GroupProject-InterimReport

Sai Deepthi Mattam, Sri Harsha Samanthula

INTRODUCTION

The client's requirement is a real-time effective model to predict final selling price of houses in the city of Ames, Iowa.

OBJECTIVE

Initial focus of the project is to gain knowledge of the data and understand the relation between each of the variables to the house's sale price. Later, further statistical analysis will be conducted to select any 5 variables which tend to effect the price the most.

Later, using the training data set, a best-fitting model will be constructed with the 5 variables as predictors of housing prices. Performance of various statistical models will be compared against each other to determine which model fits the best.

ABOUT THE DATA

The data set available on Kaggle contains 80 variables that involve in assessing home values. Out of these, 20 are continuous, 14 are discrete and the remaining 46 are categorical variables. This data has been randomized and then split in to two sets(train and test) of equal size. "SalePrice" is the outcome variable

DATA EXPLORATION

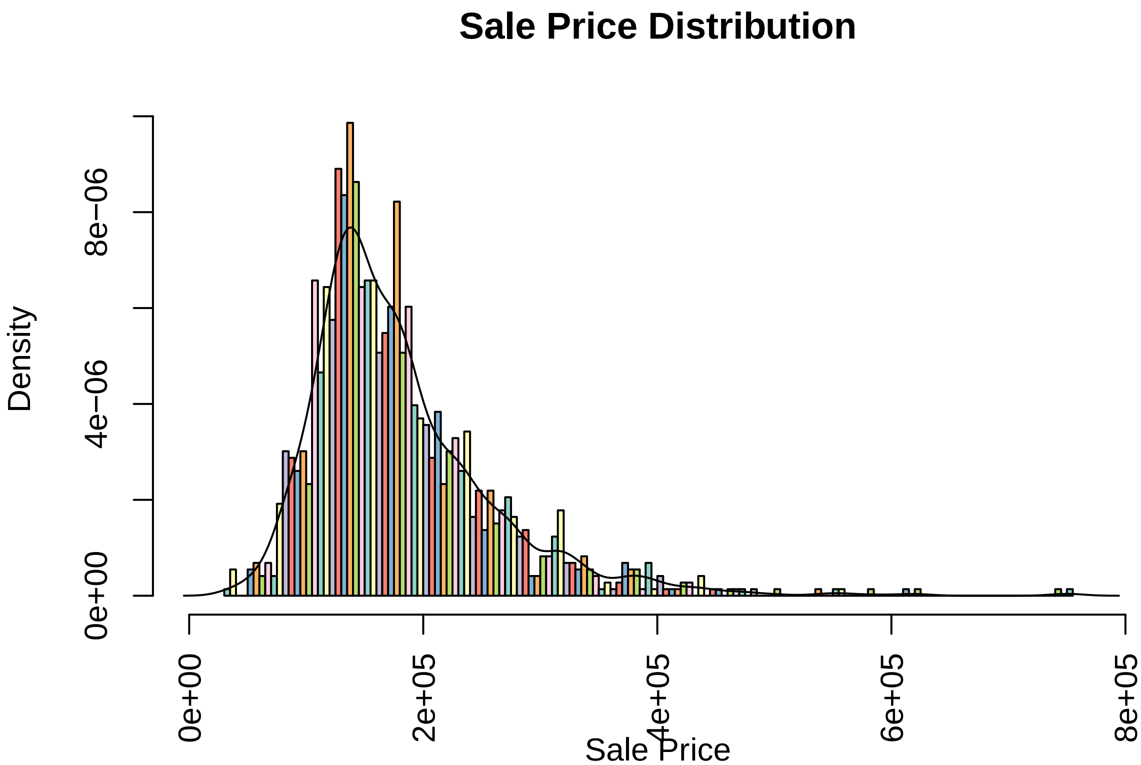
Certain columns have missing values(NAs). Below is the summary of all missing value information.

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

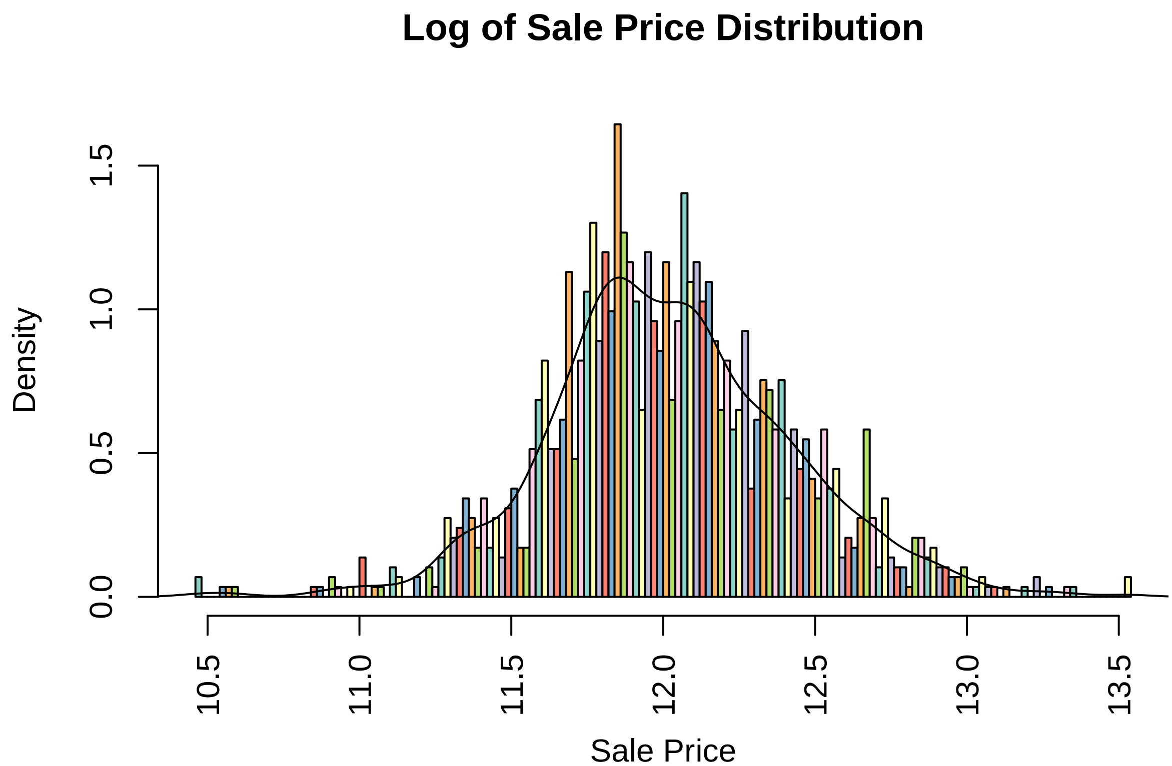
DATA VISUALIZATION

Summary statistics of Sales Price

```
##    mean_sp median_sp   sd_sp
## 1 180921.2    163000 79442.5
```



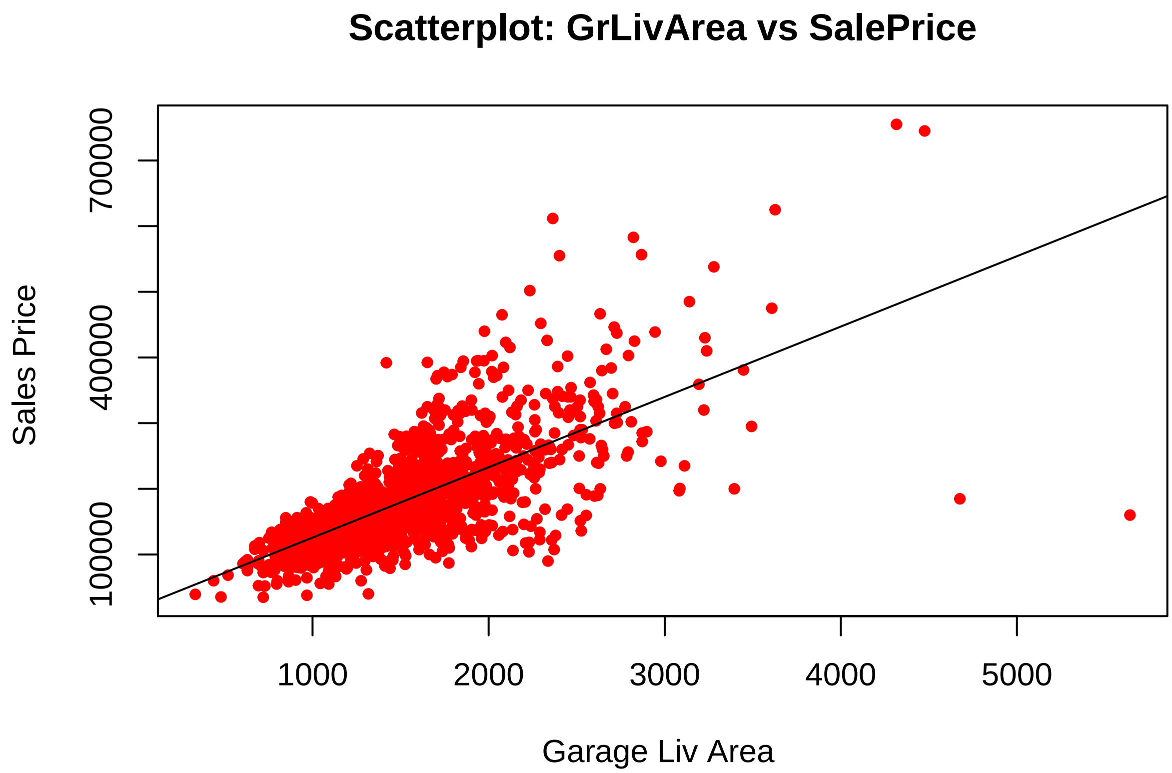
This histogram clearly shows that distribution of SalesPrice is Skewed to the right. To rectify this we need to apply log or power functions to SalesPrice variable.

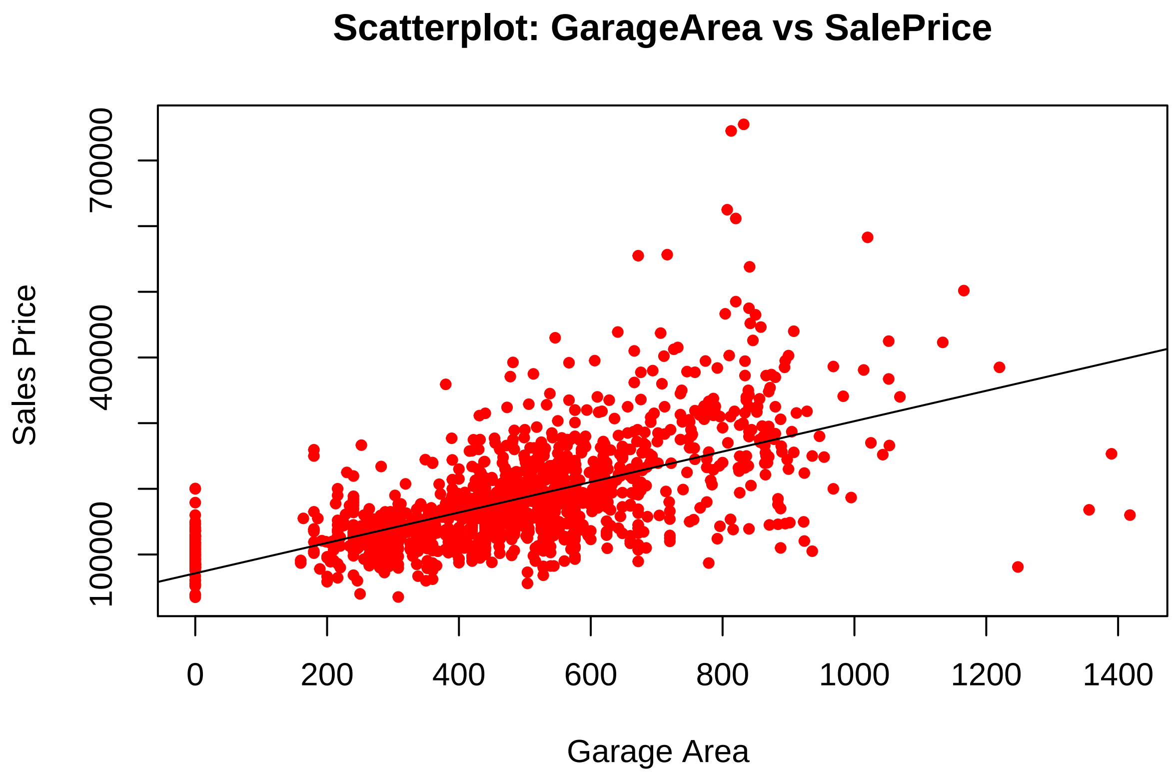


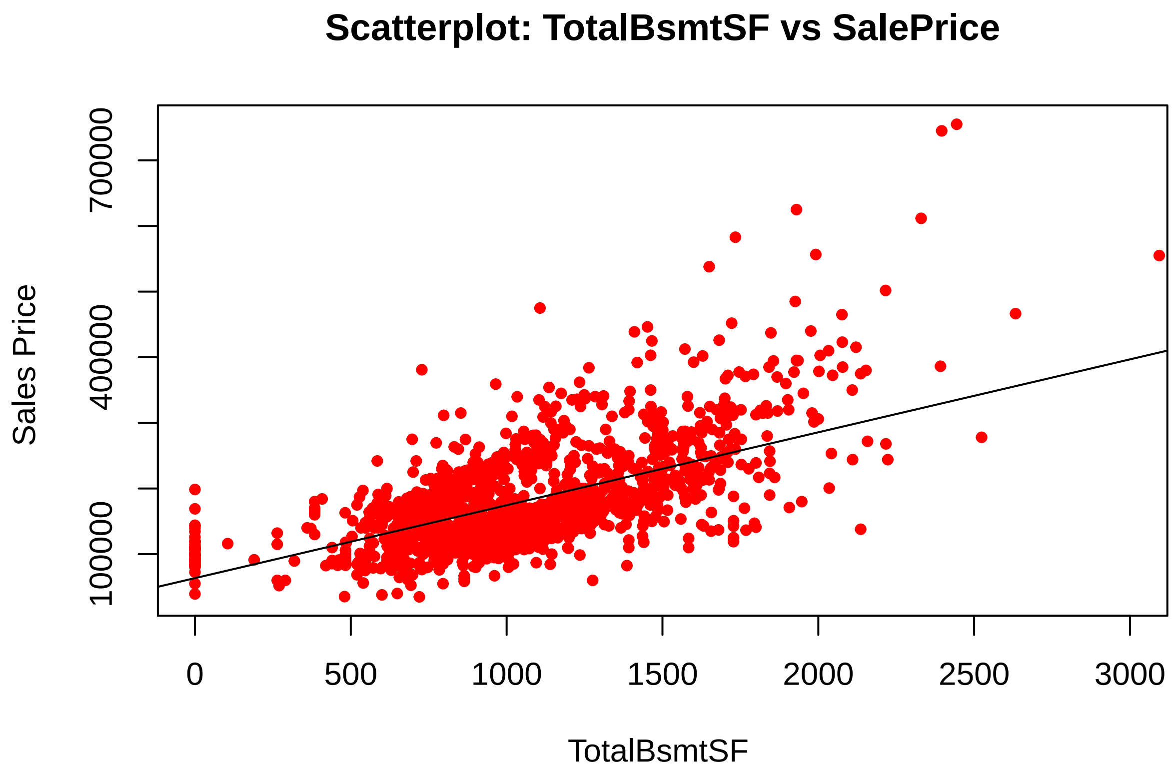
Top 5 Correlation Numerical Variables

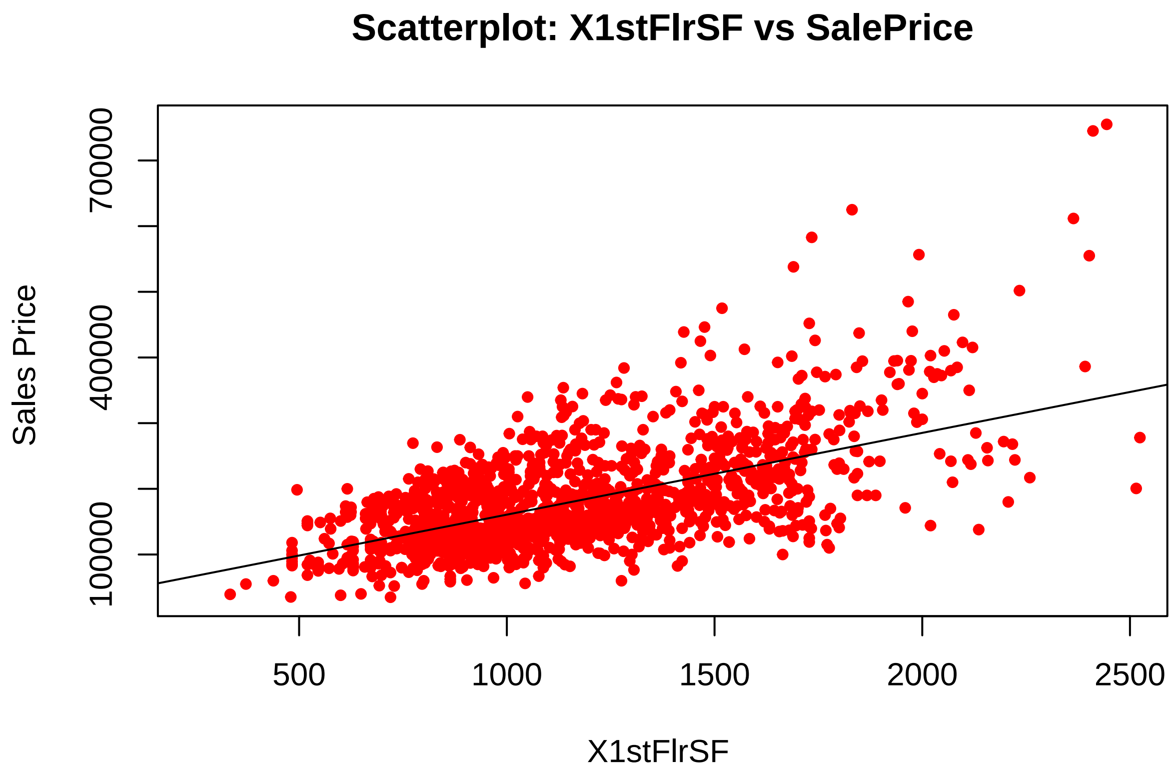
	Cors	Features
5	0.7909816	OverallQual
17	0.7086245	GrLivArea
27	0.6404092	GarageCars
28	0.6234314	GarageArea
13	0.6135806	TotalBsmtSF
14	0.6058522	X1stFlrSF

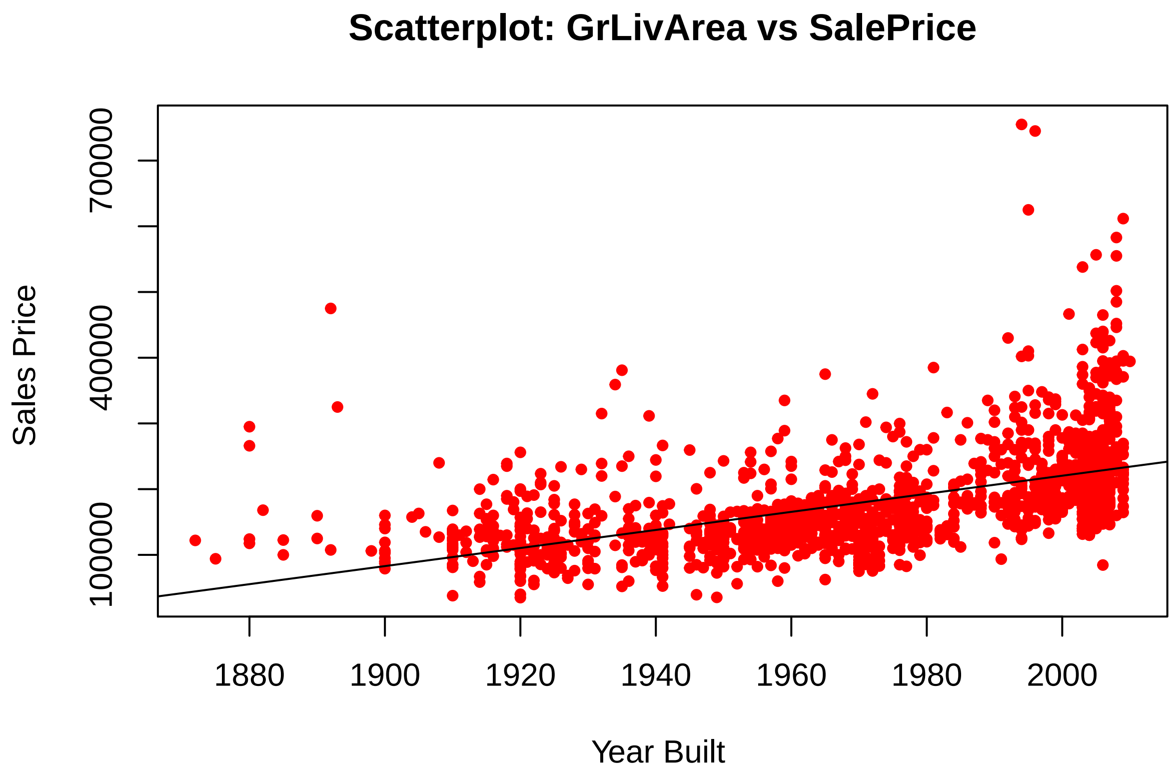
Exploring features using Scatterplots, BoxPlots etc

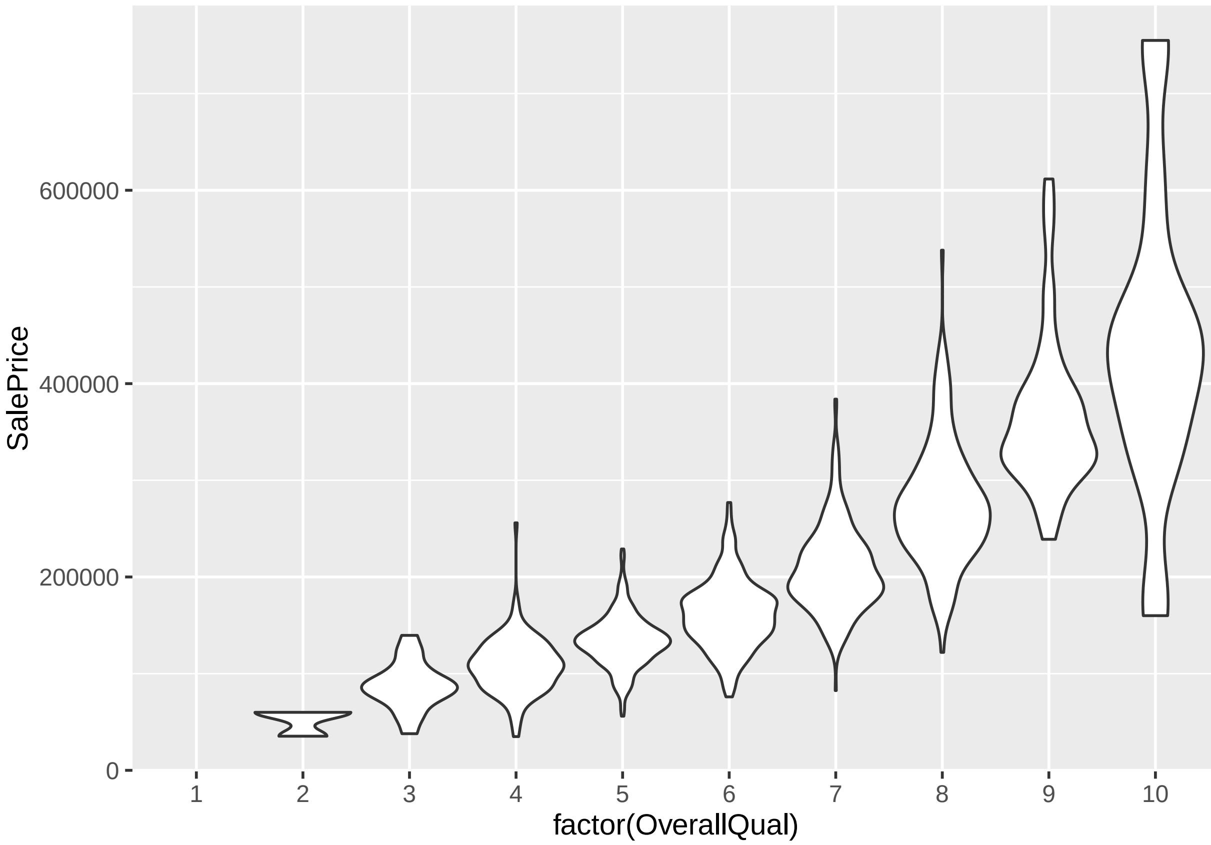

Scatterplot: GrLivArea vs SalePrice
Sales Price
100000
400000
700000
1000
2000
3000
4000
5000
Garage Liv Area


Scatterplot: GarageArea vs SalePrice
Sales Price
100000
400000
700000
0
200
400
600
800
1000
1200
1400
Garage Area

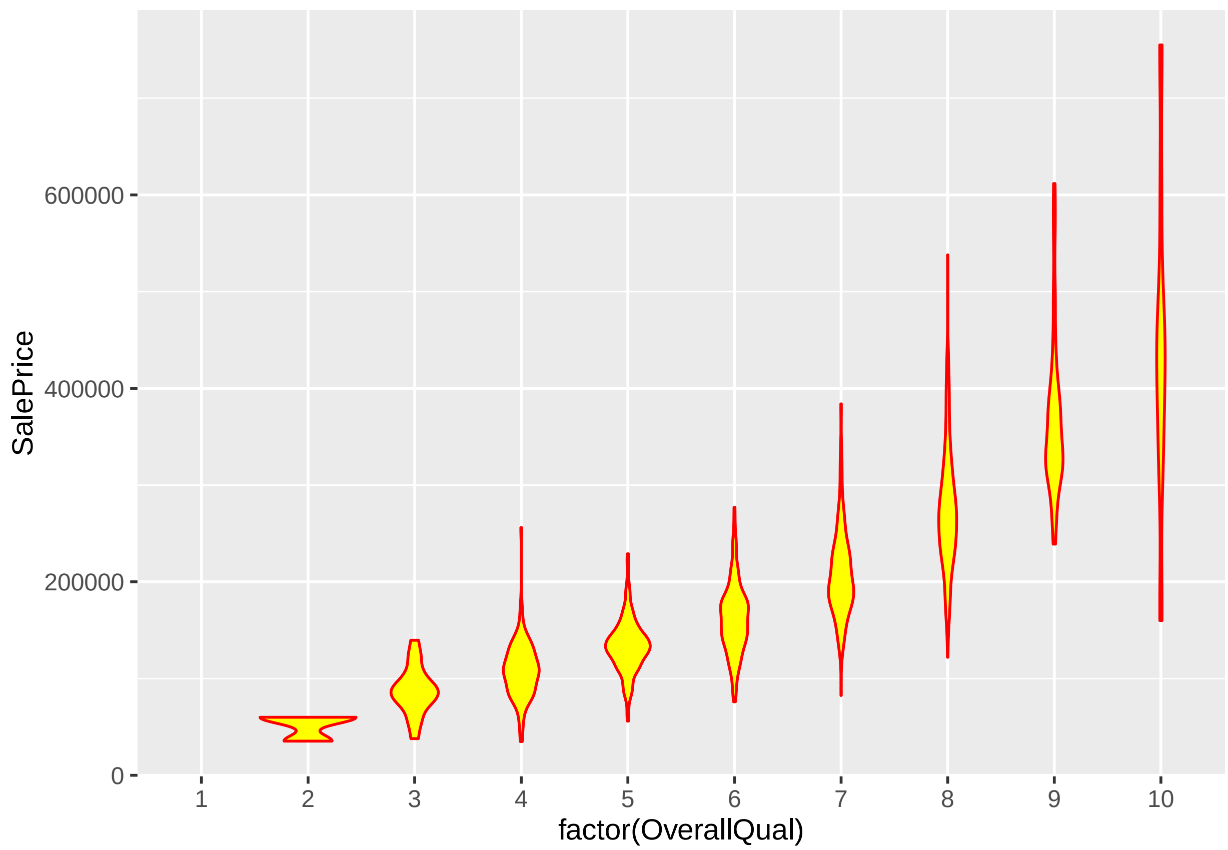

Scatterplot: TotalBsmtSF vs SalePrice
Sales Price
100000
400000
700000
0
500
1000
1500
2000
2500
3000
TotalBsmtSF

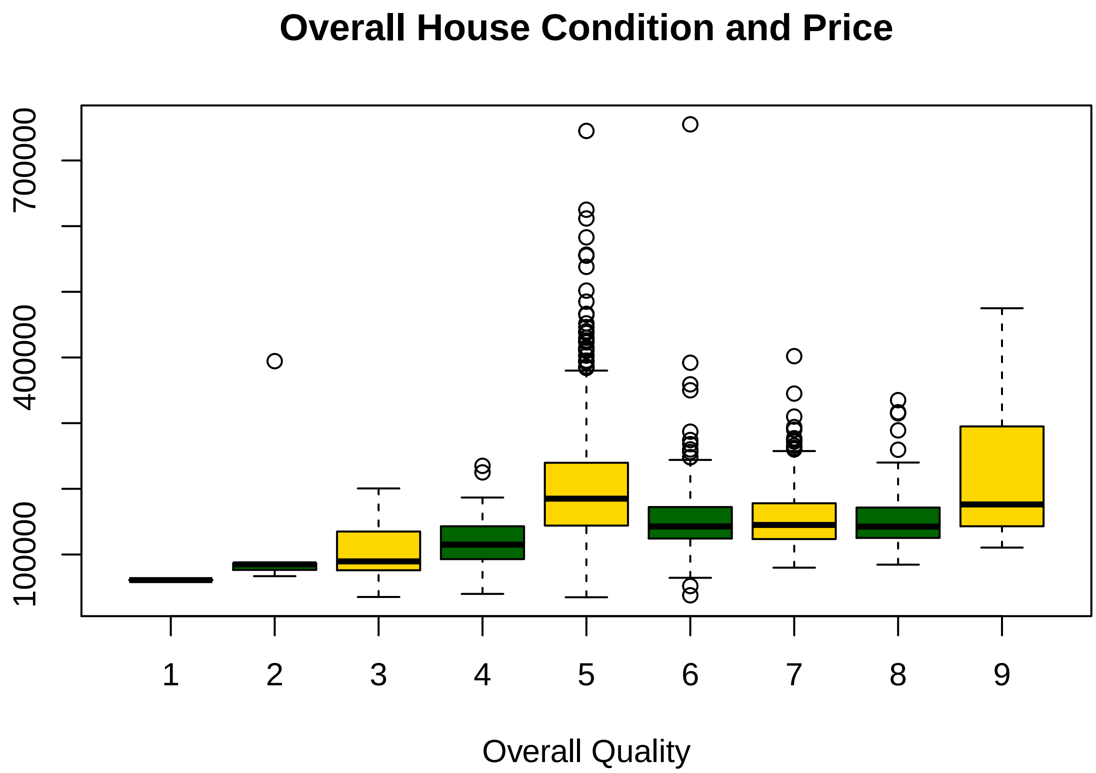

Scatterplot: X1stFlrSF vs SalePrice
Sales Price
100000
400000
700000
500
1000
1500
2000
2500
X1stFlrSF

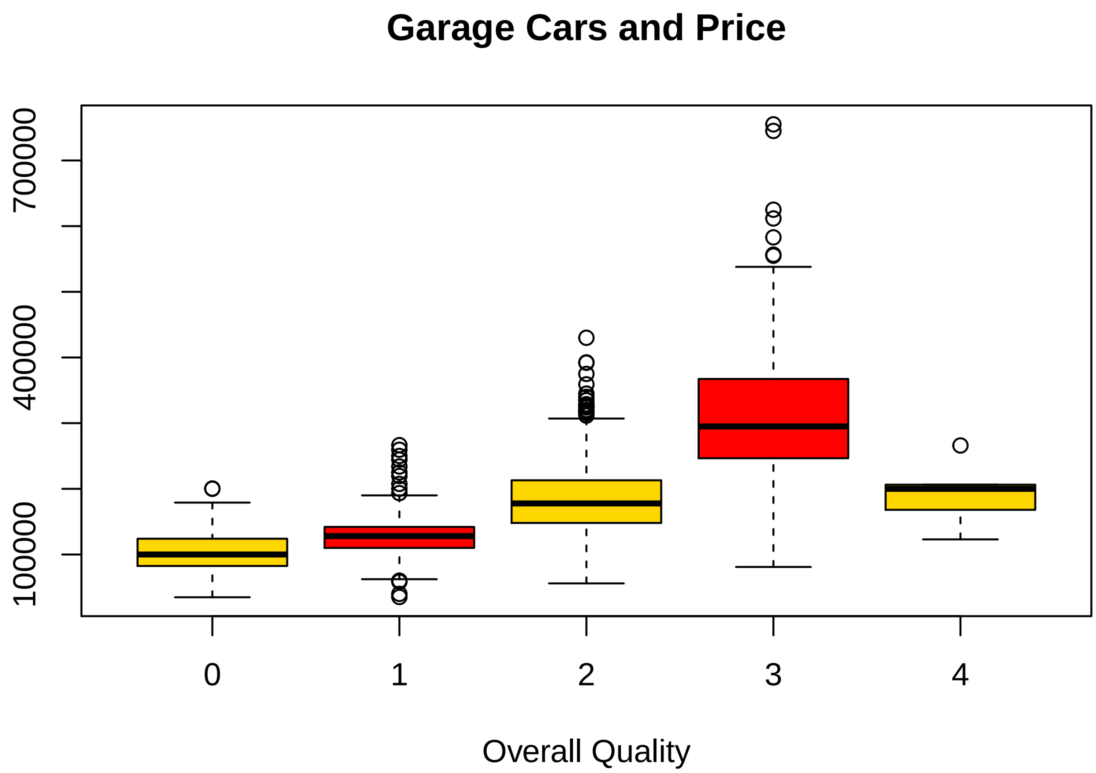

Scatterplot: GrLivArea vs SalePrice
Sales Price
100000
400000
700000
1880
1900
1920
1940
1960
1980
2000
Year Built



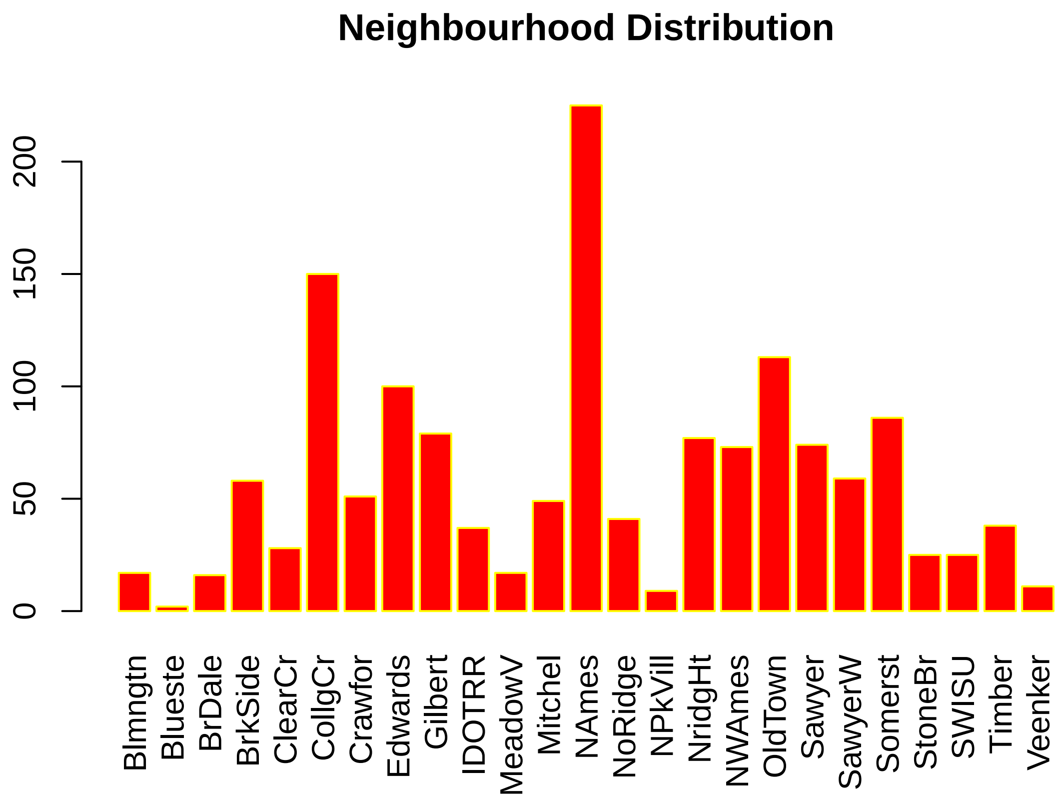
600000
400000
200000
0
SalePrice
1
2
3
4
5
6
7
8
9
10
factor(OverallQual)


SalePrice
600000
400000
200000
0
1
2
3
4
5
6
7
8
9
10
factor(OverallQual)

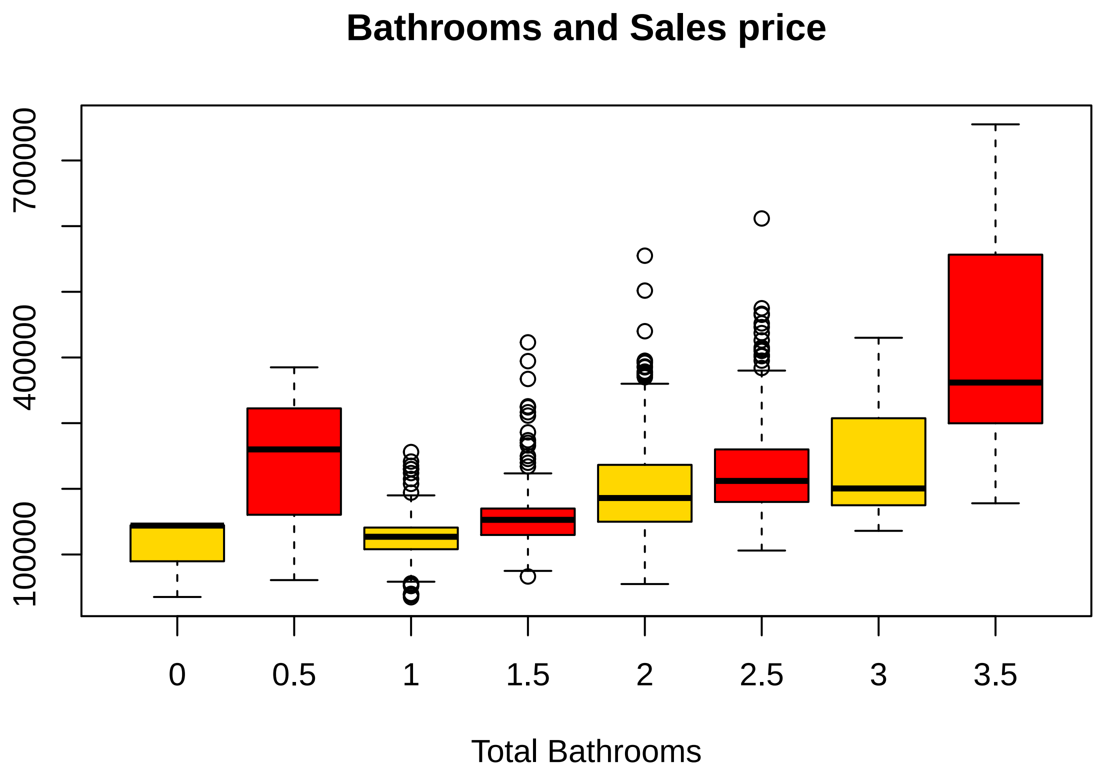

Overall House Condition and Price
700000
400000
100000
1
2
3
4
5
6
7
8
9
Overall Quality



Garage Cars and Price
700000
400000
100000
0
1
2
3
4
Overall Quality



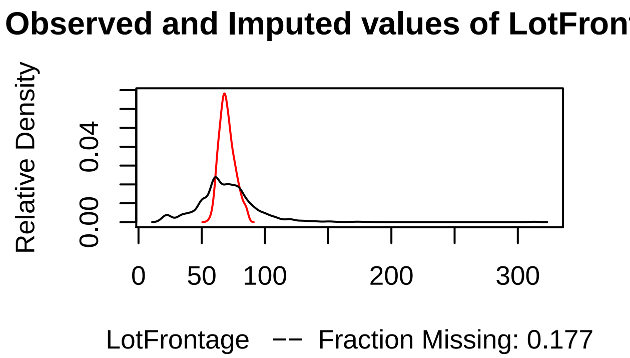
Neighbourhood Distribution
0
50
100
150
200
Blmngtn
Blueste
BrDale
BrkSide
ClearCr
CollgCr
Crawfor
Edwards
Gilbert
IDOTRR
MeadowV
Mitchel
NAmes
NoRidge
NPkVill
NridgHt
NWAmes
OldTown
Sawyer
SawyerW
Somerst
StoneBr
SWISU
Timber
Veenker

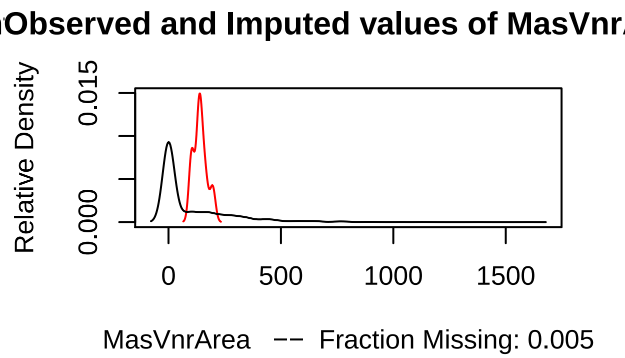


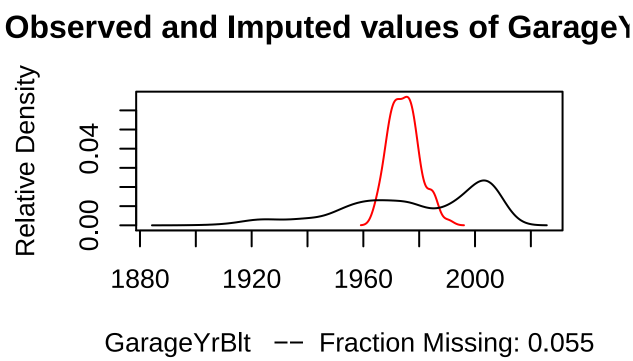
DATA CLEANING

NAs in numeric variables: Since these variables have an impact on the outcome variables, they can not be ignored. Also, the number of missing values for each variable is significantly higher which might introduce a substantial amount of bias or create reductions in efficiency. To avoid this, Imputation has been performed and Include methods on these variables. Imputation is a process of replacing missing data with an estimated value based on other available information.

Imputation with Amelia.

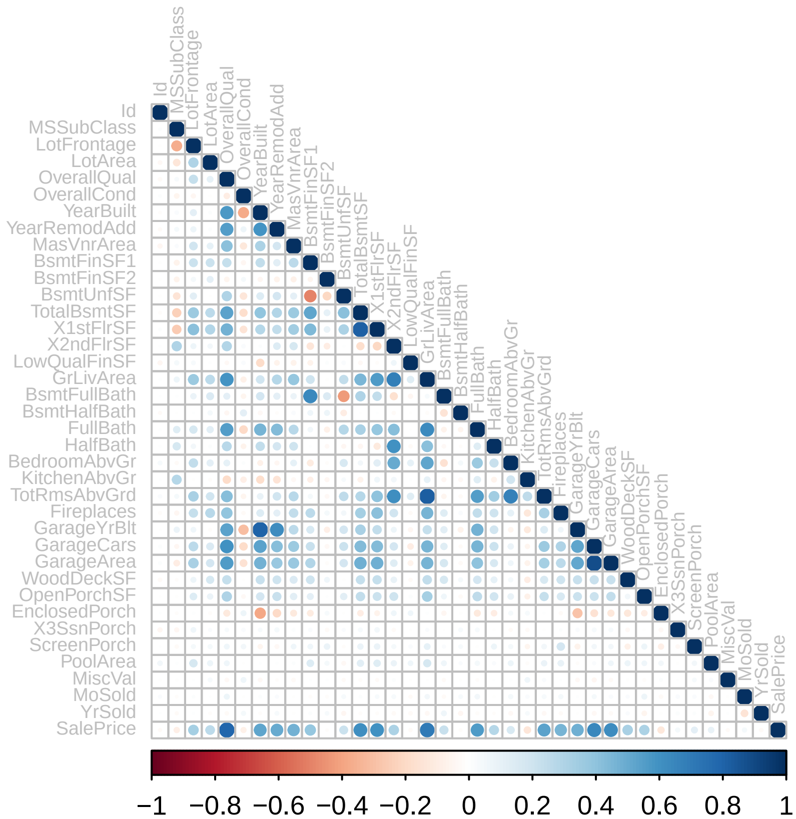






NAs in character variables: All character variables contain the category of a certain feature available in the house. As per the data description from Kaggle, NAs in such cases means absence of that feature. Hence, replacing NAs with more descriptive words.

Viewing the Correlation Plot after



Inspecting Multicolinearity between features in order to eliminate highly corelated features.

name1	name2	cor
X1stFlrSF	TotalBsmtSF	0.81953
GrLivArea	X2ndFlrSF	0.6875011
BsmtFullBath	BsmtFinSF1	0.6492118
FullBath	GrLivArea	0.6300116
HalfBath	X2ndFlrSF	0.6097073
TotRmsAbvGrd	X2ndFlrSF	0.6164226
TotRmsAbvGrd	GrLivArea	0.8254894
TotRmsAbvGrd	BedroomAbvGr	0.6766199
GarageYrBlt	YearBuilt	0.8024955
GarageYrBlt	YearRemodAdd	0.6239463
GarageCars	OverallQual	0.6006707
GarageArea	GarageCars	0.8824754

Converting character variables into factors/catergorical variables.

MODEL AND MODEL DEVELOPMENT

Creating a base Linear Model using all the predictors.

```
lm.all <- standardize(
  lm(
```

```
    SalePrice ~ MSSubClass +   MSZoning +     LotFrontage +  LotArea +      Street +
      Alley +        LotShape +
    LandContour +  Utilities +    LotConfig +    LandSlope +    Neighborhood + Condition1 +
      Condition2 +   BldgType +
    HouseStyle +   OverallQual +  OverallCond +  YearBuilt +    YearRemodAdd + RoofStyle +
      RoofMatl +     Exterior1st +
    Exterior2nd +  MasVnrType +   MasVnrArea +   ExterQual +    ExterCond +    Foundation +
      BsmtQual +     BsmtCond +
    BsmtExposure + BsmtFinType1 + BsmtFinSF1 +   BsmtFinType2 + BsmtFinSF2 +   BsmtUnfSF +
      TotalBsmtSF +  Heating +
    HeatingQC +    CentralAir +   Electrical +   X1stFlrSF +    X2ndFlrSF +    LowQualFinSF
    + GrLivArea +    BsmtFullBath +
    BsmtHalfBath + FullBath +     HalfBath +     BedroomAbvGr + KitchenAbvGr + KitchenQual
    +  TotRmsAbvGrd + Functional +
    Fireplaces +   FireplaceQu +  GarageType +   GarageYrBlt +  GarageFinish + GarageCars
    +   GarageArea +   GarageQual +
    GarageCond +   PavedDrive +   WoodDeckSF +   OpenPorchSF +  EnclosedPorch + X3SsnPorch
    +   ScreenPorch +  PoolArea +
    PoolQC +       Fence +        MiscFeature +  MiscVal +      MoSold +       YrSold +
      SaleType +     SaleCondition
    , data = dt.train
  )
)
```

```
## RMSE of the baseline model with all predictors  32484.24
```

Removing the predictor with NAs as coeffiecient, because of multi colinearity Exterior2nd, BsmtCond, BsmtFinType1, TotalBsmtSF, Electrical, GarageFinish, GarageCond, GrLivArea, GarageQual

```
lm.sel <- standardize(
  lm(
    SalePrice ~ MSSubClass +   MSZoning +     LotFrontage +  LotArea +      Street +
      Alley +        LotShape +
    LandContour +  Utilities +    LotConfig +    LandSlope +    Neighborhood +
      Condition1 +   Condition2 +   BldgType +
    HouseStyle +   OverallQual +  OverallCond +  YearBuilt +    YearRemodAdd +
      RoofStyle +    RoofMatl +     Exterior1st +
    #
      MasVnrType +   MasVnrArea +   ExterQual +    ExterCond +    Foundation +   BsmtQual +
    BsmtExposure  + BsmtFinSF1 +   BsmtFinType2 + BsmtFinSF2 +   BsmtUnfSF      +  Heating +
    HeatingQC +    CentralAir +   X1stFlrSF +    X2ndFlrSF +
      LowQualFinSF  +    BsmtFullBath +
    BsmtHalfBath + FullBath +     HalfBath +     BedroomAbvGr + KitchenAbvGr
    + KitchenQual +  TotRmsAbvGrd + Functional +
    Fireplaces +   FireplaceQu +  GarageType +   GarageYrBlt +
      GarageCars +   GarageArea  +
     PavedDrive +   WoodDeckSF +   OpenPorchSF +  EnclosedPorch + X3SsnPorch +
      ScreenPorch +  PoolArea +
    PoolQC +       Fence +        MiscFeature +  MiscVal +      MoSold +
      YrSold +       SaleType +     SaleCondition
    , data = dt.train
  )
)
```

```
## RMSE of the model after removing multicollinear variables with all predictors  20915.49
```

Picking predictors basing on the Beta coeffiencients and P values.

	Estimate	Std. Error	t value	Pr(>|t|)	estabs
(Intercept)	-674310.63	138219.316	-4.8785557	0.0000012	674310.63
RoofMatlMembran	648807.30	61021.111	10.6325055	0.0000000	648807.30
RoofMatlWdShngl	635853.18	51984.900	12.2314975	0.0000000	635853.18
RoofMatlMetal	609675.61	60516.051	10.0746100	0.0000000	609675.61
RoofMatlCompShg	562690.96	51170.382	10.9964189	0.0000000	562690.96
RoofMatlTar&Grv	559485.32	54881.530	10.1944192	0.0000000	559485.32
RoofMatlWdShake	555503.91	53425.679	10.3976949	0.0000000	555503.91
RoofMatlRoll	552409.08	56459.476	9.7841695	0.0000000	552409.08
PoolQCNoPool	271450.75	117366.143	2.3128540	0.0208932	271450.75
Condition2PosN	-236131.67	26910.703	-8.7746377	0.0000000	236131.67
PoolQCFa	-167287.44	39332.949	-4.2531120	0.0000227	167287.44
PoolQCGd	-133427.31	35861.150	-3.7206645	0.0002076	133427.31
Condition2RRAe	-117856.56	64281.507	-1.8334442	0.0669754	117856.56
RoofStyleShed	86358.89	33942.024	2.5443059	0.0110696	86358.89
z.PoolArea	57713.29	17356.136	3.3252382	0.0009092	57713.29
z.X2ndFlrSF	55596.71	4892.465	11.3637420	0.0000000	55596.71
NeighborhoodStoneBr	36738.62	8096.229	4.5377446	0.0000062	36738.62
FunctionalSev	-36327.86	29247.449	-1.2420865	0.2144388	36327.86
Condition2PosA	34643.94	36757.024	0.9425122	0.3461135	34643.94
z.X1stFlrSF	34506.61	4307.948	8.0099859	0.0000000	34506.61

New model with just the strong predictors picked from above, and strongly corelated variables.

```
## RMSE of the model with selected variables 33266.17
```

Using FSelector, and performing Chisquare test to pick important features.

Features obtained:

```
## SalePrice ~ FullBath + Fireplaces + OverallQual + GarageCars +
##     Neighborhood
## <environment: 0x0000000050accd70>
```

Using CFS test to pick important numercial variables.

Features obtained.

```
## SalePrice ~ OverallQual + TotalBsmtSF + GrLivArea + GarageCars
## <environment: 0x00000000540b2b00>
```

For Feature selections we used chi.squared which will find weights of discrete attributes.This shows us the most important features out of all available variables. The features obtained according to this test are : OverallQual, FullBath, Neightbourhood, Fireplace, GarageCar . So, these are most influential categorical variables. Correlation based feature selection has also been used to identity the most important numerical variables. Numerical variables obtained in this test are : Overall Qual, GarageCar, TotalBasment, GrLivArea

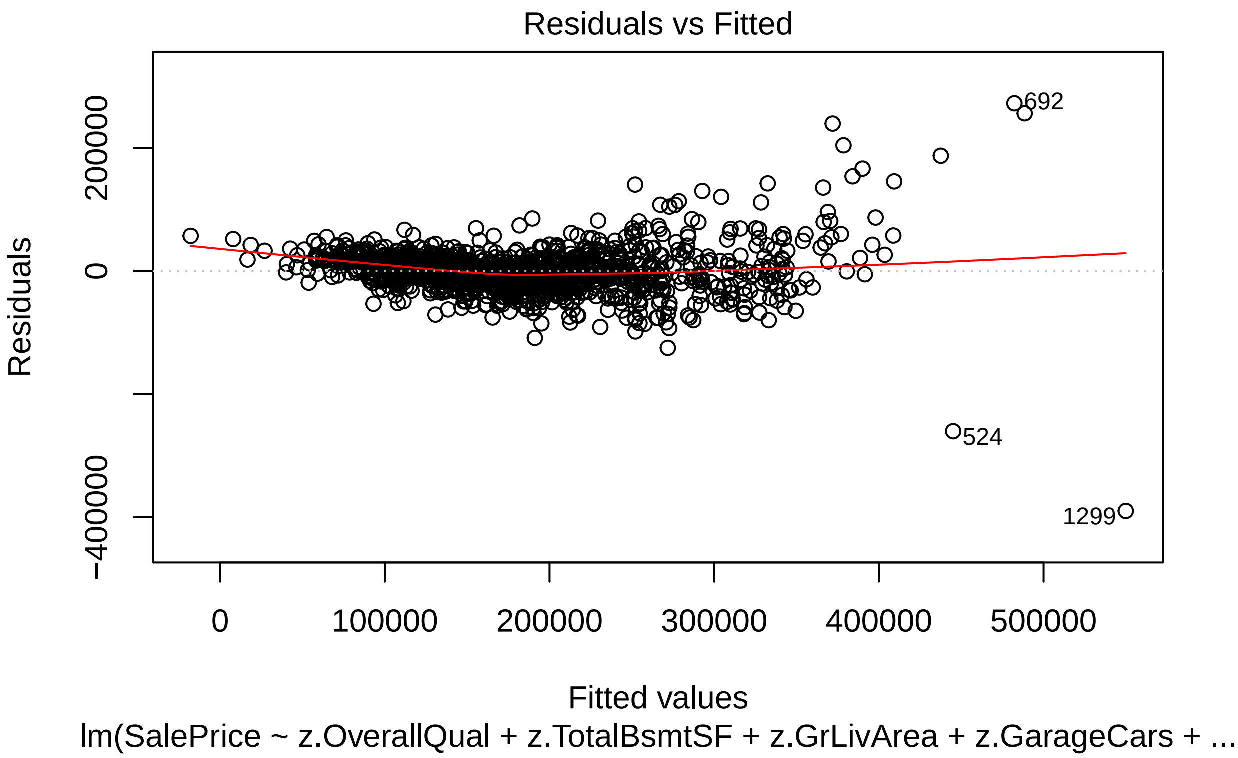
Final Model with just the Top 5 predictors.

```r
lm.sel4 <- standardize(lm(SalePrice ~ OverallQual  + TotalBsmtSF
                          + GrLivArea + GarageCars + Neighborhood ,data=dt.train))
```

```
## RMSE of the final model 34805.43
```

After brainstorming about general features considered by people to make a decision about a house, conclusion have been made that above features are considered more often than other available variables

Exploring the residual plot of the final model

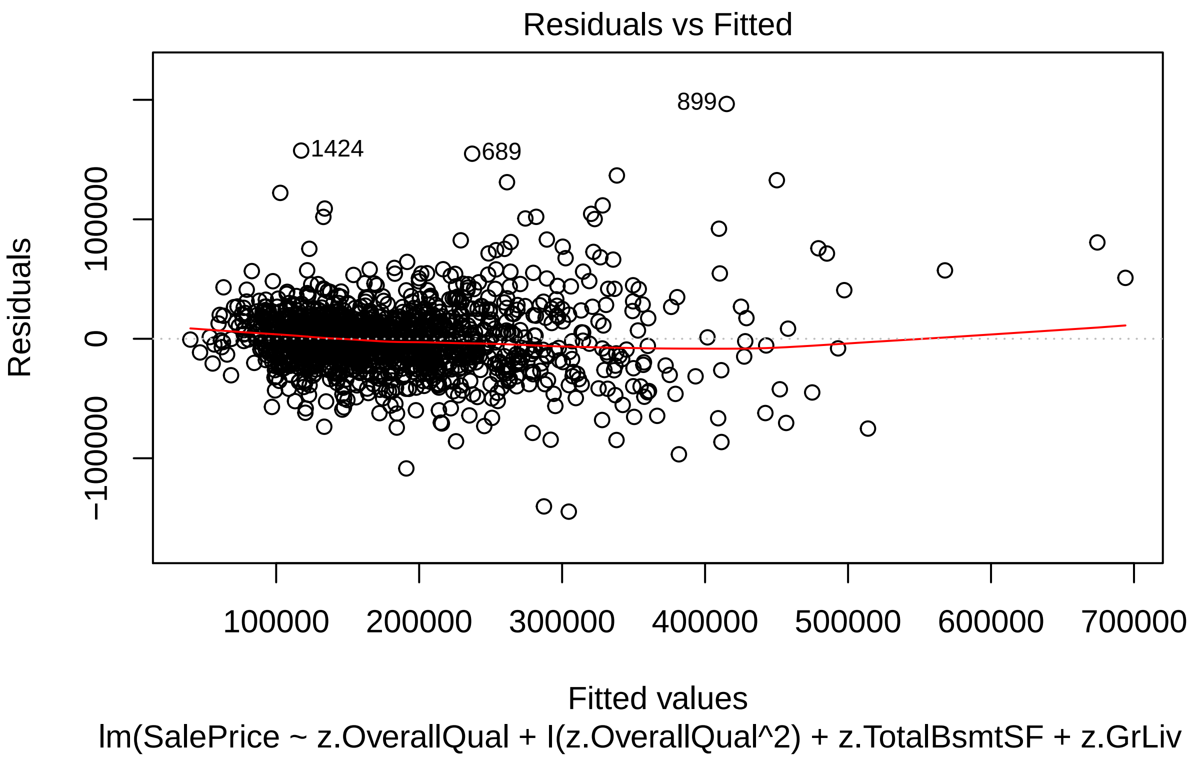


including the quadratic term of Quality variables to address the non linearity

```r
cat("RMSE of the final model with quadratic term and interaction", rmse(dt.train$SalePrice, predict(lm.s
```

```
## RMSE of the final model with quadratic term and interaction 27746.46
```

Residual plot of the final model after adding the quadratic variable and interaction term



NEXT STEPS

After the initial attempts and computations, these following steps have been planned to improve the model

1. Use ensemble to improve the model performance
2. Try various combinatons of interactions between variables and try building model with various forms such as quadratic, power forms.