Zhaowei Liang
Udacity – Data Analyst Nanodegree
Feb 19 2018

Wrangle report

Data gathering

For this data wrangling projects, there are 3 sources of data: Manually downloaded WeRateDogs Twitter archive, programmatically downloaded image predictions result and API queried tweets' status data. After this three data files are ready to go, they are stored in 3 separated pandas DataFrames for data cleaning.

Data Assessing and Cleaning

These three Dataframes are named as shown in the table below:

Source	Raw dataframe(before cleaning)	Cleaned dataframe (after cleaning procedure)
Manually downloaded	archive_df	archive_df_clean
Programmatically downloaded	image_df	image_df_clean
From API	status_df	status_df

For all these three dataframes, couple pandas method was used includes: info(), value_counts, head(5) and sample(). The .info method

After the assessing process, the issues below were identified:

Quality issues

1. Retweets: some of the tweets in this dataframe are retweet. as mentioned in the project detail, These retweets are not supposed to be included in the analysis.
2. Unnecessary information: text, sources are not needed for analysis. Retweeted_status_id and retweeted_status_user_id, and retweetd_status_timestamp are not needed after data cleaning procedure.
3. Wrong data type for tweet id. Since no calculations will be applied on tweet ID, the tweet ID needs to be str instead of int64.
4. The timestamp column has wrong data type.
5. Non-descriptive column headers: p1, p1_conf, p1_dog, p2, p2_dog, p3, p3_conf, p3_dog etc.
6. Some of the dog breeds has first letter capitalized and some are not.
7. Many ratings' denominator are not 10, even the numerator which greater than 10 is the feature of this twitter account, keep the denominator same is vital to later analysis.
8. Date and time are in the same column, this is not necessarily a tidyness issue, because its nothing wrong with putting this two in same column. It still need tobe parsed because we will perfome analysis around date and time later.

Zhaowei Liang
Udacity – Data Analyst Nanodegree
Feb 19 2018

Tidiness issues:

1. One observation unit does not form a table. At least, retweet_count and favorite_count needs to be part of the archive dataframe to form a complete observation unit.
2. Dog stages are not in one column, instead, they are divided into 4.

These issues are corrected through multiple method include change the type of columns, extend/specify non-descriptive column names, inner-join different dataframe to form an observation unit.

The final observation unit, so called twitter_df_merged, is used for visualization and analyzation later.

Data Analysis

3 insights and 1 visualization were conducted in this data analysis procedure. Firstly, the conditional mean feature of pandas was used to find the most popular dog breeds, most favorited (liked) breeds and most popular dog names. These results are found by sort the dataframe and display the table contains relevant data.

Then, the pyplot date plot feature was used which ploted sum of likes and retweets of WeRateDogs account through 2 years from 2015-11 to 2017-09. As shown in the graph, the WeRateDogs account are generally going up trend.