Wrangle Report

Gathering Data

Three dataframes were extracted from different sources to do the analysis, and they are:

- A csv file which contains the WeRateDogs twitter archive, this was saved into a dataframe using the pd.read_csv function
- A tsv file which is hosted on Udacity's servers, this was downloaded using the request library, and saved into a dataframe using the pd.read_csv function
- Due to not having access to Twitter's API, a json text file was used to save the required columns (tweet id, retweet count, and favorite count) into a dataframe

Assessing Data

Quality issues

Twitter archive:

- Timestamp column's data type is object instead of datetime
- Some entries where the rating denominator is 0 which is not logical, or values other than 10
- Some columns are not needed for this analysis
- Remove rows that has invalid values in the name column (Such as a, an, and by)
- Tweets which have no name are represented as None instead of Nulls
- Convert tweet_id data type to string
- All None values in the last four columns which represent dog stage should be space ' '

Image predictions:

- Image prediction columns need to have a proper naming
- Some columns are not needed for this analysis
- Convert tweet_id data type to string

Tidiness issues

- Archive's data frame should have one column identifying the dog's stage instead of four (dogo, floofer, pupper, puppo)
- All three datasets can be merged to form one table

Cleaning Data

After solving all the quality and tidiness issues mentioned in the section above using pandas and numpy functions, the final dataframe consists of one table that has all the required data for the analysis, this had been saved to a csv file name twitter_archive_master.csv.