Keep on Closing

Sagar Bansal

SECTION I: ANALYSIS SUMMARY

1) Understanding the relationships:

To understand how the relationship of square footage with house price will change after taking architecture style into consideration, we started with some data exploration to see if we have any evidence to believe that architecture style can be correlated with the House price. We have used Architecture Style and Type interchangeably in the report. Also, numerical summaries of the variables can be found in Table 1 and 2 in the Appendix.

We produced three box plots of Price variable for each of the architectural type and noticed that all three boxplots are considerably different from each other (see Figure 1 in Appendix). Each box plot gives us a sense of the Price distribution by showing us the minimum, first quartile below which 25% of the data lies, median below which half of the data lies, third quartile below which 75% of the data lies and maximum values of the Price for a particular architecture type. From the box plots, we found that the median price of Victoria type is highest, followed by the median price of Craftsman type and then, the median price of Queen Anne type. Similar sequences were observed for minimum, maximum, first and third quartile values of the Price among the three types. This gave us a strong indication that the type variable can be correlated with our response variable so we decided to include the variable term in our model.

Moreover, we also generated a scatter plot with Price on Y-axis, Sqft on X-axis and data points colored by Type variable (see Figure 2 in Appendix). We noticed that Sqft can have a positive linear correlation with our response so we agreed to include it in our model. Interestingly, when we added reference line for each of the three type of houses in the scatter plot, we found that all of them suggests a linear correlation with the response. Also, it was visible that reference lines are slanted to each other (different slopes). In other words, it showed different house types may change the relationship between square footage and house prices differently. It motivated us to further investigate this effect by including interaction terms between these two variables in the model.

After estimating the coefficients, we revealed three important correlations of Square Footage and Architecture Type with House Price according to our model (see Table 5 in Appendix). First, a 100 unit increase in the Square footage is associated with 3251.20 dollars increase in the average house price for Queen Anne house types. Second, a 100 unit increase in the Square footage is associated with 12907.5 dollars increase in the average house price for Craftsman house types. Third, a 100 unit increase in the Square footage is associated with 22569.30 dollars increase in the average house price for Victorian house types.

For the house types, the model estimates that the change in average house prices is roughly 14915.89 dollars higher for Victorian types than Queen Anne types when square footage is zero. In addition, the model estimates that the change in average house prices is 29749.83 dollars higher for Craftsman types than Queen Anne types when square footage is zero. Also, the

average house price for Queen Anne house types is roughly 166596.77 dollars when square footage is set to zero. Please note that a house of zero square footage is practically infeasible. These values are theoretical to help the client understand the individual correlation of different house types with the house prices. In practice, we will always have some value of square footage for a particular house.

One more thing, only the intercept and interaction terms are found to be statistically significant at 5% significance level. By this we mean that the relationship between the square footage and house price is affected by the type of house which we also expected during the data exploration phase. However, the individual predictor terms do not provide values relative to the variability in the House Price i.e., their specific quantities are not statistically different from 0.

2) Limitations of the analysis:

Based on our analysis, it is crucial to keep in mind that the relationships determined are only valid for house area values (Sqft) ranging from roughly 813.80 ft^2 to 3284.65 ft^2 and architecture types of Queen Anne, Victorian and Craftsman. In addition, 91.3% of the variability in the model is explained by this model after accounting for the model size. This is a great value and we will not actively encourage our client to collect additional features for future analysis considering the costs associated with data collection. However, if the client wishes to further improve the variability explained by the model or to understand the relationship between a new variable and the House Price, he may consider some of the following factors for that purpose: 1) Distance from frequently visited places such as Supermarket, Hospital and Airport, 2) Age of the property 3) Current state of local real estate market (Recession or not) and 4) Condition of the house.

SECTION II: APPENDIX

1) Statistical analysis:

We summarized the data and found that the (mean, minimum, maximum) values of Price and Sqft are roughly (475218.29, 142585.53, 939780.40) and (1907.50, 813.80, 3284.65), respectively (see Table 1). We discovered that 42 percent of all the houses in the dataset are Victorian type, 40 percent of all the houses are Craftsman type and rest of the houses are of type Queen Anne (see Table 2).

After accounting all information, we hypothesized the following Multiple Linear Regression model for further study:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D1_i + \beta_3 D2_i + \beta_4 D1_i X_i + \beta_5 D2_i X_i + \varepsilon_i$$

where Y_i is the price in USD of the ith house, X_i is the area in square foot of the ith house, $D1_i$ and $D2_i$ are two dummy variables where 1 stands for Craftsman and Victorian type, respectively and Queen Anne type is the reference level for the ith house, $\beta_4 D1_i X_i$ and $\beta_5 D2_i X_i$ are the interaction terms between X, D1 and X, D2, respectively, β_0, β_1, β_2 and β_3 are the intercept, slope and coefficients of two dummy variables, β_4 and β_5 are interaction effects, ε_i is the error term with the following assumptions: $\varepsilon_i \sim$ indp. Normal(0, σ).

After running the regression for this model, we got the estimated co-efficient β_0, β_1, β_3, β_4, β_5 and β_6 as roughly 166596.77, 32.51, 14915.89, 29749.830, 96.56 and 193.181, respectively (see Table 5). From table 3, we can note that the R square for this model came out to be 0.922 which means that 92.2% of the variability in this Price is explained by this model. Also, adjusted R square value is 0.913 which means that 91.3% of the variability is explained by our model after accounting for the model size. Note that the values of R square and adjusted R square are very close indicating that the additional terms do not have a big penalty relative to our sample size of 50. The standard error of the estimate is roughly 58336.03 which is also good relative to the mean house price of 475218.29 USD approximately. Finally, as reported in Table 4, the result of F-test shows us that our model is statistically significant with a p-value of nearly 0.000 which is less than 0.05 at 5% significance level i.e., our model is found to be useful as a whole. These different criteria strongly suggest that the model should be considered useful.

To check the validity of our model, we first looked at the Unstandardized Residual vs Unstandardized Predicted value plot (see Figure3). The plot doesn't show any clear pattern demonstrating that the model has no heterogeneity issue and it fits well. In addition, in Figure 4 we see that Q-Q plot is aligned with the normal distribution diagonal line suggesting the error terms are almost normal i.e., the model has no normality issue. Lastly, we did not find any evidence of time series structure in our data so no need to perform Durbin-Watson test i.e., no independence issue. Hence, we do not have any evidence to reject the model based on validity issues.

Furthermore, we conducted hypothesis testing for the intercept, slope and interaction coefficient at 5% significance level (see Table 5). For β_0, the null and alternative hypotheses were as follows: H_0: $\beta_0 = 0$; H_A: $\beta_0 \neq 0$. Since the P-value (0.010) is less than 0.05 at 5% significance level, we reject the null hypothesis i.e., β_0 is statistically significant. For β_1, β_2 and β_3, the null and alternative hypotheses were as follows: H_0: $\beta_1 = 0$; H_A: $\beta_1 \neq 0$, H_0: $\beta_2 = 0$; H_A: $\beta_2 \neq 0$ and H_0: $\beta_3 = 0$; H_A: $\beta_3 \neq 0$, respectively. Since the P-values for β_1 (0.265), β_2 (0.836) and β_3 (0.688) are more than 0.05 at 5% significance level, we fail to reject the null hypotheses i.e., β_1, β_2 and β_3 are not very different from what we would expect if they were zero, according to the model. For β_4 and β_5, the null and alternative hypotheses were as follows: H_0: $\beta_4 = 0$; H_A: $\beta_4 \neq 0$ and H_0: $\beta_5 = 0$; H_A: $\beta_5 \neq 0$, respectively. The P-values came out to be more than 0.05 for both β_4 (0.010) and β_5 (0.000) i.e., the interaction effects are statistically significant at 5% significance level and there is some interaction between our dummy variables and Sqft that will change the relationship of Sqft with House Price as D1 (Victoria Type) or D2 (Craftsman Type) changes.

2) Supporting figures and tables:

Table 1: Descriptive Statistics – Price and Sqft

	N	Minimum	Maximum	Mean	Std. Deviation
Price	50	142585.5285	939780.4047	475218.2909	198161.8876
Sqft	50	813.7991103	3284.654868	1907.502007	700.2886270
Valid N (listwise)	50				

Table 2: Descriptive Statistics – Type

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	"Queen Anne"	9	18.0	18.0	18.0
	Craftsman	20	40.0	40.0	58.0
	Victorian	21	42.0	42.0	100.0
	Total	50	100.0	100.0	

Table 3: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.960[a]	.922	.913	58336.02694

a. Predictors: (Constant), SqftV, Sqft, TypeC, SqftC, TypeV

b. Dependent Variable: Price

Table 4: ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.774E+12	5	3.549E+11	104.282	.000[b]
	Residual	1.497E+11	44	3403092039		
	Total	1.924E+12	49			

a. Dependent Variable: Price

b. Predictors: (Constant), SqftV, Sqft, TypeC, SqftC, TypeV

Table 5: Coefficients

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	166596.766	61941.918		2.690	.010
	Sqft	32.512	28.775	.115	1.130	.265
	TypeV	14915.893	71592.822	.038	.208	.836
	TypeC	29749.830	73576.512	.074	.404	.688
	SqftC	96.563	35.629	.472	2.710	.010
	SqftV	193.181	33.459	1.067	5.774	.000

a. Dependent Variable: Price

Figure 1: Box plot – Price vs Architecture Type

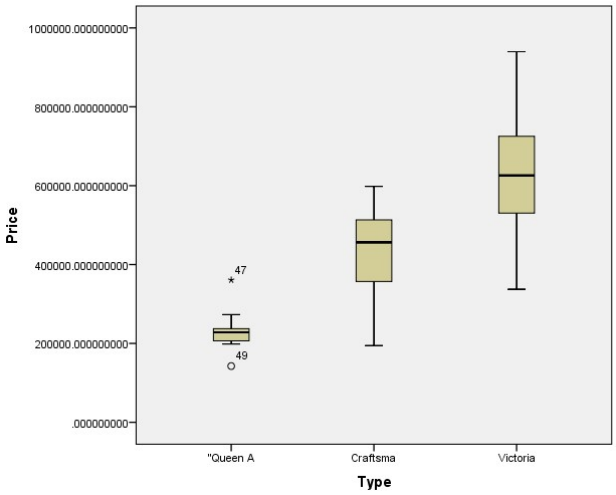

Figure 2: Scatter plot – Price vs Sqft (Coloured by Type)

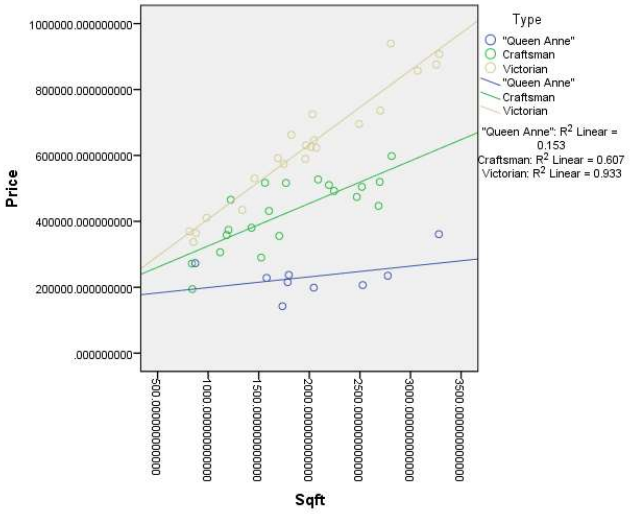

Figure 3: Scatter plot – Unstandardized Residual vs Unstandardized Predicted Value

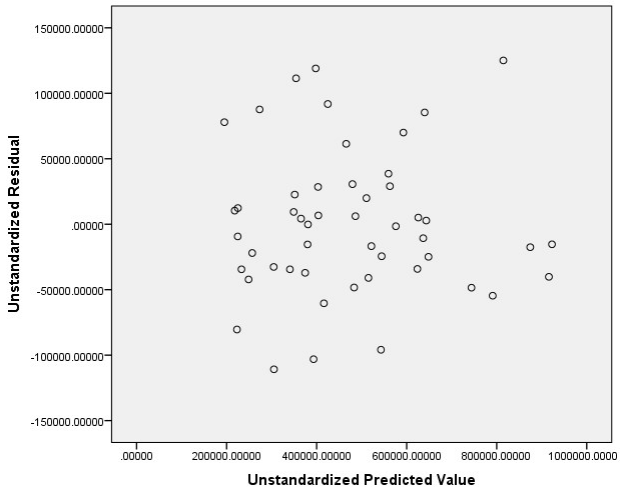

Figure 4: Normal Q-Q Plot of Unstandardized Residual