We Rate Dogs

Data Gathering, Assessing

& Cleaning

1- Data Gathering:

In this phase, we gather data from 3 different sources with 3 different formats.

- Downloading the twitter archive file manually which contains the data archive of that twitter account, as its format is (.csv).
- Getting the image prediction data file from a URL using requests, as its format is (.tsv).
- Using twitter API to get retweets and favourite counts of the required tweets and download them in a text file (.txt)

2- Assessing:

In this phase, we extract issues from the data to be cleaned.

Visual assessment: Using info(), value_counts, head(), and other pandas or data frame functions.

Programmatic assessment: Using Jupiter notebook and pandas' functions.

2.1 Quality: We focus on the data accuracy, completeness, consistency... etc.

achive_df:

- Time stamp column has "+0000" at the end of each value which is better to be removed.
- Time stamp column should have a datetime dtype instead of string.
- Tweet_id column should be string instead of int to make it easier to join the archive table with the api data frame table on the tweet id column which is common in both tables.
- The expanded URL column has missing values which means that these tweets have no images and should be removed.
- Some values in retweeted_status_id column have nan values which means that they are not retweets, so those who have no nan values will be removed.
- There are some values in in_reply_to_status_id which means they are replies and not original tweets so they should be removed.
- In text column, some tweet's text has "&apm;" instead of "&" and should be replaced.
- In name column, some names are lower case, and some have first uppercase letters or are not names.
- In name column, some values are None instead of Nan as it should be replaced if we need to perform any Nan operations on them or even found them calculated as missing values in .info () method, ex: dropna, fillna, etc.
- In name column, some dog names are wrongly extracted as they are written as none but in the text column.
- Some ratings are overrated as a dog is 660/10 rated, so it is better to scale or normalize these high ratings to a boundary limit ex: rating above 30 will be scaled down to 30.
- Tweet id 786709082849828864 has wrong rating (wrong float extraction).
- Some ratings are not out of 10 so it is better to scale down these rating denominators to 10.
- There is wrong rating for the tweet_id 810984652412424192 as it specifies that the dog smiles 24/7 and it is considered as rating.

2.2 Tidiness: focusing on the data frame structure, ex: rows, columns.

archive_df:

- there are many columns for dog types --> instead it should be a column called dog_type and filled with the required type.
- archive_df and api_df are better to be as one unit (table) which contains the data of each tweet.
- we are focusing on original tweets only so the retweet's data columns (retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp) should be removed and replies data columns also (in_reply_to_status_id, In_reply_to_user_id)

image_df and archive_df:

- we should remove any tweets that its image is not in the image_df table, by merging the merged archive and api data frame by image prediction data frame on the intersected tweet ids.

image_df:

- using the first prediction of image as the appropriate one, by taking the true values of its prediction and dropping all unwanted other prediction columns

3- Cleaning: "fixing the extracted issues"

This phase consists of 3 main parts:

- **Define:** Defining the issue and suggest a solution.
- **Code:** Writing the code which will solve the issue.
- **Test:** Testing code that ensures that the solution is done.