UDACITY Data professional track

Project: Wrangle and analyze “WeRateDogs” Data

THE FIRST STAGE OF THIS PROJECT IS GATHERING DATA, MORE DETAILS OF GATHERING IN BELOW.

Gathering

The project is consisting of 3 datasets which need to be collected from 3 different location as follows:

- The first file is located at “UDACITY” server and need to be downloaded manually, the file is contain basic information which has been collected from twitter account called “WeRateDogs”, the has been downloaded manually and imported to the project through pandas.
- Second file is located on the cloud and has to been downloaded programmatically through python “request” library, this file contains predictions of the dogs breeds for each dog image in the tweets
- Third file and final file has to be collected through twitter API by “tweetpy” python library, this file has additional information regarding tweets (retweets counts, favorite counts).

Second stage is assessing the data, more details on assessing as below.

Assessing

Data has been assessed on both visual and programmatically methods, I found several missing values in several columns, duplicated rows and invalid data type for some columns. Aside to that I found some tidiness issues, during assessment process I started to understand more about the data meaning and its representation. Knowing that I do not have any pervious knowledge on the domain of the data.

The third stage and the final stage of wrangling process is cleaning the data. More details in below.

Cleaning

First I had to take a copy of current data so I make sure that the original data is safe and I can go back for it for any reason, then I started fix the issues which was addressed in the assessment stage, the process was enjoyable when you saw the data become more meaningful. Finally, I had to save the cleaned version of the data for future use.