Data Wrangling Report

Introduction

Real-world data rarely come clean. Using Python and its libraries, I gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling.

The dataset that I wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. And the numerators almost always greater than 10. 12/10, 13/10, etc. Why? Because "they're good dogs Brent."

- ## Gather Data

Data is gathered from three different resources:

1- From a given file:

Use pandas read_csv() to read data from an existing file called twitter-archive-enhanced.csv and save it as twitter_archive.

2- Download file using Requests library and URL:

Download file image_prediction.tsv programmatically from the Internet and save it as image_pre_df.

3- Gather data from Twitter API:

By using Tweepy library get and save the data in json_df.

- ## Assess and Clean Data

Assess data process done first programmatically and visually to detect the data defects, then I correct them in the cleaning data process. To clarify the quality and tidiness issues that I found look at the tables below:

Quality Issues	Issue Solution
1- In tweeter_archive dataset, incorrect data type and values for in_reply_to_status_id and in_reply_to_status_id.	Convert data type for in_reply_to_status_id and in_reply_to_status_id from float to integer by first replace NaN to 0s and then convert it to integer.
2- In tweeter_archive dataset, incorrect data type for the timestamp.	Convert data type for timestamp to datetime.
3- In tweeter_archive dataset, source is difficult to read. So, change the source to more readable content.	Replace the tag and URL to four readable sources.

4- In tweeter_archive dataset, remove retweets.	Remove the retweets by select only the records where 'retweeted_status_id' is null.
5- In tweeter_archive dataset, incorrect values for the rating_numerator it has to be more than 10.	Remove all rows which have rating_numerator <10 to keep only rows that has numerator more than or equal to 10.
6- In tweeter_archive dataset, incorrect values for the rating_denominator it has to be equal 10.	Remove all rows which have rating_denominator not equal to 10 to keep only rows that have denominator equal to 10.
7- In tweeter_archive dataset, incorrect dog names (a, an, such, the, etc.)	Replace all lowercase values of name column with None to remove the (a, an, such, the, etc.)
8- In tweeter_archive dataset, wrong represented of nulls value as None.	Replace all None in 'doggo','floofer','pupper'and 'puppo' culomns with NaN.
9- There are many tweets with no images, and we only care about ratings with images.	Drop the rows in twitter_clean which don't have dog image.
10- In tweeter_archive dataset, there are many dogs have more than one name.	For those dogs which has more than one name, I'll separate between these names by a comma.

Tidiness Issues	Issue Solution
1- In tweeter_archive dataset, remove retweets because it duplicates. and we only care about original ratings. So, the other retweets columns are not useful.	Drop all retweets columns.
2- In tweeter_archive dataset, the doggo, floofer, pupper, puppo columns show one variable. So, it should be merged into one column called stage.	Merge the doggo, floofer, pupper, puppo columns and put them in a new column called 'stage' and then drop these 4 columns.
3- retweet_count and favorite_count columns from json_df should be part of the tweeter_archive dataset.	Join twitter_clean with json_clean by tweet_id to make twitter_clean data set have retweet & favorite columns.

- ## Store Data

  After cleaning data, I stored it in a file called twitter_archive_master.csv file.