Wrangle Report

Data Gathering:

First, I gather the data in the three data frames:

- **df_tw_archive:** that has The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything this is load from csv file.
- **df_image_predictions:** that contains the predictions of dog breeds after running every image in WeRateDogs Twitter archive through a neural network.
- **df_tweets:** that contains retweet count and favorite count for each tweet obtained via the Twitter API.

Data Assessment:

After assessing the data in the three data frames above I found the following issues:

Quality Issues:

- Nulls represented as 'None' in columns 'name', 'doggo', 'floofer', 'pupper','puppo'
- Replace name incorrect names such as 'a', 'an', and 'the' etc. also missing names with correct name (extracted from tweet text) or NaN value if value not available
- Tweets text has value encoding issues '&' instead of '&'
- Only want original ratings so I'll drop any reply and retweets
- Source column values are not clear need to formatted
- Timestamp column contains extra "+0000"
- drop rows where text has 'don't send' as they are not ratings of dogs
- Data type are not appropriate and needs to be changed such as timestamp
- Incorrect rating denominator values taken form wrong #/# pattern (fix related rating_numerator value alongside of it)
- Incorrect rating denominator values in the float format of #.##/# (only value after decimal point were captured)
- Drop record from df_image_predictions where p1_dog if false (I'm only considering the first prediction as it is the one with highest confidence)
- Rename columns to be more appropriate, for example:
    - "timestamp" : "tweet_timestamp"
    - "text" : "tweet_text"
    - "rating_numerator" : "rating_out_of_ten"
    - "name" : "dog_name"
    - "p1" : "dog_breed_prediction"
    - "p1_conf" : "prediction_confidence"

Tidiness Issues:

- Data are scattered in 3 dataframes data need to be combined into one dataframe has the data we need

- Data in the four dog stages columns 'doggo', 'floofer', 'pupper' and 'puppo' need to be combined into one category column 'dog_stage'
- The columns 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'in_reply_to_status_id' and 'in_reply_to_user_id' of df_tw_archive will not be useful after I drop any reply and retweets. Also drop 'doggo', 'floofer', 'pupper' and 'puppo' columns
- As I'm are only considering the first prediction as it is the one with highest confidence. So I'll drop 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', and 'p3_dog' columns form df_image_predictions.

Data Cleaning:

We tacked the issues found in the assessment:

1. **Issue:** Data are scattered in 3 data frames data need to be combined into one data frame has the data we need

    - **Solved by:** Joining the three data frames in one data frame based on tweet id

2. **Issue:** Nulls represented as 'None' in columns 'name', 'doggo', 'floofer', 'pupper','puppo'

    - **Solved by:** Replacing 'None' with np.nan in 'doggo', 'floofer', 'pupper','puppo' colunms . 'name' column will be handled when incorrect names are replaced

3. **Issue:** Replace name incorrect names such as 'a', 'an', and 'the' etc. also missing names with correct name (extracted from tweet text) or NaN value if value not available

    - **Solved by:** Applying a function to get name from text that contains 'named xxx' or 'name is xxx' where the name in the data frame is incorrect ('None' or lower letter such as 'a', 'an', 'the', 'such' etc.) and replace missing name ('None') to NaN

4. **Issue:** Tweets text has value encoding issues '&' instead of '&'

    - **Solved by:** Replacing every '&' with '&' in the tweet text column

5. **Issue:** Only want original ratings so I'll drop any reply and retweets

    - **Solved by:** Dropping rows with 'in_reply_to_status_id' is not NaN then 'retweeted_status_id' is not NaN

6. **Issue:** Source column values are not clear need to formatted

- **Solved by:** Replacing the Source column values with the extracted string between the HTML tags

7. **Issue:** Timestamp column contains extra "+0000"

    - **Solved by:** Applying the function to drop "+0000" from timestamp

8. **Issue:** Data type are not appropriate and needs to be changed such as timestamp

    - **Solved by:** Converting timestamp to datetime using to_datetime function

9. **Issue:** drop rows where text has 'don't send' as they are not ratings of dogs

    - **Solved by:** Fetching rows where text doesn't have 'don't send'

10. **Issue:** Incorrect rating denominator values taken form wrong #/# pattern (fix related rating_numerator value alongside of it)

    - **Solved by:** Doing the following:
        - tweet with id (810984652412424192) has no rating (drop it)
        - tweet with id (740373189193256964) has wrong rating 9/11 instead of 14/10
        - tweet with id (722974582966214656) has wrong rating 4/20 instead of 13/10
        - tweet with id (716439118184652801) has wrong rating 50/50 instead of 11/10
        - tweet with id (666287406224695296) has wrong rating 1/2 instead of 9/10

11. **Issue:** Incorrect rating denominator values in the float format of #.##/# (only value after decimal point were captured)

    - **Solved by:** Doing the following:
        - tweet with id (883482846933004288) has wrong rating_numerator 5 instead of 13.5 which I'll round up to 14
        - tweet with id (786709082849828864) has wrong rating_numerator 75 instead of 9.75 which I'll round up to 10
        - tweet with id (778027034220126208) has wrong rating_numerator 27 instead of 11.27 which I'll round down to 11
        - tweet with id (680494726643068929) has wrong rating_numerator 26 instead of 111.26 which I'll round down to 11
    -

12. **Issue:** Drop record from df_image_predictions where p1_dog if false (I'm only considering the first prediction as it is the one with highest confidence)

    - **Solved by:** Fetching only the rows that has True in p1_dog

13. **Issue:** Data in the four dog stages columns 'doggo', 'floofer', 'pupper' and 'puppo' need to be combined into one category column 'dog_stage

    - **Solved by:** Applying a function that assign the correct dog stage to each tweet, ‘multiple' if we have multiple dog stage for tweet and nan if there is no stage

14. **Issue:** Found after testing the solved issue above number **13**:
    - tweet with id (887101392804085760) do not have a picture of a dog (drop it)
    - some tweets contain 'floof' or 'floofs' in text has incorrect dog_stage (nan instead of floofer') and some of them contain 'puppers' in text has incorrect dog_stage (nan instead of pupper) and
    - tweet with id (817777686764523521) was a pupper wrongly classified as doggo
    - tweet with id (855851453814013952) was a puppo wrongly classified as doggo
    - tweet with id (854010172552949760) was a floofer wrongly classified as doggo

    - **Solved by:** dropping tweet with id (887101392804085760) and changing value of dog_stage to correct classification

15. **Issue**: The columns 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'in_reply_to_status_id' and 'in_reply_to_user_id' of df_tw_archive will not be useful after I drop any reply and retweets. Also drop 'doggo', 'floofer', 'pupper' and 'puppo' columns

Solved by: Dropping columns 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'in_reply_to_status_id', 'in_reply_to_user_id', 'doggo', 'floofer', 'pupper' and 'puppo'

-

16. **Issue:** As I'm are only considering the first prediction as it is the one with highest confidence. So I'll drop 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog' and 'p1_dog' columns form df_image_predictions

    - **Solved by:** Dropping columns 'p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog' and 'p1_dog'

17. **Issue: - Rename columns to be more appropriate, for example:**
    - "timestamp" : "tweet_timestamp"
    - "text" : "tweet_text"
    - "rating_numerator" : "rating_out_of_denominator"
    - "name" : "dog_name"
    - "p1" : "dog_breed_prediction"
    - "p1_conf" : "prediction_confidence"

    - **Solved by:** Renaming columns:

- "timestamp" : "tweet_timestamp"

- "text" : "tweet_text"

- "rating_numerator" : "rating_out_of_denominator"

- "name" : "dog_name"

- "p1" : "dog_breed_prediction"

- "p1_conf" : "prediction_confidence"