Analysis and Visualization

Using the cleaned data set twitter_archive_master.csv I proceeded to explore the different values in some of the columns.

- For 'favorite_count' and 'retweet_count' all rows have the same value, so these columns are not useful for correlation analysis.
- The column ‘source’ sows that most pictures are uploaded from an iPhone.
- The most common name is Charlie.
- I was interested to find out the distribution of dogs among their stages and this dataset is formed mostly (More than 50%) by dogs in the pupper stage (figure 1).

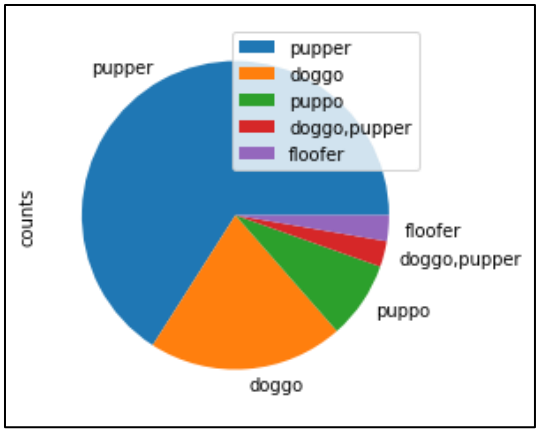

Figure 1, dog stage distribution.

- Next, I was interested to find out how the dog’s stage affected the given rating. The data shows that doggo, puppo and doggo/pupper receive the higher ratings suggesting that people who participates in this polling system prefers older dogs. Puppers fall behind other dogs in ratings so it seems people that rate these dogs are not particularly fond of younger dogs (figure 2).

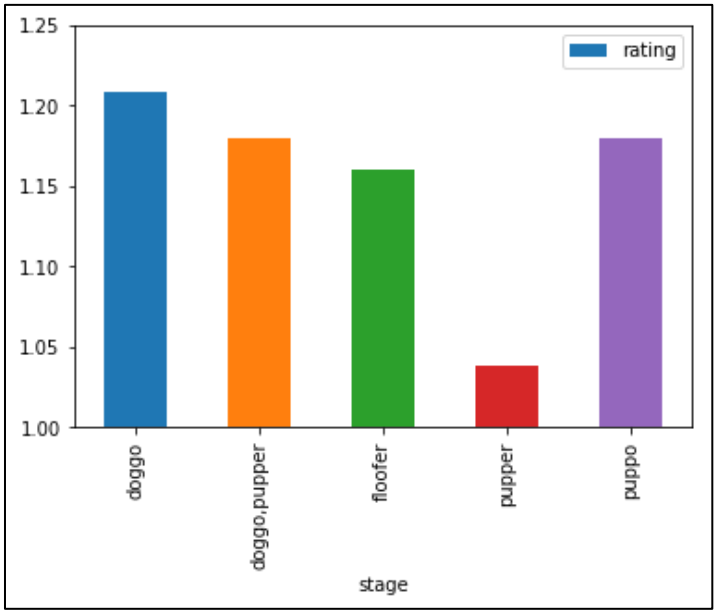

Figure 2, dog stage vs. rating.

- The project mentioned that the rating system is flexible and that users can potentially use any numerator and denominator they want, so I got interested in how the values have changed over time. The data shows that almost all users keep 10 as denominator but, the rating numerator increased from an average of 9 up to 13 until January 2017 when this value stabilized between 12 and 13, the rating followed the same pattern as the rating numerator with ranges between 0.9 to 1.3 (figure 3).

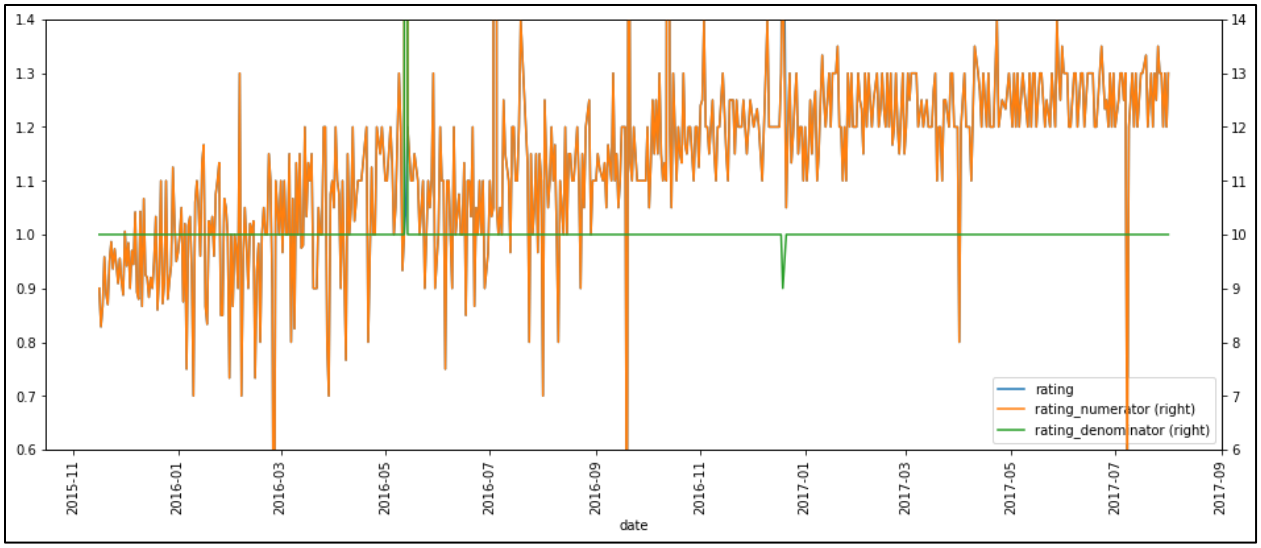

Figure 3, Rating system evolution over time.