Wrangle report for data wrangling of WeRateDogs Twitter data

July 6, 2019

1 Introduction

This goal of this project is to perform data wrangling on the WeRateDogs Twitter data to create interesting and truthworthy analysis and visualizations.

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. and numerator almost always greater than 10: 11/10, 12/10, 13/10.

The data wrangling consisted of the following steps:

2 Data gathering

First, the following data have been gathered from three different sources:

- The WeRateDogs Twitter archive was read from *twitter_archive_enhanced.csv* file and stored in **tweet_data**. It contains basic tweet data for 2356 tweets, which have the rating information.
- The tweet image predictions were downloaded programmatically from the Internet using Requests library and URL and stored in **predictions**.
- Retweet count and favorite count data,which were missing in the original Twitter achive, were gathered from Twitter API data using Python's tweepy library and stored in **tweets_api**.

3 Data assessment

Then after data have been gathered, I did data assessment to find possible data quality and tidiness issues.

Data quality issues:

For tweet_data:

- Some tweets are retweets
- Some tweets are in-replies.
- For some tweets rating denominator was not extracted correctly from 'text' column.
- For some tweets rating numerator was not extracted correctly from 'text' column.
- Columns shoud be renamed: 'timestamp': 'tweet_timestamp', 'text': 'tweet_text','name': 'dog_name', 'source': 'tweet_source'

- Timestamp is not datetime format
- For source column url should be extracted from html tags.

For predictions:

- columns should be renamed: p1 - prediction1, p1_conf- prediction1_CI, p1_dog - prediction1_dog, p2 - prediction2, p2_conf- prediction2_CI, p2_dog - prediction2_dog, p3 - prediction3, p3_conf- prediction3_CI, p3_dog - prediction3_dog.

Tidiness issues

- For dog stage we have 4 columns: Doggo, floofer, puppo, pupper. We can create only one column 'dog_stage', where doggo, puppo, pupper are the values.
- Column 'rating numerator' and 'rating denominator' can be merge in one column called 'rating'. Column 'rating_denominator' will be removed because all values are 10.
- Merging all the datasets into 1 table as 1 observational unit (tweets info) should be in 1 table only as rule 3 (Each type of observational unit forms a table.).

4 Data cleaning

Before data cleaning, *tweet_data*, *predictions* and *tweets_api* tables were copied to *tweet_data_clean*, *predictions_clean* and *tweets_api_clean* tables.

Data cleaning included correcting all quality and tideness issues found during data assessment.

After all quality and tideness issues were solved , all three tables were merged on tweet_id in a single dataframe **twitter_archive_master** , which was stored in ***twitter_archive_master.csv*** file.

This final merged dataset was used for data analysis and visualization.

5 Conclusion

After the data were analysed, the following conclusions have been drawn:

- Golden retvier is the most common dog's breed, following by Labrador retriever, pembroke and chihuahua.
- Soft-coated wheaten terrier has the highest average rating, followed by West-Highland white terrier.
- Saluki again has the highest favorite count, followed by French bulldog.
- Standard poodle has the highest retweet count.
- Favorite and retweet count have a strong positive relationship.
- Overall, the favorite count is higher that retweet count for all dog stages. It can be also observed that doggo and puppo have the highest favorite and retweet counts, while pupper have the lowest both favorite and retweet counts.