ST. XAVIER'S COLLEGE(AUTONOMOUS), KOLKATA

DEPARTMENT OF STATISTICS

House Price Prediction Using Linear Regression

Dissertation

Name: Saheli Datta

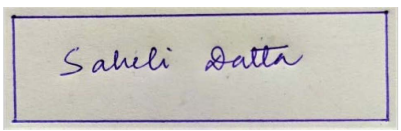
Roll Number: 415

Supervisor: Prof. Dr. Durba Bhattacharya

Session: 2019 - 2022

Date of Submission: 13 April 2022

I affirm that I have identified all my sources and that no part of my dissertation paper uses unacknowledged materials.



Student's Signature

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 -

1. To build a suitable linear regression model for predicting house price.
2. To find the housing features which have significant effect in determining house prices.

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. DURBIN-WATSON TEST:

Durbin-Watson Test is a test used to detect the presence of autocorrelation at lag 1 in the residuals from a regression analysis.

If errors e_t is the residual obtained from the regression model and

$$e_t = \rho \mathrm{e}_{\mathrm{t}-1} + \upsilon_t$$

Here, the test is given by,

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

The test statistic for this test is given by,

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n}(e_t)^2}$$

We reject the null hypothesis if $d < d_{L,\alpha}$ or $(4-d) < d_{L',\alpha}$.

Where $d_{L,\alpha}$ and $d_{L',\alpha}$ are obtained from appendices of statistical texts.

4. SHAPIRO-WILK NORMALITY TEST:

The Shapiro-Wilk test is a test for normality. It tests whether a sample x1, x2,..., xk came from a normally distributed population.

The testing hypothesis:

H_0 : The underlying population is normaly distributed

H_1 : The underlying population is not normaly distributed

The test statistic is

$$W = \frac{(\sum_{t=1}^{n} a_i x_{(i)})^2}{\sum_{t=1}^{n} (x_i - \bar{x})^2}$$

where, $x_{(i)}$ is the ith order statistic.

$\bar{x}$ is the sample mean.

The coefficients a_i are given by,

$$W = \frac{m^T V^{-}1}{\sqrt{(m^T V^{-}1 V^{-}1 m)}}$$

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. 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.

Clearly, 'SalePrice' is a continous variable. The unit of measurement of the prices is Dollar.

Table 2: Summary Table of 'SalePrice'(in Dollars)

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
135751	168703	179209	179184	186789	281644

Below the histogram of 'SalePrice' is shown:

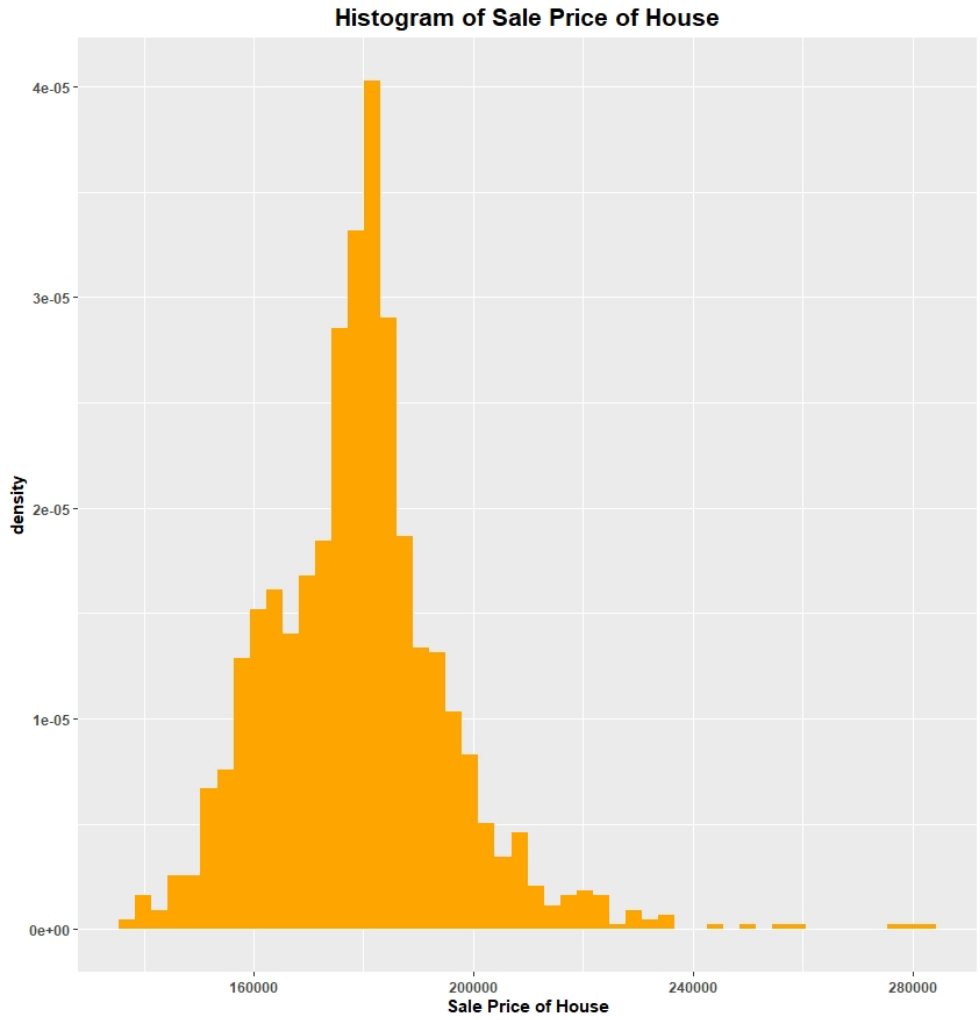


Figure 1: Histogram of 'SalePrice'

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

5.2 Analysis of Predictors

Given the dataset, we have 79 different housing parameters to be used in the regression model. Firstly, we will take a look on some of the numerical features, who intuitively, might have a great impact on determining House Prices, such as, Area of the Living Space ('GrLivArea'), size of property ('LotArea'), number of bedrooms in the house ('BedroomAbvGr')

1. GrLivArea:

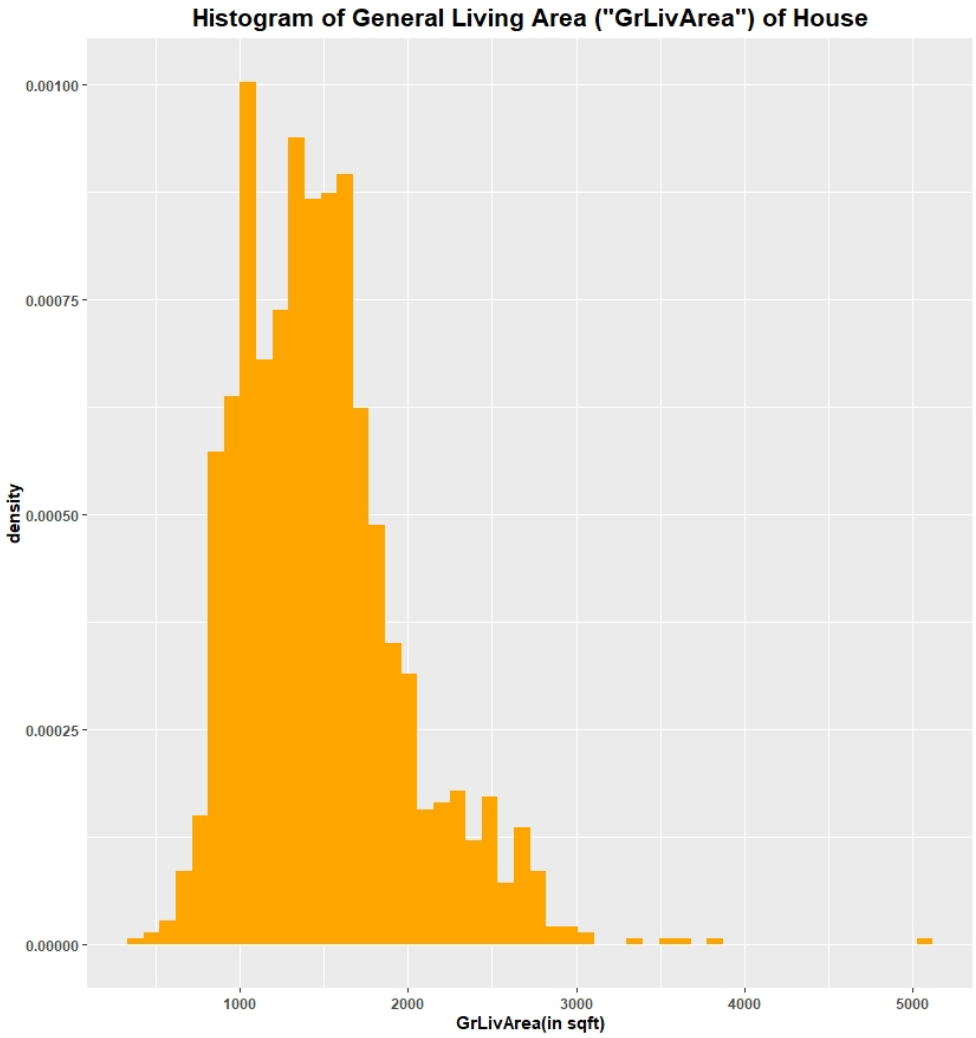


Figure 2: Histogram of 'GrLivArea'

From the histogram in Figure 2, we observe that the explanatory variable 'GrLivArea' is positively skewed.

2. LotArea:

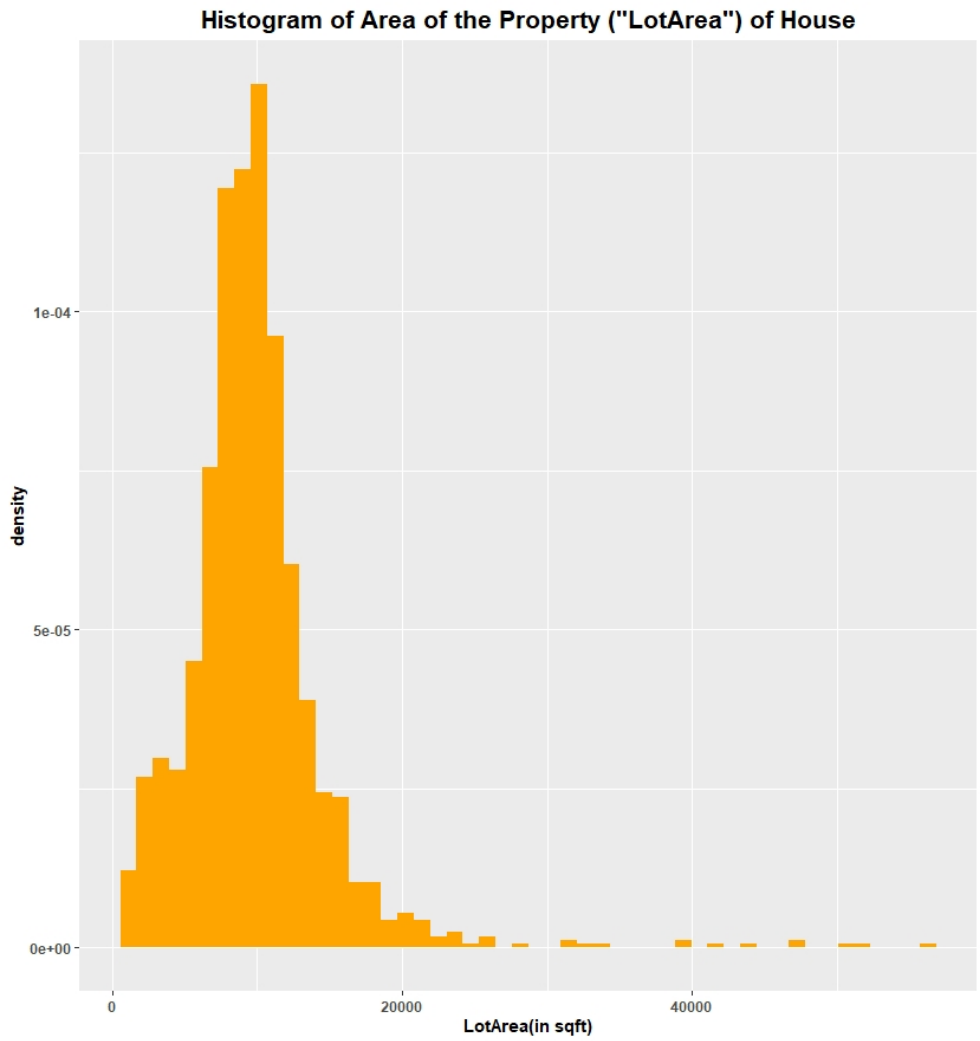


Figure 3: Histogram of 'LotArea'

From the histogram in Figure 3, we observe that the explanatory variable 'LotArea' is positively skewed.

3. BedroomAbvGr:

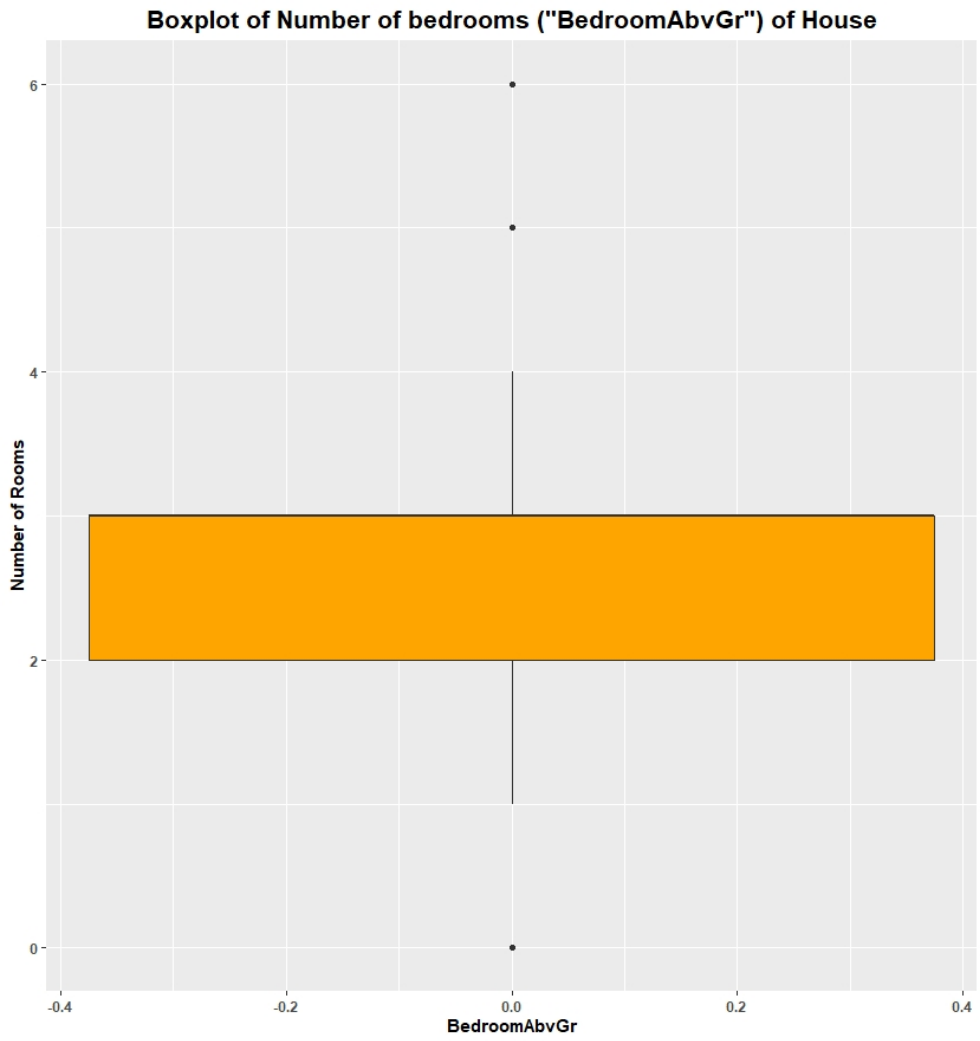


Figure 4: Boxplot of 'BedroomAbvGr'

From the Figure 4, we can see that most of the houses have 2-3 bedrooms and there is only one house which do not have a bedroom.

Missing Observations

Ames Housing contains huge amount of missing observations. Below is the graph representing the percentage amount of missing observations in each of the mentioned housing parameter group:

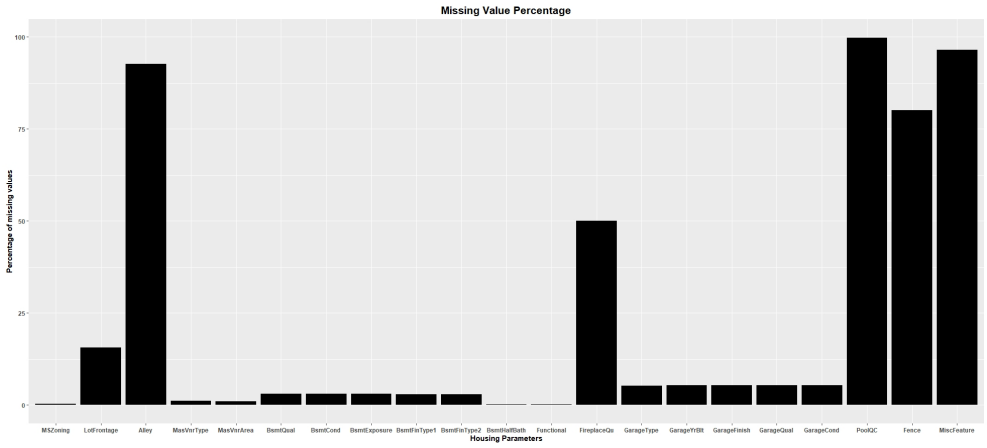


Figure 5: Plot of Missing Observation Percentage in the data

From Figure 5, we can observe that 'PoolQc','Fence','MiscVal' etc. explanatory variable has high percentage of missing observation. We need to impute the missing observation.

Skewed Features:

Given below the table that represents explanatory variables and their corresponding skewness:

Table 3: Table representing skewness of numerical explanatory variables:

Variable	Skewness
GarageYrBlt	-3.94795
YearRemodAdd	-0.39949
GarageCars	-0.10988
FullBath	0.295534
GarageArea	0.295986
BedroomAbvGr	0.436174
LotFrontage	0.616062
BsmtFullBath	0.651195
HalfBath	0.713993
TotalBsmtSF	0.804238
Fireplaces	0.819015
TotRmsAbvGrd	0.841731
2ndFlrSF	0.911944
BsmtUnfSF	0.918977
GrLivArea	1.12924
BsmtFinSF1	1.16513
1stFlrSF	1.556592
WoodDeckSF	2.128569
MasVnrArea	2.546947
OpenPorchSF	2.685015
LotArea	3.112013
BsmtHalfBath	3.779085
ScreenPorch	3.784349
BsmtFinSF2	4.038796
KitchenAbvGr	4.07486
EnclosedPorch	4.664371
3SsnPorch	12.51134
LowQualFinSF	16.15063
MiscVal	20.05454
PoolArea	20.17612

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 in the Dataset

In order to choose a suitable linear regression model it is necessary to grab an idea about how the response('SalePrice') and the explanatory variables are related. We may use graphical representations and correlation heatmap to understand how the variables are interrelated. In this way we also may get an overview if there are outliers, leverage points present in the dataset. 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 SalePrice vs LotArea:

Figure 6: Scatterplot of sale price of house('SalePrice') vs Area of the property('LotArea')

From Figure 6, we observe that house price is highly positively correlated with area of the property, hence we must include 'LotArea' as a predictor in the model.

2. Scatterplot of SalePrice vs GrLivArea:

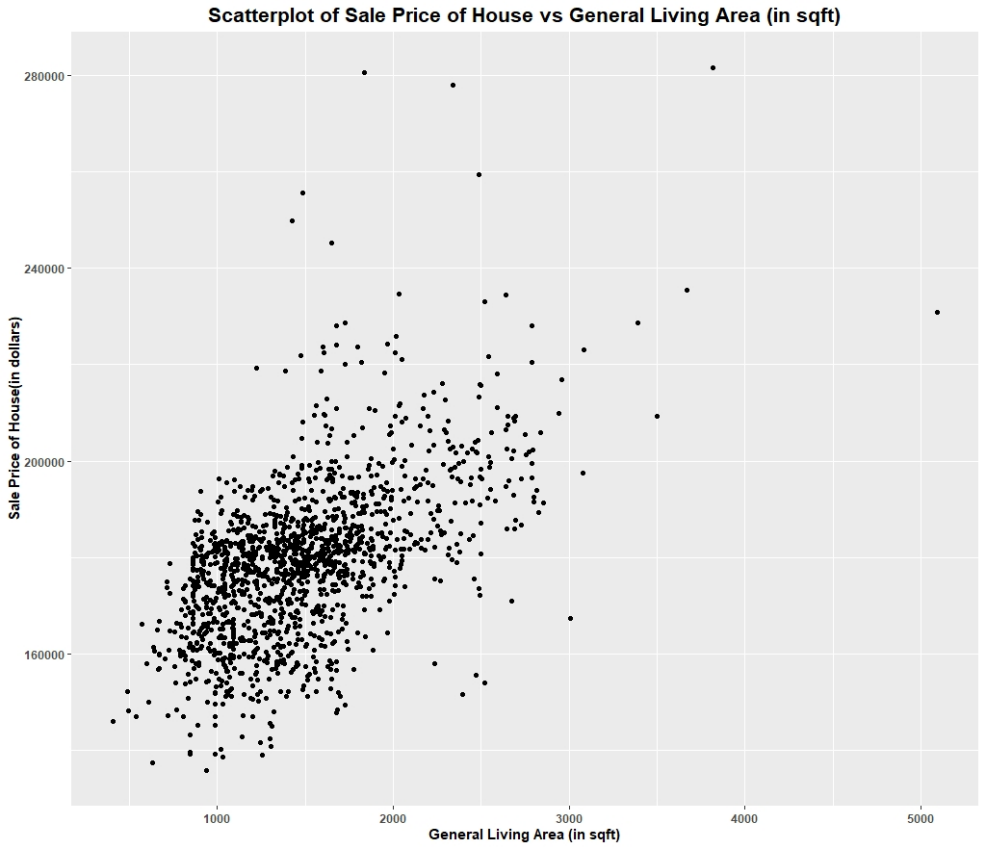


Figure 7: Scatterplot of sale price of house('SalePrice') vs Area of the Living Space('GrLivArea')

From Figure 7, we observe that house price is highly positively correlated with area of the living space of house, hence we must include 'GrLivArea' as a predictor in the model. Also, we see some potential influential points in the graph.

3. Boxplot of SalePrice vs BedroomAbvGr:

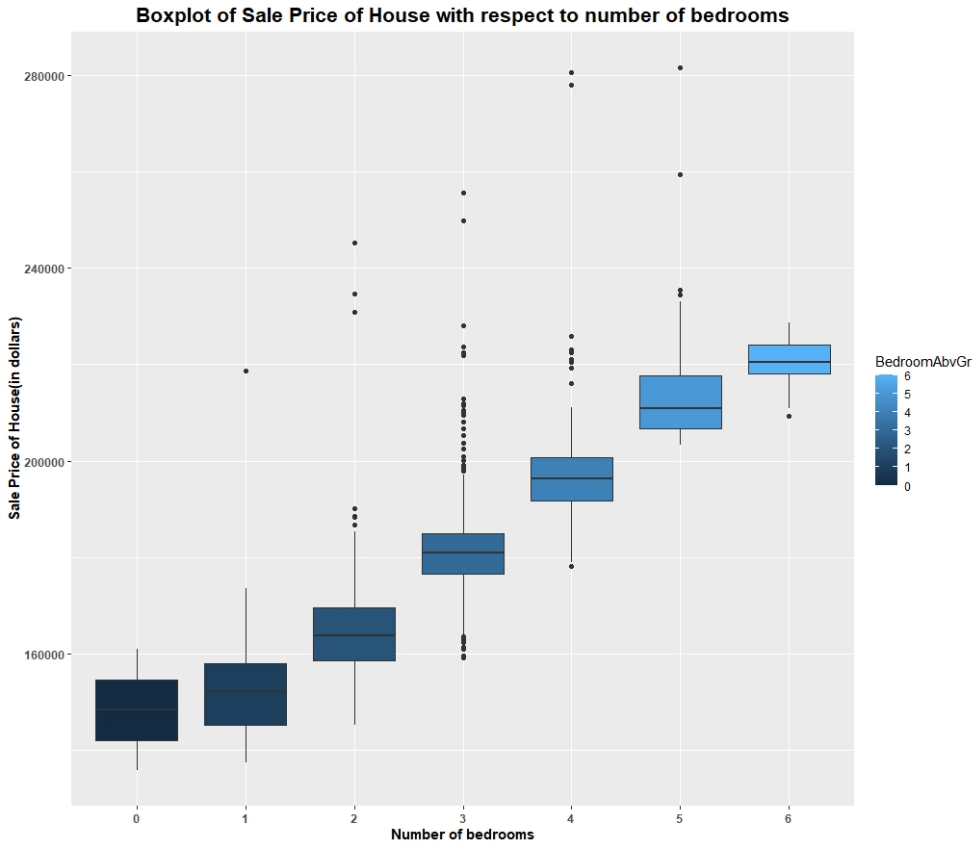


Figure 8: Boxplot of sale price of house('SalePrice') vs Number of bedrooms('BedroomAbvGr')

From Figure 8, we observe that as number of bedrooms increases, the average house price also increases, which implies number of bedrooms and price of house are very much positively correlated.

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 multi-collinearity in the model. Hence, we may ignore one of the variable which have high correlation with another predictor.

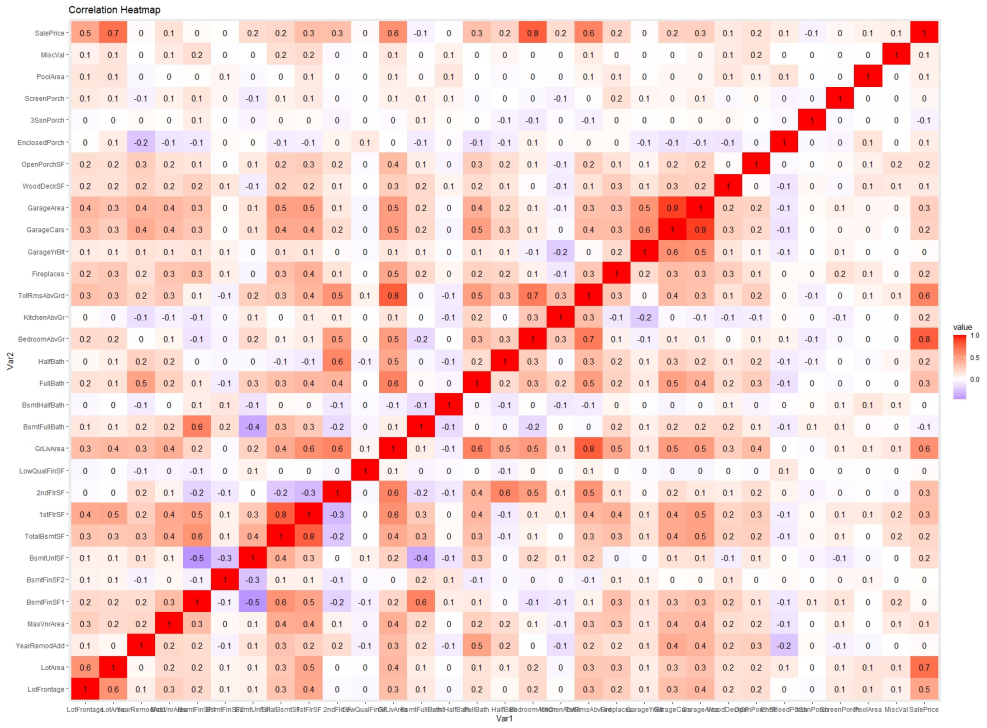


Figure 9: Correlation Heatmap

[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.

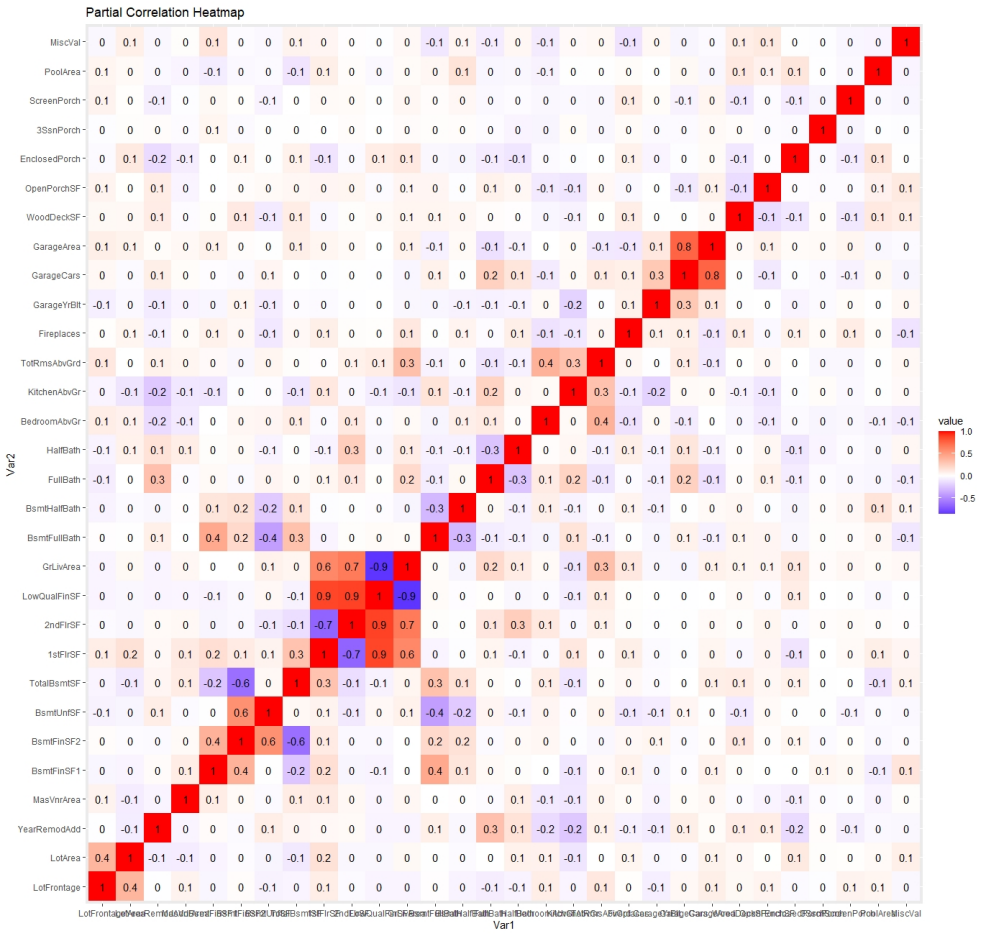


Figure 10: Partial Correlation Heatmap

[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 Data Modification and Transformation

From Table 3, we have seen that many of the predicting features are highly skewed in nature. One of the basic assumptions of multiple linear regression is that,the errors are Normally Distributed, which implies that the error distribution is symmetric in nature. Having skewed features in the model may violate such assumption and model accuracy may get reduced.

To prevent such model assumption violation, it is wise to reduce the skewness of the highly skewed predictors and response and then include them in the model. One way to reduce skewness of variables is to transform the variables.

In this case we will be using two major transformation technique:

a. Log Transformation: Transformed feature = log(skewed feature)

b. Square-root Transformation: Transformed feature = $\sqrt{('skewed\,feature')}$

Note:

1. Log transformation is possible if and only if the skewed feature takes positive values.

2. Square-root Transformation is possible if and only if the skewed feature takes non-negative values.

Given below the table showing skewness of the variables before and after transformation.

Table 4: Table representing skewness of numerical variables (response and predictors) before and after transformation:

Variable	before Transformation	after Transformation	Transformation Type
SalePrice	0.9284042	0.3784614	Log
YearRemodAdd	-0.39949	-.4063642	Log
FullBath	0.295534	0.09835095	Log
GarageArea	0.295986	-0.1572788	Log
BedroomAbvGr	0.436174	-0.02527827	Log
LotFrontage	0.616062	-0.2985982	square-root
GrLivArea	1.12924	.03759557	Log
1stFlrSF	1.556592	0.05247267	square-root
WoodDeckSF	2.128569	0.5282867	square-root
MasVnrArea	2.546947	1.084573	square-root
OpenPorchSF	2.685015	0.601642	square-root
LotArea	3.112013	0.7353648	square-root
ScreenPorch	3.784349	2.93836	square-root
EnclosedPorch	4.664371	2.235721	square-root
3SsnPorch	12.51134	10.99729	square-root
LowQualFinSF	16.15063	11.84683	square-root
MiscVal	20.05454	9.408631	square-root
PoolArea	20.17612	17.01934	square-root

We use these transformed variables in our regression model.

Also, we create a new variable namely, 'age', where,

age = YrSold - YearBuilt

7 Model Building and Model Modification

After the variable transformation and data modification, our response variable becomes $\log(SalePrice)$. Out of 79 explanatory variable in the original dataset, we now use 78 variables including the 'age' variable and excluding 'YrSold' and 'YearBuilt'. These 78 explanatory variables also includes 17 transformed variable.

7.1 Finding the Principal Components:

We observe that there is a huge number of numerical predictors in the given dataset. Clearly it is difficult to include all those variable in the model. Hence, one way to reduce the number of predictor variable 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 Partial Correlation heatmap (Figure: 10) 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 22 principal components.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7
LotFrontage	0.203326	-0.06322	0.258549	-0.14124	0.244533	-0.15787	0.168732
LotArea	0.207419	-0.0536	0.268735	-0.18204	0.26937	-0.08559	0.184723
YearRemodAdd	0.198942	0.006511	-0.24509	0.258146	-0.17793	0.158382	-0.03525
MasVnArea	0.217374	-0.09756	-0.02209	-0.01573	-0.11861	0.093373	0.098364
BsmtFinSF1	0.178288	-0.34364	-0.08409	-0.22267	-0.27464	-0.04067	0.060921
BsmtFinSF2	0.006476	-0.12656	-0.03047	-0.23648	0.170924	0.007486	-0.33765
BsmtUnfSF	0.090845	0.193354	0.327975	0.491213	0.047851	0.115257	0.038638
TotalBsmtSF	0.275122	-0.21246	0.224924	0.161576	-0.1666	0.074874	-0.03388
1stFlrSF	0.2931	-0.1963	0.304847	0.06245	-0.08403	0.038626	-0.03678
2ndFlrSF	0.115546	0.419782	-0.29025	-0.19932	0.001092	0.036643	0.054869
LowQualFinSF	-0.01065	0.044045	0.112643	0.011086	0.068788	-0.00309	-0.0344
GrLivArea	0.338801	0.215755	0.002904	-0.09338	-0.05747	0.029095	-0.00875
BsmtFullBath	0.111263	-0.3041	-0.12507	-0.2265	-0.29681	-0.27689	-0.0275
BsmtHalfBath	-0.02017	-0.0837	0.035922	-0.1171	0.161344	0.487812	-0.3245
FullBath	0.258955	0.176657	-0.03193	0.150014	-0.16271	0.01272	-0.21812
HalfBath	0.133322	0.251049	-0.33386	-0.16146	-0.00534	0.088883	0.204127
BedroomAbvGr	0.138882	0.35498	0.14203	-0.17698	-0.00465	-0.08146	-0.09056
KitchenAbvGr	0.02096	0.181128	0.240425	-0.07802	-0.27773	-0.34277	-0.34613
TotRmsAbvGrd	0.268266	0.296558	0.124131	-0.15539	-0.07823	-0.05028	-0.05994
Fireplaces	0.218797	-0.07574	-0.05459	-0.20894	0.094859	0.068434	0.026771
GarageYrBlt	0.1387	-0.11809	-0.27491	0.202053	0.410369	-0.16428	-0.01618
GarageCars	0.30788	-0.05005	-0.19139	0.200498	0.204514	-0.15672	-0.09835
GarageArea	0.303848	-0.09319	-0.15741	0.195292	0.247385	-0.14455	-0.05276
WoodDeckSF	0.148653	-0.09558	-0.13831	-0.12526	0.014377	0.224617	-0.31173
OpenPorchSF	0.159209	0.032097	-0.00062	0.023517	-0.10005	0.275288	0.332621
EnclosedPorch	-0.04401	0.066314	0.205533	-0.20136	0.251444	-0.02298	0.061167
ScreenPorch	0.001659	-0.07875	-0.03264	0.010775	-0.09902	0.02727	-0.14699
ScreenPorch	0.040203	-0.08729	0.003197	-0.14917	0.087882	-0.11267	0.314287
PoolArea	0.034911	-0.02856	0.092328	-0.17451	0.249048	0.260884	-0.25368
MiscVal	0.055842	-0.05611	0.103014	-0.08409	-0.12168	0.418232	0.252735

Variable	PC8	PC9	PC10	PC11	PC12	PC13	PC14
LotFrontage	-0.13447	-0.29588	0.095754	0.052726	0.193626	0.125876	0.159066
LotArea	-0.04649	-0.28383	0.057319	0.081205	0.198292	0.098903	0.105215
YearRemodAdd	0.058002	-0.12934	0.160627	-0.19368	0.142494	0.073303	0.165167
MasVnArea	-0.0934	0.161172	-0.08789	-0.00408	0.059965	-0.02811	-0.43457
BsmtFinSF1	0.066009	0.08135	-0.02176	0.159096	-0.20813	0.138271	0.049892
BsmtFinSF2	-0.1699	-0.07022	-0.20624	-0.43054	0.303459	-0.56033	-0.01616
BsmtUnfSF	0.011845	0.052077	0.049776	-0.05824	0.107285	0.012141	-0.20341
TotalBsmtSF	0.011744	0.106864	-0.05541	-0.06564	0.013048	-0.06932	-0.15578
1stFlrSF	-0.02681	0.052733	0.012523	-0.05158	-0.00474	-0.02938	-0.06839
2ndFlrSF	0.032524	0.007245	0.030068	0.041813	-0.01501	-0.02881	-0.02139
LowQualFinSF	0.503725	0.459471	-0.31899	-0.00288	0.316716	0.1068	0.453139
GrLivArea	0.060789	0.114662	0.041011	-0.0242	0.017331	-0.03805	-0.03206
BsmtFullBath	0.174608	-0.07203	-0.00989	-0.14398	-0.05282	-0.00943	0.059189
BsmtHalfBath	-0.33652	0.26912	0.021547	0.192117	-0.16992	-0.03186	0.150574
FullBath	-0.01303	0.051332	0.107697	-0.0749	-0.06947	-0.01938	0.184396
HalfBath	-0.00416	-0.05053	-0.05412	0.003944	0.090015	-0.00286	-0.18807
BedroomAbvGr	-0.14819	0.021401	-0.08426	0.104272	0.009614	0.010292	0.064492
KitchenAbvGr	-0.10287	-0.05231	-0.08923	0.069838	-0.26711	-0.07085	0.086855
TotRmsAbvGrd	-0.01818	0.038201	-0.06659	0.055031	-0.02538	-0.00408	0.026547
Fireplaces	0.088958	0.272377	0.144793	-0.0268	0.109362	0.04799	-0.27332
GarageYrBlt	0.000626	0.027555	-0.20259	0.144712	-0.18716	-0.09756	0.05704
GarageCars	0.012403	-0.0117	-0.07075	0.07878	-0.15038	-0.03306	0.035638
GarageArea	-0.00376	-0.01606	-0.08661	0.099769	-0.14921	-0.06665	0.051848
WoodDeckSF	0.067418	-0.17312	-0.12071	0.018262	0.328895	0.389605	-0.04175
OpenPorchSF	0.1333	-0.14905	0.150309	-0.28386	-0.1638	-0.37042	0.350659
EnclosedPorch	0.483576	0.079084	-0.04564	-0.02685	-0.25375	-0.21204	-0.30448
ScreenPorch	0.216068	-0.06172	0.426145	0.641729	0.296418	-0.44519	0.018951
ScreenPorch	-0.35324	0.502221	0.269717	-0.05405	0.01284	-0.01205	0.216163
PoolArea	0.24731	-0.13047	0.372538	-0.1779	-0.36292	0.219066	0.045313
MiscVal	-0.03599	-0.20523	-0.5094	0.286075	-0.15385	-0.1065	0.105776

Variable	PC15	PC16	PC17	PC18	PC19	PC20	PC21	PC22
LotFrontage	-0.14214	-0.11324	0.082042	0.150888	-0.07354	-0.01015	-0.36546	0.195432
LotArea	-0.13119	0.054626	-0.00076	0.159945	0.056202	0.032625	0.349488	-0.05438
YearRemodAdd	-0.10975	0.202853	0.25489	0.281404	0.260811	-0.02173	-0.45578	-0.34056
MasVnArea	-0.06074	-0.52819	0.371623	0.09895	-0.33758	-0.25311	-0.10065	-0.01697
BsmtFinSF1	-0.20781	0.043524	0.017173	-0.09431	0.061048	0.047087	0.130042	-0.19705
BsmtFinSF2	0.096102	-0.01077	0.070433	0.001929	0.143722	-0.04114	0.050586	-0.02062
BsmtUnfSF	0.156976	0.054815	-0.02906	-0.13044	0.038879	0.221362	-0.12058	0.286798
TotalBsmtSF	-0.02023	0.094422	0.017025	-0.22462	0.158234	0.250186	0.034724	0.072319
1stFlrSF	0.047249	-0.02098	-0.08505	0.002292	0.084175	-0.00365	0.178944	-0.29052
2ndFlrSF	-0.00756	0.05231	0.079282	-0.04995	-0.03291	-0.06366	0.00338	0.305995
LowQualFinSF	-0.01244	-0.26545	0.035664	0.100475	0.025161	0.058417	0.00503	0.011089
GrLivArea	0.017296	0.033072	-0.01317	-0.0426	0.074347	-0.04567	0.145377	-0.01248
BsmtFullBath	-0.00655	0.020632	0.067503	-0.16909	0.118931	0.154015	-0.2055	0.504373
BsmtHalfBath	-0.38764	0.016044	0.00173	0.125089	-0.04935	0.275718	-0.12055	0.162057
FullBath	-0.0792	0.202508	0.101308	0.184491	-0.04269	-0.39248	0.378337	0.20081
HalfBath	0.091572	-0.17066	-0.03004	0.18039	0.192171	0.577811	0.173302	-0.13731
BedroomAbvGr	-0.18925	0.011439	0.063967	-0.50443	0.159065	-0.06331	-0.1363	-0.14085
KitchenAbvGr	0.27973	-0.07459	-0.15023	0.367844	-0.22729	0.206672	-0.16524	-0.13593
TotRmsAbvGrd	0.005774	-0.05565	-0.0176	-0.12228	0.025915	-0.04487	-0.20224	-0.14
Fireplaces	0.042798	0.106187	-0.66206	0.226821	0.092898	-0.23719	-0.22706	0.094741
GarageYrBlt	0.082673	-0.02674	-0.15205	-0.2654	0.002314	-0.13577	-0.19279	-0.23418
GarageCars	0.023432	-0.01692	0.055539	0.132992	-0.0287	0.092201	0.071764	0.163921
GarageArea	0.020288	-0.02389	0.064815	0.063099	-0.06757	0.123921	0.081793	0.096843
WoodDeckSF	0.271329	0.28748	0.012989	-0.20402	-0.5079	0.104632	0.007856	-0.06829
OpenPorchSF	-0.07526	-0.10508	-0.26007	-0.15746	-0.45795	0.095642	-0.03128	-0.05569
EnclosedPorch	-0.14372	0.445884	0.306034	0.132345	-0.15734	0.074159	-0.10851	-0.10055
ScreenPorch	0.130704	-0.06021	0.059212	-0.05741	0.009124	0.006421	-0.01713	-0.01899
ScreenPorch	0.455206	0.230119	0.269645	-0.02805	-0.06831	0.054386	-0.03273	-0.04113
PoolArea	0.341852	-0.34763	0.13042	-0.09265	0.23642	-0.05299	0.014964	0.004275
MiscVal	0.369928	0.114066	0.052251	0.127331	0.210052	-0.20712	-0.0796	0.140613

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 22 principal components.

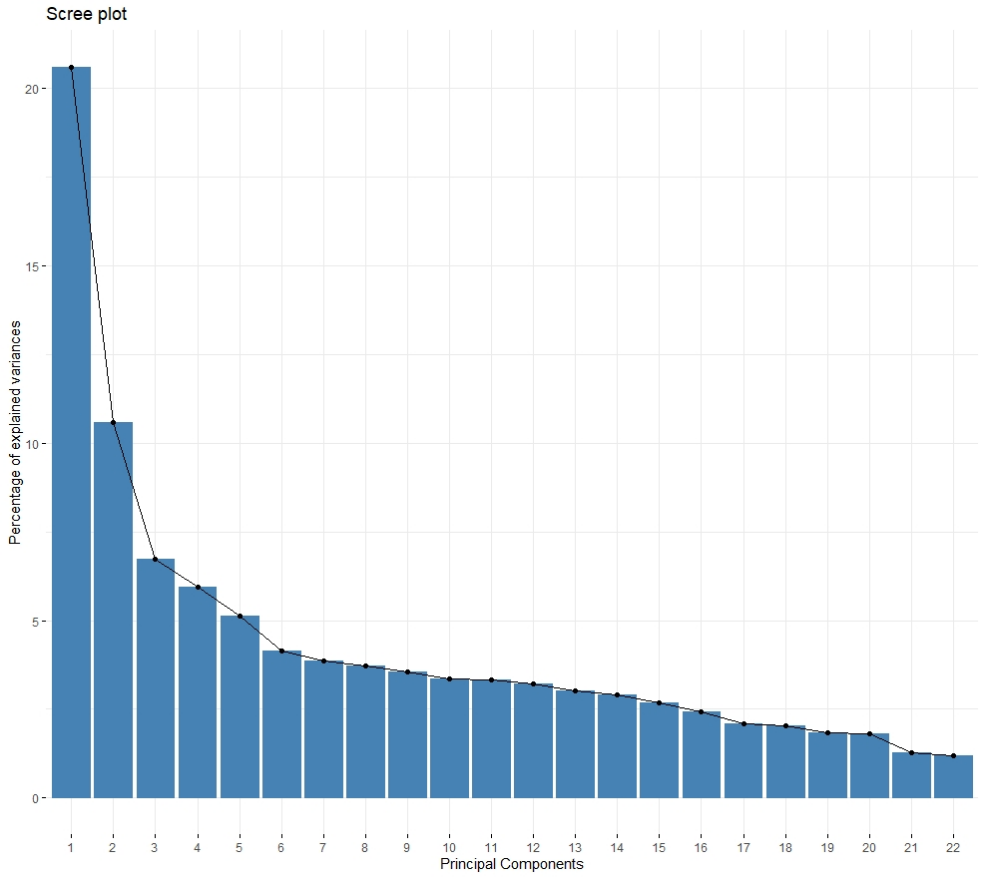


Figure 11: Scree Plot of the 22 Principal Components

From the screeplot, we observe that the first principal component more than 20% of the total variability and the 22th component explains almost 3% of total variability.

7.2 Fitting the Model and Results

Here we have total 22 principal components which explains 95 percent variation of the response which is explained by the numerical features only.

Therefore, we have total 69 predictors to include in the model, of which 46 are categorical variables, 1 numerical modified variable and 22 principal components.

Here we use the Multiple Linear Regression with log-transformed response variable to predict the House Prices.

Note that, the 46 categorical variables are dummyfied into binary variables (variables that take only two values, 1 if a particular character is present or 0 otherwise). We, choose a base lavel 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:

Coefficients	estimate	std.error	statistic	p.value
(Intercept)	12.02403	0.0822	146.2771	0*
MSSubClass150	-0.09418	0.033054	-2.84929	0.004457*
MSSubClass160	-0.00256	0.00993	-0.25827	0.796241
MSSubClass180	0.028298	0.015334	1.845378	0.065231
MSSubClass190	0.035028	0.029308	1.195189	0.232251
MSSubClass20	0.022616	0.014643	1.544443	0.122747
MSSubClass30	0.028712	0.015461	1.856981	0.063561
MSSubClass40	0.034857	0.025487	1.367609	0.171692
MSSubClass45	0.013717	0.024504	0.559771	0.575741
MSSubClass50	0.012193	0.017515	0.696128	0.486485
MSSubClass60	-0.00837	0.016652	-0.5025	0.615408
MSSubClass70	0.001625	0.017113	0.09495	0.92437
MSSubClass75	-0.00818	0.021028	-0.38902	0.697329
MSSubClass80	0.004279	0.023991	0.178377	0.858457
MSSubClass85	0.030203	0.018104	1.668331	0.095513
MSSubClass90	0.072032	0.017111	4.209722	2.75E-05*
MSZoningFV	0.000124	0.012686	0.009752	0.99222
MSZoningRH	0.001839	0.013811	0.133167	0.894084
MSZoningRL	0.00041	0.010586	0.03872	0.96912
MSZoningRM	-0.00573	0.010019	-0.57233	0.567204
StreetPave	0.023406	0.014161	1.652861	0.098623
AlleyNone	0.000862	0.004438	0.194346	0.845938
AlleyPave	0.004907	0.006965	0.704507	0.481255
LotShapeIR2	0.018795	0.004959	3.789814	0.000158*
LotShapeIR3	0.021908	0.011409	1.920176	0.055074
LotShapeReg	-0.00189	0.001828	-1.03473	0.301007
LandContourHLS	-0.00062	0.005874	-0.10557	0.915942
LandContourLow	0.031195	0.008063	3.868942	0.000115*
LandContourLvl	0.001141	0.004395	0.259646	0.795182
UtilitiesNone	0.081539	0.023104	3.52927	0.000433*
LotConfigCulDSac	0.020458	0.003934	5.200082	2.34E-07*
LotConfigFR2	0.015545	0.005072	3.064806	0.002227*
LotConfigFR3	0.008121	0.009069	0.895499	0.370701
LotConfigInside	0.005135	0.002091	2.455818	0.014198*
LandSlopeMod	-0.00428	0.004774	-0.89678	0.370019
LandSlopeSev	0.099851	0.020417	4.890528	1.14E-06*

Coefficients	estimate	std.error	statistic	p.value
NeighborhoodBlueste	0.021861	0.014174	1.542364	0.123251
NeighborhoodBrDale	0.044514	0.013766	3.233506	0.001256*
NeighborhoodBrkSide	0.018251	0.01107	1.648684	0.099476
NeighborhoodClearCr	0.000859	0.012048	0.071313	0.943161
NeighborhoodCollgCr	0.014669	0.008967	1.635923	0.10212
NeighborhoodCrawfor	0.018777	0.010128	1.853986	0.063988
NeighborhoodEdwards	0.017587	0.009644	1.823681	0.068451
NeighborhoodGilbert	0.022579	0.009324	2.421553	0.015603*
NeighborhoodIDOTRR	0.021895	0.011491	1.905475	0.056959
NeighborhoodMeadowV	0.041296	0.013692	3.016008	0.002616*
NeighborhoodMitchel	0.018802	0.009702	1.937961	0.052864
NeighborhoodNAmes	0.014807	0.00951	1.556963	0.119745
NeighborhoodNoRidge	0.002998	0.010626	0.282105	0.777912
NeighborhoodNPkVill	0.020782	0.022281	0.93269	0.351169
NeighborhoodNridgHt	0.001255	0.009449	0.132769	0.894398
NeighborhoodNWAmes	0.00174	0.009831	0.176959	0.859571
NeighborhoodOldTown	0.018567	0.010903	1.702951	0.088838
NeighborhoodSawyer	0.014073	0.00978	1.438975	0.15042
NeighborhoodSawyerW	0.010471	0.009413	1.112454	0.266167
NeighborhoodSomerst	0.00299	0.010654	0.28069	0.778997
NeighborhoodStoneBr	0.018634	0.010786	1.727583	0.084322
NeighborhoodSWISU	0.01049	0.011562	0.907215	0.364476
NeighborhoodTimber	0.007614	0.009857	0.772458	0.439996
NeighborhoodVeenker	-0.00697	0.012633	-0.55173	0.581238
Condition1Feedr	0.005407	0.005606	0.964444	0.335019
Condition1Norm	-0.00131	0.004659	-0.28036	0.779247
Condition1PosA	0.00815	0.010318	0.789935	0.429723
Condition1PosN	0.012495	0.007971	1.567452	0.117275
Condition1RRAe	0.018308	0.008254	2.218194	0.02673*
Condition1RRAn	0.010951	0.007684	1.425165	0.154371
Condition1RRNe	0.00767	0.01513	0.506935	0.612294
Condition1RRNn	0.009645	0.014489	0.665644	0.505767
Condition2Feedr	0.016113	0.019189	0.839666	0.401264
Condition2Norm	0.03416	0.016399	2.083116	0.037454*
Condition2PosA	0.009763	0.025641	0.380742	0.703462
Condition2PosN	-0.03036	0.025856	-1.17425	0.240528
BldgType2fmCon	-0.00247	0.025812	-0.09586	0.923645
BldgTypeTwnhs	0.038726	0.015444	2.50753	0.01229*
BldgTypeTwnhsE	0.029626	0.01443	2.053053	0.040285*
HouseStyle1.5Unf	0.000903	0.019225	0.046989	0.96253
HouseStyle1Story	-0.00771	0.009655	-0.79894	0.424485
HouseStyle2.5Unf	0.023286	0.013793	1.688253	0.091624
HouseStyle2Story	0.003865	0.009227	0.41885	0.675401
HouseStyleSFoyer	-0.02704	0.012355	-2.18853	0.028825*
HouseStyleSLvl	-0.00256	0.019556	-0.13091	0.895871

Coefficients	estimate	std.error	statistic	p.value
OverallQual2	-0.05721	0.057195	-1.00018	0.317424
OverallQual3	-0.04964	0.055814	-0.88932	0.374012
OverallQual4	-0.06432	0.055484	-1.15923	0.246594
OverallQual5	-0.06453	0.055663	-1.15927	0.246578
OverallQual6	-0.06476	0.055851	-1.15959	0.24645
OverallQual7	-0.06855	0.055891	-1.22643	0.22028
OverallQual8	-0.07215	0.055932	-1.28992	0.19733
OverallQual9	-0.08161	0.056144	-1.45366	0.146303
OverallQual10	-0.0898	0.057062	-1.5737	0.115822
OverallCond2	-0.03242	0.021515	-1.50701	0.132073
OverallCond3	-0.02914	0.01581	-1.84307	0.065567
OverallCond4	-0.01862	0.015294	-1.21732	0.223723
OverallCond5	-0.02962	0.015219	-1.94606	0.051882
OverallCond6	-0.0263	0.015334	-1.71492	0.08662
OverallCond7	-0.02104	0.015413	-1.36512	0.172474
OverallCond8	-0.01476	0.01551	-0.95158	0.341501
OverallCond9	-0.00761	0.016889	-0.45033	0.652557
RoofStyleGable	0.071304	0.018776	3.797544	0.000154*
RoofStyleGambrel	0.07188	0.020838	3.449436	0.000581*
RoofStyleHip	0.075207	0.018905	3.978069	7.37E-05*
RoofStyleMansard	0.058963	0.024261	2.430399	0.015229*
RoofStyleShed	0.094595	0.027678	3.417656	0.000653*
RoofMatlTarGrv	0.039249	0.013394	2.930349	0.00345*
RoofMatlWdShake	0.028116	0.014723	1.909614	0.056423
RoofMatlWdShngl	-0.17174	0.04051	-4.23944	2.41E-05*
Exterior1stAsphShn	0.033067	0.030862	1.071449	0.284185
Exterior1stBrkComm	-0.05759	0.023896	-2.41011	0.016099*
Exterior1stBrkFace	-0.02471	0.017128	-1.44236	0.149465
Exterior1stCBlock	-0.05102	0.059983	-0.85053	0.395203
Exterior1stCemntBd	0.041447	0.02774	1.494136	0.135405
Exterior1stHdBoard	-0.01722	0.016092	-1.06984	0.284907
Exterior1stMetalSd	-0.02294	0.017867	-1.28373	0.199487
Exterior1stPlywood	-0.01224	0.015681	-0.78078	0.435087
Exterior1stSdng	0.033842	0.034739	0.974167	0.330171
Exterior1stStucco	-0.02684	0.018415	-1.45732	0.145291
Exterior1stVinylSd	-0.02739	0.018301	-1.49655	0.134776
Exterior1stWd Sdng	-0.01561	0.015869	-0.98339	0.325615
Exterior1stWdShing	-0.02565	0.017037	-1.50529	0.132514

Coefficients	estimate	std.error	statistic	p.value
Exterior2ndBrk Cmn	0.001575	0.026135	0.060283	0.951941
Exterior2ndBrkFace	0.00042	0.01899	0.022114	0.982361
Exterior2ndCBlock	-0.00135	0.029782	-0.04534	0.963844
Exterior2ndCmentBd	-0.05615	0.028264	-1.98672	0.047183*
Exterior2ndHdBoard	0.003859	0.017388	0.221929	0.824407
Exterior2ndImStucc	0.013184	0.021004	0.627689	0.530328
Exterior2ndMetalSd	0.010343	0.019005	0.544203	0.586404
Exterior2ndPlywood	0.005943	0.016767	0.354437	0.723074
Exterior2ndStone	0.027343	0.034602	0.79023	0.429551
Exterior2ndStucco	0.007965	0.01944	0.409717	0.682087
Exterior2ndVinylSd	0.013421	0.019514	0.687755	0.491741
Exterior2ndWd Sdng	0.002703	0.017236	0.156834	0.875402
Exterior2ndWd Shng	0.011577	0.018041	0.641728	0.521173
MasVnrTypeBrkFace	0.007557	0.009058	0.834301	0.404279
MasVnrTypeNone	-0.0028	0.008999	-0.31133	0.755608
MasVnrTypeStone	0.001021	0.009342	0.1093	0.912983
ExterQualFa	-0.02733	0.010441	-2.61715	0.008979*
ExterQualGd	-0.01802	0.006042	-2.98252	0.002917*
ExterQualTA	-0.0137	0.0067	-2.04457	0.041117*
ExterCondFa	-0.0005	0.011467	-0.04359	0.965239
ExterCondGd	-0.0036	0.009886	-0.36402	0.715909
ExterCondPo	-0.07582	0.033416	-2.2689	0.023453*
ExterCondTA	-0.00367	0.009887	-0.37153	0.710311
FoundationCBlock	0.001929	0.003532	0.546135	0.585076
FoundationPConc	-0.00239	0.003846	-0.62063	0.534959
FoundationSlab	0.015401	0.010699	1.439544	0.150259
FoundationStone	0.026962	0.014213	1.896992	0.058071
FoundationWood	-0.01601	0.019715	-0.81222	0.41683
BsmtQualFa	0.003717	0.006459	0.575519	0.565049
BsmtQualGd	0.001319	0.003676	0.358828	0.719788
BsmtQualNone	-0.06067	0.026482	-2.29107	0.022133*
BsmtQualTA	0.002372	0.00474	0.500447	0.616853
BsmtCondGd	0.009456	0.005674	1.666493	0.095878
BsmtCondNone	-0.01586	0.017028	-0.93116	0.351961
BsmtCondPo	0.01814	0.017441	1.040069	0.298519
BsmtCondTA	0.007528	0.004294	1.753093	0.079843
BsmtExposureGd	-0.00594	0.003372	-1.76313	0.078135
BsmtExposureMn	-0.00374	0.003314	-1.12794	0.259571
BsmtExposureNo	-0.00346	0.002569	-1.34718	0.178178
BsmtExposureNone	-0.01256	0.018505	-0.67886	0.497355*

Coefficients	estimate	std.error	statistic	p.value
BsmtFinType1BLQ	-0.00683	0.003301	-2.06875	0.038785
BsmtFinType1GLQ	0.004489	0.00285	1.575014	0.115519
BsmtFinType1LwQ	-0.00325	0.003982	-0.81591	0.414714
BsmtFinType1None	0.081684	0.037939	2.153036	0.031516*
BsmtFinType1Rec	-8.60E-05	0.003237	-0.02669	0.978714
BsmtFinType1Unf	0.00338	0.003251	1.039639	0.298719
BsmtFinType2BLQ	0.003094	0.006854	0.451344	0.651824
BsmtFinType2GLQ	-0.00197	0.008441	-0.23286	0.815909
BsmtFinType2LwQ	-0.01156	0.006782	-1.70527	0.088405
BsmtFinType2Rec	0.005476	0.006414	0.853789	0.393394
BsmtFinType2Unf	0.00693	0.006392	1.084248	0.278474
HeatingGasW	-0.00835	0.010904	-0.76537	0.444204
HeatingGrav	-0.00923	0.022538	-0.40944	0.682289
HeatingQCFa	-0.00567	0.005065	-1.12038	0.262777
HeatingQCGd	0.000388	0.002309	0.167992	0.866618
HeatingQCPo	0.003128	0.028194	0.110931	0.91169
HeatingQCTA	-0.00182	0.002275	-0.79964	0.424078
CentralAirY	0.001754	0.004062	0.431885	0.665903
ElectricalFuseF	-0.00325	0.007269	-0.44772	0.65444
ElectricalFuseP	0.009212	0.013785	0.668219	0.504123
ElectricalSBrkr	0.00443	0.003331	1.330098	0.183741
KitchenQualFa	0.009205	0.007516	1.224725	0.220921
KitchenQualGd	0.006069	0.004545	1.335345	0.182019
KitchenQualTA	0.007164	0.004948	1.447836	0.147926
FunctionalMaj2	-0.00201	0.020452	-0.09831	0.921707
FunctionalMin1	-0.01931	0.015136	-1.27566	0.202326
FunctionalMin2	-0.00409	0.015374	-0.26626	0.790084
FunctionalMod	-0.00465	0.015872	-0.29279	0.769732
FunctionalSev	-0.08393	0.032557	-2.57783	0.010062*
FunctionalTyp	-0.01052	0.014311	-0.73531	0.462295
FireplaceQuFa	0.013091	0.007961	1.644357	0.100366
FireplaceQuGd	0.004609	0.006494	0.709667	0.47805
FireplaceQuNone	-0.00927	0.007348	-1.26215	0.207142
FireplaceQuPo	0.007082	0.008623	0.821275	0.411654
FireplaceQuTA	0.00516	0.00664	0.777052	0.437282
GarageTypeAttchd	-0.01362	0.007348	-1.85379	0.064017
GarageTypeBasment	-0.00844	0.010465	-0.80681	0.419938
GarageTypeBuiltIn	-0.01681	0.00801	-2.09877	0.036047*
GarageTypeCarPort	0.004711	0.013395	0.35169	0.725133
GarageTypeDetchd	-0.00352	0.007362	-0.47745	0.633129
GarageTypeNone	0.002233	0.025471	0.087673	0.930152
GarageFinishNone	-0.03798	0.039197	-0.969	0.332742
GarageFinishRFn	0.001559	0.002191	0.711481	0.476926
GarageFinishUnf	0.002146	0.002579	0.832094	0.405523

Coefficients	estimate	std.error	statistic	p.value
GarageQualGd	0.018167	0.01148	1.582415	0.11382
GarageQualPo	0.045164	0.024295	1.858948	0.063281
GarageQualTA	-0.00364	0.004036	-0.90255	0.366945
GarageCondFa	-0.02833	0.029221	-0.96961	0.332439
GarageCondGd	-0.04964	0.030022	-1.6534	0.098513
GarageCondPo	-0.02222	0.031574	-0.70381	0.48169
GarageCondTA	-0.0293	0.028724	-1.01999	0.30794
PavedDriveP	0.005462	0.005885	0.928205	0.353489
PavedDriveY	-0.00126	0.003907	-0.32184	0.747634
PoolQCGd	0.022458	0.045801	0.490332	0.623989
PoolQCNone	0.024604	0.022022	1.117227	0.264123
FenceGdWo	0.00067	0.005117	0.130852	0.895915
FenceMnPrv	0.000559	0.00421	0.132842	0.89434
FenceMnWw	0.021597	0.026611	0.811557	0.417208
FenceNone	0.000224	0.003785	0.05912	0.952867
MiscFeatureNone	-0.02242	0.01817	-1.23394	0.217467
MiscFeatureOthr	-0.02602	0.02516	-1.03398	0.301353
MiscFeatureShed	-0.01589	0.017948	-0.88519	0.376235
MoSold2	0.000121	0.004682	0.025785	0.979433*
MoSold3	0.006342	0.004297	1.476083	0.140186*
MoSold4	0.015904	0.004243	3.748445	0.000186*
MoSold5	0.020649	0.004036	5.116522	3.62E-07*
MoSold6	0.026004	0.003915	6.64275	4.67E-11*
MoSold7	0.034117	0.00399	8.551484	3.68E-17*
MoSold8	0.043359	0.004358	9.949268	1.86E-22*
MoSold9	0.051654	0.004476	11.53933	2.82E-29*
MoSold10	0.060262	0.004698	12.82772	2.24E-35*
MoSold11	0.055139	0.004925	11.19573	9.82E-28*
MoSold12	0.065919	0.005486	12.01487	1.80E-31*
SaleTypeCon	0.037512	0.017055	2.199535	0.028031*
SaleTypeConLD	-0.01129	0.00867	-1.30243	0.193021
SaleTypeConLI	-0.00524	0.014868	-0.35256	0.724482
SaleTypeConLw	-0.01354	0.016627	-0.81437	0.415594
SaleTypeCWD	0.005263	0.010504	0.501006	0.61646
SaleTypeNew	0.011959	0.017826	0.670873	0.502432
SaleTypeOth	-0.0034	0.014	-0.24274	0.80825
SaleTypeWD	0.002162	0.00465	0.465008	0.642011
SaleConditionAdjLand	-0.00789	0.011593	-0.68027	0.496463
SaleConditionAlloca	0.027014	0.009846	2.743623	0.006168
SaleConditionFamily	0.016616	0.00641	2.592387	0.009648
SaleConditionNormal	0.0025	0.003485	0.717311	0.473323
SaleConditionPartial	-0.00568	0.017364	-0.32692	0.74379

Coefficients	estimate	std.error	statistic	p.value
PC1	0.019682	0.000849	23.17762	2.16E-98*
PC2	0.01993	0.001174	16.97821	4.59E-58*
PC3	0.020711	0.00128	16.18279	2.15E-53*
PC4	-0.02048	0.001175	-17.4289	9.02E-61*
PC5	0.018068	0.001455	12.41506	2.28E-33*
PC6	-0.00664	0.001031	-6.44355	1.69E-10*
PC7	0.008724	0.001053	8.287326	3.09E-16*
PC8	-0.01042	0.00087	-11.9737	2.81E-31*
PC9	-0.01502	0.001031	-14.5683	2.30E-44*
PC10	-0.00109	0.001091	-0.99549	0.319701
PC11	0.008222	0.001041	7.899059	6.35E-15*
PC12	0.012956	0.001177	11.00556	6.75E-27*
PC13	0.004528	0.0011	4.117515	4.09E-05*
PC14	0.011348	0.001045	10.86399	2.79E-26*
PC15	-0.01999	0.001065	-18.7609	5.27E-69*
PC16	0.00274	0.001118	2.450681	0.014402*
PC17	0.007936	0.001623	4.889347	1.15E-06*
PC18	-0.02691	0.001416	-19.0093	1.41E-70*
PC19	0.013382	0.001208	11.07524	3.34E-27*
PC20	-0.00295	0.001303	-2.26478	0.023705*
PC21	0.011935	0.001541	7.74585	2.02E-14*
PC22	-0.00767	0.001577	-4.86201	1.32E-06*
age	-0.00036	0.000105	-3.40557	0.000682*

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 = 0.9373 ; Residual Standard error = 0.02495

Interpretation:

The R-squared value implies that approximately 94 % of the total variation in the response is explained by the chosen model.

7.3 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:

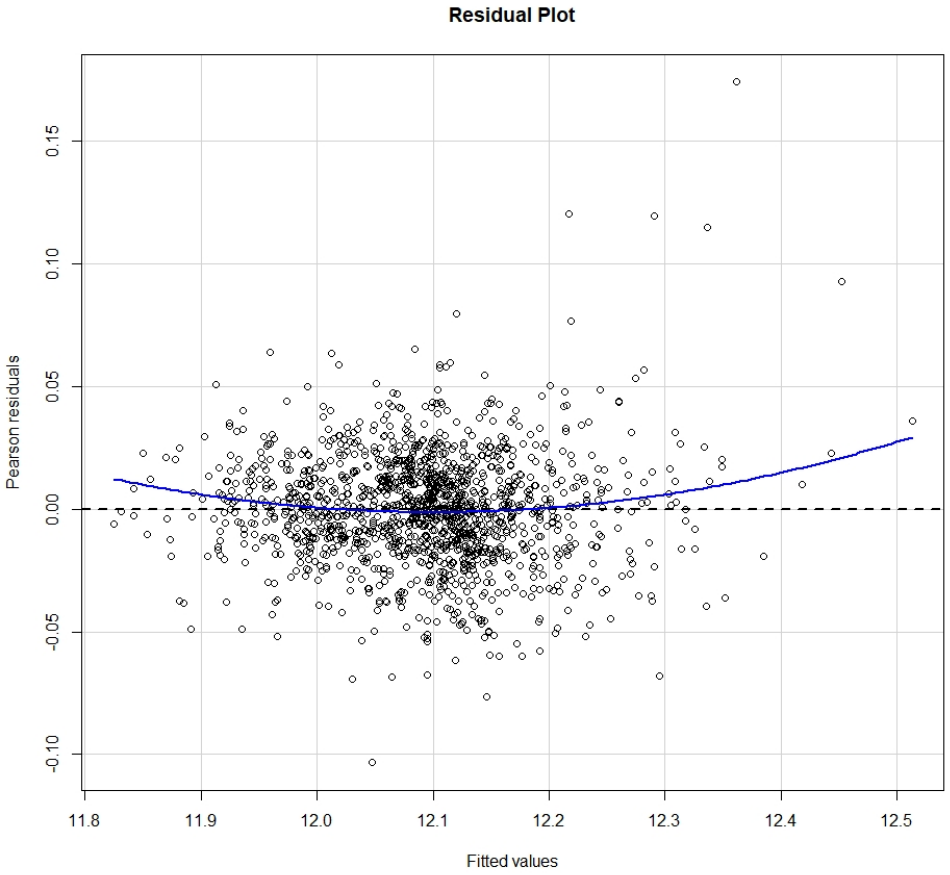


Figure 12: Residual Plot

From the above residual plot, we see that the residuals are clustered around 0 line and there are some visible outliers and leverage points present in the residuals.

2. Q-Q Plot:

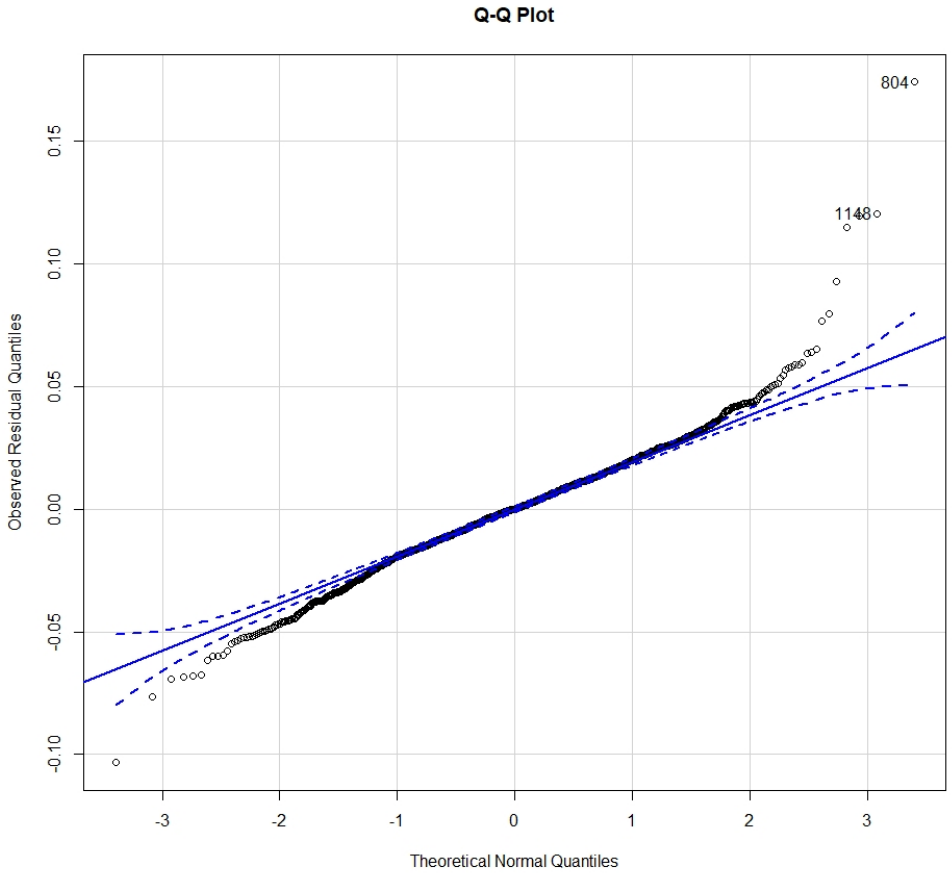


Figure 13: Q-Q Plot of Residuals

From the above q-q plot, we observe that the distribution of residuals is 'Highly Positively Skewed'. This violates the 'Normality assumption of Errors' in linear regression model.

To check whether the model really violates the normality assumption we need to test the for non-normality of the residuals.

3.Shapiro-Wilk Normality Test:

Here, the testing hypothesis is:

H_0 : The error distribution is normally distributed

H_1 : The error distribution is not normally distributed

Testing Results:

Observed Test Statistic = 0.9664 ; P-Value = 2.2e-16 (approx.)

Decision:

We reject the Null Hypothesis at 0.05 level of significance.

Conclusion:

In the light of the given data, we may conclude that The error distribution is not Normal under the chosen model.

7.4 Rebuilding the Model with appropriate predictors

Our initial model violates the assumptions of Linear Model. In order to improve the model it is important to reduce the skewness of the residuals. It is to be noted that our initial model includes many predicting features which are highly skewed

in nature, even after applying transformation techniques. It will be wise to exclude those predicting features and rebuild the model.

While rebuilding the model, the predicting features that we exclude from the model are:
'3SsnPorch','LowQualFinSF',' MiscVal', 'PoolArea', 'PoolQc', 'KitchenAbvGr', 'EnclosedPorch', 'OpenPorchSF', 'ScreenPorch', 'BsmtHalfBath', 'WoodDckSF', 'OverallQual', 'MoSold', 'BsmtFinSF2', 'MasVnrArea', 'GarageYrBuilt'.
Now we have (69 - 16) = 53 predicting features to include in the model.

Finding New Principal Components:

After removing the skewed features we now have 19 numerical features left to use in model. To reduce number of numerical predictors and moreover to eliminate multicollinearity from the model, we use Principal Component Analysis. We create a new set of orthogonal predictors which explain 95% of the variability of response that was previously explained by those 19 non-skewed variables.
After PCA transformation, we get 12 Principal Components.
Given below the transformation that is used to make those 12 principal components:

Variable	PC1	PC2	PC3	PC4	PC5	PC6
LotFrontage	-0.20281	-0.07817	0.351562	0.195118	0.392673	0.28939
LotArea	-0.20684	-0.06441	0.365582	0.228568	0.35609	0.162725
YearRemodAdd	-0.22302	-0.00942	-0.32218	-0.28234	0.013152	0.25955
BsmtFinSF1	-0.1738	-0.36669	-0.17195	0.315418	-0.23261	0.078739
BsmtUnfSF	-0.10611	0.148668	0.336013	-0.5404	-0.08892	-0.17718
TotalBsmtSF	-0.28225	-0.2609	0.154459	-0.16472	-0.30404	-0.10012
1stFlrSF	-0.29537	-0.24463	0.261645	-0.05994	-0.19317	-0.10422
2ndFlrSF	-0.12287	0.464024	-0.21922	0.201546	0.000647	-0.0091
GrLivArea	-0.35061	0.20933	0.022367	0.097771	-0.17473	-0.09889
BsmtFullBath	-0.10804	-0.3315	-0.2206	0.295423	-0.21635	0.205516
FullBath	-0.28332	0.145945	-0.09246	-0.18581	-0.19236	0.216007
HalfBath	-0.14357	0.288083	-0.30457	0.160742	0.175862	-0.20526
BedroomAbvGr	-0.15038	0.356391	0.203517	0.164755	-0.23062	0.279509
TotalRmsAbvGrd	-0.27786	0.282142	0.147045	0.152787	-0.21541	0.07212
Fireplaces	-0.21718	-0.06152	-0.01287	0.240562	-0.02165	-0.70631
GarageCars	-0.31967	-0.05364	-0.14263	-0.11682	0.356821	-0.0814
GarageArea	-0.31049	-0.0932	-0.09023	-0.10111	0.388357	-0.08799
age	0.248622	0.099448	0.337549	0.278705	-0.05996	-0.16269

Variable	PC7	PC8	PC9	PC10	PC11	PC12
LotFrontage	-0.2265	0.053236	0.099234	0.269876	-0.02032	0.614311
LotArea	-0.27066	0.055303	-0.00079	-0.34598	0.169474	-0.58638
YearRemodAdd	-0.40722	0.221916	0.166779	0.081938	-0.58995	-0.22649
BsmtFinSF1	0.070086	-0.07654	-0.33579	-0.11252	-0.17341	0.053731
BsmtUnfSF	-0.10165	-0.2461	0.320143	0.061353	0.108233	-0.02388
TotalBsmtSF	-0.02813	-0.30584	0.001339	-0.02009	-0.02781	0.004187
1stFlrSF	-0.04875	-0.02054	-0.17295	-0.20823	-0.08186	0.08312
2ndFlrSF	0.051456	0.029449	0.24919	-0.15862	-0.03812	0.074251
GrLivArea	0.006925	0.032112	0.078661	-0.2706	-0.08714	0.113897
BsmtFullBath	0.030975	-0.18546	0.677581	0.129819	0.287084	-0.09717
FullBath	0.102007	0.500568	-0.06345	-0.25126	0.486104	0.126108
HalfBath	-0.35362	-0.5434	-0.15892	-0.15252	0.150328	0.07714
BedroomAbvGr	0.166343	-0.12741	-0.2537	0.553515	0.052038	-0.35854
TotalRmsAbvGrd	0.083926	-0.04122	0.039787	0.02508	-0.27051	0.131327
Fireplaces	-0.21156	0.411976	0.046748	0.38683	0.076652	-0.09837
GarageCars	0.438657	-0.03557	0.04851	0.063414	-0.02084	-0.05065
GarageArea	0.464152	-0.09055	0.010659	0.027306	-0.11392	-0.08575
age	0.254493	0.072179	0.299835	-0.27844	-0.35782	-0.04393

The PC1 - PC12 are the new set of principal components that explains 95% of the variation that was previously explained by the 19 non-skewed numerical predictors.

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 New model coefficients:

Coefficients	estimate	std.error	statistic	p.value
(Intercept)	12.17731	0.042952	283.5111	0 *
MSSubClass150	0.029017	0.026124	1.110771	0.266885
MSSubClass160	0.008174	0.007756	1.053847	0.292161
MSSubClass180	0.001222	0.012147	0.100579	0.919901
MSSubClass190	-0.00559	0.023114	-0.2418	0.808974
MSSubClass20	-0.00827	0.011545	-0.71623	0.473984
MSSubClass30	-0.00372	0.012132	-0.30674	0.759097
MSSubClass40	0.01546	0.019969	0.774197	0.438964
MSSubClass45	-0.05291	0.019198	-2.75587	0.005941*
MSSubClass50	-0.01806	0.013769	-1.3119	0.1898
MSSubClass60	-0.01293	0.013054	-0.99067	0.322045
MSSubClass70	-0.01037	0.013403	-0.77407	0.439041
MSSubClass75	-0.01517	0.016597	-0.91378	0.361012
MSSubClass80	-0.03935	0.018954	-2.07628	0.038076*
MSSubClass85	-0.01527	0.01419	-1.07641	0.281955
MSSubClass90	-0.01599	0.012493	-1.2801	0.200751
MSZoningFV	-0.00056	0.009886	-0.05657	0.954895
MSZoningRH	-0.00268	0.010831	-0.24778	0.804347
MSZoningRL	-0.00282	0.008206	-0.34361	0.731195
MSZoningRM	-0.00499	0.00774	-0.64497	0.51907
StreetPave	0.01261	0.0112	1.125875	0.260439
AlleyNone	-0.00307	0.003507	-0.8761	0.381149
AlleyPave	-0.01147	0.005459	-2.1012	0.035827 *
LotShapeIR2	0.006055	0.003956	1.530701	0.126102
LotShapeIR3	0.007395	0.009017	0.820074	0.412334
LotShapeReg	0.001736	0.001443	1.203228	0.229121
LandContourHLS	-0.00612	0.004557	-1.3429	0.179553
LandContourLow	0.008229	0.006282	1.310007	0.190439
LandContourLvl	-0.00169	0.003421	-0.49425	0.621219
UtilitiesNone	0.015017	0.018397	0.816266	0.414507
LotConfigCulDSac	-0.00634	0.003157	-2.00705	0.044965 *
LotConfigFR2	0.004047	0.004023	1.005827	0.314698
LotConfigFR3	0.002758	0.007194	0.383386	0.7015
LotConfigInside	0.000561	0.001635	0.343358	0.731388
LandSlopeMod	-0.00458	0.003724	-1.23079	0.218638
LandSlopeSev	0.061169	0.016037	3.814247	0.000143*

Coefficients	estimate	std.error	statistic	p.value
NeighborhoodBlueste	-0.0071	0.011055	-0.64258	0.520619
NeighborhoodBrDale	-0.0104	0.010625	-0.97842	0.328059
NeighborhoodBrkSide	-0.0079	0.00855	-0.92453	0.355391
NeighborhoodClearCr	-0.02285	0.009478	-2.41072	0.016068*
NeighborhoodCollgCr	-0.01353	0.007049	-1.91885	0.055236
NeighborhoodCrawfor	-0.00786	0.007907	-0.99386	0.320489
NeighborhoodEdwards	-0.01286	0.007511	-1.71224	0.087106
NeighborhoodGilbert	-0.01298	0.007337	-1.76875	0.077186
NeighborhoodIDOTRR	-0.00388	0.008952	-0.43384	0.664482
NeighborhoodMeadowV	-0.01092	0.010544	-1.03562	0.300583
NeighborhoodMitchel	-0.01364	0.007618	-1.79024	0.073663
NeighborhoodNAmes	-0.00886	0.007422	-1.19359	0.232869
NeighborhoodNoRidge	-0.00993	0.008038	-1.23531	0.216953
NeighborhoodNPkVill	0.000695	0.017451	0.039835	0.968231
NeighborhoodNridgHt	-0.02171	0.007251	-2.99455	0.002804 *
NeighborhoodNWAmes	-0.01465	0.007713	-1.89891	0.057811
NeighborhoodOldTown	-0.00363	0.008409	-0.43121	0.666389
NeighborhoodSawyer	-0.01291	0.00766	-1.68483	0.092276
NeighborhoodSawyerW	-0.01531	0.00737	-2.07775	0.03794 *
NeighborhoodSomerst	-0.01306	0.008218	-1.58919	0.112276
NeighborhoodStoneBr	-0.00896	0.008266	-1.08362	0.278748
NeighborhoodSWISU	-0.00992	0.008996	-1.10293	0.270276
NeighborhoodTimber	-0.01569	0.007679	-2.04267	0.041299 *
NeighborhoodVeenker	-0.0278	0.009547	-2.91149	0.003662*
Condition1Feedr	-0.0004	0.004424	-0.08972	0.928523
Condition1Norm	-0.00365	0.00366	-0.99591	0.31949
Condition1PosA	-0.00125	0.008188	-0.15299	0.878428
Condition1PosN	0.004526	0.006277	0.720983	0.471058
Condition1RRAe	0.003036	0.006564	0.462412	0.643868
Condition1RRAn	-0.00398	0.006086	-0.65472	0.512772
Condition1RRNe	-0.00046	0.011986	-0.03848	0.96931
Condition1RRNn	-0.01478	0.011474	-1.28772	0.198086
Condition2Feedr	0.020438	0.015204	1.344195	0.179135
Condition2Norm	0.009648	0.012973	0.743697	0.457203
Condition2PosA	0.002953	0.019161	0.154097	0.877558
Condition2PosN	0.000698	0.020376	0.034271	0.972667
BldgType2fmCon	-0.01023	0.020316	-0.50378	0.61451
BldgTypeTwnhs	0.016772	0.012176	1.377426	0.168633
BldgTypeTwnhsE	0.005887	0.011355	0.518491	0.60421
HouseStyle1.5Unf	0.045692	0.01514	3.018004	0.002597 *
HouseStyle1Story	-0.01496	0.007582	-1.97293	0.048728 *
HouseStyle2.5Unf	0.0113	0.010802	1.046168	0.29569
HouseStyle2Story	-0.00231	0.007263	-0.3185	0.750158
HouseStyleSFoyer	-0.01027	0.009502	-1.08122	0.279813
HouseStyleSLvl	0.012811	0.015508	0.826068	0.408927
OverallCond2	-0.0329	0.01646	-1.99871	0.045861 *

Coefficients	estimate	std.error	statistic	p.value
OverallCond3	-0.00609	0.012444	-0.48928	0.624731
OverallCond4	-0.00374	0.012068	-0.31013	0.756512
OverallCond5	-0.00935	0.011928	-0.78409	0.433138
OverallCond6	-0.00772	0.012006	-0.64262	0.520594
OverallCond7	-0.00327	0.012021	-0.27184	0.785791
OverallCond8	-0.00137	0.012055	-0.11344	0.909697
OverallCond9	0.001604	0.012984	0.123518	0.901717
RoofStyleGable	-0.00056	0.014714	-0.03827	0.969482
RoofStyleGambrel	0.002631	0.016386	0.160562	0.872465
RoofStyleHip	0.001255	0.014857	0.0845	0.932673
RoofStyleMansard	0.014806	0.019119	0.774409	0.438839
RoofStyleShed	0.031092	0.02171	1.43219	0.152345
RoofMatlTarGrv	0.005951	0.010584	0.562235	0.574059
RoofMatlWdShake	-0.00322	0.011574	-0.27842	0.780738
RoofMatlWdShngl	-0.03599	0.032104	-1.12096	0.262525
Exterior1stAsphShn	-0.01218	0.023629	-0.51558	0.606242
Exterior1stBrkComm	-0.00433	0.018938	-0.22886	0.819018
Exterior1stBrkFace	-0.00815	0.013574	-0.60067	0.548171
Exterior1stCBlock	0.012342	0.026795	0.46059	0.645175
Exterior1stCemntBd	-0.05881	0.021751	-2.70389	0.006948 *
Exterior1stHdBoard	-0.01112	0.012763	-0.87128	0.383771
Exterior1stMetalSd	-0.00835	0.014145	-0.59005	0.555269
Exterior1stPlywood	-0.00929	0.012419	-0.74835	0.454393
Exterior1stSdng	-0.02766	0.02736	-1.01087	0.312279
Exterior1stStucco	-0.00409	0.014521	-0.28151	0.778366
Exterior1stVinylSd	-0.00127	0.014435	-0.08792	0.929955
Exterior1stWd Sdng	-0.0102	0.012582	-0.81041	0.417862
Exterior1stWdShing	-0.00932	0.013422	-0.69466	0.487402
Exterior2ndBrk Cmn	-0.01312	0.020616	-0.63622	0.524755
Exterior2ndBrkFace	0.008221	0.015032	0.54687	0.584568
Exterior2ndCBlock	-0.00561	0.023298	-0.2407	0.809828
Exterior2ndCmentBd	0.050857	0.022201	2.290778	0.022146 *
Exterior2ndHdBoard	0.001489	0.013753	0.108285	0.913787
Exterior2ndImStucc	0.005168	0.016625	0.310852	0.755966
Exterior2ndMetalSd	-0.00096	0.015016	-0.06393	0.949034
Exterior2ndPlywood	0.004173	0.01325	0.314974	0.752835
Exterior2ndStone	0.119386	0.054321	2.197774	0.028152*
Exterior2ndStucco	-0.00266	0.015353	-0.17319	0.862533
Exterior2ndVinylSd	-0.01112	0.015286	-0.72747	0.46708
Exterior2ndWd Sdng	0.001886	0.013622	0.138451	0.889907
Exterior2ndWd Shng	0.003226	0.014196	0.227248	0.820269
MasVnrTypeBrkFace	-0.00677	0.007147	-0.94694	0.343857
MasVnrTypeNone	-0.00707	0.007099	-0.99575	0.319568
MasVnrTypeStone	-0.01075	0.007388	-1.45573	0.145724

Coefficients	estimate	std.error	statistic	p.value
ExterQualFa	-0.00496	0.007805	-0.63504	0.525518
ExterQualGd	-0.00073	0.004151	-0.17515	0.860991
ExterQualTA	0.00181	0.00472	0.383444	0.701457
ExterCondFa	0.000819	0.00897	0.091336	0.92724
ExterCondGd	0.003369	0.007778	0.433149	0.664983
ExterCondPo	-0.00971	0.026222	-0.37046	0.711105
ExterCondTA	0.00509	0.007767	0.655311	0.512391
FoundationCBlock	0.000318	0.002757	0.115511	0.908059
FoundationPConc	-0.00207	0.002957	-0.70104	0.483409
FoundationSlab	-0.00624	0.00826	-0.75495	0.450422
FoundationStone	-0.01771	0.011263	-1.57254	0.116085
FoundationWood	0.007933	0.015613	0.508115	0.611464
BsmtQualFa	-0.00882	0.00505	-1.74679	0.080925
BsmtQualGd	-0.00288	0.002867	-1.00487	0.315156
BsmtQualNone	0.01833	0.021032	0.871524	0.383639
BsmtQualTA	-0.00313	0.003702	-0.84595	0.397744
BsmtCondGd	-0.00632	0.004484	-1.40929	0.159005
BsmtCondNone	-0.00334	0.013332	-0.25073	0.802066
BsmtCondPo	-0.00109	0.013018	-0.08365	0.933347
BsmtCondTA	-0.00584	0.003372	-1.7311	0.083686
BsmtExposureGd	-0.00092	0.002635	-0.35063	0.725927
BsmtExposureMn	-0.00416	0.002595	-1.60436	0.108893
BsmtExposureNo	-0.00428	0.002024	-2.11567	0.034575 *
BsmtExposureNone	-0.0167	0.014647	-1.13999	0.254512
BsmtFinType1BLQ	-0.0046	0.002592	-1.77324	0.076438
BsmtFinType1GLQ	-0.00137	0.002234	-0.61255	0.540289
BsmtFinType1LwQ	-0.00198	0.003109	-0.63549	0.525225
BsmtFinType1None	-0.00966	0.029741	-0.32485	0.745348
BsmtFinType1Rec	-0.00013	0.002547	-0.05189	0.958621
BsmtFinType1Unf	0.000723	0.002559	0.282552	0.777568
BsmtFinType2BLQ	0.003868	0.005306	0.72906	0.466104
BsmtFinType2GLQ	0.000485	0.006565	0.073891	0.941109
BsmtFinType2LwQ	-0.00633	0.005238	-1.20934	0.226767
BsmtFinType2Rec	0.002159	0.004978	0.433688	0.664591
BsmtFinType2Unf	0.002212	0.00405	0.546124	0.585081
HeatingGasW	0.001278	0.00862	0.148295	0.882134
HeatingGrav	-0.0223	0.017575	-1.2687	0.204788
HeatingWall	-0.06008	0.04392	-1.36791	0.171592
HeatingQCFa	-0.00276	0.003977	-0.69394	0.487853
HeatingQCGd	0.001053	0.001814	0.580283	0.561831
HeatingQCPo	0.019113	0.022183	0.861568	0.389094
HeatingQCTA	0.000266	0.001797	0.147842	0.882491
CentralAirY	-0.0013	0.00319	-0.4079	0.683419
ElectricalFuseF	-0.01536	0.005714	-2.6876	0.007294 *
ElectricalFuseP	-0.01236	0.010863	-1.13769	0.255474
ElectricalSBrkr	-0.00058	0.002617	-0.22232	0.8241
KitchenQualFa	0.010322	0.005726	1.802798	0.071666
KitchenQualGd	-0.00043	0.003357	-0.12713	0.89886
KitchenQualTA	0.000143	0.003702	0.038517	0.969282

Coefficients	estimate	std.error	statistic	p.value
FunctionalMaj2	0.00634	0.015978	0.396815	0.691573
FunctionalMin1	0.006308	0.011696	0.539355	0.58974
FunctionalMin2	0.007915	0.01181	0.670194	0.502861
FunctionalMod	0.00872	0.012142	0.718152	0.472801
FunctionalSev	0.023061	0.025886	0.890863	0.373178
FunctionalTyp	0.002351	0.011036	0.213072	0.831306
FireplaceQuFa	-0.00227	0.006279	-0.36111	0.718077
FireplaceQuGd	-0.00299	0.005076	-0.58834	0.556414
FireplaceQuNone	0.001532	0.00572	0.267775	0.788918
FireplaceQuPo	-0.00359	0.006705	-0.53515	0.592646
FireplaceQuTA	-0.00069	0.005208	-0.13301	0.894207
GarageTypeAttchd	-0.00842	0.005784	-1.45552	0.145783
GarageTypeBasment	-0.00967	0.008317	-1.163	0.245056
GarageTypeBuiltIn	-0.00754	0.0063	-1.19625	0.231832
GarageTypeCarPort	-0.00216	0.010547	-0.20453	0.837977
GarageTypeDetchd	-0.00747	0.005796	-1.28928	0.197544
GarageTypeNone	-0.01331	0.020188	-0.65943	0.509743
GarageFinishNone	-0.02353	0.02983	-0.78864	0.430477
GarageFinishRFn	0.000686	0.001729	0.396818	0.691571
GarageFinishUnf	0.003713	0.002036	1.824111	0.068379
GarageQualGd	-0.00148	0.008881	-0.16677	0.867578
GarageQualPo	-0.03721	0.019121	-1.94591	0.051894
GarageQualTA	2.46E-05	0.003188	0.007719	0.993842
GarageCondFa	-0.03075	0.022976	-1.33827	0.181058
GarageCondGd	-0.02459	0.023383	-1.05144	0.293265
GarageCondPo	-0.01945	0.024754	-0.78588	0.432092
GarageCondTA	-0.03388	0.022581	-1.50055	0.13373
PavedDriveP	-0.00132	0.004649	-0.28319	0.777077
PavedDriveY	-0.00519	0.003054	-1.70107	0.089183
FenceGdWo	-0.00864	0.004021	-2.14931	0.031806 *
FenceMnPrv	-0.0031	0.003277	-0.94718	0.343737
FenceMnWw	-0.00573	0.021097	-0.27153	0.786031
FenceNone	-0.00217	0.002934	-0.73923	0.459907
MiscFeatureNone	0.003681	0.011957	0.307891	0.758218
MiscFeatureOthr	0.009242	0.019723	0.468596	0.639442
MiscFeatureShed	0.002891	0.012384	0.233488	0.815422
SaleTypeCon	-0.01327	0.013457	-0.98625	0.324208
SaleTypeConLD	-0.00233	0.006705	-0.34749	0.728283
SaleTypeConLI	0.016263	0.011654	1.395523	0.163111
SaleTypeConLw	0.004675	0.012951	0.360986	0.718172
SaleTypeCWD	0.016654	0.008297	2.007314	0.044936*
SaleTypeNew	0.001903	0.014127	0.134693	0.892876
SaleTypeOth	-0.01095	0.011051	-0.99053	0.322111
SaleTypeWD	0.004505	0.003648	1.23487	0.217116

Coefficients	estimate	std.error	statistic	p.value
SaleConditionAdjLand	0.01185	0.009076	1.305724	0.191892
SaleConditionAlloca	0.001441	0.00777	0.185406	0.852941
SaleConditionFamily	0.005843	0.005054	1.156167	0.247839
SaleConditionNormal	-0.00162	0.00274	-0.5919	0.554029
SaleConditionPartial	0.012104	0.013721	0.882098	0.377897
PC1	-0.02299	0.000716	-32.1156	1.30E-164 *
PC2	0.016108	0.000916	17.57758	7.67E-62*
PC3	0.031282	0.000924	33.86965	6.30E-178 *
PC4	0.021434	0.000917	23.38452	2.90E-100*
PC5	0.005626	0.000818	6.873562	9.96E-12 *
PC6	0.025964	0.001227	21.15942	6.29E-85 *
PC7	-0.00514	0.000999	-5.14313	3.14E-07*
PC8	-0.00513	0.00102	-5.03238	5.57E-07*
PC9	-0.02102	0.001147	-18.3307	1.67E-66*
PC10	0.018707	0.001259	14.85543	5.11E-46*
PC11	0.017167	0.001243	13.80626	2.11E-40*
PC12	-0.05528	0.00122	-45.2968	7.70E-264*

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 = 0.9588 ; Residual Standard error = 0.01995

Interpretation:

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

Definitely, the exclusion of those highly skewed predictors improves the model such that the model explains the total variation in a better way.

7.5 Residual Diagnostics of the New Model

1. Residual Plot:

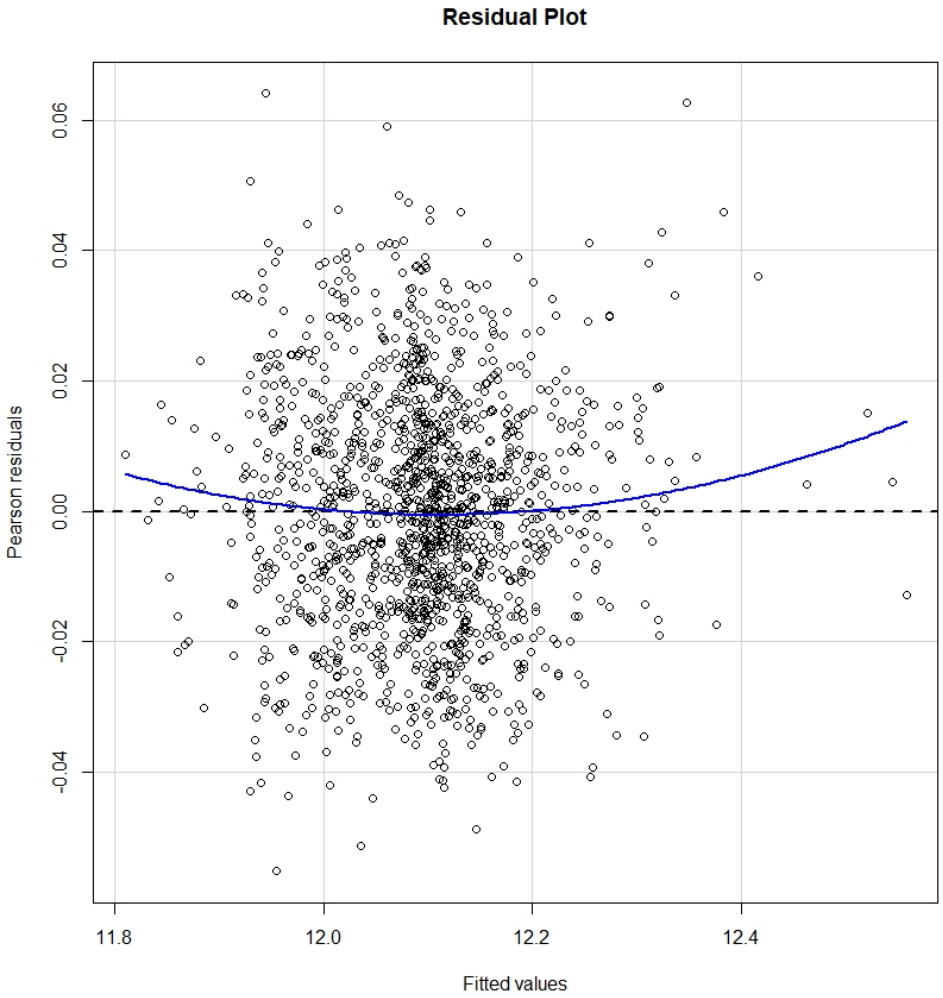


Figure 14: Residual Plot

From the above residual plot, we see that the residuals are randomly scattered around 0 line. This indicates that the error distribution is homoskedastic which satisfies the linear regression assumption that is 'The errors are homoskedastic in

nature'.

2. Q-Q Plot:

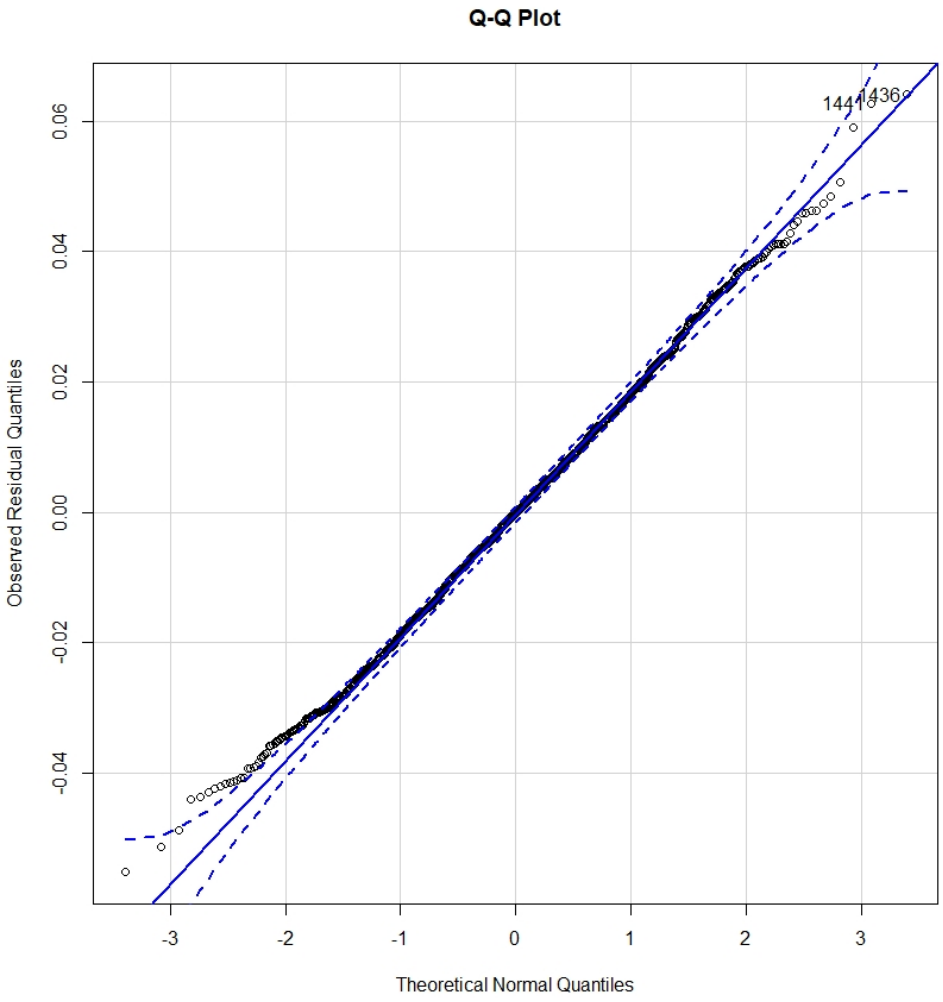


Figure 15: Q-Q Plot of Residuals

From the above q-q plot, we observe that the sample quantiles of residuals almost coincides with theoretical normal quantiles. This implies that the errors are

more likely to be normally distributed.

3.Shapiro-Wilk Normality Test:

Here, the testing hypothesis is:

H_0 : The error distribution is normally distributed

H_1 : The error distribution is not normally distributed

Testing Results:

Observed Test Statistic = 0.99804 ; P-Value = 0.0807

Decision:

We accept the Null Hypothesis at 0.05 level of significance.

Conclusion:

In the light of the given data, we may conclude that The error distribution is Normal under the chosen model.

The New Model does not violate the 'Normality assumption of Error Distribution of Linear Regression Model'.

4. Durbin-Watson Test:

Here, the testing hypothesis is:

H_0 : The errors are correlated

H_1 : The error are not correlated

Testing Results:

Observed Test Statistic = 1.7082 ; P-Value = 0.479

Decision:

We accept the Null Hypothesis at 0.05 level of significance.

Conclusion:

In the light of the given data, we may conclude that The errors are not correlated under the chosen model.

The New Model does not violate the assumption 'Errors are Independently distributed in Linear Regression Model'.

Therefore, this modified linear regression model satisfies all the basic assumptions of Linear Regression Model.

Table 5: Summary Table of The residuals

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
-0.05520	-0.01314	0	0	0.01235	0.06407

From this new model, we obtain the predicted value of 'SalePrice', say $\hat{y}$, where,

$$\hat{y} = e^{\beta_0 + \Sigma_i \beta_i x i}$$

Given below the histograms of Actual House Price(y) and Predicted House Price($\hat{y}$):

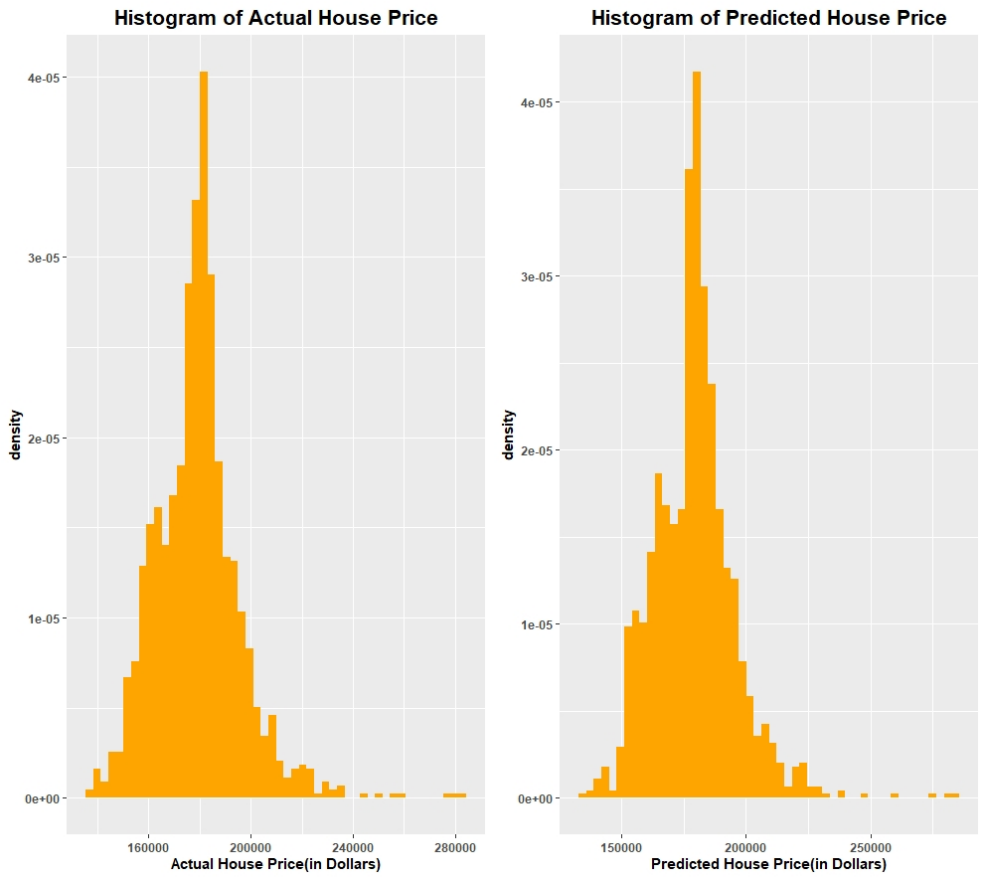


Figure 16: Histograms of Actual House Price (left) and Predicted House Price (right)

The histograms of Actual House Price and Predicted House Price seems almost identical which implies that the predicted values are likely to be close to the actual value in most of the cases.

Comparison of summary table of Actual House Price and Predicted House Price:

Table 6: Summary Table of Actual House Price and Predicted House Price

Measure	Actual House Price(in Dollars)	Predicted House Price(in Dollars)
Minimum	135751	134573
1st Quantile	168703	168410
Median	179209	179232
Mean	179184	179153
3rd Quantile	186789	186787
Maximum	281644	284254

From Table 6 and from Figure 15, we may say that based on the given data, the new fitted model predicts House Price with better accuracy with lesser number of predictors than the initial model. Hence, in order to predict House Price in Ames region, one may choose the new model for better prediction purposes.

8 Impact of Significant Predicting Features

Here, the model is log-linear where the target variable is the natural log of "SalePrice". To compute the impact of a unit change in each variable on the "SalePrice", we would first need the exponent of each coefficient. Moreover, the coefficients of a log-linear model imply the percentage change in the target variable for small unit changes in the explanatory variable. To get the specific dollar impact, it is best to use the mean "SalePrice" as the reference comparison, i.e.,

We compute, $\hat{\bar{y}} = \bar{y}\exp(\beta_i)$

where,

$\hat{\bar{y}}$: mean predicted House Price(in Dollars) for 1 unit change in ith predictor when other predictors remain fixed.

$\bar{y}$: mean House Price (in Dollars)

β_i : Coefficient of ith predictor

Given below the table showing the dollar impact in 1 unit change in the significant predicting features:

Significant predictors	impact in 1 unit change(in dollars)
MSSubClass45	-9233.68
MSSubClass80	-6914.59
AlleyPave	-2043.63
LotConfigCulDSac	-1131.63
LandSlopeSev	11302.68
NeighborhoodClearCr	-4047.52
NeighborhoodNridgHt	-3848.68
NeighborhoodSawyerW	-2722.87
NeighborhoodTimber	-2788.75
NeighborhoodVeenker	-4912.2
HouseStyle1.5Unf	8377.167
HouseStyle1Story	-2660.51
OverallCond2	-5799.04
Exterior1stCemntBd	-10234.4
Exterior2ndCmentBd	9348.442
Exterior2ndStone	22721.32
BsmtExposureNo	-765.809
ElectricalFuseF	-2730.68
FenceGdWo	-1541.75
SaleTypeCWD	3009.139
PC1	-4072.74
PC2	2909.594
PC3	5693.823
PC4	3882.079
PC5	1010.846
PC6	4713.326
PC7	-918.543
PC8	-917.594
PC9	-3726.96
PC10	3383.563
PC11	3102.671
PC12	-9637

The **dollar impact of a one-unit change in each explanatory variable on the average house price** in Ames is listed in the table on the left.

Effect of the significant housing parameters:

1. A “Paved Alley” will reduce the average house price by $2043, all else being fixed.

2. A ”1-1/2 STORY - UNFINISHED ALL AGES” will reduce the average house price by $9233, all else being fixed.

3. A ”SPLIT OR MULTI-LEVEL” will reduce the average house price by $6914, all else being fixed.

4. A ”CulDsac Lot configuration” will reduce the average house price by $1131, all else being fixed.

5. A ”ClearCr Neighbourhood” will reduce the average house price by $4047, all else being fixed.

6. A ”NridgHt Neighbourhood” will reduce the average house price by $3848, all else being fixed.

7. A ”SaywerW Neighbourhood” will reduce the average house price by $2722, all else being fixed.

8. A ”Timber Neighbourhood” will reduce the average house price by $2788, all else being fixed.

9. A ”Veenker Neighbourhood” will reduce the average house price by $4912, all else being fixed.

10. A ”One and one-half story: 2nd level unfinished House” will rise the average house price by $8733, all else being fixed.

11. A ”One storey House” will reduce the average house price by $2600, all else being fixed.

12. A ”Poor Conditioned House” will reduce the average house price by $5799, all else being fixed.

13. A ”Cement Board Exterior covering” will bump the average house price by $9348, all else being fixed.

14. A ”Stone Exterior Covering” will rise the average house price by $22721, all else being fixed.

15. A ”Basement with no garden Level walls” will reduce the average house price by $765, all else being fixed.

16. A ”60 AMP Fuse Box and mostly Romex wiring” will reduce the average house price by $2730, all else being fixed.

17. A ”Wood Fence” will reduce the average house price by $1541, all else being fixed.

18. A ”Warranty Deed - Cash Sale” will rise the average house price by $3009, all else being fixed.

19. A ”Severely sloped House property” will rise the average house price by $11302, all else being fixed.

Here, we note that the 12 significant principal components are made of the linear combination of 19 different numerical predicting features. Hence, the individuality of the original predictors is lost in the principal component. Hence, it is very complicated to interpret how the price changes in one unit change of the principal components, in terms of the original variables.

9 Acknowledgement

I would like to express my gratitude to the Rector and Principal of my college, Rev. Dr Dominic Savio, S.J. for giving me the opportunity to work on this project.
I would also like to thank the Vice Principal of Arts and Science department, Prof Betram Da'Silva and The dean of Science Dr Tapati Dutta.
I am also very grateful to the Head of the Department of Statistics and also my project supervisor Dr. Durba Bhattacharya for guiding me with the guidelines of the project. She constantly supported me and guided me with her valuable advice whenever I was stuck with my project. She took notice of the problems I faced, all the concepts I did not have. She also provided me with necessary materials and references that helped me complete this project and also helped me gaining knowledge about different aspects of theoretical statistics and its application.
I would like to thank my family and my friends as well. They kept me motivated to work constantly on my project. They also helped me understanding some of the concepts and helped me writing some codes. My work was made easier by my family and my friends.

10 Appendix

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

```
rm(list=ls())
library(ggplot2)
library(MASS)
library(reshape2)
library(tidyverse)
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)
par(mfrow=c(1,1))
data_new = readxl::read_xlsx('C:/Users/Saheli/Desktop/
```

```
   dissertation/project data.xlsx')
attach(data_new)
nrow(data_new)

#analysis of the variables
#-----------------------------------------
summary(SalePrice)
ggplot(data=data_new,aes(x=SalePrice,y=..density..))+
  geom_histogram(bins=50,fill='orange')+
  labs(title = 'Histogram of Sale Price of
       House',x='Sale Price of House')+
  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'))

qqPlot(data_new$SalePrice, main =
         'qqplot of saleprice',xlab='Theoretical
       normal constants',
       ylab='Sample quantiles of Sale Price')
```

```
47
48 #Analysis of Predictors
49 #-----------------------------------------------
50 #1.GrlivArea
51 ggplot(data=data_new,aes(x=GrLivArea,y=..density..))+
52   geom_histogram(bins=50,fill='orange')+
53   labs(title = 'Histogram of General
54        Living Area ("GrLivArea") of House',
55        x='GrLivArea(in sqft)')+
56   theme(plot.title =
57           element_text(size=
58                          16,
59                        hjust=.5,face='bold'),
60         plot.subtitle = element_text(
61           size=14,hjust=.5,face='italic'
62         ),
63         legend.title = element_text(hjust=.5),
64         axis.title = element_text(face='bold'),
65         axis.text=element_text(face='bold'))
66 #2. LotArea
67 ggplot(data=data_new,aes(x=LotArea,y=..density..))+
68   geom_histogram(bins=50,fill='orange')+
69   labs(title = 'Histogram of Area of the
70        Property ("LotArea") of House',
71        x='LotArea(in sqft)')+
72   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'))
#3. BedroomAbvGr
ggplot(data=data_new,aes(y=BedroomAbvGr))+
  geom_boxplot(fill='orange')+
  labs(title = 'Boxplot of Number of bedrooms
       ("BedroomAbvGr") of House',
       x='BedroomAbvGr',y='Number of Rooms')+
  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'))

```

```
#missing value percentage
#---------------------------------------
group=missval=array(0)
for(i in 1:ncol(data_new))
{
  count = 0
  for(j in 1:nrow(data_new))
  {
  if(data_new[j,i]=='NA'| is.na(data_new[,i]
                                )[j]== 'TRUE')
  {
    count = count+1
  }
  }
  group[i] = colnames(data_new)[i]
  missval[i] = count*100/nrow(data_new)
}
m= data.frame(group1=factor(group,levels = group),
   missval)
miss.val.ratio = ( m %>% arrange(m$missval))


ggplot(miss.val.ratio[-(1:60),],aes(x=group1,y=missval))
   +
  geom_col(fill='black')+
  labs(title = 'Missing Value Percentage',
```

```
123         x='Housing Parameters',
124         y='Percentage of missing values')+
125   theme(plot.title =
126           element_text(size=
127                          16,
128                        hjust=.5,face='bold'),
129         plot.subtitle = element_text(
130           size=14,hjust=.5,face='italic'
131         ),
132         legend.title = element_text(hjust=.5),
133         axis.title = element_text(face='bold'),
134         axis.text=element_text(face='bold'))
135 #relationships between predictor and response
136 #---------------------------------------
137 ggplot(data=NULL,aes(x=GrLivArea,y=SalePrice))+
138   geom_point()+
139   labs(title = 'Scatterplot of Sale Price of House
140         vs General Living Area (in sqft)',x='General
141         Living Area (in sqft)',y='Sale Price
142         of House(in dollars)')+
143   theme(plot.title =
144           element_text(size=
145                          16,
146                        hjust=.5,face='bold'),
147         plot.subtitle = element_text(
148           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=NULL,aes(x=LotArea,y=SalePrice))+
  geom_point()+
  labs(title = 'Scatterplot of Sale Price of House
       vs Area of the property(in sqft)',x=
       'General Living Area
       (in sqft)',y='Sale Price of House(in
       dollars)')+
  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'))

ggplot(data=NULL,aes(x=as.factor(BedroomAbvGr),y=
   SalePrice
                     ,fill=BedroomAbvGr))+
  geom_boxplot()+
```

```
174   labs(title = 'Boxplot of Sale Price of House
175        with respect to number of bedrooms'
176        ,x='Number of bedrooms',
177        y='Sale Price of House(in dollars)')+
178   theme(plot.title =
179           element_text(size=
180                          16,
181                        hjust=.5,face='bold'),
182         plot.subtitle = element_text(
183           size=14,hjust=.5,face='italic'
184         ),
185         legend.title = element_text(hjust=.5),
186         axis.title = element_text(face='bold'),
187         axis.text=element_text(face='bold'))
188
189 #Imputing Missing Val
190 #---------------------------------------
191 data_new$PoolQC[data_new$PoolQC=='NA']= 'None'
192 data_new$MiscFeature[data_new$MiscFeature=='NA']= 'None'
193 data_new$Alley[data_new$Alley=='NA']= 'None'
194 data_new$Fence[data_new$Fence=='NA']= 'None'
195 data_new$FireplaceQu[data_new$FireplaceQu=='NA']= 'None'
196 data_new$LotFrontage[data_new$LotFrontage=='NA']= 'None'
197
198 data_new$LotFrontage= as.numeric(data_new$LotFrontage)
199 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]
  }}}

data_new$LotFrontage[c] = d #missing values replaced by
   median

data_new$GarageCond[data_new$GarageCond=='NA']= 'None'
data_new$GarageQual[data_new$GarageQual=='NA']= 'None'
data_new$GarageType[data_new$GarageType=='NA']= 'None'
data_new$GarageArea[is.na(data_new$GarageArea)== TRUE]=
   0
data_new$GarageFinish[data_new$GarageFinish=='NA']= '
   None'

data_new$GarageYrBlt = as.numeric(data_new$GarageYrBlt)
data_new$GarageYrBlt[is.na(data_new$GarageYrBlt)== TRUE
   ]= 0
data_new$GarageCars = as.numeric(data_new$GarageCars)
```

```
222 data_new$GarageCars[is.na(data_new$GarageCars)== TRUE]=
      0
223 data_new$BsmtFinSF1 = as.numeric(data_new$BsmtFinSF1)
224 data_new$BsmtFinSF1[is.na(data_new$BsmtFinSF1)== TRUE]=
      0
225 data_new$BsmtFinSF2 = as.numeric(data_new$BsmtFinSF2)
226 data_new$BsmtFinSF2[is.na(data_new$BsmtFinSF2)== TRUE]=
      0
227 data_new$BsmtUnfSF = as.numeric(data_new$BsmtUnfSF)
228 data_new$BsmtUnfSF[is.na(data_new$BsmtUnfSF)== TRUE]= 0
229 data_new$TotalBsmtSF = as.numeric(data_new$TotalBsmtSF)
230 data_new$TotalBsmtSF[is.na(data_new$TotalBsmtSF)== TRUE
      ]= 0
231 data_new$BsmtFullBath = as.numeric(data_new$BsmtFullBath
      )
232 data_new$BsmtFullBath[is.na(data_new$BsmtFullBath)==
      TRUE]= 0
233 data_new$BsmtHalfBath = as.numeric(data_new$BsmtHalfBath
      )
234 data_new$BsmtHalfBath[is.na(data_new$BsmtHalfBath)==
      TRUE]= 0
235
236 data_new$BsmtQual[data_new$BsmtQual=='NA']= 'None'
237 data_new$BsmtCond[data_new$BsmtCond=='NA']= 'None'
238 data_new$BsmtExposure[data_new$BsmtExposure=='NA']= '
      None'
```

```
239 data_new$BsmtFinType1[data_new$BsmtFinType1=='NA']= '
    None'
240 data_new$BsmtFinType2[data_new$BsmtFinType2=='NA']= '
    None'
241 data_new$YearBuilt=as.factor(data_new$YearBuilt)
242
243
244 data_new$MasVnrType[data_new$MasVnrType=='NA']= 'None'
245 data_new$MasVnrArea = as.numeric(data_new$MasVnrArea)
246 data_new$MasVnrArea[is.na(data_new$MasVnrArea)== TRUE]=
    0
247
248 data_new$MSZoning[data_new$MSZoning=='NA']= 'RL'
249 data_new$Utilities[data_new$Utilities=='NA']= 'None'
250 data_new$Functional[data_new$Functional=='NA']= 'Typ'
251 data_new$KitchenQual[data_new$KitchenQual=='NA']= 'TA'
252 data_new$Exterior1st[data_new$Exterior1st=='NA']= 'Sdng'
253 data_new$Exterior2nd[data_new$Exterior2nd=='NA']= '
    VinylSd'
254 data_new$SaleType[data_new$SaleType=='NA']= 'WD'
255 data_new$MSSubClass[data_new$MSSubClass=='NA']= 'None'
256
257 any(is.na(data_new))  #no missing values
258
259 #feature engineering
260 #----------------------------------------------
```

```
261 data_new$MSSubClass = as.factor(data_new$MSSubClass)
262 data_new$OverallCond = as.factor(data_new$OverallCond)
263 data_new$OverallQual = as.factor(data_new$OverallQual)
264 data_new$YrSold = as.factor(data_new$YrSold)
265 data_new$MoSold = as.factor(data_new$MoSold)
266
267 summary(data_new)
268
269 #correlation heatmap
270 #------------------------------------------
271 data.na.omit = na.omit(data_new[,-80])
272 corr = data.matrix(cor(data.na.omit[sapply(data.na.omit,
273                                            is.numeric)])
                                               )
274 mel = melt(corr)
275 ggplot(mel, aes(Var1,Var2))+geom_tile(aes(fill=value)) +
276   geom_text(aes(label = round(value, 1)))+
277   scale_fill_gradient2(low='blue',mid = 'White' ,high='
        red')
278 + labs(title = 'Correlation Heatmap')
279
280 #skewed freatures
281 #------------------------------------------
282 sk =name =array(0)
283 for(i in 1:ncol(data_new[sapply(data_new,is.numeric)]))
284 {
```

```
285
286
287   sk[i] = skewness(data_new[sapply(data_new,is.numeric)
          ][,i])
288   name[i] = colnames(data_new[sapply(data_new,is.numeric
          )])[i]
289 }
290  t = data.frame(name, sk)
291  t1 = t %>% arrange(t$sk)
292
293 writexl::write_xlsx(t1,'C:/Users/Saheli/
294                     Desktop/dissertation/
295                     Latex Dissertation/skewness.xlsx')
296
297 #Partial Correlation heatmap
298 #------------------------------------------------
299 partial.cor_new = corpcor::cor2pcor(cov(
300   data_new[,which(sapply(data_new[,-c(80,81)],is.numeric
          ))]))
301
302
303 colnames(partial.cor_new)=colnames(
304   data_new[,which(sapply(data_new[,-c(80,81)],is.numeric
          ))])
305 rownames(partial.cor_new)=colnames(
306   data_new[,which(sapply(data_new[,-c(80,81)],is.numeric
```

```
    ))])
mel.partial_new = melt(data.matrix(partial.cor_new))
ggplot(mel.partial_new, 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 = 'Partial Correlation Heatmap')

#transforming skewed features
#-------------------------------------------------
data_new$GarageArea = log(max(data_new$GarageArea)
                          +data_new$GarageArea )
skewness(data_new$GarageArea)
data_new$LotFrontage =sqrt(data_new$LotFrontage)
skewness(data_new$LotFrontage)
data_new$FullBath =log(max(FullBath)+data_new$FullBath)
skewness(data_new$FullBath)
data_new$BedroomAbvGr =log(max(data_new$BedroomAbvGr)
                           +data_new$BedroomAbvGr)
skewness(data_new$BedroomAbvGr)
data_new$LotArea=sqrt(data_new$LotArea)
skewness(data_new$LotArea)
data_new$GrLivArea=log(data_new$GrLivArea)
skewness(data_new$GrLivArea)
data_new$YearRemodAdd=log(data_new$YearRemodAdd)
```

```
skewness(data_new$YearRemodAdd)
data_new$SalePrice=log(data_new$SalePrice)
skewness(data_new$SalePrice)
skewness(SalePrice)
skewness(sqrt(data_new$WoodDeckSF))
skewness(sqrt(data_new$MasVnrArea))
skewness(sqrt(data_new$ScreenPorch))
skewness(sqrt(data_new$LowQualFinSF))
skewness(sqrt(data_new$MiscVal))
skewness(sqrt(data_new$PoolArea))
skewness(sqrt(data_new$OpenPorchSF))
skewness(sqrt(data_new$EnclosedPorch))

#PCA of whole data
data.pca = prcomp(data_new[,-c(80,81)][ , unlist(
  lapply(data_new[,-c(80,81)], is.numeric)) ] ,
  center = TRUE,scale. = TRUE)
summary(data.pca)
trans.data1 <- preProcess(data_new[,-c(80,81)], method
   = "pca")
PC = as.data.frame(trans.data1$rotation)

writexl::write_xlsx(PC,'C:/Users/Saheli/Desktop /
   dissertation/Latex Dissertation/PCwholedata.xlsx')
transformedData1 <- predict(trans.data1,data_new[,-c
   (80,81)] )
```

```
nrow(transformedData1)
colnames(transformedData1)
factoextra::fviz_eig(data.pca,ncp= 22,
                     xlab='Principal Components')
age.pca = as.numeric(data_new$YrSold)-
  as.numeric(data_new$YearBuilt)
indep.pca = cbind(select(transformedData1,
                         -one_of(

                   'YearBuilt','YrSold')),age.pca)
fit.pca1 = lm(data_new$SalePrice~.,indep.pca)
options(max.print = 2000)
summary(fit.pca1)
fit.pca1.data = broom::tidy(fit.pca1)
writexl::write_xlsx(fit.pca1.data,'C:/Users/Saheli/
   Desktop/dissertation/Latex Dissertation/summarypca1.
   xlsx')

residualPlot(fit.pca1, main = 'Residual Plot')
qqPlot(resid(fit.pca1), main = 'Q-Q Plot',
       xlab = 'Theoretical Normal Quantiles',
       ylab = 'Observed Residual Quantiles')
hist(resid(fit.pca1),freq=F)

#correlation between errors
DurbinWatsonTest(fit.pca1,alternative='two.sided')
```

```
378
379 #checking for normality- Shapiro-Wilk test, qqplot
380 shapiro.test(resid(fit.pca1))
381
382 qqPlot(resid(fit.pca1))
383
384 #PCA excluding skewed features
385 dd.pca.out = subset(data_new,
386                     select = -c(
387                     `3SsnPorch`,LowQualFinSF,MiscVal,
388                     PoolArea, PoolQC,
389                     KitchenAbvGr,EnclosedPorch,
                        OpenPorchSF,
390                     ScreenPorch,BsmtHalfBath,WoodDeckSF,
391                     OverallQual,MoSold,
392                     BsmtFinSF2,MasVnrArea,GarageYrBlt,id
                        ,
393                    YearBuilt,  SalePrice,YrSold))
394 age.pca.out = as.numeric(data_new$YrSold)-
395   as.numeric(data_new$YearBuilt)
396 dd.pca1.out = cbind(dd.pca.out,age.pca.out)
397 prcomp(dd.pca1.out[ , unlist(lapply(
398   dd.pca1.out, is.numeric)) ] ,
399   center = TRUE,scale. = TRUE)
400 trans.data.out <- preProcess(dd.pca1.out[ ,
401     unlist(lapply(dd.pca1.out, is.numeric)) ] ,
```

```
    method  = "pca")
transformedData.out <- predict(trans.data.out, dd.pca1.
   out)
trans.data.out$rotation
PC.out= as.data.frame(trans.data.out$rotation)
rownames(PC.out)
writexl::write_xlsx(PC.out,'C:/Users/Saheli/
                    Desktop/dissertation/Latex
                       Dissertation
                    /PC.outwholedata.xlsx')
fit.pca.out = lm(data_new$SalePrice~.,
                 data=transformedData.out)
summary(fit.pca.out)
fit.pca.out.data = broom::tidy(fit.pca.out)
writexl::write_xlsx(fit.pca.out.data,'C:/
                    Users/Saheli/Desktop/
                    dissertation/Latex Dissertation
                    /summarypca2.xlsx')

residualPlot(fit.pca.out, main ='Residual Plot')
residualPlots(fit.pca.out)

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

```
DurbinWatsonTest(fit.pca.out,alternative = 'two.sided')
summary(resid(fit.pca.out))
resi.pca.out =  SalePrice - exp(predict(fit.pca.out))
data.frame(Actual_Price= SalePrice,
           Fitted_Price = exp(predict(fit.pca.out)))
summary(resi.pca.out)
summary(exp(predict(fit.pca.out)))
A = data.frame(id,actual_price =SalePrice,
               Fitted_price=exp(predict(fit.pca.out)))

p1 = ggplot(data=NULL,aes(x=SalePrice,y=..density..))+
  geom_histogram(bins=50,fill='orange')+
  labs(title = 'Histogram of Actual House Price',
       x='Actual House Price(in Dollars)')+
  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'))
p2 = ggplot(data=NULL,aes(x=exp(predict(
  fit.pca.out)),y=..density..))+
```

```
452   geom_histogram(bins=50,fill='orange')+
453   labs(title = 'Histogram of Predicted House
454        Price',x='Predicted House Price(in Dollars)')+
455   theme(plot.title =
456           element_text(size=
457                          16,
458                        hjust=.5,face='bold'),
459         plot.subtitle = element_text(
460           size=14,hjust=.5,face='italic'
461         ),
462         legend.title = element_text(hjust=.5),
463         axis.title = element_text(face='bold'),
464         axis.text=element_text(face='bold'))
465
466 grid.arrange(p1,p2,ncol=2)
467
468 coeff.pval = data.frame(summary(fit.pca.out)$
      coefficients[,c(1,4)] )
469 coeff.sig = data.frame(signi.pred=rownames(coeff.pval[
      which(coeff.pval[,2]<0.05),]),oneunitchange=mean(
      SalePrice)*exp(coeff.pval[which(coeff.pval[,2]<0.05)
      ,][,1])-mean(SalePrice))
470
471 writexl::write_xlsx(coeff.sig,'C:/Users/Saheli/Desktop/
      dissertation/Latex Dissertation/significantpreds.xlsx
      ')
```