Data Wrangling Report

To: Udacity Reviewer

From: Miles Murphy

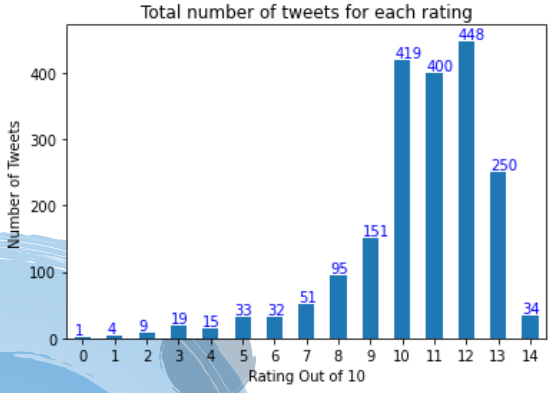
CC: GitHub or other Udacity Viewers

The following represents a summary of the wrangling objectives and activities related to the twitter_archive_enhanced.csv, image_predictions.tsv, and tweet_json.txt datasets.

The first step of the data wrangling process required that these three datasets be properly gathered by collecting any data which was not provided (in this case utilizing the Twitter API) and importing it all into pandas. This was conducted in three different manners due to the different content and file type of each dataset. Once these master dataframes were created copies were made to be used for the assessing and cleaning process.

The assessing process consisted of both programmatic (using methods like '.info()') and visual assessment. Each dataframe was evaluated to determine first its tidiness and then its quality. From the initial visual assessment, it was immediately determined that all three of these dataframes could be merged with one another, as their data was related and could be viewed together, once thoroughly cleaned for other tidiness and quality issues. Beyond combining the dataframes, three other tidiness issues were noted, though one did arise from the merging of the dataframes. One column possessed more that one data entry and another data category was needlessly split into four independent columns.

While there were relatively few data tidiness issues, there were many data quality issues. Throughout the assessing and cleaning process over 20 data quality cleaning steps were taken. Some of these steps arose from other quality or tidiness cleaning steps and others existed from the start. Several columns which were not useful for the final analysis (source, name, img_num, etc.) were dropped. Additionally, the assignment required that all the tweets retained in the dataframe have an image, not be a reply, and not be a retweet. Around 400 tweets were dropped due to their inability to meet one or more of these requirements.



Additional steps were taken to clean other quality issues such as addressing datatype issues, fixing errors in rating data, making data more concise when possible, and cleaning up columns with repetitive or incomplete information. The results of the cleaning process were a dataframe ('twitter_archive_master') featuring 1961 tweets from the original 2356 tweets. This dataframe had a handful of basic data analysis and visualization techniques performed on it which, among other details, determined that the average dog rating for the cleaned data set was 10.54/10.

[Miles Murphy - GitHub]