"WeRateDogs" WRANGLING REPORT

The project aimed to analyse data from a Twitter account, "WeRateDogs". It is part of the capstone project of the Data Analyst Nanodegree Program by Udacity. WeRateDogs account rates people's dogs with a funny comment about the dog. For this project, data were wrangled, analysed, and the reports were presented in visualisation. However, this report mainly focuses on the various approaches undertaken to wrangle the data.

1.0 Data Gathering

Gathering data is the first step of data wrangling, and it involves collecting important and needful data for data analysis. For this project, data was gathered using three different approaches.

A. **Twitter Archive File -** WeRateDogs Twitter archive file, which was a little bit cleaned, was provided by the instructor for this project as a csv file. This archive contains 2356 tweet data (tweet ID, timestamp, text, etc.). I then downloaded this file on my laptop and loaded it on the python notebook.

B. **Image Prediction File**: The tweet image predictions file contains information on the breed of dog in each tweet as predicted by a neural network. It was hosted on an online server and had to be downloaded programmatically using the Url and python's request library. I then read this file as image_prediction_df.

C. **Twitter API — JSON File:** I could not create a tweet Id to query Twitter API. However, I downloaded the tweet_json.txt provided instead and read it into a panda DataFrame with only the tweet_id, favorite_counts, and retweet_count was included.

2.0 Assessing Data

After gathering the data, the three tables were saved and assessed Visually and Programmatically. With both the assessments, I identified 8 data quality and 6 data tidiness issues in all the three DataFrames,

3.0 Data Cleaning

I cleaned the data accordingly using some commonly used inbuilt python functions. These data cleaning processes can be categorised into;

I. Changing from a data type to another e.g. string to datetime
II. Dropping columns that are not needed
III. Renaming columns
IV. Extracting specific information from a string of text
V. Joining columns together to create a unique column
VI. Removing unnecessary character(s) from a string
VII. Correcting incorrect values and dropping the ones that cannot be corrected.

The detailed steps are provided on the HTML file provided with this report.

At the end of the cleaning process, a new and cleaned table containing information important to make necessary analysis was formed. However, not all issues were resolved, but the important ones to guide my analysis were resolved.