MA334-SP-7 Final Project (2023-2024)

Huu Ann Tran (2322761)

This report aims to investigate the factors affecting house prices in the USA, using a data set sourced from Dr. Kelley Pace, Department of Finance, Louisiana State University.

1. Data Exploration

The data set gives information on 856 different houses from the area Baton Rouge, LA, during mid-2005 about ten different attributes; that is to say, it gives 856 observations of ten variables. The variables are as follows: in terms of continuous data, we have the sale price (in dollars) and the total area (in square feet); for discrete data we have the number of bedrooms, number of bathrooms, the age (in years), and the number of days on the market; and in terms of qualitative data (or more specifically, nominal data) we have whether or not it has a pool, the style of the house, whether or not it has a fireplace, as well as whether it is on the waterfront.

The following tables show the descriptive statistics for the housing data with the most relevant statistics for each variable.

Table 1

	Mean	Standard Deviation	Range	Mode	Median
Price	160826.51	134310.19	22654, 1580000		132250
Area (sqft)	2373.46	1054.44	662, 7897		2224
Bedrooms	3.213		1, 8	3	
Bathrooms	1.984		1, 5	2	
Age (years)	19.208	17.394	1, 80		18
Days on market	72.116	91.839	0, 728		39

Table 2

Variable	Frequency
POOL	66
STYLE	
Traditional	523
Townhouse	63
Ranch	0
New Orleans	11
Mobile Home	0
Garden	0
French	87
Cottage	61
Contemporary	0

Variable	Frequency
Colonial	17
Acadian	94
FIREPLACE	481
WATERFRONT	60

As shown by Table 1 the mean area of the houses was calculated to be 2373.46 sqft, with mostly 3 bedrooms and 2 bathrooms. At the time of data collection, the houses were 19.208 years old on average, though it ranges from newly built houses to houses that were 80 years old. There is also a large range for the number of days the house was on the market (0, 728), but on average houses were on the market for 72.116 days. Table 2 shows that out of the 856 houses, 66 have a pool, the most common type of house was traditional, just over half of houses have a fireplace, and 60 are on the waterfront.

In terms of sales prices, the mean sales price was calculated to be 160826.51 USD, however, the large standard deviation (134310.19 USD) and range (22654, 1580000) show that there is a lot of variability in the data. As the median sales price was calculated to be 132250 USD, this suggests the presence of extreme values which are resulting in a significantly higher mean. To investigate this, a histogram was plotted which can be seen in Figure 1.

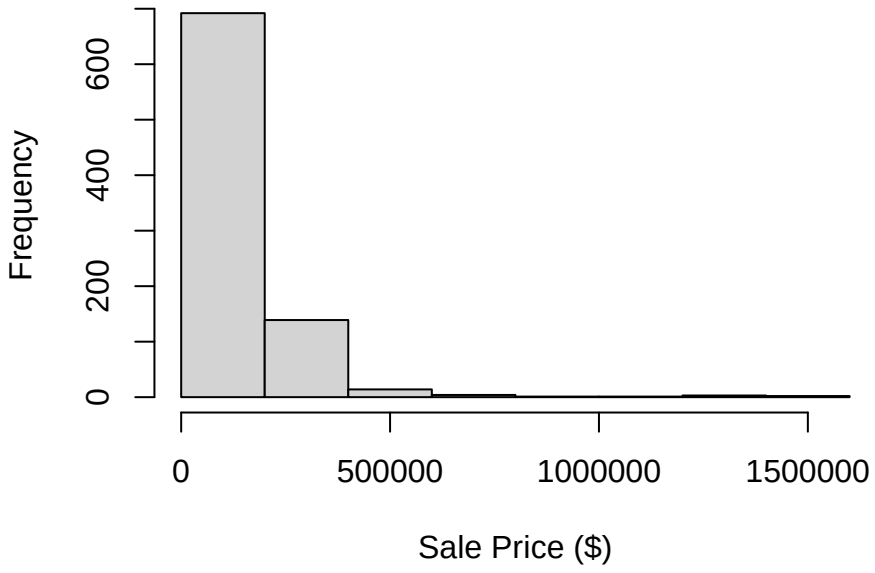


Figure 1: Histogram of Sale Price

As shown by the histogram, the sale price is positively skewed, meaning that there are in fact extreme values that are making the mean higher than expected. To this end the median may be a more accurate representation of the average sales price.

Correlations between variables

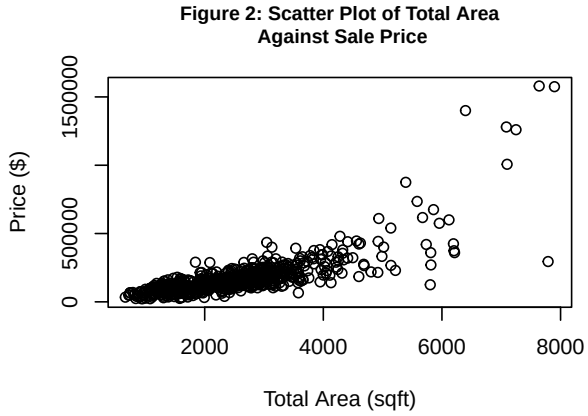


Figure 2: Scatter Plot of Total Area Against Sale Price

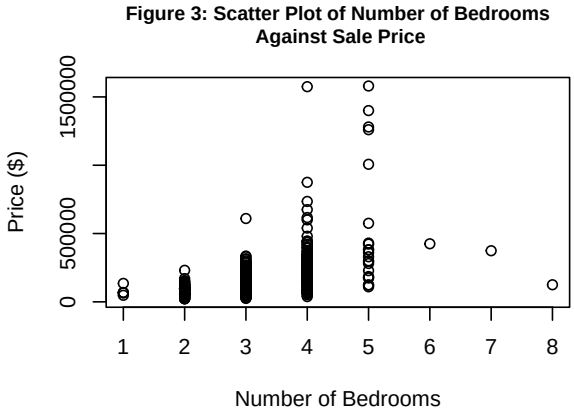


Figure 3: Scatter Plot of Number of Bedrooms Against Sale Price

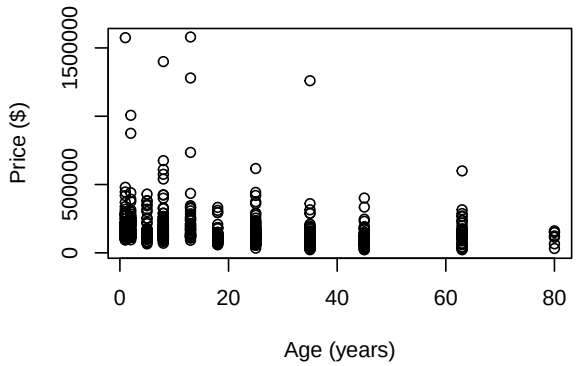


Figure 4: Scatter Plot of Age Against Sale Price

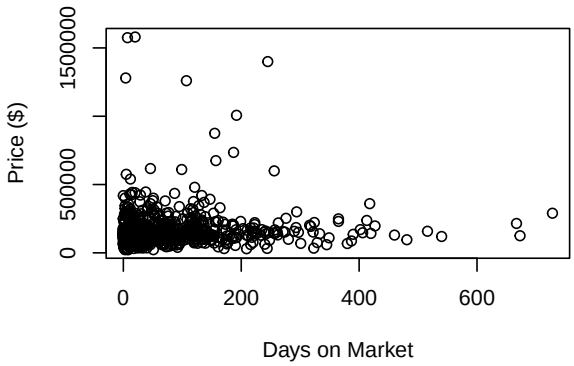


Figure 5: Scatter Plot of Days on Market Against Sale Price

Figures 2 to 5 determine whether there is any correlation between some of the variables and the sale price. In general, it is shown from Figures 2 and 3 that there is a positive correlation between the total area and the price, and between the number of bedrooms and price, i.e. the higher the number of bedrooms and the higher the total area, the higher the sale price. For the age, there is a slight negative correlation, meaning that newer houses tend to be slightly more expensive than older houses, as shown by Figure 4. In contrast, Figure 5 shows that there is no clear correlation between the days on the market and the price. It is important to note, however, that correlation does not mean causation, and in the case of housing prices it is a range of combined factors that determine the ultimate price.

2. Probability, probability distributions and confidence intervals

Given that the frequency of a property having a pool is 66, the probability that a house chosen at random from the data set has a pool is 0.077. Similarly, the probability that the property would have a fireplace given that it has a pool is 0.562.

The probability that, out of 10 houses chosen at random from the data set, at least 3 will have a pool is 0.036. This was achieved through the following calculation: as the probability that 3 or more houses have a pool ($P(X \geq 3)$) and the probability that less than 3 houses have a pool ($P(X \leq 2)$) add up to 1, the probability that at least 3 houses have a pool is given by $1 - P(X \leq 2)$ (where X is the random variable "has a pool"). This, in turn is the same as $1 - (P(X = 2) + P(X = 1) + P(X = 0))$.

Then, the probability of each event was found using the concept of binomial probability, which can be used

when the following are true: (1) “there is a fixed number of independent identical trials” (in this case 10), (2) “each trial results in either”success” or a “failure”” (having a pool or not having a pool), and (3), “the probability of success is the same for each trial” (0.077)(1, p.49). This was calculated using the dbinom() function in R for each event.

Assuming the data set provides a random sample of houses in the USA, a 95% confidence interval on the mean house price in the USA is (151829.045, 169823.975 USD). Loosely, this means that there is a 95% degree of confidence that the true mean price for the USA lies between these values.

3. Contingency tables and hypothesis tests

Is the mean house price (over all house styles) greater if a house is on the waterfront? To answer this question I will use the one-sample t-test, which determines whether the sample mean and the mean of the population differ. As the t-test assumes independence, the null hypothesis (H_0) and the alternative hypothesis (H_1) will be the following:

H_0 : The mean house price over all styles is the same as the mean price of houses on a waterfront.

H_1 : The mean house price is greater if a house is on the waterfront than the mean house price of all styles.

Upon carrying out the t-test in R, we get a p-value of 0. With a significance level of 5%, this means that there is strong evidence to reject the null hypothesis, i.e. we can assume the mean house price is in fact greater if a house is on the waterfront than the mean house price of all styles.

Table 3: Contingency table showing relative frequencies for “Pool” and “No pool” according to whether a house has or has not got a fireplace.

	Pool	No Pool	Total
Fireplace	52	429	481
No Fireplace	14	361	375
Total	66	790	856

To test whether a house having a fireplace is independent of whether it has a pool requires the chi-squared test of independence. Let the hypotheses be as follows:

H_0 : A house having a fireplace is independent of whether it has a pool.

H_1 : A house having a fireplace is dependent of whether it has a pool.

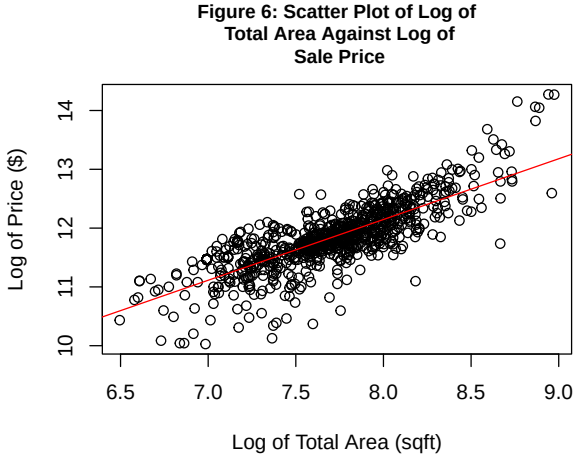
According to the results, the chi-squared statistic (which measures how much the observed frequencies differ from the expected frequencies if the two variables are unrelated) came to be 13.855 which is higher than the critical value of 3.841, and the corresponding p-value was 0, meaning that at a 5% significance level there is overwhelming evidence to reject the null hypothesis and thus it is safe to assume that whether a house has a fireplace is in fact related to whether it has a pool.

4. Simple Linear Regression

Table 4: Table showing fitted linear regression model of ln(price) as response variable and ln(sqft) as predictor variable.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.88	0.207	18.8	0
logsqft	1.03	0.027	38.5	0

When a simple linear regression is performed with the logarithm of the house price as the response variable and the logarithm of the total area as the predictor variable, we get a p-value of 0. This means that the total area of the house is indeed a significant predictor of the house price. Furthermore, the coefficient from the output tells us that for every 1% increase in the total area, there is roughly a 1.03% increase in the house price.



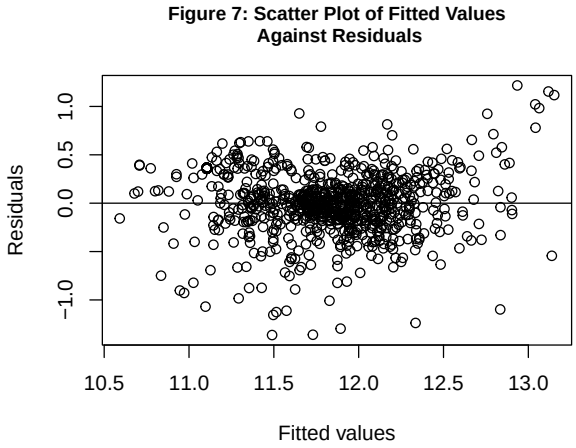


Figure 6 demonstrates the linear relationship between the log of the total area and the log of the price, and the way that the residuals are relatively evenly distributed above and below the reference line for zero residuals in Figure 7 suggest that the variances of the error terms are equal. Additionally, both plots also show the presence of extreme values.

5. Multiple Linear Regression

Below is the full linear regression model of the log price against all the predictor variables (also using the log of the total area). For the qualitative variables, dummy variables were used to represent the data, with the first dummy variable removed from each column in order to avoid multicollinearity.

Table 5: Table showing multiple linear regression model of ln(price) against all variables (i.e. full model).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.628	0.261	21.572	0.000
Log_sqft	0.748	0.040	18.530	0.000
Bedrooms	-0.003	0.020	-0.145	0.885
Baths	0.217	0.023	9.366	0.000

	Estimate	Std. Error	t value	Pr(>|t|)
Age	-0.004	0.001	-6.501	0.000
DOM	0.000	0.000	-0.264	0.792
Pool	0.039	0.038	1.040	0.299
Townhouse	0.274	0.041	6.642	0.000
New_Orleans	0.346	0.086	4.045	0.000
French	0.146	0.035	4.178	0.000
Cottage	0.131	0.041	3.213	0.001
Colonial	-0.012	0.068	-0.170	0.865
Acadian	-0.032	0.031	-1.022	0.307
Fireplace	0.087	0.022	3.965	0.000
Waterfront	0.109	0.039	2.819	0.005

From the full model, the p-values indicate that the total area, number of bathrooms, age, certain styles (townhouse, New Orleans, French, and cottage), presence of a fireplace, and presence of a waterfront, are statistically significant predictors of the sale price, as they are all less than 0.05. Surprisingly, the presence of a pool is not a significant predictor of price, as indicated by its p-value of 0.299, which fails to reach statistical significance. Furthermore, when stepwise selection for the AIC value is performed (which selects the model that resembles the true model the most from those available (1, p.99), it is the model with these exact variables which give the lowest AIC value. Thus, these variables can be used to produce a reduced model which is roughly as follows:

$$\begin{aligned}\ln(\text{Price}) = {} & 5.630 + 0.745(\ln(\text{sqft})) + 0.220(\text{Bathroom Number}) \\ & - 0.004(\text{Age}) + 0.284(\text{Townhouse}) + 0.368(\text{New Orleans}) \\ & + 0.147(\text{French}) + 0.134(\text{Cottage}) + 0.086(\text{Fireplace}) \\ & + 0.109(\text{Waterfront})\end{aligned}$$

whereby the quantitative variables are replaced by 1 if it is present, or 0 if not present.

Using a k-fold cross validation, we find that overall, the reduced model obtained following feature selection performs better than the full model. This is because it results in a smaller AIC value (219.674 for the reduced model, and 227.585 for the full model), despite the fact that the reduced model has a slightly higher RMSE than the full model (0.277 for the reduced, 0.28 for the full). This is due to the fact that there are 5 less variables in the reduced model.

Conclusion

In conclusion, there are many factors which affect house prices in the USA based on this data set, with the most prominent predictors being the total area, number of bathrooms, age, style of house, presence of a fireplace, and presence of a waterfront.

Bibliography

1. Upton G, Brawn D. *Data Analysis* [Internet]. First. Oxford: Oxford University Press; 2023 [cited 2024 Apr 17]. Available from: https://read.kortext.com/reader/pdf/2415060/iv

2. Kaplan J, Schlegel B. (2023). *fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables.* Version 1.7.1. URL: https://github.com/jacobkap/fastDummies.

3. Venables WN, Ripley BD. (2002). *Modern Applied Statistics with S.* Fourth edition. New York: Springer. [ISBN 0-387-95457-0]. Available from: https://www.stats.ox.ac.uk/pub/MASS4/.

4. R Core Team. (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. Available from: https://www.R-project.org/.

5. Signorell A (2024). *DescTools: Tools for Descriptive Statistics.* R package version 0.99.54, https://CRAN.R-project.org/package=DescTools.

6. Kuhn, M. (2008). *Building Predictive Models in R Using the caret Package.* Journal of Statistical Software, 28(5), 1–26. https://doi.org/10.18637/jss.v028.i05