Wrangling Report

Gathering

The wrangling Process starts by gathering the data, in this project we have three data sources which are twitter-archive-enhanced.csv file downloaded from the project web-page and placed in the same directory of my Jupyter notebook to be later read as a data frame called archive_df We also have the TSV file downloaded from the URL shared in the project webpage and then it was read taking into consideration the separator / delimiter as a tab ;to be stored in a dataframe called Image_predictions_df and finally the API JSON file was also downloaded due to lake of Twitter Developer account and it was read line by line and having it stored in the last dataframe called api_df

Assessment

Visual Assessment was performed as a first step to identify the existing columns in each dataframe and identify which ones could be useful to perform our data analysis which will direct our cleaning efforts afterwards Archive_df contains valuable info regarding the Tweet ID, Tweet Time and Date, Dog Rating, Name and Stage but it has some issues

1.Not all tweets are Original ones, i.e. we have some tweets which are replies and retweets

2.Not all tweets contain images

3.Dog Name was derived via “text” column parsing where it's either Meet XYZ or This is XYZ , so if the text is " This is a dog .... “, The name will be incorrectly derived as "a" , if no similar pattern is found , Name is put as "none"

Image_predictions_df source was mainly an output of another project to perform prediction for the tweet images and identify whether it is for a dog or not , so it contains the prediction of 3 algorithms which decide whether this picture is for a dog or not

API dataframe contains mainly the count of retweets / favorites

Below Quality and Tidiness issues were found

1. Archive_df
    A. rating_denominator/rating_numerator are not uniform/consistent , so rating is sometimes from 10 , others from 20 , others are from very big values like 170
    B. rating_denominator = 0 will cause the rate to be infinity
    C. Some names are missing, entered as "none" or "a" , "an" , "the"
    D. expanded_urls are missing in some entries so this means that no picture was included in the tweet
    E. doggo, floofer, pupper, and puppo as columns with Value = "None"
    F. Dog Stage as a tidiness issue should be used as a separate column and have the value doggo, floofer, pupper, or puppo ..etc instead of having a separate column for each stage
    G. timestamp is not in datetime format
2. Image_predictions_df

    A. p1,p2,p3 ...etc. as column names are not descriptive
    B. Some pictures are not for dogs
3. API_df
    A. lang column name is not descriptive
    B. "Created At" is not in datetime format
4. General
    A. Some Tweet ID's exist in one Dataframe (Data Source) but not in the others
    B. It was decided that Archive_df and API_df will be merged togther and use their merge in the analysis then do further merging for Image_predictions_df in the future as I didn't perform any analysis on the Dog Breed

Cleaning¶

It was decided to concatenate the archive and API data frames and we fixed the above issues as follows

1. Make a copy of each dataframe

2. Make "timestamp" and "Created At" as a datetime format

3. Drop the entries in Archive_df copy with NaN expanded_urls

4. Drop the entries in Archive_df copy which is having in_reply_to_status_id Or retwee t_status_id

5. Since Image_predictions_df contains pictures only so we get the tweet ID's from it a nd filter out Archive_df using the Tweet ID and do the same for API_df

6. last four columns of the archive dataframe , if value is "a" , "an" , "the" , "None" , the y are to be replaced with (NaN)

7. Combine the dog Stages in one new column called "Stage"

8. Image Prediction dataframe columns headers are fixed to be more descriptive

9. Columns with very high numerator / denominator are omitted to make the rating an alysis correct

As an output all datafrmes were containing 1953 entries and archive and API data frames were merged to have Twitter_archive_master.csv generted upon which the analysis was done