wrangle_report

February 19, 2019

1 WeRateDogs Twitter Data Wrangling

The stages for WeRateDogs data are described in detail below:

1.1 Gathering Data

In this project data is gathered from 3 different sources:

- From flat file using Pandas' read_csv method as `twitter_archive_df`
- Downloading flat file (.tsv) from Internet using request library and using read_csv method with sep ='' option as `image_predictions_df`
- Data is gathered via the Twitter API: Firstly, developer account is obtained. Then, various keys are generated to get api object. Using get_status method, the data is gathered for each tweet in twitter_archive_df as a line and saved as a text file.Then line by line these tweets are read as JSON format,and queried for tweet_id, retweet count and favorite count,finally gathering as a dataframe: `retweet_favorite_df`

1.2 Assessing Data

Visual Assessment (Quality)

twitter_archive_df:

- Almost all values in the columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp are missing. (missing values)
- denominator_rating for tweet_id: 666287406224695296 is 2 (WeRateDogs has a unique rating system which has denominator_rating as 10)
- Dog names are: 'None','such','quite','a','an','the'

image_predictions_df:

- Dog breeds sometimes start with uppercase sometimes all lowercase.

Programmatic Assessment (Quality)

- Using `.info()` method, number of non-missing values and data types can be observed: some of the datatypes are erroneous.
- Using `twitter_archive_df.rating_denominator.value_counts()` and `twitter_archive_df.rating_numerator.value_counts()`, denominator_ratings other than 10, and very high numerator_ratings are checked one by one. Some of these ratings are found to be misinterpreted, some of them includes multiple ratings.
- Some of the tweets are retweets or replied tweets. (duplicate)
- Looking at `text` of some tweets it is found that some dog stages are missed in dog stage columns. (doggo,floofer,pupper,puppo)

Visual Assessment (Tidiness)

- All tables can be merged into one table
- columns doggo, floofer, pupper, puppo can be merged as stage column

1.3 Cleaning Data

- The tables are copied before cleaning the data.
- First it's needed to clean duplicate tweets (retweets or replies other than NaN) because using them in the analysis will result in inaccurate analysis. Then colums related to retweets or replies need to be dropped using `.drop` method because all of the values will ne NaN.
- Dog stages are extracted from the text column because some values were missing.Then these columns are merged to dog_stage column using .join, dropping NaN values.
- All three tables merged on all_clean table.
- Tweets does not include dog ratings are dropped. Tweets with misinterpreted ratings' text are read and one by one ratings replaced with correct ratings using `.at` method . Ratings of tweets with multiple dogs are divided by number of dogs. Then rating_denominator column is dropped as all of them are 10.
- Incorrect dog names are replaced with NaN using .replace
- Erroneous data types are converted to correct ones using `.astype` and `.to_datetime`
- Name of dog breed are made lowercase using `.str.lower()`

All data is stored in 'twitter_archive_master.csv' file.