wrangle_report

April 5, 2022

0.1 Reporting: wragle_report

The process of wrangling data was composed by 3 steps: gather, access and clean.

On this project, those steps were not followed, and it was started by gathering.

On gathering, it was used three types of gather data methods. A Read CSV file from pandas library, a Request library and a Twitter API. Reading CSV file (of a file provided by Udacity and uploaded on the workspace) and use request library was simple, Twitter API were a bit more complex and took more time to process, since it required keys and token used to connect to Twitter, both were created on the Twitter Developer page and that information not appear on the project steps.

When accessing the data by visual and programmatic method it was easy to see quality and tidiness issues, although it was necessary to step back on gather to extract more data with the API since it was on accessing that it was possible to get some idea of what information it was needed. But the issues raised on the part were the follows: Columns like “tweet id” and “retweet status timestamp” were not with the correct datatypes, Columns names were not easily describe such as “p1”, “p2”, The dog stage that are a variable has 4 columns instead of one, for example. All items observed on this section are listed here:

1. retweeted_status_timestamp, timestamp should be datetime instead of object (string). (twitter_archive)
2. tweet_id should be object(string) (twitter_archive)
3. We only want original ratings "no retweets" those ones have image. (twitter_archive)
4. Change p1 to Prediction1 and other columns that is related to the prediction such as (p1_conf, p1_dog etc...) to (Prediction_conf etc..)
5. The ratings are not extracted correctly especially decimals
6. The name column has many invalid values like , a, an, the
7. Rating data can’t be compare since the demoninator are not the same.
8. Change tweet_id to an object datatype
9. the dog stage has 4 stages
10. Combine the datasets together.

The final part of gather was cleaning. On this step it was used pandas functions to manipulate the data sets and correct the raised issues on the earlier step. Functions like: drop, to datetime, rename, melt, merge, info and head were used to code and test.

With the wrangle process finished the data set was stored to further analyze. The result was a table with tweet id, rating, 3 possible predictions, the dog breed and stage, image, url of the tweet, times of favorites and retweets.

In [ ]: