Project Name: Data Wrangling of Twitter data (WeRateDogs)

By Chen Yi Fei

- Introduction to the project task:

The dataset is originally sourced from famous social media Twitter. My target is wrangle "WeRatedogs" which is one of the most popular forums on it. The collection of data source is mainly about user's rating info of each particular photo of dogs and other relevant information for example, retweet counts, favorite counts and image prediction done by machine learning method.

- Data Wrangling Methodology and Steps:

As mentioned, the data source is very original one from web site so I need to handle 3 different source of format or sources. For example, I try to get 3 data file from twitter API, local CSV and JSON file. Each of them needs different skills to handle. Once I get all of them ready in my local place, I start to assess the data quality and identify key issues to be cleaned up later time.

At assessing phase, I try to view and assess data set from different angles by using pandas. My way is to start with overview of data structure by using pd.info. It allows me to have overview of data type of each variable and check any inappropriate data types or missing values for example, user id shall be in the format of string instead of numerical type. The next step is to dive into any interested variables to assess any potential data issue for example, duplicated content in one variable, incorrectly extracted name, ratings etc.. I have identified various issues regarding quality and tidiness. Please refer to the details in the wotking of wrangle_report.jpynb

At cleaning phase, I start to clean the issues I have identified. It costs me a lot of time to do. Some are easy and some are difficult especially the column of 'name' and 'rating' relevant. Both of them have to use Regex expression to identify certain patters and replace with correct values.

- Conclusion of this project

I am glad to go through the project. For myself, I learn much about the key skills of data wrangling. The first is about gathering data from different sources of file format as what happens in real life. When I do my assessment to identify issues, I also learn about the quality issues can affect the final output of analyze. For example, the wrong rating numerator and denominator can lead to incorrect interpretation of insights. Finally, the cleaning phase, I polish my skills to use powerful pandas package to perform it with so many real examples. For example, I also realize that if I don't assign correct data type of rating which is float64, I cannot even update the value successfully.