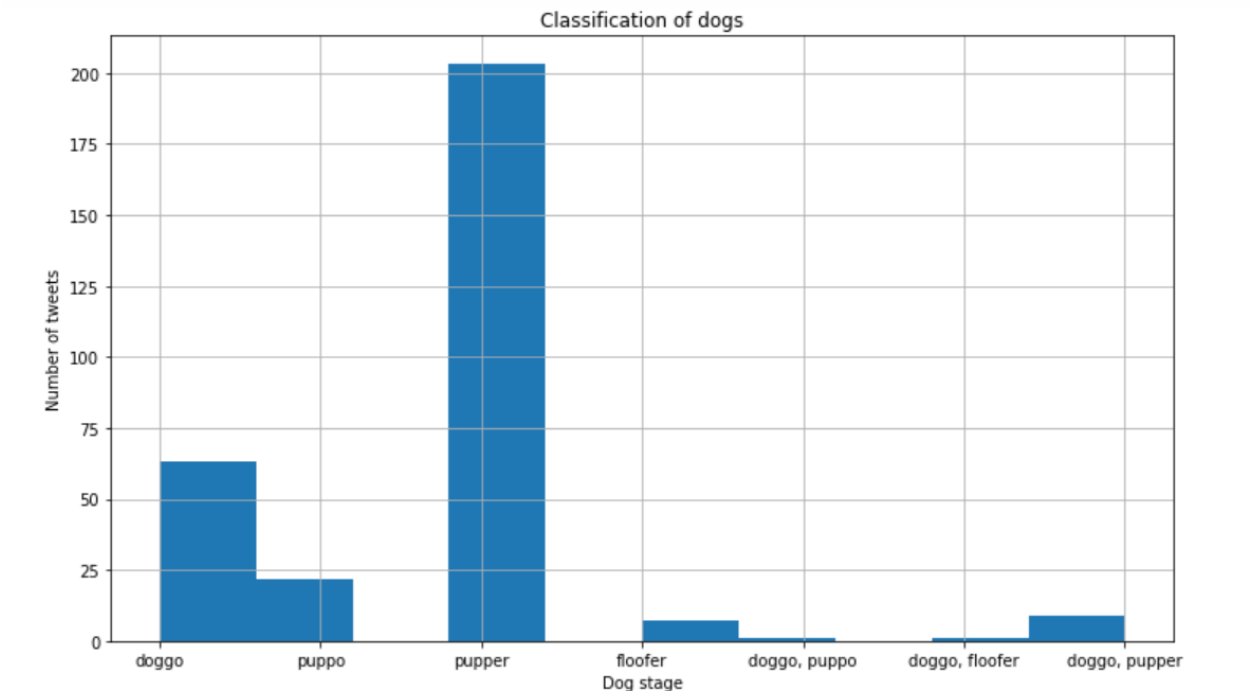
Introduction

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.

Visualization and reporting:



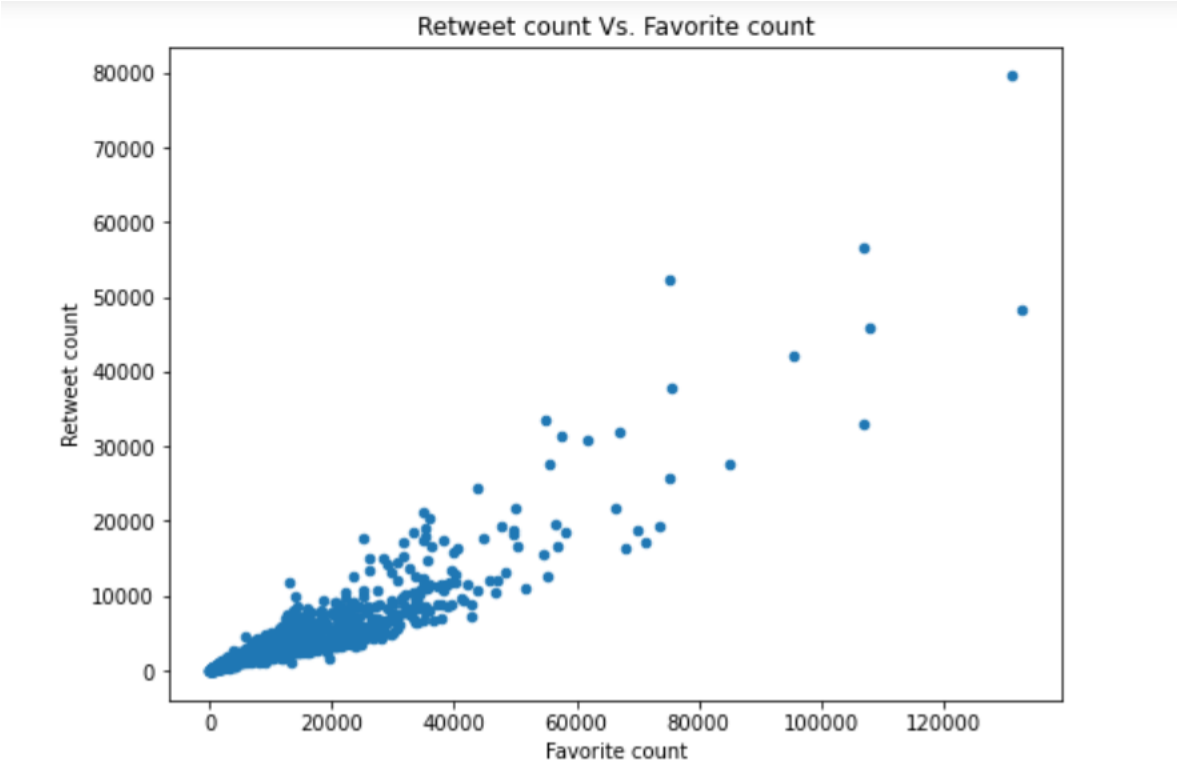
dog_stage	rating_numerator	
doggo	13.0	19
	11.0	16
	12.0	12
	10.0	9
	14.0	7
doggo, floofer	11.0	1
doggo, pupper	12.0	5
	10.0	2
	11.0	1
	13.0	1
doggo, puppo	13.0	1
floofer	13.0	3
	12.0	2
	10.0	1
	11.0	1
pupper	10.0	86
	11.0	50
	12.0	47
	13.0	12
	14.0	7
	27.0	1
puppo	13.0	9
	12.0	6
	10.0	4
	11.0	2
	14.0	1

Looking at the visualization above, it's obvious that the dog class "pupper" is the most common class among the classification of dogs and when we go deeper and look at the number of tweets ratings we can conclude that most of the bad rated dogs "10" that equal to 88 tweets are classified under the dog class "pupper" while the most of the good rated dogs "14" are equally distributed between two classes "doggo" and "pupper" and some have multiple classes of dogs for example: one has "doggo, floofer" class, five have "doggo, pupper" and one has "doggo, puppo" class.

:

	tweet_id	rating_numerator	rating_denominator	img_num	p1_conf	p2_conf	p3_conf	retweet_count	favorite_count
count	1.994000e+03	1994.000000	1994.000000	1994.000000	1994.000000	1.994000e+03	1.994000e+03	1994.000000	1994.000000
mean	7.358508e+17	12.860582	10.532096	1.203109	0.593941	1.344195e-01	6.024848e-02	2766.753260	8895.725677
std	6.747816e+16	41.431360	7.320710	0.560777	0.271954	1.006807e-01	5.089067e-02	4674.698447	12213.193181
min	6.660209e+17	10.000000	2.000000	1.000000	0.044333	1.011300e-08	1.740170e-10	16.000000	81.000000
25%	6.758475e+17	10.000000	10.000000	1.000000	0.362857	5.393988e-02	1.619283e-02	624.750000	1982.000000
50%	7.084748e+17	11.000000	10.000000	1.000000	0.587635	1.174550e-01	4.950530e-02	1359.500000	4136.000000
75%	7.877873e+17	12.000000	10.000000	1.000000	0.846285	1.951377e-01	9.159438e-02	3220.000000	11308.000000
max	8.924206e+17	1776.000000	170.000000	4.000000	1.000000	4.880140e-01	2.734190e-01	79515.000000	132810.000000

Looking at "rating_numerator" column and as we consider the rating is almost always out of 10, we can conclude from describe the average rating equals to "12.86" and the maximum rate number is "1776" and 50% of these rating equals to "11".



This visualization shows us that there is a strong positive correlation between Retweet count and Favorite count, which indicates that there is a such evidence when someone favorite a tweet, he is more likely to retweet this tweet, so as the favorite count increases, the retweet count will increase as well.