Liz 136 Project Step 2 Data Clean up - NA

Hyunkyung Kim

November 3, 2018

## Loading required package: lattice

## Loading required package: ggplot2

## -- Attaching packages ------------------------------------------------------------------------------------------------- tidyverse 1.2.1 --

## v tibble 1.4.2 v purrr 0.2.4  
## v tidyr 0.8.0 v dplyr 0.7.7  
## v readr 1.1.1 v stringr 1.3.0  
## v tibble 1.4.2 v forcats 0.3.0

## -- Conflicts ---------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::arrange() masks plyr::arrange()  
## x purrr::compact() masks plyr::compact()  
## x dplyr::count() masks plyr::count()  
## x dplyr::failwith() masks plyr::failwith()  
## x dplyr::filter() masks stats::filter()  
## x dplyr::id() masks plyr::id()  
## x dplyr::lag() masks stats::lag()  
## x purrr::lift() masks caret::lift()  
## x dplyr::mutate() masks plyr::mutate()  
## x dplyr::rename() masks plyr::rename()  
## x dplyr::summarise() masks plyr::summarise()  
## x dplyr::summarize() masks plyr::summarize()

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

## Loading required package: Matrix

##   
## Attaching package: 'Matrix'

## The following object is masked from 'package:tidyr':  
##   
## expand

## Loading required package: foreach

##   
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':  
##   
## accumulate, when

## Loaded glmnet 2.0-16

##   
## Attaching package: 'mice'

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

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

### Step 1 was used with test data only, but this step includes cleaning up of test set and train set.

##### Read train and test dataset and combine

H\_train<-read.csv("C:\\Users\\Hyunkyung Kim\\Desktop\\CKME999\\136\\dataset\\all\\train.csv")  
H\_test<-read.csv("C:\\Users\\Hyunkyung Kim\\Desktop\\CKME999\\136\\dataset\\all\\test.csv")  
  
H\_Orig<-rbind.fill(H\_train,H\_test) #rbind.fill does fill with NA values if column is missing. in here Saleprice missing for test data.  
H\_Working<-H\_Orig # Save a copy  
  
tail(H\_Orig)

## Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape  
## 2914 2914 160 RM 21 1526 Pave <NA> Reg  
## 2915 2915 160 RM 21 1936 Pave <NA> Reg  
## 2916 2916 160 RM 21 1894 Pave <NA> Reg  
## 2917 2917 20 RL 160 20000 Pave <NA> Reg  
## 2918 2918 85 RL 62 10441 Pave <NA> Reg  
## 2919 2919 60 RL 74 9627 Pave <NA> Reg  
## LandContour Utilities LotConfig LandSlope Neighborhood Condition1  
## 2914 Lvl AllPub Inside Gtl MeadowV Norm  
## 2915 Lvl AllPub Inside Gtl MeadowV Norm  
## 2916 Lvl AllPub Inside Gtl MeadowV Norm  
## 2917 Lvl AllPub Inside Gtl Mitchel Norm  
## 2918 Lvl AllPub Inside Gtl Mitchel Norm  
## 2919 Lvl AllPub Inside Mod Mitchel Norm  
## Condition2 BldgType HouseStyle OverallQual OverallCond YearBuilt  
## 2914 Norm Twnhs 2Story 4 5 1970  
## 2915 Norm Twnhs 2Story 4 7 1970  
## 2916 Norm TwnhsE 2Story 4 5 1970  
## 2917 Norm 1Fam 1Story 5 7 1960  
## 2918 Norm 1Fam SFoyer 5 5 1992  
## 2919 Norm 1Fam 2Story 7 5 1993  
## YearRemodAdd RoofStyle RoofMatl Exterior1st Exterior2nd MasVnrType  
## 2914 1970 Gable CompShg CemntBd CmentBd None  
## 2915 1970 Gable CompShg CemntBd CmentBd None  
## 2916 1970 Gable CompShg CemntBd CmentBd None  
## 2917 1996 Gable CompShg VinylSd VinylSd None  
## 2918 1992 Gable CompShg HdBoard Wd Shng None  
## 2919 1994 Gable CompShg HdBoard HdBoard BrkFace  
## MasVnrArea ExterQual ExterCond Foundation BsmtQual BsmtCond  
## 2914 0 TA TA CBlock TA TA  
## 2915 0 TA TA CBlock TA TA  
## 2916 0 TA TA CBlock TA TA  
## 2917 0 TA TA CBlock TA TA  
## 2918 0 TA TA PConc Gd TA  
## 2919 94 TA TA PConc Gd TA  
## BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2  
## 2914 No Unf 0 Unf 0  
## 2915 No Unf 0 Unf 0  
## 2916 No Rec 252 Unf 0  
## 2917 No ALQ 1224 Unf 0  
## 2918 Av GLQ 337 Unf 0  
## 2919 Av LwQ 758 Unf 0  
## BsmtUnfSF TotalBsmtSF Heating HeatingQC CentralAir Electrical  
## 2914 546 546 GasA TA Y SBrkr  
## 2915 546 546 GasA Gd Y SBrkr  
## 2916 294 546 GasA TA Y SBrkr  
## 2917 0 1224 GasA Ex Y SBrkr  
## 2918 575 912 GasA TA Y SBrkr  
## 2919 238 996 GasA Ex Y SBrkr  
## X1stFlrSF X2ndFlrSF LowQualFinSF GrLivArea BsmtFullBath BsmtHalfBath  
## 2914 546 546 0 1092 0 0  
## 2915 546 546 0 1092 0 0  
## 2916 546 546 0 1092 0 0  
## 2917 1224 0 0 1224 1 0  
## 2918 970 0 0 970 0 1  
## 2919 996 1004 0 2000 0 0  
## FullBath HalfBath BedroomAbvGr KitchenAbvGr KitchenQual TotRmsAbvGrd  
## 2914 1 1 3 1 TA 5  
## 2915 1 1 3 1 TA 5  
## 2916 1 1 3 1 TA 6  
## 2917 1 0 4 1 TA 7  
## 2918 1 0 3 1 TA 6  
## 2919 2 1 3 1 TA 9  
## Functional Fireplaces FireplaceQu GarageType GarageYrBlt GarageFinish  
## 2914 Typ 0 <NA> <NA> NA <NA>  
## 2915 Typ 0 <NA> <NA> NA <NA>  
## 2916 Typ 0 <NA> CarPort 1970 Unf  
## 2917 Typ 1 TA Detchd 1960 Unf  
## 2918 Typ 0 <NA> <NA> NA <NA>  
## 2919 Typ 1 TA Attchd 1993 Fin  
## GarageCars GarageArea GarageQual GarageCond PavedDrive WoodDeckSF  
## 2914 0 0 <NA> <NA> Y 0  
## 2915 0 0 <NA> <NA> Y 0  
## 2916 1 286 TA TA Y 0  
## 2917 2 576 TA TA Y 474  
## 2918 0 0 <NA> <NA> Y 80  
## 2919 3 650 TA TA Y 190  
## OpenPorchSF EnclosedPorch X3SsnPorch ScreenPorch PoolArea PoolQC  
## 2914 34 0 0 0 0 <NA>  
## 2915 0 0 0 0 0 <NA>  
## 2916 24 0 0 0 0 <NA>  
## 2917 0 0 0 0 0 <NA>  
## 2918 32 0 0 0 0 <NA>  
## 2919 48 0 0 0 0 <NA>  
## Fence MiscFeature MiscVal MoSold YrSold SaleType SaleCondition  
## 2914 GdPrv <NA> 0 6 2006 WD Normal  
## 2915 <NA> <NA> 0 6 2006 WD Normal  
## 2916 <NA> <NA> 0 4 2006 WD Abnorml  
## 2917 <NA> <NA> 0 9 2006 WD Abnorml  
## 2918 MnPrv Shed 700 7 2006 WD Normal  
## 2919 <NA> <NA> 0 11 2006 WD Normal  
## SalePrice  
## 2914 NA  
## 2915 NA  
## 2916 NA  
## 2917 NA  
## 2918 NA  
## 2919 NA

#### Check for duplicates

nrow(H\_Working[,-1])

## [1] 2919

nrow(unique(H\_Working[,-c(1,81)]))

## [1] 2917

* We have 2 pairs of duplicates. Both are exact same except one is in the training, one is in the test set. Will leave as is for now.
* ID 194/2866 and 830/2714 appears to be the same.

### DATA CLEANING & Working with N/As

#### CHeck for N/As

NAs<-colSums(is.na(H\_Working))  
  
# Percentage  
NAs[NAs>0]

## MSZoning LotFrontage Alley Utilities Exterior1st   
## 4 486 2721 2 1   
## Exterior2nd MasVnrType MasVnrArea BsmtQual BsmtCond   
## 1 24 23 81 82   
## BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2   
## 82 79 1 80 1   
## BsmtUnfSF TotalBsmtSF Electrical BsmtFullBath BsmtHalfBath   
## 1 1 1 2 2   
## KitchenQual Functional FireplaceQu GarageType GarageYrBlt   
## 1 2 1420 157 159   
## GarageFinish GarageCars GarageArea GarageQual GarageCond   
## 159 1 1 159 159   
## PoolQC Fence MiscFeature SaleType SalePrice   
## 2909 2348 2814 1 1459

round(NAs[NAs>0]/nrow(H\_Working)\*100,digits=2)

## MSZoning LotFrontage Alley Utilities Exterior1st   
## 0.14 16.65 93.22 0.07 0.03   
## Exterior2nd MasVnrType MasVnrArea BsmtQual BsmtCond   
## 0.03 0.82 0.79 2.77 2.81   
## BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2   
## 2.81 2.71 0.03 2.74 0.03   
## BsmtUnfSF TotalBsmtSF Electrical BsmtFullBath BsmtHalfBath   
## 0.03 0.03 0.03 0.07 0.07   
## KitchenQual Functional FireplaceQu GarageType GarageYrBlt   
## 0.03 0.07 48.65 5.38 5.45   
## GarageFinish GarageCars GarageArea GarageQual GarageCond   
## 5.45 0.03 0.03 5.45 5.45   
## PoolQC Fence MiscFeature SaleType SalePrice   
## 99.66 80.44 96.40 0.03 49.98

#### Below are the items to change from factors to numerics

Col Name type N/A(%) R - Output Description FireplaceQu F 47% Factor w/ 5 levels “Ex”,“Fa”,“Gd”,..: NA 5 5 3 5 NA 3 5 5 5 … Fireplace quality ExterCond F 0% Factor w/ 5 levels “Ex”,“Fa”,“Gd”,..: 5 5 5 5 5 5 5 5 5 5 … Evaluates the present condition of the material on the exterior GarageCond F 6% Factor w/ 5 levels “Ex”,“Fa”,“Gd”,..: 5 5 5 5 5 5 5 5 5 5 … Garage condition GarageQual F 6% Factor w/ 5 levels “Ex”,“Fa”,“Gd”,..: 5 5 5 5 5 5 5 5 2 3 … Garage quality HeatingQC F 0% Factor w/ 5 levels “Ex”,“Fa”,“Gd”,..: 1 1 1 3 1 1 1 1 3 1 … Heating quality and condition ExterQual F 0% Factor w/ 4 levels “Ex”,“Fa”,“Gd”,..: 3 4 3 4 3 4 3 4 4 4 … Evaluates the quality of the material on the exterior KitchenQual F 0% Factor w/ 4 levels “Ex”,“Fa”,“Gd”,..: 3 4 3 3 3 4 3 4 4 4 … Kitchen quality BsmtQual F 3% Factor w/ 4 levels “Ex”,“Fa”,“Gd”,..: 3 3 3 4 3 3 1 3 4 4 … Evaluates the height of the basement PoolQC F 100% Factor w/ 3 levels “Ex”,“Fa”,“Gd”: NA NA NA NA NA NA NA NA NA NA … Pool quality BsmtCond : Factor w/ 4 levels “Fa”,“Gd”,“Po”,..: 4 4 4 2 4 4 4 4 4 4 …

Functional F 0% Factor w/ 7 levels “Maj1”,“Maj2”,..: 7 7 7 7 7 7 7 7 3 7 … BsmtFinType2 F 3% Factor w/ 6 levels “ALQ”,“BLQ”,“GLQ”,..: 6 6 6 6 6 6 6 2 6 6 … BsmtFinType1 F 3% Factor w/ 6 levels “ALQ”,“BLQ”,“GLQ”,..: 3 1 3 1 3 3 3 1 6 3 … Fence F 81% Factor w/ 4 levels “GdPrv”,“GdWo”,..: NA NA NA NA NA 3 NA NA NA NA … BsmtExposure F 3% Factor w/ 4 levels “Av”,“Gd”,“Mn”,..: 4 2 3 4 1 4 1 3 4 4 … PavedDrive F 0% Factor w/ 3 levels “N”,“P”,“Y”: 3 3 3 3 3 3 3 3 3 3 … LandSlope F 0% Factor w/ 3 levels “Gtl”,“Mod”,“Sev”: 1 1 1 1 1 1 1 1 1 1 … GarageFinish F 6% Factor w/ 3 levels “Fin”,“RFn”,“Unf”: 2 2 2 3 2 3 2 2 3 2 …

### For factors including Ex, Fa, Gd, Po, TA ones:

Function created to automate some of ordinals to numerics and check before and after the transformation.

Exorder<-function(x){ # Reorder Ex,Fa,Gd,Po,TA order ones into 1,2,3,4,5 and check before and after. Retiring this since it somehow doesn't work.  
H\_Working[,x]<-as.numeric(recode(H\_Orig[,x],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
print(table(H\_Orig[,x],useNA = 'ifany'))  
print(table(H\_Working[,x],useNA ='ifany'))  
}  
  
  
# BnF - This is to compare before and after transformation. Need quotation before and after.  
BnF<-function(x){  
   
print(table(H\_Orig[,x],useNA = 'ifany'))  
print(table(H\_Working[,x],useNA ='ifany'))  
  
}

FireplaceQu F 47% Factor w/ 5 levels “Ex”,“Fa”,“Gd”,..: NA 5 5 3 5 NA 3 5 5 5 … Fireplace quality ExterCond F 0% Factor w/ 5 levels “Ex”,“Fa”,“Gd”,..: 5 5 5 5 5 5 5 5 5 5 … Evaluates the present condition of the material on the exterior GarageCond F 6% Factor w/ 5 levels “Ex”,“Fa”,“Gd”,..: 5 5 5 5 5 5 5 5 5 5 … Garage condition GarageQual F 6% Factor w/ 5 levels “Ex”,“Fa”,“Gd”,..: 5 5 5 5 5 5 5 5 2 3 … Garage quality HeatingQC F 0% Factor w/ 5 levels “Ex”,“Fa”,“Gd”,..: 1 1 1 3 1 1 1 1 3 1 … Heating quality and condition ExterQual F 0% Factor w/ 4 levels “Ex”,“Fa”,“Gd”,..: 3 4 3 4 3 4 3 4 4 4 … Evaluates the quality of the material on the exterior KitchenQual F 0% Factor w/ 4 levels “Ex”,“Fa”,“Gd”,..: 3 4 3 3 3 4 3 4 4 4 … Kitchen quality BsmtQual F 3% Factor w/ 4 levels “Ex”,“Fa”,“Gd”,..: 3 3 3 4 3 3 1 3 4 4 … Evaluates the height of the basement PoolQC F 100% Factor w/ 3 levels “Ex”,“Fa”,“Gd”: NA NA NA NA NA NA NA NA NA NA … Pool quality

H\_Working[,"FireplaceQu"]<-as.numeric(recode(H\_Orig[,"FireplaceQu"],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
# Imputing 1460NA - matches with 0 fireplaces.  
H\_Working$FireplaceQu[is.na(H\_Orig$FireplaceQu)]<-0  
BnF('FireplaceQu')

##   
## Ex Fa Gd Po TA <NA>   
## 43 74 744 46 592 1420   
##   
## 0 1 2 3 4 5   
## 1420 46 74 592 744 43

# no NA  
H\_Working[,"ExterCond"]<-as.numeric(recode(H\_Orig[,"ExterCond"],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
BnF('ExterCond')

##   
## Ex Fa Gd Po TA   
## 12 67 299 3 2538   
##   
## 1 2 3 4 5   
## 3 67 2538 299 12

# Garage Items will look together  
H\_Working[,"GarageCond"]<-as.numeric(recode(H\_Orig[,"GarageCond"],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
BnF('GarageCond')

##   
## Ex Fa Gd Po TA <NA>   
## 3 74 15 14 2654 159   
##   
## 1 2 3 4 5 <NA>   
## 14 74 2654 15 3 159

H\_Working[,"GarageQual"]<-as.numeric(recode(H\_Orig[,"GarageQual"],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
BnF('GarageQual')

##   
## Ex Fa Gd Po TA <NA>   
## 3 124 24 5 2604 159   
##   
## 1 2 3 4 5 <NA>   
## 5 124 2604 24 3 159

# no NA  
H\_Working[,"HeatingQC"]<-as.numeric(recode(H\_Orig[,"HeatingQC"],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
BnF('HeatingQC')

##   
## Ex Fa Gd Po TA   
## 1493 92 474 3 857   
##   
## 1 2 3 4 5   
## 3 92 857 474 1493

# no NA  
H\_Working[,"ExterQual"]<-as.numeric(recode(H\_Orig[,"ExterQual"],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
BnF('ExterQual')

##   
## Ex Fa Gd TA   
## 107 35 979 1798   
##   
## 2 3 4 5   
## 35 1798 979 107

# Replacing NA with TA (most common item - Kitchen exists for this row)  
H\_Working[,"KitchenQual"]<-as.numeric(recode(H\_Orig[,"KitchenQual"],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
H\_Working$KitchenQual[is.na(H\_Orig$KitchenQual)]<-0  
BnF('KitchenQual')

##   
## Ex Fa Gd TA <NA>   
## 205 70 1151 1492 1   
##   
## 0 2 3 4 5   
## 1 70 1492 1151 205

# 3 rows NA - PoolArea >0 but NA on pool condition. Will impute good=4 (Good and Ex ties).  
# Rest of NAs will be 0  
H\_Working[,"PoolQC"]<-as.numeric(recode(H\_Orig[,"PoolQC"],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
H\_Working$PoolQC[is.na(H\_Orig$PoolQC) & H\_Orig$PoolArea>0]<-4  
H\_Working$PoolQC[is.na(H\_Working$PoolQC)]<-0  
BnF('PoolQC')

##   
## Ex Fa Gd <NA>   
## 4 2 4 2909   
##   
## 0 2 4 5   
## 2906 2 7 4

# Will work with Bmst Nas together  
H\_Working[,"BsmtCond"]<-as.numeric(recode(H\_Orig[,"BsmtCond"],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
BnF('BsmtCond')

##   
## Fa Gd Po TA <NA>   
## 104 122 5 2606 82   
##   
## 1 2 3 4 <NA>   
## 5 104 2606 122 82

H\_Working[,"BsmtQual"]<-as.numeric(recode(H\_Orig[,"BsmtQual"],Ex=5,Fa=2,Gd=4,Po=1,TA=3))  
BnF('BsmtQual')

##   
## Ex Fa Gd TA <NA>   
## 258 88 1209 1283 81   
##   
## 2 3 4 5 <NA>   
## 88 1283 1209 258 81

#Exorder('FireplaceQu')  
#Exorder('ExterCond')  
#Exorder('GarageCond')  
#Exorder('GarageQual')  
#Exorder('HeatingQC')  
#Exorder('ExterQual')  
#Exorder('KitchenQual')  
#Exorder('BsmtQual')  
#Exorder('PoolQC')  
#Exorder('BsmtCond')

They look good.

* Functional F 0% Factor w/ 7 levels “Maj1”,“Maj2”,..: 7 7 7 7 7 7 7 7 3 7 … Also Impute 2 missing value with most common value (over 90%)

7 Typ Typical Functionality 6 Min1 Minor Deductions 1 5 Min2 Minor Deductions 2 4 Mod Moderate Deductions 3 Maj1 Major Deductions 1 2 Maj2 Major Deductions 2 1 Sev Severely Damaged 0 Sal Salvage only

levels(H\_Orig$Functional)

## [1] "Maj1" "Maj2" "Min1" "Min2" "Mod" "Sev" "Typ"

So order should be 3, 2, 6, 5, 4, 1, 7

H\_Working$Functional<-c(3,2,6,5,4,1,7)[as.numeric(H\_Orig$Functional)]  
  
# Majority are Typ so will impute to that for 2 NAs  
H\_Working$Functional[is.na(H\_Orig$Functional)]<-7  
  
table(H\_Orig$Functional,useNA = 'ifany')

##   
## Maj1 Maj2 Min1 Min2 Mod Sev Typ <NA>   
## 19 9 65 70 35 2 2717 2

table(H\_Working$Functional,useNA = 'ifany')

##   
## 1 2 3 4 5 6 7   
## 2 9 19 35 70 65 2719

* Other Basment Related ordinal variables BsmtFinType1: BsmtFinType1: Rating of basement finished area BsmtFinType2: Rating of basement finished area (if multiple types)

6 GLQ Good Living Quarters 5 ALQ Average Living Quarters 4 BLQ Below Average Living Quarters 3 Rec Average Rec Room 2 LwQ Low Quality 1 Unf Unfinshed 0 NA No Basement

levels(H\_Orig$BsmtFinType2)

## [1] "ALQ" "BLQ" "GLQ" "LwQ" "Rec" "Unf"

levels(H\_Orig$BsmtFinType1)

## [1] "ALQ" "BLQ" "GLQ" "LwQ" "Rec" "Unf"

Order should be 5,4,6,2,3,1

H\_Working$BsmtFinType1<-c(5,4,6,2,3,1)[as.numeric(H\_Orig$BsmtFinType1)]  
  
H\_Working$BsmtFinType2<-c(5,4,6,2,3,1)[as.numeric(H\_Orig$BsmtFinType2)]  
  
BnF('BsmtFinType1')

##   
## ALQ BLQ GLQ LwQ Rec Unf <NA>   
## 429 269 849 154 288 851 79   
##   
## 1 2 3 4 5 6 <NA>   
## 851 154 288 269 429 849 79

#table(H\_Orig$BsmtFinType1,useNA ='ifany')  
#table(H\_Working$BsmtFinType1,useNA ='ifany')  
BnF('BsmtFinType2')

##   
## ALQ BLQ GLQ LwQ Rec Unf <NA>   
## 52 68 34 87 105 2493 80   
##   
## 1 2 3 4 5 6 <NA>   
## 2493 87 105 68 52 34 80

#table(H\_Orig$BsmtFinType2,useNA ='ifany')  
#table(H\_Working$BsmtFinType2, useNA = 'ifany')

* BsmtExposure: Refers to walkout or garden level walls

4 Gd Good Exposure 3 Av Average Exposure (split levels or foyers typically score average or above) 2 Mn Mimimum Exposure 1 No No Exposure 0 NA No Basement

levels(H\_Orig$BsmtExposure)

## [1] "Av" "Gd" "Mn" "No"

Again, looking at data itself, I can see that one row is a mistake in putting in NA instead of No.

To fix this I’m going to use two conditions. All BsmtFinType1 Unf and BsmtExposure NA to 1 (No) then I will move rest to 0 (NA). 3 houses associated.

H\_Working$BsmtExposure<-c(3,4,2,1)[as.numeric(H\_Orig$BsmtExposure)]  
H\_Working$BsmtExposure[is.na(H\_Orig$BsmtExposure) & H\_Orig$BsmtFinType1=='Unf']<-1  
H\_Working$BsmtExposure[is.na(H\_Working$BsmtExposure)]<-0  
  
table(H\_Orig$BsmtExposure,useNA='ifany')

##   
## Av Gd Mn No <NA>   
## 418 276 239 1904 82

table(H\_Working$BsmtExposure,useNA='ifany')

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

* Look at other values in Bsmt ALL related items are transformed into numerical values for Bsmt. Now impute rest of the missing values regarding basement. It looks like 79 itmes are related to actually not having basements. Rest are mistakes.

BsmtQual 81 - 79NA to 0 , 2 use most common item BsmtCond 82 - 79NA to 0 , 2 use most common item + BsmtExposure 82 - 79NA to 0, 3 to No=0 (no exposure - applied above ) + BsmtFinType1 79 - 79NA to 0 + BsmtFinSF1 1 - to 0 (typo - no basement) + BsmtFinType2 80 -79NA one to most common item + BsmtFinSF2 1 - to 0 (typo - no basement) + BsmtUnfSF 1 - to 0 (typo - no basement) + TotalBsmtSF 1 - to 0 (typo - no basement) BsmtFullBath 2 BsmtHalfBath 2

Among this,

BsmtQual BsmtCond BsmtExposure BsmtFinType1 BsmtFinSF1 BsmtFinType2 BsmtFinSF2 BsmtUnfSF TotalBsmtSF NA NA NA NA NA NA NA NA NA

This row is responsible for - they should all be 0

BsmtFinSF1 1  
BsmtFinSF2 1 BsmtUnfSF 1 TotalBsmtSF 1

H\_Working$BsmtFinSF1[is.na(H\_Orig$BsmtFinSF1)]<-0  
H\_Working$BsmtFinSF2[is.na(H\_Orig$BsmtFinSF2)]<-0  
H\_Working$BsmtUnfSF[is.na(H\_Orig$BsmtUnfSF)]<-0  
H\_Working$TotalBsmtSF[is.na(H\_Orig$TotalBsmtSF)]<-0  
  
# Check Before and After for Each - coutn of 0 should increase by one.  
sum((H\_Orig$BsmtFinSF1==0),na.rm=T)

## [1] 929

sum((H\_Working$BsmtFinSF1==0))

## [1] 930

sum((H\_Orig$BsmtFinSF2==0),na.rm=T)

## [1] 2571

sum((H\_Working$BsmtFinSF2==0))

## [1] 2572

sum((H\_Orig$BsmtUnfSF==0),na.rm=T)

## [1] 241

sum((H\_Working$BsmtUnfSF==0))

## [1] 242

sum((H\_Orig$TotalBsmtSF==0),na.rm=T)

## [1] 78

sum((H\_Working$TotalBsmtSF==0))

## [1] 79

# Here, this NA for H\_Working$BsmtFinType2 is a typo for 1 row. Looking at the data itself, 479Sf of Basement 2 exists.

# For this row, I will impute the value to the most frequent item when the Basement 2 exist which is Rec from the table.

ALQ BLQ GLQ LwQ Rec Unf 19 33 14 46 54 1256 38

This row had unique 479 square foot for BsmtFinSF2, so I will use this condition to impute that first. So that row will have 3(Rec) for its value. Rest will have value 0 for having no basement

H\_Working$BsmtFinType2[H\_Working$BsmtFinSF2==479]<-3  
H\_Working$BsmtFinType2[is.na(H\_Working$BsmtFinType2)]<-0  
  
# Also fill in 79 NA for Type1  
H\_Working$BsmtFinType1[is.na(H\_Working$BsmtFinType1)]<-0  
  
table(H\_Orig$BsmtFinType2,useNA ='ifany')

##   
## ALQ BLQ GLQ LwQ Rec Unf <NA>   
## 52 68 34 87 105 2493 80

table(H\_Working$BsmtFinType2, useNA = 'ifany')

##   
## 0 1 2 3 4 5 6   
## 79 2493 87 106 68 52 34

table(H\_Orig$BsmtFinType1,useNA ='ifany')

##   
## ALQ BLQ GLQ LwQ Rec Unf <NA>   
## 429 269 849 154 288 851 79

table(H\_Working$BsmtFinType1,useNA ='ifany')

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

Can see one increased from 105(Rec) to 106(3) and NA decreased from 80(NA) to 79(0) for BsmtFinType2, and BsmtFinType1 NA replaced by 0.

* BsmtQual 81 - 79NA to 0 , 2 use most common item - TA (3)
* BsmtCond 82 - 79NA to 0 , 3 use most common item - TA (3)

H\_Working$BsmtQual[is.na(H\_Orig$BsmtQual) & !is.na(H\_Orig$BsmtCond)]<-3  
H\_Working$BsmtQual[is.na(H\_Working$BsmtQual)]<-0  
BnF('BsmtQual')

##   
## Ex Fa Gd TA <NA>   
## 258 88 1209 1283 81   
##   
## 0 2 3 4 5   
## 79 88 1285 1209 258

H\_Working$BsmtCond[is.na(H\_Orig$BsmtCond) & !is.na(H\_Orig$BsmtQual)]<-3  
H\_Working$BsmtCond[is.na(H\_Working$BsmtCond)]<-0  
BnF('BsmtCond')

##   
## Fa Gd Po TA <NA>   
## 104 122 5 2606 82   
##   
## 0 1 2 3 4   
## 79 5 104 2609 122

* Basement Bathrooms They are from no basement house data, so will impute 0 for both.

H\_Working$BsmtFullBath[is.na(H\_Working$BsmtFullBath)]<-0  
H\_Working$BsmtHalfBath[is.na(H\_Working$BsmtHalfBath)]<-0  
  
BnF('BsmtFullBath')

##   
## 0 1 2 3 <NA>   
## 1705 1172 38 2 2   
##   
## 0 1 2 3   
## 1707 1172 38 2

BnF('BsmtHalfBath')

##   
## 0 1 2 <NA>   
## 2742 171 4 2   
##   
## 0 1 2   
## 2744 171 4

# PavedDrive: Paved driveway

2 Y Paved 1 P Partial Pavement 0 N Dirt/Gravel

levels(H\_Orig$PavedDrive)

## [1] "N" "P" "Y"

H\_Working$PavedDrive<-c(0,1,2)[as.numeric(H\_Orig$PavedDrive)]  
  
table(H\_Orig$PavedDrive, useNA='ifany')

##   
## N P Y   
## 216 62 2641

table(H\_Working$PavedDrive, useNA='ifany')

##   
## 0 1 2   
## 216 62 2641

# LandSlope: Slope of property

3 Gtl Gentle slope 2 Mod Moderate Slope 1 Sev Severe Slope

levels(H\_Orig$LandSlope)

## [1] "Gtl" "Mod" "Sev"

H\_Working$LandSlope<-c(3,2,1)[as.factor(H\_Orig$LandSlope)]  
table(H\_Orig$LandSlope, useNA='ifany')

##   
## Gtl Mod Sev   
## 2778 125 16

table(H\_Working$LandSlope, useNA = 'ifany')

##   
## 1 2 3   
## 16 125 2778

##### Utilities: Type of utilities available, and impute 2 NA to common value

4 AllPub All public Utilities (E,G,W,& S) 3 NoSewr Electricity, Gas, and Water (Septic Tank) 2 NoSeWa Electricity and Gas Only 1 ELO Electricity only

levels(H\_Orig$Utilities)

## [1] "AllPub" "NoSeWa"

H\_Working$Utilities<-c(4,1)[as.numeric(H\_Orig$Utilities)]  
#impute common value -4  
H\_Working$Utilities[is.na(H\_Orig$Utilities)]<-4  
table(H\_Orig$Utilities,useNA = 'ifany')

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

table(H\_Working$Utilities,useNA ='ifany')

##   
## 1 4   
## 1 2918

Two levels - didn’t really need to be changed to ordinal since the rest didn’t exist

# BldgType: Type of dwelling

5 1Fam Single-family Detached 4 2FmCon Two-family Conversion; originally built as one-family dwelling 3 Duplx Duplex 2 TwnhsE Townhouse End Unit 1 TwnhsI Townhouse Inside Unit

levels(H\_Orig$BldgType)

## [1] "1Fam" "2fmCon" "Duplex" "Twnhs" "TwnhsE"

H\_Working$BldgType<-c(5,4,3,1,2)[H\_Orig$BldgType]  
table(H\_Orig$BldgType, useNA = 'ifany')

##   
## 1Fam 2fmCon Duplex Twnhs TwnhsE   
## 2425 62 109 96 227

table(H\_Working$BldgType,useNA = 'ifany')

##   
## 1 2 3 4 5   
## 96 227 109 62 2425

Not sure if I should combine duplex and 2fmCon

###### GarageFinish: Interior finish of the garage

3 Fin Finished 2 RFn Rough Finished 1 Unf Unfinished 0 NA No Garage

# Will Impute 159 NA into 0 too

H\_Working$GarageFinish<-c(3,2,1)[as.numeric(H\_Orig$GarageFinish)]  
H\_Working$GarageFinish[is.na(H\_Working$GarageFinish)]<-0  
table(H\_Orig$GarageFinish,useNA='ifany')

##   
## Fin RFn Unf <NA>   
## 719 811 1230 159

table(H\_Working$GarageFinish,useNA = 'ifany')

##   
## 0 1 2 3   
## 159 1230 811 719

#### Ordinal changes from factors to numerics are complete. Now do the rest of NA imputation.

Below are N/As because they do not have Garage. Each has mostly 157 to 159 NAs.

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

There are 3 more items that has 2 more NAs than GarageType. Look into this.

Here we have 2 extra N/As for GarageYrBlt/GarageQual/GarageFinish from GarageType

GarageType GarageYrBlt GarageFinish GarageCars GarageArea GarageQual GarageCond ROW1- Detchd NA NA 1 360 NA NA ROW2- Detchd NA NA NA NA NA NA

First one seems to have Garage Area and # of GarageCars so looks valid. Will impute the GarageYrBlt as BuiltYear, GarageFinish/Qual/Con most common ones

table(H\_Orig$GarageFinish, useNA = 'ifany')

##   
## Fin RFn Unf <NA>   
## 719 811 1230 159

table(H\_Orig$GarageQual, useNA = 'ifany')

##   
## Ex Fa Gd Po TA <NA>   
## 3 124 24 5 2604 159

table(H\_Orig$GarageCond, useNA='ifany')

##   
## Ex Fa Gd Po TA <NA>   
## 3 74 15 14 2654 159

Unf/TA/TA are the most common items. Changes : GarageYrBlt->Builtyear, GarageFinish->Unf, GarageQual<-TA, GarageCon<-TA

#H\_Working$GarageYrBlt[is.na(H\_Orig$GarageYrBlt)]<-0  
H\_Orig$YearBuilt[is.na(H\_Orig$GarageYrBlt) & H\_Orig$GarageArea==360]

## [1] 1910 NA

H\_Working$GarageYrBlt[is.na(H\_Orig$GarageYrBlt) & H\_Orig$GarageArea==360]<-min(H\_Orig$YearBuilt[is.na(H\_Orig$GarageYrBlt) & H\_Orig$GarageArea==360], na.rm=T)  
  
#CHECK  
BnF('GarageYrBlt')

##   
## 1895 1896 1900 1906 1908 1910 1914 1915 1916 1917 1918 1919 1920 1921 1922   
## 1 1 6 1 1 10 2 7 6 2 3 1 33 5 8   
## 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935 1936 1937   
## 6 8 15 15 5 7 2 27 4 4 1 4 8 7 6   
## 1938 1939 1940 1941 1942 1943 1945 1946 1947 1948 1949 1950 1951 1952 1953   
## 11 21 25 14 6 1 10 9 5 19 14 51 17 16 23   
## 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968   
## 37 24 41 34 42 36 37 31 35 34 35 34 39 36 48   
## 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983   
## 32 32 24 27 29 35 28 50 66 41 35 32 15 9 11   
## 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998   
## 19 18 12 18 20 19 26 17 27 49 39 35 40 44 58   
## 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2207 <NA>   
## 54 55 41 53 92 99 142 115 115 61 29 5 1 159   
##   
## 1895 1896 1900 1906 1908 1910 1914 1915 1916 1917 1918 1919 1920 1921 1922   
## 1 1 6 1 1 11 2 7 6 2 3 1 33 5 8   
## 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935 1936 1937   
## 6 8 15 15 5 7 2 27 4 4 1 4 8 7 6   
## 1938 1939 1940 1941 1942 1943 1945 1946 1947 1948 1949 1950 1951 1952 1953   
## 11 21 25 14 6 1 10 9 5 19 14 51 17 16 23   
## 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968   
## 37 24 41 34 42 36 37 31 35 34 35 34 39 36 48   
## 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983   
## 32 32 24 27 29 35 28 50 66 41 35 32 15 9 11   
## 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998   
## 19 18 12 18 20 19 26 17 27 49 39 35 40 44 58   
## 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2207 <NA>   
## 54 55 41 53 92 99 142 115 115 61 29 5 1 158

H\_Working$GarageFinish[is.na(H\_Orig$GarageYrBlt) & H\_Orig$GarageArea==360]<-"Unf"  
H\_Working$GarageQual[is.na(H\_Orig$GarageYrBlt) & H\_Orig$GarageArea==360]<-"TA"  
H\_Working$GarageCond[is.na(H\_Orig$GarageYrBlt) & H\_Orig$GarageArea==360]<-"TA"  
  
#CHeck  
BnF('GarageFinish')

##   
## Fin RFn Unf <NA>   
## 719 811 1230 159   
##   
## 0 1 2 3 Unf   
## 158 1230 811 719 1

BnF('GarageQual')

##   
## Ex Fa Gd Po TA <NA>   
## 3 124 24 5 2604 159   
##   
## 1 2 3 4 5 TA <NA>   
## 5 124 2604 24 3 1 158

BnF('GarageCond')

##   
## Ex Fa Gd Po TA <NA>   
## 3 74 15 14 2654 159   
##   
## 1 2 3 4 5 TA <NA>   
## 14 74 2654 15 3 1 158

#H\_Orig$GaragYrBlt  
#H\_Working$GarageYrBlt

Second one has all NAs, so this is probably a typing error of detached instead of NA. Changes : Detchd-> NA, GarageCars ->0, GarageArea->0 (Can change Along with other NAs, GarageQual->0 GarageCond->0 later)

H\_Working$GarageType[is.na(H\_Orig$GarageArea) & H\_Orig$GarageType=='Detchd']<-NA  
H\_Working$GarageCars[is.na(H\_Orig$GarageArea) & H\_Orig$GarageType=='Detchd']<-0  
H\_Working$GarageArea[is.na(H\_Orig$GarageArea) & H\_Orig$GarageType=='Detchd']<-0  
BnF('GarageType')

##   
## 2Types Attchd Basment BuiltIn CarPort Detchd <NA>   
## 23 1723 36 186 15 779 157   
##   
## 2Types Attchd Basment BuiltIn CarPort Detchd <NA>   
## 23 1723 36 186 15 778 158

BnF('GarageCars')

##   
## 0 1 2 3 4 5 <NA>   
## 157 776 1594 374 16 1 1   
##   
## 0 1 2 3 4 5   
## 158 776 1594 374 16 1

# BnF('GarageArea') # GarageArea ==0 increaed in a number. Should check in more easier way.

Looks like worked as expected. This increased # of GarageType NA to 158.

# GarageType: Garage location

2Types More than one type of garage  
 Attchd Attached to home  
 Basment Basement Garage  
 BuiltIn Built-In (Garage part of house - typically has room above garage)  
 CarPort Car Port  
 Detchd Detached from home  
 NA No Garage

Imptue NA to NoGarage

# Garage Type NA change to NoGarage. Using different way to add another factor into the level.  
H\_Working$GarageType<-as.character(H\_Working$GarageType)  
H\_Working$GarageType[is.na(H\_Working$GarageType)]<-"NoGarage"  
  
H\_Working$GarageType<-as.factor(H\_Working$GarageType)  
  
table(H\_Orig$GarageType,useNA = 'ifany')

##   
## 2Types Attchd Basment BuiltIn CarPort Detchd <NA>   
## 23 1723 36 186 15 779 157

table(H\_Working$GarageType,useNA ='ifany')

##   
## 2Types Attchd Basment BuiltIn CarPort Detchd NoGarage   
## 23 1723 36 186 15 778 158

H\_Working$GarageQual[is.na(H\_Working$GarageQual)]<-0  
H\_Working$GarageCond[is.na(H\_Working$GarageCond)]<-0  
BnF('GarageQual')

##   
## Ex Fa Gd Po TA <NA>   
## 3 124 24 5 2604 159   
##   
## 0 1 2 3 4 5 TA   
## 158 5 124 2604 24 3 1

BnF('GarageCond')

##   
## Ex Fa Gd Po TA <NA>   
## 3 74 15 14 2654 159   
##   
## 0 1 2 3 4 5 TA   
## 158 14 74 2654 15 3 1

Looks as intended.

##### GarageYrBlt

This is a ordinal value (year), so I have decided to give the same year as year built. ##### Also the 2207 is impossible value so impute that to also the year built.

allGarageYrBlt)] <- allGarageYrBlt)]

H\_Working$GarageYrBlt[is.na(H\_Working$GarageYrBlt)]<-H\_Orig$YearBuilt[is.na(H\_Working$GarageYrBlt)]  
  
H\_Working$GarageYrBlt[H\_Orig$GarageYrBlt==2207]<-H\_Working$YearBuilt[which(H\_Orig$GarageYrBlt==2207)]  
BnF('GarageYrBlt')

##   
## 1895 1896 1900 1906 1908 1910 1914 1915 1916 1917 1918 1919 1920 1921 1922   
## 1 1 6 1 1 10 2 7 6 2 3 1 33 5 8   
## 1923 1924 1925 1926 1927 1928 1929 1930 1931 1932 1933 1934 1935 1936 1937   
## 6 8 15 15 5 7 2 27 4 4 1 4 8 7 6   
## 1938 1939 1940 1941 1942 1943 1945 1946 1947 1948 1949 1950 1951 1952 1953   
## 11 21 25 14 6 1 10 9 5 19 14 51 17 16 23   
## 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965 1966 1967 1968   
## 37 24 41 34 42 36 37 31 35 34 35 34 39 36 48   
## 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979 1980 1981 1982 1983   
## 32 32 24 27 29 35 28 50 66 41 35 32 15 9 11   
## 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998   
## 19 18 12 18 20 19 26 17 27 49 39 35 40 44 58   
## 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2207 <NA>   
## 54 55 41 53 92 99 142 115 115 61 29 5 1 159   
##   
## 1872 1875 1890 1895 1896 1900 1902 1905 1906 1907 1908 1910 1911 1912 1914   
## 1 1 2 3 1 9 1 1 1 1 1 21 1 3 6   
## 1915 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929   
## 10 8 2 4 3 42 5 13 10 10 18 16 5 7 2   
## 1930 1931 1932 1933 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1945   
## 30 6 4 1 4 10 8 6 12 21 31 16 6 1 13   
## 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960   
## 13 9 19 16 51 18 16 23 38 30 42 34 44 39 38   
## 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975   
## 33 37 34 36 35 39 38 49 32 42 29 29 29 36 31   
## 1976 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990   
## 50 67 42 35 32 15 9 11 19 19 12 20 20 19 27   
## 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005   
## 18 28 49 40 35 40 44 58 54 55 41 54 92 102 145   
## 2006 2007 2008 2009 2010   
## 119 117 61 29 5

Impute some Factors NA - Fence, Alley, MiscFeature into NoFence NoAlley NoMiscFeature

levels(H\_Working$Fence)<-c(levels(H\_Working$Fence),"NoFence")  
H\_Working$Fence[is.na(H\_Orig$Fence)]<-"NoFence"  
BnF("Fence")

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

levels(H\_Working$Alley)<-c(levels(H\_Working$Alley),"NoAlley")  
H\_Working$Alley[is.na(H\_Orig$Alley)]<-"NoAlley"  
BnF("Alley")

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

levels(H\_Working$MiscFeature)<-c(levels(H\_Working$MiscFeature),"NoMiscFeature")  
H\_Working$MiscFeature[is.na(H\_Orig$MiscFeature)]<-"NoMiscFeature"  
BnF("MiscFeature")

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

#### MS zoning - Majority are RL - impute to RL

#table(H\_Orig$MSZoning) - this was to check majority  
H\_Working$MSZoning[is.na(H\_Orig$MSZoning)]<-'RL'  
BnF('MSZoning')

##   
## C (all) FV RH RL RM <NA>   
## 25 139 26 2265 460 4   
##   
## C (all) FV RH RL RM   
## 25 139 26 2269 460

Exterior1st: Exterior covering on house

#table(H\_Orig$Exterior1st) #this was to check majority, VinylSd for both.  
#table(H\_Orig$Exterior2nd)  
  
H\_Working$Exterior1st[is.na(H\_Orig$Exterior1st)]<-'VinylSd'  
H\_Working$Exterior2nd[is.na(H\_Orig$Exterior2nd)]<-'VinylSd'  
BnF('Exterior1st')

##   
## AsbShng AsphShn BrkComm BrkFace CBlock CemntBd HdBoard ImStucc MetalSd   
## 44 2 6 87 2 126 442 1 450   
## Plywood Stone Stucco VinylSd Wd Sdng WdShing <NA>   
## 221 2 43 1025 411 56 1   
##   
## AsbShng AsphShn BrkComm BrkFace CBlock CemntBd HdBoard ImStucc MetalSd   
## 44 2 6 87 2 126 442 1 450   
## Plywood Stone Stucco VinylSd Wd Sdng WdShing   
## 221 2 43 1026 411 56

BnF('Exterior2nd')

##   
## AsbShng AsphShn Brk Cmn BrkFace CBlock CmentBd HdBoard ImStucc MetalSd   
## 38 4 22 47 3 126 406 15 447   
## Other Plywood Stone Stucco VinylSd Wd Sdng Wd Shng <NA>   
## 1 270 6 47 1014 391 81 1   
##   
## AsbShng AsphShn Brk Cmn BrkFace CBlock CmentBd HdBoard ImStucc MetalSd   
## 38 4 22 47 3 126 406 15 447   
## Other Plywood Stone Stucco VinylSd Wd Sdng Wd Shng   
## 1 270 6 47 1015 391 81

VinylSd increased by 1 for both.

##### Left Over NAs

LotFrontage MasVnrType MasVnrArea Electrical SaleType  
486 24 23 1 1

One row has Area but no MasVnrType - will impute that row with majority item- BrkFace. Rest will be Type - none, Area 0

MasVnrType: Masonry veneer type -> to None MasVnrArea: Masonry veneer area in square feet -> to 0  
BrkCmn Brick Common BrkFace Brick Face CBlock Cinder Block None None Stone Stone

table(H\_Orig$MasVnrType)

##   
## BrkCmn BrkFace None Stone   
## 25 879 1742 249

H\_Working$MasVnrType[!is.na(H\_Orig$MasVnrArea) & is.na(H\_Orig$MasVnrType)]<-'BrkFace'  
H\_Working$MasVnrType[is.na(H\_Working$MasVnrType)]<-"None"  
H\_Working$MasVnrArea[is.na(H\_Working$MasVnrArea)]<-0  
  
BnF('MasVnrType')

##   
## BrkCmn BrkFace None Stone <NA>   
## 25 879 1742 249 24   
##   
## BrkCmn BrkFace None Stone   
## 25 880 1765 249

BnF('MasVnrArea')

##   
## 0 1 3 11 14 16 18 20 22 23 24 27 28 30 31   
## 1738 3 1 1 4 11 3 4 2 4 2 1 2 4 1   
## 32 34 36 38 39 40 41 42 44 45 46 47 48 50 51   
## 4 1 2 2 1 8 3 3 7 3 1 1 1 7 3   
## 52 53 54 56 57 58 60 62 63 64 65 66 67 68 69   
## 3 2 4 2 1 2 7 1 1 1 2 2 2 5 1   
## 70 72 74 75 76 80 81 82 84 85 86 87 88 89 90   
## 4 11 4 2 7 9 1 5 7 4 3 1 5 2 6   
## 91 92 94 95 96 97 98 99 100 101 102 104 105 106 108   
## 1 2 4 3 4 1 5 4 5 3 2 4 2 7 11   
## 109 110 112 113 114 115 116 117 118 119 120 121 122 123 124   
## 1 3 6 3 2 3 3 2 1 2 15 1 3 3 1   
## 125 126 127 128 130 132 134 135 136 137 138 140 141 142 143   
## 3 4 1 9 6 8 2 3 5 1 2 7 1 2 6   
## 144 145 146 147 148 149 150 151 153 154 156 157 158 160 161   
## 11 6 2 2 5 4 5 1 3 1 3 3 3 5 3   
## 162 163 164 165 166 167 168 169 170 171 172 174 175 176 177   
## 5 2 7 3 4 1 5 3 8 2 5 7 1 13 1   
## 178 179 180 182 183 184 186 187 188 189 190 192 194 196 197   
## 8 1 12 5 4 3 7 1 3 3 3 4 5 9 1   
## 198 199 200 202 203 204 205 206 207 208 209 210 212 214 215   
## 6 1 13 2 7 2 3 5 1 3 2 9 4 1 3   
## 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230   
## 12 1 3 1 4 1 1 1 1 1 4 2 2 1 2   
## 232 233 234 235 236 237 238 240 242 243 244 245 246 247 248   
## 6 2 2 1 3 1 4 7 4 2 2 2 6 1 4   
## 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264   
## 4 1 7 1 2 1 8 1 2 2 7 2 1 1 3   
## 265 266 268 270 272 274 275 276 278 279 280 281 283 284 285   
## 2 2 5 7 5 1 3 1 2 1 4 2 1 3 3   
## 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300   
## 2 1 6 3 3 1 2 1 2 3 2 1 3 1 7   
## 302 304 305 306 308 309 310 312 315 318 320 322 323 324 327   
## 8 3 3 6 1 2 3 3 1 2 7 1 1 1 1   
## 328 332 333 335 336 337 338 340 342 344 348 350 351 352 353   
## 2 1 1 2 4 1 2 10 2 2 1 3 2 2 1   
## 355 356 359 360 361 362 364 365 366 368 370 371 372 375 376   
## 1 2 2 7 1 2 2 2 2 2 1 1 1 1 1   
## 378 379 380 381 382 383 385 387 388 391 394 396 397 399 400   
## 2 1 2 1 1 2 1 1 1 1 1 1 1 1 1   
## 402 405 406 408 410 412 415 418 420 422 423 424 425 426 428   
## 2 1 1 1 2 1 1 1 7 2 3 2 3 1 1   
## 430 432 434 435 436 438 440 442 443 444 448 450 451 452 456   
## 2 2 1 1 1 1 1 3 1 1 1 4 1 1 7   
## 459 464 466 468 470 472 473 479 480 481 491 492 495 500 501   
## 1 1 3 2 1 3 3 1 4 1 1 2 1 2 1   
## 502 504 506 509 510 513 514 515 518 519 522 525 526 528 530   
## 1 6 2 1 2 5 1 1 1 1 1 2 1 1 1   
## 532 541 549 550 554 562 564 567 568 571 572 573 576 579 584   
## 1 1 1 1 3 1 1 2 2 1 1 1 1 1 1   
## 594 600 603 604 615 616 621 630 632 634 640 647 650 651 652   
## 1 3 1 1 1 1 2 1 2 1 1 1 2 1 1   
## 653 657 660 662 664 668 673 674 680 692 705 710 714 724 726   
## 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1   
## 730 731 734 738 748 754 760 762 766 768 771 772 788 796 816   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   
## 860 870 877 886 894 902 921 922 945 970 975 1031 1047 1050 1095   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   
## 1110 1115 1129 1159 1170 1224 1290 1378 1600 <NA>   
## 1 1 1 1 1 2 1 1 1 23   
##   
## 0 1 3 11 14 16 18 20 22 23 24 27 28 30 31   
## 1761 3 1 1 4 11 3 4 2 4 2 1 2 4 1   
## 32 34 36 38 39 40 41 42 44 45 46 47 48 50 51   
## 4 1 2 2 1 8 3 3 7 3 1 1 1 7 3   
## 52 53 54 56 57 58 60 62 63 64 65 66 67 68 69   
## 3 2 4 2 1 2 7 1 1 1 2 2 2 5 1   
## 70 72 74 75 76 80 81 82 84 85 86 87 88 89 90   
## 4 11 4 2 7 9 1 5 7 4 3 1 5 2 6   
## 91 92 94 95 96 97 98 99 100 101 102 104 105 106 108   
## 1 2 4 3 4 1 5 4 5 3 2 4 2 7 11   
## 109 110 112 113 114 115 116 117 118 119 120 121 122 123 124   
## 1 3 6 3 2 3 3 2 1 2 15 1 3 3 1   
## 125 126 127 128 130 132 134 135 136 137 138 140 141 142 143   
## 3 4 1 9 6 8 2 3 5 1 2 7 1 2 6   
## 144 145 146 147 148 149 150 151 153 154 156 157 158 160 161   
## 11 6 2 2 5 4 5 1 3 1 3 3 3 5 3   
## 162 163 164 165 166 167 168 169 170 171 172 174 175 176 177   
## 5 2 7 3 4 1 5 3 8 2 5 7 1 13 1   
## 178 179 180 182 183 184 186 187 188 189 190 192 194 196 197   
## 8 1 12 5 4 3 7 1 3 3 3 4 5 9 1   
## 198 199 200 202 203 204 205 206 207 208 209 210 212 214 215   
## 6 1 13 2 7 2 3 5 1 3 2 9 4 1 3   
## 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230   
## 12 1 3 1 4 1 1 1 1 1 4 2 2 1 2   
## 232 233 234 235 236 237 238 240 242 243 244 245 246 247 248   
## 6 2 2 1 3 1 4 7 4 2 2 2 6 1 4   
## 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264   
## 4 1 7 1 2 1 8 1 2 2 7 2 1 1 3   
## 265 266 268 270 272 274 275 276 278 279 280 281 283 284 285   
## 2 2 5 7 5 1 3 1 2 1 4 2 1 3 3   
## 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300   
## 2 1 6 3 3 1 2 1 2 3 2 1 3 1 7   
## 302 304 305 306 308 309 310 312 315 318 320 322 323 324 327   
## 8 3 3 6 1 2 3 3 1 2 7 1 1 1 1   
## 328 332 333 335 336 337 338 340 342 344 348 350 351 352 353   
## 2 1 1 2 4 1 2 10 2 2 1 3 2 2 1   
## 355 356 359 360 361 362 364 365 366 368 370 371 372 375 376   
## 1 2 2 7 1 2 2 2 2 2 1 1 1 1 1   
## 378 379 380 381 382 383 385 387 388 391 394 396 397 399 400   
## 2 1 2 1 1 2 1 1 1 1 1 1 1 1 1   
## 402 405 406 408 410 412 415 418 420 422 423 424 425 426 428   
## 2 1 1 1 2 1 1 1 7 2 3 2 3 1 1   
## 430 432 434 435 436 438 440 442 443 444 448 450 451 452 456   
## 2 2 1 1 1 1 1 3 1 1 1 4 1 1 7   
## 459 464 466 468 470 472 473 479 480 481 491 492 495 500 501   
## 1 1 3 2 1 3 3 1 4 1 1 2 1 2 1   
## 502 504 506 509 510 513 514 515 518 519 522 525 526 528 530   
## 1 6 2 1 2 5 1 1 1 1 1 2 1 1 1   
## 532 541 549 550 554 562 564 567 568 571 572 573 576 579 584   
## 1 1 1 1 3 1 1 2 2 1 1 1 1 1 1   
## 594 600 603 604 615 616 621 630 632 634 640 647 650 651 652   
## 1 3 1 1 1 1 2 1 2 1 1 1 2 1 1   
## 653 657 660 662 664 668 673 674 680 692 705 710 714 724 726   
## 1 1 2 1 1 1 1 2 1 1 1 1 1 1 1   
## 730 731 734 738 748 754 760 762 766 768 771 772 788 796 816   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   
## 860 870 877 886 894 902 921 922 945 970 975 1031 1047 1050 1095   
## 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1   
## 1110 1115 1129 1159 1170 1224 1290 1378 1600   
## 1 1 1 1 1 2 1 1 1

# Electrical

This probably is a typo since this is a newly built building and other facilities are there.It has central air, and gas, all utilities so it should have something decent.

I haved decided to look at this for anything that were built after 2000 and impute the most common one from there

SBrkr Standard Circuit Breakers & Romex  
 FuseA Fuse Box over 60 AMP and all Romex wiring (Average)  
 FuseF 60 AMP Fuse Box and mostly Romex wiring (Fair)  
 FuseP 60 AMP Fuse Box and mostly knob & tube wiring (poor)  
 Mix Mixed

table(H\_Orig$Electrical[H\_Orig$YearBuilt>=2000])

##   
## FuseA FuseF FuseP Mix SBrkr   
## 0 0 0 0 782

After 2000, evertyhing was Sbrkr

H\_Working$Electrical[is.na(H\_Working$Electrical)]<-'SBrkr'  
table(H\_Orig$Electrical,useNA='ifany')

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

table(H\_Working$Electrical,useNA='ifany')

##   
## FuseA FuseF FuseP Mix SBrkr   
## 188 50 8 1 2672

##### SaleType

#table(H\_Orig$SaleType)#To find out majority'  
H\_Working$SaleType[is.na(H\_Orig$SaleType)]<-'WD'  
BnF('SaleType')

##   
## COD Con ConLD ConLI ConLw CWD New Oth WD <NA>   
## 87 5 26 9 8 12 239 7 2525 1   
##   
## COD Con ConLD ConLI ConLw CWD New Oth WD   
## 87 5 26 9 8 12 239 7 2526

#### Lot Frontage Imputation

This is done using mice package.

H\_Working\_1<-H\_Working[,-80] # Exclude SalePrice  
H\_Working\_2<- mice(H\_Working\_1, m=1, method='cart', printFlag=FALSE)

## Warning: Number of logged events: 13

Imputed Value

xyplot(H\_Working\_2,LotFrontage~LotArea)

