Reporting: wrangle_report

Table of Contents

Gathering

The Gathering Phase is where the data are collected. For this analysis it is collected from three source which are:

- **twitter_enhanced_archive** which is provided to download and it is a csv(comma seperated file) and read by using Pandas read_csv
- **image-predictions.tsv** which is gathered through the request library from the url where the tsv file is being stored and the contents are being derived through the use of Beautiful soup library which is then saved as a tsv(tab separated value) file and read using the Pandas read_csv and
- **additional_tweets_info** is gathered through the use of Twitter API which requires for the use of developer account and the API is query by using the *tweets_id* column in the **twitter_enhanced_archive** dataset, then it is saved as a text file in the name *tweet_json.txt*. The *tweet_json.txt*** *is read and the keys with values that are extracted are the* id*,* favorite_count *and* retweet_count* which is then stored as a csv file and read by using the Pandas read_csv.

Accessing

`twitter_archive_enhanced`

The head of the dataframe is being checked which is the first six columns, the shape of the dataframe is checked, the info, null values was checked, the summary statistics was checked and duplicates values. Some null values were being spotted, no duplicates values.

The value counts of the *source* column is checked which shows the column should be split into two a be a category datatype, value counts and unique of *name* column was checked which leads to spotting some issues like invalid names and null names which is in None(represent missing values for object or string).

The *rating_denominator* was checked and issues like some values greater than 10 and the *rating_denominator* column also was checked were some issue like the values way above the acceptable value for numerator

`image_prediction`

The head of the dataframe is being checked which is the first six columns, the shape of the dataframe is checked, the info, null values was checked, the summary statistics was checked and duplicates values. No null values was spotted, no duplicates values also.

The *p1*, *p2* and *p3* columns was checked by using the value counts and inconsistent names was spotted which are some names having underscore, entirely lowercase and some not entirely in lowercase

`additional_tweets_info`

The head of the dataframe is being checked which is the first six columns, the shape of the dataframe is checked, the info, null values was checked, the summary statistics was checked and duplicates values. No null values was spotted, no duplicates values also.

Cleaning

Quality

`twitter_archive_enhanced`

Issue	Solution
timestamp datatype is in object instead of datetime	the datatype of the timestamp was changed to datetime using pandas datetime
The columns such as the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id needs to be drop	the columns were dropped because they were not needed for the analysis
the column *name* contains value such as None, a, by and so on	the column *name* value were all being dropped because it impossible for a dog to name a single, double characther and not having no name
the rating_denominator have some values greater than or less than 10 which should be all 10	the rating_denominator with values greater than or less than 10 was being replaced with 10 because the values are ranking and is based on 10 for the denominator
the numerator_rating have some values that are way above the scale which is something like 920 and so	the *numerator_rating* columns with such odd or above the scale values was being replaced with the median

`image_predictions`

Issue	Solution
The column p1 and p3 contains values that are inconsistents which are some name written in underscore, some not in underscore, some the entire letters in lowercase and some are not in entirely in lowercase	the columns are clean by replacing names match together with underscore with space and all names being in the same format which is the first letter

in captial letter

`twitter_clean`

Issue	Solution
the datatype of the source_name and source_link should be category instead of object	the datatype of the columns source_name and source_link was changed the category and this occurs after the tidiness has being solved

Tidiness

Issue	Solution
The column *source* in the `twiter_archive` should table be splits into two new columns called source_name and source_link	The column *source* was split into two different columns which are the *source_name* which contains the name of the source sending the tweets and *source_link** the link to the source sending the tweets

Issue	Solution
The columns such as doggo, floofer, pupper and puppo in `twitter_archive_table` should be combine into a single columns called dog_stage	the columns are converted into a single column called *all_dogs* which is done by first adding the columns into together the form one column and the values of the column are counted which are addedd into a new single column called *dog_stage* and *all_drogs* columns was dropped hereafter.
The *twitter_archive*, *image_predictions*, and *additional_tweets_info* table needs to be merged	the three tables are merged together by using the Pandas merge function and was named *twitter_clean* table which was used to solve a quality issue which then was copied to the *twitter_archive_master* and then saved as *twitter_archive_master.csv*

After all this processing the dataset is now in a single table called **twitter_archive_master** which is then analysed.