Reporting: wrangle_report

Agenda

1. Introduction

2. Data Gathering

3. Data Assessing

4. Data Cleaning

5. Conclusion

1. Introduction

The dataset that we wrangled (analyzed and visualized) 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.

2. Data Gathering

we gathered three pieces of data for this project with 3 differents methods and load them in the notebook.

1. We directly download the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv) into a DataFrame

1. We Use the Requests library to download the tweet image prediction (image_predictions.tsv) and load it in a DataFrame

1. We use the Tweepy library to query additional data via the Twitter API save it in a text file (tweet_json.txt), then read the file and load data in a DataFrame:

3. Data Assessing

In this section, We used both visual assessment and programmatic assessement to assess the data and document eight (**9**) quality issues and five (**5**) tidiness issues:

Quality issues

1. In `twitter_archive` and `images` datasets `tweet_id` is in int64 datatype

2. There are retweets rows in the `twitter_archive` table

3. `in_reply_to_status_id` , `in_reply_to_user_id` , `retweeted_status_id` , and `retweeted_status_user_id` , contain more than 80% of nulls values

4. `name` , `doggo` , `flooter` , `pupper` and `puppo` have *None* values

5. `timestamp` variable is in object datatype not in datetime

6. `source` contains 4 unique values containing : *iPhone*, *Vine*, *Web* and *TweetDeck* .

7. Some `name` are incorect, and start in lowercase.

8. Minimum `rating_denominator` is 0 instead of 10, and the maximum is 170 istead of 10 - Multiple `rating_denominator` instead of one unique.

9. Minimum `rating_numerator` is 0, and the maximum is 1176. There are `rating_numerator` bigger than 15.

Tidiness issues

1. `tweet_count` and `images` tables should be part of the `twitter_archive` table.

2. `expanded_urls` , `jpg_url` , and `img_num` are useless column.

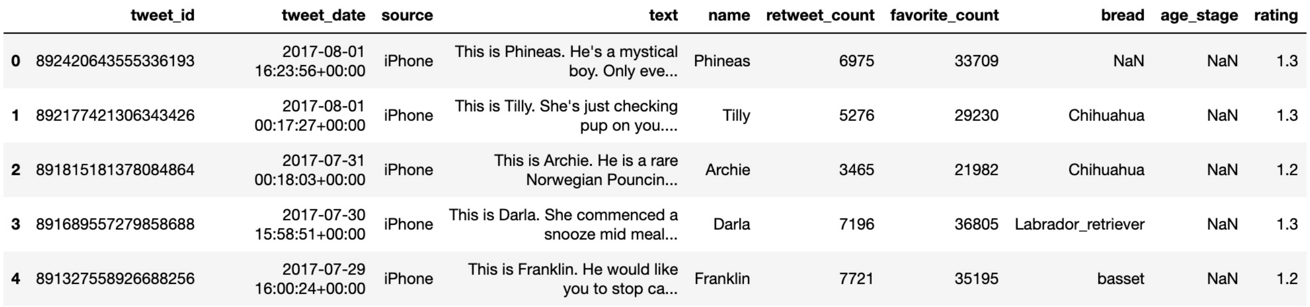
3. `doggo` , `floofer` , `pupper` and `puppo` should be grouped in one column `age_stage` .

4. `p1` , `p2` , `p3` , `p1_conf` , `p2_conf` , `p3_conf` , `p1_dog` , `p2_dog` , `p3_dog` , should be used to extract the only one `bread` column.

5. `rating_numerator` and `rating_denominator` should be grouped in one column `rating` .

4. Data Cleaning

In this section, We cleaned **all** of the issues documented while assessing. We first made a copy of the original dataset, then following the **Define**, **Code**, and **Test** process, We cleaned issues one by one and came out with one cleaned master dataset that was saved to a CSV file named "twitter_archive_master.csv" with 9 features which are: `tweet_id` , `tweet_date` , `source` , `text` , `name` , `retweet_count` , `favorite_count` , `bread` , `age_stage` and `rating` .



	tweet_id	tweet_date	source	text	name	retweet_count	favorite_count	bread	age_stage	rating
0	892420643555336193	2017-08-01 16:23:56+00:00	iPhone	This is Phineas. He's a mystical boy. Only eve...	Phineas	6975	33709	NaN	NaN	1.3
1	892177421306343426	2017-08-01 00:17:27+00:00	iPhone	This is Tilly. She's just checking pup on you....	Tilly	5276	29230	Chihuahua	NaN	1.3
2	891815181378084864	2017-07-31 00:18:03+00:00	iPhone	This is Archie. He is a rare Norwegian Pouncin...	Archie	3465	21982	Chihuahua	NaN	1.2
3	891689557279858688	2017-07-30 15:58:51+00:00	iPhone	This is Darla. She commenced a snooze mid meal...	Darla	7196	36805	Labrador_retriever	NaN	1.3
4	891327558926688256	2017-07-29 16:00:24+00:00	iPhone	This is Franklin. He would like you to stop ca...	Franklin	7721	35195	basset	NaN	1.2

5. Conclusion

During the wrangled phase of the project, We gathered, assessed, and cleaned the 3 datasets of the twitter account **WeRateDogs** archive. While doing this, some considerations were made such as:

- Considering original ratings (no retweets) that have images.
- Assessing and cleaning at least 8 quality issues and at least 2 tidiness issues in this dataset. Not all qualities and tidiness issues were assessed and cleaned.
- The fact that the rating numerators are greater than the denominators does not need to be cleaned.
- No gathering the tweets beyond August 1st, 2017.
- Droping rows with rating denominator different to 10, and rating numerator bigger than 15.

With differents considerations, or more informations on the datasets, the results (the master dataset) might be different.

```
In [ ]:
```