WeRateDogs' Wrangling Report

A full report of my Data Wrangling process on the WeRateDogs' data.

My notebook included these sections :

1. Data Wrangling :
    - Gathering the data from multiple sources
    - Assessment of the data
    - Cleaning the issues
2. Storing the data
3. Analyzing and visualizing

1- Data Gathering

Here, I imported the data from different sources using different methods and libraries.
Here is a brief description of my data gathering process :

- Twitter Archive Enhanced CSV : I downloaded the data manually from the classroom and then used Pandas library to read the data and store it in a variable as a DataFrame object.
- Image Predictions TSV : I used the Requests library to perform a *GET* request on the resource url for the data provided in the classroom. After writing response's content to a file, I loaded it as a Pandas dataframe
- Tweets JSON : Unfortunately, I couldn't use the Tweepy api for getting the data due to multiple obstacles (limited posts per month, 15 minutes rate limit and limited endpoints) which were all caused by the new access levels of the X api.
  As a result, I manually downloaded the text file containing json data
  Secondly, I imported the json data using pandas and assigned it to a DataFrame
  Lastly, I assigned the required columns of the DataFrame to a variable for usage in the notebook.

Data Assessment & Cleaning

In these two sections, I firstly assessed the data thoroughly using both Visual and Programmatic assessment and made a conclusion for all the issues I noticed in the data
In Data Cleaning, I fixed all issues mentioned in Data Assessment and ensured that I have clean data that can be depended on in analysis.
Below is a table of the addressed issues in Data Assessment and how I fixed them in Data Cleaning

Issue	Fix
Presence of duplicated rows (retweets)	I filtered the rows which are retweets using the 3 conditions I made and dropped them from the data
Missing / Inaccurate dog names	Since all the inaccurate names were lowercase, I replaced all missing names and lowercase names with ‘Unknown’ which indicates absence of name in tweet’s text
Missing Expanded URLs	Got automatically fixed in issue 1 after removing retweets
ID columns interpreted as integer in all DataFrames	Transforming the Tweet_id in the 3 DataFrames into a string and dropping other ID columns
Timestamp isn’t in date format	Transforming the column using pandas
Inaccurate dog ratings	Re-assign the ratings columns with the extracted rating from the text. In addition to manually fixing some inaccurate ratings.
Missing Dog Stages	It got already fixed in issue 12 where I replaced all dog stages columns with one
Inconsistency of predictions	Making all predictions in title casing
Underscores in predictions	Replacing underscores with a space
66 duplicate images in predictions	Dropping the 66 duplicates
Missing values in likes / retweets	Filling with the mean of the column
All dog stages columns are one variable and must be in one column	Replacing all dog stages columns with one column for the dog stage extracted from the tweet’s text
Timestamps can be splitted into other values (Year,month,day and day name)	Using .dt accessor and making 4 new columns for these values. Then, removing the timestamp
Existence of tweet’s link inside its text	Filtering the tweet’s link and assigning it to its own column, then removing it from the text
The img_num column isn’t needed	Dropping the column from the predictions
All of our DataFrames are on one topic and must be in one dataframe	Merging the 3 DataFrames on the tweet_id using left method for keeping matches in the archive only