Wrangle Report

Data wrangling process consisted of 3 main steps: Gathering, Assessing, and finally Cleaning.

1- Gathering Data:

In this project, the data we needed existed in 3 different forms:

1- A .csv file extracted from twitter archive called “twitter-archive-enhanced”.

2- A .tsv file created using a neural network called “image-predictions”. This one was hosted on Udacity’s servers.

3- Data stored in twitter archive.

So, for every Data form mentioned, we used the proper method for **Gathering** then **Reading**, as follows:

1- For “twitter-archive-enhanced.csv” file:
   - it was available for direct download through a URL given by Udacity side.
   - Then we used “Pandas” library to read it inside our jupyter notebook as a Data Frame.

2- For “image-predictions.tsv” file:
   - we used a library called “Requests” to programmatically download it from Udacity’s servers.
   - Then used “Pandas” library to read it inside the notebook as a Data Frame.

3- For the data stored in twitter archive:
   - we used twitter API to access it.
   - Inside the notebook, we used “Tweepy” access library to access twitter API.
   - Then we used “Json” library to store the extracted data in a .json file so that it’s easy to extract only the required tweet status from that file.
   - Then we used “Pandas” to convert the data stored in this .json file into a Data Frame in order to use it during assessing process.

2- Assessing Data:

During this step, we started exploring the gather data both and programmatically in order to check for any Quality or Tidiness issues.
We actually found some of both types. Most of quality issues found existed in the first Data Frame which was made of “twitter-archive-enhanced.csv” file.
Quality & Tidiness issues found were as follows:

Quality

Twitter archive **data:**
1- There are retweets (Only tweets are required).
2- After retweet entries are dropped, ('retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp') features' values are all Null.
3- A lot of data is missing in 2 features ('in_reply_to_status_id', 'in_reply_to_user_id').
4- Wrong values in "rating_denominator" column (Higher than 10).
5- Short URLs column is missing. They can be found at the end of 'text' column entries.
6- Short URLs at the end of 'text' column entries.
7- 'timestamp' column ending with an extra text " +0000".

Image prediction data*:*
8- Some entries with all 3 predictions being False.

Tidiness

1- Last 4 columns (doggo - floofer - pupper - puppo) in "twitter_arch_df" representing only 1 feature (dog stage).
2- The 3 data frames better be merged. Given that "tweet id" is a key feature in all of them.

3- Cleaning Data:

In this step, we made our cleaning through 4 steps for each issue found.

1- Define:
   We define the issue which we are going to solve in this part.
2- Code:
   The coding part where we explore parts of data affected by the issue and solve it.
3- Test:
   We check our Data Frame again after solving the issue to make sure it’s generally solved through the whole Frame.
4- Store:
   We store a clean version of data after solving a particular issue, so that we can re-call it before solving the next issue.

Finally, we store our clean version of data in a .csv file called “twitter_archive_master.csv”. This will be the file to use in our analysis.