Wrangling Report

There are three major steps in data wrangling.

1. Gather
2. Assess
3. Clean

The data gathered here is from three sources, source number one the provided .csv file, the next file was stored on a server which was downloaded using the requests library by providing it the url this file contained image predication, and the third data was to be gathered using the Tweepy API (Application Programming Interface) but unfortunately I couldn't get access to it, so I used the data that was provided on the website, the text file.

After gathering the data from three sources, I initially started with visual assessment, just by looking at all of the data frames, trying to find which values are Null. But this didn't give a much information about the datasets that were being used, so programmatic assessment is a better choice, for all of the datasets, I have tried to find which columns contain null values, which columns data type is not right, the number of rows and columns the datasets contain, found out important statistical values such as mean, standard deviation, min, max and quartile values, tried to get a random row to see if there are any values that are not correct, and duplicated rows. The functions that I used are head(), sample(), info(), describe(), isnull(), duplicated(), sum().

After assessing the data, I noted down two tidiness issues and eight quality issues, the first tidiness issue was that the data is differentiating across three sets of data, so combining it into one and using that dataframe for our wrangling would be a good idea, the second issue was that the dog stages were not proper so they were combined into a single dog_stages column. When these two issues were solved I later moved towards cleaning of quality issues, the issues that I found here were relating to data type issues, for example 'timestamp' column was not of type datetime, there were columns that were unnecessary and were not required for data analysis, so they were dropped for example, 'retweeted_status_id' was not required for our data analysis. The other was with the column 'source' which contained hyperlinks, so I removed those hyperlinks and just kept the text which was useful for our analysis. I also used regular expressions to combine numerator and denominator into a single column. The 'name' column contained a lot of words that were not names, so that were removed from the dataframe. After cleaning the dataset I saved it and started with analysis by creating visualizations that could provide insights about out dataset.