Data Wrangling-Twitter Data-WeRateDogs

I decided to work with this data in my own jupyter notebook environment. First I had to download the 'twitter-archive-enhanced.csv', which to me was a csv file. Then I had to download 'image-predictions.tsv' using the url. This was done programmatically using the request library.
The twitter data was downloaded as a json file and we needed to import tweepy . I had to extract the list of tweet_id through a loop and ,each tweet_id and query twitter API to get each tweets json data. This tweet got saved in 'tweet-json.txt'.After the code was executed the data was saved in a text file.I read the text file line by line which was then appended to an empty list.
Lastly I converted the dictionaries into a dataframe:'api'

Now we have all the tables ready for assessment . I used .info(),.describe(),.head(),.tail(),.value_counts() to assess the tables . They were few quality and tidiness issues identified .

Quality issues	How I resolved it
-keep original tweets(no retweets)	Used isnull() on retweet_status_user_id
-Error in dog names (e.g a,an,actually) are not a dog's name.	Used .unique() to find out the unique dog names. ANd the error names to 'None'
-Erroneous data type fix	Changed tweet_id to str and source to category
-timestamp to make datetime	Changed timestamp to datetime
-drop columns not needed for	Dropped columns which were

our analysis **-doggo, floofer, pupper and puppo columns in 'final' table** **-URLs are not clear** **-Drop tweets with no images**	**not required. Used .drop()** **Created a new column for dog stage. Created a def function .** **Removed long urls** **SOme tweets had no images so dropped that row.**

<u>Tidiness</u>	**<u>How I resolved it</u>**
-joining 3 dataframes **-drop doggo, floofer, pupper and puppo columns** **-clean the source column**	**Used merge() to join all tables. Used left join** **Dropped 4 columns since it was not required now** **Used extract() to clean up the source column.**