Wrangle Report

Karina Condeixa

Description

The datasets used in this project are from the Twitter account WeRateDogs. My tasks are to gather the data, analyze the data, visualize the data, document my efforts in this wrangling report and prepare a post about my project.

Getting Started

Tools and requirements: Python 3.6 and Libraries:

- pandas
- numpy
- tweepy
- requests
- json
- os
- re
- matplotlib.pyplot
- datetime

Gather

This study hat three ways to gathering data: the given *twitter_archive_enhanced.csv*, Image Predictions (URL: https://twitter.com/dog_rates/status/889531135344209921) as *image_predictions.tsv,* and Twitter (API) by a setup access to the twitter API using tweepy.

Findings

Particularities: Is a problem to have numerator lower or higher than denominator?
NO. This is part of the unique rating system. The rates above 10 means that the dog are super doggo, *pupper*,*puppo* or/and *floofer.*

Quality

df_twitter_archive (2356, 17)

1. Wrong type in timestamp, tweet_id, name.
2. Retweets should be removed (rows). We only want original ratings.
3. These irrelevant columns should be removed:

- in_reply_to_user_id
- in_reply_to_status_id
- retweeted_status_id

- retweeted_status_user_id
- retweeted_status_timestamp

4. Missing expanded_urls: the total is 2356 rows
5. There weird names, in lower case, that are probably wrong and 745 rows missing names (none).

df_image_predictions (2075, 12)

6. Lower case in p1, p2 and p3 values.

df_additional_tweet_data (2333, 3)

7. retweet_count does not matter to us
8. Tweet_id should be string

Tidy

df_twitter_archive

9. Join the Series doggo,pupper, puppo or floofer, in just one that are slangs which represents a the cuteness of a dog.

df_image_predictions

10. p1, p2 and p3, p1_conf, p2_conf and p3_conf, and p1_dog, p2_dog and p3_dog don't are good names for head.
    They should describe the variable (such as prediction #1, confidence #1, breed #1).

General

11. the dataframes should be merged

Insights and data visualization

- Insite 1: Number of Retweets for each cuteness assortment
- Insite 2: The most common dog breeds
- Data Visualization:
  I plot the top 10 dog breeds in a bar chart, considering the attributes: *'retweet_count', 'favorite_count', 'rating_denominator', 'rating_numerator'.*
- Insite 3: The most popular breeds
- Data Visualization:
  I plot the top 10 dog breeds considering *retweets*, and the top 10 dog breeds considering *favorites*.

Description of my efforts

For now, this was the most challenging project for me. I had to deal with many new issues for me at the same project.

References:

https://github.com/nanakoohashi/Wrangle-analyze-twitter-posts

https://github.com/DanaCody/Wrangling-Doggo-Data/blob/master/wrangle_report.pdf

https://github.com/MrGeislinger/UdacityDAND_Proj_WrangleAndAnalyzeData/blob/master/act_report.pdf

http://lindsaymoir.com/wp-content/uploads/2018/06/wrangle_act-1.html

https://www.geeksforgeeks.org/python-pandas-series-str-count/

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop.html