Wrangle report

This dataset is from the tweet archive of twitter user @dog_rates, which known as We Rate Dogs. We Rate Dogs is an account which displays people rating of their dogs and their comments.

Wrangling was proceed in the following stages.

- Gathering data
- Accessing data
- Cleaning data

and then,

- Storing, visualizing, analysing wrangled data
- Reporting on wrangle efforts and wrangle act report(conclusions)

Gathering data

After gathering data from three different sources:

1. twitter-archive-enhanced.csv
   this file was read by pandas function pandas.read_csv
2. image- predictions.tsv
   this file was downloaded programmatically and then readed using pandas function.
3. tweet-json.txt

Accessing

After accessing data visually and programmatically I found the following quality issues:

in twitter archive.csv

- retweets must be dropped
- replies must be dropped
- tweets without images must be dropped
- numerator and denominator msut be floats
- invalid denominators like (0,15,70,7,11,150,11,170,50,80,90,etc..)
- time stamp must changed its's to datetime64

- tweet_id better cahnged to object as it's not used to make calculations
- missing values in name cloumn
- * missing info in doggo floofer pupper puppo cloumns **in image_prediction.tsv**
- some of p1,p2 and p3 values begins with upper case letter while the rest besgins with lowercased letter.
- 66 duplicate in jpg_url column

And the following tidiness issues :

- three tables must gathered together in one table

in ` twitter_archive.csv`

- * table columns (doggo floofer pupper puppo) must tided into one column
- rating(numerator and denominator) must tided into one cloumn called "rating"

cleaning data

after accessing data, I cleaned dataset by fixing quality and tidiness issues each by following this steps define, code and test.

In twitter-archive-enhanced.csv

- drop rows with missing values in expanded_urls cloumn in `twitter_archive.csv`
- correcring "not 10" denomirators
- drop unnesecsary columns
- replace none with Nan in `doggo floofer pupper puppo` in `twitter_archive.csv` and tide the columns "doggo floofer pupper puppo" by adding, then drop unnecessary column
- Extract pet name information from text

- change timestamp type
- change tweet_id type

`image_predection.tsv` table

- lowercase all records in p1,p2 and p3 cloumns
- change tweet_id type
- change tweet_id type
- drop jpg_url duplicates

tweet_json

- renaming id column in tweet_json

then stored data in twitter-archive-master after merging the the three data frames in one dataframe, and started analysing