Wrangle and Analyze Data

- Introduction

In this project, we will do the wrangle with 3 dataset from Twitter user WeRateDogs. 1 of the dataset is tweet archive document including tweeter_id,ratings information and the dog stages, etc; One is image prediction file which include the url of the dog image, the prediction result and confidence level of the predictions; Another one is json file which include tweet_id, the number of retweet and the number of favourite to show how popular each twitter is. The goal of this project is to go through the 3 important steps of data wrangling which are gathering, assessing and cleaning. Also, I will explore some interesting insights about the dog and the dog related twitter to answer some questions like what's the popular dog names, which breed that the neural networking predict the most and what's the distribution of the ratings.

- Gather

We have 3 types of file, archive file is a csv file, image file is in tsv format and tweet is in json format. I used pandas package to read this 3 dataset and load it into the jupyter for future analysis.

- Assess

In this project, I divide the assessing process into 2 parts, one is about quality, another one is about tidiness.

1. Quality:
    a. Missing value and completeness
    b. Invalid data
    c. Inaccurate data
    d. Inconsistent
2. Tidiness
    a. Column headers are values, not variable names.
    b. Multiple variables are stored in one column
    c. Variables are stored in both rows and columns.
    d. Multiple types of observational units are stored in the same table.
    e. A single observational unit is stored in multiple tables.

I go through each step to assess the data and find a lot of data issues. For example, in the twitter archive dataset, I found:

1. All the item ends with _id should be string type.

2. There are 59 null values in expanded_url. It's ok if there are missing values about retweets and in_reply since it's very likely that this tweet doesn't have any retweets and replies.
3. For rating_numerator and rating_denominator, there are some outliers. All rating_denominator should be 10.
4. There are some outlier in the rating numerator; Also, rating numerator and denominator are separated so we need to create a new column to calculate the rating.
5. There are a lot of None value in doggo, floofer, pupper and puppo and we should replace it with NaN.
6. In the name column, there are some names like a or None which should be processed.
7. the data type for timestamp is wrong.
8. Non-null retweet_status_id should be removed since it stands for retweet.
9. There are multiple values in the same cell like source. We need to extract useful information from it for further analysis.

In the tweet-json table, I found some issues like:

1. tweet_id and id is not consistent

2. The datatype of id is not correct

- Clean

I created all comments under each data and clean the data related to the issues step by step. For example, In the tweet archive dataset, I remove the records with not null retweet_status_id tag since we only want to explore the insight about the original tweet. Then I encoding the wrong format records into the right format and some commands I used are like replace, rename, etc. Another example is that we have 4 columns: doggo, fluffo, pupper and puppo but they are actually the they refer to the same thing: dog_stage, so I first merge them into the same cell, and the encode them as different type and stored the new value into a new column called dog_stage and then remove the old 4 redundant columns.

- Summary/Conclusion

In this project, I deal with the real world dataset and it helps me realize how messy the data is in the real world.

1. We may get data from different data sources and they could be in the different file types like csv, json, tsv, etc. pandas is a great package to deal with a lot of different file types.
2. Data wrangling includes 2 different parts, we always need to consider the data issues from quality aspect and tidiness aspect.