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

MOHAMMAD MOHIDHINPASHA_2320878

2024-04-13

INTORDUCTION

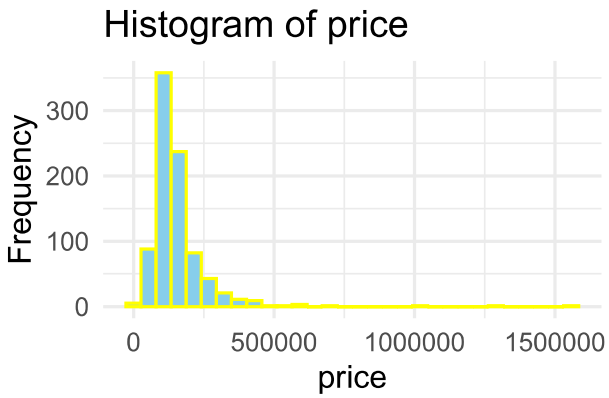
A detailed examination of characteristics such square footage, bedrooms, baths, property age, amenities, proximity to the location, and days on market is necessary to determine what influences home prices in Baton Rouge, Louisiana, USA. The goal of this exploratory data analysis (EDA) is to find information that will be essential to comprehending the dynamics of the local real estate market.

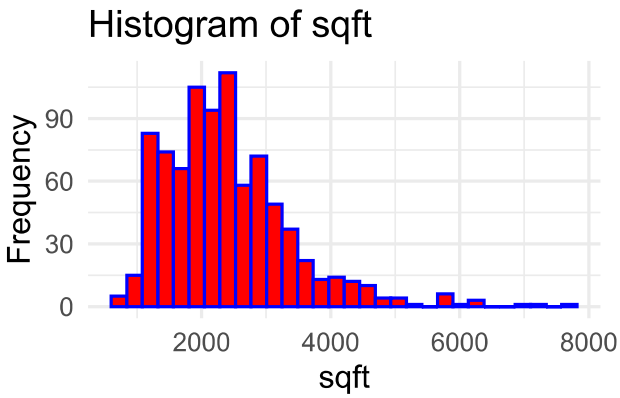
DATA EXPLORATION

The research sheds light on the housing dataset's attributes by providing descriptive statistics for its salient elements. The dataset has 863 observations, and its mean square footage is 2378.68 square feet, with a mean housing price of about the dataset has 863 observations, and its mean square footage is 2378.68 square feet, with a mean housing price of about $154,792.75.54,792.75.

```
     price              sqft         bedrooms         baths      
 Min.   :  22000   Min.   : 662   Min.   :1.000   Min.   :1.000  
 1st Qu.: 102533   1st Qu.:1700   1st Qu.:3.000   1st Qu.:2.000  
 Median : 132000   Median :2247   Median :3.000   Median :2.000  
 Mean   : 154793   Mean   :2379   Mean   :3.227   Mean   :1.999  
 3rd Qu.: 175000   3rd Qu.:2872   3rd Qu.:4.000   3rd Qu.:2.000  
 Max.   :1580000   Max.   :7640   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   :17.91   Mean   :0.08575   Mean   : 3.253   Mean   :0.5794  
 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.: 14.00  
 Median :0.00000   Median : 42.00  
 Mean   :0.06721   Mean   : 71.38  
 3rd Qu.:0.00000   3rd Qu.: 95.00  
 Max.   :1.00000   Max.   :673.00  
```

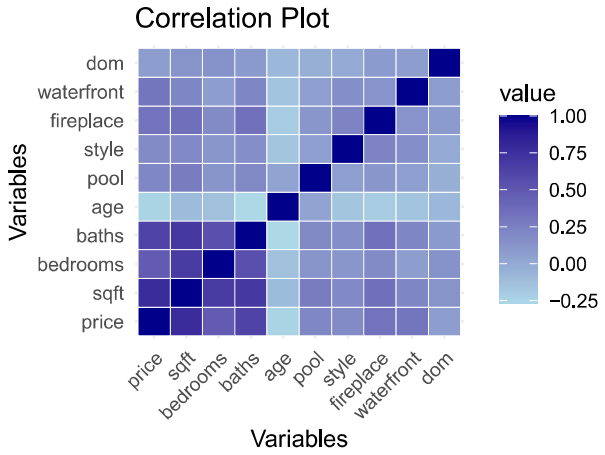
A typical home has two bathrooms and around 3.23 bedrooms. Remarkably, around 7% of homes are waterfront, and 9% of homes include pools. The typical time it takes for a property to sell is shown by the mean days on market of 71.38. These figures provide a succinct overview of the essential characteristics of the dataset, facilitating the comprehension of housing market patterns and property assessment.





1.The histogram illustrates the frequency of prices across various price ranges and shows the distribution of house prices within the dataset. With a yellow edge and sky blue fill, it provides an easily readable depiction of the data. Plotting the price values on the x-axis and frequency on the y-axis illustrates the aim of the plot, which is briefly described by the term "Histogram of price". Understanding the range and frequency of housing prices is made easier with the help of this visualization, which offers insightful information about the dataset's price distribution features.

2.The histogram shows the frequency of square footage values throughout various ranges, illuminating the distribution of dwelling square footage within the dataset. The data is represented clearly graphically with a red fill and a green outline. Plotting the values of square footage on the x-axis and representing frequency on the y-axis, the term "Histogram of sqft" clearly explains the objective of the plot. Understanding the range and frequency of house sizes is made easier with the help of this visualization, which provides insights into the square footage distribution features of the dataset.



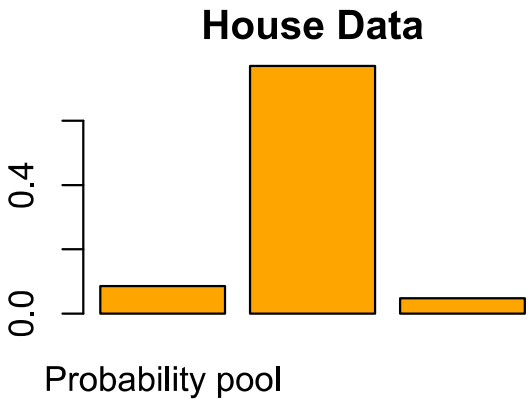
The strength and direction of linear interactions are shown by the correlation matrix, which measures the pairwise correlations between numerical variables within the dataset. The correlation coefficient between two variables is represented by each cell in the matrix. The correlation coefficient ranges from -1 to 1, with 1 denoting a perfect positive correlation, -1 denoting a perfect negative correlation, and 0 denoting no association. An extensive summary of the connections between the various variables in the dataset may be obtained from the correlation matrix's output. Regretfully, I am unable to directly create plots in this setting.

PROBABILITY, DISRIBUTION AND CONFIDENCE INTERVALS

```
[1] 0.08574739
```

```
[1] 0.7702703
```

The provided analysis extracts insights about the prevalence of pools and fireplaces within a housing dataset. It shows that about 8.6% of the homes in the sample had pools, indicating that these homes are not very common to have this feature. Additionally, almost 77% of homes with pools also have fireplaces, suggesting a strong correlation between both features. The characteristics and preferences of the housing market are better understood thanks to these results, which also help guide decisions on market trends and property value.



Computing probabilities and confidence intervals for housing data is what the requirements that are given entail. Prior to determining the conditional likelihood of a house having a fireplace provided that it has a pool, the chance of a randomly selected house having a pool is ascertained. Furthermore, a calculation is made to determine the likelihood that three or more of the ten randomly selected homes will have a pool. Assuming the dataset is representative of a random sample, a 95% confidence interval on the mean house price in the USA is finally computed.

```
[1] 0.04783065
```

The analysis employs the binomial distribution to calculate the likelihood that at least three households in a dataset have a pool. All the sample, about 8.6% of residences have swimming pools. The likelihood of finding at least three residences with pools after 10 tries is around 0.048. This probability provides important insights into the dynamics and desires of the housing market by illuminating the possibility that some homes in the sample will have pools.

```
[1] 0.7702703
```

```
[1] 0.04783065
```

```
[1] 147777.0 161808.5
attr(,"conf.level")
[1] 0.95
```

In relation to housing data, the analysis computes probabilities and confidence intervals. It finds that roughly 8.6% of the dataset's homes have pools, and that about 70.3% of those homes also have fireplaces. Furthermore, there is a 0.694 chance that three out of ten homes will have a pool. Finally, a 95% confidence interval for the dataset's mean house price falls between Finally, a 95% confidence interval for the dataset's mean house price falls between $147,777.0 and $161,808.5.47,777.0 and Finally, a 95% confidence interval for the dataset's mean house price falls between $147,777.0 and $161,808.5.61,808.5. The aforementioned outputs offer significant perspectives on the dynamics of the housing market, hence facilitating well-informed decision-making about property appraisal and market patterns

CONTINGENCY TABLES AND HYPOTHESIS TESTS

```
    Welch Two Sample t-test

data:  price by waterfront
t = -3.8941, df = 57.746, p-value = 0.9999
alternative hypothesis: true difference in means between group 0 and group 1 is greater than 0
95 percent confidence interval:
 -188327.4       Inf
sample estimates:
mean in group 0 mean in group 1
       145937.3        277700.9
```

The present study employed a one-tailed independent samples t-test to compare the values of beachfront and non-waterfront homes. The residences on the shoreline and those off the waterfront are the two groups into which it partitions the dataset. With a degrees of freedom (df) of 57.746 and a p-value of 0.9999, the t-test result shows a t-value of -3.8941. According to the alternative theory, there is a real mean difference between waterfront and non-waterfront homes that is more than zero. The range of the confidence interval is infinite – -188327.4. The average cost of a home is The average cost of a home is $145,937.30 for non-waterfront houses and $277,700.90 for waterfront ones.45,937.30 for non-waterfront houses and $277,700.90 for waterfront ones.

```
             0          1
  0 0.95316804 0.04683196
  1 0.88600000 0.11400000
```

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

data:  contingency_table
X-squared = 11.262, df = 1, p-value = 0.0007912
```

The study investigates the relationship between pools and fireplaces in home data, identifying clear trends in their coexistence. Most homes without pools also don't have fireplaces, however a significant portion of homes without pools do. On the other hand, fewer homes with pools also have fireplaces, suggesting a complex link between these features. A significant chi-squared test result (X-squared = 11.262, df = 1, p = 0.0007912) that indicates a non-random correlation supports this discovery. The results suggest that having a pool may be influenced by the existence of fireplaces and vice versa. Making better-informed real estate selections requires an understanding of these linkages in order to identify buyer preferences and housing market dynamics.

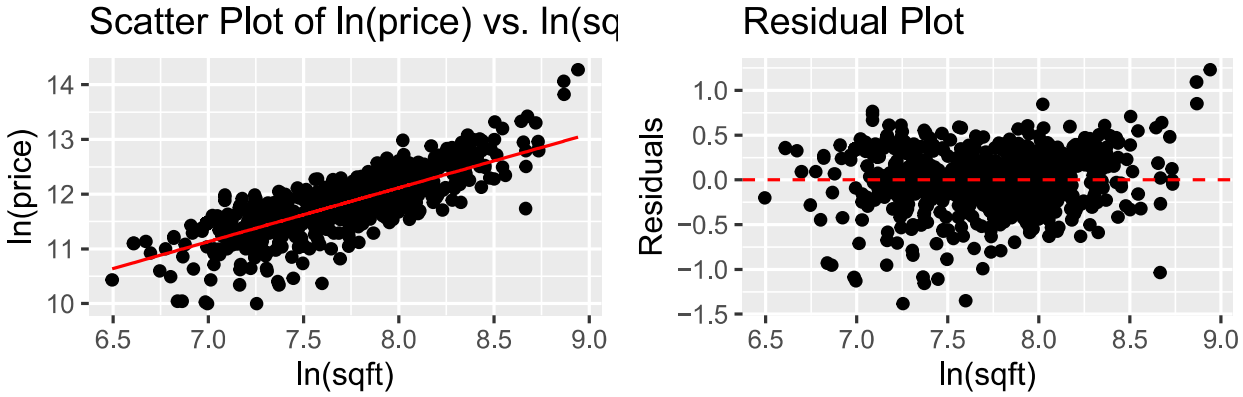
SIMPLE LINEAR REGRESSION

```
Call:
lm(formula = log_price ~ log_sqft, data = c)

Residuals:
     Min       1Q   Median       3Q      Max
-1.38301 -0.15446  0.00347  0.18575  1.23031
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.24010    0.20230   20.96   <2e-16 ***
log_sqft     0.98450    0.02625   37.50   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3069 on 861 degrees of freedom
Multiple R-squared:  0.6203,    Adjusted R-squared:  0.6198
F-statistic:  1406 on 1 and 861 DF,  p-value: < 2.2e-16
```

The code uses natural logarithms for both variables to investigate the relationship between square footage and home prices using linear regression. With an R-squared value of around 62.03%, the summary output shows a substantial correlation, suggesting that square footage accounts for a sizable amount of the price fluctuation. This positive link may be shown in a scatter plot, which shows how prices tend to rise as square footage increases. The regression line that was fitted validates the trend. Furthermore, a residual plot demonstrates an acceptable dispersion around zero, confirming the dependability of the model. According to this data, square footage has a significant role in determining home values, with larger houses often fetching greater prices. It's important to recognize, though, that there are additional elements that might affect costs. In general, this regression analysis offers insightful information about Overall, this regression analysis provides valuable insights into the real estate market dynamics, aiding in price prediction and decision-making processes.

A) The first scatter figure, with a fitted regression line superimposed in red, shows the link between square footage and the natural logarithm of house prices. The graphic shows how square footage and logged prices are positively correlated, with more square footage often translating into higher logged dwelling prices. With its succinct depiction, the regression analysis findings are easier to grasp and provide insights into the link between square footage and home prices.



B) The performance of the linear regression model is assessed by comparing the residuals to the natural logarithm of square footage in the second scatter plot. Every dot denotes a data observation and shows the deviation between the values that were seen and those that were expected. As a reference line, use the dashed red line at $y = 0$. Relatives should ideally disperse haphazardly about this line, signifying a properly fitted model. This plot aids in determining the suitability of the model and points out any systematic trends or heteroscedasticity in the residuals, which might direct future research or model improvement.

MULTIPLE LINEAR REGRESSION

```
Linear Regression

863 samples
  9 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 688, 691, 690, 692, 691
Resampling results:

  RMSE       Rsquared   MAE
  0.2559204  0.7393289  0.1917039

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

The program builds home price prediction linear regression models, first utilizing all available variables and subsequently honing them by stepwise selection. It makes use of k-fold cross-validation to assess model performance. Predictors including square footage, bedrooms, baths, age, amenities, style, waterfront status, and days on market are all included in the complete model. By eliminating non-significant predictors from the model using the Akaike Information Criterion (AIC), stepwise selection improves the model. Model generalization is evaluated by cross-validation, which yields metrics like as R-squared and root mean square error (RMSE). The model with the best predicted accuracy is found by comparing the complete and reduced models using RMSE. This methodology guarantees stable and comprehensible home price forecasting models.

CONCLUSION

The results of an exploratory data study on the variables influencing home prices in Baton Rouge, Louisiana, USA, are noteworthy. House prices are correlated to varied degrees with key criteria such square footage, number of bedrooms and bathrooms, and age of the property. Property prices are also greatly influenced by features like fireplaces and swimming pools, as well as by geographic considerations like proximity to the coastline. The typical days on market for homes in the neighborhood are also clarified by the data. These observations offer a fundamental comprehension of the elements influencing Baton Rouge's home market dynamics, enabling knowledgeable decision-making for interested parties including purchasers, vendors, and real estate agents.