Project Analysis - Always Be Closing

Sagar Bansal

SECTION I: ANALYSIS SUMMARY

We did some numerical and visual exploration to know that how the data looks like and what does it mean in general sense. We found that the predictor (Sqft) has a minimum and maximum value of 850 and 3262 with the average value of 1876.56 (Figure 1 & Table 1). Similarly, response variable (Price) ranges from \$307,500 to \$840,000 with a mean value of \$507502.78 (Figure 2 & Table 1). In addition, we can say that the square foot is positively correlated with the Price (Figure 3) i.e., for every single unit increase in the square foot, the average price would be approximately \$179.14 higher according to the model (Table 3).

Yes, this data can be used as an evidence to persuade your client that the estimated price of \$300,000 is less for a house of 1500 square foot (Table 4). The average estimate range is roughly between \$404,926.84 and \$475,164.32 that is more than your client's estimate. This means that on average houses of 1500 square foot area are sold in the range of \$404,926.84 and \$475,164.32. Hence, your client should consider re-evaluating his/her judgement.

Unfortunately, the model is restricted to make predication for houses of area that ranges from 850 to 3262. The house in question with the area of 3750 square foot is larger than houses in the provided data. Conceptually, there could be other factors involved in predicting house prices of this big area such as house layout, useable space and potential future. For instance, the buyer of a large house may want to check the layout so that he/she doesn't need to spend more on renovating the place which can be a really huge amount. In contrast, buyer of a small house may not need to worry as any modification won't be that expensive given the house area. This might alternate the linear correlation that we see in the data currently. Thus, the model may mislead your second client.

SECTION II: APPENDIX

1) Statistical analysis:

We visualized the predictor and response in individual box plots to get the five-number summary (Minimum, First quartile, Median, Third quartile, Maximum). We discovered that the five number summaries of Price and Square foot are (307500, 394250, 502500, 585625, 840000) and (850, 1419, 1739.5, 2306.25, 3262), respectively. As the variables are linearly correlated (Figure 3), we defined the linear relationship between the two as the following:

$$Y = \beta_0 + \beta_1 X$$

where Y is the price in USD and X is the house area in square foot. β_0 and β_1 are the intercept and slope of the predictor, respectively.

Upon fitting the model, we got the estimated co-efficient β_0 and β_1 as 171331.47 and 179.14, respectively (Table 3). From table 2, we can note that the R square for this model came out to be .604 which means that 60.4% of the variability in this response is explained by this model. Although it is not ideal, we can still use the model to make the predictions. Similarly, the standard error of the estimate is 87111.55 which is not great either but still allowable relative to the mean of the house price.

We then did hypothesize testing for both the intercept and slope at 5% significance levels (Table 3). For β_0, the null and alternative hypothesis were as follows: H_0: $\beta_0 = 0$; H_A: $\beta_0 \neq 0$. Since the P-value (0.001) is less than 0.05 at 5% significance level, we reject the null hypothesis i.e., β_0 is statistically significant and cannot be equated to zero. For β_1, the null and alternative hypothesis were as follows: H_0: $\beta_1 = 0$; H_A: $\beta_1 \neq 0$. Since the P-value for β_1 (0.000) is less than 0.05 at 5% significance level, we reject the null hypothesis i.e., β_1 is statistically significant and there is some correlation between Square foot and Price.

It is important to note that this model is only limited to predictor values ranging from 850 to 3262 i.e., making predictions for anything less than or greater than that would be an attempt of extrapolation. Furthermore, only 60.4% of the variability in the model is explained by this model i.e., we may need to add more features to our analysis. Some of the additional variables to consider could be: 1) Distance from frequently visited places such as Supermarket, Hospital and Airport, 2) Age of the property and 3) Current state of local real estate market.

2) Supporting figures and tables:

Figure 1: Box plot - Sqft

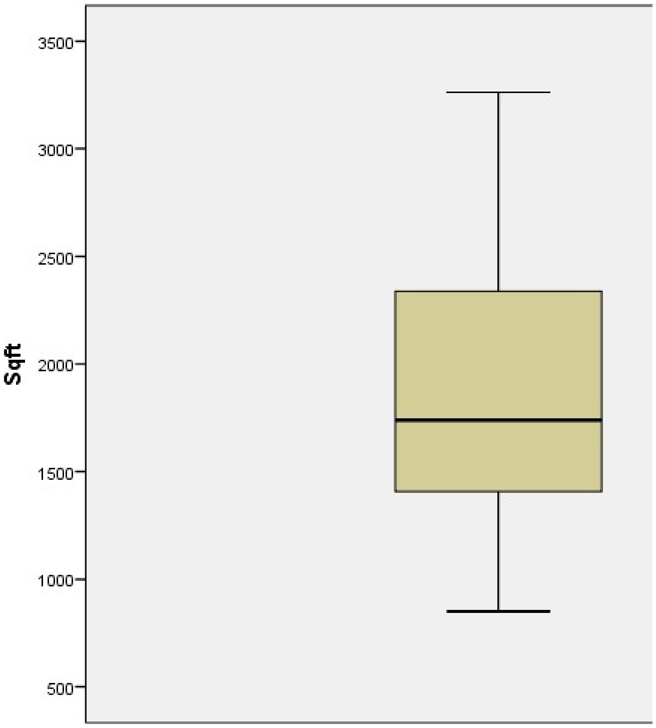


Figure 2: Box plot – Price

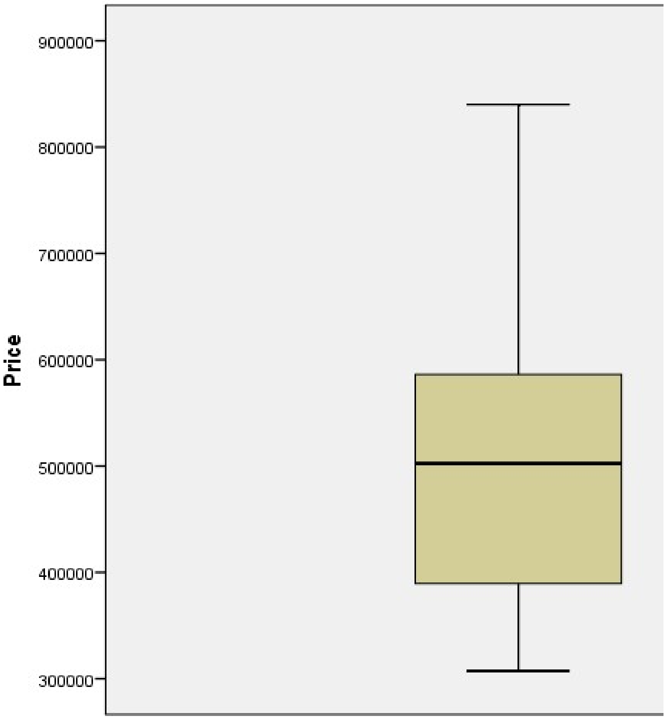


Figure 3: Scatter plot – Price vs Sqft

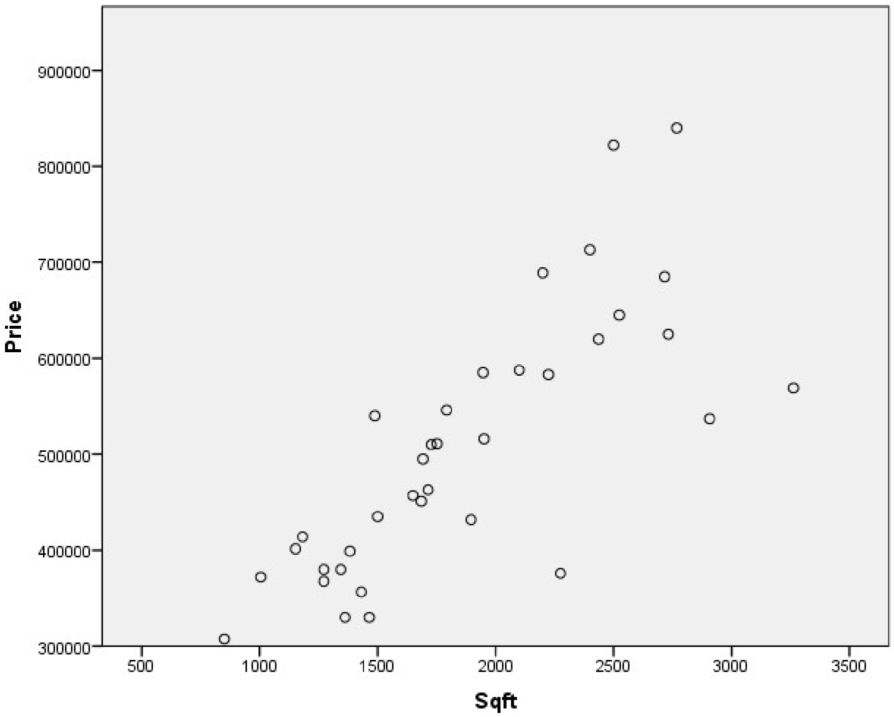
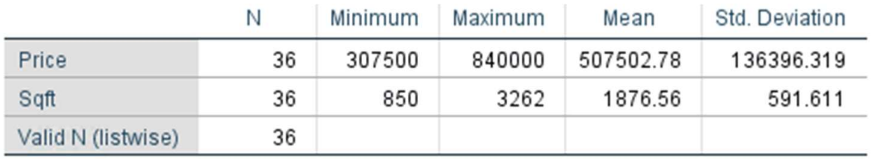
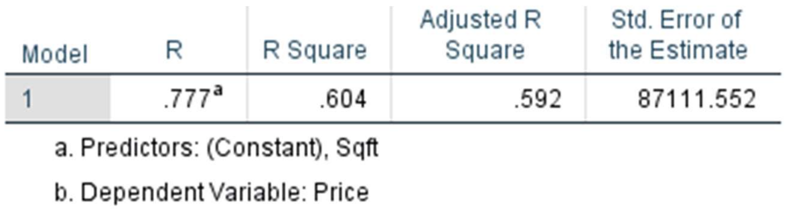


Table 1: Descriptive Statistics – Price, Sqft



	N	Minimum	Maximum	Mean	Std. Deviation
Price	36	307500	840000	507502.78	136396.319
Sqft	36	850	3262	1876.56	591.611
Valid N (listwise)	36				

Table 2: Model Summary

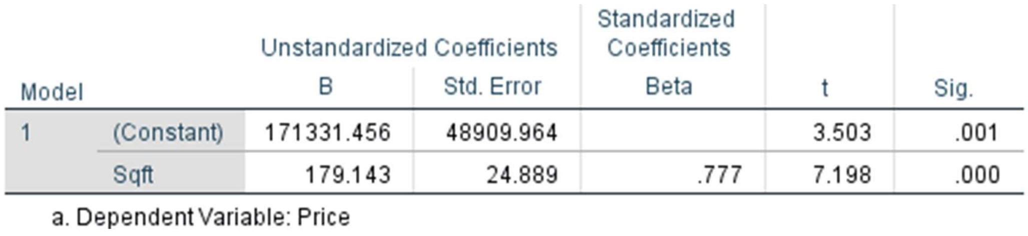


Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.777[a]	.604	.592	87111.552

a. Predictors: (Constant), Sqft

b. Dependent Variable: Price

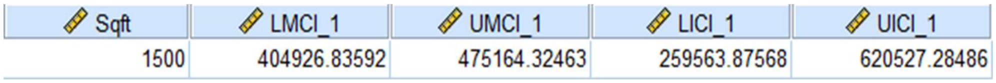
Table 3: Coefficients



Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	171331.456	48909.964		3.503	.001
	Sqft	179.143	24.889	.777	7.198	.000

a. Dependent Variable: Price

Table 4: 95% Confidence and Prediction Intervals for House Area of 1500 Sqft



Sqft	LMCI_1	UMCI_1	LICI_1	UICI_1
1500	404926.83592	475164.32463	259563.87568	620527.28486