WeRateDog Twitter Wrangling Report

Gathering Data

The Data for this project combined from three sources:

1. The WeRateDogs Twitter archive csv file I download this file manually and uploaded it to my jupyter notebook using pandas red_csv finction.

2. The tweet image predictions file is hosted in udacity's server i downloaded it programmatically using request library to makes a HTTP request and return the required content by using request.get,then saccess it using .content and save it to my computer in the same folder with name image-predictions.tsv, then open this file using pandas read_csv function

3. Favorite_count,retweet is conducted by accessing twitter API using tweepy library then search favorite and retweet counts for each tweets_id in twitter archeive t, I stored the JSON data in a text file, then loaded what I needed into a pandas dataframe.

Assessing Data

After gathering our data i visually and programmatically assess the data using value counts, number of non-null entries, and numeric summaries to detect and inspect any quality or tidiness issue .

I notice that there is a lack of tidiness such as:

1. There is three table when actually it sould be only one.

2. Some columns should be comined into one column to make table easy to analyze like dog_stage,dog_type predicted.

Also there are many quality issue for example:

1. Wrong dogs name such as (a,the, very,...)

2. Wrong rating

3. unrelated twitter

4. Erroneous data types

Cleaning

First I make a copy for each data to be able to look back to the messy data ,then i started to clean each issue was defined in the assessing step programmatically as possible unless the issue are one-off occurrence .The cleaning step is mainly construct with three main steps:

1. define

2. code

3. test