UDACITY PROJECT
WRANGLE AND ANALYZE DATA

18.06.2020

OVERVIEW

1. Project Background and Description

Real-world data rarely comes clean. Using Python and its libraries, you will 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. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

2. Software Requirements

The entirety of this project can be completed inside the Udacity classroom on the **Project Workspace: Complete and Submit Project** page using the Jupyter Notebook provided there. (Note: This Project Workspace may not be available in all versions of this project, in which case you should follow the directions below.)

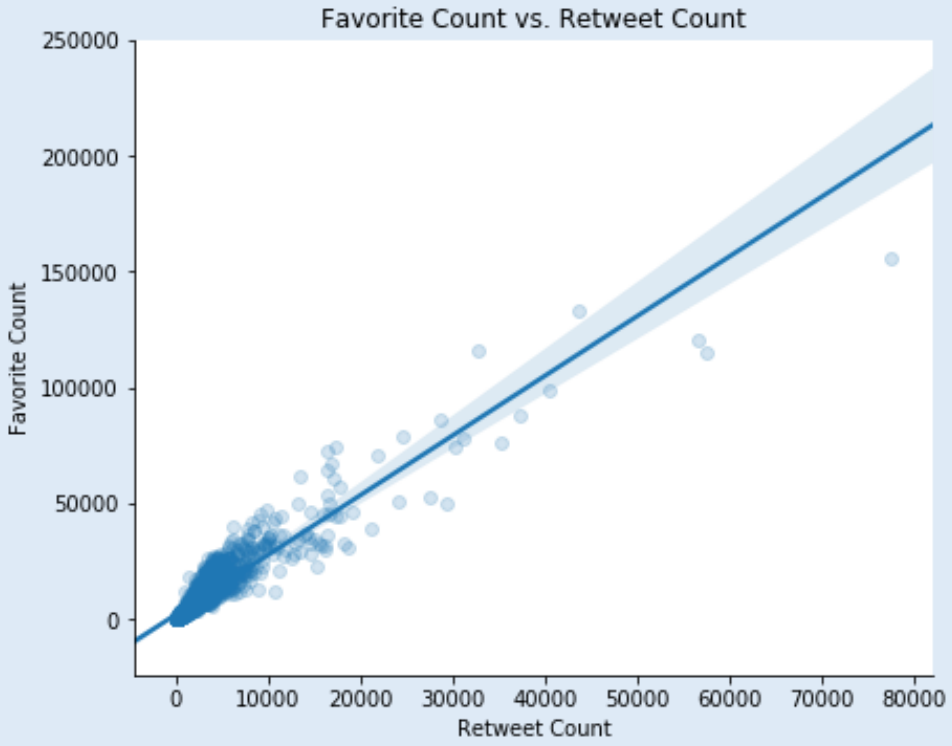
If you want to work outside of the Udacity classroom, the following software requirements apply:

- You need to be able to work in a Jupyter Notebook on your computer. Please revisit our Jupyter Notebook and Anaconda tutorials earlier in the Nanodegree program for installation instructions.
- The following packages (libraries) need to be installed. You can install these packages via conda or pip. Please revisit our Anaconda tutorial earlier in the Nanodegree program for package installation instructions.
    - pandas
    - NumPy
    - requests
    - tweepy
    - json
- You need to be able to create written documents that contain images and you need to be able to export these documents as PDF files. This task can be done in a Jupyter Notebook, but you might prefer to use a word processor like Google Docs, which is free, or Microsoft Word.
- A text editor, like Sublime, which is free, will be useful but is not required

3. ANALYSIS AND INSIGHTS

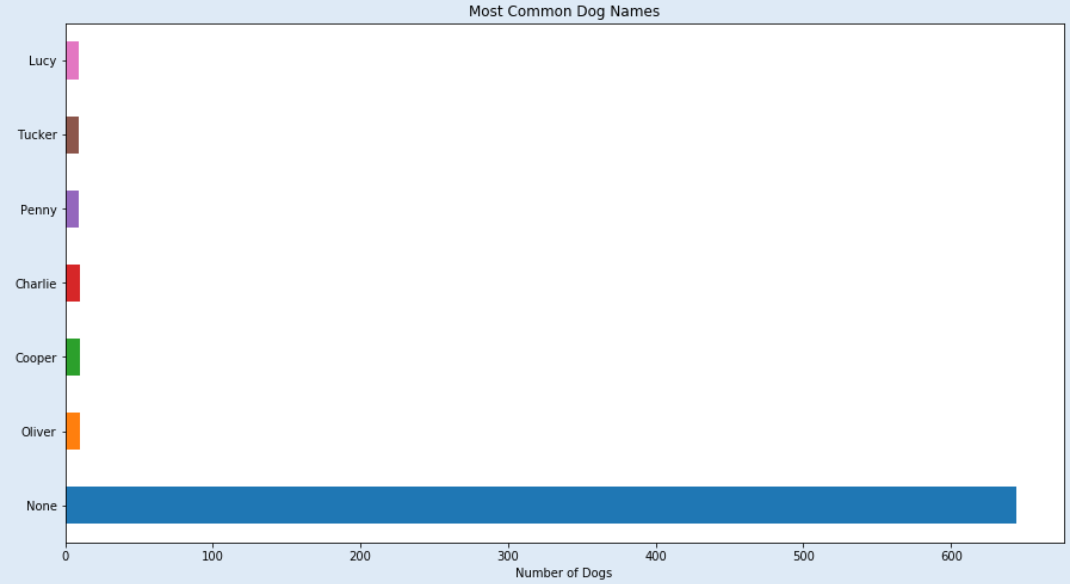
FAVORITE COUNT VS RETWEET

WeRateDogs had more than 4 million followers at the time this data was collected; therefore, their tweets are likely to get many favorites and retweets. In fact, if they are part of international news coverage or go viral, there might be some tweets which are highly common. Figure shows that favourite and retweet counts are highly correlated positively



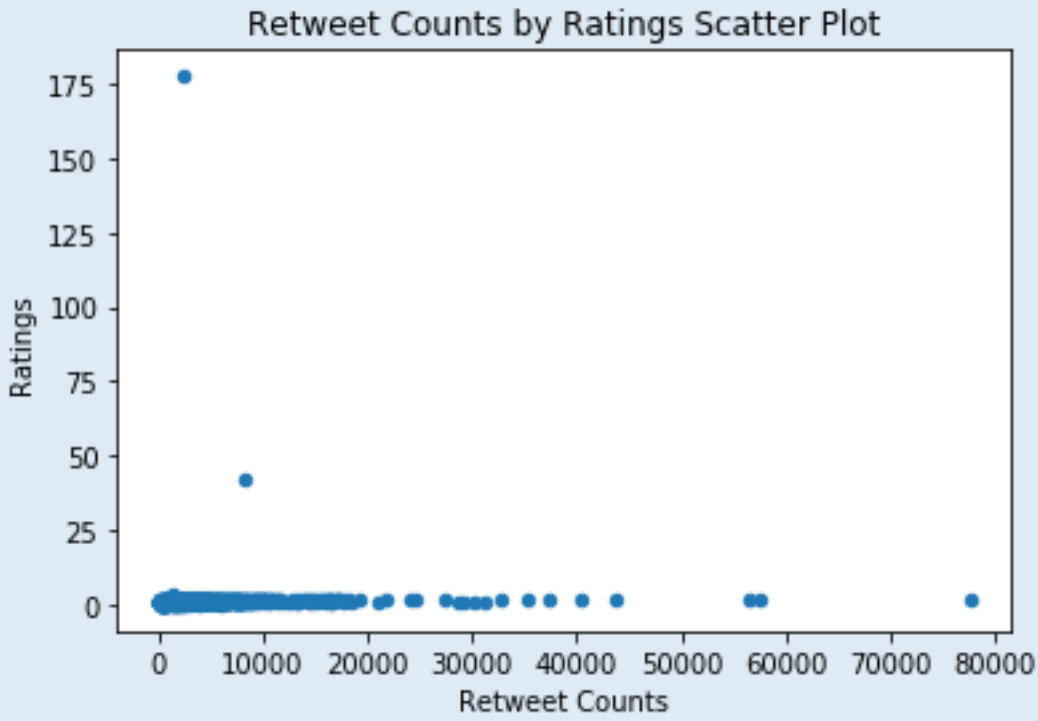
Dog Names Common

Most popular dog names were Oliver, Cooper etc.



Retweet Counts

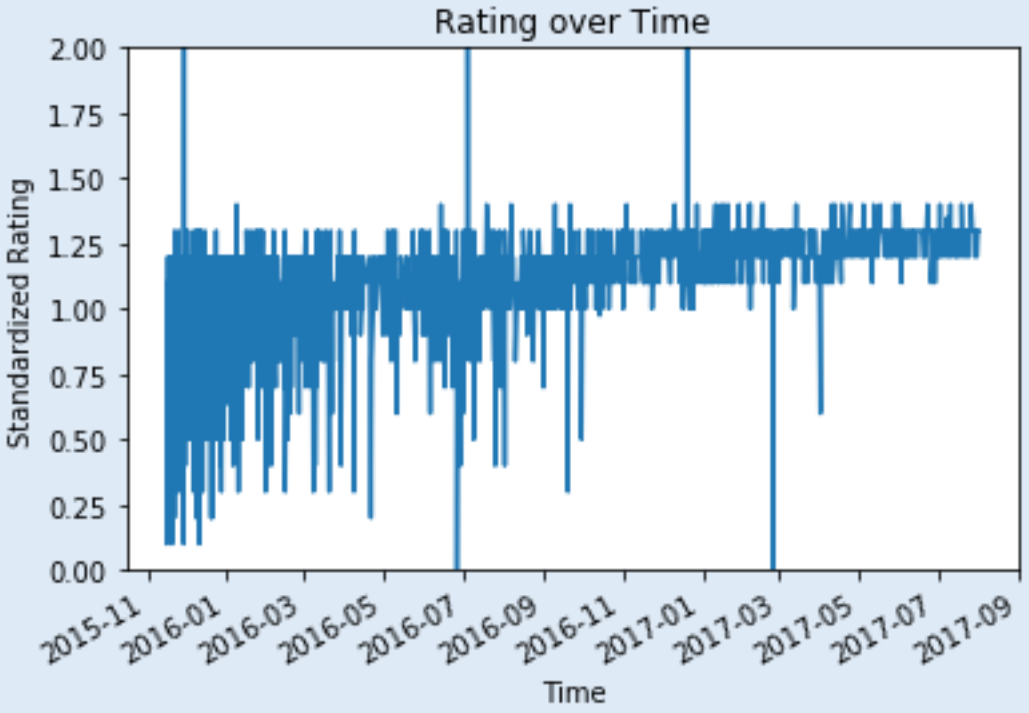
Many tweets have been retweeted more times. So there is no clear relationship between ratings and the retweets. The highest ratings do not receive the most retweets.



STANDARDIZED RATING OVER TIME

The idea behind the WeRateDogs account is that they ask people to send them photos of their dogs and, with humorous comments, they will rate them on a scale of 1-10; however, they are often given ratings above 10. I assumed that nearly all the dogs had a rating higher than 10/10 but I was surprised to notice many with ratings below 10/10.

Additionally, there were many ratings with no denominator of 10. Therefore I calculated a numbered value divided by denominator to standardize the ratings.



Indeed, the frequency of ratings below 1 appears to decrease overtime. There were many ratings below 1, before 2016-11, while there were barely any after that time.