Data Analytics and Visualization

**Housing Prices Project**

Taqee Hussain

March 22, 2025

West Kelowna in BC. North of Gellatly Rd and South-West of Westside Rd.



* Based off the House Price vs Floor Space, there is a linear relationship and that all linear regression assumptions are met.



* The House Price vs Land Size graph has a weak linear relationship. There does not seem to be constant variance throughout the graph.



* The House Price vs Numbers of Bedrooms graph has a linear relationship. There could be an argument made that there is some non-constant variance, but conclude that all assumptions are met.



* The House Price vs Number of Bathrooms has a linear relationship. Although it could be argued that there is some non-constant variance, the conclusion is that all assumptions are met.



* The House Price vs Age of Building has a generally negative linear relationship. However, the variance in prices for new building (>20 years old) and the flattening pattern for older ones suggest potential heteroscedasticity and non-linearity



* Based on this graph comparing House Price to Number of Storeys there is a very weak positive linear relationship. The vertical clustering and wide price spread indicate that the Number of Storey’s is not a good linear predictor.

Correlation Matrix



* **X1 and X4 (Interior Floor Space and Number of Bathrooms)**: Correlation of **0.8575** is very strong and indicates a significant risk of multicollinearity. These two variables are somewhat redundant and may cause problems in the regression analysis.
* **X3 and X4 (Number of Bedrooms and Number of Bathrooms)**: Correlation of **0.6861** suggests potential multicollinearity concerns and redundancy between these two predictors. Having highly correlated predictors increases the standard error of the model, which can cause problems.

**Regression Model**



**Without X5**



X5 ('Age of Building') was removed first because it had the highest p-value (0.7436), clearly indicating it did not significantly contribute to explaining variations in Selling Price

**Without X3, X5**



X3 ('Number of Bedrooms') was removed because its high p-value (0.2777) clearly indicated it did not significantly contribute to predicting Selling Price. Removing X3 simplified the model without notably reducing the Adjusted R².

**Without X3, X4, X5**



X4 ('Number of Bathrooms') was removed because its high p-value (0.0899) indicated it did not significantly influence Selling Price. Removing X4 simplified the model and maintained a strong Adjusted R² without substantially impacting its predictive accuracy.

**Final Reduced Model**



I removed X6 ('Number of Storeys') because it had a high p-value (0.1191), indicating it wasn't significantly helpful in predicting Selling Price, and removing it had minimal impact on the Adjusted R²

b. The overall regression model is statistically significant, as indicated by the very small p-value (Significance F = 7.66e-32 ≈ 0.0000). This provides very strong evidence that at least one predictor (Interior Floor Space or Land Size) has a significant effect on the Selling Price of homes.

3. f) Y=β0 +β1(X1) +β2(X2)  
Where:

Y: Selling Price ($CAD)

β0: Y-intercept (House sale price when the square footage (X₁) and the number of bedrooms (X₂) is = 0)

X1: Interior Floor Space (Square feet)

X2: Land Size (Acres)

**g)** What is your estimate of the final TRUE REGRESSION model?   
ŷ = -165529.8874 + 461.6224682(X1) + 238298.2533(X2)

**h)** Clearly state the meaning of the regression coefficients in part g) in the context of this problem.  
  
β1: For each additional square foot of interior floor space, the predicted selling price increases by approximately $461.62 on average, assuming land size remains constant.

β2: For each additional acre of land, the predicted selling price increases by approximately $238,298.25 on average, assuming interior floor space remains constant.

J)

The fit of our model is determined by the Adjusted R Square value. Our Adjusted R Square value is 0.7635.

This indicates that approximately 76.4% of the variation in house prices can be explained by the predictors in our model: floor space (in square feet) and land size (in acres).

k)

I want to estimate a 95% confidence interval for the price of a house with 3500 square feet in floor space and 0.60 acres in land size.

X1 = 2700

X2 = 0.55

ŷ = -165529.8874 + 461.6224682(2700) + 238298.2533(0.55) = $1,593,127.70

Confidence interval: ŷ ± 2SE

1,593,127.70 + 2(374779.125386464)  
1,593,127.70 - 2(374779.125386464)

Upper limit: $2,342,685.95

Lower limit: $843,569.45

I am 95% confident that for a house with 3500 square feet in floor space and 0.6 acres in land size, the average price would be between $843,569.45 and $2,342,685.95.

L)

Based on statistical evidence from the regression analysis, the model does appear to be a good predictor of selling price for houses in the West Kelowna area. However, there are areas for improvement.

The R^2 value of 0.7653 indicates that approximately 76.53% of the variance in selling price can be attributed to the two predictors (Interior floor space and Land size). This suggests a strong linear relationship between the two variables and selling price. Moreover, the p-values of 4.47E-20 and 2.48E-16 for floor space and land size respectively are very low (<0.05), indicating statistical significance when predicting house price. The F-statistic further validates this with a significance level of 7.66E-32.

When it comes to limitations and areas for improvements to our model, I noticed a high standard error of 374,779, indicating the possibility for large variations in individual house prices. This could be improved by including more house listing samples in the data collection from the region. The expansion of the sample size would lead to lower standard error as the effect of outliers in the dataset is reduced, making the model more stable.