Wrangle Report

This report summarizes the project's data wrangling activities. The tweet archive of Twitter user @dog rates, also known as WeRateDogs was wrangled (and analyzed and visualized). This 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.

Google Colab Research Workspace was used to complete the entire project, and Google Docs was used to prepare and export the reports as PDFs.

There are three steps in the wrangling process:

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

Below is a more detailed explanation of each stage.

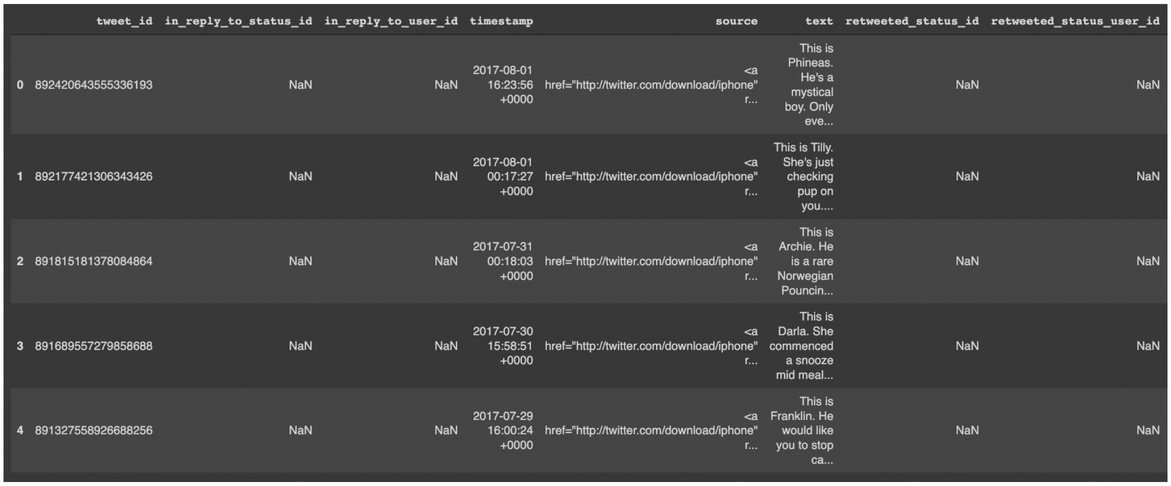
1. **Gathering Data**

   The data used was gathered from three different sources:

   a. **Enhanced Twitter Archive**

      Holds data that was programmatically retrieved from tweet data that WeRateDogs gave to Udacity through email just for this project.

      The information in the data includes the rating, the dog's name, its stage, and some other relevant details.



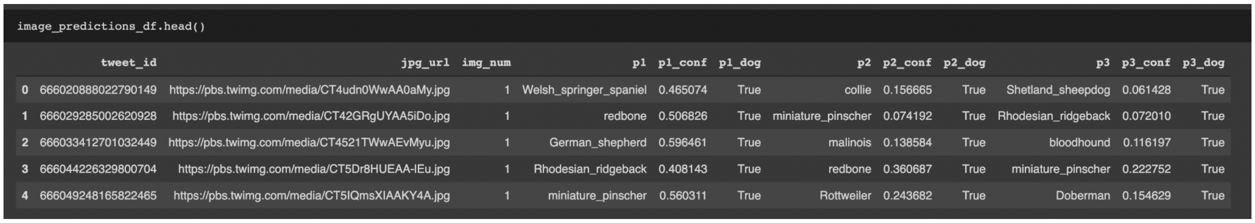
	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	retweeted_status_id	retweeted_status_user_id
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" r...	This is Phineas. He's a mystical boy. Only eve...	NaN	NaN
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you....	NaN	NaN
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" r...	This is Archie. He is a rare Norwegian Pouncin...	NaN	NaN
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http://twitter.com/download/iphone" r...	This is Darla. She commenced a snooze mid meal...	NaN	NaN
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="http://twitter.com/download/iphone" r...	This is Franklin. He would like you to stop ca...	NaN	NaN

Data which was obtained from the text of each tweet.

b. **Image Predictions File**

A neural network that recognizes dog breeds was used to process each photograph in the WeRateDogs Twitter collection to achieve the desired output. With each tweet ID, image URL, and image number that matched the most certain prediction, this algorithm produced a table full of image predictions (the top three only) (numbered 1 to 4 since tweets can have up to four images).

This file is hosted on Udacity's 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



```
image_predictions_df.head()
```

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074	True	collie	0.156665	True	Shetland_sheepdog	0.061428	True
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826	True	miniature_pinscher	0.074192	True	Rhodesian_ridgeback	0.072010	True
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461	True	malinois	0.138584	True	bloodhound	0.116197	True
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143	True	redbone	0.360687	True	miniature_pinscher	0.222752	True
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311	True	Rottweiler	0.243682	True	Doberman	0.154629	True

Tweet image prediction data

c. **Additional Data via Twitter API**

The instructions given was to extract this information through Twitter's API and then save it in a text file called tweet-json, but because I had trouble setting up my Twitter developer account, I had to manually download the tweet-json file provided by Udacity.

The ready-made version was used in this work and was read line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count, and was later saved to a 'tweet_data.csv' file for future use.

2. **Assessing Data**

Each of the aforementioned items of data was gathered, and after that, quality and tidiness issues were assessed visually and programmatically.

The following findings were concluded (Only the cleaned ones are mentioned):

a. **Tidiness:**

i. There are four columns of data for the dog stage.

ii. Despite being split into three different dataframes, all the data are related.

b. **Quality**:

i. Inaccurate tweet id (integer instead of string) and Invalid timestamp data type (String not DateTime).

ii. The Twitter enhanced archive data contains 181 retweets.

iii. There are 78 replies to tweets in the Twitter enhanced archive data.

iv. There are 440 Numerators less than 10.

v. Incorrect dog names like None, a, an, & the.

vi. Predicted photo data are not all complete (2075 instead of 2356).

vii. Instead of using spaces, the picture names use underscore in the p1, p2, and p3 columns.

viii. Inconsistent title case for p name.

3. Cleaning Data

A high-quality DataFrame was generated after the earlier issues were
cleaned as necessary.