Wrangle Report

The dataset that I wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Gathering

The data come from these 3 different sources:

- The file twitter_archive_enhanced.csv that was manually downloaded from "click here for link" and stored in the local hard drive.
- The file image_predictions.tsv that was downloaded programmatically using the python request library from "click here for link" and saved in the local hard drive.
- The tweeter API using python's Tweepy library and the results were stored in the file tweet_json.txt in the local hard drive.

Assessing

Quality

for quality the basic issues that I tried to detect were:

- Non-original tweets or retweets, this include images that show up in more than one tweet.
- Duplicated records.
- Unusable records, records that don't have all the info that I need like dog name or with invalid an unrecoverable info like invalid dog names.
- Unnecessary columns, columns that I won't use for my analysis.
- Wrong but recoverable value in columns like wrong rating_numerator and rating_denominator.
- Wrong data types, for example Id columns as integer and not as string.

Tidiness

- Each variable forms a column, example: dog stage used 4 columns when it should use only 1 column
- Each type of observational unit forms a table, all the data gathered forms part of the same observational unit but was spread among 3 tables

Cleaning

- The first step was to copy all the data frames into new data frames to preserve the original data.
- Each issue found during the assessment face was addressed and fixed using pandas or python functions instead of procedures or loops.
- After fixing each issue a test was performed to ensure that the issue was in fact fixed.
- Iterate, in one case fixing one issue highlighted the existence of another, such was the case of transforming the dog stages into one column which highlighted the fact that some dogs had more than one stage, for these cases the stages were stored in one column but separated by comma.
- Next, I merged all 3 tables into only one table called twitter_archive_clean
- Finally, twitter_archive_clean was stored in the file twitter_archive_master.csv