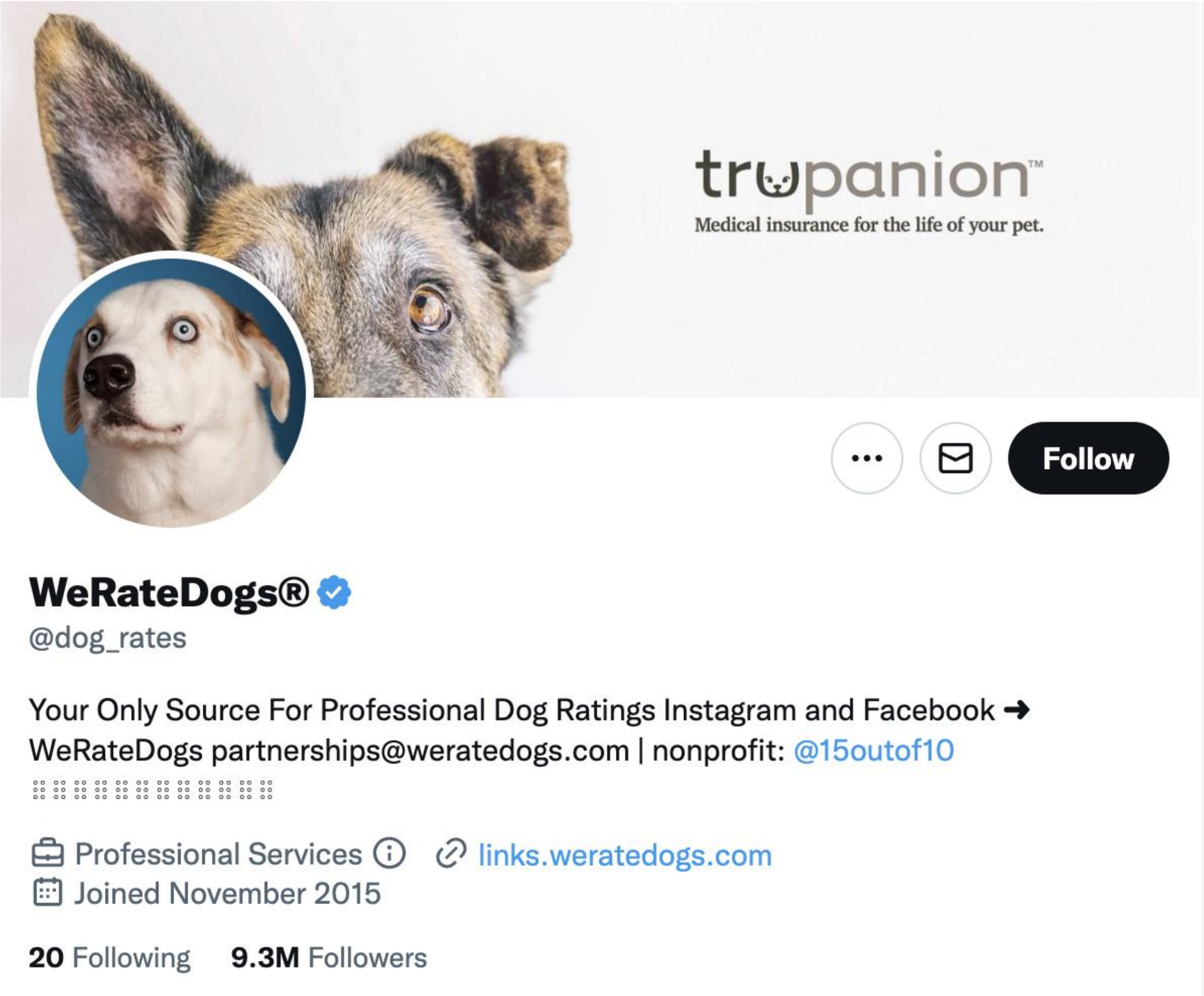

WeRateDogs®

@dog_rates

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

Professional Services ⓘ links.weratedogs.com
Joined November 2015

20 Following 9.3M Followers

This project featured the tweet archive from the @WeRateDogs Twitter account. I gathered data from two primary sources – the Udacity classroom and Twitter API. Data provided in the Udacity classroom (twiter_archive_enhanced.csv) contained the tweet id, ratings, & texts associated with each tweet, and neural network data (image_predictions.txt) specifying breed types. The image_predictions.txt was added to the notebook programmatically. Additional data, such as the retweet count, favorite count, and quote status of the tweet, were gathered using the Twitter API. The total data collected from the Twitter API contained 2325 unique tweet ids.

After gathering, I assessed the data frames visually and programmatically using Microsoft Excel & pandas and documented the cleanliness and tidiness issues. The data frame from the

Twitter API was clean and tidy. However, the tweet archive obtained from the Udacity classroom was messy and untidy. Among the cleanliness issues documented were the presence of retweets in addition to originals, missing dog names (a, an, None), denominator rating greater than 10 for ratings, missing values, null values, and inappropriate data types. Furthermore, a significant tidiness issue was found for the dog stages columns in the 'twiter_archive_enhanced.csv' file, as multiple columns were created for dog stages even though the maximum stages a dog can belong to simultaneously is just two.

Next, I performed cleaning operations on the identified issues. Each issue was clearly stated, defined, coded, and tested. I tried to solve nearly all quality issues, but the name column containing ambiguous dog names (a, an, None) was not fixed. Additionally, some tweet ids had to be dropped during the cleaning, giving a total row count of 1958. I merged the three data frames to form a complete observation unit, and the cleaned data was stored in a CSV file and re-named as 'twitter_archive_master.csv'. Lastly. I carried out some analysis and visualizations on the cleaned data.

Image: https://twitter.com/dog_rates