Wrangle Report
Kyle Santana
3/23/2018

# Data Wrangling of We Rate Dogs Twitter Data

I was given access to the following three sources of information: A file twitter_archive_enhanced.csv that I downloaded, image_predictions.tsv which I pulled from a url via a request, and tweet_json.txt which I pulled via the twitter API using Tweepy.

There were many issues with the data that I could have addressed. For the purposes of this project I worked on the following (8) quality issues and two (2) tidiness issues.

## List of Quality issues:

1. Replace & amp; in text with just &.
2. Convert id to string in tweet info dataframe.
3. Convert tweet_id to a string in twitter archive dataframe.
4. Convert tweet_id to a string in image predictions dataframe.
5. Rename tweet info id to tweet_id to merge it with the other two dataframes.
6. Convert datetime from string to datetime.
7. Remove columns that contain no information, and the redundant dog stage columns.
8. Some of the name records in Twitter Archive contain articles (the, an, a) instead of actual names. I will rename them to None for consistency.

## List of Tidiness Issues

1. Merge all lists into a master list.
2. Combine Dog Stages into one column.

## Issues that I ran into.

Most of the issues were pretty straightforward to resolve. There were three issues that were difficult to resolve. First was every time I ran the jupyter notebook cells I kept adding on to the dataframe. I finally included logic that would check if the file exist locally and skip the import. The next issue was converting datetime to the datetime datatype. To resolve the datetime issue I had to figure out the proper format that would allow the column to be properly converted. Finally was merging the 3 dataframes into 1. When trying to merge the 3 databases I ran into a lot of issues of the datatypes not matching. I went through and converted all the columns to the same datatype and after everything matched I merged all of the datasets together and saved it locally as twitter_archive_master.csv. To put together my insights and visualization I put together a dataframe of mean of the favorite and retweet counts compared to count of dog breeds.