Wrangle_report

This data wrangling process mainly contains five parts, which respectively are Gather, Assess, Clean, Reassessment and Save the csv.

Gather

Three data sets have been gethered, which are Twitter_archive, Image_predictions and Twitter_JSON

- Twitter_archive was downloaded in udacity website
- Image_predictions - requests library was used to download the data, and then saved the csv in the local file system.
- Twitter_JSON - also used requests library to download the data, then extract ID, retweet_count and favorite_count to form a dataframe

Assess

Used both visual assessment and programmatical assessment, for visual assessment, mainly used df.info(), df.sample(), for programmatical assessment, mainly used for loop, df.value_counts(), indexing, df.duplicated() and so on, issues found in those data sets are as follows:

Quality

archive

- lowercase name(the lowercase names are not real names)
- wrong data type(tweet_id)
- outliners in columns of rating_numerator and rating_denominator
- E-notation in columns of in_reply_to_status_id & in_reply_to_user_id & retweeted_status_id & retweeted_status_user_id
- missing data(dog types & name & expanded_urls) cannot be wrangled

predictions

- wrong data type(tweet_id, img_num)
- lowercase dog category(p1, p2, p3)

twitter_json

- wrong data type(tweet_id)
- 'favorite_count' contains 0 which is abnormal

Tidiness

archive

- dog type(doggo floofer pupper puppo)takes four columns(Each variable forms a column.)
- combine the three data sets

Clean

To cope with above questions, first, copies were made fro those three data sets. Solved each question once at a time follwing the procedure of define - code - test. Below are the solutions for all issues according to the order in the wrangling process:

- Tidiness - dog type - convert four columns of dog type into one column(doggo floofer pupper puppo)
- Quality - archive - lowercase name - There are lower-case names in the column of 'name', which are not the real name. The solution is to convert those lower-case names into 'None'
- Quality - Archive - wrong data type(tweet_id) - change column 'tweet_id' data type to 'str'
- Quality - Archive - outliners in columns of rating_numerator and rating_denominator - Normally the rating_denominator should be 10, but some of the pictures contain several dogs, then the rating_denominator should be the multiple of ten, check the denominators which are not the multiple of ten, then see if we can find the correct denominator in relevant text.
- Quality - predictions - wrong data type(tweet_id, img_num) - change the data type of 'tweet_id' to str, and the data type of 'img_num' to category
- Quality - predictions - lowercase dog category(p1, p2, p3) - convert the lowercase dog categories to uppercase
- Quality - twitter_json - wrong data type(tweet_id) - convert the data type of tweet_id to str
- Tidiness - Combine the three data sets
- Quality - archive - E-notation in columns of in_reply_to_status_id & in_reply_to_user_id & retweeted_status_id & retweeted_status_user_id - first fill na with 0, then convert the data type to int
- Quality - twitter_json - 'favorite_count' contains 0 - Remove rows of favorite_count containing 0

Save the csv

Used the df.to_csv to save the wrangled dataset in the local file system

Reassessment

The new combined data set needs to be reassessed. All the issues found in the new data set were processed right after the assessment, below are the issues and relevant solutions:

Quality - combined_all

- wrong data types of 'retweet_count' & 'favorite_count'
- wrong data type of 'type'
- drop nan
- some of the values in columns of 'rating_numerator' & 'rating_denominator' are for several dogs, such as some of the denominators are not 10, but a multiple of 10, which would cause difficulties in analyzing the topic related with those ratings

Solutions

- drop nan according to column 'p1'
- change the data types of 'retweet_count' & 'favorite_count' to int
- change the data type of 'type' to category
- remove na in columns of rating_numerator & rating_denominator, then use convert those denominators which are not 10 into 10, and do the relevant modification to the related numerators. After the modification, since all the denominators are 10, then remove the column of 'rating_denominator', keep the modified column of 'rating_numerator'