Wrangle Report - Project: WeRateDogs

Introduction

This report record the all steps of the project - WeRateDogs in the data wrangling process. The goal is to classify all types of relative data and then set up the datasets, and then assess the dataset to clean it so that it can reach to a standardized format for the further process.

The data wrangling process compasses three main stages:

1. Gathering data;
2. Assessing data;
3. Cleaning data.

Gathering

The data of this project was collected from three resources showed below:

1. The twitter record of WeRateDogs, which is called: **twitter_archive_enhanced.csv**
   This .csv file is already gathered, we can use it directly.
2. The image prediction data record, which is called: **image_predictions.tsv**
   This is according to neural network, predicting dog's breed in each twitter, and this file needs to use Requests library in Python and specific URL to download by programming.

By ZhenwenXu

3. The extra attached data of each twitter including almost 'retweet_count' and 'favorite_count', which is called: **tweet_json.txt**

   This .txt file needs to use Tweepy library in Python to search for each twitter's JSON data in API, and then store all JSON data to the **tweet_json.txt.**

These datasets will be read to separate Pandas Dataframes - df_archive, df_prediction, and df_tweet.

Wrangle Report - Project: WeRateDogs

Assessing

After gathering three datasets above, assessing all datasets about Quality and Tidiness through visual and programmatic methods. And then listing at least 8 quality issues and 2 tidiness issues.

Completeness

df_archive table

1. Many rows have no value for the dog stage(all columns about dog stage are None in the row), and some rows have more than one entry for dog stages

Quality

df_archive table

visual assessment:

2. It has link at the end of each item in the text column
3. Sometimes items in the text column start with 'RT @dogs_rates:'
4. Information about retweets are not needed, like retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp
5. Some values in *rating_numerator* and *rating_denominator* are too large, like 90 or even over 100

Wrangle Report - Project: WeRateDogs

programming assessment:

6. Columns in_reply_to_status_id and in_reply_to_user_id are float data type
7. Columns timestamp and retweeted_status_timestamp are string not datetime data type
8. Missing name of the dog in name column (I cannot solve this one)

df_prediction table

visual assessment:

9. The names in columns p1, p2 and p3 have unconsistent spelling (like 'Lakeland_terrier' and 'sea_lion')
10. The names in columns p1, p2 and p3 have '_' and '-' between each word
11. The values of columns p1_conf, p2_conf and p3_conf need to be rounded

programming assessment:

12. 66 jpg image URLs are duplicated
13. The data only has 2075 tweet ids compared to 2356 tweet ids in df_archive

df_tweet table

programming assessment:

14. The data only has 2341 tweet ids compared to 2356 tweet ids in df_archive

By ZhenwenXu

Wrangle Report - Project: WeRateDogs

Tidiness

df_archive table

1. Multiple columns for dog stages (*doggo*, *floofer*, *pupper*, *puppo*) are not needed

df_prediction table

2. Summarize three predictions into one column

df_prediction table

3. df_tweet, df_prediction and df_archive should be combined to form one single dataframe

By ZhenwenXu

Wrangle Report - Project: WeRateDogs

Cleaning

This stage is about verifying and fixing the Quality and Tidiness issues in the previous stage, and before the cleaning, it is needed to copy three datasets into three new dataframes which are used to operate for the following process

Cleaning process contains three steps in each issue:

1. Define - Define each cleaning actions in details
2. Code - Write code to fix the issues programmatically
3. Test - Ensure whether the issues has been solved

After the cleaning, these three dataframes are combined into one dataset, and it is called ‘df_combined’

Conclusion

This wrangling is a very complex and time-consuming process, and this part spend almost two weeks (5hours/day) to complete all the issues I have listed.

The assessment stage includes all issues I found through visual and programmatic ways, but some of them cannot be solved in this project like missing name of dogs because the text doesn’t mention it.

By the way, some issues are dealt with together in the cleaning stage.

By ZhenwenXu