wrangle_report

September 27, 2020

1 Data Wrangling Report

By Eman Abd Elhalim

This report illustrates the main steps involved in the data-wrangling of Twitter account "WeRate- Doges".

1.1 Data Gathering

In this step, collecting data for this project from three sources :

- Twitter_archive_enhanced.csv file, this file downloaded manually to my working directory.Then imported into working environment using pandas function "pd.read.csv".
- Image_prediction.tsv is the second file, it hosted from webpage and downloaded it from its relevant URL using Requests library get function and pd.read.csv pandas function. This file is containing image predictions for the dogs breeds obtained through a neural network on most of the tweets in the archive file.
- The third dataset was gathered form twitter REST API via the tweepy library by querying the API to obtain extra information pertinent to the tweets' ids in twitter_archive_enhanced.csv file, e.g. retweets count and favorite count .

1.2 Data assessment

in this step, I investigate the datasets both visually and programmatically for quality and tidiness issues.

- I do the visual assessment on excel spreadsheet. Then, the programmatic assessment is donein Jupyter notebook.
- First I addressed the missing values. Then addressed the tidiness issues second to facilitate the tackling of the rest of quality issues that were validity, accuracy and consistency classes of the quality aspects and extracts these assessment point :

1.2.1 Quality aspects:

archive table

- retweets and replies not requird ('in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp')
- some expanded_urls doesn't have pictures

- missing values in name, doggo,floofer,pupper,puppo
- rating_numerator have values greater than 15 and smaller than 6 and some contain decimals rating_demoniator have values not equal 10
- Erroneous datatypes (timestamp, tweet_id)
- weird names in name column

image_prediction table

- Erroneous datatypes (tweet_id)
- there're tweets_id not have image

api table

- Erroneous datatypes (tweet_id)

1.2.2 Tidiness aspects:

- one column for doggo and floffer, pupper and puppo
- columns headers are values not varibale name in image prediction
- Api table isn't an observational unit to have its own table

1.3 Data Cleaning

first make a copy from original tables

1.3.1 Quality issues

archive table

- change timestamp type from string to date using astype('datetime64')
- change tweet_id type from int to string using astype('str')
- Removed the rows that have no expanded_url entry
- fixing rating_numerator values that greater than 15 and smaller than 6 by slice their tweets and investigate texts and extract their values from texts
- changing all inconsistent value such as a, an and any name less than 3 letters in name column

image prediction table

- changing datetype of tweet_id from int to str in image prediction table

API table

- changing datetype of tweet_id from int to str in API table

1.3.2 Tidiness issues

- combine four dog stages into one column dog_stage and drop doggo,floofer,pupper,puppo columns where :

  at first assigned the last four columns of the archive dataframe to its new value records without None and Getting all the tweets where the value of both 'doggo' and 'pupper' is not none, Extracted only those the columns of interest and investigate its head. and combined 4 dog stages using addition operation. then,replaced empty string with np.nan. after that,separated the combined stages with a hyphen then dropped dogs type cloumns and tested this step using 'info()' to confirm that all columns combined well.

- drop records not have images then drop replies and retweets and drop retweets and replies from image table :

  at first create a list of tweet_ids with images "tweets_with_image" and confirming that all the tweets with images exist in the archive dataset length. then, extracted the retweets that include data in the retweet_status_id. and dropped the retweets from the archive data set. and Extracted replies entries that include data. then checked image_prediction table for extra tweet ids not in archive table after that,dropped retweets and replies ids from image prediction dataframe and tested this issues using 'shape' method for archive table and image table and I confirmed that the two tables are identical rows.

- reshaping image_prediction cloumn using pd.wide_to_long to do that first edit the name of columns in image_prediction table

  at first renamed the dataset columns with clear names, then reshaping the dataframe using'pd.wide_to_long'

- merage api table and archive table together

  in this step I meraged api and archive tables using left join to keep all tweets id in archive table

- delete retweets and replies from table

  I deleted retweets and replies from dataframe

1.3.3 Output:

two tables :

- twitter_archive_master table 1981 record
- prediction table 5943 entries