Wrangle report

In this project, I wrangled the tweet archive of Twitter user @dog_rates from February 2015 to August 2017 and a table of image predictions alongside each tweet ID of the tweet archive.Twitter user @dog_rates, also known as WeRateDogs, is a Twitter account that rates people's dogs with a humorous comment about the dog.
This tweet archive contains more than 2000 tweets. I gathered additional information then assessed and cleaned the datasets following the below steps:

Step1: Gathering data

The whole dataset has 3 data resources. I used 3 different ways to gather these data and read them into 3 pandas dataframs.

(1) Download the *twitter_archive_enhanced.csv* file from Udacity website manually, load the csv file into a pandas datafram.
(2) Use requests module to programmatically download the *image_predictions.tsv* file from a URL supplied by Udacity, load the tsv file into a pandas datafram.
(3) Query the Twitter API for every tweet_id value in *twitter_archive_enhanced.csv* file with Tweepy library. Store each tweet's entire set of JSON data in a file called tweet_json.txt. Load useful data from every line of JSON data into a pandas datafram.

Step2: Assessing Data

After loaded all 3 dataframs, I used several ways to assess these data, and listed issues to be cleaned and tided.

(1) Print these dataframs in Jupyer notebook, then visually assess for quality. In this step, I found several issues such as missing data, redundant characters, wrong values.
(2) Using pandas methods and functions to check every column of each datafram. methods and functions include duplicated(), isnull(), filter, value_counts(), isin(), indexing, info(), sum(), count() and so on. Most of the issues to be cleaned were in this phase. These issues includes unnecessary rows and columns, wrong values, wrong datatype and so on.
(3) Base on the table structure and logical relationship, find out two issues to be tidied. One is about combine two dataframs, one is about adjusting columns.

I totally figured out 13 quality and tidiness issues.

Step3: Cleaning Data

After identified all necessary issues, I used different technology to clean and tidy the dataframs.

(1) Firstly, I make a copy of each datafram, named them *twitter_enhanced_clean, image_predictions_clean* and *supplement_clean*.
(2) There is a lot of missing data, but there is no idea to deal with this information missing, unless there are additional data sources. I have to leave them there.
(3) It's better to deal with the tidy issues, in order to make quality cleaning work easily. But there are several cleaning jobs should be done before tidying. I coverted the

data type of tweet_id column in *twitter_archive* to string, and dealed with several dog stage value missing and wrong issues.

(4) Then I finished the two tidiness issues. I let retweet_count and favourte_count from the *supplement_clean* be part of *twitter_enhanced_clean.* I also combine all 4 columns about dog stages in *twitter_enhanced_clean* into 1 column named 'stage'.

(5) I finished all left quality issues, included dropping redundant rows, correcting wrong values, getting rid of html tags, changing column datatype.

Step4: Storing Data

After finishing all cleaning and tidying works, I stored the two final datafram as two csv files, named twitter_archive_master.csv and image_predictions_master.csv.
I also set a database called 'dog_tweet.db' and save the two final datafram in this database using sqlalchemy library.

Now I am ready to analysing and visualizing data.