Wrangle and Analyze « WeRateDogs » Data

Wrangle Report

Kévin Péricart

Introduction

The dataset that we will bre wrangling (and analyzind and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent". WeRateDogs as over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for us to us in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Our goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

Different Steps

1. Gathering Data : We gathered data from Twitter Archive, Image Prediction and Twitter API ;
2. Assessing Data : We assessed the three datasets individually ;
3. Cleaning Data : We cleaned data in order to make better analyzes and visualizations ;
4. Merging Data : We merged the three datasets into a single one ;
5. Analyzing and Visualizing Data : We analyzed and plotted the merged dataset.

1. Gathering Data

Twitter Archive : The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, used to extract rating, dog name, and dog 'stage' to make this Twitter archive 'enhanced' ;

Image Prediction : The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. The file is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL : https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv ;

Twitter API Data : Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

2. Assessing Data

We assessed the three loaded datasets individually.

Twitter Archive :

- We observed the reply and retweet ids, and counted the number of replies and retweets ;
- We counted the dog names ;
- We counted dogs categories ;
- We observed the rating numerator and denominator. We also observed outliers values.

Quality and Tidiness issues in Twitter Archive :

- A number of numerators and denominators do not comply with the rules laid down ;

- We can observe that dogo stage is splitted in 4 columns (doggo, floofer, pupper, puppo) ;
- Some denominators are greater than 10, which is not in accordance with the scoring rules ;
- We can observe 745 'None' entries and 55 'a' entries in the colume name. We can also observe different invalid names that contain less than three characters and start with lower letter ;
- Columns numerator and denominator expresses the same observation, i.e. the rating ;
- A number of numerators are overvalued and have outliers.

Image Prediction :

- We observed jpg_url for duplicates and jpg_url format ;
- We observed p1, p2 and p3 predictions.

Quality and Tidiness issues in Image Prediction :

- df_archive counts 2356 entries while df_image counts 2075 entries, we can observe a mismatch with missing entries ;
- We can find 66 jpg_url duplicates ;
- 1532 images have been classified true as dog images in p1 predictions, 1553 have been classified true as dog images in p2 predictions, 1499 have been classified true as dog images in p3 predictions ;
- We will need to merge the three levels of prediction into one column.

Twitter API Data :

Quality and Tidiness issues in Twitter API Data :

- df_archive counts 2356 entries while df_api counts 2354 entries, we can observe a mismatch with missing entries ;
- Apart from the mismatch of the number of entries, the three datasets seem to be linked and can be merged into one single ;
- A number of tweets are present several times within the three datasets. Since only original tweets are to be considered for the project, these duplicate tweets will have to be purged.

3. Cleaning Data

Thereafter, we cleaned and created copies of the three datasets

Twitter Archive :

- We replaced None values to « » and then to NaN in the columns doggo, pupper, puppo and floofer. We finally merged the three columns into one « dog_stage » and removed the precedent columns ;
- We dropped null values for retweets ;
- We replaced the differents error names in the name column ;
- We replaced name to None and then to NaN ;
- We converted timestamp to datetime format ;
- We created new column rating by calculating numerator/denominator and dropped colums/outliers.

Image Prediction :

- We dropped jpg_url duplicates.

Twitter API Data :

- We did nothing.

4. Merging Data

We merged the three datasets into one using the tweet_id column. We also stored the final dataset in a CSV file « twitter_archive_master.csv ».

5. Analyzing and Visualizing Data

- We observed the merged dataset ;
- We plotted histograms of the full merged dataset ;
- We plotted ratings, favorites and retweets opposed to « dog_stage » ;
- We computed a linear regression between retweets and favorites and the correlation between the two ;
- We counted the differents doggo categories and plotted the distribution ;
- We plotted the 10 most popular names in the merged dataset.