Data Wrangling Project

Gathering data

For this project, the data was gathered from three souces.

1. CSV File: twitter-archive-enhanced.csv. The dataframe was named df.
2. Image prediction file from the Udacity Website. The dataframe was named df_ip.
3. File from twitter api. Since I do not have a twitter account, I read the data from the json text file. The dataframe was name df_tweets.

Assessing data

To understand the quality and tidiness issues with the dataset, visual and programmatic assessment was used.

twitter-archive-enhanced.csv

To visually assess the data, the *.head()* and *.describe()* function was used to take a look at the data in the column. To make sure that the inspection was not just random, I sorted the data with descending *rating_denominator* to look for patterns. Just a visual inspection made clear the following quality issues

1. Outliers in the numerator and denominator
2. Missing stages of dog
3. None instead of nan in dog name
4. Lower case dog names which were actually not dog names
5. The datatype for timestamp is incorrect

Next, I did a programmatic assessment of the dataframe to find the exact issues with the data.

First, I looked at the names of the dogs and found the following issue

1. Not all the text in the dataframe was in the same format, therefore there were missing names. Some of the missing names had either "*named*" or "*name is*" in front of them, and later this will be used to find them.

First, I looked at the ratings, and found the following issues

1. Some tweets did not contain ratings for the dogs but were just tweets
2. To find the rating, the first instance of '/' was used. However, in some cases, there was text such as 9/11, 24/7. Here this resulted in erroneous ratings
3. Some ratings were for multiple dogs, and were therefore multiples of a single rating
4. Some tweets did not have any ratings
5. Some of the rating of the numerator were too high, for example, 1776, which is an outlier.
6. Some ratings have decimal points.

Next, I looked at the image prediction data frame, and found the following issues

1. There is no column that specifies a species of dog, but multiple columns are given that give a probability and associated species
2. There are columns where the image is not that of a dog

Lastly, the data has following tidiness issues

1. There are 4 columns for dog stages, we just need one
2. The three dataframes need to be combined

Cleaning data

A copy of all the data frames was made.

First I started with cleaning the dataframe (df_clean) that had the data from twitter-archive-enhanced.csv

1. Remove the retweeted rows and then the columns
2. Dropped the unnecessary columns
3. Extracted all the stages of the dogs from the string and appended it to a new column named stage, and tested to make sure that this indeed happened.
    a. Also, made rows labeled Multiple if there were more than one stage of the dog.
4. Cleaning names
    a. For the dogs with lowercase names (or not names), we have some dogs whose names are given in the following way
        i. Preceded by *named*
        ii. Preceded by *name is*
    b. For extracting the names based on the above conditions, we will
        i. Make two lists with the indices of text that contain *named* and *name is*
    c. Extract the names
    d. Append these names to the dataframe
    e. Replace other lower case items in the names column with Nan
5. Cleaning Ratings
    a. Make lists of indices for the following
        i. Ratings for multiple dogs
        ii. Ratings that have another '/' preceding the actual rating
        iii. Invalid Rating
    b. Extract the correct ratings for the first two cases, and update the dataframe with the correct ratings
    c. A list of ratings with decimal was made, and it was updated manually because there were only 5 ratings.
6. Cleaning dog_images dataframe
    a. For this purpose, I dropped all the entries which were not dogs, (p1_dog, p2_dog and p3_dog were false)
    b. Made a new column with the species of dog, based on the highest probability

c. Dropped unnecessary columns.

Once, the data was cleaned I merged the three data frames to a single data frame for final analysis.