Liudmila Semenova
UDACITY Data Analyst Nanodegree Program
2019, Jule 10

Wrangle and Analyze Data Project
Wrange Report

Gathering Data

There were three data sources:

1. **twitter_archive_enhanced.csv** which is the WeRateDogs twitter account archive in csv-format and can be downloaded manually here.

2. **image_predictions.tsv** with predictions what breed of dog (or other object, animal, etc.) is present in each tweet. This file should be downloaded programmatically using the Requests library from the following url.

3. **tweet_json.txt** which should contain entire tweets from WeRateDogs Twitter archive since twitter_archive_enhanced.csv contains only basic information. tweet_json.txt should be created programmatically by querying the Twitter API for entire tweet's JSON data using the tweet IDs in the WeRateDogs Twitter archive and Python's Tweepy library, then storing all these JSON data in a txt-file.

Twitter Archive

I downloaded **twitter_archive_enhanced.csv** manually from the given url, then used *pd.read_csv* to create **archive** dataframe.

Image Predictions

I downloaded **image_predictions.tsv** programmatically from the given url using *requests*; the data were written to the same-name file. Then I used *pd.read_csv* to create **images** dataframe.

Tweets

I created access to Twitter, then downloaded tweets to **tweet_json.txt**, counting how many tweets were downloaded and how many tweets were not found. I also wrote ids of tweets which were not found to **errors.txt**. Then I used *pd.read_json* to create **images** dataframe.

Assessing Data

Quality

archive table

- there are 22 rows with tweets which no longer exist.
- *tweet_id* is integer instead of string.
- *in_reply_to_status_id*, *in_reply_to_user_id*, *retweeted_status_user_id*, *retweeted_status_timestamp* columns are redundant since they are not useful for further analysis.
- *timestamp* is string instead of datetime.
- *source* column contains values surrounded by html-tags.
- *retweeted_status_id* is not null for 181 which means that all these tweets were retweeted.
- *retweeted_status_id* will be redundant after using this column for deleting retweeted tweets.
- some *expanded_urls* values contain the same url more than once.
- some *expanded_urls* values contain several urls without spaces so these urls do not work properly.
- *rating_numerator* contains 24 values which are greater than 20 and 440 which are less than 10.
- *rating_denominator* contains 23 values which are not equal to 10.
- *rating_numerator* and *rating_denominator* could have been extracted with errors from 33 tweets .
- there are 109 values in the *name* column that do not seem like real names.
- some of rows where the *name* values do not seem like real names in fact contain a dog's name follows the word "named" in the *text* column.
- there are no any dog stage for 2326 tweets.
- *doggo*, *floofer*, *pupper*, *puppo* are string instead of categorical.

Issues which were discovered but left without cleaning:

- *text* values does not contain urls or contain broken urls or not twitter urls for 250 tweets.
- *expanded_urls* are empty for 59 tweets.
- some *expanded_urls* occurs two times in the different rows with the different tweet_id.

images table

- *tweet_id* is integer instead of string.
- *jpg_url* columns contains broken links (images no longer exist):
    - https://pbs.twimg.com/media/CWDbv2yU4AARfeH.jpg
    - https://pbs.twimg.com/media/C52pYJXWgAA2BEf.jpg
    - https://pbs.twimg.com/media/C6RkiQZUsAAM4R4.jpg
- some values in the *p1*, *p2*, *p3* columns are capitalised words while others are lowercase.
- *p1*, *p2*, *p3* values which consist of more than one word have underscores between words instead of spaces.
- *p1_conf*, *p2_conf*, *p3_conf* are in proportion forms instead of percentage.
- *jpg_url*, *img_num*, *p1*, *p2*, *p3*, *p1_conf*, *p2_conf*, *p3_conf*, *p1_dog*, *p2_dog*, *p3_dog* are not informative column names.

tweets table

- *contributors*, *coordinates* and *geo* columns do not contain any values.
- *place* column contains only one value.
- *id* is integer instead of string.
- *id_str* is redundant since it contains the same information as *id*.
- *id* and *created_at* should be renamed to *tweet_id* and *timestamp*.
- *display_text_range*, *entities*, *extended_entities*, *favorited*, *in_reply_to_screen_name*, *in_reply_to_status_id*, *in_reply_to_status_id_str*, *in_reply_to_user_id*, *in_reply_to_user_id_str*, *is_quote_status*, *possibly_sensitive*, *possibly_sensitive_appealable*, *quoted_status*, *quoted_status_id*, *quoted_status_id_str*, *quoted_status_permalink*, *retweeted*, *truncated*, *user* columns are redundant since they contain metadata information or the data which are not useful for further analysis.
- *source* is redundant since there is the same column in **archive** table.

all tables

- tables contain different numbers of rows: **archive** - 2356, **images** - 2075, **tweets** - 2334.

Tidiness

archive table

- *rating_numerator* and *rating_denominator* should be one variable - *rating*.
- *doggo*, *floofer*, *pupper*, *puppo* should be in one column as they are values of *dog_stage* variable.

all tables

- **archive, tweets** and **images** tables should be joined into one table since they have the same observational unit (a tweet).

Cleaning Data

archive | archive_clean

I created a copy of an **archive** dataframe named **arcive_clean** to clean the data.

ISSUE

There are 22 rows with tweets which no longer exist.

DEFINE

Delete all rows from the **archive_clean** with tweet ids from the dictionary created and saved to the file named *errors* during gathering data for *tweet_json.txt* and contains tweet ids for nonexistent tweets and their error codes.

ISSUE

tweet_id is integer instead of string.

DEFINE

Convert *tweet_id* from integer to string using astype.

ISSUE

timestamp is string instead of datetime.

DEFINE

Delete the column since the **tweets** table contains the same variable with a proper datatype and will be joined to **archive** table later.

ISSUE

source column contains values surrounded by html-tags.

DEFINE

Delete all html-tags including "<" and ">" symbols.

ISSUE

retweeted_status_id is not null for 181 which means that all these tweets were retweeted.

DEFINE

Delete all rows which contain retweeted tweets.

ISSUES

- *in_reply_to_status_id*, *in_reply_to_user_id*, *retweeted_status_user_id*, *retweeted_status_timestamp* columns are redundant since they are not useful for further analysis.
- *retweeted_status_id* is redundant after using this column for deleting retweeted tweets.

DEFINE

Delete all these columns.

ISSUES

- some *expanded_urls* values contain the same url more than once.
- some *expanded_urls* values contain several urls without spaces so these urls do not work properly.

DEFINE

Find *expanded_urls* values that are urls without spaces and separate them from each other, then delete duplicate urls.

ISSUES

- *rating_numerator* contains 24 values which are greater than 20 and 440 which are less than 10.
- *rating_denominator* contains 23 values which are not equal to 10.
- *rating_numerator* and *rating_denominator* could have been extracted with errors from 33 tweets.

DEFINE

- since *rating_numerator* and *rating_denominator* values could have been extracted with errors from 33 tweets, extract them in a different way and replace existing values in *rating_numerator* and *rating_denominator* columns taking into account that:
    - ▪ *rating_numerator* and *rating_denominator* values can contain decimals;
    - ▪ there are 33 tweets containing more than one number sequence which look like the rating (two numbers with a slash);
    - ▪ all these 33 tweets have the number sequence with 10 as a denominator as the last number sequence in the tweet;
- if there are rating for a bunch of dogs or there are no any rating delete these rows.

ISSUE

rating_numerator and *rating_denominator* should be one variable - *rating.*

DEFINE

- create new column *rating* where values will be equal to *rating_numerator*/*rating_denominator*;
- delete *rating_numerator* and *rating_denominator* columns.

ISSUES

- there are 109 values in the *name* column that do not seem like real names.
- some of rows where the *name* values do not seem like real names in fact contain a dog's name follows the word "named" in the *text* column.

DEFINE

- extract the names from *text* where a dog's name follows the word "named" using regex;
- then analyse remaining tweets visually and:
    - if there are any names assign these names to the values in the *name* column.
    - if there is no name assign *None* to the value in the *name* column.

ISSUES

- there are no any dog stage for 2326 tweets.
- *doggo*, *floofer*, *pupper*, *puppo* should be in one column as they are values of *dog_stage* variable.

DEFINE

- check how many any dog stages and their combination in *archive_clean* after previous cleaning;
- for the rows with more than one dog stage:
    - if there are any extracting errors for dog stages correct these dog stages;
    - if there are rows with tweets about not a dog drop these rows.
    - replace *None* values with empty strings, then create a new column *dog_stage* and fill it with values from *doggo*, *floofer*, *pupper*, *puppo* columns;
    - if there are any multistage values split them in *doggo*, *floofer*, *pupper* or *puppo*.
- replace empty string values in the *dog_stage* column with *None*;
- delete *doggo*, *floofer*, *pupper*, *puppo* columns.

ISSUE

doggo, *floofer*, *pupper*, *puppo* are string instead of string.

DEFINE

Convert *doggo*, *floofer*, *pupper*, *puppo* columns to categorical datatype using astype.

images | images_clean

I created a copy of an **images** dataframe named **images_clean** to clean the data.

ISSUE

tweet_id is integer instead of string.

DEFINE

Convert *tweet_id* from integer to string using astype.

ISSUES

jpg_url column contains broken links (images no longer exist):

- https://pbs.twimg.com/media/CWDbv2yU4AARfeH.jpg
- https://pbs.twimg.com/media/C52pYJXWgAA2BEf.jpg
- https://pbs.twimg.com/media/C6RkiQZUsAAM4R4.jpg

DEFINE

Delete the rows with tweets which contains the following links in *jpg_url* column:

- https://pbs.twimg.com/media/CWDbv2yU4AARfeH.jpg
- https://pbs.twimg.com/media/C52pYJXWgAA2BEf.jpg
- https://pbs.twimg.com/media/C6RkiQZUsAAM4R4.jpg

ISSUES

- some values in the *p1*, *p2*, *p3* columns are capitalised words while others are lowercase;
- *p1*, *p2*, *p3* values which consist of more than one word have underscores between words instead of spaces.

DEFINE

Replace underscores with spaces and capitalize all words in the *p1*, *p2*, *p3* values.

ISSUE

p1_conf, *p2_conf*, *p3_conf* are in proportion forms instead of percentage.

DEFINE

Convert *p1_conf*, *p2_conf*, *p3_conf* values to percentage form.

ISSUE

jpg_url, *img_num*, *p1*, *p2*, *p3*, *p1_conf*, *p2_conf*, *p3_conf*, *p1_dog*, *p2_dog*, *p3_dog* are not informative column names.

DEFINE

Rename *jpg_url*, *img_num,* *p1*, *p2*, *p3*, *p1_conf*, *p2_conf*, *p3_conf*, *p1_dog*, *p2_dog*, *p3_dog* to *image_url, image_order_number, prediction_1*, *prediction_2*, *prediction_3*, *confidence_percentages_1*, *confidence_percentages_2*, *confidence_percentages_3*, *dog_or_not_1*, *dog_or_not_2*, *dog_or_not_3*.

tweets | tweets_clean

I created a copy of an **tweets** dataframe named **tweets_clean** to clean the data.

ISSUES

- *contributors*, *coordinates* and *geo* columns do not contain any values;
- *place* column contains only one value;
- *display_text_range*, *entities*, *extended_entities*, *favorited*, *in_reply_to_screen_name*, *in_reply_to_status_id*, *in_reply_to_status_id_str*, *in_reply_to_user_id*, *in_reply_to_user_id_str*, *is_quote_status*, *possibly_sensitive*, *possibly_sensitive_appealable*, *quoted_status*, *quoted_status_id*, *quoted_status_id_str*, *quoted_status_permalink*, *retweeted*, *truncated*, *user* columns are redundant since they contain metadata information or data which are not useful for further analysis;
- *id_str* column is redundant since it contains the same information as *id*;
- *source* column is redundant since there is the same column in **archive** table.

DEFINE

Delete *contributors*, *coordinates*, *geo*, *place*, *display_text_range*, *entities*, *extended_entities*, *favorited*, *in_reply_to_screen_name*, *in_reply_to_status_id*, *in_reply_to_status_id_str*, *in_reply_to_user_id*, *in_reply_to_user_id_str*, *is_quote_status*, *possibly_sensitive*, *possibly_sensitive_appealable*, *quoted_status*, *quoted_status_id*, *quoted_status_id_str*, *quoted_status_permalink*, *retweeted*, *truncated*, *user*, *id_str*, *source* columns.

ISSUE

id is integer instead of string.

DEFINE

Convert *id* from integer to string using astype.

ISSUE

id and *created_at* should be renamed to *tweet_id* and *timestamp*.

DEFINE

Rename *id* to *tweet_id* and *created_at* to *timestamp*.

all tables | df_master

ISSUE

- **archive**, **images** and **tweets** tables should be joined into one table since they have the same observational unit (tweet);
- tables contain different numbers of rows: **archive** - 2356, **images** - 2057, **tweets** - 2334.

DEFINE

- merge **archive_clean, images_clean** and a**rchive_clean** tables into one table **df_master** on *tweet_id* column using only common values from all three dataframes;
- check if *text* and *full_text* column contain the same data and if it is true delete the *full_text* column;
- reorder the columns for clarity.

Finally, I saved **df_master** dataframe to **twitter_archive_master.csv.**