Wrangle Report-2

February 5, 2019

0.1 Wrangle Report

The purpose of the wrangling process is to gather, asses and clean data in order to investigate the data and to draw conclusion from the data. In this case I would like to investigate the tweets from the WeRateDogs (Twitter account that rates people's dogs with a humorous comment).

In particular I would like investigate if the rating from WeRateDogs predicts the amount of retweets and faved tweets and if the users and the creators of the account have favorite breeds.

This questions should lead the wrangling process since we are not supossed to clean the whole data set

0.1.1 1.Gather

The first of the Data wrangling process is gathering of the needed the data. In this project I gathered data from 3 different sources which led to 3 different tables - The WeRateDogs Twitter archive which was provided by the WeRateDogs as CSV and which I importet with the pd.read_csv into the Jupyter Notebook (`twitter_archive`) - tweet image predictions (determind what breed is present in each tweet according to a neural network) which I downloaded programmatically with the request libary (`tweet_images`) - Additional Data via the Twitter API (retweet count and favorite count). Since I never granted access from Twitter to use the API, I was forced top access the project data without a Twitter account: I read the tweet_json.txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count as outcome (`tweet_measurements`)

0.1.2 2. Asses

In the first step I did a visual assesment of the three tables: `twitter_archive`, `tweet_images` & `tweet_measurements` and collected my findings.

In the next step I further investigated the tables programmatically using the following panda functions: - pd.info() - pd.duplicated - pd.describe() - pd.value_counts() - etc.

Through this I identified the following quality and tidiness issues (I just named the ones here that I acutally cleaned in the data cleaning phase of the wrangling project; see the complete list in the jupyter note book `wrangle_act`):

Quality

`twitter_archive` **table**

- Erroneous datatype (doggo, floofer, pupper, puppo, timestamp, tweet_id, in_reply_to_status_id, in_reply_to_user_id)
- wrong value in the rating_denominator column (0 in row 313)
- 181 tweets are retweets

`tweet_images` **table**

- wrong value in the p1 values: "shopping_cart", "convertible"
- undescriptive column names in this table
- different formats of the breed-values for the dogs
- Erroneous datatype (tweet_id (str instead), p1, p2, p3 (category instead)

`tweet_measurements` **table**

- The column name in the beginning of all values in all columns of this table
- wrong values in the favorite_count and retweet_count column. "is_quote_status": false"(67 entries total) and "contributors": null" (25 entries total), "lang": "en"}(41 entries total) instead of favorite count and retweet count in for ex. row 27, 1148, 1168. In total 67 entries.
- Erroneous datatype (retweeted, favorited)

Tidiness

- tweet_measurements result of image Predictions should be part of the twitter_archive table
- rating_numerator & rating_denominator should be one column in (each info in one column)
- dog stages floofer, pupper, puppo, doggo are spread in 4 columns. they should be merged into one.

0.1.3 3. Cleaning

In the first step I will created copies of the tables twitter_archive, tweet_images, tweet_measurements: - twitter_archive_clean - tweet_images_clean - tweet_measurements_clean

1. `tweet_measurements` **table: The column name in the beginning of all values in all columns of this table** cutting off the text in the begining of every row of the all columns in table tweet_measurements: tweet_id, favorite_count, retweet_count, retweeted and favorited

2. `tweet_measurements` **table: wrong values in the favorite_count and retweet_count column** identifying the wrong values and filling the missing values with the information from the txt file. 1. identifying the wrong values and storing them in a new data frame called wrong_values 2. dropping the rows with the wrong values from tweet_measurement 3. appending the fixed values (wrong_value table) to the df and dropping the rows I just need for the cleaning process

3. `twitter_archive` **table: 181 tweets are retweets (we just want to have original tweets)** Deleting/excluding the rows/tweets that are retweets from twitter_archive and then deleting the columns concerning the retweets from the table twitter_archive

4. `twitter_archive` **table: Erroneous datatype (doggo, floofer, pupper, puppo, timestamp, tweet_id, in_reply_to_status_id, in_reply_to_user_id)** Convert doggo, floofer, pupper, puppo will category data type. Convert timestamp to datetime and tweet_id, n_reply_to_status_id and in_reply_to_user_id to string data type.

5. `tweet_images` **table: Erroneous datatype (tweet_id (str instead), p1, p2, p3 (category instead)** Convert p1, p2, p3 to category data type. Convert tweet_id to string data type

6. `twitter_archive` **table: wrong value in the rating_denominator column (0 in row 313)** Changing the null value in the denominator column (row 13) to 13 (as it is correct)

7. `tweet_images` **table: - undescriptive column names in this table** Change the following columns: - img_num = number_images - p1 = prediction_1 - p2 = prediction_2 - p3 = prediction_3 - p1_conf = confidence_prediction_1 - p2_conf = confidence_prediction_2 - p3_conf = confidence_prediction_3 - p1_dog = breed_prediction_1 - p2_dog = breed_prediction_2 - p3_dog = breed_prediction_3

8. `tweet_images` **table: different formats of the breed-values for the dogs** changes all breed names to lower case and take away the _ in all breed names in prediction_1, prediction_2 and prediction_3

Tidiness

1. `tweet_images` **table: result of image Predictions should be in twitter_archive table** integrate the column 'prediction_1','breed_predicted_1, etc. in table "twitter_archive_clean by merging the column with the twitter_archive on Tweet_id

Merge the favorite_count and retweet_count column to the twitter_archive_clean table, joining on tweet_id

2. `twitter_archive`**: dog stages floofer, pupper, puppo, doggo are spread in 4 columns. they should be merged into one.** creating a new dataframe with the dog stages floofer, pupper, puppo, doggo in order to merge the 4 columns with a lambda function into one column named dog_stages. Then merging the dog_stages column with the twitter_archive column and dropping the columns floofer, pupper, puppo, doggo from the table

In the last step I stored the clean dataframe in a CSV file named twitter_archive_master.

```
In [ ]:
```