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.

The following were the steps applied within the wrangling process

> 1. Gathering of the data
> 2. Cleaning of the data
> 3. Analysis of the data
> 4. Reporting on the data

Gathering process

> The data was gathered from multiple sources:
>
> 1. Mined twitter data using the tweepy library
> 2. Obtaining the image data through a given URL
> 3. Reading data from `.csv` files

Within the cleaning process the following issues were identified

Quality Issues

> 1. Not all `rating_numeritor` values are less than 10 given majority of the rating is out of 10
> 2. Not all `rating_denominator` values are equal than 10 given majority of the rating is out of 10
> 3. Under the name column the value `a` isn't a name yet it appears 55 times (`None` can be used in place)
> 4. Multiple columns are missing data within majority of their rows within the twitter archive dataset
> 5. Unecessary retweets rows
> 6. Unecessary "in reply to users' tweet" rows
> 7. Wrong Data types of "timestamp", "tweet_id"
> 8. The `tweet_id` , `source` , `in_reply_to_status_id` and `in_reply_to_user_id` columns are duplicated in all the tables
> 9. A large number of the predictions contain erronious entries in the image prediction dataset

Tidiness Issues

> 1. The `doggo` , `floofer` , `pupper` an `puppo` columns can be represented in a single column known as stage (i.e. stages of dog).
> 2. The `retweet count` and `favorite count` columns are not in the twitter archive dataset where they would ideally fit into

Process

The process taken to handle the wrangling was as follows:

> 1. Once the data was gathered , it was assessed visually and programatically using various techniques that looked into each table, column and row to try identify the issues stated above.
> 2. The `Define-Code-Test` framework was then applied to the issues to try solve them
> 3. The final clean data was then saved as a copy to keep using for the next process i.e the analysis
> 4. The analysis stage handled to looking at the data in detail and gathering insights from it that would seem relevant

A number of insights such as:

Research question: What dog stage got the most favorite counts?

> This highlighted the dog stage that recieved the most favorite counts to try identify what stage the population rating tends to prefer.
>
> ### Research Question: What dog stage got the most retweet counts?
>
> This highlighted the dog stage that recieved the most retweet counts to try identify what stage the population rating tends to prefer.
>
> ### Research Question: What are the top 5 favorite tweets?
>
> This highlighted the top tweets that had the most favorite counts
>
> ### Research Question: What are the top 5 retweet tweets?
>
> This highlighted the top tweets that had the most retweet counts
>
> ### Research Question: What is the common dog stage?
>
> This highlighted the most common dog stage based on the number of tweets accounted to the stage