Wrangle Report

Gabriela Garcia

This project required collecting data from multiple data sources in order to do the analysis. Two data sources we're given to us to use with data about the Dog Rating Twitter account and information about predictions of pictures tweeted by the same account. Then we we're asked to use the Twitter Developers API to get the retweet count and the like count of the tweets provided by the other two data sources.

The quality issues I came across are as follows:

1. df.retweet_count should be an integer
2. df.like_count should be an integer
3. ids_df.timestamp should be a datetime object
4. retweeted tweets included in the data
5. in-reply tweets included in the data
6. not all image predictions are dogs
7. NaN values in the expanded url column
8. missing tweets not included in some of the data sources and included in others

The tidiness issues I came across are as follows:

1. doggo, floofer, pupper, and puppo columns in the given data should be combined into one variable.
2. data is spread out among 3 different sources and should be joined on tweet_id

Thankfully all data sources had a unique tweet id. This tweet id allowed me to use all data sources when making my analysis. I found the csv file to be very helpful.

The tsv file was also very helpful since the majority of the tweets did not mention details about the dogs' looks or breed. I was surprised to see that a lot of the tweets did not have a good prediction as far as what kind of dog was in the picture, I decided not to use these tweets since some of my analysis was interested in the prediction of what type of dog was in the picture.

I was excited to use the twitter API since this is something I've been wanting to explore for a while. Using the provided json for each tweet I was able to get the retweet and favorite counts for each tweet.