Summary of wrangling process

The process to do data wrangling can be separated into 3 main steps staring form gathering data, then assesing quality of it, and cleaning it using programmatic approach.

First gathering data

The data has been received from three different sources: twitter, provided csv, and download direct from Udacity server.

I have started the process by import necessary python library consisted of tweepy, pandas, json, requests, numpy, and % matplotlib inline

Then

- Gather data from twitter using code written. Save it in 'twitter_json.txt', read tweet_json.txt file line by line into dataframe called it df.
- Read provided csv ('twitter-archive-enhanced.csv') store it in dataframe called df1.
- Gather dog breed classification data from udacity server stored it in dataframe called img_pred

After that I take a quick overview of what information provided inside using both python and excel.

Second Accessing data

In this process, the data has been checked step by step. Few main things I am looking for was whether it has missing value or not, how is the data quality, and how is data tidiness. In case of issues detected, I will document and classify clearly into either quality issues or tidiness issues.

Twitter data (df): 6 quality issues and 2 tidiness issues have been identified.

Quality

- 3 columns (contributors, coordinates, geo) have no data at all.
- Missing values columns are extended_entities, in_reply_to_screen_name, in_reply_to_status_id, in_reply_to_status_id_str, in_reply_to_user_id, in_reply_to_user_id_str, place, possibly_sensitive, possibly_sensitive_appealable, quoted_status, quoted_status_id, quoted_status_id_str,retweeted_status)
- 15 columns have data completed for all rows (created_at, display_text_range, entities, favorite_count, favorited, full_text, id, id_str, is_quote_status, lang, retweet_count, retweeted, source, truncated, user)
- Column 'created_at' has incorrect datatype. It should be date time
- Column 'Source' has information of source (twitter for iphone, twitter web client) together with non related information
- Column 'display_text_range' has stored as object

Tidiness

- Not need:
  'contributors','coordinates','entities','extended_entities','favorited','full_text','user','id_str', 'truncated', 'entities', 'extended_entities','in_reply_to_status_id','in_reply_to_user_id', 'in_reply_to_status_id_str', 'in_reply_to_user_id_str', 'in_reply_to_screen_name', 'geo','coordinates', 'place', 'contributors', 'is_quote_status', 'possibly_sensitive','possibly_sensitive_appealable','quoted_status','quote_status_id','quoted_status_id_str','retweeted','retweeted_status'
- Need to combine data from 3 table using 'id' as key to create final table

Twitter-archive-enhanced.csv (stored as df1). Identified 2 qualities issue and 2 tidiness issues.

Quality

- Column 'rating_numerator' has wrong information (not equal to 10)
- Column 'timestampt' has +0000 inside

Tidiness

- Not need column: 'in_reply_to_status_id', 'in_reply_to_user_id', 'timestamp','retweeted_status_timestamp', 'expanded_urls'

Dog breed classification (img_pred): 2 tidiness issues have been identified.

Tidiness

- Not need column: jpg_url, img_num
- Interested only p1 prediction. Remove p2 and p3

Final steps clean the data

The cleaning process starting from define what to clean, write a code, and then test the code to make sure everything work as expected.

Twitter file (df): remove not necessary columns, change column 'created_at' to date time, remove unnecessary text to collect twitter source from column 'source', and change text lenght to integer from column 'display_text_range'

Twitter-archive-enhanced.csv (stored as df1): remove not necessary column, change column 'rating_denominator' to 10 when figure is not equal to 10

Dog breed classification (img_pred): remove not necessary column

Combine table is the last step

Finally, I have combined 3 files together to create final table in order to further use for analysis. First combine df with df1 in order to get rating, numerator, denominator, name, and type of dogs. After that, combine it with img_pred to get dog breed information.

After finished combining data. The file ‘twitter_archive_master.csv’ has been exported for further reference.