‘WeRateDogs’ Data Analysis

Wrangle report

By Imesha Kuruppu

4/28/2021

Introduction

‘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. But, the numerator is always greater than 10. 11/10, 12/10, 13/10, etc. This rating system is unique to the 'WeRateDogs'. This unique rating system is a big part of the popularity of WeRateDogs.

The goal of this project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Project details

In this project, I have gathered, assessed, and cleaned data related to the 'WeRateDogs' Twitter account. Data have been gathered from three sources(CSV file, tsv file downloaded from a given URL, and Twitter API). Then, assessed the data visually and programmatically to find quality and tidiness issues. Discovered issues were cleaned in the data cleaning step and finally, obtained a high quality and tidy master pandas DataFrame. This data frame was saved as the 'twitter_archive_master.csv' and used for analyzing, and visualizing data.

Data wrangling process

Gathering Data:

1) 'twitter-archive-enhanced.csv' file was given by Udacity and it was downloaded programmatically and saved into the 'twitter_archive' data frame.

2) The tweet image predictions(what breed of dog is present in each tweet according to a neural network.) file, 'image-predictions.tsv' was hosted on Udacity's servers. This file was downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. Then, the data was stored in the 'image_prediction' data frame.

3) Additional data was gathered using Twitter API using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt. Data in this file was stored in the df_tweet data frame.

Assessing Data:

Data frames 'twitter_archive', 'image_prediction', and df_tweet were accessed visually and programmatically. For visual assessment data files were downloaded to excel files. jupyter notebook was used to access visually and programmatically. While assessing I checked for quality issues (completeness, validity, accuracy, and consistency issues) and tidiness issues (each variable should form a column, each observation should form a row and each type of observational unit should form a table).

1) Twitter_archive

Initially, twitter_archive included 2356 rows and 17 columns. Initial visual assessment I recognized that the dog stage variable has split into four columns in the dataset. But, this needs to include in one column. When I checked the information of the dataset found that a lot of columns (ex: in_reply_to_status_id, retweeted_status_id) have missing values. Some columns (ex: timestamp, dog stage columns) have wrong data types. This data set didn't contain data beyond 2017-08-01. No duplicated rows in the dataset. Using the describe function on the dataset found that there are outliers in the rating numerator and denominator columns. According to the rating system, all values in the rating_denominator should equal 10. Box plot was plotted to find the outliers in the rating_numerator. All dog names didn't have meaningful names. 56 rows had a single character as a dog name.

2) Image_prediction

Image_prediction data frame has 2075 rows and 12 columns. From all 12 columns, 3 columns were used to store 3 predictions and another 3 columns were used to store the prediction confidence and another 3 columns were used to indicate whether the prediction is true/false. This prediction variable should only form one column. Likewise, prediction confidence and its' correctness should form separate columns instead of three columns. There are 66 duplicated image URLs in this data frame. It seems img_num column values are not necessary for data analysis.

3) df_tweet

We have extracted only important columns from the Twitter API. That is the favorite_count column and retweet_count column. This data frame has 2331 rows and 3 columns. There are no duplicated rows or missing values.

Finally, all three data frames need to combine as they represent all details of tweet ids. ex: tweet id details(timestamp, source, text, dog stage, etc), dogs' breed prediction details, and favorite and retweet counts.

Cleaning Data:

First, copies of all three data frames were taken before starting the cleaning process. Then removing quality and tidiness issues were started according to the following order,

1) Remove all columns that are not needed for data analysis in the Twitter archive and image prediction.

2) Concatenate doggo, floofer, pupper, and puppo columns to one column called dog_stage in the Twitter archive. Since we have concatenated four columns in the new data frame, four rows were created with the same tweet id. Thus, values were sorted, and unnecessary rows were deleted except the last row.

3) Replace rows of the Twitter archive that have a single character as the dog name by the value 'None'.

4) Correct the data types in the Twitter archive. Change the data type of timestamp and dog stage to correct types.

5) According to the rating system all records should have rating_denominator equals 10. But in the data set, there are different numbers for denominators other than 10. So changed all other numbers in the denominator to 10.

6) There were a lot of outliers in the rating_numerator column. So I decided to remove all rows that contain a value that greater than 100 as the rating numerator.

7) Remove 66 duplicated URLs from the image prediction.

8) Merged image prediction data frame to Twitter archive using the 'tweet_id' column. Before the merge, the Twitter archive contained 2345 rows and image prediction contained 2009 rows. Merged data frame(df_twitter_data1) had 2345 rows.

9) Remove rows that don't contain 'jpg_url' in the df_twitter_data1. For that, extract all the rows that have URLs to the df_twitter_data1 data frame. Now, the df_twitter_data1 data frame has only 2002 rows.

10) Concetenate dog breed prediction columns. The dog breed predictions(p1, p2,p3) should give in one column (p) , The prediction confidences(p1_conf, p2_conf, P3_conf) should give in

one column(p_conf) and p1_dog, p2_dog, p3_dog should give in one column(p_dog), and also prediction attempt need to record in a new column. Temporary three data frames(df_twitter_data_test1, df_twitter_data_test2, df_twitter_data_test3) were created and finally all merged using prediction_attempt column. Since there are 3 prediction attempts final dataframe(df_twitter_data1) now has 6006 rows(2002*3)

11) Merge df_tweet_clean data frame to df_twitter_data1 data frame using the tweet id column and used the 'inner' joining method. Now df_twitter_data data frame has 5961 rows and 13 columns.

12) Capitalize the first letter of values in the columns name, dog_stage, prediction

13) Create the new column dog_rating by dividing the rating_numerator by rating_denominator and delete those two columns.

The final data frame df_twitter_data has 5961 rows and 13 columns after solving quality and tidiness issues in the dataset. Finally, the data frame df_twitter_data was saved to a file called 'twitter_archive_master.csv'. Now, this file can be used for data analysis.