Udacity Data Analyst Nanodegree Program

wrangle and analyze data

- ## Introduction:

The goal of this project is to wrangle, clean and analyze data from three different sources: one CSV file ”twitter_archive_master” read locally, one TSV file ”image-predictions” downloaded and read programmatically and one text file ”tweet-json.txt” read line by line locally.

The output of these activities is in the “act report” document which includes 3 insights and a visualization of my findings.

- ## Efforts:

  - ### Gathering:

”twitter_archive_master” was downloaded normally through the Udacity website, while ”image-predictions” was downloaded programmatically, finally ”tweet-json.txt” was created by using The Query Twitter API to retrieve tweets and write them locally, this file is later on read line by line to create a data frame.

  - ### Assessing :

The assessment was done both visually and programmatically overall date frames which highlighted several qualitative and
tidiness issues which needed to be resolved before combining all data frames.

  - ### Cleaning :

For the ”tweet-json.txt” date frame, it didn’t have any quality or tidiness issues that needed to be clean. all the issues wherein the other data frames.

”image-predictions” date frame just had one mean tidiness issues where I had to two of the three predictions field which had the lowest confidence coefficient. Aside from that, I drop field that I consider to be useless to me like “img_num”.

”twitter_archive_master” data frame had the most quality and tidiness issues ranging from simple one like change the type of the field to String and dropping retweets form that data frame to merging all stages of a dog into one filed and using regex to find the correction name of a dog in the data frame.

- ## Conclusion:

  The conclusion of these activities is one clean and tidy CSV file called “twitter_archive_master”, which is a product of merging the previous three data frames and is perfectly ready for data analysis and visualization.