**Reporting: wrangle_report**

*** Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.**

Data Wrangling often refers to as the process of removing errors and combining sets of data to make them more accessible, meaningful and easier to analyze. It involves three major steps: gathering, assessing and cleaning respectively.

My wrangling efforts started with gathering the data from three different sources using three different wrangling methods. I downloaded the WeRateDogs Twitter Archive data directly, then used the request library to download the tweet image predictions from a tsv file. Later, I used the tweepy library to query the data via the Twitter API. Using the twitter API was a bit challenging because I had to upgrade my developers account from standard to elevated to get the right permissions to get the data which took some time to get approved.

I then went ahead to assess the data to check for quality and tidiness issues. I found the following quality issues:

1) The following columns have float datatypes instead of ints datatype:

- in_reply_to_status_id
- in-reply_to_user_id
- retweeted_status_id
- retweeted_status_user_id

2) The name, doggo, floofer, pupper and puppo columns have nonnull values i.e. The none should be NAN instead.
3) The name column has invalid names such as ‘none’, ‘a’ and ‘an’.
4) We only need original ratings with images and not all ratings have images.
5) The df2 dataset should be part of df
6) We have missing values in the df1 dataset. There should be 2356 rows, but we got 2075 rows.
7) In df1, some ratings are wrong.
8) The datatype for timestamp column is erroneous.

And the following tidiness issues:

1) df2 should be part of df
2) Doggo, floofer, pupper, puppo columns show one variable.
3) Rating_numerator and denominator should be one variable rating.

I did my assessment both manually and programmatically using various pandas methods.

The third step in data wrangling is the cleaning step where you must define you issue, code it or solve it programmatically and then test if you solved it correctly. I indicated each issue, defined, coded and tested. After I was done with cleaning, and I had to store the data in a csv file and then did some analysis and visualization.