Wrangle Report

Introduction

The dataset 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.

The tasks in this project are as follows:

- Data Wrangling
  - Gather Data
  - Assessing Data
    - Quality
    - Tidiness
  - Cleaning Data
    - Define
    - Code
    - Test
- Storing, Analyzing and Visualizing the wrangled data
- Reporting on the data wrangling efforts and data analyses as well as visualizations.

Gathering Data

A. Enhanced Twitter Archive

   i. Download the file manually: twitter_archive_enhanced.csv

B. Image_predictions

   i. This file (image_predictions.tsv) 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

C. Tweet_json

   i. This file (image_predictions.tsv) 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

Assessing Data

Quality

df_twitter_archive

1. A few columns with NaN values (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp & expanded_urls)
2. Name with (a, such, the, this, unacceptable and very), which is invalid for a dog name.
3. tweet_id data type should be str.
4. Rating denominator should be 10 only.
5. Rating numerator has maximum value of 1776, which is invalid for this case. The rating numerator should be kept from 0-14 only.
6. Rating denominator and rating numerator to be changed to float type for later calculation.

df_image_predictions

7. tweet_id data_type should be str
8. To change the values in columns p1, p2, p3 to lowercase.

df_tweet

9. tweet_id data_type should be str

Tidiness

df_twitter_archive

1. Combine table doggo, floofer, pupper and puppo to one column only.
2. Timestamp to be seperated to year, month and day.

df_image_predictions

3. To drop the columns p1,p2,p3 and respective conf. columns. Thereafter, create new columns dog_type and confidence level.

Cleaning Data

1. df_twitter_archive
    a. - Keep the original tweets only. Based on the info, we found that there are 181 retweets. I'll delete these 181 records and keep the original tweets only.
    b. Delete the columns ('in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id','retweeted_status_timestamp','retweeted_status_user_id', 'expanded_urls', 'source') that will not be used for analysis later.
    c. Replace invalid dog name with none (a, such, the, this, unacceptable and very).
    d. Change tweet_id data type to str.
    e. Drop the rows where the rating denominator is not 10.
    f. Rating numerator has maximum value of 1776, which is invalid for this case. As mentioned in the introduction earlier, (The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent."), I will keep the range to (0-14). Thus, the rating numerator not within the range will be dropped.
    g. Change the data type to float for rating_numerator and rating_denominator
    h. Change the 'timestamp' data type to datetime and split it by 'year', 'month', 'day'
    i. Combine table doggo, floofer, pupper and puppo to one column 'dog_stage' only.

2. df_image_predictions
    a. Change the tweet_id data type to str.
    b. To change the values in columns p1, p2, p3 to lowercase.
    c. To drop the columns p1,p2,p3 and respective conf. columns. Thereafter, create new columns dog_type and confidence level.

3. df_tweet
    a. To change the tweet_id data type to str