Wrangle report

December 28, 2018

1 Data Wrangling: We Rate Dogs

1.0.1 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. In this project, I used Python and its libraries to gather data from 3 sources and in a variety of formats, assess its quality and tidiness, then clean it. The data wrangling, analysis, and visulization processes were documented in Jupyter Notebook.

1.0.2 Data Gathering

Data for this project came from the following 3 sources: * the original twitter archive data was provided through email:twitter-archive-enhanced.csv * image prediction data for classifying breeds of dogs was programmatically downloaded from Udacity: image_predictions * addition twitter data was collected from the twitter API using Tweepy: tweet_api

1.0.3 Data Assessment

Quality Issues:

1. Retweets are included in the dataset
2. Text column includes both text and short version of URL to the tweet
3. Large missing values in columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, and expanded_urls
4. The rating_numerator and rating_denominator of some rows do not make sense, the minimum rating_numerator and rating_denominator are 0s, some rows the denominator is larger than 10, eg: index 1121, rating_numerator=204, rating denominator=170;the maximum numerator is 1776
5. Some names such as "a", "an", "the" are not names
6. Nulls are identified as "None" for name, doggo, floofer, pupper, and puppo columns
7. The format of column 'timestamp' is not clean
8. tweet_id is incorrectly labeled as ID in api table

Tidiness Issues

1. Doggo, floofer, pupper and puppo columns reppresent the types of dogs, so they could be combined into a single column named dog_type.

2. There are multiple columns containing the same type of data, e.g. p1, p2, p3 all contain dog breed predictions, they could be combined into one single colunm named breed_pred
3. The api table could be merged to the archive table
4. Part of predictions and api info is missing, the archive table contains 2356 observations, while the prediction table contains 2075 observations in total, the api table contains 2342 observations

1.0.4 Data Cleaning

Quality Issues:

1. remove rows are retweets and replies.
2. remove shortened URL from the end of the "text" column in the df_info_clean dataframe
3. remove unneccessary columns containing large number of misssing values, including columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp
4. extract the numerator and denominator of the score from the text column and save them in the correct column ("numerator", "denominator")
5. clean the missing value and wrong values for 'name' column
6. redifine None values as nulls for doggo, floofer, pupper, and puppo columns
7. change time stamp to datatime api_clean table and df_clean table
8. change the column name 'ID' to 'tweet_id' in api table

Tidiness Issues

1. change 'None' values as nulls for doggo, floofer, pupper, and puppo columns; create a new column 'dog_stage' from 'doggo', 'floofer', 'pupper', 'puppo' columns where value is not null;find the dog_stage from 'text' column, name the new column dog_type, and match the values of the dog_type column; merge dog_stage and dog_type column to a new column dog_stages, it fill the null values of dog_type with dog types found from the text column; drop 'doggo', 'floofer', 'pupper', 'puppo', 'dog_type', 'dog_stage' columns
2. remove rows in predictions_clean where all prediciton are False;creat 3 new cloumns: prediction, confidence, dog_breed, remove rows if p1_dog, p2_dog, p3_dog are all False, if p1_dog if True, append p1, p1_conf, p1_dog in new columns, else if p2_dog is True append p2, p2_conf, p2_dog in new columns, else if p3_dog is True append p3, p3_conf, p3_dog in new columns;merge new_pred with predictions_clean;drop columns p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, p3_dog
3. merge the cleaned archive, prediction, and api dataframes into one dataframe, and save it to csv file named 'twitter_archive_master.csv'

1.0.5 Result

- Final data frame have 2085 rows and 18 columns
- Data analysis and visulization are conducted to find the insight of the data, including the most popular type of the dog, how accurate the prediction of neural network is, the image of the dog with highest retweet counts and rating