Wrangling The WeRateDogs Twitter archive

The wrangling process was separate in three stages: gather, assess and clean.

The gathering considered three sources of data. First it was obtained the file tweet_json.txt using Twitter's API after registering as a developer. This file was useful to obtain the reweeted and favorite count as well as the text length. Secondly, the df_twitter_archive_enhanced which was given as well as the third source which was the image-predictions.tsv. Each of these sources were loaded in a different way.

Once the data was loaded into dataframes it was assessed to find possible missing data, duplicated, redundant data, or data in the wrong format. The idea was to collect all the necessary evidence to pass to the next stage of cleaning. It also used Google Docs Sheets to assess the correct values of the rating numerators as some of them with decimals showed wrong data at first.

In the cleaning stage it was clear what to do and here it was really when hands got dirty. The mission was to fix at least eight quality and two tidiness issues.

QUALITY ISSUES
1. Delete columns unnecessary columns because they do not add value to the analysis
2. Fix timestamp it is a string and should be a date
3. Clean column source and leave only the name of the source
4. Create only one column for rating
5. display_text_range only use the end as that counts the number of characters
6. In columns p1, p2 and p3 some capitalized and others lower case
7. In df_tweet_json_clean table change column name id to tweet_id as the other tables
8. Change the column name to something more descriptive text_length

1. Delete columns unnecessary columns because they do not add value to the analysis
   Many columns were empty,had little values, or just did not add much information to what other columns provided so that they were dropped.

2. Fix timestamp it is a string and should be a date. This is a common problem when loading csv files and it is important to change to date because it is easier to manipulate dates rather than strings.

3. The source column was dirt with html tags so that a regex expression was needed to clean the data.

4. The rating was divided into two columns numerator and denominator. It was seen that both denominator and numerator should have been greater than ten but there were some values less than ten. The denominator did not add much value as most values were greater than ten so that this column was dropped. Only the numerator was left but after a few fixes including replacing wrong numbers because they had decimals and changing the column name to just rating.

5. Display text was initially shown as a range with a start and end number. It was assessed that the start number was not needed and only the right value was considered as it represented the text length.

6. In order to keep all values even in columns p1,p2 and p3 it was needed to lower case the values as they had some strings capitalized and some not.

7. In one of the sources df_tweet_json_clean the column representing the id of the tweet was named just id. It was renamed to tweet_id in order to match the other dataframes.

8. The old name of the column was obsolete and had to be changed to a more descriptive name such as text_length.

TIDINESS ISSUES
1. create a new variable dog_stage and delete all other columns: doggo, floofer, pupper, puppo
2. create master archive - Merge the three tables(df_twitter_archive_enhanced_clean,df_image_predictions_clean ,df_tweet_json_clean) into one master data file using merge 'twitter_archive_master.csv'

1. According to the principle that each variable forms a column. It was decided that it was more convenient to have only one column for the existing columns: doggo, floofer, pupper and puppo. However, there was the issue that some rows had doggo and floofer or other combination. Here the solution was to add or create a new value of the mix. For instance, “doggo-floofer”.

2. The final step after cleaning all issues was to merge all three data frames into one master dataset so that it can be easily analysed.