

'WeRateDogs'



INTRODUCTION

'WeRateDogs' is a Twitter account which became popular for its smile-evoking comments and ratings of dog pictures

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 9 million followers and has received international media coverage.

This project is focused on data wrangling and analysis. However, visualisations were made and insights were gathered

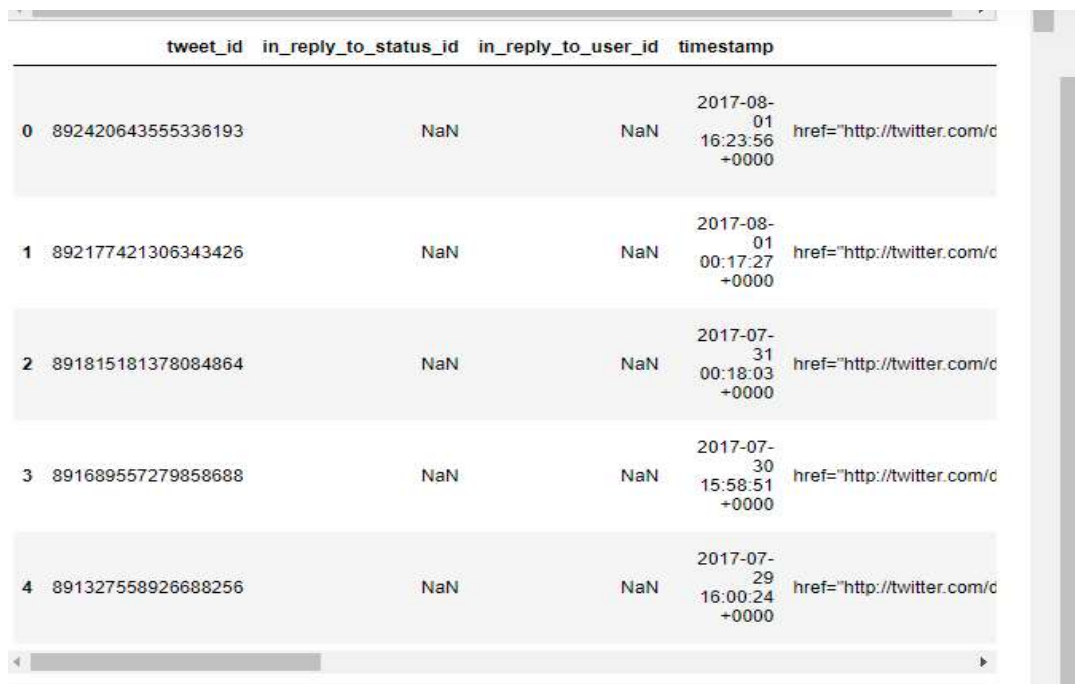
DATA GATHERING

Data was gathered from three data sources and read into dataframes

Data source 1:

'twitter_archive' file

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which was used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, they have been filtered for those tweets with ratings only (there are 2356). This is the 'twitter-archive-enhanced' file which is then read into the twitter_archive dataframe



	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp
0	89242064355336193	NaN	NaN	2017-08-01 16:23:56 +0000
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000

Data source 2:

'image-predictions.tsv'

The 'image_predictions.tsv' file contained data on image predictions of the dogs from the tweets. It was extracted from the given URL into a folder and then read into 'image_pred' dataframe

	tweet_id	jpg_url	img_num	
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	(
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-IEu.jpg	1	Rhc
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	r
5	666050758794694657	https://pbs.twimg.com/media/CT5Jof1WUAEuVxN.jpg	1	Berne
6	666051853826850816	https://pbs.twimg.com/media/CT5KoJ1WoAAJash.jpg	1	
7	666055525042405380	https://pbs.twimg.com/media/CT5N9tpXIAAifs1.jpg	1	
8	666057090499244032	https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg	1	
9	666058600524156928	https://pbs.twimg.com/media/CT5Qw94XAAA_2dP.jpg	1	
		https://pbs.twimg.com/media/CT5Vg_vwXIAAXfnj.jpg	1	

Data source 3:

'tweet-json.txt'

The given 'tweet-json.txt' file contained additional columns on retweet counts and favourite counts. It was read into a 'tweet' dataframe

	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

ASSESSING DATA

Visual and programmatic assessments were used to identify quality and tidiness issues

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	retweeted_status_id	retweeted_s
0	89242064355336193	NaN	NaN	2017-08-01 16:23:56 +0000	href="http://twitter.com/download/iphone" r...	This is Phineas. He's a mystical boy. Only eve...	NaN	
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you....	NaN	
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	href="http://twitter.com/download/iphone" r...	This is Archie. He is a rare Norwegian Pouncin...	NaN	
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	href="http://twitter.com/download/iphone" r...	This is Darla. She commenced a snooze mid meal...	NaN	
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	href="http://twitter.com/download/iphone" r...	This is Franklin. He would like you to stop ca...	NaN	

Identified issues:

Quality issues

1. Datatype issues: tweet_id column is stored as integer in 'twitter_archive', 'image_prediction' and 'tweet' dataframes also timestamp column in 'twitter_archive' is stored as string datatype
2. Wrong dog names, the wrong dog names are noted to all be in lowercase
3. Rows with replies and rows with retweets
4. Irrelevant columns: 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', 'retweeted_status_user_id' and 'retweeted_status_timestamp' contain replies and retweet and are not relevant to analysis
5. Some rating_denominators are not 10
6. Unclear column names description in image_pred dataframe
7. Some dog rating_numerators were wrongly extracted
8. From image_pred dataframe, the three image predictions (p1_dog, p2_dog and p3_dog) of some tweets were not dogs

Tidiness issues

1. Dog stage variable in four columns
2. All the data should be on one table

CLEANING DATA

The identified issues were addressed during the cleaning process

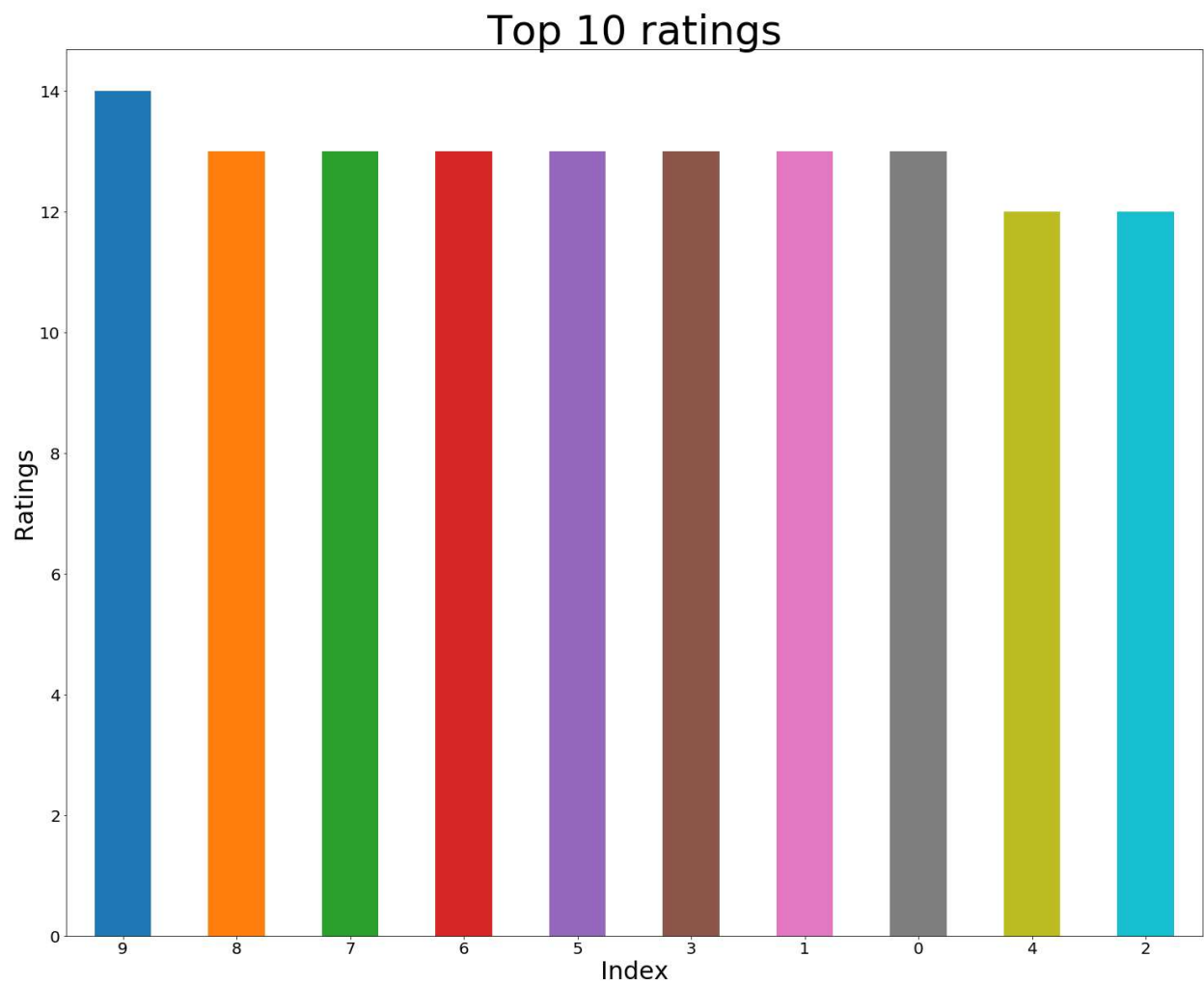
The Define-Code-Test cycle was used to clean each issue identified

SAVING DATA

The master dataset was saved into a csv file named 'twitter_archive_master.csv'

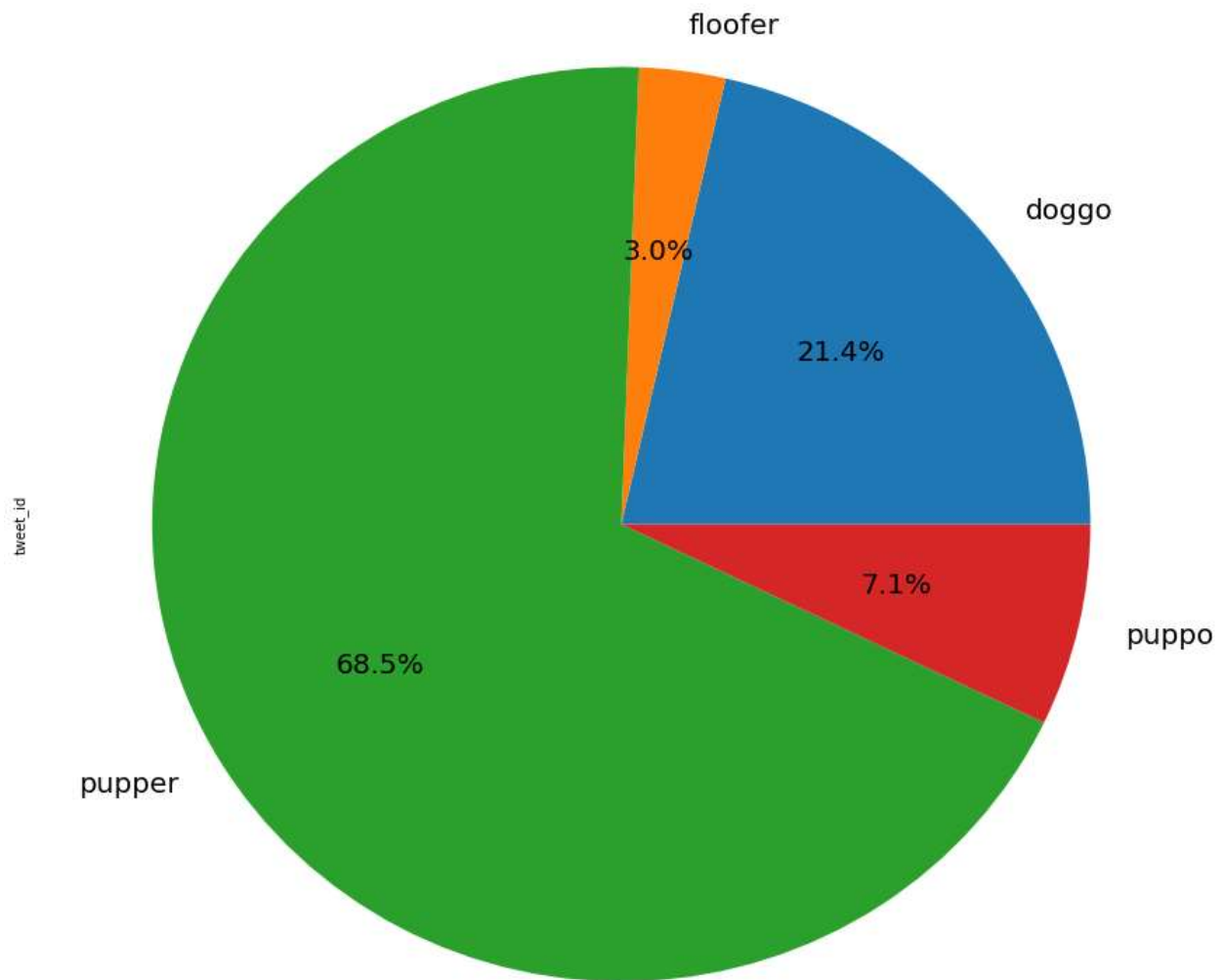
ANALYSIS AND VISUALISATIONS

The analysis and visualisations made are as below:



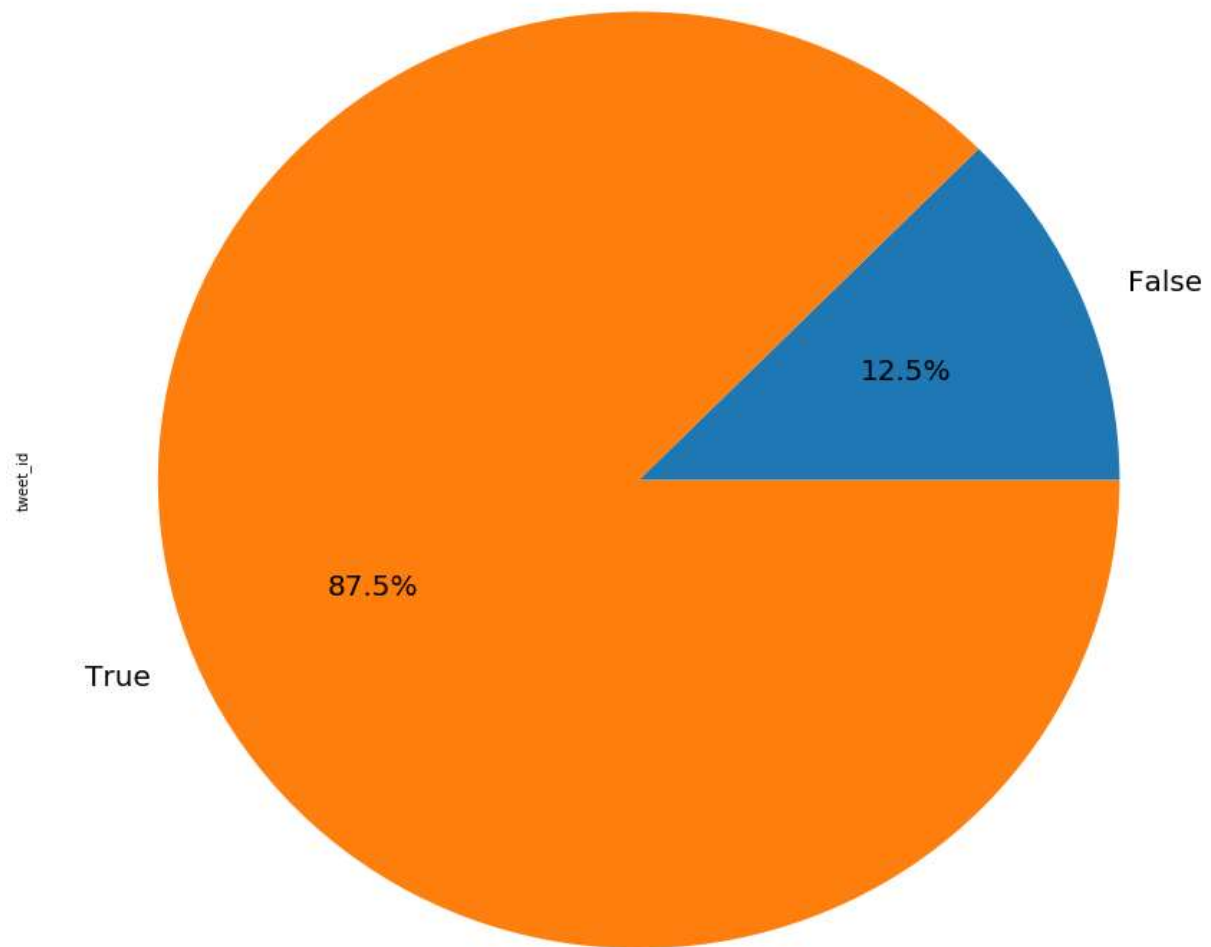
● This showed that the highest rating is 14/10

Dog stages



- Of the tweets with dog stages, the pupper stage constituted about two-thirds

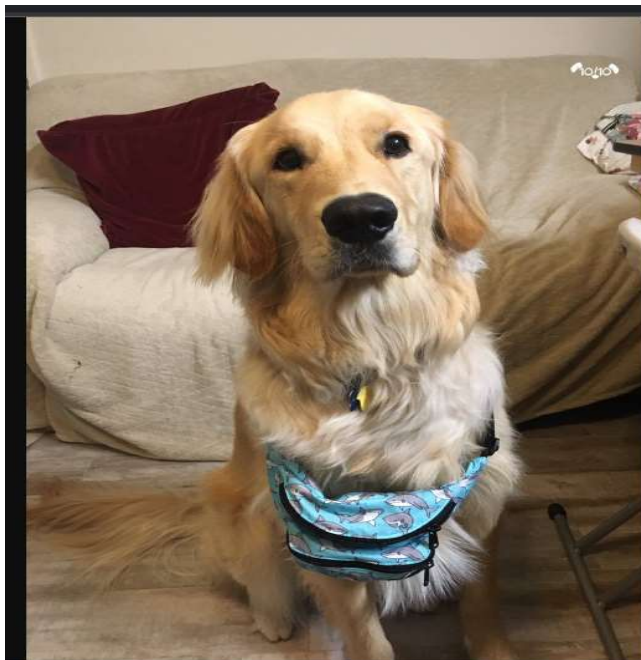
Image Predictions



- Over three-fourths of the first image predictions by the algorithm turned out to be dogs



My name is Obie. I have the highest favorite count



Hello, my name is Stuart. I'm a golden retriever

INSIGHTS

Insights:

- The highest value of rating is 14
- Only 336 of the 2356 tweets had dog_stages
- Of the tweets that had stages, pupper was the most, consisting about two thirds of the total
- Of the tweets with Dog names, Obie's tweet seemed to have the highest number of retweets and favorites count
- For the image predictions, over three-fourths of the first image predictions by the algorithm turned out to be dogs
- Out of the correct first predictions, 'golden_retriever' is the most predicted dog breed

CONCLUSION

The 'WeRateDogs' tweets served as a good data source to practice data wrangling and analysis

The most tweeted dog stage are those in the 'pupper' stage

The algorithm used to predict the dog types was had about 87.5% correct dog prediction rate for the first prediction

In-depth assessment of the dataframes is needed to identify all the possible issues and properly wrangle and analyse the data for better insight.