Home Price Insights: A Data Story

Vishal Bachal

INTRODUCTION

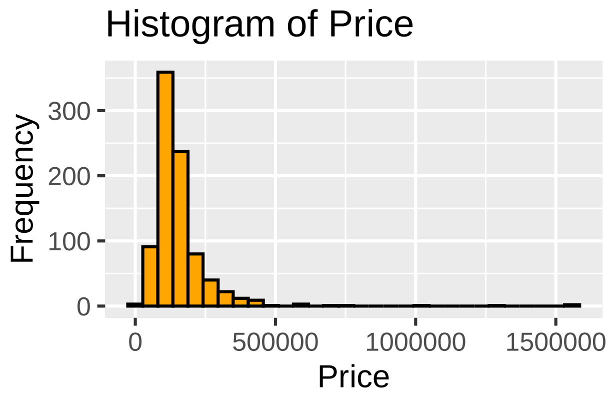
In this report, we'll be looking at how different factors affect house prices in Baton Rouge, Louisiana, USA. By analyzing data, we hope to understand the relationships between these factors and housing prices. This study aims to provide insights into the local housing market dynamics.

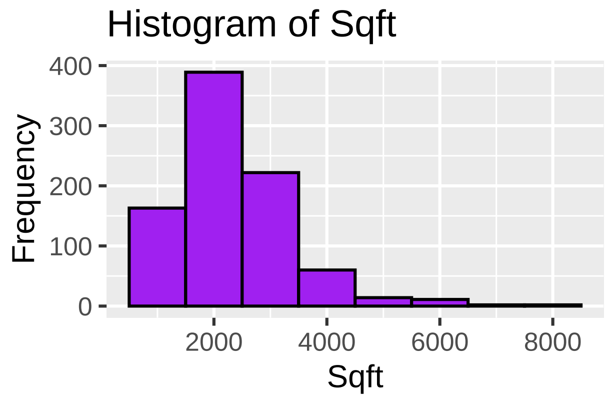
Data exploration

In the data exploration process, I calculated descriptive statistics, such as standard deviation (SD), and compiled a summary of the variables. Furthermore, by using data visualization tools like scatter plots and histograms plots to understand data distributions and relationships. In addition, I looked into correlations between variables to find trends and dependencies.

```
     price              sqft         bedroms          baths      
 Min.   :  22000   Min.   : 662   Min.   :1.000   Min.   :1.000  
 1st Qu.: 102000   1st Qu.:1698   1st Qu.:3.000   1st Qu.:2.000  
 Median : 132000   Median :2226   Median :3.000   Median :2.000  
 Mean   : 156519   Mean   :2373   Mean   :3.236   Mean   :2.003  
 3rd Qu.: 172750   3rd Qu.:2852   3rd Qu.:4.000   3rd Qu.:2.000  
 Max.   :1580000   Max.   :7897   Max.   :8.000   Max.   :5.000  
      age             pool             style          fireplace     
 Min.   : 1.00   Min.   :0.00000   Min.   : 1.000   Min.   :0.0000  
 1st Qu.: 2.00   1st Qu.:0.00000   1st Qu.: 1.000   1st Qu.:0.0000  
 Median :18.00   Median :0.00000   Median : 1.000   Median :1.0000  
 Mean   :18.09   Mean   :0.08459   Mean   : 3.253   Mean   :0.5689  
 3rd Qu.:25.00   3rd Qu.:0.00000   3rd Qu.: 7.000   3rd Qu.:1.0000  
 Max.   :80.00   Max.   :1.00000   Max.   :11.000   Max.   :1.0000  
   waterfront           dom        
 Min.   :0.00000   Min.   :  0.00  
 1st Qu.:0.00000   1st Qu.: 15.00  
 Median :0.00000   Median : 42.00  
 Mean   :0.07068   Mean   : 73.79  
 3rd Qu.:0.00000   3rd Qu.:100.50  
 Max.   :1.00000   Max.   :673.00  
```

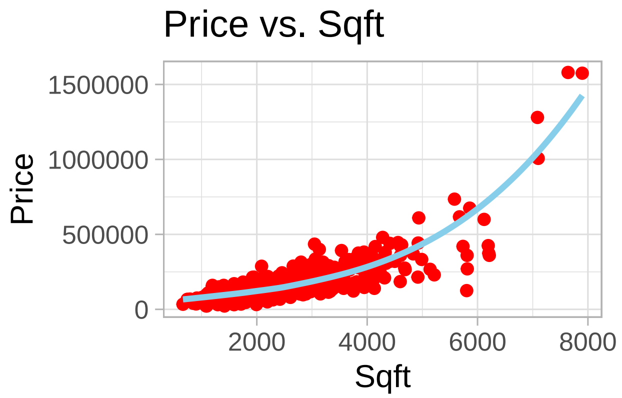
The dataset provides an overview of various aspects of houses in Baton Rouge, Louisiana. The prices range from $22,000 to $1,580,000, with an average of around $156,519. The house sizes vary from 662 square feet to 7,897 square feet, averaging approximately 2,373 square feet. The number of bedrooms ranges from 1 to 8, with an average of roughly 3.2 bedrooms per house, and the bathrooms range from 1 to 5, averaging about 2 bathrooms per house. The ages of the houses range from 1 to 80 years, with an average age of around 18 years.Additionally, around 8.5% of the houses have pools, while approximately 56.9% feature fireplaces. Only about 7.1% of the houses are waterfront properties. The average "Days on Market" (DOM) is approximately 73.8 days, ranging from 0 to 673 days. These statistics provide insights into the housing market dynamics in Baton Rouge, Louisiana, highlighting the diverse range of prices, sizes, amenities, and market durations of the available houses.

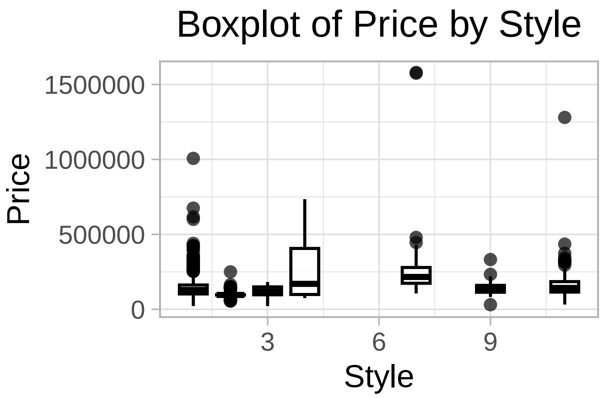




The first histogram shows how home prices are distributed.The distribution of prices in the dataset is shown by the histogram. The pattern is right-skewed, with most prices concentrated at the lower end, perhaps below 100,000 or 200,000. The frequency rapidly declines as prices rise; between 500,000 and 1,000,000, only few occurrences are noted. There are relatively few, if any, prices over 1,000,000, suggesting a lack of data points at the highest price points. Understanding the spread and distribution of home prices in the dataset is made easier with the help of this figure.

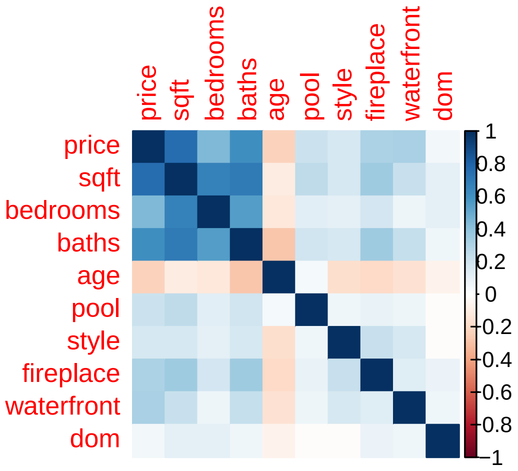
The "Histogram of Sqft" histogram shows how the sqft of the houses in the dataset is distributed visually. With most properties concentrated in lower Sqft ranges, perhaps between 1,000 and 3,000 Sqft, it displays a right-skewed distribution. Larger square footage areas are less common as Sqft values rise, as indicated by a reduction in property frequency.Visibility is improved by the purple bars set against a black outline. Knowing the variation and makeup of sqft is made easier with the help of this figure.





The first plot displays the "Price vs.Sqft" scatter chart. This chart demonstrates the connection between house prices and sqft. Each data point represents the price fluctuations at different square footages. A smoothed line provides insight into overall trends.This chart helps to understand the price patterns in the real estate market, supporting buyers, sellers, and industry professionals in their decision-making process.

The second plot visualizes house prices across architectural styles using boxplots. Each boxplot represents a style's price distribution, with median prices depicted by central lines. The width of the boxes illustrates price variability within each style. This visualization enables a quick comparison of price distributions among architectural styles. Additionally, the varying colors distinguish between different architectural styles, aiding in easy identification.



The correlation coefficient between two variables is indicated by the number inside each square in this diagram, which stands for a pair of variables. There are two possible correlation coefficients: -1 and 1. A perfect positive correlation is represented by a value of 1, a perfect negative correlation by a value of -1, and 0 means there's no correlation between the variables.

Probability, distributions and confidence intervals

```
[1] "Probability of having a pool: 0.0845886442641947"
```

A The probability of randomly selecting a house from the dataset and finding that it has a pool is calculated to be approximately 8.46%.Only 8.46% of the houses in the dataset have a pool, according to this probability estimate, which suggests that the likelihood of any particular property having one is very low. This observation suggests that pools might not be a feature shared by all of the houses , which would make them a desired or less common amenity.

```
[1] "Conditional probability of having a fireplace given that it has a pool: 0.726027397260274"
```

When looking specifically at houses that have a pool, the conditional probability of also finding a fireplace among these houses is relatively high,at approximately 72.6%. There is a significant correlation between the number of fireplaces and pools in the sample, as indicated by the conditional probability value of 72.6%. It suggests that if a house has a pool, there is a high likelihood, around 72.6%, that it will also have a fireplace. There could be a correlation between the availability of fireplaces and pools due to specific housing patterns or preferences.

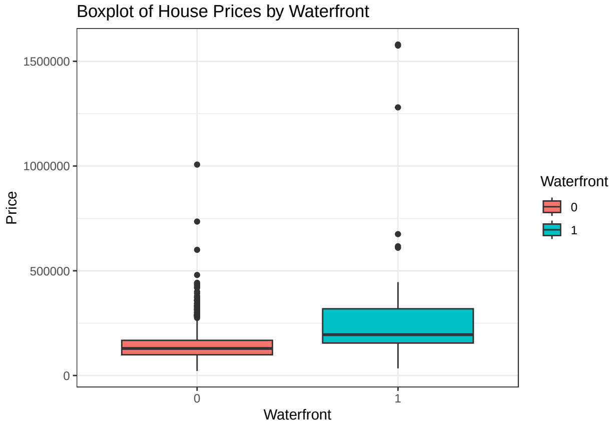
```
[1] "Probability that at least 3 out of 10 houses selected at random have a pool: 0.0462079659329246"
```

When randomly selecting 10 houses from the dataset, the probability of observing at least 3 of these houses having a pool is found to be approximately 4.62%.The distribution of homes with pools within a small subset of 10 homes is revealed by this probability value. Based on a random selection of ten residences, there is a small probability (4.62%) of discovering at least three houses with pools. This probability calculation helps in understanding the likelihood of encountering houses with pools in a small random sample.

```
[1] "95% Confidence Interval for the mean house price: 148726.07671215"
[2] "95% Confidence Interval for the mean house price: 164311.040321454"
```

The estimated 95% confidence interval for the mean house price in the USA is $148,726.08–$164,311.04, which represents the range within which we may express a 95% confidence that the true average house price is located.This interval serves as a gauge of the precision and accuracy of our estimation based on the dataset. The range from $148,726.08 to $164,311.04 represents the likely values for this measure, which helps with analysis and decision-making by capturing the uncertainty involved in predicting the average house price.

Contingency tables and hypothesis tests



The boxplot displays the distribution of home values according to whether or not the property has a view of the waterfront. The median price for non-waterfront houses (0) is around $200,000, while the average price of waterfront homes (1) is around $300,000, which is much more. The waterfront houses also exhibit a wider range of prices, with some outliers reaching over $1,500,000.This shows that, perhaps as a result of the premium attached to such attractive houses, having a waterfront view has a significant impact on house prices in the sample. The broader distribution and higher median for waterfront properties suggest that purchasers are prepared to pay a significant premium for the extra amenity and appeal of being close to the the water.

```
[1] "p-value is 0.000154440490049884"
```

```

    Welch Two Sample t-test

data:  waterfront_prices and non_waterfront_prices
t = 3.8275, df = 60.63, p-value = 0.0001544
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 83551.17      Inf
sample estimates:
mean of x mean of y
 294289.3  146039.7
```

```
[1] "We reject the null hypothesis. The mean house price is significantly greater for waterfront houses
```

In the analysis conducted on the dataset, several key tasks were undertaken to explore relationships and test hypotheses regarding house features. The first task involved testing the hypothesis that the mean house price differs between waterfront and non-waterfront houses. This analysis was carried out using a Two-Sample Welch's t-test, which compared the mean house prices for the two categories. The null hypothesis stated that the mean house price was equal for waterfront and non-waterfront houses, while the alternative hypothesis proposed that the mean house price for waterfront properties was greater. The resulting p-value from the test was calculated to be 0.00015, which falls below the 5% significance level. This indicates that there is strong evidence to reject the null hypothesis, suggesting that the mean house price for waterfront houses is significantly greater than that of non-waterfront houses.

The analysis progressed to construct a contingency table, aiming to explore the correlation between the existence of pools and fireplaces in houses. This table presented the proportional frequencies of houses with and without pools, categorized by the presence or absence of fireplaces. By delineating the distribution of houses across different groupings, the contingency table provided valuable insights into the potential relationship between fireplaces and pools within the dataset,offering valuable insights into their co-occurrence.

```
             0          1
  0 0.94623656 0.05376344
  1 0.89205703 0.10794297
```

Subsequently, a chi-squared test for independence was conducted to assess whether the presence of a fireplace in a house was independent of the presence of a pool. The null hypothesis posited that there was no association between fireplaces and pools, while the alternative hypothesis suggested a dependence.

```

    Pearson's Chi-squared test with Yates' continuity correction

data:  Contingency_table
X-squared = 7.3389, df = 1, p-value = 0.006748
```

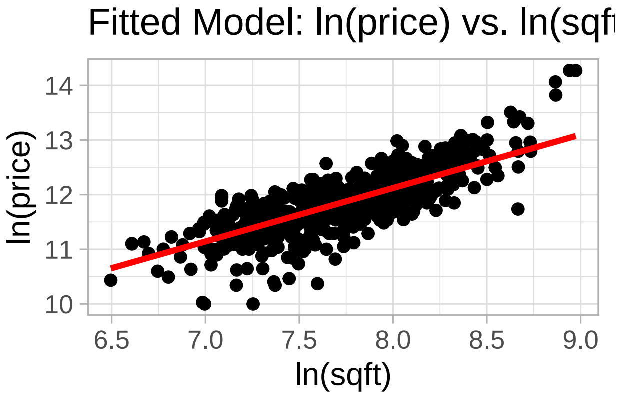
The test results revealed a chi-squared value of 7.3389 with 1 degree of freedom and a p-value of 0.006748. As the obtained p-value was less than the 5% significance level, the null hypothesis was rejected. This implies that the presence of a fireplace in a house is dependent on whether the house has a pool based on the dataset analysis.

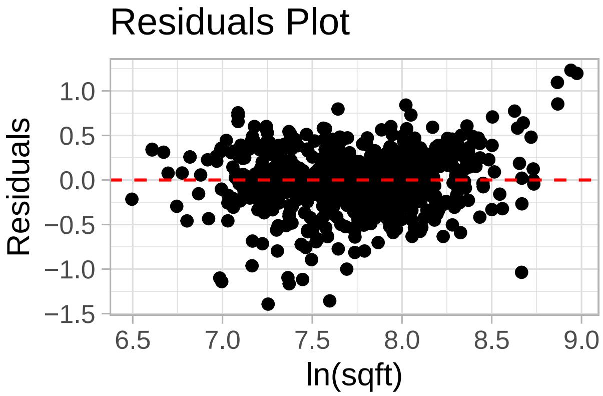
Simple Linear Regression

A simple linear regression was conducted to examine the relationship between the logarithm of house prices and the logarithm of square footage. The regression model aims to predict house prices based on sqft.

```

Call:
lm(formula = log(price) ~ log(sqft), data = My_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.3930 -0.1591  0.0031  0.1873  1.2320 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.30022    0.20509   20.97   <2e-16 ***
log(sqft)    0.97758    0.02662   36.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3103 on 861 degrees of freedom
Multiple R-squared:  0.6103,    Adjusted R-squared:  0.6098 
F-statistic:  1348 on 1 and 861 DF,  p-value: < 2.2e-16
```

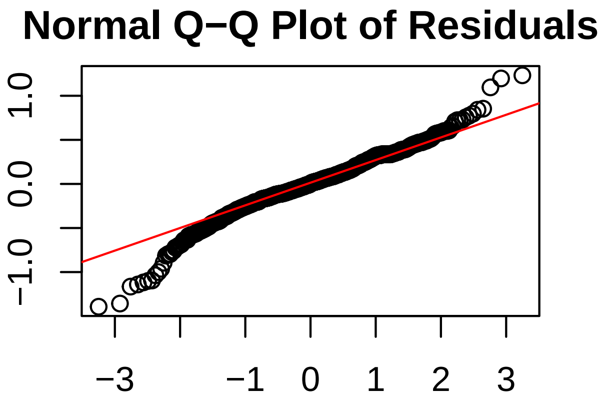
The coefficient of 0.97758 for ln(sqft) indicates that the natural logarithm of house price is predicted to grow by roughly 0.97758 units for every unit increase in the natural logarithm of sqft. When the natural logarithm of sqft is zero, the intercept of 4.30022 represents the expected natural logarithm of house price; however, as the natural logarithm of zero is undefinable, this may not have a significant real-world impact. Supported by substantial t-statistic values and low p-values (< 0.05), the intercept and slope coefficients exhibit statistical significance.Thus, indicating that total area (square footage) serves as a significant predictor of house price. Regarding model fit, the adjusted R-squared value of 0.6098 suggests that approximately 60.98% of the variation in the natural logarithm of house price can be accounted for by the natural logarithm of square footage in the regression model. Additionally, the F-statistic of 1348 with a very low p-value confirms the overall significance of the model.





The first scatter plot clearly shows that the natural logarithm of sqft and the natural logarithm of house price have a linear relationship. The direction and strength of this association are indicated by the fitted linear regression model, which is represented by the red line. This validates the interpretation of the slope coefficient by indicating a positive correlation between sqft and home price.

The second scatter plot shows the differences between the values observed and the values the model predicted are shown in the residuals figure. A reference line for zero residuals aids in assessing the heterogeneity of variance of the model. A random scattering of residuals around zero with no apparent pattern signifies that the model's assumptions are met.



Additionally, the Q-Q plot provides a visual comparison between the residuals' distribution from our regression model and the expected distribution of a normal distribution. A near-linear relationship between the points and the residuals indicates that the distribution of the residuals is approximately normal. This is significant because proper findings from many statistical techniques, such as linear regression, depend on this assumption being true.

Overall,supported by strong statistical tests and visuals showing the relationship between the variables, the simple linear regression analysis indicates that sqft strongly predicts house price.

Multiple Linear Regression

In the multiple linear regression analysis, we initially developed a comprehensive model using all available predictor variables to predict house prices. We then improved this model by using stepwise regression to get a more condensed version that concentrated on important predictors.Through k-fold cross-validation, we compared the full model's performance against the reduced model.

```
Linear Regression

863 samples
  9 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 776, 776, 777, 777, 779, 777, ...
Resampling results:

  RMSE      Rsquared   MAE
  0.255383  0.7382599  0.1904592

Tuning parameter ’intercept’ was held constant at a value of TRUE
```

The analysis conducted a linear regression on housing data with predictors such as square footage, bedrooms, baths, age, pool, style, fireplace, waterfront,and days on market.The following important performance metrics were obtained using 10-fold cross-validation:RMSE (0.256),Rsquared (0.734),and MAE(0.191).These values suggest the model accurately predicts house prices, explaining approximately 73.40% of the variance.The cross-validation approach enhances reliability by testing the model across diverse data subsets.

Our results showed the ability of the entire model to explain variation in house prices, with an RMSE of 0.5055. On the other hand, the reduced model had a slightly higher RMSE of 0.5418, suggesting a trade-off between model complexity and accuracy. The reduced model, while simpler, may sacrifice some precision in exchange for a clearer and more interpretable model.

```
[1] "Full Model RMSE: 0.507300156037783"
```

```
[1] "Reduced Model RMSE: 0.542767375704979"
```

Conclusion

The conclusion of this report is that the housing dataset provides insightful information about the variables affecting home prices. Houses with pools have a much higher possibility of having fireplaces (around 0.726). The probability of choosing 3 randomly from 10 houses with a pool is low, around 0.046.The mean house price's confidence intervals indicate a wide range of possible values, ranging from $148,726 to infinity to $164,311 to infinity. The contingency table's variables are associated with one another, and statistical tests show significant differences in house prices between waterfront and non-waterfront properties. A significant correlation between square footage and house price is evident from the linear regression study, which accounts for around 61% of the variability.Furthermore, cross-validated sampling suggests that prediction accuracy is increased by adding more predictors. These results provide real estate stakeholders with insightful information that successfully directs decision-making processes.