Logo

Description automatically generated

**College of Professional Studies**

**Northeastern University San Jose**

**MPS Analytics**

**Course: ALY6015: Intermediate Analytics**

**Assignment:**

Module 1 Assignment - Regression Diagnostics with R

**Submitted on:**

January 15, 2023

**Submitted to:**  **Submitted by:**

Professor: PAROMITA GUHA NIKSHITA RANGANATHAN

# **INTRODUCTION**

A statistical method called regression analysis is used to understand how the dependent variable changes when one or more of the independent variables are related. It is also used to predict the value of the dependent variable when the values of the independent variables are known. Regression analysis is used in many fields, including economics, finance, marketing, psychology, sociology, and engineering.

Applications for regression analysis include forecasting stock prices, predicting sales, and studying consumer behavior. It is also used to identify trends in data and assess the impact of certain factors on outcomes. Regression analysis is an effective data analytics approach that may be applied to provide conclusions.

**Ordinary Least Squares:**

The parameters of a linear regression model are estimated using a form of linear regression approach known as ordinary least squares (OLS) estimators. In order to estimate the model's parameters, OLS estimators employ the least squares approach to reduce the sum of squared residuals, or the difference between the predicted and actual values.

There are some conditions that must be met in order for the OLS method to produce reliable and valid results. These assumptions include linearity, homoscedasticity, normality, no multicollinearity, and independence of errors.

**About the dataset:**

# The title of the dataset is “Ames Housing”.

The data set includes details from the Ames Assessor's Office used to determine assessed values for specific residential properties sold in Ames, Iowa, between 2006 and 2010.

The Ames Housing dataset contains 82 explanatory variables and 2930 properties that describe the various aspects of residential homes in Ames, Iowa.

**Purpose:**

The Ames Housing dataset (Kaggle) has been chosen to analyze the houses located in Ames, Iowa to find out which variables of a house contribute most to its sale price. To predict housing prices in Ames, Iowa, our focus is on utilizing multiple housing price indicators, including factors related to the size and location of the living spaces.

The dataset includes the following features:

|  |  |  |
| --- | --- | --- |
| **No** | **Feature** | **Dictionary** |
| 1 | Order | Number of observations |
| 2 | PID | Parcel identification number, which can be used on the city website for parcel review |
| 3 | MS SubClass | Specifies the category of home that is being sold |
| 4 | MS Zoning | Determines the sale's broad zoning classification |
| 5 | Lot Frontage | Linear feet of street connected to the property |
| 6 | Lot Area | Lot size (square feet) |
| 7 | Street | Road access type to the property |
| 8 | Alley | Type of property's alley access |
| 9 | Lot Shape | The property's general layout |
| 10 | Land Contour | The building's flatness |
| 11 | Utilities | utility options |
| 12 | Lot Config | Lot arrangement |
| 13 | Land Slope | Property's slope |
| 14 | Neighborhood | Physical places within the boundaries of Ames |
| 15 | Condition 1 | Proximity to variety of conditions |
| 16 | Condition 2 | Proximity to variety of conditions (if more than one) |
| 17 | Bldg Type | Type of dwelling |
| 18 | House Style | Style of dwelling |
| 19 | Overall Qual | Evaluates the home's overall construction quality and finish |
| 20 | Overall Cond | Evaluates the home’s overall condition |
| 21 | Year Built | Date of Construction |
| 22 | Year Remod/Add | Date of Renovation (same as construction date if no renovations or additions) |
| 23 | Roof Style | Type of roof |
| 24 | Roof Matl | Roof material |
| 25 | Exterior 1 | House Exterior covering material 1 |
| 26 | Exterior 2 | House Exterior covering material 2 (if more than one material) |
| 27 | Mas Vnr Type | Masonry veneer construction type |
| 28 | Mas Vnr Area | Area of Masonry veneer (square feet) |
| 29 | Exter Qual | Evaluates the external material's quality |
| 30 | Exter Cond | Evaluates the external material's present condition |
| 31 | Foundation | Foundation type |
| 32 | Bsmt Qual | Evaluates the basement’s height |
| 33 | Bsmt Cond | Evaluates the basement's overall condition |
| 34 | Bsmt Exposure | Denotes the walkout or garden-level walls |
| 35 | BsmtFin Type 1 | Evaluation of the finished basement space |
| 36 | BsmtFin SF 1 | Square feet of Type 1 finished |
| 37 | BsmtFinType 2 | Evaluation of finished basement space (if there are many types) |
| 38 | BsmtFin SF 2 | Square feet of Type 2 finished |
| 39 | Bsmt Unf SF | Area of the Unfinished basement (square feet) |
| 40 | Total Bsmt SF | Total Area of the basement (square feet) |
| 41 | Heating | Type of Heating system in the house |
| 42 | HeatingQC | Condition of the heating system in the house |
| 43 | Central Air | Central air conditioning |
| 44 | Electrical | Electrical system |
| 45 | 1st Flr SF | Square feet on the first floor |
| 46 | 2nd Flr SF | Square feet of the second Floor |
| 47 | Low Qual Fin SF | Square feet of Low-quality finished area (all floors) |
| 48 | Gr Liv Area | Area of Above grade ground living (square feet) |
| 49 | Bsmt Full Bath | Full bathrooms in the basement |
| 50 | Bsmt Half Bath | Half bathrooms in the basement |
| 51 | Full Bath | Full bathrooms - above grade |
| 52 | Half Bath | Half bathrooms - above grade |
| 53 | Bedroom | Bedrooms - above grade (basement bedrooms not included) |
| 54 | Kitchen | Kitchens above grade |
| 55 | KitchenQual | Kitchen Quality |
| 56 | TotRmsAbvGrd | Rooms above grade in total (bathrooms not included) |
| 57 | Functional | Functionalities used in the home |
| 58 | Fireplaces | Number of fireplaces |
| 59 | FireplaceQu | Fireplace quality |
| 60 | Garage Type | Location of the Garage |
| 61 | Garage Yr Blt | Year garage was constructed |
| 62 | Garage Finish | The garage’s interior finish |
| 63 | Garage Cars | Area of the garage (car capacity) |
| 64 | Garage Area | Area of the garage (square feet) |
| 65 | Garage Qual | Garage quality |
| 66 | Garage Cond | Garage condition |
| 67 | Paved Drive | Paved driveway |
| 68 | Wood Deck SF | Area of Wood deck (square feet) |
| 69 | Open Porch SF | Area of Open porch (square feet) |
| 70 | Enclosed Porch | Area of the enclosed porch (square feet) |
| 71 | 3-Ssn Porch | Area of Three-season porch (square feet) |
| 72 | Screen Porch | Area of Screen porch (square feet) |
| 73 | Pool Area | Area of the pool (square feet) |
| 74 | Pool QC | Pool quality |
| 75 | Fence | Fence quality |
| 76 | Misc Feature | Other features not included in the columns |
| 77 | Misc Val | $Value of miscellaneous feature |
| 78 | Mo Sold | Month Sold |
| 79 | Yr Sold | Year Sold |
| 80 | Sale Type | Type of sale |
| 81 | Sale Condition | Condition of sale |
| 82 | Sale Price | Sale Price of Houses |

*Table 1: Features of the Ames Housing Data Set with their dictionary*

**ANALYSIS & INTERPRETATION**

* Installing the libraries

R requires downloading and loading libraries in order to access datasets and functions that are not a part of the core R package.

**1. Loading the Ames housing dataset.**

* Importing the CSV file

<housing> vector contains information about the various properties.

**2. Perform Exploratory Data Analysis and summarize the data using descriptive statistics.**

* Understanding the housing dataset

With the help of str(), different datatypes can be noticed. The datatypes in this dataset are characters, integers, and numerals.

glimpse() is part of the dplyr package, and we can see the preview of columns of the dataset with the help of this function.

**Text, letter

Description automatically generated**

**Text, letter

Description automatically generated**

**Figure 1 – str() and glimpse()**

dim() views the number of columns and rows of a dataset. There are 2930 observations and 82 features in this case.

**A picture containing logo

Description automatically generated**

**Figure 2 – dim()**

* Descriptive statistics

An alternative to summary() that quickly gives a general overview of a data frame is skim().

skim() is useful for getting a statistical summary of the features.

There are 43 character and 39 numerical attributes in the dataset. Skim also returns the total missing data in each column.

**Table

Description automatically generated**

**Figure 3 – skim()**

To obtain the descriptive statistics of the data set's features, I utilized the describe() function from the package "PSYCH."

The mean of SalePrice (Body weight) is approximately $ 180,796, with a standard deviation of 79886.69. $12,789 is the minimum value and $755,000 is the maximum value.

Two metrics used to determine a distribution's shape are skew and kurtosis. Skew measures the asymmetry of a distribution, while kurtosis assesses how "peaked" a distribution is.

The SalePrice column has a skew value greater than 1 means it is right-skewed and that the mean is greater than the median. A kurtosis value of 5.1189 means data is highly peaked (Leptokurtic distribution).

**Table

Description automatically generated**

**Figure 4 –describe()**

* Exploratory Data Analysis

If the data is normal, a bell-shaped pattern will appear on the density plot.

The plots for Sale Price, Lot Footage, and Lot area show a right-skewed distribution.

We can observe that Years Built and Years Sold have many peaks in the distribution.

The density plot for Living Area has a multi-nodal and right-skewed distribution.

**Chart, surface chart

Description automatically generated**

**Figure 5 –Density Plots**

A boxplot is a depiction of numerical data that shows the first, third quartile, median, and extreme values of a dataset. Boxplots are useful for quickly identifying outliers and other features of the distribution of the data. In a boxplot, the quartiles should be uniformly spaced apart, and the median should be in the center of the box for normal distribution.

The box plots between Sale prices and House style reveals that 2Story and 2.5Fin (Two and one-half story: 2nd level finished) are priced more with respect to others.

**Chart, box and whisker chart

Description automatically generated**

**Figure 6 –Boxplots**

Ridge charts are a type of line chart used to visualize the relationship between two variables. They are similar to line charts, but instead of plotting points along a line, they plot points along a curved line.

The graph shows that Meadow Village has the cheapest housing, while neighbourhoods Northridge Heights and Stone Brook have luxurious homes.

**A picture containing diagram

Description automatically generated**

**Figure 7 – Ridge Chart**

Below bar chart shows fluctuations in the sales prices for each category of Building type. The year 2007 had the most expensive Single-family Detached homes. In the year 2009, Duplex homes were the costliest.

**Chart, bar chart

Description automatically generated**

**Figure 8 – Grouped Bar chart**

* Checking and removing duplicate rows

Duplicate rows should be eliminated to ensure that the data is accurate. To search for duplicate rows in the dataset, we utilized R's anyDuplicated() function. This function gives the index of the first duplicate encountered, confirming that the dataset contained some duplicate values. There are no duplicate rows.

A picture containing logo

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

**Figure 9- Duplicate rows**

**3. Data Imputation by Mean or other values**

* Checking for null,<NA> values

There are 5.8% of missing values in the dataset.

NA values are removed from the analysis because they can lead to inaccurate results and can create bias in the analysis. I was able to count the NA values within every column with the aid of the colSums() method.

Table

Description automatically generated

Table

Description automatically generated

Logo

Description automatically generated with low confidence

Icon

Description automatically generated

**Figure 10 -<NA> values**

* Imputing missing values

Data imputation by mean or other values is done in R to fill in missing data points. This is done to make sure that the data set is complete, and that the analysis is not skewed by the missing values.

I added the mean value to all missing values in the Lot Frontage column

I used mutate(), ifelse() and is.na() functions to replace N/A values with a string in other columns. For example, I substituted the NA values in the Alley column with “No Alley”.

Mutate() is a part of the dplyr package.

**4. Correlation matrix with cor() function**

The degree of association between variables is measured by correlation. A positive correlation between variables signifies that if one rises, the other increases as well. When there is a negative correlation, one variable rises while the other variable declines.

The correlation coefficients:

* -1 to -0.8: Very Strong negative
* -0.79 to -0.6: Strong negative
* -0.59 to -0.4: Moderate negative
* -0.39 to -0.2: Weak negative
* -0.19 to -0.01: Very weak negative
* 0 to 0.19: Very weak positive
* 0.2 to 0.39: Weak positive
* 0.4 to 0.59: Moderate positive
* 0.6 to 0.79: Strong positive
* 1 to 0.8: Very Strong positive
* Correlation Matrix

A table displaying the correlation between each of the columns in the data frame is known as a correlation matrix.

**Table

Description automatically generated**

**Figure 11 – Correlation Matrix**

Sale Price has a very strong positive relationship with Overall Qual (Correlation Coefficient – 0.8)

Sale Price has a positive relationship with Total Bsmt SF, 1st Flr SF, Gr Liv Area, Garage Cars, and Garage Area (Correlation Coefficient – 0.63,0.62,0.71,0.65 and 0.64 respectively)

**5. Correlation Plot**

An example of a graphical display that demonstrates the correlation between various variables is a correlation plot.

**Chart

Description automatically generated**

**Figure 12 – Correlation plot**

The color coding of a correlation plot shows the strength of the relationship. Red color depicts a negative correlation, and the blue color depicts a positive correlation.

**6. Make a scatter plot for the X continuous variable with the highest correlation with SalePrice. Do the same for the X variable that has the lowest correlation with SalePrice. Finally, make a scatter plot between X and SalePrice with a correlation closest to 0.5. Interpret the scatter plots and describe how the patterns differ.**

The variable with the highest correlation with SalePrice is Overall Qual(Correlation Coefficient – 0.8). The graph is close to a straight line and has a steep slope which represents a strong correlation.

**Chart, line chart

Description automatically generated**

**Figure 13 – Scatterplot between Sale price and Overall Quality**

BsmtFinType 2 has the lowest connection with SalePrice of all the variables (Correlation Coefficient – 0.01). The points in the graph are scattered in a random pattern with no clear linear relationship.

**Chart, scatter chart

Description automatically generated**

**Figure 14 – Scatterplot between Sale price and Rating of basement finished area**

The variable with a correlation closest to 0.5 with SalePrice is Masonry veneer area(Correlation Coefficient – 0.5). represented in the graph is less steep (not as steeply as the 0.8 correlation).

**Chart, scatter chart

Description automatically generated**

**Figure 15 – Scatterplot between Sale price and Masonry veneer area**

**7. Regression model with at least 3 continuous variables**

* Regression Model

In a regression model, a continuous outcome variable (or dependent variable) is predicted based on one or more predictor factors (or independent variables).

In this case, I have taken 5 continuous variables with strong correlation

Independent variable: Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area + Garage.Area + BsmtFin.SF.1

Dependent variable: SalePrice

**Table

Description automatically generated**

**Figure 16 – Model 1**

The Adjusted R2 value of **0.691396** is a relatively high R2 value, indicating that the model is a good fit for the data

****

**Figure 17 – Model 1 (Adjusted R2)**

AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are two statistical measures used to compare the relative goodness of fit of different models

Text

Description automatically generated with medium confidence

**Figure 18 – AIC and BIC**

**8. Regression model Equation**

**Y = -25511.583 + (45.068) Total.Bsmt.SF + (-5.111) X1st.Flr.SF + (71.178) Gr.Liv.Area + (101.264) Garage.Area + (23.159) BsmtFin.SF.1**

-25511.583: This is the intercept of the model, which is the predicted sale price when all of the other variables(independent) are equivalent to zero.

45.068: This coefficient represents the effect of Total Basement Square Feet on the sale price.

-5.111: This coefficient represents the effect of First Floor Square Feet on the sale price.

71.178: This coefficient represents the effect of Above grade (ground) Living Area on the sale price.

101.264: This coefficient represents the effect of the Garage Area on the sale price.

23.159: This coefficient represents the effect of Type 1 finished square feet on the sale price.

**9. Diagnostic plots -plot() function**

A model's validity is evaluated using diagnostic plots. They can help identify potential issues with the model, such as outliers, non-linearity, heteroscedasticity, and other problems.

* Residuals vs Fitted –

The difference between an observed value and a regression model's predicted value is depicted visually in the residuals vs. fitted graph. It is used to assess the accuracy of the model by showing if there is any pattern in the residuals. The residuals vs. fitted graph for this model shows a curved pattern, which shows that the data does not match a linear model.

* QQ Plots –

A quantile-quantile plot is a graph used to check whether a given data set follows a normal distribution. We can notice that the points form a straight line. So, the data is normally distributed. Outliers are also seen in the graph.

* Scale-Location –

The scale location graph is used to assess the homoscedasticity of a data set. It is a combination of a scatter plot and a line graph and can be used to identify outliers, trends, and correlations between variables. Additionally, it can be used to compare several groups. The points in the graph formed a curved shape, this means the data is not homoscedastic. This graph also clearly represents some of the outliers.

* Residuals vs Leverage –

It is a scatterplot of the residuals of the regression model against the leverage of each observation. Leverage tells us how far one observation is from the mean of all observations. It can also be utilized to detect outliers. The data points are in the shape of a cluster. Outliers should be further examined because they can significantly affect the model.

**Chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Chart

Description automatically generatedChart, scatter chart

Description automatically generated**

**Figure 19 – Diagnostic Plots**

**10. Check your model for multicollinearity and report your findings. What steps would you take to correct multicollinearity if it exists?**

The vif() function from the car package can be used to test for multicollinearity in R. VIF is less than 5 for all of the model's variables. The data is not significantly multicollinear.

**Text

Description automatically generated**

**Text

Description automatically generated**

**Figure 20 – Multicollinearity checks**

If multicollinearity exists in a dataset, the first step is to identify which variables are causing the issue. This can be done by calculating the variance inflation factor (VIF) for each variable. If the VIF is greater than 10, the variable is likely to be the cause of the multicollinearity.

Once the variables causing the issue have been identified, there are several steps that can be taken to correct it. Removing one of the highly linked variables from the model is the most typical strategy. Another option is to combine the two variables into one, by creating a new variable.

**11. Check your model for outliers and report your findings. Should these observations be removed from the model?**

OutlierTest() in R is a statistical function used to identify outliers in a dataset. It can be used to identify extreme values that may indicate errors or anomalies in the data.

**A screenshot of a computer

Description automatically generated with low confidence**

**Figure 21 – Outliertest()**

Outliers can skew the results of the regression model, making it difficult to accurately interpret the results. Outliers can also increase the variance of the model, making it difficult to identify the true relationship between the variables. We should remove extreme outliers. Removing outliers can help reduce the variance of the model and make it easier to interpret the results.

**12. Attempt to correct any issues that you have discovered in your model. Did your changes improve the model, why or why not?**

I have performed around 6 iterations and eliminated extreme outliers mentioned in the outliertest().

**Text, letter

Description automatically generated**

**Figure 22 – Removal of outliers**

When I ran the regression model again with the same variables (Total.Bsmt.SF + X1st.Flr.SF + Gr.Liv.Area + Garage.Area + BsmtFin.SF.1), this time the adjusted R2 increased from **0.6914 to 0.7542**. The adjusted R2 value increased because the outliers are no longer skewing the results.

**A screenshot of a computer

Description automatically generated with low confidence**

**Icon

Description automatically generated**

**Figure 23 – Model 1 after the removal of outliers**

The lower the AIC and BIC scores, the better the model. When outliers are removed from a model, the AIC and BIC scores decreased, and the model is likely to fit the data better.

**Text

Description automatically generated**

**Figure 24 – AIC and BIC of Model 1 after the removal of outliers**

**13. Use the all subsets regression method to identify the "best" model. State the preferred model in equation form.**

I have considered all the features in the second model for feature selection method.

Table

Description automatically generated with low confidence

Text

Description automatically generated

**Figure 25 – Model 2**

* Forward Selection: In this method, we begin with a model that contains no features. The feature that best enhances our model is added in each iteration until the performance of the model is not improved by the addition of a new variable.

Text

Description automatically generated

A close-up of a document

Description automatically generated with medium confidence

**Figure 26 – Forward Selection**

* Text, letter

  Description automatically generatedBackward Selection This is also an iterative procedure, where we begin with all the features and eliminate the least important feature at each iteration to enhance the model's functionality. We keep on doing this until the features are removed and no improvement is shown.

A screenshot of a computer

Description automatically generated with low confidence

**Figure 27 – Backward Selection**

* Stepwise Selection: This selection is a combination of forward selection and backward selection. We perform forward selection and then backward selection to come up with the best set of features.

Text, letter

Description automatically generated

A screenshot of a computer

Description automatically generated with low confidence

**Figure 28 – Stepwise Selection**

* Best Subset Selection: Best subset selection is an exhaustive search technique that evaluates all possible combinations of features and selects the best subset of features that provides the best performance.

**Table

Description automatically generated**

**Figure 29 – Best Subset Selection**

The adjusted R2 for the model after doing the best subset selection was the highest (0.8639).

According to the results, that proves the model fits the data well.

The best Subset selection helped me find the best model.

Independent variables -Subclass+ Overall.Qual+ Year.Built + BsmtFin.SF.1+ Gr.Liv.Area + Bedroom.AbvGr+ Garage.Area

**Table

Description automatically generated**

**Figure 30 – High Adjusted R2**

**Y = -7.366\*10^5 + (-2.132\*10^2) MS.SubClass + (1.951\*10^4) Overall.Qual + (3.54\*10^2) Year.Built + (3.177\*10^1) BsmtFin.SF.1 + (7.489\*10^1) Gr.Liv.Area + (-1.075\*10^4) Bedroom.AbvGr + (3.564\*10^1) Garage.Area**

**14. Compare the preferred model from step 13 with your model from step 12. How do they differ? Which model do you prefer and why?**

The two models vary in terms of their ability to explain the variation in the data. The model with R2 0.754 explains 75% of the variation in the data, while the model with R2 0.8639 explains 86% of the data variation.

I would prefer the model with an R2 of 0.86 because that model is more reliable and is therefore more likely to produce more accurate predictions.

**CONCLUSION**

If a person wants to buy a house, then it is his life’s significant milestone. There have been choices of individuals and there is a wide range of what is considered wants and needs based on individual interests in a home which is unique to each home buyer

In this assignment, we did data cleaning, EDA, plotted the correlation matrix, used regression models to determine the factors that influence the sale price of the houses, and performed regression diagnostics which helped us identify potential problems with the model, such as non-linearity, outliers, heteroscedasticity, and multicollinearity

The key findings are :

* By using a correlation matrix, we were able to understand how the sale price related to other parameters.
* The data variation can be defined by the R-square values of all the models.
* In order to deploy the model that best fits the data set, a variety of regression models were generated by experimenting with the variables.
* We tested for linearity, normality, homoscedasticity, multicollinearity, outliers, etc in the dataset and got rid of the extreme outliers which improved our initial model
* Lastly with the help of the Best subset method, we could recognize the “best” model with an adjusted R squared value of 0.8639.

**REFERENCES**

Bluman, A. G. (2018). Elementary Statistics, 10th ed. McGraw Hill.

Kabacoff, R. I. (2011). R in action: Data analysis and graphics with R. Manning Publications Co.

Ames Housing Dataset. (2018, September 10). Kaggle. <https://www.kaggle.com/datasets/prevek18/ames-housing-dataset>

Best Subsets Regression Essentials in R. (2018, March 11). Articles - STHDA. http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/155-best-subsets-regression-essentials-in-r/

**APPENDIX: CODE**

#---------------------- Week\_1\_Module\_1 R Script ----------------------#

print("Author : Nikshita Ranganathan")

print("Module 1 Assignment - Regression Diagnostics with R")

print("Course Name - ALY6015: Intermediate Analytics")

# Installing and loading the libraries

install.packages("psych")

library(car)

library(ggcorrplot)

library(psych)

library(GGally)

library(corrgram)

library(corrplot)

library(ggplot2)

library(leaps)

library(ochRe)

library(dplyr)

library(skimr)

library(visdat)

library(gridExtra)

library(naniar)

library(hrbrthemes)

library(magrittr)

install.packages("ggridges")

library(ggridges)

install.packages("olsrr")

library("olsrr")

# Importing the dataset

getwd()

housing<-read.csv("AmesHousing.csv")

# Descriptive statistics & EDA

str(housing)

summary(housing)

headTail(housing)

dim(housing)

skim(housing)

describe(housing,quant = c(0.25, 0.75),IQR = T)

glimpse(housing)

# Density Plots

d1<-ggplot(housing,aes(x=`SalePrice`))+geom\_density(fill = "#925E9F99")+xlab("Sale Price of Houses")+ggtitle("Density plot A")

d2<-ggplot(housing,aes(x=`Lot.Frontage`))+geom\_density(fill = "#79AF97FF")+xlab("Lot Frontage")+ggtitle("Density plot B")

d3<-ggplot(housing,aes(x=`Lot.Area`))+geom\_density(fill = "#FFDC91FF")+xlab("Lot Area")+ggtitle("Density plot C")

d4<-ggplot(housing,aes(x=`Year.Built`))+geom\_density(fill = "cornflowerblue")+xlab("Year Built")+ggtitle("Density plot D")

d5<-ggplot(housing,aes(x=`Yr.Sold`))+geom\_density(fill = "tomato3")+xlab("Year Sold")+ggtitle("Density plot E")

d6<-ggplot(housing,aes(x=`Gr.Liv.Area`))+geom\_density(fill = "#FF9DA7")+xlab("Living Area")+ggtitle("Density plot F")

grid.arrange(d1,d2,d3,d4,d5,d6)

# Boxplots

options(scipen=5)

ggplot(housing, aes(x=House.Style, y=SalePrice, fill=House.Style)) +

geom\_boxplot(outlier.size = 1) +theme\_ipsum()+scale\_fill\_ochre("qalah") +ggtitle("Boxplots")+xlab("House Style")+ylab("Sale Price of Houses")+labs(fill = "House Style")

# Ridge Chart

ggplot(housing, aes(x=SalePrice, y=Neighborhood, fill=Neighborhood)) +

geom\_density\_ridges() +theme(legend.position="none")+theme\_ipsum() +ggtitle("Ridge Chart")+xlab("Sale Price of Houses")+ylab("Neighbourhood")+labs(fill = "Neighbourhood")

# Grouped Bar Chart

options(scipen=5)

ggplot(housing, aes(Yr.Sold,SalePrice,fill=Bldg.Type)) +

geom\_col(position = "dodge")+

scale\_fill\_ochre(palette="tasmania")+ggtitle("Grouped Bar chart")+xlab("Year Sold")+ylab("Sale Price of Houses")+labs(fill = "Building Type")+theme(axis.text.x = element\_text( hjust = 1))

# Changing the datatypes

housing$Garage.Yr.Blt<-as.numeric(housing$Garage.Yr.Blt)

# Visualization and Checking NA values

vis\_miss(housing)

gg\_miss\_var(housing)

gg\_miss\_which(housing)

sum(is.na(housing))

sum(is.null(housing))

colSums(is.na(housing))

# Checking for duplicated rows and removing them

duplicated(housing)

anyDuplicated(housing)

# Imputation Model (Mean value and other values)

housing$Lot.Frontage[which(is.na(housing$Lot.Frontage))]<-mean(housing$Lot.Frontage,na.rm=TRUE)

housing <-housing %>% mutate(Alley = ifelse(is.na(Alley), "No Alley",Alley)) %>%

mutate(Mas.Vnr.Area = ifelse(is.na(Mas.Vnr.Area), 0,Mas.Vnr.Area)) %>%

mutate(Bsmt.Qual = ifelse(is.na(Bsmt.Qual), "No Bsmt",Bsmt.Qual)) %>%

mutate(Bsmt.Cond = ifelse(is.na(Bsmt.Cond), "No Bsmt",Bsmt.Cond)) %>%

mutate(Bsmt.Exposure = ifelse(is.na(Bsmt.Exposure), "No Bsmt",Bsmt.Exposure)) %>%

mutate(BsmtFin.Type.1 = ifelse(is.na(BsmtFin.Type.1), "No Bsmt",BsmtFin.Type.1)) %>%

mutate(BsmtFin.SF.1 = ifelse(is.na(BsmtFin.SF.1), 0,BsmtFin.SF.1)) %>%

mutate(BsmtFin.Type.2 = ifelse(is.na(BsmtFin.Type.2), "No Bsmt",BsmtFin.Type.2)) %>%

mutate(BsmtFin.SF.2 = ifelse(is.na(BsmtFin.SF.2), 0,BsmtFin.SF.2)) %>%

mutate(Bsmt.Unf.SF = ifelse(is.na(Bsmt.Unf.SF), 0,Bsmt.Unf.SF)) %>%

mutate(Total.Bsmt.SF = ifelse(is.na(Total.Bsmt.SF), 0,Total.Bsmt.SF)) %>%

mutate(Bsmt.Half.Bath = ifelse(is.na(Bsmt.Half.Bath), 0,Bsmt.Half.Bath)) %>%

mutate(Bsmt.Full.Bath = ifelse(is.na(Bsmt.Full.Bath), 0,Bsmt.Full.Bath)) %>%

mutate(Fireplace.Qu = ifelse(is.na(Fireplace.Qu), "No Fireplace",Fireplace.Qu)) %>%

mutate(Garage.Type = ifelse(is.na(Garage.Type), "No Garage",Garage.Type)) %>%

mutate(Garage.Yr.Blt = ifelse(is.na(Garage.Yr.Blt),0, Garage.Yr.Blt)) %>%

mutate(Garage.Finish = ifelse(is.na(Garage.Finish), "No Garage",Garage.Finish)) %>%

mutate(Garage.Cars = ifelse(is.na(Garage.Cars), 0,Garage.Cars)) %>%

mutate(Garage.Area = ifelse(is.na(Garage.Area), 0,Garage.Area)) %>%

mutate(Garage.Qual = ifelse(is.na(Garage.Qual), "No Garage",Garage.Qual)) %>%

mutate(Garage.Cond = ifelse(is.na(Garage.Cond), "No Garage",Garage.Cond)) %>%

mutate(Pool.QC = ifelse(is.na(Pool.QC), "No Pool",Pool.QC)) %>%

mutate(Fence = ifelse(is.na(Fence), "No Fence",Fence)) %>%

mutate(Misc.Feature=ifelse(is.na(Misc.Feature),"None",Misc.Feature))

# cor() function - Correlation Matrix

corr <- select\_if(housing, is.numeric)

drop<-c("Order","PID")

corr<-corr[,!(names(corr) %in% drop)]

cormatrix<-round(cor(corr,method = "pearson"),digits=2)

View(cormatrix)

# Correlation plot

ggcorrplot(cormatrix,

outline.color = "white",

ggtheme = theme\_bw(),

colors = c("#D53E4F","#FFFFBF","#3288BD"),tl.cex=7,title = "Correlation Plot",legend.title = "Correlation")

# Scatterplots

options(scipen=5)

ggplot(housing, aes(x=Overall.Qual, y=SalePrice)) +

geom\_point(alpha = 0.7,color="#00A9FF") +

geom\_smooth(color="#F8766D", fill="darkgrey", se=TRUE) +

theme\_ipsum\_rc(grid="XY") +labs(title="Scatterplot A",x="Overall Quality", y="House Sale Price")

ggplot(housing, aes(x=BsmtFin.SF.2, y=SalePrice)) +

geom\_point(alpha = 0.7,color="#00BF7d") +

geom\_smooth(color="#F8766D", fill="darkgrey", se=TRUE) +

theme\_ipsum\_rc(grid="XY") +labs(title="Scatterplot B",x=" Rating of basement finished area (if multiple types)", y="House Sale Price")

ggplot(housing, aes(x=Mas.Vnr.Area, y=SalePrice)) +

geom\_point(alpha = 0.7,color="#925e9fff") +

geom\_smooth(color="#F8766D", fill="darkgrey", se=TRUE) +

theme\_ipsum\_rc(grid="XY") +labs(title="Scatterplot C",x=" Masonry veneer area (sqft)", y="House Sale Price")

# Regression Model

model1<-lm(formula = SalePrice ~ Total.Bsmt.SF+X1st.Flr.SF+Gr.Liv.Area+Garage.Area+BsmtFin.SF.1, data = corr)

summary(model1)

summary(model1)$adj.r.squared

AIC(model1)

BIC(model1)

# Equation = Y=-25511.583+(45.068)Total.Bsmt.SF+(-5.111)X1st.Flr.SF+(71.178)Gr.Liv.Area+(101.264)Garage.Area+(23.159)BsmtFin.SF.1

# Diagnostic Plots

plot(model1)

qqPlot(model1, labels = rownames(cormatrix), simulate = TRUE, main = "Q-Q Plot")

crPlots(model=model1)

spreadLevelPlot(model1)

vif(model1)

# Multicollinearity

ols\_vif\_tol(model1)

# Outlier Detection & Removal

outlierTest(model = model1)

corr <- corr[!(row.names(corr) %in% c("1499","2181","2182","1768","45","434","1064","1761","2333","433")),]

model1<-lm(formula = SalePrice ~ Total.Bsmt.SF+X1st.Flr.SF+Gr.Liv.Area+Garage.Area+BsmtFin.SF.1, data = corr)

outlierTest(model = model1)

corr <- corr[!(row.names(corr) %in% c("2593","1638","2331","2446","2451","2335","1641","1183")),]

model1<-lm(formula = SalePrice ~ Total.Bsmt.SF+X1st.Flr.SF+Gr.Liv.Area+Garage.Area+BsmtFin.SF.1, data = corr)

outlierTest(model = model1)

corr <- corr[!(row.names(corr) %in% c("2342")),]

model1<-lm(formula = SalePrice ~ Total.Bsmt.SF+X1st.Flr.SF+Gr.Liv.Area+Garage.Area+BsmtFin.SF.1, data = corr)

outlierTest(model = model1)

corr <- corr[!(row.names(corr) %in% c("1643")),]

model1<-lm(formula = SalePrice ~ Total.Bsmt.SF+X1st.Flr.SF+Gr.Liv.Area+Garage.Area+BsmtFin.SF.1, data = corr)

outlierTest(model = model1)

corr <- corr[!(row.names(corr) %in% c("424")),]

model1<-lm(formula = SalePrice ~ Total.Bsmt.SF+X1st.Flr.SF+Gr.Liv.Area+Garage.Area+BsmtFin.SF.1, data = corr)

summary(model1)

summary(model1)$adj.r.squared

AIC(model1)

BIC(model1)

# Feature Selection

library(MASS)

model2<-lm(formula = SalePrice ~ ., data = corr)

summary(model2)

# Forward Stepwise Selection

stepAIC(model2, direction = "forward")

# Backward Stepwise Selection

stepAIC(model2, direction = "backward")

# Stepwise Selection

stepAIC(model2, direction = "both")

# Best Subset Selection

leaps <- regsubsets(SalePrice ~ ., data = corr, nbest = 4)

summary(leaps)

with(summary(leaps),data.frame(rsq,adjr2,cp,rss,outmat))

plot(leaps, scale = "adjr2",main = "Adjusted R^2")

final <- lm(formula = SalePrice ~ MS.SubClass +Overall.Qual + Year.Built +

BsmtFin.SF.1 + Gr.Liv.Area + Bedroom.AbvGr + Garage.Area,

data = corr)

summary(final)