Wrangle and Analyze Data

Project Report

by Markus Müller

In reality, it is rarely possible to analyze data sets directly without making any changes. This project shows the individual steps to process multiple datasets in such a way that the data can be analyzed without problems.

The data wrangle process is divided into three phases: (1) gather data, (2)assess data and (3)clean data. Each step has a specific task :

(1) Gathering Data

In the gather phase the individual data sets are downloaded and specific dataframe are created from the datasets

I have downloaded data from the following sources:

- Twitter archive Data from @WeRateDogs, which was provided by Udacity to download
- The image predictions from a neural net, which were also provided by Udacity but were downloaded programmatically
- Twitter API data from Tweepy to gather retweet counts and favorite counts for each tweet. I have downloaded it from Udacity, because I had problems creating a developer account.

(2) Assessing Data

In the assessment phase datasets are visually and programmatically evaluated to find quality and tidiness issues. The individual observations are written down and used as the basis for the final phase: cleaning the datasets.

I found the following quality and tidiness issues:

Quality Issues:

twitter_archive:

- 181 retweets, which aren‘t original and therefore should be removed from the data
- 78 replies, the same as for the retweets
- There are a few irrelevant columns related to retweets and replies
- Tweet_id is an integer, which could be change since we don’t calculate with it
- Timestamp is a string. It should be a datetime object to make further analysis easier regarding the time period
- The extracted rating is wrong sometimes (decimals in the rating and false extraction)
- There are inconsistent rating denominators that are unequal to 10 and some outliers for the rating numerators
- The name column contains strings that aren’t related to dog names

twitter_api_data:

- Again tweet_id is an integer

image_predictions:

- There are 66 duplicates in the column jpg_url
- A few irrelevant columns for the predictions of the neural network
- The prediction of the dog type is inconsistent in terms of capitalization

Tidiness issues

twitter_archive:

- The dog stages variable is split into 4 columns (doggo, floofer, pupper & puppo)

general:

- All dataframes are part of the same observational unit so they should be merged into one dataframe

(3) Cleaning Data

In the cleaning phase the observations are defined into exact steps, than the steps are transformed into code and finally tested whether the cleaning process has produced the expected result.

1. I dropped replies and retweets from twitter_archive to make sure that only original data is used for the analysis.
2. Dropped the related columns for replies and retweets since they don’t provide any value anymore
3. Transformed the twitter_id from an integer into a string for all dataframes
4. Transformed the timestamp into a Datetime object
5. Removed the issues regarding the inaccurate rating and stored the right values for the numerators and denominators
6. Removed inconsistent ratings for denominators and outliers for numerators
7. Replaced unrelated strings for dog names in the name column with ‘None’
8. Droped duplicated in the jpg_url columns in image_predictions
9. Created a column in image_predictions for the first right dog type prediction and its confidence level
10. Removed 4 columns for the dog stages in twitter_archive and created one variable with the four dog stages as a category with regular expression.
11. Created one dataframe and rearranged columns to enhance readability