Reporting: wrangle report

Objectives

Data wrangling is a process that entails gathering, assessing and cleaning data. These form the objectives of the data wrangling process of the project:

- **Gathering** - Collecting and loading data from various endpoints
- **Assessing** - Visually and programmatically looking at the data for various issues
- **Cleaning** - Correcting the issues from the assess stage

This process is also iterative.

Step 1: Gather Data

In this step, data from three different points were to be gathered using various methods:

1. `twitter_archive_enhanced.csv` which was manually downloaded and loaded into the notebook as required.
2. `image_predictions.tsv` which was loaded to the notebook programmatically.
3. Retweet count and favorite data accessed from the Twitter API through Tweepy library and saved to `tweet_json.txt` and loaded to the notebook. This was possible through a script that used tweet IDs from archive data to get the data required seemlessly.

Step 2 and 3: Assessing and Cleaning Data

When assessing data, one faces issues to do with quality and tidiness. This is done through visually or programmatically assessing the data.

Quality Issues

Dataset	Observation	Solution
Archive data	Data contains replies and retweets instead of orginal tweets	Removed retweeted and reply tweets and kept original tweets only
	The columns `doggo` , `floofer` , `pupper` and `puppo` have `None` representing missing values	Changed `None` values to np.nan values
	`timestamp` is object data type instead of datetime	Data type changed from object to datetime data type
	`text` has links in them	Removed the links
	The `rating_numerator` has incorrect values and datatypes	Extracted the numerator rating values again from the text column
	The `rating_denominator` has incorrect values	Extracted the denominator rating values again from the text column
		Extracted the source values again from

	The `source` column values are closed within `<a>` tags	`<a>` tags in the source column
	Some columns are not necessary for analysis	Removed the unnecessary columns
Image predictions data	There are duplicated image url's in `jpg_url`	Removed the duplicated image url rows

Tidiness Issues

Dataset	Observation	Solution
Archive data	The columns `doggo` , `floofer` , `pupper` and `puppo` should be in one column i.e `dog_stage`	Melted the four columns into one column
Image predictions data	The columns `p1` `p1_conf` `p1_dog` `p2` `p2_conf` `p2_dog` `p3` `p3_conf` and `p3_dog` should be in two columns i.e `breed` and `conf`	Picked the greatest true p1 confidence level value and corresponding dog breeds into new columns while dropping these columns
General	All datasets should be combined into one dataset	Merged all datasets into one dataset using tweet ids

Results

The result was a final dataset that merged data from the three sources after effectively cleaning the data. This data was stored into a csv file called `twitter_archive_master.csv` .