Ahmed Mousa

s-ahmed.mosa@zewailcity.edu.eg

Udacity data wrangling project report

Gathering

There were three different datasets, one for tweets archive, one for image predictions, and another for API.

Archive

This was gathered using read_csv to read the file: twitter-archive-enhanced.csv, and then turned into a dataframe (df).

Image

Was gathered from this url using requests library: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

It was then read similar to the archive df but with separator set to '\t'.

API

Was gathered from the prepared file, not with accessing API data using a developer account. It was read using json library. Only retweet_count, favorite_count, and tweet_id were selected.

Assessment

Quality

Archive

The following issues were found:

- 55 names have 'a' as name and 7 as 'an', 8 as 'the': validity and accuracy issue

Visual assessment discovered a name 'a' in the name column. Then by searching visually and programmatically, they were determined to be 55. Then 7 rows named 'an' and 8 named 'the' were discovered.

- missing names in name column: completeness issue

Caught by visual assessment, then confirmed programmatically.

- some values in rating_denominator are not 10: consistency and accuracy issue

Confirmed visually and programmatically.

- existence of replies and retweets: validity issue

Confirmed visually and programmatically.

- existence of tweets without image predictions: validity issue

Confirmed programmatically with clear difference in entries between the two dfs.

- timestamps are objects instead of datetime: validity issue

Confirmed programmatically.

- tweets containing photos of non-dogs

Confirmed visually when searching for tweets with low ratings and invalid names.

Image

- some predictions for unoriginal tweets might exist: validity issue

Based on the existence of unoriginal tweets in archive.

- three predictions per one image: consistency issue

The first prediction has higher confidence as confirmed visually. Hence, no need for the other two predictions.

- some predictions show the images don't belong to dogs

As confirmed in the p_dog columns visually and programmatically.

API

No significant issues found.

Tidiness

Archive

- doggo, floofer, pupper, puppo should be only one column determining the stage of dog

These columns represent only one observation.

All dfs

- the 3 dfs contain observation of a shared subject, but are separated into 3 dfs

The 3 dfs contain shared tweet ids. It would be more consistent and informative merging them into one df.

Cleaning

Quality

Archive

- 55 names have 'a' as name and 7 as 'an', 8 as 'the'

Fixed using replace function.

- missing names in name column

Cannot be fixed since no way in my hands to obtain these names.

- some values in rating_denominator are not 10

Some denominators are multiples of 10 reflecting the existence of several dogs in the same image. Yet, some values such as 2 and 15 are indivisible by 10. Thus, they were removed.

We need to make numerators consistent too, i.e., per dog, as follows:

Creating a dog_count column using the denominator column. Then dividing numerators by dog counts to give the true rating per dog for images with multiple dogs. Then, any abnormal values were checked afterwards.

Unexpected values resulted including values less than 1. Manual check for the urls of a case of rating 165/150 revealed that the rating was correct according to the tweet text. Two values were zeroes and both were correct in tweet texts. Next, manually checked two samples for ratings >

100. They turned out to be true. However, one, which was rated 420, was not of a real dog. This was what lead to discovering the issue of non-dog photos. This was confirmed by other manual checks too.

- existence of replies and retweets

Cleaning requires removing all tweets which have non-null values in any of the following five columns:

- in_reply_to_status_id
- retweeted_reply_to_status_id
- retweeted_status_id
- retweeted_status_user_id
- retweeted_status_timestamp

When focusing on the two columns, retweeted_status_id and in_reply_to_status_id, after attempting removing non-nulls in the first column, all corresponding values in retweeted_status_user_id and retweeted_status_timestamp were removed. Thus, the same step was attempted with in_reply_to_status_id. The attempt was successful and all entries in in_reply_to_user_id were removed. The final result was a df with 2093 entries. All the five columns were then dropped.

- existence of tweets without image predictions

All tweets without images were then removed from the df. Then the resultant archive df was 1967 entries.

- timestamps are objects instead of datetime

Using pd.to_datetime, the dtype of this column was changed.

Image

- some predictions for unoriginal tweets might exist

Now that archive df only contains originals, archive can be used to remove unoriginals in images. However, no changes were found in entries. Thus, no unoriginals were in image df.

- three predictions per one image

Since prediction 1 usually has the highest confidence, as confirmed by visual assessment, it would be logical to remove the other two predictions, along with their related columns.

- some predictions show the images don't belong to dogs

Since we now have only highest confidences, it is more likely that the non-dogs here would be true. So, removing these would give us acceptable results with only few mistakes. The resultant was a df with 1543.

Archive continued

- tweets containing photos of non-dogs

This final issue could only be fixed after the previous fixes in image df.

Now, we use the same method as before to keep only columns with dog images in archive. The resultant archive df was 1460 entries. This means 507 out of 543 dogs without images were in archive df.

Tidiness

Archive

- doggo, floofer, pupper, puppo should be only one column determining the stage of dog

A stage column was created, and these columns were all then dropped.

All dfs

- the 3 dfs contain observation of a shared subject, but are separated into 3 dfs

We want only dog image predictions in archive. For api, we want only ones with image predictions and for original tweets. So, we first remove any images not in the last version of archive. Then compare and remove the extras from api not in image, and not in archive. Then the 3 dfs were merged into a df called twitter_archive.

Merging the dfs was with the aid with the stackoverflow answer here:

https://stackoverflow.com/questions/44327999/python-pandas-merge-multiple-dataframes#:~:text=Just%20simply%20merge%20with%20DATE,to%20get%20all%20the%20data).&text=Now%2C%20basically%20load%20all%20the,using%20merge%20or%20reduce%2

0function.&text=Note%3A%20you%20can%20add%20as,frames%20inside%20the%20above%20list.

Saving

The master df was then saved into a csv file called: twitter_archive_master.csv