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

Use the Big Zillow Fall 2021 available on the course webpage to answer the following questions.

1. To begin, there are 114 data points sorted from lowest to highest Zestimate. Graph **and print** the data with the trend line for just the square feet (as the x-variable) versus the Zestimate (as the y-variable) for all 114 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?

Equation and r^2 on graph. $460 per square foot then refund of $145k. Yes three obvious outliers at least.



1. Now let’s eliminate some of those outliers to make a more consistent data set.
	1. Use the standard Tukey’s method (using =quartile.inc()) to determine which data are outliers for either the Zestimate or for the square footage. What are ranges are typical for both Zestimate and square footage? Cost under $2,256,125 and square footage under 5797. This should identify 10 total points that are outliers -- *eliminate* them from further consideration. As a check, the average square footage of all data should now be 2366.65. Interpret the depressing news that Q1 tells new home buyers about home prices in Durango. The bottom 25% of houses are at $572k.
	2. Re-graph just square feet versus the Zestimate for these 104 data points. Redraw the graph with the regression line and equation and from your graph **and print it**.
	3. 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?

Base cost of a house is $62k then $345 per square foot.

* 1. Calculate the coefficient of correlation r and the coefficient of determination r2 and interpret them both. Square footage determines 67% of housing cost. r = 0.82, really good fit for real data.
	2. About how much is my 1441 square foot house worth according to part b)? $558,925
	3. 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? $367 per square foot.
1. Using just the linear equation you found in 2b),
	1. Forecast the cost of each of the 104 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. 24% and 13%. Not great, but simple model.
	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… 6.9% nationwide and 4.8% for Denver.
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)? 0.83 and 70%. Slightly better than before.
	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? 0%, model works.
	3. List the multiple linear regression equation.

y = -57129 + 268\*square footage +54128\*bedrooms + 54288\*bathrooms – 745 age

* 1. About how much is my house worth with 1441 square feet, 3 bedrooms, 1.5 bathrooms, and 42 years old? $541852
	2. 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? 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. 62%, 0%, 11%, 11%, 38%. Intercept is zero.
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? 0%, 10%, 12%, 18%
	2. How do the r value and r2 values compare to what you got in 4a) and 2c)? Way better, r = 0.97 and r^2 = 94%.
	3. What is the new “Significance F” and what does it tell us? 0%. Good fitting model.
	4. List the multiple linear regression equation **and interpret** all parts of it. Y = 268\*sqft +44894\*bedrooms + 49732\*bathrooms – 967\*age. Square footage, bedrooms and bathrooms all increase the price by the amounts listed but age makes it decrease by about $1000.
	5. About how much is my house worth (sqft =1441, bed = 3, bath = 1.5, age = 42)? $545,370
	6. Use the multiple linear regression equation to forecast all 104 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)? 24% and 17%, WORSE than the simple model!!!
2. Eliminate the x-variable with the worst (highest) p-value and rerun the regression. Do this one more time until you only have two variables remaining; both should have very low p-values at this point. Repeat parts a) – f) of 5).

p-values are 0% and 2%.

r and r^2 same as previous.

y = 293\*sqft + 67275\*bathroom

$523,529

Slightly better than previous with 25% and 17% but not better than just sqft.

1. 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 two x-variables. Compare the MAPE, Median APE and r-values for each. 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? In this case, simple model is superior. Adding other variables increase complexity but makes error percentages worse.
2. In part 5), the coefficient for Age is negative meaning that older houses are worth less; this makes sense as newer house have newer fixtures, etc. This is the first year that this has happened in all the years the class has collected data; maybe something is changing? Until now, there has been a paradox in part 5) with a positive slope for Age meaning **the older a house is, the more it is worth**. Explain this Durango paradox (when it happens). In Durango, age is a proxy for LOCATION. Older houses are on the desirable grid, newer houses are away from town.