Wrangle and Analyze Data

Jamal Alanazi
July 20, 2020

Introduction:
We have three files and we have three step in project Data Wrangling Gathering, Assessing and Cleaning. Report of wrangle and Analyze Data. I will write report detailed report on the most important detailed steps on the project, step by step.

First step Gathering:

In the gathering, we are import pandas, NumPy, random, matplotlib, JSON, os, and requests. Next step, we writ API testing code but we could not be run the code Because

Will add each available tweet JSON to df_list. Create DataFrames 'JSON' and save it in a file. Read the first file, 'JSON,' and print. Read tweet_df information.

Second step Assessing:

Read the file 'twitter-archive-enhanced.' Dimensions twi_arc_enh file and getting the file information. Reading the next file 'image_predictions.tsv', general statistics and getting the file information.

I do eight Quality Problems:

1- Drop Null values from the twt_arch_enh file.

2- Drop the columns 'in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp', from twi_arc_enh table.

3- Transform colums p1_dog, p2_dog, p3_dog from bool type to int.

4- Dropp the zero in colums p3_dog.

5- In the image columns, jpg_url should be dropped.

6- Drop P1_conf >= 0.1.

7- imp_pre (p1) change to lowercase.

8- imp_pre (p2) change to lowercase.

9- imp_pre (p3) change to lowercase.

I do Tidiness Problems:

1- Change timestamp in twi_arc_enh from object to datetime.

2- The three files should be merged.

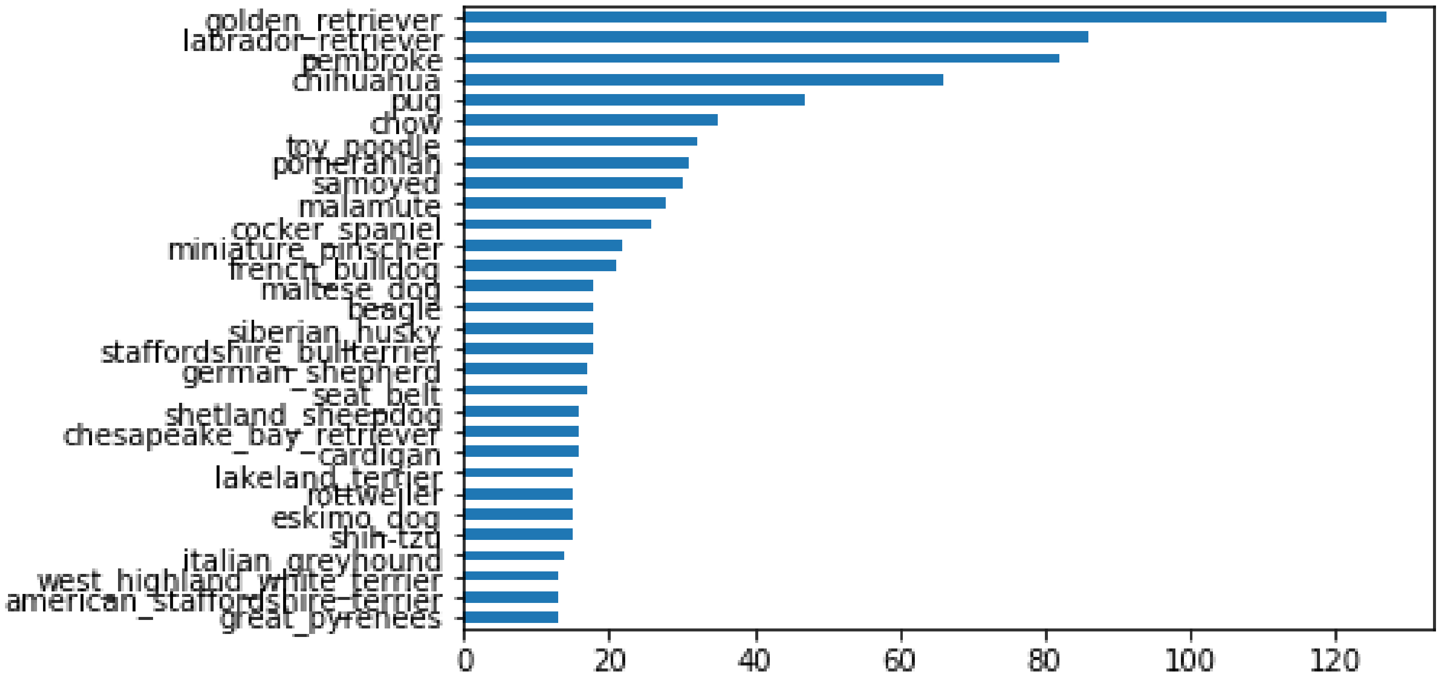
Third step Cleaning:
I clean all Quality Problems. Drop the columns 'in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp', from twi_arc_enh table. Locating columns with values <200. After that test.
Next step, drop the located columns from above after that test.
Chang time type from object to pandas DateTime. After that test and test the file after dropping.
Check the number of unique values and test the time column type.
Change is from bool to int, after that test.
Drop the zero in columns p3_dog.
In the image columns, jpg_url should be dropped after that test.
Drop P1_conf >= 0.1 and test.
imp_pre (p1,p2,p3) change to lowercase and test.

The last step in the Cleaning, I will merge the three files that should be merge together.

Analyzing the data:
Read the file 'master_data' and information.
I put three questions and answer:
1- What dog type received the highest retweet count?
2- What are the dog types, and how many dogs are there for each?
3- What is the average rating for the Golden Retriever dog type?

Q1: Hence, the Labrador Retriever attained the highest retweet count!

Q2: The highest counted type is the Golden Retriever, with more than 120 counts. Also, we have various kinds of sharing similar count values.



Q3: get the sum and len. = 13.36