Data Wrangling Steps

The first thing that I looked at in my data set were the null values in all of my features. When the % of null values was greater than 90%, I deemed that that was insufficient data for that features, so I plan on not looking at that feature in my analysis.

Next, I looked at features that had a large amount of null values (70%+) and also contained records with a value of 0. I deemed that these features also could not be used for analysis. The reason being that it would not be a safe assumption to assign the null values to 0, given that 0 was an actual option for a record to be.

Moving on, I then looked at null values where I figured that I could adjust appropriately. I split these features that had null values into 2 scenarios:

1. Features that should not have null values and have under 10% of records containing nulls
2. Features that have valid null values

For all records that fall in Scenario #1, I deleted those records.

For all records that fall in Scenario #2, I changed the null values -> 0.

I then saw features that had records consisting of Boolean True/False. In these situations, the ‘False’ value was null, so I changed that to 0. The ‘True’ value I changed to 1. My thought process is that this would be easier to work with down the road.

I then changed all ID/year/descriptive information from Float -> Int. Considering I will not be running any quantitative analysis on these fields, I figured it was best to change them.

Next, I looked at outliers. Because I changed some of the data types, I looked at outliers on all fields with a float datatype. That would be 21 features. I currently have a count of the number of outliers in a given field as well as the % of total records that are outliers. I have code that can easily identify which parcels have outliers in any of these 21 features, however I have done nothing with outlier records to this point. I figure to leave them in my dataset until their impact can be fully analyzed.