Project: Wrangling and Analyzing data

For the ALX-Udacity Data Analysis with Python Nanodegree, we were given a project to wrangle and analyze three datasets containing information on tweets about rating dogs.

Wrangling Data

The three datasets were gathered differently.

The first dataset was provided and I downloaded it directly from the Udacity website and loaded it into pandas.

For the second dataset, I had to download programmatically from a website using the requests library.

The third dataset was queried from Twitter API, tweepy.

The three datasets were then merged for tidiness and unnecessary columns were removed. After assessment, eight quality issues were detected and cleaned.

After wrangling, the clean master dataset was stored as a CSV.

Insights and Visualization

The following insights were derived from analyzing this dataset.

1. On average, posts on the dog ratings were retweeted 2765 times.

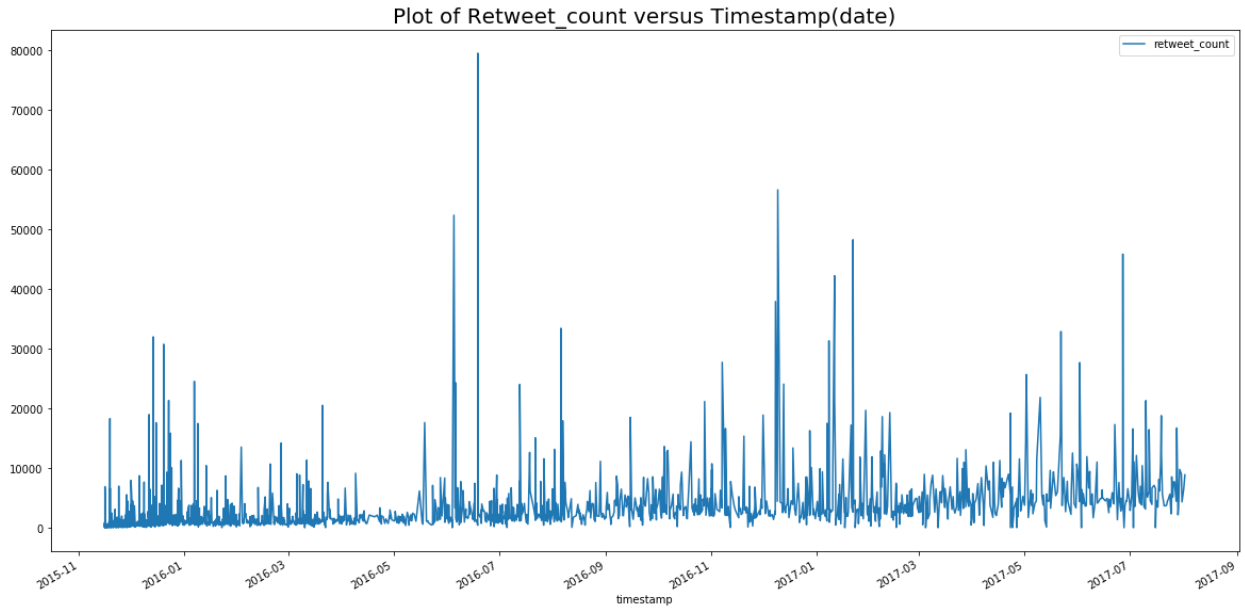
2. The minimum rating given to a dog was 0/10. This must have been a very bad dog.

The maximum rating given was 1776/10. This is either a typo error or we have a super dog on our hands. The average rating of dogs though was 12/10.

3. The dog with the highest favorite count was at puppo stage and had a rating of 13/10. That's higher than average.

4. The dog with the highest retweet count was at doggo stage, most likely a Labrador retriever from predictions and had a rating of 13/10.

For visualization, a line plot of retweet count versus timestamp was done to know which day had the highest retweet count.



From the plot above, you can see that the highest retweet count was recorded in July, 2016.

Was this because of how good the dogs were, or people are generally more active in July?

Whatever reason, further analysis will tell us.