Wrangle Act

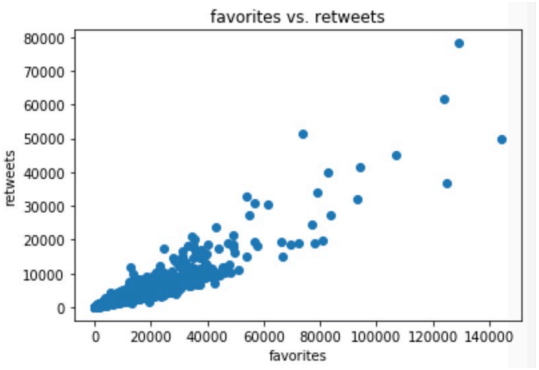
I gathered data from three different sources for this data cleaning. The WeRateDogs Twitter archive and The tweet image predictions are downloaded from udacity website directly. And Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data. Then I merged all these three sources data into one, then began the data assessing and data cleaning.

Once I had successfully gathered all the data, I copied the files for the assessment and data cleaning processes. I evaluated the dataframes looking for quality and tidiness issues in excel visually and also displayed in the Jupyter Notebook for visual assessment purposes. According the project motivation, I know there are some problems with the ratings and dog names and dog stages, so I take a quick look of these at first. So I find these problems did exist. And there are also other problems as below:
* The timestamp column should be datetime instead of object.
* The tweet_id column should be string instead of int.
* The predict data for p1, p2, p3 the case are not consistent
* The name column has many cells which do not look like correct names. such as "a", "bo" and so on.
* we only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.
* Some columns have missing values, such as in_reply_to_status, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.
* In several columns, null values are not shown as null values.
* get the rating instead of numerator and dominator
* a lot of columns are irrelevant
* Reindex the all three tables using the tweet_id
* Parse the datetime information into seperate columns date and time
* The dog stages should be values instead of columns.

Then I began to fix them. I began the cleaning process by deleting the unrelated columns, change the datatime datatype. addressing missing data and mislabeled information. I then converted columns to a proper data values, condense 'pupper', 'puppo', 'doggo', 'floof' into dogtype. Then I extract the exact rating and name from the dataset. And deleted all the incorrect and irrelevant columns.

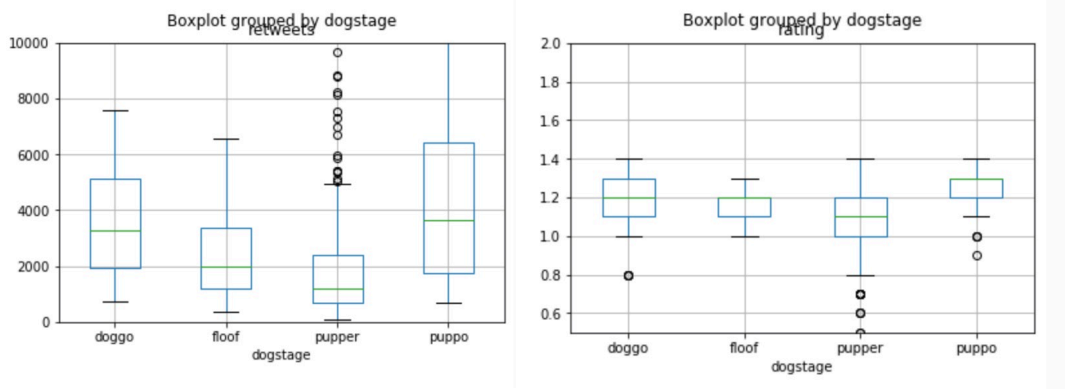
After clean all the data, I made several plots to do some simple research.
First I want to know if more favorites means more retweets and high ratings. So I used correlation matrix and scatter plots to identify it.

	favorites	rating	retweets
favorites	1.000000	0.022577	0.916974
rating	0.022577	1.000000	0.023412
retweets	0.916974	0.023412	1.000000



from the correlation matrix and scatter plot. we can find there is a strong relationship between retweets and favorites. Which is very reasonable.

Then I wonder is there any differences in rating and retweets for different dogstage



So from these two boxplot, we can see, the pupper has the lowest rating and retweets.