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

Use the Durango Big Data Set available on the course webpage to answer the following questions. Thanks to JCP for putting this together.

1. 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 171 data points. What is the equation of the line? Interpret what the y-intercept and slope tell us about the cost of a house in Durango.
2. Now let’s eliminate some outliers to make a more consistent data set.
	1. Use the Revised Tukey’s method to determine which data are outliers for either square footage or zestimate. What are the typical ranges for square footage and zestimate? This should identify eight total points that are outliers -- *eliminate* them from further consideration.
	2. Re-graph **and print** just square feet versus the zestimate with the equation of the line. Interpret the slope **and** y-intercept of the line. According to this line, what does a square foot of housing cost in Durango?
	3. Calculate the coefficient of correlation r and the coefficient of determination r2 and interpret them both.
	4. About how much is my 1441 square foot house worth according to part b)?
	5. Try forcing 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?
3. Using just the line you found in 2b),
	1. Forecast the cost of each of the 163 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.
4. Now, run multiple linear regression 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?
	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?
	3. List the multiple linear regression equation.
	4. About how much is my house worth with 1441 square feet, 3 bedrooms, 1.5 bathrooms, and 39 years old?
	5. Convert the p-values for the y-intercept and x-variables into percentages. Which factor seems to be the least relevant to our model?
5. Eliminate the x-variable with p-value > 20%, keep the others and rerun the multiple linear regression. (A high p-value tends to indicate that a variable *is not* relevant towards the model, should not be included in it and is more like “noise”; a low p-value tends to indicate that a variable *is* relevant towards the model and should be included.)
	1. What are the individual p-values now?
	2. Interpret both the r value and the r2 value.
	3. What is the new “Significance F” and what does it tell us?
	4. List the multiple linear regression equation **and interpret** all parts of it.
	5. About how much is my house worth (sqft =1441, bed = 3, bath = 1.5, age = 39)?
	6. Use the multiple linear regression equation to forecast all 163 home values and then calculate the *mean* absolute percentage error (MAPE) and *median* absolute percentage error (Median APE) for this method.
6. Compare your results from parts 2) and 3) to part 5). Parts 2) and 3) are the simplest model using only one x-variable whereas part 5) has three x-variables. 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?
7. There is seemingly a paradox in part 5) – one of your relevant x-variables should be Age with a positive slope, meaning the older a house is, the more it is worth. *Usually*, a newer house is worth *more* since it has newer fixtures, newer plumbing, newer electrical, etc. Explain this Durango paradox.