MA334-SP-7_2322905

Ajinkya_Thokal_2322905

2024-04-26

Data exploration

Introduction:

This comprehensive analysis aims to provide valuable insights into the housing landscape by leveraging a rich dataset encompassing various property characteristics. Through descriptive statistics, probability distributions, hypothesis testing, and regression modeling, we uncover the central tendencies, relationships, and underlying patterns that govern this intricate market. The findings shed light on the impact of square footage, number of rooms, age, and amenities such as pools and fireplaces on property values. Additionally, we explore the influence of location-specific attributes, such as waterfront access, on pricing dynamics. By dissecting these multifaceted variables, this study serves as a foundation for informed decision-making and a deeper understanding of the residential real estate domain.

Descriptive Statistics:

```
     price              sqft         bedrooms         baths      
 Min.   :  22000   Min.   : 662   Min.   :1.000   Min.   :1.000  
 1st Qu.:  99975   1st Qu.:1616   1st Qu.:3.000   1st Qu.:2.000  
 Median : 132000   Median :2205   Median :3.000   Median :2.000  
 Mean   : 158839   Mean   :2348   Mean   :3.204   Mean   :1.987  
 3rd Qu.: 176750   3rd Qu.:2805   3rd Qu.:4.000   3rd Qu.:2.000  
 Max.   :1580000   Max.   :7897   Max.   :7.000   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   :18.98   Mean   :0.07593   Mean   : 3.498   Mean   :0.5572  
 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.: 13.00  
 Median :0.00000   Median : 39.00  
 Mean   :0.06893   Mean   : 73.34  
 3rd Qu.:0.00000   3rd Qu.: 97.00  
 Max.   :1.00000   Max.   :728.00  
```

Trimmed mean:

```
       price         sqft     bedrooms        baths          age 
140109.75364   2227.34694      3.22449      1.94898     16.62682 
```

The dataset provides comprehensive information about residential properties, including prices, square footage, number of bedrooms and bathrooms, age, and additional features like pools, fireplaces, architectural styles, and waterfront status.

Price and Square Footage: The prices exhibit a wide range, from $22,000 to $1,580,000, with a median of $132,000 and a mean of $158,839, suggesting the presence of some high-priced outliers. The trimmed mean of $140,109.75 provides a more robust estimate of the central tendency, mitigating the influence of these outliers. Similarly, the square footage varies from 662 to 7,897 square feet, with a median of 2,205 square feet and a trimmed mean of 2,227.35 square feet.

Bedrooms and Bathrooms: The number of bedrooms ranges from 1 to 7, with a median and trimmed mean of 3. The number of bathrooms ranges from 1 to 5, with a median of 2 and a trimmed mean of 1.95, indicating a skew towards properties with fewer bathrooms.

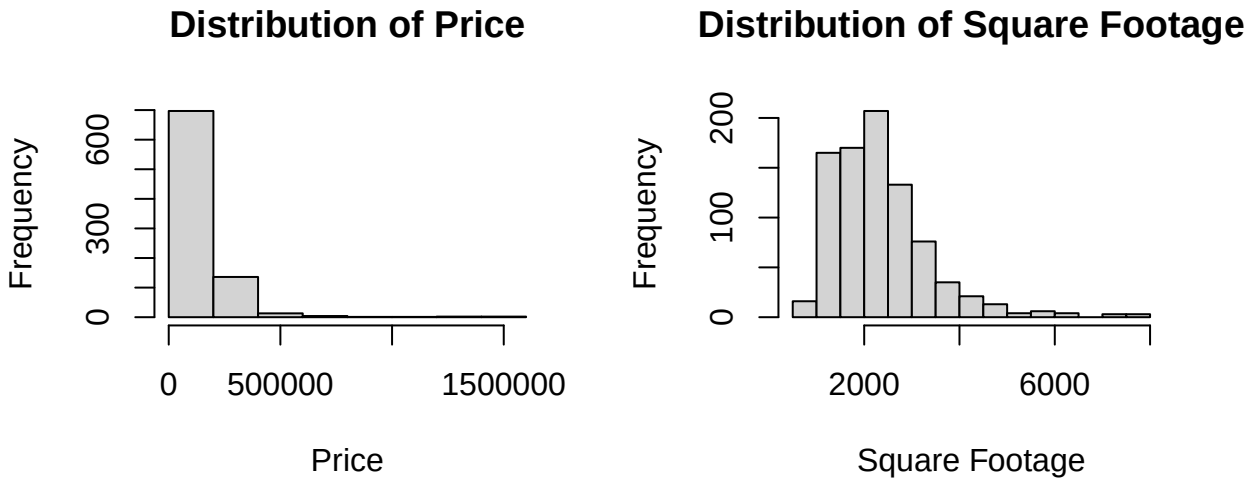
Age: The age of properties varies from 1 to 80 years, with a median of 18 years and a trimmed mean of 16.63 years, suggesting a mix of newer and older properties.

Additional Features: The dataset also provides information on the presence of pools (7.6% of properties), fireplaces (55.7%), and waterfront status (6.9%), as well as architectural styles and days on the market.

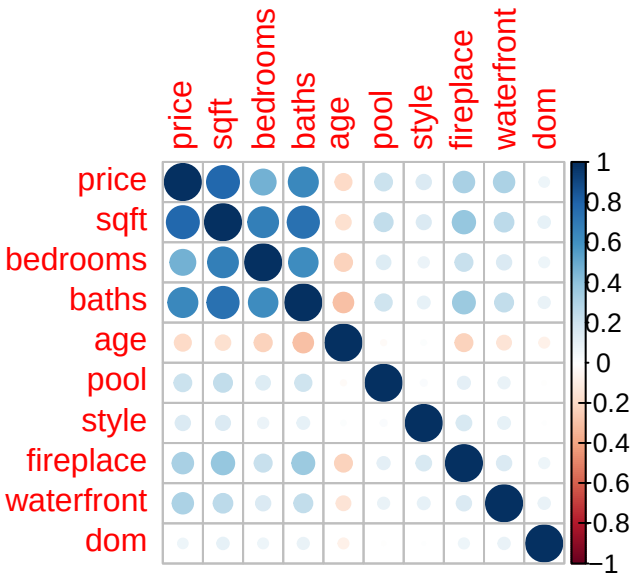
These descriptive statistics and trimmed means provide valuable insights into the central tendencies and distributions of key variables, enabling further analysis and modeling of residential property characteristics and their relationships with prices.

Distribution of variables:

The distribution of real estate properties' prices and square footage is shown in the plot. With a few extremely high outliers, the majority of prices are low, suggesting a right-skewed distribution. The distribution of square footage is more even, with most units ranging between 2,000 and 6,000 square feet. This implies a mixture of residential buildings, with a few more opulent ones accounting for the premium costs.



Correlation Analysis:



The links between house features are shown in the correlation matrix. Square footage, beds, baths, and the inclusion of a pool all positively correlate with price, suggesting that larger, more feature-rich homes sell for more money. Values of detached single-family houses and waterfront locations are positively correlated with those of newer properties. The presence of a fireplace and home style have weak connections. There may be commonality in larger homes as bedrooms, baths, and square footage all have favorable correlations with these characteristics, as do pools.

Probability, probability distributions and confidence intervals:

1. Calculating Basic Probabilities:

```
[1] "Probability of a pool: 0.0759345794392523"
```

```
[1] "Conditional probability of a fireplace given a pool: 0.753846153846154"
```

The probability of finding a pool in homes is approximately 7.6%. Interestingly, if a home has a pool, there's a notably high chance, around 75.4%, of it also having a fireplace. This indicates a significant correlation between pool and fireplace presence in homes.

2. Probability Analysis of Presence of Pools in Randomly Selected Houses:

```
[1] 0.03503515
```

In order to determine the likelihood that at least three of ten houses will have a pool, this code first uses the binomial distribution to compute the cumulative probability of 0, 1, and 2 successes. It then subtracts this cumulative probability from 1 in order to determine the likelihood of at least three successes, which comes out to be roughly 3.5%.

3. Confidence Interval for House Prices:

```
95% Confidence Interval for the mean house price in the USA: [ 150323.6 , 167354.8 ]
```

The interval we calculated ranges from approximately $150,324 to $167,355. This means we can say with 95% confidence that the average house price in the dataset is between these two figures.

Contingency tables and hypothesis tests

1. Two-sample t-test for house prices:

```
  Welch Two Sample t-test

data:  house_price_waterfront and house_price_not_waterfront
t = 4.0073, df = 58.802, p-value = 8.759e-05
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 93507.44      Inf
sample estimates:
mean of x mean of y
 308182.2  147783.7
```

The two-sample t-test compares the average (mean) house prices for houses on the waterfront versus those not on the waterfront.

Result: The mean house price for waterfront houses ($308,182.2) is significantly greater than that for non-waterfront houses ($147,783.7) with a p-value of 8.759×10^{-5} at a 95% confidence level. This suggests strong evidence that waterfront houses tend to have higher prices.

2. Contingency table for pool and fireplace:

```
[1] "Contingency Table - Relative Frequencies for Pool and No Pool by Fireplace Presence"
```

	No_Fireplace	Fireplace
No_Pool	45.89128	54.10872
Pool	24.61538	75.38462

The contingency table shows the relative frequencies of houses with and without a pool, categorized by the presence or absence of a fireplace. The row "No_Pool" indicates that among houses without a pool, 45.89% did not have a fireplace, while 54.11% had a fireplace. On the other hand, the row "Pool" shows that among houses with a pool, only 24.62% did not have a fireplace, while a majority of 75.38% had a fireplace. This suggests that houses with a pool are more likely to have a fireplace as well, compared to houses without a pool. The higher relative frequency of 75.38% for houses with both a pool and a fireplace indicates a potential association between these two features, which could be further explored through statistical tests of independence.

3. Chi-squared test of independence:

```
[1] "Chi-squared Test of Independence:"

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

data:  contingency_table
X-squared = 10.175, df = 1, p-value = 0.001424
```

The chi-squared test determines whether the presence of a fireplace and the presence of a pool are independent of each other.

Result: The test yields a p-value of 0.001424, which is less than the significance level of 0.05. This suggests that there is a significant association between having a fireplace and having a pool in houses. In other words, the presence of a fireplace and the presence of a pool are not independent of each other; they tend to occur together more often than expected by chance.

These results provide valuable insights into the relationships between house features, such as waterfront location, presence of fireplaces, and presence of pools, which can be useful for understanding housing market dynamics and buyer preferences.

Simple Linear Regression:

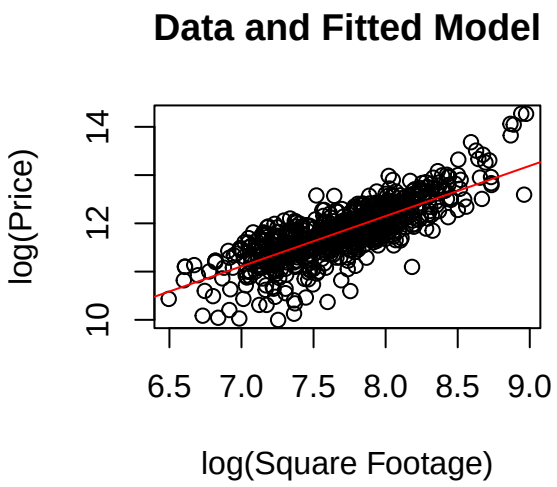
1. Simple Linear Regression Analysis: ln(Price) and ln(Sqft) as Predictors:

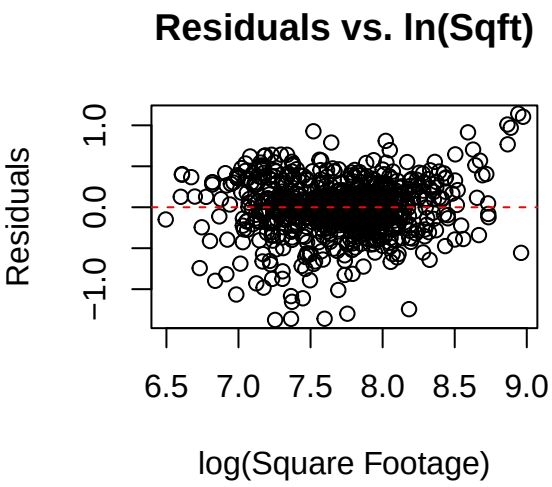
```
Call:
lm(formula = ln_price ~ ln_sqft, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.37501 -0.15673  0.01193  0.19059  1.14202

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.81807    0.20934   18.24   <2e-16 ***
ln_sqft      1.04157    0.02723   38.25   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3268 on 854 degrees of freedom
Multiple R-squared:  0.6314,  Adjusted R-squared:  0.631
F-statistic:  1463 on 1 and 854 DF,  p-value: < 2.2e-16
```

The data demonstrates that a house's square footage, or size, has a significant role in determining its cost. We discovered that a house's price generally rises with its size using a linear regression model. The model can account for almost 63.14% of the variations in pricing based just on house size, demonstrating the strength of this link.

2. Scatter Plot and Residual Analysis: Fitted Model vs. Data:

Data and fitted model on the left side, a scatter plot illustrates the relationship between house prices and square footage, showing a positive correlation with some variability. On the right side, residuals plotted against the logarithm of square footage reveal potential heteroscedasticity, suggesting varying spread across square footage ranges.





Multiple Linear Regression:

1. Full Model Heading:

```
lm(formula = ln_price ~ ln_sqft + bedrooms + baths + age + pool +
    style + fireplace + waterfront + dom, data = data)

           1            2            3            4            5            6
-0.009916437 -0.500205993  0.564842937  0.009651335  0.016295890  0.278237381
           7            8            9           10
 0.142584235  0.043599728 -0.149158268  0.594158854

[1] 0.2756128

[1] 0.7418432

[1] 0.7375457

   value    numdf    dendf
172.6222  14.0000 841.0000
```

In this analysis, I fitted a multiple linear regression model to examine the relationship between the natural logarithm of house prices (ln_price) and various predictor variables, including square footage (ln_sqft), number of bedrooms and bathrooms, age of the house, presence of a pool, house style, fireplace, waterfront view, and days on the market (dom). The fitted model summary provided the following key insights:

R-squared (0.7418) and Adjusted R-squared (0.7375):These values indicate that the model explains around 73.75% of the variation in ln_price, suggesting a reasonably good fit. F-statistic (172.6 on 14 and 841 df) with a very small p-value: This highly significant F-statistic implies that the overall model is statistically significant in explaining the variation in house prices.

Notably, variables such as square footage (ln_sqft), number of bathrooms, age, certain house styles, presence of a fireplace, and waterfront view emerged as significant predictors of house prices based on their associated p-values.

2. Feature selection using stepwise regression:

```
lm(formula = ln_price ~ ln_sqft + baths + age + pool + style +
    fireplace + waterfront, data = data)
```

```
[1] 0.275394
```

```
[1] 0.7416401
```

```
[1] 0.7379624
```

```
   value    numdf    dendf
201.6575  12.0000 843.0000
```

I performed stepwise regression using the `stepAIC` function from the `MASS` package to select the most relevant predictors for the house price model. This approach uses the Akaike Information Criterion (AIC) to identify the subset of variables that best explains the variation in the natural log of house prices (`ln_price`) while avoiding overfitting.

The selected model includes predictors: `ln_sqft` (log square footage), `baths` (number of bathrooms), `age`, `pool`, `style` (house style categories), `fireplace`, and `waterfront`.

Key insights: - Residual standard error: 0.2754 (typical distance between observed and predicted `ln_price` values) - R-squared: 0.7416, Adjusted R-squared: 0.738 (model explains ~73.8% variation in `ln_price`) - Highly significant F-statistic (201.7 on 12 and 843 df, p-value $<$ 2.2e-16)

Stepwise regression helped simplify the model by retaining only the most relevant predictors based on the AIC criterion

3. K-fold cross-validation:

```
RMSE: 0.2795633 Rsquared: 0.7266244 MAE: 0.2014314
```

```
RMSE: 0.2787653 Rsquared: 0.7324301 MAE: 0.2010294
```

The code performs k-fold cross-validation to evaluate the performance of the full and reduced linear regression models for predicting house prices. Cross-validation provides a reliable estimate of a model's generalization ability by splitting the data into multiple folds and iteratively training and testing on different subsets. The results show that the reduced model, obtained through feature selection, has slightly better performance metrics (lower RMSE of 0.2788 and MAE of 0.2010, higher R-squared of 0.7324) compared to the full model (RMSE: 0.2796, MAE: 0.2014, R-squared: 0.7266). The improved metrics suggest that the reduced model, with fewer predictors, generalizes better to unseen data.