wrangle_report

December 17, 2018

1 Wrangling Report

1.1 Gathering the data

All the data were imported from different envifonments and stored in three DataFrames: - Twitter archive imported from given csv file - images imported from imported from url - tweet informations imported from twitter Api

1.1.1 Gathering the data from given file

To gather the data from given csv file I used pd.read_csv() file name: twitter-archive-enhanced.csv; to store it in twitter_archive.

1.1.2 Taking the ID of tween from url in column 'expanded_urls'

The columns tweet_id in twitter_archive had wrong datatype and value so what we needed to do is to extract tweet_id from expaned_urtls.

1.1.3 Downloading image_predictions file using requests

Thanks to requests libraries we can download images from url and store it in image predictions

1.1.4 Gather data from twitter via API

Gthering data from the Twitter api using developer acces by consumer_key, consumer_secret, access_token, access_secret. To acces the data we needed to create developers accoun and use tweepy.

1.2 Assesing the data

1.2.1 Quality of the data

- in twitter_archive columns 'in_reply_to_status_id' and 'in_reply_to_user_id' there are erronous datatypes and valuses
- in twitter_archive we have tweet_id in wrong datatype it should be string
- Columns 'retweeted_status_id', 'retweeted_status_user_id' and 'retweeted_status_timestamp' are not quite usefull, we don't need retweets
- in twitter_archive we have difficoult to read sources

- in twitter_archive a part of dog names is incorrect
- in twitter_archive tweet_id:810984652412424192 doesn't contain a rating
- in twitter_archive we would like to have only original ratings
- in twitter_archive timestamp there is wrong datatype
- in twitter_archive not all tweets have images (the shape is not equal for rows)
- in image_predications there are wierd dogos breads as 'fur_coat', it is possible that it isn't a dog ;)
- dogos breads are written in different ways
- after compbaingin puppest and floffers - in one variable it should be a category type
- the archive data should contain also images

1.2.2 Tidiness of the data

- in twitter_archive columns 'doggo', 'floofer', 'pupper','puppo' are one variable but in 4 columns and some of them has two values
- tweet_info should be joined to twitter_archive data
- rating numerator and dominator should be shown as a one variable

1.2.3 Cleaning the data

Making copies of the data Making copies of the Data before cleaning and naming it for first two letters of names before copying - twitter_archive into 'ta' - image_predictions into 'ip' - tweet_info into 'ti'

Problem #1 Quality: columns 'retweeted_status_id', 'retweeted_status_user_id' and 'retweeted_status_timestamp' are not quite usefull, we don't need retweets

Define Delete columns 'retweeted_status_id', 'retweeted_status_user_id' and 'retweeted_status_timestamp'

Problem #2 Quality: tweets should have images but there is less rows for images than for tweets

Define Delete rows there is no match for image

Problem #3 Quality: column 'timestamp' in twitter archive has wrong datatype

Define Convert the column to proper datatype

Problem #4 - #5 Tidiness: in twitter_archive columns 'doggo', 'floofer', 'pupper','puppo' are one variable but in 4 columns and some of them has two values.

Define Take only data with one value and create 'dog_stage' variable which is made by extracting the dog stage variables from the text column when available and drop all the previous ones.

Problem #6 Quality: the dog stage is a category not an usual object - the datatype should be changed

Define Change the datatype for 'dog_stage' to a category

Problem #7 Quality: change a view of sources becouse it's hard to read

Define Remove urlis signs before acctual source

Problem #8 - #9 Quality: - In ta, nulls represented as 'None' in columns 'name', - Some values are wrong in name. Names that varen't capitalized are wrong.

Define Set the value wrong names to those from text and replace 'None' with np.nan.

Problem #10 Quality: In ta (twitter archive), some ratings are wrong.
Tidiness: Rating_numerator and denominator should be one variable rating.

Define

- Change the rating_numerator and rating_denominator for oberservations with wrong value
- Dropping the: '810984652412424192' because it doesn't have a valid rating
- Create new column rating which is an division of rating_numerator/rating_denominator, getting rid of rating_numerator and rating_denominator.
- Droping oberservations with extreme ratings.

Problem #11 Quality: In ip (image predictions), some predictions are not dogs, there is no column for the most possible breed of a dog and the confidence.

Define Create new columns predicted_breed and predicted_conf for the most possible breed of a dog and the confidence.

Problem #13 Tidiness: Twittes, tweet informations and images predictions should be together

Define Join those schemas

Problem #14 Quality: column 'timestamp' in twitter archive has wrong datatype

Define Convert the column to proper datatype

1.3 Store data

Storing cleaned data info a csv named 'twitter_archive_master.csv'