**BA 355: Short Assignment 3 – Big Zillow Raw Data Set**

In the accompanying data set you will find file with 155 rows of data with information about houses in Durango. In this assignment, I want you to think about how you would clean the data; I was going to make you actually do it, but 1) it’s a tough slog and 2) I want us to all have the same data for Case 3.2 so I have done it for you, but I still want you to think about what would need to be done.

1) Take a look at the data and see if you can spot some potential errors or issues. What are they? a) Consider column F – how do some rows differ from the others? What needs to be done on the last 10? b) What’s up in rows 87 through 96? c) How would you fix column F so that all the numbers are consistently the Age?

2) Assume you fixed the stuff from part 1). Now, you need to eliminate repeats (this is no one’s fault, just happens randomly). Consider how you would search efficiently for repeats. You can sort the data and look for the same numbers occurring; you shouldn’t do this with the price, however – why not? You could sort by address, but not everyone entered them the exact same way so might be hard to tell. So, assume you’ve sorted by square footage and are looking for repeats; what’s a good way to identify where the numbers are the same without just looking at each one?

3) Note that some of the addresses say “trailer” or “#202” or “apt”. The original goal here was to look at single family, detached homes. While there’s nothing wrong with buying an apartment, our data set was only supposed to include houses. Describe a method to identify (and then eliminate) these data points that are not houses systematically – sure, you could eyeball each one then decide, but this would not work if you had 110 million of these like Zillow does. Is there a systematic way to separate the apartments from the houses?