Kaggel\_Final\_Notebook\_Team Machine

Prabhudatta Mohapatra

Enni Su

Hunter Conrad

Shreya Chawla

2023-04-16

Table of Contents

[Introduction 2](#_Toc132570421)

[Project Goal 2](#_Toc132570422)

[Loading Required Libraries 2](#_Toc132570423)

[Importing Data Files From Kaggle 2](#_Toc132570424)

[Trainging Data Summary, Missing Value Imputation, and Data Transformation 3](#_Toc132570425)

[Data Summary & Missingness Assessment 3](#_Toc132570426)

[Variables Distribution 7](#_Toc132570427)

[Data Transformation & Missing Value Imputation 15](#_Toc132570428)

[Correlation Plot 17](#_Toc132570429)

[Train\_fold and Validation\_fold Data Preparation 18](#_Toc132570430)

[RMSE and R2 Function 19](#_Toc132570431)

[Variable (Feature) Selection For Modeling 19](#_Toc132570432)

[Data Description 19](#_Toc132570433)

[Modeling 20](#_Toc132570434)

[Train\_fold Model 20](#_Toc132570435)

[Train\_fold Prediction 23](#_Toc132570436)

[Validation\_fold Prediction 23](#_Toc132570437)

[Submission Model 24](#_Toc132570438)

[Validation 26](#_Toc132570439)

[Test Data Summary, Missing Value Imputation, and Data Transformation 29](#_Toc132570440)

[Data Summary & Missingness Assessment 29](#_Toc132570441)

[Variables Distribution 32](#_Toc132570442)

[Data Transformation & Missing Value Imputation 37](#_Toc132570443)

[Predicting SalePrice For Test Data 39](#_Toc132570444)

[Kaggle Submission Report 39](#_Toc132570445)

[Contributors & Contributions 39](#_Toc132570446)

# Introduction

The Kaggle House Prices - Advanced Regression Techniques competition provides a platform to perform statistical analyses to understand what features of a house determine the sale price and how to use them to predict sale price accuratley. Kaggle provides sample data set for Ames, Iowa which has train dataset having 81 explanatory variables and 1460 observations and test data set having 80 explanatory variables and 1459 observations, describing (almost) every aspect of residential homes (dimensions, neighborhoods, sale prices etc). Insights gained will be helpful to the individuals in the decision making process trying to purchase a house.

# Project Goal

Our project goal is to applying concepts of Exploratory Data Analysis, visualization, data cleaning, preprocessing, and linear models to predict house prices given the features of the house, and interpret the linear models to find out features that add value to a house price. The data set is multivariate and the most important features will be selected to predict the sale price of each home using linear regression technique.Train and test data sets provided by Kaggle will be used for the project.

# Loading Required Libraries

#loading libraries  
library(tidyverse)  
library(ggplot2)  
library(dplyr)  
library(corrplot)  
library(gridExtra)  
library(grid)  
library(skimr)

# Importing Data Files From Kaggle

#importing data files  
train <- read\_csv("train.csv")  
test <- read\_csv("test.csv")

# Trainging Data Summary, Missing Value Imputation, and Data Transformation

## Data Summary & Missingness Assessment

#summary of training data set  
skim(train)

Data summary

|  |  |
| --- | --- |
| Name | train |
| Number of rows | 1460 |
| Number of columns | 81 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 43 |
| numeric | 38 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

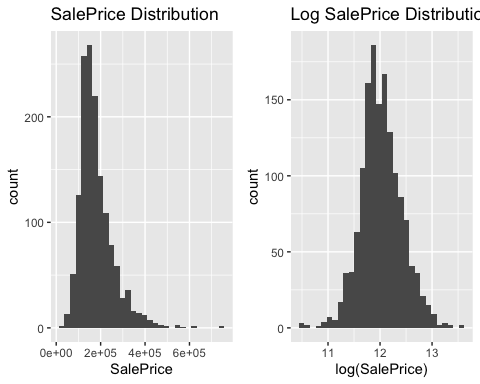
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| MSZoning | 0 | 1.00 | 2 | 7 | 0 | 5 | 0 |
| Street | 0 | 1.00 | 4 | 4 | 0 | 2 | 0 |
| Alley | 1369 | 0.06 | 4 | 4 | 0 | 2 | 0 |
| LotShape | 0 | 1.00 | 3 | 3 | 0 | 4 | 0 |
| LandContour | 0 | 1.00 | 3 | 3 | 0 | 4 | 0 |
| Utilities | 0 | 1.00 | 6 | 6 | 0 | 2 | 0 |
| LotConfig | 0 | 1.00 | 3 | 7 | 0 | 5 | 0 |
| LandSlope | 0 | 1.00 | 3 | 3 | 0 | 3 | 0 |
| Neighborhood | 0 | 1.00 | 5 | 7 | 0 | 25 | 0 |
| Condition1 | 0 | 1.00 | 4 | 6 | 0 | 9 | 0 |
| Condition2 | 0 | 1.00 | 4 | 6 | 0 | 8 | 0 |
| BldgType | 0 | 1.00 | 4 | 6 | 0 | 5 | 0 |
| HouseStyle | 0 | 1.00 | 4 | 6 | 0 | 8 | 0 |
| RoofStyle | 0 | 1.00 | 3 | 7 | 0 | 6 | 0 |
| RoofMatl | 0 | 1.00 | 4 | 7 | 0 | 8 | 0 |
| Exterior1st | 0 | 1.00 | 5 | 7 | 0 | 15 | 0 |
| Exterior2nd | 0 | 1.00 | 5 | 7 | 0 | 16 | 0 |
| MasVnrType | 8 | 0.99 | 4 | 7 | 0 | 4 | 0 |
| ExterQual | 0 | 1.00 | 2 | 2 | 0 | 4 | 0 |
| ExterCond | 0 | 1.00 | 2 | 2 | 0 | 5 | 0 |
| Foundation | 0 | 1.00 | 4 | 6 | 0 | 6 | 0 |
| BsmtQual | 37 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtCond | 37 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtExposure | 38 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtFinType1 | 37 | 0.97 | 3 | 3 | 0 | 6 | 0 |
| BsmtFinType2 | 38 | 0.97 | 3 | 3 | 0 | 6 | 0 |
| Heating | 0 | 1.00 | 4 | 5 | 0 | 6 | 0 |
| HeatingQC | 0 | 1.00 | 2 | 2 | 0 | 5 | 0 |
| CentralAir | 0 | 1.00 | 1 | 1 | 0 | 2 | 0 |
| Electrical | 1 | 1.00 | 3 | 5 | 0 | 5 | 0 |
| KitchenQual | 0 | 1.00 | 2 | 2 | 0 | 4 | 0 |
| Functional | 0 | 1.00 | 3 | 4 | 0 | 7 | 0 |
| FireplaceQu | 690 | 0.53 | 2 | 2 | 0 | 5 | 0 |
| GarageType | 81 | 0.94 | 6 | 7 | 0 | 6 | 0 |
| GarageFinish | 81 | 0.94 | 3 | 3 | 0 | 3 | 0 |
| GarageQual | 81 | 0.94 | 2 | 2 | 0 | 5 | 0 |
| GarageCond | 81 | 0.94 | 2 | 2 | 0 | 5 | 0 |
| PavedDrive | 0 | 1.00 | 1 | 1 | 0 | 3 | 0 |
| PoolQC | 1453 | 0.00 | 2 | 2 | 0 | 3 | 0 |
| Fence | 1179 | 0.19 | 4 | 5 | 0 | 4 | 0 |
| MiscFeature | 1406 | 0.04 | 4 | 4 | 0 | 4 | 0 |
| SaleType | 0 | 1.00 | 2 | 5 | 0 | 9 | 0 |
| SaleCondition | 0 | 1.00 | 6 | 7 | 0 | 6 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1.00 | 730.50 | 421.61 | 1 | 365.75 | 730.5 | 1095.25 | 1460 | ▇▇▇▇▇ |
| MSSubClass | 0 | 1.00 | 56.90 | 42.30 | 20 | 20.00 | 50.0 | 70.00 | 190 | ▇▅▂▁▁ |
| LotFrontage | 259 | 0.82 | 70.05 | 24.28 | 21 | 59.00 | 69.0 | 80.00 | 313 | ▇▃▁▁▁ |
| LotArea | 0 | 1.00 | 10516.83 | 9981.26 | 1300 | 7553.50 | 9478.5 | 11601.50 | 215245 | ▇▁▁▁▁ |
| OverallQual | 0 | 1.00 | 6.10 | 1.38 | 1 | 5.00 | 6.0 | 7.00 | 10 | ▁▂▇▅▁ |
| OverallCond | 0 | 1.00 | 5.58 | 1.11 | 1 | 5.00 | 5.0 | 6.00 | 9 | ▁▁▇▅▁ |
| YearBuilt | 0 | 1.00 | 1971.27 | 30.20 | 1872 | 1954.00 | 1973.0 | 2000.00 | 2010 | ▁▂▃▆▇ |
| YearRemodAdd | 0 | 1.00 | 1984.87 | 20.65 | 1950 | 1967.00 | 1994.0 | 2004.00 | 2010 | ▅▂▂▃▇ |
| MasVnrArea | 8 | 0.99 | 103.69 | 181.07 | 0 | 0.00 | 0.0 | 166.00 | 1600 | ▇▁▁▁▁ |
| BsmtFinSF1 | 0 | 1.00 | 443.64 | 456.10 | 0 | 0.00 | 383.5 | 712.25 | 5644 | ▇▁▁▁▁ |
| BsmtFinSF2 | 0 | 1.00 | 46.55 | 161.32 | 0 | 0.00 | 0.0 | 0.00 | 1474 | ▇▁▁▁▁ |
| BsmtUnfSF | 0 | 1.00 | 567.24 | 441.87 | 0 | 223.00 | 477.5 | 808.00 | 2336 | ▇▅▂▁▁ |
| TotalBsmtSF | 0 | 1.00 | 1057.43 | 438.71 | 0 | 795.75 | 991.5 | 1298.25 | 6110 | ▇▃▁▁▁ |
| 1stFlrSF | 0 | 1.00 | 1162.63 | 386.59 | 334 | 882.00 | 1087.0 | 1391.25 | 4692 | ▇▅▁▁▁ |
| 2ndFlrSF | 0 | 1.00 | 346.99 | 436.53 | 0 | 0.00 | 0.0 | 728.00 | 2065 | ▇▃▂▁▁ |
| LowQualFinSF | 0 | 1.00 | 5.84 | 48.62 | 0 | 0.00 | 0.0 | 0.00 | 572 | ▇▁▁▁▁ |
| GrLivArea | 0 | 1.00 | 1515.46 | 525.48 | 334 | 1129.50 | 1464.0 | 1776.75 | 5642 | ▇▇▁▁▁ |
| BsmtFullBath | 0 | 1.00 | 0.43 | 0.52 | 0 | 0.00 | 0.0 | 1.00 | 3 | ▇▆▁▁▁ |
| BsmtHalfBath | 0 | 1.00 | 0.06 | 0.24 | 0 | 0.00 | 0.0 | 0.00 | 2 | ▇▁▁▁▁ |
| FullBath | 0 | 1.00 | 1.57 | 0.55 | 0 | 1.00 | 2.0 | 2.00 | 3 | ▁▇▁▇▁ |
| HalfBath | 0 | 1.00 | 0.38 | 0.50 | 0 | 0.00 | 0.0 | 1.00 | 2 | ▇▁▅▁▁ |
| BedroomAbvGr | 0 | 1.00 | 2.87 | 0.82 | 0 | 2.00 | 3.0 | 3.00 | 8 | ▁▇▂▁▁ |
| KitchenAbvGr | 0 | 1.00 | 1.05 | 0.22 | 0 | 1.00 | 1.0 | 1.00 | 3 | ▁▇▁▁▁ |
| TotRmsAbvGrd | 0 | 1.00 | 6.52 | 1.63 | 2 | 5.00 | 6.0 | 7.00 | 14 | ▂▇▇▁▁ |
| Fireplaces | 0 | 1.00 | 0.61 | 0.64 | 0 | 0.00 | 1.0 | 1.00 | 3 | ▇▇▁▁▁ |
| GarageYrBlt | 81 | 0.94 | 1978.51 | 24.69 | 1900 | 1961.00 | 1980.0 | 2002.00 | 2010 | ▁▁▅▅▇ |
| GarageCars | 0 | 1.00 | 1.77 | 0.75 | 0 | 1.00 | 2.0 | 2.00 | 4 | ▁▃▇▂▁ |
| GarageArea | 0 | 1.00 | 472.98 | 213.80 | 0 | 334.50 | 480.0 | 576.00 | 1418 | ▂▇▃▁▁ |
| WoodDeckSF | 0 | 1.00 | 94.24 | 125.34 | 0 | 0.00 | 0.0 | 168.00 | 857 | ▇▂▁▁▁ |
| OpenPorchSF | 0 | 1.00 | 46.66 | 66.26 | 0 | 0.00 | 25.0 | 68.00 | 547 | ▇▁▁▁▁ |
| EnclosedPorch | 0 | 1.00 | 21.95 | 61.12 | 0 | 0.00 | 0.0 | 0.00 | 552 | ▇▁▁▁▁ |
| 3SsnPorch | 0 | 1.00 | 3.41 | 29.32 | 0 | 0.00 | 0.0 | 0.00 | 508 | ▇▁▁▁▁ |
| ScreenPorch | 0 | 1.00 | 15.06 | 55.76 | 0 | 0.00 | 0.0 | 0.00 | 480 | ▇▁▁▁▁ |
| PoolArea | 0 | 1.00 | 2.76 | 40.18 | 0 | 0.00 | 0.0 | 0.00 | 738 | ▇▁▁▁▁ |
| MiscVal | 0 | 1.00 | 43.49 | 496.12 | 0 | 0.00 | 0.0 | 0.00 | 15500 | ▇▁▁▁▁ |
| MoSold | 0 | 1.00 | 6.32 | 2.70 | 1 | 5.00 | 6.0 | 8.00 | 12 | ▃▆▇▃▃ |
| YrSold | 0 | 1.00 | 2007.82 | 1.33 | 2006 | 2007.00 | 2008.0 | 2009.00 | 2010 | ▇▇▇▇▅ |
| SalePrice | 0 | 1.00 | 180921.20 | 79442.50 | 34900 | 129975.00 | 163000.0 | 214000.00 | 755000 | ▇▅▁▁▁ |

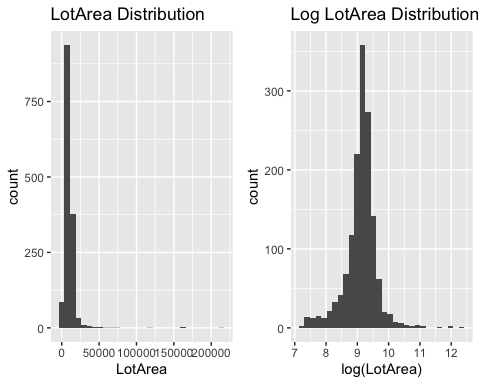
## Variables Distribution

#SalePrice Distribution Vs Log transformed SalePrice Distribution  
pl1<-ggplot(train, aes(SalePrice)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "SalePrice Distribution")  
pl2<-ggplot(train, aes(log(SalePrice))) +  
 geom\_histogram(bins = 30) +  
 labs(title = "Log SalePrice Distribution")  
grid.arrange(pl1, pl2, ncol = 2, nrow = 1)



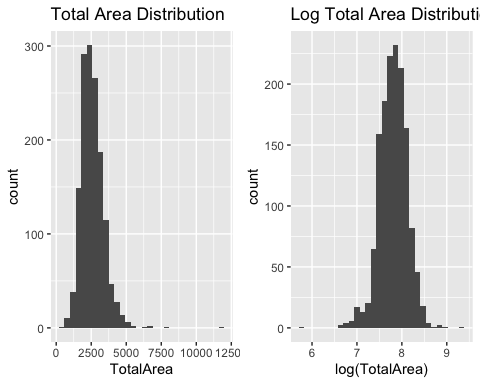
From the above graph it can be observed that the SalePrice is not normally distributed and right-skewed with some outliers. So SalePrice is log transformed for modeling. After log transformation it looks more like a bell-curved normal distribution.

#LotArea Distribution Vs Log transformed LotArea Distribution  
pl3<-ggplot(train, aes(LotArea)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "LotArea Distribution")  
pl4<-ggplot(train, aes(log(LotArea))) +  
 geom\_histogram(bins = 30) +  
 labs(title = "Log LotArea Distribution")  
grid.arrange(pl3, pl4, ncol = 2, nrow = 1)



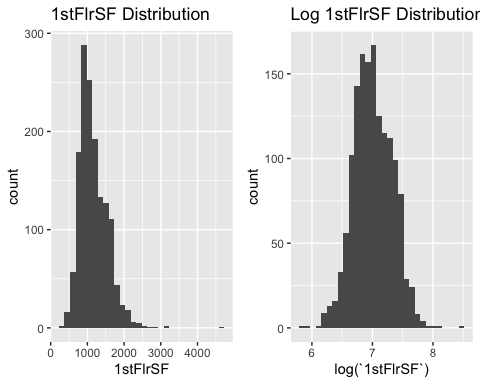
From the above graph it can be observed that the LotArea is not normally distributed and highly right-skewed with some outliers. So LotArea is log transformed for modeling. After log transformation it looks more like a normal distribution however having a right tail.

#Total Area Distribution Vs Log transformed Total Area Distribution  
TotalArea = (train$TotalBsmtSF + train$GrLivArea)  
pl5<-ggplot(train, aes(TotalArea)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "Total Area Distribution")  
pl6<-ggplot(train, aes(log(TotalArea))) +  
 geom\_histogram(bins = 30) +  
 labs(title = "Log Total Area Distribution")  
grid.arrange(pl5, pl6, ncol = 2, nrow = 1)



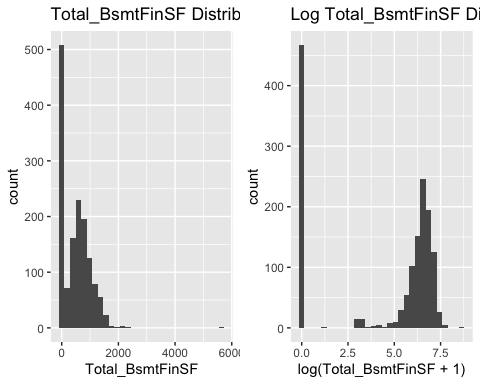
TotalBsmtSF and GrLivArea are added to create a new variable TotalArea of the house. From the above graph it can be observed that the Total Area is not normally distributed and right-skewed with some outliers. So Total Area is log transformed for modeling. After log transformation it looks more like a normal distribution.

#1stFlrSF Distribution Vs Log transformed 1stFlrSF Distribution  
pl7<-ggplot(train, aes(`1stFlrSF`)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "1stFlrSF Distribution")  
pl8<-ggplot(train, aes(log(`1stFlrSF`))) +  
 geom\_histogram(bins = 30) +  
 labs(title = "Log 1stFlrSF Distribution")  
grid.arrange(pl7, pl8, ncol = 2, nrow = 1)



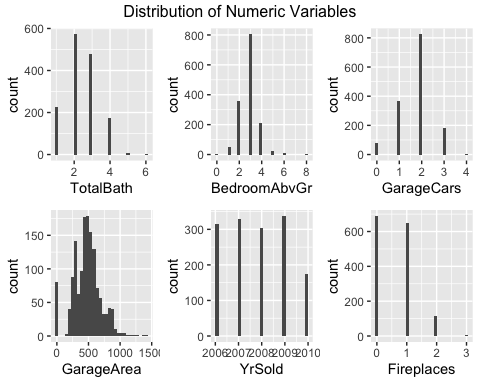
From the above graph it can be observed that the 1stFlrSF is not normally distributed and right-skewed with some outliers. So 1stFlrSF is log transformed for modeling. After log transformation it looks more like a normal distribution.

#Total\_BsmtFinSF Distribution Vs Log transformed Total\_BsmtFinSF Distribution  
Total\_BsmtFinSF = (train$BsmtFinSF1 + train$BsmtFinSF2)  
pl9<-ggplot(train, aes(Total\_BsmtFinSF)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "Total\_BsmtFinSF Distribution")  
pl10<-ggplot(train, aes(log(Total\_BsmtFinSF+1))) +  
 geom\_histogram(bins = 30) +  
 labs(title = "Log Total\_BsmtFinSF Distribution")  
grid.arrange(pl9, pl10, ncol = 2, nrow = 1)



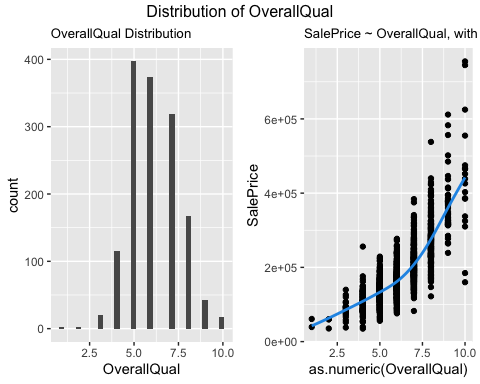
BsmtFinSF1 and BsmtFinSF2 are added to get the total basement finished Square feet area (Total\_BsmtFinSF). From the above graph it can be observed that the Total\_BsmtFinSF is not normally distributed and right-skewed with some outliers with most of the values are zero. So Total\_BsmtFinSF is log transformed for modeling. After log transformation it looks more like a normal distribution apart from observations having a value of zero .

#TotalBath Distribution  
TotalBath = (train$HalfBath + train$FullBath + train$BsmtFullBath + train$BsmtHalfBath)  
pl11<-ggplot(train, aes(TotalBath)) +  
 geom\_histogram(bins = 30)   
  
#BedroomAbvGr Distribution  
pl12<-ggplot(train, aes(BedroomAbvGr)) +  
 geom\_histogram(bins = 30)   
  
#GarageCars Distribution  
pl13<-ggplot(train, aes(GarageCars)) +  
 geom\_histogram(bins = 30)   
  
#GarageArea Distribution  
pl14<-ggplot(train, aes(GarageArea)) +  
 geom\_histogram(bins = 30)   
  
#YrSold Distribution  
pl15<-ggplot(train, aes(YrSold)) +  
 geom\_histogram(bins = 30)   
  
#Fireplaces Distribution  
pl16<-ggplot(train, aes(Fireplaces)) +  
 geom\_histogram(bins = 30)   
margin = theme(plot.margin = unit(c(1,1,1,1), "cm"))  
grid.arrange(pl11, pl12, pl13, pl14, pl15, pl16, ncol = 3, nrow = 2, top=textGrob("Distribution of Numeric Variables"))



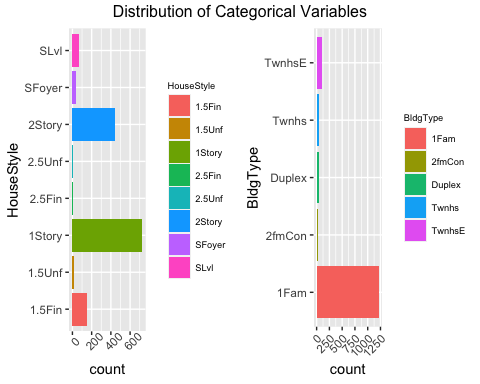
All the bathrooms available in a house are added to create TotalBath variable. From the above vizulization it can be observed that TotalBath, BedroomAbvGr, GarageCars, and GarageArea are almost look like normal distribution with GarageArea having mamy zeros. Number of houses sold is equally distributed for all the years except 2010 and most houses do not have a Fireplace with few houses having 3 Fireplaces. So, no transformations will be done to these varaibles for modeling purpose.

#OverallQual Distribution, need to factor the overall quality  
p1 <- ggplot(train, aes(OverallQual)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "OverallQual Distribution") +  
 theme(plot.title = element\_text(size = 10))   
  
p2 <- ggplot(train, aes(as.numeric(OverallQual), SalePrice)) +  
 geom\_point() +  
 geom\_smooth(se = F, col = 4) + # Local regression named LOESS  
 labs(title = "SalePrice ~ OverallQual, with local regression") +  
 theme(plot.title = element\_text(size = 10))   
  
margin = theme(plot.margin = unit(c(1,1,1,1), "cm"))  
grid.arrange(p1, p2, ncol = 2, nrow = 1, top=textGrob("Distribution of OverallQual"))



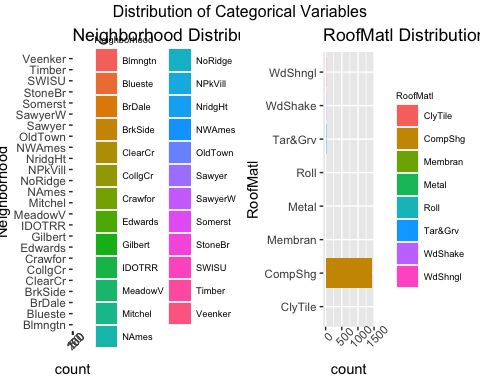
As observed from the above graphs Overall Quality cannot be described by a line (as it looks more like an exponential distribution ) so it will be converted to factor for modeling purpose.

# House style and BldgType distribution  
pl17 <- ggplot(train, aes(y = HouseStyle, fill = HouseStyle)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 45),  
 legend.title = element\_text(size = 7),   
 legend.text = element\_text(size = 7))  
pl18 <-ggplot(train, aes(y = BldgType, fill = BldgType)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 45, vjust = 1, hjust = 1),  
 legend.title = element\_text(size = 7),   
 legend.text = element\_text(size = 7))   
grid.arrange(pl17,pl18, ncol = 2, nrow = 1, top=textGrob("Distribution of Categorical Variables"))



Above graphs represent the distribution of House style and Building type from which we can observe that most houses are one storey and single family house. Both House Style and BldgType will be factored for modeling purpose.

# Neighborhood distribution  
pl19 <- ggplot(train, aes(y = Neighborhood, fill = Neighborhood)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 45),  
 legend.title = element\_text(size = 7),   
 legend.text = element\_text(size = 7)) +  
 labs(title = "Neighborhood Distribution")  
  
pl20 <- ggplot(train, aes(y = RoofMatl, fill = RoofMatl)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 45),  
 legend.title = element\_text(size = 7),   
 legend.text = element\_text(size = 7)) +  
 labs(title = "RoofMatl Distribution")  
grid.arrange(pl19, pl20, ncol = 2, nrow = 1, top=textGrob("Distribution of Categorical Variables"))



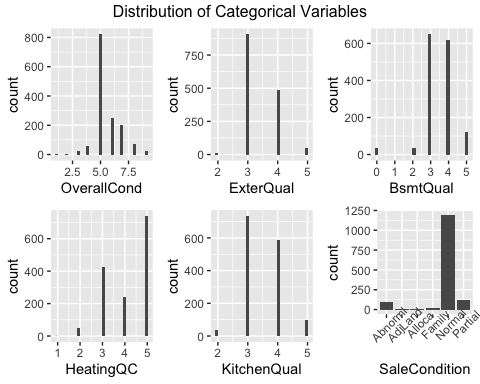
Above graph represents the distribution of neighborhoods where houses were sold and it can be observed that NAmes (North Ames) area has the most number of houses sold. Also, most of the houses has the RoofMatl of Standard (Composite) Shingle. Neighborhood and RoofMatl will be factored for modeling purpose.

## Data Transformation & Missing Value Imputation

# Data Transformation & Missing Value Imputation  
train<-train %>%   
 mutate(log\_TotalArea = log(TotalBsmtSF + GrLivArea),  
 log\_SalePrice = log(SalePrice),  
 log\_LotArea = log(LotArea),  
 log\_1stFlrSF = log(`1stFlrSF`),   
 TotalBath = (HalfBath + FullBath + BsmtFullBath + BsmtHalfBath),  
 ExterQual= case\_when(ExterQual == "Ex" ~ 5,  
 ExterQual == "Gd" ~ 4,  
 ExterQual == "TA" ~ 3,  
 ExterQual == "Fa" ~ 2,  
 ExterQual == "Po" ~ 1),  
 ExterQual = ifelse(is.na(ExterQual), 0, ExterQual),  
 BsmtQual = case\_when(BsmtQual == "Ex" ~ 5,  
 BsmtQual == "Gd" ~ 4,  
 BsmtQual == "TA" ~ 3,  
 BsmtQual == "Fa" ~ 2,  
 BsmtQual == "Po" ~ 1),  
 BsmtQual = ifelse(is.na(BsmtQual), 0, BsmtQual),  
 log\_Total\_BsmtFinSF = log(BsmtFinSF1 + BsmtFinSF2 + 1),  
 HeatingQC= case\_when(HeatingQC == "Ex" ~ 5,  
 HeatingQC == "Gd" ~ 4,  
 HeatingQC == "TA" ~ 3,  
 HeatingQC == "Fa" ~ 2,  
 HeatingQC == "Po" ~ 1),  
 HeatingQC = ifelse(is.na(HeatingQC), 0, HeatingQC),  
 KitchenQual = case\_when(KitchenQual == "Ex" ~ 5,  
 KitchenQual == "Gd" ~ 4,  
 KitchenQual == "TA" ~ 3,  
 KitchenQual == "Fa" ~ 2,  
 KitchenQual == "Po" ~ 1),  
 KitchenQual = ifelse(is.na(KitchenQual), 0, KitchenQual))

Total Area (TotalBsmtSF + GrLivArea), Total\_BsmtFinSF = (BsmtFinSF1 + BsmtFinSF2), and TotalBath = (HalfBath + FullBath + BsmtFullBath + BsmtHalfBath) are created. TotalArea, SalePrice, LotArea, Total\_BsmtFinSF, and 1stFlrSF are log transformed to make sure these variables distribution looks more like a normal distribution. Qualitative variables ExterQual, BsmtQual, HeatingQC, and KitchenQual are converted from category to ordinal variable to better predict the sale price. Also, if any qualitative variable value is missing it is assigned as zero.

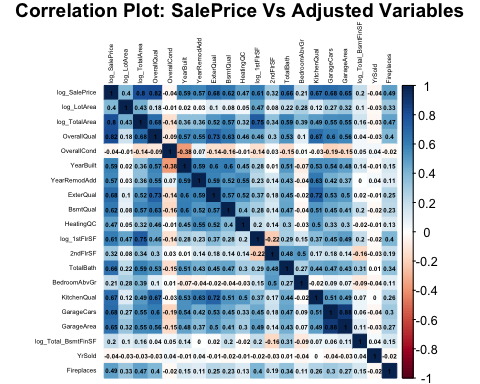
#OverallCond Distribution  
pl21<-ggplot(train, aes(OverallCond)) +  
 geom\_histogram(bins = 30)   
  
#ExterQual Distribution  
pl22<-ggplot(train, aes(ExterQual)) +  
 geom\_histogram(bins = 30)   
  
#BsmtQual Distribution  
pl23<-ggplot(train, aes(BsmtQual)) +  
 geom\_histogram(bins = 30)   
  
#HeatingQC Distribution  
pl24<-ggplot(train, aes(HeatingQC)) +  
 geom\_histogram(bins = 30)   
  
#KitchenQual Distribution  
pl25<-ggplot(train, aes(KitchenQual)) +  
 geom\_histogram(bins = 30)   
  
#SaleCondition Distribution  
pl26<-ggplot(train, aes(SaleCondition)) +  
 geom\_bar() +  
 theme(axis.text.x = element\_text(angle = 45))  
  
margin = theme(plot.margin = unit(c(1,1,1,1), "cm"))  
grid.arrange(pl21, pl22, pl23, pl24, pl25, pl26, ncol = 3, nrow = 2, top=textGrob("Distribution of Categorical Variables"))



Above graph represents all the qualitative variables: OverallCond, ExterQual, BsmtQual, HeatingQC, KitchenQual, and Salecondition. The qualitative variables are converted to ordinal variables to better predict the sale price. These will be converted to factors for modeling purpose.

# Correlation Plot

#correlation Plot  
corrdata <- select(train, log\_SalePrice, log\_LotArea, log\_TotalArea, OverallQual, OverallCond, YearBuilt , YearRemodAdd, ExterQual, BsmtQual, HeatingQC,log\_1stFlrSF, `2ndFlrSF`, TotalBath, BedroomAbvGr, KitchenQual, GarageCars , GarageArea, log\_Total\_BsmtFinSF, YrSold, Fireplaces )  
  
corrplot(cor(corrdata), method="color", addCoef.col = "black", tl.col="black", tl.cex=.4, number.cex=.4, title = "Correlation Plot: SalePrice Vs Adjusted Variables", mar=c(0,0,1,0))



Above plot shows the correlation bewtween the adjusted numeric variables and the log transformed SalePrice. From the above correlation plot we can observe that the selected varaibles are highly correlated with log sale price (e.g. correlation between OverallQual and LogSalePrice is 0.82 and correlation between log\_TotalArea and LogSalePrice is 0.8) which would be helpful predict the sale price accurately.

## Train\_fold and Validation\_fold Data Preparation

#splitting the data into train and test data set  
set.seed(123)  
index <- sample(x = 1:nrow(train), size = nrow(train)\*.7, replace = F)  
  
head(index)

## [1] 415 463 179 526 195 938

# Subset train using the index to create train\_fold  
train\_fold <- train[index, ]  
  
# Subset the remaining row to create validation fold.  
validation\_fold <- train[-index, ]

## RMSE and R2 Function

# Create functions for calculating RMSE and R-squared  
rmse <- function(observed, predicted) sqrt(mean((observed - predicted)^2))  
  
R2 <- function(observed, predicted){  
 TSS <- sum((observed - mean(observed))^2)  
 RSS <- sum((observed - predicted)^2)  
 1- RSS/TSS  
}

# Variable (Feature) Selection For Modeling

Here is a list of predictor variables used for modeling: log\_LotArea, LotConfig, Neighborhood, log\_TotalArea, BldgType, HouseStyle, OverallQual, OverallCond, YearBuilt, YearRemodAdd, RoofMatl, ExterQual, BsmtQual, HeatingQC, log\_1stFlrSF, 2ndFlrSF, TotalBath, BedroomAbvGr, KitchenQual, GarageCars, GarageArea, SaleCondition, log\_Total\_BsmtFinSF, YrSold, and Fireplaces. The variables which are highly correlated with the sale price and help increase the R-square value of train-fold, validation\_fold, and submission model while decreasing the RMSE value are selected to predict the Sale Price of the houses. An interaction term between YearBuilt and YearRemodAdd has been added to the model to describe a situation in which the effect of YearBuilt variable on the Sale Price depends on the YearRemodAdd which is a second causal variable (that is, when effects of the two causes are not additive).

# Data Description

1. SalePrice: the property’s sale price in dollars. This is the target variable that you’re trying to predict
2. LotArea: Lot size in square feet
3. LotConfig: Lot configuration
4. Neighborhood: Physical locations within Ames city limits
5. TotalBsmtSF: Total square feet of basement area
6. GrLivArea: Above grade (ground) living area square feet
7. BldgType: Type of dwelling
8. HouseStyle: Style of dwelling
9. OverallQual: Overall material and finish quality
10. OverallCond: Overall condition rating
11. YearBuilt: Original construction date
12. YearRemodAdd: Remodel date
13. RoofMatl: Roof material
14. ExterQual: Exterior material quality
15. BsmtQual: Height of the basement
16. HeatingQC: Heating quality and condition
17. 1stFlrSF: First Floor square feet
18. 2ndFlrSF: Second floor square feet
19. BsmtFullBath: Basement full bathrooms
20. BsmtHalfBath: Basement half bathrooms
21. FullBath: Full bathrooms above grade
22. HalfBath: Half baths above grade
23. Bedroom: Number of bedrooms above basement level
24. KitchenQual: Kitchen quality
25. GarageCars: Size of garage in car capacity
26. GarageArea: Size of garage in square feet
27. SaleCondition: Condition of sale
28. BsmtFinSF1: Type 1 finished square feet
29. BsmtFinSF2: Type 2 finished square feet
30. YrSold: Year Sold
31. Fireplaces: Number of fireplaces

# Modeling

## Train\_fold Model

# Train model using the train\_fold  
train\_model <- (lm(log\_SalePrice ~ log\_LotArea + factor(LotConfig) + factor(Neighborhood) + log\_TotalArea + factor(BldgType) + factor(HouseStyle) + factor(OverallQual) + factor(OverallCond) + (YearBuilt \* YearRemodAdd) + factor(ExterQual) + factor(BsmtQual) + factor(HeatingQC) + log\_1stFlrSF + `2ndFlrSF` + TotalBath + BedroomAbvGr + factor(KitchenQual) + GarageCars + GarageArea + SaleCondition + log\_Total\_BsmtFinSF + YrSold + Fireplaces + factor(RoofMatl) , data = train\_fold))  
#Summary of the train model  
summary(train\_model)

##   
## Call:  
## lm(formula = log\_SalePrice ~ log\_LotArea + factor(LotConfig) +   
## factor(Neighborhood) + log\_TotalArea + factor(BldgType) +   
## factor(HouseStyle) + factor(OverallQual) + factor(OverallCond) +   
## (YearBuilt \* YearRemodAdd) + factor(ExterQual) + factor(BsmtQual) +   
## factor(HeatingQC) + log\_1stFlrSF + `2ndFlrSF` + TotalBath +   
## BedroomAbvGr + factor(KitchenQual) + GarageCars + GarageArea +   
## SaleCondition + log\_Total\_BsmtFinSF + YrSold + Fireplaces +   
## factor(RoofMatl), data = train\_fold)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.07292 -0.05388 0.00475 0.05541 0.40262   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.166e+01 4.320e+01 -2.122 0.034109 \*   
## log\_LotArea 6.374e-02 1.368e-02 4.661 3.61e-06 \*\*\*  
## factor(LotConfig)CulDSac 2.238e-02 1.899e-02 1.179 0.238741   
## factor(LotConfig)FR2 -4.603e-02 2.244e-02 -2.051 0.040524 \*   
## factor(LotConfig)FR3 -4.477e-02 8.693e-02 -0.515 0.606651   
## factor(LotConfig)Inside -1.655e-02 1.034e-02 -1.601 0.109821   
## factor(Neighborhood)Blueste -8.839e-02 1.263e-01 -0.700 0.484063   
## factor(Neighborhood)BrDale -1.081e-01 5.879e-02 -1.839 0.066303 .   
## factor(Neighborhood)BrkSide -4.755e-02 4.779e-02 -0.995 0.320055   
## factor(Neighborhood)ClearCr -8.087e-02 5.024e-02 -1.609 0.107852   
## factor(Neighborhood)CollgCr -5.977e-02 3.945e-02 -1.515 0.130107   
## factor(Neighborhood)Crawfor 7.784e-02 4.515e-02 1.724 0.085078 .   
## factor(Neighborhood)Edwards -1.459e-01 4.295e-02 -3.397 0.000710 \*\*\*  
## factor(Neighborhood)Gilbert -5.973e-02 4.227e-02 -1.413 0.158017   
## factor(Neighborhood)IDOTRR -1.411e-01 5.157e-02 -2.736 0.006336 \*\*   
## factor(Neighborhood)MeadowV -1.932e-01 5.668e-02 -3.408 0.000682 \*\*\*  
## factor(Neighborhood)Mitchel -9.808e-02 4.423e-02 -2.218 0.026826 \*   
## factor(Neighborhood)NAmes -8.149e-02 4.159e-02 -1.960 0.050343 .   
## factor(Neighborhood)NoRidge 5.989e-03 4.557e-02 0.131 0.895459   
## factor(Neighborhood)NPkVill -5.318e-02 6.014e-02 -0.884 0.376740   
## factor(Neighborhood)NridgHt 4.245e-02 4.045e-02 1.050 0.294198   
## factor(Neighborhood)NWAmes -1.309e-01 4.355e-02 -3.006 0.002716 \*\*   
## factor(Neighborhood)OldTown -1.199e-01 4.506e-02 -2.661 0.007918 \*\*   
## factor(Neighborhood)Sawyer -1.134e-01 4.333e-02 -2.619 0.008975 \*\*   
## factor(Neighborhood)SawyerW -7.780e-02 4.210e-02 -1.848 0.064884 .   
## factor(Neighborhood)Somerst 1.001e-02 3.914e-02 0.256 0.798125   
## factor(Neighborhood)StoneBr 8.274e-02 4.626e-02 1.789 0.074007 .   
## factor(Neighborhood)SWISU -4.293e-02 5.287e-02 -0.812 0.417022   
## factor(Neighborhood)Timber -4.778e-02 4.447e-02 -1.075 0.282821   
## factor(Neighborhood)Veenker -3.875e-02 5.559e-02 -0.697 0.485975   
## log\_TotalArea 3.968e-01 6.007e-02 6.606 6.66e-11 \*\*\*  
## factor(BldgType)2fmCon -5.645e-02 2.962e-02 -1.906 0.056975 .   
## factor(BldgType)Duplex -1.023e-01 2.608e-02 -3.922 9.44e-05 \*\*\*  
## factor(BldgType)Twnhs -5.653e-02 3.729e-02 -1.516 0.129896   
## factor(BldgType)TwnhsE -1.778e-02 2.057e-02 -0.864 0.387678   
## factor(HouseStyle)1.5Unf -2.553e-02 4.588e-02 -0.556 0.578067   
## factor(HouseStyle)1Story 3.213e-03 2.173e-02 0.148 0.882466   
## factor(HouseStyle)2.5Fin -1.214e-02 6.113e-02 -0.199 0.842649   
## factor(HouseStyle)2.5Unf 2.857e-02 5.431e-02 0.526 0.598960   
## factor(HouseStyle)2Story -3.480e-02 1.862e-02 -1.869 0.061947 .   
## factor(HouseStyle)SFoyer 2.900e-02 3.618e-02 0.802 0.423040   
## factor(HouseStyle)SLvl 3.699e-02 2.786e-02 1.328 0.184534   
## factor(OverallQual)2 4.356e-02 1.718e-01 0.254 0.799935   
## factor(OverallQual)3 -7.213e-02 1.415e-01 -0.510 0.610465   
## factor(OverallQual)4 1.796e-02 1.425e-01 0.126 0.899750   
## factor(OverallQual)5 2.747e-02 1.438e-01 0.191 0.848532   
## factor(OverallQual)6 6.355e-02 1.444e-01 0.440 0.660091   
## factor(OverallQual)7 1.186e-01 1.451e-01 0.817 0.414008   
## factor(OverallQual)8 2.049e-01 1.462e-01 1.401 0.161402   
## factor(OverallQual)9 3.070e-01 1.492e-01 2.058 0.039889 \*   
## factor(OverallQual)10 2.711e-01 1.535e-01 1.767 0.077588 .   
## factor(OverallCond)2 1.181e-01 1.926e-01 0.613 0.539726   
## factor(OverallCond)3 6.386e-02 1.780e-01 0.359 0.719882   
## factor(OverallCond)4 1.509e-01 1.834e-01 0.823 0.410942   
## factor(OverallCond)5 2.150e-01 1.824e-01 1.178 0.238925   
## factor(OverallCond)6 2.673e-01 1.824e-01 1.466 0.143066   
## factor(OverallCond)7 3.126e-01 1.828e-01 1.710 0.087635 .   
## factor(OverallCond)8 3.400e-01 1.834e-01 1.854 0.064010 .   
## factor(OverallCond)9 3.840e-01 1.865e-01 2.059 0.039813 \*   
## YearBuilt 5.231e-02 2.183e-02 2.397 0.016737 \*   
## YearRemodAdd 4.956e-02 2.133e-02 2.324 0.020347 \*   
## factor(ExterQual)3 1.201e-01 5.202e-02 2.310 0.021122 \*   
## factor(ExterQual)4 1.160e-01 5.402e-02 2.148 0.031999 \*   
## factor(ExterQual)5 1.165e-01 6.068e-02 1.920 0.055184 .   
## factor(BsmtQual)2 -1.125e-01 5.182e-02 -2.171 0.030171 \*   
## factor(BsmtQual)3 -1.349e-01 4.663e-02 -2.894 0.003893 \*\*   
## factor(BsmtQual)4 -1.215e-01 4.897e-02 -2.480 0.013309 \*   
## factor(BsmtQual)5 -8.722e-02 5.322e-02 -1.639 0.101598   
## factor(HeatingQC)2 7.272e-02 1.314e-01 0.553 0.580200   
## factor(HeatingQC)3 5.449e-02 1.305e-01 0.418 0.676354   
## factor(HeatingQC)4 6.859e-02 1.305e-01 0.526 0.599359   
## factor(HeatingQC)5 9.489e-02 1.303e-01 0.728 0.466699   
## log\_1stFlrSF 5.295e-02 5.174e-02 1.023 0.306377   
## `2ndFlrSF` 1.045e-04 2.874e-05 3.636 0.000292 \*\*\*  
## TotalBath 3.084e-02 7.424e-03 4.154 3.57e-05 \*\*\*  
## BedroomAbvGr -1.014e-03 7.336e-03 -0.138 0.890062   
## factor(KitchenQual)3 1.637e-02 2.930e-02 0.559 0.576498   
## factor(KitchenQual)4 3.457e-02 3.162e-02 1.094 0.274409   
## factor(KitchenQual)5 7.790e-02 3.669e-02 2.123 0.033997 \*   
## GarageCars 3.299e-02 1.287e-02 2.563 0.010538 \*   
## GarageArea 3.790e-05 4.333e-05 0.875 0.381979   
## SaleConditionAdjLand 1.215e-01 7.436e-02 1.633 0.102762   
## SaleConditionAlloca 4.731e-03 4.456e-02 0.106 0.915476   
## SaleConditionFamily 1.377e-02 3.511e-02 0.392 0.695014   
## SaleConditionNormal 4.748e-02 1.571e-02 3.022 0.002584 \*\*   
## SaleConditionPartial 7.462e-02 2.209e-02 3.379 0.000759 \*\*\*  
## log\_Total\_BsmtFinSF 9.786e-03 1.659e-03 5.900 5.09e-09 \*\*\*  
## YrSold -2.796e-03 2.947e-03 -0.949 0.342969   
## Fireplaces 3.431e-02 7.748e-03 4.427 1.07e-05 \*\*\*  
## factor(RoofMatl)CompShg 1.449e+00 1.346e-01 10.763 < 2e-16 \*\*\*  
## factor(RoofMatl)Metal 1.514e+00 1.847e-01 8.195 8.32e-16 \*\*\*  
## factor(RoofMatl)Roll 1.450e+00 1.827e-01 7.935 6.05e-15 \*\*\*  
## factor(RoofMatl)Tar&Grv 1.428e+00 1.405e-01 10.169 < 2e-16 \*\*\*  
## factor(RoofMatl)WdShake 1.519e+00 1.525e-01 9.962 < 2e-16 \*\*\*  
## factor(RoofMatl)WdShngl 1.623e+00 1.426e-01 11.382 < 2e-16 \*\*\*  
## YearBuilt:YearRemodAdd -2.513e-05 1.097e-05 -2.291 0.022203 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1185 on 925 degrees of freedom  
## Multiple R-squared: 0.9183, Adjusted R-squared: 0.9099   
## F-statistic: 109.5 on 95 and 925 DF, p-value: < 2.2e-16

Residual standard error for Train\_fold model: 0.1185 (0.12)

Multiple R-squared for Train\_fold model: 0.9183 (0.92)

## Train\_fold Prediction

# Get predictions for the train fold  
predictions\_train <- predict(train\_model, newdata = train\_fold)  
  
#calculating rmse and r2 for the training data set   
rmse(train\_fold$log\_SalePrice, predictions\_train)

## [1] 0.1128045

R2(train\_fold$log\_SalePrice, predictions\_train)

## [1] 0.9183293

RMSE for Train\_fold Prediction: 0.1128045 (0.11)

R2 for Train\_fold Prediction: 0.9183293 (0.92)

## Validation\_fold Prediction

# Get predictions for the validation fold  
validation\_fold <- validation\_fold %>% filter(validation\_fold$RoofMatl !="Membran")  
predictions\_validation <- predict(train\_model, newdata = validation\_fold)  
  
#calculating rmse and r2 for the validation data set   
rmse(validation\_fold$log\_SalePrice, predictions\_validation)

## [1] 0.1166372

R2(validation\_fold$log\_SalePrice, predictions\_validation)

## [1] 0.9190226

In the RoofMatl variable there is only one observation having the category of “Membran”. So it can either be in train\_fold or validation\_fold. So, the row having the value of “Membran” for RoofMatl is removed to able to predict the validation fold sale price (and to avoid the error that the same categories are not avaiable in both the train fold data and validation fold data).

RMSE for Test\_fold Prediction: 0.1166372 (0.12)

R2 for Test\_fold Prediction: 0.9190226 (0.92)

## Submission Model

#submission model  
submission\_model <- (lm(log\_SalePrice ~ log\_LotArea + factor(LotConfig) + factor(Neighborhood) + log\_TotalArea+ factor(BldgType) + factor(HouseStyle) + factor(OverallQual) + factor(OverallCond) + (YearBuilt \* YearRemodAdd) + factor(ExterQual) + factor(BsmtQual) + factor(HeatingQC) + log\_1stFlrSF + `2ndFlrSF` + TotalBath + BedroomAbvGr + factor(KitchenQual) + GarageCars + GarageArea + SaleCondition + log\_Total\_BsmtFinSF + YrSold + Fireplaces + factor(RoofMatl), data = train))  
  
#Submission model summary  
summary(submission\_model)

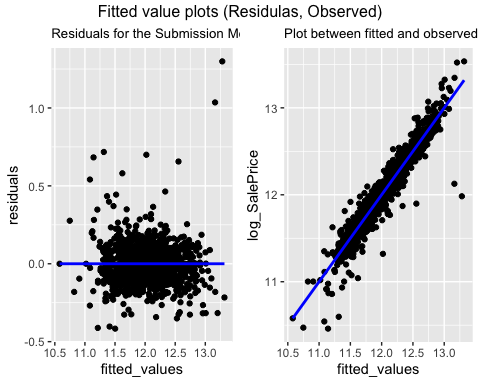
##   
## Call:  
## lm(formula = log\_SalePrice ~ log\_LotArea + factor(LotConfig) +   
## factor(Neighborhood) + log\_TotalArea + factor(BldgType) +   
## factor(HouseStyle) + factor(OverallQual) + factor(OverallCond) +   
## (YearBuilt \* YearRemodAdd) + factor(ExterQual) + factor(BsmtQual) +   
## factor(HeatingQC) + log\_1stFlrSF + `2ndFlrSF` + TotalBath +   
## BedroomAbvGr + factor(KitchenQual) + GarageCars + GarageArea +   
## SaleCondition + log\_Total\_BsmtFinSF + YrSold + Fireplaces +   
## factor(RoofMatl), data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.16514 -0.05502 0.00301 0.05681 0.41303   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -8.338e+01 3.327e+01 -2.506 0.012331 \*   
## log\_LotArea 7.473e-02 1.088e-02 6.870 9.75e-12 \*\*\*  
## factor(LotConfig)CulDSac 1.716e-02 1.483e-02 1.157 0.247315   
## factor(LotConfig)FR2 -4.685e-02 1.915e-02 -2.446 0.014559 \*   
## factor(LotConfig)FR3 -9.925e-02 5.971e-02 -1.662 0.096702 .   
## factor(LotConfig)Inside -1.212e-02 8.244e-03 -1.470 0.141706   
## factor(Neighborhood)Blueste -4.464e-02 8.952e-02 -0.499 0.618075   
## factor(Neighborhood)BrDale -4.020e-02 4.812e-02 -0.836 0.403567   
## factor(Neighborhood)BrkSide -3.345e-04 3.916e-02 -0.009 0.993186   
## factor(Neighborhood)ClearCr -8.547e-03 4.225e-02 -0.202 0.839714   
## factor(Neighborhood)CollgCr -2.678e-02 3.383e-02 -0.792 0.428753   
## factor(Neighborhood)Crawfor 9.575e-02 3.820e-02 2.506 0.012314 \*   
## factor(Neighborhood)Edwards -9.937e-02 3.639e-02 -2.731 0.006396 \*\*   
## factor(Neighborhood)Gilbert -3.523e-02 3.621e-02 -0.973 0.330730   
## factor(Neighborhood)IDOTRR -1.350e-01 4.190e-02 -3.222 0.001301 \*\*   
## factor(Neighborhood)MeadowV -1.366e-01 4.544e-02 -3.007 0.002687 \*\*   
## factor(Neighborhood)Mitchel -8.091e-02 3.781e-02 -2.140 0.032547 \*   
## factor(Neighborhood)NAmes -5.556e-02 3.559e-02 -1.561 0.118676   
## factor(Neighborhood)NoRidge 4.808e-02 3.858e-02 1.246 0.212882   
## factor(Neighborhood)NPkVill -2.726e-02 5.131e-02 -0.531 0.595348   
## factor(Neighborhood)NridgHt 5.141e-02 3.504e-02 1.467 0.142558   
## factor(Neighborhood)NWAmes -9.040e-02 3.688e-02 -2.451 0.014366 \*   
## factor(Neighborhood)OldTown -9.439e-02 3.789e-02 -2.491 0.012857 \*   
## factor(Neighborhood)Sawyer -7.891e-02 3.730e-02 -2.115 0.034576 \*   
## factor(Neighborhood)SawyerW -5.210e-02 3.619e-02 -1.439 0.150258   
## factor(Neighborhood)Somerst 4.100e-02 3.375e-02 1.215 0.224674   
## factor(Neighborhood)StoneBr 9.116e-02 3.898e-02 2.339 0.019494 \*   
## factor(Neighborhood)SWISU -6.233e-03 4.397e-02 -0.142 0.887296   
## factor(Neighborhood)Timber -2.599e-02 3.847e-02 -0.676 0.499417   
## factor(Neighborhood)Veenker -7.861e-03 4.849e-02 -0.162 0.871243   
## log\_TotalArea 3.950e-01 4.373e-02 9.033 < 2e-16 \*\*\*  
## factor(BldgType)2fmCon -3.736e-02 2.295e-02 -1.628 0.103753   
## factor(BldgType)Duplex -8.791e-02 2.065e-02 -4.257 2.21e-05 \*\*\*  
## factor(BldgType)Twnhs -4.407e-02 2.680e-02 -1.645 0.100243   
## factor(BldgType)TwnhsE -1.145e-02 1.732e-02 -0.661 0.508780   
## factor(HouseStyle)1.5Unf 1.328e-02 3.604e-02 0.369 0.712534   
## factor(HouseStyle)1Story 1.346e-02 1.699e-02 0.792 0.428479   
## factor(HouseStyle)2.5Fin 1.559e-02 4.669e-02 0.334 0.738415   
## factor(HouseStyle)2.5Unf 6.131e-03 3.830e-02 0.160 0.872849   
## factor(HouseStyle)2Story -1.177e-02 1.454e-02 -0.810 0.418241   
## factor(HouseStyle)SFoyer 4.428e-02 2.659e-02 1.665 0.096145 .   
## factor(HouseStyle)SLvl 4.693e-02 2.144e-02 2.189 0.028744 \*   
## factor(OverallQual)2 -1.192e-01 1.389e-01 -0.858 0.391069   
## factor(OverallQual)3 -1.700e-02 1.294e-01 -0.131 0.895546   
## factor(OverallQual)4 5.269e-02 1.292e-01 0.408 0.683527   
## factor(OverallQual)5 8.187e-02 1.299e-01 0.630 0.528717   
## factor(OverallQual)6 1.224e-01 1.304e-01 0.939 0.348048   
## factor(OverallQual)7 1.846e-01 1.308e-01 1.411 0.158344   
## factor(OverallQual)8 2.556e-01 1.316e-01 1.942 0.052293 .   
## factor(OverallQual)9 3.546e-01 1.341e-01 2.645 0.008263 \*\*   
## factor(OverallQual)10 3.455e-01 1.370e-01 2.522 0.011790 \*   
## factor(OverallCond)2 1.228e-01 1.793e-01 0.685 0.493616   
## factor(OverallCond)3 8.245e-02 1.694e-01 0.487 0.626580   
## factor(OverallCond)4 2.034e-01 1.726e-01 1.178 0.238865   
## factor(OverallCond)5 2.599e-01 1.722e-01 1.510 0.131378   
## factor(OverallCond)6 3.068e-01 1.722e-01 1.782 0.075019 .   
## factor(OverallCond)7 3.644e-01 1.724e-01 2.115 0.034654 \*   
## factor(OverallCond)8 3.707e-01 1.727e-01 2.146 0.032016 \*   
## factor(OverallCond)9 4.172e-01 1.748e-01 2.387 0.017132 \*   
## YearBuilt 4.902e-02 1.681e-02 2.916 0.003608 \*\*   
## YearRemodAdd 4.643e-02 1.642e-02 2.827 0.004765 \*\*   
## factor(ExterQual)3 3.931e-02 3.742e-02 1.051 0.293663   
## factor(ExterQual)4 3.589e-02 3.912e-02 0.917 0.359059   
## factor(ExterQual)5 3.929e-02 4.592e-02 0.856 0.392334   
## factor(BsmtQual)2 -1.288e-01 3.798e-02 -3.391 0.000717 \*\*\*  
## factor(BsmtQual)3 -1.458e-01 3.362e-02 -4.336 1.55e-05 \*\*\*  
## factor(BsmtQual)4 -1.317e-01 3.492e-02 -3.773 0.000168 \*\*\*  
## factor(BsmtQual)5 -8.217e-02 3.844e-02 -2.138 0.032721 \*   
## factor(HeatingQC)2 5.554e-02 1.231e-01 0.451 0.651940   
## factor(HeatingQC)3 5.912e-02 1.222e-01 0.484 0.628523   
## factor(HeatingQC)4 7.557e-02 1.223e-01 0.618 0.536565   
## factor(HeatingQC)5 9.112e-02 1.222e-01 0.746 0.455892   
## log\_1stFlrSF 6.381e-02 3.793e-02 1.682 0.092789 .   
## `2ndFlrSF` 9.616e-05 2.304e-05 4.174 3.18e-05 \*\*\*  
## TotalBath 3.173e-02 5.980e-03 5.307 1.30e-07 \*\*\*  
## BedroomAbvGr -2.941e-03 5.762e-03 -0.510 0.609899   
## factor(KitchenQual)3 2.038e-02 2.182e-02 0.934 0.350480   
## factor(KitchenQual)4 3.726e-02 2.382e-02 1.564 0.118028   
## factor(KitchenQual)5 8.584e-02 2.874e-02 2.987 0.002871 \*\*   
## GarageCars 3.516e-02 1.030e-02 3.414 0.000658 \*\*\*  
## GarageArea 4.825e-05 3.544e-05 1.362 0.173527   
## SaleConditionAdjLand 1.505e-01 6.266e-02 2.402 0.016438 \*   
## SaleConditionAlloca 4.952e-02 3.835e-02 1.291 0.196827   
## SaleConditionFamily 3.382e-02 2.897e-02 1.167 0.243240   
## SaleConditionNormal 7.874e-02 1.249e-02 6.304 3.91e-10 \*\*\*  
## SaleConditionPartial 1.138e-01 1.816e-02 6.267 4.93e-10 \*\*\*  
## log\_Total\_BsmtFinSF 8.530e-03 1.329e-03 6.420 1.87e-10 \*\*\*  
## YrSold -3.868e-03 2.386e-03 -1.621 0.105248   
## Fireplaces 2.941e-02 6.227e-03 4.723 2.56e-06 \*\*\*  
## factor(RoofMatl)CompShg 1.541e+00 1.262e-01 12.216 < 2e-16 \*\*\*  
## factor(RoofMatl)Membran 1.544e+00 1.740e-01 8.874 < 2e-16 \*\*\*  
## factor(RoofMatl)Metal 1.566e+00 1.744e-01 8.978 < 2e-16 \*\*\*  
## factor(RoofMatl)Roll 1.520e+00 1.730e-01 8.784 < 2e-16 \*\*\*  
## factor(RoofMatl)Tar&Grv 1.522e+00 1.305e-01 11.669 < 2e-16 \*\*\*  
## factor(RoofMatl)WdShake 1.560e+00 1.370e-01 11.385 < 2e-16 \*\*\*  
## factor(RoofMatl)WdShngl 1.672e+00 1.333e-01 12.542 < 2e-16 \*\*\*  
## YearBuilt:YearRemodAdd -2.353e-05 8.447e-06 -2.786 0.005408 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.115 on 1363 degrees of freedom  
## Multiple R-squared: 0.9226, Adjusted R-squared: 0.9171   
## F-statistic: 169.2 on 96 and 1363 DF, p-value: < 2.2e-16

Residual standard error for Submission Model: 0.115 (0.12)

Multiple R-squared for Submission Model: 0.9226 (0.92)

## Validation

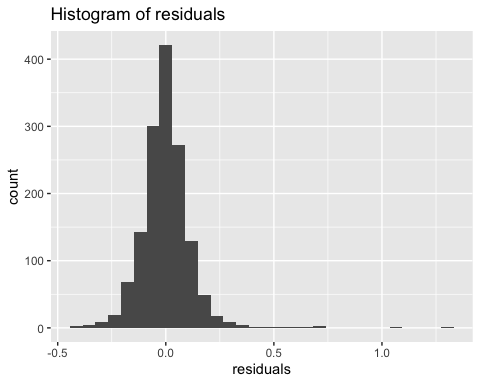
#extracting fitted values  
fitted\_values<-fitted(lm(log\_SalePrice ~ log\_LotArea + factor(LotConfig) + factor(Neighborhood) + log\_TotalArea + factor(BldgType) + factor(HouseStyle) + factor(OverallQual) + factor(OverallCond) + (YearBuilt \* YearRemodAdd) + factor(ExterQual) + factor(BsmtQual) + factor(HeatingQC) + log\_1stFlrSF + `2ndFlrSF` + TotalBath + BedroomAbvGr + factor(KitchenQual) + GarageCars + GarageArea + SaleCondition + log\_Total\_BsmtFinSF + YrSold + Fireplaces, data = train))   
residuals = fitted\_values-train$log\_SalePrice  
#residual plot   
pl25<-ggplot(train, aes(fitted\_values, residuals)) +  
 geom\_point() +  
 geom\_smooth(formula=y~x,method = "lm", se = F, col = "blue") +  
 labs(title = "Residuals for the Submission Model") +  
 theme(plot.title = element\_text(size=10))  
  
#fitted values vs observed values plot for log Sale Price  
pl26<- ggplot(train, aes(fitted\_values, log\_SalePrice)) +  
 geom\_point() +  
 geom\_smooth(formula=y~x,method = "lm", se = F, col = "blue")+  
 labs(title = "Plot between fitted and observed SalePrice Values") +  
 theme(plot.title = element\_text(size=10))  
  
margin = theme(plot.margin = unit(c(1,1,1,1), "cm"))  
grid.arrange(pl25, pl26, ncol = 2, nrow = 1, top=textGrob("Fitted value plots (Residulas, Observed)"))



From the residual plot of the Submission Model, we can observe that the residuals are randomly distributed around the summary line or residual line = 0. There is no visual pattern in residual distribution which satisfies the condition for selecting the correct regression technique for modeling and confirms a good fit for the data.

From the fitted vs observed plot of Sale Price values, we can observe that the linearity assumption condition of the data is satisfied and so, implemented linear regression method is justified. Data is accurately presented with a line that satisfies the conditions of a good model. We can see that most of the points are pretty close to the summary line,which confirms a good fit of data.

# Histogram of residuals  
ggplot(train, aes(residuals)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "Histogram of residuals")



Histogram of the residuals is used to verify normal distribtution of the variance. A close observation of the plot shows a symmetric bell-shaped histogram that is evenly distributed around zero indicates that the residuals are normally distributed and the assumption that variance is normally distributted is true. This also confirms that the implemented linear regression model is the right model for the data.

# Test Data Summary, Missing Value Imputation, and Data Transformation

## Data Summary & Missingness Assessment

# summary of test data set  
skim(test)

Data summary

|  |  |
| --- | --- |
| Name | test |
| Number of rows | 1459 |
| Number of columns | 80 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 43 |
| numeric | 37 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

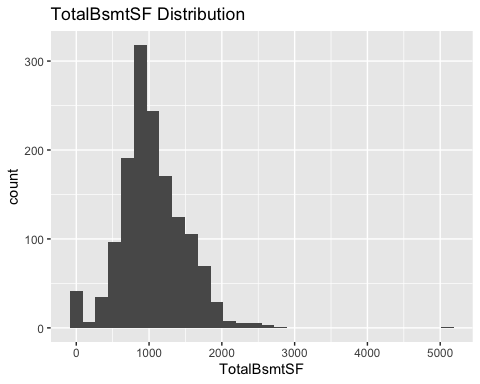
| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| MSZoning | 4 | 1.00 | 2 | 7 | 0 | 5 | 0 |
| Street | 0 | 1.00 | 4 | 4 | 0 | 2 | 0 |
| Alley | 1352 | 0.07 | 4 | 4 | 0 | 2 | 0 |
| LotShape | 0 | 1.00 | 3 | 3 | 0 | 4 | 0 |
| LandContour | 0 | 1.00 | 3 | 3 | 0 | 4 | 0 |
| Utilities | 2 | 1.00 | 6 | 6 | 0 | 1 | 0 |
| LotConfig | 0 | 1.00 | 3 | 7 | 0 | 5 | 0 |
| LandSlope | 0 | 1.00 | 3 | 3 | 0 | 3 | 0 |
| Neighborhood | 0 | 1.00 | 5 | 7 | 0 | 25 | 0 |
| Condition1 | 0 | 1.00 | 4 | 6 | 0 | 9 | 0 |
| Condition2 | 0 | 1.00 | 4 | 6 | 0 | 5 | 0 |
| BldgType | 0 | 1.00 | 4 | 6 | 0 | 5 | 0 |
| HouseStyle | 0 | 1.00 | 4 | 6 | 0 | 7 | 0 |
| RoofStyle | 0 | 1.00 | 3 | 7 | 0 | 6 | 0 |
| RoofMatl | 0 | 1.00 | 7 | 7 | 0 | 4 | 0 |
| Exterior1st | 1 | 1.00 | 6 | 7 | 0 | 13 | 0 |
| Exterior2nd | 1 | 1.00 | 5 | 7 | 0 | 15 | 0 |
| MasVnrType | 16 | 0.99 | 4 | 7 | 0 | 4 | 0 |
| ExterQual | 0 | 1.00 | 2 | 2 | 0 | 4 | 0 |
| ExterCond | 0 | 1.00 | 2 | 2 | 0 | 5 | 0 |
| Foundation | 0 | 1.00 | 4 | 6 | 0 | 6 | 0 |
| BsmtQual | 44 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtCond | 45 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtExposure | 44 | 0.97 | 2 | 2 | 0 | 4 | 0 |
| BsmtFinType1 | 42 | 0.97 | 3 | 3 | 0 | 6 | 0 |
| BsmtFinType2 | 42 | 0.97 | 3 | 3 | 0 | 6 | 0 |
| Heating | 0 | 1.00 | 4 | 4 | 0 | 4 | 0 |
| HeatingQC | 0 | 1.00 | 2 | 2 | 0 | 5 | 0 |
| CentralAir | 0 | 1.00 | 1 | 1 | 0 | 2 | 0 |
| Electrical | 0 | 1.00 | 5 | 5 | 0 | 4 | 0 |
| KitchenQual | 1 | 1.00 | 2 | 2 | 0 | 4 | 0 |
| Functional | 2 | 1.00 | 3 | 4 | 0 | 7 | 0 |
| FireplaceQu | 730 | 0.50 | 2 | 2 | 0 | 5 | 0 |
| GarageType | 76 | 0.95 | 6 | 7 | 0 | 6 | 0 |
| GarageFinish | 78 | 0.95 | 3 | 3 | 0 | 3 | 0 |
| GarageQual | 78 | 0.95 | 2 | 2 | 0 | 4 | 0 |
| GarageCond | 78 | 0.95 | 2 | 2 | 0 | 5 | 0 |
| PavedDrive | 0 | 1.00 | 1 | 1 | 0 | 3 | 0 |
| PoolQC | 1456 | 0.00 | 2 | 2 | 0 | 2 | 0 |
| Fence | 1169 | 0.20 | 4 | 5 | 0 | 4 | 0 |
| MiscFeature | 1408 | 0.03 | 4 | 4 | 0 | 3 | 0 |
| SaleType | 1 | 1.00 | 2 | 5 | 0 | 9 | 0 |
| SaleCondition | 0 | 1.00 | 6 | 7 | 0 | 6 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Id | 0 | 1.00 | 2190.00 | 421.32 | 1461 | 1825.50 | 2190.0 | 2554.50 | 2919 | ▇▇▇▇▇ |
| MSSubClass | 0 | 1.00 | 57.38 | 42.75 | 20 | 20.00 | 50.0 | 70.00 | 190 | ▇▅▂▁▁ |
| LotFrontage | 227 | 0.84 | 68.58 | 22.38 | 21 | 58.00 | 67.0 | 80.00 | 200 | ▃▇▁▁▁ |
| LotArea | 0 | 1.00 | 9819.16 | 4955.52 | 1470 | 7391.00 | 9399.0 | 11517.50 | 56600 | ▇▂▁▁▁ |
| OverallQual | 0 | 1.00 | 6.08 | 1.44 | 1 | 5.00 | 6.0 | 7.00 | 10 | ▁▁▇▅▁ |
| OverallCond | 0 | 1.00 | 5.55 | 1.11 | 1 | 5.00 | 5.0 | 6.00 | 9 | ▁▁▇▅▁ |
| YearBuilt | 0 | 1.00 | 1971.36 | 30.39 | 1879 | 1953.00 | 1973.0 | 2001.00 | 2010 | ▁▂▃▆▇ |
| YearRemodAdd | 0 | 1.00 | 1983.66 | 21.13 | 1950 | 1963.00 | 1992.0 | 2004.00 | 2010 | ▅▂▂▃▇ |
| MasVnrArea | 15 | 0.99 | 100.71 | 177.63 | 0 | 0.00 | 0.0 | 164.00 | 1290 | ▇▁▁▁▁ |
| BsmtFinSF1 | 1 | 1.00 | 439.20 | 455.27 | 0 | 0.00 | 350.5 | 753.50 | 4010 | ▇▂▁▁▁ |
| BsmtFinSF2 | 1 | 1.00 | 52.62 | 176.75 | 0 | 0.00 | 0.0 | 0.00 | 1526 | ▇▁▁▁▁ |
| BsmtUnfSF | 1 | 1.00 | 554.29 | 437.26 | 0 | 219.25 | 460.0 | 797.75 | 2140 | ▇▆▂▁▁ |
| TotalBsmtSF | 1 | 1.00 | 1046.12 | 442.90 | 0 | 784.00 | 988.0 | 1305.00 | 5095 | ▇▇▁▁▁ |
| 1stFlrSF | 0 | 1.00 | 1156.53 | 398.17 | 407 | 873.50 | 1079.0 | 1382.50 | 5095 | ▇▃▁▁▁ |
| 2ndFlrSF | 0 | 1.00 | 325.97 | 420.61 | 0 | 0.00 | 0.0 | 676.00 | 1862 | ▇▃▂▁▁ |
| LowQualFinSF | 0 | 1.00 | 3.54 | 44.04 | 0 | 0.00 | 0.0 | 0.00 | 1064 | ▇▁▁▁▁ |
| GrLivArea | 0 | 1.00 | 1486.05 | 485.57 | 407 | 1117.50 | 1432.0 | 1721.00 | 5095 | ▇▇▁▁▁ |
| BsmtFullBath | 2 | 1.00 | 0.43 | 0.53 | 0 | 0.00 | 0.0 | 1.00 | 3 | ▇▆▁▁▁ |
| BsmtHalfBath | 2 | 1.00 | 0.07 | 0.25 | 0 | 0.00 | 0.0 | 0.00 | 2 | ▇▁▁▁▁ |
| FullBath | 0 | 1.00 | 1.57 | 0.56 | 0 | 1.00 | 2.0 | 2.00 | 4 | ▁▇▇▁▁ |
| HalfBath | 0 | 1.00 | 0.38 | 0.50 | 0 | 0.00 | 0.0 | 1.00 | 2 | ▇▁▅▁▁ |
| BedroomAbvGr | 0 | 1.00 | 2.85 | 0.83 | 0 | 2.00 | 3.0 | 3.00 | 6 | ▁▃▇▂▁ |
| KitchenAbvGr | 0 | 1.00 | 1.04 | 0.21 | 0 | 1.00 | 1.0 | 1.00 | 2 | ▁▁▇▁▁ |
| TotRmsAbvGrd | 0 | 1.00 | 6.39 | 1.51 | 3 | 5.00 | 6.0 | 7.00 | 15 | ▅▇▃▁▁ |
| Fireplaces | 0 | 1.00 | 0.58 | 0.65 | 0 | 0.00 | 0.0 | 1.00 | 4 | ▇▇▁▁▁ |
| GarageYrBlt | 78 | 0.95 | 1977.72 | 26.43 | 1895 | 1959.00 | 1979.0 | 2002.00 | 2207 | ▂▇▁▁▁ |
| GarageCars | 1 | 1.00 | 1.77 | 0.78 | 0 | 1.00 | 2.0 | 2.00 | 5 | ▅▇▂▁▁ |
| GarageArea | 1 | 1.00 | 472.77 | 217.05 | 0 | 318.00 | 480.0 | 576.00 | 1488 | ▃▇▃▁▁ |
| WoodDeckSF | 0 | 1.00 | 93.17 | 127.74 | 0 | 0.00 | 0.0 | 168.00 | 1424 | ▇▁▁▁▁ |
| OpenPorchSF | 0 | 1.00 | 48.31 | 68.88 | 0 | 0.00 | 28.0 | 72.00 | 742 | ▇▁▁▁▁ |
| EnclosedPorch | 0 | 1.00 | 24.24 | 67.23 | 0 | 0.00 | 0.0 | 0.00 | 1012 | ▇▁▁▁▁ |
| 3SsnPorch | 0 | 1.00 | 1.79 | 20.21 | 0 | 0.00 | 0.0 | 0.00 | 360 | ▇▁▁▁▁ |
| ScreenPorch | 0 | 1.00 | 17.06 | 56.61 | 0 | 0.00 | 0.0 | 0.00 | 576 | ▇▁▁▁▁ |
| PoolArea | 0 | 1.00 | 1.74 | 30.49 | 0 | 0.00 | 0.0 | 0.00 | 800 | ▇▁▁▁▁ |
| MiscVal | 0 | 1.00 | 58.17 | 630.81 | 0 | 0.00 | 0.0 | 0.00 | 17000 | ▇▁▁▁▁ |
| MoSold | 0 | 1.00 | 6.10 | 2.72 | 1 | 4.00 | 6.0 | 8.00 | 12 | ▅▆▇▃▃ |
| YrSold | 0 | 1.00 | 2007.77 | 1.30 | 2006 | 2007.00 | 2008.0 | 2009.00 | 2010 | ▇▇▇▇▃ |

## Variables Distribution

## Total Basement SF distribution and summary for missing value imputation   
ggplot(test, aes(TotalBsmtSF)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "TotalBsmtSF Distribution")

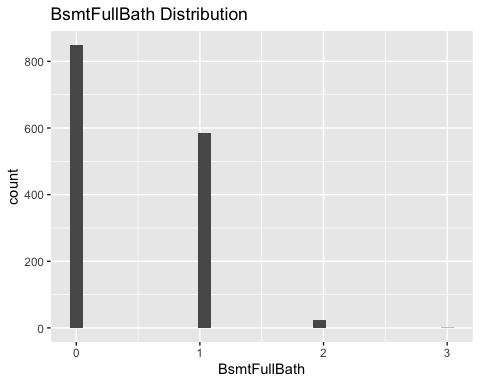


summary(test$TotalBsmtSF)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0 784 988 1046 1305 5095 1

As TotalBsmtSF is used to calculate the total area and predict the sale price in submission model,the same will be calculated to predict the sale price for test data. There is a missing value and it will be imputed using the mean of TotalbsmtSF as the above graph of TotalbsmtSF looks almost like a normal distribution with a couple of outliers.

## BsmtFullBath distribution and summary for missing value imputation   
ggplot(test, aes(BsmtFullBath)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "BsmtFullBath Distribution")

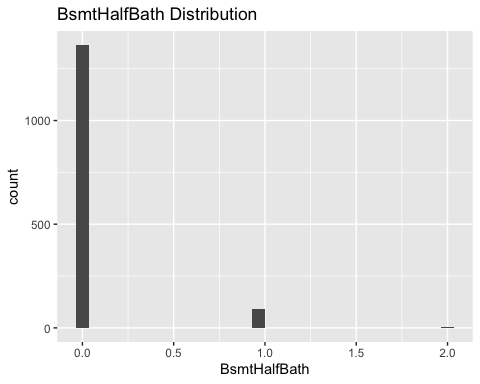


summary(test$BsmtFullBath)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0000 0.0000 0.0000 0.4345 1.0000 3.0000 2

BsmtFullBath is used to calculate the TotalBath variable and to predict the sale price in submission model and will be used to predict the sale price for test data. There is a missing value and it will be imputed using the median of BsmtFullBath as the above graph of BsmtFullBath is right skewed and majority is zero(0).

## BsmtHalfBath distribution and summary for missing value imputation   
ggplot(test, aes(BsmtHalfBath)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "BsmtHalfBath Distribution")

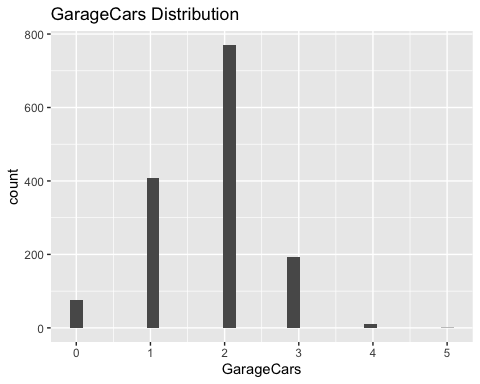


summary(test$BsmtHalfBath)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0000 0.0000 0.0000 0.0652 0.0000 2.0000 2

BsmtHalfBath is used to calculate the TotalBath variable and to predict the sale price in submission model and will be used to predict the sale price for test data. There is a missing value and it will be imputed using the median of BsmtHalfBath as the above graph of BsmtHalfBath is right skewed and majority is zero(0).

## GarageCars distribution and summary for missing value imputation   
ggplot(test, aes(GarageCars)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "GarageCars Distribution")

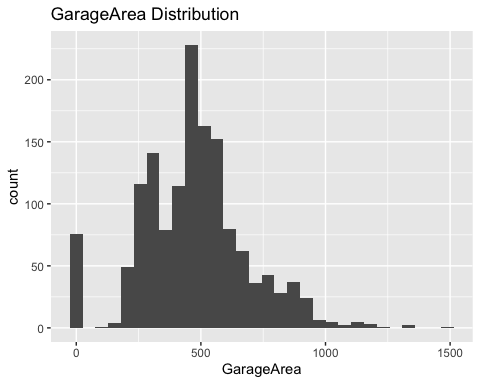


summary(test$GarageCars)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.000 1.000 2.000 1.766 2.000 5.000 1

GarageCars is used to predict the sale price in submission model and will be used to predict the sale price for test data. There is a missing value and it will be imputed using the mean of GarageCars as the above graph of GarageCars looks almost like a normal distribution.

## GarageArea distribution and summary for missing value imputation   
ggplot(test, aes(GarageArea)) +  
 geom\_histogram(bins = 30) +  
 labs(title = "GarageArea Distribution")



summary(test$GarageArea)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0 318.0 480.0 472.8 576.0 1488.0 1

GarageArea is used to predict the sale price in submission model and will be used to predict the sale price for test data. There is a missing value and it will be imputed using the median of GarageCars as the above graph of GarageArea is right skewed.

## Data Transformation & Missing Value Imputation

#Data Transformation & Missing Value Imputation  
test<- test %>%   
 mutate(TotalBsmtSF = ifelse(is.na(TotalBsmtSF), mean(TotalBsmtSF, na.rm=TRUE), TotalBsmtSF),  
 log\_TotalArea = log(TotalBsmtSF + GrLivArea),  
 log\_LotArea = log(LotArea),  
 log\_1stFlrSF = log(`1stFlrSF`),  
 BsmtFullBath = ifelse(is.na(BsmtFullBath), median(BsmtFullBath, na.rm=TRUE), BsmtFullBath),  
 BsmtHalfBath = ifelse(is.na(BsmtHalfBath), median(BsmtHalfBath, na.rm=TRUE), BsmtHalfBath),  
 TotalBath = (FullBath + HalfBath + BsmtFullBath + BsmtHalfBath),  
 GarageCars = ifelse(is.na(GarageCars), mean(GarageCars, na.rm=TRUE), GarageCars),  
 GarageArea = ifelse(is.na(GarageArea), median(GarageArea, na.rm=TRUE), GarageArea),  
 ExterQual= case\_when(ExterQual == "Ex" ~ 5,  
 ExterQual == "Gd" ~ 4,  
 ExterQual == "TA" ~ 3,  
 ExterQual == "Fa" ~ 2,  
 ExterQual == "Po" ~ 1),  
 ExterQual = ifelse(is.na(ExterQual), 0, ExterQual),  
 BsmtQual = case\_when(BsmtQual == "Ex" ~ 5,  
 BsmtQual == "Gd" ~ 4,  
 BsmtQual == "TA" ~ 3,  
 BsmtQual == "Fa" ~ 2,  
 BsmtQual == "Po" ~ 1),  
 BsmtFinSF1 = ifelse(is.na(BsmtFinSF1), median(BsmtFinSF1, na.rm=TRUE), BsmtFinSF1),  
 BsmtFinSF2 = ifelse(is.na(BsmtFinSF2), median(BsmtFinSF2, na.rm=TRUE), BsmtFinSF2),  
 log\_Total\_BsmtFinSF = log(BsmtFinSF1 + BsmtFinSF2 + 1),  
 BsmtQual = ifelse(is.na(BsmtQual), 0, BsmtQual),  
 HeatingQC= case\_when(HeatingQC == "Ex" ~ 5,  
 HeatingQC == "Gd" ~ 4,  
 HeatingQC == "TA" ~ 3,  
 HeatingQC == "Fa" ~ 2,  
 HeatingQC == "Po" ~ 1),  
 HeatingQC = ifelse(is.na(HeatingQC), 0, HeatingQC),  
 KitchenQual = case\_when(KitchenQual == "Ex" ~ 5,  
 KitchenQual == "Gd" ~ 4,  
 KitchenQual == "TA" ~ 3,  
 KitchenQual == "Fa" ~ 2,  
 KitchenQual == "Po" ~ 1),  
 KitchenQual = ifelse(is.na(KitchenQual), 3, KitchenQual), # median value imputed  
 Exterior1st = ifelse(is.na(Exterior1st), "VinylSd", Exterior1st))   
#sapply(test, function(x) sum(is.na(x)))

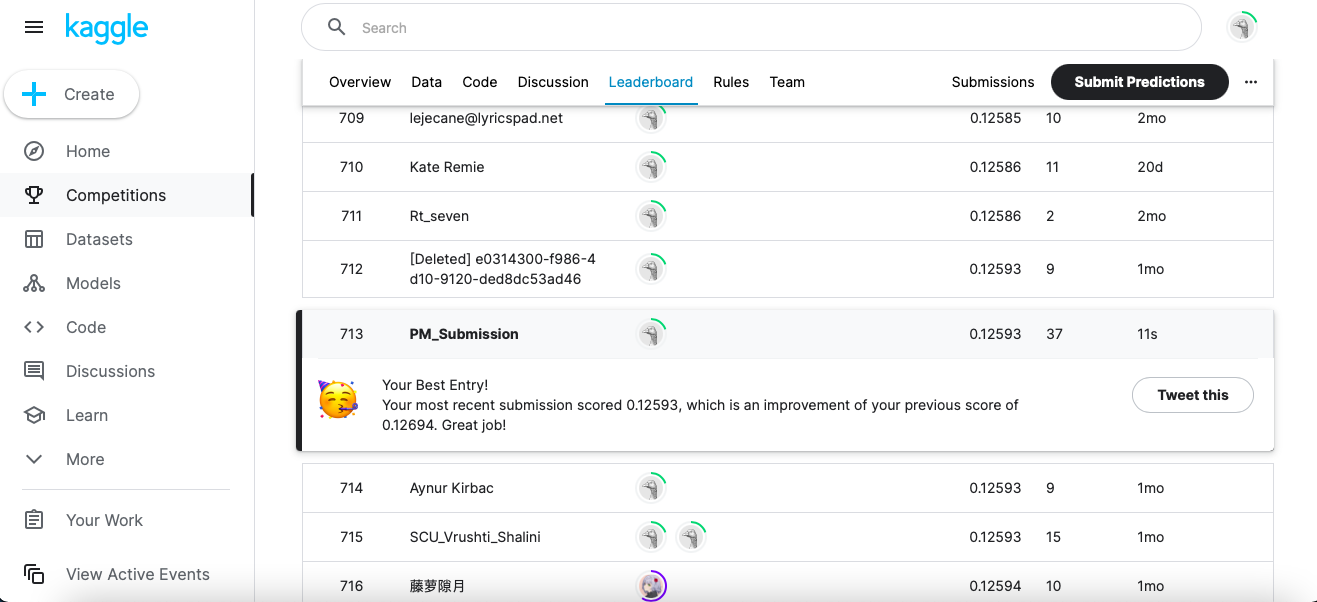
Missing values are imputed for the variables that are used in the submission model and will be used to predict the sale price in the test data. Also, all other data transformations that are implemented in the training data set are executed in the test data set. Total Area (TotalBsmtSF + GrLivArea), Total\_BsmtFinSF = (BsmtFinSF1 + BsmtFinSF2), and TotalBath = (HalfBath + FullBath + BsmtFullBath + BsmtHalfBath) are created. TotalArea, LotArea, Total\_BsmtFinSF, and 1stFlrSF are log transformed to make sure these variables distribution looks more like a normal distribution. Qualitative variables ExterQual, BsmtQual, HeatingQC, and KitchenQual are converted from category to ordinal variable to better predict the sale price. Also, if any qualitative variable value is missing it is assigned as zero.

# Predicting SalePrice For Test Data

# predicting sale price  
SalePrice<-predict(submission\_model, test)  
SalePrice<-exp(SalePrice)  
Id<-test$Id  
submission<-(cbind(Id, SalePrice))  
write.csv(submission, "submission.csv", row.names = F)

# Kaggle Submission Report

knitr::include\_graphics("kaggle\_submission.png")



Kaggle score or log RMSE: 0.12593 (0.13)

Kaggle rank: 713

# Contributors & Contributions

contribution<-"  
| Contributors | Contributions |  
| ---------------------- | ------------------------------- |  
| Prabhudatta Mohapatra | EDA, Data Wrangling, Modeling |  
| Enni Su | EDA, Data Wrangling, Modeling |  
| Hunter Conrad | EDA, Data Wrangling, Modeling |  
| Shreya Chawla | EDA, Data Wrangling, Modeling |  
"  
cat(contribution)

##   
## | Contributors | Contributions |  
## | ---------------------- | ------------------------------- |  
## | Prabhudatta Mohapatra | EDA, Data Wrangling, Modeling |  
## | Enni Su | EDA, Data Wrangling, Modeling |  
## | Hunter Conrad | EDA, Data Wrangling, Modeling |  
## | Shreya Chawla | EDA, Data Wrangling, Modeling |