**Documentation for Data Wrangling Steps**

Data wrangling can be broken into three phase : gathering, assesing, cleaning. This document is about the report regarding the process.

1. Gathering Phase

Gathering phases is required to obtain the data required to data analysis. There are three documents in this project, twitter\_archive\_enhanced.csv, image\_predictions.tsv, tweet\_json.txt. , Twitter\_archive\_enhanced.csv is given from the project and can be downloaded manually. Image\_predictions.tsv is in tsv format is downloaded using python requests modul. Tweet\_json.txt is in json format and required twitter api to gather the data. In addition, to acquire the data required, information regarding relevant tweets\_id is provided using twitter\_archive\_enhanced.csv data.

Then, the file is imported to python using Pandas library. Twitter\_archive\_enhanced.csv is already open to get the tweets\_id, image\_predictions.tsv is opened using “tab” as separator for the data, Tweet\_json.txt is opened using read\_json and set the lines parameter to True to read file line by line. Only tweets\_id, retweet, and favorite count is used. The other columns in tweet\_json.txt is dropped.

1. Asseing Phase

Assesing phase is done in two steps, visually and programmatically. Visually is done by seeing the data visually. Example of visual assesing is done in Figure 2 and Figure 2.

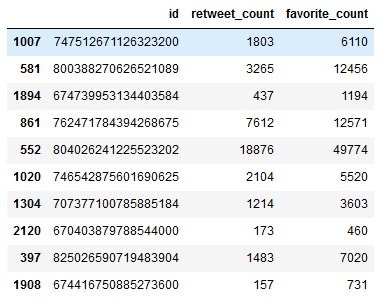


Figure visual asseing tweets table

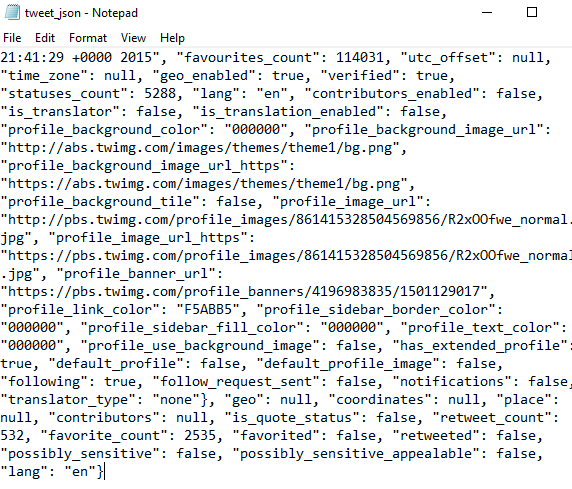


Figure visual assesing tweet\_json

Programmatical assesing is done using code to show the data characteristics. Programmatical is done using info module or query some data that is interesting. Example of programmatical assesment is in Figure 3 and Figure 4.

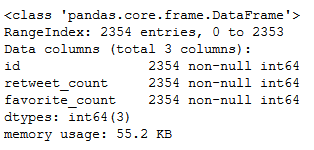


Figure tweet.info

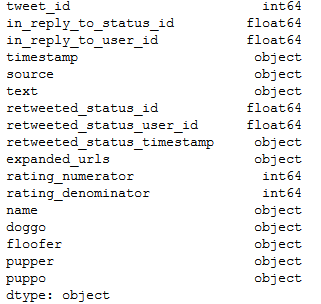


Figure archive.info

Identified data problem is documented is identified in this phase. Some problems identified in this phase is classified into two groups, quality and tidiness. Quality problem is in Figure 5. Tidiness problem is in Figure 6.

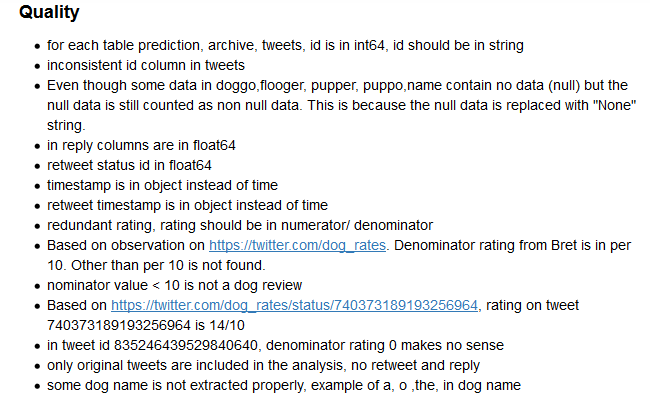


Figure Quality problem

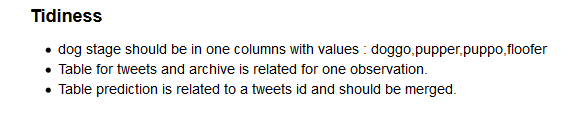


Figure Tidiness problem

Assesment is done to identified the data problems. This data problems are solved in the next phase, cleaning phase.

1. Cleaning Phase

Cleaning phase is done to solve the problem in assesment phase. Cleaning is done in define-code-test framework. Define is defining the problem and the solution in human words. Code is implementing the solution in python code. Test is validating the code to solve the problems. Cleaning phase solved the quality and tidiness problems identified. The resulting cleaned data information is in Figure 7. Some columns that is not needed for analysis is dropped like model 2 and 3 in prediction, and more. Details regarding cleaning process is in the wrangle\_act.ipynb code.

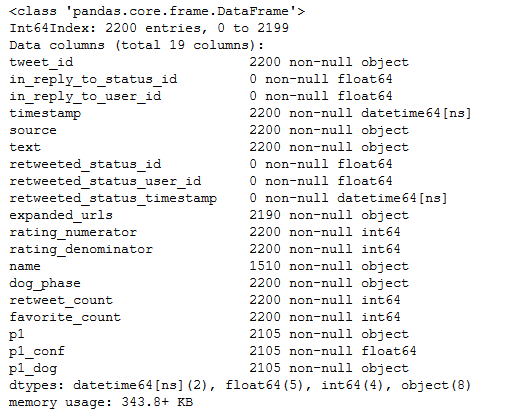


Figure Cleaned data info