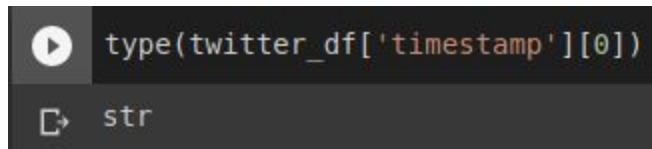


Wrangle Report

WeRateDogs Twitter archive:

- Quality Issues:

- 'timestamp' column is string not datetime.



```
type(twitter_df['timestamp'])[0])
```

```
str
```

- Converted timestamp column from string to datetime.



```
timestamp      2356 non-null  datetime64[ns, UTC]
```

- We only want original ratings, not retweets or replies.
 - Removed retweets and replies by removing tweets where 'in_reply_to_status_id' or 'retweeted_status_id' is not null.
- We do not need columns related to tweets and replies.
 - Removed Columns related to retweets and replies:
'in_reply_to_status_id',
'in_reply_to_user_id','retweeted_status_id',
'retweeted_status_user_id','retweeted_status_timestamp'.
- 'source' column is not useful.
 - Removed 'source' column.
- Dog stage is None instead of NaN.
 - Replaced entries where 'stage' column is "None" with Nan.

- Some entries have more than one stage.

doggo	floofer	pupper	puppo
doggo	NaN	NaN	puppo
doggo	floofer	NaN	NaN
doggo	NaN	pupper	NaN
doggo	NaN	pupper	NaN
doggo	NaN	pupper	NaN

- Removed entries with more than one stage.

- Some entries have invalid names. ('None', 'a', 'the', ...etc)
 - Replaced invalid names with NaN.

- Tidiness Issues:

- There should be a column for the date and a column of the time.
 - Split the 'timestamp' column into date and time.

date	time
2017-08-01	16:23:56
2017-08-01	00:17:27
2017-07-31	00:18:03
2017-07-30	15:58:51
2017-07-29	16:00:24

- The dog stage should be a single column.

```
doggo floofer pupper puppo
```

- Changed the stages into a single column called 'stage'.

```
stage
doggo
puppo
puppo
pupper
doggo
```

- Rating should be a single column.

```
rating_numerator rating_denominator
13 10
13 10
12 10
13 10
12 10
```

- Merged the rating into a single column called 'rating'.

```
rating
13/10
13/10
12/10
13/10
12/10
```

Tweet Image Predictions:

- **Quality Issues:**
 - Entries where the highest confidence is non-dog should be removed.
 - Created a row with the class of the highest confidence between 'p1_conf', 'p2_conf', 'p3_conf' for each row and named it 'breed'.
 - Kept only rows corresponding to 'p1_dog', 'p2_dog', 'p3_dog' is True.
 - Dog breed should be titled and separated using spaces.

breed
Welsh_springer_spaniel
redbone
German_shepherd
Rhodesian_ridgeback
miniature_pinscher

- Split the breed name at '_', titled each word and joint them using spaces.

breed
Welsh Springer Spaniel
Redbone
German Shepherd
Rhodesian Ridgeback
Miniature Pinscher

- Merged all data frames into one by merging them using 'tweet_id'.
- Removed unnecessary columns such as 'source' and 'tweet_id'.