FACTORS AFFECTING HOUSE PRICES IN THE USA (BATON ROUGE, LOUISIANA)

LILIAN OKPO

2024-08-20

SET A CRAN MIRROR FOR KNITTING

INSTALL R PACKAGES

DATA CLEANING AND PREPROCESSING

OBSERVATION: There are no missing values in the dataset.

QUESTION ONE: DATA EXPLORATION

(1a) A brief summary of the variables in the dataset:

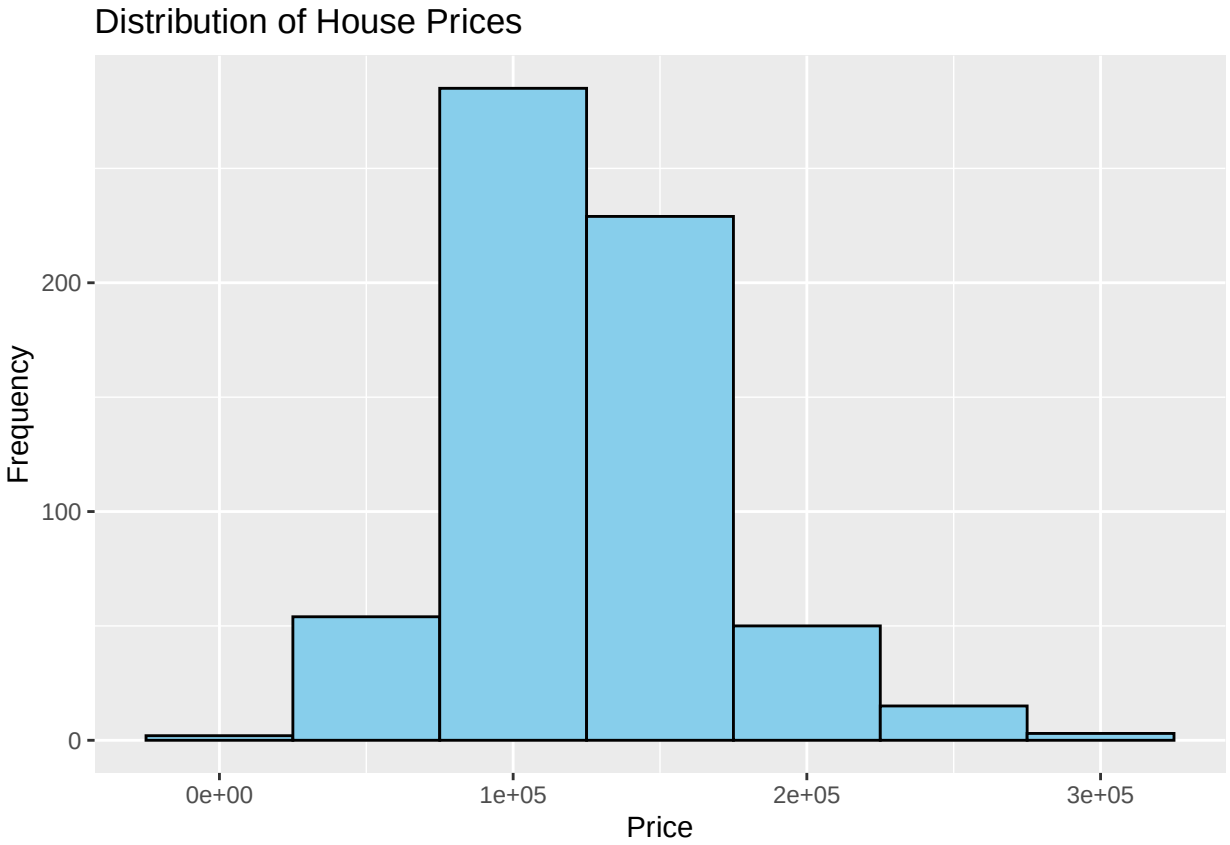
OBSERVATIONS: There are 793 Observations and 10 variables in the dataset and the data types are Integers. There are no qualitative variables.

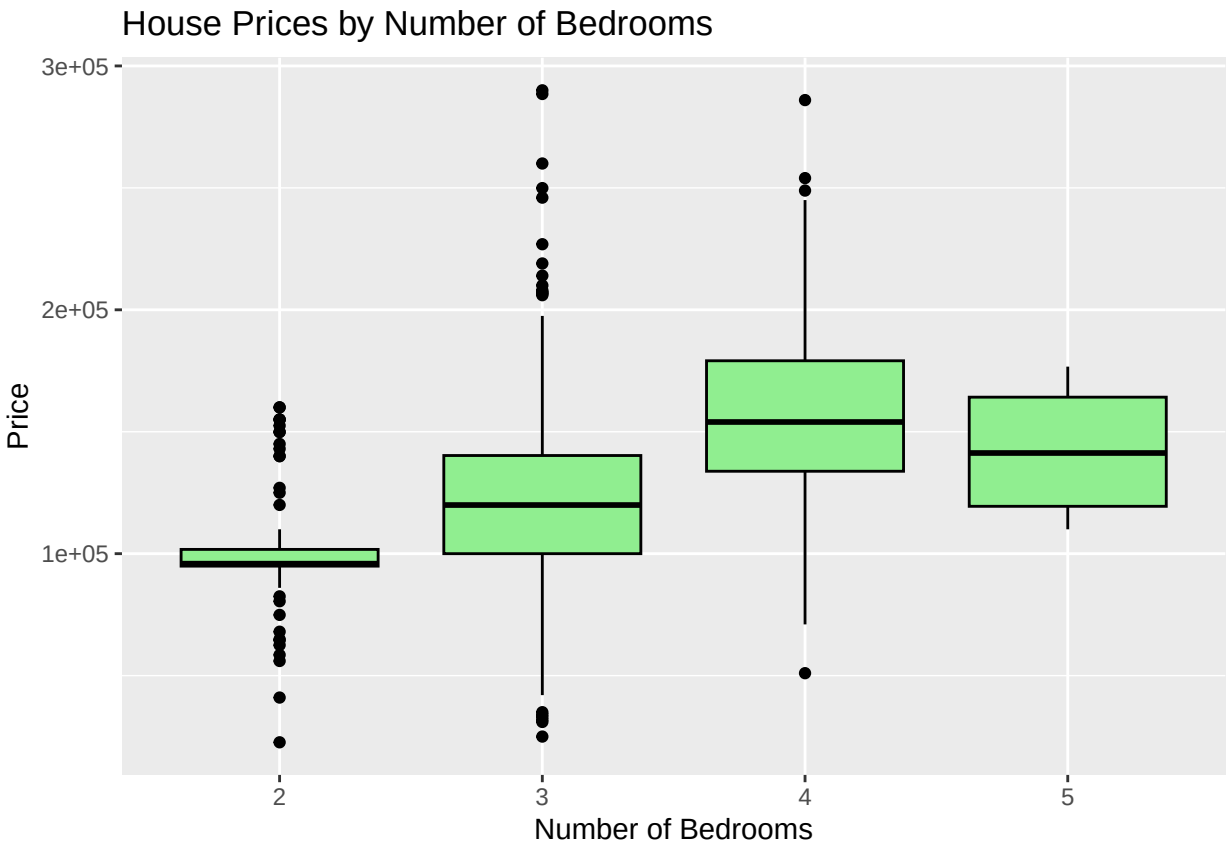
(1b) Descriptive statistics:

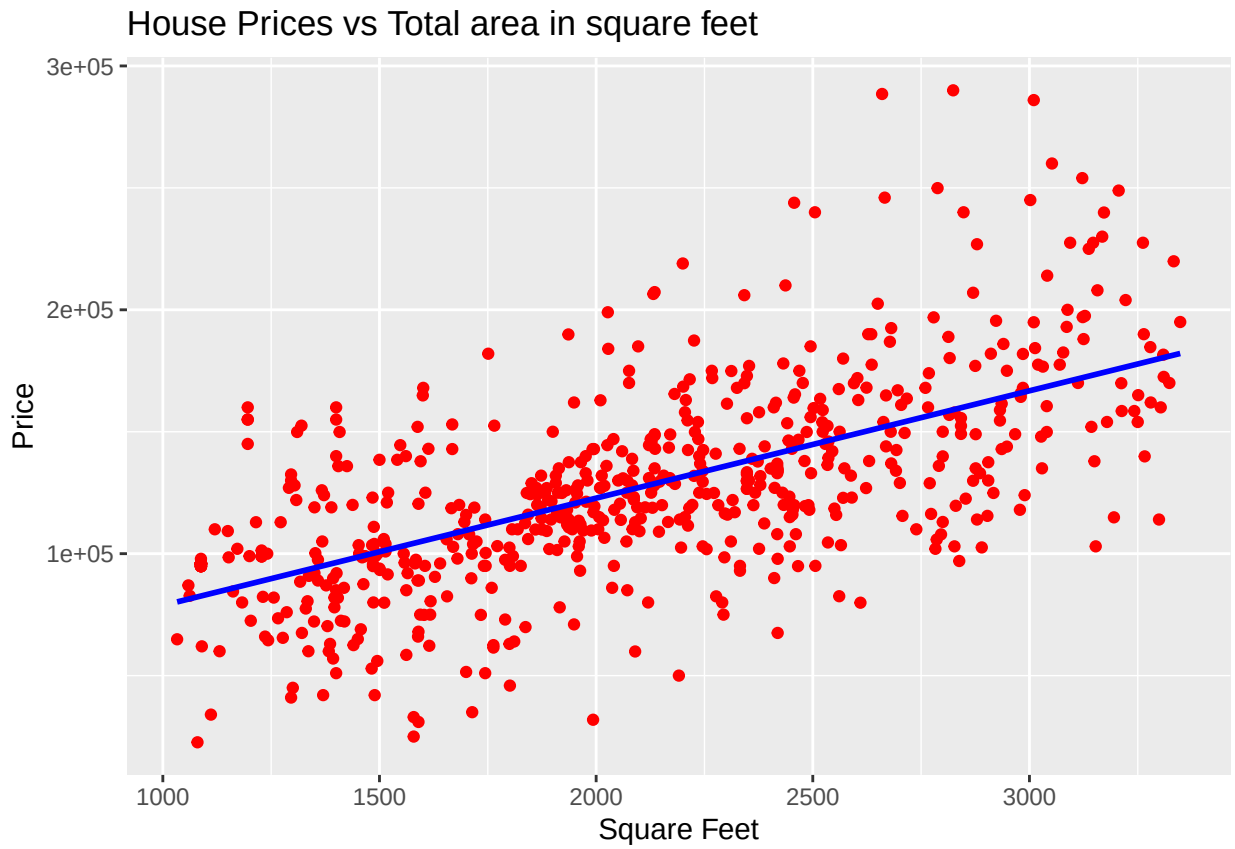
RESULTS OF THE DESCRIPTIVE STATISTICS TABLE

OBSERVATIONS: The trimmed means across variables show that in many cases, extreme values (outliers) significantly influence the raw mean. Particularly in `price`, `sqft`, and `dom`, outliers pull the mean higher, while the trimmed mean provides a more robust and perhaps realistic estimate of central tendency. This suggests that the data set includes some properties that are significantly larger, more expensive, or take much longer to sell, which may not be representative of the overall market.

(1c) Appropriate plots visualizing the distributions of some of the variables:


Distribution of House Prices
Frequency
200
100
0
0e+00
1e+05
2e+05
3e+05
Price


House Prices by Number of Bedrooms
3e+05
2e+05
1e+05
Price
2
3
4
5
Number of Bedrooms



OBSERVATIONS: - From the Histogram the majority of the houses in Baton Rouge, Louisiana are priced between 100,000 and 200,000. There are fewer houses priced below 100,000 and above 200,000 with the frequency decreasing as the price increases. This suggests that most house in Baton Rouge, Louisiana are modereately priced,with fewer house at the lower and higher end of the price spectrum.

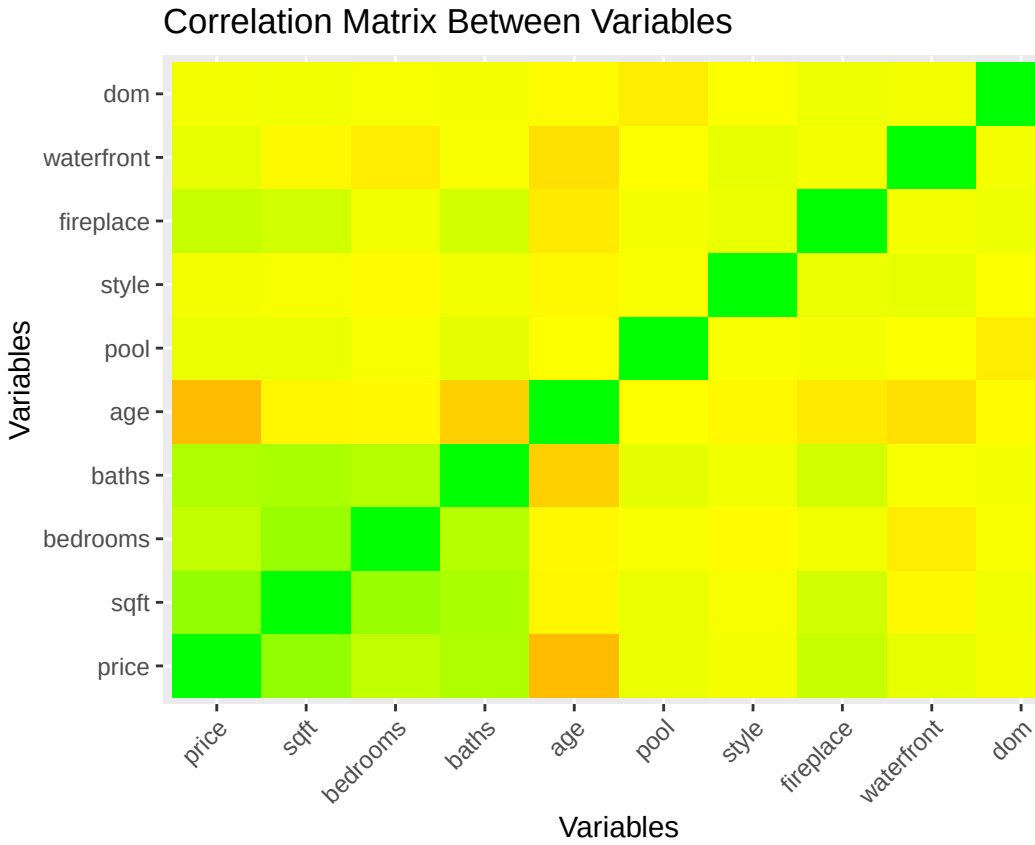
- The boxplot shows how house prices vary based on the number of bedrooms. 1-Bedroom Houses are generally lower prices, with a median price below 100,000. 2-Bedroom Houses Prices are higher than 1-bedroom houses, with a median around 150,000. The median price for 3-Bedroom Houses is higher, around 200,000, with a wider range of prices. 4 and 5-Bedroom Houses have the highest prices, with medians close to or above 250,000. The present of outlier at this price range , indicate some very high-priced houses. This suggests that house prices increase with the number of bedrooms, and there is more variability in prices for larger houses.
- The scatter plot shows a positive correlation between house size and price. As the total area of a house increases from 1000 to 3000 square feet, the price also tends to increase, with most prices falling below $300,000. This implies that larger houses generally cost more.

INTERACTIVE VISUALIZATION PLOTS

OBSERVATIONS: -Interactive Histogram for Price Distribution: Most houses are priced between $100,000 and $200,000, with fewer properties above or below this range.

-Interactive Boxplot for Price by Number of Bedrooms: House prices generally increase with the number of bedrooms, with greater price variability in larger houses.

-Interactive Scatter Plot for Price vs. Total Area in Square Feet: There is a positive correlation between house size and price; larger houses tend to be more expensive.



(1d) Correlation between variables
OBSERVATIONS: The strong positive correlation between `sqft` and `price` shows that larger houses are more expensive than smaller house. The number of bathrooms `baths` and `bedrooms` also positively correlate with the price, indicating that houses with more bathrooms and bedrooms generally cost more. Features like having a fireplace or a pool show weaker correlations with the price, suggesting these features don't significantly impact the house price.

QUESTION TWO: PROBABILITY, PROBABILITY DISTRIBUTIONS AND CONFIDENCE INTERVALS

(2a)

OBSERVATIONSS: -The probaility that a house has been in the market for greater than 100 days is 0.1410658 -The probability that a house is in traditional style is 0.661442

(2b)

OBSERVATIONS: The Probability that out of 20 houses chosen at random, more than half will be in the "Traditional" style is approximately 0.8993316

(2c) 90% Confidence Interval on the Mean Total Area (sqft) of houses in Baton Rouge:

FINDINGS: We are 90% confident that the true mean total area of all houses in Baton Rough is at least 2045.495 square feet and the true mean total area of all houses in Baton Rough is no more than 2124.128. Therefore the mean total area is approximately 2045 and 2124 square feet.

QUESTION THREE: HYPOTHESIS TEST

(3a) - Comparing mean house price with fireplace presence:

INTERPRETATION: -A two tail t-test is used because we are comparing the sample means of two independent group, house with or without a fireplace. -**Null Hypothesis (H0)**: The mean house price for houses with a fireplace is equal to the mean house price for houses without a fireplace. -**Alternative Hypothesis**

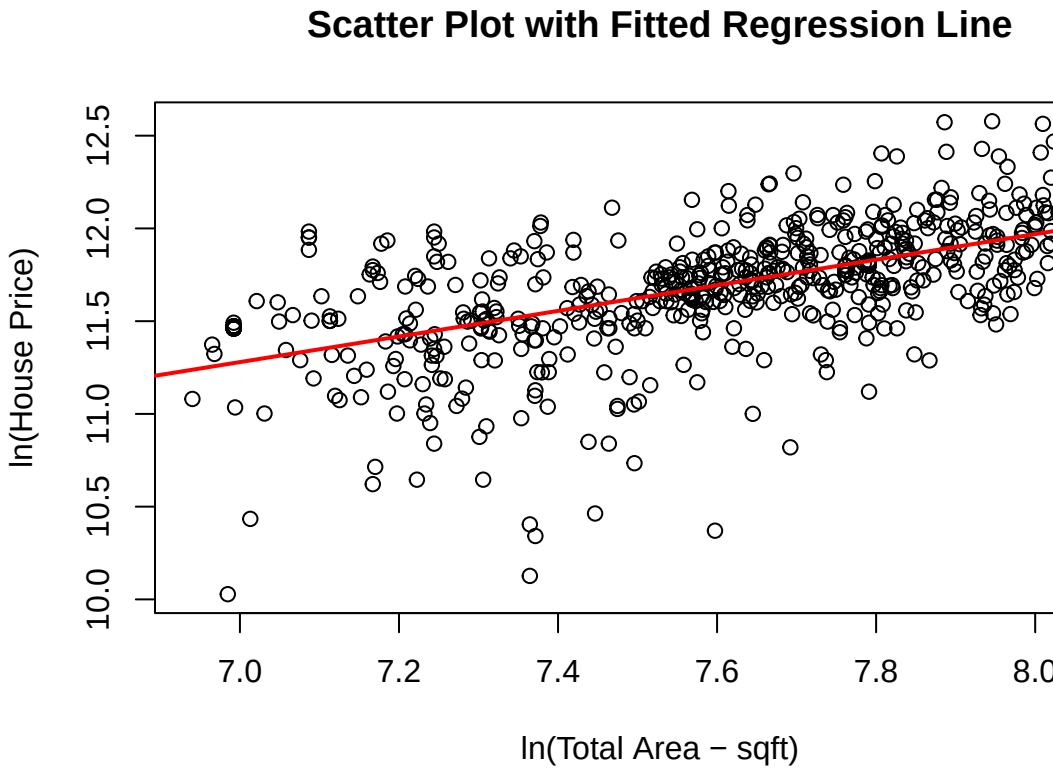
(H1): The mean house price for houses with a fireplace is different from the mean house price for houses without a fireplace. - The p-value is less than 0.01, we will reject the null hypothesis and accept the alternative hypothesis. This implies that we have a very strong evidence to justify that we are 99% confident that the presence of a fireplace significantly increases the price of a house in Baton Rough. On average , houses with a fireplace are priced between $21,479.45 and $37,197.61 higher than those without a fireplace. House with a fireplace has a higher mean price of $140,810.0 compare to house without a fireplace price of $111,471.5.

(3b) Hypothesis test on independence of having a pool and days on the market(dom):

INTERPRETATION: - A chi-square sample test was used to test if having a pool is independent of the days on the market is high or low or not. - **Null Hypothesis (H0)**: Having a pool is independent of whether the days on the market (dom) is high or low. - **Alternative Hypothesis (H1)**: Having a pool is not independent of whether the days on the market (dom) is high or low. - The p-value of 0.09964 is greater than the significant level of 5%. We will fail to reject the null hypothesis. This implies that, whether a house has a pool does not seem to affect how long it stays on the market. There is no significant evidence to suggest that the presence of a pool is related to the number of days a house remains listed.

QUESTION FOUR: SIMPLE LINEAR REGRESSION (4a) price as response and sqft as predictor:

OBSERVATION: for every 1 unit increase in the square feet of a house the price is expected to increase by $0.69008. With a p-value of 2e-16 *** is very small and less than 0.05 This implies that total area is a significant predictor of house price.



(4b) Scatter plot on fitted model:
OBSERVATIONS:- The scatter plot of the fitted regression model shows a positive relationship between house prices and the total area in square feet. This implies that as the total area increase , house price will also increase. The regression line indicates a strong positive correlation between these two variables. - The r- squared value of 0.3663329 indicate that about 37% variability in house price can be explain by the total area square of the house. The r squared has a very low value less than 60% what this mean is that

total area in square feet is not the only determinate of price. Because the model only explain 36.63% of the variation in house prices, other factor also plays a crucial role in determining house prices.

QUESTION FIVE: MULTIPLE LINEAR REGRESSION (5a) lm(price) against all predictor variables:

INTERPRETATION: - The results shows that The full model shows that house prices in Baton Rough are significantly influenced by predictor such as square footage, the number of bedrooms and bathrooms, the age of the property and dom. The age negative value is a pointer that older houses are tend to be less expensive because the house may have depreciate due to effusion on time. The number of bedrooms in a house does not really determine if it will be expensive or not as there may be houses with more bedroom but lack taste.The property that have stayed longer in the market is expected to be expensive compare to those that is just advertised. - The high R squared value of the full model of 0.9274453 compare to a very low R square value of simple learner model of 0.3663329 indicate that about 93% variability in house price can be explain by the entire predictor variables. What this means is that it provides a very good fit for the data, far better than a simple model that only considers total area in square feet(sqft).

(5b) The variable that is statistically significant:

OBSERVATION: Since we set our significant levels at 0.05. The following variables `price`,`sqft`,`bedrooms`,`baths`,`age`,`log_of_s` are statistically significant because thier p-value is less than 5%. We are 95% confidence that they plays a very signicant factors in determining the price of houses in Baton Rouge.

(5c) Stepwise selection:

OBSERVATIONS: - The stepwise feature selection process removed non-significant variables like `Pool`, `Fireplace`,`Style`, `dom` (Days on Market), and `High/Low Category`. This suggests that these predictive variables do not provide substantial additional predictive value for the price when the other variables are accounted for. - The adjusted R-squared value of the reduced model (0.9266) is very close to that of the full model (0.9261). This indicates that the reduced model is almost as effective in explaining the variability in house prices as the full model, despite using fewer variables.

(5d) Compare the full model and the reduced model r^2

OBSERVATION: The coefficient of determination between the full model and Reduce model is very close. This indicate that instead of considering all features in fullmodel that may not add much value, it will be statistically wise to adopt the reduce model in determining house prices in Baton Rouge.