MA334-SP-7 Final Project (2023-24)

Kuldeep Paul (2322439)

Introduction

This report examines the key factors that impact house prices, with a specific focus on the Baton Rouge housing market. Comprehending these elements is crucial for stakeholders such as purchasers, vendors, and decision-makers. Statistical methods are utilized to examine housing data, pinpointing patterns, and investigating correlations between factors like price and square footage. Hypothesis testing is used to assess the influence of waterfront locations on prices. Afterwards, linear regression models are employed to evaluate the impact of various factors on housing costs, including a multiple regression analysis.

Data Exploration

Data Set Summary

- When we load the data set and view its structure, we can see that the data has **10 variables** and **770 observations**
- Of the 10 variables in the data set, all of which have an integer data type: **Numeric Variables:** price, sqft, bedrooms, baths, age, dom; **Categorical Variables:** pool, style, fireplace, waterfront
- By examining the dataset, we can find out the categrories of the qualitative variables and have listed the same in the table below.

Table 1: Categories of the Qualitative variables

	Variable_Context	Categories
pool	Whether pool is present	0, 1
style	Architectural Style	1, 2, 3, 4, 6, 8, 9
fireplace	Whether fireplace is present	0, 1
waterfront	Whether waterfront is present	0, 1

Descriptive Statistics

Since, none of the variables have NA values and thus we would not need to clean the data for our further analysis.

We would now calculate the descriptive statistics for all the numeric variables, this would include the mean, median, minimum, maximum, standard deviation and the trimmed mean (Winsorized Mean with 10% trimming) for the numeric variables.

Table 2: Descriptive Statistics of the numeric variables

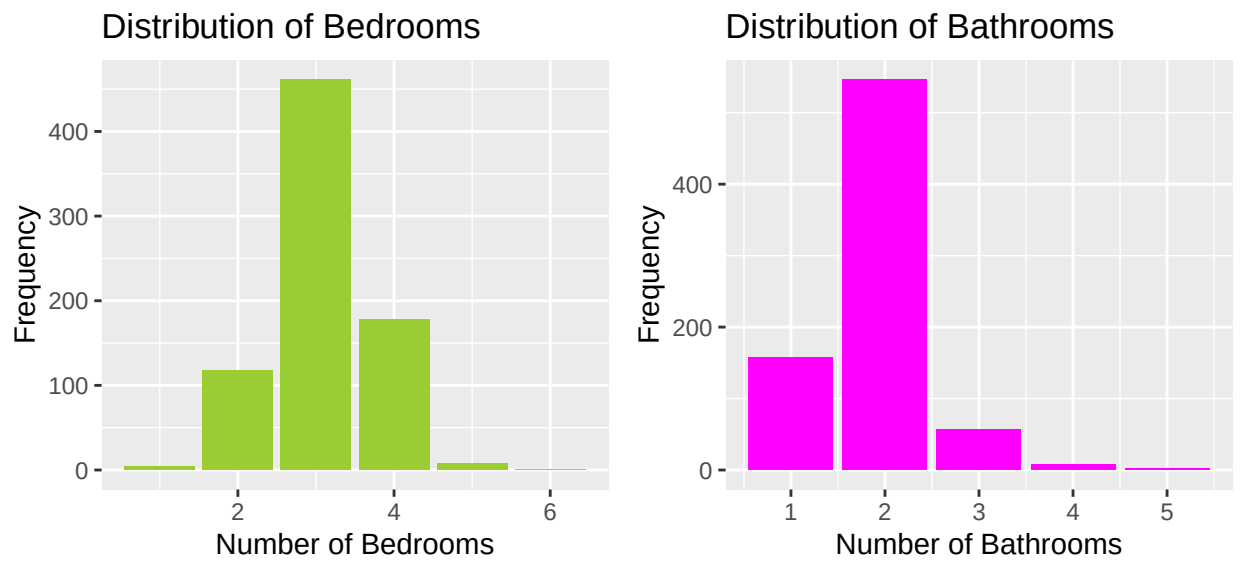
	Mean	Median	Min	Max	Std_Dev	Trimmed_Mean
price	138830.89	125000.0	22000	1007000	83049.84	127060.74
sqft	2188.76	2087.0	735	7099	891.32	2096.78
bedrooms	3.09	3.0	1	6	0.67	3.11
baths	1.89	2.0	1	5	0.57	1.87
age	21.61	18.0	1	80	17.32	19.96
dom	73.93	36.5	0	728	96.60	53.69

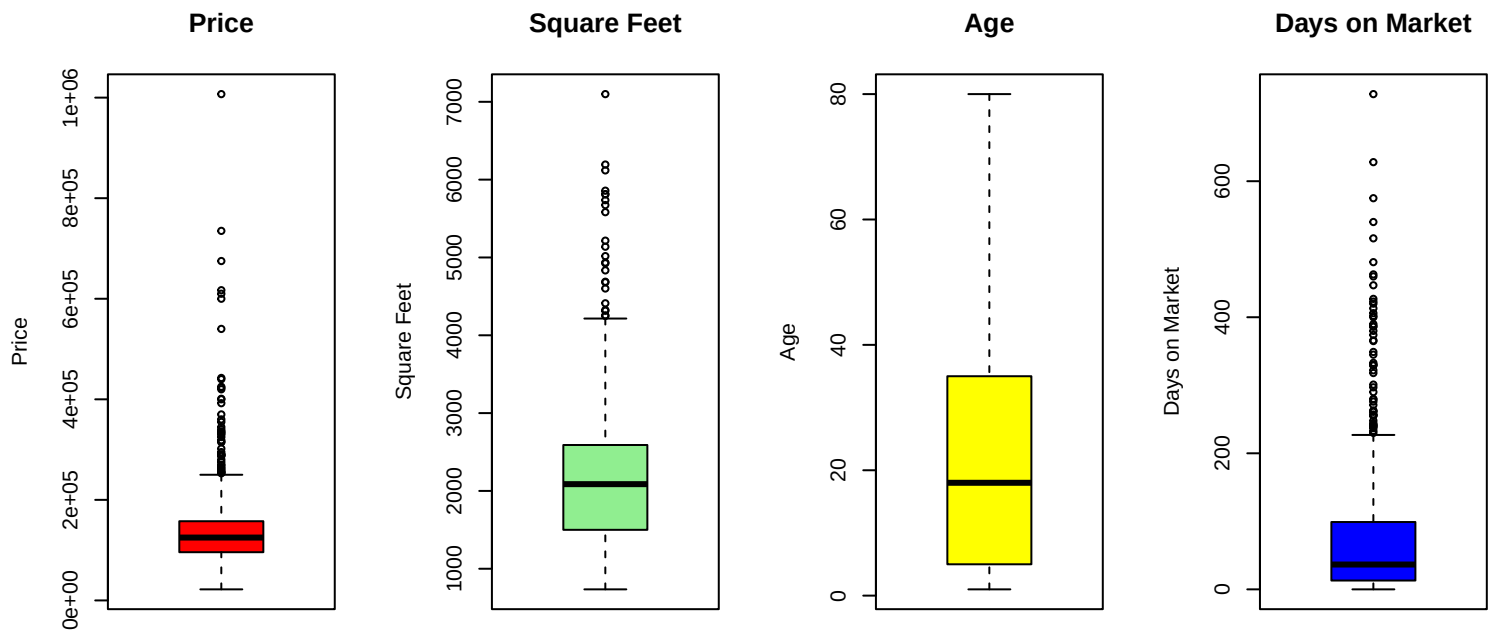
Significant observations from the Descriptive Analysis:

- Price Variability: The dataset exhibits significant variability in prices, reflecting a diverse housing market catering to various budgets.
- Price Outliers: The trimmed mean is slightly lower than the mean, likely due to a few high-priced outliers.
- Size Variability: There's a preference for spacious living with moderate variation in house sizes.
- Age: The housing market is relatively young, featuring both established neighborhoods and newer developments.
- Days on Market (dom): Indicates a moderately active market with significant variation in selling times.

Visualisation of distribution of certain variables in the dataset

We are visualising the distribution of the following variables: **bedrooms, baths, price, sqft, age, dom**

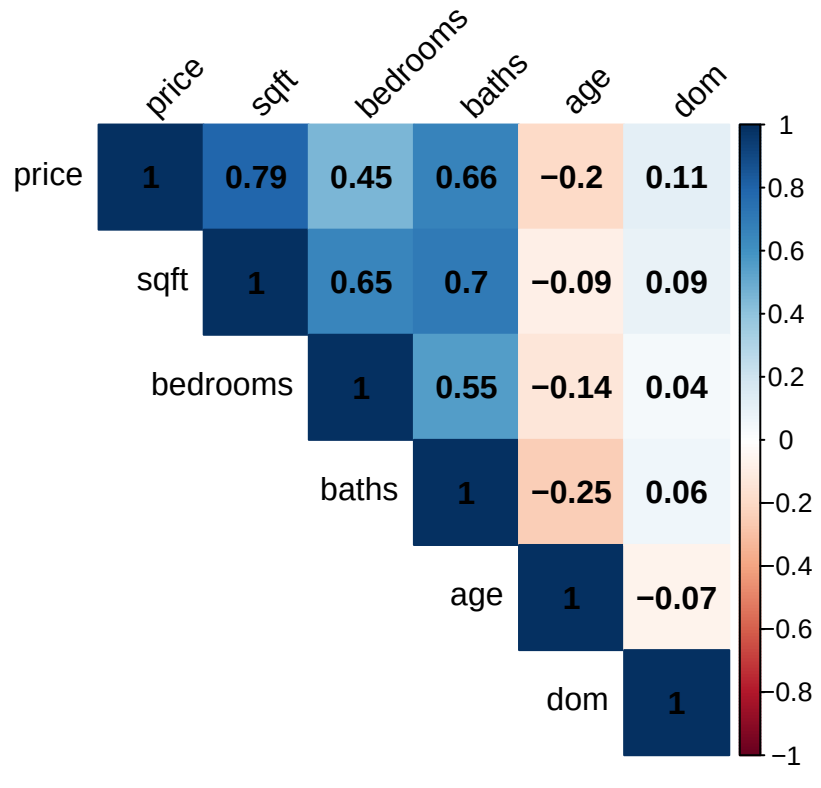




With the above histograms and boxplots we can infer the following:

- Majority of the homes have 3 bedrooms in the given dataset (More than 450)
- Majority of the homes have 2 bathrooms in the given dataset (More than 500)
- Skew: The distribution for the dom variable is skewed towards the lower values, while the distribution is relatively symmetrical for the variables price, sqft and age
- Outliers: The boxplots show several outliers in the price, sqft and dom variables

Correlation between variables



Observations

- From the above plot we can evidently see that there is a strong positive correlation between the following variables: *price, bedrooms, sqft, baths*
- There is negative correlation between: *price and age*
- There is very little correlation between: *dom and other variables*

The inference we can draw from the correlation matrix is that houses with more bedrooms would have more bathrooms, would have higher sqft and would have higher price and also, older houses would have lower price. The matrix also shows that the dom variable does not correlate to any other variable in the data this could signify that none of the selected variables have a significant impact on how fast or slow a house would sell after being listed on the market. This signifies a very active housing market and shows that is significant demand in the market of houses with varied range and size.

Probability, Probability Distributions and Confidence Intervals

Calculating the probability of a house chosen at random from the data set has a pool:

- Houses in the dataset with a pool: 59
- Total houses in the dataset: 770
- The probability of a house having a pool: **0.0766**

Calculating the conditional probability that it has a fireplace, given that it has a pool:

- Houses having a fireplace and pool: 44
- Total houses having a pool: 59
- Conditional probability of having a fireplace given it has a pool: **0.7458**

Calculating the probability that, out of 10 houses chosen at random from your data set, at least 3 will have a pool:

Probability of out of 10 houses chosen atleast 3 will have a pool: **0.0358**

Explanation: To determine the likelihood of at least 3 out of 10 randomly chosen houses having a pool from a dataset, we utilized the binomial distribution method. This method models the number of successes (houses with a pool) in a fixed number of trials (houses chosen), each with a given probability of success (probability of a house having a pool). By summing the probabilities of getting at least 3 successes out of 10 trials, we calculated the desired probability. This approach provides a quantitative measure to assess the probability of a specific event occurring, facilitating informed decision-making and analysis in real-world scenarios.

Calculating a 95% confidence interval on the mean house price in the USA, assuming the data set provides a random sample of houses in the USA:

The 95% confidence interval on the mean price in the USA is: **132955.65 to 144706.13**

Contingency Tables and Hypothesis Test

To test the hypothesis that the mean house price is greater for houses on the waterfront compared to those not on the waterfront, we can use a two-sample t-test (Welch Two Sample t-test)

- **Null Hypothesis (H0):** There is no difference in the mean house prices between houses on the waterfront and those not on the waterfront.
- **Alternative Hypothesis (H1):** The mean house price for houses on the waterfront is greater than the mean house price for houses not on the waterfront.

Using the Welch two sample t-test we get the following output:

t = 4.5859, df = 46.795, p-value = 1.692e-05, alternative hypothesis: true difference in means is greater than 0, 95 percent confidence interval: 58552.78 - Inf sample estimates:mean of x = 225657.6, mean of y = 133314.3

Interpretation:

- Since the p-value (1.692e-05) is much smaller than the significance level of 0.05, we reject the null hypothesis.
- This indicates that there is strong evidence to suggest that the mean house price for waterfront houses is significantly greater than the mean house price for non-waterfront houses.
- The 95% confidence interval confirms that the true difference in means is likely to be at least $58552.78, with waterfront houses having higher prices on average.
- The sample estimates provide the mean house prices for both waterfront and non-waterfront houses, showing a substantial difference between the two groups.

Hence, we can safely conclude that mean house price for houses on the waterfront is higher than houses not on the waterfront

Creating a contingency table showing relative frequencies for "Pool" and "No pool" according to whether a house has or hasn't got a fireplace:

Table 3: Contingency Table - Relative Frequencies:

	No Fireplace	Fireplace
No Pool	0.5302391	0.4697609
Pool	0.2542373	0.7457627

- Houses without Pools:
  - Among houses without pools, the majority (approximately 53.02%) do not have a fireplace.
  - A notable proportion (approximately 46.98%) of houses without pools have a fireplace.
- Houses with Pools:
  - Among houses with pools, a relatively smaller proportion (approximately 25.42%) do not have a fireplace.
  - A majority percentage (approximately 74.58%) of houses with pools have a fireplace.

Inferences:

- A majority of the houses with a pool have a fireplace thereby signalling a strong correlation between the two features.
- In houses without a pool, the relative frequencies of the presence or absence of a fireplace are not that far away and are around the 50% mark.

Using a 5% significance level, testing whether a house having a fireplace is independent of whether it has a pool

To test the independence between having a fireplace and having a pool, we can conduct a chi-squared test of independence. We'll compare the observed frequencies in the contingency table to the frequencies we would expect if fireplace and pool were independent

- **Null Hypothesis (H0):** Fireplace and pool are independent of each other.
- **Alternative Hypothesis (H1):** Fireplace and pool are not independent; there is a relationship between them.

At a significance level of 0.05, if the p-value is less than 0.05, we reject the null hypothesis, indicating that there is a significant relationship between having a fireplace and having a pool.

Using a Pearson's Chi-squared test with Yates' continuity correction, we get:

- Chi-squared statistic (X-squared): 15.52
- Degrees of freedom (df): 1
- The p-value: 8.165e-05

With a p-value much smaller than the significance level of 0.05, there is strong evidence to reject the null hypothesis. Therefore, based on this test, we can conclude that **there is a significant relationship between the presence of a pool and the presence of a fireplace in houses.** In simpler terms, the test suggests that the occurrence of having a pool is not independent of having a fireplace in the dataset. This indicates that there is likely some association or dependency between the presence of these two features in houses.

Simple Linear Regression

Performing Simple Linear Regression on the natural log of price and sqft from the dataset, we get the following performance:

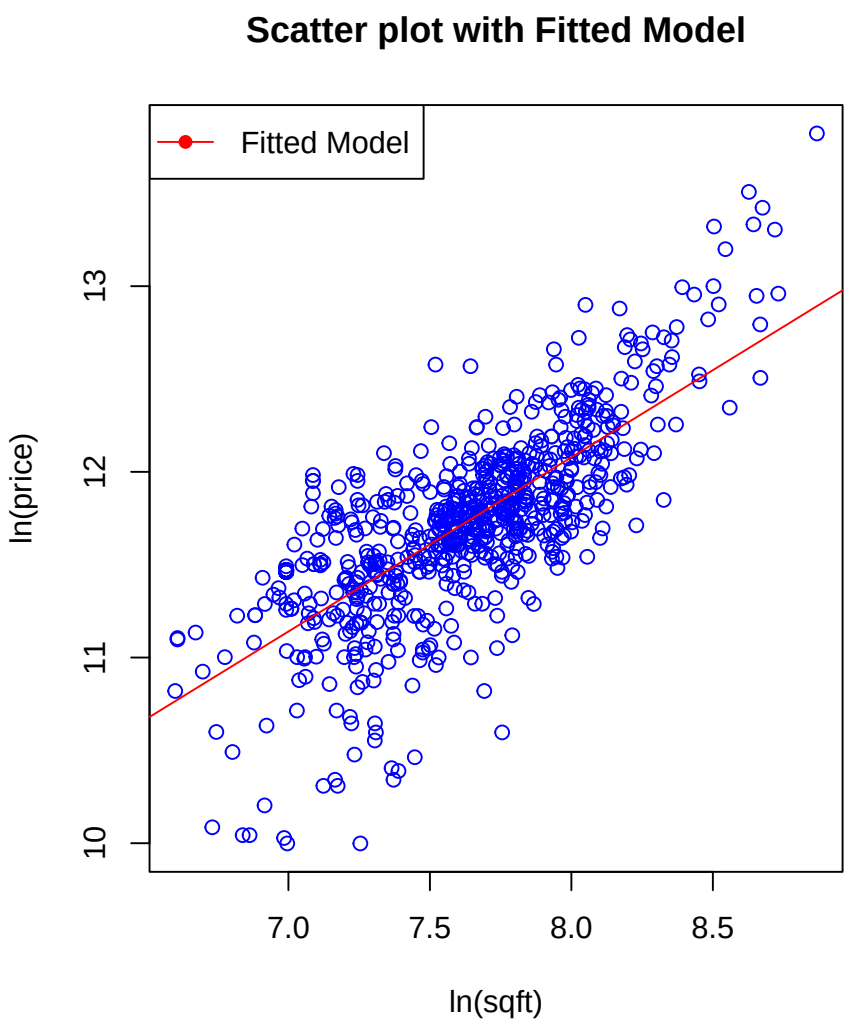
- The coefficient of ln(sqft) is 0.9381.
- The Intercept is 4.5735, but is of minimal significance as the area of a house cannot be zero.
- The standard error associated with the coefficient of ln(sqft) is 0.0302.
- The t-value associated with the coefficient of ln(sqft) is 31.06.
- The p-value associated with the coefficient of ln(sqft) is <2e-16, essentially zero.
- Since the p-value associated with the coefficient of ln(sqft) is essentially zero (<2e-16), which is much smaller than the significance level of 0.05, we reject the null hypothesis and conclude that ln(sqft) is a significant predictor of ln(price).
- Additionally, the R-squared value of 0.5568 suggests that approximately 55.68% of the variability in ln(price) can be explained by ln(sqft) in the model. This indicates a moderately strong relationship between ln(sqft) and ln(price).

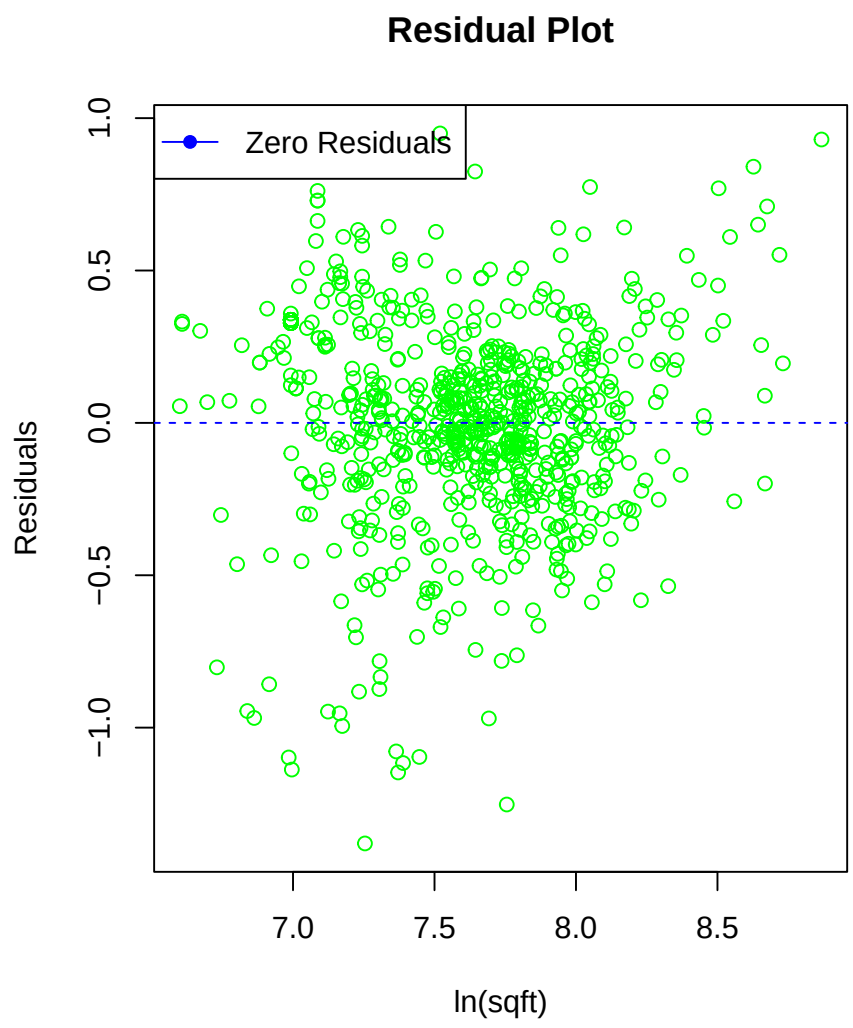
Interpretation of the slope coefficient (0.9381):

- The slope coefficient of ln(sqft) represents the estimated change in ln(price) for a one-unit increase in ln(sqft).
- In this case, for every one percent increase in total area (sqft), the ln(price) is expected to increase by approximately 0.9381 units on the natural logarithmic scale.
- Since ln(price) is in logarithmic scale, the interpretation of the slope coefficient in percentage terms is as follows: For every one percent increase in total area (sqft), the house price is expected to increase by approximately 0.9381 percent in the logarithmic scale.

In summary, based on the output provided, we can conclude that total area (represented by ln(sqft)) is a significant predictor of house price (ln(price)), and for every one percent increase in total area, the house price is expected to increase by approximately 0.9381 percent on the natural logarithmic scale. Hence, we can conclude that the total area is a significant predictor of house price.

Scatter Plot of the data along with the fitted model along with the plot of the residuals:



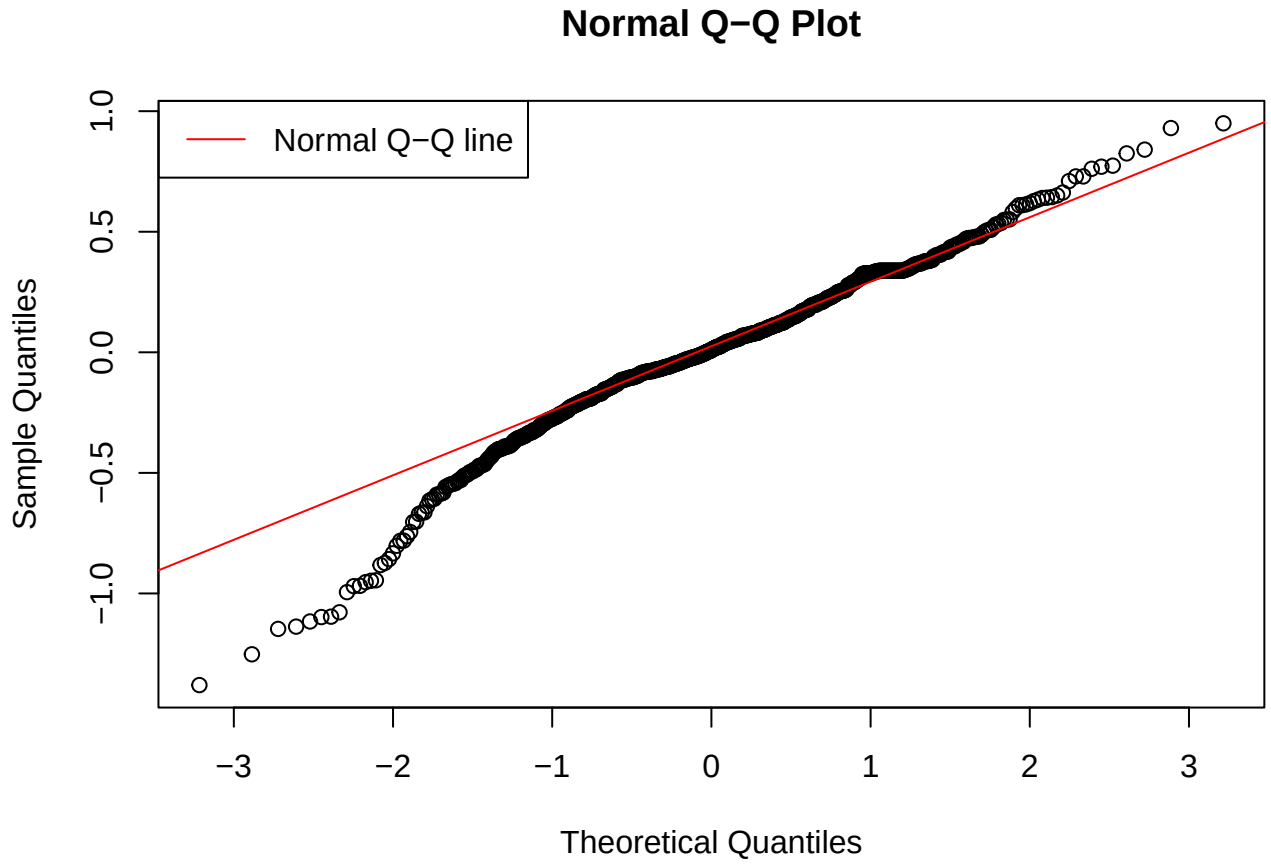


Scatter Plot with Fitted Model:

- The scatter plot shows the relationship between the predictor variable (ln(sqft)) and the response variable (ln(price)). The fitted model line represents the linear relationship estimated by the regression model.
- Ideally, the points should be distributed randomly around the line, indicating that the model captures the relationship between the variables well. If the points deviate systematically from the line, it suggests that the linear model might not be the best fit for the data.
- In the above plot indicates a positive linear relationship between the natural log of sqft and the natural log of price.
- There are a few instances in which the data points are significantly far away from the fitted line but there is no systematic deviation observed from the line, thereby showing that the linear model fits the data.

Residual Plot with the Zero Residual line:

- The residual plot shows the distribution of residuals (differences between observed and predicted values) against the predictor variable (ln(sqft)).
- Ideally, the residuals should be randomly scattered around the horizontal reference line at zero. If there are systematic patterns (e.g., increasing or decreasing spread of residuals with the predictor variable), it suggests that the linear regression model might not be appropriate for the data.
- In the above residual plot, the residuals are randomly distributed around the zero residual line, indicating that the assumption of constant variance is reasonable and the linear model is appropriate.



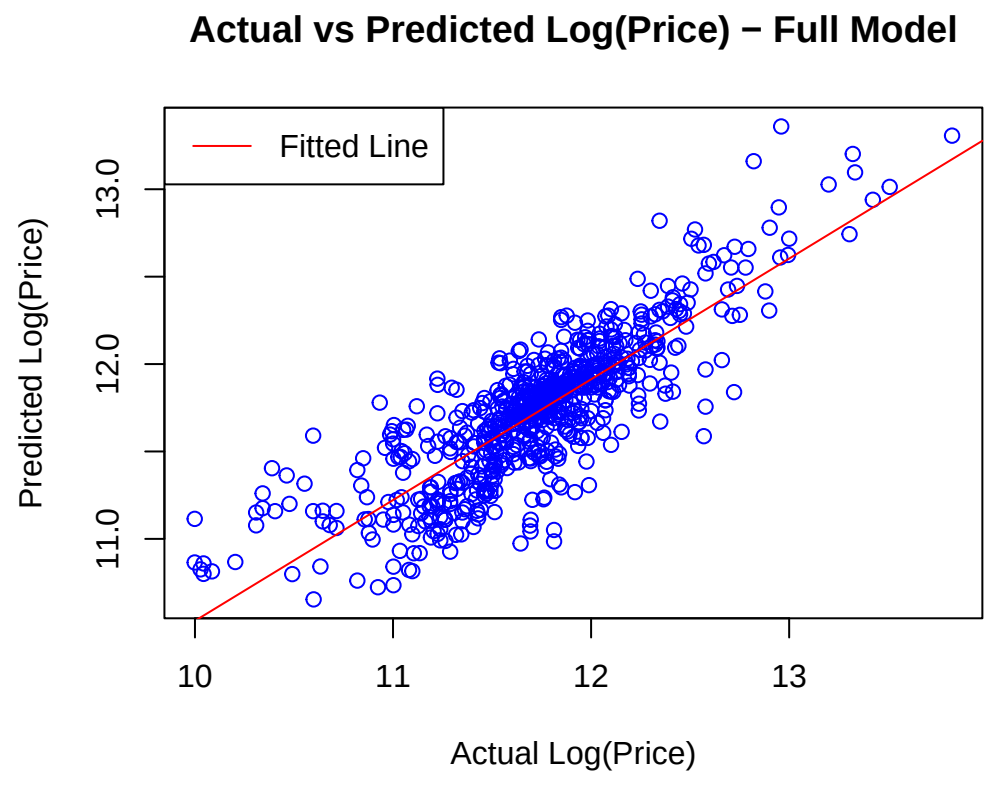
QQ Plot of Residuals:

- The QQ plot compares the distribution of residuals to a theoretical normal distribution. If the points approximately fall along a straight line, it suggests that the residuals are normally distributed.
- In the above QQ Plot majority of the points lie on the Normal QQ Line with minor deviations near the tail and the head indicating that the model generally captures the overall trend of the data.

Multiple Linear Regression

Performing a multiple linear regression of ln(price) against all the predictor variables:

In order to perform the Multiple Linear Regression on all the predictor variables, it is necessary for us to on-hot encode the "style" variable which is a categorical variable with 7 levels. We also convert the sqft variable to its natural log. To do this we convert the "style" variable into a factor and then encode the variables using dummy variables. Once this is done, we use the encoded dataframe and train the **full_model** for the price variable against all the predictor variables.



The above plot shows the Actual Log(Price) vs Predicted Log(Price) of the full model and we can clearly see that the model is able to make quite good predictions and there are very few outliers that can be observed on the plot.

Interpretation:

The linear regression model fitted to the encoded data provides valuable insights into the determinants of house prices in the Baton Rouge housing market. Here are the key findings from the analysis:

- Intercept: When all other predictors are zero, the expected log of the house price is approximately 5.936.
- Coefficients: Square footage log(sqft) has a significant positive effect on house prices, with an estimated coefficient of 0.718. For each unit increase in square footage, the expected log of the house price increases by approximately 0.718, holding other variables constant.

- Bedrooms, Baths, Age, Pool, Style, Fireplace, Waterfront, and Days on Market (dom): While some variables such as baths, age, style2, fireplace, and waterfront show statistically significant coefficients, others like bedrooms, pool, style1, style3, style4, style6, style8, and dom do not exhibit statistical significance based on their p-values.
- Residuals: The residuals, representing the differences between observed and predicted log house prices, appear to be normally distributed with a mean close to zero and a standard error of approximately 0.2738.
- Model Fit: The multiple R-squared value of approximately 0.6908 suggests that around 69.08% of the variance in the log house prices is explained by the model. The adjusted R-squared value, accounting for the number of predictors, is approximately 0.685.
- F-statistic: With a value of 120.5 on 14 and 755 degrees of freedom, and a p-value less than 2.2e-16, the model is considered statistically significant.

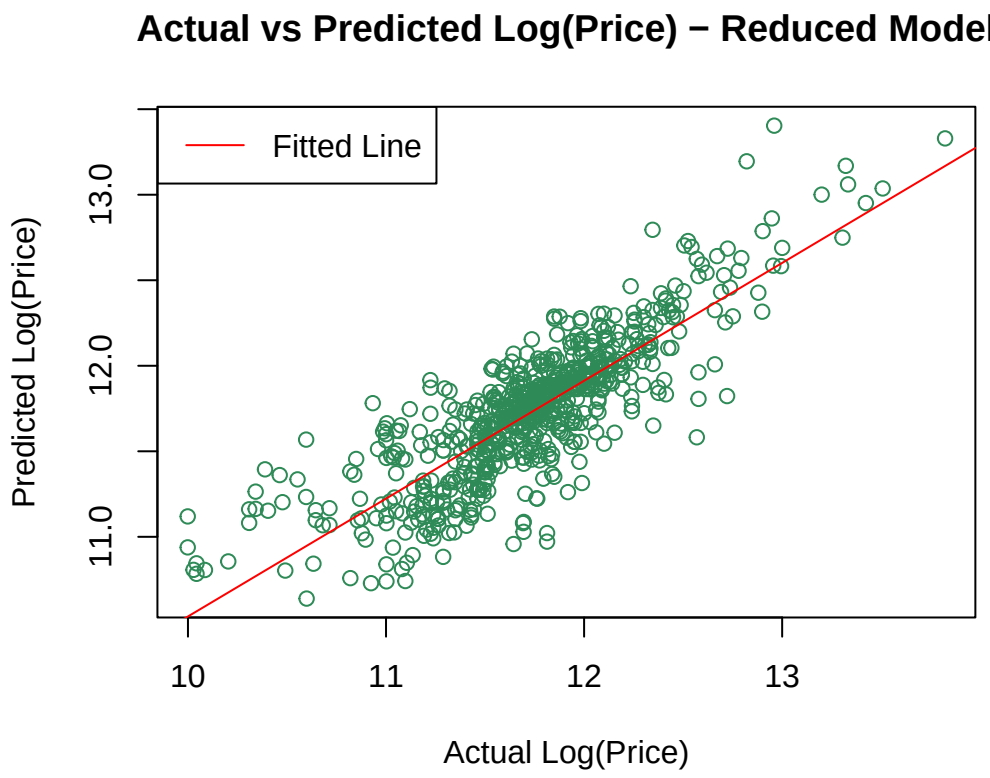
In summary, square footage, number of bathrooms, age of the house, specific architectural styles, presence of a fireplace, and waterfront location are significant predictors of house prices in the Baton Rouge housing market. This analysis provides valuable insights for stakeholders involved in real estate transactions and policy-making in the region.

Using feature selection to produce a reduced model:

The purpose of feature selection is to identify and retain only the most important predictors in the model while discarding those that are less relevant or redundant. Backward stepwise regression is a common technique for feature selection in linear regression models. It starts with a full model containing all potential predictors and iteratively removes predictors that contribute the least to the model until no further improvement in the model fit is observed.

On performing Backward stepwise regression on the full model, we get the following performance:

- The residual standard error (0.2737) represents the standard deviation of the residuals, which are the differences between observed and predicted values of the response variable.
- The multiple R-squared (0.689) indicates the proportion of variability in the response variable that is explained by the model's predictors.
- The adjusted R-squared (0.6853) adjusts the multiple R-squared value for the number of predictors in the model.
- The F-statistic (187.1) and its associated p-value ($< 2.2e\text{-}16$) test the overall significance of the model, indicating whether at least one predictor variable has a non-zero coefficient.



The above plot shows the Actual Log(Price) vs Predicted Log(Price) of the reduced model and we can clearly see that the model is able to make quite good predictions and there are very few outliers that can be observed on the plot. We can also see that the reduced model's plot is very similar to the Full Model's plot and this signifies that using the feature selection method we reduced the complexity of the model but this did not cause significant deterioration of the Model's Predictions.

Interpretation:

Inference:

- Based on the coefficients and their significance, we can infer that square footage(log(sqft)), number of bathrooms, property age, architectural style, presence of a fireplace, and waterfront location are all important factors in determining the price of a property.
- The model suggests that larger square footage, more bathrooms, newer properties, specific architectural styles, presence of a fireplace, and waterfront location tend to be associated with higher property prices.

Justification:

- Backward stepwise regression is a well-established method for feature selection in linear regression models.
- It systematically evaluates the contribution of each predictor to the model and removes those that do not significantly improve the model's fit.
- By starting with the full model and iteratively removing predictors, backward stepwise regression ensures that only the most important predictors are retained in the reduced model.
- This method helps prevent overfitting by avoiding the inclusion of unnecessary predictors that may lead to model complexity without improving predictive performance.

Using k-fold cross validation to compare the two models:

Using the k-fold cross-validation method we split the dataset into k=10 folds and train both the full and the reduced model. We evaluate the performance of both the models and compare their performances on multiple metrics such as MAE, RMSE and R-Squared.

While training the reduced model we are using the variables - **log(sqft), baths, age, style2, style4, fireplace, waterfront**

The summary of the comparison is given below:

```
##
## Call:
## summary.resamples(object = compare_models)
##
## Models: Full_Model, Reduced_Model
## Number of resamples: 10
##
## MAE
##                     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## Full_Model    0.1830893 0.1894528 0.1987059 0.1998347 0.2106264 0.2179510    0
## Reduced_Model 0.1703988 0.1842487 0.1989082 0.1984566 0.2121445 0.2267791    0
##
## RMSE
##                     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## Full_Model    0.2448733 0.2652714 0.2754677 0.2776751 0.2946288 0.3124925    0
## Reduced_Model 0.2304954 0.2524413 0.2826170 0.2753442 0.2954585 0.3147867    0
##
## Rsquared
##                     Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## Full_Model    0.5889480 0.6601587 0.6797115 0.6799170 0.7136204 0.7321847    0
## Reduced_Model 0.6194201 0.6494863 0.6809578 0.6835306 0.7091139 0.7524687    0
```

Observations:

Mean Absolute Error (MAE): The reduced model slightly outperforms the full model in terms of MAE, indicating that it makes slightly more accurate predictions on average.

Root Mean Squared Error (RMSE): Again, the reduced model shows slightly better performance, with lower RMSE compared to the full model.

R-squared (Rsquared): The reduced model exhibits a higher mean R-squared value, indicating that it explains a slightly higher proportion of the variance in the target variable compared to the full model.

Thus, based on the MAE, RMSE and R-Squared metrics the reduced model appears to perform mildly better than the full model in terms of predictive accuracy. And although the difference in small, we should choose the reduced model as it offers reduced model complexity than the full-model.

Conclusion

Our comprehensive analysis of the Baton Rouge housing market has yielded significant insights into the factors that shape property values. By employing a multifaceted approach that included data exploration, correlation analysis, and multiple regression modeling (including both backward stepwise regression for feature selection and k-fold cross-validation for performance evaluation), we have identified key determinants that significantly influence house prices.

Square footage, number of bathrooms, property age, architectural style, fireplace presence, and waterfront location emerged as crucial predictors of house price variation. These findings provide valuable guidance for stakeholders across the Baton Rouge real estate landscape, including prospective buyers, sellers, and policymakers.

Our modeling approach has yielded not only a comprehensive understanding of the housing market dynamics but also a streamlined predictive model. The reduced model, achieved through backward stepwise regression, offers a balance between model simplicity and effectiveness in house price prediction. This fosters enhanced decision-making capabilities within the real estate sector.

Furthermore, k-fold cross-validation confirms the efficacy of the reduced model, demonstrating slightly superior predictive accuracy while maintaining model parsimony compared to the full model.

In essence, this report offers a robust and informative resource for all those involved in the Baton Rouge housing market. It delivers actionable insights alongside a reliable methodology for comprehending and navigating the intricacies of property valuation. By leveraging these findings, stakeholders can make well-informed decisions that align with their specific goals and contribute to the sustainable growth of the Baton Rouge housing market.