Act Report

The dataset that you will be wrangling (and analysing 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.

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.

WeRateDogs™ @dog_rates · Apr 16

Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16

A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

The Jupyter Notebook has an intuitive, easy-to-follow logical structure. The code uses comments effectively and is interspersed with Jupyter Notebook Markdown cells. The steps of the data wrangling process (i.e. gather, assess, and clean) are clearly identified with comments or Markdown cells, as well.

It's great to see that you have organized the notebook in the 4 distinct sections of GATHER / ASSESS / CLEAN and ANALYZE. The notebook is interspersed with code and markdown text. This helps anyone in following along the work and can also understand the process flow that you have taken.

Two types of assessment are used:

- Visual assessment: each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes. Once displayed, data can additionally be assessed in an external application (e.g. Excel, text editor).
- Programmatic assessment: pandas' functions and/or methods are used to assess the data.
- `info(), describe(), value_counts(), sum() and duplicated()` to explore more about the data.

  Back to the basic-ness of Twitter archives: retweet count and favorite count are two of the notable column omissions. Fortunately, this additional data can be gathered by anyone from Twitter's API. Well, "anyone" who has access to data for the 3000 most recent tweets, at least. But you, because you have the WeRateDogs Twitter archive and specifically the tweet IDs within it, can gather this data for all 5000+. And guess what? You're going to query Twitter's API to gather this valuable data.

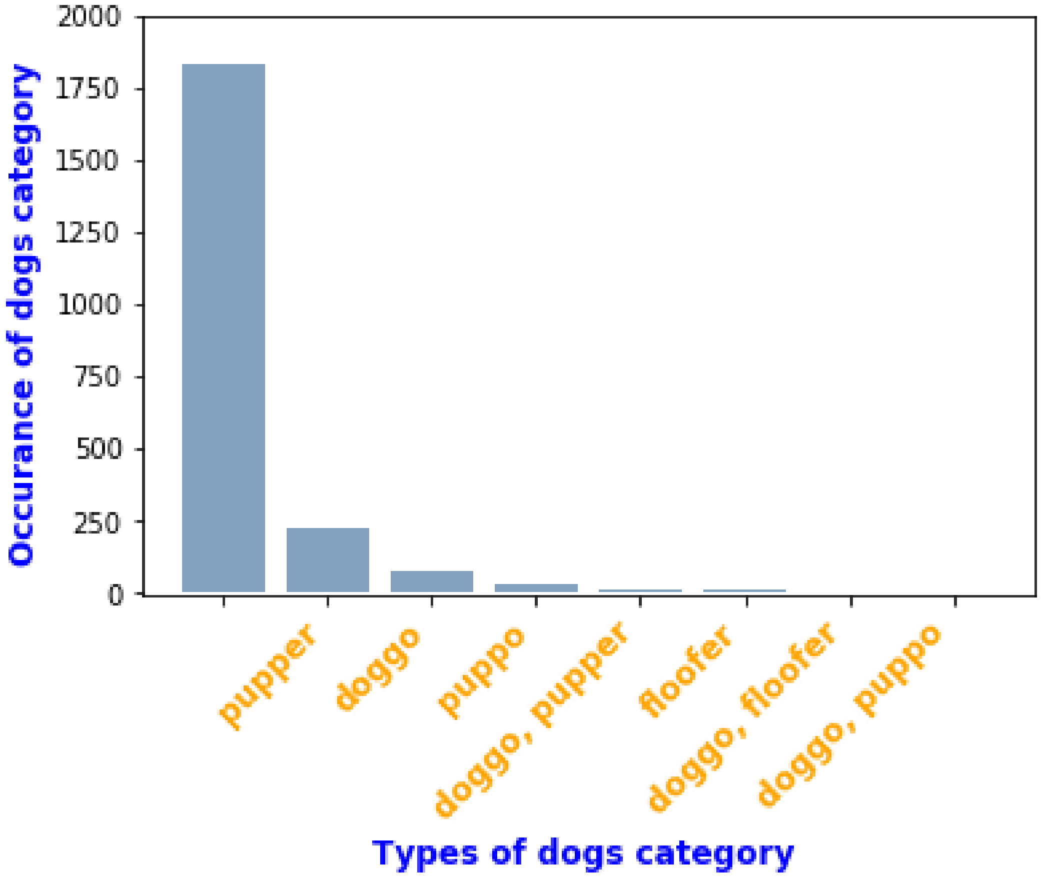
Key Points

Key points to keep in mind when data wrangling for this project:

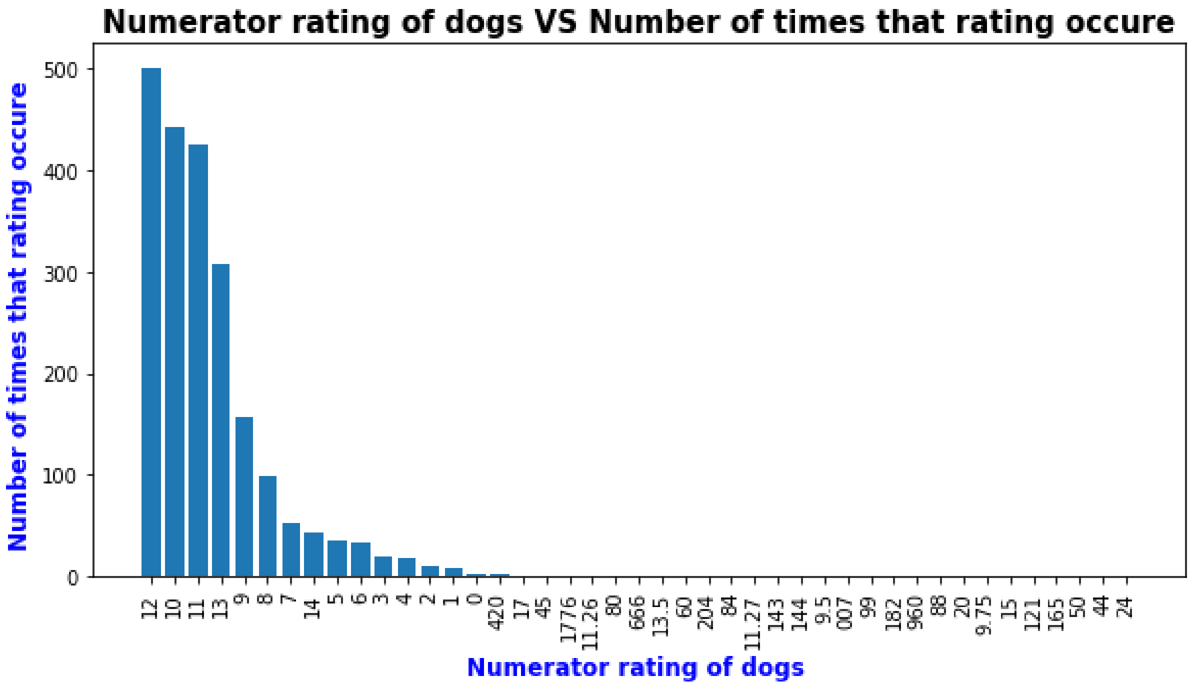
- You only want original ratings (no retweets) that have images. Though there are 5000+ tweets in the dataset, not all are dog ratings, and some are retweets.
- Assessing and cleaning the entire dataset completely would require a lot of time and is not necessary to practice and demonstrate your skills in data wrangling. Therefore, the requirements of this project are only to assess and clean at least 8 quality issues and at least 2 tidiness issues in this dataset.
- Cleaning includes merging individual pieces of data according to the rules of tidy data.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs.
- You do *not* need to gather the tweets beyond August 1st, 2017. You can but note that you won't be able to gather the image predictions for these tweets since you don't have access to the algorithm used.

After cleaning and analysing data, I get following outputs:

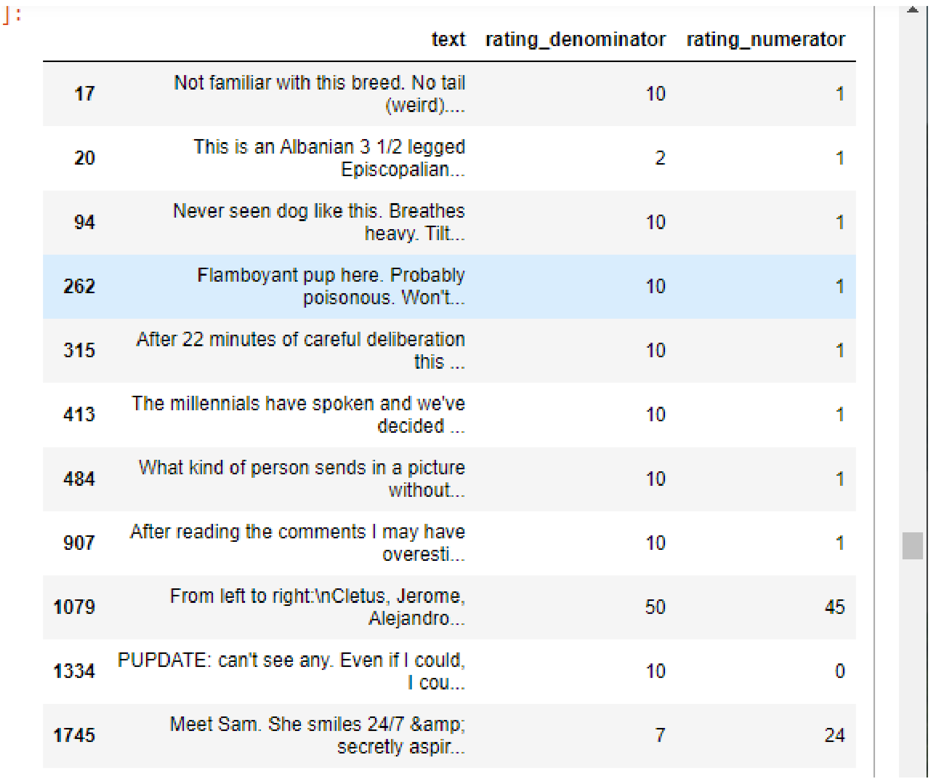
1) Maximum dogs are classified in pupper and minimum are in floofer. Total of 16.69% of dogs are classified others are not.



2) From graph I can conclude that the most number of rating is 12 and the other ratings from 420 to 204 on x axis are singly rated

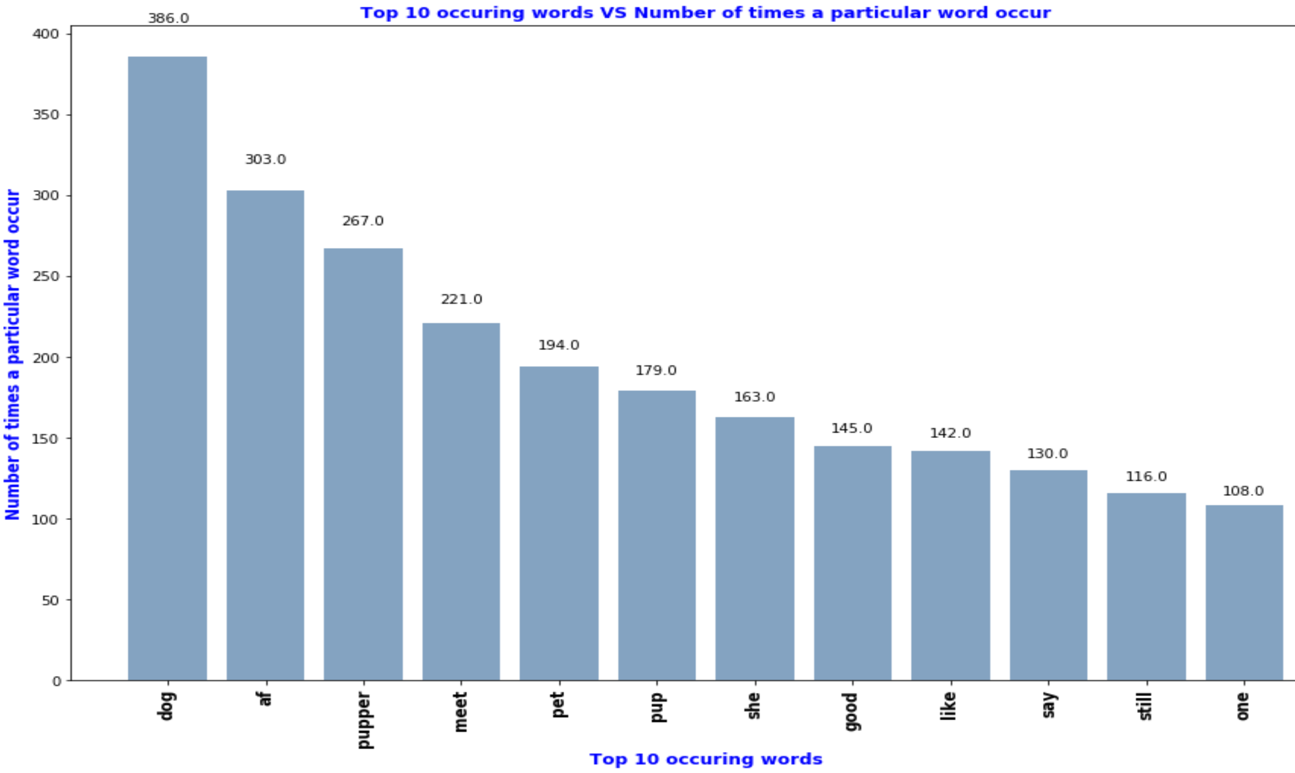


3) Many ratings are not clear and few tweets are done from two accounts .



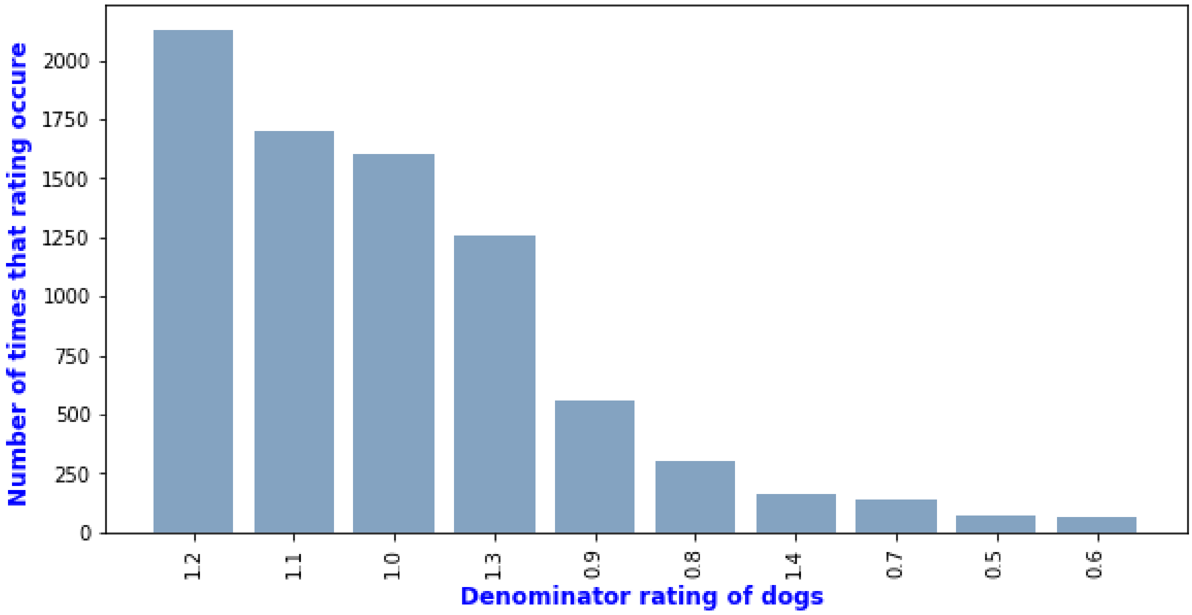
	text	rating_denominator	rating_numerator
17	Not familiar with this breed. No tail (weird)....	10	1
20	This is an Albanian 3 1/2 legged Episcopalian...	2	1
94	Never seen dog like this. Breathes heavy. Tilt...	10	1
262	Flamboyant pup here. Probably poisonous. Won't...	10	1
315	After 22 minutes of careful deliberation this ...	10	1
413	The millennials have spoken and we've decided ...	10	1
484	What kind of person sends in a picture without...	10	1
907	After reading the comments I may have overesti...	10	1
1079	From left to right:\nCletus, Jerome, Alejandro...	50	45
1334	PUPDATE: can't see any. Even if I could, I cou...	10	0
1745	Meet Sam. She smiles 24/7 &amp; secretly aspir...	7	24

4)



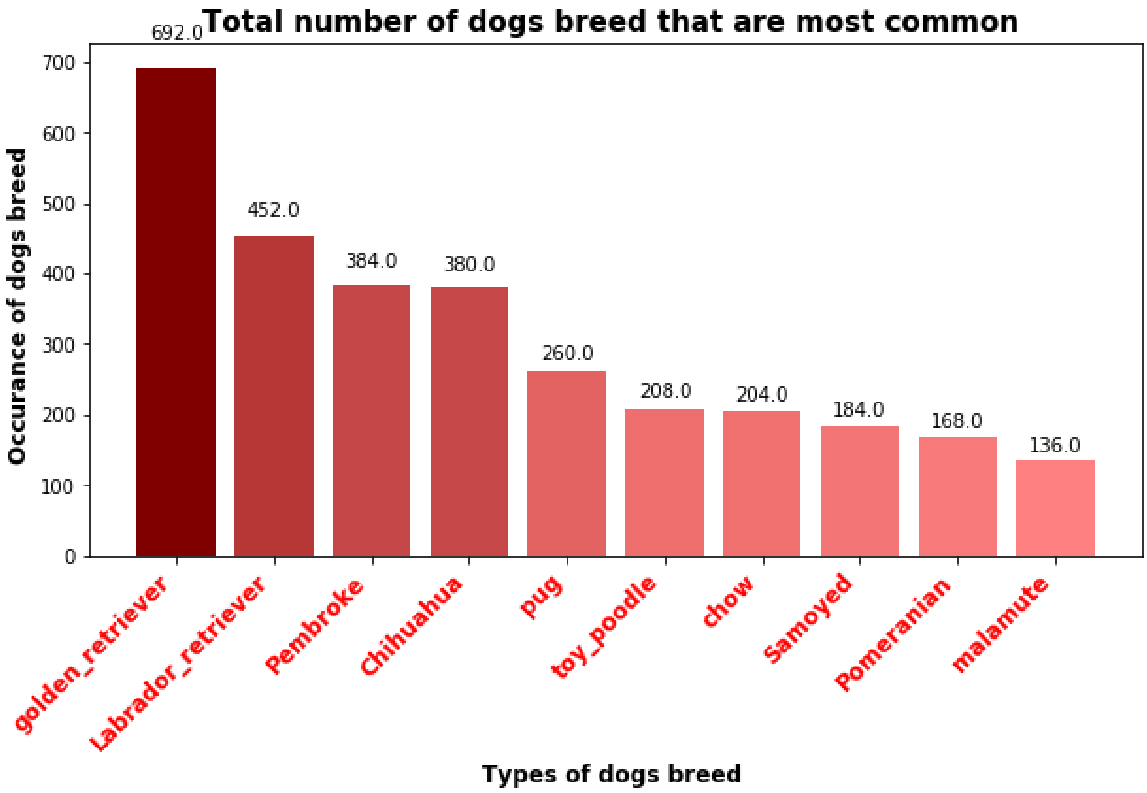
From the above analysis we can say that the most used words are pupper pup meet ect.

5)



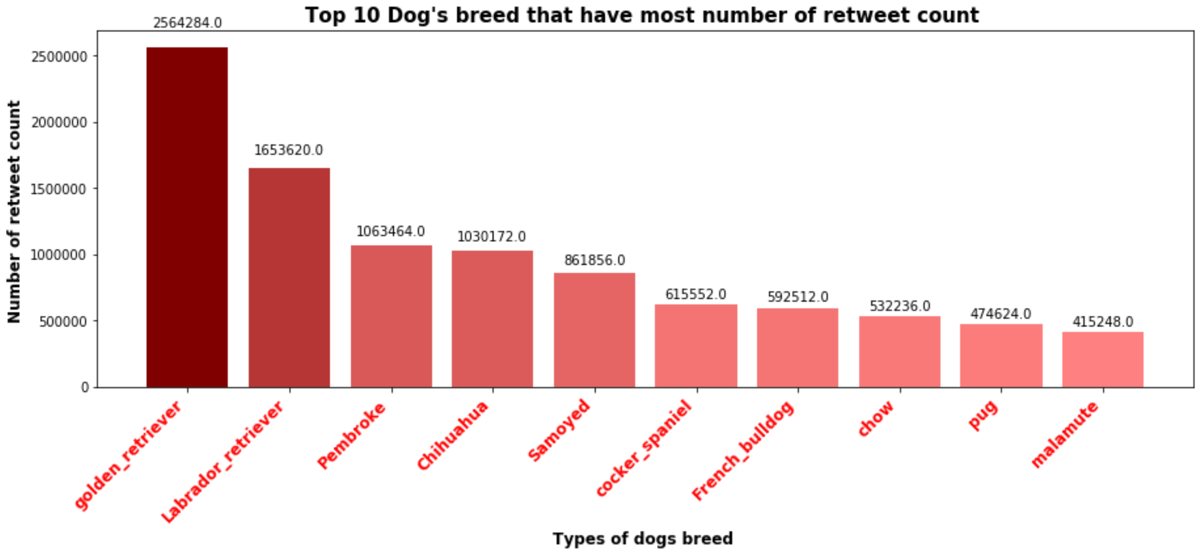
Most number of are rated 1.2 and 1.1 as we can see from the graph.

6)



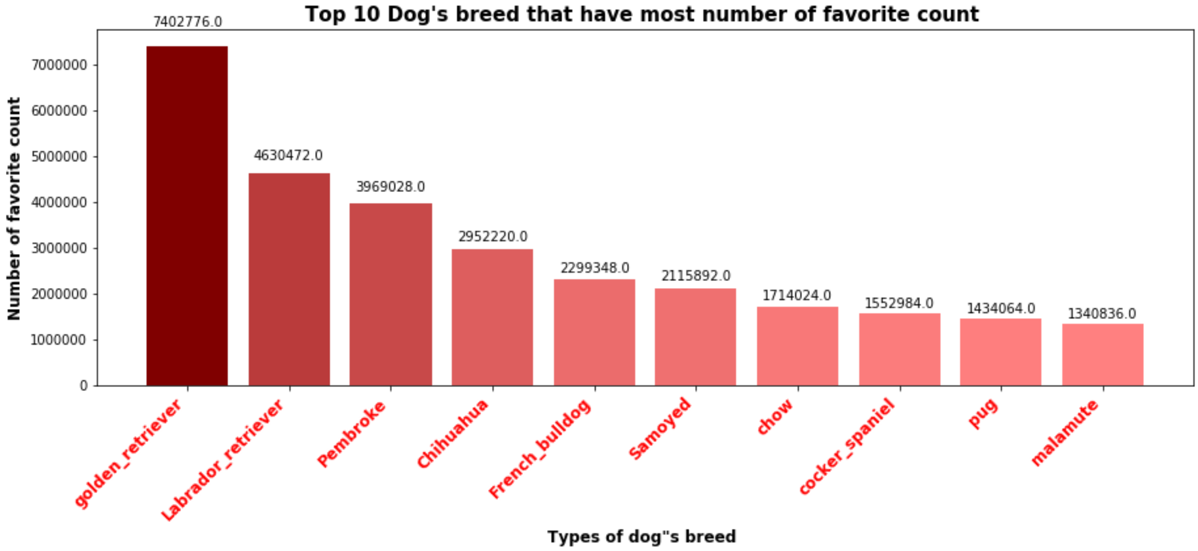
Maximum number of dog breed is for golden retriever

7)



Maximum number of retweets counts is for golden retriever

8)



Maximum number of favourite counts is for golden retriever