WRANGLE REPORT

For this project, I have worked on 3 datasets : "**twitter-archive-enhanced.csv**", "**image_predictions.tsv**", "**tweet_json.txt**". My tasks were to Gathering Data, Assessing Data, Cleaning Data and Storing Data.

To begin, I imported different librairies which I need to work. There were pandas, numpy, seaboard, matplotlib, requests, os, tweedy, json…After I have imported all the libraries I need, I started the data gathering. My data gathering has been about my three datasets. I first read the master file "twitter-archive-enhanced" as usual by putting it in a data frame called df_archive. On this data frame, I do some operations to know what informations it contains (number of rows, number of columns, types of attributes. I also read the dataset "image_predictions.tsv" and did the same operations to have more informations about it. I had a problem the API of Twitter. Twitter didn't allow me to access to it. I used then the code given by Udacity. But when I wanted to use it, it returned me an error. To continue my project, I created a list in which I put the attributes of tweet_json I want to use in my work (id, retweet_count, favorite_count, followers_count, listed_count. I then put it in a Data Frame, read it and did the same operations as above for the two other dataset.

After my data gathering, I did the data assessing. The goals of my data gathering were to detect at least eight quality issues and two tidiness issue. To assess my 3 datasets, I used both visual assessment and programmatic assessment. So for each of my datasets, I check if there are some null values, some duplicate rows. I also checked the types of the different attributes of my datasets. My assessing data show me some issues in my datasets, quality issues and tidiness issues. My eight quality issues are :

1. **In the dataset df_archive, Timestamp is in string instead of datetime.**
2. **Some columns as "in_reply_to_status_id" and "in_reply_to_user_id" are not necessary.**
3. **We want original valuations, not retweets.**
4. **The column img_num is not necessary.**
5. **The column favorites_count in df_file_json have some null values.**
6. **There are some null values in df_archive**
7. **Some tweets Id in df_archive are missed in df_images**

8. **Some columns as "in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id" in df_archive are in the wrong types.**

My two tidiness issues are :

1. **In Df_archive, Columns doggo, flooder, pupper and puppo can be merged into one column.**
2. **The three datasets can be merged into one**

After I have finished to assess my datasets, I started the cleaning phase. In what this phase consist ? This phase consist into fixing the quality and tidiness issues I have found when I was doing the data assessing. For example, I changed the type of timestamp into datetime. One by one, I have resolved all issues I found. I finished the second project by storing the data. After that, I have done the analyzing and visualizing, steps which I will talk about in my second report.