DATA ANALYSIS AND VISUALIZATION REPORT ON WERATEDOGS

By: Timothy Yaji

Contents

Introduction

Real-world data rarely come clean. Using Python and its libraries, i will gathered data from various sources and in a variety of formats, assess its quality and tidiness, then clean it this is called data wrangling.

The dataset that I wrangled 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. Details here

Software Used for Wrangling

The entirety of this wrangling project was in Jupyter Notebook with the following packages (libraries) installed: pandas, NumPy, requests, tweepy ,JSON, os, and seaborn.

Step1: Data Gathering

In this project we used three different pieces of data from various sources using different methods to gather each data for the purpose of wrangling.

- **The WeRateDogs Twitter archive** I am given this file in my Udacity work space. I downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv. Once it was downloaded and saved in C:/Users/USER/3D Objects/Udacity_Project_02/wrangling_analyzing_data/twitter-archive-enhanced.csv, i upload it and read the data into a pandas DataFrame.

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

- **The tweet image predictions** This file (image_predictions.tsv) is hosted on Udacity's servers and i 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

- **Additional data from the Twitter API** I gather each tweet's retweet count and favorite ("like") count a Using the tweet IDs in the WeRateDogs Twitter archive, I 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 was then written to its own line and read as .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

	tweet_id	retweet_count	favorite_count
0	892420643555336193	6969	33696
1	892177421306343426	5272	29223
2	891815181378084864	3464	21977
3	891689557279858688	7191	36788
4	891327558926688256	7717	35183
5	891087950875897856	2586	17749
6	890971913173991426	1647	10331

Step 2: Assessing Data

After gathering all three pieces of data. I assessed the data using the following two types of assessment: **Visual assessment:** Each piece of gathered data was displayed in the Jupyter Notebook for visual assessment purposes. This was achieved by setting the display limits in pandas DataFrame (pd.set_option("display.max_colwidth",150) and pd.options.display.max_rows = 3000) **Programmatic assessment:** I used pandas' functions and/or methods to assess the data.

In order to meet project specifications, I made sure that:

- Only original ratings (no retweets) that have images were retained.
- I gathered the tweets and image predictions for these tweets up to August 1st, 2017 only.
- I identified 8 quality and 5 tidiness issues which I resolved them.

Step 3: Cleaning Data

In this section, we will clean all the issues documented while assessing. This also includes merging individual pieces of data according to the rules of tidy data.

During the cleaning process i used the defined Define-Code-Test Framework to clearly document the process. The table below is the transposed version of the cleaned data.

tweet_id	676776431406465024	829449946868879360
tweet_date	2015-12-15 14:50:49+00:00	2017-02-08 22:00:52+00:00
rating_numerator	10	11
rating_denominator	10	10
dog_name	None	None
tweet_source	Twitter for iPhone	Twitter for iPhone
tweets_text	When someone yells "cops!" at a party and you ...	Here's a stressed doggo. Had a long day. Many ...
tweet_url	https://t.co/4rMZi5Ca1k	https://t.co/fmRS43mWQB
image_url	https://pbs.twimg.com/ext_tw_video_thumb/67677...	https://pbs.twimg.com/media/C4LMUf8WYAkWz4I.jpg
image_number	1	1
dog_breed	Unknown Breed	Labrador Retriever
confidence_probabilty	0	0.315163
retweets	1796	1854
likes	4504	9846
dog_stage	NaN	Doggo

Step 4: Analysis and Visualization

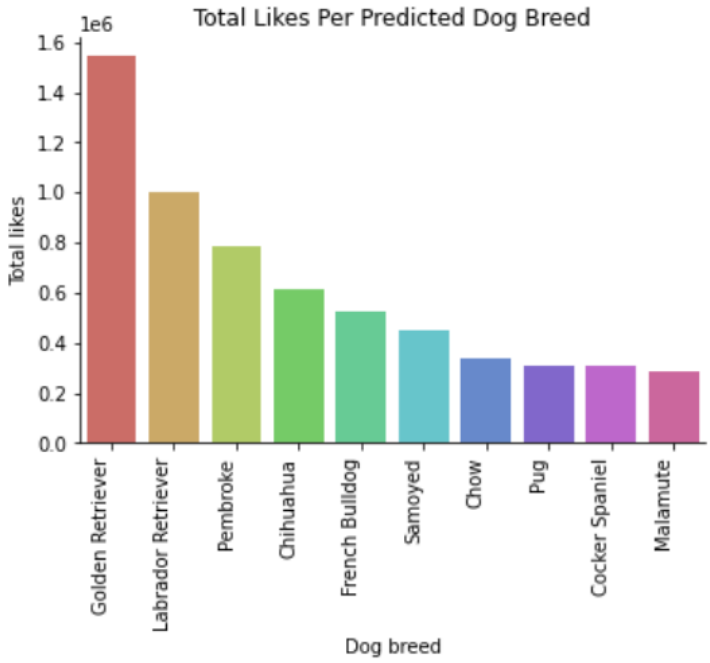
Research question1:

What are the top 10 dog breed that are most likes?

	Dog breed	Total likes
0	Golden Retriever	1543006
1	Labrador Retriever	998092
2	Pembroke	785309
3	Chihuahua	610428
4	French Bulldog	522028
5	Samoyed	449760
6	Chow	339040
7	Pug	311738
8	Cocker Spaniel	309705
9	Malamute	283122

The table above shows the top ten dog predicted breeds.

Visual



Insight 1:

- The **Golden retriever** is the most popular breed, in terms of total likes (1351679). The **Labrador Retriever Pembroke** follows behind with 916617 likes. **Pembroke** occupied the third position. Other notable breeds include the Chihuahua, Samoyed, Frech Bulldog, Pug, Malamute, Chow and cardigan in that order**.**

Conclusion:

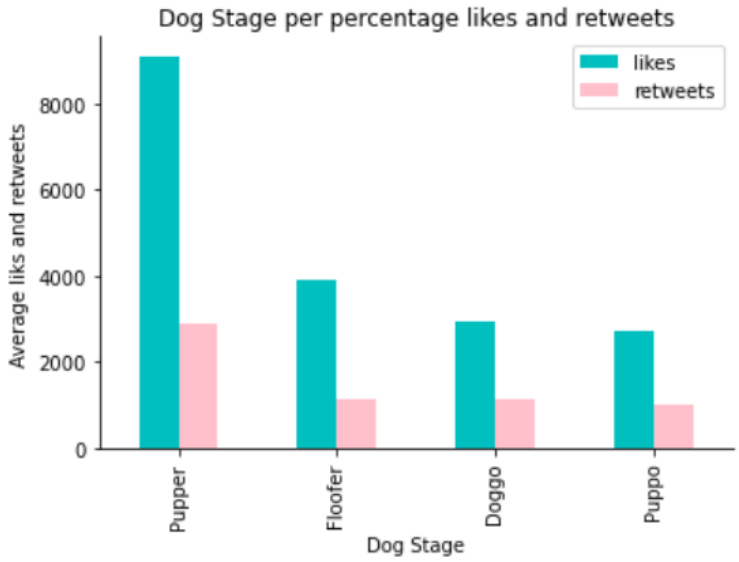
The **Golden retriever** is the most popular breed, in terms of total likes

Research question2:

What dog stage has the most likes and retweets

dog_stage	likes	retweets
Pupper	9098.0	2875.0
Floofer	3909.0	1116.0
Doggo	2930.0	1117.0
Puppo	2707.0	983.0

Visuals



Insight 2:

- Puppers are the people's favorite and with more engagements leading with average likes of 1998 and 2875 retweets. Followed by Floofer in terms of likes But Doggo tweets showed edged Floofer slightly in terms of user engagements on average with (1117) against Floofer(1116). Pupp0 comes last both on average likes and retweets.

Conclusion:

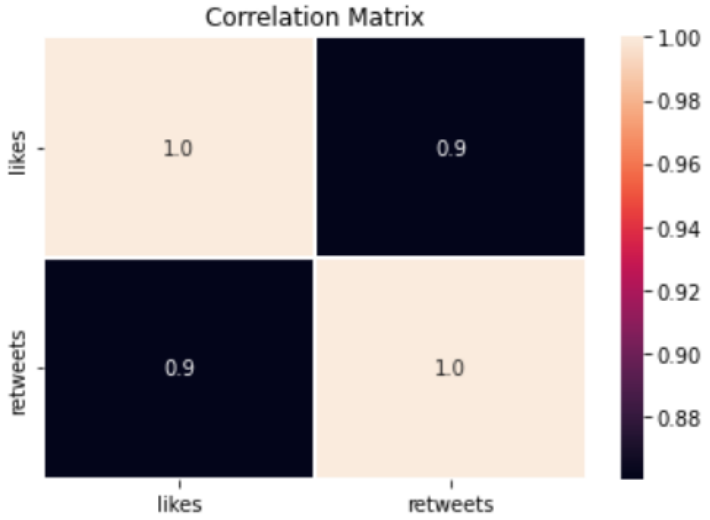
Puppers are the people's favorite and attracts more user engagement than the other dog stages.

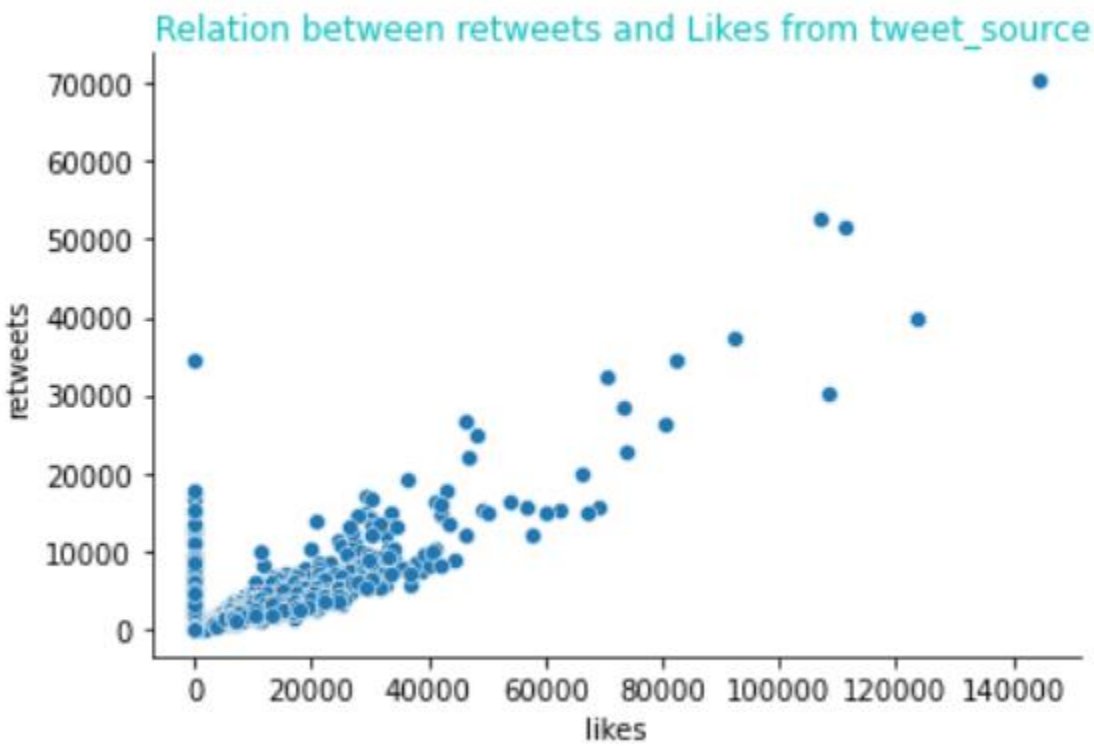
Research question 3:

Is there any relationship between likes and retweets?

	likes	retweets
likes	1.00000	0.86037
retweets	0.86037	1.00000

Visuals:





Insight 3:

Retweets and likes show **strong positive correlation** of **0.92** as evidenced by the correlation matrix plot and the scatter plot.

Conclusion:

There is a positive relationship between retweets and likes for all tweets.