WRANGLE_REPORT THAT BRIEFLY DESCRIBES MY WRANGLING EFFORTS.

The project consist of three data that was combined to form a single dataset which helped in the analysis. The first step of Data wrangling is gathering data; raw data collected for this project were from various sources are were also in different formats which was not suitable for further analysis and modelling. The gathered dataset were with different file type (html, CSV and TSV), the data were gathered from their respective sources programmatically using different python libraries and read all this into a data frame.

The next step was to assess the data, this was done to evaluate the datasets, determine the quality and tidiness issues. I came up with 8 quality issues and 2 tidiness issues with two different approach in achieving this; programmatically and visually. This highlights the shortcomings of the datasets and help to know what to address when cleaning the data. It helped with direction and precision of the final cleaned dataset. Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. From the quality and tidiness issues observed, I was able to carry out a coordinated and direct approach to cleaning. Each dataset had difference approach adopted in the cleaning process.

The cleaning process involve firstly making a copy of the data frame and create a dataframe named cleaned so that there will not be confusion and errors running the codes. Also, columns that are not relevant to the analysis such as the reply columns. There was also some columns that was converted to strings first before dis associating it into multiples column. Every cleaning process had the clear definition of what will be done, the code that was used to execute the cleaning process and a test done to affirm that the cleaning stage has been completed. The final stage was the merging of the cleaned dataset into a single data frame and this was used for visualization.