Wrangling Data Act Report

The Twitter account WeRateDogs (@dog_rates) is devoted to humorously reviewing pictures of dogs doing adorable poses. Dogs are rated on a scale of one to ten, but are invariably given ratings in excess of the maximum, such as "16/10". WeRateDogs has over 6million followers and has received international media coverage.

After finishing the gathering data process, I assessed data using various functions and find out that the data has various quality and tidiness issues :

Quality

Low quality data is commonly referred to as dirty data. Dirty data has issues with its content. The Data Quality Dimensions are Completeness, Validity, Accuracy and Consistency

- tweet_id is an integer
- Several columns have empty values, like in_reply_to_status, in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp.
- The name column has many entries which had incorrect values. The most frequent entry in name column is "a", which is not a name.
- Tweets with no images
- The timestamp column is an object so it has to be a datetime object.
- There are 2075 rows in the image_predictions dataframe and 2356 rows in the archive dataframe.
- In doggo, floofer, pupper, and puppo columns, null values are not represented as "NaN" values.
- Found an instance of a name being "O" instead of "O'Malley"
- Source is in HTML format with a and \a tags surrounding the text

Tidiness

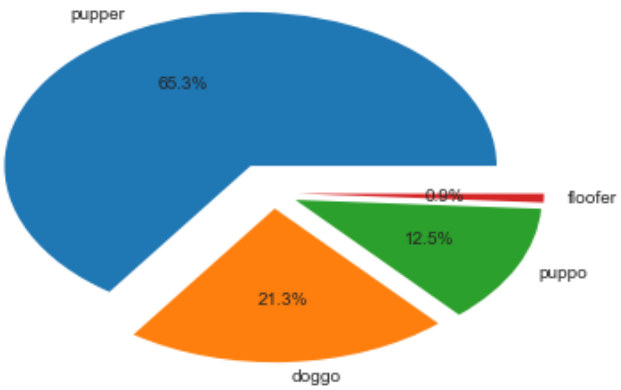
Untidy data is commonly referred to as “messy” data. Messy data has issues with its structure.

- Merging 'tweet_info' and 'image_predictions' to 'twitter_archive'
- Doggo, floofer, pupper, puppo should be column values but are instead column headers.
- Two values in the timestamp column: date and time.
- Adding the favourite and retweet columns to twitter_archives table from the tweets_info table.

After finishing wrangling part which encompass the given data, we want then to answer some questions by using visualization diagrams.

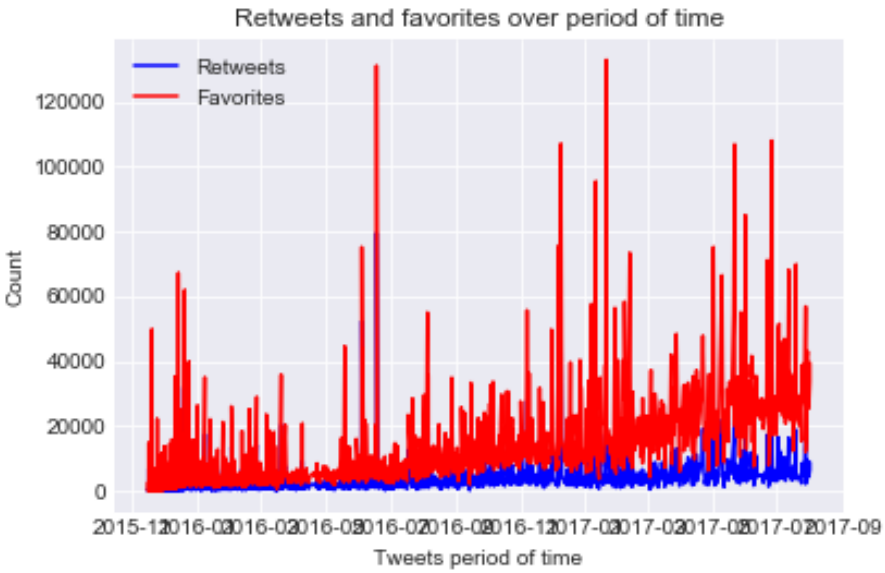
Most common dog category

We used a pie chart to find the percentage of the most prevalent dog categories and their percentage. As we can see from the list of dogs categories like doggo, floofer, pupper, puppo the highest favorites rate was wone by pupper. Seems like the puppers where the most appreciated and has the highest percentage of 65.3 %.



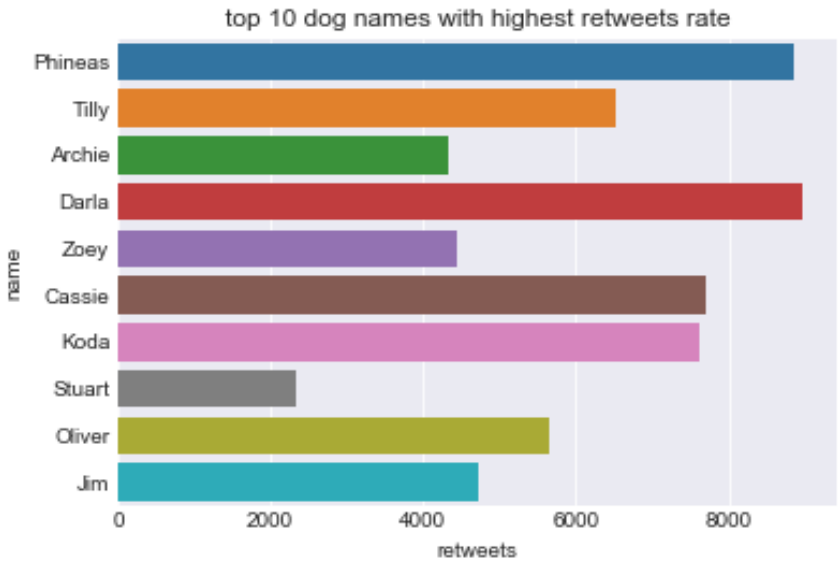
Discover the ratio of dog rating distribution:

I've created a rating_ratio variable by dividing the rating numerator by the rating denominator to normalize scores which are not out of 10 .



Visualisation out the top 10 of dog names with retweets rate.

For a better visualization I removed the none values and then proceed to find out the top 10 of dog names which have recorded the highest retweets rating. As we can see from the chart the name of dog "Darla", which is very closed to “Phineas” recorded the highest rate of retweets.



Resources:

"https://medium.com/ub-women-data-scholars/we-rate-dogs-twitter-data-analysis-672e1a8903b4"

“#http://docs.tweepy.org”s

"https://stackoverflow.com/questions/24813673/split-datetime-column-into-a-date-and-time-python"