WeRateDogs®
@dog_rates

Your Only Source For Professional Dog Ratings Instagram and Facebook ⇨ WeRateDogs partnerships@weratedogs.com

「 DM YOUR DOGS 」 weratedogs.com Joined November 2015

Wrangling Report.

-Introduction: This report discussing the process of wrangling data and the efforts that have been made to get one-well-cleaned CSV file that contains all the data needed.

Data Gathering.

I gathered data from four different sources:

1 - A pre-given CSV file.
2 - A TSV file that is hosted on Udacity's servers.
3 - AKC website .
4 - Twitter.

First, I started by loading the pre-given CSV file from my local computer into a DataFrame . Then I downloaded the TSV file programmatically to my local and loading it into a second DataFrame.I got the third DataFrame from scraping AKC (American Kennel Club) website. Finally, I used twitter API to get my last DataFrame.

Data Assessing & Cleaning.

After gathering the four data sets the main goal I was chasing was having a master data set that contains all useful and well-cleaned data from each data set.

So, I started with checking what these data sets have in common, which was common tweets for twitter_archive,tweet_status and image_predictions and common dog breed for akc_breeds with these three data sets combined .

But there were missing tweets in the first three data sets so I dropped all the tweets that don't intersect.

Then after having three data sets with the same tweets I started assessing and cleaning each one separately.

- I started with twitter_archive , from this data set the useful information I cleaned were the four dog stage ,dog name,rating numerator and rating denominator columns:
    1. For the dog stage columns i melt these columns into one column named as dog_stage and i cleaned the tweets that have two dog stages.
    2. For dog name there were wrong names that i chose to set them as NaN's Values cause the dog name for these tweets wasn't mentioned.
    3. Finally, for rating denominator there were values that are not 10 Some of them (larger than 10) are due to the tweets that have more than one dog to rate so WeRateDogs multiplied the actual rating denominator and numerator by the number of dogs for these tweets,so I chose to set both the rating numerator and rating denominator back to the actual rating for these tweets (average rating).The others that have rating denominator has been check by me manually and i figured out that some of their rating has been missgathered so I cleaned them manually.
       For the rating numerator i wanted to have tweets with the same rating style Which is(>10/10) so i excluded all the tweets that have rating numerator less than 10, by checking some of these tweets out i figured out that the majority of them are other animals or not real dogs, but there were still outliers (some values far larger than 10) needed to be cleaned, so I checked them manually and cleaned them manually
- For image_predictions all I cared about is to get a one column to know what is the dog breed for each tweet so cleaned this data set into a one useful column that gives the name of the predicted breed with the highest algorithm confidence and if them algorithm predicted no dog then return None.
- For tweets_status
    1. I dropped all the useless columns
    2. I cleaned the source column which was containing data of HTML strings to a useful string tells the source of the tweet
    3. From display_text_range column i cleaned the values of lists to one integer values represents the length of the tweet
    4. From the entities and extended_entities column i picked the tweets' images urls
    5. The full text column had values of strings of the tweet text and tweet url, so i got them separated into two different columns
- For akc_breeds

    1. I started with checking if all the dog breed names in image_predictions are existed in akc_breeds so i can join this data set to the others on this column and to make sure that all the breeds are included, so I started with uniting the dog breed style for both of the data sets, making them both lowercase and replacing the dashes and the spaces with underscores and , after doing this there were still some different dog names that are not in akc_breeds and by checking them manually i figured out that some of the names are misspelled and others have more than one name pointing to the same breed , so i chose to fix them manually
    2. Some breeds were have missing weight,height values that i checked them online and have them fixed manually.

- Finally , I joined all the data sets together to get a data set of 1400 tweets with variables describes:
  1 - The tweet id.
  2 - The time when the tweet was tweeted.
  3 - The full text of tweet.
  4 - The tweet length.
  5 - The source of tweet.
  6 - The retweet and favorite counts.
  7 - The tweet url and the url of the images the tweet contains.
  8 - The predicted dog breed of tweet.
  9 - The confidence of the algorithm of the prediction.
  10 - The rating numerator of the dog, the dog's name and its stage .