Report

On

Data Wrangling Steps: Gather, Assess, and Cleaning for WeRateDogs® Twitter Project.

By:

Oluwatosin Olanrewaju

22-07-2022

Wrangle Report:

The wrangled data in this project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs®. WeRateDogs® is a twitter account that rates people's dogs with humorous comment about the dog.

The WeRateDogs® currently has over 9 million followers on twitter. Their dog ratings are based on a denominator 10 and numerator of 10 upwards which makes the account to have a wide twitter presence and engagement.

The project goals include:

- Wrangling the twitter data through the following processes:
  - Gathering the Data
  - Assessing the Data
  - Cleaning the Data
- Storing, analyzing and visualizing your wrangled data
- Reporting on the data wrangling efforts and data analyse and visualization

Gathering Data:

My wrangling efforts for the project included gathering data from three sources, which were:

• The WeRateDogs® Twitter enhanced archive: The 'twitter_archive_enhanced.csv' file was provided by Udacity to students. This archive provided the basic tweet data such as tweet ID, timestamp, text, numerator and denominator rating etc. This is for over 5000 of their tweets as at August 1, 2017. The file was downloaded through a provided link, which was then imported into the provided Udacity jupyter work space using Pandas read_csv.

Out[2]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	timestamp	source	text	retweeted_status
0	892420643555336193	NaN	NaN	2017-08-01 16:23:56 +0000	<a href="http://twitter.com/download/iphone" r...	This is Phineas. He's a mystical boy. Only eve...	N
1	892177421306343426	NaN	NaN	2017-08-01 00:17:27 +0000	<a href="http://twitter.com/download/iphone" r...	This is Tilly. She's just checking pup on you....	N
2	891815181378084864	NaN	NaN	2017-07-31 00:18:03 +0000	<a href="http://twitter.com/download/iphone" r...	This is Archie. He is a rare Norwegian Pouncin...	N
3	891689557279858688	NaN	NaN	2017-07-30 15:58:51 +0000	<a href="http://twitter.com/download/iphone" r...	This is Darla. She commenced a snooze mid meal...	N
4	891327558926688256	NaN	NaN	2017-07-29 16:00:24 +0000	<a href="http://twitter.com/download/iphone" r...	This is Franklin. He would like you to stop ca...	N
5	891087950875897856	NaN	NaN	2017-07-29 00:08:17 +0000	<a href="http://twitter.com/download/iphone" r...	Here we have a majestic great white breaching ...	N
6	890971913173991426	NaN	NaN	2017-07-28 16:27:12 +0000	<a href="http://twitter.com/download/iphone" r...	Meet Jax. He enjoys ice cream so much he gets ...	N

• The image-predictions file: this file was generated through a link provided by Udacity. The data from the file was imported to the jupyter notebook using requests function. The provides information on: what breed of dog (or another object, animal, etc.) and the accompanying url.

Out[4]:

	tweet_id	jpg_url	img_num	p1	p1_conf	p1_dog	p2	p2_conf	p2_c
0	666020888022790149	https://pbs.twimg.com/media/CT4udn0WwAA0aMy.jpg	1	Welsh_springer_spaniel	0.465074	True	collie	0.156665	T
1	666029285002620928	https://pbs.twimg.com/media/CT42GRgUYAA5iDo.jpg	1	redbone	0.506826	True	miniature_pinscher	0.074192	T
2	666033412701032449	https://pbs.twimg.com/media/CT4521TWwAEvMyu.jpg	1	German_shepherd	0.596461	True	malinois	0.138584	T
3	666044226329800704	https://pbs.twimg.com/media/CT5Dr8HUEAA-lEu.jpg	1	Rhodesian_ridgeback	0.408143	True	redbone	0.360687	T
4	666049248165822465	https://pbs.twimg.com/media/CT5IQmsXIAAKY4A.jpg	1	miniature_pinscher	0.560311	True	Rottweiler	0.243682	T

• Gathered Twitter API: Tweet-Json.txt file: I wasn't able to get an elevated access to twitter developer's but i couldnt get the required api. I used the Udacity provided twitter-json.txt. I was able to gather each tweet's retweet count and favorite ("like") count at minimum.

Out[7]:

```
[{'created_at': 'Tue Aug 01 16:23:56 +0000 2017',
  'id': 892420643555336193,
  'id_str': '892420643555336193',
  'full_text': "This is Phineas. He's a mystical boy. Only ever appears in the hole of a donut. 13/10 https://t.co/MgUWQ76dJU",
  'truncated': False,
  'display_text_range': [0, 85],
  'entities': {'hashtags': [],
   'symbols': [],
   'user_mentions': [],
   'urls': [],
   'media': [{'id': 892420639486877696,
     'id_str': '892420639486877696',
     'indices': [86, 109],
     'media_url': 'http://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg',
     'media_url_https': 'https://pbs.twimg.com/media/DGKD1-bXoAAIAUK.jpg',
     'url': 'https://t.co/MgUWQ76dJU',
     'display_url': 'pic.twitter.com/MgUWQ76dJU',
     'expanded_url': 'https://twitter.com/dog_rates/status/892420643555336193/photo/1',
     'type': 'photo',
     'sizes': {'large': {'w': 540, 'h': 528, 'resize': 'fit'},
      'thumb': {'w': 150, 'h': 150, 'resize': 'crop'},
      'small': {'w': 540, 'h': 528, 'resize': 'fit'},
```


Assessing Data:

Once the data was gathered, I assessed the data on both quality and tidiness issues, visually and programmatically using python functions.

Quality issues

Assessing the twitter Archive Enhanced File

1. There are 181 retweets in the the file.
2. Presence of invalid dog names (None, a, The, an, etc.)
3. Numerators with ratings less than 10 about 440

4. Denominators with zero rating
5. Columns in wrong data type object to datetime
6. Tweet id data type is integer instead of spring

Assessing the Image Prediction File

7. The jpg_urls are duplicated
8. Some of the tweet_ids have no images total (2075 rows instead 2356)
9. Some of the dog's names 'p's start with small letters and others capital letters

Assessing the Tweet_json File

10. Missing entries (Only 2354 entries, instead of 2356).

Tidiness issues

The dog data is separated into four different columns.

The data files are related but are in different data frames divide.

Cleaning Data:

After the assessment, the issues noted under quality and tidiness were cleaned through the Define, Code and Test process.

Define, Code and Test Process includes:

1. Made copies of original pieces of data.
2. Merged the dog's data are in four separate columns into a new single column 'dog_states' (doggo, floofer, pupper, puppo) and deleted the unnecessary columns.
3. Merged the three files of archived data, image data, and twitter_counts data into one dataframe.
4. Deleted retweets.
5. Remove columns no longer needed: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp.
6. Change tweet_id from an integer to a string.
7. Change the timestamp to correct datetime format.
8. Correct naming issues and Standardize dog ratings.