@WeRateDog Data Wrangling Project

BY Lydia Fang

February 18, 2019

Introduction

This Data Wrangling project is one of the Udacity Data Analyst nanodegree projects. In this project, we archived the basic tweet data (tweet ID, timestamp, text, etc.) from @WeRateDog for all 5000+ of their tweets as they stood on August 1, 2017. Also, through the image prediction URL and the tweepy API, we got the image links of these dogs and the prediction results of these images through a neural network, the favorite count and the retweet count data of all the posts were also included in this dataset. In this projects, I conducted several wrangling efforts to this dataset, then got a final clean dataset, and did some preliminary analyses and visualization on the dataset.

Data Gathering

This dataset includes three separate datasets from different sources.

- Data1 is the "twitter_archieve-enhanced.csv" data, which was already provided.
- Data 2 is the "image-predictions.csv" data that downloaded from the url https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv using the requests function of python.
- Data 3 is the retweet and favorite information from the twitter API. In order to acquire the data, I created a twitter account and got the developer access from twitter, then using the tweepy API for python and json to obtain the tweet_json.txt file. After that, I chose the useful information, including 'id', 'favorite count', and 'retweet count', and save these information into "retweet_fav.csv" file.

Data Assessing and Cleaning

After gathering the three dataset, I started assess each data and found out all quality issues and tidiness issues that need to be further cleaned. Here I describe the assessing and cleaning steps together.

- **Data1-Assessing and Cleaning**

Assessing: Data 1 “twitter-archive-enhanced.csv” has lots of information, and had the most quality and tidiness issues, such as including the retweet and reply data, missing url link, missing names, wrong rating denominator, extreme rating numerator, wrong timestamp type, inappropriate dog stage column structure, and missing dog stages information etc.

Cleaning: To clean this dataset. (1) I removed the retweet and reply data points that should not be included and removed the url link column that is not necessary for my analyses. (2) Using lambda function, I re-categorize the source column to make the information more clear and easier to read. (3) Using lambda function, I re-categorize the four dog stage columns (re; doggo, floofer, puppo, pupper) into one column. However, because a large number of data point does not have the stage information, so the dog stage sample size is significantly reduced in the further analyses. (4) Although it is not important, but I still changed all meaningless names (e.g. an, the, Bo, a) to “None”. (5) Make sure all rating denominators are 10. (6) Although the rating numerator is allowed to be larger than 10, but after assessing, I found some of the rating numerator was not correct based on the text comments, so I changed those wrong rating numerators into correct ones according to the associated texts. (7) Finally, I changed the timestamp into datetime type using pd.to_datetime function.

- **Data 2 Assessing and Cleaning**

  Assessing: The “image_predictions.csv” file is much more easier. One of the issues is that according to the three prediction and images, some of the images are actually not dogs. Also, the image number information is not necessary for my further analyses.

  Cleaning: First, I removed all images that are not dogs (i.e. the p1_dog, p2_dog, and p3_dog are all “False”). Then, I removed the ‘img_num’ column.

- **Data 3 Assessing and Cleaning**

  Change the column “id” into “tweet_id” to make it consistent with the other two dataframe.

Save and store the final clean data

After finishing the above steps, I merged three dataset using the pd.merge function, inner join on the “tweet_id” and got the final clean dataframe. I saved it as ‘ twitter_archive_master.csv’ file. The final dataframe size is 1084*20.

Data Visualization

Finally, I did several preliminary analyses and visualization using the “twitter_archive_master.csv” file. The detailed report is presented in the “act_report.pdf” file.