wrangle_report

May 18, 2020

0.0.1 Gathering Data

- Gathering and loading the data of "twitter_archive_enhanced.csv", "image_predictions.tsv", "tweet_json.txt" files

0.0.2 Assessing Data

Quality

1) twitter_archive table

- 'name' is all written by the string including Nones
- 'name' sometimes do not have the right name i.e. "a"
- 'doggo' 'floofer' 'pupper' 'puppo' are written by string including Nones
- 'rating_numerator' values sometimes have outliers
- 'tweet_id' is integer
- we do not need the tweets beyond August 1st, 2017
- delete unnecessary columns regarding to image_predictions and tweets_df table

2) tweets_df table

- 'tweet_id' is integer

3) image_predictions table

- 'p1' 'p2' 'p3' have lower cases
- 'tweet_id' is integer
- it is easier to read the names of 'p1', 'p2', 'p3' without deliminators

Tidiness

- We need a new combined csv file
- Make a new column 'dogs' to show the types of dogs at once

0.0.3 Cleaning

Quality

1) twitter_archive table

- 'name' is all written by the string including Nones -> Replace "None" strings to null values
- 'name' sometimes do not have the right name i.e. "a" -> Replace "a" to null values
- 'doggo' 'floofer' 'pupper' 'puppo' are written by string including Nones -> Replace "None" strings to null values
- 'rating_numerator' values sometimes have outliers -> Omit the outliers using the Outlier Formula
- 'tweet_id' is integer -> Change the data type to string
- we do not need the tweets beyond August 1st, 2017 -> Change the data type to datetime, but no data to be deleted
- delete unnecessary columns regarding to image_predictions and tweets_df table -> Erase the columns that we didn't need

2) tweets_df table

- 'tweet_id' is integer -> Change the data type to string

3) image_predictions table

- 'p1' 'p2' 'p3' have lower cases -> Use the str.title() method
- 'tweet_id' is integer -> Change the data type to string
- it is easier to read the names of 'p1', 'p2', 'p3' without deliminators -> Erase the delimiters "-"

Tidiness

- Merge all 3 tables and save it as csv file
- Make a new column 'dogs' to show the types of dogs at once