Our Best Friends For Ever

By Fady Maher Mazhor Barh

Data wrangling report of weratingdogs

By Fady Maher Mazhor Barh

December 2020

As an assignment for Udacity Data Analyst Nanodegree , this Report illustrates the main steps involved in the data wrangling of twitter account " weratedogs ".

Data Gathering

In this step we collected data from three main sources for data to deal with :

1. Twitter-archive-enhanced.csv , We manually uploaded this file to your working directory that we got through workspace, and then importing it in the project by using pandas function " pd.read_csv() "
2. Image-predictions.tsv is the second file, we download this file from url link by using requests library get function then we importing into to work enviroment by using pandas library function " pd.read_csv() ", this file includes image predictions for the dogs.
3. The final dataset we collected was from api twitter account by using tweepy library then we did extract information from this api in the form of "tweet-json.txt", then we read it and line by line and extract from each line informations ex (id_str, retweet_count , favorite_count) , then we create by using this information dataframe called api_df.

Data Assessment

This is the second step in his process Data Wrangling , Where we do visual and programmatic of the data and extract the quality and tidiness issues.

The Visual Assessment we done on spreadsheet application, then the programmatic Assessment we done on jupiter notebook.

QUALTIY DIMENSIONS

1- completeness:

1.1 We can see in In Table "archive_df" many missing values in several columns like (retweeted_status_id retweeted_status_user_id, retweeted_status_timestamp) .

1.2 We can see in In Table "archive_df" many missing values in several columns like (in_reply_to_status_id, missing values in in_reply_to_user_id).

1.3 we notice that the tweets that do not have pictures (archive_df , image_prediction_df) should be deleted from datasets.

1.4 dropping the re-tweets from the archive_df set , image_predictions_df and api_df sets.

2- validity:

2.1 we notice In Table (image_predictions_df, archive_df) tweet_id should be is a string not a integer .

2.2 we notice In Table (archive_df) rating_denominator column is invalid .

3- accurancy:

3.1 we notice in table "archive_df" Erroneous datatype (timestamp) should be convert from string to datatime.

4- Data types(consistency issues):

4.1 we notice in table "archive_df" (doggo - floofer - pupper - puppo) should be convert from None value to nan value.

Tidiness issues

1. We notice in table "Archive_df" values are column names (doggo - floofer - pupper - puppo).

2. We notice in table "image_predictions" of the type Column headers are values, not variable names.

3.we notice that tweet_id columns in 'archive_df' table duplicated in 'image_predictions_df' and 'api_df' tables.

Data cleaning

This is final step of Data Wrangling steps , Where in these steps we perform cleaning of the data through several steps :

1. Firstly , We made copies of the datasets into new tables ("archive_df_clean" ,"image_predictions_df_clean","api_df_clean") by using copy() function.

2. Then we do operations Evaluation of qaultity and tidiness issues through three steps :

1. **Define** : here we define the problem and what we want to get to it.
2. **Code** : here we are writing the code that solves that problem.
3. **Test** : Here, we do a test for the code step , if we have reached a

Valid solution or not , in the case if the solution is correct, the problem has been solved, but if the solution is a mistake, we go back to the step of the code and make sure of it.

3. We dropped columns (in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id) in table "archive_df" because there are too many missing values in it.

4. We deleted the tweets that do not have pictures in table "archive_df_clean" so that we have every tweet we have a picture and thus become the number of rows in table "archive_df_clean" equal in table "image_prediction_df_clean" .

5.we dropping the re-tweets from the archive_df_clean set , image_predictions_df_clean and api_df_clean so that we have unique tweets, there is no repetition in it.

6-we dropping the rows that include value of "rating_denominator" not equal 10 in archive_df_clean table because its value rating_denominator must be 10.

7-we converted timestamp column in "archive_df_clean" from string to datatime, so that we can deal with it as a date and thus it is easy for us to extract from it (year, month, day, ... Etc.).

8- we converted values (None) of (doggo - floofer - pupper - puppo) columns to (empty string) in table archive_df_clean because it is wrong values.

9- we notice in table "Archive_df_clean" values are column names (doggo - floofer - pupper - puppo) so we use pandas libray melt function to to convert (doggo - floofer - pupper - puppo) into inputs to new column called "dog_stage".

10- we notice that in "image_predictions_df_clean" of the type Column headers are values, not variable names Therefore, we used pandas library wide_to_long function to reshape the table so that we created only three new columns ('prediction', 'confidence', 'breed') and insert the values of the 9 old columns in them

11- finally we created Twitter_archive_master, so we can merged archive_df_clean, image_predictions_df_clean, api_df_clean tables.

Conclusion

We have made a report on the weratedogs and explained the 3 main steps in it, which are represented in gathering, assessment, cleaning data and we clarified how to gathering data and the tools used for that, in the second step we explained how We carried out an assessment of the data and quality by do visual and programmatic of the data, in final step we clarified the cleaning processes that were carried out on the data.