$LandSlope,c(’Gtl’ = 2,’Mod’ = 1,’Sev’ = 0)))
combined$ExterQual <- as.integer(revalue(combined$ExterQual,encoding_levels))
```

```
## The following ‘from‘ values were not present in ‘x‘: None, Po
```

```
combined$ExterCond <- as.integer(revalue(combined$ExterCond,encoding_levels))
```

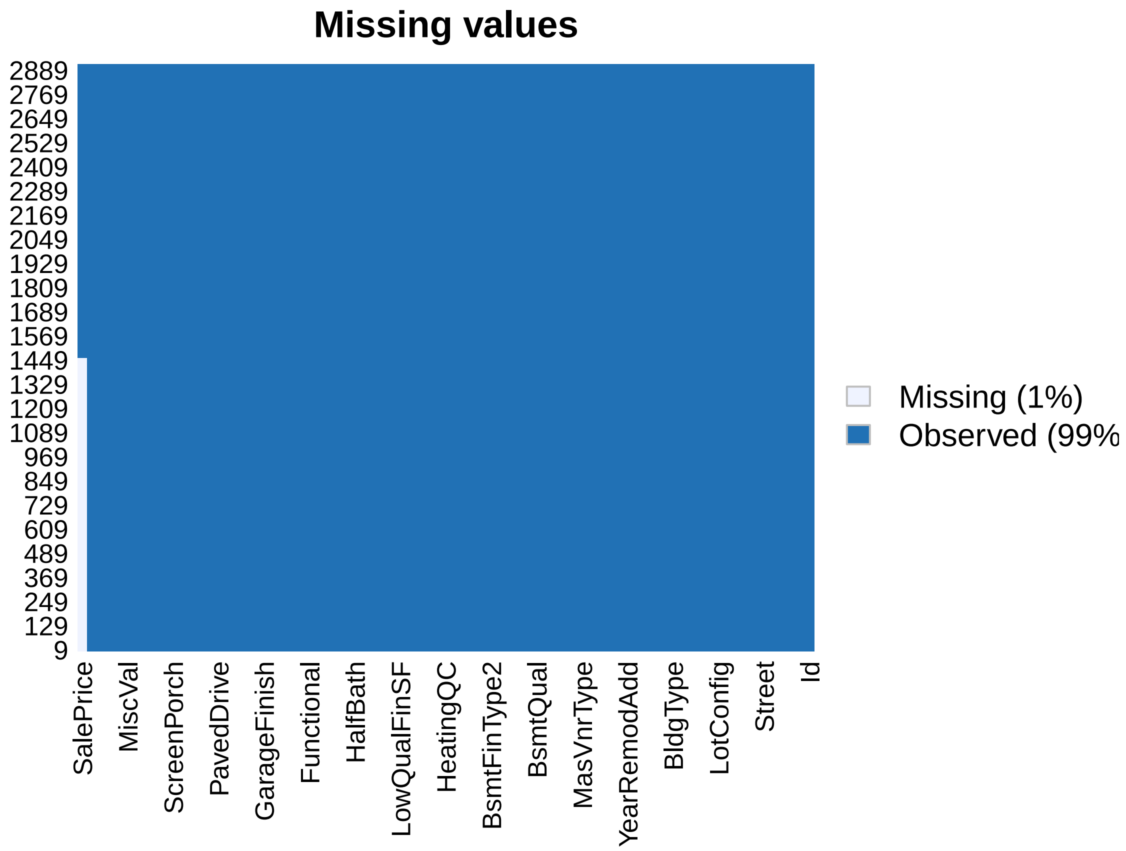
```
## The following ‘from‘ values were not present in ‘x‘: None
```

```
combined$HeatingQC <- as.integer(revalue(combined$HeatingQC,encoding_levels))
```

```
## The following ‘from‘ values were not present in ‘x‘: None
```

```
combined$PavedDrive <- as.integer(revalue(combined$PavedDrive,c(’Y’=2,’P’=1,’N’=0)))
```

```
misscounts <- sapply(combined,function(x) sum(is.na(x)))
missmap(combined, main = "Missing values")
```



As we can see there are no missing values except in SalePrice as this indicates the obeservations for test data.

```
num_vars <- which(sapply(combined,is.numeric))
factor_vars <- which(sapply(combined,is.factor))
cat(’numeric variables: ’, length(num_vars),’ and categorical variables: ’,length(factor_vars),’\n’)
```

```
## numeric variables:  53  and categorical variables:  28
```

```
str(combined)
```

```
## 'data.frame':    2919 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : Factor w/ 16 levels "20","30","40",..: 6 1 6 7 6 5 1 6 5 16 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
##  $ LotFrontage  : num  65 80 68 60 84 85 75 74 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/ 3 levels "Grvl","None",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ LotShape     : int  3 3 2 2 2 2 3 2 3 3 ...
##  $ 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    : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ 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  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  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   : num  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : int  4 3 4 3 4 3 4 3 3 3 ...
##  $ ExterCond    : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
##  $ BsmtQual     : int  4 4 4 3 4 4 5 4 3 3 ...
##  $ BsmtCond     : int  3 3 3 4 3 3 3 3 3 3 ...
##  $ BsmtExposure : int  1 4 2 1 3 1 3 2 1 1 ...
##  $ BsmtFinType1 : int  6 5 6 5 6 6 6 5 1 6 ...
##  $ BsmtFinSF1   : num  706 978 486 216 655 ...
##  $ BsmtFinType2 : int  1 1 1 1 1 1 1 4 1 1 ...
##  $ BsmtFinSF2   : num  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : num  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : num  856 1262 920 756 1145 ...
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ HeatingQC    : int  5 5 5 4 5 5 5 5 4 5 ...
##  $ 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 : num  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : num  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  : int  4 3 4 4 4 3 4 3 3 3 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : int  7 7 7 7 7 7 7 7 6 7 ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : int  0 3 3 4 3 0 4 3 3 3 ...
##  $ GarageType   : Factor w/ 7 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 : int  2 2 2 1 2 1 2 2 1 2 ...
##  $ GarageCars   : num  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : num  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : int  3 3 3 3 3 3 3 3 2 4 ...
##  $ GarageCond   : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ PavedDrive   : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ 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       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Fence        : Factor w/ 5 levels "GdPrv","GdWo",..: 5 5 5 5 5 3 5 5 5 5 ...
##  $ MiscFeature  : Factor w/ 5 levels "Gar2","None",..: 2 2 2 2 2 4 2 4 2 2 ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : Factor w/ 12 levels "1","2","3","4",..: 2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : Factor w/ 5 levels "2006","2007",..: 3 2 3 1 3 4 2 4 3 3 ...
##  $ 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 ...
```

EDA