Wrangling report

Gathering

The first step in the wrangling phase is gathering, this project involved three different datasets with a different type of .csv, .txt and. tsv, all three already gathered in Udacity and ready to use. Since each one file has a different type, they have been read in a different way. First file twitter-archive-enhanced.csv have been read through the typical technique pd.read_csv. The second file, this file image_predictions.tsv have been read in a different technique, Request library used to download image prediction file. The third file was tweet-json.txt, this file has been read using read_json. Since the file was already in Udacity I skipped the API technique.

Assessing

I applied both the visual and programmatic assessing techniques. Jupyter notebook used for visual assess with pandas library, this provides a quick scan for the files. Also, the first is the largest dataset, so I used Numbers software to see all the rows and columns. Programmatic assess used pandas library with multiple functions such as .info (),.describe(), .value_counts() and .duplicated().

In this phase, so many issues in the quality and tidiness found, issues that categorize as invalid, inaccurate, inconsistency, duplicate and inaccurate content goes under quality issues. The issues in the structure such as unrelated data and two columns values in one column go under the tidiness. The detected problems all mentioned in wrangle_act.jpynb at the assessing part.

Cleaning

The last step in the wrangling phase, the detected problems solved in three steps: define, code, and test. this helps in clarifying the solution and the used techniques. Also, testing the solution to check. Issues cleaned programmatically in Jupyter notebook with pandas library through using some functions such as dropping, filtering, rename…etc.

I did not solve all the problems in the three files as mention I have to solve at least 8, so I left the rest for future additions.
In the end, the ***twitter_archive_master.csv*** has been reached after merging the three files altogether, remove unnecessary columns and rename others, after that the merged data frame export to a CSV file.