Contents

PART A-

1. Exploratory Study on Property Valuation in the USA:

Predictive modelling cannot be fully understood without first having a firm grasp of the field of property value. Following are some of the significant factors:

Location: A property's location significantly affects its value. criminal rates, criminal proximity, and a property's distance from parks, schools, and public transportation are a few factors that might affect property values.

Features of the property: Size, number of bedrooms and baths, lot size, and general condition are all significant factors influencing a property's value.

Market Trends: A number of elements, such as interest rates, local real estate market trends, and prevailing economic conditions, have an impact on property valuation.

Government regulations: Zoning laws, local government regulations, and property tax rates may have an impact on a property's value.

Services around the area: Areas with closer proximity to Schools, Univerities or Hospitals generally play a very important role in predicting the pricing of the property

2. Identifying Data Sources:

Property assessment databases are-

Real estate databases: These historical sources compile details on past real estate deals, such as sale prices, property details, and transaction dates. Zillow, Redfin, and Realtor.com are a few instances.

Government Records: Local governments typically keep records of real estate transactions, tax assessments, and property information.

Property Websites: Sites like Zillow and Trulia offer details about listed properties, such as asking prices.

Environmental Information: Environmental information can have an impact on real estate values. Examples include pollution levels, flood plains, and accessibility to natural attractions.

Economic Indicators: You can assess the state of the economy that affects how much a property is worth by looking at data on interest rates, unemployment rates, and GDP growth.

Potential Information Formats

- Fields for the property address, sale price, square footage, number of bedrooms and bathrooms, and other essential information are included in the traditional structure of real estate transaction data.
- Economic and census data are typically available in tabular form.
- GeoJSON and shapefiles are two GIS file types that can both be used to store geospatial data.

Blockers in obtaining data

- In order to access it, it may be essential to obtain historical property sales data from public or private data sources.
- There could be exclusive data sources with licencing or subscription costs.
- Data consistency and quality can vary between sources.

3. Variables for Predictive Model:

When creating a prediction model for property price assessment, the following factors should be considered:

- Lot size, year of construction, square footage, number of bedrooms and bathrooms, and general condition are all characteristics of a property.
- The location is defined by its geographic coordinates or address, community characteristics, ease of access to amenities, and school district.
- Interest rates, local real estate market movements, and economic information are market trends.
- Government regulations include things like zoning regulations, property tax rates, and any new rules.
- Demographic data refers to details on the local population, income levels, and demography.

The efficacy of your predictive model will depend on how accurate and relevant the collected data is, as well as the modelling techniques applied. Thus, Domain knowledge is crucial for making informed decisions throughout the project.

PART B-

1. Numerical and Categorical Variables-

Variables are considered to be numerical variables when they have a quantifiable number and can answer questions like ‘how many’ or ‘how much’. Numerical variables can be divided in Discrete- Countable values or Continuous- Measurable values. (“Variables | Australian Bureau of Statistics,” 2023)

Variables are considered to be categorical variables when they describe a quantity or characteristic of a data unit 'what type' or 'which category'. Categorical variables can be divided into Nominal- Variables without a logical sequence and Ordinal- Variables with a ranking or sequence. (“Variables | Australian Bureau of Statistics,” 2023)

Categorical		Numerical	
Nominal	**Ordinal**	**Discrete**	**Continuous**
Id	LotShape	FullBath	LotArea
LandContour	Slope	HalfBath	
Utilities	OverallQuality	BedroomAbvGr	TotalBSF
LotConfig	OverallCondition	KitchenAbvGr	LowQualFinSF
DwellClass	ExteriorCondition	TotalRmsAbvGrd	LivingArea
GarageType	BasementCondition	Fireplaces	GarageCars
CentralAir	KitchenQuality	SalePrice	PoolArea
PavedDrive	MoSold		OpenPorchSF
	YrSold		
	YearBuilt		

b. **Summary Statistics-**

- Min, Median, Max for numerical variables

```
   FullBath        HalfBath       BedroomAbvGr    KitchenAbvGr   TotalRmsAbvGrd
Min.   :0.000   Min.   :0.0000   Min.   :0.000   Min.   :0.000   Min.   : 2.00
1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 5.00
Median :2.000   Median :0.0000   Median :3.000   Median :1.000   Median : 6.00
Mean   :1.566   Mean   :0.3831   Mean   :2.869   Mean   :1.047   Mean   : 6.52
3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:3.000   3rd Qu.:1.000   3rd Qu.: 7.00
Max.   :3.000   Max.   :2.0000   Max.   :8.000   Max.   :3.000   Max.   :14.00

  Fireplaces       SalePrice        LotArea          TotalBSF     LowQualFinSF
Min.   :0.0000   Min.   : 34900   Min.   :  1300   Min.   :   0   Min.   :  0.000
1st Qu.:0.0000   1st Qu.:130000   1st Qu.:  7544   1st Qu.: 799   1st Qu.:  0.000
Median :1.0000   Median :163250   Median :  9478   Median : 999   Median :  0.000
Mean   :0.6142   Mean   :181112   Mean   : 10521   Mean   :1065   Mean   :  5.869
3rd Qu.:1.0000   3rd Qu.:214000   3rd Qu.: 11604   3rd Qu.:1304   3rd Qu.:  0.000
Max.   :3.0000   Max.   :755000   Max.   :215245   Max.   :6110   Max.   :572.000
                                                   NA's   :49
  LivingArea     GarageCars       PoolArea       OpenPorchSF
Min.   : 334   Min.   :0.000   Min.   :  0.00   Min.   :  0.00
1st Qu.:1131   1st Qu.:1.000   1st Qu.:  0.00   1st Qu.:  0.00
Median :1467   Median :2.000   Median :  0.00   Median : 25.00
Mean   :1517   Mean   :1.771   Mean   :  2.77   Mean   : 46.37
3rd Qu.:1780   3rd Qu.:2.000   3rd Qu.:  0.00   3rd Qu.: 68.00
Max.   :5642   Max.   :4.000   Max.   :738.00   Max.   :547.00
NA's   :22
```

- Skewness for numerical variables-

```
  FullBath       HalfBath   BedroomAbvGr   KitchenAbvGr TotalRmsAbvGrd     Fireplaces
0.03750844     0.67497576     0.21459872     4.47322624     0.67818529     0.64716146
 SalePrice        LotArea       TotalBSF   LowQualFinSF     LivingArea     GarageCars
1.89164556    12.17391591             NA     8.98288957             NA    -0.33510866
  PoolArea    OpenPorchSF
14.78224984     2.26606930
```

- Standard Deviation-

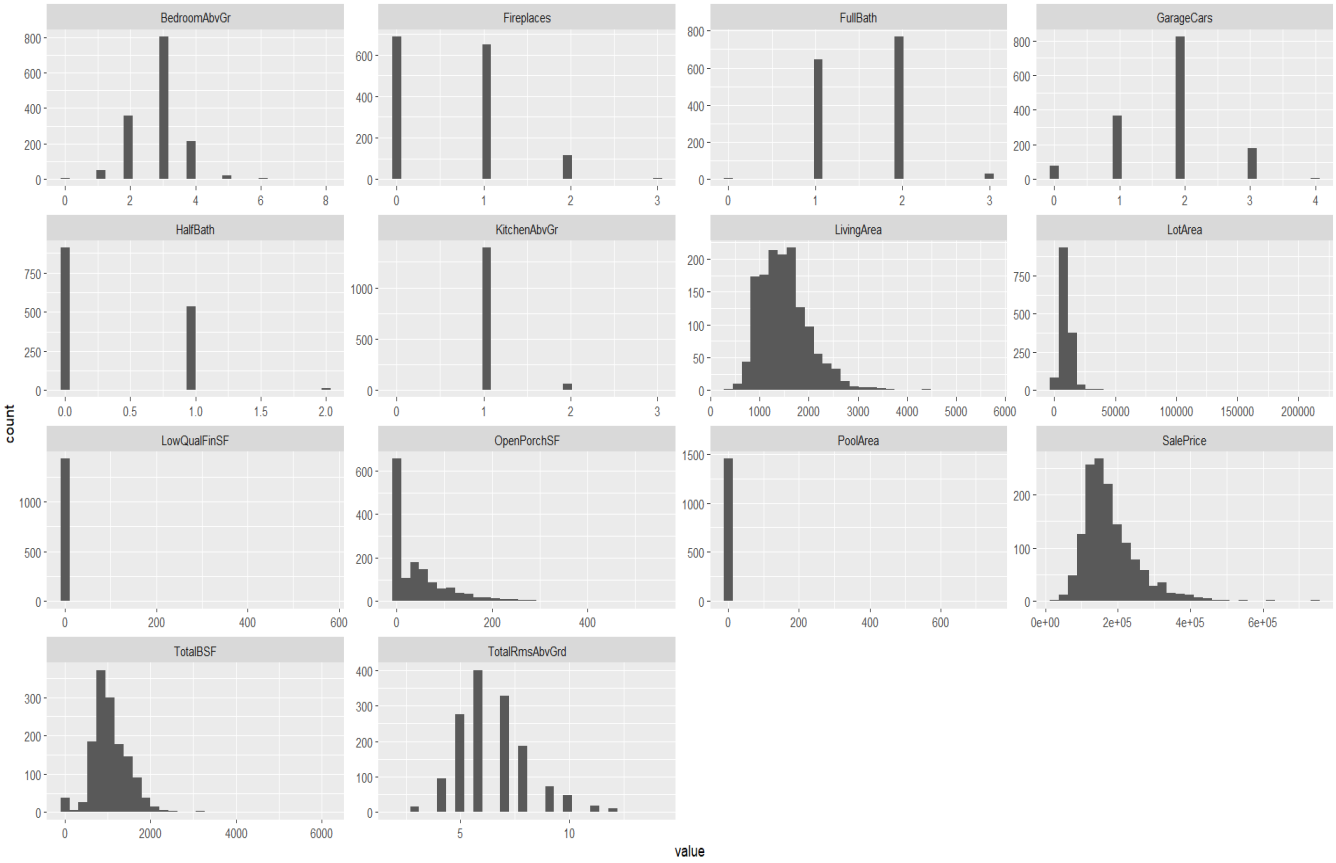
```
    FullBath      HalfBath   BedroomAbvGr   KitchenAbvGr TotalRmsAbvGrd     Fireplaces
5.485146e-01  5.030012e-01  8.153339e-01   2.207723e-01   1.624596e+00  6.450340e-01
   SalePrice       LotArea      TotalBSF   LowQualFinSF     LivingArea     GarageCars
7.933169e+04  1.000046e+04             NA  4.872192e+01             NA  7.442069e-01
    PoolArea    OpenPorchSF
4.025978e+01  6.513858e+01
```

- Frequency of Categorical Variables

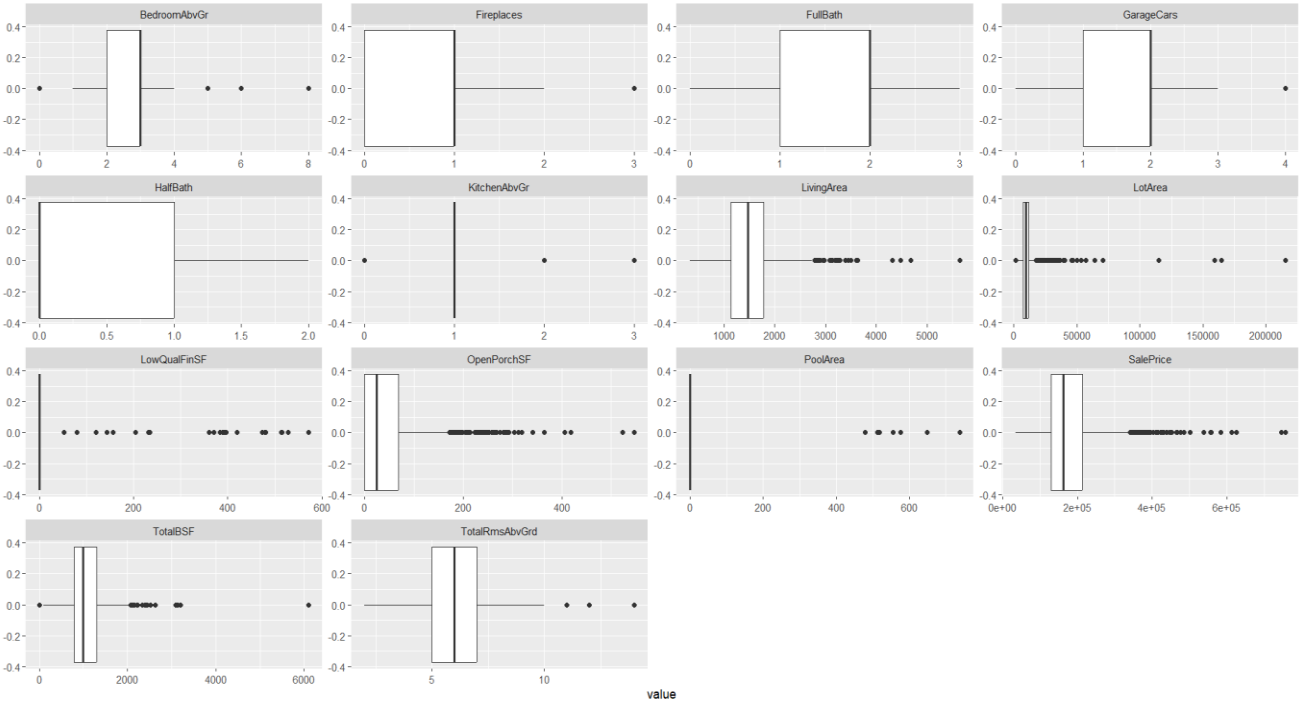
$LandContour	Bnk	HLS	Low	Lvl								
	63	50	35	1306								
$Utilities	AllPub	NoSeWa										
	1453	1										
$LotConfig	Corner	CulDSac	FR2	FR3	Inside							
	262	94	47	4	1047							
$DwellClass	1Fam	2fmCon	Duplex	Twnhs	TwnhsE							
	1214	31	52	43	114							
$GarageType	2Types	Attchd	Basment	BuiltIn	CarPort	Detchd	NA					
	6	870	19	88	9	384	78					
$CentralAir	N	Y										
	94	1360										
$PavedDrive	N	P	Y									
	88	30	1336									
$LotShape	1	2	3	4								
	10	41	482	921								
$Slope	Gtl	Mod	Sev									
	1377	65	12									
$OverallQuality	1	2	3	4	5	6	7	8	9	10		
	1	3	19	116	395	374	318	168	43	17		
$OverallCondition	2	3	4	5	6	7	8	9				
	3	25	57	821	252	205	72	19				
$ExteriorCondition	Fa	Gd	TA									
	26	146	1282									
$BasementCondition	Fa	Gd	NB	TA								
	45	65	37	1307								
$KitchenQuality	Ex	Fa	Gd	TA								
	100	37	584	733								
$MoSold	1	2	3	4	5	6	7	8	9	10	11	12
	58	52	105	141	203	252	233	122	63	88	78	59
$YrSold	2006	2007	2008	2009	2010							

	314	328	302	336	174							

2.a. Histogram for Numerical variables-



2.b. Boxplot for numerical variables-



c. Explanation-

i. Largest variability Variables-

We can define variability using standard deviations. The variables having the largest variability-

- Saleprice- 79331.69
- Lot area- 10000
- LowqualFinSF- 48.72
- PoolArea- 40.25
- OpenPorchSF- 65.13

ii. Skewed Variables-

We can check the skewness of a variable using the Skewness values and the histogram. Acceptable skewness values vary between -3 and +3. The variables having the most skewed values are-

- KitchenAbvGrd- 4.47 (Right Skewed)
- LotArea-12.17 (Right Skewed)
- LowQualityFinSF- 8.98 (Right Skewed)
- PoolArea- 14.78 (Right Skewed)

iii. Variables with Extreme Values-

We can check for the extreme values of variables using the box plot and the outliers in the plot. If a plot has multiple outliers, it represents extreme values in a plot According to the data the following variables have the most outliers or extreme values-

- LivingArea
- LotArea
- LowQualFinSF
- OpenPorchSF
- PoolArea
- SalePrice
- TotalBSF

3.a. Variables with Missing values-

According to the results below, we see that TotalBSF and LivingArea in numeric variables has missing values.

```
  FullBath       HalfBath   BedroomAbvGr   KitchenAbvGr TotalRmsAbvGrd     Fireplaces
         0              0              0              0              0              0
 SalePrice        LotArea      TotalBSF   LowQualFinSF     LivingArea     GarageCars
         0              0             49              0             22              0
  PoolArea    OpenPorchSF
         0              0
```

According to the results below, we see that there are no missing values in categorical data.

```
            Id       LandContour         Utilities         LotConfig        DwellClass
             0                 0                 0                 0                 0
    GarageType        CentralAir        PavedDrive         LotShape             Slope
             0                 0                 0                 0                 0
OverallQuality  OverallCondition ExteriorCondition BasementCondition    KitchenQuality
             0                 0                 0                 0                 0
        MoSold            YrSold         YearBuilt
             0                 0                 0
```

b. Methods to handle missing values-

- Add zero to the missing values- We can add the value '0' to the missing values in order to handle the missing values
- Add mean to the missing values- We can add the mean of the values without the empty values to the missing places.
- Delete the values- The last method is to remove the rows containing the missing values to handle the data.

c. Handling missing values in Housing data-

- Add Zero to missing values-
  R- Script-

  ```
  num_col_zero<-num_col
  is.na(num_col_zero)
  num_col_zero[is.na(num_col_zero)]<- 0
  ```

colSums(is.na(num_col_zero))

- Add Mean to missing values-

R-Script-

num_col_mean<-num_col

colSums(is.na(num_col_mean))

num_col_mean$TotalBSF[is.na(num_col_mean$TotalBSF)]<-
mean(num_col_mean$TotalBSF,na.rm = TRUE)

num_col_mean$LivingArea[is.na(num_col_mean$LivingArea)]<-
mean(num_col_mean$LivingArea,na.rm = TRUE)

- Delete missing values-

R-Script-

num_col_del<-num_col[complete.cases(num_col),]

d. Comparison of Summary Statistics after handling missing values-

- Summary Statistics for TotalBSF-

```
> summary(num_col$TotalBSF)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0     799     999    1065    1304    6110      49
> summary(num_col_zero$TotalBSF)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   780.0   982.5  1028.6  1285.5  6110.0
> summary(num_col_del$TotalBSF)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0     799    1002    1066    1307    6110
> summary(num_col_mean$TotalBSF)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   806.2  1017.0  1064.5  1285.5  6110.0
```

- Skewness for TotalBSF-

```
> skewness(num_col$TotalBSF)
[1] NA
> skewness(num_col_zero$TotalBSF)
[1] 1.043557
> skewness(num_col_del$TotalBSF)
[1] 1.538457
> skewness(num_col_mean$TotalBSF)
[1] 1.569715
```

- Standard Deviation for TotalBSF-

```
> sd(num_col$TotalBSF)
[1] NA
> sd(num_col_zero$TotalBSF)
[1] 473.714
> sd(num_col_del$TotalBSF)
[1] 442.098
> sd(num_col_mean$TotalBSF)
[1] 432.9874
```

- Summary Statistics for Living Area

```
> summary(num_col$LivingArea)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    334    1131    1467    1517    1780    5642      22
> summary(num_col_zero$LivingArea)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    1120    1456    1494    1775    5642
> summary(num_col_del$LivingArea)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    334    1130    1465    1517    1784    5642
> summary(num_col_mean$LivingArea)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    334    1136    1474    1517    1775    5642
```

- Skewness for Living Area

```
> skewness(num_col$LivingArea)
[1] NA
> skewness(num_col_zero$LivingArea)
[1] 0.9713158
> skewness(num_col_del$LivingArea)
[1] 1.387363
> skewness(num_col_mean$LivingArea)
[1] 1.384482
```

- Standard Deviation for Living Area

```
> sd(num_col$LivingArea)
[1] NA
> sd(num_col_zero$LivingArea)
[1] 553.6328
> sd(num_col_del$LivingArea)
[1] 528.8984
> sd(num_col_mean$LivingArea)
[1] 521.7111
```

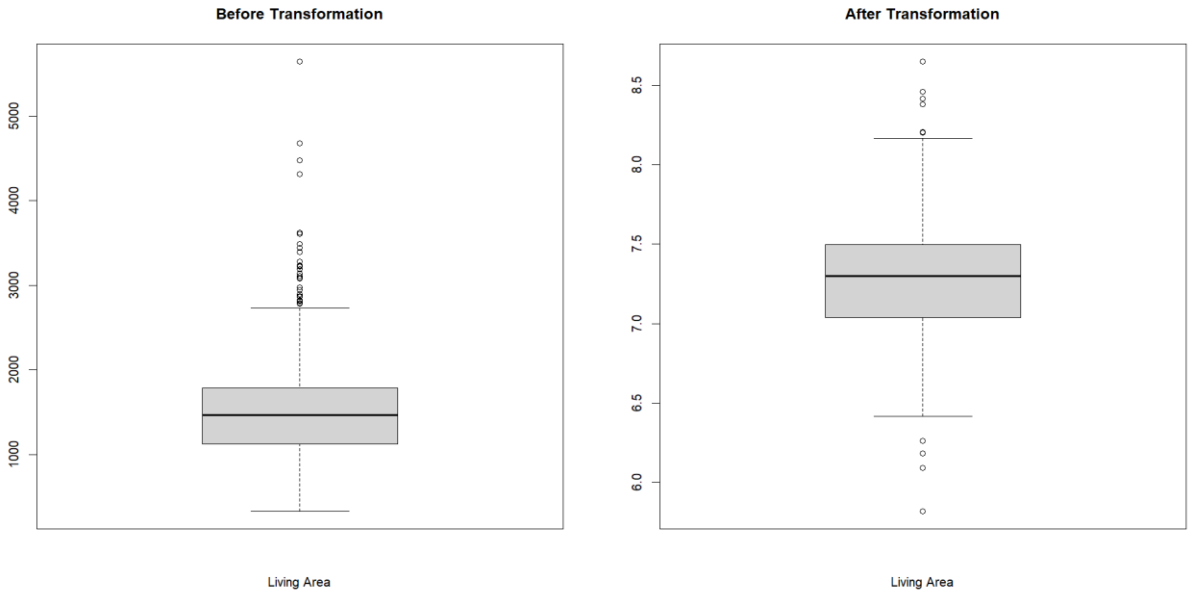
Comparing the data after handling the missing values, we see that the least deviation in the mean, median for TotalBSF and Living Area occurs when the values are deleted. Thus, deleting the values is the best method to handle the missing values for this dataset.

e. Transformation of Right Skewed Variables-

In this task, we will perform the transformation of two right skewed variables- SalePrice and LivingArea. We will be performing the transformation of the right skewed variables using the Logarithmic transformation. In this method, we add a log to the variable and add a constant to avoid values reaching infinity.

- Transformation of SalePrice-

  Boxplot before and after transformation-

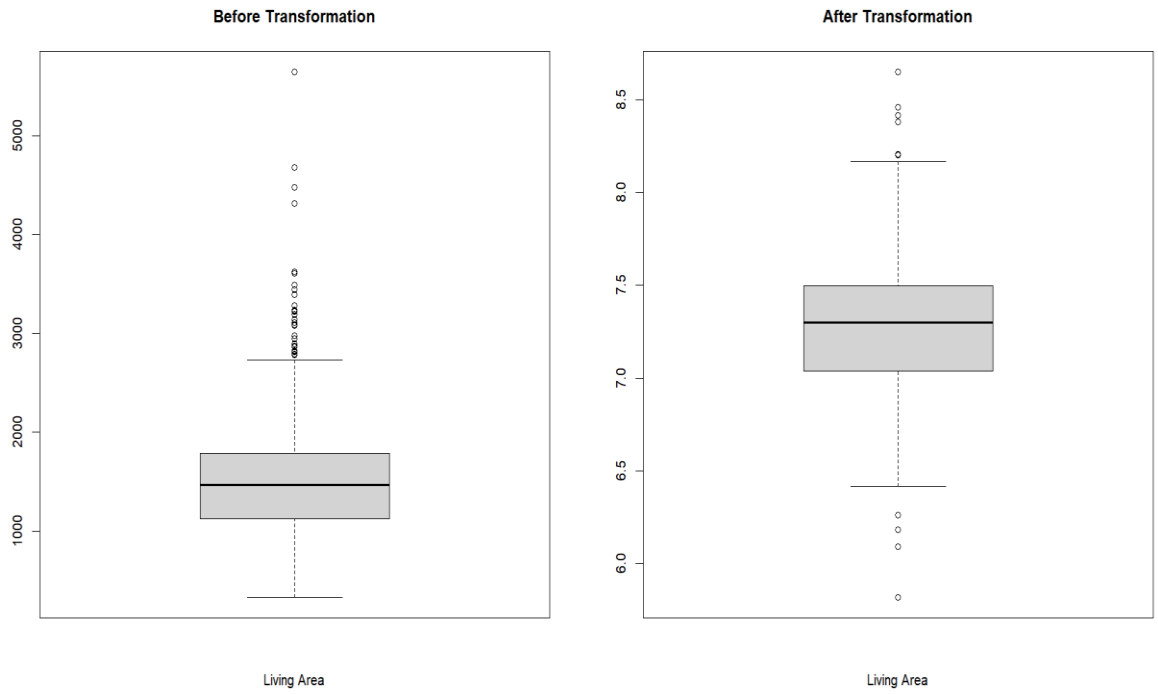


  On performing the transformation we get the skewness as, Before Transformation- 1.89

  After Transformation- 0.13 which is almost a normally distributed data

- Transformation of Living Area-

  Boxplot before and after transformation-



On performing the transformation, we get the skewness, before transformation as 1.38 and after transformation as -0.0003 which is a normally distributed data.

4.a. Methods of transforming categorical variables to numerical variables-

- One hot encoding- In one-hot encoding, each unique category value is assigned a binary value of 1 or 0. For example, in our dataset, we have CentralAir, in this if we consider N as 0 and Y as 1, then if a data has the value N, we put it as 10 and for the value Y, we put it as 01.
- Integer Encoding- In Integer encoding, we assign each unique category value with an integer. For example, for the variable ExteriorCondition in our dataset, we will consider Ex-1, Gd-2, TA as 3, Fa as 4 and Po as 5, then the converted data will be Integer encoded

b. Performing transformation of categorical variables to numerical variables for the dataset

- In this case, we will be performing Integer Encoding-

```
transform_data<-housing_data%>% mutate(LandContour=as.numeric(factor(LandContour)))%>%
mutate(Utilities=as.numeric(factor(Utilities)))%>%
mutate(LotConfig=as.numeric(factor(LotConfig)))%>%
mutate(Slope=as.numeric(factor(Slope)))%>%
mutate(DwellClass=as.numeric(factor(DwellClass)))%>%
mutate(ExteriorCondition=as.numeric(factor(ExteriorCondition)))%>%
mutate(BasementCondition=as.numeric(factor(BasementCondition)))%>%
mutate(KitchenQuality=as.numeric(factor(KitchenQuality)))%>%
```

mutate(GarageType=as.numeric(factor(GarageType)))%>%

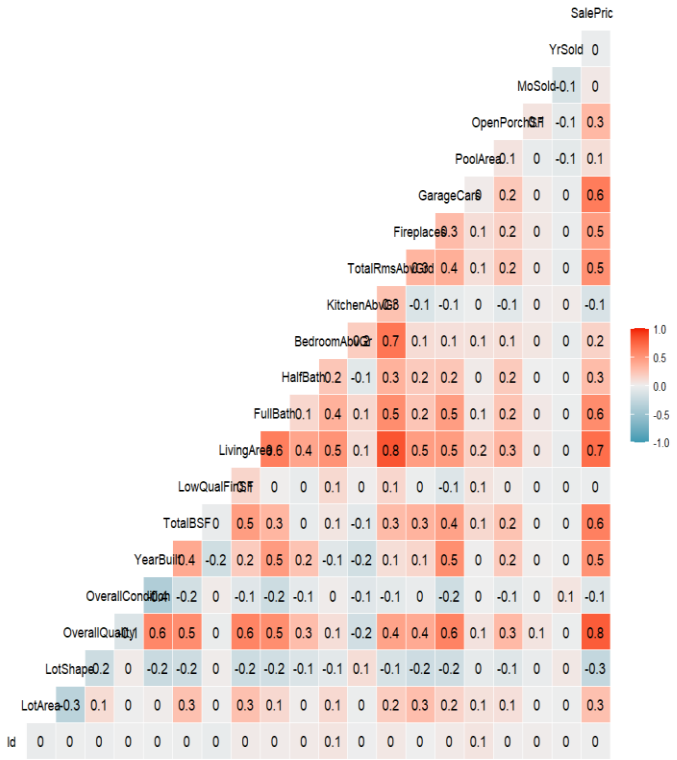
mutate(PavedDrive=as.numeric(factor(PavedDrive)))%>%

mutate(CentralAir=as.numeric(factor(CentralAir)))

- In this case, we are not converting the MoSold, YrSold and YearBuilt as they are already numeric categorical values.

5.a. Correlation between variables-

- Correlation Matrix-

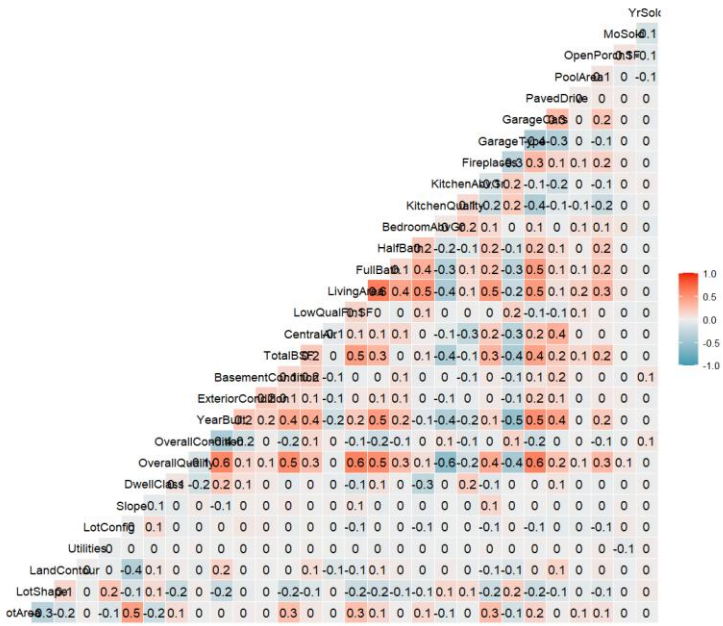


In this matrix, we see a high correlation between the variables LivingArea and TotalRmsAbvGrd of 0.8 and OverallQuality and Saleprice of 0.8. In this process, using R, we find out the variables with the highest correlation to perform dimensionality reduction.

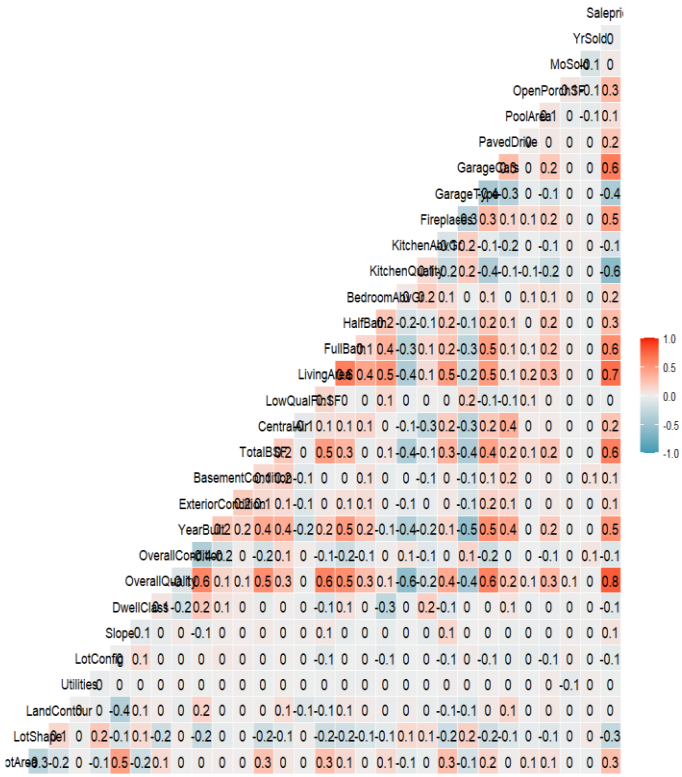
b. Performing Dimensionality Reduction-

We find the highly correlated variables to be "LivingArea" "OverallQuality" "YearBuilt" "TotalRmsAbvGrd"

In this data, we can remove TotalRmsAbvGrd as it is highly correlated to BedroomAbvGrd, FullBath and LivingArea.



c. Distribution of selected variables against the target variable(Saleprice)-



In this matrix, we see a high correlation of the Saleprice with the GarageCars, KitchenQuality, FullBath, LivingArea and OverallQuality, i.e., these factors highly affect the Saleprice of the property.

PART C-

1. Building Linear Model

a. Evaluating regression metrics-

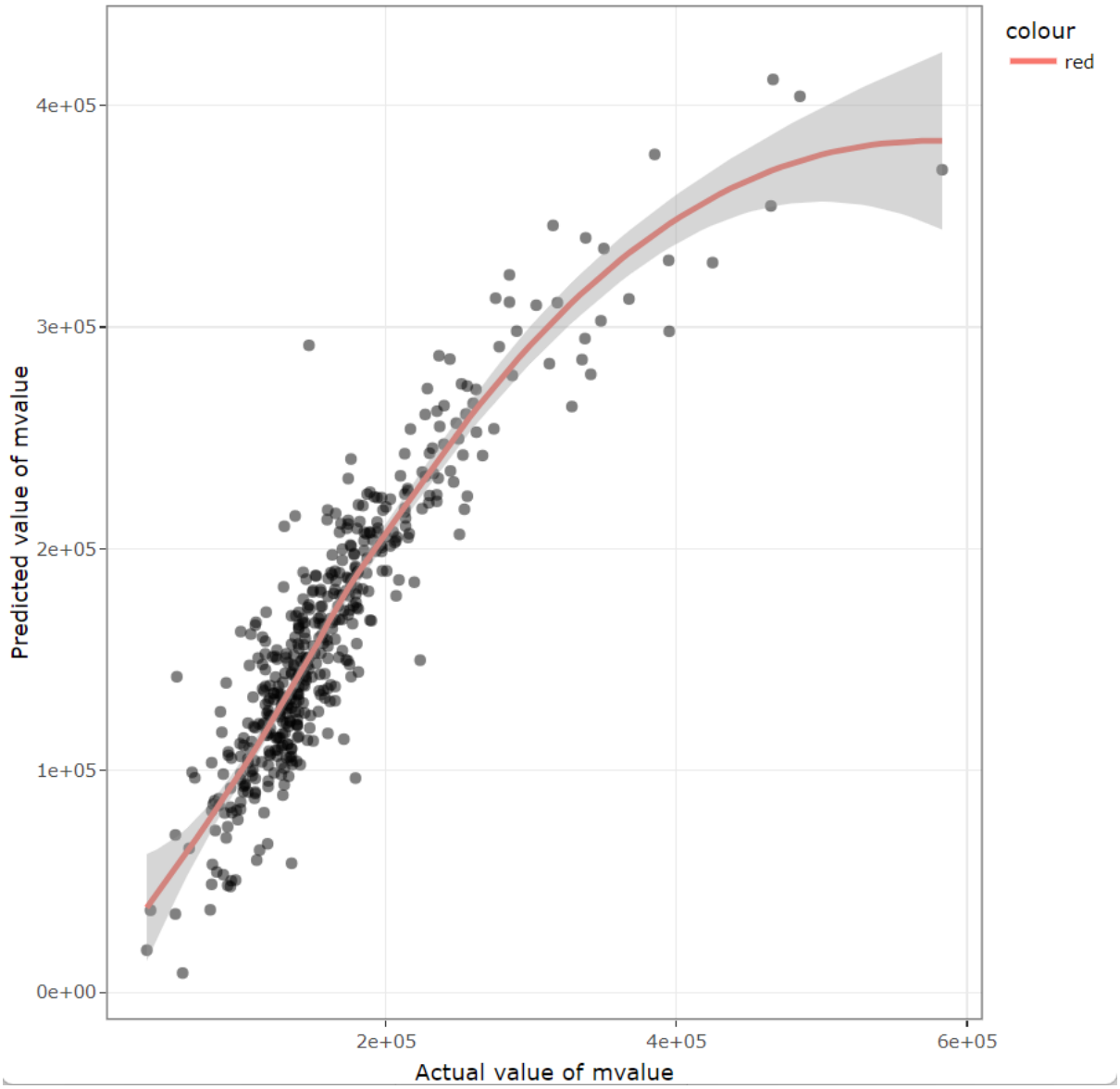
- In this model, we will be using the data set after removing the TotalRmsAbvGrd variable. We have created a linear regression model in order to predict the SalePrice.

In order to create the R model, we use the R-code below-

```
#set sample configuration

smp_size <- floor(2/3 * nrow(housing_data_selected))

set.seed(2)


housing.train <- housing_data_selected[1:smp_size, ]

housing.test <- housing_data_selected[(smp_size+1):nrow(housing_data_selected),]


#building predictive model

library(stats)

formula = Saleprice~.


model <- lm(formula = formula, data = housing.train)

# Display the coefficients of the linear regression model

summary(model)$coefficients


#applying regression formula

as.formula(
  paste0("y ~ ", round(coefficients(model)[1],2), " + ",
         paste(sprintf("%.2f * %s",coefficients(model)[-1],
                       names(coefficients(model)[-1])),
               collapse=" + ")
  )
)


#predicted probabilities


housing.train$predicted.Saleprice <- predict(model, housing.train) #use type="class" for logarithmic regression
```

```
housing.test$predicted.Saleprice <- predict(model, housing.test)
print("Actual Values")
head(housing.test$Saleprice[1:5])
print("Predicted Values")
head(housing.test$predicted.SalePrice[1:5])

library(plotly)
plot1 <-housing.test %>%
  ggplot(aes(Saleprice,predicted.Saleprice)) +
  geom_point(alpha=0.5) +
  stat_smooth(aes(colour='red')) +
  xlab('Actual value of mvalue') +
  ylab('Predicted value of mvalue')+
  theme_bw()
```

ggplotly(plot1)



In order to perform the evaluation of the prediction, we use the Root Mean Square Error (RMSE), upon checking the RMSE we get the value as 29111.35

In order to fetch this we use the R code-

error1 <- housing.test$Saleprice-housing.test$predicted.Saleprice

rmse1 <- sqrt(mean(error1^2))

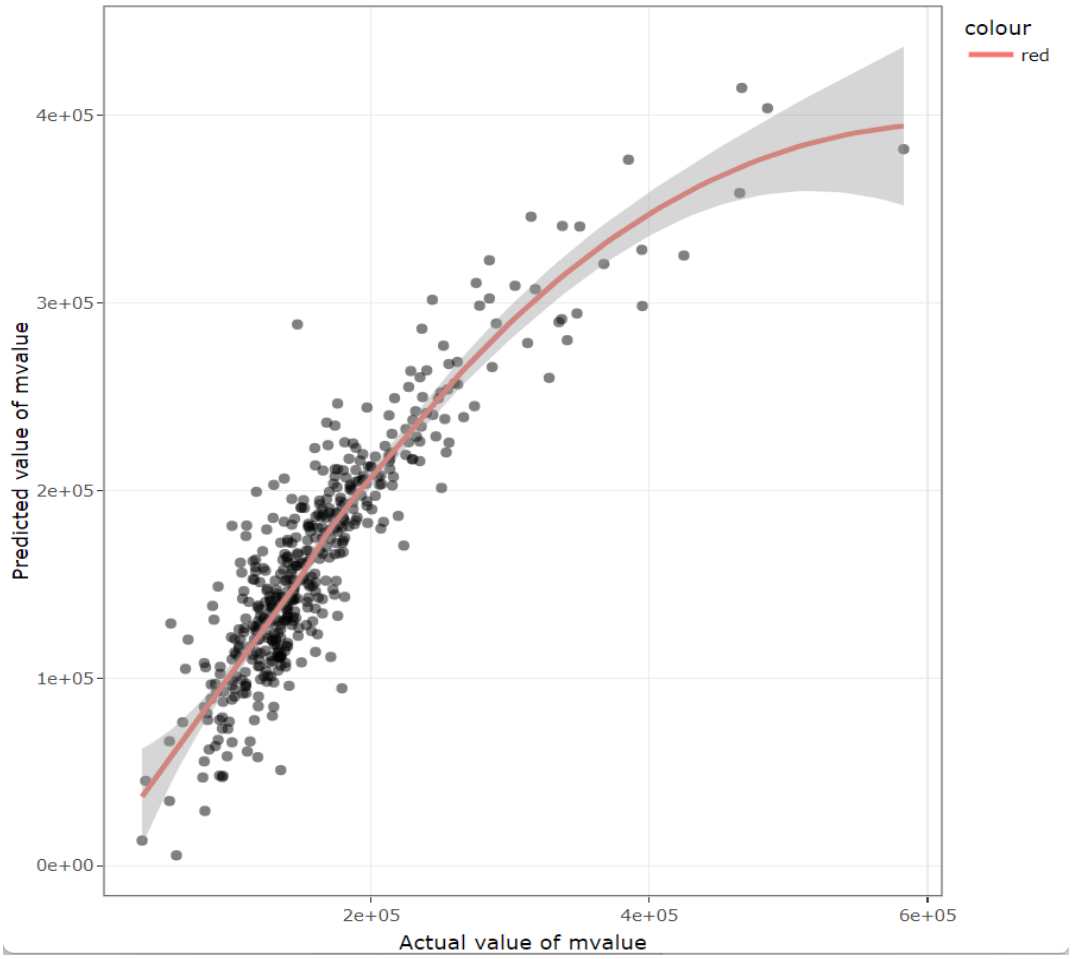
print(paste("Root Mean Square Error: ", rmse1))

b. Performing Feature Selection-

In order to consider the best feature, we need to perform dimensionality reduction on the variables with the highest correlation.

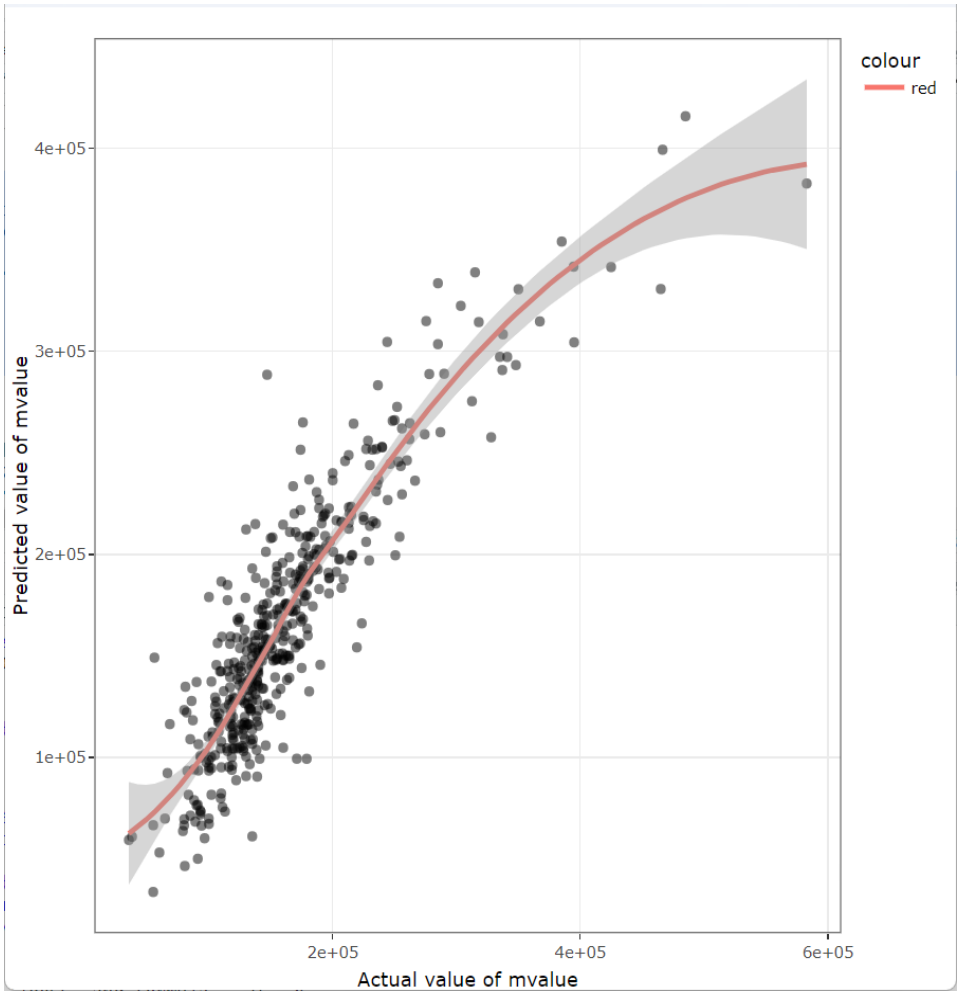
We will be performing prediction and evaluating the prediction by removing the other variables i.e. YearBuilt and OverallQuality.

- Linear Model 1- Removing YearBuilt



Upon calculating the RMSE, we get the value as 30354.27

- Linear Model 2- Removing OverallQuality



Upon calculating the RMSE, we get the value as 30557.67.

b. Comparison of Linear Models-

Upon reviewing the linear models, we see that the model created after removing the TotalRmsAbvGrd gives us the least RMSE value of 29111.35. We use the formula-

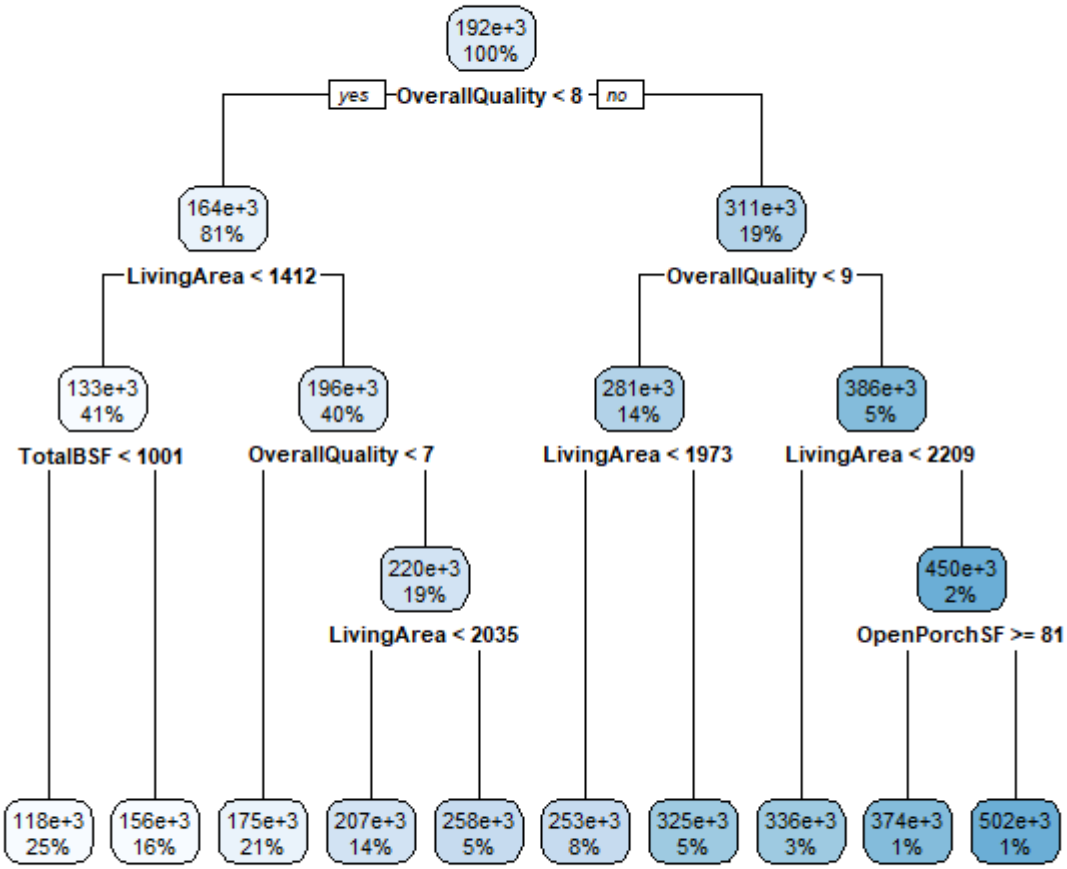
error1 <- housing.test$Saleprice-housing.test$predicted.Saleprice

rmse1 <- sqrt(mean(error1^2))

2. Decision Tree Modelling-

a. Building a decision tree

In order to create a decision tree we will be considering the model without Pruning. On performing decision tree modelling, we get the following model-



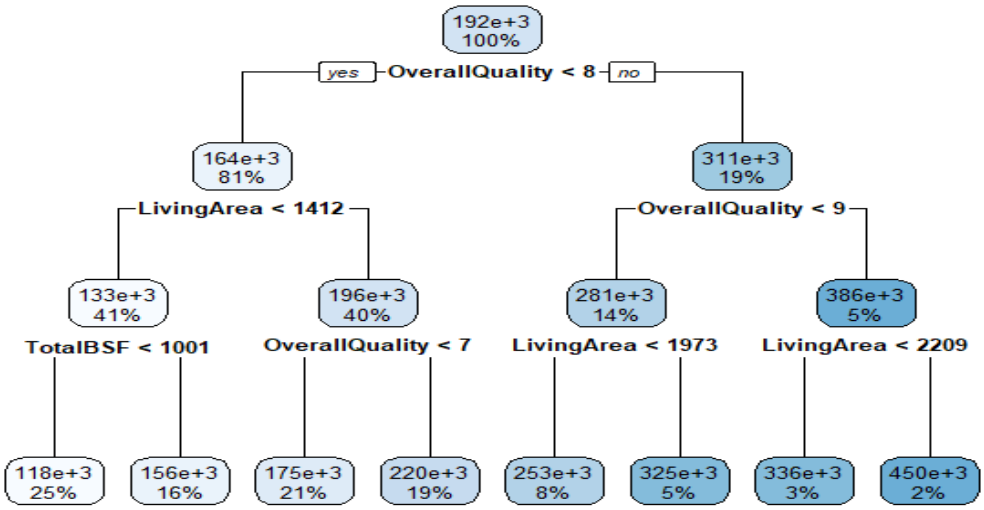
Upon performing the RMSE, we get the following value 38977.59

b. Testing Decision Tree Models

In order to consider the best feature, we need to perform dimensionality reduction on the variables with the highest correlation.

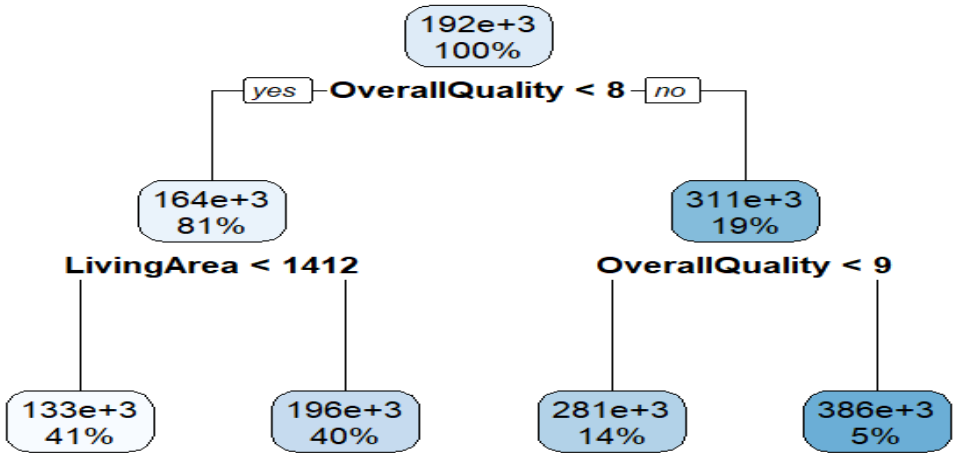
We will be performing prediction and evaluating the prediction by pruning the decision tree with CP 0.02 and CP 0.04

i. Decision Tree 1- Cp=0.02



Upon performing the RMSE, we get the following value 104028

ii. Decision Tree 2- CP=0.04



Upon performing the RMSE, we get the following value 50997.36

c. **Comparing the decision tree models-**

Upon review of the decision tree models, we see that the tree with CP=0.04 with 50997.36 RMSE can be considered. Upon pruning and running the test data for testing the model, we get the least RMSE while using the CP=0.04 model. Hence, we can consider this model for our predictions.

We can calculate this value using the formula-

residuals <- housing.train$Saleprice - test_predictions

rmse <- sqrt(mean(residuals^2))

3. Model Comparison-

a. **Reason to create multiple models-**

We use multiple models to improve the performance of the model. By building several models and checking their predictions, we can estimate the value would get from a single model. This is because each model has a different method for prediction.

Using multiple models reduces overfitting of the data. When a training data is overfitted, it becomes probable for the model to generalize new data. By building several models, we select the one that performs the best with the test data thus reducing the the risk of overfitting.

Linear regression and decision trees make different assumptions of the data. By building assumptions, we gain a better understanding of the data and the best way to model it.

Decision trees are more interpretable than linear regression models as by building several decision trees with different depths, we can gain a better understanding of the relationships between the features and the target variable.

In general, it is a good practice to build several models in both linear regression and decision trees before selecting the final model. This will help to ensure that we are getting the best possible results.

b.**Comparison of linear and decision tree modelling-**

According to the linear and decision tree models, we cannot conclude the complete reliance on a particular model without some trade off. In this case, the RMSE value is lower for the linear model than the decision tree. We will be considering the linear model because our data has multiple variables affecting the Sale Price. As there are a large number of data to be considered, a linear model is better to avoid overfitting of the data. Thus, after identifying the correlated variables using dimensionality reduction, we have considered to move with the linear model.

References-

Gomez, J. (2022, June 3). Opendoor | Sell your home the minute you're ready. Retrieved from opendoor.com website: https://www.opendoor.com/articles/factors-that-influence-home-value

St, S. W., 24/7 Wall. (n.d.). 11 home features buyers will pay extra for. Retrieved from USA TODAY website: https://www.usatoday.com/story/money/personalfinance/2013/04/28/24-7-home-features/2106203/

Variables | Australian Bureau of Statistics. (2023, February 2). Retrieved from www.abs.gov.au website: https://www.abs.gov.au/statistics/understanding-statistics/statistical-terms-and-concepts/variables#:~:text=Numeric%20variables%20have%20values%20that