Wrangling report

The dataset that wrangled (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.

The wrangling process is divided to three steps:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

1. gathering data
   1.1. gathering data from twitter-archive-enhanced.csv as t_archive contains data extracted programmatically from twitter data sent by WeRateDoge to Udacity via email exclusively to be used in this project the data provides the rating, dog name, and dog stage and some other information.

#Data gathered by this code

2.1. gathering data from twitter-archive-enhanced.csv as t_archive

```
l]: t_archive=pd.read_csv('twitter-archive-enhanced.csv')
```

the gathered data will be like this

```
In [123]: #exploring the first five rows from the data
          t_archive.head()
```

Out[123]:

_user_id	retweeted_status_timestamp	expanded_urls	rating_numerator	rating_denominator	name	doggo	floofer	pupper	puppo
NaN	NaN	https://twitter.com/dog_rates/status/892420643...	13	10	Phineas	None	None	None	None
NaN	NaN	https://twitter.com/dog_rates/status/892177421...	13	10	Tilly	None	None	None	None
NaN	NaN	https://twitter.com/dog_rates/status/891815181...	12	10	Archie	None	None	None	None
NaN	NaN	https://twitter.com/dog_rates/status/891689557...	13	10	Darla	None	None	None	None
NaN	NaN	https://twitter.com/dog_rates/status/891327558...	12	10	Franklin	None	None	None	None

1.2. gathering data from image-predictions.tsv as image_p
this file is hosted on Udacity servers and was downloaded programmatically using the requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

#Data gathered by this code

2.2. gathering data from image-predictions.tsv as image_p

In [72]:
```
url='https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'
response=requests.get(url)
with open('image-predictions.tsv',mode='wb') as file:
    file.write(response.content)
image_p=pd.read_csv('image-predictions.tsv',sep='\t')
```

the gathered data will be like this

In [79]:
```
#exploring random 10 rows from the data
image_p.sample(10)
```

Out[79]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog
96	667724302356258817	https://pbs.twimg.com/media/CUQ7tv3W4AA3KlI.jpg	1	ibex	0.619098	False	bighorn	0.125119	False
1751	824297048279236611	https://pbs.twimg.com/media/C3B9ypNWEAM1bVs.jpg	2	teddy	0.588230	False	jigsaw_puzzle	0.028910	False
1557	793226087023144960	https://pbs.twimg.com/media/CwIa5CjW8AErZgL.jpg	1	wire-haired_fox_terrier	0.456047	True	Lakeland_terrier	0.273428	True
302	671504605491109889	https://pbs.twimg.com/media/CVGp4LKWoAAoD03.jpg	1	toy_poodle	0.259115	True	bath_towel	0.177669	False
1669	813066809284972545	https://pbs.twimg.com/media/C0iX8OOVEAEIpMC.jpg	1	toy_terrier	0.776400	True	Pembroke	0.115034	True
1456	777641927919427584	https://pbs.twimg.com/media/CmoPdmHW8AAi8Bl.jpg	1	golden_retriever	0.964929	True	Labrador_retriever	0.011584	True
1684	814153002265309185	https://pbs.twimg.com/media/C0xz04SVIAAeyDb.jpg	1	golden_retriever	0.490068	True	Labrador_retriever	0.291956	True
1528	789137962068021249	https://pbs.twimg.com/media/CvOUw8vWYAAzJDq.jpg	2	Chihuahua	0.746135	True	Pekinese	0.070383	True
959	705591895322394625	https://pbs.twimg.com/media/CcrEFQdUcAA7CJf.jpg	1	basenji	0.877207	True	Italian_greyhound	0.047854	True
967	706310011488698368	https://pbs.twimg.com/media/Cc1RNHLW4AACG6H.jpg	1	Pembroke	0.698165	True	Chihuahua	0.105834	True

2. gathering data from json file and making a data frame using empty list
this data gathered using the file provided by Udacity

#the data gathered using this code

2.4. gathering data from json file and making a data frame using empty list

In [74]:
```
# a.3. gathering tweet data from json file
df_list=[]
with open ("tweet-json.txt")as file:
    for l in file:
        df_list.append(json.loads(l))

tweet_df=pd.DataFrame(df_list,columns=['id','retweet_count','favorite_count'])
tweet_df=tweet_df.rename(columns={'id':'tweet_id'})
tweet_df.to_csv('tweet_df.csv',index=False)
```

#the gathered data will be like this

```
In [72]: #exploring the first five rows from the data
         tweet_df.head()
```

Out[72]:

	tweet_id	retweet_count	favorite_count
0	892420643555336193	8853	39467
1	892177421306343426	6514	33819
2	891815181378084864	4328	25461
3	891689557279858688	8964	42908
4	891327558926688256	9774	41048

2. Assessing data

1. Tidiness

1. merge four columns (doggo & floofer & pupper & puppo) in one column dog_stage

2. all data sets is related but separated into 3 df.

2. Quality

1. t_archive

1. For dog stage, there are some rows with multiple dog stages
2. there are 181 retweeted_status_id to be removed
3. 23 rating_denominator not equal 10
4. convert tweet_id from int to object
5. Convert timestamp from object to datetime
6. remove columns ('retweeted_status_id' & 'retweeted_status_user_id')
7. from name column remove rows depend on None & a value

2. image_p

1. missing picture for some ID's "remove any ID without photo"
2. some P names starting with uppercase letter "convert all letter to be lowercases"
3. missing entries 2075 instead of 2356

3. tweet_df

1. missing entries 2354 instead of 2356

cleaning steps

all the cleaning steps Define >> Code >> Test are maintained

for example, the issue number 4 (removing rating denominator not equal to 10)

Define issue 4

removing rating_denominator not equal 10

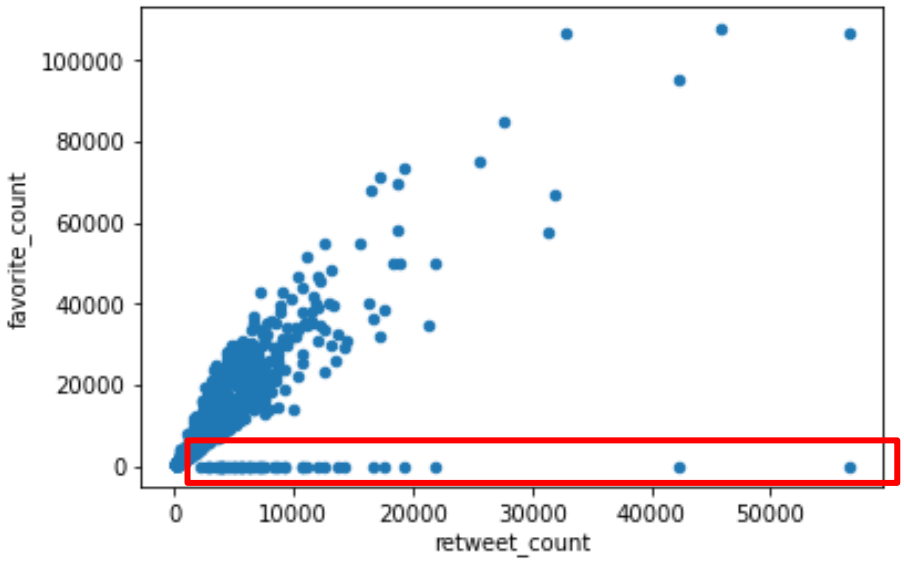
Code issue 4

```
In [149]: df=df[df['rating_denominator']==10]
```

Test issue 4

```
In [150]: df[df['rating_denominator']!=10].count()[0]
Out[150]: 0
```

After the visualization phase I found that we can clean the data more to make better visualization



I defined new data frame to has all the data except (favorite_count equal zero)

And the steps are

Define issue 11

Removing favorite count equal zero to make the visualization better

Code issue 11

```
n [151]: df_cleanded=df[df['favorite_count']!= 0]
```

Test issue 11

```
n [203]: import seaborn as sns
         sns.set_style('darkgrid')
         df_cleanded.plot(x='retweet_count',y='favorite_count',kind='scatter',figsize=(8,8));
         #scatter plot showing the forwerd relation between favorite_acounts and retweet_acounts
```