Wrangling Report

Data wangling Steps:

- Gathering
- Assessing
- Cleansing

1-Gathering

- The first file **“twitter-archive-enhanced.csv” imported with pandas CSV read method “pd.read_csv() “and the data stored in a new data frame “twitter_archive”.**
- The second file **“'image-predictions.tsv” imported also with pandas CSV read method “pd.read_csv()” with additional parameter in the method “sep='\t' ”and the data stored in a new data frame “image_predictions”.**
- The third file was **“tweet_json_copy.json” imported with pandas CSV read method “pd. read_json ()” and the data stored in a new data frame “tweet_jso” , I didn’t use twitter API .**

2-Assessing:

- **I started assessing every data-frame by looking by using head() function to see sample of the data.**
- **I start looking on every column in every data-frame look on its values and its unique values to see if I can use it in my analysis or it have a useful meaning.**
- **Also investigate every data type for every column to see if it need to be converted to another suitable data type.**

3-Cleansing:

Quality issues:

- **twitter_archive.source fetch the source from the URL and remove string outliers.**
- **Drop useless columns [retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_user_id, expanded_urls, text] from twitter_archive**
- **Convert twitter_archive['timestamp'] to date time.**
- **Drop useless columns [p2,p2_conf,p2_dog ,p3,p3_conf,p3_dog,img_num,jpg_url,p1_conf,p1_dog] from image_predictions.**
- **Drop all columns except tweet_id, retweet count, lang**
- **The name attribute has to issue 577 record with 'NON' replaced with 'No-Name', and 57 with only 'a' Litter replaced with 'Abby'.**
- **rating_numerator can't be bigger than rating_denominator so any rating_numerator bigger than 10 it turned to 10 .**
- **filter p1 column in image_predictions to remove the uncleaned values that not belong to any dog type .**

Tidiness:

- **combine: puppo, doggo, floofer, pupper columns in one column called dog-stage.**
- **Convert all values in image_predictions['p1'] to lower case.**
- **Rename tweet_jso['id'] to tweet_jso['tweet_id'] .**
- **Then merge all the data frames in one called twitter_archive_master**