Data Wrangling Report

WeRateDog is a Twitter account comprising of people's ratings about their dogs, humorous comments, and images of the dogs. For this project, different features of these tweets are to be gathered, accessed, and cleaned.

Data wrangling is the process of cleaning, organizing, and structuring data to make it fit for analysis.

Data wrangling steps are:
1. Gathering the data: data is gathered from various sources.
2. Assessing the data: data is assessed visually and programmatically.
3. Cleaning the data: modifying the data to make it clean.

Gathering Data:
WeRateDog data is gathered from three sources:

- Twitter_enhanced_archive CSV file. This csv file is imported using Pandas.
- Image_predictions stored online. This is done using the requests library.
- Additional data which contains important features such as favorite and retweet count are obtained from Twitter API. The tweet is read line by line and saved in a dataframe.

Assessing Data:
Data is accessed visually and programmatically for structural(tidiness) and quality issues. After detecting issues, they are documented before cleaning begins.

Visual assessment: done by visually looking at the data to spot errors and irregularities.
Programmatic assessment: done using pandas functions and methods such as .info(), .value_counts, .sample, .describe(), .head() etc. This gives more insight into the content of the dataset and minimizes the repetition of tasks.

The structural issue is known as messy data. These issues have to do with the columns, rows, or tables of the data. They make the data look untidy. In this project, some structural issues were found and addressed.

The quality issue also known as dirty data or low-quality data is the presence of inaccurate, duplicate, corrupted data in the dataset. They are not supposed to be in the data and are needed to be removed or cleaned.

Cleaning Data:
This is the last step in the wrangling process. A define-code-test format is used to clean all structural and quality issues identified while assessing the data.

Define-Code-Test Format
Define - the issues and how they will be cleaned are defined.
Code - this is the next step that converts the step above to an executable code.
Test - test code used to ensure the code above was implemented correctly.

Before cleaning, a copy of all the dataframes is made to work with, this is to retain the original copy of the datasets. Pandas and Numpy are python libraries used for cleaning the data.