WRANGLE REPORT BY ABDULLAHI MOHAMMED AHMED

This report was made to highlight my data wrangling process for tweets from the user @dog_rates.
This account rates tweets of dogs with humorous and occasionally viral tweets.
My wrangling process included 4 core stages:

- Gathering the data
- Assessing the data
- Cleaning the data
- Visualizing the data

GATHERING THE DATA:
The data was gathered from three different sources
1. I downloaded and uploaded twitter_archive_enhanced.csv file and read it into a pandas dataframe
2. I downloaded the image_prediction.tsv file from the provided url using the request library
3. From Tweeter API and its access library called Tweepy to get the retweet counts and favourite counts of tweets. After which I read the resulting text file line by line into a pandas dataframe.

ASSESSING THE DATA
After gathering the data from their respective sources into the notebook. It was time to assess the data for data quality and tidiness issues. I first created a copy of each dataframe before employing both visual and programmatic methods to assess the data and found out 10 quality issues and 4 tidiness issues

i. Twitter archive file copy was renamed tweets
ii. Image predictions copy was renamed imgs_copy
iii. Df_json copy was renamed json_tweets
iv. Tweets had 2356 rows ad 17 columns
v. Imgs_copy had 2075 rows and 12 columns
vi. Json_tweets had 2354 rows and 4 columns

QUALITY ISSUES:

1. Source column in twitter_archive has hyperlinks:
   Source columns had html hyperlinks which need to be cleaned.
2. Twitter_archive name column having 'a' as dog names:
   Name column had invalid values which needs to be rectified
3. Twitter_archive expanded_urls column contains repetitive urls:
   Repetition of urls in expanded_urls.
4. Twitter_archive timestamp column is of wrong datatype:
   Timestamp column is of object datatype
5. Twitter_archive has some retweets:
   Retweets are duplicates, needs to be handled
6. P1,P2, P3 columns in image_predictions have underscore in the names
   These columns contained values with underscores
7. Tweet_id column is of integer datatype instead of string:
8. Tweet_archive ratings denominator above 10:
   Some ratings denominator were above or below 10
9. Rename p1 column to dog_breed:
   p1 column in image prediction dataframe will be renamed

10. Remove rows from imgs_copy dataframe that arent dog breeds
    Some rows in the imgs_copy dataframe have value of False in the p1_dog column indicating it isnt a dog breed hence irrelevant to our analysis.

TIDINESS ISSUES
1. Json date column is not tidy, needs to be cleaned
2. Dog_stages column created from four dog stages column in tweets dataframe
3. Dropping unnecessary columns
4. All three dataframes needs to be combined into one.

CLEANING THE DATA

Using Define - Clean - Test framework I was able to rectify most of the issues outlined above and exported the final dataframe to a csv file named twitter_archive_master.csv

- ✓ Created a new column for the source links without the hyperlinks
- ✓ Replaced 'a' values with 'Unknown'
- ✓ Split repetitive_urls columns and selected only the first item as the new value
- ✓ Used pd.to_datetime to convert timestamp column to appropriate datatype
- ✓ Created a subset of tweets dataframe extracting only original tweets
- ✓ Replaced underscores with spaces
- ✓ Converted tweet_id columns datatype to string
- ✓ Renamed p1 column to dog_breed
- ✓ Selected subset of imgs_copy dataframe by querying rows where p1_dog is True
- ✓ Created day, month and year columns from date column json_tweets dataframe
- ✓ Created a single column containing all four dog stages by creating a dataframe containing all four stages and tweet_id, then proceeding to unpivot these columns using pd.melt().
- ✓ Removed in_reply_to_status_id,in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp,doggo,floofer,pupper,puppo,variable columns from tweets_main dataframe.
- ✓ Merged all dataframes into one dataframe and exported as a csv file.