WeRateDogs – Insights into Twitter page

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment. Has over 8.8 million followers and has received international media coverage.

This project focus on wrangle process, start with gather from different sources and format, assessing by note and documents data issues(quality, tidiness), end wrangle by clean these issues. “Data scientists spend 60% of their time on cleaning and organizing data.” By Gil Press, Forbes post

After data wrangle there are visualizations and observations from the analysis as well.

Gather Data

This project gathered data from the following sources:

- **Enhanced Twitter Archive**. The WeRateDogs downloaded their Twitter archive and sent it to Udacity via email to use in this project. Twitter archive file format in csv. Archive contains basic tweet data (tweet ID, timestamp, text, etc.) their tweets as they stood on August 1, 2017
- **Image Predictions File**. WeRateDogs Twitter archive 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 4 since tweets can have up to four images). Download it programmatically from Udacity's servers using the Requests library. The file format in tsv.
- **Tweet json.** Each tweet's retweet count and favorite ("like") count. Udacity provided alternative to twitter API for student have troubles in request and registration. The file in txt format

Assessing Data

In this process, data analyst notes and documents issues (quality, tidiness).

There are four main data quality issues:

- Completeness: missing data?
- Validity: does the data make sense?
- Accuracy: inaccurate data? (wrong data can still show up as valid)
- Consistency: standardization?

There are three requirements for tidiness:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

Quality Issues

twitter_arch table

- Missing data in columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls)

- tweet_id column data type is `int` instead of `string`

- timestamp column data type is `object` instead of `datetime`

- timestamp has time zone (e.g. 2017-08-01 16:23:56 **+0000**)

- In dog stages rows(doggo, floofer, pupper, puppo) has None instead of NaN, duo to that 1976 wasn't shown as missing data

- 14 rows have more than one stage(invalid data)

- There are names like (the, this, very, unacceptable) which is inaccurate names. Also, these names have lowercase characters

imgpred table

- Missing 100 image

tweet_json table

- To match other table **id** insted of **tweet_id**

Tidiness Issues

- All table observe rating but observation store into multiple tables
- The variable (Dog stage) stored in 4 columns(doggo, floofer, pupper, puppo)
- The observation is rating from original tweets, but retweets are stored in the same table
- The observation is rating but there are unnecessary columns(in_reply_to_status_id, in_reply_to_user_id, source, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls) stored in the same table

Cleaning Data

Following the assessing, the clean process begins with a method Define, Code and Test.

1- Copy to new dataframe before cleaning.

2- Combine tables into one table to describe the observation.

3- In dog stages rows(doggo, floofer, pupper, puppo) has None instead of NaN, duo to that 1976 wasn't shown as missing data.

4- 14 rows have more than one stage(invalid data).

	doggo	floofer	pupper	puppo	count
0					1976
1				puppo	29
2			pupper		245
3		floofer			9
4	doggo				83
5	doggo			puppo	1
6	doggo		pupper		12
7	doggo	floofer			1

After Cleaning

	doggo	floofer	pupper	puppo	count
0					1976
1				puppo	30
2			pupper		245
3		floofer			10
4	doggo				83

5- The variable (Dog stage) stored in 4 columns(doggo, floofer, pupper, puppo). Combine them in one column dog_stage.

6- tweet_id type need convert to string format

7- Proper Timestamp format(without time zone)

8- Timestamp type need convert to datetime format

9- Delete retweets from the dataframe

10- Delete rows no longer needed ('in_reply_to_status_id', 'in_reply_to_user_id','retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp', 'expanded_urls', 'source', 'doggo', 'floofer', 'pupper', 'puppo')

11- Fix inaccurate names(e.g. dog name "a, an, the, very") , luckly inaccurate name has lowercase as patren.

12- Fix incorrect rating_denominator. For rows with rating_denominator != 10, there are cases where they are valid ratings and there are also invalid ones.

	tweet_id	text	rating_numerator	rating_denominator
313	835246439529840640	@jonnysun @Lin_Manuel ok jomny I know you're excited but 960/00 isn't a valid rating, 13/10 is tho	960	0
342	832088576586297345	@docmisterio account started on 11/15/15	11	15
433	820690176645140481	The floofs have been released I repeat the floofs have been released. 84/70 https://t.co/NIYC820tmd	84	70
516	810984652412424192	Meet Sam. She smiles 24/7 & secretly aspires to be a reindeer. \nKeep Sam smiling by clicking and sharing this link:\nhttps://t.co/98tB8y7y7t https://t.co/LouL5vdvxx	24	7
902	758467244762497024	Why does this never happen at my front door... 165/150 https://t.co/HmwrdfEfUE	165	150
1068	740373189193256964	After so many requests, this is Bretagne. She was the last surviving 9/11 search dog, and our second ever 14/10. RIP https://t.co/XAVDNDaVgQ	9	11
1165	722974582966214656	Happy 4/20 from the squad! 13/10 for all https://t.co/eV1diwds8a	4	20
1202	716439118184652801	This is Bluebert. He just saw that both #FinalFur match ups are split 50/50. Amazed af. 11/10 https://t.co/Kky1DPG4iq	50	50
1228	713900603437621249	Happy Saturday here's 9 puppers on a bench. 99/90 good work everybody https://t.co/mpvaVxKmc1	99	90
1254	710658690886586372	Here's a brigade of puppers. All look very prepared for whatever happens next. 80/80 https://t.co/0eb7R1Om12	80	80
1274	709198395643068416	From left to right:\nCletus, Jerome, Alejandro, Burp, & Titson\nNone know where camera is. 45/50 would hug all at once https://t.co/sedre1ivTK	45	50
1433	697463031882764288	Happy Wednesday here's a bucket of pups. 44/40 would pet all at once https://t.co/HppvrYuamZ	44	40
1598	686035780142297088	Yes I do realize a rating of 4/20 would've been fitting. However, it would be unjust to give these cooperative pups that low of a rating	4	20
1634	684225744407494656	Two sneaky puppers were not initially seen, moving the rating to 143/130. Please forgive us. Thank you https://t.co/kRK51Y5ac3	143	130
1635	684222868335505415	Someone help the girl is being mugged. Several are distracting her while two steal her shoes. Clever puppers 121/110 https://t.co/1zfnTJLt55	121	110
1662	682962037429899265	This is Darrel. He just robbed a 7/11 and is in a high speed police chase. Was just spotted by the helicopter 10/10 https://t.co/7EsP8LmSp5	7	11
1663	682808988178739200	I'm aware that I could've said 20/16, but here at WeRateDogs we are very professional. An inconsistent rating scale is simply irresponsible	20	16
1779	677716515794329600	IT'S PUPPERGEDDON. Total of 144/120 ...I think https://t.co/ZanVtAtvIq	144	120
1843	675853064436391936	Here we have an entire platoon of puppers. Total score: 88/80 would pet all at once https://t.co/y93p6FLvVw	88	80

13- Fix incorrect rating_numerator. Some extraction wasn't done correctly, e.g. tweet_id 786709082849828864 has numerator 75 instead of 9.75.