Data Wrangling Report

Mahmoud Khayrallah

Introduction :

This process consists of 3 steps :

- gathering
- assessing (visual and programmatic)
- cleaning

Data gathering :

This process passed through three steps :

- downloading ‘twitter_archive_master.csv’ manually from the classroom and then loading it to local_df
- downloading ‘image_predictions.tsv’ programmatically using requests library then loading the data to image_df
- downloading retweets and favorites data through the twitter API then load them into json_df using tweepy and json modules (in this step, I faced problems with my internet connection, so I left the code in the notebook and used the ready file from the classroom)

Assessing :

This step aims to assess the data and determine the needed modifications . It contains two inner steps :

- visual assessing : in this step , I loaded the three dataframes in the notebook and looked on their structure and variable values in general , why ? It gave me better impression about the data and deeper understanding, I was also able to find clear problems in the data (quality and tidiness) like missing values and variable types
- programmatic assessing : deeper assessing using the understanding I built in the first step and making use of pandas and numpy functions

Note : I repeated this step twice , once before merging the tables and the other after merging with some mini unit iterations through assessing and cleaning processes

Issues Found :

all issues :

quality :

- source variable contains the whole link html not only the source (accuracy)
- there are some retweets that need to be deleted(accuracy)
- the "None" word needs to be replaced by 0 in the four types of the dogs variables in order not to affect the visualizations and analysis(consistency)
- the expanded url has some missing data(completeness)
- timestamp is a string not a date (consistency)
- "impossible to get" tweets should be removed from the dataset (completeness)
- wrongly parsed tweets and tweets contains dates and fractions should be modified (accuracy)
- overrating tweets should be deleted for accurate analysis(accuracy)
- modify dog names to have the first letter only capitalized (consistency)

> all quality issues are in the local_df dataframe
>
> the retweets issue can be classified as tidiness issue since our table unit is tweets not retweets, but in this problem after merging we will study dogs ratings not the tweets itself so retweets are considered repeated (invalid) values

tidiness :

- the dog type four columns(doggo, floofer, ...) represent one categorical variable, they are values not variables(local_df)
- numerator and donominator both represent the rating variable
- p1_dog, p2_dog, p3_dog all represent 1 variable that can be named "p_dog" - is this a dog ? -(image_df)
- p1, p2, p3 contains only one variable which is breed(if the pic contains a dog ofc)(image_df)
- pics without dogs should be excluded (for better analysis)(image_df)
- rows are not equal in the three tables
- the three tables represent dogs, they should be merged

In the second iteration I found that names variable is still not accurate and tried making it better.

During the merging step, I deleted some columns from each and let he ones I'll use in the analysis only

Cleaning :

This step consists of 3 levels :

- define
- code
- test

I used the same 3 steps and my approach was :

1. defining cleaning steps
2. coding and testing every single step of them as a single unit in 2 code chunks ,one for code and one for test and sometimes the step took more than two chunks (This approach made testing easier)

Cleaning define step :

define

> the steps here will be done in the code chunks below, each step will have 2 chunks with its number in a comment (one for coding and one for testing)

1. copy the three dataframes to local_df_copy, image_df_copy, json_df_copy
2. extract the real source source variable in local_df_copy
3. find missing expanded urls using tweet id and expandurl api(not needed in the analysis step)
4. delete retweets in local_df_copy
5. delete overrating tweets and tweets with no possible logical rating in local_df_copy
6. extract the datetime from the time stamp in local_df
7. modify wrong ratings in local_df_copy
8. capitalize the first letter of the names only in local_df_copy
9. convert doggo, floofer, pupper, puppo variables to boolean in local_df_copy
10. construct a column "type" in local_df_copy(excluded from the process)
11. construct a column "rating" in local_df_copy equals num/den * 10
12. construct a column "p_dog" in image_df_copy
13. delete rows with no dog pics in image_df_copy
14. construct a column "breed" in image_df_copy
15. merge three tables in a df master_df

> I began with completeness issues and after that I should look at consistency but the order is a little different since there are steps depending on other steps.
> **example** : rating column depends on modifying ratings

The other two steps are included with comments in the project notebook

Reflection :

This step took more time than I expected, in the beginning I didn’t understand why would the wrangling process take 70-80% of the project time. But after working on this project I understood its details and had too much fun. Hope that my work was good enough .