Wrangler Report:

1. Gather :

We Gather Data Via 3 Methods.

- Downloading twitter_archived_enhanced.csv from udacity project page.
- Programitically hitting the site and downloading image_predictions.tsv from the link provide in Udacity Project.
- Programitically hitting API and getting the info like retweets count and favorites count.

2. Assessing

Quality issue

- In df_twitter dataframe, columns like (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp) contains mostly null value and cannot be used for analysis.
- In df_images dataframe, img_num column is not required.
- Time stamp data type is object instead as date time.
- Few names are mention as article 'a', 'the','an'
- rating_denominator column for one of the record is '0' which is i am assuming is because of some error.
- Count of tweeter id in all the dataset is different.
- Numerator and denominator rating are not correct.

Tideness Issue

- Doggo,floofer,pupper,puppo is used as columns.
- There is 2 columns rating_numerator and rating_denominator which can be merged as 'rating' instead.
- The columns for dog breed predictions can be condensed.
- df_images and df_tweets, df_twitter can be merged into one dataset.

3. Cleaning:

- Issue 1: Timestamp column has data type as object(string)
  Define: Change data type as ‘time_stamp’

- Issue 2: Incorrect names like name as ‘a’,’an’,’the’ etc
  Define: replace incorrect name with None
- Issue 3: rating_denominator and rating denominator value are incorrect.
  Define: programitically replaces the correct value.
- Issue 4: Condesing stages of dog.
  Define : Created a stage row in df_twitter dataframe.
- Issue 5: Inncorrect data type for stage column
  Define: Convert object data type to categorical data type.
- Issue 6: There is 2 columns rating_numerator and rating_denominator which can be merged as 'rating' instead
  Define created a new columns called rating whose value will be rating_numerator/rating_denominator.
- Issue 7: Remove unnecessary row like ‘text’,’url’ etc
  Define : Drop the column from data frame.
- Issue 8: merge p_conf and p to find the breed of dog
  Define : Create a column for breed of dog and store a value using confidence interval.
- Issue 9: convert breed to categorical variable.
  Define: Convert object data type to categorical data type.
- Issue 10: multiple data frame
  Define: Merge all 3 data frame and create in df_clean.

Stored the data in twitter_archive_master.csv after processing.