Data wrangling Report

Gathering:

- I imported twitter-archive-enhanced.csv, by pandas.
- Imported image-predictions.tsv by requests library then created path using os library.
- Collected data from tweet-json.txt without tweeter developer account, then created panda data frame and saved as favs_tweets_data.

Assessing and Cleaning:

Visual assessing for the csv files done by excel file.

Tidiniess and missing data:

- **Twitter_archive_enhanced:**
  - dogs classifications (doggo, puppo, etc….) are distributed over three columns while it should be in one column using melt that gives value observation which is done under the name dog classification.
  - Dropped columns like (retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp), are not complete(empty), removed as they don't add value.
- **Image_prediction:**
  - columns of p1, p2, p3 and p1dog,..etc related to pics prediction were reformatted to have two columns one for the breed type and other for confidence, the reformatting done by for loops based on priorities for p1 = True if not then p2=True if not, then p3 = True otherwise the prediction is failed then concatenated all together.

    - The resulted two columns of prediction diverted to Dataframe then merged with **twitter_archive_enhanced** under name **archived_breeds_clean.**
    -
- **favs_tweets_data:**
    - tweets and favorite counts merged with **archived_breeds_clean** forming **archived_breeds_favs_retweets** final dataframe which I used for quality assessing and cleaning.

Quality:

- missing some names (745 of none values), replaced None with nan.
- Ratings numerators out of range for the rating_numerator as max is 1776, neglected all values bigger than 20, I took the numerator as a scale over ten.
- timestamp type is not correct, converted to to_datetime type to be accessible.
- tweet_id should be string , converted using as type.
- Removed duplicated tweet_id .
- columns of p1, p2, p3 and p1dog,..etc need to be more discriptive (already done in tidiness).
- Dropping non predicted breeds and replacing them with failed prediction.

Final DataFrame was stored as **archived_breeds_favs_retweets_final.csv**

And conducted visualization and analysis.