Wrangle and Analyze data: WeRateDogs

In this project we want to wrangle data from twitter account WeRateDogs, we will gather, assess and clean the data.

1. **Gathering**:

First step of wrangling, we will gather the data in three ways:

- Using 'twitter-archive-enhanced.csv' file.
- Using a URL.
- Using twitter API (I created developer account to get tokens and secret keys).

For the first way I just downloaded the file, for the second way from Udacity server I downloaded the file ‘image_predictions.tsv’ programmatically, to do that we can see from Jupyter file you can see that request library was used.

2. Assessing

Second step of wrangling, we found in the dataset many issues and I have finished 8 quality and two tidiness issues.

Quality issues:

-I noticed that in the first dataset (ds),the tweet_ID has wrong data type and value. I extracted the tweet_ID from expanded_urls, eventhough some tweet_ID are missing.
- Wrong datatypes + values for in_reply_to_status_id,in_reply_to_user_id
- For ds, the denominator must be 10.
- For ds, wrong datatype for timestamp
- For ds, many dog names are not correct.
- For (img_data), many predictions are not dogs.
- For ds, after we get rid of retweets the colmuns 'retweeted_status_id' 'retweeted_status_user_id' and 'retweeted_status_timestamp' are not useful.
- Columns (doggo, floofer, pupper, puppo) has None for missing values.

Tidiness:

- We can notice that In ds, the columns 'doggo', 'floofer', 'pupper','puppo' shows one variable.
- We can notice that the all datasets must be combined

- we have the following columns (p1, p1_dog, ...etc) need to be just breed and confidence.

3. Cleaning:

In this part there are three steps:

- Define: What is the problem that you want to fix
- Code: Use cleaning methods to fix the problem (sort, drop, duplicate)
- Test: show the data after cleaning

I have dropped some columns and join also colmns such as ('doggo', 'floofer', 'pupper','puppo''), and you can check the jupyter document to see the cleaning.

4. Conclusion:

In this section at Udacity, I understood what is the data wrangling and in the project I applied concepts of wrangling (Gathering, Assessing, Cleaning). In the next steps I need to improve myself in gathering, for example gather from other sources rather than twitter, I also need to improve my cleaning skills, also I in the next phase I going to enhance the visualizations.