Amanda Zhou
August 28th, 2020

Project: Wrangle and Analyze Data

(Internal Document)

Introduction

Real-world data rarely comes clean. Using Python and its libraries, I will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it.

The dataset that I will be wrangling (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.

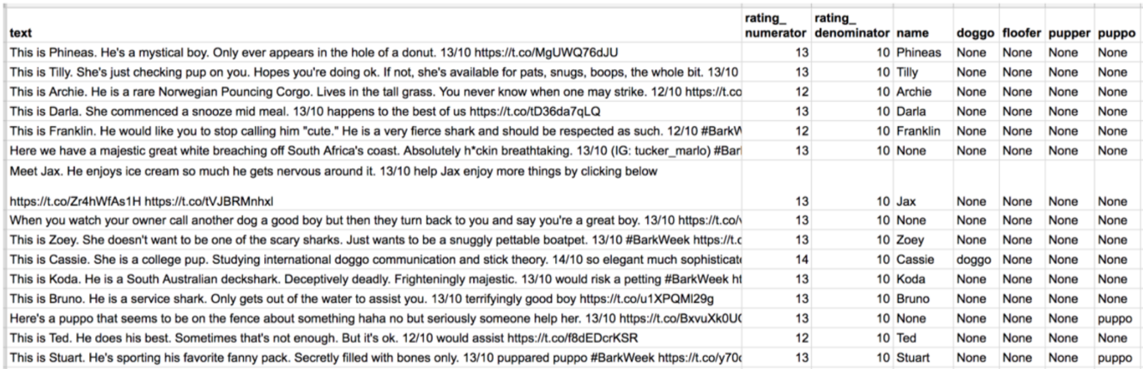
WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for me to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

My goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required.

Gather Data

1. **The WeRateDogs Twitter Archive File**
   1) Download “twitter_archive_enhanced.csv” file manually by clicking the given link from Udacity, then upload this csv file to Jupyter Notebook workspace.

   2) Information about this 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 were filtered for tweets with ratings only (there are 2356). The ratings probably aren't all correct. Same goes for the dog names and probably dog stages too. I might need to assess and clean these columns if I use them for analysis and visualization.



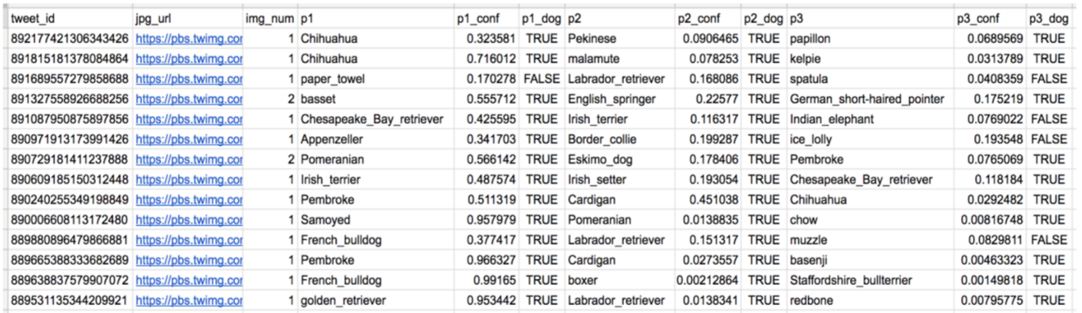
text	rating_ numerator	rating_ denominator	name	doggo	floofer	pupper	puppo
This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU	13	10	Phineas	None	None	None	None
This is Tilly. She's just checking pup on you. Hopes you're doing ok. If not, she's 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/BxvuXk0U(	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 extracted data from each tweet's text

2. The Image Predictions File

1) This file (image_predictions.tsv) is hosted on Udacity's servers and will be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

2) Information about this file: this file is about what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network.



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

Tweet image prediction data

So for the last row in that table:

- tweet_id is the last part of the tweet URL after "*status/*" → https://twitter.com/dog_rates/status/889531135344209921
- p1 is the algorithm's #1 prediction for the image in the tweet → **golden retriever**
- p1_conf is how confident the algorithm is in its #1 prediction → **95%**
- p1_dog is whether or not the #1 prediction is a breed of dog → **TRUE**
- p2 is the algorithm's second most likely prediction → **Labrador retriever**
- p2_conf is how confident the algorithm is in its #2 prediction → **1%**
- p2_dog is whether or not the #2 prediction is a breed of dog → **TRUE**
- etc.

3. **The Twitter API File**
   1) I'll be using Tweepy to query Twitter's API for additional data beyond the data included in the WeRateDogs Twitter archive. After querying each tweet ID, I will write its JSON data to a tweet_json.txt file with each tweet's JSON data on its own line. I will then read this file, line by line, to create a pandas DataFrame that I will assess and clean.

   2) Information about this file: This additional data will include retweet count and favorite count.

Assess Data

1. **Quality**
   - ***df_archive table*** (twitter_archive_enhanced.csv)
     1) Variable tweet_id is an integer not a string
     2) Variables in_reply_to_status_id and in_reply_to_user_id are float not string
     3) Variables timestamp and retweeted_status_timestamp are objects not datetime
     4) Variables in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp and expanded_urls are missing data
     5) Variable source contains HTML code
     6) This dataset contains retweets, so it has duplicated data
     7) Names are extracted inaccurately, such as "a", "an", “the”, “very”, etc. Looking at their corresponding texts from Excel file, the pattern is that these incorrect names are lowercase and followed by “This is…” or “Here is…”. Otherwise, the output is “None”
     8) Values in columns doggo, floofer, pupper and puppo are using "None" instead of "NaN"
     9) Many dogs have not been classified their stages
     10) There is a large proportion that rating numerator is greater than its denominator, which means dogs were given a rating of 100% and above. That is a unique rating system which is a big part of the popularity of WeRateDogs. However, for those with extreme numerator or denominator, we can make some adjustment for better analysis and consistency.

   - ***df_image table*** (image-predictions.tsv)
     1) Missing tweets compared to the 2356 tweets in the df_archive table
     2) Variable tweet_id is an integer not a string
     3) For consistency, should standardise words in columns p1, p2 and p3 since some are lowercase and others are capitalized
     4) Should use space instead of underscore in columns p1, p2 and p3
     5) Picture contains the item is not a dog

- ***df_api table*** (tweet-json.txt)
  1) Missing tweets compared to the 2356 tweets in the df_archive table
  2) Variable tweet_id is an integer not a string

2. Tidiness

- ***df_archive table*** (twitter_archive_enhanced.csv)
  1) A variable Dog Stage, which includes doggo, floofer, pupper and puppo, can be written in one column instead of four

- ***df_image table*** (image-predictions.tsv)
  1) Image predictions should be combined with df_archive table since their information are related to the same tweet
  2) Prediction of the dog breed can be obtained based on p1_dog, p2_dog and p3_dog columns, and their level of confidence columns

- ***df_api table*** (tweet-json.txt)
  1) Variables retweet_count and favorite_count should be part of df_archive table since information is extracted from same tweet

Clean Data

1. Tidiness 1: Combine 4 columns (doggo, floofer, pupper and puppo) into 1 column as dog_stage
2. Quality 1: Replace "None" with "NaN" for new column dog_stage
3. Tidiness 2: Combine 3 DataFrames with merge
4. Tidiness 3: Create dog breed column
5. Quality 2: Replace underscore in new dog breed column
6. Quality 3: Capitalize each word in new dog breed column
7. Quality 4: Change datatypes
8. Quality 5: Remove retweets
9. Quality 6: Missing and Unnecessary Data
10. Quality 7: Remove HTML code
11. Quality 8: Replace inaccurate name
12. Quality 9: Standardize rating
13. Store the clean DataFrame(s) in a CSV file named as twitter_archive_master.csv

References

1. https://stackoverflow.com/questions/16476924/how-to-iterate-over-rows-in-a-dataframe-in-pandas
2. https://www.kite.com/python/answers/how-to-replace-spaces-with-underscores-in-python#:~:text=Use%20str.,spaces%20with%20underscores%20in%20str%20.
3. https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.loc.html
4. https://stackoverflow.com/questions/23668427/pandas-three-way-joining-multiple-dataframes-on-columns
5. https://stackoverflow.com/questions/27881366/regular-expressions-and
6. https://stackoverflow.com/questions/14463277/how-to-disable-python-warnings