Wrangling Report

January 2023

CONFIDENTIAL

Internal Use Only

Written By
Singhanart Nakpongphun

01 Introduction

This wrangle report explained the wrangling, analyzing and visualizing process of tweet archive of Twitter user @dog_rates, also known as WeRateDogs, dataset. The WeRateDogs is a Twitter account that rates people's dogs with a hilrarious comment about the dog.

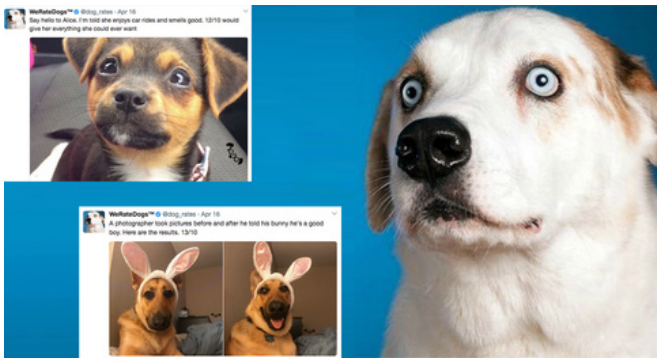
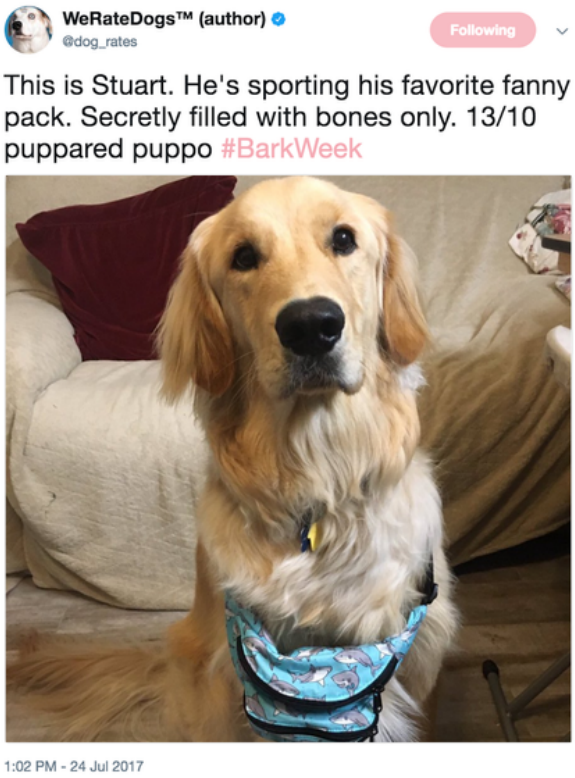


Image from Boston Magazine via Udacity

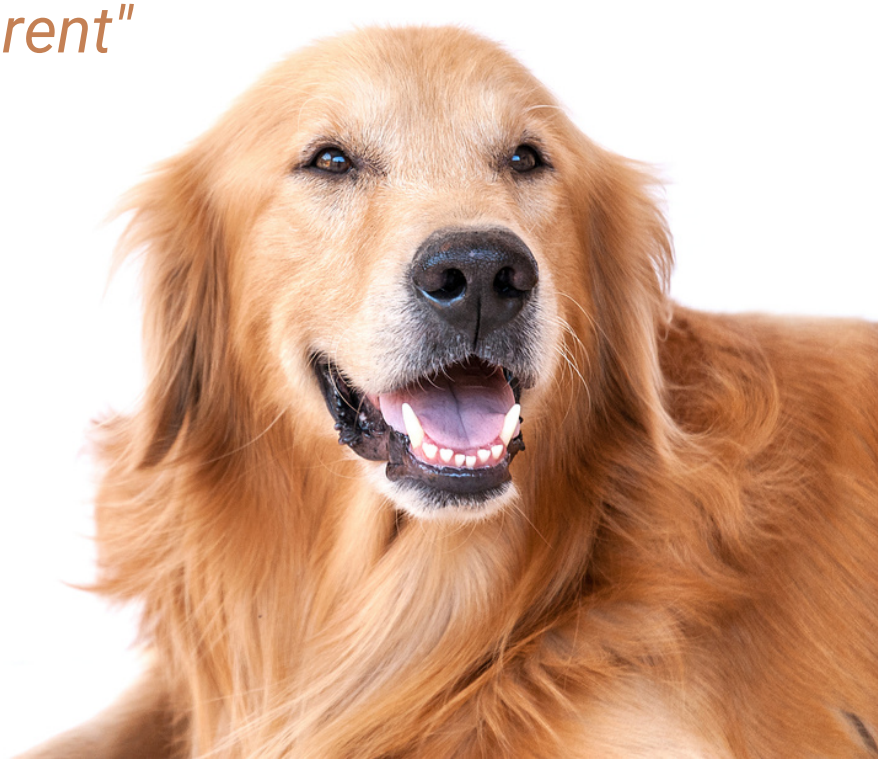


A Sample tweet from WeRateDogs
Image from Udacity

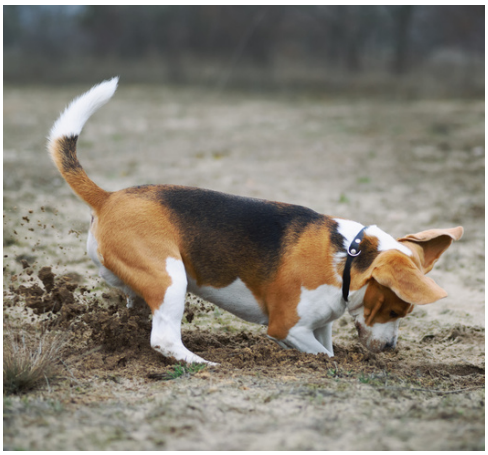
"They're good dogs Brent"

WeRateDog, 2016

There are three parts in this report which are gathering, assessing, and cleaning data. Each will be described respectively per the actual wrangling process.



Gathering Data 02



This process, I gathered data from 3 sources in the following:

1. Enhanced Twitter Archive: This file the tweets information. I downloaded the file from the Udacity platform and opened with Pandas library.
2. Image Prediction: This file contain the results of image prediction from neural network machine learning. I downloaded the file via the Request library.
3. Tweet's JSON data: This file contain the information about retweet and favorite data. I also downloaded the file via the Request library.

Assessing Data 03

After gathered all data. I both assessed all datasets visually and programmatically with Python. After the assessment completion, I found several issues regarding to the quality (content) and tidiness (structure) of all dataset which will be explained in the next page.

***For the denominator !=10 issue , It seems like it is s a part of content creation. For example, https://twitter.com/dog_rates/status/713900603437621249 or https://twitter.com/dog_rates/status/675853064436391936. Additionally, according to https://knowyourmeme.com/memes/theyre-good-dogs-brent, I found the context of the rating system is sound more creative than calculative purpose. Thus I will leave it as it is.

Assessing Data (cont.) 03

SUMMARY OF QUALITY AND TIDINESS ISSSUES

QUALITY ISSSUES

twitter_archive: quality issue

- Timestamp is an object type. It should be DateTime type
- Value in the source column is derived from HTML code which is hard to read
- Name column have an inconsistency of capitalization
- Name column contains some article word such as the, a, an
- There are missing value in columns (such as name, floofer, pupper) that use None instead of NaN
- This project only want original ratings (no retweets). The repiles and retweet should be removed
- There are unnessary columns that should be dropped.

image_prediction: quality issue

- There are 66 duplicated images
- The value format of p1,p2,p3 are inconsistency of capitalization.

tweet_json: quality issue

- The column name 'id' should change into 'tweet_id' to match with the other two datasets.

TIDINESS ISSSUES

twitter_archive: tidiness issue

- The dog stages (['doggo','floofer','pupper','puppo']) are now in seperated columns, It should combine in a single feature (column)

image_prediction: tidyness issue

- As it is the same observation from the twitter_archive, it should be combine in the same table.

tweet_json: tidyness issue

- It is also the same observation from the twitter_archive, it should be combine in the same table.

Cleaning Data 04

For the cleaning process, I started by make a copy of each DataFrame and apply the Define, Code, Test framework to tackle for each issue. The sequence of cleaning is respect to the order mentioned in assessing data session. After the cleaning was done, I exported a cleaned DataFrame to a single csv file (twitter_archive_master.csv) which ready to be used in an analysis process.

Thank you very much

Written by Singhanart Nakpongphun