Internal document – My wrangling efforts

First I gathered the data required from 3 separate sources provided in the instructions. The twitter archive was downloaded as a CSV. The image predictions file was downloaded from the URL given and read using pandas. The twitter data was read line by line using the alternative method approach outlined in the instructions using the file given in the instructions so that I did not need to make or use a twitter developer account. These three data sources each included unique data which was used in this project.

Issues addressed:

Quality issue 1: Unoriginal tweets were removed from the archive file.

Tidiness Issue 1: Several unneeded columns were deleted.

Quality Issue 2: I fixed the time stamp column to use the datetime format

Tidiness Issue 2: I combined the information contained in the doggo, floofer, pupper, and puppo columns into a single column.

Quality issue 3: I cleaned the source column so that the source was more easily read.

Quality Issue 4: I created a gender column by reading the tweets for gendered pronouns.

Quality Issue 5: Fixed the int type for the tweet ids in the twitter data.

Quality issue 6: Changed the numerator and denominator types to float.

Quality Issue 7: Identified several tweets and fixed them manually be searching for tweets that had a denominator not equal to 10.

Quality issue 8: Created new columns to track the first correct prediction and the associated confidence levels.

Tidiness Issue 3: Merged the 3 separate data sources together into a single dataframe.

Insight 1: The golden retriever and labrador retriever dogs are the most common, with the golden retriever beating out the labrador by 49. The difference between the golden retriever appears to be much greater than the difference between any two dogs that are next to each other in terms of how common they appear. There are eight dogs tied for last place in this data set in terms of how common they are.

Insight Two: Favorites and retweets appear to increase together. The curve might be geometric, suggesting that retweets increase faster as favorites increase. This could be an area for further investigation.

Insight Three: It appears that, on average, numerators tend to have a greater value than denominators. This suggests that ratings tend to be greater than the maximum value. This may suggest that people in this sample really liked the dogs they tweeted about