Wrangle Report - WeRateDogs Twitter Data Wrangling

Date: 2/8/2019 By: Alex Cox

Gathering

I was able to pull together the data from the WeRateDogs twitter account through a variety of means. I was able to import the twitter archive from the .csv file provided to me containing the full twitter archive for the WeRateDogs Twitter account. I saved the imported .csv files into a dataframe called "dogs". I evaluated the data by looking at the first five records in the dataframe and checking the count of records.

I then pulled the .tsv data containing dog image information about the tweets from a website using the requests package and a link to the file. I saved the data into a variable "images" then create a file called "weratedogs_img" with the data written to it. I read the file into a dataframe using pd.read_csv then check the first few records of that dataframe and the count of records.

The last data gathering step was to connect to Twitter via API. I connected to my twitter account and used a list of the tweet id's from the "dogs" dataframe to pull all tweet data via Tweepy's get_status() function and append the json portion of the tweet into a new dictionary called data['dog_data']. I used the json package to create a new file called 'json_data.txt' and pulled just the tweet id, retweet count, and favorite count into a new array. I then used the pd.DataFrame function to parse this into a dataframe called 'df'.

Assessment

Next, I copied the datasets to have clean version to revert to if anything went wrong with the data. I then assessed the datasets using many of the normal review functions in Python to get a sense of what needs to be cleaned: head and tail, describe, shape, value_counts, count, dtypes, duplicated and isnull. These evaluations gave me my list of quality and tidiness questions to take care of in the cleaning step.

Cleaning

I then cleaned the datasets based upon the questions I developed to create useful, useable, clean, tidy and good quality data to develop insights. I went through and removed null data, dropped rows of data that was unnecessary for the analysis, corrected data types, parsed out and corrected dog names from tweet text, corrected dog names that don’t exist to ‘null’, parsed out stages from text to correct errors in stages of dog and combined ratings into one string type column. I then tidied the datasets by combining missing data into the main dataframe, removing unneeded columns and combining dog stage data into one category data column to finish up cleaning the data into three usable datasets.