HarvardX PH125.9xData Science: Capstone Choose Your Own!(House Pricing)

Jose Quesada

27/12/2020

Introduccion

This project consists of determining the value of a house according to its location, property characteristics and payment methods using the data set from kaggle House Prices - Advanced Regression Techniques https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data (kaggle competitions download -c house-prices-advanced-regression-techniques).

Importing Data.

Files:

train.csv - the training set test.csv - the test set data_description.txt - full description of each column. sample_submission.csv - a benchmark submission from a linear regression on year and month of sale, lot square footage, and number of bedrooms

```
read_csv <- function(file){
    path_data <- "data"
    filename <- paste(path_data,file,sep="/")
    csv__ <- read.csv(filename)
    csv__
}

test_set <-read_csv('test.csv')
train_set<- read_csv('train.csv')


#Join datasets, For this project we going to join train and set data for the cleansing and EDA,
#later we going to split again by SalesPrices not null as train set and test set is null.
df<- bind_rows(train_set,test_set)
```

EDA

train_set:

* Dimensions: 1460, 81
* Memory Usage: 0.7 Mb

test_set:

* Dimensions: 1459, 80
* Memory Usage: 0.7 Mb

Comparing amount of columns between each dataset we can see that we have 1 more column in the train set vs the test set. **SalePrice** is the additional column in the train set and our **target value** for this model We going to use the train set to predict **SalePrice** on the test, first we going to make some EDA and data cleaning.

Total categorical columns: 43

Categorical Columns				
MSZoning	Street	Alley	LotShape	LandContour
Utilities	LotConfig	LandSlope	Neighborhood	Condition1
Condition2	BldgType	HouseStyle	RoofStyle	RoofMatl
Exterior1st	Exterior2nd	MasVnrType	ExterQual	ExterCond
Foundation	BsmtQual	BsmtCond	BsmtExposure	BsmtFinType1
BsmtFinType2	Heating	HeatingQC	CentralAir	Electrical
KitchenQual	Functional	FireplaceQu	GarageType	GarageFinish
GarageQual	GarageCond	PavedDrive	PoolQC	Fence
MiscFeature	SaleType	SaleCondition	MSZoning	Street
Alley	LotShape	LandContour	Utilities	LotConfig

Total numeric columns: 38

Numerical Columns			
Id	MSSubClass	LotFrontage	LotArea
OverallQual	OverallCond	YearBuilt	YearRemodAdd
MasVnrArea	BsmtFinSF1	BsmtFinSF2	BsmtUnfSF
TotalBsmtSF	X1stFlrSF	X2ndFlrSF	LowQualFinSF
GrLivArea	BsmtFullBath	BsmtHalfBath	FullBath
HalfBath	BedroomAbvGr	KitchenAbvGr	TotRmsAbvGrd
Fireplaces	GarageYrBlt	GarageCars	GarageArea
WoodDeckSF	OpenPorchSF	EnclosedPorch	X3SsnPorch
ScreenPorch	PoolArea	MiscVal	MoSold
YrSold	SalePrice	Id	MSSubClass

Im going to handle MSSubClass as categorical data although it is shown as numeric column, is actually a categorical data, the numbers is the columns are the type of dwelling involved in the sale, im removing from numeric columns and append as categorical.

Missing Values

For this analysis we going to select just the columns that have missing values, **if they not in plot or table its because they not have missing values**.

Description of columns with Missing Values:

Missing Categorical Columns	
name	prc_na
PoolQC	0.9965742
MiscFeature	0.9640288
Alley	0.9321686
Fence	0.8043851
FireplaceQu	0.4864680
GarageFinish	0.0544707
GarageQual	0.0544707
GarageCond	0.0544707
GarageType	0.0537855
BsmtCond	0.0280918
BsmtExposure	0.0280918
BsmtQual	0.0277492
BsmtFinType2	0.0274066
BsmtFinType1	0.0270641
MasVnrType	0.0082220
MSZoning	0.0013703
Utilities	0.0006852
Functional	0.0006852
Exterior1st	0.0003426
Exterior2nd	0.0003426
Electrical	0.0003426
KitchenQual	0.0003426
SaleType	0.0003426

Missing Numerical Columns	
name	prc_na
SalePrice	0.4998287
LotFrontage	0.1664954
GarageYrBlt	0.0544707
MasVnrArea	0.0078794
BsmtFullBath	0.0006852
BsmtHalfBath	0.0006852
BsmtFinSF1	0.0003426
BsmtFinSF2	0.0003426
BsmtUnfSF	0.0003426
TotalBsmtSF	0.0003426
GarageCars	0.0003426
GarageArea	0.0003426

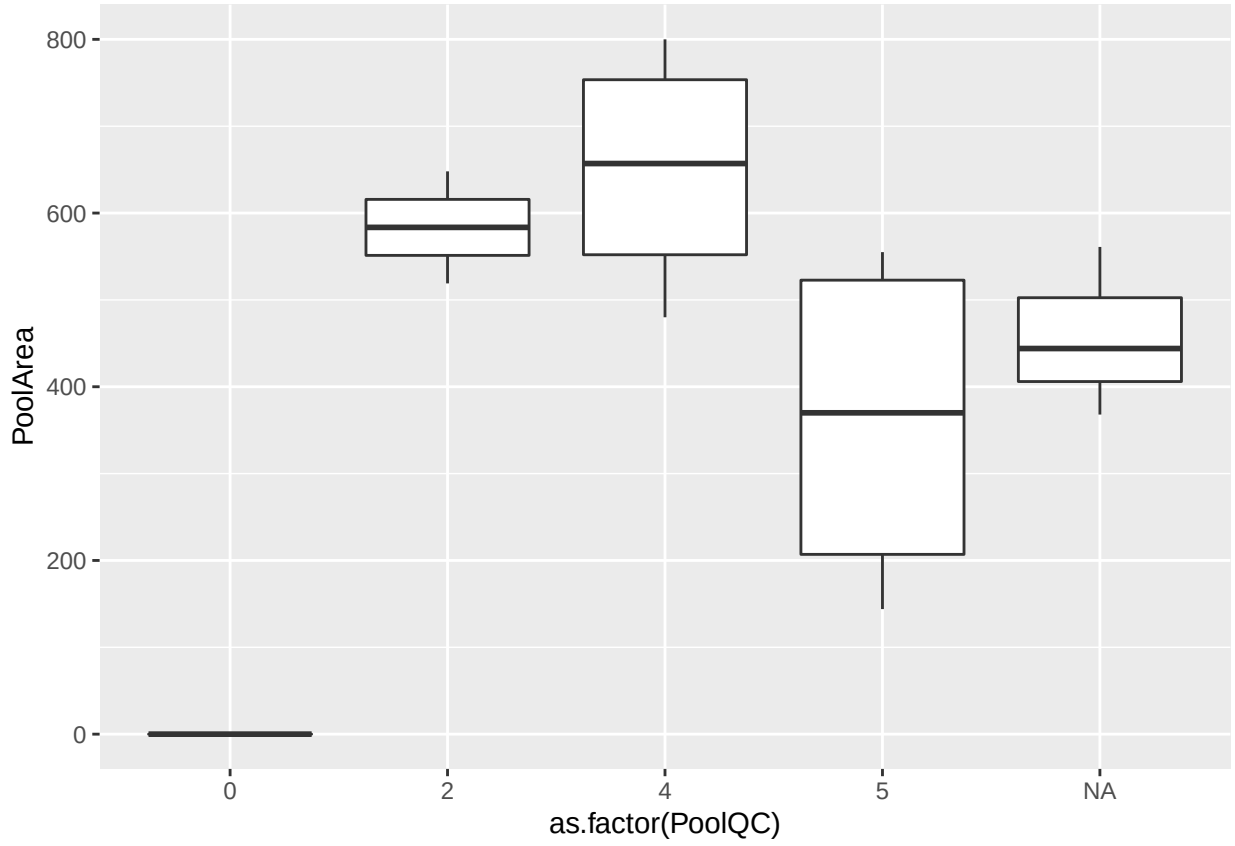
By looking in the data_description file, we can determine that we have columns that show us measurements, condition and qualities of additional features of the houses, They are defined with NA when they do not have one of them. To determine if they are really null values, we must compare multiple columns, example if we have NA PoolQC and PoolArea equal to 0 is the NA is not a Missing value, because the house doesnt have a pool, if PoolQC is NA but the PoolArea is greater than 0, we have a missing value.

Identify associated columns

Related Features		
name_features	dim_features	dtype
MasVnr		
MasVnr	MasVnrArea	numeric
MasVnr	MasVnrType	categorical
Bsmt		
Bsmt	BsmtCond	categorical
Bsmt	BsmtExposure	categorical
Bsmt	BsmtFinSF1	numeric
Bsmt	BsmtFinSF2	numeric
Bsmt	BsmtFinType1	categorical
Bsmt	BsmtFinType2	categorical
Bsmt	BsmtFullBath	numeric
Bsmt	BsmtHalfBath	numeric
Bsmt	BsmtQual	categorical
Bsmt	BsmtUnfSF	numeric
Bsmt	TotalBsmtSF	numeric
Fireplace		
Fireplace	FireplaceQu	categorical
Fireplace	Fireplaces	numeric
Pool		
Pool	PoolArea	numeric
Pool	PoolQC	categorical
Heating		
Heating	Heating	categorical
Heating	HeatingQC	categorical
Misc		
Misc	MiscFeature	categorical
Misc	MiscVal	numeric
Kitchen		
Kitchen	KitchenAbvGr	numeric
Kitchen	KitchenQual	categorical
Exterior		
Exterior	Exterior1st	categorical
Exterior	Exterior2nd	categorical
Garage		
Garage	GarageArea	numeric
Garage	GarageCars	numeric
Garage	GarageCond	categorical
Garage	GarageFinish	categorical
Garage	GarageQual	categorical
Garage	GarageType	categorical
Garage	GarageYrBlt	numeric

Pool

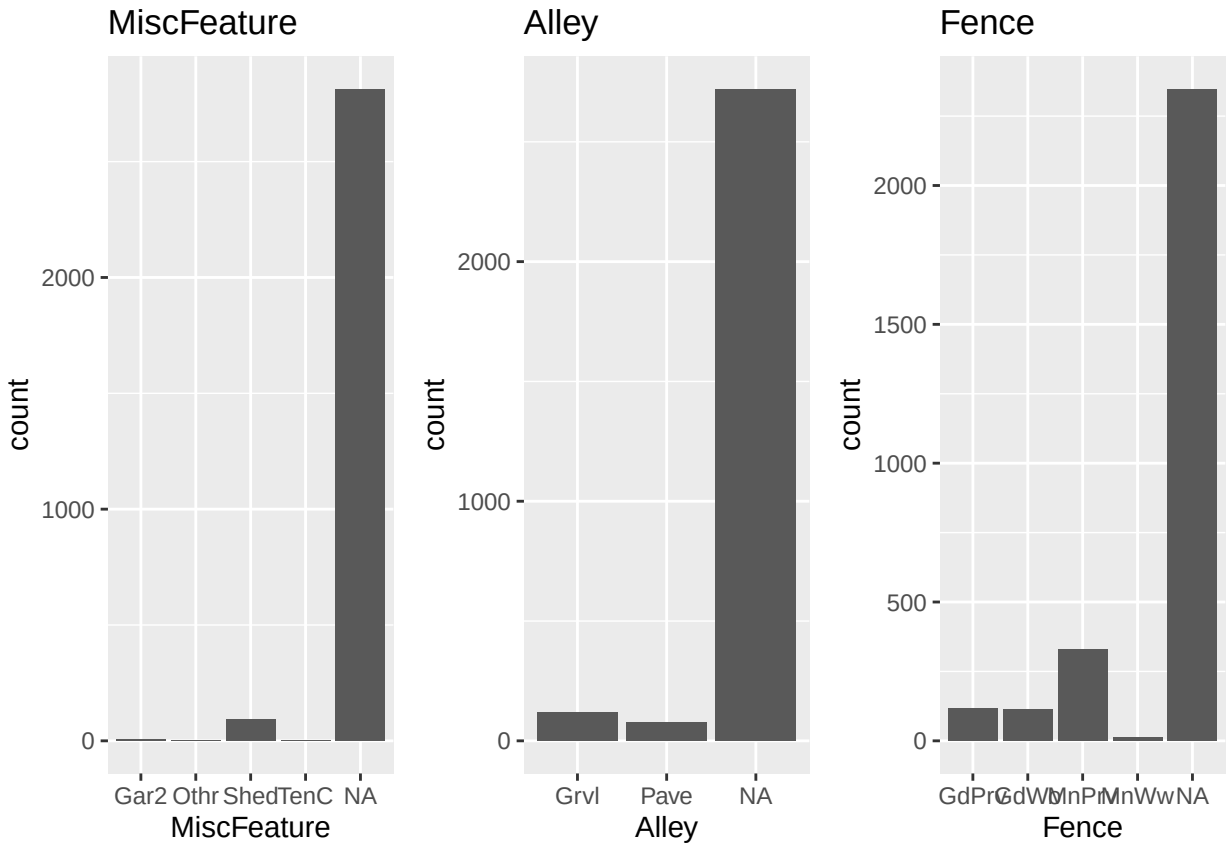
First we going to replace NA in PoolArea to 0, then im replacing PoolQC to No Pool whe Area is equal to 0 and after we going to label encoder the column(It is used to transform non-numerical labels to numerical labels (or nominal categorical variables). Numerical labels are always between 0 and n_classes-1.)



Just for looking into this graph im going to fill NA with 5.

Pool Missing Values		
name	prc_na	type
PoolArea	0	numerical
PoolQC	0	categorical

MiscFeatures, Alley & Fence



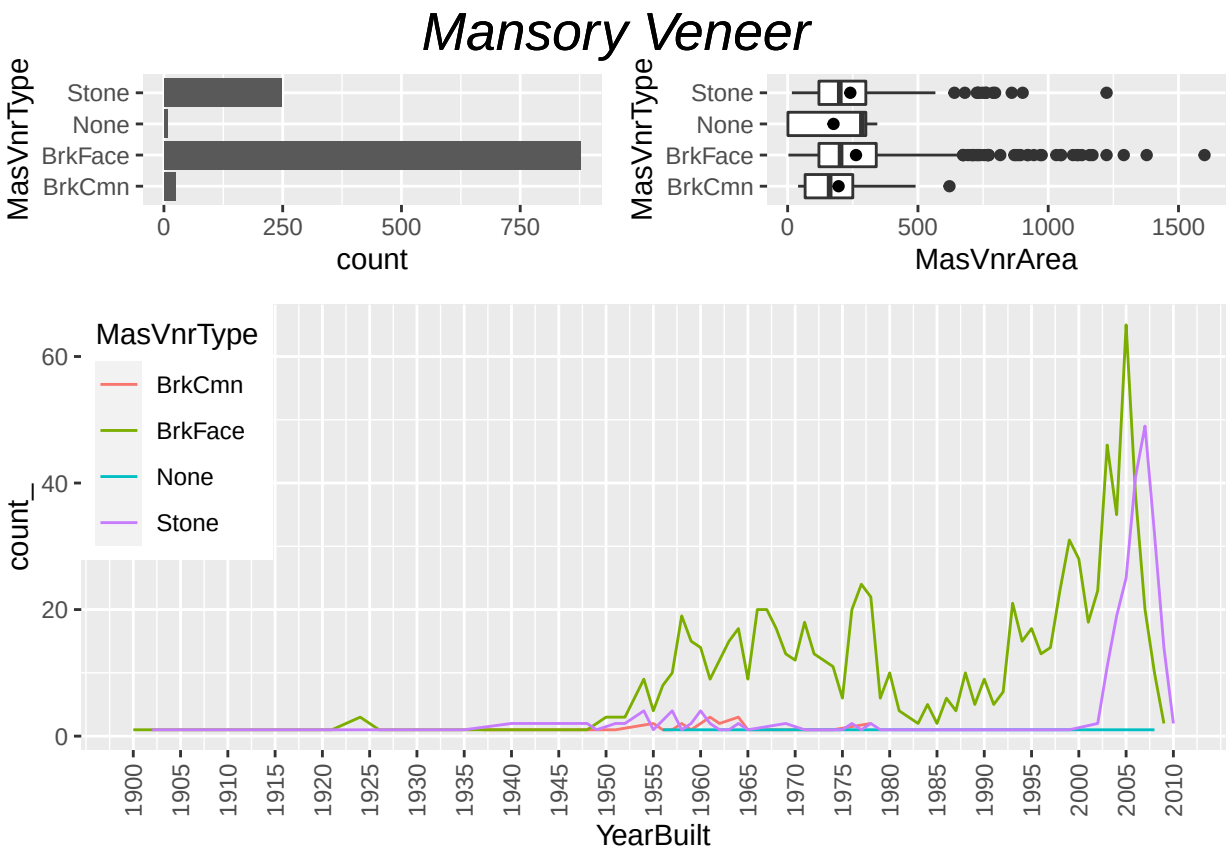
For NA in MiscFeature, its just replace by "None Feature", Alley to "No Alley" and Fence to "No Fence"

MiscF, Alley & Fence Missing Values		
name	prc_na	type
MiscFeature	0	categorical
Alley	0	categorical
Fence	0	categorical

Masonry veneer

Columns:

1. `MasVnrType`.
2. `MasVnrArea`.



This plot show us between 1950 and 2005 BrkFace was the predominant type of MasVnr and after 2005 was Stone, we going to get the mode in every yearn and fill Na values with mode by year and look how many null values we get.

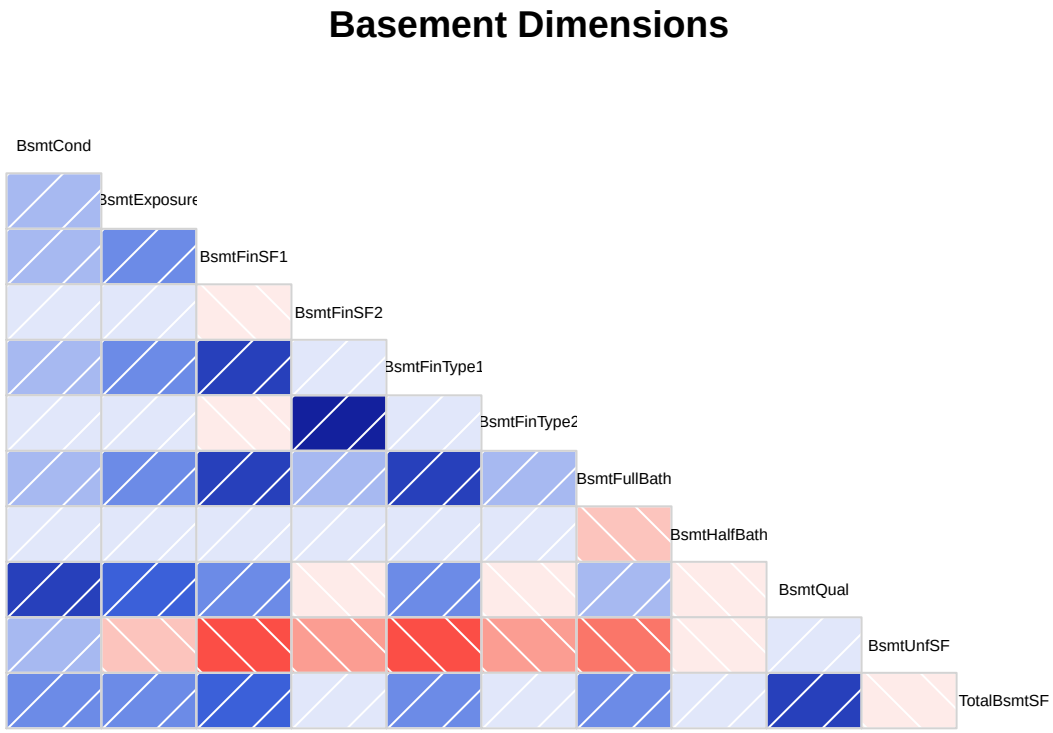
Mansory Veneer		
name	prc_na	type
MasVnrArea	0	numerical
MasVnrType	0	categorical

Basement

First we are going to replace the null values of columns Bsmt fSF, BsmtFinSF1, BsmtFinSF2 and TotalBsmtSF, by 0. After this we are going to replace the null values of the categorical columns by "No Basement" and "No Basement 1" when BsmtFinSF1 is equal to 0 and "No Basement 2" when BsmtFinSF2 is equal to 0. After the first cleaning we are going to convert the BsmtCond columns, BsmtExposure, BsmtFinType1, BsmtFinType2 and BsmtQual, in numerical values, giving as a classification based on the descriptions that are in the file "data/data_description.txt", to be able to find correlations and finish replacing the null values in these columns, we also transform the BsmtUnfSF column into a percentage of the TotalBsmtSF.

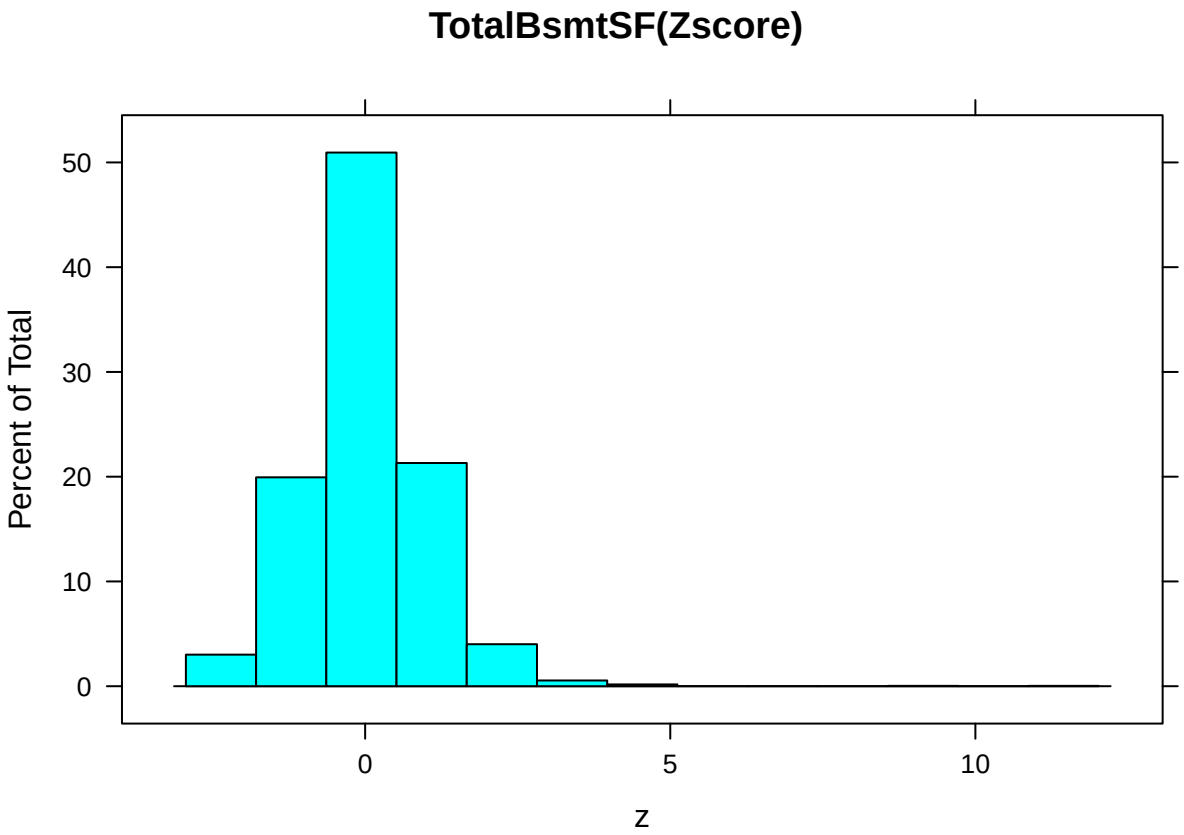
Mansory Veneer		
name	prc_na	type
BsmtCond	0.0010277	categorical
BsmtQual	0.0006852	categorical
BsmtFinType2	0.0003426	categorical
BsmtExposure	0.0000000	categorical
BsmtFinSF1	0.0000000	numerical
BsmtFinSF2	0.0000000	numerical
BsmtFinType1	0.0000000	categorical
BsmtFullBath	0.0000000	numerical
BsmtHalfBath	0.0000000	numerical
BsmtUnfSF	0.0000000	numerical
TotalBsmtSF	0.0000000	numerical

We are going to analyze variables that are highly correlated to replace the null values

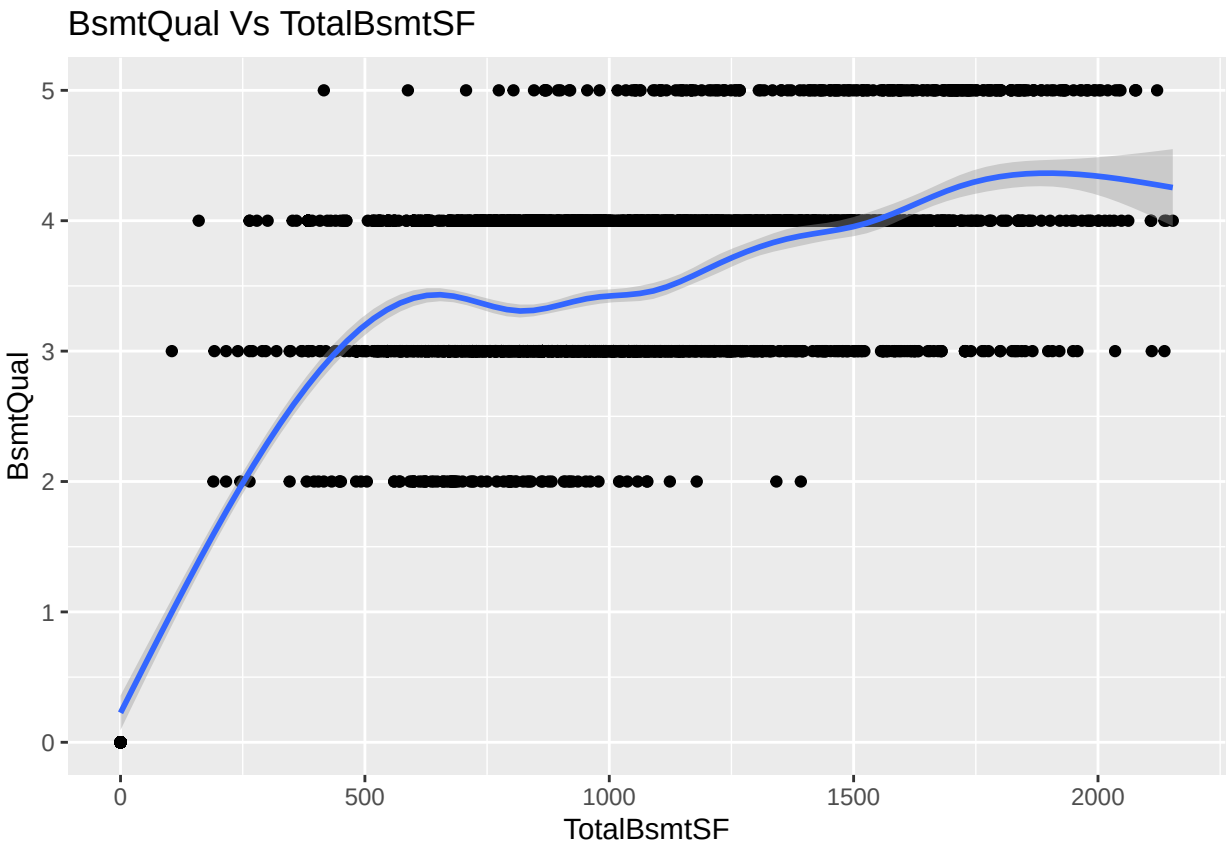


High Correlated Dim			
	Var1	Var2	value
1	BsmtFinType2	BsmtFinSF2	0.8288252
3	BsmtFinType1	BsmtFinSF1	0.7122094
5	BsmtFullBath	BsmtFinSF1	0.6394350
7	BsmtQual	BsmtCond	0.6344277
9	BsmtFullBath	BsmtFinType1	0.5877612
11	TotalBsmtSF	BsmtQual	0.5788890
13	TotalBsmtSF	BsmtFinSF1	0.5361229

To complete the null values for BsmtQual im making alinear models() BsmtQual ~ TotalBsmtSF), first calculate z socre of TotalBsmtSF the remove outliers.



For this model I am going to remove all the TotalBsmtSF that are at least 2.5 z score absolute from the average, this represents 98.8694758 percent of the data.



Linear Regression BsmtQual ~ TotalBsmtSF

```
MSE: 0.8381899
```

Now im replacing null values with BsmtQual predictions and removing the SE_(Error column) and predic(predictions column), then im making the model for replace null values at BsmtCond.

Bsmt Missing Values		
name	prc_na	type
BsmtCond	0	categorical
BsmtExposure	0	categorical
BsmtFinSF1	0	numerical
BsmtFinSF2	0	numerical
BsmtFinType1	0	categorical
BsmtFinType2	0	categorical
BsmtFullBath	0	numerical
BsmtHalfBath	0	numerical
BsmtQual	0	categorical
BsmtUnfSF	0	numerical
TotalBsmtSF	0	numerical

Fireplace

Im using similar approach as Pool NA's, if Fireplaces =0 the FireplaceQu = "No Fireplace"

Fireplace Missing Values		
name	prc_na	type
FireplaceQu	0	categorical
Fireplaces	0	numerical