GroupProject-InterimReport

Sai Deepthi Matam, Sri Harsha Samanthula

INTRODUCTION

The primary 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

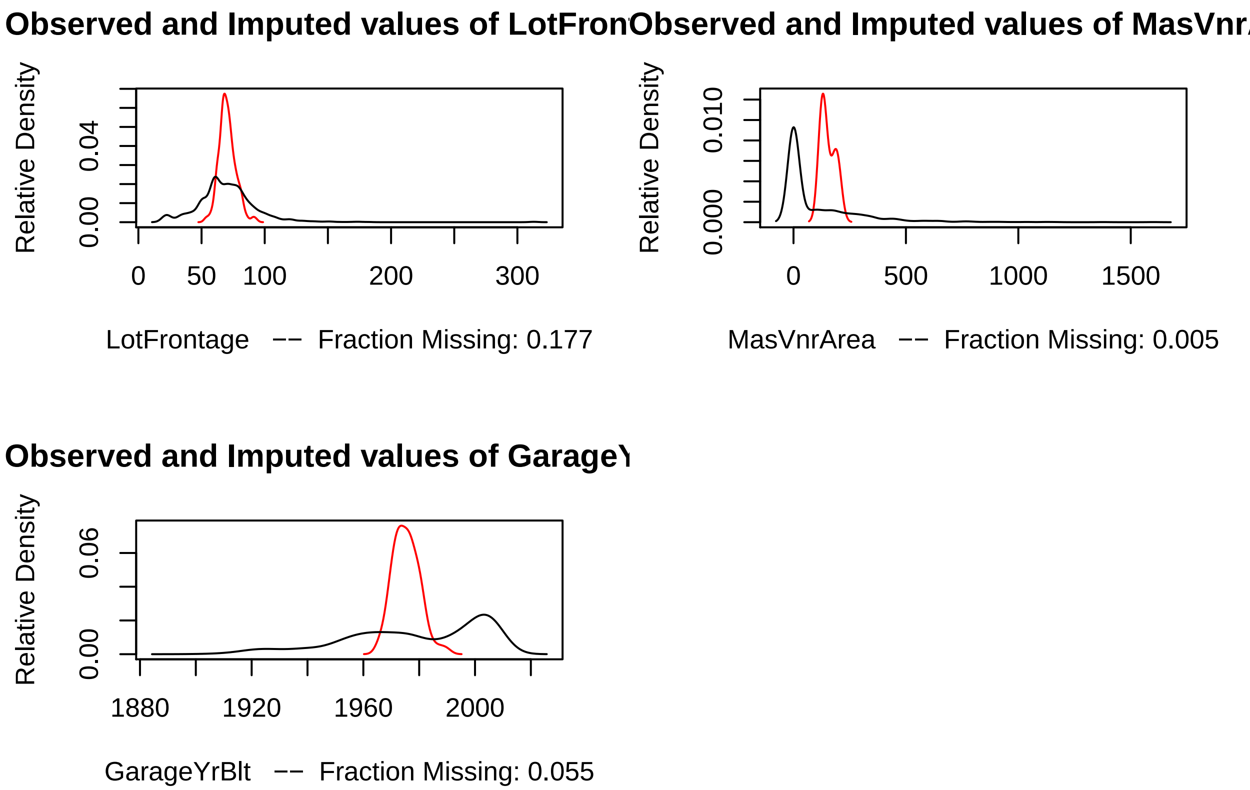
Certain columns have missing values(NAs). Below is the summary of all missing value information.

	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 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.As Amelia is known for better efficiency and reduction in bias when compared to Mean imputations, it has been used.



Here, out of 80 varaibles, there are only 3 variables that has missing values. Single imputations works well in this case. So, we used Bagimpute

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.

DATA VISUALIZATION

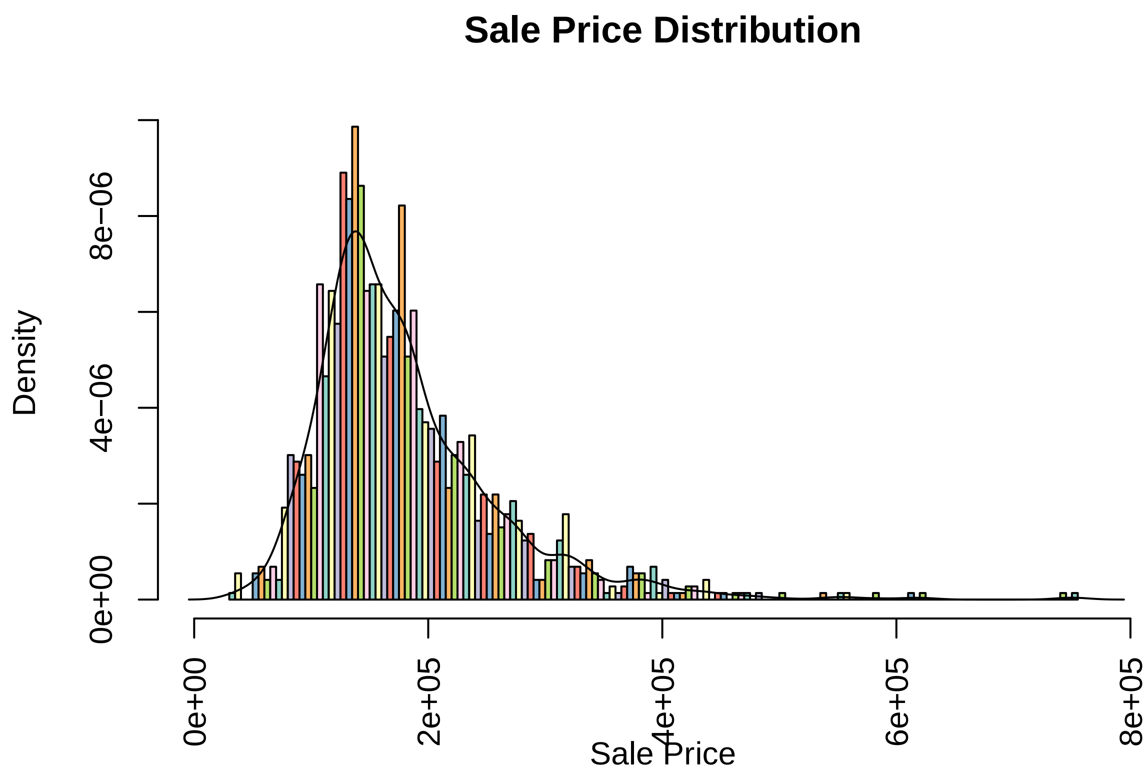
To understand the spread of the Sale Price of houses in Ames.

```
## Mean :  180921.2
```

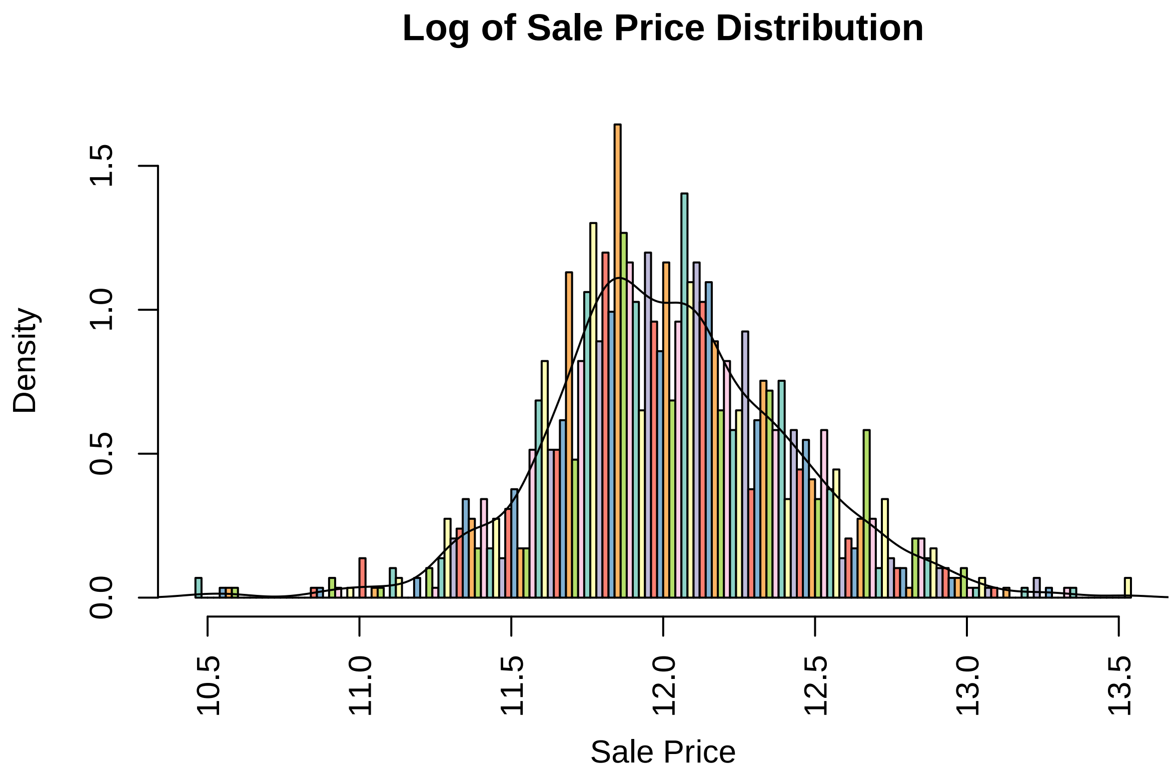
```
## Median :  163000
```

```
## Standard Deviation :  79442.5
```

Here the Mean > Median which indicates a right skew in the data. The same is also plotted below:



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.

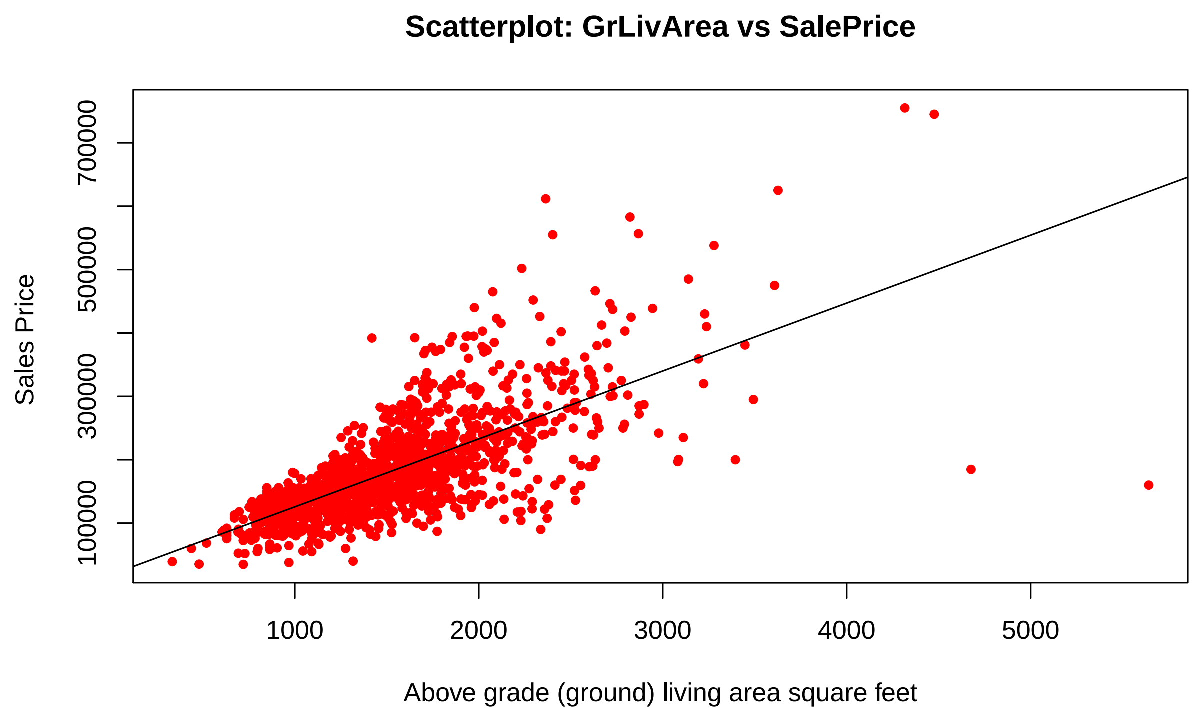


After applying the log function to the SalePrice, the distribution is closer to a normal distribution. Hence we can apply central limit theorm.

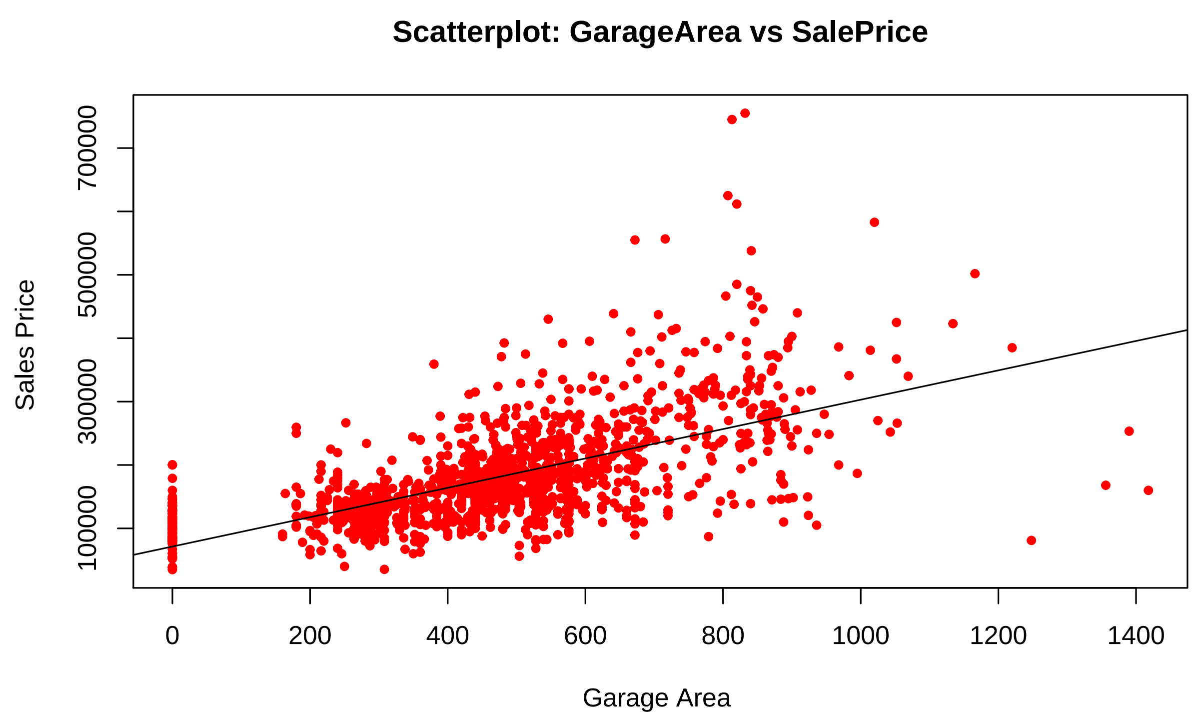
Top 5 Correlation Numerical Variables

Features	Cors
OverallQual	0.7909816
GrLivArea	0.7086245
GarageCars	0.6404092
GarageArea	0.6234314
TotalBsmtSF	0.6135806
X1stFlrSF	0.6058522

Exploring top 5 correlated features using Scatterplots, BoxPlots etc

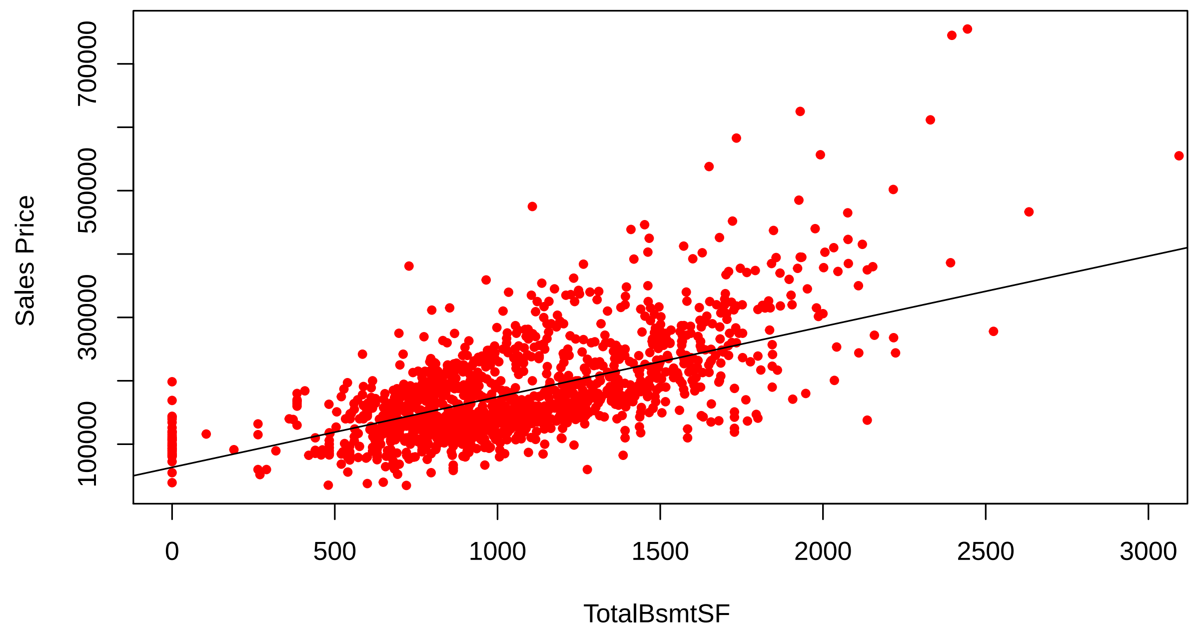


This plot clearly shows that the Living area above grade has a strong positive linear relationship with the Sale price.



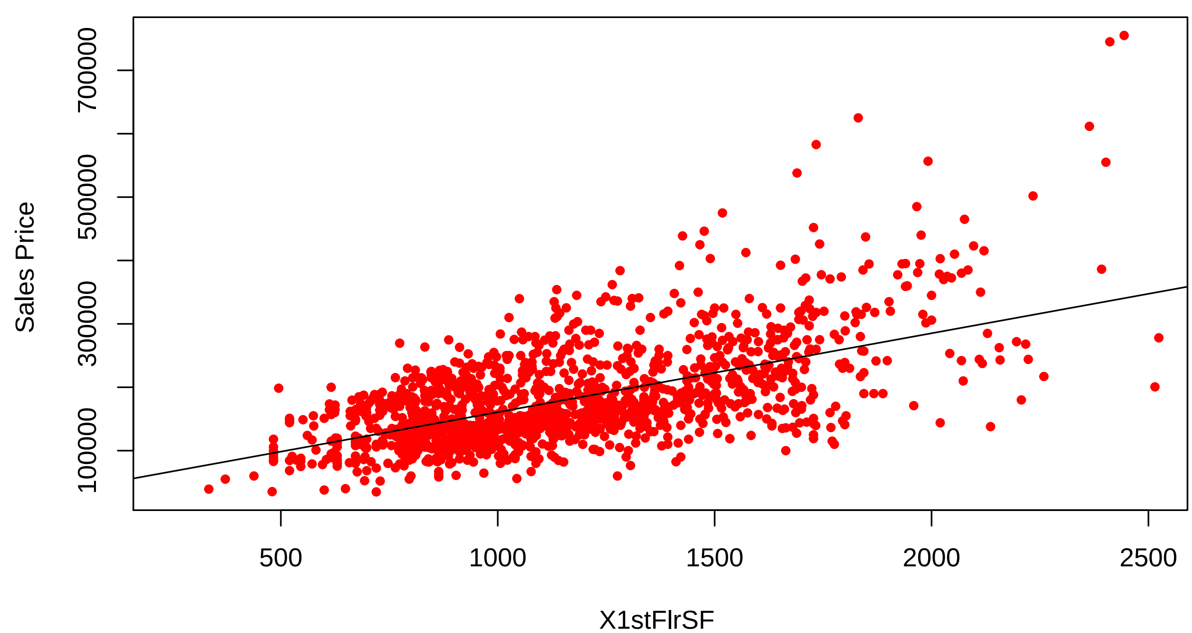
This plot clearly shows that the Garage Area has a strong positive linear relationship with the Sale price.But, this graph has lot of data points concentrated at units '0' which results in an anomaly. There are considerable amount of houses with no basement at all. That resulted in this anomaly

Scatterplot: TotalBsmtSF vs SalePrice



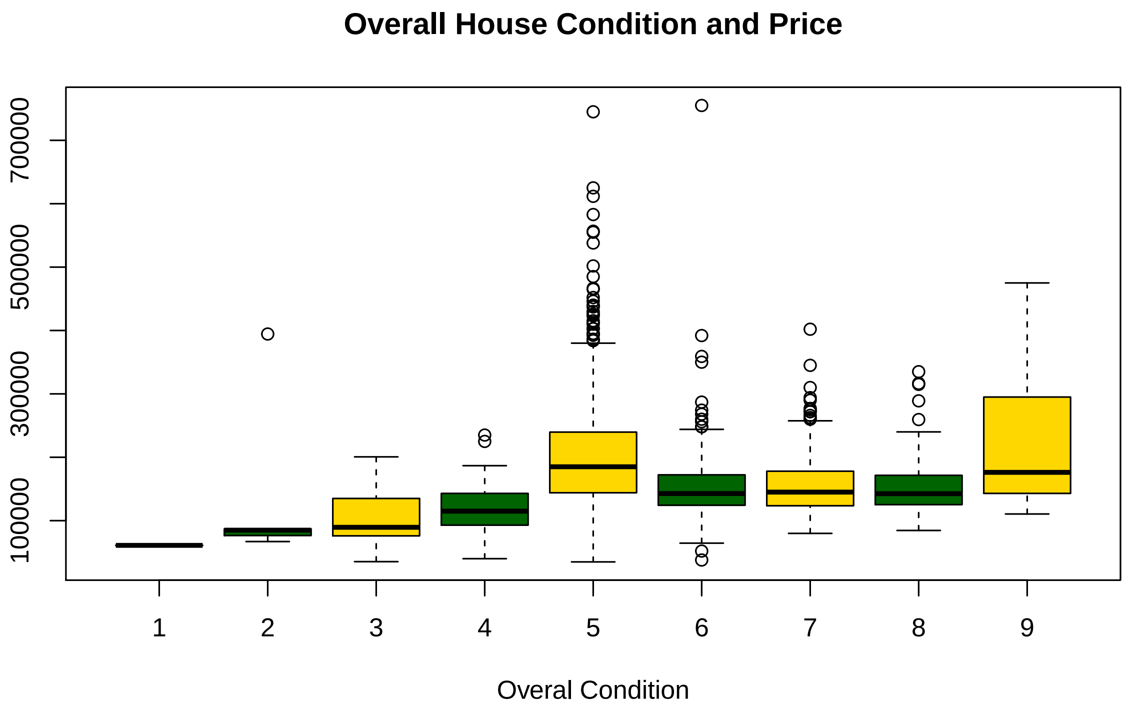
This plot clearly shows that the Total Basement Area has a strong positive linear relationship with the Sale price.But, this graph has lot of data points concentrated at units '0' which results in an anomaly. There are considerable amount of houses with no basement at all. That resulted in this anomaly

Scatterplot: X1stFlrSF vs SalePrice

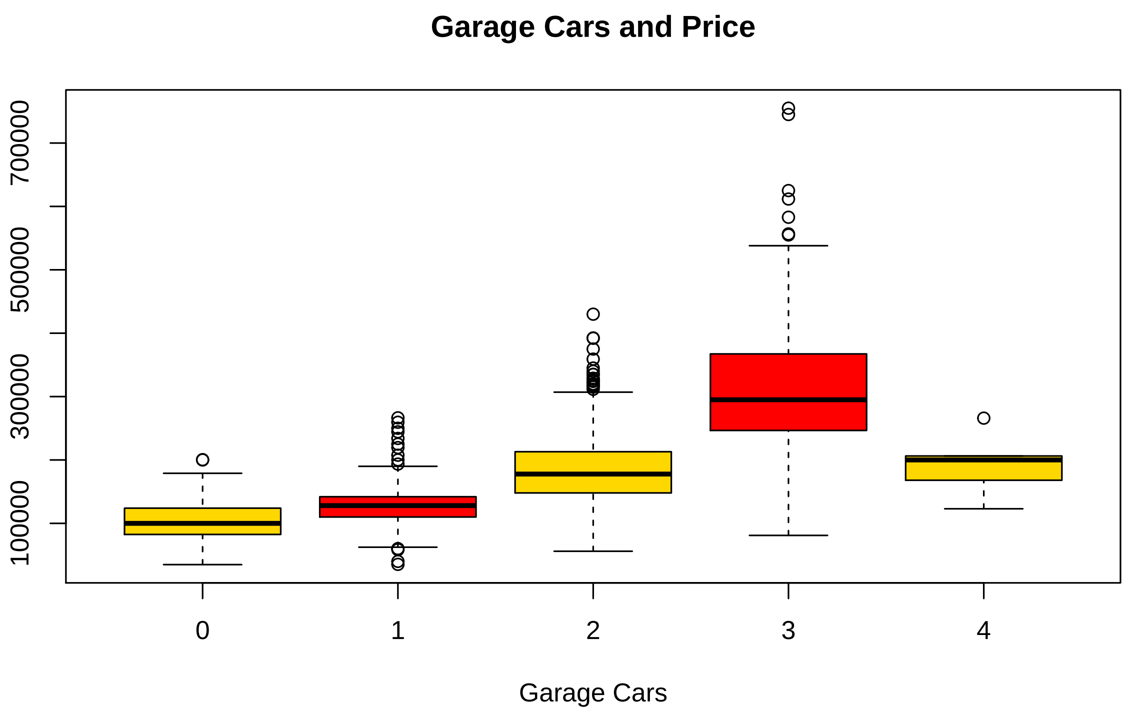


This plot clearly shows that the First Floor area has a strong positive linear relationship with the Sale price.

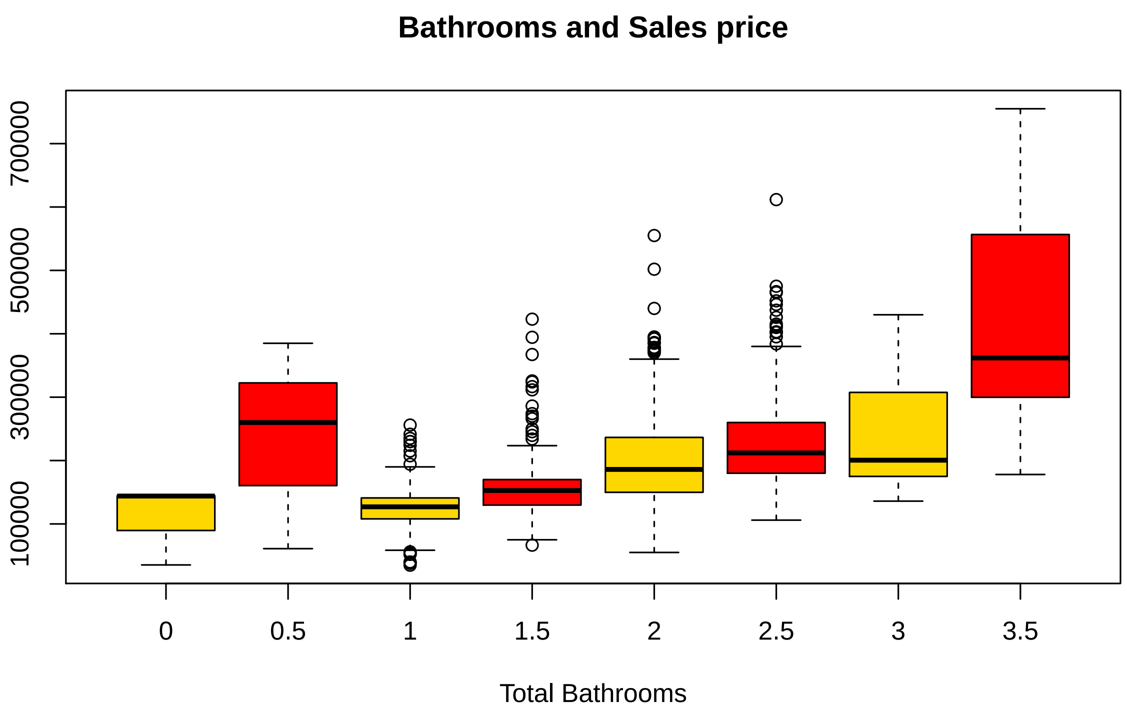
This violin plot shows probability density of the data at different values. For a house with maximum(10) Over all Quality has very high spread and distribution is close to normal where as Over all Quality with 2 has no standard probability and has minimum spread. Rest of the values has close to normal distribution with mean value increasing as the Over all Quality increase



It is quiet evident that OverallCond with 5 units has many outliers and mean sales price of houses with more than 5 rating for Over all condition is similar

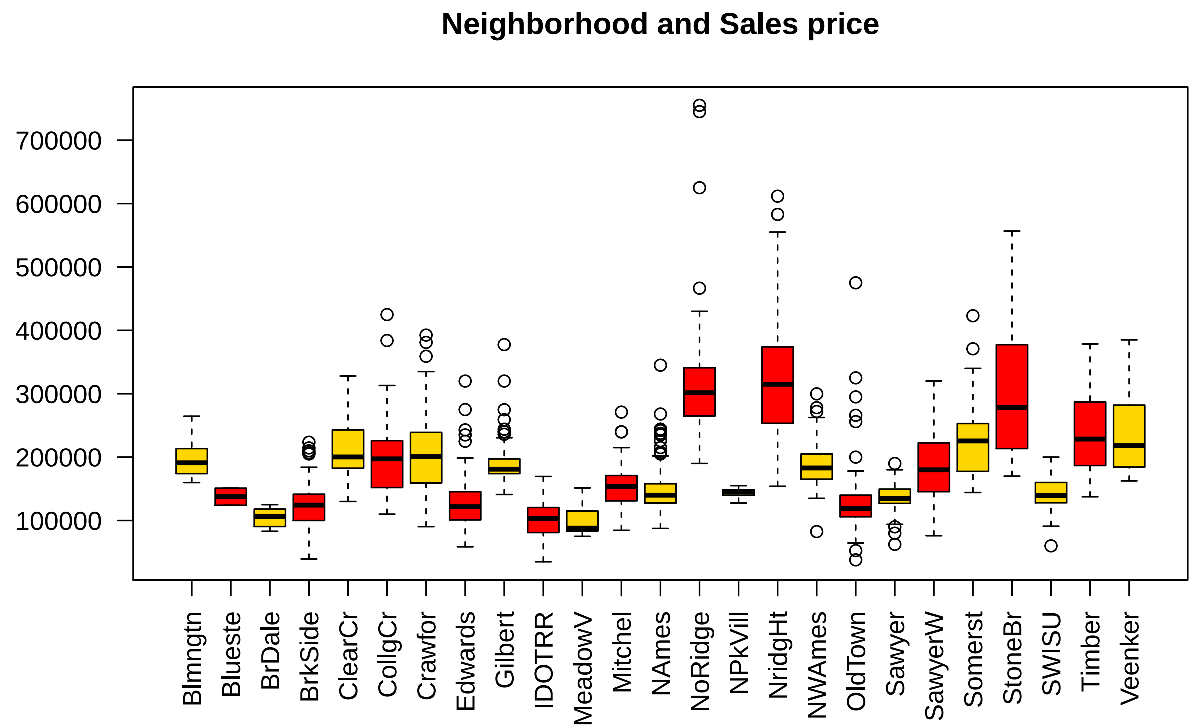


This plot shows that houses with 3 car Garage Space has suprisingly greater mean than the rest of the values

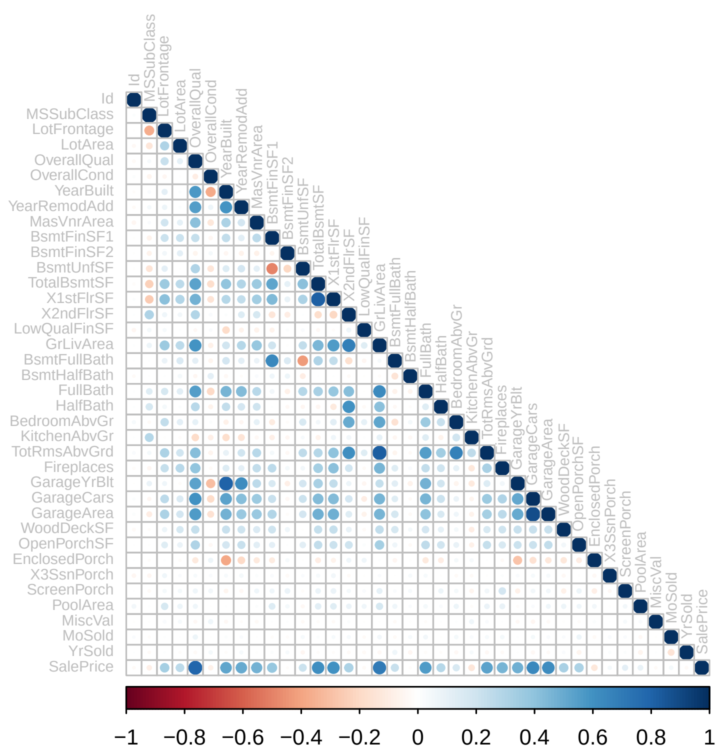


Data given has Full and Half bathrooms. Here, we combined those columns to see data so that both full and half bathroom quantity is quantized in a single value. Box plot clearly shows that prices for each value of

1,1.5, 2 and 2.5 house prices are quite similar to each other as the width of box is short



Viewing the Correlation Plot



Above Correlation heat map helps to visualize correlation between different combinations of variables

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

Following table contains the combinations of variables with highest correlation which has a minimum of 0.6 as corelation value. This will identify redundant predictors

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	0.8009778
GarageYrBlt	YearRemodAdd	0.6227175
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.

```
## RMSE of the baseline model with all predictors  32484.77
```

Base model served two purposes.

1. This helps to compare the performance of base model with the future models and see if the there is any improvement after selecting the best variables
2. This also helped in checking collinearity between categorical variables

Removing the predictor with NAs as coeffiecient because of multi colinearity. These are the predictors removed:Exterior2nd, BsmtCond, BsmtFinType1, TotalBsmtSF, Electrical, GarageFinish, GarageCond, GrLivArea, GarageQual

```
## RMSE of the model after removing multicollinear variables with all predictors  20915.33
```

Picking Top 20 predictors basing on the Beta coeffiencients and P values.

	Estimate	Std. Error	t value	Pr(>|t|)	estimate_absolute_estimates
(Intercept)	-679074.08	138425.215	-4.9057109	0.0000011	679074.08
RoofMatlMembran	650035.08	61056.197	10.6465046	0.0000000	650035.08
RoofMatlWdShngl	636276.33	51956.036	12.2464370	0.0000000	636276.33
RoofMatlMetal	610804.05	60546.158	10.0882379	0.0000000	610804.05
RoofMatlCompShg	563087.99	51141.415	11.0104107	0.0000000	563087.99
RoofMatlTar&Grv	560673.64	54918.007	10.2092860	0.0000000	560673.64
RoofMatlWdShake	555866.70	53384.769	10.4124586	0.0000000	555866.70
RoofMatlRoll	552797.21	56434.930	9.7953026	0.0000000	552797.21

	Estimate	Std. Error	t value	Pr(>|t|)	estimate_absolute_estimates
PoolQCNoPool	272708.41	117384.061	2.3232150	0.0203282	272708.41
Condition2PosN	-236158.39	26909.325	-8.7760798	0.0000000	236158.39
PoolQCFa	-167282.07	39327.646	-4.2535490	0.0000226	167282.07
PoolQCGd	-133182.53	35857.598	-3.7142067	0.0002129	133182.53
Condition2RRAe	-115621.77	64326.046	-1.7974332	0.0725093	115621.77
RoofStyleShed	87266.17	33924.016	2.5724008	0.0102147	87266.17
z.PoolArea	57853.17	17358.267	3.3328887	0.0008848	57853.17
z.X2ndFlrSF	55612.11	4892.296	11.3672823	0.0000000	55612.11
NeighborhoodStoneBr	36789.47	8095.609	4.5443730	0.0000060	36789.47
FunctionalSev	-35754.84	29267.709	-1.2216481	0.2220722	35754.84
Condition2PosA	34635.62	36750.428	0.9424548	0.3461428	34635.62
z.X1stFlrSF	34505.07	4307.566	8.0103389	0.0000000	34505.07

New model after selecting the strong predictors picked from above, and strongly corelated variables. RoofMatl, Condition2, PoolQC, OverallQual, RoofStyle, OverallCond, YearBuilt, GarageArea, GrLivArea, TotalBsmtSF

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

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

Features obtained: FullBath + Fireplaces + OverallQual + GarageCars + Neighborhood

```
## RMSE of the model with selected variables from chi-squared test 38564.29
```

Using CFS(Correlation based Feature Selection) test to pick important numercial variables.

Features obtained : OverallQual + TotalBsmtSF + GrLivArea + GarageCars

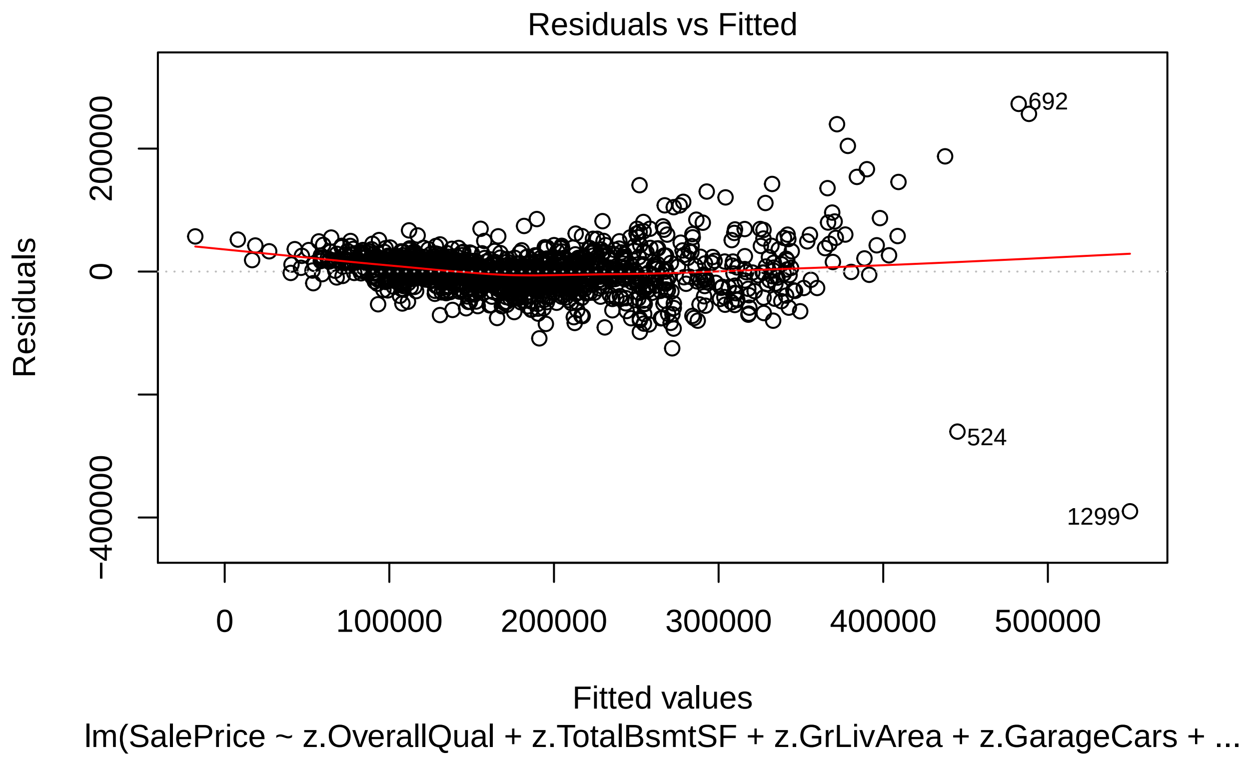
For Feature selections we used chi.squared which will find weights of discrete attributes.This shows us the most important features out of all available variables. The features obtained according to this test are : OverallQual, FullBath, Neightbourhood, Fireplace, GarageCar . So, these are most influential 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.

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

Also,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



Modifying the model futher by

1. Converting Quality variable into factor variable to take into account the bin like effect on the SalePrice
2. Adding a interaction between Neighborhood ad Quality

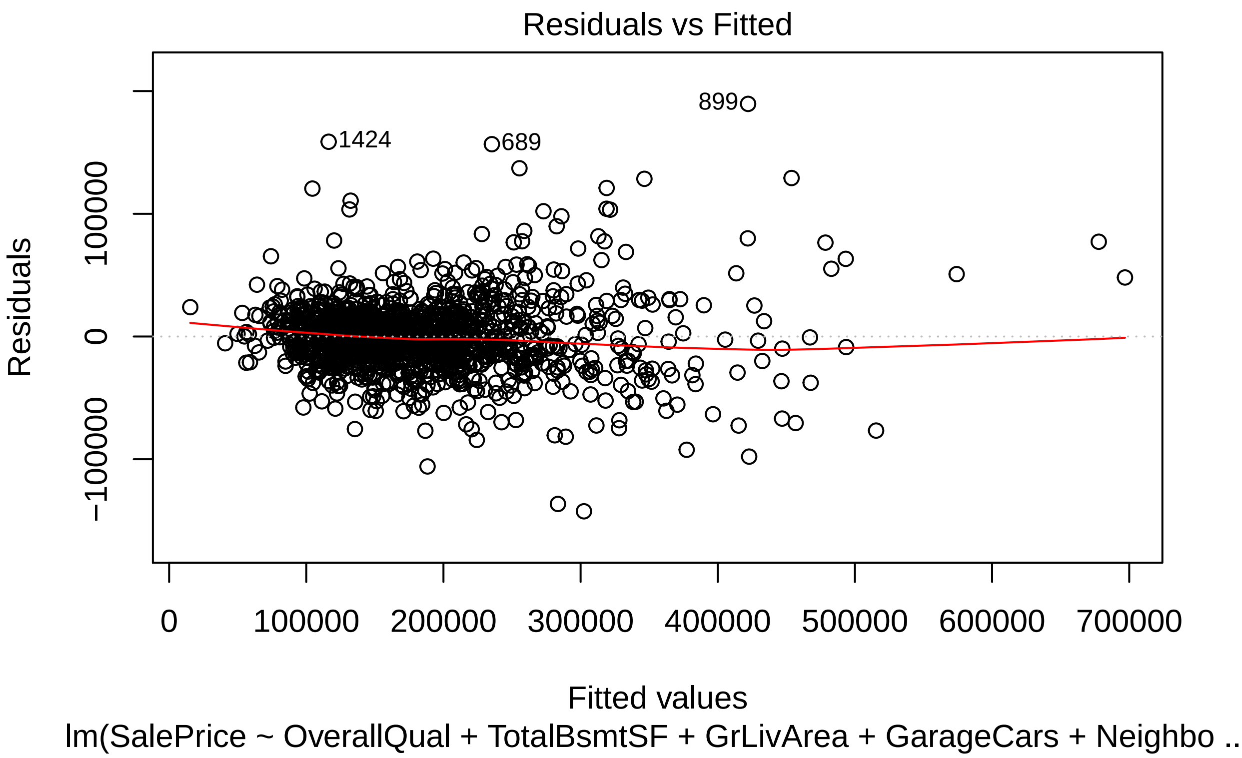
```
## RMSE of the final model with quality as factor and interaction term 27517.21
```

After analysing the model, it is evident that Neighbourhood is one of the significant factor in deciding sales price. Especially, a house in NoRidge neighbourhood with one unit more GrLivArea compared to Blmgtn results in 110061($30891-$24902+$104072) of price increase

A well known fact which is usually considered for deciding a house price is : overallqual. A house with Quality 10 cost approximately $1,62,181 more than a Quality 1 house with all other factors being same.

One unit increase in TotalBsmtSF results in an increase of approximately $26000 increase in sales prices with all other factors being same.Number of cars in Garage has good contribution to the sales price with $17640 of increase in price with every extra cars space a house has with all other factors being same.

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



Residual Plot show that the residuals and the predicted values do not follow any linear relationship. Data points are randomly distributed. This indicates that the linear model above is appropriate for the data.

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.