**BA 355: Business Analytics (30 points)**

**Case 3.1: Zillow.com**

Collect 10 in town Durango data points from Zillow.com and 10 from somewhere else (your hometown, somewhere you’d like to live someday, somewhere awful – your call). Include the address, the zestimate price estimate, the square footage, number of bedrooms, number of bathrooms and year built. Keep track of which properties are in your data set as we will revisit them later. Think about other factors, too, such as property tax, location, school district, previous sales information, etc.

1. Input the two data sets into two separate Excel sheets with the address in column A, the zestimate in column B, square footage in C, bedrooms in D, bathrooms in E and age of the house in F.
2. Examine both data sets for “outliers” – data points that seem very different from the rest of the data. (We’ll learn a method for this next week, but for now just eyeball it.) Are there one or two houses that are significantly more expensive (or less, I guess?) than the others? Decide whether to include or exclude any potential outliers from the analysis below. The benefit of including them is more data, the possible issue is they may throw off your analysis below. If you do include them, they will probably stand out on the graph you plot in part c). This is decision to include or exclude is yours – no real right or wrong answer here – but think carefully about it first.
3. On the Durango data, run linear regression letting the zestimate be the y-value and using square footage as the only x-value. ***Interpret*** the slope **and** y-intercept of this line. **Graph and include** the data points with the LR line fitted to them.
4. Now, force the y-intercept to be zero and re-interpret the slope.
5. For one more way to estimate the cost per square foot, simply divide the sum of column B by the sum of column C. How does this compare to the slopes from c) and d)?
6. Using the line from part c), determine the bestimate™ for each property and compare it to the zestimate. Calculate the mean absolute percentage error for how far our bestimate™ is from the zestimate.
7. Run multiple linear regression on the 10 Durango data points (in Excel, I can show you how) using square footage, bedrooms, bathrooms and how old the house is as your x-variables (plural) and the zestimate as your y-variable. Determine the multiple linear regression equation that predicts price based on these variables.
8. Interpret the p-value and slope for each x-variable and the y-intercept. Which factors seem relevant to the model and which don’t? (Factors with higher p-values are generally *irrelevant,* factors with lower p-values are generally *relevant.*)
9. Rerun the multiple regression, eliminating the factors that are not relevant. Using this model, determine the bestimate™ for each property and compare it to the zestimate. Calculate the mean absolute percentage error for how far our bestimate™ is from the zestimate.
10. Repeat c) – f) for the other 10 data points from elsewhere. (Skip g) through l) – multiple regression on one small data set is enough.)
11. Compare the Durango data to the other data. What are the main differences between the two areas? What do the slopes tell you about the cost of a square foot of housing, etc.?
12. Estimate how much my house is worth in Durango and what it might be worth in your other location using whichever model from above that you think is best overall. 1441 square feet, 3 bedrooms, 1.5 bathrooms, built in 1979.