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
```

```
combined <- df

str(combined)
```

```
## '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 ...
```

Missing values

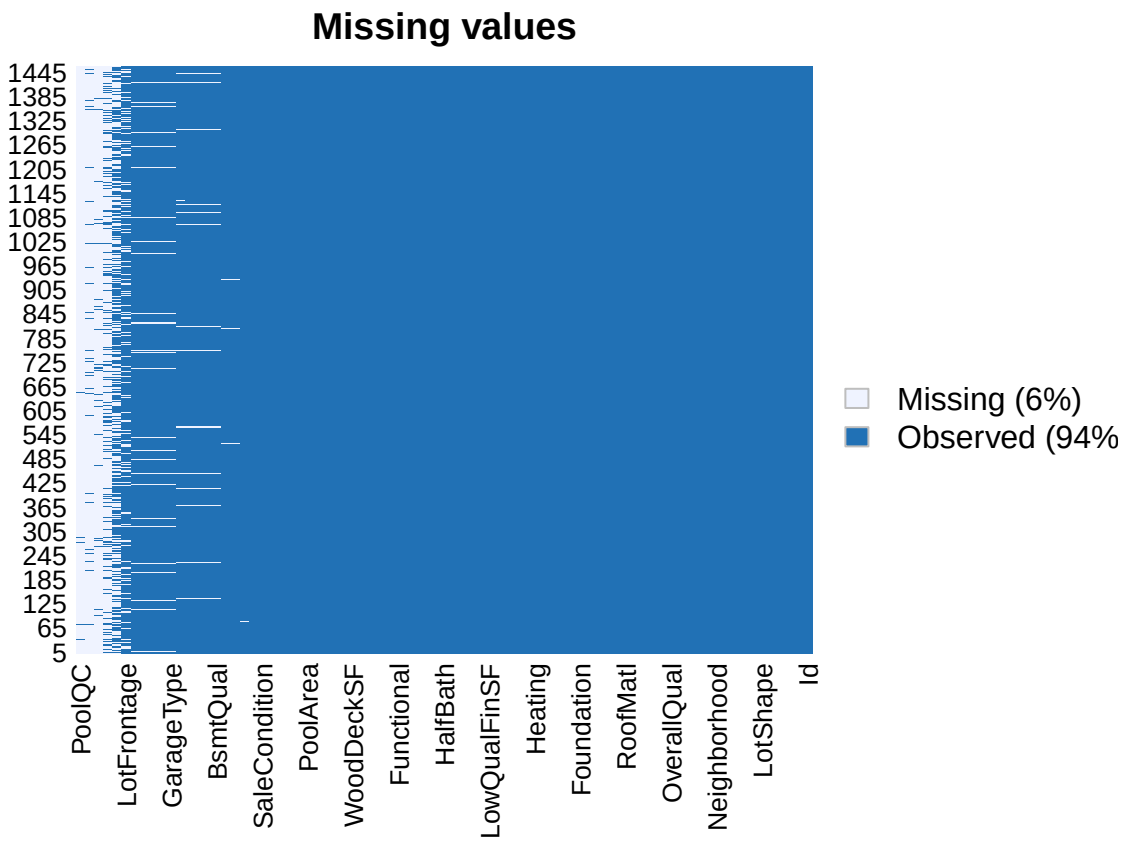
```
library(Amelia)
```

```
## 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   FireplaceQu
##          1453          1406          1369          1179           690
##   LotFrontage    GarageType   GarageYrBlt GarageFinish    GarageQual
##           259            81            81            81            81
##    GarageCond  BsmtExposure  BsmtFinType2      BsmtQual      BsmtCond
##            81            38            38            37            37
##  BsmtFinType1    MasVnrType    MasVnrArea    Electrical            Id
##            37             8             8             1             0
##    MSSubClass      MSZoning       LotArea        Street      LotShape
##             0             0             0             0             0
##   LandContour     Utilities     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
##   Exterior1st   Exterior2nd     ExterQual     ExterCond    Foundation
##             0             0             0             0             0
##    BsmtFinSF1    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   KitchenQual  TotRmsAbvGrd    Functional
##             0             0             0             0             0
##    Fireplaces    GarageCars    GarageArea    PavedDrive    WoodDeckSF
##             0             0             0             0             0
##   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
```

```
##             0             0             0             0             0
##     SalePrice
##             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
##     7  1453
```

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

```
##
##         Ex   Fa   Gd <NA>
##   0      0    0    0 1453
##   480    0    0    1    0
##   512    1    0    0    0
##   519    0    1    0    0
##   555    1    0    0    0
##   576    0    0    1    0
##   648    0    1    0    0
##   738    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")]
```

```
## [1] OverallQual PoolArea
## <0 rows> (or 0-length row.names)
```

```r
# 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', 'Po' , 'Fa', 'TA' , 'Gd', 'Ex' )

combined$PoolQC <- factor(combined$PoolQC, order = TRUE, levels = encoding_levels)

table(combined$PoolQC)
```

```
##
## None   Po   Fa   TA   Gd   Ex
## 1453    0    2    0    3    2
```

```r
str(combined$PoolQC)
```

```
##  Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 1 1 1 1 1 1 1 1 1 1 ...
```

MiscFeature variable

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

```
##
## Gar2 Othr Shed TenC <NA>
##    2    2   49    1 1406
```

In MiscFeature variable, there are 1406 missing values that have to be replaced by none.

```r
library(ggplot2)
```

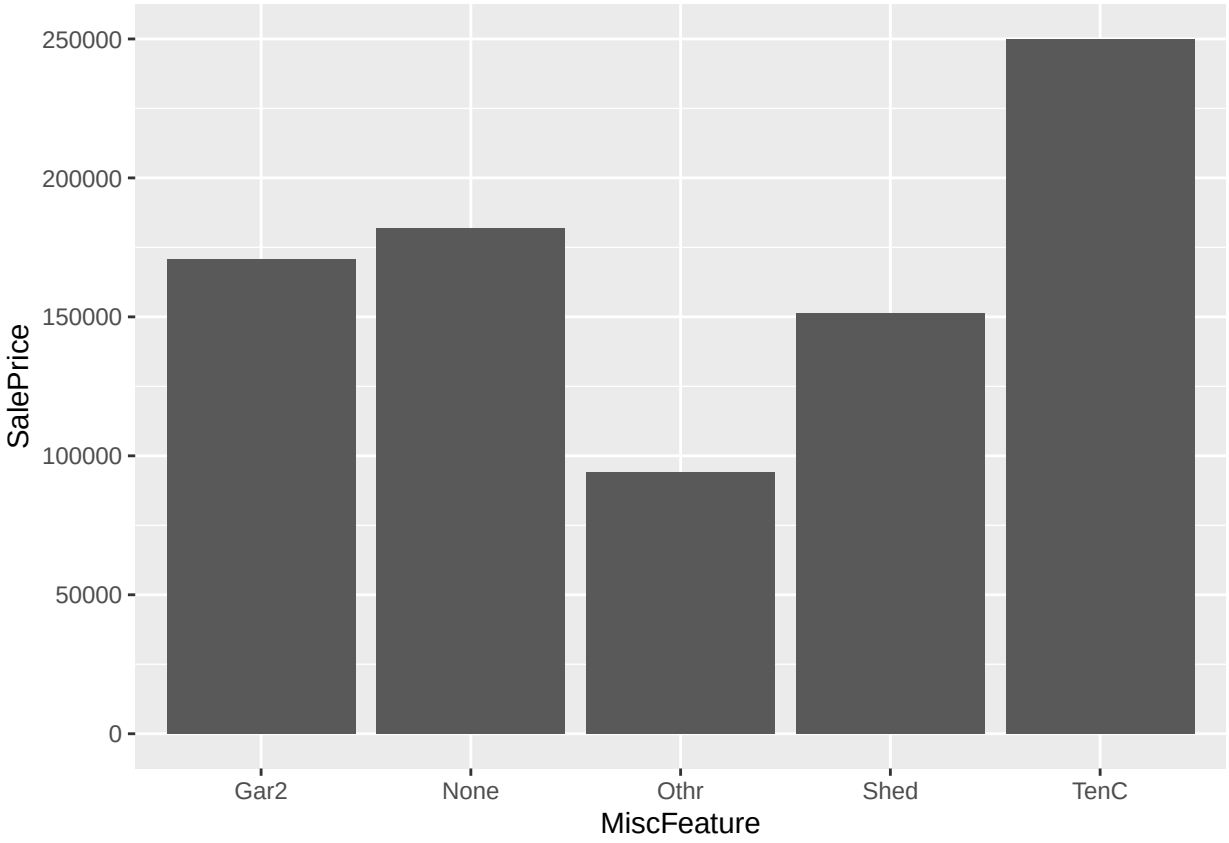
```
## Warning: package 'ggplot2' was built under R version 4.0.3
```

```r
# 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')
```

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



Alley Predictor

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

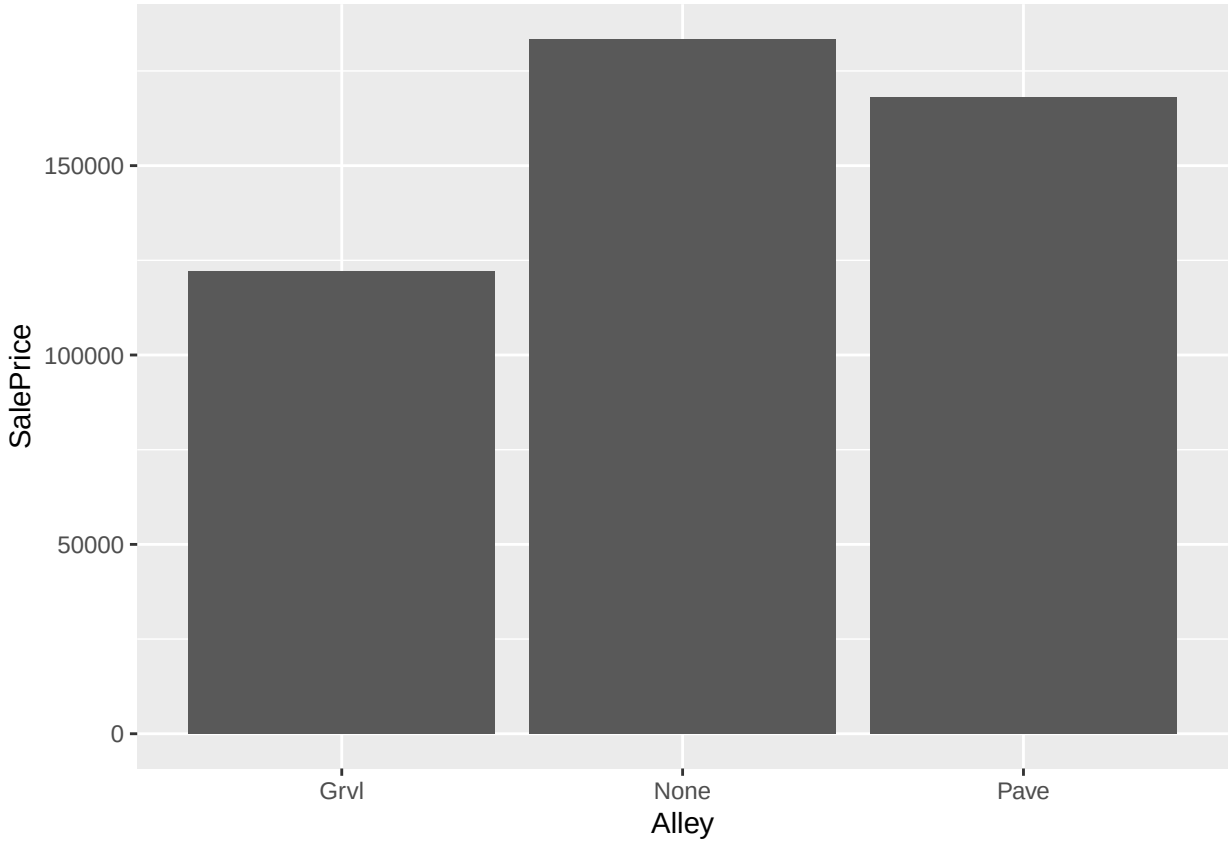
```
##
## Grvl Pave <NA>
##   50   41 1369
```

```
# 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')
```

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



Fence predictor

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

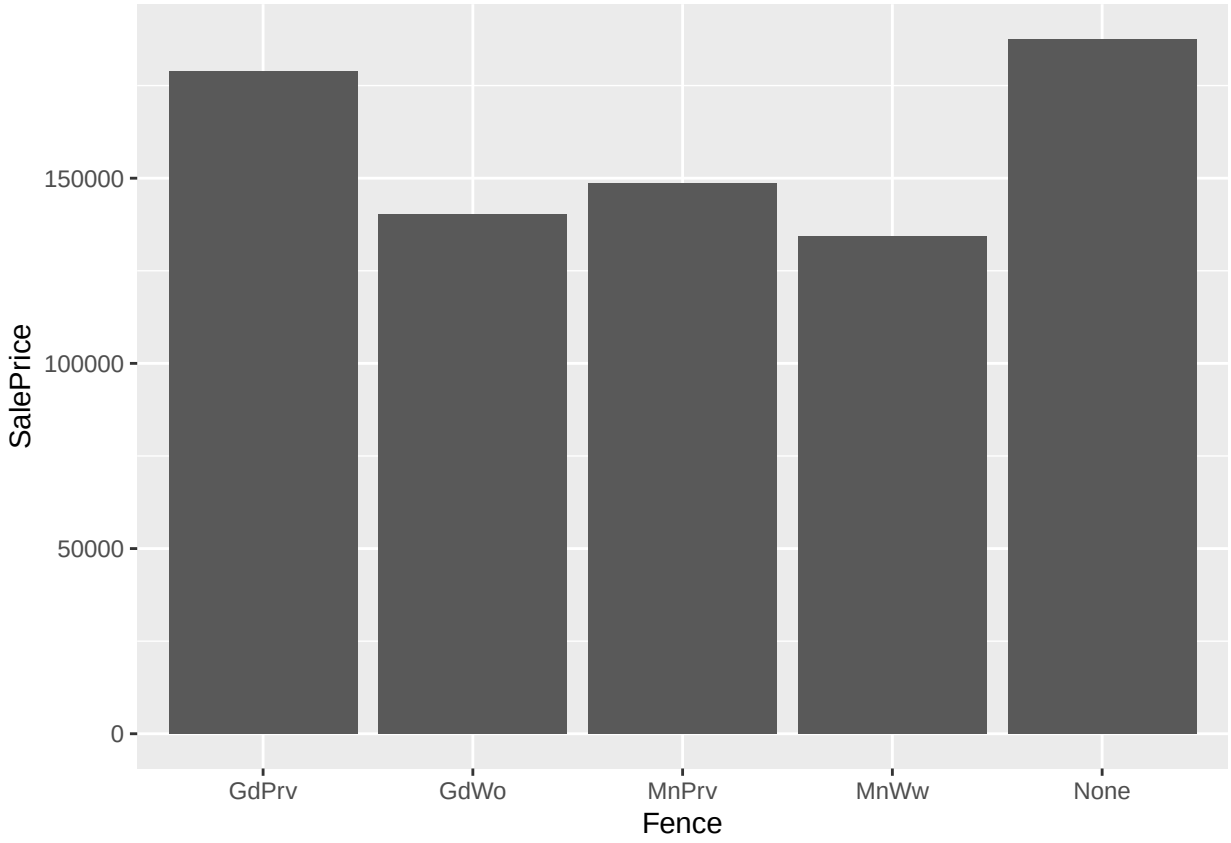
```
##
## GdPrv  GdWo MnPrv  MnWw  <NA>
##    59    54   157    11  1179
```

```
# 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')
```

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



Fireplace variables

Fireplace quality

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

```
##
##   Ex   Fa   Gd   Po   TA <NA>
##   24   33  380   20  313  690
```

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

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

# Changing and converting to factor levels from character.


combined$FireplaceQu <- factor(combined$FireplaceQu, order = TRUE, levels = encoding_levels)

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

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

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

```
##  Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 1 4 4 5 4 1 5 4 4 4 ...
```

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

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

Lot variables

LotFrontage LotShape LotConfig LotArea

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

```
##
## FALSE  TRUE
##  1201   259
```

Here we have 259 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, useNA = "ifany")
```

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

```
combined$LotShape <- factor(combined$LotShape, order = TRUE, levels = c("IR3" , "IR2" , "IR1" , "Reg" ))
table(combined$LotConfig, useNA = "ifany")
```

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

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

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

```
##    GarageType   GarageYrBlt  GarageFinish    GarageQual    GarageCond
##            81            81            81            81            81
##  BsmtExposure  BsmtFinType2      BsmtQual      BsmtCond  BsmtFinType1
##            38            38            37            37            37
##    MasVnrType    MasVnrArea    Electrical            Id    MSSubClass
```

```
##             8             8             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     ExterQual     ExterCond
##             0             0             0             0             0
##    Foundation    BsmtFinSF1    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF
##             0             0             0             0             0
##       Heating     HeatingQC    CentralAir     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageCars    GarageArea
##             0             0             0             0             0
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             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   GarageType GarageFinish   GarageQual   GarageCond   GarageCars
##           81           81           81           81           81            0
##   GarageArea
##            0
```

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

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
##       6     870      19      88       9     387      81
```

```
# convert into ordinal
Finish <- c('None', 'Unf', 'RFn', 'Fin')


combined$GarageFinish<-factor(combined$GarageFinish, order = TRUE, levels = Finish)
table(combined$GarageFinish, useNA = 'ifany')
```

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

```
combined$GarageCond<-factor(combined$GarageCond, order = TRUE, levels = encoding_levels)
table(combined$GarageCond, useNA = "ifany")
```

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

```
combined$GarageQual<-factor(combined$GarageQual, order = TRUE, levels = encoding_levels)
table(combined$GarageQual, useNA = "ifany")
```

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

```
## BsmtExposure BsmtFinType2     BsmtQual     BsmtCond BsmtFinType1 BsmtFullBath
##           38           38           37           37           37            0
## BsmtHalfBath   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF
##            0            0            0            0            0
```

```
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
##     BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 333            0       1124        479      1603        3206
## 949            0          0          0       936         936
```

```
# impute mode
combined[c(949),"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
##     BsmtHalfBath BsmtFinSF1 BsmtFinSF2 BsmtUnfSF TotalBsmtSF
## 333            0       1124        479      1603        3206
## 949            0          0          0       936         936
```

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

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

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

```
##     BsmtQual     BsmtCond BsmtExposure BsmtFinType1 BsmtFinType2 BsmtFullBath
##           37           37           37           37           37            0
## BsmtHalfBath   BsmtFinSF1   BsmtFinSF2    BsmtUnfSF  TotalBsmtSF
##            0            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$BsmtQual<-factor(combined$BsmtQual, order = TRUE, levels = encoding_levels)
table(combined$BsmtQual, useNA = "ifany")
```

```
##
## None   Po   Fa   TA   Gd   Ex
##   37    0   35  649  618  121
```

```
combined$BsmtCond<-factor(combined$BsmtCond, order = TRUE, levels = encoding_levels)
table(combined$BsmtCond, useNA = "ifany")
```

```
##
## None   Po   Fa   TA   Gd   Ex
##   37    2   45 1311   65    0
```

```
exposure <- c('None','No','Mn','Av','Gd')
combined$BsmtExposure<-factor(combined$BsmtExposure, order = TRUE, levels = exposure)
table(combined$BsmtExposure, useNA = "ifany")
```

```
##
## None   No   Mn   Av   Gd
##   37  954  114  221  134
```

```
rating <- c('None','Unf','LwQ','Rec','BLQ','ALQ','GLQ')
combined$BsmtFinType1<-factor(combined$BsmtFinType1, order = TRUE, levels = rating)
table(combined$BsmtFinType1, useNA = "ifany")
```

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

```
combined$BsmtFinType2<-factor(combined$BsmtFinType2, order = TRUE, levels = rating)
table(combined$BsmtFinType2, useNA = "ifany")
```

```
##
## None  Unf  LwQ  Rec  BLQ  ALQ  GLQ
##   37 1257   46   54   33   19   14
```

masonry variables

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

```
##    MasVnrType    MasVnrArea    Electrical            Id    MSSubClass
##             8             8             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     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     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             0             0             0             0             0
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##             0             0             0             0             0
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
```

```
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             0
```

```
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
##      15     445     872     128
```

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)
```

```
##    Electrical            Id    MSSubClass      MSZoning   LotFrontage
##             1             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     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             0             0             0             0             0
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##             0             0             0             0             0
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             0
```

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


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

```
##
## FuseA FuseF FuseP   Mix SBrkr  <NA>
##    94    27     3     1  1334     1
```

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


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
```

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

str(combined[,categorical_variables])
```

```
## 'data.frame':    1460 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/ 15 levels "20","30","40",..: 6 1 6 7 6 5 1 6 5 15 ...
##  $ 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 ...
```

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

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0             0             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0             0             0             0             0
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##             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    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             0             0             0             0             0
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##             0             0             0             0             0
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             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', 'Sev', 'Maj2', 'Maj1', 'Mod', 'Min2', 'Min1', 'Typ')
combined$Functional <- factor(combined$Functional, order = TRUE, levels = functionality)

combined[is.na(combined$KitchenQual),"KitchenQual"] <- names(sort(-table(combined$KitchenQual)))[1]
combined$KitchenQual <- factor(combined$KitchenQual, order = TRUE, levels = encoding_levels)
```

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

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0             0             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0             0             0             0             0
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
```

```
##             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    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             0             0             0             0             0
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##             0             0             0             0             0
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             0
```

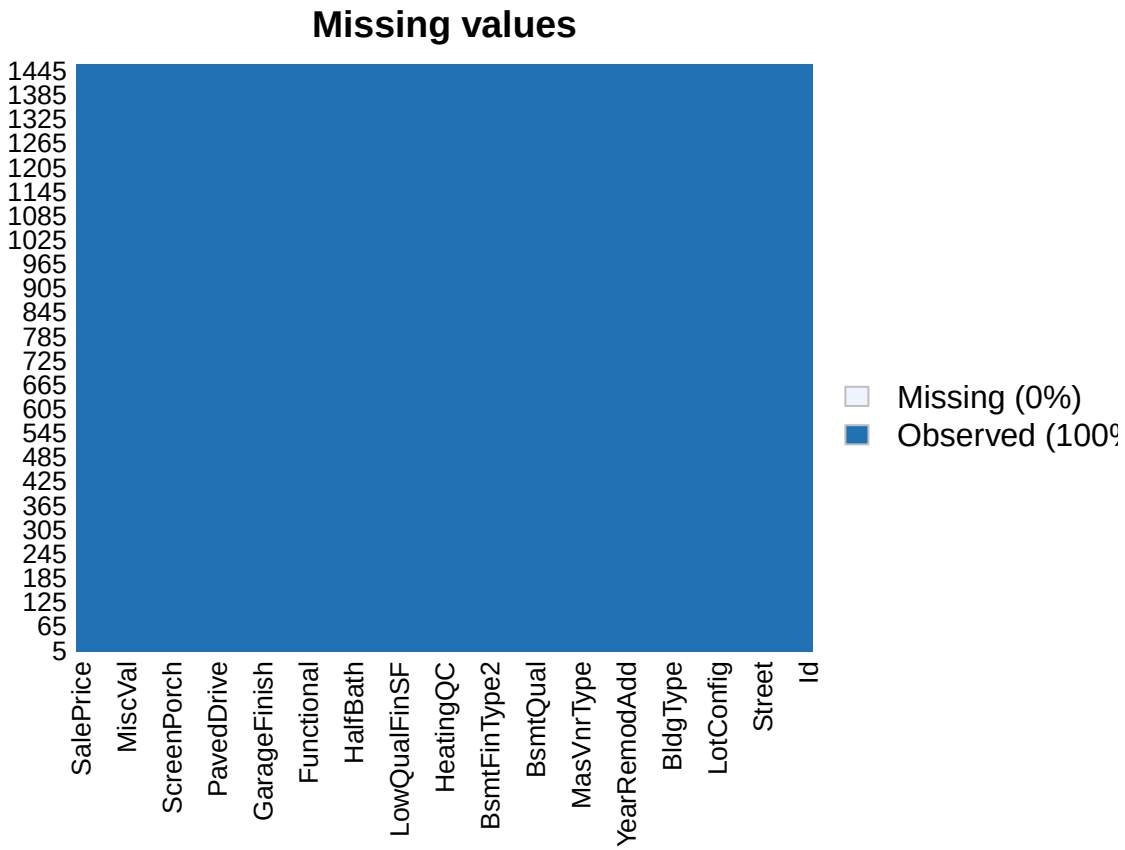
```
char_columns <- names(combined[,sapply(combined, is.character)])
char_columns
```

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

```
# convert remaining character variables into categorical

combined$LandSlope <- factor(combined$LandSlope, order = TRUE, levels = c('Sev','Mod','Gtl'))
combined$ExterQual <- factor(combined$ExterQual, order = TRUE, levels = encoding_levels)
combined$ExterCond <- factor(combined$ExterCond, order = TRUE, levels = encoding_levels)
combined$HeatingQC <- factor(combined$HeatingQC, order = TRUE, levels = encoding_levels)
combined$PavedDrive <- factor(combined$PavedDrive, order = TRUE, levels = c('N','P','Y'))
```

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




```
anyNA(combined)
```

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

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:  35  and categorical variables:  46
```

```
str(combined)
```

```
## 'data.frame':    1460 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : Factor w/ 15 levels "20","30","40",..: 6 1 6 7 6 5 1 6 5 15 ...
##  $ 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 75 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     : Ord.factor w/ 4 levels "IR3"<"IR2"<"IR1"<..: 4 4 3 3 3 3 4 3 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    : Ord.factor w/ 3 levels "Sev"<"Mod"<"Gtl": 3 3 3 3 3 3 3 3 3 3 ...
##  $ 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    : Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 5 4 5 4 5 4 5 4 4 4 ...
##  $ ExterCond    : Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
##  $ BsmtQual     : Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 5 5 5 4 5 5 6 5 4 4 ...
##  $ BsmtCond     : Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 4 4 4 5 4 4 4 4 4 4 ...
##  $ BsmtExposure : Ord.factor w/ 5 levels "None"<"No"<"Mn"<..: 2 5 3 2 4 2 4 3 2 2 ...
##  $ BsmtFinType1 : Ord.factor w/ 7 levels "None"<"Unf"<"LwQ"<..: 7 6 7 6 7 7 7 6 2 7 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : Ord.factor w/ 7 levels "None"<"Unf"<"LwQ"<..: 2 2 2 2 2 2 2 5 2 2 ...
##  $ 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    : Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 6 6 6 5 6 6 6 6 5 6 ...
##  $ 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  : Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 5 4 5 5 5 4 5 4 4 4 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : Ord.factor w/ 8 levels "Sal"<"Sev"<"Maj2"<..: 8 8 8 8 8 8 8 8 7 8 ...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 1 4 4 5 4 1 5 4 4 4 ...
##  $ 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 : Ord.factor w/ 4 levels "None"<"Unf"<"RFn"<..: 3 3 3 2 3 2 3 3 2 3 ...
##  $ 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   : Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 4 4 4 4 4 4 4 4 3 5 ...
##  $ GarageCond   : Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 4 4 4 4 4 4 4 4 4 4 ...
##  $ PavedDrive   : Ord.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       : Ord.factor w/ 6 levels "None"<"Po"<"Fa"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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 ...
```

```
write.csv(combined, "../data/processed/clean_data.csv")
```

EDA