NORTHEASTERN UNIVERSITY

COLLEGE OF PROFESSIONAL STUDIES

INTERMEDIATE ANALYTICS

ALY6015 20305

MODULE 1 ASSIGNMENT

REGRESSION DIAGNOSTICS WITH R
AMES HOUSING DATASET

Submitted to: Professor Richard He

Submitted by: Vanshika Singh Miyan

Introduction

The objective of this report is to analyze the Ames Housing dataset using regression diagnostics techniques to understand the factors that influence housing prices in Ames. Iowa. The analysis consists of exploratory data analysis, data preparation, model building, and diagnostic analysis to answer the main research question: What are the significant predictors of housing prices in Ames?

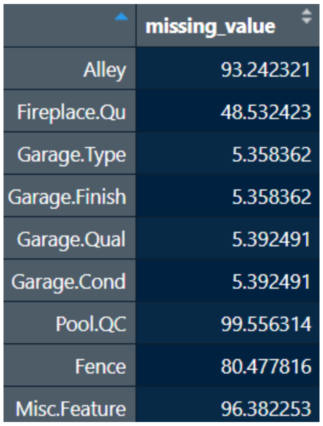
Data Preparation and exploratory Data Analysis

The Ames Housing dataset was loaded into RStudio for analysis. An exploratory data analysis was conducted to understand the data characteristics of the predictors. Various statistical summaries and visualizations were used to understand correlations, trends, and problems with the outliers and missing data.

Descriptive Statistics and Missing Data

Summary(housing) was used to understand the basic statistics of the data and the output highlighted some key features such as SalePrice with a median of $160,000 and a Mean of $180,796, pointing towards the right-skewed distribution of the data. A histogram of SalePrice was created to validate the findings and a boxplot was used to spot the potential outliers to identify the influential point is the dataset.

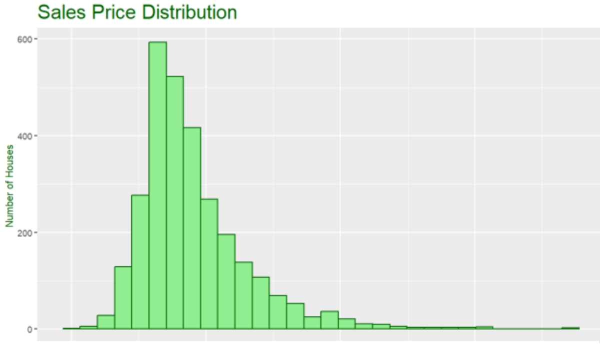
colSums(is.na(housing)) was used to identify the missing values in the dataset. The findings showed that some feature such as Misc.Feature, Pool.QC, and Alley had more than 90% missing data. Variables with more than 5% missing data were replaced by the median values and list is created.

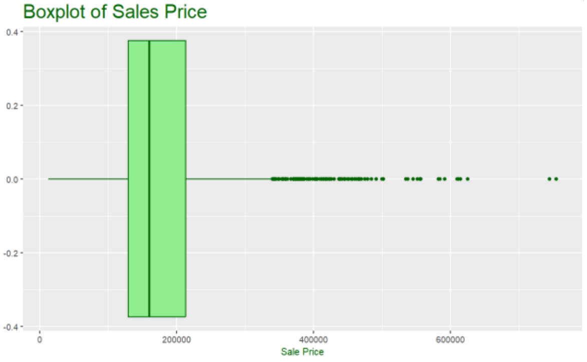


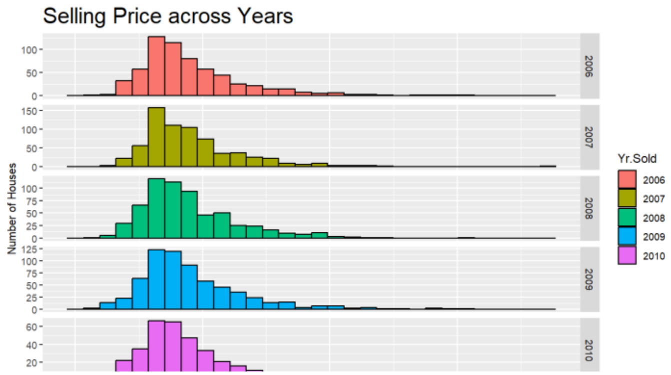
	missing_value
Alley	93.242321
Fireplace.Qu	48.532423
Garage.Type	5.358362
Garage.Finish	5.358362
Garage.Qual	5.392491
Garage.Cond	5.392491
Pool.QC	99.556314
Fence	80.477816
Misc.Feature	96.382253

Sale Price Distribution

- The histogram of the Sale Price has a right-skewed distribution, with most of the properties concentrated around the pricing range of $150,000 - $2000,000.
- There are some properties with high-value and hence create outliers.
- The mean sale price is $180,796, and the median is $160,000, highlighting the skewness in the distribution.
- There are several outliers, indicating high-end luxury properties that exceed $400,000.
- The interquartile range spreads from $120,000 to $210,000, providing insights about the typical property values

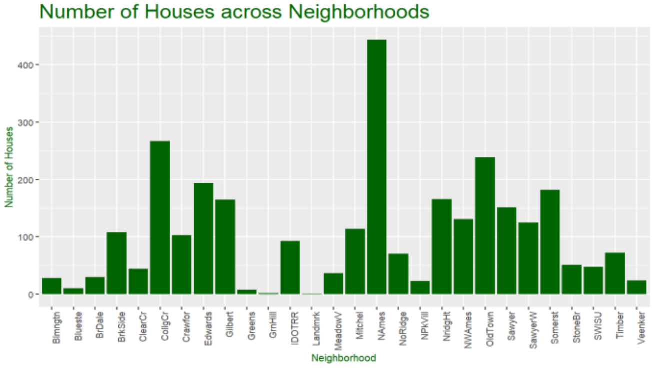






Sale Price across Neighborhoods

- There wasn’t much fluctuation is the price range from 2006 to 2010.
- There was not much sales in the year 2010.
- There was a slight increase in the selling prices in 2009 and 2010.
- The facet histogram shows that overall, the selling price had been constant with little fluctuations possibly due to inflations and market recovery.
- The Ames area has the highest number of properties as compared to the neighborhoods, pointing towards the a more social community of residents.

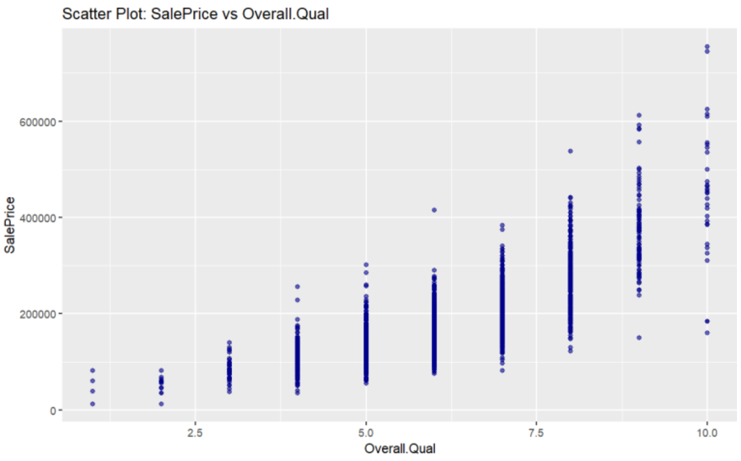


Correlation Analysis

A correlation matrix was created to understand the relationships between numerical variables. The matrix depicted that Overall.Qual has the highest positive correlation with SalePrice (0.8)**,** followed by Gr.Liv.Area(0.7) and Garage.Area(0.64). These results highlight the importance of overall quality, living area and garage space in evaluating housing prices in Ames. Moreover, a heatmap depiction was used to understand the relationship between variables, helping in feature selection for the regression model.

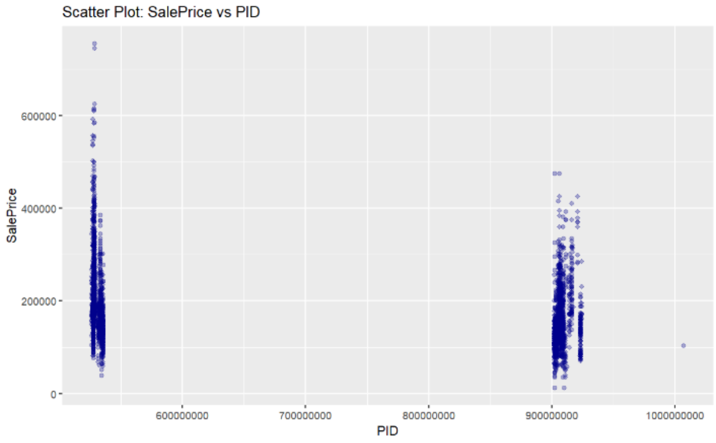
Highest correlation

- SalePrice vs. Overall.Qual (0.8)
- The plot suggests a strong relationship with a positive trend. This means that better quality homes have higher prices.
- Clusters at level 5 and 10 have mean prices of 130,000 and 300,000
- A few homes deviate marginally, possibly due to the high-quality.



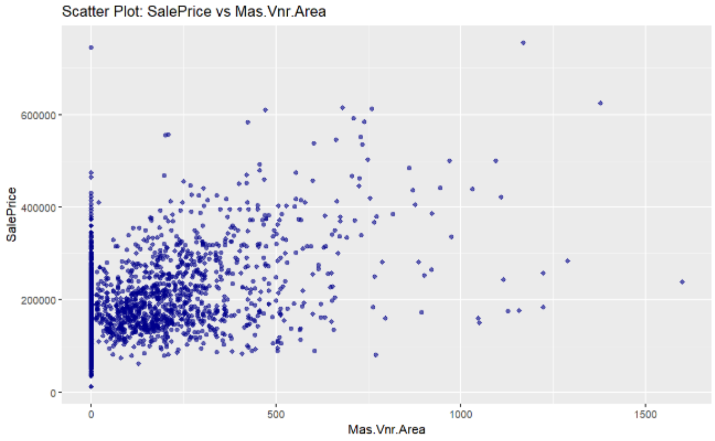
Lowest Correlation

- SalePrice vs. PID (0.01)
- No significant trend observed, suggesting of low predictive value of the variable.
- The variables are randomly distributed.
- PID is unsuitable as a predictor and should be excluded from the predictors' options.



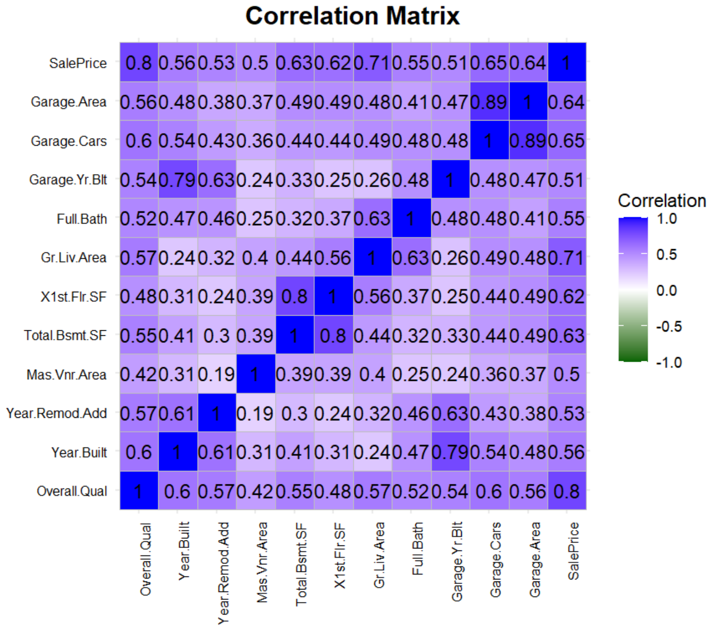
Correlation closest to 0.5:

- SalePrice vs. Mas.Vnr.Arena (0.42)
- A positive relationship is observed with high variability.
- Larger veneer areas are mostly associated with higher prices.
- Properties with larger veneer area greater than 500 sq. ft indicate the luxury end of the properties.



Insights from Correlation Heatmap:

1. Overall.Qual and Gr.Liv.Area have high correlation with SalePrice, suggesting they could be the strong predictors.
2. Multicollinearity is observed as correlations among predictors (Garage.Cars and Garage.Area) which indicates potential multicollinearity.
3. The variable PID has a weak correlation of 0.01, indicating its negligible influence on housing prices.



Regression Model:

A multiple linear regression model was built using four predictors:

Overall.Qual, Gr.Liv.Area, Garage.Cars, and Garage.Area

Model equation: SalePrice = -103818.43 + 27998.61 * Overall.Qual + 50.86 * Gr.Liv.Area + 5572.72 * Garage.Cars + 58.89 * Garage.Area

Coefficients:

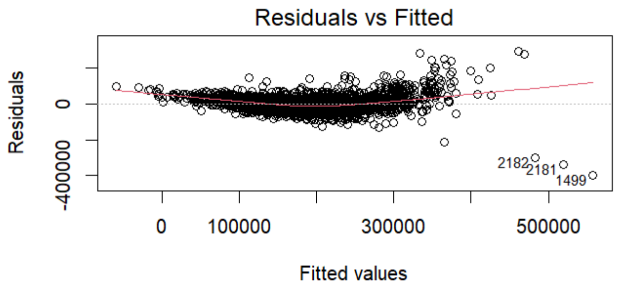
- **Intercept:** The baseline for SalePrice when all other predictors are held constant.
- **Overall.Qual:** For each unit increase in quality, the selling price increase by approximately $27,999 highlighting the strong influence of Overall quality of house on its' selling price.
- **Gr.Liv.Area:** For each additional square foot of living space, the Saleprice increases by approximately $50, highlighting the preference of bigger space among buyers.
- **Garage.Cars:** For each additional garage space, the SalePrice increases by $5573, highlighting the demand of more storage and parking among the buyers.
- **Garage.Area:** for each additional square foot in garage area, the selling price increase by approximately $59, highlighting the value of larger garages.

Model Diagnostics

Regression diagnostics were conducted to analyze model assumptions and performance:

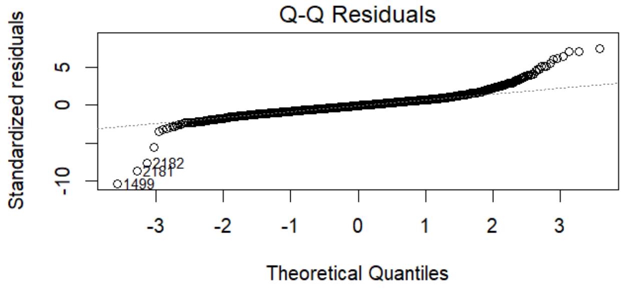
Residulas vs. Fitted

- No significant pattern observed, hence, supporting the linearity assumption.
- Visible outliers observed, indicating further review.
- Most of the residuals in the plot are close to 0.



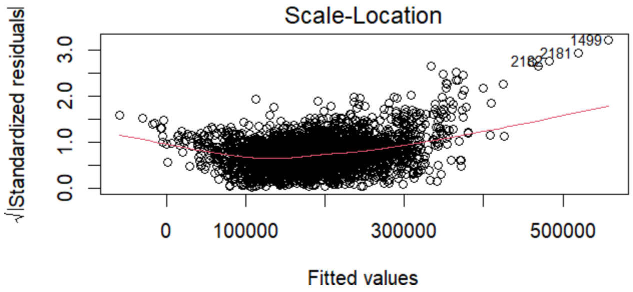
Q-Q Plot

- Residuals were observed to form normal distribution, with some deviations at the tails.
- Extreme values at both ends indicating minor issues in the values.
- Regression assumptions verified as the distribution is largely normal.



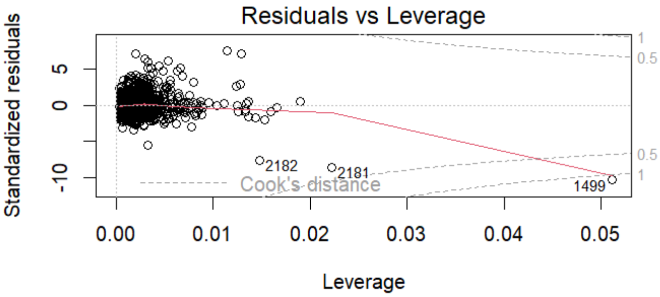
Scale-location Plot

- The plot verified homoscedasticity and residuals were observed to spread normally.
- There is consistent variance, therefore the model is constructed carefully.
- Outliers detected.



Residuals vs. leverage:

- Some influential points were observed such as 1499 and 2182.
- There's a minimal impact on Cook's distance, confirming the variables have not exceeded the threshold.

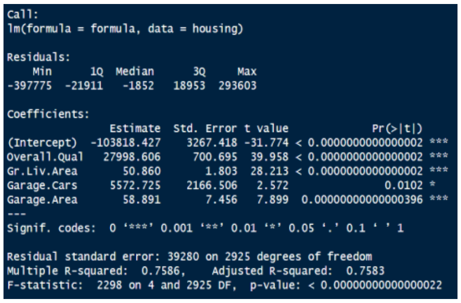


Multicollinearity Analysis:

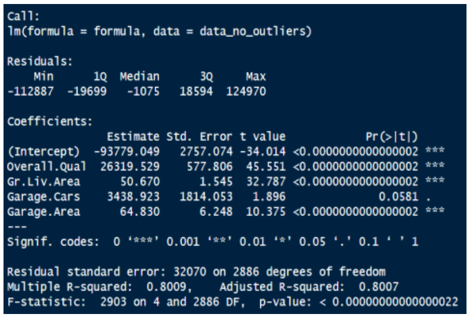
- The predictors such as Overall.Qual, Garage.Cars, Gr.Liv.Area had Variance Inflation Factor below 5, leading towards minimal multicollinearity.
- Garage.Area and Garage.Cars had a VIF of 5.15 and 4.88, highlighting the possibility of multicollinearity.
- Since, the VIF values are below the critical threshold, no changes were made to the moderate correlation between Garage.Area and Garage.Cars.

Outlier Analysis:

- Outliers were identified and removed using the standardized residuals exceeding 3 Standard Deviations.
- Excluding the outliers enhanced the model's adjusted R^2 value from 0.758 to 0.800, depicting a better fit and reliability.
- Outliers impacted the predictions significantly and had to be treated.
- Variance Inflation Factors were calculated, resulting in Garage.Cars and Garage.Area had high VIF's but did not exceed the critical threshold value.
- An all-subsets regression approach was utilized to identify the most optimal features. The features were compared to the original model and it was observed that the refined model had a higher value of adjusted R^2.



```
Call:
lm(formula = formula, data = housing)

Residuals:
    Min      1Q  Median      3Q     Max 
-397775  -21911   -1852   18953  293603 

Coefficients:
               Estimate  Std. Error t value             Pr(>|t|)    
(Intercept)  -103818.427    3267.418 -31.774 < 0.0000000000000002 ***
Overall.Qual   27998.606     700.695  39.958 < 0.0000000000000002 ***
Gr.Liv.Area       50.860       1.803  28.213 < 0.0000000000000002 ***
Garage.Cars     5572.725    2166.506   2.572               0.0102 *  
Garage.Area       58.891       7.456   7.899  0.00000000000000396 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 39280 on 2925 degrees of freedom
Multiple R-squared:  0.7586,    Adjusted R-squared:  0.7583 
F-statistic:  2298 on 4 and 2925 DF,  p-value: < 0.00000000000000022
```



```
Call:
lm(formula = formula, data = data_no_outliers)

Residuals:
    Min      1Q  Median      3Q     Max 
-112887  -19699   -1075   18594  124970 

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)    
(Intercept)  -93779.049   2757.074 -34.014 <0.0000000000000002 ***
Overall.Qual  26319.529    577.806  45.551 <0.0000000000000002 ***
Gr.Liv.Area      50.670      1.545  32.787 <0.0000000000000002 ***
Garage.Cars    3438.923   1814.053   1.896               0.0581 .  
Garage.Area      64.830      6.248  10.375 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32070 on 2886 degrees of freedom
Multiple R-squared:  0.8009,    Adjusted R-squared:  0.8007 
F-statistic:  2903 on 4 and 2886 DF,  p-value: < 0.00000000000000022
```

All-subsets Regression:

All-subsets Regression approach was used to identify the best combination of the predictors:

- **1-feature model:** Overall.Qual
- **2-feature model:** Over.Qual, Bsmt.Full.Bath
- **3-feature model:** Overall.Qual, BsmtFin.SF.1. Bsmt.Full.Bath
- **4-feature model:** Overall.Qual, Bsmt.Fin.1, Bsmt.Ful.Bath, Wood.Deck.SF

The 4-feature model equation:

SalePrice = -88087.23 + 40504.58 * Overall.Qual + 35.66 * BsmtFin.SF.1 + 1061.73 * Bsmt.Full.Bath + 61.56 * Wood.Deck.SF

Comparison of models and correction of issues:

The preferred model from step 13 differs from the model in step 12 mainly because of the selection of features. The Variables selected in the first model were based on high correlations with the selling price. On the other hand, the preferred model from step 13 selected the variables that were observed as most optimal predictors through the all-subsets regression approach. The second model also achieved a higher adjusted R^2 of 0.82 as compared to the first model (0.800).

I prefer the second model because it balances accuracy and complexity by identifying features with maximum explanatory power, and therefore reducing multicollinearity and highlighting predictors such as deck area and basement finish, which were not considered in the first model.

Conclusion:

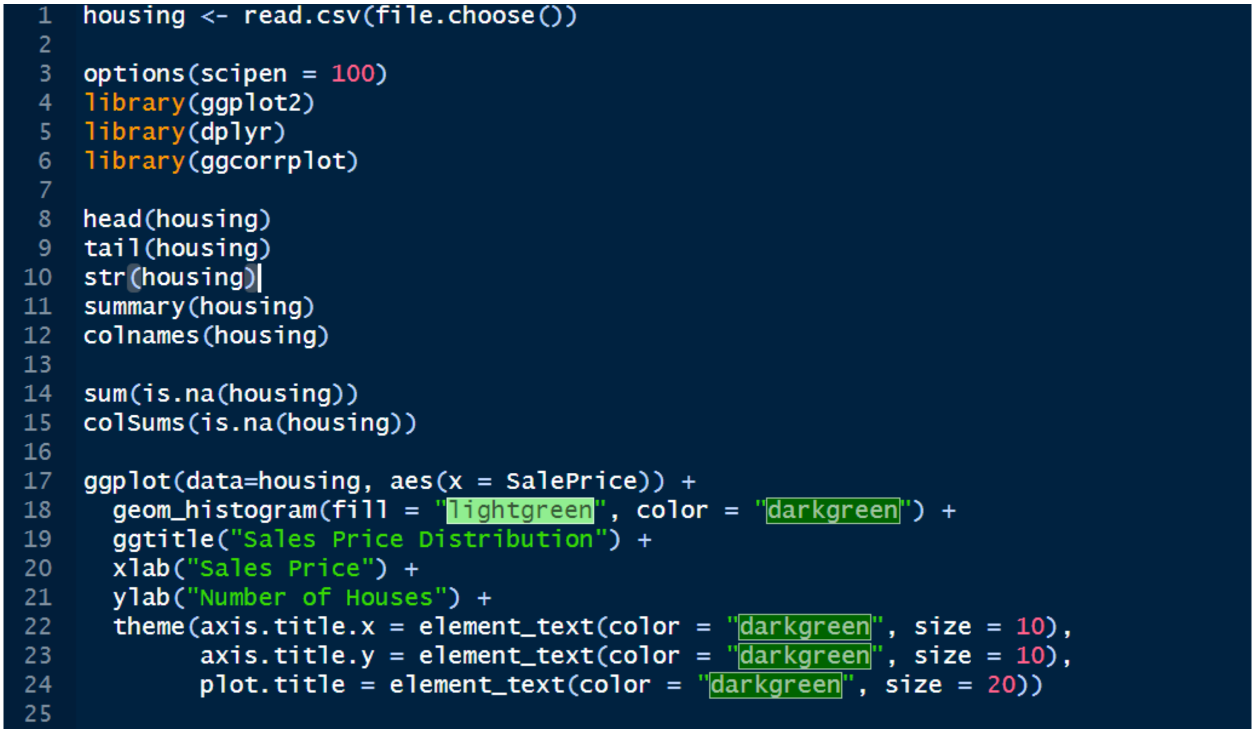
- The importance of bigger living and garage area highlights the demand for functional and spacious homes. The correlation between these features and the selling price indicates that buyers prefer good living and storage units.
- The most influential attribute while purchasing a house came out to be the Overall Quality, validating that the premium buyers prioritize the build quality. High quality homes are priced higher as compared to others and are in a popular demand.
- Buyers prefer amenities like bigger wooden decks and bathrooms in the basement. Homes with better functionalities and better outdoor seating space enhances the lifestyle and therefore are preferred.
- The selling price did not fluctuate much but had some ups and downs within a specific range, reflecting on the market condition over those years.
- Removing the outliers significantly improved the model’s performance and revealed a high concentration of houses in a specific price range.

References:

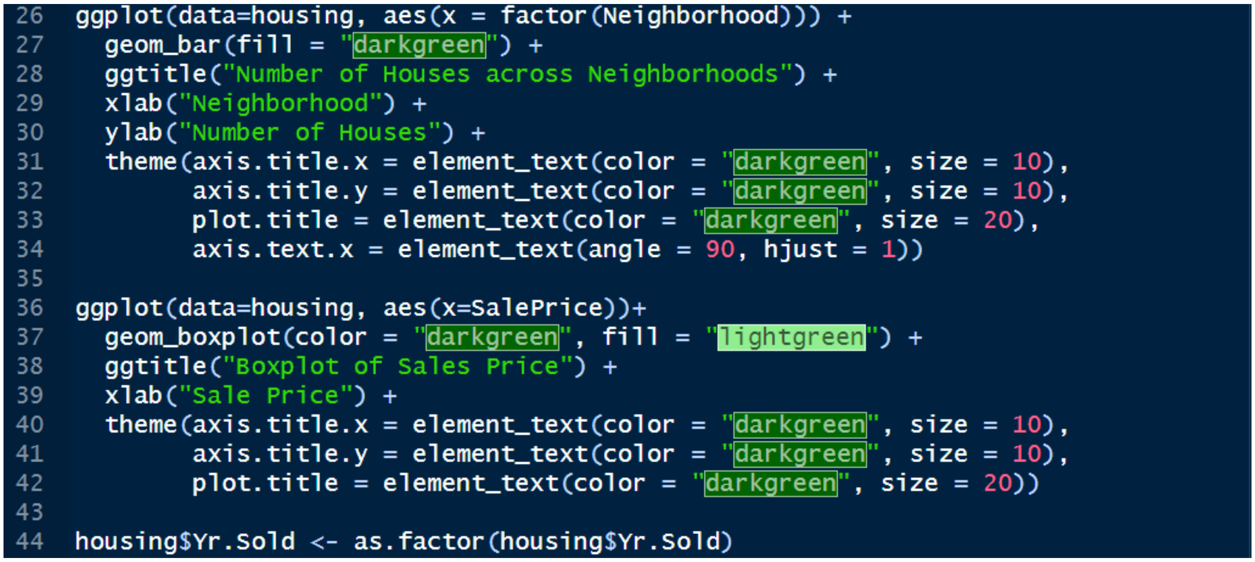
Kabacoff, R. I. (2021). *R in Tidyverse: Data visualization and analysis for the modern world*. Manning Publications

W3Schools. (n.d.). *Linear regression in R*. https://www.w3schools.com/#gsc.tab=0&gsc.q=linear%20regression

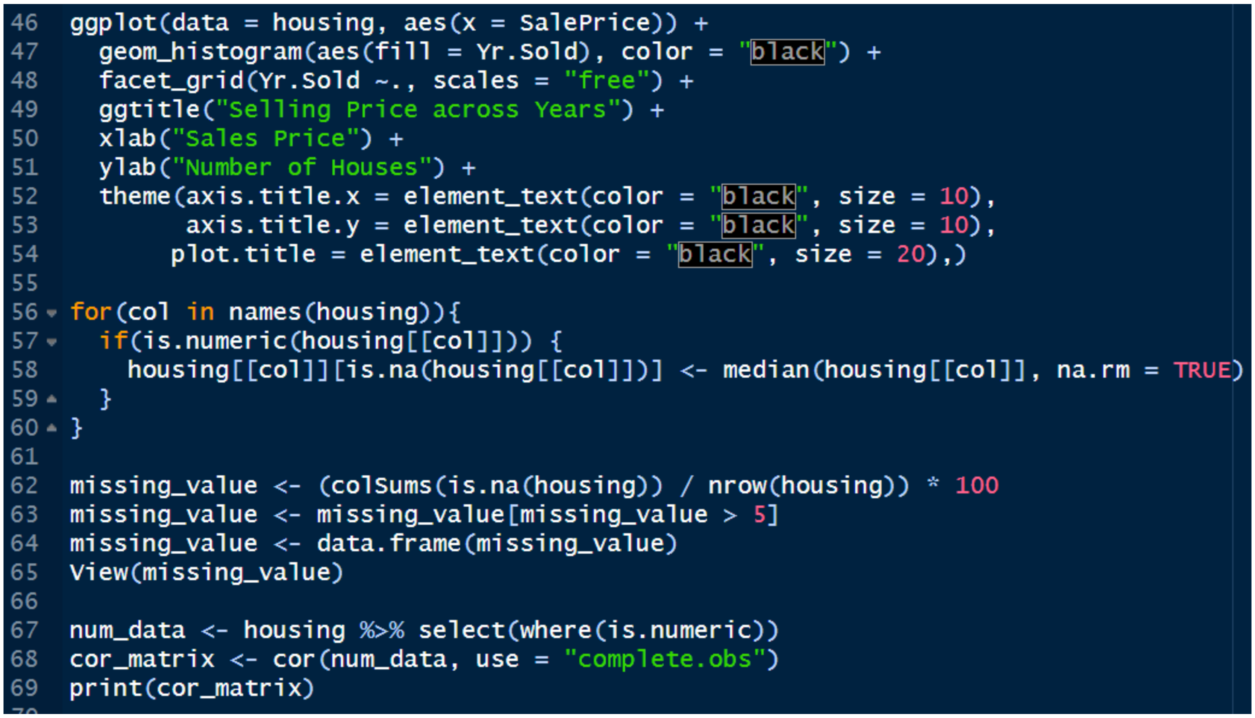
Appendix:



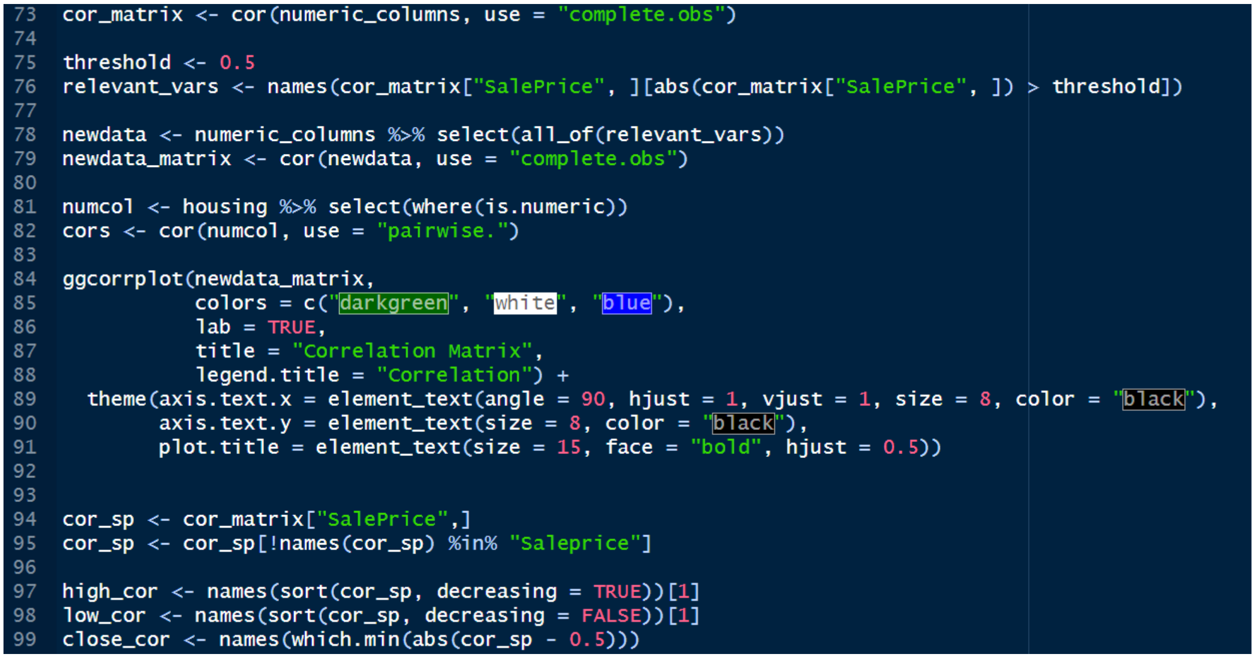
```r
housing <- read.csv(file.choose())

options(scipen = 100)
library(ggplot2)
library(dplyr)
library(ggcorrplot)

head(housing)
tail(housing)
str(housing)
summary(housing)
colnames(housing)

sum(is.na(housing))
colSums(is.na(housing))

ggplot(data=housing, aes(x = SalePrice)) +
  geom_histogram(fill = "lightgreen", color = "darkgreen") +
  ggtitle("Sales Price Distribution") +
  xlab("Sales Price") +
  ylab("Number of Houses") +
  theme(axis.title.x = element_text(color = "darkgreen", size = 10),
        axis.title.y = element_text(color = "darkgreen", size = 10),
        plot.title = element_text(color = "darkgreen", size = 20))

```



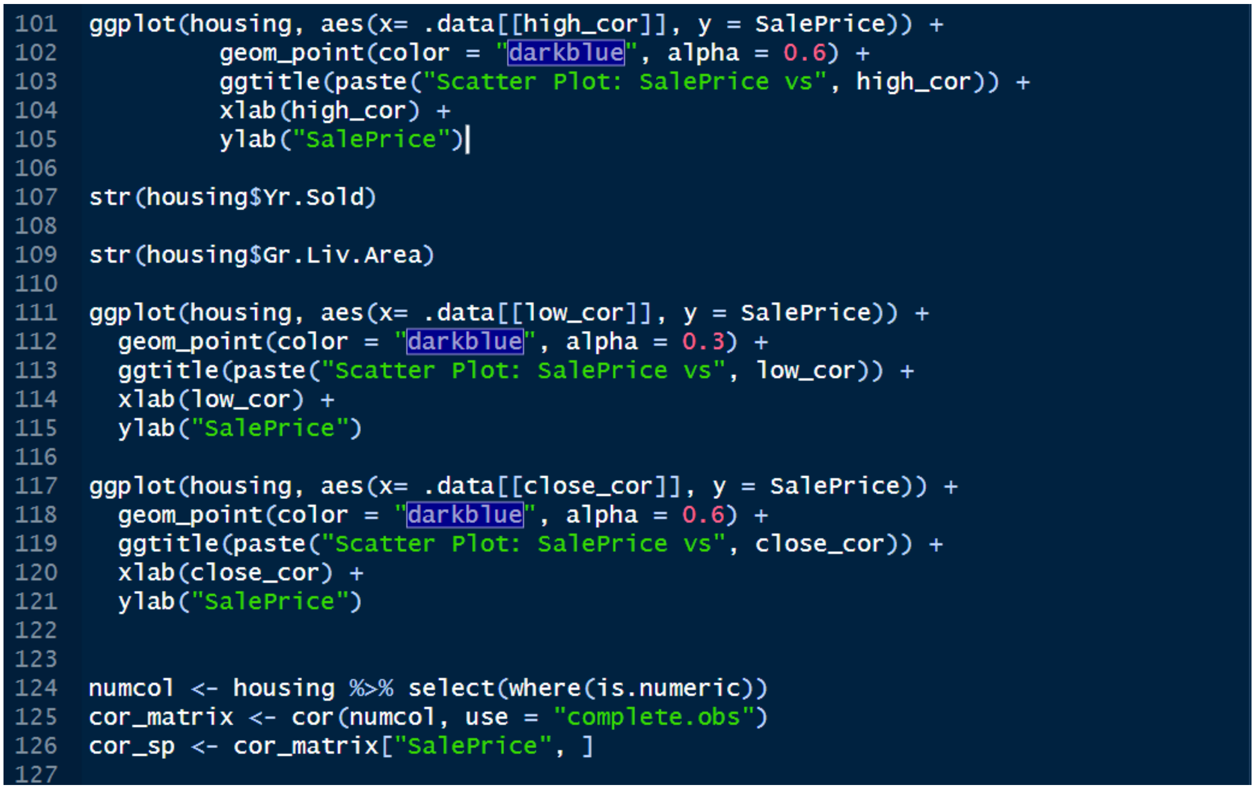
```r
ggplot(data=housing, aes(x = factor(Neighborhood))) +
  geom_bar(fill = "darkgreen") +
  ggtitle("Number of Houses across Neighborhoods") +
  xlab("Neighborhood") +
  ylab("Number of Houses") +
  theme(axis.title.x = element_text(color = "darkgreen", size = 10),
        axis.title.y = element_text(color = "darkgreen", size = 10),
        plot.title = element_text(color = "darkgreen", size = 20),
        axis.text.x = element_text(angle = 90, hjust = 1))

ggplot(data=housing, aes(x=SalePrice))+
  geom_boxplot(color = "darkgreen", fill = "lightgreen") +
  ggtitle("Boxplot of Sales Price") +
  xlab("Sale Price") +
  theme(axis.title.x = element_text(color = "darkgreen", size = 10),
        axis.title.y = element_text(color = "darkgreen", size = 10),
        plot.title = element_text(color = "darkgreen", size = 20))

housing$Yr.Sold <- as.factor(housing$Yr.Sold)
```



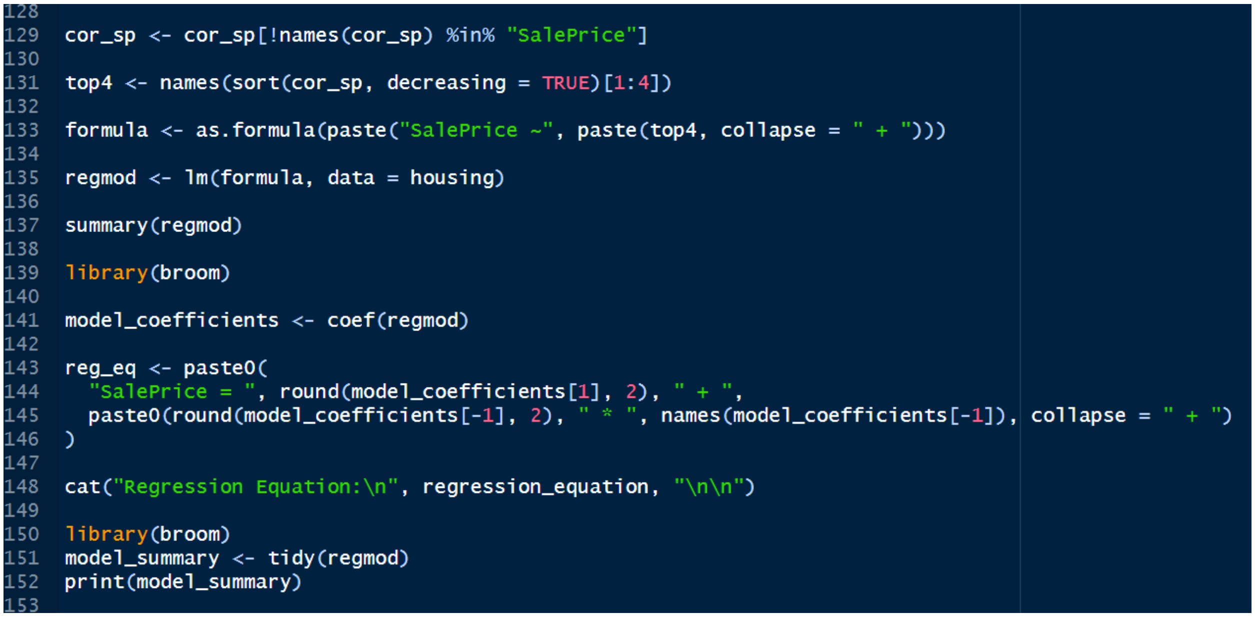
```r
ggplot(data = housing, aes(x = SalePrice)) +
  geom_histogram(aes(fill = Yr.Sold), color = "black") +
  facet_grid(Yr.Sold ~., scales = "free") +
  ggtitle("Selling Price across Years") +
  xlab("Sales Price") +
  ylab("Number of Houses") +
  theme(axis.title.x = element_text(color = "black", size = 10),
        axis.title.y = element_text(color = "black", size = 10),
       plot.title = element_text(color = "black", size = 20),)

for(col in names(housing)){
  if(is.numeric(housing[[col]])) {
    housing[[col]][is.na(housing[[col]])] <- median(housing[[col]], na.rm = TRUE)
  }
}

missing_value <- (colSums(is.na(housing)) / nrow(housing)) * 100
missing_value <- missing_value[missing_value > 5]
missing_value <- data.frame(missing_value)
View(missing_value)

num_data <- housing %>% select(where(is.numeric))
cor_matrix <- cor(num_data, use = "complete.obs")
print(cor_matrix)
```



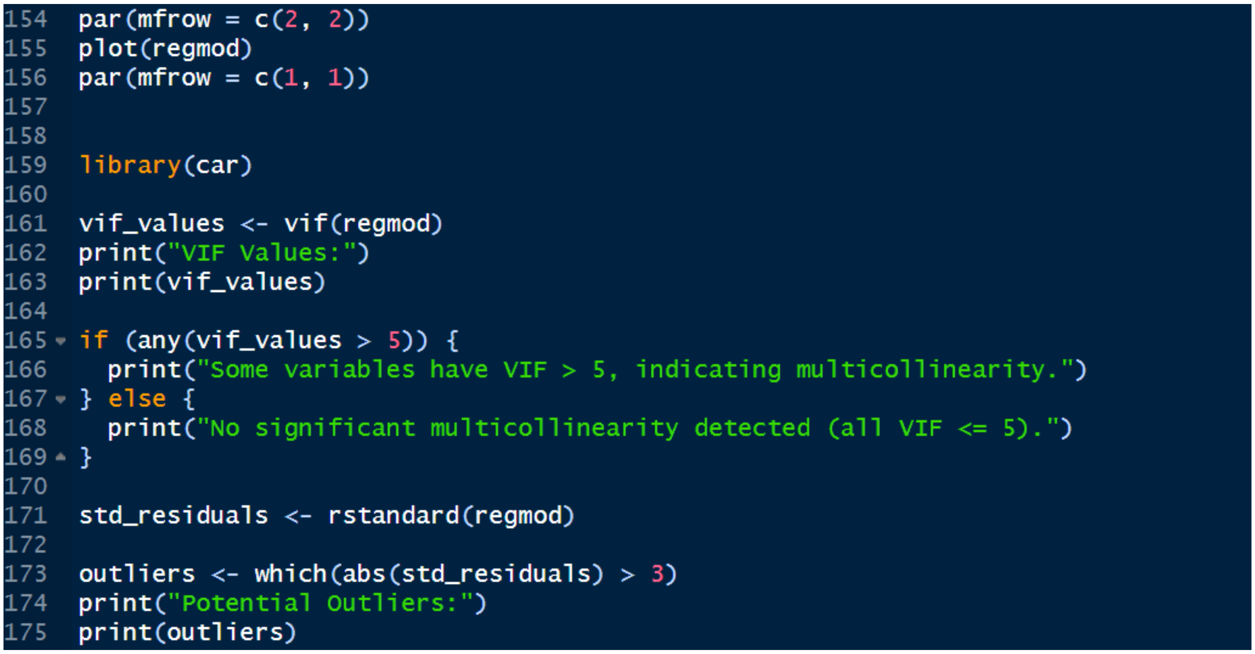
```r
cor_matrix <- cor(numeric_columns, use = "complete.obs")

threshold <- 0.5
relevant_vars <- names(cor_matrix["SalePrice", ][abs(cor_matrix["SalePrice", ]) > threshold])

newdata <- numeric_columns %>% select(all_of(relevant_vars))
newdata_matrix <- cor(newdata, use = "complete.obs")

numcol <- housing %>% select(where(is.numeric))
cors <- cor(numcol, use = "pairwise.")

ggcorrplot(newdata_matrix,
           colors = c("darkgreen", "white", "blue"),
           lab = TRUE,
           title = "Correlation Matrix",
           legend.title = "Correlation") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1, vjust = 1, size = 8, color = "black"),
        axis.text.y = element_text(size = 8, color = "black"),
        plot.title = element_text(size = 15, face = "bold", hjust = 0.5))


cor_sp <- cor_matrix["SalePrice",]
cor_sp <- cor_sp[!names(cor_sp) %in% "Saleprice"]

high_cor <- names(sort(cor_sp, decreasing = TRUE))[1]
low_cor <- names(sort(cor_sp, decreasing = FALSE))[1]
close_cor <- names(which.min(abs(cor_sp - 0.5)))
```



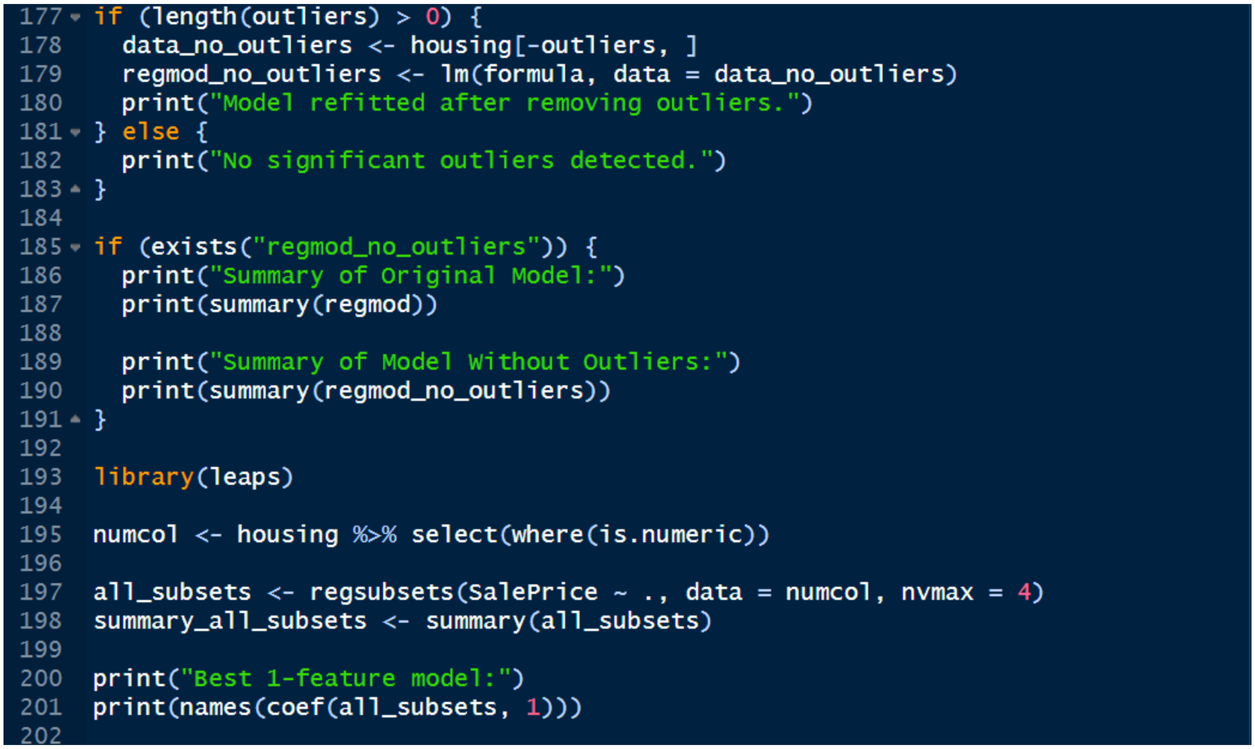
```
ggplot(housing, aes(x= .data[[high_cor]], y = SalePrice)) +
          geom_point(color = "darkblue", alpha = 0.6) +
          ggtitle(paste("Scatter Plot: SalePrice vs", high_cor)) +
          xlab(high_cor) +
          ylab("SalePrice")

str(housing$Yr.Sold)

str(housing$Gr.Liv.Area)

ggplot(housing, aes(x= .data[[low_cor]], y = SalePrice)) +
  geom_point(color = "darkblue", alpha = 0.3) +
  ggtitle(paste("Scatter Plot: SalePrice vs", low_cor)) +
  xlab(low_cor) +
  ylab("SalePrice")

ggplot(housing, aes(x= .data[[close_cor]], y = SalePrice)) +
  geom_point(color = "darkblue", alpha = 0.6) +
  ggtitle(paste("Scatter Plot: SalePrice vs", close_cor)) +
  xlab(close_cor) +
  ylab("SalePrice")


numcol <- housing %>% select(where(is.numeric))
cor_matrix <- cor(numcol, use = "complete.obs")
cor_sp <- cor_matrix["SalePrice", ]

```



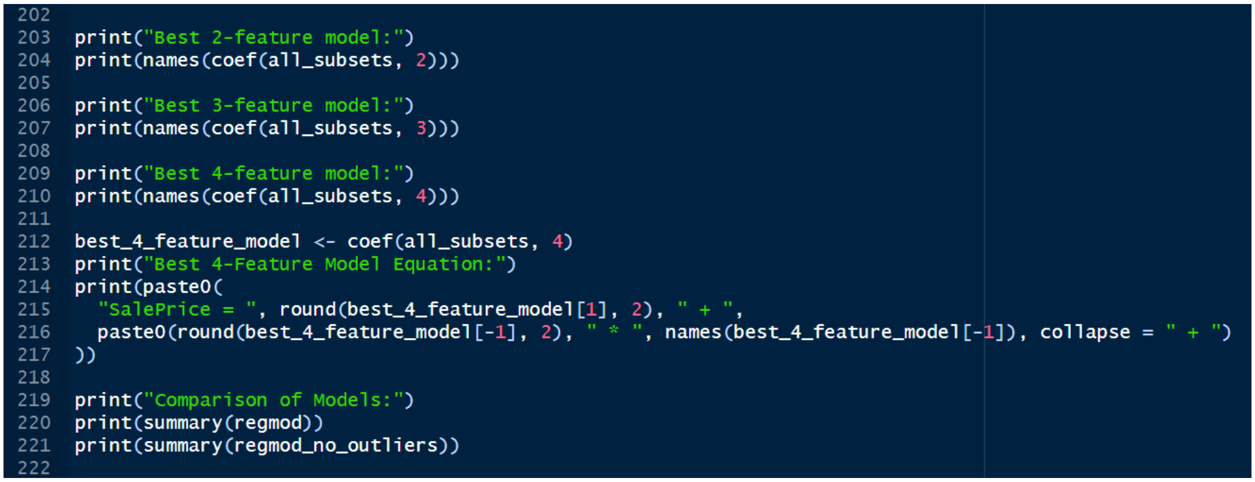
```

cor_sp <- cor_sp[!names(cor_sp) %in% "SalePrice"]

top4 <- names(sort(cor_sp, decreasing = TRUE)[1:4])

formula <- as.formula(paste("SalePrice ~", paste(top4, collapse = " + ")))

regmod <- lm(formula, data = housing)

summary(regmod)

library(broom)

model_coefficients <- coef(regmod)

reg_eq <- paste0(
  "SalePrice = ", round(model_coefficients[1], 2), " + ",
  paste0(round(model_coefficients[-1], 2), " * ", names(model_coefficients[-1]), collapse = " + ")
)

cat("Regression Equation:\n", regression_equation, "\n\n")

library(broom)
model_summary <- tidy(regmod)
print(model_summary)

```



```r
par(mfrow = c(2, 2))
plot(regmod)
par(mfrow = c(1, 1))


library(car)

vif_values <- vif(regmod)
print("VIF Values:")
print(vif_values)

if (any(vif_values > 5)) {
  print("Some variables have VIF > 5, indicating multicollinearity.")
} else {
  print("No significant multicollinearity detected (all VIF <= 5).")
}

std_residuals <- rstandard(regmod)

outliers <- which(abs(std_residuals) > 3)
print("Potential Outliers:")
print(outliers)
```



```r
if (length(outliers) > 0) {
  data_no_outliers <- housing[-outliers, ]
  regmod_no_outliers <- lm(formula, data = data_no_outliers)
  print("Model refitted after removing outliers.")
} else {
  print("No significant outliers detected.")
}

if (exists("regmod_no_outliers")) {
  print("Summary of Original Model:")
  print(summary(regmod))

  print("Summary of Model Without Outliers:")
  print(summary(regmod_no_outliers))
}

library(leaps)

numcol <- housing %>% select(where(is.numeric))

all_subsets <- regsubsets(SalePrice ~ ., data = numcol, nvmax = 4)
summary_all_subsets <- summary(all_subsets)

print("Best 1-feature model:")
print(names(coef(all_subsets, 1)))

```



```r

print("Best 2-feature model:")
print(names(coef(all_subsets, 2)))

print("Best 3-feature model:")
print(names(coef(all_subsets, 3)))

print("Best 4-feature model:")
print(names(coef(all_subsets, 4)))

best_4_feature_model <- coef(all_subsets, 4)
print("Best 4-Feature Model Equation:")
print(paste0(
  "SalePrice = ", round(best_4_feature_model[1], 2), " + ",
  paste0(round(best_4_feature_model[-1], 2), " * ", names(best_4_feature_model[-1]), collapse = " + ")
))

print("Comparison of Models:")
print(summary(regmod))
print(summary(regmod_no_outliers))

```