Evan Okin's Blog Post – Module 1 Final Project

As any data scientist knows, dealing with data before analyzing it makes up more than half of the battle. There are many steps that need to be taken with data before performing analysis on it, such as understanding the file/reading it in, combining data from different sources, or scaling variables to be standardized before analysis. One of the most important steps is understanding and deciding what to do with missing data. It is imperative that a data scientist has a grasp on how to handle missing data, because most data is messy and data is often incomplete.

For the final project for Module 1, we analyzed the KC Housing Price dataset, which has over 21,000 data-points for house prices and corresponding variables (such as the number of bedrooms or the square footage of the basement). The objective was to predict house prices based on the set of provided variables. The file has missing data, in the form of blank values or strings of "?" To start, it is important to run a code to look at the type of our variables to see if it is in-line with our expectations. For example, if we look at a column for the number of rooms, we might expect an integer, assuming that there can only be a discrete number of rooms. However, if the data has a "?" in it, then the type of the entire column will be a string, because of the faulty value. For example, take the variable for the square footage of the basement of the home. The second most common value in the dataset is a "?", with 454 entries. A naïve approach to the data cleaning would be as follows: over half of the dataset (~13,000 entries) has a 0 in this value, so we can assign the missing data to be 0's. But, if we understand the variables holistically, we will see that the data has columns for square footage of the whole lot and square footage of the upper area. Therefore, the best approach would be to (1) confirm that for the majority of rows that have full data, the square footage of the whole lot is equal to the square footage of the upper area plus the square footage of the basement, (2) calculate the missing values as the difference between the columns for the square footage of the whole lot minus the square footage of the upper area.

That said, though not appropriate above, there are cases where it is perfectly reasonable to plug in the mode of the dataset into the missing values. In this dataset, there are three other variables with missing values – waterfront view, whether or not the house was viewed, and year renovated. For waterfront view, 12% of all values are missing. However, the overwhelming majority are zeros (corresponding to "no waterfront view"). Similarly. for whether the home was viewed, the overwhelming majority are zeros (corresponding to "not viewed"). 22% of the year renovated variable are missing, and 79% of the non-missing values are zeros (corresponding to "not renovated"). In all of these cases, it makes sense to input the mode. Notice how all three of these variables are categorical. If the variables were continuous (for example – square footage), other statistical metrics, such as the mean or median, could be better substitutions for missing values.

One common method is to use the mean of a dataset, which works best when variables are continuous. Using the median of a dataset works as well, because the mean can be swayed by outliers. Another method is to eye-ball each row individually and choose what "makes sense" intuitively (although this would only work if there aren't many missing values). We can also use a machine learning algorithm from the other values to predict what the missing value should be (using machine learning to calculate missing values before running machine learning on the whole thing!). There's no "right" or "wrong" answer for what to use for missing values. Let's say that there was a missing value for the number of bathrooms. We can't make a blanket statement that using the mean is always fine - if the average number of bathrooms is 3.1415, that can't be a possible value.

If the information from a missing value is valuable for our model, it might be unreasonable to even estimate it, and then it's worth contemplating deleting the entire row. This might be completely fine if the dataset has sufficient amount of data but could be problematic if there isn't much data. Also, we might not want to delete an entire row if a different variable provides value for that data-point.

While many of our missing values were set to zeros, that's merely a coincidence, and it's important to understand the process of knowing how to handle our data. There are many possible ways to deal with missing values, which is as much of an art as a science.