we rate dogs Wrangle report.

Gathering

for this project, I gathered data from 3 different sources. The first one was manually downloaded from the udacitys server, the second one was received getting a URL and programmatically reading that tsv file and the third one I got using API. Since Twitter API was not publicly open, I requested API keys and code to us tweeters API called tweepy. And then using the API codes in the file I manually downloaded from Udacty I got the retweet and favorite count.

Assessing&Cleaning

to assessing the 3 df's I both programmatically and visually tried to identify quality issues.

1)
Chage the rating numerator to a float and implement some code that the number is correctly represented

2)the tweet id was represented as strings in all three data frames using .dtypes().similarly timestamp is not is the data type datetime

3)All names starting with a lowercaps are invalid names and need to be replaced with none

4)in the expanded URL's the Html a tag was inclouded to remove that you can extract the Html code and as a result, you will receive four groups Twitter for iPhone, Vine , Twitter Web Client, TweetDeck.

5) some tweets had the wrong ratings as proven with the text and needed to be manually fixed
6)you can drop all rows with multible dogs in it since we are only intresseted in the single dog ones
7)some tweets were retweets which needed to be removed since we are only interested in 'original tweets'
8)drop all rows without the extended url’s

9)some tweets have unrealistivly high ratings so the two highest need to be droped

10)then I merged the json_df and archive data frame, using the tweet id
11)1 value (stage) wich includes doggo floofer pupper Puppo had their separate column and needed to be collapsed in one column.
12)in tidiness we don't need the rating denominator, so I dropped that column

13)to finish off we need all this in one data frame and we only need the bread and final confidence of that breed columns