Wrangle Report

Introduction

Real-world data rarely comes clean and as a Data Analyst it’s our job to gather, assess and clean the data to make it viable for analysis.
The dataset I’m working on is a twitter archive of user @dog_rates. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.
Their archive is provided by udacity, we need to download the image_prediction file from Udacity hosted server programmatically and Download additional data using tweepy API.
After gathering the data we should asses them and find quality and tidiness issues with it, clean it and then analyse it.

1. Gather Data

    - Twitter_archive_enhanced.csv was Provided by Udacity and was loaded into a dataframe called ‘archive’
    - Image_predictions.tsv was downloaded programmatically using Requests and was loaded into a dataframe called ‘image_pred’
    - Additional data such as favorite and retweet_count was gathered by using Tweepy API and stored in a ‘tweet_json.txt’ file and data frame called ‘tweet_count’.

2. Assess Data
    - Quality Issues
        - Tables do not contain the same number of entries. This due to non-pictures and retweets included.
        - Remove tweets without images.
        - Remove tweets with retweet.
        - Remove unnecessary columns.
        - Convert non-dog names to 'None' then make title case.
        - The rating_numerator and rating_denominator have offbeat values.
        - Several columns have empty values, such as in_reply_to_status, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.

    - Convert timestamp to datetime.
    - Change source column from ulr type to text.
- Tidiness
    - Dog Stages into 1 column instead of 4.
    - The prediction column of dog breed can be simplified.
    - Join tweet_archive, image_prediction, tweet_json into one master dataset on tweetid.

3. Clean

- Made a copy of dataframe.
- Remove tweets with retweet.
    - Remove rows where retweet_status_x is not null and finally drop the columns that are not necessary.
- Dog Stages into 1 column instead of 4.
    - Combine dog stage columns (doggo, floofer, pupper, puppo) into one 'dog_stage' column.
    - Rename None with Unknown
    - Drop the 4 columns.
- Change source column from URL Text to text.
    - Replace source links to string defining them.
- The rating_numerator and rating_denominator have offbeat values.
    - Change the rating_numerator and rating_denominator for observations with wrong value or drop them if necessary.
    - Create new column rating=rating_numerator/rating_denominator.
    - Drop rating_numerator and rating_denominator.
    - Drop tweet_id 810984652412424192 as it's rating is inaccurate.
    - Drop ratings which is not realistic.
- Convert timestamp to datetime.
- Convert non-dog names to 'None' then make title case.
    - Replace names None, a, an, the with unknown.
    - Make name title case.
- The prediction column of dog breed can be simplified.
    - Condense Dog breed Column by choosing the one which is true at first as the first column has the highest percentage than the next one.
    - Drop the columns that are not necessary.
- Remove tweets without images.

- Join tweet_archive, image_prediction, tweet_json into one master dataset on tweetid.
    - Join tweet_archive, image_prediction, tweet_json into one master dataset on tweetid.
- Drop columns such as in_reply_to_status, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.

4. Store Data

- Saved the master dataframe to csv file ‘twitter_archive_master.csv’.