Wrangling Process for the WeRateDogs Dataset

I have undertaken the following steps while wrangling the data:

1) Data Gathering
2) Data Accessing
3) Data Cleaning

Let's start from Step one which is **Data Gathering**:

I used the normal read csv command for getting the twitter archive dataset.

Next, I went ahead with using requests module for obtaining the image prediction dataset which consisted of the predictions if the image contains a dog or not and what breed of dog is present.

Moving forward I utilized the tweepy module for setting up the API for pulling data from the tweeter ids in the twitter archive dataset. This step was the most challenging as I had to learn many new things such as the basics of tweepy module and also file writing operations using json.dump function to write the new data I obtained into a json_text.txt file.

Next was **Data Accessing**:

In this step I started off by using the info, head, sample functions which helped in ascertaining the details present in each of the datasets. I observed the different issues each data set had and what measures could be taken in order to fix them, following is a list of all the issues I found in the datasets:

For twitter archive dataset:

1)Unnecessary columns which won't be of much use later on such as.

2)Missing values in some of the columns.

3) Redundancy in the dog stages columns, making it difficult to analyze. Bringing that down to one column

4)tweet id column was of integer datatype, we anyways cannot perform any mathematical operations on the id as it doesn't make sense to do so, so we should convert it to string to avoid doing it inadvertently.

5) Some of the dog names were invalid such as ‘None’, ‘a’, ’an’, ‘the’, ’by’, ’my’.

6)The denominators and numerators of the ratings had inconsistencies at places, such as rating less than 10, denominator being 0. Which was later fixed using string operations for obtaining the correct rating.

For imagepred dataset:

1)img_num column was redundant and dropped later.

2) prediction columns were having multiple issues with their instances, such as underscore instead of spaces, also uppercase and lowercase were used without any pattern.

3) There were some instances with all three predictions being False meaning there is no dog in the images so should be dropped.

4) There were instances with multiple image urls meaning its either retweeted or a separate tweet with same image.

For api dataset:

1) It had no issues the isnull function was used along with seaborn’s heatmap for checking the null values in the dataset for the visual aspect.

Finally, **Data Cleaning** was done:

1. Remove the rows that contain the retweet data

- Some of the dogs are classified by using more than one dog stages, the columns could be brought down to one.

- Remove the unnecessary columns

- Incorrect numerator extraction issue

- Cleaning the invalid dog names

- Dropping columns where all predictions are false, and ones where p1 is false and p2 and p3 are less than 30%.

- Fixing the datatypes

- Merging the datasets

- Making the denominator uniform

- Fixing the inconsistencies of the predicted_breed column and optimising the breed prediction columns