WERATEDOGS DATA WRANGLING PROJECT

-Ifedayo Ayilaran

←
JustFood ForDogs.
More life. More years. More love.
trupanion
Medical insurance for the life of your pet.
•••
Follow

WeRateDogs®
15.8K Tweets

WeRateDogs®
@dog_rates

Your Only Source For Professional Dog Ratings Instagram and Facebook →
WeRateDogs partnerships@weratedogs.com | nonprofit: @15outof10

Professional Services all our links → links.weratedogs.com
Joined November 2015

21 Following 9.3M Followers

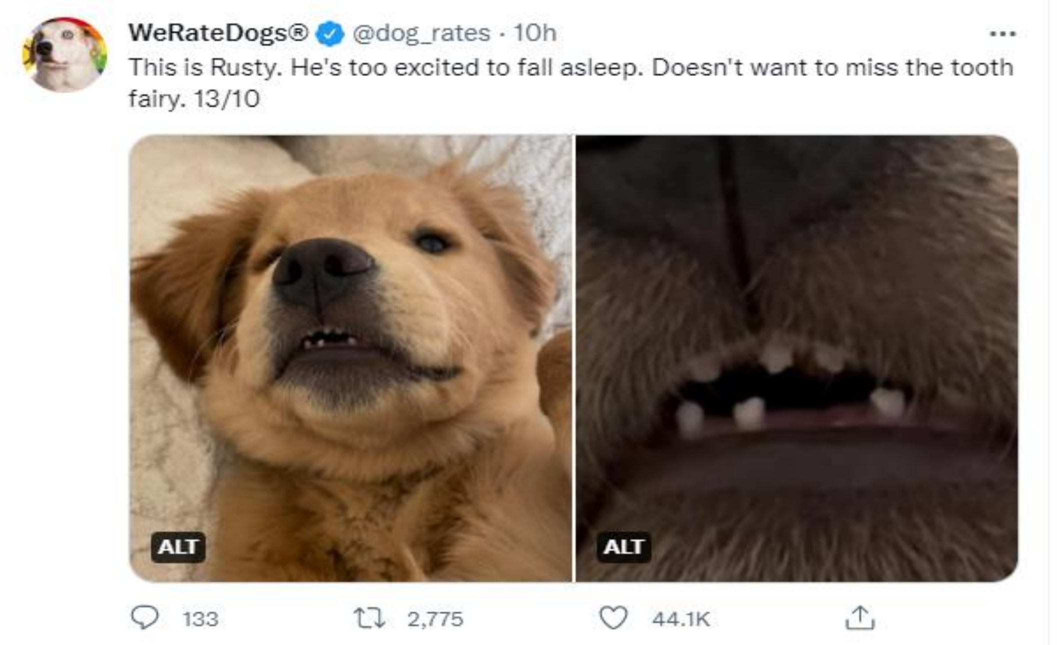
WeRateDogs is a Twitter account that posts funny comments and rates of dogs from photos submitted by their followers. Active since November 2015, and with over 7.6 million followers by January 2019, it has grown beyond being a twitter account to launch even a store with dog-related products. WeRateDogs has developed its own dog classification, based on the dog's age and appearance. Puppies are "puppers", older puppers are "puppos", older puppos are "doggos" and hairy dogs are "floofers."

For this data analysis project, data was gathered from a number of sources, including the twitter account using Twitter API. The aim was to have consolidated data about the tweets, the likes and retweets, and image predictions of the dogs posted. The purpose of the project is to focus more on data wrangling and less on EDA and/or visualization.

Data wrangling is simply the process of gathering, assessing and cleaning data. Most times, data has quality and tidiness issues. Quality issues refer to issues concerning accuracy, completeness, consistency, and validity. Tidiness issues mainly affect the structure of the data. Tidy data means each variable forms a column, each observation forms a row and each type of observational unit forms a table.

After gathering the data, I assessed the data programmatically and visually and noticed some issues which included:

- Some tweets were retweets or replies.
- Some columns were saved as the wrong data types.
- Some rating denominators were higher than 10.
- Data about dogs and tweets were in a same dataset
- Dog stage was spread in four columns and predictions were spread in three columns
- The retweets and likes data was split from the main dataset describing tweets



WeRateDogs® @dog_rates · 10h

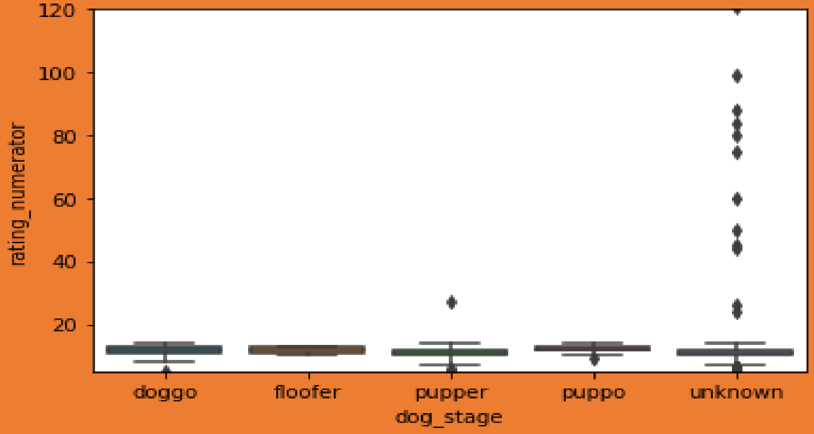
This is Rusty. He's too excited to fall asleep. Doesn't want to miss the tooth fairy. 13/10

Thankfully, the pandas library is well-equipped with data wrangling tools, so cleaning the data wasn't a difficult task. I ended up splitting the data into data about the dogs and data about the tweets. I also melted the dog stages into the same column.

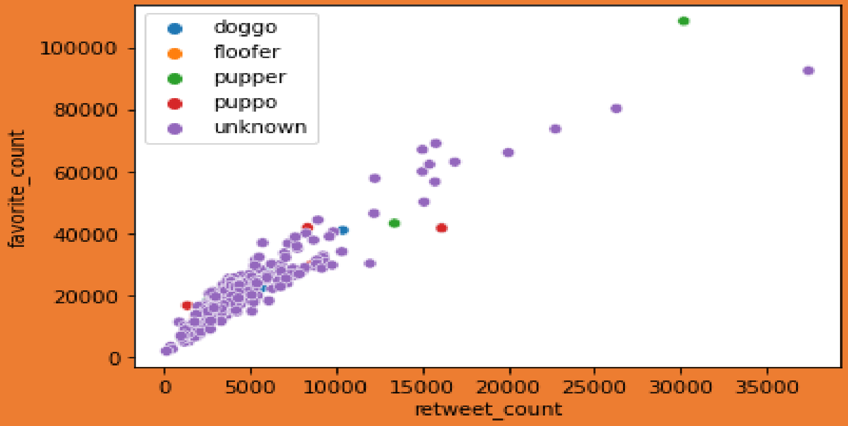
Keep in mind that this project's focus was more of wrangling than analyzing. However, I was able to derive some insights from the master data:

1. People love puppers.

Although a lot of the dog species are unknown, he data shows hat the puppers are the most highly rated of the dog predictions.

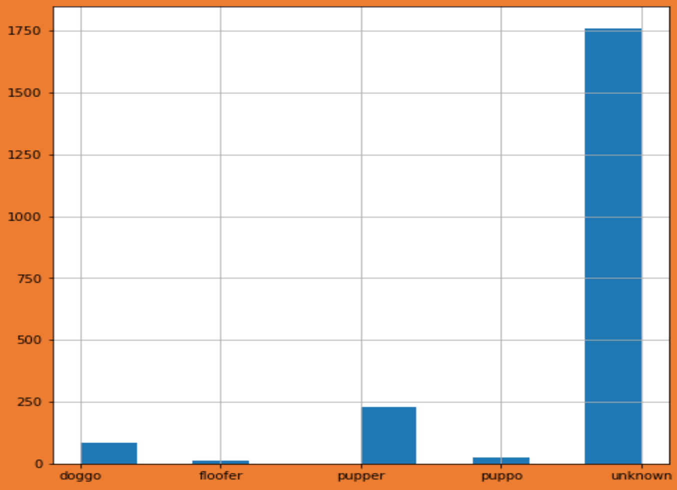


2. Retweet and likes are positively correlated



3. Puppers received more retweets of al the known species.

This is according to the distribution of the retweets.


doggo
floofer
pupper
puppo
unknown