WRANGLING REPORT

Gathering:

The gathering step is pretty straight forward and simple with first importing all the libraries that I need to use for all the steps in the wrangling & visualizations processes. And second, loading all the datasets. For the “Twiter-archive.csv” and the “image_pridction.csv” files, I use pandas.red_csv(‘file.csv’) to load, and for the “tweet.json” I use the tweep to call Twitter’s API. Unfortunately Twitter disapproved my developer account request, so I use the file “tweet_json.txt” instead.

Assessing:

First, I explore each dataset by using the functions DataFrame.sample( ), DataFrame.info( ) and DataFrame.describe( ) to look at the sample data, types of the data and the data range. This helps me understand what this dataset is about, what information there is, and what interests me to explore further. At this point, I start developing some research questions and think of the data that I would need to help answer these.

Once I get the grip of what each dataset is, I then look at if there is any missing data, and whether the data has the right type or not. After spotting the basic missing data and data types, I continue to look into some sample values to spot quality and tidiness issues. With these three data sets, I use function DataFrame[‘column’].value_counts( ) and DataFrame[‘column’].list( ) to inspect further. I also filter out the information and columns that are invaluable or won’t help answer my research questions, and only keep the wanted columns to work with.

Finally, I make action notes on how to correct each issue and divide them into two batches, quality vs tidiness issues. This is a very important step to enable my cleaning process

Cleaning:

My cleaning process includes two tasks which are coding, to fix the issue, and testing, to see if the new code works. The codes and tests are varied in term of functions and depend on the note. After, finish cleaning all the issues and ensuring that the three data sets are clean, I then combine the three datasets into one single dataset called df_clean.