WERATEDOGS TWITTER ARCHIVE WRANGLE DATA REPORT

In this report we are describing the steps and effort to gather, assess, clean and analyze the data required.

I. GATHER DATA

In this part I have gather data from 3 sources:

1. WeRateDogs twiiter archive enhanced file: The file was downloaded manually from Udacity platform and read it using pandas library
2. Images Predictions file: The file was downloaded programmatically from an url provided in the project.
3. Tweet JSON Data: We have query the data from Twitter API programmatically using Tweepy

I have then stored the 3 data in the table: archive, predictions and tweeter_json

II. ASSESS DATA AND CLEAN QUALITY/TIDINESS ISSUES

We have performed some assessments programmatically and visually in order to detect the quality and tidiness issues. We have detected at least 8 quality issues and more than 2 tidiness Issues :

- We have drop all the tweets with non-null values in the columns retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp
- We have also dropped all the tweets with non-null values in **in_reply_to_status_id,in_reply_to_user_id**
- We have checked the tweets having missing expanded url and we found that there is not an impact on the rating
- The **timestamp** was changed from string to date format
- We have analyzed the values of name and we found some inconsistent values like : a and none
- We have dropped all the **rating_denominator** >10
- We have dropped all the tweets with **rating_numerator** >15
- We have cleaned the **source** column
- We have melted the **dog stage** into only one column from 4 different columns. Tweets without specific stage were set to no_stage
- We have melted the **dog prediction** and **prediction confidence** in 2 different columns
- No issue was found in the **Json_data :** No missing data, data was tidy and all the all the columns type were well defined

III. MERGE ALL THE 3 DATA

We have the merged all the 3 data source into one master data: twitter_archive_master.csv

IV. DATA VISUALIZATION

Finally, we have analyzed the data and find more than 6 insights.

Boris MVE MANGA 09/09/2022.