wrangle_report

January 2, 2018

1 Introduction

This report describes wrangling efforts for the project 3: Wrangle and analyze data of the Udacity data analyst nanodegree. The outline of the report follows the typical steps in any data wrangling process:

- Gathering data
- Assessing data
- Cleaning data

2 Gathering data

For this project 3 sources of data were used:

1. WeRateDogs Twitter archive
2. Tweet image predictions
3. Each tweet's retweet count and favorite

The WeRateDogs archive and tweet image predictions were read into a pandas dataframe. the tweepy library was used to extract additional data from the twitter API. This resulted in 3 separate pandas dataframes.

3 Assessing data

The gathered data was assessed for quality and tidiness issues using a combination of visual and programmatic techniques. This resulted in the following quality and tidiness issues. Although much more could be identified this project is limited to 8 quality issues and 2 tidiness issues.

3.1 Quality

- None in dog stage columns and name is not recognised as null value
- Timestamp is an object and no datetime
- 181 retweets in the twitter archive. We only want original ratings.
- 78 replies in the twitter archive. We only want original ratings.
- When multiple dogs are in the picture the denominator can be a multiplication of 10 for each dog in the picture

- Some numerators and denominators are wrong extracted (810984652412424192, 740373189193256964, 682962037429899265 & 666287406224695296)
- 2 numerators are equal to 0
- Image predictions are not always dogs

3.2 Tidiness

- Dog "stage" columns, name and ratings are more linked to image predictions which would result in a table containing all information on the dogs
- retweet count and favorite count should be linked to the twitter archive

4 Cleaning data

The issues identified were cleaned in the next step. First the issues related to missing data were solved. Then tables were reorganised according to the rules of tidiness. Finally the remaining quality issues were solved. The result of this is are two separate clean and tidy datasets. The first dataset (twitter_archive_master.csv) contains all data related to the twitter post. The second dataset (dog_info.csv) contains additional data derived from the twitter post itself and related to the dogs shown in the twitter images.

One issue that was not resolved is related to the dog stages. According to the rules of tidiness they should not be in separate columns. However, an analysis of the values showed no clear distinction between the different stages. Since my knowledge of dogs is limited I choose to ignore this issue for this project.

As mentioned above not all possible issues in the dataset were resolved but the resulting dataset was sufficiently clean to gain some insights on the popularity of different dog breeds.