PART A: Problem formulation

Aspects impacting property prices in USA and home valuation methods:

Housing prices in the USA, on one hand, can be influenced by various internal aspects, which revolve around the key features of the properties. The interior patterns such total sizes, space usability, number of bedrooms and bathrooms, the property ages and conditions etc., in fact, contribute towards that valuation and purchase prices of properties (Gomez, 2019). Similarly, exterior sides such as the surrounding environmental aspects, location ideality, proximity to the CBD or even the current prices of neighborhood properties can significantly increase a home's selling price under favorable conditions. On the other hand, external factors also shape property prices, such as economic conditions that impact employment, income, and interest rates, ultimately affecting housing supply and demand. Another key factor is demographics, where increasing urbanization and population growth generally lead to higher property prices, especially for those located near the center of the USA (Leung, 2024). Governmental aspects including taxation system and subsidies program also shape the demand for housing.

Rohde (n.d.) notes that the sales comparison approach is widely favored for property valuation, as it benchmarks a house's value against recent sales of similar properties. The second common method is the income-based approach, which estimates the net present value of rental income for a property, depending on assumptions about discount and growth rates. Thirdly, a cost-based approach determines prices based on building costs, useful when recent sales data are scarce. Automated valuation models employ machine learning and algorithms to predict house prices by integrating various factors, offering advantages in processing multiple variables and enhancing accuracy through pattern recognition (Martin, n.d.).

Useful data sources:

There are available sources for extracting data of property market, one of which is the public webpages, online news, or blogs that can be highly accessible. The format of extracted data could be reports or raw dataset collection files, implying a potential issue with regards to missing values, unorganized or poorly labeled data, and lacked data dictionaries for understanding of key variables. The second type of housing data source is through governmental bodies which usually make reports and dataset public online annually (Consumer Affairs Victoria, n.d.). While the credibility and format of data are reasonably expected to be structured and presentable, the information provided might have been aggregated on a nation-wide level, posing a challenge for smaller-scaled analysis. The third potential housing data source comes from real estate agencies, which tend to be highly industry-specific and timely relevant. However, some worth-noting problems are that access might be restricted against the public and due to the agency's interest in selling more properties, the data collected might be biased towards favorable outcomes.

Variables of interest:

The combination of variables, covering both key factors influencing property prices and a mix of categorical and numeric data, will enhance the predictive model's performance.

- Numeric variables: interior (number of rooms, sizes, space), exterior (distance to CBD, prices of similar properties), economic (employment rates, income), government (house taxes, subsidies), demographics (population growth)
- Ordinal variables: interior (conditions, quality), exterior (driveway states), economic (income levels), government (zone restriction levels), demographics (age levels)
- Nominal variables: interior (room types), exterior (building types), government (subsidies options), demographics (household class)

PART B: Data exploration and cleaning

<u>Variables categorization:</u>

a. The categorizations of variables in the dataset are as follows:

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

b. For converting ordinal variables to numeric ones, integer-encoding method could be used to transform categorical values to numbers, while preserving the hierarchy or sequence of the relevant values. That is, the magnitude of the integer being assigned to one value in the ordinal sequence should be based on that ordinal value's rankings, compared to the others.
For transforming nominal variables to numeric ones, one-hot-encoding method could be used to create new binary columns in the dataset. Each column represents one nominal category and if a row belongs to a particular category, the value in that cell would be 1 and in other cells would be 0.

c. (In R script)

<u>Data exploration:</u>

a. Summary statistics of continuous variables:

```
LotArea:
Mean:  10521.13
Median:  9478.5
Max:  215245
Standard Deviation:  10000.46

TotalBSF:
Mean:  1058.357
Median:  992
Max:  6110
Standard Deviation:  439.1744

LowQualFinSF:
Mean:  5.868638
Median:  0
Max:  572
Standard Deviation:  48.72192

LivingArea:
Mean:  1517.197
Median:  1466
Max:  5642
Standard Deviation:  525.4673
```

```
LivingArea:
Mean:  1517.197
Median:  1466
Max:  5642
Standard Deviation:  525.4673

OpenPorchSF:
Mean:  46.37001
Median:  25
Max:  547
Standard Deviation:  65.13858

PoolArea:
Mean:  2.770289
Median:  0
Max:  738
Standard Deviation:  40.25978

SalePrice:
Mean:  181111.7
Median:  163250
Max:  755000
Standard Deviation:  79331.69
```

Counts of nominal variables:

```
Nominal Variables Counts:
LotConfig - Inside:  1047
LotConfig - Corner:  262
LotConfig - CulDSac:  94
LotConfig - FR2:  47
LotConfig - FR3:  4
DwellClass - 1Fam:  1214
DwellClass - 2FmCon:  31
DwellClass - Duplx:  52
DwellClass - TwnhsE:  114
DwellClass - TwnhsI:  43
CentralAir - Y:  1360
CentralAir - N:  94
GarageType - 2Types:  6
GarageType - Attchd:  870
GarageType - Basment:  19
GarageType - BuiltIn:  88
GarageType - CarPort:  9
GarageType - Detchd:  384
GarageType - NoGarage:  78
```

Counts of ordinal variables:

```
LotShape:
Reg (0):  10
IR1 (1):  41
IR2 (2):  482
IR3 (3):  921

LandContour:
Lvl (0):  1306
Bnk (1):  63
HLS (2):  50
Low (3):  35

Utilities:
AllPub (0):  1453
NoSewr (1):  NA
NoSeWa (2):  1
ELO (3):  NA

Slope:
Gtl (0):  1377
Mod (1):  65
Sev (2):  12
```

```
OverallQuality:
0 :  1
1 :  3
2 :  19
3 :  116
4 :  395
5 :  374
6 :  318
7 :  168
8 :  43
9 :  17

OverallCondition:
0 :  NA
1 :  3
2 :  25
3 :  57
4 :  821
5 :  252
6 :  205
7 :  72
8 :  19
9 :  NA

ExteriorCondition:
Ex (0):  NA
Gd (1):  146
TA (2):  1282
Fa (3):  26
Po (4):  NA
```

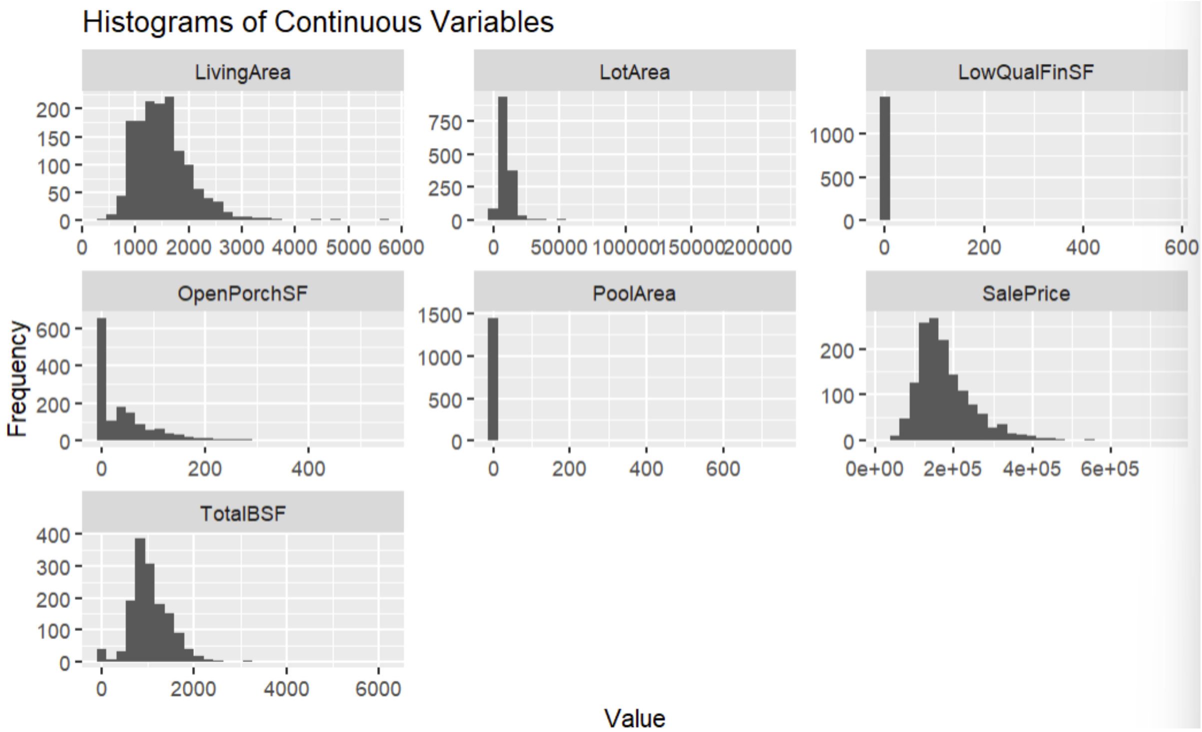
```
BasementCondition:
Ex (0):  NA
Gd (1):  65
TA (2):  1307
Fa (3):  45
Po (4):  NA
NB (5):  37

KitchenQuality:
Ex (0):  100
Gd (1):  584
TA (2):  733
Fa (3):  37
Po (4):  NA

PavedDrive:
Y (0):  1336
P (1):  30
N (2):  88
```

b. Based on the summary statistics of continuous variables, there is likely to have the presence of extreme values across **all continuous variables** because the maximum values are extremely larger than the mean or median. For instance, the LotArea variable has a maximum value of 215,245, which exceeds more than 20 times the mean value of 10,521.13 and the median of 9,478.5. The same principle applies to all remaining continuous variables, including TotalBSF, LowQualFinSF, LivingArea, OpenPorchSF, PoolArea, SalePrice, whose extreme values are their maximum.

Continuous variable distribution:




```
   LotArea              TotalBSF
 Min.   :  1300     Min.   :   0
 1st Qu.:  7544     1st Qu.: 796
 Median :  9478     Median : 992
 Mean   : 10521     Mean   :1058
 3rd Qu.: 11604     3rd Qu.:1300
 Max.   :215245     Max.   :6110

  LowQualFinSF        LivingArea
 Min.   :  0.000     Min.   : 334
 1st Qu.:  0.000     1st Qu.:1131
 Median :  0.000     Median :1466
 Mean   :  5.869     Mean   :1517
 3rd Qu.:  0.000     3rd Qu.:1777
 Max.   :572.000     Max.   :5642
                     NA's   :10

  OpenPorchSF          PoolArea
 Min.   :  0.00     Min.   :  0.00
 1st Qu.:  0.00     1st Qu.:  0.00
 Median : 25.00     Median :  0.00
 Mean   : 46.37     Mean   :  2.77
 3rd Qu.: 68.00     3rd Qu.:  0.00
 Max.   :547.00     Max.   :738.00

   SalePrice
 Min.   : 34900
 1st Qu.:130000
 Median :163250
 Mean   :181112
 3rd Qu.:214000
 Max.   :755000
```

a. The variable with the largest variability is SalePrice because it has the highest standard deviation, and the range (i.e. distance between minimum and maximum values) is also the largest compared to other continuous variables. The histogram of SalePrice also supports this conclusion because it displays a wide dispersion with a significant number of values concentrated in the lower range, but it tails off significantly as it approaches higher values, indicating a high variability. The presence of extreme values (or outliers) further extends the spread on the right side of the distribution

b. As shown in the histograms as above, all variables are skewed to the right with long tails. However, LotArea, LowQualFinSF, OpenPorchSF, and PoolArea show highest degree of right skewness, while LivingArea, SalePrice, and TotalBSF shows a slightly less pronounced right skewness problem

c. All variables are having extreme values, which are likely their maximum values. As discussed above, since all histograms of are right skewed, the extreme values are located in the right tail of the histograms, and hence, the extreme values are the upper-bound and maximum values.

Handling missing values:

a. Using the summary(). function in R gives the statistics summary of all variables. As shown below, the variables that are having missing values (i.e. NAs) are YearBuilt with 13 NAs and LivingArea with 10 NAs.

```
 OverallCondition   YearBuilt     ExteriorCondition
 Min.   :1.000    Min.   :1872    Min.   :1.000
 1st Qu.:4.000    1st Qu.:1954    1st Qu.:2.000
 Median :4.000    Median :1973    Median :2.000
 Mean   :4.576    Mean   :1972    Mean   :1.917
 3rd Qu.:5.000    3rd Qu.:2000    3rd Qu.:2.000
 Max.   :8.000    Max.   :2010    Max.   :3.000
                  NA's   :13
 BasementCondition     TotalBSF      LowQualFinSF
 Min.   :1.000      Min.   :   0    Min.   :  0.000
 1st Qu.:2.000      1st Qu.: 796    1st Qu.:  0.000
 Median :2.000      Median : 992    Median :  0.000
 Mean   :2.063      Mean   :1058    Mean   :  5.869
 3rd Qu.:2.000      3rd Qu.:1300    3rd Qu.:  0.000
 Max.   :5.000      Max.   :6110    Max.   :572.000

   LivingArea        FullBath        HalfBath
 Min.   : 334    Min.   :0.000    Min.   :0.0000
 1st Qu.:1131    1st Qu.:1.000    1st Qu.:0.0000
 Median :1466    Median :2.000    Median :0.0000
 Mean   :1517    Mean   :1.566    Mean   :0.3831
 3rd Qu.:1777    3rd Qu.:2.000    3rd Qu.:1.0000
 Max.   :5642    Max.   :3.000    Max.   :2.0000
 NA's   :10
```

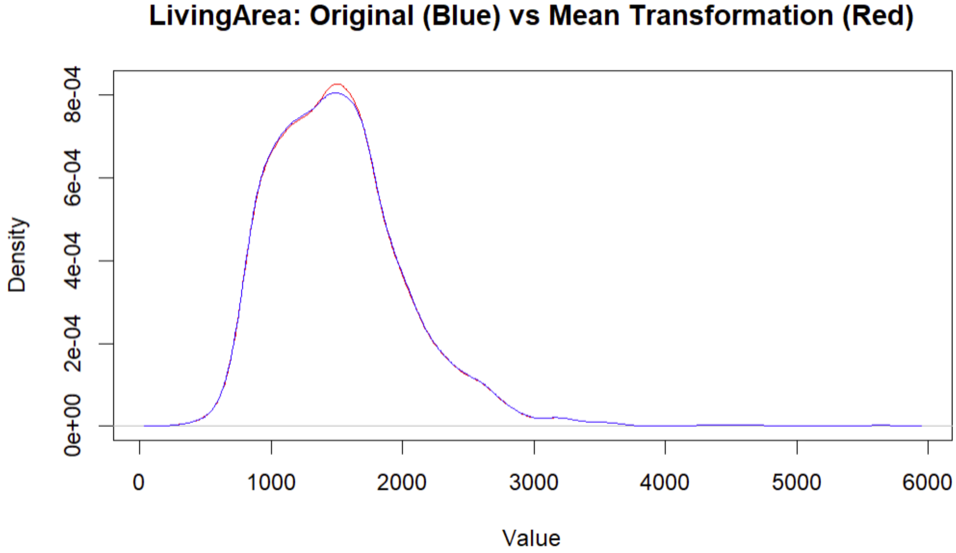
```
Summary statistics of LivingArea_Original:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    334    1131    1466    1517    1777    5642      10
Summary statistics of YearBuilt_Original:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   1872    1954    1973    1972    2000    2010      13
```

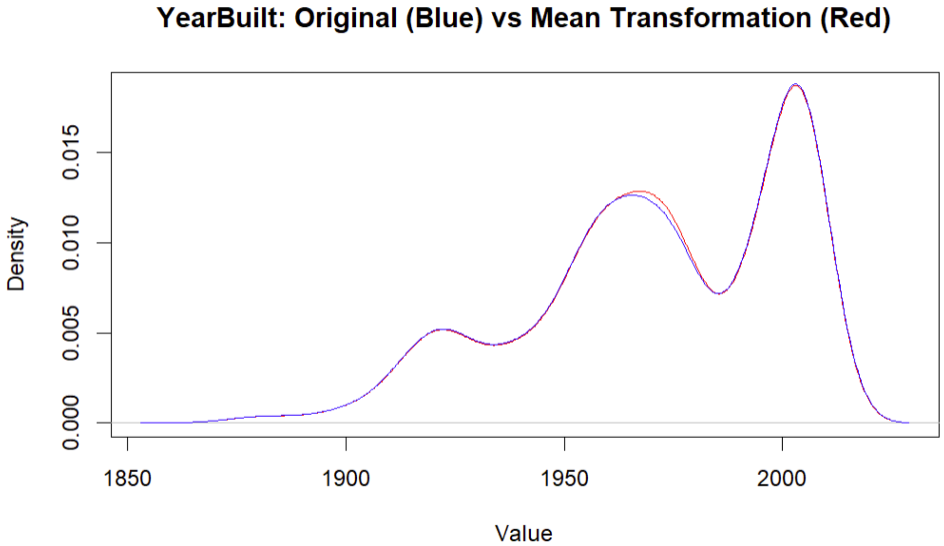
b. There are three main ways to handle missing values:
   - For numeric variables, missing values can be handled by filling the in the mean or median values. For categorical variables, the mode can be used to fill in the missing values.
   - The second way is to delete all records containing missing values.

- For numeric variables, the third method is to replace missing values with certain value, such as 0 in most cases. For categorical variables, the missing values can be set as another category, such as “undefined” or “unknown”.

c. Method 1: fill missing values with median for YearBuilt and mean for LivingArea

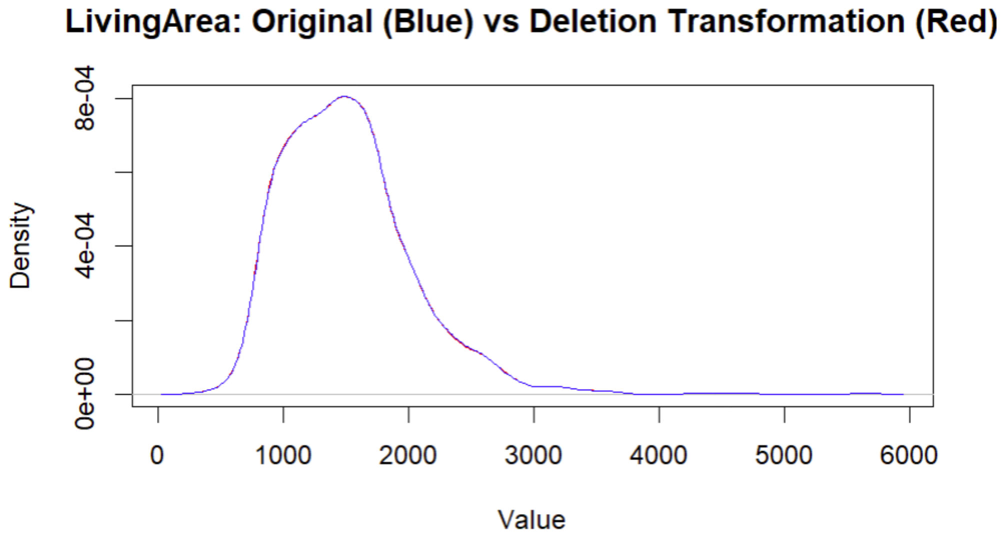
```
Summary statistics of LivingArea_Method 1:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    334    1134    1470    1517    1776    5642
Summary statistics of YearBuilt_Method 1:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1872    1954    1973    1972    2000    2010
```

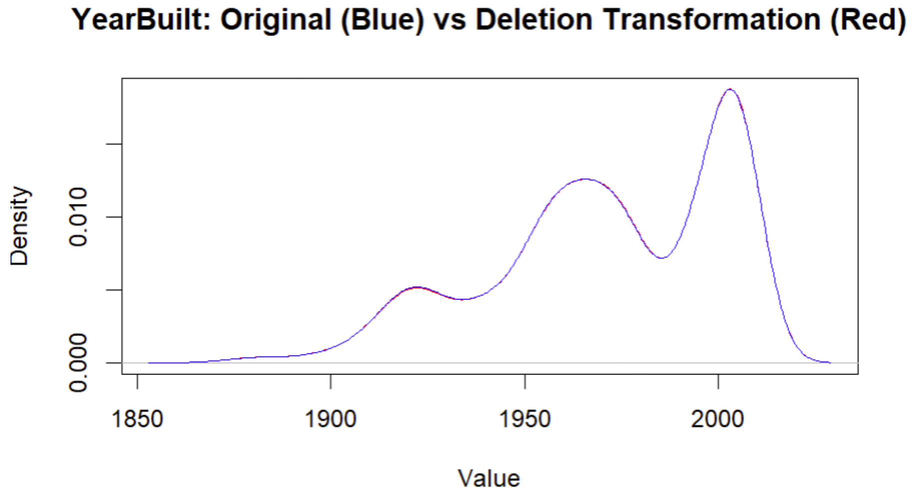




Method 2: deleting all missing values

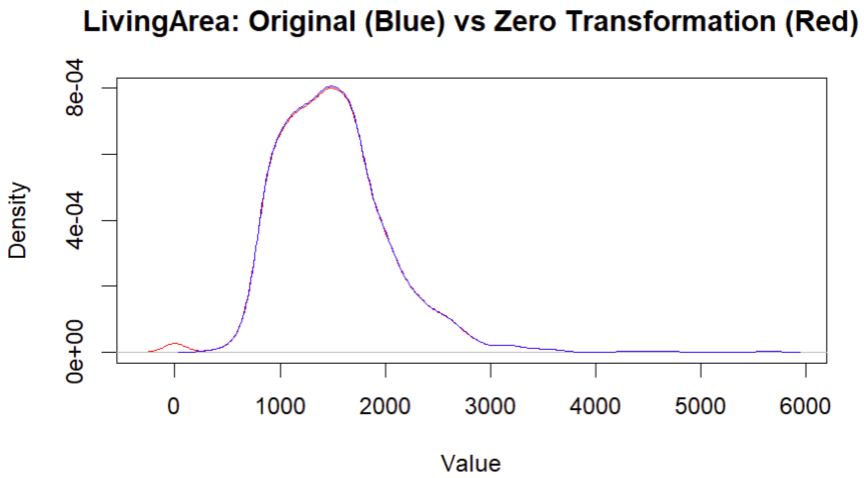
```
Summary statistics of LivingArea_Method 2:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    334    1130    1465    1516    1776    5642
Summary statistics of YearBuilt_Method 2:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1872    1954    1973    1972    2001    2010
```

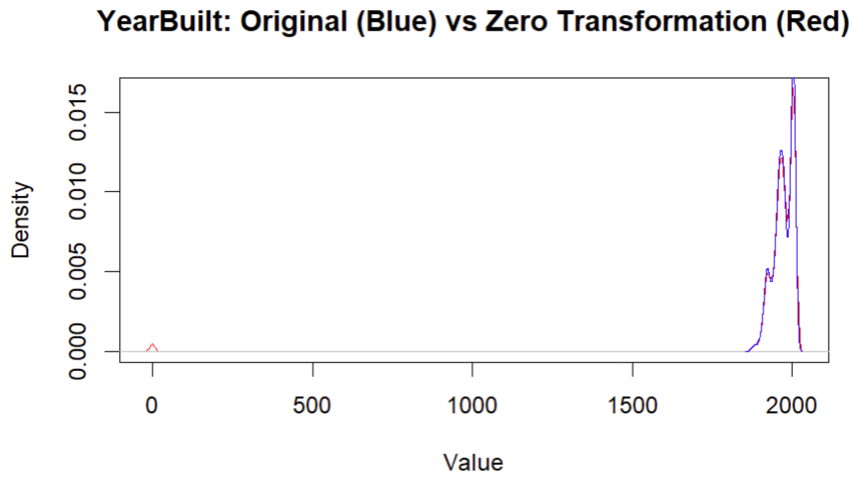




Method 3: replacing missing values with 0

```
Summary statistics of LivingArea_Method 3:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    1126    1458    1507    1776    5642
Summary statistics of YearBuilt_Method 3:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    1954    1972    1954    2000    2010
```

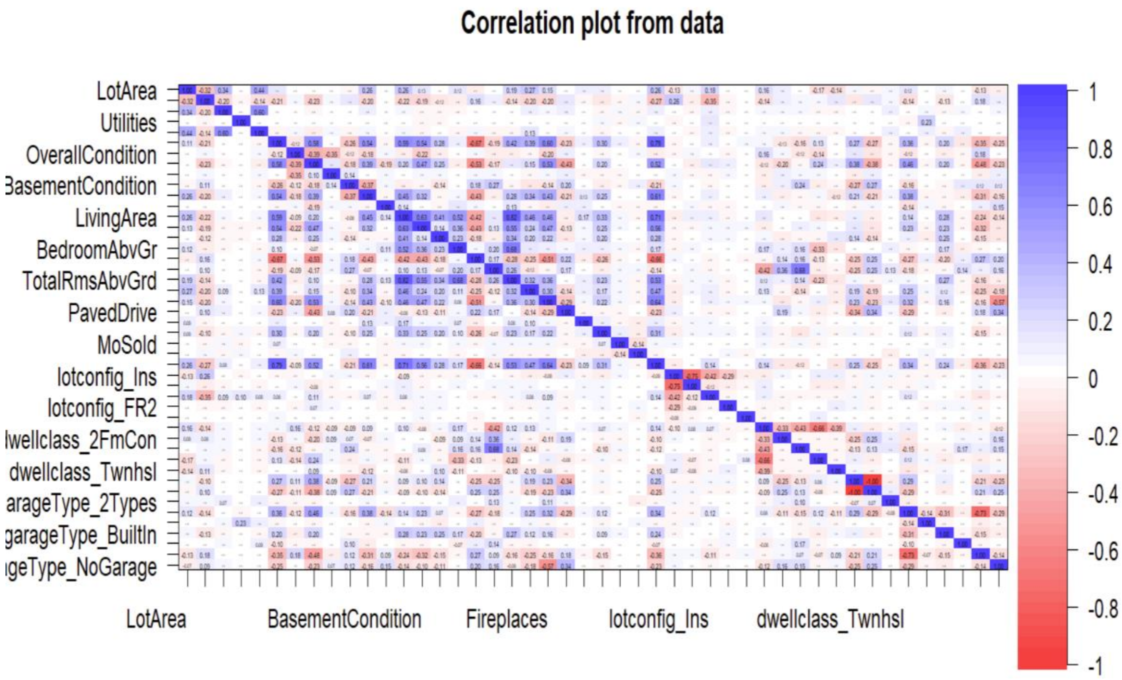


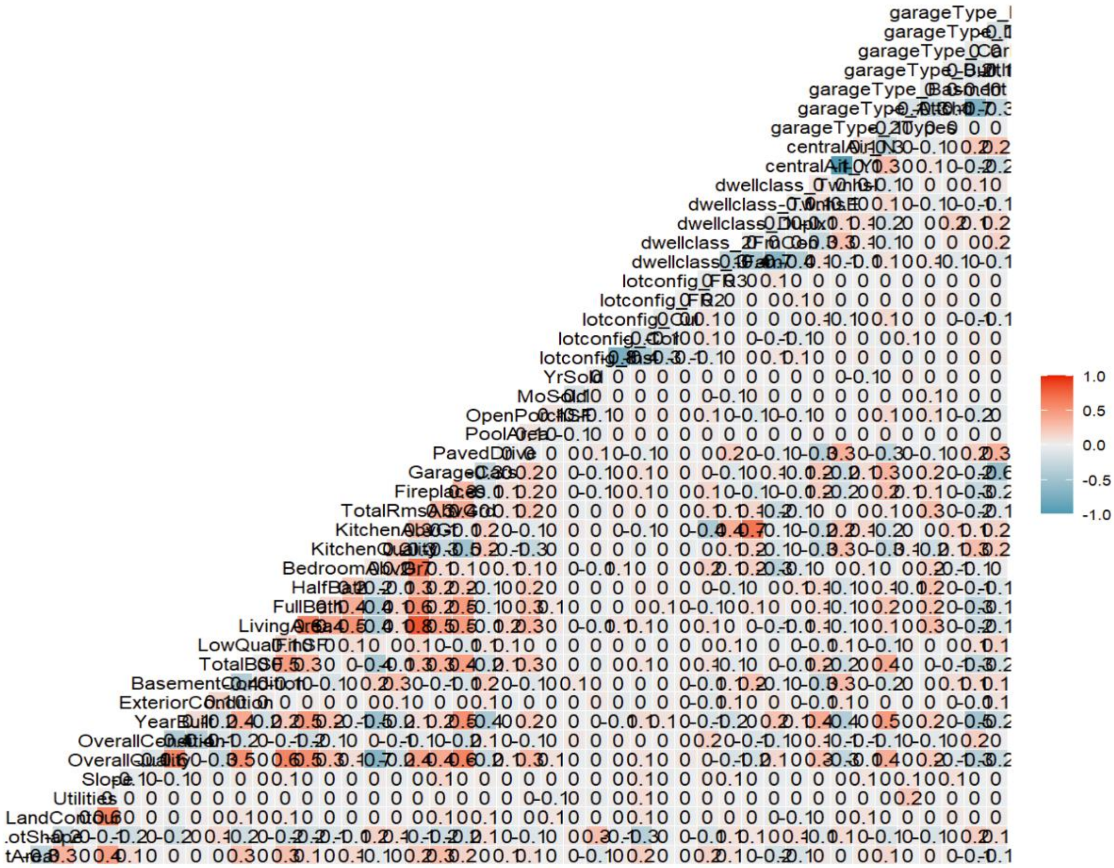


In this case, the most suitable method for handling missing values is method 1, which is to replace missing values in LivingArea with its mean, and YearBuilt with its median. This is because it allows the preservation of the dataset size and by looking at the summary statistics, compared to the original dataset, there is no big difference between both datasets. Most importantly, through the transformation plots, it can be seen that even if method 2 fits the original dataset better, it actually reduces the sample size by around 3%, which may affect the accuracy of the model or analysis power. Compared to method 3, method 1 is more suitable because in the case of YearBuilt, replacing missing values with 0s will distort the logical meaning and representation of a Year variable, as can be shown in the transformed plot.

<u>Multicollinearity detection and dimension reduction:</u>

a. After handling the missing values and determining that method 1 is strongly preferred, the new focal data frame is created, which is called "HousingValuation_mean". The correlation plot between all variables are as follows:





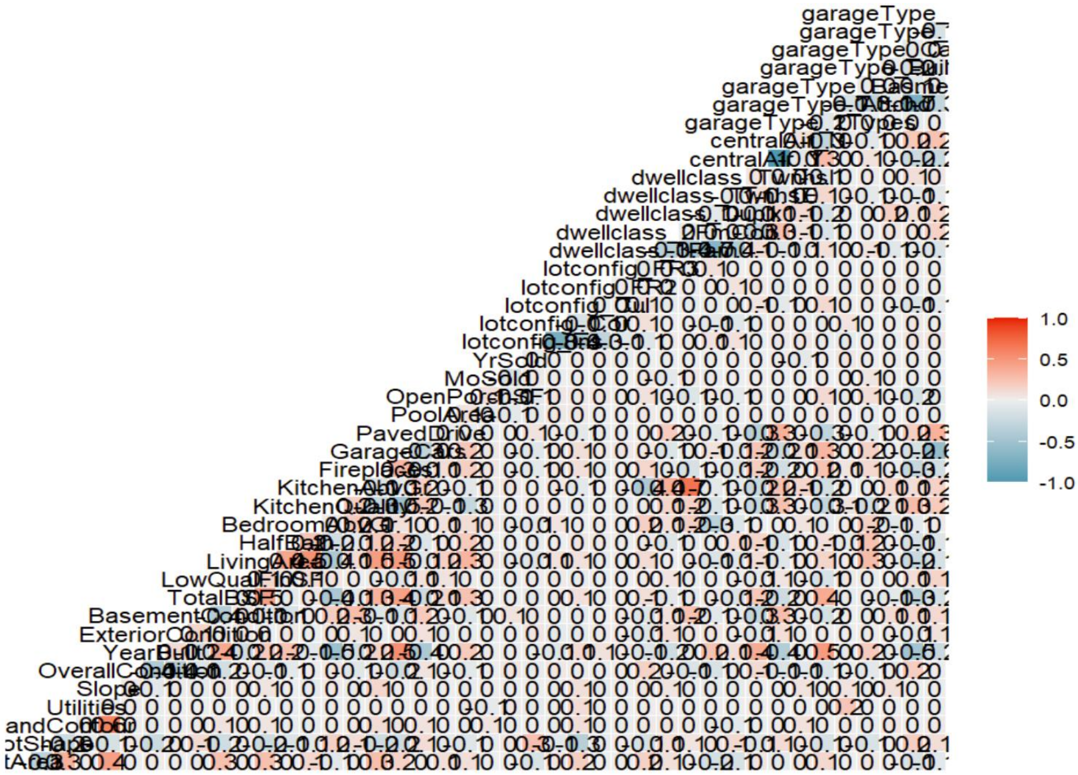
As can be seen, there are some strong correlations amongst independent variables, while some are showing minimal correlation values. For example, OverallQuality is showing strong positive correlation with OverallCondition, ExteriorCondition, TotalBSF, LivingArea, HalfBath, which is expected because the houses with better conditions of these features will likely be having greater quality ratings. Similarly, LivingArea and KitchenAbvGr are also positively correlated, as the houses with larger area of living rooms also expects more spacious kitchen sections. However, it is also worth noting that some strong correlations, regardless of being positive or negative, indicates the potential for model dimension reduction and the issue of overfitting in the predictive model. That being, the presence of highly correlated variables will often require the need to reduce dimension as one variable can represent the correlated ones in some ways to reduce model complexity and avoid overfitting, which consequently increases the performance of the predictive model.

b. Besides the correlation matrix above, R can support in finding the variables with high cross-correlations. Using the cutoff level = 0.5, there are 12 suggested highly correlated variables that can be considered to perform dimension reduction.

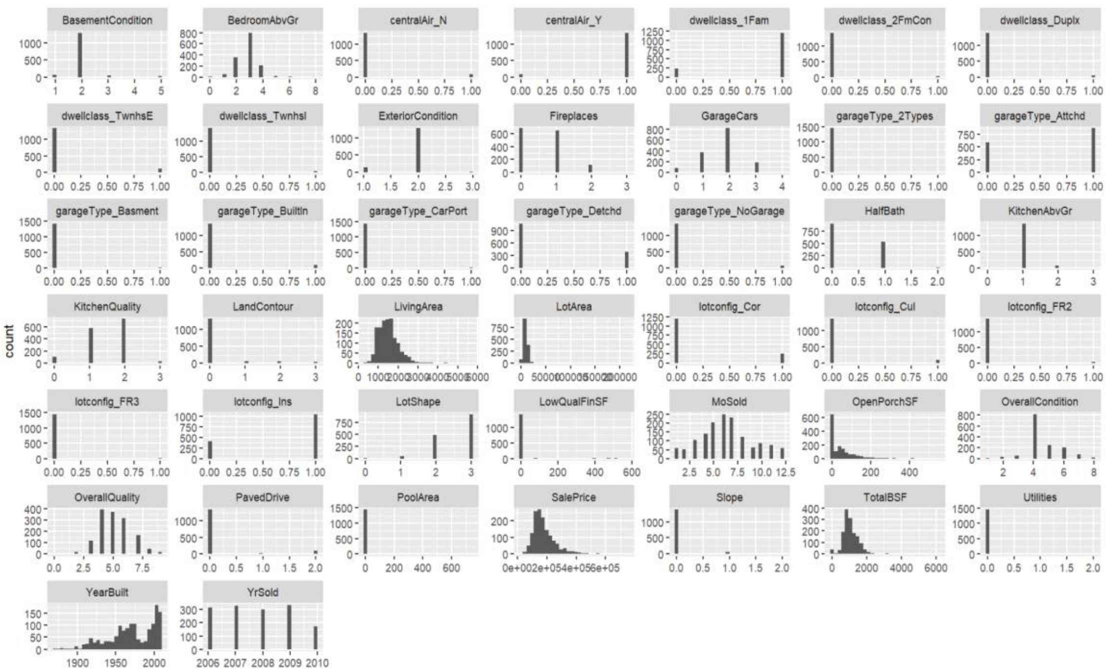
```
 [1] "OverallQuality"     "YearBuilt"          "LivingArea"
 [4] "GarageCars"         "FullBath"           "garageType_Attchd"
 [7] "TotalRmsAbvGrd"     "centralAir_Y"       "KitchenAbvGr"
[10] "dwellclass_1Fam"    "lotconfig_Ins"      "Slope"
```

As mentioned in part a), OverallQuality shows correlations with many different variables, it can be the first one to be removed for model dimension reduction. TotalRmsAbvGrd is highly correlated with LivingArea, which captures the relationship with many different variables as can be shown in the correlation matrix, meaning that removing TotalRmsAbvGrd can help reduce redundancy while preserving the model's preciseness. Finally, FullBath and HalfBath are relatively correlated with different variables, and given that FullBath is suggested by R, it will be removed for dimension reduction as well.

To sum up, OverallQuality, TotalRmsAbvGrd and FullBath are the three variables to be removed from the model and after carrying out the dimension reduction, the new correlation matrix is as follows, which shows that the strong correlations between variables have actually been reduced significantly.

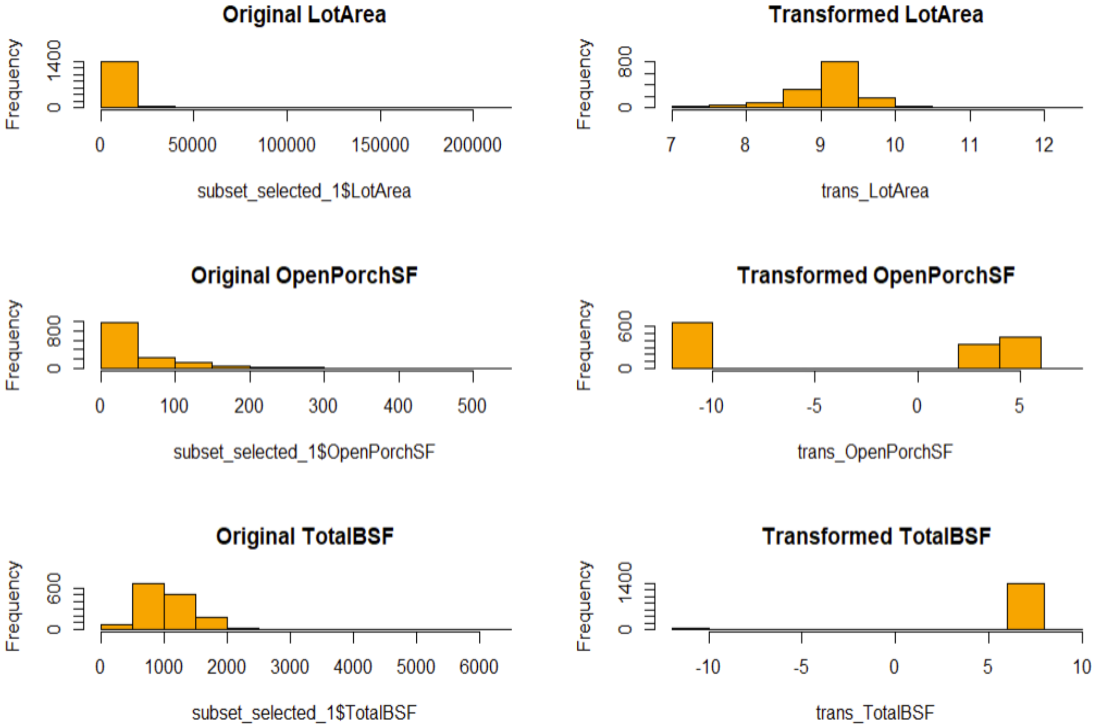


c. Through the use of histograms, the distribution of variables, against the target variable is as follows:



As can be seen, besides the target variable (i.e. SalePrice), independent variables such as LotArea, OpenPorchSF, TotalBSF are heavily right skewed and hence, it is necessary to transform them with logs function before building predictive models.

Since there are some 0s in these 3 columns, the transformation with logs will require the replacement of 0 with a very small number, such as 0.00001 being used in this case. The transformed variables, compared to the original ones, are as below:



PART C: Building predictive models

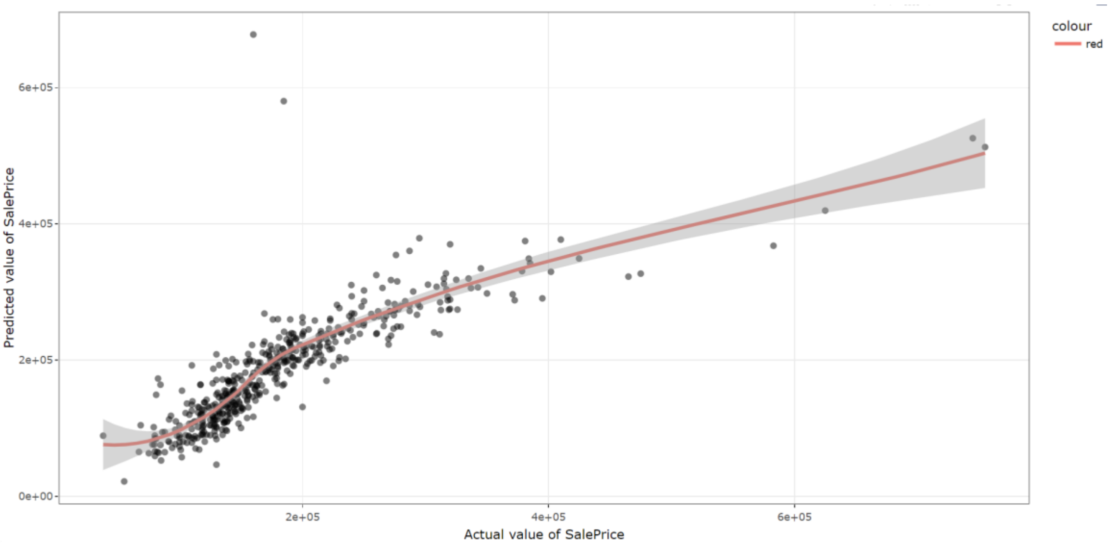
<u>Regression modelling:</u>

a. After performing dimension reduction on OverallQuality, TotalRmsAbvGrd and FullBath, the base model is as follows:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.507907e+06	1.660163e+06	-1.510638571	1.312202e-01
LotArea	1.825458e+04	3.698162e+03	4.936122913	9.441871e-07
LotShape	1.281846e+01	2.159900e+03	0.005934747	9.952661e-01
LandContour	5.283974e+03	2.226878e+03	2.372817623	1.785548e-02
Utilities	-2.167976e+04	1.712394e+04	-1.266049732	2.058123e-01
Slope	-2.552493e+03	5.131542e+03	-0.497412479	6.190158e-01
OverallCondition	5.975988e+03	1.225958e+03	4.874544613	1.281941e-06
YearBuilt	6.683584e+02	6.973816e+01	9.583826505	8.236442e-21
ExteriorCondition	7.612759e+03	3.468871e+03	2.194592853	2.843945e-02
BasementCondition	6.525038e+03	4.036386e+03	1.616554524	1.063136e-01
TotalBSF	2.540979e+03	7.377685e+02	3.444140733	5.984593e-04
LowQualFinSF	-1.164749e+02	2.707813e+01	-4.301439422	1.876303e-05
LivingArea	9.174514e+01	4.025461e+00	22.791213996	1.061992e-91
HalfBath	-1.086927e+04	2.450496e+03	-4.435538026	1.028179e-05
BedroomAbvGr	-1.666401e+04	1.829613e+03	-9.107944680	5.024574e-19
KitchenQuality	-2.166502e+04	2.189136e+03	-9.896608167	5.043846e-22
KitchenAbvGr	-2.529729e+04	8.830488e+03	-2.864767173	4.267111e-03
Fireplaces	7.770987e+03	2.005121e+03	3.875569796	1.138422e-04
GarageCars	1.897172e+04	2.374572e+03	7.989532173	3.991026e-15
PavedDrive	-2.957763e+03	2.681363e+03	-1.103081595	2.702771e-01
PoolArea	1.459235e+01	2.670284e+01	0.546471998	5.848726e-01
OpenPorchSF	-1.685835e+02	1.573547e+02	-1.071359638	2.842857e-01
MoSold	1.418423e+02	4.000999e+02	0.354517332	7.230316e-01
YrSold	5.374430e+02	8.249236e+02	0.651506395	5.148806e-01
lotconfig_Ins	8.784356e+03	1.904124e+04	0.461333171	6.446675e-01
lotconfig_Cor	6.454327e+03	1.916854e+04	0.336714556	7.364081e-01
lotconfig_Cul	9.860152e+03	1.950066e+04	0.505631603	6.132351e-01
lotconfig_FR2	1.815107e+03	1.979931e+04	0.091675280	9.269758e-01
dwellclass_1Fam	9.274131e+03	8.438744e+03	1.098994292	2.720550e-01
dwellclass_2FmCon	1.152824e+04	1.330135e+04	0.866696714	3.863318e-01
dwellclass_Duplx	8.253301e+03	1.274566e+04	0.647538283	5.174434e-01
dwellclass_TwnhsE	3.849956e+03	7.967552e+03	0.483204322	6.290644e-01
centralAir_Y	4.645322e+02	5.185625e+03	0.089580758	9.286397e-01
garageType_2Types	-5.507853e+04	1.690241e+04	-3.258620861	1.160137e-03
garageType_Attchd	-3.213073e+04	6.607507e+03	-4.862761549	1.358671e-06
garageType_Basment	-4.058612e+04	1.075701e+04	-3.772993662	1.715063e-04
garageType_BuiltIn	-2.987504e+04	8.158442e+03	-3.661856118	2.645169e-04
garageType_CarPort	-7.086184e+04	1.400396e+04	-5.060130289	5.048421e-07
garageType_Detchd	-2.983454e+04	6.429552e+03	-4.640221264	3.979364e-06

```
y ~ -2507907.02 + 18254.58 * LotArea + 12.82 * LotShape + 5283.97 *
    LandContour + -21679.76 * Utilities + -2552.49 * Slope +
    5975.99 * OverallCondition + 668.36 * YearBuilt + 7612.76 *
    ExteriorCondition + 6525.04 * BasementCondition + 2540.98 *
    TotalBSF + -116.47 * LowQualFinSF + 91.75 * LivingArea +
    -10869.27 * HalfBath + -16664.01 * BedroomAbvGr + -21665.02 *
    KitchenQuality + -25297.29 * KitchenAbvGr + 7770.99 * Fireplaces +
    18971.72 * GarageCars + -2957.76 * PavedDrive + 14.59 * PoolArea +
    -168.58 * OpenPorchSF + 141.84 * MoSold + 537.44 * YrSold +
    8784.36 * lotconfig_Ins + 6454.33 * lotconfig_Cor + 9860.15 *
    lotconfig_Cul + 1815.11 * lotconfig_FR2 + NA * lotconfig_FR3 +
    9274.13 * dwellclass_1Fam + 11528.24 * dwellclass_2FmCon +
    8253.3 * dwellclass_Duplx + 3849.96 * dwellclass_TwnhsE +
    NA * dwellclass_TwnhsI + 464.53 * centralAir_Y + NA * centralAir_N +
    -55078.53 * garageType_2Types + -32130.73 * garageType_Attchd +
    -40586.12 * garageType_Basment + -29875.04 * garageType_BuiltIn +
    -70861.84 * garageType_CarPort + -29834.54 * garageType_Detchd +
    NA * garageType_NoGarage
```

```
[1] "Actual Values:"
[1] 266500 170000 215000 120000 190000
[1] "Predicted Values:"
[1] 259203.6 182566.5 238800.4 137933.6 179967.0
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'[1] "Root Mean
Square Error:  47615.464914557"
```



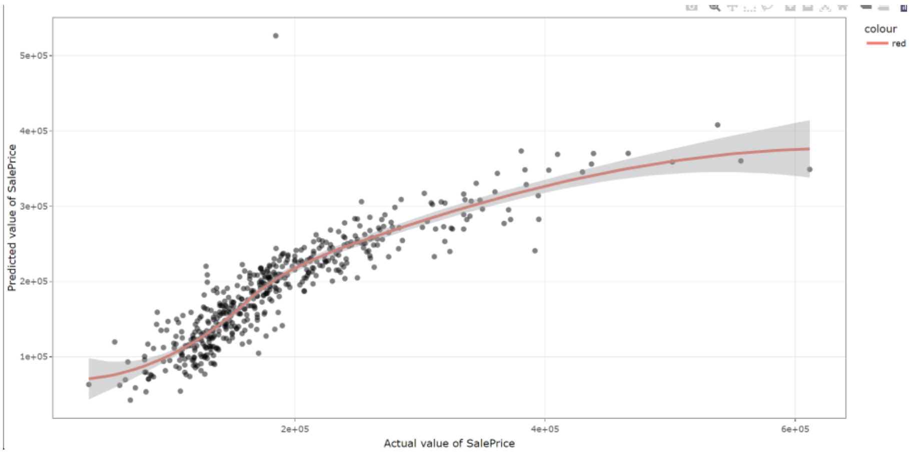
b. Looking at results above, the predicted output, as compared to the actual ones, are reasonably close to each other. However, the RMSE, which provides a measure of the model performance on predicting the actual values, is 47,615. 25. This can be interpreted as the predicted sales price will be deviating from the actual prices by around 47,615 dollars. This is a relatively significant error to have in this predictive model, which might be due to the presence of some outliers of the house prices, as shown the in histogram of SalePrice in question 5c).

In order to build a better regression model, some trials on removing different combinations of variables (i.e. feature selection) are performed.

1st model: This model is created by removing LotArea, OpenPorchSF, TotalBSF

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	400855.00408	1.903431e+06	0.21059608	8.332484e-01
LotShape	1489.35004	2.477109e+03	0.60124534	5.478225e-01
LandContour	3531.26476	2.534045e+03	1.39352886	1.637914e-01
Slope	4524.14107	6.131968e+03	0.73779599	4.608238e-01
OverallCondition	5114.05023	1.393272e+03	3.67053344	2.557607e-04
YearBuilt	599.43383	7.611519e+01	7.87535115	9.405026e-15
ExteriorCondition	1217.42759	3.980671e+03	0.30583475	7.597986e-01
BasementCondition	-7163.86807	2.319435e+03	-3.08862674	2.070286e-03
LowQualFinSF	-38.57055	2.379477e+01	-1.62096761	1.053621e-01
LivingArea	83.23356	4.118088e+00	20.21169989	1.224708e-75
HalfBath	-8135.20039	2.885314e+03	-2.81951973	4.911292e-03
BedroomAbvGr	-11234.37398	2.065011e+03	-5.44034631	6.794197e-08
KitchenQuality	-22352.35840	2.556977e+03	-8.74171174	1.051701e-17
KitchenAbvGr	-26801.05061	8.171501e+03	-3.27981962	1.077149e-03
Fireplaces	7383.39577	2.388514e+03	3.09120851	2.052552e-03
GarageCars	23741.11780	2.713282e+03	8.74996367	9.829986e-18
PavedDrive	-190.25001	2.857440e+03	-0.06658057	9.469299e-01
PoolArea	-47.69450	2.826133e+01	-1.68762411	9.181728e-02
MoSold	178.33257	4.590417e+02	0.38848882	6.977429e-01
YrSold	-742.38724	9.453040e+02	-0.78534235	4.324520e-01
lotconfig_Ins	-3819.13081	1.936922e+04	-0.19717526	8.437333e-01
lotconfig_Cor	-7369.48949	1.959531e+04	-0.37608441	7.069396e-01
lotconfig_Cul	4030.16327	2.007663e+04	0.20073902	8.409464e-01
lotconfig_FR2	-7343.50897	2.068043e+04	-0.35509454	7.225989e-01
dwellclass_1Fam	30495.27912	7.657436e+03	3.98243983	7.351058e-05
dwellclass_2FmCon	39981.66363	1.234990e+04	3.23740782	1.248695e-03
dwellclass_Duplx	31837.52718	1.168249e+04	2.72523482	6.545724e-03
dwellclass_TwnhsE	9840.03026	8.569150e+03	1.14830880	2.511351e-01
centralAir_Y	-1662.72024	5.776267e+03	-0.28785375	7.735225e-01
garageType_2Types	-70477.29564	2.421127e+04	-2.91092906	3.689199e-03
garageType_Attchd	-26443.50297	7.397040e+03	-3.57487644	3.683416e-04
garageType_Basment	-25570.04983	1.358999e+04	-1.88153529	6.020994e-02
garageType_BuiltIn	-28551.07174	9.193566e+03	-3.10554926	1.956558e-03
garageType_CarPort	-44910.19971	1.754100e+04	-2.56029873	1.061451e-02
garageType_Detchd	-28957.70739	7.144370e+03	-4.05322030	5.472192e-05

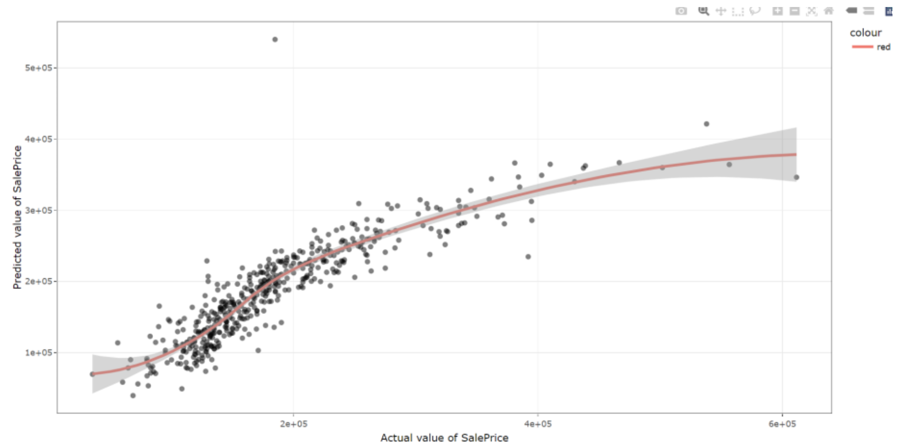
```
doubtful cases[1] "Actual Values:"
[1] 255900 239500 180000 132000 202900
[1] "Predicted Values:"
[1] 262394.9 248457.2 176787.5 113283.1 225103.6
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'[1] "Root
Mean Square Error:  37979.6495164468"
```



2nd model: <u>This model is created by removing OpenPorchSF, TotalBSF</u>

```
                       Estimate   Std. Error     t value     Pr(>|t|)
(Intercept)         240495.93268 1.879754e+06  0.12794014 8.982239e-01
LotArea              19964.76492 3.996432e+03  4.99564757 6.999062e-07
LotShape              4287.25941 2.509242e+03  1.70858755 8.785997e-02
LandContour           2312.29808 2.514028e+03  0.91975825 3.579368e-01
Slope                 -892.55118 6.151128e+03 -0.14510367 8.846604e-01
OverallCondition      5691.88689 1.380594e+03  4.12278144 4.076696e-05
YearBuilt              603.50446 7.516184e+01  8.02940006 2.937129e-15
ExteriorCondition     1974.54975 3.933503e+03  0.50198255 6.157982e-01
BasementCondition    -7324.48360 2.290475e+03 -3.19780166 1.431429e-03
LowQualFinSF           -40.20220 2.349763e+01 -1.71090430 8.743118e-02
LivingArea              79.06028 4.151194e+00 19.04518902 1.430926e-68
HalfBath             -6481.88413 2.868166e+03 -2.25994052 2.405482e-02
BedroomAbvGr        -10824.29264 2.040678e+03 -5.30426235 1.413330e-07
KitchenQuality      -22746.13516 2.526033e+03 -9.00468766 1.192559e-18
KitchenAbvGr        -22435.00650 8.115873e+03 -2.76433683 5.816278e-03
Fireplaces            6071.36408 2.373038e+03  2.55847780 1.067004e-02
GarageCars           22657.88516 2.687901e+03  8.42958407 1.304327e-16
PavedDrive           -1234.09146 2.829211e+03 -0.43619627 6.627952e-01
PoolArea               -54.34919 2.793749e+01 -1.94538546 5.202866e-02
MoSold                 268.05053 4.536212e+02  0.59091275 5.547220e-01
YrSold                -748.23039 9.334098e+02 -0.80160973 4.229829e-01
lotconfig_Ins          381.63833 1.914397e+04  0.01993517 9.840993e-01
lotconfig_Cor        -3781.61191 1.936206e+04 -0.19531039 8.451925e-01
lotconfig_Cul         5869.83469 1.982742e+04  0.29604624 7.672606e-01
lotconfig_FR2        -6413.16439 2.042106e+04 -0.31404662 7.535558e-01
dwellclass_1Fam       4659.54393 9.160563e+03  0.50865259 6.111160e-01
dwellclass_2FmCon     7540.75824 1.381578e+04  0.54580774 5.853286e-01
dwellclass_Duplx       886.67391 1.309399e+04  0.06771612 9.460261e-01
dwellclass_TwnhsE     1328.89025 8.631143e+03  0.15396458 8.776710e-01
centralAir_Y         -3353.36497 5.713615e+03 -0.58690773 5.574077e-01
garageType_2Types   -77391.42370 2.394665e+04 -3.23182699 1.273107e-03
garageType_Attchd   -31249.43645 7.367045e+03 -4.24178723 2.438494e-05
garageType_Basment  -27931.23546 1.342731e+04 -2.08018120 3.778136e-02
garageType_BuiltIn  -31436.89965 9.096243e+03 -3.45603104 5.729125e-04
garageType_CarPort  -47060.93246 1.732563e+04 -2.71626111 6.724504e-03
garageType_Detchd   -30939.91895 7.065622e+03 -4.37893772 1.327511e-05
```

```
doubtful cases[1] "Actual Values:"
[1] 255900 239500 180000 132000 202900
[1] "Predicted Values:"
[1] 262518.8 244470.2 171881.8 123637.9 233281.8
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'[1] "Root
Mean Square Error:  37883.7424403385"
```

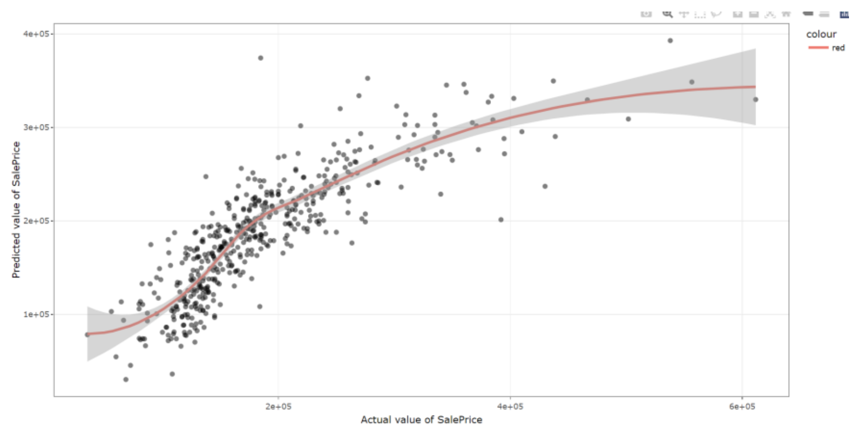


3rd model: This model is created by removing LivingArea, YearBuilt

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.911131e+05	2.226652e+06	0.31038226	7.563396e-01
LotArea	3.383253e+04	4.645075e+03	7.28352815	6.912745e-13
LotShape	1.897274e+03	2.958330e+03	0.64133258	5.214641e-01
LandContour	8.710250e+02	2.988756e+03	0.29143400	7.707842e-01
Slope	1.658573e+03	7.311242e+03	0.22685249	8.205881e-01
OverallCondition	1.992031e+03	1.506365e+03	1.32240904	1.863561e-01
ExteriorCondition	1.822469e+03	4.687888e+03	0.38876124	6.975415e-01
BasementCondition	1.545626e+03	5.538617e+03	0.27906346	7.802580e-01
TotalBSF	1.959364e+03	9.980924e+02	1.96310927	4.993021e-02
LowQualFinSF	8.998827e+00	2.703594e+01	0.33284678	7.393247e-01
HalfBath	9.869260e+03	3.270003e+03	3.01811942	2.612601e-03
BedroomAbvGr	7.322850e+03	2.120876e+03	3.45274871	5.798389e-04
KitchenQuality	-3.947954e+04	2.778293e+03	-14.20999803	1.251492e-41
KitchenAbvGr	2.244826e+03	9.348906e+03	0.24011649	8.102927e-01
Fireplaces	1.866561e+04	2.617912e+03	7.12995864	2.010912e-12
GarageCars	4.074153e+04	2.942731e+03	13.84480570	8.788066e-40
PavedDrive	-3.947503e+03	3.252135e+03	-1.21381887	2.251241e-01
PoolArea	4.080765e+01	3.252829e+01	1.25452801	2.099643e-01
OpenPorchSF	6.126423e+02	2.129094e+02	2.87747887	4.100090e-03
MoSold	-3.048399e+01	5.384425e+02	-0.05661512	9.548639e-01
YrSold	-4.176325e+02	1.110193e+03	-0.37618002	7.068686e-01
lotconfig_Ins	5.921044e+03	2.275991e+04	0.26015238	7.948036e-01
lotconfig_Cor	2.877616e+03	2.301232e+04	0.12504672	9.005135e-01
lotconfig_Cul	1.065527e+04	2.358034e+04	0.45187085	6.514671e-01
lotconfig_FR2	-8.188804e+03	2.426490e+04	-0.33747523	7.358346e-01
dwellclass_1Fam	-2.373599e+04	1.071156e+04	-2.21592297	2.693745e-02
dwellclass_2FmCon	-2.732567e+04	1.625147e+04	-1.68142832	9.301442e-02
dwellclass_Duplx	-2.931494e+04	1.544220e+04	-1.89836491	5.795639e-02
dwellclass_TwnhsE	-8.269066e+03	1.025783e+04	-0.80612249	4.203776e-01
centralAir_Y	8.280970e+02	6.564528e+03	0.12614723	8.996426e-01
garageType_2Types	-1.359166e+05	2.820493e+04	-4.81889664	1.684274e-06
garageType_Attchd	-4.748288e+04	8.705410e+03	-5.45441098	6.294775e-08
garageType_Basment	-3.761241e+04	1.592412e+04	-2.36197772	1.838227e-02
garageType_BuiltIn	-3.362157e+04	1.079951e+04	-3.11324948	1.906795e-03
garageType_CarPort	-6.843765e+04	2.049117e+04	-3.33986124	8.714229e-04
garageType_Detchd	-5.768983e+04	8.125063e+03	-7.10023204	2.467049e-12

```
y ~ 691113.14 + 33832.53 * LotArea + 1897.27 * LotShape + 871.02 *
```

```
doubtful cases[1] "Actual Values:"
[1] 255900 239500 180000 132000 202900
[1] "Predicted Values:"
[1] 229623.8 228570.2 182304.2 118917.8 220752.3
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'[1] "Root Mean
Square Error:  43458.2757979429"
```



c. As can be seen, after the dimension reduction of OverallQuality, TotalRmsAbvGrd and FullBath, the optimal model is model 2 with selecting to remove OpenPorchSF, TotalBSF after because it is having the smallest RMSE of 37,883.74. In fact, there have been lots of trials being undertaken and those 3 models are the most representative ones to showcase the process of variable removal and testing. The formulas for each regression model are as below:

1st model: This model is created by removing LotArea, OpenPorchSF, TotalBSF

```
y ~ 400855 + 1489.35 * LotShape + 3531.26 * LandContour + NA *
    Utilities + 4524.14 * Slope + 5114.05 * OverallCondition +
    599.43 * YearBuilt + 1217.43 * ExteriorCondition + -7163.87 *
    BasementCondition + -38.57 * LowQualFinSF + 83.23 * LivingArea +
    -8135.2 * HalfBath + -11234.37 * BedroomAbvGr + -22352.36 *
    KitchenQuality + -26801.05 * KitchenAbvGr + 7383.4 * Fireplaces +
    23741.12 * GarageCars + -190.25 * PavedDrive + -47.69 * PoolArea +
    178.33 * MoSold + -742.39 * YrSold + -3819.13 * lotconfig_Ins +
    -7369.49 * lotconfig_Cor + 4030.16 * lotconfig_Cul + -7343.51 *
    lotconfig_FR2 + NA * lotconfig_FR3 + 30495.28 * dwellclass_1Fam +
    39981.66 * dwellclass_2FmCon + 31837.53 * dwellclass_Duplx +
    9840.03 * dwellclass_TwnhsE + NA * dwellclass_TwnhsI + -1662.72 *
    centralAir_Y + NA * centralAir_N + -70477.3 * garageType_2Types +
    -26443.5 * garageType_Attchd + -25570.05 * garageType_Basment +
    -28551.07 * garageType_BuiltIn + -44910.2 * garageType_CarPort +
    -28957.71 * garageType_Detchd + NA * garageType_NoGarage
```

2nd model: This model is created by removing OpenPorchSF, TotalBSF

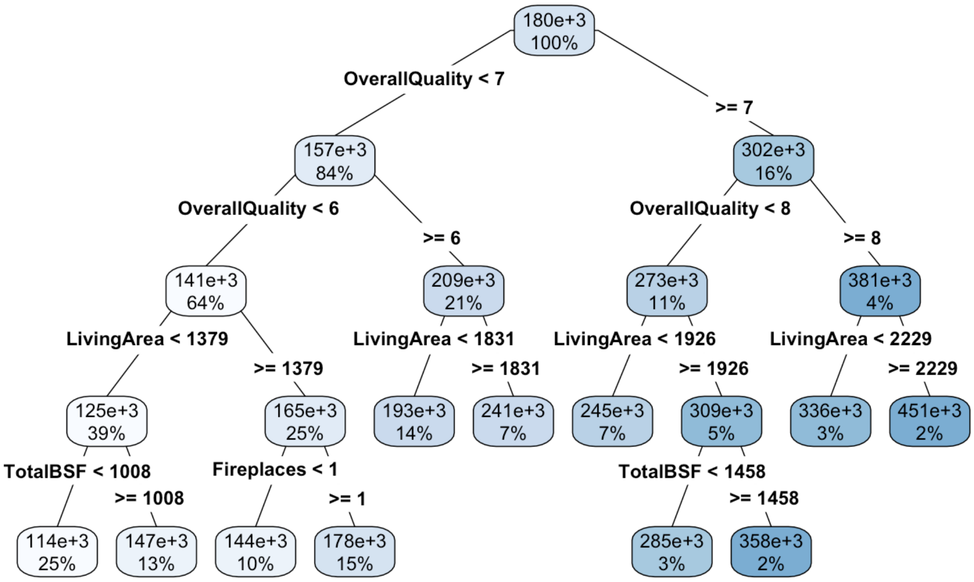
```
y ~ 240495.93 + 19964.76 * LotArea + 4287.26 * LotShape + 2312.3 *
    LandContour + NA * Utilities + -892.55 * Slope + 5691.89 *
    OverallCondition + 603.5 * YearBuilt + 1974.55 * ExteriorCondition
+
    -7324.48 * BasementCondition + -40.2 * LowQualFinSF + 79.06 *
    LivingArea + -6481.88 * HalfBath + -10824.29 * BedroomAbvGr +
    -22746.14 * KitchenQuality + -22435.01 * KitchenAbvGr + 6071.36 *
    Fireplaces + 22657.89 * GarageCars + -1234.09 * PavedDrive +
    -54.35 * PoolArea + 268.05 * MoSold + -748.23 * YrSold +
    381.64 * lotconfig_Ins + -3781.61 * lotconfig_Cor + 5869.83 *
    lotconfig_Cul + -6413.16 * lotconfig_FR2 + NA * lotconfig_FR3 +
    4659.54 * dwellclass_1Fam + 7540.76 * dwellclass_2FmCon +
    886.67 * dwellclass_Duplx + 1328.89 * dwellclass_TwnhsE +
    NA * dwellclass_TwnhsI + -3353.36 * centralAir_Y + NA *
centralAir_N +
    -77391.42 * garageType_2Types + -31249.44 * garageType_Attchd +
    -27931.24 * garageType_Basment + -31436.9 * garageType_BuiltIn +
    -47060.93 * garageType_CarPort + -30939.92 * garageType_Detchd +
    NA * garageType_NoGarage
```

3rd model: This model is created by removing LivingArea, YearBuilt

```
y ~ 691113.14 + 33832.53 * LotArea + 1897.27 * LotShape + 871.02 *
    LandContour + NA * Utilities + 1658.57 * Slope + 1992.03 *
    OverallCondition + 1822.47 * ExteriorCondition + 1545.63 *
    BasementCondition + 1959.36 * TotalBSF + 9 * LowQualFinSF +
    9869.26 * HalfBath + 7322.85 * BedroomAbvGr + -39479.54 *
    KitchenQuality + 2244.83 * KitchenAbvGr + 18665.61 * Fireplaces +
    40741.53 * GarageCars + -3947.5 * PavedDrive + 40.81 * PoolArea +
    612.64 * OpenPorchSF + -30.48 * MoSold + -417.63 * YrSold +
    5921.04 * lotconfig_Ins + 2877.62 * lotconfig_Cor + 10655.27 *
    lotconfig_Cul + -8188.8 * lotconfig_FR2 + NA * lotconfig_FR3 +
    -23735.99 * dwellclass_1Fam + -27325.67 * dwellclass_2FmCon +
    -29314.94 * dwellclass_Duplx + -8269.07 * dwellclass_TwnhsE +
    NA * dwellclass_TwnhsI + 828.1 * centralAir_Y + NA * centralAir_N +
    -135916.64 * garageType_2Types + -47482.88 * garageType_Attchd +
    -37612.41 * garageType_Basment + -33621.57 * garageType_BuiltIn +
    -68437.65 * garageType_CarPort + -57689.83 * garageType_Detchd +
    NA * garageType_NoGarage
```

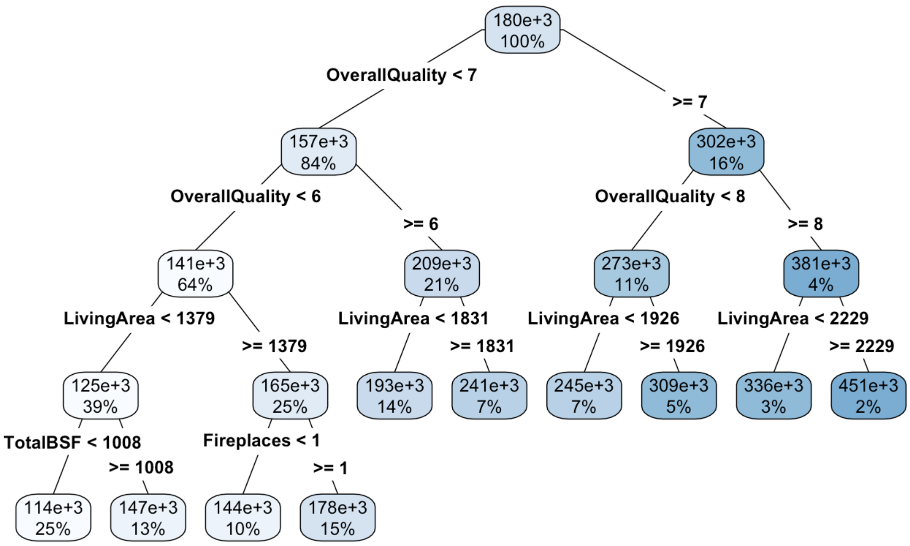
Decision tree modelling:

a. Using the dataset with dimension reduction of the decision tree built on the selected variables are as follows:

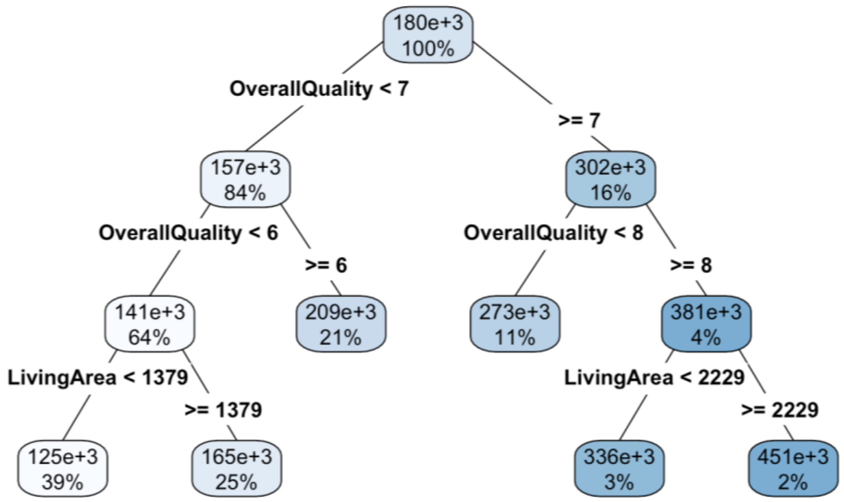


b. As can be seen, the root node starts with OverallQuality being less than 7, and the subsequent nodes are based on OverallQuality, LivingArea, TotalBSF, and Fireplace. The first split 84% of the data into OverallQuality < 7, and 16% of the data belongs to OverallQuality >= 7. The tree further splits are based on OverallQuality, LivingArea, TotalBSF and Fireplaces. The decision rule is determined by each leaf node to predict the SalePrice. For example, OverallQuality < 7, OverallQuality < 6, LivingArea < 1379, and TotalBSF < 1008 will lead to the predicted value of 125e+3 accounting for 25% of the data given.

Using pruning technique with setting cp = 0.011628, the second model is as below:



Using pruning technique with setting cp = 0.02, the third model is as below:



c. The tree plots have been provided in part b).

```
 1) root 969 5.743202e+12 179906.6
   2) OverallQuality< 6.5 817 1.945617e+12 157228.7
     4) OverallQuality< 5.5 618 8.992050e+11 140682.7
       8) LivingArea< 1378.5 377 3.198739e+11 125199.6
        16) TotalBSF< 1007.5 247 1.597443e+11 113650.3 *
        17) TotalBSF>=1007.5 130 6.458598e+10 147143.1 *
       9) LivingArea>=1378.5 241 3.475754e+11 164903.2
        18) Fireplaces< 0.5 93 1.015760e+11 143833.3 *
        19) Fireplaces>=0.5 148 1.787695e+11 178143.1 *
     5) OverallQuality>=5.5 199 3.517960e+11 208612.9
      10) LivingArea< 1831 133 1.268920e+11 192698.4 *
      11) LivingArea>=1831 66 1.233383e+11 240683.0 *
   3) OverallQuality>=6.5 152 1.118996e+12 301800.0
     6) OverallQuality< 7.5 111 4.900402e+11 272519.1
      12) LivingArea< 1925.5 63 1.555895e+11 244679.3 *
      13) LivingArea>=1925.5 48 2.215345e+11 309058.9
        26) TotalBSF< 1457.5 32 7.738354e+10 284593.6 *
        27) TotalBSF>=1457.5 16 8.668983e+10 357989.6 *
     7) OverallQuality>=7.5 41 2.761363e+11 381072.8
      14) LivingArea< 2229 25 5.514901e+10 336413.9 *
      15) LivingArea>=2229 16 9.321991e+10 450852.4 *
[1] "Actual Values"
[1] 266500 170000 215000 120000 190000
[1] "Predicted Values"
    1452      202      171     1272      128
240683.0 192698.4 244679.3 147143.1 178143.1
[1] "Root Mean Square Error (Unpruned):  44860.3541194601"

Regression tree:
rpart(formula = formula, data = tree_selected.train, method = "anova")

Variables actually used in tree construction:
[1] Fireplaces     LivingArea     OverallQuality TotalBSF
```

```
Regression tree:
rpart(formula = formula, data = tree_selected.train, method = "anova")

Variables actually used in tree construction:
[1] Fireplaces     LivingArea     OverallQuality TotalBSF

Root node error: 5.7432e+12/969 = 5926937067

n= 969

         CP nsplit rel error  xerror     xstd
1  0.466393      0   1.00000 1.00239 0.076710
2  0.120946      1   0.53361 0.57074 0.042100
3  0.061433      2   0.41266 0.41809 0.029252
4  0.040353      3   0.35123 0.38702 0.026552
5  0.022247      4   0.31088 0.33820 0.026759
6  0.019661      5   0.28863 0.32575 0.025203
7  0.017685      6   0.26897 0.32552 0.024596
8  0.016636      7   0.25128 0.32332 0.024581
9  0.011706      8   0.23465 0.28779 0.022700
10 0.010005      9   0.22294 0.28532 0.022948
11 0.010000     10   0.21294 0.27828 0.022003
[1] "Best CP:  0.01"
[1] "Root Mean Square Error (Pruned 0.011628):  45363.2074684153"
[1] "Root Mean Square Error (Pruned 0.02):  48969.6995481871"
```

With model 1, the root node starts with OverallQuality and it lasts for the next layer, which indicates that this is a very important variable in the model for predicting SalePrice. LivingArea is the second-ranked one in this model.
With model 2, the overall tree structure is similar to that of model 1. However, it is worth noting that the RMSE is slightly higher than that of model 1, indicating that it might be giving a less precise prediction. Based on the performance of model 3, the RMSE, which is the largest as compared to the other two models, implying that this is not the well-fitted prediction model. This can be explained because model 3 only takes into account the OverallQuality and LivingArea as the main variables of interests.

Model comparison:

a. In building predictive models, it is necessary to build several models for comparisons in both regression and decision trees because:
    - Since different predictive model might learn and perform better when being given particular datasets, it is necessary to try on different models so that the performance could be precisely compared and evaluated against certain metrics, for example in the case of regression model above, RMSE is used as an evaluation metrics to select the most optimal model.
    - Performing different predictive models can also help to identify the significance of certain variables in the models, because feature selection can be performed, and the evaluation metrics can be used to understand the impact of removing or adding new variables to the models.

- Overfitting or underfitting problems can be reduced significantly with several models being built. This is because there are different models to compare, rather than solely being dependent on one particular model. For example, in the case of decision trees above, an overly deep decision tree might be overfitting with the data patterns, meaning that the comparison with other 2 models are giving a better outlook on the tree performance.

b. As shown above, in terms of model accuracy:
    - The selected optimal regression model is model 2 with removing OpenPorchSF, TotalBSF with RMSE of 37,883.74, which is shown to give the most accurate fit on the price prediction
    - The optimal decision tree is model 1 with the RMSE of 44,860.35, which is the prediction model with the smallest possible errors.

In deciding the most suitable model for price predictions, based on RMSE as the key performance metrics, it can be concluded that in this particular dataset, using the regression model might be better, compared to decision tree, in terms of producing the outcome with the smaller errors. However, besides RMSE as a metric, there might be context-relevant factors to be considered. For example, in most business cases, decision trees have been widely known for understandability and easy interpretation

REFERENCES

Consumer Affairs Victoria, V. G. (n.d.). *Property data*. Www.consumer.vic.gov.au. https://www.consumer.vic.gov.au/housing/buying-and-selling-property/property-data

Gomez, J. (2019, March 27). *8 critical factors that influence a home's value | Opendoor*. Opendoor. https://www.opendoor.com/w/blog/factors-that-influence-home-value

Leung, W. (2024, May 13). *The Top Factors Affecting House Prices and How to Navigate Them*. TheOwnTeam.com. https://www.theownteam.com/blog/the-top-factors-affecting-house-prices-and-how-to-navigate-them/

Martin, E. J. (n.d.). *How much is my house worth?* Bankrate. https://www.bankrate.com/real-estate/how-much-is-my-house-worth/

Rohde, J. (n.d.). *Rental property valuation: 5 ways to value your property*. Www.stessa.com. https://www.stessa.com/blog/rental-property-valuation/