Table of Contents

Exploratory Data Analysis

Numerical -Discrete	Numerical -Continuous	Nominal
Id	LotArea	Utilities
YearBuilt	TotalBSF	LotConfig
FullBath	LowQualFinSF	DwellClass
HalfBath	LivingArea	CentralAir
BedroomAbvGr	PoolArea	GarageType
KitchenAbvGr	OpenPorchSF	LandContour
TotalRmsAbvGrd	SalePrice	PavedDrive
Fireplaces		
GarageCars		
YrSold		

Ordinal	
LotShape	- Although the data for LotShape is given in numerical form, the values represent categories with a meaningful order from Regular to Irregular. The order indicates a progression of irregularity, making it an ordinal variable
Slope	- because slope has been ordered from gentle to severe
OverallQuality	- because it has a rating scale from poor to excellent
OverallCondition	- because it has a rating scale from poor to excellent
ExteriorCondition	- because it has a rating scale from poor to excellent
BasementCondition	- because it has a rating scale from poor to excellent
KitchenQuality	- because it has a rating scale from poor to excellent
MoSold	- While the data is stored numerically, its nature is ordinal, reflecting

Please refer R Script

	Variable <chr>	Mean <dbl>	Median <dbl>	Max <int>	SD <dbl>
LotArea	LotArea	1.052113e+04	9478.5	215245	10000.46368
TotalBSF	TotalBSF	1.058357e+03	992.0	6110	439.17440
LowQualFinSF	LowQualFinSF	5.868638e+00	0.0	572	48.72192
LivingArea	LivingArea	1.517197e+03	1466.0	5642	525.46729
PoolArea	PoolArea	2.770289e+00	0.0	738	40.25978
OpenPorchSF	OpenPorchSF	4.637001e+01	25.0	547	65.13858
SalePrice	SalePrice	1.811117e+05	163250.0	755000	79331.69323

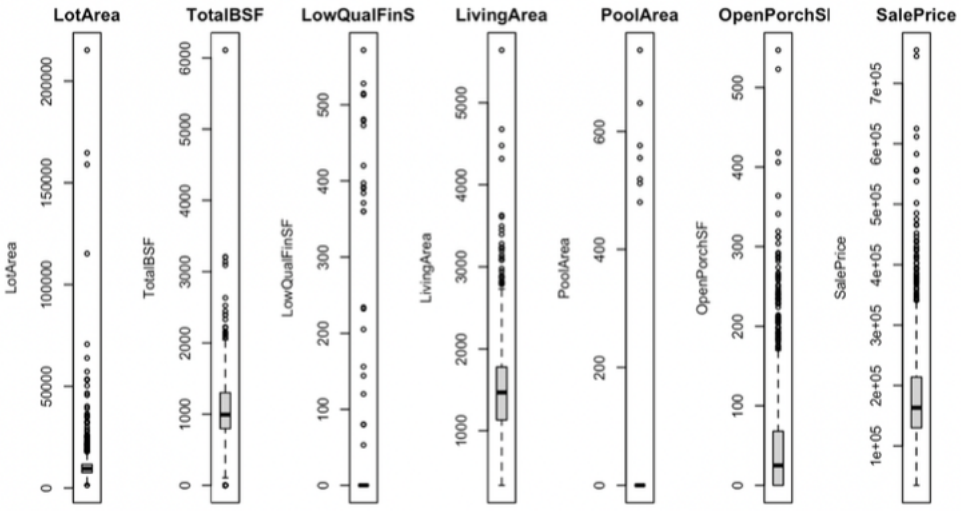
```
  45   65   37 1307
[1] "CentralAir"

   N    Y
  94 1360
[1] "KitchenQuality"

 Ex  Fa  Gd  TA
100  37 584 733
[1] "GarageType"

 2Types  Attchd Basment BuiltIn CarPort  Detchd
      6     870      19      88       9     384
[1] "PavedDrive"

   N    P    Y
  88   30 1336
[1] "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
[1] "YrSold"

2006 2007 2008 2009 2010
 314  328  302  336  174
```

Please refer R Script
In order to identify the evidence of extreme values , also called the outliers in the dataset, I used the following methods:

1. Boxplots for Visualizing Outliers:

- **Visualization:** Visual inspection of each continuous variable for possible outliers may be done with boxplots. Outliers are identified as points that fall outside of the plot's "whiskers" in a boxplot, which shows the distribution of data points.
- **Interpretation:** In each boxplot:

o The range between the first (Q1) and third quartiles (Q3) is known as the interquartile range (IQR), and it is shown by the box.
o The "whiskers" are 1.5 times the IQR from the box; any data points beyond it are regarded as possible outliers.



2. IQR-Based Outlier Detection:

• **IQR calculation:** The diNerence between Q3, the 75th percentile, and Q1, the 25th percentile, was used to get the IQR for each continuous variable.
• **Defining Outliers:** Data points that fall outside of Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are considered outliers. Extreme values are frequently identified using these criteria.
• **Outliers Summary:** The number of outliers was found and tallied by using this procedure on each continuous variable. Each variable is listed with the number of outliers found in the summary table (outliers_summary)

	Variable <chr>	Num_Outliers <int>
LotArea	LotArea	68
TotalBSF	TotalBSF	61
LowQualFinSF	LowQualFinSF	26
LivingArea	LivingArea	40
PoolArea	PoolArea	7
OpenPorchSF	OpenPorchSF	75
SalePrice	SalePrice	61

Brief Discussion on Extreme Values

Both the boxplots and the IQR-based technique provide evidence of extreme values, which implies that certain variables in the dataset contain observations that greatly depart from the data's central tendency.
• **Impact on Analysis**: Predictive modelling and statistical analysis can be significantly impacted by outliers. They could enhance standard deviations, aNect mean values, and yield false conclusions about how variables relate to one another.

Variability in a variable is often measured by its standard deviation (SD). A higher SD indicates more variability. From the summary statistics, the variable(s) with the highest SD can be identified. Accordingly, **SalePrice** has the largest standard deviation (79,331.69), indicating a wide range in property prices within the dataset. **LotArea** also shows substantial variability with a standard deviation of 10,000.46, suggesting a significant diversity in lot sizes.

Extreme values or outliers are data points that fall far outside the range of most other values. These can often be identified in histograms as isolated bars far from the main distribution. Reviewing the histograms for any bars that are far from the bulk of the data and checking the summary statistics for maximum values that are significantly higher than the median or mean, which may indicate the presence of outliers.

- **TotalBSF:** The maximum value (6,110) stands out as an outlier, particularly when compared to the mean (1,058.36) and median (992.0). This suggests that some properties have unusually large total finished basement areas.
- **LotArea:** The maximum value (215,245) is exceptionally high compared to the mean (10,521.13) and median (9,478.5), indicating the presence of significant outliers.
- **OpenPorchSF:** The maximum value (547) is much higher than the mean (46.37) and median (25.0), suggesting some properties have unusually large open porches.
- **PoolArea:** The maximum value (738) is notably larger than the median (0), indicating an extreme outlier, as most properties do not have pools.

Imputation is done in order to handle the missing values in a dataset. Here, the missing values will be replaced with some reasonable value and this can be done in three ways;

1. Fillinthemissingvalueswiththecompletecases'meanfornumericalvariables and with the mode for categorical variables.
2. Deleterecordswithmissingvalues
3. Replacemissingvalueswithaspecificvaluesuchas0

Please refer R Script

In order to check the most suitable method to handle missing values, the following actions were taken,

1. **GeneratingSummaryStatistics**:
   For each method, the summary statistics was calculated and compared the mean and standard deviation across methods. The method resulting in a mean and SD closest to the original data (before imputation) is likely preferable. Significant deviations may indicate unrealistic imputation eNects.

2. **Histograms**:
   Histograms provide a visual overview of how each method aNects the distribution of the variables. A histogram similar to the original data indicates a method that preserves the original data's characteristics better.

3. **DensityPlots**:
   Density plots show how each method alters the distribution shape. Methods that keep the density plot similar to the original one are generally better. Focus was given to the methods that avoid creating artificial peaks or altering the data spread significantly.

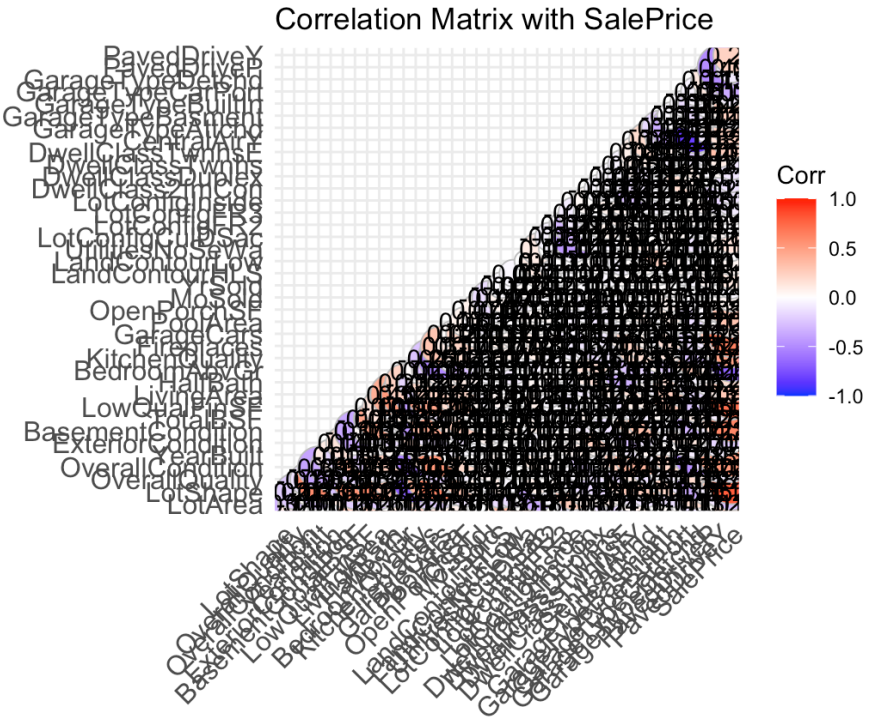
Accordingly, for the variables YearBuilt and LivingArea, it can be seen that the mean of the first method (replacing missing values with the mean for numerical variables) is closest to the original dataset's mean. The standard deviation (SD) for this method is also moderate, indicating that it does not significantly deviate from the original data (unlike method 3, which resulted in a high SD, or method 2, which resulted in a low SD). Additionally, the histograms generated from method 1 visually resemble the original distributions, indicating minimal distortion.
For GarageType, method 2 (deleting records with missing values) has notably reduced the dataset size, which might lead to a loss of valuable information. Methods 1 (imputing with the mode) and 3 (replacing with 0) retain all original records. However, method 3 results in a dataset that appears unrealistic and has a higher likelihood of distorting the data distribution due to the introduction of non-representative values (0s). In conclusion, the best method for handling missing values is to impute with the mean for numerical variables (YearBuiltand LivingArea) and with the mode for categorical variables (GarageType). This approach maintains the dataset's integrity while minimizing distortion and preserving the original data's characteristics.

Regression Analysis

Dimension reduction focuses on reducing the number of features while retaining as much information as possible. Key indicators for dimension reduction are high correlations (collinearity) among variables and potential redundancy.
To identify which variables should be used for dimension reduction, we examine the correlation matrix of the predictors. High correlations between variables (i.e., above a certain cutoN. In here, the cut oN used was 0.5) indicate multicollinearity. To avoid redundancy, we typically remove highly correlated variables, selecting a subset that minimizes inter-correlation while retaining essential information.
From the correlation matrix analysis (corrM), high correlations are identified using the findCorrelation function from the caret package.
Accordingly, the following were resulted:
OverallQuality, YearBuilt, LivingArea, FullBath, KitchenQuality, TotalRmsAbvGrd, GarageTypeDetchd, KitchenAbvGr, LandContourLvl, Slope
Among these variables, the following were chosen to be retained because they represent major aspects of a home that directly impact its market value and appeal, supported by industry insights on what drives home prices.

- **OverallQuality**: This variable is critical because it reflects the general condition and quality of the home, including materials and workmanship. High-quality finishes and construction are highly valued by both buyers and appraisers (Opendoor).
- **YearBuilt**: The age of the property is another key factor, as newer homes typically require less maintenance and may include modern design and energy-eNicient features, which increase their value(Bankrate)
- **LivingArea**: The square footage of a home is directly related to its value. Larger living areas oNer more space, which is highly desirable and directly influences the market price(HSH)
- **KitchenQuality**: The kitchen is one of the most important rooms in a home for buyers. High-quality kitchens with modern appliances and finishes are known to boost a home's value significantly((Bankrate)
- **GarageTypeDetchd**: While not as universally critical as the above factors, having a detached garage is a valuable feature, especially in areas where additional storage or workspace is prized. It can add convenience and flexibility, which buyers often appreciate(Opendoor).

Thereby, the following variables were removed: FullBath, TotalRmsAbvGrd, KitchenAbvGr, LandContourLvl, Slope
Conducting dimension reduction helped in reducing redundancy and multicollinearity, making the dataset more manageable and improving model performance.



High Positive Correlation:

- **OverallQuality (0.79)**: This variable has the highest positive correlation with SalePrice, indicating that better overall quality of the property strongly influences higher sale prices.
- **TotalBSF (0.61)**: Larger building square footage is also strongly associated with higher sale prices.
- **LivingArea (0.71)**: Similar to TotalBSF, more living area is correlated with higher property values.
- **GarageCars (0.64)**: The number of cars the garage can hold is positively correlated with sale price, suggesting that more garage space is valued by buyers.
- **YearBuilt (0.52)**: Newer properties tend to have higher sale prices, as indicated by the positive correlation.
- **Fireplaces (0.47)**: The presence of fireplaces has a moderate positive correlation with sale price, indicating that it's a desirable feature. **High Negative Correlation**:

• **KitchenQuality (-0.66)**: Interestingly, KitchenQuality has a strong negative correlation with SalePrice, which might suggest that properties with lower kitchen quality tend to have higher sale prices, although this could be due to other compensating features or specific market trends. This may require further investigation.

- **GarageTypeDetchd (-0.36)**: Detached garages are negatively correlated with sale prices, possibly indicating that buyers prefer attached garages.
- **LotShape (-0.27)**: The shape of the lot has a slight negative correlation, suggesting certain lot shapes might be less desirable. **Low or No Correlation**:
- **ExteriorCondition (-0.0047)**: Exterior condition has almost no correlation with SalePrice, implying it might not be a significant factor in property value.
- **UtilitiesNoSeWa (-0.014)**: The utility variable also shows minimal correlation, suggesting utilities are consistent across properties or not a major determinant of sale price.

- **LowQualFinSF (-0.025)** and **MoSold (0.046)**: Both variables have very low correlations with SalePrice, indicating little to no linear relationship.

Moderate Positive Correlations:

• **LotArea (0.26)** and **OpenPorchSF (0.31)**: These variables show moderate positive correlations with SalePrice, indicating that larger lots and open porch space are somewhat valued by buyers

Please refer R Script

MODEL 1 - This is the initial model before refinement.

MODEL 2 - The model after the initial refinement to remove variables with low correlation.

Based on the provided correlation values with SalePrice, you can identify which variables have low correlation. Variables with a correlation coeNicient close to 0 or those that are least correlated with SalePrice should be considered for elimination. From the list:

- LotShape: -0.2707
- OverallCondition: -0.0922
- ExteriorCondition: -0.0047
- BasementCondition: -0.2080
- LowQualFinSF: -0.0260
- HalfBath: 0.2834
- BedroomAbvGr: 0.1652
- LandContourHLS: 0.1200
- LandContourLow: 0.0527
- UtilitiesNoSeWa: -0.0144
- LotConfigCulDSac: 0.1417
- LotConfigFR2: -0.0073
- LotConfigFR3: 0.0181
- LotConfigInside: -0.0816
- DwellClass2fmCon: -0.0980
- DwellClassDuplex: -0.1155
- DwellClassTwnhs: -0.0995
- DwellClassTwnhsE: 0.0031
- CentralAirY: 0.2499
- GarageTypeAttchd: 0.2355
- GarageTypeBasment: -0.0298
- GarageTypeBuiltIn: 0.2357
- GarageTypeCarPort: -0.0708
- GarageTypeDetchd: -0.3582
- PavedDriveP: -0.0893
- PavedDriveY: 0.2311

Variables with lower absolute values of correlation are generally less useful for predicting SalePrice because they have less influence on the target variable.

MODEL 3 - The final model after removing insignificant variables from Model2.

Based on the summary of model2, the model was further refined by removing variables that are not statistically significant. From the summary output, it was identified which variables have high p-values (greater than 0.05), indicating that they are not significantly contributing to the model. The variables that were therefore removed were:

- PoolArea (p-value = 0.29346)
- OpenPorchSF (p-value = 0.76545)
- MoSold (p-value = 0.53452)

- YrSold (p-value = 0.75251)
- CentralAirY (p-value = 0.37268)
- GarageTypeDetchd (p-value = 0.14515)
- PavedDriveY (p-value = 0.72701)

QUESTION 1. C.

Model 1 (Initial Model)

Formula: SalePrice=396164.56+0.13·LotArea+4903.63·LotShape+15940.56·OverallQuality+5633.48·OverallCondition+265.73·YearBuilt+3043.11·ExteriorCondition+2738.76·BasementCondition+23.27·TotalBSF−5.8·LowQualFinSF+50.84·LivingArea−405.35·HalfBath−7861.73·BedroomAbvGr−17053.22·KitchenQuality+4650.91·Fireplaces+15300.71·GarageCars−25.89·PoolArea+24.35·OpenPorchSF−181.05·MoSold−499.62·YrSold+20180.35·LandContourHLS+17457.93·LandContourLow−63770.38·UtilitiesNoSeWa+18549.09·LotConfigCulDSac+4962.31·LotConfigFR2−1087.97·LotConfigFR3+2647.25·LotConfigInside+4275.07·DwellClass2fmCon−10387.58·DwellClassDuplex−29190.28·DwellClassTwnhs−22773.52·DwellClassTwnhsE+2920.15·CentralAirY+29891.85·GarageTypeAttchd+35908.04·GarageTypeBasment+39707.93·GarageTypeBuiltIn+4970.9·GarageTypeCarPort+24024.32·GarageTypeDetchd−9293.48·PavedDriveP+134.77·PavedDriveYSalePrice=396164.56+0.13·LotArea+4903.63·LotShape+15940.56·OverallQuality+5633.48·OverallCondition+265.73·YearBuilt+3043.11·ExteriorCondition+2738.76·BasementCondition+23.27·TotalBSF−5.8·LowQualFinSF+50.84·LivingArea−405.35·HalfBath−7861.73·BedroomAbvGr−17053.22·KitchenQuality+4650.91·Fireplaces+15300.71·GarageCars−25.89·PoolArea+24.35·OpenPorchSF−181.05·MoSold−499.62·YrSold+20180.35·LandContourHLS+17457.93·LandContourLow−63770.38·UtilitiesNoSeWa+18549.09·LotConfigCulDSac+4962.31·LotConfigFR2−1087.97·LotConfigFR3+2647.25·LotConfigInside+4275.07·DwellClass2fmCon−10387.58·DwellClassDuplex−29190.28·DwellClassTwnhs−22773.52·DwellClassTwnhsE+2920.15·CentralAirY+29891.85·GarageTypeAttchd+35908.04·GarageTypeBasment+39707.93·GarageTypeBuiltIn+4970.9·GarageTypeCarPort+24024.32·GarageTypeDetchd−9293.48·PavedDriveP+134.77·PavedDriveY

Metrics:

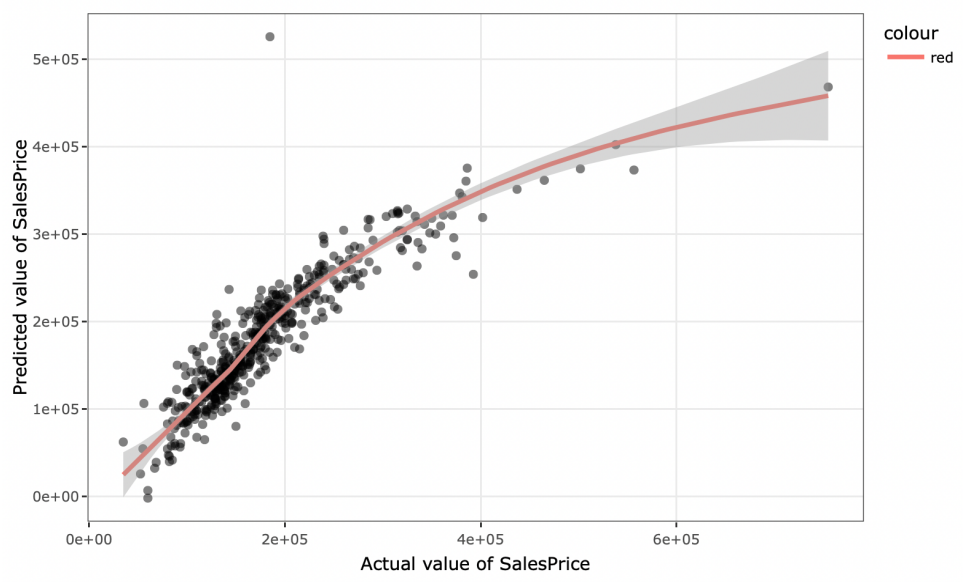
- Root Mean Square Error (RMSE): 36068.82
- R Squared (R2): 0.807

Model 2 (Refined Model)

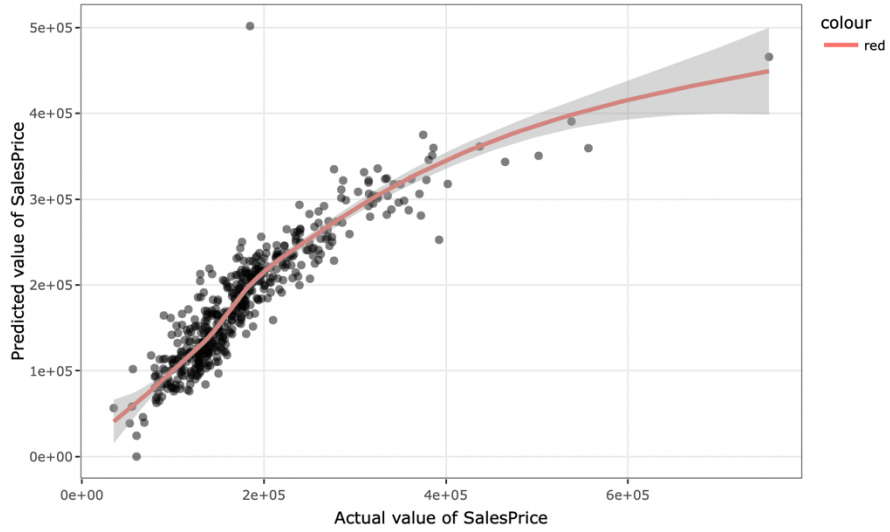
Formula: SalePrice=26984.01+0.7·LotArea+17199.68·OverallQuality+225.48·YearBuilt+57.04·LivingArea−6647.9·HalfBath−6705.8·BedroomAbvGr−18328.32·KitchenQuality+7589.4·Fireplaces+14461.84·GarageCars−26.08·PoolArea+4.84·OpenPorchSF−228·MoSold−234.64·YrSold+3969.05·CentralAirY+17630.98·GarageTypeAttchd+18444.36·GarageTypeBuiltIn+9807.74·GarageTypeDetchd+1422.06·PavedDriveYSalePrice=26984.01+0.7·LotArea+17199.68·OverallQuality+225.48·YearBuilt+57.04·LivingArea−6647.9·HalfBath−6705.8·BedroomAbvGr−18328.32·KitchenQuality+7589.4·Fireplaces+14461.84·GarageCars−26.08·PoolArea+4.84·OpenPorchSF−228·MoSold−234.64·YrSold+3969.05·CentralAirY+17630.98·GarageTypeAttchd+18444.36·GarageTypeBuiltIn+9807.74·GarageTypeDetchd+1422.06·PavedDriveY

Metrics:

• RMSE: 36391.04



• R Squared (R2): 0.784

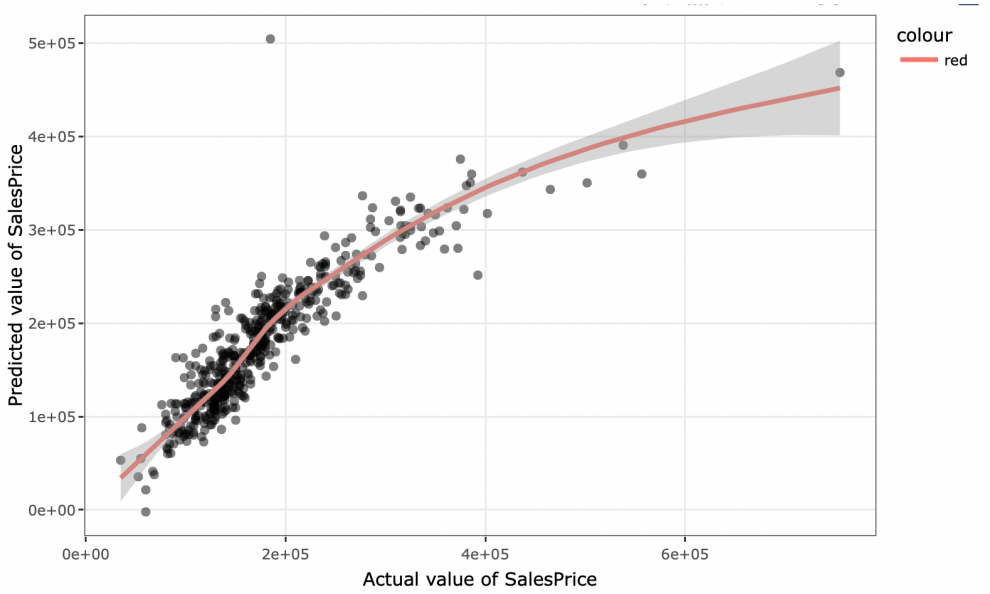


Model 3 (Final Model)

Formula: SalePrice=−447066.81+0.69·LotArea+17385.69·OverallQuality+233.4·YearBuilt+55.92·LivingArea−6183.05·HalfBath−6547.28·BedroomAbvGr−18536.34·KitchenQuality+7627.16·Fireplaces+14633.86·GarageCars+8877.62·GarageTypeAttchd+9795.05·GarageTypeBuiltInSalePrice=−447066.81+0.69·LotArea+17385.69·OverallQuality+233.4·YearBuilt+55.92·LivingArea−6183.05·HalfBath−6547.28·BedroomAbvGr−18536.34·KitchenQuality+7627.16·Fireplaces+14633.86·GarageCars+8877.62·GarageTypeAttchd+9795.05·GarageTypeBuiltIn

Metrics:

- RMSE: 36380.65
- R Squared (R2): 0.783



Comparison

1. **RootMeanSquareError(RMSE):**

o **Model 1**: 36068.82 o **Model 2**: 36391.04 o **Model 3**: 36380.65

Model 1 has the lowest RMSE, indicating it has the smallest average prediction error. Models 2 and 3 have slightly higher RMSE values, suggesting that Model 1's predictions are more accurate.

2. **RSquared(R2):**

o **Model 1**: 0.807

o **Model 2**: 0.784

o **Model 3**: 0.783

Model 1 has the highest R2, meaning it explains the largest proportion of the variance in the target variable. Models 2 and 3 have slightly lower R2 values, showing that Model 1 provides a better fit to the data.

Justification for Comparison

- **Model 1** includes all the initially considered variables and provides the best performance in terms of both RMSE and R2. This suggests that, despite potential collinearity, it better captures the relationships between the features and the target variable.
- **Model 2** is refined to include only variables with higher correlation but shows slightly worse performance compared to Model 1. This might be due to the removal of variables that, while showing low individual correlation, contribute to the model's predictive power.
- **Model 3**, the final model with further refinement to exclude insignificant variables, has similar RMSE and R2 values to Model 2. The minor drop in performance suggests that further simplification does not significantly improve the model and may even reduce its predictive capability.

Based on these metrics, **Model 1** is the most eNective in capturing the variability in the target variable and making accurate predictions.

Decision Tree Models

Please refer R script
Three pruned decision trees were developed to explore the impact of diNerent complexity parameters (CP) on model performance:
1. **DecisionTree1(PrunedwiththeBestCP(0.01))**:
o This tree was pruned using the optimal CP value identified through cross-validation, which aimed to balance model complexity and performance. 2. **DecisionTree2(PrunedwithCP>BestCP(0.03))**:
o This tree was pruned more aggressively (with a CP value greater than the best CP), resulting in a simpler tree. While this reduces overfitting, it might sacrifice predictive accuracy.
3. **DecisionTree3(PrunedwithCP<BestCP(0.001))**:
o This tree was pruned less aggressively (with a CP value lower than the best CP), retaining more branches. This could potentially improve accuracy but also increase the risk of overfitting.

Based on the RMSE values obtained:

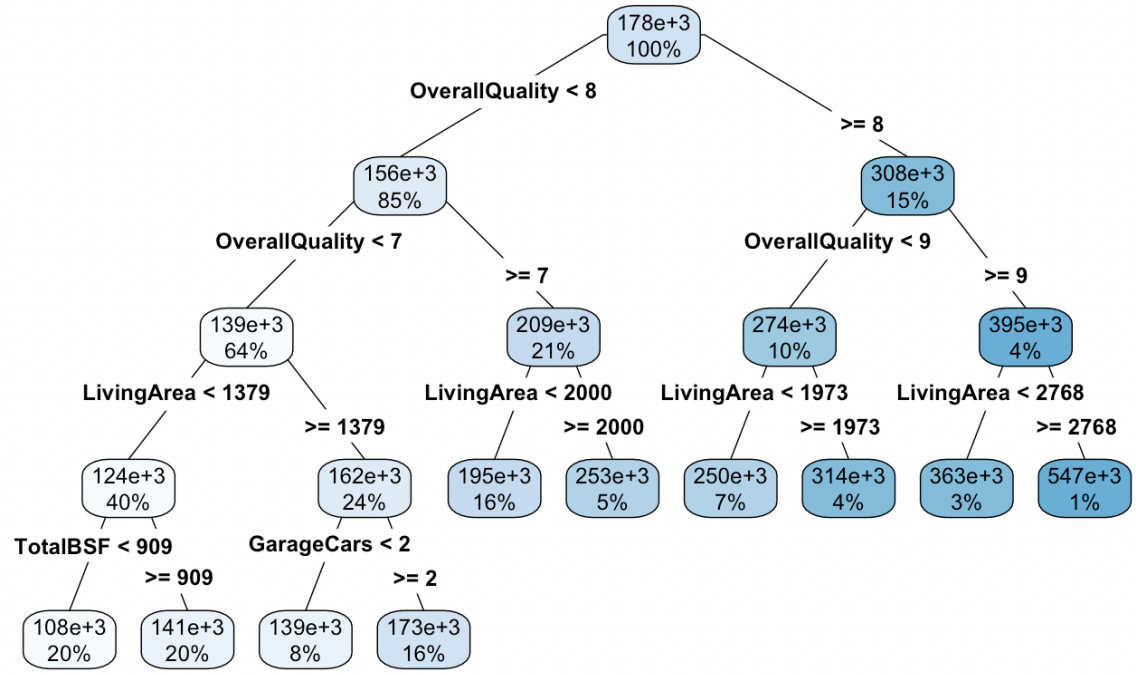
- **Initial Decision Tree RMSE:** 45424.27
- **Pruned Tree with CP = 0.01 RMSE:** 43732.39 (Lowest RMSE)
- **Pruned Tree with CP = 0.03 RMSE:** 48317.41
- **Pruned Tree with CP = 0.001 RMSE:** 43732.39 (Same as CP = 0.01)

  Optimal Tree: The pruned trees with CP = 0.01 and CP = 0.001 both have the lowest RMSE of 43732.39. Since both pruned models have the same RMSE, either could be considered optimal based on this metric. However, the tree with CP = 0.01 may be preferable due to its balance of complexity and performance.

  Model 1: Decision Tree with Pruning at CP = 0.01

• **RMSE:** 43732.39 (Best)

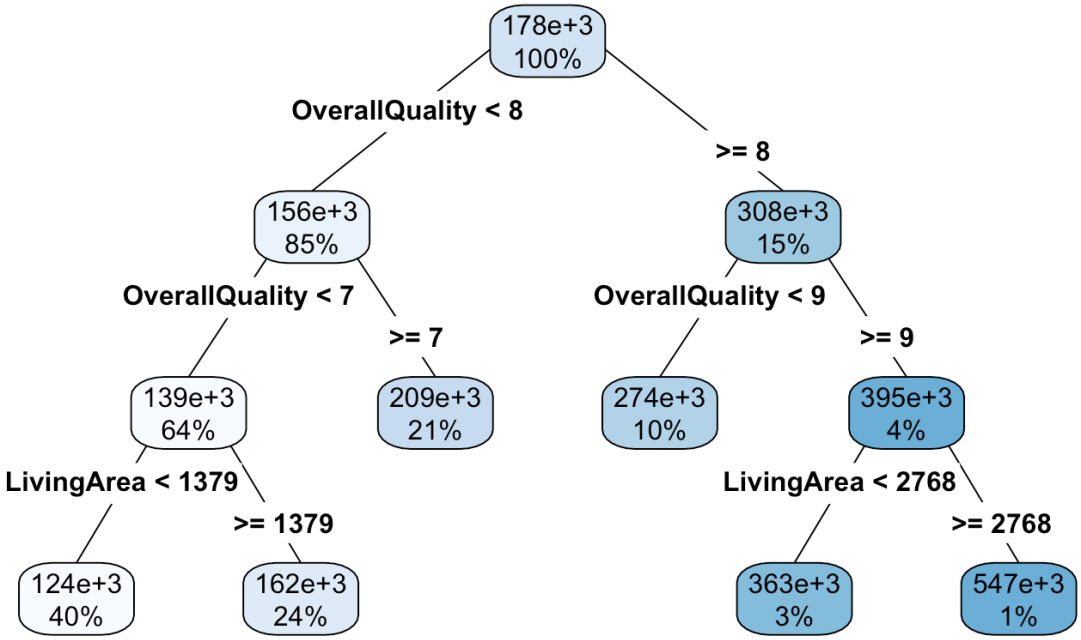
• **Tree Plot:**



Explanation: This model has the lowest RMSE, indicating it is the best in terms of prediction accuracy. The tree is pruned to avoid overfitting while capturing significant patterns in the data.

Model 2: Decision Tree with Pruning at CP = 0.03

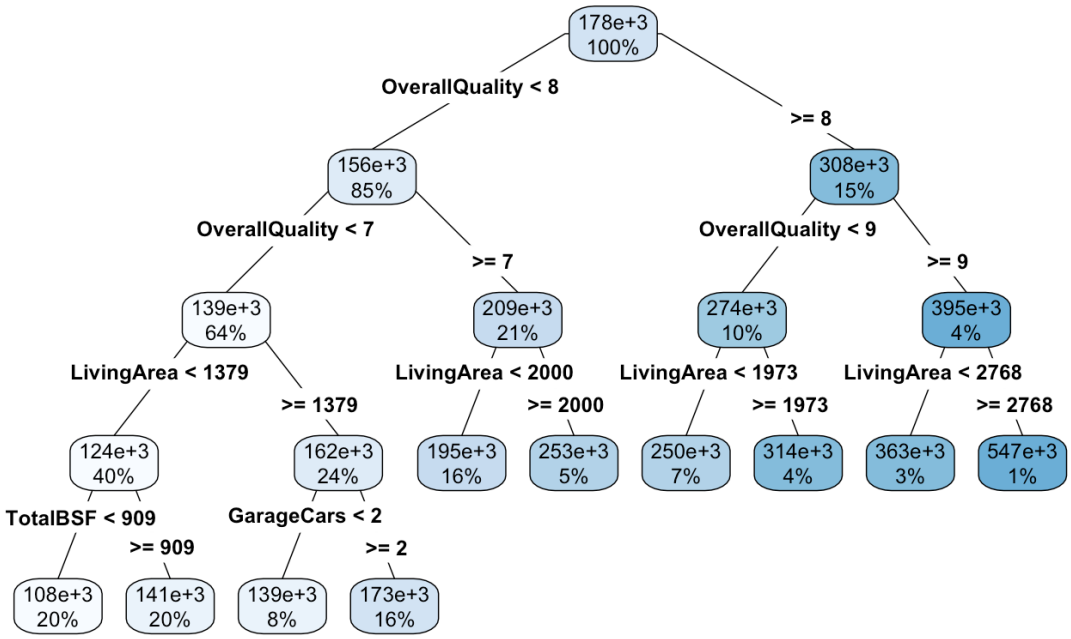
- **RMSE:** 48317.41
- **Tree Plot:**



Explanation: This model has a higher RMSE compared to the pruned model with CP = 0.01, suggesting that it might be too simplistic and missing important data patterns due to excessive pruning.

Model 3: Decision Tree with CP = 0.001

- **RMSE:** 43732.39 (Same as CP = 0.01)
- **Tree Plot:**



Explanation: This model has the same RMSE as the one with CP = 0.01, indicating that minimal pruning leads to a more complex tree without improving performance. It may still be prone to overfitting.

Model Comparison

Choosing key predictors: Certain features improve the performance of some models. You may determine which elements are most crucial for predicting the target variable and evaluate their respective eNects by testing with alternative models.
Robustness: Several models are tested to make sure the selected model is resilient and works well in a variety of situations or data sets. This aids in choosing a model that performs well when applied to unseen data.
Managing Multicollinearity: The way various models tackle feature interactions and multicollinearity varies. Understanding the impact of feature interactions on model performance is facilitated by the construction of several models.
Model Refinement and Improvement: Initial models oNer a performance baseline and expose areas in need of refinement. Developing and improving models aids in performance optimisation. The accuracy and generalizability of the model are enhanced by this iterative approach.
Bias-Variance Trade-oN: The bias and variance levels of various models may vary. For instance, simpler models may have high bias but low variance, whereas more complicated models (such as deep decision trees) may have low bias but high variance. Determining the ideal ratio of variance to bias can be achieved by comparing several models.
Evaluation of Performance: The performance of various models might diNer greatly. To choose the model that works best with your data, you may develop many models and evaluate metrics like RMSE, R2, etc.

Optimal Regression Model (Model 1): o RMSE: 36,068.82
o R2: 0.807

Optimal Decision Tree Model (Pruned at CP = 0.01):
o RMSE: 43,732.39

1. Accuracy: Theregressionmodelpredictsmorecorrectlyonaveragethanthe decision tree, as evidenced by its lower RMSE (36,068.82) over 43,732.39. Furthermore, the regression model appears to explain more variance in the target variable (SalePrice) than the decision tree, as seen by its higher R2 (0.807).

2. **ModelComplexity:**UnderstandinghoweachvariableaNectsthetargetismade easier by the regression model's simplicity and greater interpretability as a linear model. Even while the decision tree provides a more logical, rule-based structure, it is more likely to overfit, especially with little changes in the data, and can get complicated when there are numerous splits.

3. **BusinessContext:**Regressionmodellingismoreappropriateinthissituation since it fits the data better and has a higher accuracy when predicting property values. Making data-driven judgements in property assessment requires having a comprehensive understanding of the relationships between factors, which is what this tool oNers. Although helpful in addressing non-linear interactions and facilitating interpretability, the decision tree might not oNer the same degree of accuracy required for this particular business case.

Justification:

Given that both predicted accuracy and good data variability explanation are essential for a trustworthy appraisal of property prices, the regression model is the most appropriate choice for this particular business case. The simplicity and interpretability of the regression model also make it easier to communicate findings to stakeholders, supporting informed decision-making.