wrangle report
17-11

Investigation of Twitter archive of **WeRateDogs**

0.1 Table of Contents

Introduction

0.1.1 About the Dataset

The dataset that is being wrangled (and analyzed and visualized) 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 has over 4 million followers and has received international media coverage.
This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

0.1.2 Inspiration

Is it possible to find the best rated dogs based on the tweets

Questions - - Ratings of dogs based on type - Best dog according to rating - Tweets based on hour of the day - Tweets based on the days of the week - Tweets based on month - Most important factor which leads to better rating

Data Wrangling
Gathering Data Data has been gathered in 3 different formats from 3 different sources. Data files included are twitter-archive-enhanced.csv, image-predictions.tsv

from url :https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv) and retweet and favorite counts from Twitter's API as tweet-json.txt

0.1.3 Accessing Data

General Properties

- The given dataframe has 2356 rows and 17 different columns
- Columns are 'tweet_id', 'in_reply_to_status_id', 'in_reply_to_user_id', 'timestamp', 'source', 'text', 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls', 'rating_numerator', 'rating_denominator', 'name', 'doggo', 'floofer', 'pupper', 'puppo'
- in_reply_to_status_id and in_reply_to_user_id have only 78 not-null values
- retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp have only 181 not-null values
- Data types are

Column Name	Count	Data Type
tweet_id	2356 non-null	int64
in_reply_to_status_id	78 non-null	float64
in_reply_to_user_id	78 non-null	float64
timestamp	2356 non-null	object
source	2356 non-null	object
text	2356 non-null	object
retweeted_status_id	181 non-null	float64
retweeted_status_user_id	181 non-null	float64
retweeted_status_timestamp	181 non-null	object
expanded_urls	2297 non-null	object
rating_numerator	2356 non-null	int64
rating_denominator	2356 non-null	int64
name	2356 non-null	object
doggo	2356 non-null	object
floofer	2356 non-null	object
pupper	2356 non-null	object
puppo	2356 non-null	object

-	rating_numerator	rating_denominator
count	2356.000000	2356.000000
mean	13.126486	10.455433
std	45.876648	6.745237
min	0.000000	0.000000

-	rating_numerator	rating_denominator
25%	10.000000	10.000000
50%	11.000000	10.000000
75%	12.000000	10.000000
max	1776.000000	170.000000

Data Quality Issues

1-Many null values

in_reply_to_status_id

in_reply_to_user_id

retweeted_status_id

retweeted_status_user_id

2-Incorrect data types

tweet_id

in_reply_to_status_id

in_reply_to_user_id

retweeted_status_id

retweeted_status_user_id

3-datetime format for

timestamp

retweeted_status_timestamp

4-rating_denominator has minimum value as 0 which is not possible for denominators

5-Retweets need to be removed to avoid duplication in our analysis. This may be done by removing rows that have non-empty retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp

6-Add rating column as the ratio of numerator and denominator

7-unnecessary html tags in source column in place of utility name e.g. <a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone

8-Some numerators are wrongly entered. They are different as in the comments

Data Tidiness Issues

1-category column can be created to store the type of dog instead of the last 4 columns named as doggo, floofer, pupper, puppo

2-Information about one type of observational unit (tweets) is spread across three different dataframes. Therefore, these three dataframes should be merged as they are part of the same observational unit.

3-Reorder the columns into similar ones close to each other after adding or removing some extra columns.

Data Cleaning

Make a copy for the Data

Define
-Retweets need to be removed to avoid duplication in our analysis. This may be done by removing rows that have non-empty retweeted_status_id

twitter_df_clean: many tweet_id(s) of twitter_df_clean table are missing in image_df (image predictions) table

Define
Keep only those records in twitter_df_clean table whose tweet_id exists in image_df table

Define
2-Convert Incorrect data types

tweet_id

in_reply_to_status_id

in_reply_to_user_id

retweeted_status_id

retweeted_status_user_id

Define
3- Convert datetime format for

timestamp

retweeted_status_timestamp

twitter_df_clean: text column contains untruncated text instead of displayable text

Define
Using the display_text_range of tweet_df_clean table, extract displayable text from text column

Define
4-Some numerators are wrongly entered.

Exteract numerators from txt column

twitter_df_clean: rating_denominator column has values other than 10

Define
For records whose rating_denominator is greater than 10 and divisible by 10, use the quotient as the divisor to divide the rating_numerator. If the numerator turns out to be divisible (i.e. remainder=0), assign this quotient as the rating_numerator.

For the remaining records, check if the text column contains any fraction whose denominator is 10. If it does, update the rating_denominator to 10. Additionally, update the rating_numerator with the numerator value of this fraction.

Define
5- Removing the rating_denominator has minimum value as 0 which is not possible for denominators

Add rating column as the ratio of numerator and denominator
Check for Duplicated values
7-twitter_df_clean: unnecessary html tags in source column in place of utility name e.g. <a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone

Define
Strip all html anchor tags (i.e. <a..>) in source column and retain just the text in between the tags. Convert the datatype from string to categorical.

Data Tidiness Issues

Define
Drop retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp columns from twitter_df_clean table

Define
Drop rating_numerator, rating_denominator columns from twitter_df_clean table

*twitter_df_clean: erroneous dog names starting with lowercase characters (e.g. a, an, actually, by)**

Define
Replace all lowercase values of name column with None

Merge dog stages properly.
Set the None values to np.nan in all the 4 dog stage columns. Concatenate all 4 columns to 1 column dog_stage Now the multiple dog stages rows will have values combined. So replace them with code Remove the original 4 columns of dog stages.

Define
Concatenate all 4 columns to 1 column dog_stage

Add Month, Day and Hour for Tweet time
Define
Drop the columns doggo , floofer , pupper and puppo

Define Merge the dataframe image_df_clean with twitter_df_clean

The End
Home