DATA WRANGLING REPORT.

Data wrangling sometimes referred to as data remediation, data cleaning or data mingling is the process of acquiring data then transforming and mapping the acquired raw data into a different format with intentions of making it more valuable and appropriate for various downstream intentions such as visualizations and analysis.
While undertaking my nanodegree program at Udacity, I have been able to work on various projects that need the wrangling process and the journey has been amazing.
Below are the stages of wrangling that I took while working on a project "WeRateDogs" from a twitter archive.The process can either be automated or done manually and often involves 3 different steps:

- *Data gathering*
- *Assessing data*
- *Cleaning data*

1. *Data Gathering*

This is the process of collecting data from different sources with intentions of measuring the targeted variables to enable an individual evaluate outcomes though asking and answering relevant questions. Within my project I was required to put together three different types of data from three different sources. The data files include;

A. **`Twitter_archive_enhanced.csv`**: This is a csv file that was provided by Udacity in the classroom and I was able to download it from there and upload it to the Jupiter notebook work environment then reading it into a panda dataframe.

B. **`image_predictions.tsv:`** This is a csv file that is hosted in the Udacity classroom and the requirement was to download it programmatically using the requests library with a URL that was provided in the classroom.

C. **`Tweet_json.txt:`** This is a JSON text file that is that we had to options of accessing which is query the Twitter API for each tweet's JSON data using Python's Tweepy library and then and store each tweet's entire set of JSON data and then read the file into a panda dataframe or copy and paste the code and comments for the Twitter API code; twitter_api.py which was also provided in the classroom and and once you ran the code, the resulting data was tweet_json.txt.

I used the later method since I experienced some challenges when it came to having my twitter developers account approved and time was running out. Once I was done collecting all the 3 datasets required and reading them into the pandas dataframe, I moved into the assessment stage.

2. *Assessing data*

This is the stage where the gathered data goes through the process of visual or programmatic evaluation to determine any abnormalities present in a dataset, the abnormalities are usually associated with tidiness and quality. Visual assessment of data usually involves looking through the data with the eyes to see if the issues within the data can be spotted while scrolling and this

can be done on the jupyter notebook, excel, text editor etc, while programmatic assessment involves assessing using pandas functionalities .
In my project “WeRateDogs”, I was able to document more than 8 quality and tidiness issues on the 3 datasets combined. Some of the issues I was able to detect include but not limited to;

Quality Issues

- Some missing values.
- Presence of 'None' entries
- Nondescript column headers.
- Mixing of lowercases and uppercase while recording data.
- Misrepresentation of data.
- Variables in some columns are registered as integers instead of strings and vice versa.

Tidiness issues

- Different data types represented in one column rather than 2 and spreading one type of data across multiple columns instead of one.
- The tweet information is spread through 3 different data frames instead of one

3. Cleaning data

Otherwise known as data cleansing and is the process of correcting the corrupt or inaccurate parts of data assessed after collection. The process usually includes deleting, modifying or replacing the coarse data. This process is usually done programmatically by writing and implementing a few codes. In the project I worked on my cleaning process included;

- Dropping/deleting some rows/columns that did not make sense to our analysis.
- Merging the 3 data sets to enable drive meaningful insights and conclusions.
- Rename the non descriptive columns to make sense during my analysis.
- Convert the mixed case into lowercase to have a uniform data.

In conclusion, data wrangling can be the most interesting part of analysis despite being hefty and time consuming. But a good analysis with great insights is determined by the quality of data an analyst has after performing the 3 main tasks in the wrangling stage. Clean quality data is a determinant of quality analysis, hence the reason it takes much time. The process can also be iterable.