Capstone Project #1 – Data Wrangling

There are four files given for the problem: attributes.csv, descriptions.csv. train.csv and test.csv.

attributes.csv

The attributes file consists of a product id, the name of an attribute and the values for that attribute.

There were 155 missing values for the id and name columns and 2090 missing values for the value column. The 155 missing values were entire rows that were null, so these were deleted. The remaining missing values for the value column were filled with "None" since the lack of information for these attributes may prove to be valuable.

The product id column consists of a numerical id, so it was read into the data frame as a float. Since these are product id values, they are more appropriately classified as a string. After confirming that all the entries in the column consisted of six digits followed by a decimal point and a zero, the values were converted to a string and had the decimal point and zero trimmed.

The dataframe consists of text data, so there were no outliers.

descriptions.csv

The descriptions file consists of a product id and a text description of the product.

There were no missing values in the data.

Similar to the attributes file, the product id was read to the dataframe as an integer. After confirming that all values had the same format, the column was converted to strings.

The dataframe consists of text data, so there were no outliers.

train.csv & test.csv

The train and test files consist of an id, a product id, the name of the product, a search term. The train file also includes the average relevance score assigned by the raters.

There were no missing values in the data.

Similar to the previous files, the product id was read to the dataframe as an integer. After confirming that all values had the same format, the column was converted to strings.

Raters were to assign a relevance score as an integer between 1 and 3, so the average relevance should be a float between those two numbers. The values were plotted and examined and confirmed that they were in the proper range.

There were no outliers in the data.

Merging data

Since the train and test files contain a column for product id and product name, there seemed to be unnecessary duplication in these file. The product name column was extracted from these files, concatenated together and then merged with the descriptions file where it seemed more appropriate. These should remove duplication of data and save memory.