The Wrangle Report.

I performed Data wrangling of the entire project in various stepwise procedures and tools/libraries as highlighted below:

Required Libraries and tools:

- Jupyter notebook
- Numpy
- Pandas
- Matplotlib
- Seaborn
- Plotly

Procedure:

i.) Data Gathering & Understanding.

The project required three datasets that were to be obtained in different ways either programmatically or using twitter API.

The final output of this stage were three different datasets which were regarded as the originals. They also had different file extensions i.e *.tsv*, *.txt* and *.csv*.

The original datasets are saved as:

a.) ***Image-predictions.tsv***.
b.) ***Tweet-json.txt***.
c.) ***Twitter-archive-enhanced.csv***.

ii.) Data Assessing.

All the datasets were correctly loaded into jupyter notebook using the standard methods relevant for each and checked to identify any issues that may lower their data quality.

The common issues in pursuit were:

a.) Prescence of missing values and records.

b.) Prescence of inaccurate data in any column.

c.) Prescence of invalid data i.e cells with wrong or unacceptable data values.

d.) Prescence of inconsistent data - Correct conversion of datatypes.

iii.) Cleaning data.

Cleaning data entailed correcting the above identified issues to obtained a better quality of the data.

A copy of the original datasets were made and a suffix “_clean” appended for the naming convention. The copies made were used for the cleaning of data to avoid structural interference of the original data frames.

Upon completion the datasets were merged and the final output was a comprehensive homogenous dataset named:

Twitter-archive-master.csv

iv.) Analyzing and visualizing.

After cleaning, the data was ready for analysis. The analysis performed was majorly exploratory and involved visualization for key extraction of insights.

The pandas library greatly aided the analysis and such for various patterns in the dataset features which were then confirmed and presented visually using various visualizations.