wrangle_report

January 5, 2021

1 1. Wrangle and Analyze Data

1.1 1.1. Introduction

This project focused on wrangling data from the WeRateDogs Twitter account using Python, documented in a Jupyter Notebook (wrangle_act.ipynb). This Twitter account rates dogs with humorous commentary. The rating denominator is usually 10, however, the numerators are usually greater than 10. They're Good Dogs Brent wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. WeRateDogs has 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 use 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.

The goal of this project is to wrangle the WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The challenge lies in the fact that the Twitter archive is great, but it only contains very basic tweet information that comes in JSON format. I needed to gather, asses and clean the Twitter data for a worthy analysis and visualization.

1.2 1.2. Enhanced 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, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced.".We manually downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv

1.3 1.3. Image Predictions File

image_predictions.tsv contains the breed of dog (or other object, animal, etc.) which is presented in each tweet according to a neural network. File image_predictions.tsv hosted on Udacity's servers and it is downloaded programmatically using python Requests library on the following: URL of the file: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

1.4 1.4. Twitter API

Back to the basic-ness of Twitter archives retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's

API. Well, "anyone" who has access to data for the 3000 most recent tweets, at least. Given having the WeRateDogs Twitter archive and specifically the tweet IDs within it, it is possible to gather this data for all 5000+ by the query Twitter's API. Herea are few points to keep in mind during data wrangling for this project:

1) We only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings and some are retweets.

2) Fully assessing and cleaning the entire dataset requires exceptional effort so only a subset of its issues (eight (8) quality issues and two (2) tidiness issues at minimum) need to be assessed and cleaned.

3) Cleaning includes merging individual pieces of data according to the rules of tidy data.

4) The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.

2 2. Assessing Data

2.1 2.1. Quality Issues

df:

1) Missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, Retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls

2) This dataset includes retweets, which means there is duplicated data

3) Timestamp and retweeted_status_timestamp is an object and not correct datetime frame.

4) The source column still has the HTML tags

5) Dogs name have 'None', or 'a', or 'an.' and some more lower case words as names

6) Multiple dog stages occurs such as 'doggo puppo', 'doggo pupper', 'doggo floofer'

image_df:

1) Dog breeds are not consistently in p1,p2,p3 columns: this column contains dog breeds name starts with upper case or lower case.

tweet_json_df:

1) Date type of column tweet_id is an object. It should be an int in order to merge it to the master df.

2.2 2.2. Tidiness Issues

df:

1) The variable for the dog's stage (dogoo, floofer, pupper, puppo) is spread in different columns which should be in one column.

image_df:

2) This data set is part of the same observational unit as the data in the archive_df

df_tweet_json:

3) This data set is also part of the same observational unit as the data in the archive_df

3 3. Cleaning Data

3.1 Define

1) Convert the tweet_id in tweet_json_clean dataframe into int type for merging into master dataframe
2) Creates a predicted dog breed column, based on the the confidence level of minimum 20% and 'p1_dog', 'p2_dog' and 'p3_dog' statements
3) Create one column for the various dog types: doggo, floofer, pupper, puppo, 'doggo, puppo', 'doggo, pupper', 'doggo, floofer' ascolumn name ' type ' with the categorical dtype
4) Merge the copied df_clean, image_df_clean, and tweet_json_clean dataframes
5) Convert the tweet_id in master_df into object type as there is no use for maths operation in tweet_id
6) Replace 'a', 'an', 'the', 'None' and other lower case words with NaN in name column
7) Remove Inconsistency in pred_breed
8) Delete retweets
9) Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
10) Change the timestamp to correct datetime format
11) Removing HTML tags from source column
12) Dog ratings get standardized for denom of 10.