Act report – WeRateDogs

Project: Wrangling and Analyze Data

Dmitry Zmienko

1. Introduction

Real-world data rarely comes clean. And if we want to get some useful information from big data sets we have to know how to gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. We will take a real data from tweet archive of Twitter user WeRateDogs (@dog_rates), assess its quality and tidiness, clean it and then gather some useful insights.

2. Data wrangling

Whole data is gathered from three pieces:

- The twitter_archive_enhanced.csv that contains information about 2000+ tweets (data, source, URLs, dog\s name, ratings, etc.)
- The image_prediction.tsv file with results of image predictions of breed of dog (top three only) from neural network.
- Additional data from the twitter API about tweet's retweet and favorite counts stored in tweet_json.txt.

After the data was gathered it required to be assessed for quality and tidiness issues.

The four main data quality dimensions are:

- Completeness: missing data?
- Validity: does the data make sense?
- Accuracy: inaccurate data?
- Consistency: standardization?

Three requirements for tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

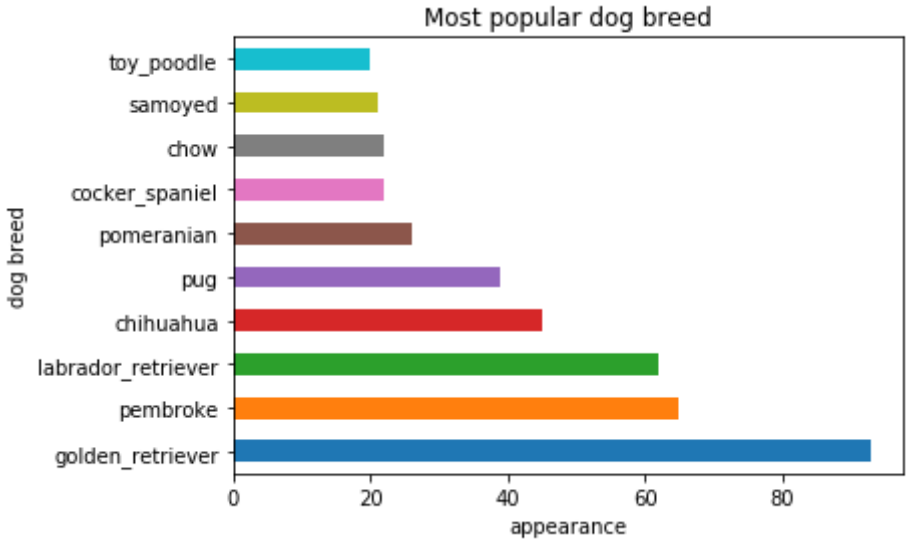
After the assessment the cleaning procedure (define – code – test) were performed. The three cleaned data frames were merged in one (also saved in twitter_archive_master.csv) for further analysis.

3. Analyzing and Visualizing Data

To check the usefulness of cleaning data let's look for these insights:

- Most popular dog breed and dog name
- Most common rating
- Most popular dogs (most favorite & retweeted)
- Tweeting activity

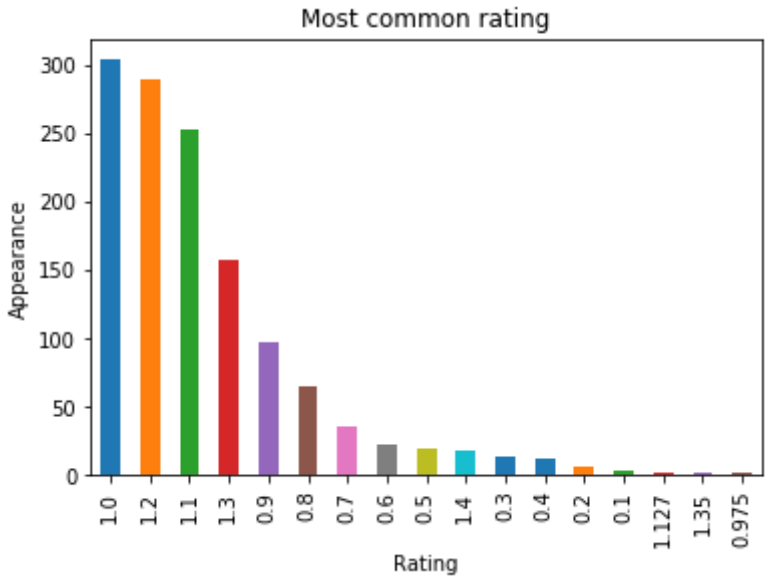
3.1 Most popular dog breed and dog name



From the plot above (the dog breed with mote than 20 appearance) we see that the Golden Retriever is most popular dog breed, while Oliver the most popular dog name (Winston and Tucker slightly less popular).

3.2 Most common rating

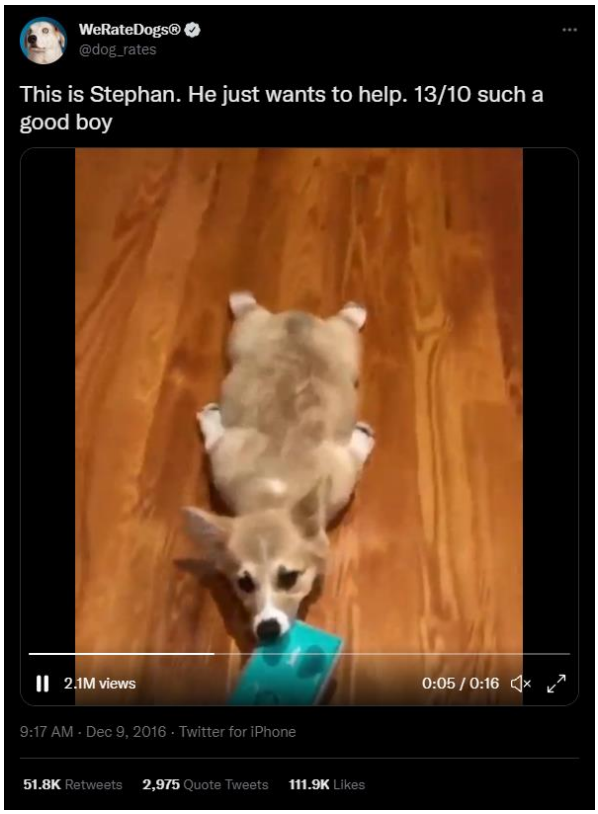
The interesting fact of WeRateDogs tweets is the rating system witch has to be from 0 to 10, but the dogs are so cute that they often get more than 10. So for analysis issues its better to use relative values ('rating_numerator' / 'rating_denominator').



The most common rating is 10/10 (as expected), but the love to puppies is so big that 12/10 and 11/10 is also often occurs

3.3 Most popular dogs (most favorite & retweeted)

If we look at tweets with most retweets and ‘likes’ we’ll se that It’s the same cute dog Stephan:



3.4 Tweeting activity

From the plot ao tweeting activity we see that at first when the tweeter start collect puppies the activity was 4 time bigger after several months. But after that the activity is just slightly decreasing and we can say about loyal subscribers.