A Data Analytics Project Udacity

We Rate Dogs Analysis

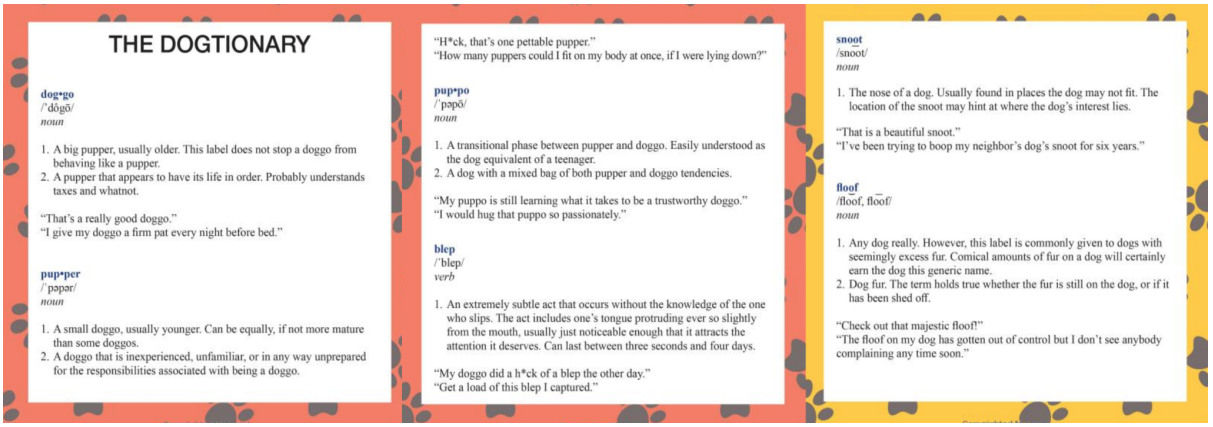
Overview:

We Rate Dogs:

WeRate Dogs is a twitter account with about 9.2M followers(at the time of publishing the article) that rates people's dogs with humorous comments about the dog. I analysed the data contained in this twitter account to get some insights to some questions like what are the popular names owners give their dogs.

This project is done to improve my DataWrangling skills. In carrying out this, I gathered three datasets for the analysis. Each dataset gathered with a different method. The first Dataset was gathered by downloading it directly, the second, by downloading programmatically and the third would be extracted from twitter, some terms were used in the article such as:

1. Doggo
2. Floofer
3. Pupper
4. puppo



GATHERING:

The separate files were gathered, they include:

1. **Twitter archive file**: download this file manually by clicking the following link: twitter_archive_enhanced.csv, this dataset was downloaded directly.

text	rating_ numerator	rating_ denominator	name	doggo	floofer	pupper	puppo
This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU	13	10	Phineas	None	None	None	None
This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's available for pats, snugs, boops, the whole bit. 13/10	13	10	Tilly	None	None	None	None
This is Archie. He is a rare Norwegian Pouncing Corgo. Lives in the tall grass. You never know when one may strike. 12/10 https://t.co	12	10	Archie	None	None	None	None
This is Darla. She commenced a snooze mid meal. 13/10 happens to the best of us https://t.co/tD36da7qLQ	13	10	Darla	None	None	None	None
This is Franklin. He would like you to stop calling him "cute." He is a very fierce shark and should be respected as such. 12/10 #BarkW	12	10	Franklin	None	None	None	None
Here we have a majestic great white breaching off South Africa's coast. Absolutely h*ckin breathtaking. 13/10 (IG: tucker_marlo) #Bar	13	10	None	None	None	None	None
Meet Jax. He enjoys ice cream so much he gets nervous around it. 13/10 help Jax enjoy more things by clicking below https://t.co/Zr4hWfAs1H https://t.co/tVJBRMnhxl	13	10	Jax	None	None	None	None
When you watch your owner call another dog a good boy but then they turn back to you and say you're a great boy. 13/10 https://t.co/	13	10	None	None	None	None	None
This is Zoey. She doesn't want to be one of the scary sharks. Just wants to be a snuggly pettable boatpet. 13/10 #BarkWeek https://t.c	13	10	Zoey	None	None	None	None
This is Cassie. She is a college pup. Studying international doggo communication and stick theory. 14/10 so elegant much sophisticat	14	10	Cassie	doggo	None	None	None
This is Koda. He is a South Australian deckshark. Deceptively deadly. Frighteningly majestic. 13/10 would risk a petting #BarkWeek ht	13	10	Koda	None	None	None	None
This is Bruno. He is a service shark. Only gets out of the water to assist you. 13/10 terrifyingly good boy https://t.co/u1XPQMl29g	13	10	Bruno	None	None	None	None
Here's a puppo that seems to be on the fence about something haha no but seriously someone help her. 13/10 https://t.co/BxvuXk0UC	13	10	None	None	None	None	puppo
This is Ted. He does his best. Sometimes that's not enough. But it's ok. 12/10 would assist https://t.co/f8dEDcrKSR	12	10	Ted	None	None	None	None
This is Stuart. He's sporting his favorite fanny pack. Secretly filled with bones only. 13/10 puppared puppo #BarkWeek https://t.co/y70	13	10	Stuart	None	None	None	puppo

2. **The tweet image predictions**: WeRateDogs Twitter archive was ran through a neural network that can classify breeds of dogs*. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 3 since tweets can have up to three images).i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network, This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_dog	p3	p3_conf	p3_dog
892177421306343426	https://pbs.twimg.cor	1	Chihuahua	0.323581	TRUE	Pekinese	0.0906465	TRUE	papillon	0.0689569	TRUE
891815181378084864	https://pbs.twimg.cor	1	Chihuahua	0.716012	TRUE	malamute	0.078253	TRUE	kelpie	0.0313789	TRUE
891689557279858688	https://pbs.twimg.cor	1	paper_towel	0.170278	FALSE	Labrador_retriever	0.168086	TRUE	spatula	0.0408359	FALSE
891327558926688256	https://pbs.twimg.cor	2	basset	0.555712	TRUE	English_springer	0.22577	TRUE	German_short-haired_pointer	0.175219	TRUE
891087950875897856	https://pbs.twimg.cor	1	Chesapeake_Bay_retriever	0.425595	TRUE	Irish_terrier	0.116317	TRUE	Indian_elephant	0.0769022	FALSE
890971913173991426	https://pbs.twimg.cor	1	Appenzeller	0.341703	TRUE	Border_collie	0.199287	TRUE	ice_lolly	0.193548	FALSE
890729181411237888	https://pbs.twimg.cor	2	Pomeranian	0.566142	TRUE	Eskimo_dog	0.178406	TRUE	Pembroke	0.0765069	TRUE
890609185150312448	https://pbs.twimg.cor	1	Irish_terrier	0.487574	TRUE	Irish_setter	0.193054	TRUE	Chesapeake_Bay_retriever	0.118184	TRUE
890240255349198849	https://pbs.twimg.cor	1	Pembroke	0.511319	TRUE	Cardigan	0.451038	TRUE	Chihuahua	0.0292482	TRUE
890006608113172480	https://pbs.twimg.cor	1	Samoyed	0.957979	TRUE	Pomeranian	0.0138835	TRUE	chow	0.00816748	TRUE
889880896479866881	https://pbs.twimg.cor	1	French_bulldog	0.377417	TRUE	Labrador_retriever	0.151317	TRUE	muzzle	0.0829811	FALSE
889665388333682689	https://pbs.twimg.cor	1	Pembroke	0.966327	TRUE	Cardigan	0.0273557	TRUE	basenji	0.00463323	TRUE
889638837579907072	https://pbs.twimg.cor	1	French_bulldog	0.99165	TRUE	boxer	0.00212864	TRUE	Staffordshire_bullterrier	0.00149818	TRUE
889531135344209921	https://pbs.twimg.cor	1	golden_retriever	0.953442	TRUE	Labrador_retriever	0.0138341	TRUE	redbone	0.00795775	TRUE

3. **Twitter API & JSON**: Each tweet's retweet count and favorite ("like") count at the minimum and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

QUALITY ISSUES:

1. Erroneous data type: tweet_id should be a string to integer datatype, timestamp should be in datetime not string datatype.

2. You only want original ratings (no retweets) that have images

3. Columns not needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp.

4. Extractions possible: Gender(M or F) of dogs could be extracted from text,(phone, vine, tweet_deck) could be extracted from source.

5. Missing values: the expanded_urls is 2297 instead of 2356.

6. invalid data entry: all rating_numerator should be => 10, all rating_denominator should be = 10.

7. Lowercase: all values in name column should be in lowercase

8. inaccurate values: values like a,an,by,none,just,his,old,the,actually etc are seen in the name column.

9. Column name/value change: should be floof instead of floofer.

10. None & Null values should be represented as Nan in all column especially name, floof, doggo, pupper, puppo columns

11. Erroneous data type: tweet_id should be a string to integer datatype.

12. Descriptive column names: jpg_url should be image_url, img_num should be image_number,p1 should be prediction_1& p1_conf should be prediction_confidence_1, p2 should be prediction_2 &

p2_conf should be prediction_confidence_2, p3 should be prediction_3 & p3_conf should be prediction_confidence_3etc...

13. Missing records: 2075 instead of 2356

14. Descriptive column name: id should be tweet_id

15. Erroneous data type: tweet_id should be a string to integer datatype.

16. Missing records: 2354 instead of 2356.

TIDINESS:

1. All three dataframes can be combined into one single dataframe.

2. doggo, floof, pupper, puppo should be in 1 column instead of 4 column.

INSIGHT:

1. Which is most source of the tweets?

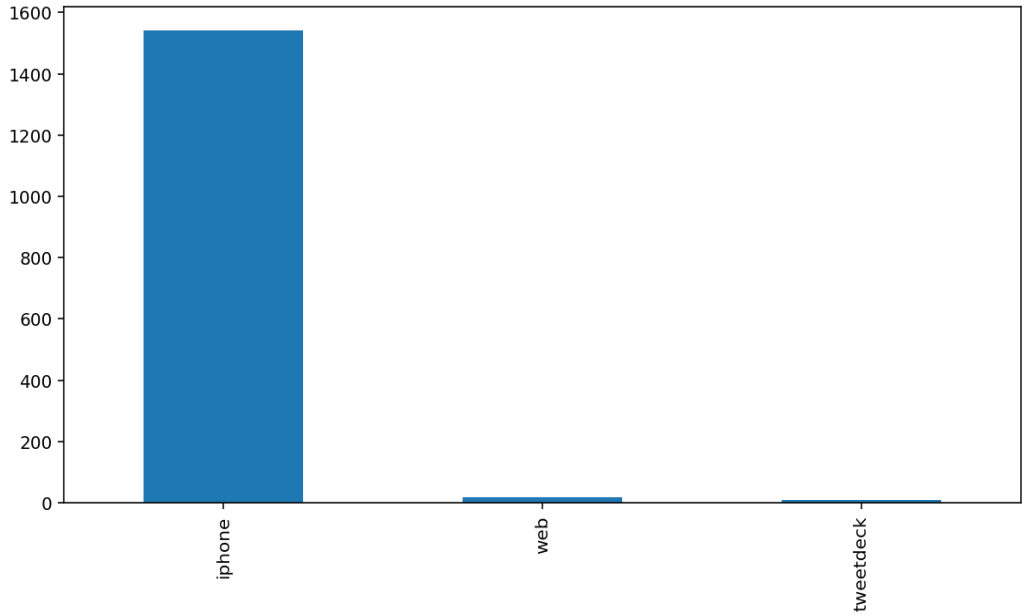
2. What are the popular names of dogs given?

3. Which is the most common dog stage?

4. Which gender is most frequently seen in the dataset

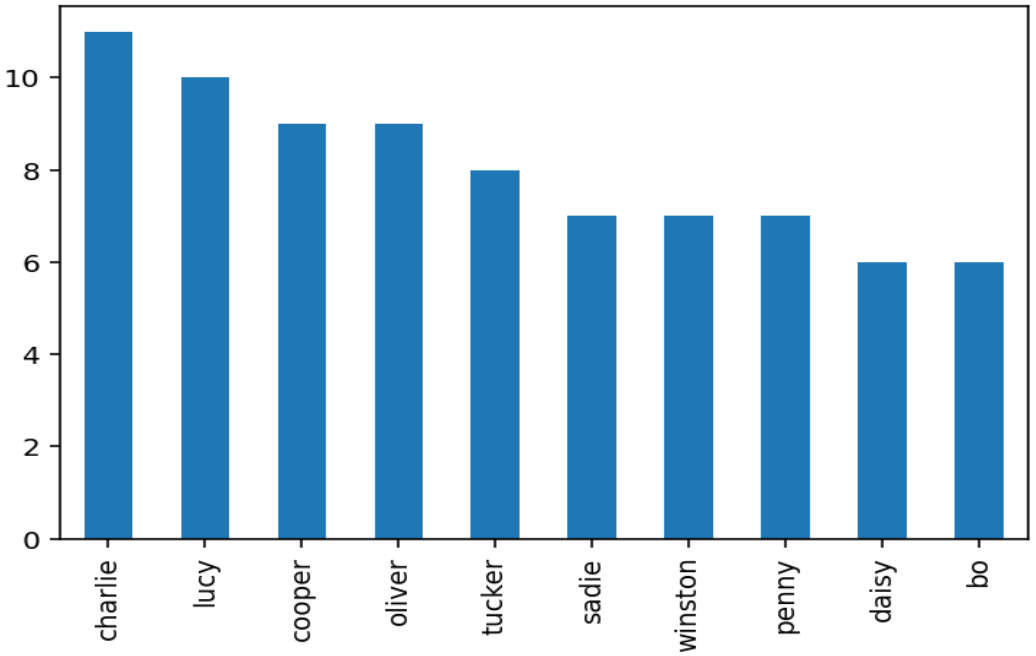
VISUALISATION:

- *What are the popular names of dogs given?*



Iphone is the most common source of tweets.

- What are the popular names of dogs given?



Charlie, Lucy, Cooper, Oliver, Tucker are the top five most popular names given by dog owners.