We Rate Dogs Data Wrangling Report

Tachin Volker Felvic Katche
13-06-2022

Objectives:

- Wrangle data from various sources
- Store and analyse the wrangled data using visualisation
- Report wrangling efforts and acting efforts

(Step 1) Gathering Data:

Data was gathered from three sources, one was provided as a csv named “twitter_archive_enhanced.csv”.
The second data was gotten from a url into a file “image_predictions.tsv) using the requests library.
The last piece of data was accessed via the twitter API using the tweepy library. I loaded all the tweets data into a “tweet_json.txt” file based on the urls I withdrew from the “twitter_archive_enhanced.csv” dataset.

(Step 2) Assessing the Data:

Each of the datasets were loaded into dataframes for visual and programmatic assessment. At least 8
The following issues were detected:

Quality issues

1. Erroneous data types for tweet_id (should be string because no arithmetics would be performed on this variable) in archive dataframe
2. timestamp in archive dataframe (should be datetime)
3. p1_conf, p2_conf and p3_conf column names should be "confidence" instead of "conf"
4. Some denominators have values other than 10
5. Columns retweeted_status_user_id, retweeted_status_id and retweeted_status_timestamp should be removed as they are not useful to our analysis
6. Erroneous data type for tweet_id in predictions dataframe
7. Some dog breeds in predictions dataframe are lowercase, others uppercase
8. Some numerators have values less than 10

Tidiness issues

1. Twitter archive data and api data should be merged into a single dataframe as they should form a single observational unit
2. The dog types "doggo, floofer, puppo, pupper" should form a single variable

(Step 3) Cleaning the Data:

The above problems were cleaned as defined below (in no particular order):

1. Change id to tweet_id in api dataframe
2. Merge three dataframes together
3. Change tweet_id datatype to str
4. Make "doggo, floofer, puppo, pupper" columns a single column¶
5. Change timestamp type to datetime
6. Drop above mentioned columns retweeted_status_user_id, retweeted_status_id and retweeted_status_timestamp
7. Replace all denominators not 10 by 10
8. Replace values less than 10 by 10
9. rename columns p1_conf, p2_conf and p3_conf
10. make all dog breeds lowercase

(Step 4) Store the Data:

The various combined datasets were stored into a “twitter_archive_master.csv” file which would be used for further analysis and visualisations