In this case, I am going to clean the data and make a master dataset by combine 3 datasets.

During data assess processing, we find 8 quality issues and 2 tidleness problems, some quality problems are occurring in more than one datasets, and other problem was occurring in some specific datasets.

Common issue:
-Incorrect datatype
(I use astype() function to make datatype correct. However, at the beginning, I tried to clean all the datatype in all three datasets, but tweepy dataset's object cannot match the CSV and TSV dataset. As an alternative option, I keep all tweet id as integer, and successfully merge all datasets. Then, I convert tweet id of master dataset.)

-Duplicate value occurs in the columns that requires distinct value.
(Use drop_duplicate() function to identify which column requires unique value(such as tweet_id and url, and filter all duplicate values, and make a test to make sure all rows with duplicate values are filtered)

Quality issues in WeRateDogs:
-filter all non-null value of retweet status id and retweet to id
(Here, I use dataframe filter to remove the rows does not meet the requirement, which is .isnull() == True)

-clean numerator and denominator number
(From project motivation we know that the values of these two columns are not always accuracy. From the text sample we know that there is always a fraction number included in the sentence, the format is numerator/denominator. However, in order to make sure all conditions are considered, I will use a critical label to mark the suspicious one. After all set, We can filter the critical row to manually add their numerator and denominator.)

-We found that there is 0 in rating denominator, which cause divide 0 error.
(use dataframe filter to remove the rows that demoninator is 0.)

-Remove the columns that does not use in the case.
(Here, I create a sub-dataset of this dataset. By the way, some of columns are not final version of dataset because some of columns are useless in the final version, but still need to use after this step.)

Quality issues in ImagePrediction:

-Some confidence number are greater than 0
(From the maximum number of p2_conf and p3_conf, we know that even the maximum number of its confident is less than 10. Hence, each value in the column divide 10, and use new data series to cover the old one. Then, we test if there is any value larger than 1, and the result is no.)

Quality issues in Tweepy:

- Convert JSON to dataframe, and remove irrelevant columns (solved in assess data part)

Tidleness issues in WeRateDogs:

-Combine dog type columns
(Here, in order to avoid any potential risks, I use the same strategy in cleaning numerator/denominator columns. I analyse the raw text of each tweet, to find the tweet is describe any specific dog types. For those cannot be clearly confirmed, I give them critical label. And then I manually find the dog type information in the tweet sentence to confirm the dog type.)

-Combine numerator and denominator as rating
(Because we already clean the suspicious value above, so we just need to use numerator divide denominator)

PS: After tidleness issues done, all irrelevant columns such as dog type dummy column and numerator/denominator will be removed.

Finally, we combine three datasets to one dataset on column tweet_id, after all set, convert the datatype of tweet_id from integer to string(object).