Data Wrangling Project

In this Project I have wrangled and analyzed the We-rate-dogs twitter archive.

There are 3 main inputs:

- We-rate-dogs twitter archive (downloaded from udacity)
- Each twetts json data, queried via twitter API using tweep
- Image predictions file, programatically downloaded from the udacity server

Cleaning:

I used the standard information functions such as .info(), .head(), .describe(), .count() to get an overview about the data. Doing this I could identify several quality and tidyness issues:

invalid names in the dataset:

- Define: The 'A' and 'None' names should be converted to N/A
- Code: The column is explored and then replaced from None, A to N/A

source column is enclosed in html tags:

- Define: The source of the tweet is enclosed with html tags
- Code: String Operations are done to extract the text

data should only contain tweets and not retweets or replies:

- Define: Remove the rows containing retweets or replies from the archive dataset
- Code: Removal can be done using the drop and notna methods

retweet and reply related columns should be dropped:

- Define: Since the data is particularly about original tweets, the retweet and reply columns are not needed in the dataset.
- Code: Using Dropna in axis 1 (column), remove: 'in_reply_to_status_id','in_reply_to_user_id','retweeted_status_id','retweeted_status_user_id',' retweeted_status_timestamp'

Expanded urls containing the links to the urls contains 3 missing values:

- Define: The missing values in the expanded urls column should be removed
- Code: The dropna method is used to remove the missing values

Doggo, Flopper, Pupper, Puppo should be melted into one column: stage

- Define: The dog stages should be combined to one column
- Code: A function is created that replaces 'None' strings with an empty string in each of the columns. The strings are then concatenated. For stages with more than one dog stages, one is selected. The empty strings are replaced with missing values.

Retweeted and Favorite columns should be added to the dataset

- Define: The favorite and retweed columns would be combined to twitter_archive data on tweet id
- Code: We could combine the two dataframes using the pd.merge function

Image Prediction data, some of the predictions were not classified as dogs

- Define: The breedPredict is created to contain the predictions of dog breeds in the imagePrediction Data. Since some dog breed predictions are not correct, p1, p2 and p3 are checked for a better prediction.
- Code: By applying conditional statements in the imagePrediction dataframe, the correct dog predictions are selected and a missing value(nan) returned when none of the predictions are classified as dogs.

predicted breed should be added to the dataset

- Define: The predicted breeds of the dogs should be added to the dataset
- Code: The breeds are added to the tweitter_archive_copy using pandas merge function on tweet id

Finally the table was saved to “twitter_archive_master.csv” and the analyzed and visualized accordingly.