Data Wrangling Report

By Ahmed Khater

1. Gather

There were three data sources that I gathered from:

- Already given file "twitter_archived_enchanced.csv" which contains all the major data such as breed, name, tweet_id and tweet text.
- Tweet image prediction file "image_prediction_tsv" is the output of breed of dog present in each tweet according to a neural network which can be requested or can be downloaded programmatically.
- The data from twitter's API was extracted, and written to a text file (tweet_json.txt) in json format. This txt file really just contained each tweet's tweet id, favorite count, and retweet count.

2. Assess

Using pandas info function we assessed the data and I found few issuse like:

QUALITY ISSUES:

Timestamp should be in datetime format.

Numerator ratings and denominator ratings did not have appropriate values esp ratings like 4.5 or 3.75 is not captured properly.

Name of dogs are not captured properly most of the values are "a" which is not a valid Dog name.

Archive Dataframe contains 2356 rows whereas images has only 2075.

There are 181 retweets found in archive dataframe which should be removed

There are tweets without images which should be removed

Unnecessary columns should be removed

Dog breeds and prediction data should be condensed

TIDINESS ISUES:

doggo, floofer, pupper, puppo columns in twitter_archive_enhanced.csv should be combined into a single column as this is one variable that identify stage of dog.

Information about one type of observational unit (tweets) is spread across three different files/dataframes. So these three dataframes should be merged as they are part of the same observational unit.

3. Clean

Timsetamp converted in date_time format

All the retweets and favorites needs to deleted so that dataset should be equal number of rows

Name of Dogs are not captured properly most of the names were "a" so we need to again extract all the names from text

Dogs breed are in multiple columns with multiple probabilities so we need to condensed in one column

Need to drop all the unnecessary columns

Ratings numerator was need to be extracted again as most of the decimal ratings was not captured properly

Ratings denominator also needed to be extracted again as most of the values were unusual.