Data Wrangling Report

Project objectives

The project main objectives were:

- Perform data wrangling (gathering, assessing and cleaning) on provided thee sources of data.
- Store, analyze, and visualize the wrangled data.
- Reporting on data wrangling efforts and data analyses and visualizations.

Step 1: Gathering Data

In this phase, the three pieces of data were gathered and represented as pandas dataframes:

- The WeRateDogs Twitter archive (file on hand, manual download of 'twitter-archiveenhanced.csv')
- The tweet image predictions ('image-predictions.tsv'). This file was being downloaded programmatically using the Requests library from a provided URL.
- Twitter API & JSON:(In my case I did the following) file on hand, manual download of 'tweet-json.txt' and read the file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Step 2: Assessing Data

Assessing data Once the three tables were obtained I assessed the data as following: Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv files in Excel. Programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, etc). Then I separated the issues encountered in quality issues and tidiness issues.

Step 2 and 3: Assessing and Cleaning Data

- **Quality issues**

archive dataset

The issue	The solution
Delete columns that won't be used for analysis.	Drop columns
Incorrected ratings on numerator and denominator columns.	Some values were checked, then modified manually. Also remove the large values for the numerator .and remove denominators not equal to 10
we keep only the original ratings (no retweets) that have images.	Delete retweets by filtering the NaN of retweeted_status_user_id
Name column have invalid names i.e 'such','the','a','an'.	Change error name in dog name to None
Incorrected datatypes on tweet-id Convert to string instead of integer.	Changing the variable type to string
Some rows have more than one dog stage.	Separate the dog stage to know which records have more than one dog stage and store the multiple dog stage with " multiple"
Incorrected datatypes on numerator and denominator columns Convert to float instead of integer.	Changing the variable type to float

images dataset

The issue	The solution
There is 2075 rows in the images dataframe compared to 2356 rows in the archive dataframe.	Drop the rows with missing images
66 duplicate photos is probably because of the retweet.	This was the last step and was resolved through previous analyzes.
Delete columns that won't be used for analysis.	Drop columns

Incorrected datatypes on tweet-id Convert to string instead of integer.	Changing the variable type to string

twitter dataset

The issue	The solution
Rename column (id to tweet_id) for merge and to make it more descriptive.	Rename column to tweet_id
Incorrected datatypes on tweet-id Convert to string instead of integer.	Changing the variable type to string

- **Tidiness issues**

The issue	The solution
Create 1 column for image prediction -dog_type-.	Create dog_type column by using a function that takes the first "true" confidence level and prediction.
Create 1 column for dog stage instead of 4 columns (doggo, floofer, pupper, puppo)-dog_stage-.	Create column dog_stage by using function to check dog stages and register multiple values with "None"
All tables should be part of one dataset	Merge the three dataset in one data set called df_merge