WeRateDogs Twitter Analysis - Wrangle Report

This report summarizes the data wrangling process of WeRateDogs Twitter Analysis. This report is sub-divided into 3:

- Data Gathering
- Data Assessment
- Data Cleaning

Data Gathering

Data used for this analysis was gotten from three sources:

- Twitter archive enhanced file: this is a csv file and was downloaded manually.

- Tweet image Predictions file: This data source contained the result of image predictions ran by Udacity through a neural network that can classify breed of dogs. It was downloaded programmatically using the Request library.

- Additional data from twitter API using tweepy: important information like favourite counts and retweet counts were missing from the twitter archive enhanced file. Twitter API was queried to get the additional data for this analysis in which the JSON data was written to a text file and then read line by line into a Dataframe.

Data Assessment

Files were assessed Visually and programmatically

Twitter archive enhanced file:

- Data consists of 2356 rows and 17 columns.
- Data contained 181 retweets and 78 replies.
- There were 59 Null values in the expanded url column in which 55 were replies and 1 retweet.
- The name column had 109 invalid names
- For the ratings, the denominators were mostly 10. Values which were larger than 10 were in multiples of 10(20,50, 60, 80etc) and it was discovered that they were for tweets with multiple dogs. Other values were wrongly extracted while some were replies or retweets.
- For the numerators, majority were above 10 and between 14 as this is the Uniqueness of WeRateDogs. As the case were for the denominators, outrageous values were for tweets about multiple dogs, and some were wrongly extracted.
- Doggo, floofer, pupper and puppo columns described the stage of the dog shown.
- Timestamp was also in object datatype instead of datetime.

Image Predictions file:

- This file consists of 2075 rows and 12 columns. The file had 3 columns of the top predictions of the tweet image ran through a neural network and with each corresponding prediction was the confidence level of each prediction and a Boolean column that explain if each prediction was a dog or not.
- Words in the predicted columns were separated by underscores instead of space

Additional Data from twitter API:

- This file consists of 2326 rows and 3 columns (tweet_id, number of retweets and number of likes).

Data Cleaning

All quality and tidiness issues found in the 3 files were addressed here:

- Firstly, a copy was made for each data to be used for cleaning: `tweet_archive_clean`, `image_predictions_clean`, `add_data_clean`
- Then, retweets and replies were removed as analysis focused on the original tweets made by WeRateDogs.
- Went ahead to remove remaining null values in the expanded url column since they weren't in the image predictions table and would be removed anyways during joining.
- Replaced invalid values in the name column with **None**
- Corrected ratings with wrong ratings and removed rows that contained ratings of multiple dogs(they weren't much so wasn't losing much data quality and would also have served as outliers if used for analysis).
- Changed timestamp to datetime
- Underscores between words in the `image predictions table` were replaced with space.
- Doggo, floofer, puppo and pupper column were unpivoted into a single column **dog_stage**.
- A column called **breed** was added to the `image predictions table` and it contained the most accurate prediction of dog from the 3 predictions, if no prediction contained a dog it is filled with 'Unknown'
- Lastly, breed column from the `image_predictions_clean` and `add_data_clean` were joined to the `tweet_archive_clean` table and finally stored to a `twitter archive master csv file`.
- Final table consists of 1943 rows and 10 columns.