Project: Wrangle and Analyze Data

The aim of this project is to gather data from different sources inorder to assess and clean them and create insights and meaningful visualizations.

The steps performed are:-

1. Gather Data
2. Assess Data
3. Clean Data
4. Analyse Data and create meaningful insights

1. Gather Data
   I gathered data from 3 different sources.
   I. Twitter_archive:I have extracted the data from Twitter archive. This data was a downloadable csv file. Hence, I imported the data using read_csv command
   II. image-predictions.tsv: I extracted the image predictions file programmatically using the requests() library
   III. Twitter API: I used the twitter API called Tweepy to read the tweets and store it in a file 'tweet_json.txt' with every new line being a record. The tweets were searched using the twitter ID for each tweet from the twitter archive. For the tweet ids whose record was successfully found was stored in the file 'tweet_json.txt'. The failed records were stored in the dictionary 'fails_dict'. These records were then read into 'twitter_api_master' dataframe. And only relevant columns are stored in 'twitter_api' dataframe

2. Assess Data
   In this section, the dataframes were assessed using .head() function. And the consistency of the datatypes was checked using the function
   Visual assessments were done using .head(),. Tail()
   Programmatic assessment was done using functions such as info(), describe() and '.dtypes'

3. Cleaning Data
   Following are the quality and tidiness issues found out based on the assessment

3.1 Quality

df_imagepred table

1. Breed names of dogs in p1, p2, p3 to be made sentence case

twitter_archive_clean

1. Rows which don't have dog name ,in the column 'name', to be removed. I know this as the rows where dog names are absent there the dog names are in lower case. Hence the rows where dog names are 'a','an' etc. are removed

2. Duplicate text column to be removed

twitter_api

1. Since only original ratings (no retweets) are asked, only the records where retweeted='False' is to be considered
2. Since the dog ratings are not extracted correctly in Enhanced Twitter Archive, the dog ratings are to be extracted
3. id to be renamed to tweet_id inorder to join the tables

twitter_archive_cons

1. The 'rating_numerator' to be reused to extract the numerator from the dog rating 'dog_extract_rating'
2. The 'rating_denominator' to be reused to extract the denominator from the dog rating 'dog_extract_rating'
3. The dog stage to be extracted from the 'text' column into a column 'dog_stage' . The name is extracted into a list initially and then into a string.

3.2 Tidiness

twitter_archive_cons

1. 'doggo,floofer','pupper' & 'puppo' columns to be deleted
2. 'retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp' to be deleted as they don't add significance