House Prices Analysis and Prediction

Moaz M. El-Essawey

April 2022

Loading Packages and dataset

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
train_df <- read_csv("./data/train.csv")
```

```
## Rows: 1460 Columns: 81
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (43): MSZoning, Street, Alley, LotShape, LandContour, Utilities, LotConf...
## dbl (38): Id, MSSubClass, LotFrontage, LotArea, OverallQual, OverallCond, Ye...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(train_df, 5)
```

```
## # A tibble: 5 x 81
##      Id MSSubClass MSZoning LotFrontage LotArea Street Alley LotShape
##   <dbl>      <dbl> <chr>          <dbl>   <dbl> <chr>  <chr> <chr>
## 1     1         60 RL                65    8450 Pave   <NA>  Reg
## 2     2         20 RL                80    9600 Pave   <NA>  Reg
## 3     3         60 RL                68   11250 Pave   <NA>  IR1
## 4     4         70 RL                60    9550 Pave   <NA>  IR1
## 5     5         60 RL                84   14260 Pave   <NA>  IR1
## # ... with 73 more variables: LandContour <chr>, Utilities <chr>,
## #   LotConfig <chr>, LandSlope <chr>, Neighborhood <chr>, Condition1 <chr>,
## #   Condition2 <chr>, BldgType <chr>, HouseStyle <chr>, OverallQual <dbl>,
## #   OverallCond <dbl>, YearBuilt <dbl>, YearRemodAdd <dbl>, RoofStyle <chr>,
## #   RoofMatl <chr>, Exterior1st <chr>, Exterior2nd <chr>, MasVnrType <chr>,
## #   MasVnrArea <dbl>, ExterQual <chr>, ExterCond <chr>, Foundation <chr>,
## #   BsmtQual <chr>, BsmtCond <chr>, BsmtExposure <chr>, BsmtFinType1 <chr>, ...
```

```
glimpse(train_df)
```

```
## Rows: 1,460
## Columns: 81
## $ Id            <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1~
## $ MSSubClass    <dbl> 60, 20, 60, 70, 60, 50, 20, 60, 50, 190, 20, 60, 20, 20,~
## $ MSZoning      <chr> "RL", "RL", "RL", "RL", "RL", "RL", "RL", "RL", "RM", "R~
## $ LotFrontage   <dbl> 65, 80, 68, 60, 84, 85, 75, NA, 51, 50, 70, 85, NA, 91, ~
## $ LotArea       <dbl> 8450, 9600, 11250, 9550, 14260, 14115, 10084, 10382, 612~
## $ Street        <chr> "Pave", "Pave", "Pave", "Pave", "Pave", "Pave", "Pave", ~
## $ Alley         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ LotShape      <chr> "Reg", "Reg", "IR1", "IR1", "IR1", "IR1", "Reg", "IR1", ~
## $ LandContour   <chr> "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", "Lvl", ~
## $ Utilities     <chr> "AllPub", "AllPub", "AllPub", "AllPub", "AllPub", "AllPu~
## $ LotConfig     <chr> "Inside", "FR2", "Inside", "Corner", "FR2", "Inside", "I~
## $ LandSlope     <chr> "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", "Gtl", ~
## $ Neighborhood  <chr> "CollgCr", "Veenker", "CollgCr", "Crawfor", "NoRidge", "~
## $ Condition1    <chr> "Norm", "Feedr", "Norm", "Norm", "Norm", "Norm", "Norm",~
## $ Condition2    <chr> "Norm", "Norm", "Norm", "Norm", "Norm", "Norm", "Norm", ~
## $ BldgType      <chr> "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", "1Fam", ~
## $ HouseStyle    <chr> "2Story", "1Story", "2Story", "2Story", "2Story", "1.5Fi~
## $ OverallQual   <dbl> 7, 6, 7, 7, 8, 5, 8, 7, 7, 5, 5, 9, 5, 7, 6, 7, 6, 4, 5,~
## $ OverallCond   <dbl> 5, 8, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 6, 5, 5, 8, 7, 5, 5,~
## $ YearBuilt     <dbl> 2003, 1976, 2001, 1915, 2000, 1993, 2004, 1973, 1931, 19~
## $ YearRemodAdd  <dbl> 2003, 1976, 2002, 1970, 2000, 1995, 2005, 1973, 1950, 19~
## $ RoofStyle     <chr> "Gable", "Gable", "Gable", "Gable", "Gable", "Gable", "G~
## $ RoofMatl      <chr> "CompShg", "CompShg", "CompShg", "CompShg", "CompShg", "~
## $ Exterior1st   <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Sdng", "VinylSd", "~
## $ Exterior2nd   <chr> "VinylSd", "MetalSd", "VinylSd", "Wd Shng", "VinylSd", "~
## $ MasVnrType    <chr> "BrkFace", "None", "BrkFace", "None", "BrkFace", "None",~
## $ MasVnrArea    <dbl> 196, 0, 162, 0, 350, 0, 186, 240, 0, 0, 0, 286, 0, 306, ~
## $ ExterQual     <chr> "Gd", "TA", "Gd", "TA", "Gd", "TA", "Gd", "TA", "TA", "T~
## $ ExterCond     <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "T~
## $ Foundation    <chr> "PConc", "CBlock", "PConc", "BrkTil", "PConc", "Wood", "~
## $ BsmtQual      <chr> "Gd", "Gd", "Gd", "TA", "Gd", "Gd", "Ex", "Gd", "TA", "T~
## $ BsmtCond      <chr> "TA", "TA", "TA", "Gd", "TA", "TA", "TA", "TA", "TA", "T~
## $ BsmtExposure  <chr> "No", "Gd", "Mn", "No", "Av", "No", "Av", "Mn", "No", "N~
## $ BsmtFinType1  <chr> "GLQ", "ALQ", "GLQ", "ALQ", "GLQ", "GLQ", "GLQ", "ALQ", ~
## $ BsmtFinSF1    <dbl> 706, 978, 486, 216, 655, 732, 1369, 859, 0, 851, 906, 99~
## $ BsmtFinType2  <chr> "Unf", "Unf", "Unf", "Unf", "Unf", "Unf", "Unf", "BLQ", ~
## $ BsmtFinSF2    <dbl> 0, 0, 0, 0, 0, 0, 0, 32, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ BsmtUnfSF     <dbl> 150, 284, 434, 540, 490, 64, 317, 216, 952, 140, 134, 17~
## $ TotalBsmtSF   <dbl> 856, 1262, 920, 756, 1145, 796, 1686, 1107, 952, 991, 10~
## $ Heating       <chr> "GasA", "GasA", "GasA", "GasA", "GasA", "GasA", "GasA", ~
## $ HeatingQC     <chr> "Ex", "Ex", "Ex", "Gd", "Ex", "Ex", "Ex", "Ex", "Gd", "E~
## $ CentralAir    <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "~
## $ Electrical    <chr> "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr", "SBrkr", "S~
## $ `1stFlrSF`    <dbl> 856, 1262, 920, 961, 1145, 796, 1694, 1107, 1022, 1077, ~
## $ `2ndFlrSF`    <dbl> 854, 0, 866, 756, 1053, 566, 0, 983, 752, 0, 0, 1142, 0,~
## $ LowQualFinSF  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ GrLivArea     <dbl> 1710, 1262, 1786, 1717, 2198, 1362, 1694, 2090, 1774, 10~
## $ BsmtFullBath  <dbl> 1, 0, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 0, 1, 0, 1, 0, 1,~
## $ BsmtHalfBath  <dbl> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ FullBath      <dbl> 2, 2, 2, 1, 2, 1, 2, 2, 2, 1, 1, 3, 1, 2, 1, 1, 1, 2, 1,~
```

```
## $ HalfBath      <dbl> 1, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1,~
## $ BedroomAbvGr  <dbl> 3, 3, 3, 3, 4, 1, 3, 3, 2, 2, 3, 4, 2, 3, 2, 2, 2, 2, 3,~
## $ KitchenAbvGr  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 1, 2, 1,~
## $ KitchenQual   <chr> "Gd", "TA", "Gd", "Gd", "Gd", "TA", "Gd", "TA", "TA", "T~
## $ TotRmsAbvGrd  <dbl> 8, 6, 6, 7, 9, 5, 7, 7, 8, 5, 5, 11, 4, 7, 5, 5, 5, 6, 6~
## $ Functional    <chr> "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", "Typ", ~
## $ Fireplaces    <dbl> 0, 1, 1, 1, 1, 0, 1, 2, 2, 2, 0, 2, 0, 1, 1, 0, 1, 0, 0,~
## $ FireplaceQu   <chr> NA, "TA", "TA", "Gd", "TA", NA, "Gd", "TA", "TA", "TA", ~
## $ GarageType    <chr> "Attchd", "Attchd", "Attchd", "Detchd", "Attchd", "Attch~
## $ GarageYrBlt   <dbl> 2003, 1976, 2001, 1998, 2000, 1993, 2004, 1973, 1931, 19~
## $ GarageFinish  <chr> "RFn", "RFn", "RFn", "Unf", "RFn", "Unf", "RFn", "RFn", ~
## $ GarageCars    <dbl> 2, 2, 2, 3, 3, 2, 2, 2, 2, 1, 1, 3, 1, 3, 1, 2, 2, 2, 2,~
## $ GarageArea    <dbl> 548, 460, 608, 642, 836, 480, 636, 484, 468, 205, 384, 7~
## $ GarageQual    <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "Fa", "G~
## $ GarageCond    <chr> "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "TA", "T~
## $ PavedDrive    <chr> "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "Y", "~
## $ WoodDeckSF    <dbl> 0, 298, 0, 0, 192, 40, 255, 235, 90, 0, 0, 147, 140, 160~
## $ OpenPorchSF   <dbl> 61, 0, 42, 35, 84, 30, 57, 204, 0, 4, 0, 21, 0, 33, 213,~
## $ EnclosedPorch <dbl> 0, 0, 0, 272, 0, 0, 0, 228, 205, 0, 0, 0, 0, 0, 176, 0, ~
## $ `3SsnPorch`   <dbl> 0, 0, 0, 0, 0, 320, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
## $ ScreenPorch   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 176, 0, 0, 0, 0, ~
## $ PoolArea      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ PoolQC        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ Fence         <chr> NA, NA, NA, NA, NA, "MnPrv", NA, NA, NA, NA, NA, NA, NA,~
## $ MiscFeature   <chr> NA, NA, NA, NA, NA, "Shed", NA, "Shed", NA, NA, NA, NA, ~
## $ MiscVal       <dbl> 0, 0, 0, 0, 0, 700, 0, 350, 0, 0, 0, 0, 0, 0, 0, 0, 700,~
## $ MoSold        <dbl> 2, 5, 9, 2, 12, 10, 8, 11, 4, 1, 2, 7, 9, 8, 5, 7, 3, 10~
## $ YrSold        <dbl> 2008, 2007, 2008, 2006, 2008, 2009, 2007, 2009, 2008, 20~
## $ SaleType      <chr> "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", "WD", "W~
## $ SaleCondition <chr> "Normal", "Normal", "Normal", "Abnorml", "Normal", "Norm~
## $ SalePrice     <dbl> 208500, 181500, 223500, 140000, 250000, 143000, 307000, ~
```

Wranlge train Dataset

Dropping NaN values

I will drop any column contains more than 200 NaN value in it. and drop any non numeric column containing any NaN value. and fill any numeric column with NaN values less than 200 with it's mean.

```
nan_cols <- train_df %>% is.na() %>% colSums()
nan_cols <- nan_cols[nan_cols > 0]
nan_cols
```

```
##  LotFrontage        Alley   MasVnrType   MasVnrArea     BsmtQual     BsmtCond
##          259         1369            8            8           37           37
## BsmtExposure BsmtFinType1 BsmtFinType2   Electrical  FireplaceQu   GarageType
##           38           37           38            1          690           81
##  GarageYrBlt GarageFinish   GarageQual   GarageCond       PoolQC        Fence
##           81           81           81           81         1453         1179
##  MiscFeature
##         1406
```

```
nan_cols_to_drop <- nan_cols[nan_cols >= 200]
train_df <- train_df %>% select(-names(nan_cols_to_drop))
```

```r
nan_cols <- train_df %>% is.na() %>% colSums()
nan_cols <- nan_cols[nan_cols > 0]
nan_cols
```

```
##   MasVnrType   MasVnrArea     BsmtQual     BsmtCond BsmtExposure BsmtFinType1
##            8            8           37           37           38           37
## BsmtFinType2   Electrical   GarageType  GarageYrBlt GarageFinish   GarageQual
##           38            1           81           81           81           81
##   GarageCond
##           81
```

```r
non_numeric_nan_cols <- colnames(train_df %>% select(names(nan_cols)) %>%  select(where(is.character)))
non_numeric_nan_cols
```

```
##  [1] "MasVnrType"   "BsmtQual"     "BsmtCond"     "BsmtExposure" "BsmtFinType1"
##  [6] "BsmtFinType2" "Electrical"   "GarageType"   "GarageFinish" "GarageQual"
## [11] "GarageCond"
```

```r
train_df <- train_df %>% select(-non_numeric_nan_cols)
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(non_numeric_nan_cols)` instead of `non_numeric_nan_cols` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```r
train_df %>% head(5)
```

```
## # A tibble: 5 x 64
##      Id MSSubClass MSZoning LotArea Street LotShape LandContour Utilities
##   <dbl>      <dbl> <chr>      <dbl> <chr>  <chr>    <chr>       <chr>
## 1     1         60 RL          8450 Pave   Reg      Lvl         AllPub
## 2     2         20 RL          9600 Pave   Reg      Lvl         AllPub
## 3     3         60 RL         11250 Pave   IR1      Lvl         AllPub
## 4     4         70 RL          9550 Pave   IR1      Lvl         AllPub
## 5     5         60 RL         14260 Pave   IR1      Lvl         AllPub
## # ... with 56 more variables: LotConfig <chr>, LandSlope <chr>,
## #   Neighborhood <chr>, Condition1 <chr>, Condition2 <chr>, BldgType <chr>,
## #   HouseStyle <chr>, OverallQual <dbl>, OverallCond <dbl>, YearBuilt <dbl>,
## #   YearRemodAdd <dbl>, RoofStyle <chr>, RoofMatl <chr>, Exterior1st <chr>,
## #   Exterior2nd <chr>, MasVnrArea <dbl>, ExterQual <chr>, ExterCond <chr>,
## #   Foundation <chr>, BsmtFinSF1 <dbl>, BsmtFinSF2 <dbl>, BsmtUnfSF <dbl>,
## #   TotalBsmtSF <dbl>, Heating <chr>, HeatingQC <chr>, CentralAir <chr>, ...
```

```r
nan_cols <- train_df %>% is.na() %>% colSums()
nan_cols[nan_cols > 0]
```

```
##  MasVnrArea GarageYrBlt
##           8          81
```

```r
train_df <- train_df %>% replace_na(
  list(MasVnrArea = mean(train_df$MasVnrArea, na.rm = TRUE),
       GarageYrBlt = mean(train_df$GarageYrBlt, na.rm = TRUE))
)

nan_cols <- train_df %>% is.na() %>% colSums()
nan_cols[nan_cols > 0]
```

```
## named numeric(0)
```

Now our data frame does not contain any **NA** values. we can start working with and inspecting it's columns in details.

Inspecting Categorical Columns

```
categorical_cols <- colnames(train_df %>% select(where(is.character)))
numerical_cols <- colnames(train_df %>% select(where(is.numeric)))

print("Categorical Columns")
```

```
## [1] "Categorical Columns"
```

```
print(categorical_cols)
```

```
##  [1] "MSZoning"      "Street"        "LotShape"      "LandContour"
##  [5] "Utilities"     "LotConfig"     "LandSlope"     "Neighborhood"
##  [9] "Condition1"    "Condition2"    "BldgType"      "HouseStyle"
## [13] "RoofStyle"     "RoofMatl"      "Exterior1st"   "Exterior2nd"
## [17] "ExterQual"     "ExterCond"     "Foundation"    "Heating"
## [21] "HeatingQC"     "CentralAir"    "KitchenQual"   "Functional"
## [25] "PavedDrive"    "SaleType"      "SaleCondition"
```

```
print("Numerical Columns")
```

```
## [1] "Numerical Columns"
```

```
print(numerical_cols)
```

```
##  [1] "Id"            "MSSubClass"    "LotArea"       "OverallQual"
##  [5] "OverallCond"   "YearBuilt"     "YearRemodAdd"  "MasVnrArea"
##  [9] "BsmtFinSF1"    "BsmtFinSF2"    "BsmtUnfSF"     "TotalBsmtSF"
## [13] "1stFlrSF"      "2ndFlrSF"      "LowQualFinSF"  "GrLivArea"
## [17] "BsmtFullBath"  "BsmtHalfBath"  "FullBath"      "HalfBath"
## [21] "BedroomAbvGr"  "KitchenAbvGr"  "TotRmsAbvGrd"  "Fireplaces"
## [25] "GarageYrBlt"   "GarageCars"    "GarageArea"    "WoodDeckSF"
## [29] "OpenPorchSF"   "EnclosedPorch" "3SsnPorch"     "ScreenPorch"
## [33] "PoolArea"      "MiscVal"       "MoSold"        "YrSold"
## [37] "SalePrice"
```

MSZoning Columns Analysis

```
train_df %>% group_by(MSZoning) %>% summarise(counts=n())  %>% arrange(-counts)
```

```
## # A tibble: 5 x 2
##   MSZoning counts
##   <chr>     <int>
## 1 RL         1151
## 2 RM          218
## 3 FV           65
## 4 RH           16
## 5 C (all)      10
```

MSZoning has a hig bias twards the **Residential Low Density** and **Residential Medium Density** zoning types.

HouseStyle Column Analysis

```
train_df %>% group_by(HouseStyle) %>% summarise(counts=n())  %>% arrange(-counts)
```

```
## # A tibble: 8 x 2
##   HouseStyle counts
##   <chr>       <int>
## 1 1Story        726
## 2 2Story        445
## 3 1.5Fin        154
## 4 SLvl           65
## 5 SFoyer         37
## 6 1.5Unf         14
## 7 2.5Unf         11
## 8 2.5Fin          8
```

HouseStyle has major proportions twards **One Story, Two Story** and **One and one-half story**: *2nd level unfinished*

SaleCondition Columns Analysis

```
train_df %>% group_by(SaleCondition) %>% summarise(counts=n()) %>% arrange(-counts)
```

```
## # A tibble: 6 x 2
##   SaleCondition counts
##   <chr>          <int>
## 1 Normal          1198
## 2 Partial          125
## 3 Abnorml          101
## 4 Family            20
## 5 Alloca            12
## 6 AdjLand            4
```

A House with **Normal** condition is dominated in selling condition.

Cleaning Categorical Columns

```
train_df <- train_df %>% mutate(
  low_density_zone = as.numeric(MSZoning == "RL"),

  one_story_type = as.numeric(HouseStyle == "1Story"),
  two_story_type = as.numeric(HouseStyle == "2Story"),
  half_story_type = as.numeric(HouseStyle == "1.5Fin"),

  normal_sale_cond = as.numeric(SaleCondition == "Normal"),
  gas_heating_sys = as.numeric(Heating == "GasA"),
  )

train_df %>% select(
  low_density_zone, one_story_type,two_story_type, half_story_type,
  normal_sale_cond, gas_heating_sys
) %>% head(5)
```

```
## # A tibble: 5 x 6
##   low_density_zone one_story_type two_story_type half_story_type
```

```
##              <dbl>          <dbl>          <dbl>          <dbl>
## 1                1              0              1              0
## 2                1              1              0              0
## 3                1              0              1              0
## 4                1              0              1              0
## 5                1              0              1              0
## # ... with 2 more variables: normal_sale_cond <dbl>, gas_heating_sys <dbl>
```

Inscpecting Numerical Columns

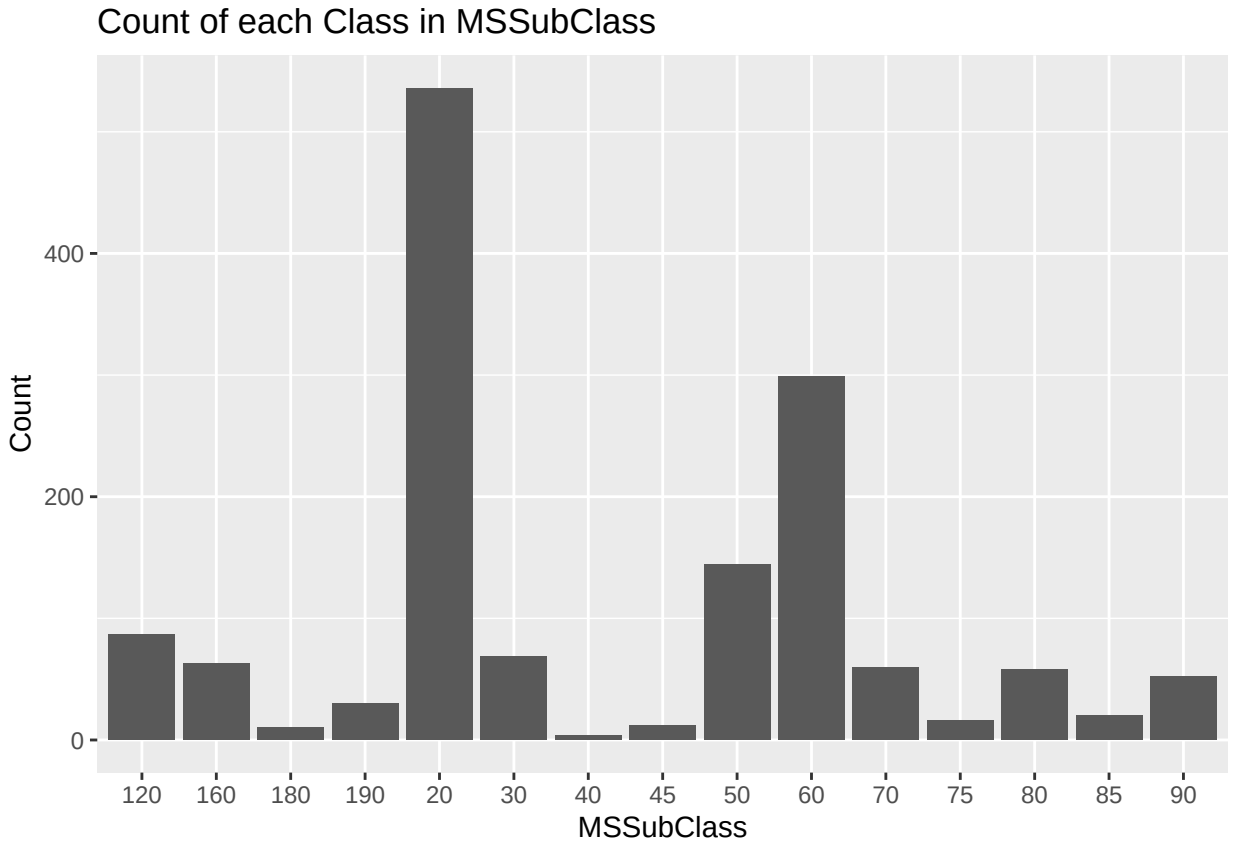
```
numerical_cols
```

```
##  [1] "Id"            "MSSubClass"    "LotArea"       "OverallQual"
##  [5] "OverallCond"   "YearBuilt"     "YearRemodAdd"  "MasVnrArea"
##  [9] "BsmtFinSF1"    "BsmtFinSF2"    "BsmtUnfSF"     "TotalBsmtSF"
## [13] "1stFlrSF"      "2ndFlrSF"      "LowQualFinSF"  "GrLivArea"
## [17] "BsmtFullBath"  "BsmtHalfBath"  "FullBath"      "HalfBath"
## [21] "BedroomAbvGr"  "KitchenAbvGr"  "TotRmsAbvGrd"  "Fireplaces"
## [25] "GarageYrBlt"   "GarageCars"    "GarageArea"    "WoodDeckSF"
## [29] "OpenPorchSF"   "EnclosedPorch" "3SsnPorch"     "ScreenPorch"
## [33] "PoolArea"      "MiscVal"       "MoSold"        "YrSold"
## [37] "SalePrice"
```

MSSubClass Column Analysis

```
## convert MSSubClass to categorical column
train_df <- train_df %>% mutate(MSSubClass = as.character(MSSubClass))

train_df %>% group_by(MSSubClass) %>%
  summarise(counts=n()) %>% arrange(-counts) %>%
  ggplot(aes(x=MSSubClass, y=counts))+
    geom_col()+
    labs(
      x="MSSubClass",
      y="Count",
      title="Count of each Class in MSSubClass"
    )
```



most people prefer those types of houses **20** : **1-STORY 1946 & NEWER ALL STYLES 60** : **2-STORY 1946 & NEWER**

YearBuild, YearRemodedAdd and YrSold Columns Analysis

```
train_df %>% mutate(
    time_taken_to_remodel = YearRemodAdd - YearBuilt,
    time_taken_to_sold = YrSold - YearBuilt,
    time_taken_to_sell_after_remodel = YrSold - YearRemodAdd) %>%

  select(time_taken_to_remodel, time_taken_to_sold, time_taken_to_sell_after_remodel) %>% summary()
```

```
##  time_taken_to_remodel time_taken_to_sold time_taken_to_sell_after_remodel
##  Min.   :  0.0         Min.   :  0.00     Min.   :-1.00
##  1st Qu.:  0.0         1st Qu.:  8.00     1st Qu.: 4.00
##  Median :  0.0         Median : 35.00     Median :14.00
##  Mean   : 13.6         Mean   : 36.55     Mean   :22.95
##  3rd Qu.: 20.0         3rd Qu.: 54.00     3rd Qu.:41.00
##  Max.   :123.0         Max.   :136.00     Max.   :60.00
```

on average - it took around **13 years** to remodel a house. - it took around **36 years** to sell a house. - it took around **22 years** to sell a house that is remodeled.

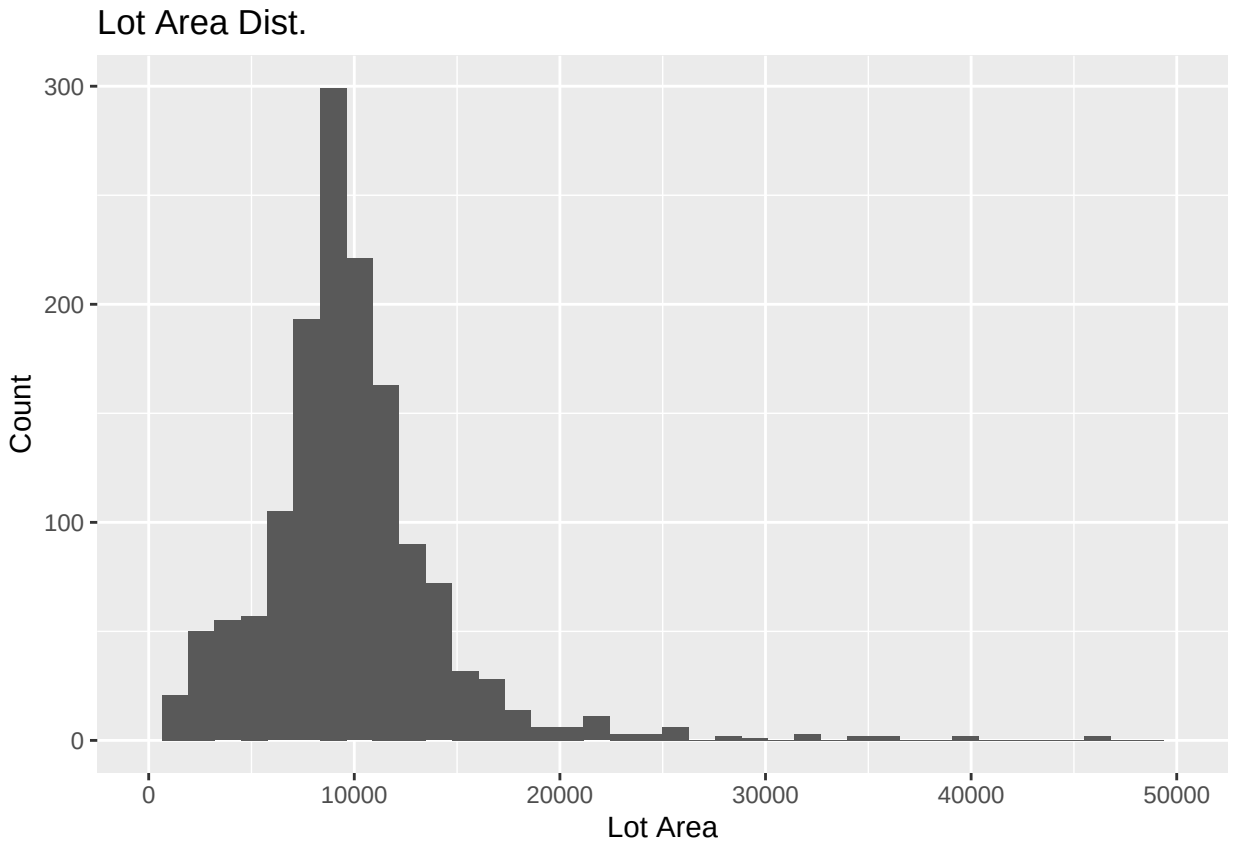
LotArea Column Analysis

```
print(train_df %>% select(LotArea) %>% summary)
```

```
##     LotArea
```

```
##  Min.   :  1300
##  1st Qu.:  7554
##  Median :  9478
##  Mean   : 10517
##  3rd Qu.: 11602
##  Max.   :215245
```

```
train_df %>% ggplot(aes(x=LotArea))+
  geom_histogram(bins=40)+xlim(0, 50000)+
  labs(
    x="Lot Area",
    y="Count",
    title="Lot Area Dist."
  )
```



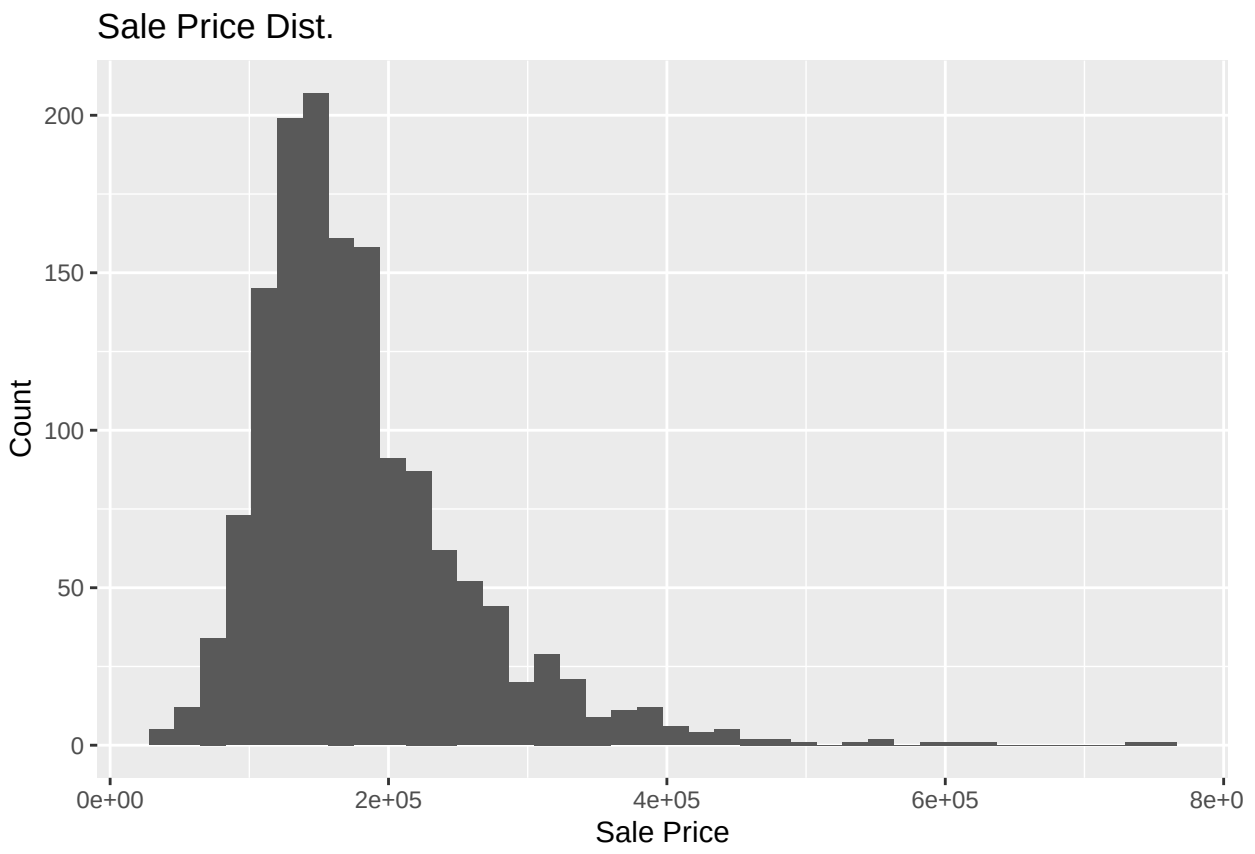
Most of the houses has a **Lot Area** around **5,000 to 15,000** square feet

SalePrice Column Analysis

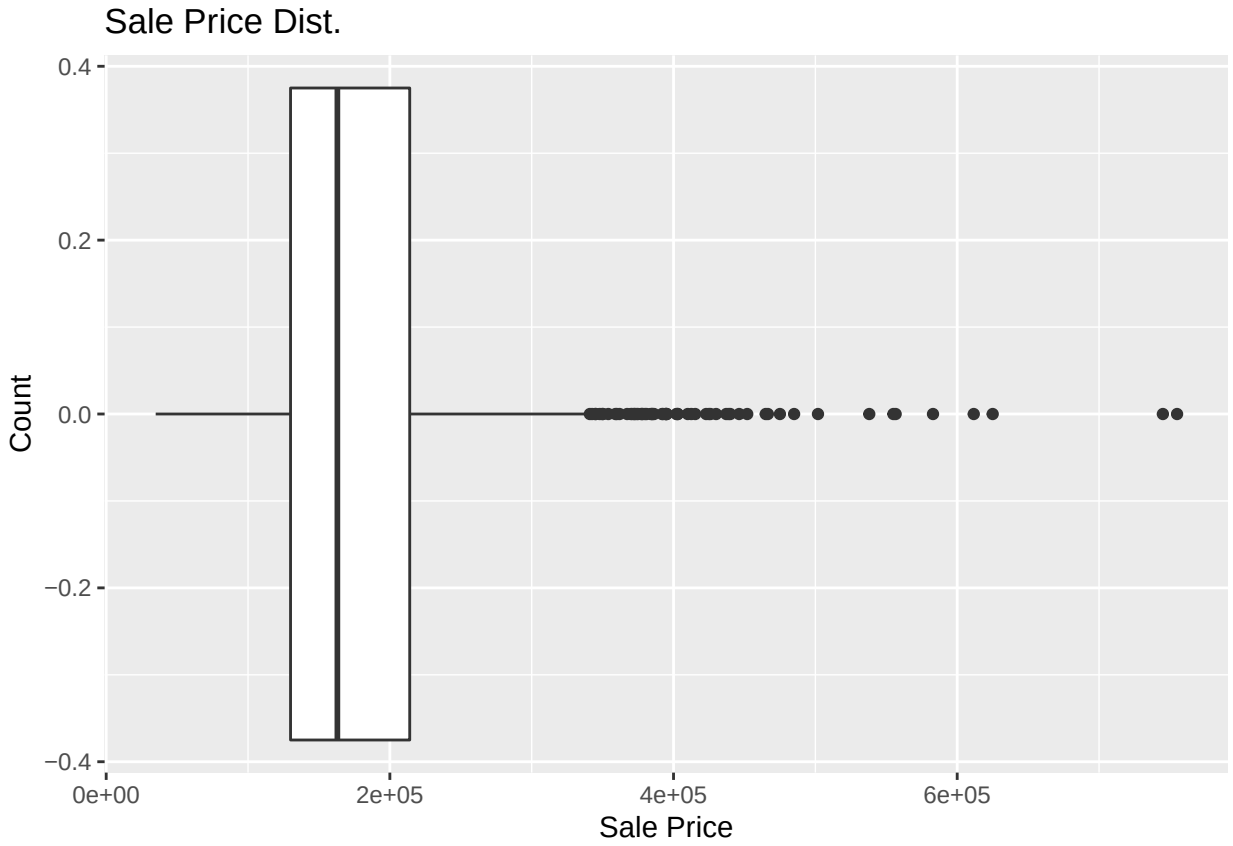
```
print(train_df %>% select(SalePrice) %>% summary)
```

```
##    SalePrice
##  Min.   : 34900
##  1st Qu.:129975
##  Median :163000
##  Mean   :180921
##  3rd Qu.:214000
##  Max.   :755000
```

```
train_df %>% ggplot(aes(x=SalePrice))+
  geom_histogram(bins=40)+
  labs(
    x="Sale Price",
    y="Count",
    title="Sale Price Dist.")
```




```
train_df %>% ggplot(aes(x=SalePrice))+
  geom_boxplot()+
  labs(
    x="Sale Price",
    y="Count",
    title="Sale Price Dist.")
```

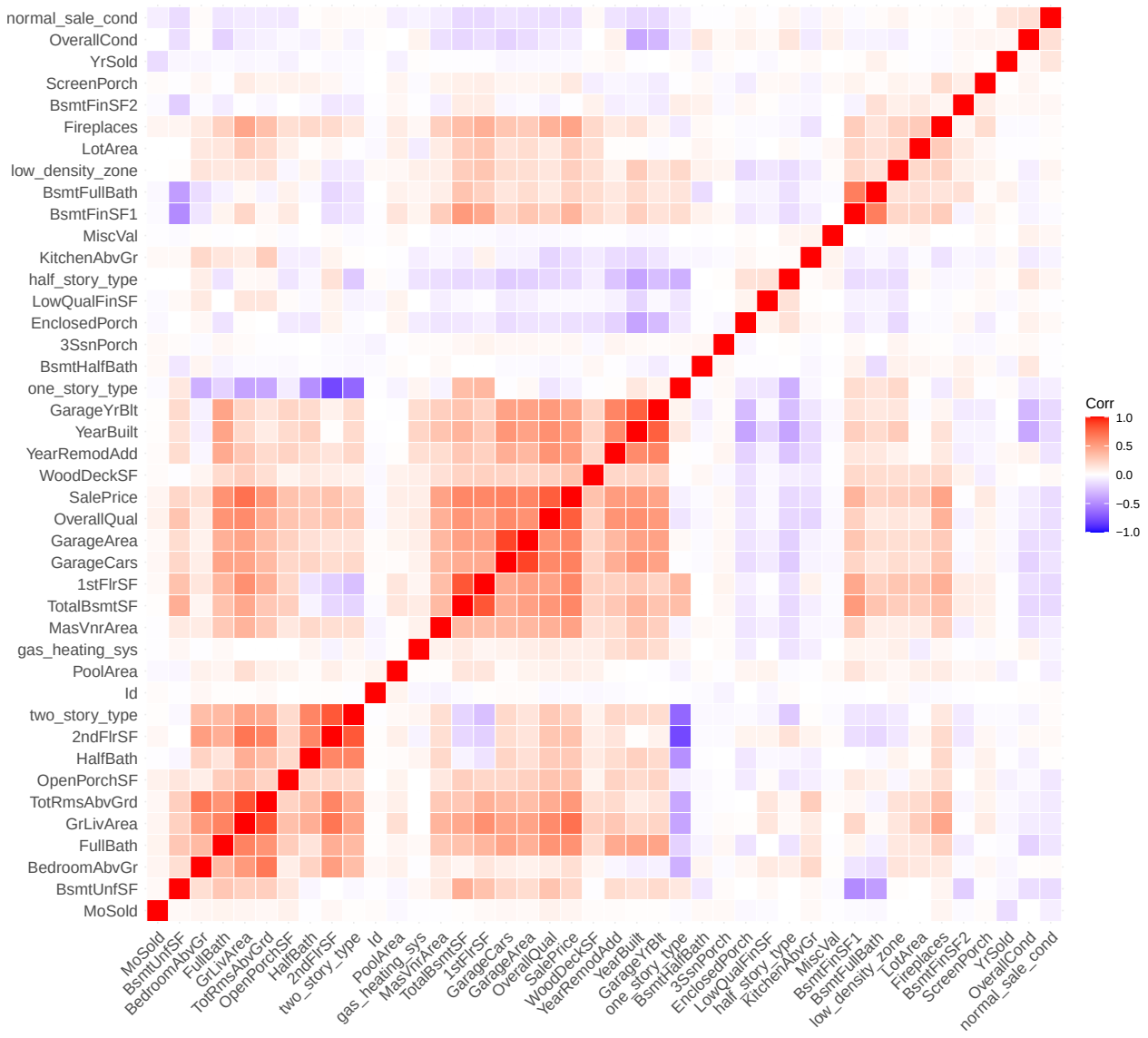


Most houses has price ranging between **1,000,000** to **2,500,000** USD.

Correlation Matrix in the Dataset

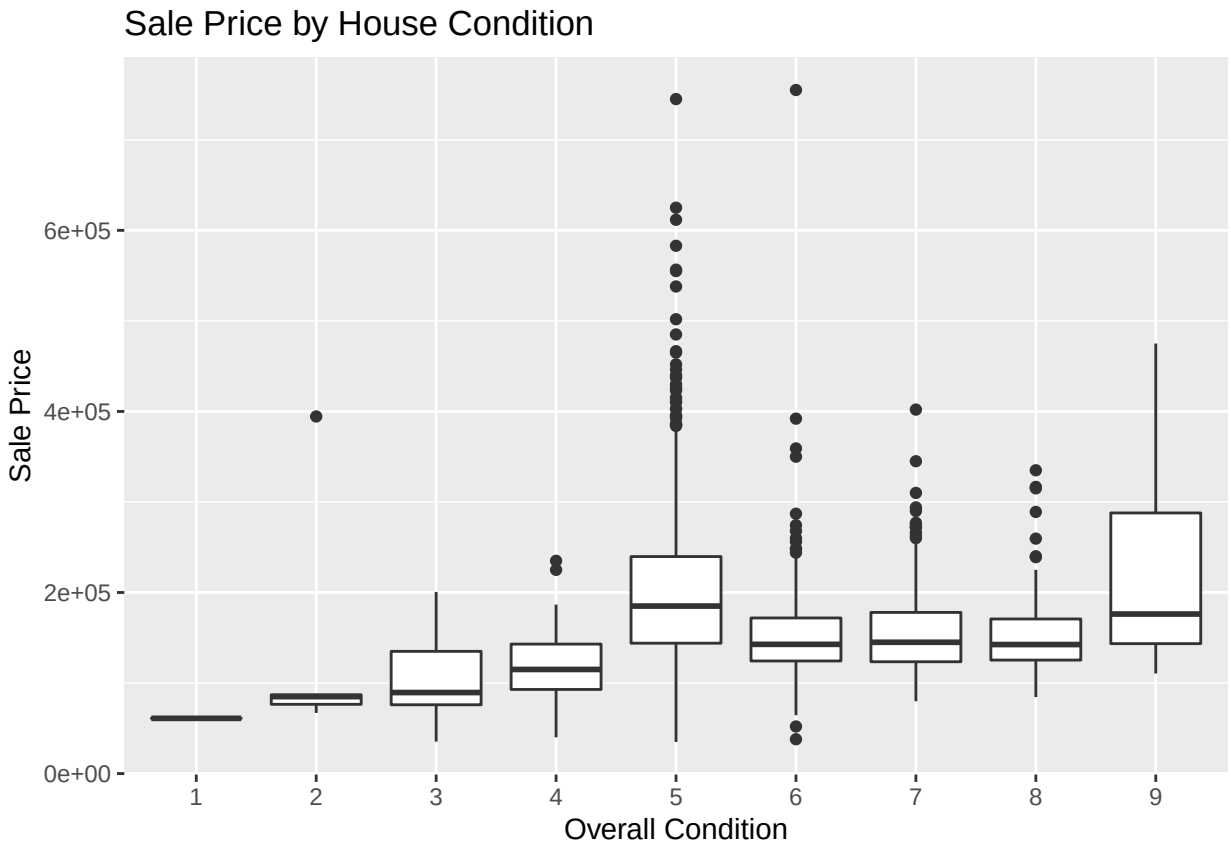
```
##install.packages("ggcorrplot")
library("ggcorrplot")
corr <- cor(train_df %>% select(where(is.numeric)))
corr[is.na(corr)] = 0

ggcorrplot(corr, hc.order = TRUE, outline.color = "white")
```



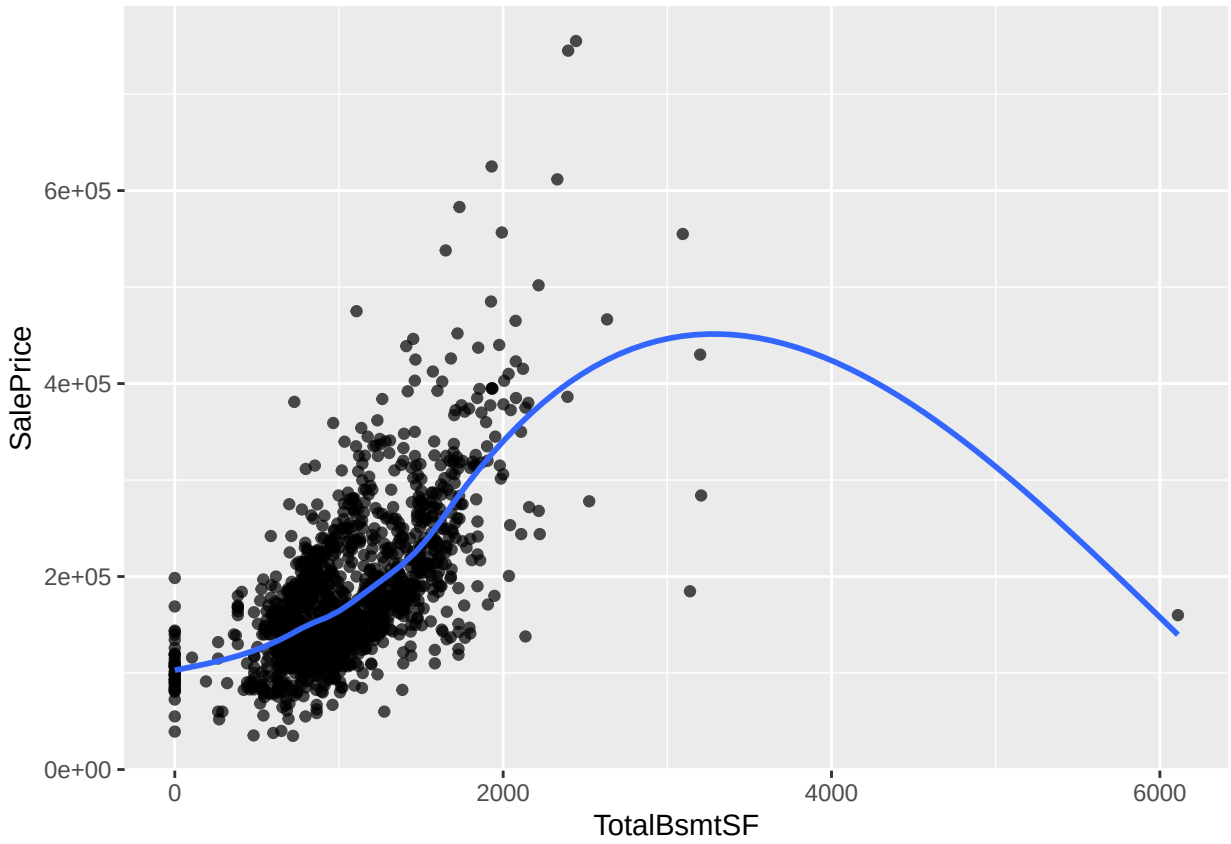
SalePrice vs OverallCond

```
train_df %>% mutate(OverallCond = as.character(OverallCond)) %>%
  ggplot(aes(x=OverallCond, y=SalePrice, group=OverallCond))+
    geom_boxplot()+
    labs(
      x="Overall Condition",
      y="Sale Price",
      title="Sale Price by House Condition"
    )
```



SalePrice vs TotalBsmtSF

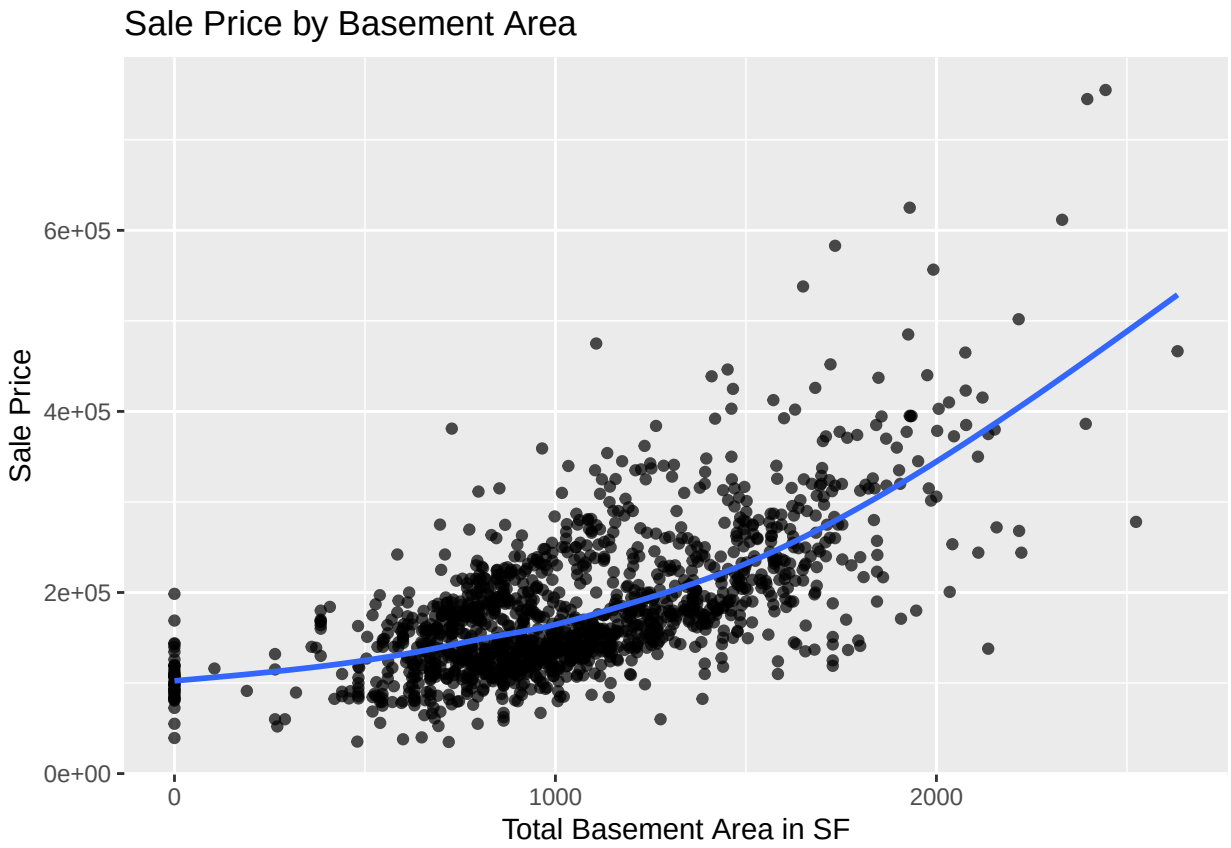
```
train_df %>% ggplot(aes(x=TotalBsmtSF, y=SalePrice))+
  geom_point(alpha=0.7)+
  geom_smooth(se=FALSE)
```



I detected an outliars at around more than **3,000** basement area, so i will choose to drop it.

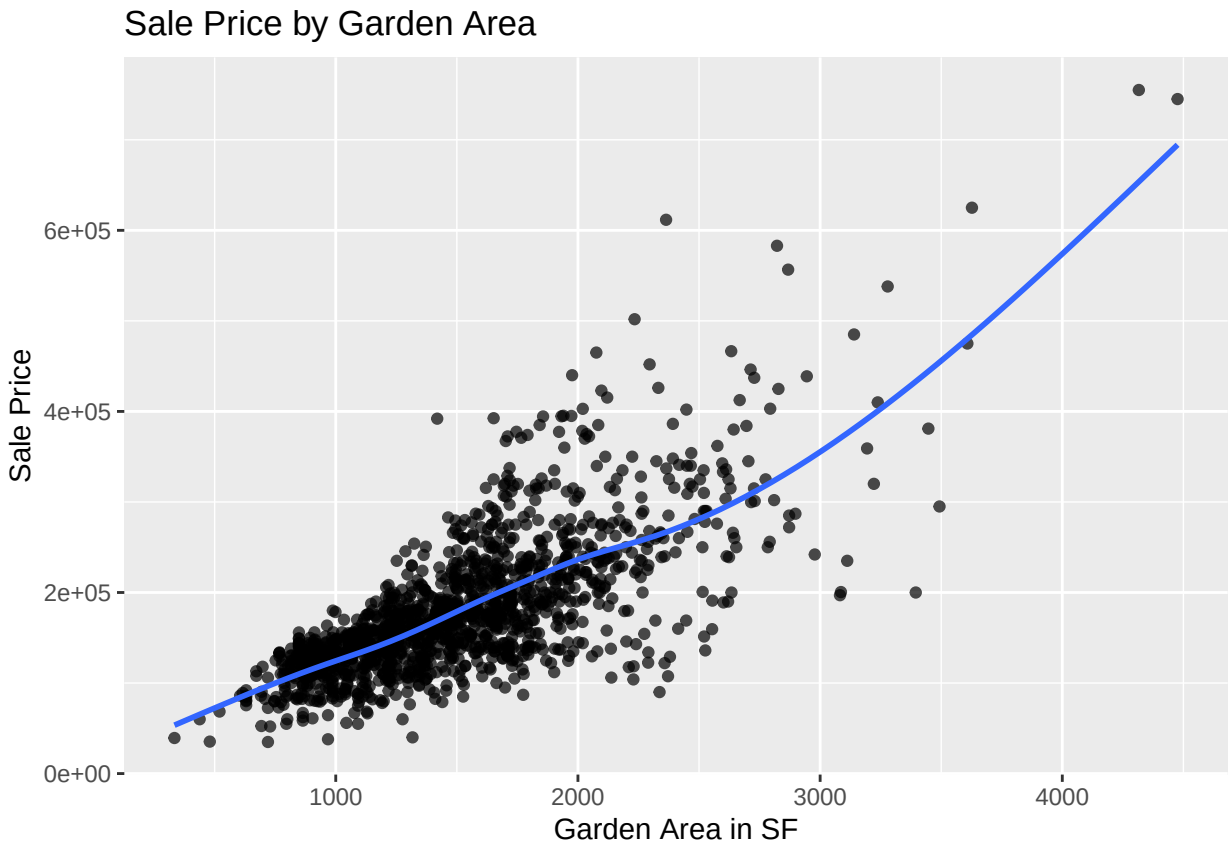
```
train_df <- train_df %>% filter(TotalBsmtSF < 3000)

train_df %>% ggplot(aes(x=TotalBsmtSF, y=SalePrice))+
  geom_point(alpha=0.7)+
  geom_smooth(se=FALSE)+
  labs(
    x="Total Basement Area in SF",
    y="Sale Price",
    title="Sale Price by Basement Area"
  )
```



SalePrice vs GrLivArea

```
train_df %>% ggplot(aes(x=GrLivArea, y=SalePrice))+
  geom_point(alpha=0.7)+
  geom_smooth(se=FALSE)+
  labs(
    x="Garden Area in SF",
    y="Sale Price",
    title="Sale Price by Garden Area"
  )
```



Both **TotalBsmtArea** and **GrLivArea** have very **strong positive** correlation with **SalePrice** columns.

Machine Learning Models Analysis

Ongoing. . . .