Wrangle Report :

This project was part of the data wrangling section of the Udacity Data Analyst Nanodegree program. It is dedicated to data wrangling concepts. In addition to deploying skills like web-scraping, for gathering data from different sources. Assessing the quality of the retrieved data and cleaning it
WeRateDogs Twitter account tweets was analyzed using Python, documented in a Jupyter Notebook (wrangle_act.ipynb).

This project included:

- Data wrangling, which consists of:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing our wrangled data
- Reporting on 1) Data wrangling and 2) Data analyses and visualizations

Assessing the Gathered Data :

After gathering the data and reading it into dataframes, assessment comes next to check data quality and tidiness. I detected and documented the following quality issues :

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be strings instead of float, because we don't want any operations on them, they also have many NaN values.
- retweeted_status_timestamp, timestamp should be datetime instead of object (string).
- name, doggo, floofer, pupper, puppo columns include None values instead of NaN, but Pandas treats None and NaN as essentially interchangeable for indicating missing or null values so it won't be an issue.
- rating_numerator, rating_denominator are integers but prefered to be float.
- Some of the elements of the name of the dog are not real name like "a","None" "an", should be replacedby null-value.

- The timestamp column is an object. It has to be a datetime object.
- There are 2075 rows in the images dataframe and 2356 rows in the archive dataframe.
- In several columns, null values are not treated as null values.
- Missing values in 'name' and dog stages showing as 'None' instead of NaN.
- tweet_archive contains retweets
- Some tweets don't include images

Tidiness Issues :

- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo instead of one column filled with their values.
- Join 'tweet_info' and 'image_predictions' to 'twitter_archive'
- We can see that all three columns share tweet_id column, all are the same type so we can merge these dataframes on that column.

Data Cleaning :

Cleaning data is the third step in data wrangling process. Where I fixed the quality and tidiness issues that were identified in the assessment step. It’s a very important step that ensures quality analysis with real and truthful insight.

It should be considered that maybe not all of the quality and tidiness issues were spotted and addressed in the assessment section but most of it had been done in the wrangle_act jupyter notebook.

Data wrangling is a core concept that every data scientist should be familiar with to draw truthful insights from the different datasets available in the world. Analyzing, visualize, or modeling unclean and messy datasets could lead to missing out on important insights or drawing wrong conclusions.