Reporting: wragle_report

Data Gathering

The data where all gathered from 3 sources, which were stored in seperate files. The datas were as follows;

1. twitter-archive-enhanced.csv file which was manually downloaded from udacity servers
2. image-prediction.tsv file which was programmatically downloaded and stored into my local work space from udacity servers.
3. tweet-json.txt file was a json file which was stored locally as a txt file which was gathered by querying twitter API using tweepy library to extract the data needed.

   pandas library was used to read the 3 data into separate dataframes namelly; twitter_arc, image-prediction and df_tweet.

Data Assessment

Each dataframe were visually and programmatically assessed for quality and tidiness issues.

Quality issues

- twitter-archive-enhanced.csv

1. The Names column had an invalid names like 'None', 'the','quite', 'a', 'an'.
2. An object datatype was asigned to timestamp column instead of datetime datatype.
3. Retweets were present in the dataframe which was not needed in the analysis since the concerns was the original tweets and also Reply tweets was not “original tweets” either; these data were stored in columns like in_reply_to_status_id and in_reply_to_user_id.
4. it was observed that in most of the columns null values were represented as 'None' instead of NaN.
5. In Some rows, there were duplicated values in their expanded_url column.
6. The Tweet_ids were stored as integers instead of strings.
7. Removing hyperlinks from tweet source column.

   - image-predictions.tsv

8. Some tweet_ids had the same jpg_url.

Tidiness Issues

- twitter-archive-enhanced.csv

1. Dog categories (doggo, floofer, pupper, puppo) were spread in different columns. which should had been in a asingle column.

- image-predictions.tsv

1. Breed Predictions, Confidence intervals and Dog texts were spread in three different columns.

- tweet-json.txt

1. twitter_arc and df_tweet had similar information about the tweet made eg thesame tweet_id. which should be melged

2. joined all the data frames together using tweet_ids and performed general cleaning on the data

Data Cleaning

The observed Quality and tidiness issues in the data were cleaned in the following procedures.

1. Wrote a for loop code to asess the text column for the missing name of the dog and created a new name column to replace the old name column. Changed names with words like 'a', 'an', 'such' to NAN values
2. Converted timestamp column from object datatype to_datetime.
3. Changed values in name, doggo,floofer, pupper and puppo columns from None to NaN values
4. Removed duplicated values from the expanded_urls
5. changed tweet_id from int to string datatype for the 3 data frames
6. Extracted the tweet-source information from the source column.
7. placed all the four dog categories (floofer,pupper,puppo,doggo) into a single column using melt method
8. dropped retweets column from the dataset as well as retweeted_status_id, retweeted_status_user, retweeted_status_timestamp, in_reply_to_status_id and in_reply_to_user_id
9. dropped tweet_ids which had duplicated jpg_urls
10. Created two new columns in image predictions dataframe called dog_breed and confidence which checked each dog breed prediction and copied the breed with the highest confidence level into the new breed column.
11. Merged twitter_arc_clean and image_prediction_clean data frame together using tweet_id and applied general cleaning
12. stored the dataframe to a csv file called 'twitter_archive_master.csv'

In [ ]: