Wrangle Report for WeRateDog Data:

There are three data sets: twitter-archive-enhanced.csv, image-predictions.tsv, tweet-json.txt

Data Wrangle:

8 quality issues and 2 tidiness issues are fixed before proceeding to data visualization and analysis.

1. Identify quality issues:

twitter_archive data:

1. There are "<a href=" and other characters in the source column, they shall be removed

2. "+0000" in the timestamp column does not provide any value, shall be removed

3. tweet_id, in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id shall be string, not float.

4. In the text, there are ratings with decimal numbers, right now only numbers after the decimal points are captured and loaded into the rating_numerator column.

5. As we don't want retweets and replies, "retweeted_status_timestamp" and other retweets and reply ralated columns becomes irrelevant , this column shall be removed.

6. Rows with 'None' in the name column has no dog name in the text

image_predictions data:

7. Entries with a False "P1_dog" value will be removed, as the model has no confidence to determine the type of dog.

8. "P1_dog" column will be removed as it does not provide information after the False "P1_dog" entries are removed.

2. Identify tidiness issues:

twitter_archive data:

1. There are 181 retweets and 78 tweet replies, those rows of entries shall be removed to ensure no duplication of the same tweet. This is following the rule of tidy data, i.e. Each observation forms a row. Therefore, we want each row to only represent a unique entry.

2. Dog types are currently in the form of wide columns, i.e. "doggo", "floofer", "pupper", and "puppo" columns, we shall combine these dog types into one single column as they are one type of variables.

Clean data set:

Fix tidiness issues first, then quality issues.

Tidiness Issues:

1. Remove rows of entries that are related to retweets or replies, they are removed to ensure no duplication of the same tweet.

2. Dog types are currently in the form of wide columns, i.e. "doggo", "floofer", "pupper", and "puppo" columns, I used lambda to join all types and then redefine the stage of dog to combine these dog types into one single column as they are one type of variables.

Quality Issues:

1. Remove extra parts of the urls in source column using the split()
2. Remove +0000 part in the timestamp using rstrip()
3. Change the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id type from float to String.
4. We can see that the rating_numerators are not correct for the text with decimals, we need to include the score on the left hand side of the decimal as well. So, let's re-extract the rating numerator with the correct regular expression
5. As we don’t want retweets and replies, “retweeted_status_timestamp” and other retweets and reply related columns becomes irrelevant, this column are removed.
6. Rows with 'None' in the name column has no dog name in the text, so replace it with np.nan
7. Entries with a False “P1_dog” value is removed, as the model has no confidence to determine the type of dog.
8. “P1_dog” column is removed as it does not provide information after the False “P1_dog” entries have been removed.