STATS 101A - Final Project Report: Data-Driven Home Valuation

Introduction

The purpose of this research project is to develop a predictive model to analyze how property features influence housing prices. Our dataset was obtained from Kaggle, containing 545 observations of a sample of households in the San Francisco Bay Area by the Bay Area Transportation Study in 1965. The following variables were included in the data: price, area, bedrooms, bathrooms, stories, and parking.

- price: Price of the house in US Dollars
- area: Area of the house in square feet
- bedrooms: Number of bedrooms in the house
- bathrooms: Number of bathrooms in the house
- stories: Number of stories the house is comprised of
- parking: Number of parking spots available for the household

Understanding how specific property features influence home prices is essential for the real estate market. Therefore, our study examines this relationship by using house price as the response variable and analyzing several key property attributes as predictor variables: area of the house, number of bedrooms, number of bathrooms, number of stories, and number of parking spots. All analysis was conducted using R. We utilized multiple linear regression to model the relationships between these property features and house prices.

This report begins with an overview of descriptive statistics for all variables examined. Our initial approach was a complete model incorporating all unmodified variables. We then explored various transformations and variable selection methods to develop models that might better capture the underlying patterns in the data. Throughout this process, each model was evaluated for the validity of model assumptions and significance of predictors. We conclude by identifying an optimal model that best explains how specific property features influence house prices in the current market.

Data Description

We begin with an examination of the summary statistics. From Figure 1, we observe that every variable in our dataset shows a right-skewed distribution.

Variable	Mean	Standard Deviation
Price	4,766,729	1,870,440
Area	5,150.541	2,170.141
Bedrooms	2.965138	0.7380639
Bathrooms	1.286239	.5024696
Stories	1.805505	.8674925
Parking	0.6935780	.8615858

Table 1. Variable Means and Standard Deviations

Figure 1. Distribution of Variables

Our summary of each variable's mean and standard deviation is found in Table 1. On average, properties are priced at around $4.77 million and cover 5,150 square feet. However, there is significant variability in property prices and size of the properties, due to the high standard deviation for prices and area. The number of bedrooms, bathrooms, stories, and parking stalls have moderate variation.

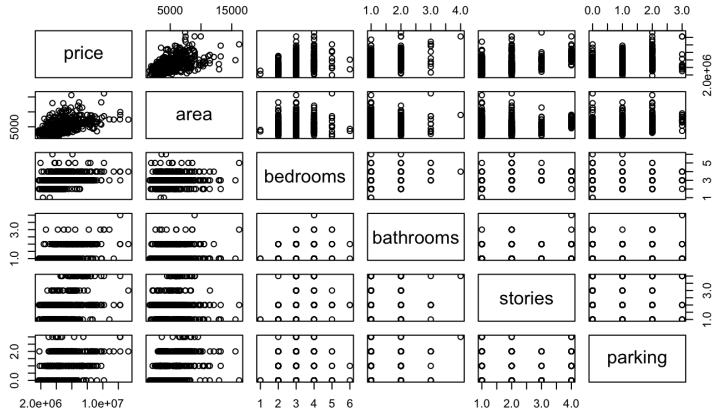

Figure 2. Scatterplot Matrix of Variables

Looking at the scatterplot matrix in Figure 2, we notice price has a positive linear relationship with area and bedrooms. However, there appears to be heteroscedasticity in the relationship between price and area, seen by the fan shape. For the other predictor variables, price does not seem to have any relationship with bedrooms, stories, and parking, but this will be further investigated through later analysis. Additionally, there does not appear to be any relationship between the predictor variables. The scatterplots appear randomly scattered, suggesting a lack of multicollinearity. Based on the observed linear relationships between certain predictor variables and the response variable, we began our analysis by fitting the data using a multiple linear regression model.

Results and Interpretation

To begin our analysis, we start with the full model including all predictor variables.

```
Call:
lm(formula = price ~ ., data = housing)

Residuals:
     Min       1Q   Median       3Q      Max
-3396744  -731825   -64056   601486  5651126

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -145734.5   246634.5  -0.591   0.5548
area            331.1       26.6  12.448  < 2e-16 ***
bedrooms     167809.8    82932.7   2.023   0.0435 *
bathrooms   1133740.2   118828.3   9.541  < 2e-16 ***
stories      547939.8    68894.5   7.953 1.07e-14 ***
parking      377596.3    66804.1   5.652 2.57e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1244000 on 539 degrees of freedom
Multiple R-squared:  0.5616,    Adjusted R-squared:  0.5575
F-statistic: 138.1 on 5 and 539 DF,  p-value: < 2.2e-16
```

Figure 3. Summary Statistics for the Full Model

Given the output from Figure 3, our model's linear regression equation is the following:

$$\widehat{price} = -145734.5 + 331.1\,area + 167809.8\,bedrooms + 1133740.2\,bathrooms + 547939.8\,stories + 377596.3\,parking$$

We notice all predictor variables have a significant p-value for their coefficients. We note an R^2 value of 0.5616, meaning only 56.16% of the variability in price is explained by the model. The overall F-test gives us the F-statistic of 138.1 and p-value less than our significance level, 0.05. We have significant evidence to reject the null hypothesis and conclude that at least one of our predictor coefficients is significant.

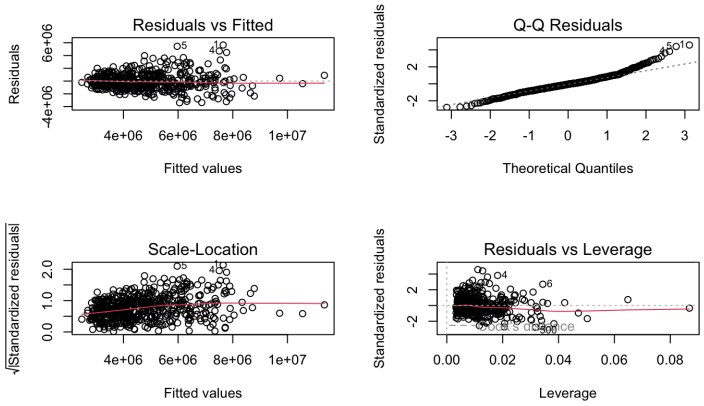

Figure 4. Diagnostic Plot for the Full Model

Based on our diagnostic plots from Figure 4, our full model fails to satisfy all the model assumptions for multiple linear regression. In the Residuals vs. Fitted plot, the red line is straight and centered around 0, upholding the assumption of linearity. However, the assumptions for normality, constant variance, and outliers, leverage, or influential points are violated. From our Normal Q-Q plot, we observe a heavy tailed distribution, violating our normality assumption. The Scale-Location plot shows that all points are relatively randomly scattered, but there is a slight upwards slope in the red line around the lower fitted values, possibly indicating that the constant variance assumption may be violated. Finally, the Residuals vs Leverage plot reveals numerous outliers and leverage points with high leverage and high standard residuals that could be looked further into, such as point 6.

bcPower Transformations to Multinormality

	Est Power	Rounded Pwr	Wald Lwr Bnd	Wald Upr Bnd
price	0.0237	0.0	-0.1350	0.1823
area	-0.1239	0.0	-0.2867	0.0390
bedrooms	0.2784	0.5	0.0377	0.5192
bathrooms	-4.4981	-4.5	-5.0067	-3.9896
stories	-0.4823	-0.5	-0.6924	-0.2721

	LRT <dbl>	df <int>	pval <chr>
LR test, lambda = (0 0 0 0 0)	439.9112	5	< 2.22e-16
LR test, lambda = (1 1 1 1 1)	1234.494	5	< 2.22e-16

Figure 5. Power Transformation Results for the Full Model

We investigated whether transformation is necessary using the Box-Cox method to improve the validity of the model. As seen in Figure 5, the likelihood ratio test that all transformation parameters are equal to 0, or require log transformation, resulted in a p-value less than 2.22e-16, which is less than $\alpha = 0.05$. Therefore, we concluded that there is not enough evidence for a log transformation of all variables. The likelihood ratio test that all transformation parameters are equal to 1 resulted in a p-value less than 2.22e-16, which is less than $\alpha = 0.05$. Therefore, we also concluded that there is not enough evidence to

recommend leaving all variables untransformed. In the Box-Cox output, price and area have a rounded power of 0, therefore we followed the recommendation to log transform price and area. λ = 0 is not within the bounds for other predictor variables, so we left those untransformed regardless of their rounded power, in order to ensure that the final model is easily interpretable.

The new multiple linear regression model post-transformation is:

$$\widehat{log(price)} = 11.250 + 0.407\, log(area) + 0.045\, bedrooms + 0.193\, bathrooms + 0.103\, stories + 0.063\, parking$$

```
Call:
lm(formula = log(price) ~ log(area) + bedrooms + bathrooms +
    stories + parking, data = housing)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76912 -0.14660  0.01159  0.16071  0.66863

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.24961    0.23908  47.055  < 2e-16 ***
log(area)    0.40716    0.02884  14.118  < 2e-16 ***
bedrooms     0.04474    0.01641   2.726  0.00661 **
bathrooms    0.19264    0.02354   8.182 2.02e-15 ***
stories      0.10316    0.01365   7.556 1.79e-13 ***
parking      0.06262    0.01327   4.718 3.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2464 on 539 degrees of freedom
Multiple R-squared:  0.5657,    Adjusted R-squared:  0.5616
F-statistic: 140.4 on 5 and 539 DF,  p-value: < 2.2e-16
```

Figure 6. Summary Statistics for the Transformed Model

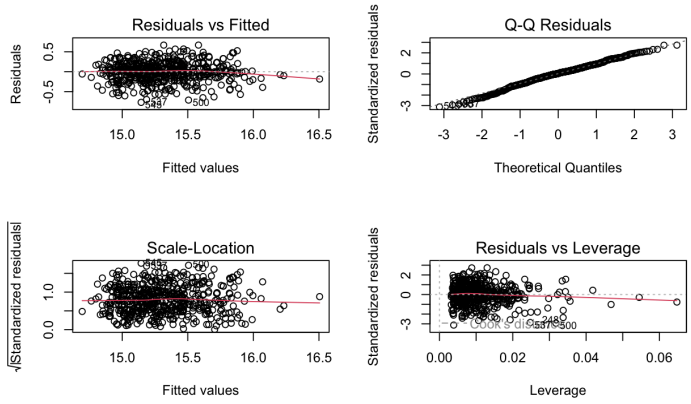

Figure 7. Diagnostic Plot for the Transformed Model

Figure 6 shows that for this transformed model, all predictor variables except bedrooms were significant at the α = 0.001 level. Bedrooms was significant at the α = 0.01 level. R^2 improved from the original model to 56.57%. The F-statistic is also significant, resulting in a p-value less than α = 0.05, meaning that at least one of the predictors of the model is significant.

The diagnostic plots in Figure 7 reveal that the model assumptions are upheld, although there may be several outliers as shown by the numerous points with standardized residuals outside of the interval (-2, 2) in the Residuals vs Leverage plot. The Residuals vs Fitted plot confirms that the linearity of the model's relationship is upheld because the points are randomly scattered around 0 with no distinct pattern. The

Normal Q-Q plot follows the normal reference line, affirming that the normality of the error term is upheld. The Scale-Location plot also shows a random scatter with a straight red line, affirming that the constant variance of the error term is upheld.

From our findings of the transformed model, we wanted to assess the effect of each predictor and find the best subset of the model. We first wanted to examine the VIF values to check for multicollinearity. As seen in Table 2, all VIF values are below 5, meaning we can abstain from saying that each coefficient is poorly estimated.

Variable	log(area)	bedrooms	bathrooms	stories	parking
VIF	1.182057	1.314613	1.253925	1.256730	1.171545

Table 2. Variance Inflation Factor (VIF) Values

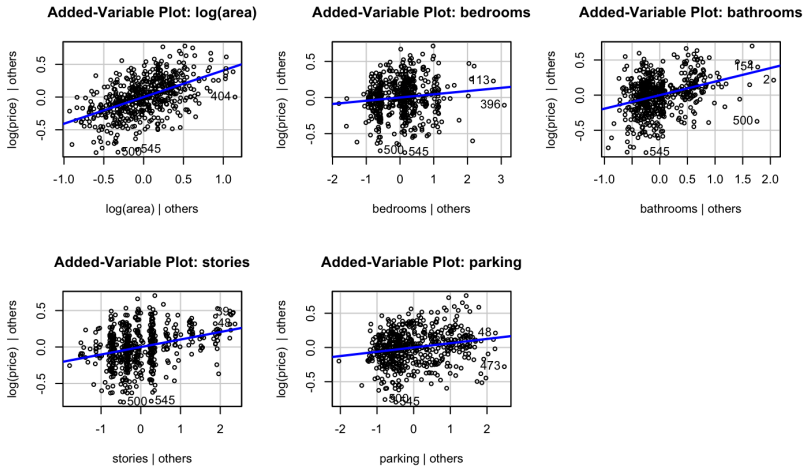

Figure 8. Added Variable Plots

With no predictor variables demonstrating multicollinearity, we proceeded with variable selection by examining the added variable plots. Figure 8 reveals that all predictors have explanatory power since no slopes are flat, and this is supported by Figure 6 since the p-values of these variables' coefficients are smaller than a significance level of 0.05. The coefficients from the summary are correctly reflected in the plots, affirming that our plots are consistent. Since all predictors have explanatory power, we will not remove any from our model.

Size	**R^2_{adj}**	**AIC**	**AICc**	**BIC**
1	0.3352771	-1297.949	-1297.741	-1299.176
2	0.4759348	-1426.529	-1426.321	-1428.37
3	0.5376248	-1493.791	-1493.583	-1496.246
4	0.5564069	-1515.4	-1515.192	-1518.469
5	0.5616289	-1520.864	-1520.656	-1524.546

Table 3. R^2_{adj}, AIC, AIC Corrected, BIC

In addition to variables, we explored the best possible models from the transformed model by considering all possible subsets, forward stepwise regression, and backward stepwise regression. By

exploring all possible subsets, we found the optimal models for all sizes and the R^2_{adj}, AIC, AIC corrected, and BIC suggest that the subset with p = 5 to be the best. When conducting forward and backward stepwise regression, we also concluded that the best subset was the same one with p = 5.

Finally, we concluded that our final model should be transformed but not reduced:

$$\widehat{log(price)} = 11.250 + 0.407\, log(area) + 0.045\, bedrooms + 0.193\, bathrooms + 0.103\, stories + 0.063\, parking$$

Following Figure 6, all predictors are significant with p-values less than 0.05, the adjusted R^2 improved from the original model to 56.16%, and the F-statistic is also significant with a p-value less than $\alpha = 0.05$. Since we did not end with a model different from our transformed model, the diagnostic plots for our finalized model are the same as the ones shown in Figure 7. Despite the remaining outliers shown in the Residuals vs Leverage plot, our model assumptions are upheld. The Residuals vs Fitted plot satisfies the linearity assumption with randomly scattered plots around 0, the Normal Q-Q plot affirms the normality assumption since the points follow the normal reference line, and the Scale-Location plot also shows a randomly scattered plot with a straight red line, satisfying the constant variance assumption.

Discussion

Our final model indicates a 1% increase in the square footage for area results in a 0.407% increase in housing prices. Adding a bedroom results in a 4.474% increase in housing price, adding a bathroom results in a 21.245% increase in housing price, adding a story results in a 10.867% increase in housing price, and adding a parking spot results in a 6.462% increase in housing price.

These findings are reasonable in a real-world situation, as housing prices tend to increase with more property features. Larger homes are generally more expensive, and in land-constrained cities like San Francisco, adding more stories to maximize space while adding more bedrooms, bathrooms, and parking stalls is a reasonable explanation for prices to increase as well. Another research study using multiple regression analysis found living space and land area were significant predictors of housing prices, aligning with our conclusion that house area, number of stories, and bedrooms are key factors (Zhou).

The primary limitations of this model lie in the context of the dataset. Although our findings align with general real estate principles, they no longer apply to the housing market of San Francisco today since the economy and demographics of San Francisco have drastically changed since 1965. Overtime, San Francisco has become overcrowded and urbanized, meaning the value of real estate has greatly increased. If we were to use our model today, it would not be accurate in estimating housing prices. However, if we were estimating prices for houses in 1965, we could improve our model by transforming it further to reduce the outliers. Future research could investigate if any other predictors can increase the predictability of the model, such as additional amenities.

References

Zhou, Shengjie. (2024). Investigation of Influential Factors of Housing Price. Advances in Economics, Management and Political Sciences. 105. None-None. 10.54254/2754-1169/105/20241950.