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

Sukanya Das (2321248)

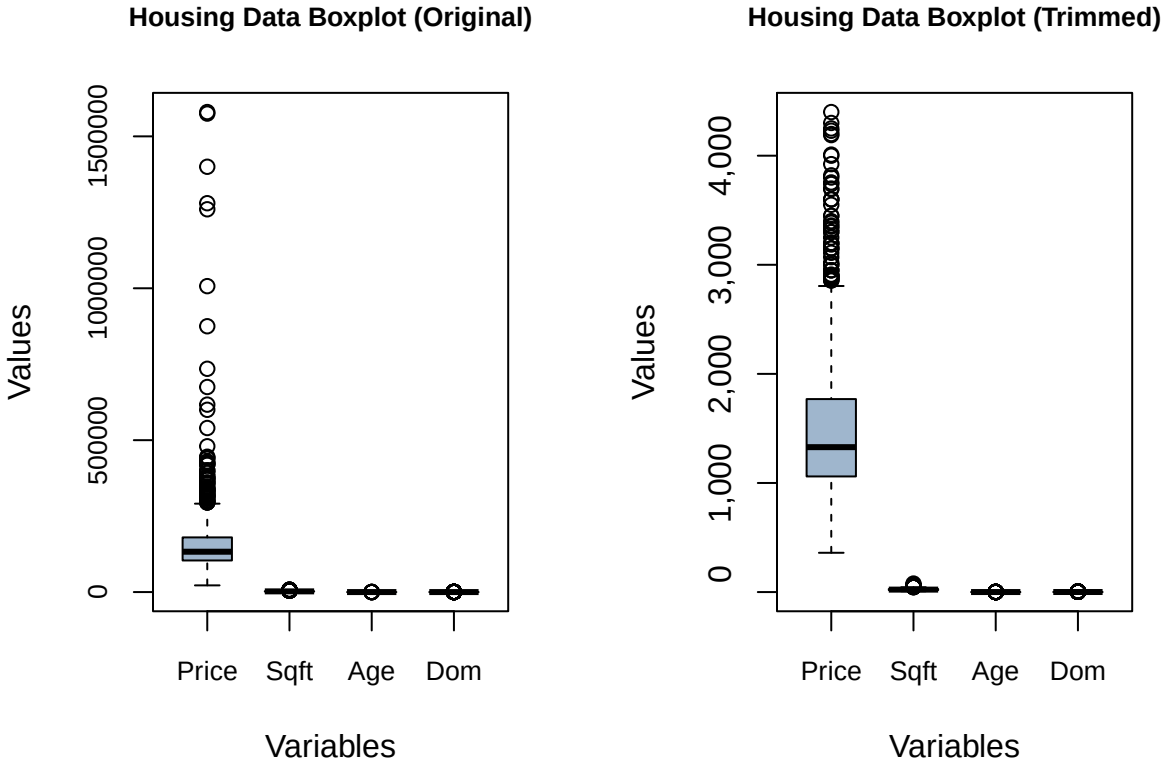
1. Data Exploration

Summary of the dataset

The housing dataset used represents the factors affecting house prices in the USA (Baton Rouge, Louisiana). It includes a total of 10 variables and 820 observations, both qualitative and quantitative.

- Qualitative variables that represent categories are pool, style, fireplace, and waterfront. If the house has a pool, it's a 1 or else 0, and the same goes for a fireplace and waterfront. The style of the house is differentiated based on certain criteria, such as Traditional, Townhouse, Ranch, New Orleans etc.
- Quantitative variable that represent quantities are price, sqft, bedrooms, baths, age, and dom.

Descriptive Statistics



As shown in the above boxplot "Housing Data Boxplot(Original)", there are extreme values or outliers for the price variable. These outliers are removed from the original dataset as shown in "Housing Data Boxplot(Trimmed)" and a new dataset named "trimmed_dataset" is created, which is used in all the subsequent scenarios.

For descriptive statistics, the mean, median and standard deviation of price, sqft, age, and dom are calculated, and below are the observations:

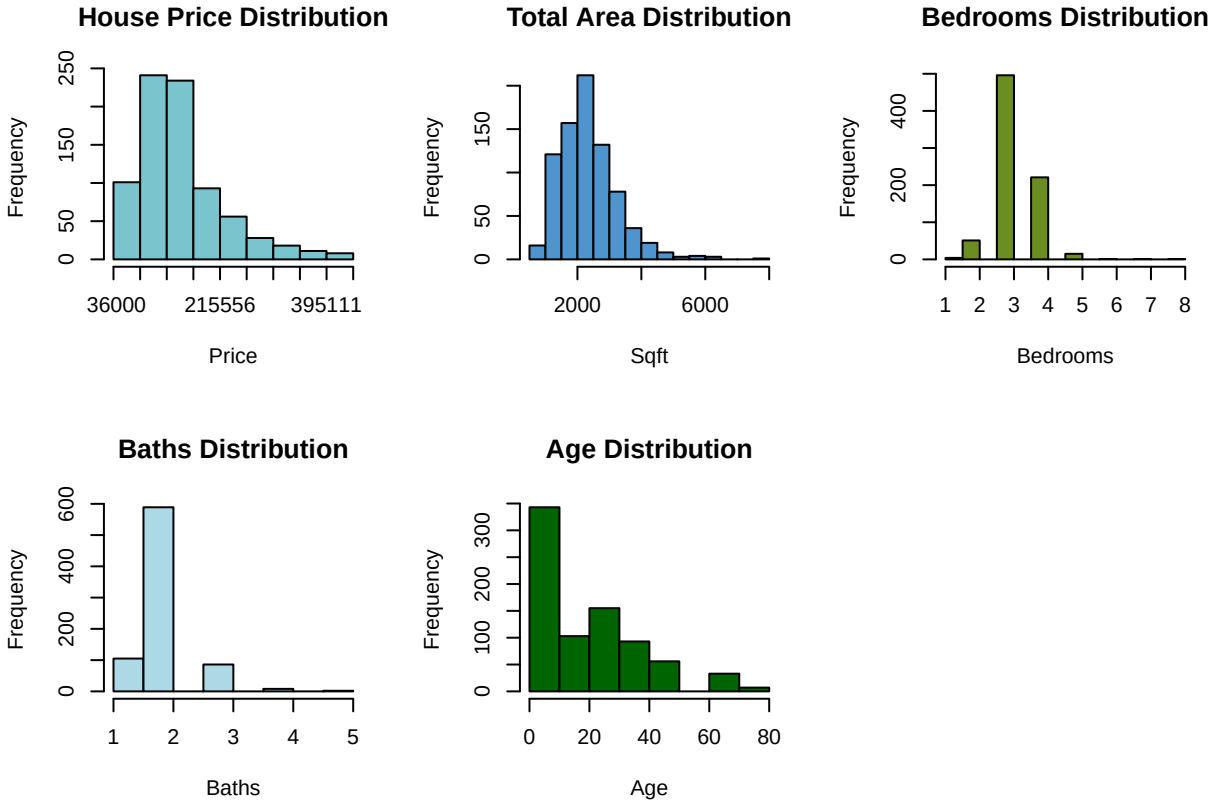
- In this scenario, a trimmed mean price is considered, which is marginally higher than the median, indicating that the price distribution is skewed to the right. The high standard deviation indicates that there's a variation in housing prices, with some values being much higher or lower than the mean.
- The median of the sqft shows that half of the properties have a size less than or equal to the mean, while the rest of the half have a size greater than or equal to the mean. The standard deviation indicates that there's a moderate variation in the house's size.
- There's a slight skewness towards newer properties as the mean age is slightly higher than the median. The standard deviation represents that there's a variation in the age of houses, with some being significantly older or newer than the average.
- There's a right skewed distribution of length of time on the market, as per the calculated mean, which is higher than the median. The standard deviation shows that there's a significant range in the amount of time for houses to sell, where some properties are selling quite quickly, while others are taking longer than expected.

Housing Variables	Mean	Median	Standard Deviation
Price (in dollars)	141259.7	132835	71604.34
Sqft (in square feet)	2361.54	2245	896.77
Age (in years)	19.17	18	17.57
Dom (in days)	75.59	42	96.68

Plots

Suitable plots are generated to represent the distributions of the variables.

- Price distribution indicates that the majority of the properties lie in the price range below 215556(in dollars). It also suggests that the high frequency of lower-priced properties is grouped towards the left side of the graph.
- Sqft distribution shows that the highest frequency of properties is observed around the lower to middle sqft range, which decreases as the square feet increase.
- Bedrooms distribution represents that the majority of the property has three bedrooms, followed by four and then two.
- Baths distribution indicates that houses with two bathrooms are the most common, and there's a drop in the frequency as the number of bathrooms increases.
- Age distribution shows that with older properties, the frequency decreases. Also, newer properties are located, with a high frequency for houses that are around 10 years.



Correlation

The following observations are made after calculating the correlations between some of the variables.

- The correlation between price and sqft is 0.79, which shows that there's a strong positive linear relationship between these variables. This implies that the variation in the total area is linked to a comparatively steady variation in the house's price.
- The correlation between age and dom is -0.07, which represents a very weak linear relationship. This suggests that there may be a slight tendency for newer houses to spend less time on the market.
- The correlation between price and dom is 0.11, which shows a weak positive correlation. There could be a little tendency for higher house prices to be linked with slightly longer days on the market, or the other way around.

2. Probability, probability distributions and confidence intervals

Probability of having a pool is 0.07 and conditional probability of having a fireplace given a pool is 0.77.

Based on the probability of a pool, 'dbinom' determines the probability of having precisely 3, 4, 5, up to 10 houses with a pool out of a sample size of 10 houses. The sum is used to obtain the total probability with at least 3 pools out of 10 properties. Thus, the probability of having at least 3 pools out of 10 houses is 0.03.

The 95% confidence interval on the mean house price in the USA is [145460.6, 155462.3] (in $).

3. Contingency tables and hypothesis tests

Hypothesis Test

A hypothesis test is performed using a 5% significance level that the mean house price (over all house styles) is greater if a house is on the waterfront. In this scenario, a two-sample t-test is used to conclude the result.

Furthermore, the Null hypothesis: the mean house price is the same for both waterfront and non-waterfront houses, and the Alternative hypothesis: the mean house price (over all house styles) is greater if a house is on the waterfront, are considered.

Based on the result, i.e., ~1, we can accept the null hypothesis as there's not enough evidence to suggest that the mean house price is greater for houses on the waterfront compared to the non-waterfront houses.

Contingency table

Contingency Table showing relative frequencies for "Pool" and "No pool" according to whether a house has or hasn't got a fireplace.

Parameters	No Fireplace	Fireplace
No Pool	0.4	0.53
Pool	0.02	0.05

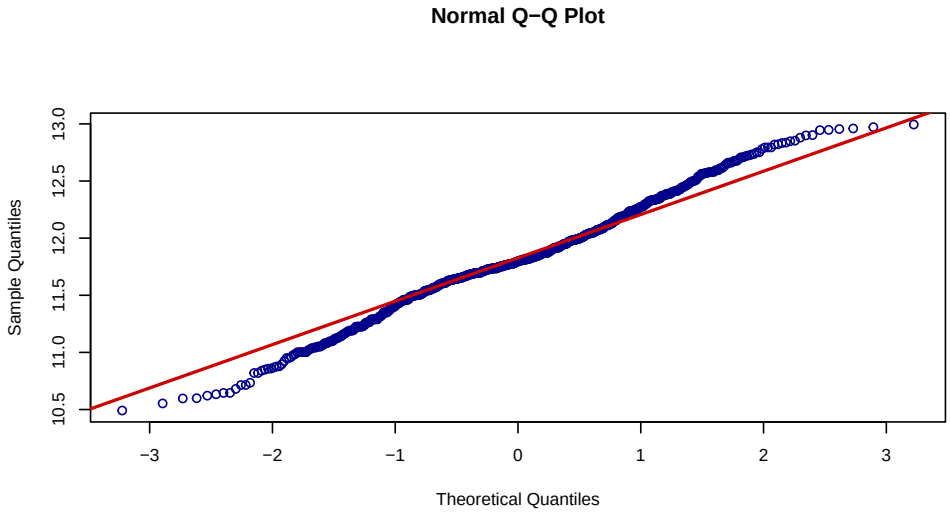
Test for Fireplace and Pool's Independence

Using a 5% significance level, a test has been conducted to show whether a house with a fireplace is independent of whether it has a pool. For this purpose, Pearson's chi-squared test is used to conclude the result.

Also, the Null hypothesis that the fireplace is independent of the existence of a pool and the Alternative hypothesis that the fireplace is not independent of whether it has a pool are considered.

Based on the result, i.e., ~0.01, we can reject the null hypothesis as there's enough evidence to prove that the fireplace is not independent of the presence of a pool.

4. Simple Linear Regression



From the above graph, we can observe that the data is almost normalized. So, both the Linear Regression can be performed based on this assumption.

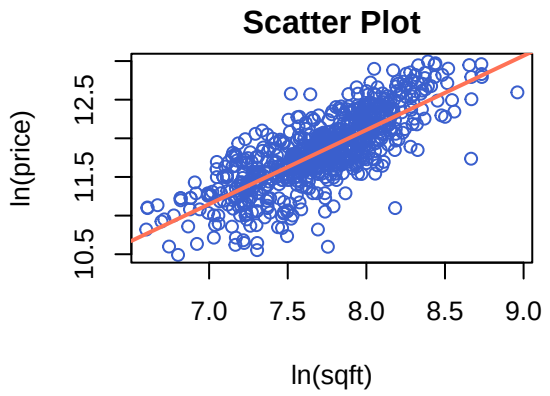
Note : In R, log() can be used to compute the natural logarithm (ln) of a numeric value. The default base for the log() function in R is "e".

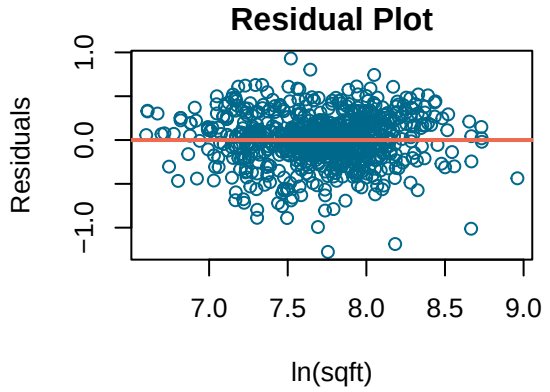
Total area significance: As shown below, it can be observed that the total area is a significant predictor of house price at a 5% significance level. The small p-value of the total area related with the coefficient estimate implies that the predictor variable is highly significant.

Slope Interpretation: From the below observation, slope coefficient indicates that doubling the total area is linked with ~0.96 times increase in the house's price, with all other variables unchanged. Thus, it suggests that the ln(price) increases by 0.96 for each unit increase in the ln(sqft).

Parameters	Estimate	Std. Error	t Value	p Value
Intercept	4.42	0.21	21.39	2.227276e-80
ln(sqft)	0.96	0.03	35.82	1.632069e-167

Scatter plot represents the relationship between the total area of a house (ln(sqft)) on the x-axis and house prices (ln(price)) on the y-axis. In this plot, there's a positive correlation between ln(sqft) and ln(price), as the slope of the fitted line is positive. This suggests that as the size of the house increases, so does its price.





Residual plot indicates that the residuals are randomly scattered around the zero line, which suggests that the model is quite well-fitted to the data. But the points are more dispersed for mid range values of the total area (ln(sqft)) and not much for the extreme data.

In general, the plots suggest that this model is a suitable fit for this dataset. Although additional analysis may be required, the model shows that it fits well with the data.

5. Multiple Linear Regression

Multiple linear regression of ln(price) against all the predictor variables is performed. The fitted model summary shows that the predictor variables are significantly related to the ln(price). As per the summary data, a one-unit increase in the ln(sqft) corresponds to a 0.69 increase in the ln(price). Also, other factors that have a major impact on house prices are baths, age, dom, fireplace, and style. Hence, the model shows the adjusted R-squared value 0.73 of the variance in house prices, suggesting a comparatively good fit. Moreover, the residuals are normally distributed and do not exhibit any systematic patterns.

Afterwards, feature selection is used to produce a reduced model. Step-wise selection with the AIC method is applied in order to get the most effective subset of predictors. The direction is specified as "both" to perform both forward and backward selection.

Thereafter, using k-fold cross validation, the performance of the full and reduced models is evaluated. In this case, 10 fold cross-validation is performed to get more accurate estimate of the model performance. The performance of the full and reduced models is evaluated using Root Mean Squared Error as the performance metric. Based on the below criteria, model performance are compared:

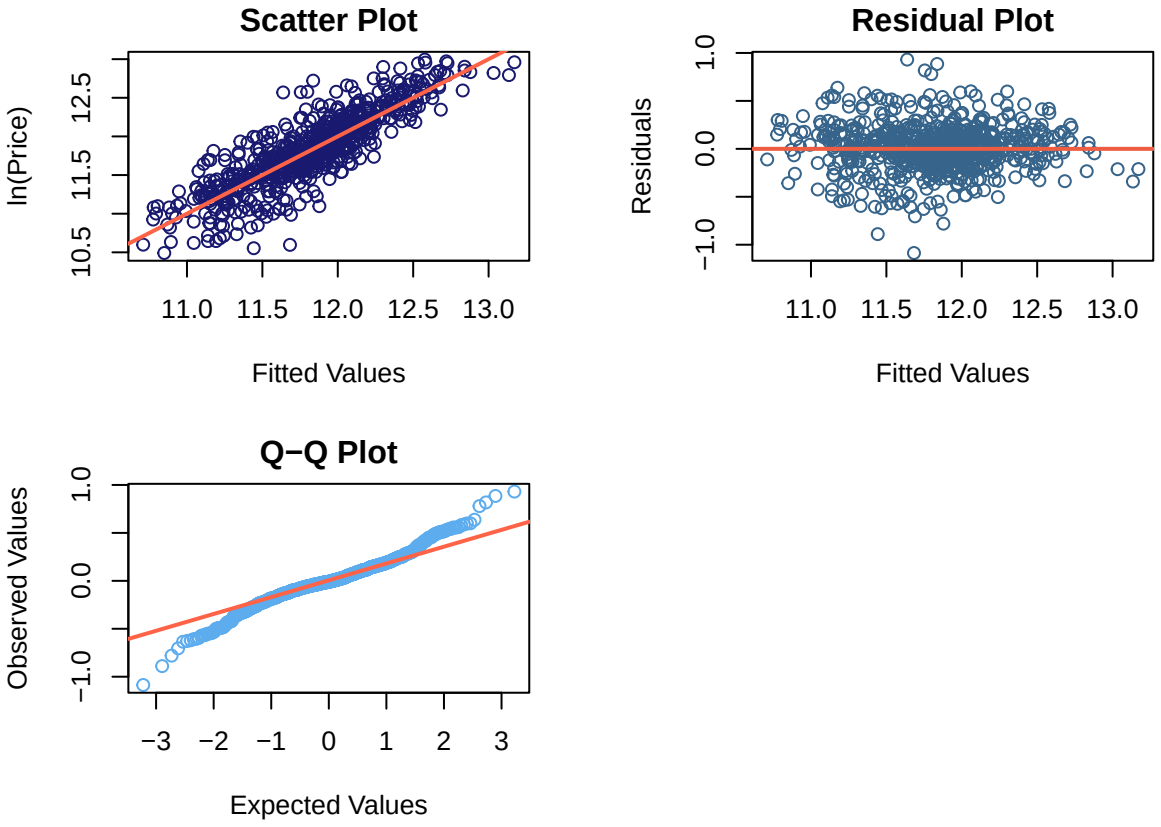
- Model performance is better if the RMSE values are lower.

- Lower AIC values suggest a better model fit, so the model with the lower AIC value is preferred.

As per the below table, we can conclude that the full model has a lower RMSE than the reduced model. Hence, the full model performs better as compared to the reduced model when it comes to predicting house prices. Also, the full model has a lower AIC value, which suggests that the full model is a better fit for the data compared to the reduced model.

Models	RMSE	AIC
Full	0.23	-45.98
Reduced	0.26	136.61

Scatter Plot shows the relation between the fitted values on the x-axis and the house prices ln(price) on the y-axis. The data points are scattered around a line, showing a positive linear relationship.This implies that the model's predictions are roughly proportional to the actual house prices using natural logarithm.



Residual Plot illustrates that the points are randomly scattered around the zero line, suggesting the model fits the data quite well. But the residuals are slightly spread out with increase in fitted values, suggesting instances where the variance of the residuals may not be constant.

Q-Q Plot represents that the points are mostly mapped with the line and normally distributed. However, there's slight deviation from the line at both the lower and upper ends of the plot, showing issues with the normality of the residuals.

Overall, the model seems to be relatively accurate in predicting house prices and a good fit for this dataset.