Wrangling Efforts Of WeRateDogs Data Analysis

Data Gathering :

- First, I imported the libraries needed which were; pandas, json, requests, os, numpy, matplotlib, and matploblib magic lines.
- Second, I downloaded all the files provided by Udacity programmatically which were; twitter-archive-enhanced.csv, tweet-json (in case of need), image predictions.tsv, twitter.api.py, twitter.api.rtf using **requests** and **os libraries** to request the links and write them locally to the laptop.
- Third, I ran the tweepy script provided by Udacity and saved the tweets gathered into a txt file named tweet-json.txt.
- Fourth, I ran a for loop to read the contents of the txt file using **json library** with function **load** and made a list with the contents to load them into a pandas dataframe.
- **Finally**, I've loaded all the previous files, twitter-archive-enhanced.csv, image_predictions.tsv, and tweet-json.txt into pandas dataframes called archive_df, image_predicitons_df, and api_now_df.

Data Assessment :

- First, I opened the twitter-archive-enhanced.csv file in **excel** to assess it visually and found out that the names column has false values like; None, a, aa, and aaa.
- Second, I moved on to programmatic assessment which was by exploring each dataframe using the methods, **head(), info(), shape(), and list(columns).**
- I found out several quality and tidiness issues, 11 quality issues and 5 tidiness issues to be particular. They're list below and categorized.
- **Quality Issues :**

  archive_df:

  1. Timestamp is in **object** type not datetime
  2. Tweet IDs in **int** type, not **str.** There's not point of having them as integers because there will be no mathematical operations done with them.
  3. Found out that there were some retweets and replies among the dataframe while the project criteria was having original tweets only within the analysis.
  4. Columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,

retweeted_status_timestamp) weren't needed as they won't contribute in the analysis.

5. Found out that the source of the tweets was tagged in a href tag.
6. There were tweets with missing images, so they need to be disposed of.
7. NaN values in dog stages were the word "None" while they shouldn't be as this is considered as a string.
8. Dogs' names column was including some false values like; lowercase words that weren't actual names(Assessed Visually).
9. Some of the Rating Numerators were incorrectly extracted.

api_now_df:

1. Tweet IDs were int and they shouldn't be integers for the same reason mentioned before in archive_df. Also, if they stayed this way, it'll affect later on the merge process with archive_df.

image_predictions_df:

1. Retweets which were removed from archive_df needs to be removed from this df as well.
2. Tweet IDs should be converted to str.
3. Dropped predictions with False results in the 3 predictions.

Tidiness Issues:

archive_df:

1. Dog stages should be one column not 4.

api_now_df:

1. Df should be a part of archive_df to help with the analysis more.

image_predictions_df:

1. Df should be a part of archive_df

Data Cleaning:

- **I followed the Define , Code, and Test work flow but then removed the testing lines later on to reduce the space I have to scroll within the notebook each time I want to reach something previous.**
- **In the below table will be each quality issue and it's resolution with the methods used.**

DF Name	Quality Issue	Resolution	Method
archive_df	Time stamp wrong dtype	Converted the column using column indexing	pd.to_datetime() function
	Removing tweets with missing images	Acquired the tweets in the image_predictions_df to use them as a guide by comparing the tweets in both dfs and excluding any other tweet.	list,unique, indexing
	Tweet IDs wrong dtype	Converted the column using column indexing	astype() function
	Removing retweets and replies from archive_df	Removed them by indexing the dataframe excluding any row that has a value in the retweeted_status_id column and in_reply_to_status_id column	indexing, masking, notnull
	Removed the unnecessary columns	Dropped the 4 column of the dataframe	pd.drop()
	Extracting sources from href tags	imported BeautifulSoup library, made an empty list, ran through each row and extracted the source from the tags, added the extracted value to the list, and defined the source column using the list	BeatifulSoup library, for loop, find, append, indexing.
	None values in dog stages columns	Replaced all the None values in the dog stages columns with NaN values.	indexing using loc, replace
	False names in Dogs' Names column	Replaced all the lowercase values of being false with NaN values	indexing, replace, np.nan
	Decimal ratings and another value were extracted incorrectly	Replaced them by their indexes with the correct values	Indexing, regex, replace
api_now_df	Tweets IDs wrong dtype	Converted the column using column indexing	astype() function
	Tweets IDs wrong dtype	Converted the column using column indexing	astype() function

image_predictions_df	Removing retweets that was originally removed from archive_df	Compared the values in this df and archive_df and removed any rows that weren't common	indexing, np.logical, isin, list, masking
	Removed the predictions which had False bool in all the 3 predictions by querying the df for the false values and dropped them using indexing	query, indexing, drop	

- **Below will be the tidiness issues table.**

DF Name	**Tidiness Issue**	**Resolution**	**Method**
archive_df	Dog stages should be in one column	Appended all the columns in a new column, extracted the names from the NaN values, and deleted the original stages columns	Indexing, iloc, extract, regex, drop
api_now_df	Mergin the df with archive_df	Merged the dataframe with archive_df as it will be helpful to have it merged for analysis using merge and conerted the retweet and favorite counts to int as they were merged with float dtype	merge, astype
image_predictions_df	Merging with archive_df	Merged the dataframe with archive_df as it will be helpful to have it merged for analysis using merge.	indexing, masking, melt, drop

Finally: I saved the master file to twitter_archive_master.csv