WRANGLE REPORT FOR WERATEDOGS TWITTER ARCHIVE PROJECT

The aim of this report is to show the different wrangling efforts and methods involved in getting the clean data for the data analysis project.

Basically, the wrangling stage was divided into 5 stages:

1. **The Data Gathering Stage:**
   Here the data needed to perform the analysis was gathered and saved on the local system. The data was gathered through the following methods.
   - A Comma Separated Value (CSV) file titled "twitter-archive-enhanced" was provided by Udacity for direct download. This file contained information like the various tweet ids of users, the sources of the images from Twitter, the various dog ratings as well as the dog stages. After this file was downloaded, the file was converted into a Dataframe using the Pandas library of Python.
   - A Tabs Separated Value (TSV) file titled "image_predictions.tsv" was downloaded using the Request library of Python. The link for downloading was provided by Udacity. Once this file was downloaded, it was converted into a Dataframe using the Pandas library in Python, taking note of the fact that it was a TSV file .
   - Twitter's API was queried for JSON data using the Tweepy library of Python. To do this, a request had to be sent to Twitter for a developer's account to get access to their API. Once this was done, the API keys were used to download the necessary data and this data was stored in a Text file named "tweet_json.txt". When the download was completed, the data was then downloaded into the work environment using the loads method of the Json library of Python. Then each line was read into a dataframe using the Pandas library of Python.

2. **The Assessment Stage:**
   Here, the downloaded data was assessed and checked for both quality and tidiness issues. The assessment was done through two methods:
   - Visual Assessment: The files were opened in text editors and inspected visually and making use of the functions built into Excel to check for any quality and structural issues. All issues detected like the dog stages being in different columns rather than being in one column for the twitter-archive file or the spelling of the names in the image_predictions file not being descriptive were quickly identified and documented for cleaning later.
   - Programmatic Assessment: The data frames that had been uploaded were analysed using Pythons in-built functions for Pandas. The functions used include the info() method to get a quick overview of the count of the data and the data types and the describe () methods for a descriptive analysis of the data. Issues observed during this programmatic assessment include the presence of null values in the data as well wrong data types for some of the data like the date format being in the wrong data type. These were also documented for cleaning later.

3. **The Cleaning Stage:**
   In this stage, each data frame was cleaned of all the issues identified during the Assessment stage. For instance, columns that were not required or not relevant to the data analysis were removed and incorrect data types were corrected. Also, incorrect entries were corrected using different Pandas methods. Finally, the three data frames were merged into one data frame.
4. **The Data Storing Stage:**
   After cleaning and merging the cleaned data into one data frame, the resultant data frame was stored in a CSV file for future references
5. **The Visualization Stage:**
   After the data was stored, different visualization methods were used on the data in order to get insights and to provide clarity on the data that had been analysed.