WRANGLING REPORT

Data wrangling can be defined as the process of cleaning, organizing, and transforming raw data into the desired format for analysts to engage in prompt decision-making.

This is a detailed report of my efforts in wrangling and analyzing The WeRateDogs Twitter Archive. Completing this project required these data wrangling processes:

1. Gather the data
2. Access the data
3. Clean the data.

Gathering Data

In this project, I had three different data to gather. I downloaded the files and read them into pandas dataframes.

1. Twitter-archive-enhanced.csv was downloaded manually and read into a dataframe as df
2. image-predictions.tsv was downloaded programmatically using the Requests library from a provided URL. The dataframe was assigned the name image_pred
3. I used Python's Tweepy library to query the Twitter API using the tweet IDs from the WeRateDogs Twitter archive, and then I stored the complete set of JSON data for each tweet in a file called twitter_data.

Assessing Data.

After gathering, I proceeded to access the three data visually and programmatically. During this assessment, I noticed the data had quality and tidiness issues. I made note of these observations and proceeded to rectify these issues in the cleaning phase. Below is a list of the issues I observed from my assessment.

- **Tidiness Issues**

1. image_pred and twitter_data2 should be part of the df table
2. The columns 'doggo, floofer, pupper, puppo' should be combined into one column called Dog Stage

- **Quality Issues**

1. Erroneus datatype (TimeStamp)
2. The column, source, is not concise
3. For this project, I don't need tweets beyond August 1st, 2017
4. Missing values in the columns (in_reply_to_status_id, in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp,expanded_urls)
5. Drop irrelevant columns: Retweets
6. NaN is the right way to indicate missing values in the dataframe. Replace "None" with NaN
7. Some of the column names are not descriptive
8. drop irrelevant columns:'jpg_url','img_num', 'p2','p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog'

Cleaning Data

In this phase of the data wrangling process, I addressed the issues noted during the assessment. I addressed the structural issues first (TIDINESS) before addressing the content issues (QUALITY). After the cleaning, I stored it as twitter_archive_main.csv (clean_data).

In the table below, I have tabulated a concise summary of the issues observed during the assessment and the cleaning measures applied to it. You can find details on my line of codes in my worksheet.

Assessment Issues	**Cleaning Measures**
image_pred and twitter_data2 should be part of the df table	I achieved this using the merge() method
The columns 'doggo,floofer,pupper,puppo' should be combined into one column called Dog Stage	I resolved this issue using the pandas series str. extract() function
Erroneus datatype (TimeStamp)	The appropriate datatype is "Datetime"
DRop columns with retweet information	Achieved this using the loc() function and subsequently the drop function.
Replace "None" with NaN	I used the replace() to resolve this issue
Remove irrelevant columns in image_pred2 table	I created one table for prediction (p1) and confidence level, therefore dropping the other columns
The column, source, is not concise	I got rid of the HTML tags using the str_replace() function

Store data

At the end of the data wrangling process, we got a clean data that is a sharp contrast from the one we started with. This cleaned data is better to understand, clear, readable, and easier to draw information from. We stored it as Twitter_archive_main.csv. After saving the data, we proceeded to visualization and analysis.