Wrangle Report

In this project, data of various types which are heterogeneous, were collected, evaluated and cleaned. Then it is stored as a master csv file which is ready for analysis, after that, some analyzes are done on the cleaned csv master file. Here is a summary of each stage**:**

Gather:

Data were gathered from three resources:
-The first one was given as a csv file, namely twitter-archive-enhanced.csv, which was provided by We Rate Dogs twitter account to Udacity and then to us. Then, the data was imported as a pandas dataframe dftwitter.The have columns:
1- tweet_id: the unique identifier for each tweet.
2- in_reply_to_status_id: if the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's ID.
3- in_reply_to_user_id: if the represented Tweet is a reply, this field will contain the integer representation of the original Tweet's author ID.
4- timestamp: time when this Tweet was created.
5- source: utility used to post the Tweet, as an HTML-formatted string. e.g. Twitter for Android, Twitter for iPhone, Twitter Web Client.
6- text: actual UTF-8 text of the status update.
7- retweeted_status_id: if the represented Tweet is a retweet, this field will contain the integer representation of the original Tweet's ID.
8- retweeted_status_user_id: if the represented Tweet is a retweet, this field will contain the integer representation of the original Tweet's author ID.
9- retweeted_status_timestamp: time of retweet.
10- expanded_urls: tweet URL.
11- rating_numerator: numerator of the rating of a dog. Note: ratings almost always greater than 10.
12- rating_denominator: denominator of the rating of a dog. Note: ratings almost always have a denominator of 10.
13- name: name of the dog.
14- doggo: one of the 4 dog "stage".
15- floofer: one of the 4 dog "stage".
16- pupper: one of the 4 dog "stage".
17- puppo: one of the 4 dog "stage".

-The second one was programmatically fetched from the URL https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv with which the response has been processed and saved into a tsv file, namely image-predictions.tsv. Then, the data was imported as a pandas dataframe dfimage.The have columns:
1- tweet_id: the unique identifier for each tweet.
2- jpg_url: dog's image URL.

3- img_num: the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).
4- p1: algorithm's #1 prediction for the image in the tweet.
5- p1_conf: how confident the algorithm is in its #1 prediction.
6- p1_dog: whether or not the #1 prediction is a breed of dog.
7- p2: algorithm's #2 prediction for the image in the tweet.
8- p2_conf: how confident the algorithm is in its #2 prediction.
9- p2_dog: whether or not the #2 prediction is a breed of dog.
10- p3: algorithm's #3 prediction for the image in the tweet.
11- p3_conf: how confident the algorithm is in its #3 prediction.
12- p3_dog: whether or not the #3 prediction is a breed of dog.

-The third one was through Twitter API, we fetched the data of those tweets we have and then saved them as a JSON file with UTF-8 encoding, namely tweet-json.txt, Then, the data was extracted as a pandas dataframe dftweetsjson. The have columns :
1- tweet_id: the unique identifier for each tweet.
2- favorites: refers to topics or subjects that users are most interested in.
3- retweets:Tweet is a retweet.

Assess:

Two types of assessments were performed, visual assessment and programmatic assessment. And the following issues were found:

Quality:

dfimage table:

1)tweet_id should convert to str.
2)delete the img_num and jpg_url values.

dftwitter table:

1)None values in name of dogs and incorrect dogs name.
2)Useless columns such as in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id ,retweeted_status_timestamp should be removed.
3)timestamp should convert to timestamp datatype.
4)solve missing value in dog type.
5)clean text.

dftweetsjson:

1)convert tweet_id, retweets, and favorites in table to int datatype.

Tidiness :

We have 3 datasets, dfimage , dftwitter, and dftweetsjson dataset , we should merge it in one dataframe.

We have three seperate columns of dog categories, drop the puppo, pupper,floofer and doggo after add to new culmn.

Clean:

All the mentioned issues were resolved in the evaluation stage programmatically. For example, unwanted columns were removed in the analysis, some types of erroneous column data were changed to the appropriate type, the tweet text containing some links, punctuation marks and repeated messages was dealt with, and incorrect or incorrectly entered dog names were dealt with And some of them were dealt with invalid data such as some invalid dog classifications found in the specified data etc. Finally, the clean data files are combined as one to make the data ready for analysis.