Wrangling Report : WeRateDogs Data Analysis

Kaori Ishikawa

1 Introduction

I wrangled WeeRateDogs Twitter data for the further analyses and visualizations. I will briefly describe the detail of wrangling effort.

2 Gathering

2.1 **The WeRateDogs twitter archive**

It is manually downloaded from the Udacity website.

2.2 **The tweet image predictions**

It is programmatically downloaded from the Udacity's servers with the Requests library.

2.3 Retweet/favourite count data

Retweet/favourite count data is downloaded from twitter. By sending queries the twitter API by Tweepy library, set of JSON data is downloaded and a pandas DataFrame is created based on this.

3 Accessing and Cleaning

3.1 The WeRateDogs twitter archive

3.1.1 Text column included not only text but also URL. (Tidiness issue) I made a new column for URL and split the data.

3.1.2 Data type of timestamp was object instead of time series data. Also, five columns related to ID were integer instead of string. (Quality issue) Therefore, I corrected these the data type.

3.1.3 Dog stage was spread over the four columns, while it should be in one column. (Tidiness issue) I melted four columns into one. As some dogs have multiple stages, some values in that column are combined stage.

3.1.4 Retweet/reply were included in the data. (Tidiness issue) I removed the rows with the ID of retweet or reply in order to keep only the original tweet.

3.1.5 The summary statistics showed that some of the denominator were not 10. After the review of the suspicious data such as too big or 0, indeed some of them were incorrect; in general, denominator is 10 time the number of the dog(s) in the picture. Also, when the denominator is incorrect, numerator is in principle also incorrect together. (Quality issue) After filtering the suspicious data as mentioned above, I manually updated the data to the correct one.

3.1.6 Ratio of rating was missing in the table. As the denominator of rating is not united, ratio is helpful. (Quality issue) I made a new column of ratio of rating (nominator/denominator) , whose datatype is float.

3.1.7 “Floof” is the same as “floofer”, one of the stage of dog. However, floofer columns didn’t reflect this correctly. (Quality issue) I updated the value of “floofer” when the text contains “floof”.

3.1.8 Some of the dogs’ name were incorrect. By visually assessing the suspicious name such as shorter than 2 alphabet, pronoun and articles, I also found a pattern that incorrect names starts from lower case. (Quality issue) I changed the incorrect data in the name column into None.

3.2 Image prediction data

3.2.1 Some predictions were not the breed of dog. (Quality issue)

3.2.2 There were three predictions with the different level of confidence in one table. (Tidiness issue) To solve 3.2.1 and 3.2.2, I made a new prediction column which includes only the breed of dog with the highest level of confidence.

3.2.3 As data type of Tweet ID was integer, it was changed into string. (Quality issue)

3.3 Retweet/favourite count data

As data type of Tweet ID was integer, it was changed into string. (Quality issue)

3.4 General

3.4.1 Data was stored in three tables, although it is a single observational unit. (Tidiness issue) Therefore, I merged three tables into one.

3.4.2 As unnecessary columns were included in some tables such as URL of each pictures and expanded URL of tweet, these were removed. (Tidiness issue)

4 Summary

As briefly discussed above, I cleaned 6 tidiness issues and 8 quality issues. And, there is only one table containing the data which was spread to three tables.