INDIAN STATISTICAL INSTITUTE, NEW DELHI

MASTERS OF STATISTICS

House Price Prediction Using Linear Regression

Project

Name: Saheli Datta

Roll Number: MD2213

Professor: Dr. Depayaan Sarkar

Date of Submission: 06 November 2022

Contents

1 Introduction

People are careful when they are trying to buy a new house with their budgets and market strategies. Usually, the change in price of residential houses depends on various factors such as number of rooms, neighbourhood, area of the housing property etc. Hence, it is worth analysing how price changes in order to predict house price. In this study, we will consider a secondary dataset, The Ames Housing dataset which consists of 79 different explanatory variables. Ames housing dataset is enriched with data on almost every aspect of a house. These explanatory housing parameters focus on the quality and quantity of many physical attributes of the property. Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property (e.g. When was it built? How big is the lot? How many square feet of living space is in the dwelling? Is the basement finished? How many bathrooms are there?). This paper deals with data analysis with visual representation, data imputation, data modification together with the fitting of a suitable linear regression in order to predict house prices. Based on the appropriate model we will then draw conclusions about the housing features which have significant impact on the house price.

2 Objectives

While buying a house, the most important questions a buyer may ask himself are - 'What is the price of the house?' and 'With such price, what features I may get?'. On a general note, a buyer looks for a house which is within the budget and offers as many features as it can. Based on these queries it is relevant to look for the features which are affecting the house prices and to observe how the features are affecting the prices.

Considering the Ames Housing Dataset, here our primary objective is -
To build a suitable linear regression model for predicting house price .

3 Description of the Dataset

The Ames Housing Dataset is obtained from GitHub: Ames Housing Dataset.

MSSubClass: Identifies the type of dwelling involved in the sale.

20 1-STORY 1946 and NEWER ALL STYLES

30 - 1-STORY 1945 and OLDER

40 - 1-STORY W/FINISHED ATTIC ALL AGES

45 - 1-1/2 STORY - UNFINISHED ALL AGES

50 - 1-1/2 STORY FINISHED ALL AGES

60 - 2-STORY 1946 and NEWER

70 - 2-STORY 1945 and OLDER

75 - 2-1/2 STORY ALL AGES

80 - SPLIT OR MULTI-LEVEL

85 - SPLIT FOYER

90 - DUPLEX - ALL STYLES AND AGES

120 - 1-STORY PUD (Planned Unit Development) - 1946 and NEWER

150 - 1-1/2 STORY PUD - ALL AGES

160 - 2-STORY PUD - 1946 and NEWER

180 - PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 - 2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

A - Agriculture

C - Commercial

FV - Floating Village Residential

I - Industrial

RH - Residential High Density

RL - Residential Low Density

RP - Residential Low Density Park

RM - Residential Medium Density

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Grvl - Gravel

Pave - Paved

Alley: Type of alley access to property

Grvl - Gravel

Pave - Paved

NA - No alley access

LotShape: General shape of property

Reg - Regular

IR1 - Slightly irregular

IR2 - Moderately Irregular

IR3 - Irregular

LandContour: Flatness of the property

Lvl - Near Flat/Level

Bnk - Banked - Quick and significant rise from street grade to building

HLS - Hillside - Significant slope from side to side

Low - Depression

Utilities: Type of utilities available

AllPub All public Utilities (E,G,W, and S)

NoSewr Electricity, Gas, and Water (Septic Tank)

NoSeWa Electricity and Gas Only

ELO Electricity only

LotConfig: Lot configuration

Inside - Inside lot

Corner - Corner lot

CulDSac - Cul-de-sac

FR2 - Frontage on 2 sides of property

FR3 - Frontage on 3 sides of property

LandSlope: Slope of property

Gtl Gentle slope

Mod Moderate Slope

Sev Severe Slope

Neighborhood: Physical locations within Ames city limits

Blmngtn - Bloomington Heights

Blueste - Bluestem

BrDale - Briardale

BrkSide - Brookside

ClearCr - Clear Creek

CollgCr - College Creek

Crawfor - Crawford

Edwards - Edwards

Gilbert - Gilbert

IDOTRR - Iowa DOT and Rail Road

MeadowV - Meadow Village

Mitchel - Mitchell

Names North - Ames

NoRidge - Northridge

NPkVill - Northpark Villa

NridgHt - Northridge Heights

NWAmes - Northwest Ames

OldTown - Old Town

SWISU - South and West of Iowa State University

Sawyer - Sawyer

SawyerW - Sawyer West

Somerst - Somerset

StoneBr Stone Brook

Timber - Timberland

Veenker - Veenker

Condition1: Proximity to various conditions

Artery - Adjacent to arterial street

Feedr - Adjacent to feeder street

Norm - Normal

RRNn - Within 200’ of North-South Railroad

RRAn - Adjacent to North-South Railroad

PosN - Near positive off-site feature–park, greenbelt, etc.

PosA - Adjacent to postive off-site feature

RRNe - Within 200’ of East-West Railroad

RRAe - Adjacent to East-West Railroad

Condition2: Proximity to various conditions (if more than one is present)

Artery - Adjacent to arterial street

Feedr - Adjacent to feeder street

Norm - Normal

RRNn - Within 200’ of North-South Railroad

RRAn - Adjacent to North-South Railroad

PosN - Near positive off-site feature–park, greenbelt, etc.

PosA - Adjacent to postive off-site feature

RRNe - Within 200’ of East-West Railroad

RRAe - Adjacent to East-West Railroad

BldgType: Type of dwelling

1Fam - Single-family Detached

2FmCon - Two-family Conversion; originally built as one-family dwelling

Duplx - Duplex

TwnhsE - Townhouse End Unit

TwnhsI - Townhouse Inside Unit

HouseStyle: Style of dwelling

1Story - One story

1.5Fin - One and one-half story: 2nd level finished

1.5Unf - One and one-half story: 2nd level unfinished

2Story - Two story

2.5Fin - Two and one-half story: 2nd level finished

2.5Unf - Two and one-half story: 2nd level unfinished

SFoyer - Split Foyer

SLvl - Split Level

OverallQual: Rates the overall material and finish of the house

10 - Very Excellent

9 - Excellent

8 - Very Good

7 - Good

6 - Above Average

5 - Average

4 - Below Average

3 - Fair

2 - Poor

1 - Very Poor

OverallCond: Rates the overall condition of the house

10 - Very Excellent

9 - Excellent

8 - Very Good

7 - Good

6 - Above Average

5 - Average

4 - Below Average

3 - Fair

2 - Poor

1 - Very Poor

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

Flat - Flat

Gable - Gable

Gambrel - Gabrel (Barn)

Hip - Hip

Mansard - Mansard

Shed - Shed

RoofMatl: Roof material

ClyTile - Clay or Tile

CompShg - Standard (Composite) Shingle

Membran - Membrane

Metal - Metal

Roll - Roll

Tar&Grv - Gravel and Tar

WdShake - Wood Shakes

WdShngl - Wood Shingles

Exterior1st: Exterior covering on house

AsbShng - Asbestos Shingles

AsphShn - Asphalt Shingles

BrkComm - Brick Common

BrkFace - Brick Face

CBlock - Cinder Block

CemntBd - Cement Board

HdBoard - Hard Board

ImStucc - Imitation Stucco

MetalSd - Metal Siding

Other - Other

Plywood - Plywood

PreCast - PreCast

Stone - Stone

Stucco- Stucco

VinylSd - Vinyl Siding

Wd Sdng - Wood Siding

WdShing - Wood Shingles

Exterior2nd: Exterior covering on house (if more than one material)

AsbShng - Asbestos Shingles

AsphShn - Asphalt Shingles

BrkComm - Brick Common

BrkFace - Brick Face

CBlock - Cinder Block

CemntBd - Cement Board

HdBoard - Hard Board

ImStucc - Imitation Stucco

MetalSd - Metal Siding

Other - Other

Plywood - Plywood

PreCast - PreCast

Stone - Stone

Stucco- Stucco

VinylSd - Vinyl Siding

Wd Sdng - Wood Siding

WdShing - Wood Shingles

MasVnrType: Masonry veneer type

BrkCmn - Brick Common

BrkFace - Brick Face

CBlock - Cinder Block

None - None

Stone - Stone

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

Ex - Excellent

Gd - Good

TA - Average/Typical

Fa - Fair

Po - Poor

ExterCond: Evaluates the present condition of the material on the exterior

Ex - Excellent

Gd - Good

TA - Average/Typical

Fa - Fair

Po - Poor

Foundation: Type of foundation

BrkTil - Brick & Tile

CBlock - Cinder Block

PConc - Poured Contrete

Slab - Slab

Stone - Stone

Wood - Wood

BsmtQual: Evaluates the height of the basement

Ex - Excellent (100+ inches)

Gd - Good (90-99 inches)

TA - Typical (80-89 inches)

Fa - Fair (70-79 inches)

Po - Poor (¡70 inches)

NA - No Basement

BsmtCond: Evaluates the general condition of the basement

Ex - Excellent

Gd - Good

TA - Typical - slight dampness allowed

Fa - Fair - dampness or some cracking or settling

Po - Poor - Severe cracking, settling, or wetness

NA - No Basement

BsmtExposure: Refers to walkout or garden level walls

Gd - Good Exposure

Av - Average Exposure (split levels or foyers typically score average or above)

Mn - Mimimum Exposure

No - No Exposure

NA - No Basement

BsmtFinType1: Rating of basement finished area

GLQ - Good Living Quarters

ALQ - Average Living Quarters BLQ - Below Average Living Quarters

Rec - Average Rec Room

LwQ - Low Quality

Unf - Unfinshed

NA - No Basement

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

GLQ - Good Living Quarters

ALQ - Average Living Quarters BLQ - Below Average Living Quarters

Rec - Average Rec Room

LwQ - Low Quality

Unf - Unfinshed

NA - No Basement

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

Heating: Type of heating

Floor - Floor Furnace

GasA - Gas forced warm air furnace

GasW - Gas hot water or steam heat

Grav - Gravity furnace

OthW - Hot water or steam heat other than gas

Wall - Wall furnace

HeatingQC: Heating quality and condition

Ex - Excellent

Gd - Good

TA - Average/Typical

Fa - Fair

Po - Poor

CentralAir: Central air conditioning

N - No

Y - Yes

Electrical: Electrical system

SBrkr : Standard Circuit Breakers & Romex

FuseA : Fuse Box over 60 AMP and all Romex wiring (Average)

FuseF : 60 AMP Fuse Box and mostly Romex wiring (Fair)

FuseP : 60 AMP Fuse Box and mostly knob & tube wiring (poor)

Mix : Mixed

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Ex - Excellent

Gd - Good

TA - Average/Typical

Fa - Fair

Po - Poor

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

Functional: Home functionality (Assume typical unless deductions are warranted)

Typ - Typical Functionality

Min1 - Minor Deductions 1

Min2 - Minor Deductions 2

Mod - Moderate Deductions

Maj1 - Major Deductions 1

Maj2 - Major Deductions 2

Sev - Severely Damaged

Sal - Salvage only

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

Ex - Excellent - Exceptional Masonry Fireplace

Gd - Good - Masonry Fireplace in main level

TA - Average - Prefabricated Fireplace in main living area or Masonry Fireplace in basement

Fa - Fair - Prefabricated Fireplace in basement

Po - Poor - Ben Franklin Stove

NA - No Fireplace

GarageType: Garage location

2Types - More than one type of garage

Attchd - Attached to home

Basment - Basement Garage

BuiltIn - Built-In (Garage part of house - typically has room above garage)

CarPort - Car Port

Detchd - Detached from home

NA - No Garage

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

Fin - Finished

RFn - Rough Finished

Unf - Unfinished

NA - No Garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

Ex - Excellent

Gd - Good

TA - Average/Typical

Fa - Fair

Po - Poor

NA - No Garage

GarageCond: Garage condition

Ex - Excellent

Gd - Good

TA - Average/Typical

Fa - Fair

Po - Poor

NA - No Garage

PavedDrive: Paved driveway

Y - Paved

P - Partial Pavement

N - Dirt/Gravel

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Ex - Excellent

Gd - Good

TA - Average/Typical

Fa - Fair

Po - Poor

NA - No Pool

Fence: Fence quality

GdPrv - Good Privacy

MnPrv - Minimum Privacy

GdWo - Good Wood

MnWw - Minimum Wood/Wire

NA - No Fence

MiscFeature: Miscellaneous feature not covered in other categories

Elev - Elevator

Gar2 - 2nd Garage (if not described in garage section)

Othr - Other

Shed - Shed (over 100 SF)

TenC - Tennis Court

NA - None

MiscVal: Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

WD - Warranty Deed - Conventional

CWD - Warranty Deed - Cash

VWD - Warranty Deed - VA Loan

New - Home just constructed and sold

COD - Court Officer Deed/Estate

Con - Contract 15 percent Down payment regular terms

ConLw - Contract Low Down payment and low interest

ConLI - Contract Low Interest

ConLD - Contract Low Down

Oth - Other

SaleCondition: Condition of sale

Normal - Normal Sale

Abnorml - Abnormal Sale - trade, foreclosure, short sale

AdjLand - Adjoining Land Purchase

Alloca - Allocation - two linked properties with separate deeds, typically condo with a garage unit

Family - Sale between family members

Partial - Home was not completed when last assessed (associated with New Homes)

The Nature of the 79 predicting features and 1 response variable i.e. Sale Price of House is given below -

Table 1: Nature of the Variables in the Dataset

Variable Type	Number of Variables
Continuous	20
Discreet	14
Categorical(nominal)	23
Categorical(ordinal)	23
Total	80

The 20 continuous variables relate to various area dimensions for each observation. In addition to the typical lot size and total dwelling square feet found on most common home listings, other more specific variables are quantified in the data set. Area measurements on the basement, main living area, and porches are broken down into individual categories based on quality and type.

The 14 discrete variables typically quantify the number of items occurring within the house. Most are specifically the number of kitchens, bedrooms, and bathrooms (full and half) located in the basement and above grade (ground) living areas of the home. Additionally, the garage capacity and construction/remodelling dates are also recorded.

There are a large number of categorical variables (23 nominal, 23 ordinal) associated with this data set. They range from 2 to 28 classes with the smallest being STREET (gravel or paved) and the largest being NEIGHBOURHOOD (areas within the Ames city limits). The nominal variables identify various types

of dwellings, garages, materials, and environmental conditions while the ordinal variables typically rate various items within the property.

The dataset lists 1459 house sale prices, starting from 2006 to 2010, with houses having stand-alone garages, condos and storage areas.

4 Methodology

STEP 1 :

we analyse the whole dataset containing our response variable House Price and 79 predicting features to get an idea about the features and the interrelationships that are present between them. For this, we use different data visualization techniques such as histogram, scatterplot etc. and descriptive measures namely, skewness, correlation etc.

STEP 2 :

Second step is to modify the dataset which includes imputing missing observations, redefining some categorical levels, transforming some of the predictors to our preferences.

STEP 3 :

we build a suitable linear regression model by the method of 'Ordinary Least Squares' using the transformed/ untransformed predictors. We check the accuracy, presence of multicollinearity and validity of the primary predicting model via different diagnostics tests and plots. After finding out the key factors that are affecting the primary model, a model modification is done (if needed) in such a way that it validates the model assumptions and predicts house prices with higher accuracy. Then, we run a hypothesis testing on the predicting parameters and conclude the predictors which have a significant effect on determining House Prices.

- **Multiple linear regression:**

  Regression analysis is a technique used in statistics for investigating and modelling the relationship between variables.Simple linear regression is a model with a single regressor x that has a relationship with a response y that is a straight line. This simple linear regression model can be expressed as

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x} + \varepsilon$$

  where the intercept β_0 and the slope β_1 are unknown constants and ε is a random error component .

  If there is more than one regressor, it is called multiple linear regression. In general, the response variable y may be related to k regressors, x1, x2,. . . ,xk, so that

$$\mathrm{y} = \beta_0 + \beta_1 \mathrm{x1} \quad + \ldots + \beta_k \mathrm{xk} + \varepsilon$$

  Model Assumptions:

  Linearity: The Model is linear in parameters.

  Homoscedasticity: Error variance remains constant for different values of predictors.

  Independence: The errors are identically distributed.

  Normality: The errors are normally distributed.

- **Ordinary Least Squares Estimation:**

  The method of least squares is used to estimate β_0, β_1,... β_k. That is, we

estimate β_i's so that the sum of the squares of the differences between the observations yi and predicted values of yi's is a minimum

- **R-squared:**

  R-squared is a measure in statistics of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determinations for multiple regression. It is the percentage variation of the response variable variation that is explained by a linear model.

$$R-squared = \frac{Explained\ Variation}{Total\ Variation}$$

  R-squared is always between 0 and 100%. 0% means the model explains none of the variability of the response data around its mean. 100% indicates that the model explains all the variability of the response data around its mean. Generally, the higher the R-squared, the better the model fits the data.

- **Testing of Hypothesis on Individual Regression Coefficients (t test):**

  Statistical hypothesis are statements about relationships. The statistical hypothesis testing is the use of statistics to determine the probability that a given hypothesis is true. The null hypothesis is denoted by H_0. The alternative hypothesis is the negation of the null hypothesis, denoted by H_1.

  The t-test is used to check the significance of individual regression coefficients in the multiple linear regression model. Adding a significant variable to a regression model makes the model more effective, while adding an unimportant variable may make the model worse. The hypothesis statements to test the significance of a particular regression coefficient, β_j , are:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

The test statistic for this test, under null hypothesis, has the t-distribution with parameter (n-2):

$$T = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

where the estimate of the standard error of β_j is se($\hat{\beta}_j$). We reject the null hypothesis if the test statistic lies in the critical region:

$$W : \{|T_{obs}| > t_{\frac{\alpha}{2},n-2}\}$$

Where α is the desired level of significance.

The rejection of this Null Hypothesis implies that the corresponding predicting variable has significant effect on predicting response variable y.

- **P-value:**

  P-value is a measure of the probability that an observed difference could have occurred just by random chance. A P-value lesser than α implies that the test rejects the null hypothesis. For two-sided test,

$$Pvalue = 2min\{P(T > T_{obs}), P(T < T_{obs})\}$$

- **Residual Diagnostics:**

  1. RESIDUAL PLOT:

A residual plot is a graph that shows the residuals on the vertical axis and the predicted values of the response on the horizontal axis. If the points in a residual plot are randomly dispersed over the graph, a linear regression model is free from heteroscedasticity i.e., the errors are not correlated. Otherwise, if a pattern is noticed in the plot, we may say that the errors are correlated.

2. Q-Q PLOT:

Q-Q plot is a graph that shows the observed sample quantiles on vertical axis corresponding to theoretical normal quantiles on horizontal axis. Q-Q plot is used to detect whether the residuals are normally distributed.

3. LASSO:

LASSO stands for "Least Absolute Shrinkage and Selection Operator". The dataset contains a large number of predictors and some of them are highly correlated. Now, to find the most important predictors that explains most of the variability of response, one should use penalised regression such as LASSO. LASSO incorporates a penalty term with the Ordinary Least square Loss Function. The penalised term is such that it shrinks the regression coefficients estimators (which are correlated with others or do not have much influence over response) to zero, thus reducing the number of covariates.
The LASSO estimator given by solving a$L1$ penalized regression with a penalty .

$$\hat{\beta} = argmin \frac{\sum_{i=1}^{n} (y_i - {x_i}^T \beta)^2}{2} + \lambda \sum_{j=1}^{p} |\beta_j|$$

where λ is the tuning parameter obtained by suitable Cross Validation method.

5 Analysis of the Dataset

In this section, we will analyse the whole dataset using suitable measures and graphical representations.

5.1 Analysis of the Response

Here, our objective is to predict the House Price per Square feet.
We have data on Price and also have data on the size of the property. From there we can easily find out the price per square feet of an individual house.

Clearly, House Price is the response variable in the model. In the given dataset, the House Price is denoted by 'SalePrice' of the house. In order to fit a suitable regression model it is necessary to analyse our response variable to figure out what characteristics do the response variable posses.
The response is a continuous variable. The unit of measurement of the prices is Dollar.

Below the histogram of 'SalePrice' is shown:

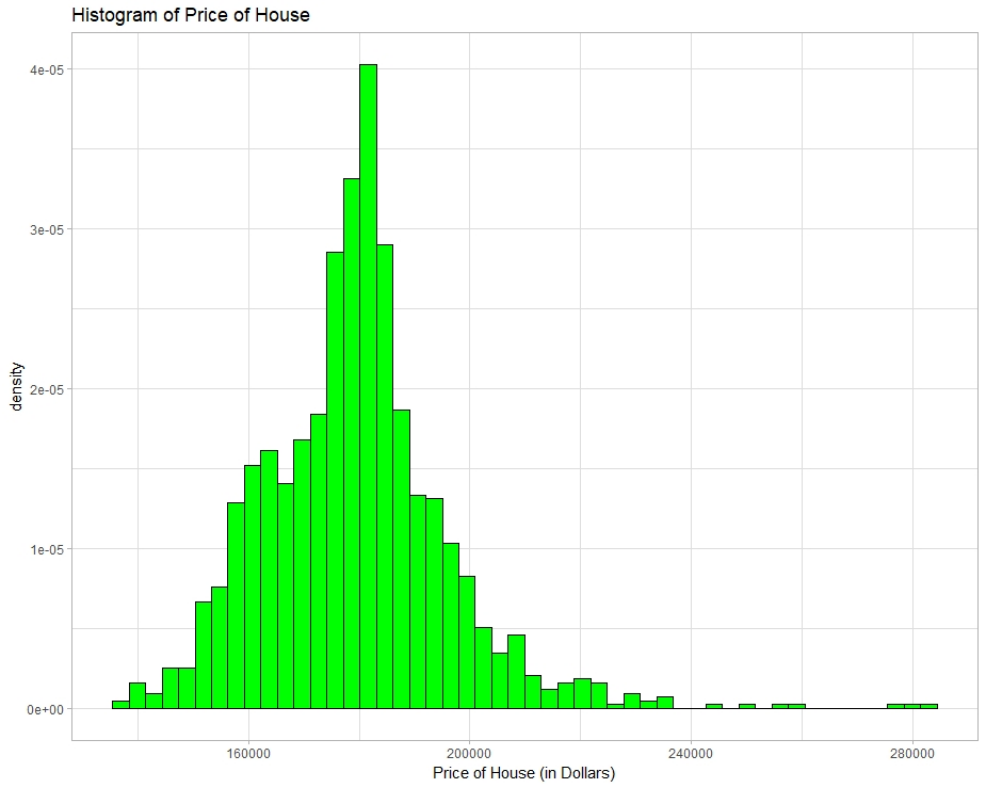

Figure 1: Histogram of Price of House

From the histogram in Figure 1, we observe that the response variable is positively skewed.

5.2 Analysis of Predictors

Given the dataset, we have 78 different housing parameters to be used in the regression model. We must take a look on the numerical features, who might have a great impact on determining House Prices.

A Distributional description of the numerical Features:

A summary of the numerical predictors are given below:

Table 2: Summary of the numerical predictors:

predictor	0%	25%	50%	75%	100%	Mean	SD	skenewss	kurtosis
LotFrontage	0	1956	1977	2001	2207	1871.99	445.79	-3.95	16.65
YearRemodAdd	1950	1963	1992	2004	2010	1983.66	21.13	-0.4	1.59
MasVnrArea	0	1	2	2	5	1.76	0.78	-0.11	3.24
BsmtFinSF1	0	1	2	2	4	1.57	0.56	0.3	2.76
BsmtFinSF2	0	317.5	480	576	1488	472.44	217.33	0.3	3.95
BsmtUnfSF	0	2	3	3	6	2.85	0.83	0.44	4.68
TotalBsmtSF	21	60	70	80	200	68.96	21	0.62	6.08
1stFlrSF	0	0	0	1	3	0.43	0.53	0.65	2.36
2ndFlrSF	0	0	0	1	2	0.38	0.5	0.71	2.01
LowQualFinSF	0	784	988	1304	5095	1045.4	443.59	0.8	8.15
GrLivArea	0	0	0	1	4	0.58	0.65	0.82	3.38
BsmtFullBath	3	5	6	7	15	6.39	1.51	0.84	4.51
BsmtHalfBath	0	0	0	676	1862	325.97	420.61	0.91	2.72
FullBath	0	219	460	797.5	2140	553.92	437.35	0.92	3.33
HalfBath	407	1117.5	1432	1721	5095	1486.05	485.57	1.13	5.91
BedroomAbvGr	0	0	350	752	4010	438.9	455.26	1.17	5.66
KitchenAbvGr	407	873.5	1079	1382.5	5095	1156.53	398.17	1.56	11.02
TotRmsAbvGrd	0	0	0	168	1424	93.17	127.74	2.13	13.21
Fireplaces	0	0	0	162	1290	99.67	177	2.55	11.45
GarageYrBlt	0	0	28	72	742	48.31	68.88	2.69	15.96
GarageCars	4.9	15.92	19.22	23.59	112.8	22.97	14.9	3.33	15.7
GarageArea	0	0	0	0	2	0.07	0.25	3.78	16.53
WoodDeckSF	0	0	0	0	576	17.06	56.61	3.78	20.18
OpenPorchSF	0	0	0	0	1526	52.58	176.7	4.04	20.62
EnclosedPorch	0	1	1	1	2	1.04	0.21	4.07	20.41
3SsnPorch	0	0	0	0	1012	24.24	67.23	4.66	42.99
ScreenPorch	0	0	0	0	360	1.79	20.21	12.51	172.61
PoolArea	0	0	0	0	1064	3.54	44.04	16.15	310.62
MiscVal	0	0	0	0	17000	58.17	630.81	20.05	472.9

Clearly, from the above table it can be observed that many of the explanatory variables are highly positively skewed.

5.3 Relationship Between Different Housing Parameters and Response in the Dataset

In order to choose a suitable linear regression model it is necessary to grab an idea about how the response and the explanatory variables are related. We may use graphical representations and correlation heatmap to understand how the variables are interrelated. We may also get a brief idea whether the explanatory variables are related.

Such overview of the whole dataset is important because in this was we may know how to treat the variables or how to transform them or how to choose proper explanatory variables in the regression model.

1. Scatterplot of Price vs Area of the property:

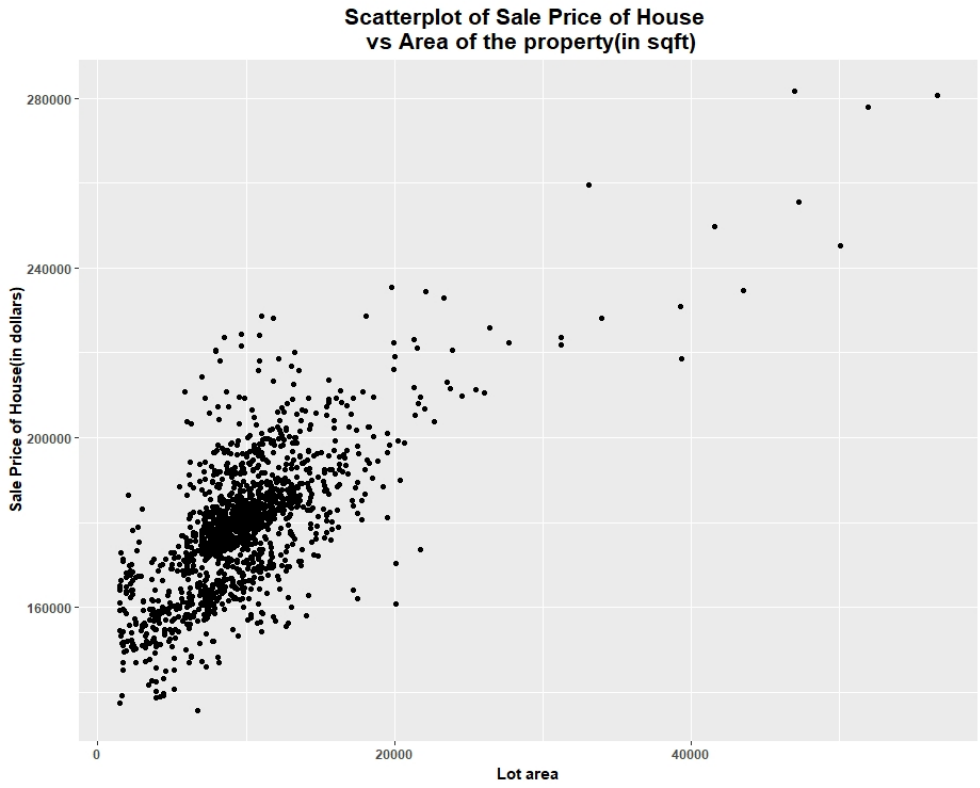

Figure 2: Scatterplot of Price of House vs Area of the property

Figure 2 shows that price of house increases as area of the property increases.

2. Boxplot of Price vs Number of Bedrooms in the House:

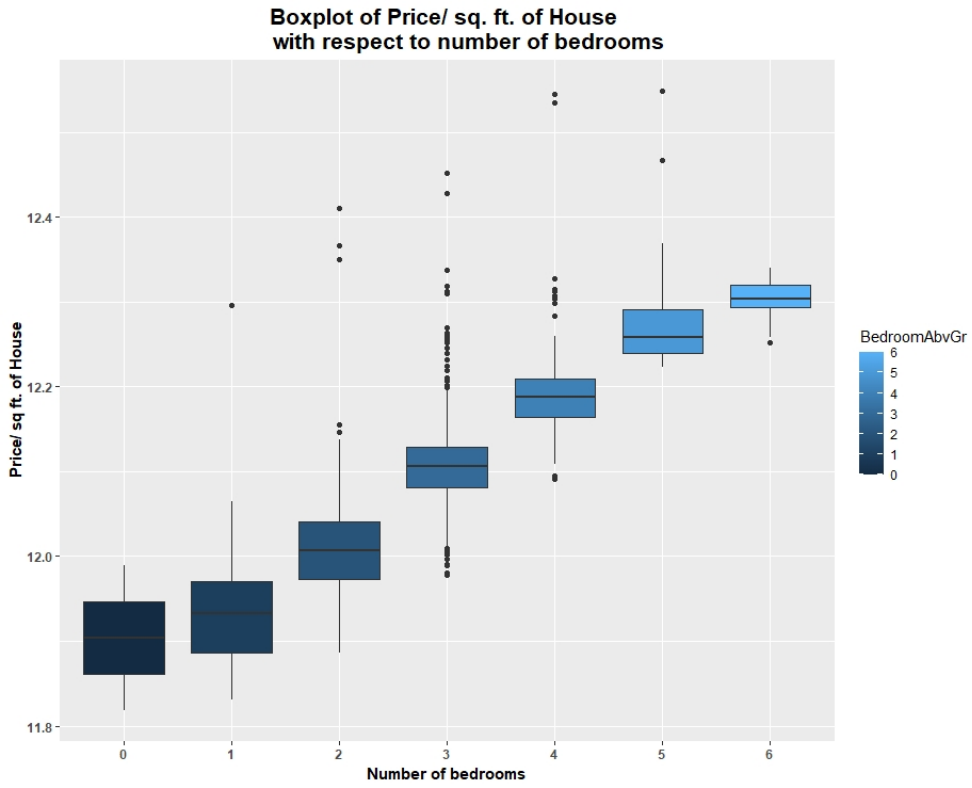

Figure 3: Boxplot of Price of House vs Number of Bedrooms

Figure 3 shows that houses with 6 bedrooms have average price higher than houses with lesser bedrooms. House price increases as number of bedrooms increases.

3. Boxplot of Price vs Neighborhood of the House:

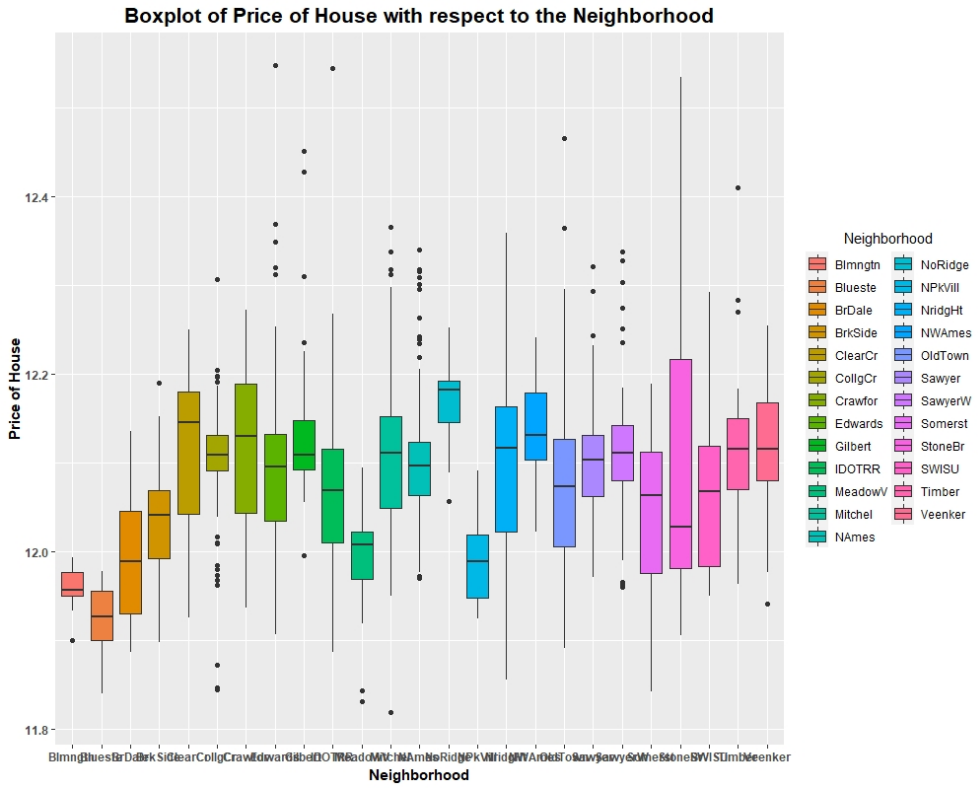

Figure 4: Boxplot of Price of House vs Neighborhood of the House

From Figure 4, we observe that houses around MeadowV neighborhood have significantly high selling price/ sqft.

4. Boxplot of Price vs Quality of House:

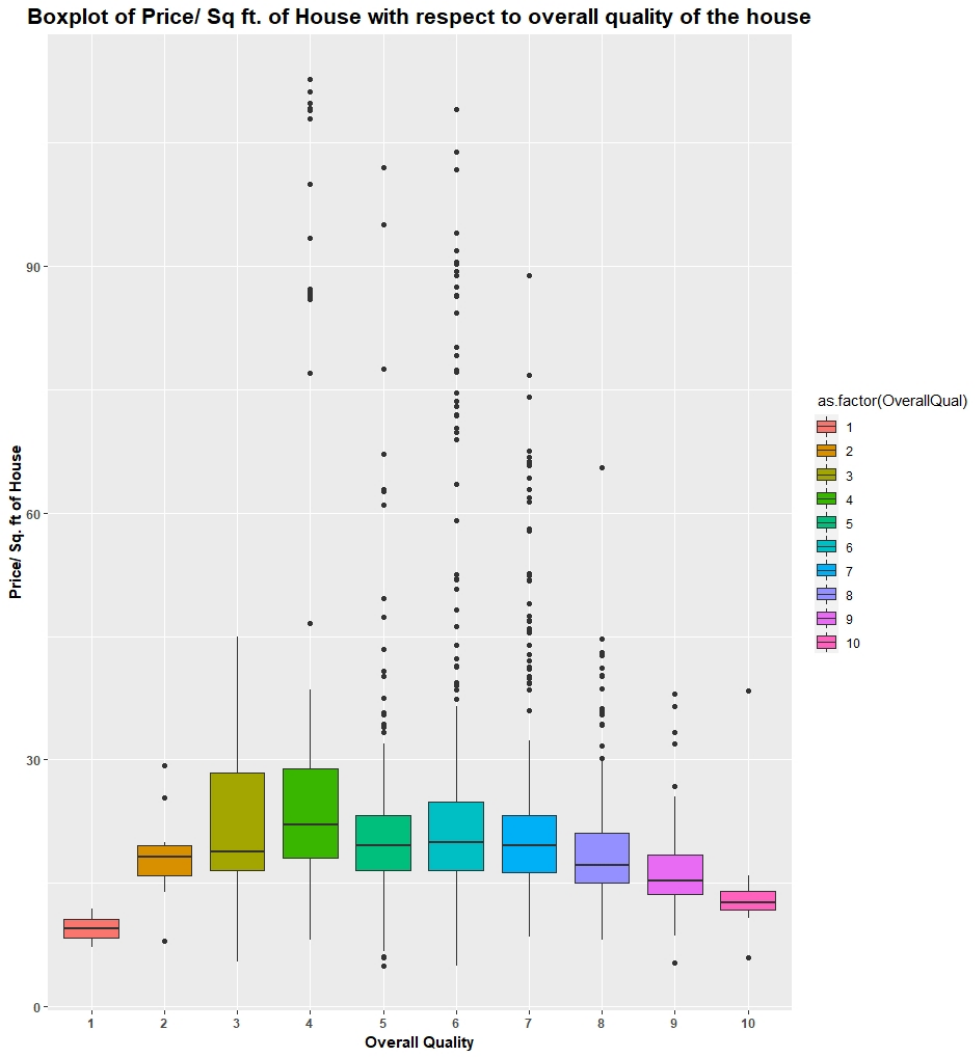

Figure 5: Boxplot of Price of House vs Quality of House

From Figure 5, we observe that excellent quality house has average price lower than average quality houses!

5. Boxplot of Price vs Type of Dwelling:

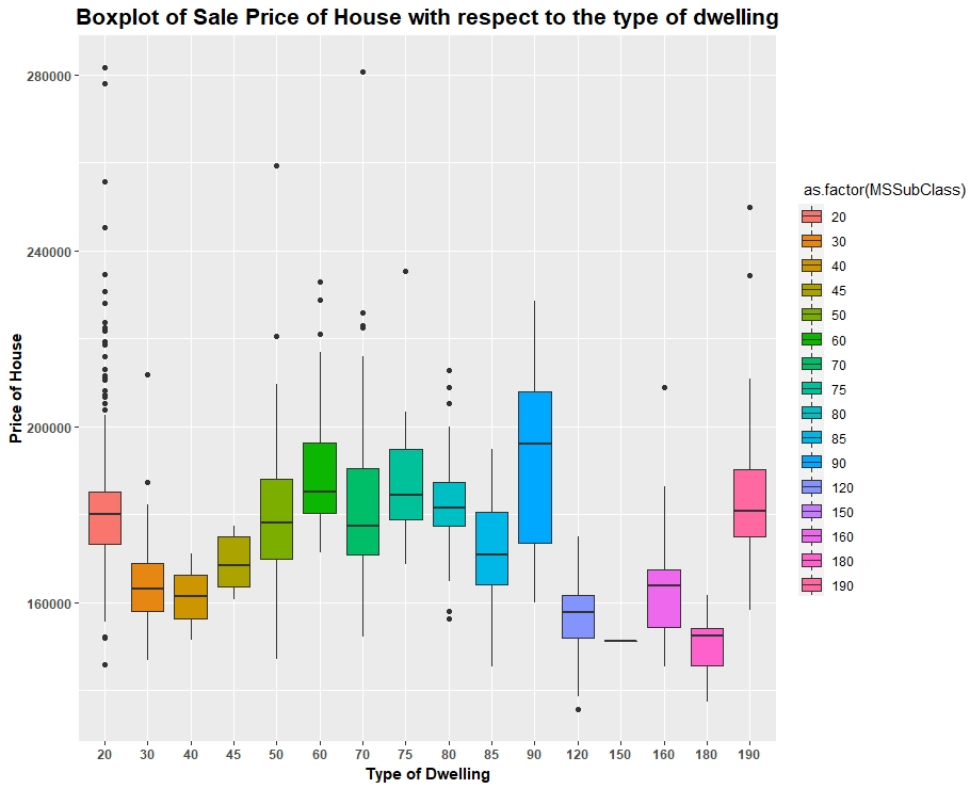

Figure 6: Boxplot of Price of House vs Type of Dwelling

From the above figure we observe that,2-STORY PUD and PUD - MULTI-LEVEL houses have per sqft price very high.

Analysis of Price of House with respect to the Age of the House

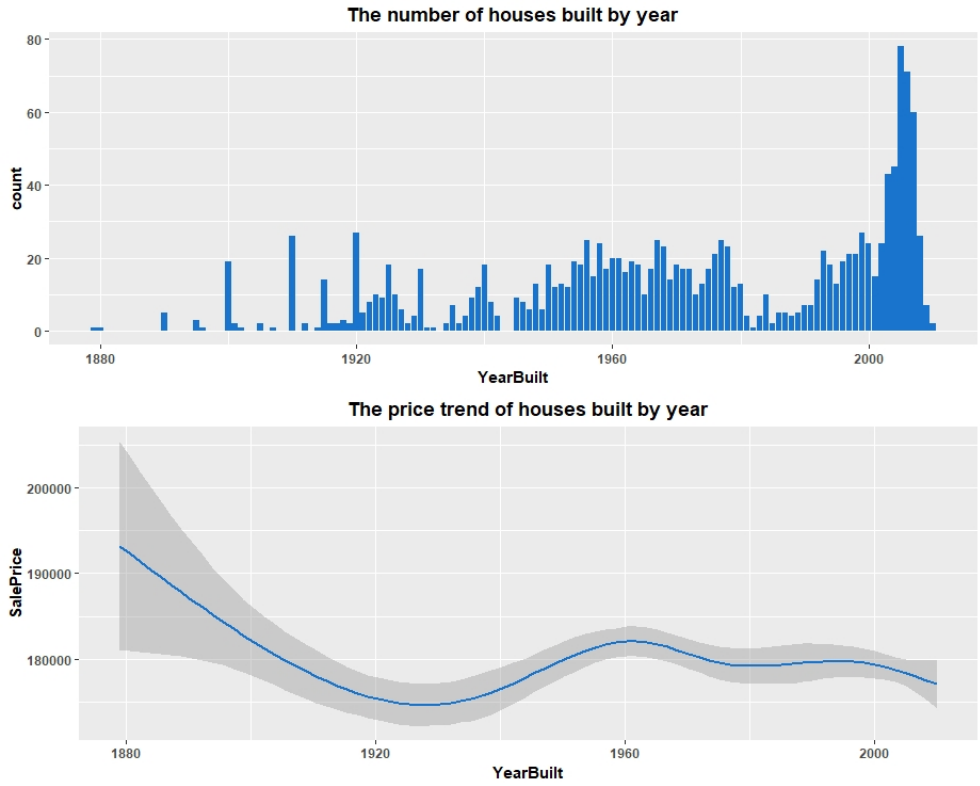

Figure 7: The number of houses built over years and their prices

In recent years that is around the 2000s, the making of house had increased in Ames. Also Older houses have sale prices very high comparatively to the new ones.

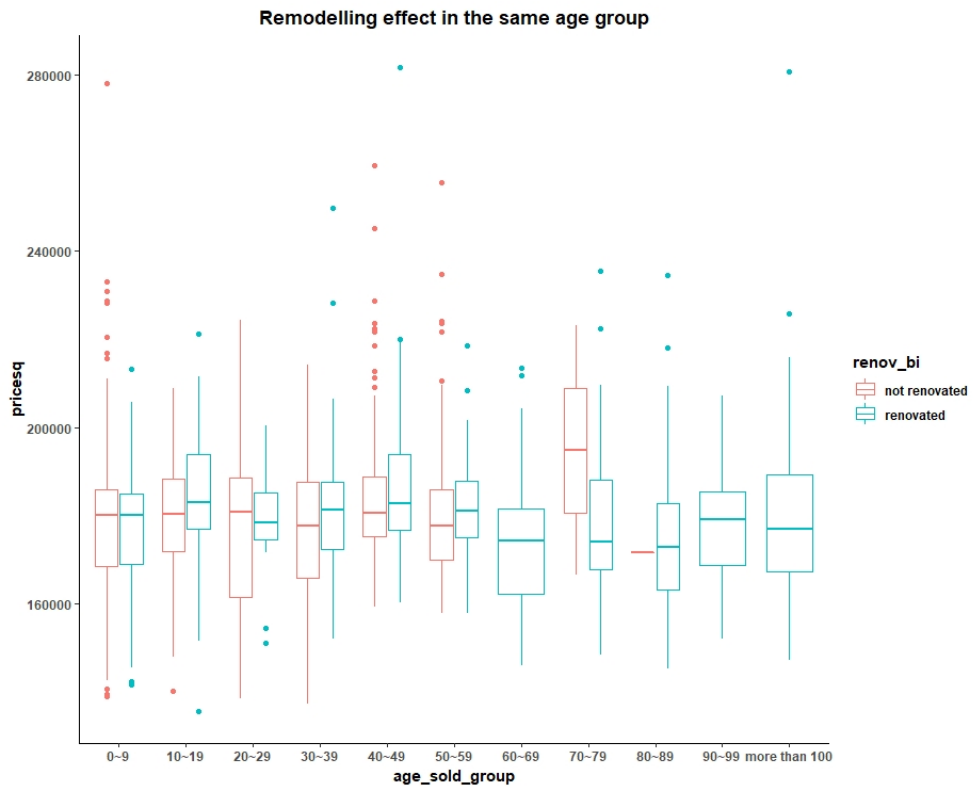

Figure 8: Boxplot of House Prices with respect to different age groups of House and whether they were renovated or not

It can noticed that average price remodelled house and not remodelled house is more or less same.

Correlation Heatmap:

Here, the Correlation Heatmap is a graphical representation of total correlation between all possible numerical variables, present in the dataset. We cannot detect multicollinearity directly from correlation heatmap. But those predictors which have a high magnitude of correlation among themselves, may cause multicollinearity in the model. Hence, we may ignore one of the variable which have high correlation with another predictor.

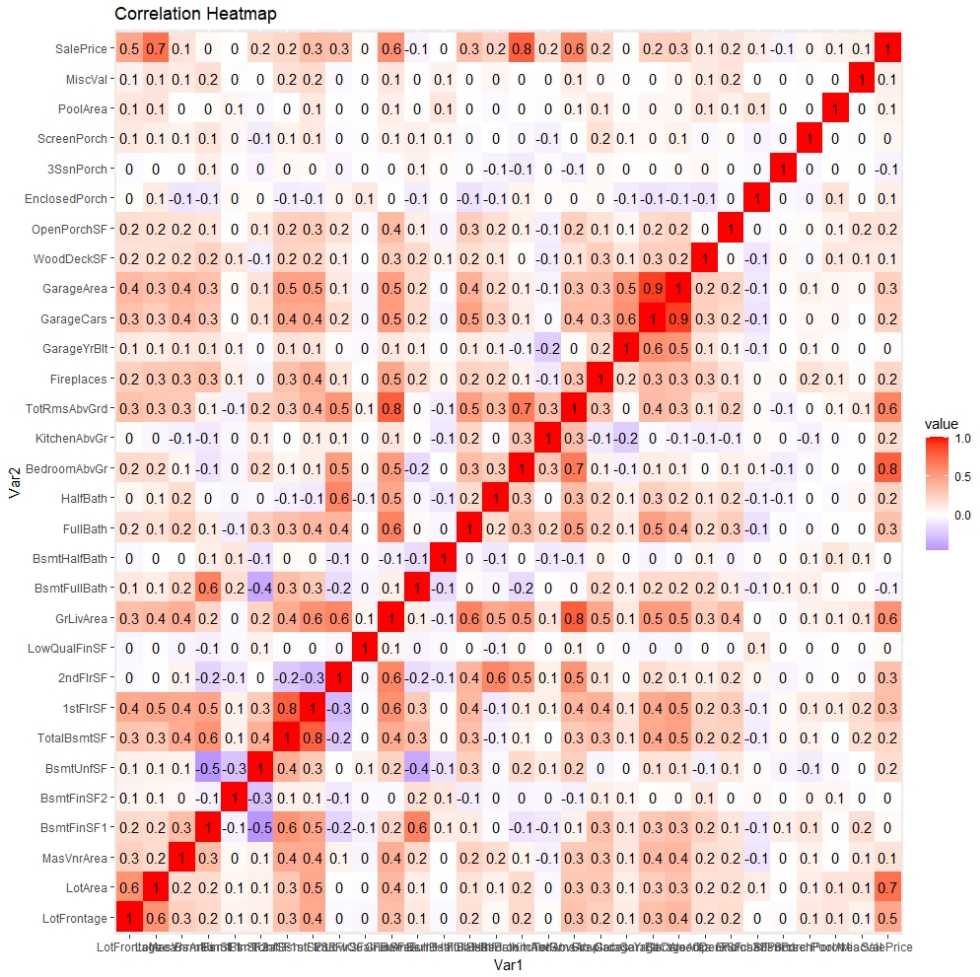

Figure 9: Correlation Heatmap

[The white box denotes 0 correlation. The more red it gets, the higher the positive correlation and the more it gets blue, the lower the negative correlation]

From Figure 9, we can observe that,

1. number of cars that can be put in a garage('GarageCars') and area of the garage ('GarageArea') have a correlation of 0.9, which is very high. That implies that the two variables are highly correlated, and hence, may cause multicollinearity in the model.

2. Total rooms above ground('TotroomAbvGr') and number of Bedrooms above ground ('BedroomAbvGr') have a high correlation of 0.7.

3. House Price is highly correlated with area of the property, total number of bedrooms and area of the living space.

Partial Correlation Heatmap:

Here, the Partial Correlation Heatmap is a graphical representation of Partial correlations of highest order between all possible numerical explanatory variables, present in the dataset. Those predictors which have a high magnitude of partial correlation, are the cause of multicollinearity in the model.

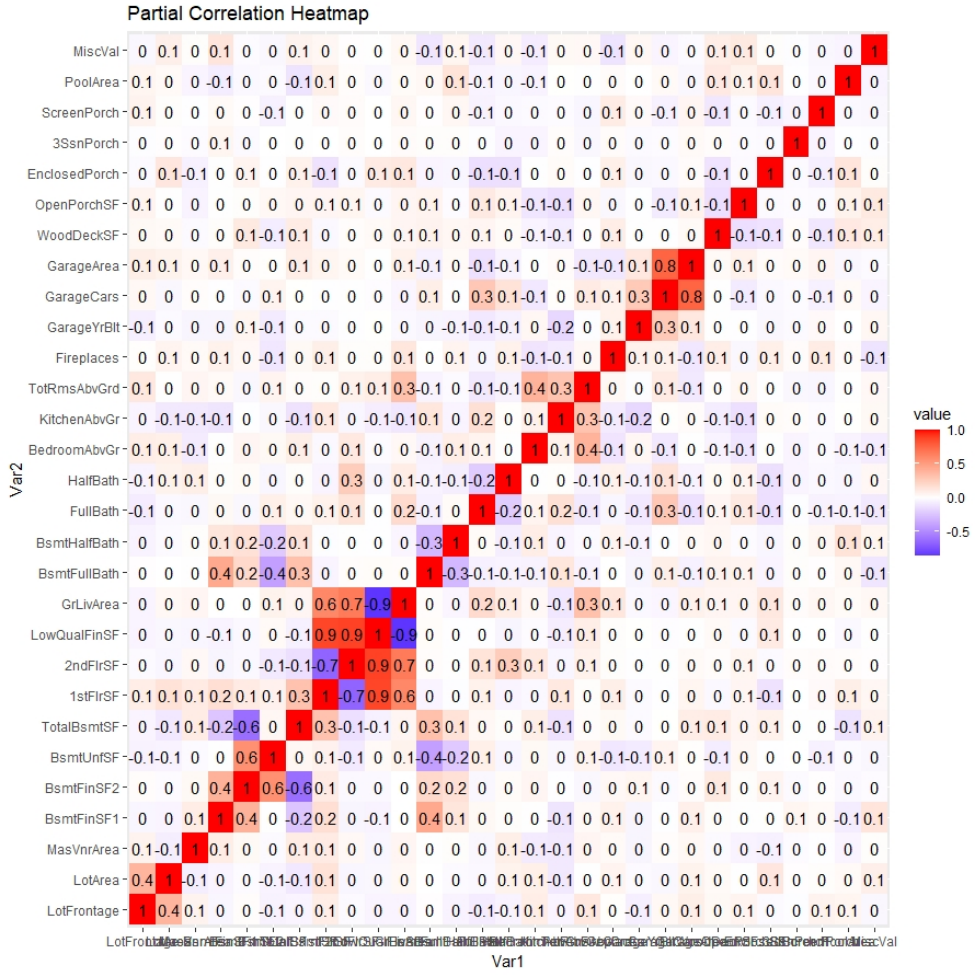

Figure 10: Partial Correlation Heatmap

[The white box denotes 0 correlation. The more red it gets, the higher the positive correlation and the more it gets blue, the lower the negative correlation]

From Figure 10, we observe that,

1. Capacity of Garage and Area of Garage are strongly associated.

2. Total rooms in a house and number of rooms are strongly associated.

3. Some of the basement surface area measurements are stringly related.

Hence, Multicollinearity is present in the dataset and it is necessary to remove

those multicollinearity before fitting the model.

6 Model Fitting Process

Since our Response variable is highly positively skewed and Least Squares assumes the normality of error, we log transform our response Price per Square feet.

our response variable becomes log('SalePrice').

We introduce a new variable called 'agegr' which is a categorical variable consists of 11 levels:
"40 49" "50 59" "10 19" "20 29" "30 39" "0 9" "90 99" "60 69" "70 79" "80 89" "more than 100"

We introduce a new variable called 'agegr' which is a categorical variable consists of 11 levels:
"Renovated" "Not Renovated"

Out of 78 explanatory variable in the original dataset, we now use 77 variables including the 'agegr' variable and excluding 'YrSold' and 'YearBuilt and 'YearRemmod'

We split the dataset into training set and test set consist of 80 percent and 20 percent of the total observations respectively.

6.1 Variable Selection: LASSO

It was observed from the correlation heatmap that a lot of predictors were linearly related. Multicollinearity leads to inestimability of coefficients in the model. To remove multicollinearity, we use LASSO as a variable selection method and collect the predictors that have non zero coefficients.

Choice of lambda :

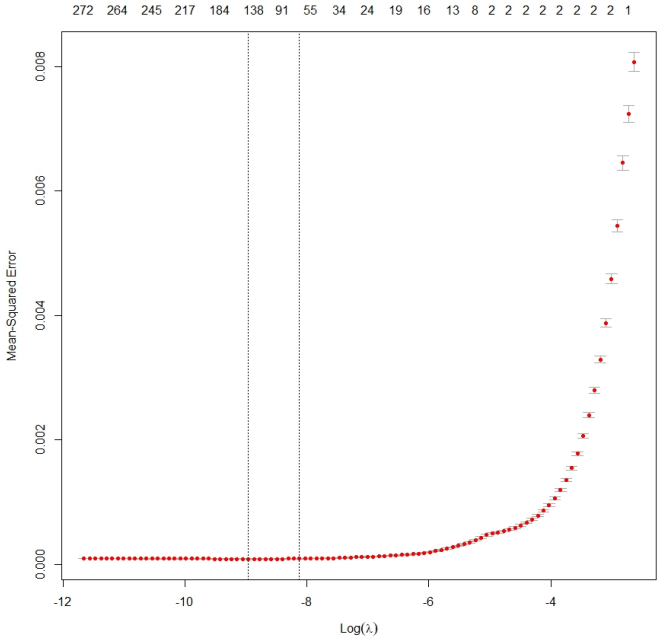

Figure 11: Choice of Lambda

We choose 1se λ for which the number of non zero predictors is coming out to be 65.

The sparse matrix gives-

PREDICTOR	LAMBDA	WEIGHT
(Intercept)	lambda.1se	3.43304
LotFrontage	lambda.1se	-0.00488
1stFlrSF	lambda.1se	-0.00014
GrLivArea	lambda.1se	-0.00003
BedroomAbvGr	lambda.1se	0.03623
Fireplaces	lambda.1se	-0.00384
GarageArea	lambda.1se	-0.00011
OpenPorchSF	lambda.1se	-0.00001
EnclosedPorch	lambda.1se	-0.00027
MSSubClass 90	lambda.1se	0.07428
MSSubClass 150	lambda.1se	0.5526
MSSubClass 160	lambda.1se	0.24395
MSZoning RH	lambda.1se	-0.03826
MSZoning RL	lambda.1se	-0.14477
MSZoning RM	lambda.1se	0.01369
LotShape IR2	lambda.1se	-0.16486
LotShape IR3	lambda.1se	-0.1017
LotShape Reg	lambda.1se	0.05458
LandContour Low	lambda.1se	-0.17026
LotConfig CulDSac	lambda.1se	-0.12386
LotConfig FR2	lambda.1se	-0.07635
LandSlope Sev	lambda.1se	-0.11854
Neighborhood Blueste	lambda.1se	0.05326
Neighborhood BrDale	lambda.1se	0.04498
Neighborhood BrkSide	lambda.1se	0.06639
Neighborhood Crawfor	lambda.1se	-0.00663
Neighborhood Gilbert	lambda.1se	-0.02684
Neighborhood IDOTRR	lambda.1se	-0.03988
Neighborhood MeadowV	lambda.1se	0.21477
Neighborhood Mitchel	lambda.1se	-0.04701
Neighborhood NoRidge	lambda.1se	0.05364
Neighborhood NPkVill	lambda.1se	0.26406
Neighborhood StoneBr	lambda.1se	-0.00998
Neighborhood SWISU	lambda.1se	0.09031
Neighborhood Veenker	lambda.1se	-0.03646

PREDICTOR	LAMBDA	WEIGHT
Condition1 RRAe	lambda.1se	-0.0483
Condition1 RRAn	lambda.1se	-0.01403
BldgType Duplex	lambda.1se	0.00107
BldgType Twnhs	lambda.1se	0.64367
BldgType TwnhsE	lambda.1se	0.39794
OverallQual 2	lambda.1se	-0.05851
OverallQual 4	lambda.1se	0.0072
OverallQual 7	lambda.1se	0.01494
OverallCond 2	lambda.1se	-0.00588
Exterior1st Sdng	lambda.1se	-0.21901
Exterior1st Stucco	lambda.1se	0.05462
Exterior1st Wd Sdng	lambda.1se	-0.00273
Functional Mod	lambda.1se	-0.08423
Functional Sev	lambda.1se	0.40883
Functional Typ	lambda.1se	0.06558
GarageType Basment	lambda.1se	0.01241
PavedDrive Y	lambda.1se	0.02926
MiscFeature None	lambda.1se	0.01547
MoSold 4	lambda.1se	-0.00942
MoSold 10	lambda.1se	0.00298
SaleType Con	lambda.1se	-0.0583
SaleCondition Family	lambda.1se	-0.00801
agegr 50 59	lambda.1se	-0.02238
agegr more than 100	lambda.1se	-0.01804

Plot of Number of Non zero coefficients and fraction deviance explained by the predictors

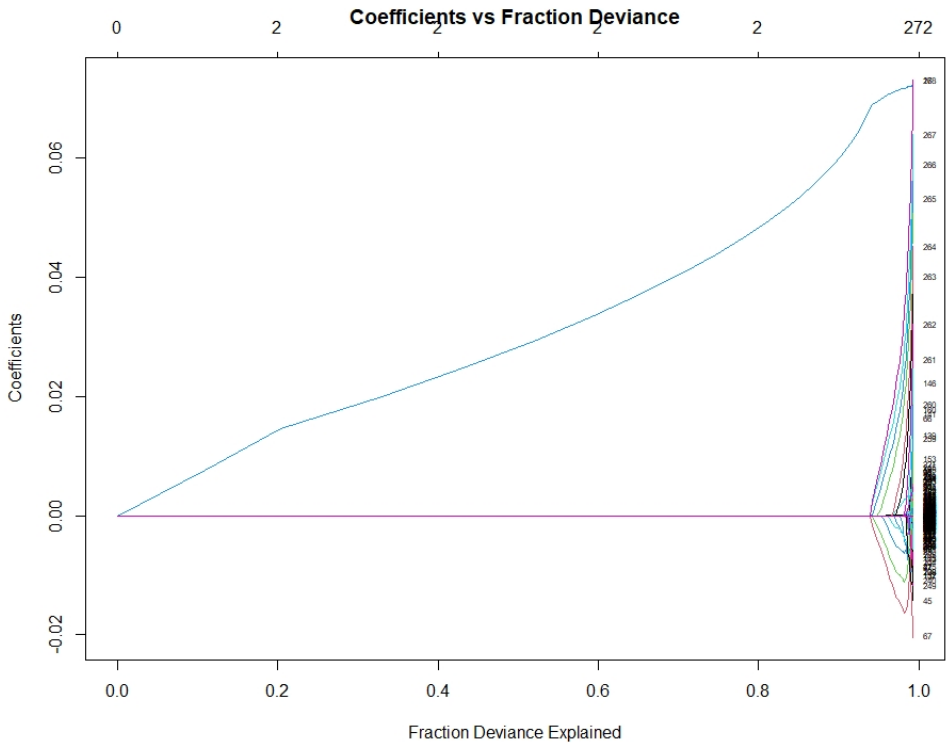

Figure 12: Plot of number of non zero coefficients and fraction deviance explained by the predictors

Almost 50 predictors are needed to explain 80% of the total variation.

6.1.1 Fitting Least Square Model on the Training Dataset

we have total 59 predictors including the dummy categorical predictors and a log transformed response

The Model used:

$$\log y = \beta_0 + \sum_i \beta_i xi + \varepsilon$$

Where,

y : House Price (in Dollars/sqft)

xi : Covariates

ε : error term

Results of Testing of Hypothesis of the model coefficients:

predictors	Estimate	Standard Error	t.value	P.value
(Intercept)	3.553112	0.060826	58.41443	0*
LotFrontage	-0.0042	0.000448	-9.36931	3.93E-20*
‘1stFlrSF‘	-0.00017	2.28E-05	-7.41162	2.46E-13*
GrLivArea	-1.20E-07	1.96E-05	-0.00611	0.995127*
Fireplaces	-0.02614	0.011532	-2.26631	0.023624*
GarageArea	-0.00017	3.84E-05	-4.48483	8.05E-06*
EnclosedPorch	-0.00029	9.69E-05	-3.02284	0.002561*
OpenPorchSF	-0.00016	0.000106	-1.5138	0.13036
MSSubClass 90	0.14985	0.033435	4.481794	8.17E-06*
MSSubClass 150	0.807068	0.223994	3.603085	0.000328*
MSSubClass 160	0.235823	0.050314	4.687047	3.11E-06*
MSZoning RL	-0.18384	0.030434	-6.04054	2.09E-09*
MSZoning RM	-0.00102	0.034228	-0.02984	0.976199
MSZoning RH	-0.24854	0.075905	-3.27438	0.001092*
LotShape IR2	-0.19224	0.040876	-4.703	2.89E-06*
LotShape IR3	-0.09299	0.094607	-0.98292	0.325862
LotShape Reg	0.064838	0.014848	4.366799	1.38E-05*
LandContour Low	-0.28276	0.052029	-5.43457	6.75E-08*
LotConfig CulDSac	-0.13898	0.02987	-4.6528	3.67E-06*
LotConfig FR2	-0.16761	0.04105	-4.08306	4.76E-05*
LandSlope Sev	-0.34034	0.125041	-2.72187	0.006593*
Neighborhood Blueste	0.169841	0.091488	1.856431	0.063656
Neighborhood BrkSide	0.084853	0.038492	2.204413	0.027699*
Neighborhood BrDale	0.091762	0.092538	0.991611	0.321603
Neighborhood Gilbert	-0.06805	0.028646	-2.37556	0.017691*
Neighborhood IDOTRR	-0.10559	0.038132	-2.76914	0.005714*
Neighborhood Mitchel	-0.11993	0.028768	-4.16887	3.30E-05*
Neighborhood StoneBr	-0.09066	0.046115	-1.96598	0.049549*
Neighborhood NoRidge	0.120382	0.047434	2.537858	0.011289*
Neighborhood SWISU	0.218487	0.056239	3.884976	0.000108*
Neighborhood Veenker	-0.15557	0.075426	-2.06248	0.039394
Neighborhood MeadowV	0.217682	0.072567	2.999733	0.002762
Neighborhood NPkVill	0.407598	0.08838	4.611904	4.45E-06*

predictors	Estimate	Standard Error	t.value	P.value
Condition1 RRAe	-0.17534	0.060883	-2.87994	0.004054*
Condition1 RRAn	-0.09807	0.0535	-1.83312	0.067051*
BldgType Twnhs	0.637074	0.061606	10.34109	5.38E-24*
BldgType TwnhsE	0.392115	0.030737	12.7572	6.92E-35
OverallQual 2	-0.25481	0.077848	-3.27316	0.001096*
OverallQual 4	0.059315	0.025793	2.299608	0.021655*
OverallCond 2	-0.02376	0.111573	-0.21298	0.831382
OverallQual 7	0.044953	0.016672	2.696358	0.007116*
Exterior1st Sdng	-0.66905	0.207886	-3.21833	0.001327*
Exterior1st Stucco	0.192605	0.057229	3.365531	0.00079
GarageType Basment	0.091439	0.057053	1.60269	0.109287
MoSold 4	-0.06289	0.020798	-3.02368	0.002554*
MoSold 10	0.059504	0.02665	2.232766	0.025763*
SaleCondition Family	-0.07271	0.04521	-1.60815	0.108086
‘agegr 50 59‘	-0.09333	0.021298	-4.38184	1.29E-05*
‘agegr more than 100‘	-0.10604	0.04168	-2.54427	0.011085*
Functional Mod	-0.23303	0.067395	-3.45765	0.000565*
Functional Sev	0.668466	0.212173	3.15057	0.001673*
Functional Typ	0.054212	0.029584	1.832496	0.067145
PavedDrive Y	0.078683	0.024878	3.162759	0.001605*
SaleType Con	-0.06585	0.206441	-0.319	0.749786

NOTE: The coefficients corresponding the * have p- value less than the desired level of significance which is 0.05, implying that those predictors are significant in determining house prices in this linear model.

R - squared = 0.9585 ; Residual Standard error = 0.0186

Interpretation:

The R-squared value implies that approximately 96 % of the total variation in the response is explained by this new model.

6.1.2 Residual Diagnostics of the Model

1. Residual Plot:

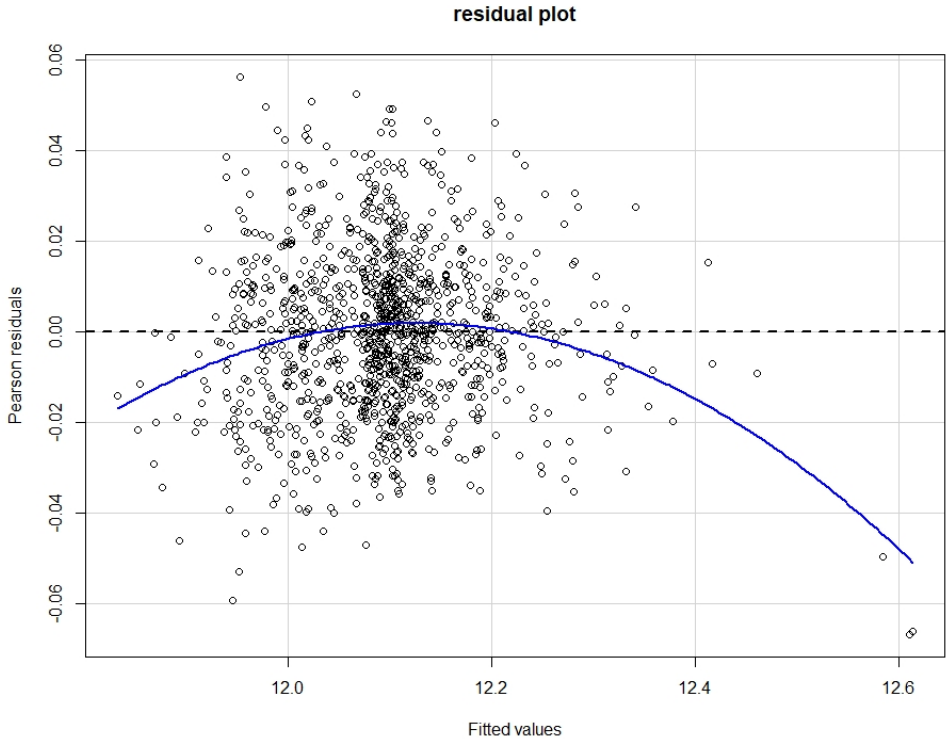

Figure 13: Residual Plot

From the above residual plot, we see that the residuals are randomly scattered around 0 line. We can 2-3 outliers are present in the data.

2. Q-Q Plot:

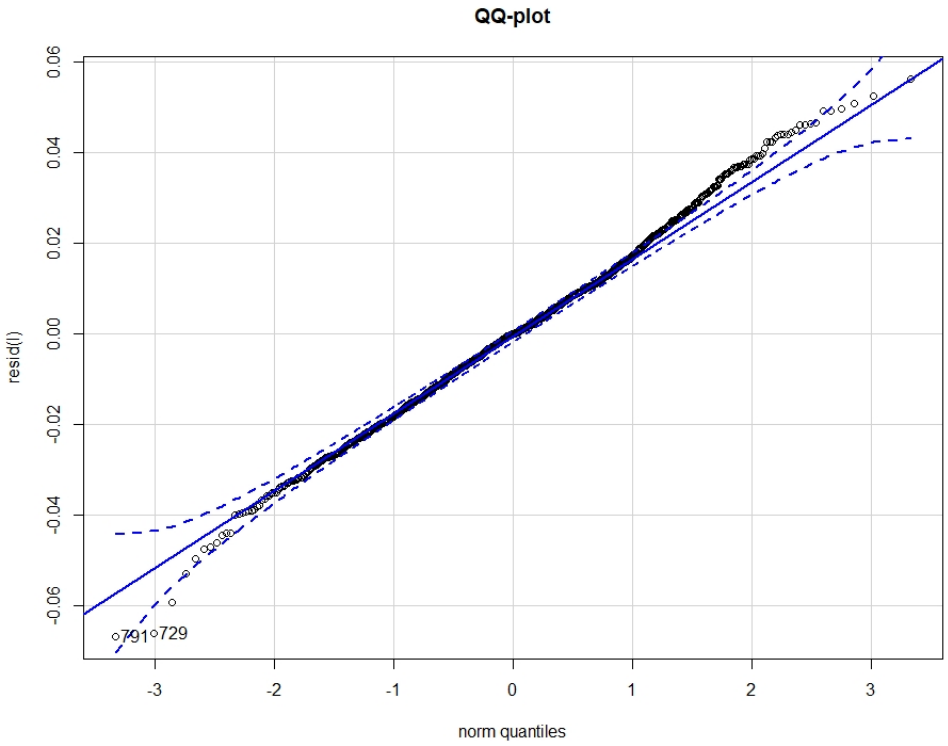

Figure 14: Q-Q Plot of Residuals

From the above q-q plot, we observe that the residuals quantiles do not exactly coincides with theoretical normal quantiles. This implies that the errors are not normally distributed.

6.1.3 Prediction over Test Dataset

Now, we fit the model on the test dataset and then predict the price.

Given below, the comparison between density plots of actual price and the predicted price

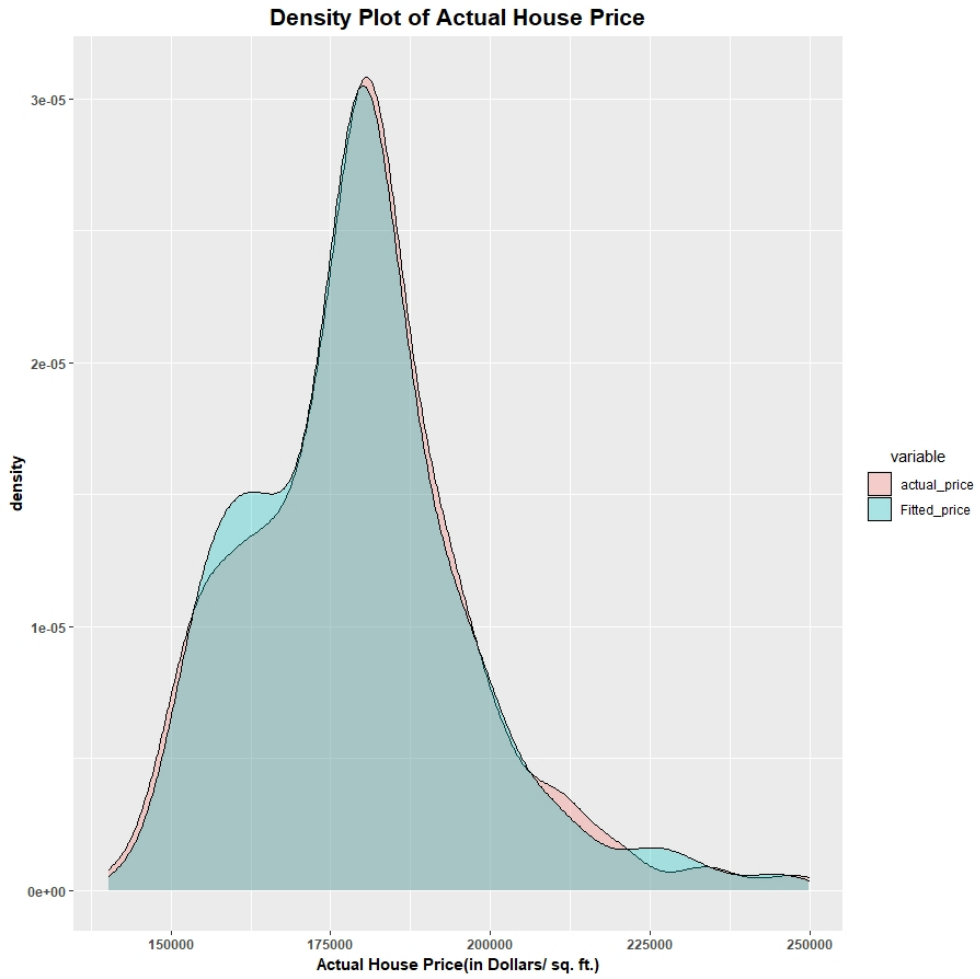

Figure 15: Density plots of actual and predicted price

6.2 Model Respecification: Principal Component Analysis

One way to reduce the number of predictor variables is Principal Component Analysis(PCA). Here, we transform the numerical predictors into orthogonal set of predicting variables such that the new predicting variables explain 95% of the total variation that was previously explained by the original set of numerical predicting variables and thus we reduce the number of predictors.

$$PC = \sum_i l_i x_i$$

Where, PC denotes the principal component and x_i's are numerical predictors. These PC's are transformed in a way such that the PC's are orthogonal to each other.

We choose the principal components one after another and continue to do so till the total variance of the response explained by them is 95% of the variation explained by the previously used numerical predictors. Clearly, number of principal components is less than that of the original set of numerical predictors.

Also, from the Correlation heatmap we definitely suspect that some of the explanatory variables that are included in the model are intercorrelated. This correlation between explanatory variables cause multicollinearity in the model which affects the model in a way that the coefficients of correlated predictors cannot be determined.

Since Principal Component Analysis Technique uses orthogonal transformation of the explanatory variables, which removes the multicollinearity from the dataset. Note that, PCA can be done only on the numerical predictors. We cannot remove the association between the categorical variables that may be present in the

dataset, using PCA technique.

Given below the table of the transformation used to make the 21 principal components.

Predictor	PC1	PC2	PC3	PC4	PC5	PC6	PC7
LotFrontage	0.19618	-0.06149	0.139217	-0.12443	0.209231	-0.11079	0.090075
YearRemodAdd	0.208289	-0.00717	-0.13386	0.29148	-0.25468	0.191335	-0.14393
MasVnrArea	0.22702	-0.10237	0.006523	-0.05327	-0.05678	0.094802	0.120258
BsmtFinSF1	0.182475	-0.34735	-0.10688	-0.25684	-0.22658	-0.01476	0.085826
BsmtFinSF2	0.003749	-0.12307	-0.10904	-0.19577	0.246767	-0.08623	-0.31825
BsmtUnfSF	0.09116	0.189601	0.468151	0.368317	0.05152	0.10964	0.03885
TotalBsmtSF	0.278644	-0.21858	0.308439	0.021557	-0.08345	0.058605	-0.00039
1stFlrSF	0.292692	-0.19743	0.338463	-0.07269	0.008835	0.01704	-0.00816
2ndFlrSF	0.127276	0.417361	-0.34253	-0.08637	-0.01601	0.047241	0.045258
LowQualFinSF	-0.01161	0.044804	0.115484	-0.02751	0.139071	-0.04652	0.175381
GrLivArea	0.349205	0.203702	-0.00869	-0.13692	0.005995	0.050674	0.048417
BsmtFullBath	0.114806	-0.30603	-0.15127	-0.24383	-0.29133	-0.24931	0.017541
BsmtHalfBath	-0.02037	-0.08178	0.001382	-0.12092	0.305688	0.37076	-0.34347
FullBath	0.268022	0.169074	0.043793	0.102068	-0.14661	-0.02388	-0.2169
HalfBath	0.142443	0.24655	-0.3751	-0.02997	-0.05444	0.132317	0.143983
BedroomAbvGr	0.141667	0.360011	0.085881	-0.21153	0.034785	-0.11975	-0.08395
KitchenAbvGr	0.024329	0.186363	0.24692	-0.21955	-0.20778	-0.38628	-0.27826
TotRmsAbvGrd	0.273974	0.295637	0.088152	-0.20186	-0.00405	-0.07214	-0.02153
Fireplaces	0.220249	-0.0791	-0.11103	-0.17243	0.178447	0.026273	0.120035
GarageYrBlt	0.137755	-0.12508	-0.22888	0.343984	0.313682	-0.20239	0.053333
GarageCars	0.314211	-0.06011	-0.12383	0.279409	0.146363	-0.19828	-0.05064
GarageArea	0.311043	-0.09818	-0.08114	0.253553	0.173543	-0.17172	-0.00844
WoodDeckSF	0.152542	-0.0975	-0.16528	-0.07808	0.051245	0.171118	-0.38485
OpenPorchSF	0.164412	0.02935	0.019125	0.003409	-0.08775	0.34058	0.256337
EnclosedPorch	-0.05144	0.073504	0.114022	-0.2094	0.352503	-0.07681	0.23779
3SsnPorch	0.0018	-0.07969	-0.02071	0.002574	-0.11575	0.034413	-0.18094
ScreenPorch	0.038212	-0.08708	-0.04744	-0.12855	0.120443	-0.1007	0.403668
PoolArea	0.033009	-0.02603	0.026559	-0.16507	0.387577	0.15001	-0.19986
MiscVal	0.058949	-0.05001	0.086204	-0.14428	-0.07014	0.479432	0.120201

Predictor	PC8	PC9	PC10	PC11	PC12	PC13	PC14
LotFrontage	0.09707	-0.1982	-0.09415	-0.08101	-0.13989	0.198554	0.341696
YearRemodAdd	-0.10117	-0.05739	-0.12139	0.261838	0.07549	-0.08535	0.010282
MasVnrArea	0.132535	0.070421	0.082438	0.033387	-0.13259	0.27771	-0.27966
BsmtFinSF1	-0.05112	0.100175	-0.01436	-0.17013	0.25399	-0.05512	-0.12618
BsmtFinSF2	0.129362	-0.23487	0.26429	0.391974	-0.53303	-0.24999	-0.03138
BsmtUnfSF	0.005095	0.022542	-0.03443	0.100298	-0.10121	0.184301	-0.02459
TotalBsmtSF	0.004084	0.031476	0.056599	0.080418	-0.05144	0.025557	-0.16624
1stFlrSF	0.038677	-0.01353	-0.01536	0.041959	-0.03755	0.03292	-0.01903
2ndFlrSF	-0.02841	0.032203	-0.03644	-0.02878	-0.01383	-0.01479	-0.04321
LowQualFinSF	-0.43308	0.470394	0.365186	0.349353	0.164076	-0.22835	0.286181
GrLivArea	-0.03218	0.059471	-0.01104	0.041167	-0.02789	-0.00653	-0.02708
BsmtFullBath	-0.18365	-0.13116	-0.00037	0.081195	0.048687	-0.07418	0.000125
BsmtHalfBath	0.32647	0.308453	-0.03086	-0.09172	0.224955	-0.29987	-0.26997
FullBath	-0.01095	0.053789	-0.10187	0.116199	0.14633	-0.23463	-0.05627
HalfBath	0.016394	-0.03328	0.057245	-0.00342	-0.11212	0.170631	-0.03536
BedroomAbvGr	0.129718	0.058375	0.072561	-0.07516	0.00232	-0.02789	-0.00364
KitchenAbvGr	0.06302	-0.06165	0.038216	-0.1913	0.097873	-0.15444	0.10221
TotRmsAbvGrd	0.017621	0.036659	0.0487	-0.04698	0.005834	-0.00816	0.0392
Fireplaces	-0.03117	0.198587	-0.12488	0.163976	-0.04135	0.201554	-0.18124
GarageYrBlt	0.011926	-0.00496	0.175365	-0.24977	0.013521	-0.13288	0.051322
GarageCars	-0.01861	-0.01585	0.0476	-0.15598	0.055708	-0.09162	-0.00071
GarageArea	0.002884	-0.02199	0.050527	-0.18147	0.017628	-0.09776	0.017338
WoodDeckSF	-0.15081	0.069038	0.174544	0.131263	0.071292	0.410814	0.295064
OpenPorchSF	-0.10789	-0.33945	-0.18824	0.102794	-0.11817	-0.48728	0.091987
EnclosedPorch	-0.42604	-0.08871	-0.00684	-0.12289	-0.06318	-0.05444	-0.43097
3SsnPorch	-0.26378	0.476353	-0.38848	-0.3275	-0.59361	-0.09197	0.149008
ScreenPorch	0.482948	0.241872	-0.25921	0.183326	0.035356	-0.11452	0.345507
PoolArea	-0.26494	-0.2681	-0.43789	-0.02205	0.290458	0.073984	0.228509
MiscVal	0.017783	-0.09739	0.45765	-0.44109	-0.05853	-0.07599	0.273471

Predictor	PC15	PC16	PC17	PC18	PC19	PC20	PC21
LotFrontage	-0.50493	-0.30956	-0.11247	0.365498	0.069855	0.035353	-0.03433
YearRemodAdd	0.045612	-0.15546	-0.2869	0.229906	0.313107	-0.00616	-0.55615
MasVnrArea	-0.2405	0.508569	-0.34873	-0.00124	-0.34419	-0.28958	-0.19935
BsmtFinSF1	-0.09317	-0.09841	-0.00417	-0.09625	0.054371	0.042752	0.059487
BsmtFinSF2	0.094528	0.090988	-0.06792	-0.02691	0.135832	-0.03166	0.060415
BsmtUnfSF	0.167148	0.022961	0.039215	-0.15785	0.01002	0.207757	-0.02261
TotalBsmtSF	0.106825	-0.04212	0.007329	-0.26512	0.119787	0.236099	0.062826
1stFlrSF	0.03208	0.03469	0.088472	0.006885	0.082354	-0.00168	0.054676
2ndFlrSF	0.041989	-0.03022	-0.07354	-0.07312	-0.03182	-0.08027	0.135665
LowQualFinSF	-0.22741	0.186362	-0.04305	0.06365	0.026821	0.062225	-0.00113
GrLivArea	0.042051	0.019172	0.004941	-0.05192	0.042396	-0.06526	0.162249
BsmtFullBath	0.026894	-0.011	-0.04364	-0.21163	0.090925	0.134964	-0.04812
BsmtHalfBath	-0.17909	-0.0584	-0.03437	0.082314	-0.04783	0.261061	-0.08777
FullBath	0.134076	-0.08533	-0.11419	0.167403	-0.02251	-0.39911	0.330896
HalfBath	0.006187	0.231748	0.002668	0.082195	0.161381	0.595352	0.038383
BedroomAbvGr	-0.19673	-0.22662	0.018715	-0.44024	0.145525	-0.06447	-0.16653
KitchenAbvGr	0.221727	0.293555	0.091601	0.329998	-0.23678	0.215641	-0.27412
TotRmsAbvGrd	-0.06369	-0.01654	0.043513	-0.09973	0.043778	-0.04664	-0.2005
Fireplaces	0.187495	0.006083	0.62661	0.338437	0.128215	-0.16898	-0.11005
GarageYrBlt	0.054388	0.034836	0.216821	-0.26892	-0.02954	-0.17197	-0.4271
GarageCars	0.02479	0.055299	-0.07279	0.114477	-0.02022	0.100369	0.198737
GarageArea	-0.01179	0.028589	-0.10598	0.080328	-0.0678	0.153502	0.249082
WoodDeckSF	0.210204	-0.29746	-0.0116	-0.10209	-0.50999	0.083702	-0.03237
OpenPorchSF	-0.14894	-0.02419	0.270436	-0.06348	-0.47144	0.082379	-0.06761
EnclosedPorch	0.278732	-0.29033	-0.34409	0.13253	-0.12628	0.064785	-0.14003
3SsnPorch	-0.01083	0.029548	-0.04508	-0.05142	0.0095	0.00325	-0.02574
ScreenPorch	0.403519	-0.03933	-0.27202	-0.06626	-0.07682	0.038429	-0.06543
PoolArea	0.057405	0.421885	-0.08586	-0.19212	0.199999	-0.0566	0.008464
MiscVal	0.282821	0.051568	-0.06386	0.121843	0.222677	-0.17654	0.006915

Scree Plot:

A scree plot is a line plot of the eigenvalues of factors or principal components in an analysis. The scree plot represents the percentage variation of the total variation that is explained by each of the 21 principal components.

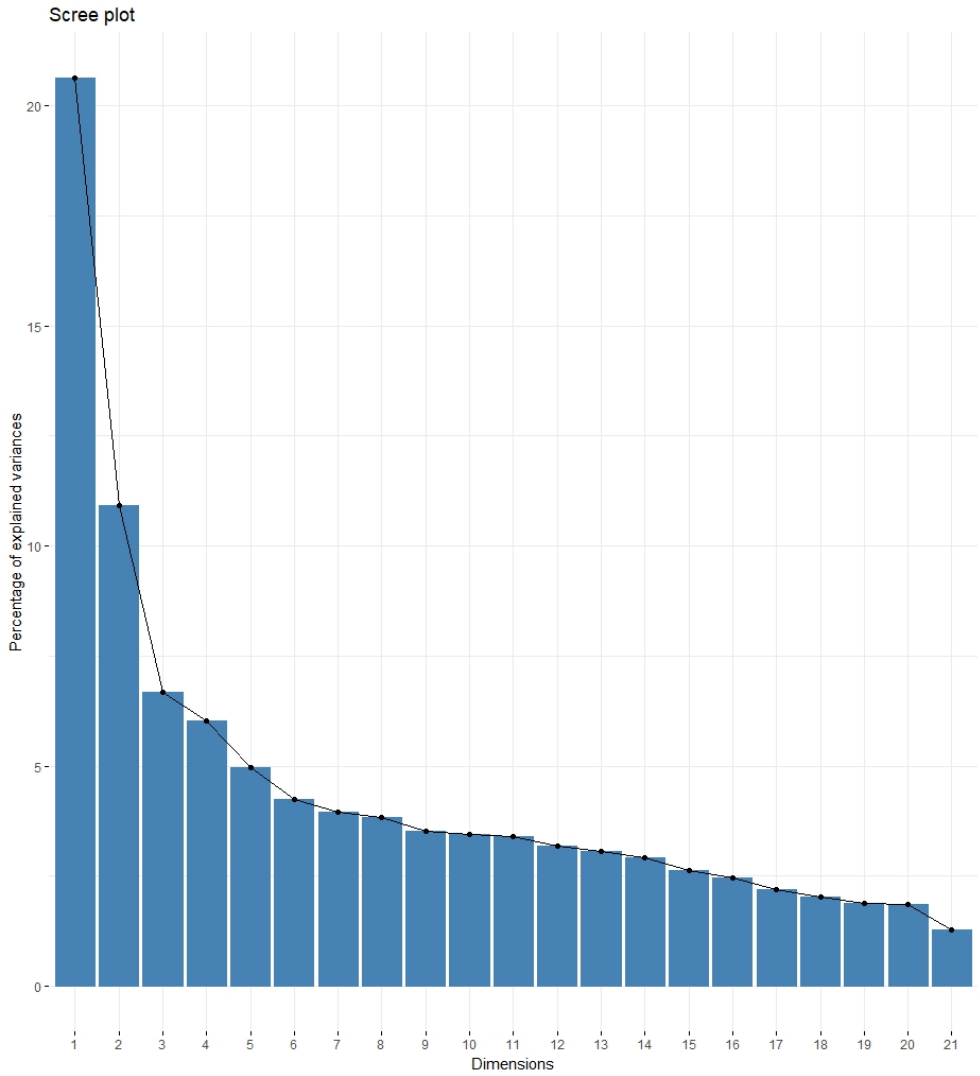

Figure 16: Scree Plot of the 21 Principal Components

From the screeplot, we observe that the first principal component more than 20% of the total variability and the 21st component explains almost 3% of total variability.

6.2.1 Fitting the Least Square Model on the Training Dataset

Here we have total 21 principal components which explains 95 percent variation of the response which is explained by the numerical features only.

We, choose a base level for each category and remove them from predictor variables to avoid singularity in the model.
The Model used:

$$\log y = \beta_0 + \sum_i \beta_i xi + \varepsilon$$

Where,

y : House Price (in Dollars)

xi : Covariates

ε : error term

Results of Testing of Hypothesis of the model coefficients:

R - squared = 0.9721 ; Residual Standard error = 0.0169

Interpretation:

The R-squared value implies that approximately 97 % of the total variation in the response is explained by the chosen model.

predictors	Estimate	Std Error	t.value	P value
(Intercept)	2.677608	0.707559	3.784287	0.000164
PC1	-0.05711	0.00818	-6.98221	5.65E-12
PC2	0.017272	0.010608	1.628119	0.10385
PC3	-0.04192	0.011803	-3.55134	0.000403
PC4	0.022569	0.013243	1.704168	0.088696
PC5	0.005592	0.013698	0.408224	0.683206
PC6	-0.01992	0.010058	-1.98012	0.047995
PC7	-0.00722	0.008919	-0.8098	0.418269
PC8	0.019464	0.007257	2.682171	0.007449
PC9	0.018118	0.008246	2.197088	0.028269
PC10	0.049367	0.009999	4.937361	9.44E-07
PC11	-0.02383	0.01139	-2.09175	0.036741
PC12	-0.00036	0.009478	-0.03823	0.969511
PC13	-0.02856	0.009166	-3.11557	0.001894
PC14	-0.01895	0.009569	-1.98054	0.047948
PC15	0.003148	0.010252	0.307093	0.758844
PC16	0.00698	0.011383	0.613175	0.539916
PC17	0.02049	0.015662	1.308229	0.19113
PC18	-0.09229	0.013404	-6.88515	1.08E-11
PC19	-0.00474	0.010856	-0.43627	0.662743
PC20	-0.03166	0.011755	-2.69292	0.007215
PC21	-0.06211	0.017564	-3.53635	0.000426
MSSubClass 30	-0.05904	0.051918	-1.13711	0.255797
MSSubClass 40	-0.18009	0.170644	-1.05533	0.291556
MSSubClass 45	-0.10273	0.171651	-0.5985	0.54966
MSSubClass 50	-0.13424	0.088263	-1.52087	0.128644
MSSubClass 60	-0.02257	0.072414	-0.31172	0.755324
MSSubClass 70	0.006439	0.08341	0.077203	0.938479
MSSubClass 75	-0.04801	0.134912	-0.35585	0.722038
MSSubClass 80	-0.10747	0.151112	-0.71121	0.477139
MSSubClass 85	0.12405	0.089057	1.392923	0.163988
MSSubClass 90	0.234456	0.083376	2.812033	0.00503
MSSubClass 120	0.139616	0.117398	1.189253	0.234654
MSSubClass 150	1.064141	0.293272	3.628509	0.000301
MSSubClass 160	0.429435	0.139767	3.07251	0.002187
MSSubClass 180	0.288324	0.164007	1.757993	0.079089
MSSubClass 190	-0.29147	0.358297	-0.81349	0.416154

predictors	Estimate	Std Error	t.value	P value
MSZoning FV	-0.21514	0.113151	-1.90134	0.057577
MSZoning RH	-0.44121	0.11899	-3.708	0.000222
MSZoning RL	-0.34106	0.096637	-3.52933	0.000438
MSZoning RM	-0.20312	0.091935	-2.20941	0.027398
Street Pave	0.259941	0.127596	2.037227	0.04192
Alley None	0.028612	0.042624	0.671254	0.502231
Alley Pave	0.042652	0.064424	0.662048	0.50811
LotShape IR2	-0.21454	0.042028	-5.10481	4.04E-07
LotShape IR3	-0.17651	0.09687	-1.82214	0.068767
LotShape Reg	0.046175	0.015824	2.918047	0.00361
LandContour HLS	0.041812	0.051146	0.817511	0.413853
LandContour Low	-0.20761	0.073467	-2.82581	0.004821
LandContour Lvl	0.036808	0.039893	0.92269	0.356417
LotConfig CulDSac	-0.13337	0.035237	-3.78479	0.000164
LotConfig FR2	-0.1832	0.045085	-4.0634	5.26E-05
LotConfig FR3	-0.13686	0.080023	-1.71021	0.087572
LotConfig Inside	6.00E-05	0.018726	0.003202	0.997446
LandSlope Mod	0.083346	0.040232	2.07164	0.038584
LandSlope Sev	-0.46067	0.161588	-2.85087	0.00446
Neighborhood Blueste	-0.23436	0.128087	-1.82967	0.06763
Neighborhood BrDale	-0.35421	0.131022	-2.70342	0.006993
Neighborhood BrkSide	-0.26199	0.100487	-2.60721	0.00928
Neighborhood ClearCr	-0.45508	0.104418	-4.35824	1.46E-05
Neighborhood CollgCr	-0.49478	0.079576	-6.21772	7.72E-10
Neighborhood Crawfor	-0.5035	0.090358	-5.57233	3.32E-08
Neighborhood Edwards	-0.47919	0.086798	-5.52078	4.42E-08
Neighborhood Gilbert	-0.52861	0.082571	-6.40188	2.47E-10
Neighborhood IDOTRR	-0.51668	0.105366	-4.90372	1.12E-06
Neighborhood MeadowV	-0.25932	0.128048	-2.02519	0.043144
Neighborhood Mitchel	-0.54009	0.085681	-6.30355	4.56E-10
Neighborhood NAmes	-0.39477	0.084931	-4.64816	3.85E-06
Neighborhood NoRidge	-0.36275	0.097116	-3.73521	0.000199
Neighborhood NPkVill	-0.16506	0.206587	-0.79899	0.424507
Neighborhood NridgHt	-0.54695	0.083584	-6.54365	1.01E-10
Neighborhood NWAmes	-0.49167	0.087933	-5.59143	2.99E-08
Neighborhood OldTown	-0.37157	0.097502	-3.81092	0.000148
Neighborhood Sawyer	-0.44926	0.087288	-5.14686	3.25E-07
Neighborhood SawyerW	-0.45145	0.084245	-5.35875	1.07E-07
Neighborhood Somerst	-0.45458	0.092082	-4.93669	9.47E-07
Neighborhood StoneBr	-0.6315	0.093839	-6.72959	3.03E-11
Neighborhood SWISU	-0.18425	0.107648	-1.71164	0.087309
Neighborhood Timber	-0.45742	0.08879	-5.15173	3.17E-07
Neighborhood Veenker	-0.52207	0.115031	-4.5385	6.44E-06

predictors	Estimate	Std Error	t.value	P value
Condition1 Feedr	0.056742	0.050258	1.129011	0.259195
Condition1 Norm	0.001509	0.041785	0.036105	0.971207
Condition1 PosA	-0.01789	0.098058	-0.18241	0.855299
Condition1 PosN	0.01865	0.067643	0.275706	0.782838
Condition1 RRAe	-0.18982	0.075423	-2.51678	0.012017
Condition1 RRAn	-0.11806	0.07126	-1.6567	0.097929
Condition1 RRNe	0.15193	0.140957	1.077848	0.281391
Condition1 RRNn	-0.07098	0.11484	-0.61812	0.536656
Condition2 Feedr	0.204243	0.158942	1.285019	0.199117
Condition2 Norm	0.088483	0.129486	0.683335	0.494571
Condition2 PosA	0.152477	0.210647	0.723849	0.469347
Condition2 PosN	0.075847	0.202293	0.374937	0.707796
BldgType 2 mCon	0.343548	0.353658	0.971414	0.331604
BldgType Twnhs	0.502257	0.127776	3.930769	9.12E-05
BldgType TwnhsE	0.275895	0.114848	2.402249	0.016497
HouseStyle 1.5Unf	-0.03976	0.16759	-0.23727	0.812499
HouseStyle 1Story	-0.09012	0.086361	-1.04352	0.29699
HouseStyle 2.5Unf	-0.03957	0.131441	-0.30103	0.763463
HouseStyle 2Story	-0.0952	0.089686	-1.06144	0.288776
HouseStyle SFoyer	-0.20198	0.106602	-1.89467	0.058458
HouseStyle SLvl	-0.044	0.160406	-0.2743	0.783916
OverallQual 2	0.33492	0.480017	0.697725	0.48553
OverallQual 3	0.360583	0.461152	0.781916	0.43447
OverallQual 4	0.406861	0.457505	0.889303	0.374078
OverallQual 5	0.333393	0.460559	0.723889	0.469322
OverallQual 6	0.377863	0.462472	0.817051	0.414116
OverallQual 7	0.434216	0.462713	0.938413	0.348285
OverallQual 8	0.433343	0.462968	0.93601	0.349519
OverallQual 9	0.456551	0.464642	0.982585	0.326076
OverallQual 10	0.47177	0.473788	0.995742	0.319644
OverallCond 2	-0.0632	0.216843	-0.29146	0.770765
OverallCond 3	0.147197	0.160705	0.915949	0.35994
OverallCond 4	0.085002	0.151892	0.55962	0.575878
OverallCond 5	0.141926	0.152819	0.928715	0.353286
OverallCond 6	0.163249	0.153602	1.062801	0.288158
OverallCond 7	0.183728	0.154763	1.187162	0.235478
OverallCond 8	0.213713	0.155976	1.370164	0.170978
OverallCond 9	0.234618	0.164498	1.426266	0.154139

predictors	Estimate	Std Error	t.value	P value
RoofStyle Gble	-0.3292	0.151634	-2.171	0.030192
RoofStyle Gambrel	-0.0971	0.175439	-0.55344	0.580099
RoofStyle Hip	-0.33969	0.152965	-2.2207	0.02662
RoofStyle Mansard	-0.33108	0.244817	-1.35237	0.176596
RoofStyle Shed	-0.71772	0.220299	-3.25792	0.001164
RoofMatl WdShake	0.005834	0.120817	0.048285	0.9615
RoofMatl WdShngl	0.813727	0.3187	2.553266	0.010836
Exterior1st BrkComm	-0.6469	0.2661	-2.43106	0.01525
Exterior1st BrkFace	-0.03218	0.158894	-0.20253	0.839551
Exterior1st CemntBd	-0.50853	0.222914	-2.2813	0.022764
Exterior1st HdBoard	-0.18605	0.150767	-1.23402	0.217519
Exterior1st MetalSd	0.033949	0.16285	0.208467	0.834912
Exterior1st Plywood	-0.20892	0.147892	-1.41263	0.15811
Exterior1st Sdng	-0.52712	0.287643	-1.83255	0.067201
Exterior1st Stucco	0.153899	0.170672	0.901728	0.367443
Exterior1st VinylSd	-0.07981	0.169485	-0.47089	0.637832
Exterior1st WdShing	-0.07065	0.159118	-0.44401	0.65714
Exterior2nd BrkFace	0.049503	0.181121	0.273315	0.784674
Exterior2nd CBlock	0.14005	0.261527	0.535508	0.592431
Exterior2nd CmentBd	0.604062	0.231492	2.609427	0.00922
Exterior2nd HdBoard	0.24784	0.170236	1.455862	0.14578
Exterior2nd ImStucc	0.11774	0.197874	0.595024	0.551978
Exterior2nd MetalSd	0.03496	0.179923	0.194307	0.84598
Exterior2nd Plywood	0.213862	0.165399	1.293007	0.196341
Exterior2nd Stone	0.062762	0.298305	0.210394	0.833408
Exterior2nd Stucco	0.133868	0.187953	0.712243	0.476499
Exterior2nd VinylSd	0.141823	0.188151	0.753773	0.451183
MasVnrType BrkFace	-0.01511	0.086853	-0.17393	0.861959
MasVnrType None	0.003091	0.086072	0.035914	0.971359
MasVnrType Stone	0.021604	0.089309	0.241906	0.808908
ExterQual Fa	0.097171	0.097229	0.999402	0.317869
ExterQual Gd	0.007972	0.055604	0.143366	0.886033
ExterQual TA	0.024453	0.061317	0.398793	0.690141
ExterCond Fa	-0.10391	0.1022	-1.0167	0.309568
ExterCond Gd	-0.10198	0.083797	-1.21694	0.223945
ExterCond Po	0.38042	0.277073	1.372993	0.170097
ExterCond TA	-0.09205	0.083623	-1.10083	0.271267
Foundation CBlock	-0.08222	0.03253	-2.52758	0.011656
Foundation PConc	-0.05002	0.03376	-1.48155	0.138811
Foundation Slab	-0.07168	0.106824	-0.67103	0.502375
Foundation Stone	-0.17415	0.145248	-1.19897	0.230855
Foundation Wood	0.03265	0.154477	0.21136	0.832654

predictors	Estimate	Std Error	t.value	P value
BsmtQual Fa	0.035413	0.057444	0.616477	0.537736
BsmtQual Gd	-0.02275	0.031615	-0.71975	0.471868
BsmtQual None	0.195504	0.340563	0.574062	0.56607
BsmtQual TA	0.007162	0.042016	0.170448	0.864696
BsmtCond Gd	-0.06249	0.049556	-1.26096	0.207652
BsmtCond None	0.05413	0.229178	0.236191	0.813338
BsmtCond Po	-0.37447	0.169305	-2.21184	0.027229
BsmtCond TA	-0.04061	0.037511	-1.0827	0.279232
BsmtExposure Gd	-0.05204	0.029801	-1.74613	0.08113
BsmtExposure Mn	-0.04609	0.029687	-1.55261	0.120867
BsmtExposure No	-0.02147	0.023532	-0.91228	0.361866
BsmtExposure None	-0.16953	0.205478	-0.82506	0.409554
BsmtFinType1 BLQ	-0.0182	0.028931	-0.62917	0.529398
BsmtFinType1 GLQ	-0.01666	0.025364	-0.65678	0.511489
BsmtFinType1 LwQ	0.054932	0.036438	1.507575	0.132015
BsmtFinType1 None	-0.20589	0.47177	-0.43643	0.662632
BsmtFinType1 Rec	-0.02461	0.028665	-0.85866	0.390755
BsmtFinType1 Unf	0.004529	0.029043	0.155922	0.876129
BsmtFinType2 BLQ	0.006096	0.062778	0.097103	0.922666
BsmtFinType2 GLQ	-0.1293	0.071097	-1.8187	0.06929
BsmtFinType2 LwQ	0.030116	0.058866	0.51161	0.60905
BsmtFinType2 Rec	0.039078	0.057174	0.683487	0.494475
BsmtFinType2 Unf	0.023149	0.055932	0.413889	0.679055
Heating GasW	0.010101	0.09396	0.107503	0.914414
Heating Grav	0.163762	0.249886	0.655346	0.512413
HeatingQC Fa	-0.04741	0.052133	-0.90944	0.363363
HeatingQC Gd	-0.01088	0.020342	-0.53499	0.592786
HeatingQC Po	-0.36172	0.226479	-1.59716	0.110582
HeatingQC TA	0.002574	0.020423	0.126036	0.899731

predictors	Estimate	Std Error	t.value	P value
CentralAir Y	0.011634	0.037293	0.311963	0.755141
Electrical FuseF	-0.05902	0.07623	-0.77426	0.438981
Electrical FuseP	0.100966	0.148214	0.681215	0.495911
Electrical SBrkr	-0.02353	0.029141	-0.8074	0.419652
KitchenQual Fa	0.08457	0.063008	1.342202	0.17987
KitchenQual Gd	0.008278	0.03806	0.217499	0.827869
KitchenQual TA	0.040556	0.041385	0.979971	0.327364
Functional Maj2	0.487244	0.24813	1.963667	0.049877
Functional Min1	0.767591	0.240297	3.194347	0.00145
Functional Min2	0.722381	0.244356	2.956268	0.003195
Functional Mod	0.415761	0.244769	1.698584	0.089744
Functional Sev	1.383579	0.285369	4.848378	1.47E-06
Functional Typ	0.683739	0.237504	2.878849	0.004086
FireplaceQu Fa	0.057264	0.078178	0.732485	0.464064
FireplaceQu Gd	-0.029	0.066024	-0.43922	0.660607
FireplaceQu None	-0.05006	0.072756	-0.688	0.491633
FireplaceQu Po	-0.01138	0.089002	-0.12783	0.898312
FireplaceQu TA	-0.01134	0.067445	-0.16819	0.866468
GarageType ttchd	0.077541	0.067673	1.145825	0.252173
GarageType Basment	0.119648	0.091761	1.303908	0.192599
GarageType BuiltIn	0.104763	0.073465	1.426019	0.15421
GarageType CarPort	0.087432	0.119695	0.730461	0.465299
GarageType Detchd	0.108475	0.067337	1.610942	0.107544
GarageType None	0.338054	0.277113	1.219913	0.222818
GarageFinish RFn	0.013297	0.019073	0.697194	0.485862
GarageFinish Unf	-0.02821	0.022642	-1.24603	0.213079
GarageQual Gd	0.085281	0.099117	0.860402	0.389797
GarageQual Po	-0.09758	0.245342	-0.39773	0.690926
GarageQual TA	0.106469	0.037728	2.822028	0.004877
GarageCond Fa	-0.06598	0.233151	-0.28298	0.77726
GarageCond Gd	-0.05242	0.246358	-0.21279	0.831541
GarageCond Po	-0.05113	0.254748	-0.20069	0.840985
GarageCond TA	-0.20879	0.228457	-0.91392	0.361007
PavedDrive P	0.082552	0.055416	1.489672	0.136662
PavedDrive Y	0.085848	0.036168	2.373595	0.017825
PoolQC None	-0.16688	0.22954	-0.72704	0.467393
Fence dWo	0.010363	0.032127	0.322548	0.747112
Fence MnPrv	-0.01491	0.021588	-0.69073	0.489915
MiscFeature None	0.162017	0.142265	1.13884	0.255073
MiscFeature Othr	0.506354	0.196259	2.580031	0.010037
MiscFeature Shed	0.100082	0.140858	0.71052	0.477566

predictors	Estimate	Std Error	t.value	P value
MoSold 2	.035014	0.041131	0.851279	0.394841
MoSold 3	0.036854	0.038522	0.956701	0.338976
MoSold 4	-0.01012	0.038582	-0.26221	0.793218
MoSold 5	0.030418	0.037163	0.818501	0.413288
MoSold 6	0.038059	0.035574	1.069831	0.284983
MoSold 7	0.039987	0.036513	1.09515	0.273744
MoSold 8	0.063842	0.039905	1.599845	0.109985
MoSold 9	0.053777	0.040796	1.31819	0.187776
MoSold 10	0.100136	0.043001	2.328688	0.020097
MoSold 11	0.086561	0.045285	1.911461	0.056263
MoSold 12	0.038082	0.048237	0.789475	0.430043
SaleType Con	-0.13977	0.207918	-0.67221	0.50162
SaleType ConLD	-0.07273	0.080228	-0.90657	0.364877
SaleType ConLI	0.073383	0.135992	0.539617	0.589595
SaleType ConLw	-0.13392	0.162411	-0.82456	0.409841
SaleType CWD	-0.10187	0.103721	-0.98214	0.326297
SaleType New	-0.20268	0.167002	-1.21365	0.225199
SaleType Oth	-0.1821	0.127224	-1.43136	0.152676
SaleType WD	-0.05091	0.040796	-1.24788	0.212402
SaleCondition AdjLand	0.267942	0.115026	2.329394	0.020059
SaleCondition Alloca	0.005047	0.08321	0.060659	0.951645
SaleCondition Family	-0.04803	0.054551	-0.88039	0.378884
SaleCondition Normal	0.0184	0.030524	0.602793	0.546799
SaleCondition Partial	0.17519	0.162941	1.075174	0.282586
‘agegr 10 19‘	-0.04088	0.032308	-1.26522	0.206122
‘agegr 20 29‘	-0.11458	0.056496	-2.02808	0.042847
‘agegr 30 39‘	-0.11182	0.057453	-1.94636	0.051924
‘agegr 40 49‘	-0.18375	0.063733	-2.88306	0.004032
‘agegr 50 59‘	-0.2799	0.066914	-4.18295	3.16E-05
‘agegr 60 69‘	-0.22115	0.077522	-2.85269	0.004435
‘agegr 70 79‘	-0.22136	0.084595	-2.61675	0.009026
‘agegr 80 89‘	-0.18201	0.085453	-2.12995	0.033447
‘agegr 90 99‘	-0.28107	0.08716	-3.22476	0.001306
‘agegr more than 100‘	-0.36836	0.0949	-3.88159	0.000111

6.2.2 Residual Diagnostics

Here, we will do some tests and will do some graphical representation of the residuals obtained from the model, in order to check whether the basic assumptions of simple linear regression model is satisfied.

1. Residual Plot:

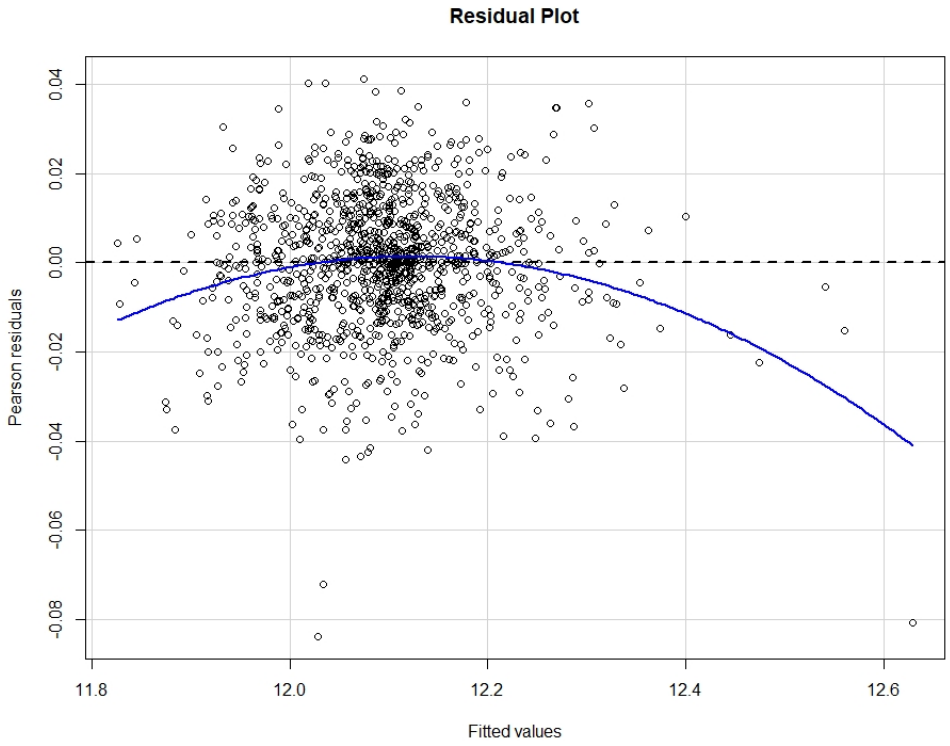

Figure 17: Residual Plot

From the above residual plot, we see that the residuals are randomly scattered around the 0 line. No pattern is visible.

2. Q-Q Plot:

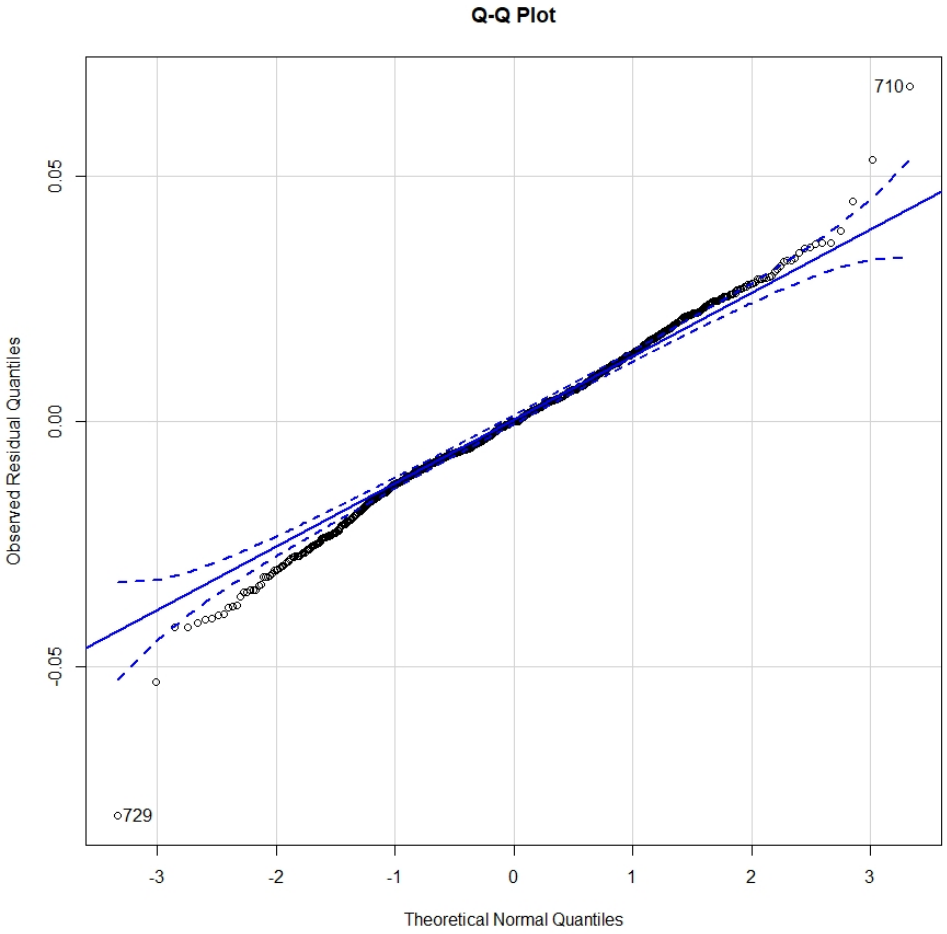

Figure 18: Q-Q Plot of Residuals

From the q-q plot, it can be noticed that residual distribution are skewed.

6.2.3 Prediction over Test Dataset

Now, we fit the model on the test dataset and then predict the price.

Given below, the comparison between density plots of actual price and the predicted price.

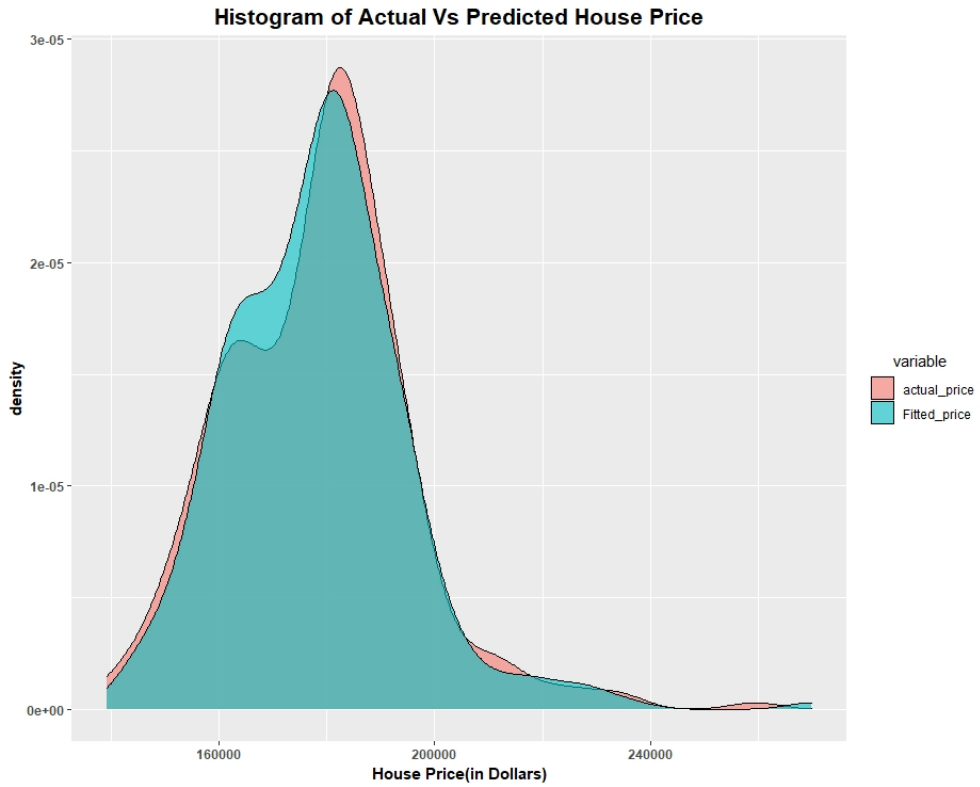

Figure 19: Density plots of actual price and predicted price

7 Conclusion

Some observations:

- Area of the property is a highly significant predictor in determining price of house.

- Older houses have a high selling price in with respect to the time they were sold.

- Poor Quality houses have selling price higher than excellent quality houses.

8 Appendix

Listing 1: R code used for the analysis and model fitting

```
1 rm(list=ls())
2 library(ggplot2)
3 library(MASS)
4 library(reshape2)
5 library(tidyverse)
6 library(dplyr)
7 library(lattice)
```

```
library(caret)
library(car)
library(trafo)
library(moments)
library(glmnet)
library(lmtest)
library(fastDummies)
library(DescTools)
library(ppcor)
library(dgof)
library(corpcor)
library(gridExtra)
library(factoextra)
library(caTools)
library(qpcR)

par(mfrow=c(1,1))
data = readxl::read_xlsx('C:/Users/Saheli/Desktop/SAHELI
   /Mstat 2022/Year 1 -Delhi/ISI-Assignments and notes/
   Deepayan sir/SEM 1 project/project data original.xlsx
   ')
attach(data)
colnames(data)
nrow(data)
data_new=data[,-c(80)]
colnames(data_new)
```

```
#missing value
#---------------------------------------
any(is.na(data_new))

#Imputing Missing Val
#---------------------------------------
data_new$PoolQC[data_new$PoolQC=='NA']= 'None'
data_new$MiscFeature[data_new$MiscFeature=='NA']= 'None'
data_new$Alley[data_new$Alley=='NA']= 'None'
data_new$Fence[data_new$Fence=='NA']= 'None'
data_new$FireplaceQu[data_new$FireplaceQu=='NA']= 'None'
data_new$LotFrontage[data_new$LotFrontage=='NA']= 'None'

data_new$LotFrontage= as.numeric(data_new$LotFrontage)
c = which(is.na(data_new$LotFrontage)== 'TRUE')
a = aggregate(data_new$LotFrontage[-c],by=
                list(data_new$Neighborhood[-c]),median)
b = data_new$Neighborhood[c]
d = array(0)
for(i in 1:length(b))
{ for(j in 1:nrow(a))
{
  if(b[i]==a$Group.1[j]){
    d[i]= a$x[j]
  }}}

```

```
57 data_new$LotFrontage[c] = d #missing values replaced by
      median
58
59 data_new$GarageCond[data_new$GarageCond=='NA']= 'None'
60 data_new$GarageQual[data_new$GarageQual=='NA']= 'None'
61 data_new$GarageType[data_new$GarageType=='NA']= 'None'
62 data_new$GarageArea[is.na(data_new$GarageArea)== TRUE]=
      0
63 data_new$GarageFinish[data_new$GarageFinish=='NA']= '
      None'
64
65 data_new$GarageYrBlt = as.numeric(data_new$GarageYrBlt)
66 data_new$GarageYrBlt[is.na(data_new$GarageYrBlt)== TRUE
      ]= 0
67 data_new$GarageCars = as.numeric(data_new$GarageCars)
68 data_new$GarageCars[is.na(data_new$GarageCars)== TRUE]=
      0
69 data_new$BsmtFinSF1 = as.numeric(data_new$BsmtFinSF1)
70 data_new$BsmtFinSF1[is.na(data_new$BsmtFinSF1)== TRUE]=
      0
71 data_new$BsmtFinSF2 = as.numeric(data_new$BsmtFinSF2)
72 data_new$BsmtFinSF2[is.na(data_new$BsmtFinSF2)== TRUE]=
      0
73 data_new$BsmtUnfSF = as.numeric(data_new$BsmtUnfSF)
74 data_new$BsmtUnfSF[is.na(data_new$BsmtUnfSF)== TRUE]= 0
75 data_new$TotalBsmtSF = as.numeric(data_new$TotalBsmtSF)
```

```
76 data_new$TotalBsmtSF[is.na(data_new$TotalBsmtSF)== TRUE
   ]= 0
77 data_new$BsmtFullBath = as.numeric(data_new$BsmtFullBath
   )
78 data_new$BsmtFullBath[is.na(data_new$BsmtFullBath)==
   TRUE]= 0
79 data_new$BsmtHalfBath = as.numeric(data_new$BsmtHalfBath
   )
80 data_new$BsmtHalfBath[is.na(data_new$BsmtHalfBath)==
   TRUE]= 0
81
82 data_new$BsmtQual[data_new$BsmtQual=='NA']= 'None'
83 data_new$BsmtCond[data_new$BsmtCond=='NA']= 'None'
84 data_new$BsmtExposure[data_new$BsmtExposure=='NA']= '
   None'
85 data_new$BsmtFinType1[data_new$BsmtFinType1=='NA']= '
   None'
86 data_new$BsmtFinType2[data_new$BsmtFinType2=='NA']= '
   None'
87 data_new$YearBuilt=as.factor(data_new$YearBuilt)
88
89
90 data_new$MasVnrType[data_new$MasVnrType=='NA']= 'BrkFace
   '
91 data_new$MasVnrArea = as.numeric(data_new$MasVnrArea)
92 data_new$MasVnrArea[is.na(data_new$MasVnrArea)== TRUE]=
```

```
   0
93
94 data_new$MSZoning[data_new$MSZoning=='NA']= 'RL'
95 data_new$Functional[data_new$Functional=='NA']= 'Typ'
96 data_new$KitchenQual[data_new$KitchenQual=='NA']= 'TA'
97 data_new$Exterior1st[data_new$Exterior1st=='NA']= 'Sdng'
98 data_new$Exterior2nd[data_new$Exterior2nd=='NA']= '
   VinylSd'
99 data_new$SaleType[data_new$SaleType=='NA']= 'WD'
100 data_new$MSSubClass[data_new$MSSubClass=='NA']= 60
101 data_new$Utilities[data_new$Utilities=='NA']= 'AllPub'
102
103 any(is.na(data_new))  #no missing values
104
105 #feature engineering
106 #-----------------------------------------------
107 data_new$MSSubClass = as.factor(data_new$MSSubClass)
108 data_new$OverallCond = as.factor(data_new$OverallCond)
109 data_new$OverallQual = as.factor(data_new$OverallQual)
110 data_new$YrSold = as.factor(data_new$YrSold)
111 data_new$MoSold = as.factor(data_new$MoSold)
112
113 #Observations on the variables
114 factor_var=data_new[sapply(data_new, is.factor)]
115 num_var=subset(data_new[sapply(data_new, is.numeric)])
116 char_var=data_new[sapply(data_new, is.character)]
```

```
117
118
119 distribution <- as.data.frame(t(sapply(num_var, quantile
      )))
120 distribution$Mean <- sapply(num_var, mean)
121 distribution$SD <- sapply(num_var, sd)
122 distribution$skenewss <- sapply(num_var, skewness)
123 distribution$kurtosis <- sapply(num_var, kurtosis)
124 distribution<-round(distribution, 2)
125 distribution
126
127 t = data.frame(distribution)
128
129 t1 = cbind(predictor=rownames(t),t %>% arrange(t$sk))
130 writexl::write_xlsx(t1,'C:/Users/Saheli/Desktop/SAHELI/
      MSTAT 2022/Year 1 -Delhi/ISI-Assignments and notes/
      Deepayan sir/SEM 1 project/distribution.xlsx')
131 #AGE OF THE HOUSES
132
133 # the age of houses
134 age_sold<-YrSold-YearBuilt
135 age_sold_group<-ifelse(age_sold<10,"0~9",
136                        ifelse(age_sold<20,"10~19",
137                               ifelse(age_sold<30,"20~29"
                                      ,
138                                      ifelse(age_sold<40,
```

```
                                        "30~39",
139                                         ifelse(age_
                                               sold<50,"
                                               40~49",
140                                                 ifelse
                                                    (
                                                    age
                                                    -
                                                    sold
                                                    <60,
                                                    "
                                                    50
                                                    ~
                                                    59
                                                    ",
141                                                     ifelse
                                                        (
                                                        age
                                                        -
                                                        sold
                                                        <70,
                                                        "
                                                        60
                                                        ~
                                                        69
                                                        "
```

```
,
ifelse
(
age
-
sold
<80,
"
70
~
79
"
,
if
```

144

```
data_new=subset(data_new,select=-c(YearBuilt,YrSold))
data_new$agegr=age_sold_group
unique(data_new$agegr)
v1 <- ggplot(data_new, aes(YearBuilt)) +
  geom_bar(fill = "dodgerblue3") +
  ggtitle("The number of houses built by year") +

  theme(text = element_text(face ="bold"),
        plot.title = element_text(hjust = 0.5))


v2 <- ggplot(data_new, aes(YearBuilt, SalePrice)) +
  geom_smooth(se = TRUE,colour = "dodgerblue3") +
  ggtitle("The price trend of houses built by year") +

  theme(text = element_text(face = "bold"),
        plot.title=element_text(hjust=0.5))


```

```
grid.arrange(v1,v2)

data.frame(YearRemodAdd,YearBuilt,YrSold)
renov_bi= as.factor(ifelse(YearRemodAdd>YearBuilt &
   YearRemodAdd<YrSold,'renovated','not renovated'))
pricesq=data_new$SalePrice
ggplot(NULL,aes(age_sold_group,pricesq,color=renov_bi))+
  geom_boxplot()+
  ggtitle("Remodelling effect in the same age group") +
  theme_classic() +
  theme(text = element_text(face ="bold"),
        plot.title = element_text(hjust = 0.5))

data_new = cbind(subset(data_new, select= -c(
   YearRemodAdd)),renov_bi)
#Graphical analysis of the variables
#-----------------------------------------
summary(SalePrice)
ggplot(data=data_new,aes(x=SalePrice,y=..density..))+
  geom_histogram(bins=50,fill='green',col='black')+
  labs(title = 'Histogram of Price of House',x='Price of
      House (in Dollars)')+
  theme(plot.title =
          element_text(size=
                         16,
                       hjust=.5,face='bold'),
```

```
188         plot.subtitle = element_text(
189           size=14,hjust=.5,face='italic'
190         ),
191         legend.title = element_text(hjust=.5),
192         axis.title = element_text(face='bold'),
193         axis.text=element_text(face='bold'))+theme_light
              ()
194
195 #relationships between predictor and response
196 #---------------------------------------
197
198 ggplot(data=NULL,aes(x=LotArea,y=SalePrice))+
199   geom_point()+
200   labs(title = 'Scatterplot of Sale Price of House \n vs
        Area of the property(in sqft)',x=
201           'Lot area',y='Sale Price of House(in dollars)')
              +
202   theme(plot.title =
203           element_text(size=
204                          16,
205                        hjust=.5,face='bold'),
206         plot.subtitle = element_text(
207           size=14,hjust=.5,face='italic'
208         ),
209         legend.title = element_text(hjust=.5),
210         axis.title = element_text(face='bold'),
```

```
211         axis.text=element_text(face='bold'))
212
213 ggplot(data=data_new,aes(x=as.factor(BedroomAbvGr),y=
      SalePrice
214                          ,fill=BedroomAbvGr))+
215   geom_boxplot()+
216   labs(title = 'Boxplot of Price of House
217        with respect to number of bedrooms',
218        x='Number of bedrooms',
219        y='Price of House')+
220   theme(plot.title =
221           element_text(size=
222                          16,
223                        hjust=.5,face='bold'),
224         plot.subtitle = element_text(
225           size=14,hjust=.5,face='italic'
226         ),
227         legend.title = element_text(hjust=.5),
228         axis.title = element_text(face='bold'),
229         axis.text=element_text(face='bold'))
230
231
232 ggplot(data=data_new,aes(x=Neighborhood,y=SalePrice,
233                          fill=Neighborhood))+
234   geom_boxplot()+
235   labs(title = 'Boxplot of Price of House with respect
```

```
    to the Neighborhood',
236       x='Neighborhood',
237       y='Price of House')+
238   theme(plot.title =
239           element_text(size=
240                        16,
241                       hjust=.5,face='bold'),
242         plot.subtitle = element_text(
243           size=14,hjust=.5,face='italic'
244         ),
245         legend.title = element_text(hjust=.5),
246         axis.title = element_text(face='bold'),
247         axis.text=element_text(face='bold'))
248
249
250 ggplot(data=data_new,aes(x=as.factor(MSSubClass),y=
     SalePrice
251                          ,fill=as.factor(MSSubClass)))+
252   geom_boxplot()+
253   labs(title = 'Boxplot of Sale Price of House with
      respect to the type of dwelling',
254       x='Type of Dwelling',
255       y='Price of House')+
256   theme(plot.title =
257           element_text(size=
258                        16,
```

```
                        hjust=.5,face='bold'),
        plot.subtitle = element_text(
          size=14,hjust=.5,face='italic'
        ),
        legend.title = element_text(hjust=.5),
        axis.title = element_text(face='bold'),
        axis.text=element_text(face='bold'))

ggplot(data=data_new,aes(x=as.factor(OverallQual),y=
   SalePrice
                         ,fill=as.factor(OverallQual)))+
  geom_boxplot()+
  labs(title = 'Boxplot of Price of House \n with
     respect to overall quality of the house',
       y='Price of House',
       x='Overall Quality')+
  theme(plot.title =
          element_text(size=
                         16,
                       hjust=.5,face='bold'),
        plot.subtitle = element_text(
          size=14,hjust=.5,face='italic'
        ),
        legend.title = element_text(hjust=.5),
        axis.title = element_text(face='bold'),
        axis.text=element_text(face='bold'))
```

```
283
284
285 ggplot(data=data_new,aes(x=BldgType,y=SalePrice
286                          ,fill=BldgType))+
287   geom_boxplot()+
288   labs(title = 'Boxplot of Price of House \n with
      respect to Typr of Dwelling',
289        y='Price of House',
290        x='Type of Dwelling')+
291   theme(plot.title =
292           element_text(size=
293                          16,
294                        hjust=.5,face='bold'),
295         plot.subtitle = element_text(
296           size=14,hjust=.5,face='italic'
297         ),
298         legend.title = element_text(hjust=.5),
299         axis.title = element_text(face='bold'),
300         axis.text=element_text(face='bold'))
301
302
303 summary(data_new)
304
305 #correlation heatmap
306 #-------------------------------------------
307 data.na.omit = na.omit(data_new)
```

```
corr = data.matrix(cor(data.na.omit[sapply(data.na.omit,
                                           is.numeric)])
                                              )
mel = melt(corr)
ggplot(mel, aes(Var1,Var2))+geom_tile(aes(fill=value)) +
  geom_text(aes(label = round(value, 1)))+
  scale_fill_gradient2(low='blue',mid = 'White' ,high='
     red') +
  labs(title = 'Correlation Heatmap')

#Partial Correlation heatmap
#-------------------------------------------------
partial.cor_new = corpcor::cor2pcor(cov(
  data_new[,which(sapply(data_new[,-c(77)],is.numeric))
     ]))


colnames(partial.cor_new)=colnames(
  data_new[,which(sapply(data_new[,-c(77)],is.numeric))
     ])
rownames(partial.cor_new)=colnames(
  data_new[,which(sapply(data_new[,-c(77)],is.numeric))
     ])
mel.partial_new = melt(data.matrix(partial.cor_new))
ggplot(mel.partial_new, aes(Var1,Var2))+geom_tile(
  aes(fill=value)) +
```

```
329   geom_text(aes(label = round(value, 1)))+
330   scale_fill_gradient2(low='blue' ,mid='white',high='red
        ') +
331   labs(title = 'Partial Correlation Heatmap')
332
333 #transforming skewed features
334 #------------------------------------------------
335
336 data_new$SalePrice=log(SalePrice)
337
338 ggplot(data=data_new,aes(x=SalePrice,y=..density..))+
339   geom_histogram(bins=50,fill='green',col='black')+
340   labs(title = 'Histogram of Price after Log
        transformation',x='Price of House (in Dollars)')+
341   theme(plot.title =
342           element_text(size=
343                          16,
344                        hjust=.5,face='bold'),
345         plot.subtitle = element_text(
346           size=14,hjust=.5,face='italic'
347         ),
348         legend.title = element_text(hjust=.5),
349         axis.title = element_text(face='bold'),
350         axis.text=element_text(face='bold'))+theme_light
              ()
351
```

```
352 data_pred_scale = subset(data_new, select= -c(SalePrice)
    )
353
354 dum_pred = dummy_cols(data_pred_scale, remove_first_
    dummy = TRUE)
355
356
357 dum_pred_data = subset(dum_pred, select=-c(
358   MSSubClass,MSZoning,Street,Alley,LotShape,LandContour,
      Utilities,LotConfig,
359   LandSlope,Neighborhood,Condition1,Condition2,BldgType,
      HouseStyle,OverallCond,
360   RoofStyle,RoofMatl,Exterior1st,Exterior2nd,MasVnrType,
      ExterQual,ExterCond,
361   Foundation,BsmtQual,BsmtCond,BsmtExposure,BsmtFinType1
      ,BsmtFinType2,Heating,
362   HeatingQC,CentralAir,Electrical,KitchenQual,Functional
      ,FireplaceQu,GarageType,
363   GarageFinish,GarageQual,GarageCond,PavedDrive,Fence,
      MiscFeature,SaleType,
364   SaleCondition,MoSold,agegr,PoolQC
365 ))
366
367
368
369 #Lasso for variable selection
```

```
370 fit.lasso= cv.glmnet(data.matrix(dum_pred_data),
371                      data_new$SalePrice,nfolds=10)
372 plot(fit.lasso)
373 fm.lasso=glmnet(data.matrix(dum_pred_data),data_new$
      SalePrice , alpha = 1)
374 plot(fm.lasso, xvar = "dev", label = TRUE,main='
      Coefficients vs Fraction Deviance')
375 s.cv <- c(lambda.min = fit.lasso$lambda.min, lambda.1se
      = fit.lasso$lambda.1se)
376 r=round(coef(fit.lasso, s = s.cv), 5)
377 summ=summary(r)
378 sparse=data.frame(Origin      = rownames(r)[summ$i],
379                   Destination = colnames(r)[summ$j],
380                   Weight      = summ$x)
381 writexl::write_xlsx(sparse,'C:/Users/Saheli/Desktop/
      SAHELI/Mstat 2022/Year 1 -Delhi/ISI-Assignments and
      notes/Deepayan sir/SEM 1 project/sparse.xlsx')
382
383
384
385
386 #significant predictors
387 set.seed(1234)
388 rownames(coef(fit.lasso, s = 'lambda.1se'))[coef(fit.
      lasso, s = 'lambda.1se')[,1]!= 0]
389 lasso_pred = subset(dum_pred_data,
```

```
390                     select= c(
391                       LotArea,LotFrontage,HalfBath,
                             BedroomAbvGr,
392                       TotRmsAbvGrd,EnclosedPorch,
393                       MSSubClass_30,MSSubClass_60,
                             MSSubClass_80,
394                       MSSubClass_85,MSSubClass_90,
                             MSSubClass_120,
395                       MSSubClass_180,MSZoning_FV,
                             MSZoning_RL,
396                       LotShape_IR2,Alley_Pave,LotShape_
                             IR3,
397                       LotConfig_Inside,LandSlope_Mod,
                             LandSlope_Sev,
398                       Neighborhood_Blueste,Neighborhood_
                             BrkSide,
399                       Neighborhood_CollgCr,Neighborhood_
                             Crawfor,
400                       Neighborhood_SWISU,Condition1_PosN
                             ,Condition1_RRAe,
401                       BldgType_Twnhs,BldgType_TwnhsE,
                             HouseStyle_2.5Unf,
402                       OverallQual_2,OverallCond_4,
                             OverallCond_6,
403                       OverallCond_7,Exterior2nd_BrkFace,
404                       Exterior2nd_Plywood,
```

```
                        Foundation_CBlock,Heating_Grav,
                           HeatingQC_Gd,HeatingQC_Po,
                        Functional_Min1,FireplaceQu_TA,
                           GarageType_Attchd,
                        GarageQual_Gd,GarageCond_Po,
                           PavedDrive_P,MoSold_10,SaleType
                           _CWD,
                        SaleType_New,SaleType_Oth,`agegr_
                           20~29`,`agegr_60~69`,
                        `agegr_70~79`,`agegr_more than
                           100`


                      ))
lasso_data = cbind(lasso_pred,price=data_new$SalePrice)
split = sample.split(data_new$SalePrice, SplitRatio =
   0.8)
price_train2 = subset(data_new$SalePrice, split == TRUE)
training_set2 = cbind(subset(lasso_pred, split == TRUE),
   price=price_train2)
price_test2 = subset(data_new$SalePrice, split == FALSE)
test_set2 = cbind(subset(lasso_pred, split == FALSE),
   price=price_test2)

l=lm(price~.,training_set2)
s1=data.frame(summary(l)$coefficients) #.9588
```

```
s2=cbind(predictors=rownames(s1),s1)

writexl::write_xlsx(s2,'C:/Users/Saheli/Desktop/SAHELI/
   Mstat 2022/Year 1 -Delhi/ISI-Assignments and notes/
   Deepayan sir/SEM 1 project/lassofit.xlsx')
residualPlot(l,main='residual plot')
qqPlot(resid(l),main='QQ-plot')
h1=hatvalues(l)
e1=residuals(l)
PRESS1=sum((e1/(1-h1))^2)
PRESS1
which(h1==1)
training_set2[c(300,544,844,1087),]
h1=hatvalues(l)
r1=studres(l)

plot(r1,h1,main='Studentised Residual vs Hatvalues',xlab
   =
       'Studentised Residual',ylab='Hatvalue')

A2 = data.frame(actual_price =exp(price_test2),
                Fitted_price=exp(predict(l, newdata=test
                   _set2)))

meltdata2=melt(A2)
ggplot(data=meltdata2,aes(value,fill=variable))+
```

```
444   geom_density(alpha=.3)+
445   labs(title = 'Density Plot of Actual House Price',
446        x='Actual House Price(in Dollars)')+
447   theme(plot.title =
448           element_text(size=
449                          16,
450                        hjust=.5,face='bold'),
451         plot.subtitle = element_text(
452           size=14,hjust=.5,face='italic'
453         ),
454         legend.title = element_text(hjust=.5),
455         axis.title = element_text(face='bold'),
456         axis.text=element_text(face='bold'))
457
458
459 ## PCA
460 dd.pca.out = subset(data_new,
461                     select = -c(
462                       SalePrice))
463 dd.pca1.out=dd.pca.out
464 trans.data.out <- preProcess(dd.pca1.out[ ,
465                                           unlist(lapply(
                                              dd.pca1.out
                                              , is.
                                              numeric)) ]
                                               ,
```

```
                                method  = "pca") #
                                   preprocess scales and
                                   centers

transformedData.out <- predict(trans.data.out, dd.pca1.
   out)
factoextra::fviz_eig(prcomp(dd.pca1.out[ , unlist(lapply
   (
  dd.pca1.out, is.numeric)) ] ,
  center = TRUE,scale. = TRUE),ncp = 21)
PC.out= as.data.frame(trans.data.out$rotation)
PC.out$name=rownames(PC.out)
writexl::write_xlsx(PC.out,'C:/Users/Saheli/Desktop/
   SAHELI/Mstat 2022/Year 1 -Delhi/ISI-Assignments and
   notes/Deepayan sir/SEM 1 project/pca comp.xlsx')


dum = dummy_cols(transformedData.out, remove_first_dummy
    = TRUE)

dumdata = subset(dum, select=-c(
  MSSubClass,MSZoning,Street,Alley,LotShape,LandContour,
     LotConfig,
  LandSlope,Neighborhood,Condition1,Condition2,BldgType,
     HouseStyle,
```

```
483   RoofStyle,RoofMatl,Exterior1st,Exterior2nd,
484   Foundation,Heating,OverallQual,OverallCond,
485   ExterQual,ExterCond,MasVnrType,BsmtQual,BsmtExposure,
486   PoolQC,Fence,Utilities,BsmtFinType1,BsmtFinType2,
487   HeatingQC,CentralAir,Electrical,KitchenQual,Functional
          ,FireplaceQu,GarageType,
488   GarageFinish,GarageQual,GarageCond,PavedDrive,
          MiscFeature,SaleType,
489   SaleCondition,agegr,MoSold,BsmtFinType2_None,BsmtCond,
          Exterior2nd_AsphShn,
490   Exterior1st_CBlock,BldgType_Duplex,Utilities_,Fence_
          None,GarageCond_None,
491   GarageQual_None,GarageFinish_None,Heating_Wall,PoolQC_
          Gd,Fence_MnWw,
492   Exterior1st_AsphShn
493 ))
494 colnames(dumdata)
495
496 price_train = subset(data_new$SalePrice, split == TRUE)
497 training_set = cbind(subset(dumdata, split == TRUE),
       price=price_train)
498 price_test = subset(data_new$SalePrice, split == FALSE)
499 test_set = cbind(subset(dumdata, split == FALSE),price=
       price_test)
500
501 fit.pca.out = lm(price~.,
```

```
                  data=training_set)
options(max.print = 2000)
p=summary(fit.pca.out) #.9721
p1=data.frame(p$coefficients)
p2=cbind(predictors=rownames(p1),p1)

writexl::write_xlsx(p2,'C:/Users/Saheli/Desktop/SAHELI/
   Mstat 2022/Year 1 -Delhi/ISI-Assignments and notes/
   Deepayan sir/SEM 1 project/pcafit.xlsx')
residualPlot(fit.pca.out, main ='Residual Plot')

qqPlot(resid(fit.pca.out), main = 'Q-Q Plot',
       xlab = 'Theoretical Normal Quantiles',
       ylab = 'Observed Residual Quantiles')

A = data.frame(actual_price =exp(price_test),
               Fitted_price=exp(predict(fit.pca.out,
                  newdata=test_set)))

meltdata=melt(A)
ggplot(data=meltdata,aes(value,fill=variable))+
  geom_density(alpha=.6)+
  labs(title = 'Histogram of Actual Vs Predicted House
     Price',
       x='House Price(in Dollars)')+
  theme(plot.title =
```

```
524           element_text(size=
525                          16,
526                        hjust=.5,face='bold'),
527         plot.subtitle = element_text(
528           size=14,hjust=.5,face='italic'
529         ),
530         legend.title = element_text(hjust=.5),
531         axis.title = element_text(face='bold'),
532         axis.text=element_text(face='bold'))
```