WRANGLE REPORT

As part of this project I did Data Wrangling, here I started with gathering data from many sources and in several formats.

1. First Dataset is downloaded manually, it is a .csv file named 'twitter_archive_enhanced.csv' and stored it in 'archive' table
2. Secondly, I used the Requests python library to download programmatically a '.tsv' file named 'tweet-image-predictions.tsv' and I stored it in 'images' table. This file holds the results of a Neural Network's analysis which predicts a dog's breed based on images.
3. Then, I wrote an API object that I used to programmatically download a JSON file contains additional Twitter data stored as 'twitter_counts' table.

After this part, I did Data assessing.

- Denominator is not 10 for 23 tweets
- Unnecessary HTML tags in source column in place of utility name.
- Erroneous datatype: timestamp and retweeted_status_timestamp should be DateTime object
- Erroneous datatype: *tweet_id* should be of String datatype
- *Doggo*, *floofer*, *pupper*, and *puppo* should be categories

Twitter_counts Table:

1. Erroneous datatype: *tweet_id* should be of String DataType
2. Then I examined tidiness issues, Tidiness issues pertain to the structure of data. The requirements for tidy data are:
    - Each variable forms a column.
    - Each observation forms a row.
    - Each type of observational unit forms a table.

Images Table :

- Erroneous Datatypes: tweet_id should be of String Datatype.
- ('p1', 'p1_conf', 'p1_dog', 'p2','p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog')

  Columns names aren't informative.

  Names should be changed to become more informative.

- Erroneous Datatypes: p1, p2 and p3 should be categorical.

Tidiness issues that were found:

- For the archive table: *doggo, floofer, pupper, puppo* should be categories of a single variable named "*dog_stage*".

- Archive and twitter_counts can be consolidated into a single table for which the observational units are tweets. Images can be left as-is, because images are the units of observation.

Using Python and its libraries, I structured and cleaned dirty data in the final section of the wrangling process into the desired format for better analysis and visualisation. I have defined the actions to take for each identified issue before translating those actions into lines of code. I also tested each code to check the cleanup result.