Data Wrangling and Analysing Report - WeRateDogs Twitter

BY YANG, Linjing

2019/02/10

Introduction

Real-world data rarely comes clean. Using Python and its libraries, I gathered data from a variety of sources and in 3 different formats, assessed its quality and tidiness, and then cleaned it, which is called data wrangling.

Goal

Practice data wrangling using **WeRateDogs Twitter** data in order to create interesting and trustworthy analyses an visualizations.

Tools

I documented my wrangling efforts in a Jupyter Notebook, plus showcased them through analyses and visualizations using Python and its libraries (pandas, numpy, requests, json, tweepy, etc.).

Dataset

My dataset consists of data from 3 different sources.

1. Twitter Archive

The first one is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**, which is a Twitter account that rates people's dogs with a humorous comment about the dog.

WeRateDogs has over 4 million followers and has received international media coverage.

Its ratings almost always have a denominator of 10, but the numerators are always greater than 10 (e.g. 11/10, 12/10, 13/10, etc). Why? Because "they're good dogs Brent."

2. Additional Information via the Twitter API

The second one is retweet count and favorite count of each tweet gathered from Twitter's API by Tweetpy Library in a Json format.

3. Image Prediction File

The third is the prediction of dog images and breeds downloaded programmatically by Requests Library from the neural network.

What I did

- Examine the 3 datasets (6,787 total entries)
- Use Python to wrangle and analyze them
- Create a custom visualization to communicate observations

Data Wrangling

Data Gathering

Data was gathered from 3 different sources.

1. Twitter Archive

The Twitter archive #WeRateDogs, a csv file that contains various kinds of information, such as dog ratings, stages, tweet ids, posted date, etc.

2. Additional Information via the Twitter API

The retweet and favorite count of each tweet gathered from Twitter's API by Tweetpy Library in a Json format.

3. Image Prediction File

The prediction of dog images and breeds downloaded programmatically by Requests Library from the neural network.

Data Accessing

Data was accessed based on both of quality and tidiness issues.

Dataset 1

Quality issues

Completeness:

- in_reply_to_status_id: 78 out of 2356 is non-null
- in_reply_to_user_id: 78 out of 2356 is non-null
- retweeted_status_id: 181 out of 2356 is non-null
- retweeted_status_user_id: 181 out of 2356 is non-null
- retweeted_status_timestamp: 181 out of 2356 is non-null
- name: string 'None' should be replaced by Null
- Since only original ratings (no retweets) that have images, the rows of retweets / replys could not contribute to this analysis and should be deleted.
- After retweet / reply rows are deleted, the following columns could be dropped:
  - in_reply_to_status_id
  - in_reply_to_user_id
  - retweeted_status_id

  - retweeted_status_user_id
  - retweeted_status_timestamp

Validity:

- rating_denominator contains invalid values (e.g. denominator = 0, index = 313)
- invalid names like *an, the, only*, etc.

Accuracy:

- rating_numberator contains extremely large values (e.g. 1776 when denominator = 10)
- Since numerators are extracted from 'text', the numerators with decimals were wrongly extracted (index = 45, 340, 695, 763, 1689, 1712)

Data Types:

- tweet_id: int -> obj
- timestamp: obj -> datetime
- in_reply_to_status_id: float -> object
- in_reply_to_user_id: float -> object
- retweeted_status_id: float -> object
- retweeted_status_user_id: float -> object

Tidiness Issues

- doggo / floofer / pupper / puppo: could be combined into 1 column as categorical data.
- source: the long name with HTML tag could be shorten.

Dataset 2

Quality issues

Completeness:

- retweet & favourite count: "Nan" should be replaced by Null.
- retweet & favourite count: 16 missing values

Data Type:

- retweet & favourite count: obj -> int

Dataset 3

Quality Issues

Consistency:

- The predicted names are written in both upper and lower cases.

Data Types:

- tweet_id: int -> obj

The column names could be more clear.

Data Cleaning

- The three datasets were combined into one with more clear column names.
- The columns / rows that could not contribute to analysis were dropped.
- The dog stages were combined into one column as categorical data.
- The HTML tags in the ‘source’ column were deleted.
- The row with invalid rating denominator was deleted.
- The wrongly extracted rating numerators were corrected according to the information in *text* column.
- The rating outliers were deleted.
- Missing values were presented as Null.
- The data types of all columns were corrected.
- Unclear column names were renamed.
- The predicted dog types were changed to lowercase.

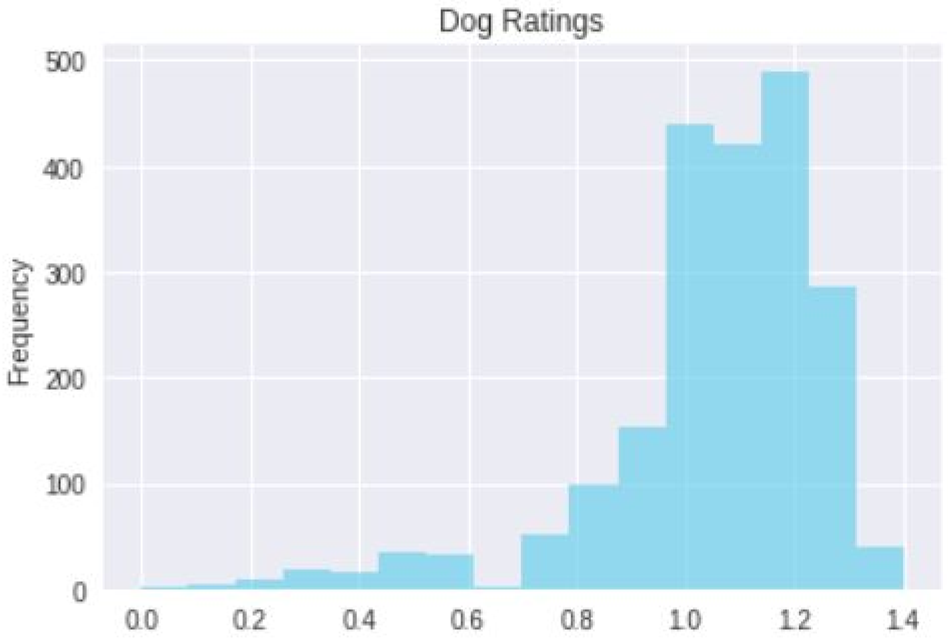
Data Analysing

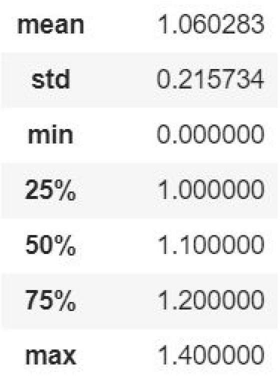
1. Dog Ratings

Since there are various rating denominators (although most of them are 10), the rating of dogs is calculated by dividing the numerator by the denominator.

The outliers have been deleted in Data Cleaning section.

It shows that the ratings of dogs are left-skewed distributed, with the mean of 1.06 and the median of 1.10. Besides, 25% of the dogs are rated equal to or more than 100%. Thus, it can be found that most dogs are considered to be better than perfect.

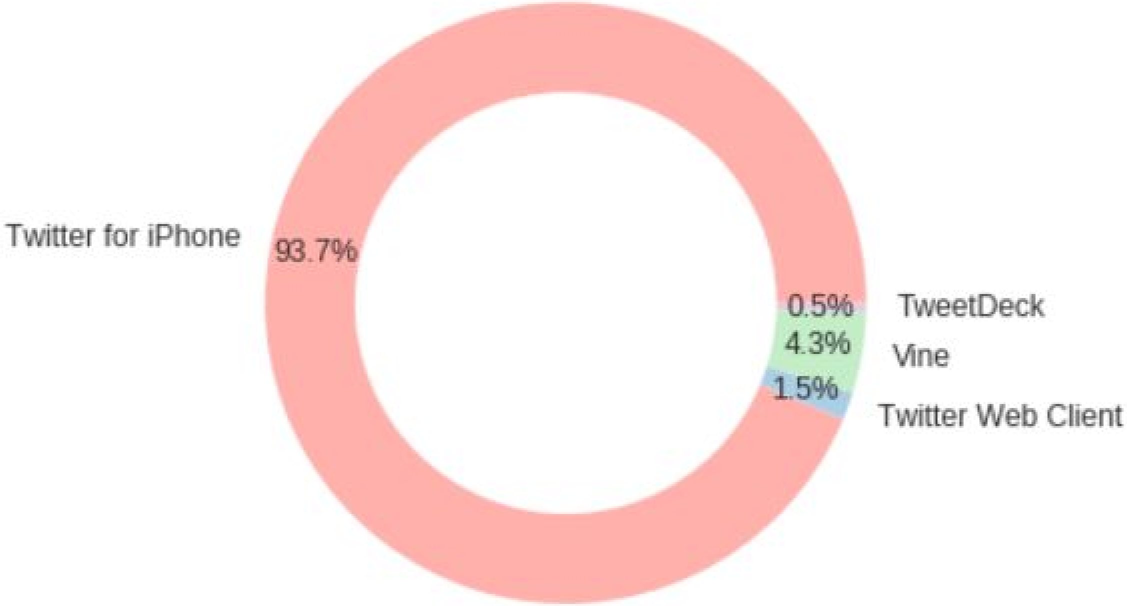




mean	1.060283
std	0.215734
min	0.000000
25%	1.000000
50%	1.100000
75%	1.200000
max	1.400000

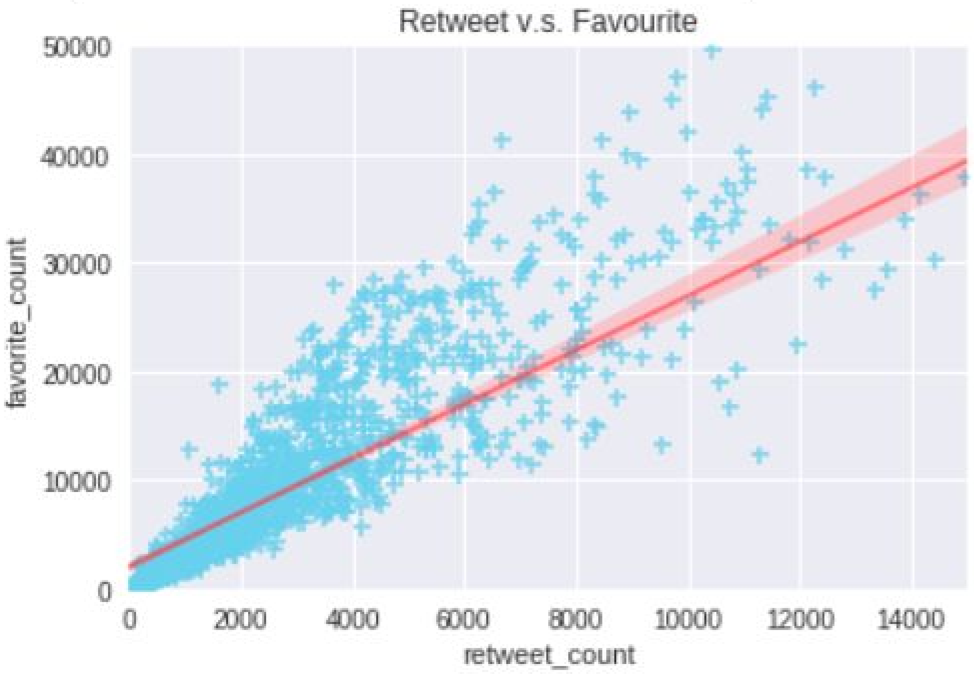
2. Source of Tweet

From the pie chart, it presents that the dominant source is from iPhone, which is 93.7%. Only a few people use Vine (4.3%), Website (1.5%) and TweetDect (0.5%) to browse WeRateDogs Tweet.



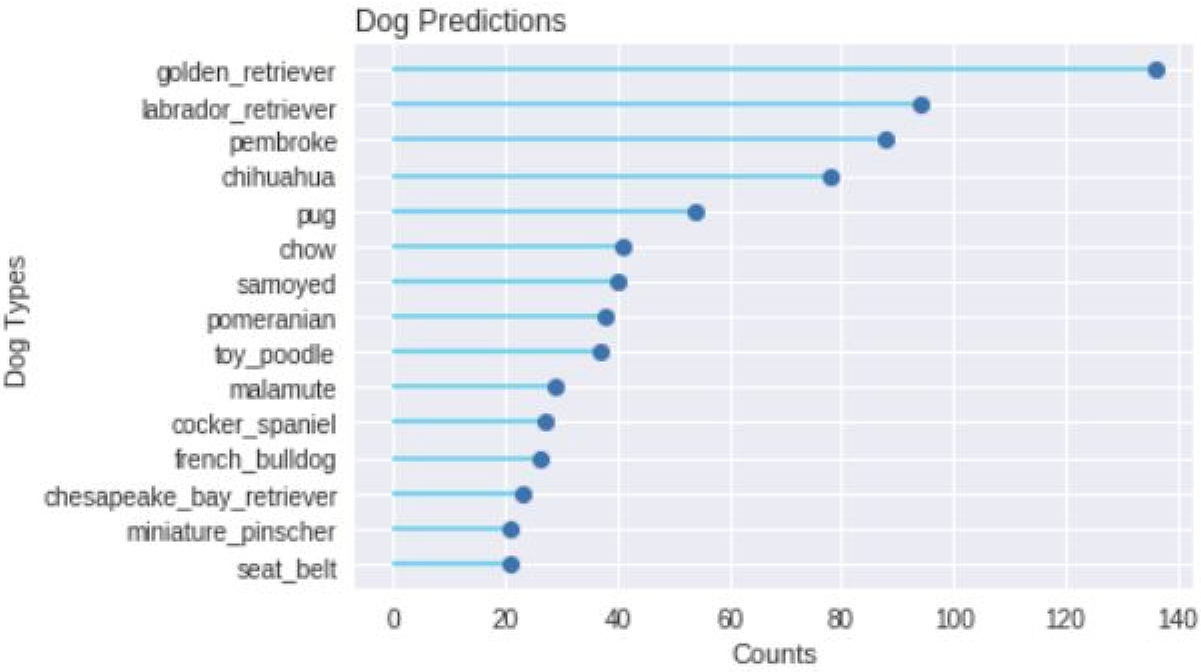
3. Relationship between Retweet and Favourite

The count of retweet and favourite are highly positively correlated (*r* = 0.927). Thus, we could say that the more people like a tweet, the more they retweet it.



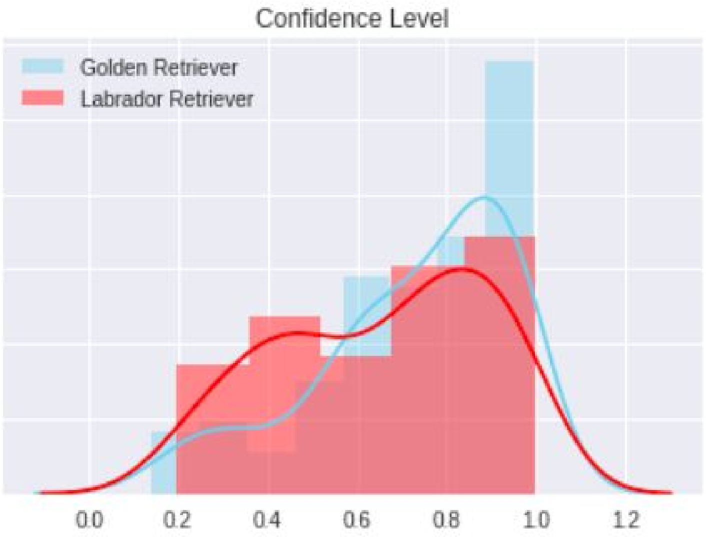
4. Breeds

The top 15 predicted dogs are shown in the plot.



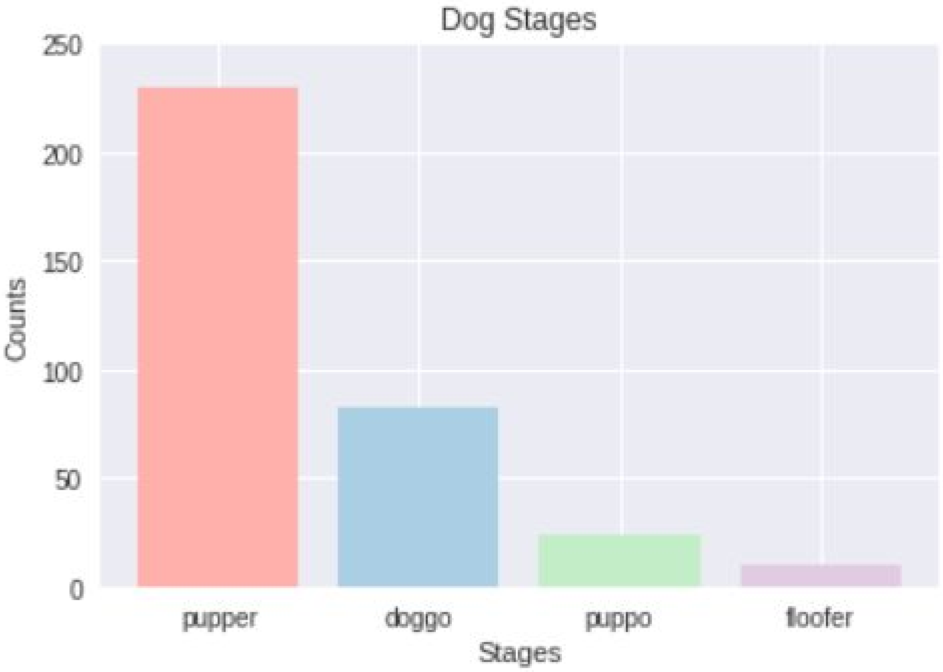
It can be seen that **Golden Retriever** is the No.1 predicted dog, which has been predicted 150 times with high confidence. The confidence of predicting it is left skewed, with the median of 0.78 and mean of 0.72.

The second most predicted dog is **Labrador Retriever** (100 times), the high confidence of which is also left skewed with the median of 0.71 and mean of 0.67.

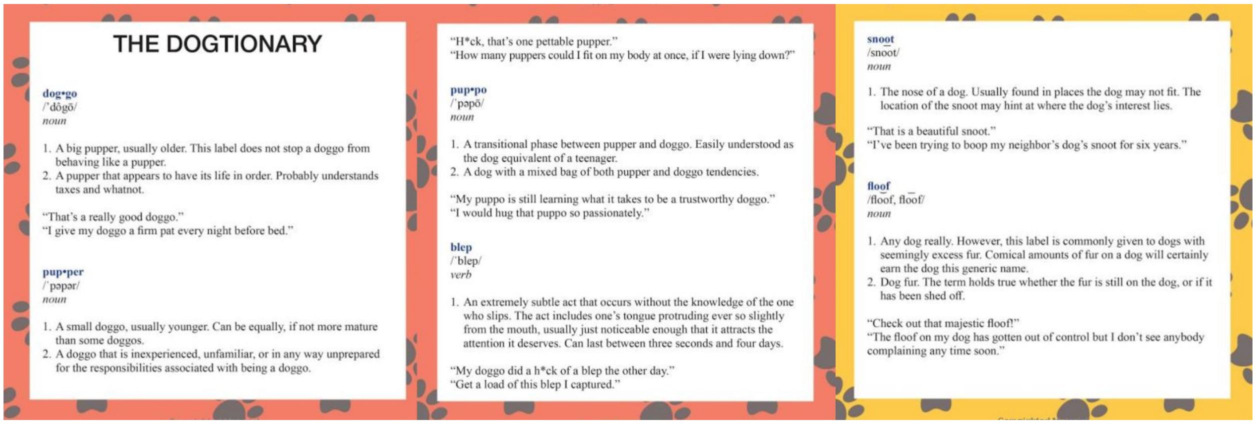


5. Dog Stages

Only 336 out of 2094 dogs have their stages presented in the dataset. The most common stage is pupper (230), followed by doggo (83) among the dogs whose stage has been presented.



The explanation of dog stages is shown as follows.



THE DOGTIONARY

dog•go
/ˈdôgō/
noun

1. A big pupper, usually older. This label does not stop a doggo from behaving like a pupper.
2. A pupper that appears to have its life in order. Probably understands taxes and whatnot.

"That's a really good doggo."
"I give my doggo a firm pat every night before bed."

pup•per
/ˈpəpər/
noun

1. A small doggo, usually younger. Can be equally, if not more mature than some doggos.
2. A doggo that is inexperienced, unfamiliar, or in any way unprepared for the responsibilities associated with being a doggo.

"H*ck, that's one pettable pupper."
"How many puppers could I fit on my body at once, if I were lying down?"

pup•po
/ˈpəpō/
noun

1. A transitional phase between pupper and doggo. Easily understood as the dog equivalent of a teenager.
2. A dog with a mixed bag of both pupper and doggo tendencies.

"My puppo is still learning what it takes to be a trustworthy doggo."
"I would hug that puppo so passionately."

blep
/ˈblep/
verb

1. An extremely subtle act that occurs without the knowledge of the one who slips. The act includes one's tongue protruding ever so slightly from the mouth, usually just noticeable enough that it attracts the attention it deserves. Can last between three seconds and four days.

"My doggo did a h*ck of a blep the other day."
"Get a load of this blep I captured."

snoot
/snoot/
noun

1. The nose of a dog. Usually found in places the dog may not fit. The location of the snoot may hint at where the dog's interest lies.

"That is a beautiful snoot."
"I've been trying to boop my neighbor's dog's snoot for six years."

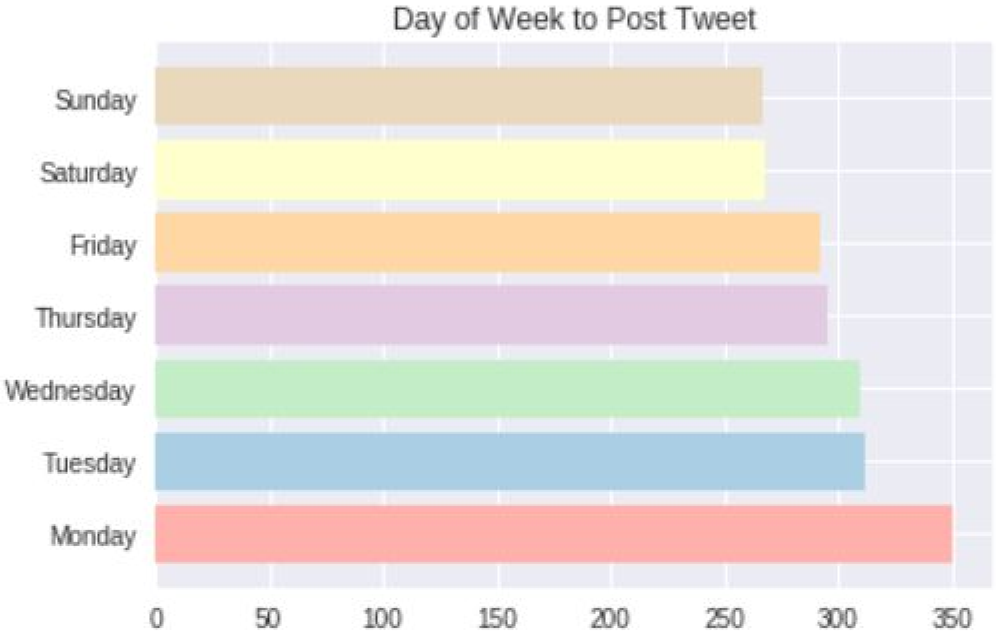
floof
/floof, floof/
noun

1. Any dog really. However, this label is commonly given to dogs with seemingly excess fur. Comical amounts of fur on a dog will certainly earn the dog this generic name.
2. Dog fur. The term holds true whether the fur is still on the dog, or if it has been shed off.

"Check out that majestic floof!"
"The floof on my dog has gotten out of control but I don't see anybody complaining any time soon."

6. Post Tweet Day

From Monday to Sunday, the number of posted tweet decreases. WeRateDogs followers may expect to see more new tweets on Monday.



Conclusion

1. 25% of the dogs are rated equal to or more than 100%. Thus, most dogs are considered to be better than perfect.
2. The dominant source is from iPhone (93.7%).
3. The count of retweet and favourite are highly positively correlated - the more people like a tweet, the more they retweet it.
4. The top 1 predicted dog is golden retriever, followed by labrador retriever.
5. The most common stage is pupper, followed by doggo.
6. From Monday to Sunday, the number of posted tweet decreases.