Wragling Report - WeRateDogs Twitter

BY YANG, Linjing

2019/02/10

## Data Gathering

Data was gathered from 3 different sources.

### **1. Twitter Archive**

The Twitter archive #WeRateDogs, a csv file that contains various kinds of information, such as dog ratings, stages, tweet ids, posted date, etc.

### **2. Additional Information via the Twitter API**

The retweet and favorite count of each tweet gathered from Twitter's API by Tweetpy Library in a Json format.

### **3. Image Prediction File**

The prediction of dog images and breeds downloaded programmatically by Requests Library from the neural network.

## Data Accessing

Data was accessed based on both of quality and tidiness issues.

### Dataset 1

#### Quality issues

Completeness:

* in\_reply\_to\_status\_id: 78 out of 2356 is non-null
* in\_reply\_to\_user\_id: 78 out of 2356 is non-null
* retweeted\_status\_id: 181 out of 2356 is non-null
* retweeted\_status\_user\_id: 181 out of 2356 is non-null
* retweeted\_status\_timestamp: 181 out of 2356 is non-null
* name: string 'None' should be replaced by Null
* Since only original ratings (no retweets) that have images, the rows of retweets / replys could not contribute to this analysis and should be deleted.
* After retweet / reply rows are deleted, the following columns could be dropped:
  + in\_reply\_to\_status\_id
  + in\_reply\_to\_user\_id
  + retweeted\_status\_id
  + retweeted\_status\_user\_id
  + retweeted\_status\_timestamp

Validity:

* rating\_denominator contains invalid values (e.g. denominator = 0, index = 313)
* invalid names like *an, the, only*, etc.

Accuracy:

* rating\_numberator contains extremely large values (e.g. 1776 when denominator = 10)
* Since numerators are extracted from 'text', the numerators with decimals were wrongly extracted (index = 45, 340, 695, 763, 1689, 1712)

Data Types:

* tweet\_id: int -> obj
* timestamp: obj -> datetime
* in\_reply\_to\_status\_id: float -> object
* in\_reply\_to\_user\_id: float -> object
* retweeted\_status\_id: float -> object
* retweeted\_status\_user\_id: float -> object

#### Tidiness Issues

* doggo / floofer / pupper / puppo: could be combined into 1 column as categorical data.
* source: the long name with HTML tag could be shorten.

### Dataset 2

#### Quality issues

Completeness:

* retweet & favourite count: "Nan" should be replaced by Null.
* retweet & favourite count: 16 missing values

Data Type:

* retweet & favourite count: obj -> int

### Dataset 3

#### Quality Issues

Consistency:

* The predicted names are written in both upper and lower cases.

Data Types:

* tweet\_id: int -> obj

The column names could be more clear.

## Data Cleaning

* The three datasets were combined into one with more clear column names.
* The columns / rows that could not contribute to analysis were dropped.
* The dog stages were combined into one column as categorical data.
* The HTML tags in the ‘source’ column were deleted.
* The row with invalid rating denominator was deleted.
* The wrongly extracted rating numerators were corrected according to the information in *text* column.
* The rating outliers were deleted.
* Missing values were presented as Null.
* The data types of all columns were corrected.
* Unclear column names were renamed.
* The predicted dog types were changed to lowercase.