Data Wrangling Report

Steps followed to wrangle data of 'WeRateDogs' twitter archive

Data Gathering

Data was collected from different sources:

1. Twitter_archive_enhanced.csv file
2. Image_predictions.tsv file which was downloaded programmatically using requests library. This file has data about predictions of dog breeds which was processed using a neural network
3. Using Twitter API with the help of tweepy library to collect additional relevant data like retweet count, favorite count, and followers count. The gathered data was in json format and was saved in a text file which was used to save the data into a dataframe named api_df

Data Assessment

1. Visual Assessment:

   Each piece of gathered data is displayed in the Jupyter Notebook for visual assessment

   Data also assessed in an external application like Excel or text editor.

2. Programmatic Assessment:

   Using different pandas methods and attributes like .info(), .shape, .value_counts(), etc.

Quality Issues

Archive Dataset

- ***Data Types consisteny issues:***
    - id type is integer instead of string
    - Timestamp is of object type instead of datetime
    - Prescence of retweets and replies and the rewuired is the original data only
    - Inconsistent representaion of null values as None in name column and dog stages columns
- ***Completeness Issues:***
    - Missing image urls in expanded_urls
    - Incomplete or mistakenly extracted dog names in name column
- ***Accuracy Issues:***
    - Some mistakenly extracted rating_denominator values; 7, 2, 0,11
    - Rating denominator for images that has more than one dog is multiplied by the number of dogs
    - integer numerator and it holds some weird values

Image Predictions Dataset:

- id type is integer instead of string
- Inconsistently capitalized names in p1,p2,p3
- Non-descriptive column names

api_df dataset:

- Possibly has retweets and replied within the records
- Tweet_id is integer instead of string

Tidiness Issues

- ***archive***
    - Dog stages should be in one column instead of 4 columns
    - dog stages column names ; puppo, pupper are values
- ***img_predictions***
    - Three values in three columns
    - Column names are values

Cleaning quality and tidiness issues:

By firstly making copies of the 3 datasets to avoid loss of data if something went wrong.

A. Quality issues

1. Removal of Null expanded url values by dropping cells
2. Cleaning tweets without images in archive dataframe with the help of img_predictions dataframe
3. Cleaning retweets and replies from archive-clean img_predictions_clean and api_clean dataframes
4. Converting timestamp into Datetime in archive_clean
5. Converting Id into object data type
6. Cleaning the name column issues
7. Fixing numerator and denominator issues

B. Tidiness Issues:

1. Dropping The following columns: (‘in_reply_to_status_id’, in_reply_to_user_id’, 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp') as they are of no importance now.
2. Combining all dog stage columns in one column named dog_stage and dropping the 4 columns (“puppo”,”duggo”, “floofer”, “pupper”)

3. Joining api_clean and archive_clean_2 into one major dataset and saving it into a csv file
4. Changing non- descriptive column names in image_predictions_clean dataframe and reshaping it using wide_to_long method and finally saving the dataframe into a csv file