Wrangle Data

- ## Intrduction :

The dataset that wrangled is Tweet archive of twitter user @WeRateDog the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10.

- ## Project details :

The tasks of this project are as follows:

1. Gathering data
2. Assessing data
3. Cleaning data

Gathering data

The data for this project consist on three different dataset that were obtained as following:

• **Twitter archive file:** the twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.

• **The tweet image predictions,** i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information

• **Twitter API & JSON:** by using the tweet IDs in the WeRateDogs Twitter archive, I stored each tweet's entire set of JSON data in a file

called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count.

Assessing data

Once the three tables were obtained I assessed the data programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc)

Then I separated the issues encountered in 9 quality issues and 2 tidiness issues. Key points to keep in mind for this process was that original ratings with images were wanted.

Cleaning data

This part of the data wrangling was divided in three parts: Define, code and test the code. These three steps were on each of the issues described in the assess section. First and very helpful step was to create a copy of the three original dataframes. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original.

Whenever I made a mistake, I could create another copy of the dataframes and continue working on the cleaning part.

There were a couple of cleaning steps that were very challenging. One of them was in the image prediction table. I had to marge t the dog stages in one column instead of four columns as original presented in twitter archive.

- Conclusion

Data wrangling is a core skill that whoever handles data should be familiar with.

I have used Python programming language and some of its packages.

• For gathering data there are several packages that help scraping data off the web, that help using APIs to collect data (json for json file) or to communicate with SQL databases. • It is strong in dealing with big data (much better than Excel).

• It can deal with a large variety of data (unstructured data like JSON (Tweets) .

.• Handling, assessing, cleaning and visualizing of data is possible programmatically using code.