**BA 355: Business Analytics, Case 3.2**

Use the Clean Data file available on the course webpage to answer the following questions.

1. To begin, there are 113 data points. Graph **and include here** the data with the trend line for just the square feet (as the x-variable) versus the Zestimate (as the y-variable) for all 113 data points. What is the equation of the line and the r2 value? Interpret what the y-intercept and slope tell us about the cost of a house in Durango. Note: The y-intercept is weird. Just by eyeballing the graph, do there appear to be any outliers?



Y = 575x – 327419. 60%. Yes outliers.

1. Now let’s eliminate some of those outliers to make a more consistent data set.
	1. Use the standard Tukey’s method to determine which data are outliers for either the Zestimate, square footage or age. What are ranges are typical for the Zestimate, square footage and age? This should identify 13 total points that are outliers -- *eliminate* them from further consideration. [Hint: Use conditional formatting to find the outliers.] As a check, the average square footage of all data should now be 2353.57. Interpret the depressing news that Q1 tells new home buyers about home prices in Durango.

(-209372, 2184687), (-445, 5227), (-56,144). Even a “cheap” house in DRO costs $688,400.

* 1. Re-graph just square feet versus the Zestimate for these 100 data points. Redraw the graph with the regression line and equation and from your graph **and include it here**.
	2. Interpret the slope **and** y-intercept of the line from part b). According to this line, what does a square foot of housing cost in Durango?

Just the land costs $280k then $289 per square foot. $289.

* 1. Calculate the coefficient of correlation r and the coefficient of determination r2 and interpret them both.

r = .72, r2 = 52%. Decent fit, square footage explains a little over half of the cost of a house.

* 1. About how much is my 1441 square foot house worth according to part b)?

$696k – I’m rich!

* 1. Force the y-intercept to 0. (Don’t need to print this one). What does the slope say now about the cost of a square foot?

$391.

1. Using just the linear equation you found in 2b),
	1. Forecast the cost of each of the 100 homes.
	2. Calculate the absolute percentage error for each data point then compute the *mean* absolute percentage error (MAPE) and *median* absolute percentage error (Median APE – sounds like a terrible movie) for this method. 20% and 16%
	3. Go to Zillow ([What is a Zestimate? Zillow's Zestimate Accuracy | Zillow](https://www.zillow.com/z/zestimate/)) and find what the median error rate is for off-market nationwide and for the state of Colorado. Our model has a ways to go…
2. Now, run multiple linear regression (available with the data analysis package in Excel, it’s an Add-In like Solver) with the zestimate as the y-variable and all four other columns (square footage, bedrooms, bathrooms and age) as the x-variables.
	1. What do the “Multiple R and R Square” values at the top of the output indicate about how well our model is preforming? How do they compare to your answers in 2c)? Better
	2. The “Significance F” is really the overall p-value for the whole model – **if it’s close to 0% it means the model works; if it’s closer to 100%, the model doesn’t work**. What is this number – as a percentage – and what does it say about our model? Model works
	3. List the multiple linear regression equation.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |
|  |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |
| Multiple R | 0.7475 |  |  |  |  |
| R Square | 0.558756 |  |  |  |  |
| Adjusted R Square | 0.540177 |  |  |  |  |
| Standard Error | 256487 |  |  |  |  |
| Observations | 100 |  |  |  |  |
|  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 4 | 7.91E+12 | 1.98E+12 | 30.07511 | 0.0000000000000003649 |
| Residual | 95 | 6.25E+12 | 6.58E+10 |  |  |
| Total | 99 | 1.42E+13 |   |   |   |
|  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* |
| Intercept | 22360 | 117288.4 | 0.190637 | 85% | -210487.3507 |
| Sq Ft | 252 | 38.77605 | 6.505594 | 0% | 175.2810407 |
| Bed | 47436 | 36475.75 | 1.300481 | 20% | -24977.50533 |
| Bath | 42649 | 40620.26 | 1.04995 | 30% | -37992.17204 |
| Age | 1298 | 777.7055 | 1.669158 | 10% | -245.8272654 |

* 1. About how much is my house worth with 1441 square feet, 3 bedrooms, 1.5 bathrooms, and built in 1979.

$648k.

* 1. Convert the p-values for the y-intercept and x-variables into percentages and list them. Which factor seems to be the least relevant to our model? The intercept. Generally, p-values close to 100% represent *irrelevant* factors that should not be included in the model and are just noise; small p-values near 0% represent *relevant* factors.
1. Eliminate the variable with the worst (highest) p-value, keep the others and rerun the multiple linear regression. If this is the intercept, click the “constant is zero” check box.
	1. What are the individual p-values now, listed as percentages?
	2. How do the r value and r2 values compare to what you got in 4a) and 2c)?
	3. What is the new “Significance F” and what does it tell us?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SUMMARY OUTPUT |  |  |  |  |
|  |  |  |  |  |  |
| *Regression Statistics* |  |  |  |  |
| Multiple R | 0.970146 |  |  |  |  |
| R Square | 0.941183 |  |  |  |  |
| Adjusted R Square | 0.928928 |  |  |  |  |
| Standard Error | 255196.4 |  |  |  |  |
| Observations | 100 |  |  |  |  |
|  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 4 | 1E+14 | 2.5E+13 | 384.0446 | 1.02E-57 |
| Residual | 96 | 6.25E+12 | 6.51E+10 |  |  |
| Total | 100 | 1.06E+14 |   |   |   |
|  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* |
| Intercept | 0 | #N/A | #N/A | #N/A | #N/A |
| Sq Ft | 253 | 38.51708 | 6.560371 | 0% | 176.2305 |
| Bed | 50871 | 31554 | 1.6122 | 11% | -11762.8 |
| Bath | 44753 | 38895.84 | 1.150585 | 25% | -32454.7 |
| Age | 1352 | 721.1033 | 1.874747 | 6% | -79.4923 |

* 1. List the multiple linear regression equation **and interpret** all parts of it.
	2. About how much is my house worth (sqft =1441, bed = 3, bath = 1.5, built in 1979)?

$642.

* 1. Use the multiple linear regression equation to forecast all 100 home values and then calculate the *mean* absolute percentage error (MAPE) and *median* absolute percentage error (Median APE) for this method. How do these compare to what you got in part 3b)?

18% and 14%.

1. Eliminate the x-variable with the worst (highest) p-value and rerun the regression. Repeat parts a) – f) of 5).
2. We now have three models for predicting the Zestimate. One (in parts 2) and 3)) uses the square footage as the only x-variable. One (in part 5)) uses everything but the intercept and one (in part 6)) that uses only three x-variables. Compare the MAPE and Median APE for each model. Which model do you think is best? There is typically a trade-off between simplicity and accuracy. In this case, is it worth adding the extra variables – does it increase the accuracy of our forecasts enough to justify the increased complexity?

Answers will vary, but for the most part the simple model based only on square footage is almost as good and really easy to use.

1. In part 5), the coefficient for Age is positive meaning that **the older a house is, the more it is worth**; this seems counter-intuitive or even paradoxical – generally newer houses (with newer everything) are worth more and older houses – with old electrical, plumbing, etc. that need repairs – are worth less. Explain this Durango paradox. Think like a real estate agent here.

Age is a proxy for location. The oldest houses are in the most valuable neighborhoods.