Data Gathering

Three different dataset types had to be downloaded from three different sources. Before we could load these files into the workspace some library packages had to be downloaded, these included pandas, numpy, matplotlib.pyplot, requests and json. The first file to be loaded was the "twitter-archive-enhanced.csv" which was available in the local library. Since this file was already provided in the local library it was read into workspace using the *pd.read _ csv("twitter-archive-enhanced.csv")*.

The next dataset to be loaded was the image_predictions.tsv dataset. In order to get this file the *requests.get(url)* function was executed, the *response.content* to view the file contents. But to perform operations on the file it needed to be available in the local workspace, json operations and functions were used to achieve this.*with open(os.path.join(workspace, 'predictions.tsv'), mode='wb') as file: file.write(response.content)* was run to write the file to the current working directory.

For the third dataset, it was supposed to be accessed from the Twitter API, but due to the expired access keys and difficulties in opening a twitter developer account, the *tweet-jason.txt* file was made available for use. Using json the file was opened and three columns favorite_countre, tweet_count and id were appended to an empty dataframe list *df_list*. This list was then converted to a dataframe titled *twt _ api*

Assessing Data

This part of the process highlighted some issues with the datasets, tidiness issues and quality issues. From the twitter-archive-enhanced.csv a total of 181 retweets were present and tweets beyond 2017 could not be used therefore could not be accessed. The following columns had a lot of NaN and none values: in_reply_to_status_id, in_reply_to_user_id have, retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp. The timestamp column was in the incorrect data type. Most users did not classify in what category their dogs belonged in either doggo, floofer, pupper or puppo. A lot of dog names were also missing; this could have been caused by errors in the data capturing process.

Image predictions dataset provided three columns that gave the predictions of the dogs and the breed of the dog. P1 provided the highest level of confidence for dog prediction, p2 being the second highest and p3 being the least confident. Some of the images provided buy users were not dog images and some of the p1 false predictions were actually dogs.

From tweet-json.txt only 3 columns could be extracted. From the extracted columns no NaN values were detected and the columns were in the correct data type.

Cleaning Data

The 1st step was to drop all the the retweets and only analyse original tweets which was done by *archive_cp.loc[archive_cp["retweeted_status_id"].isnull()],* which retained all the rolls that had no retweet user ID value. Next was to change the timestamp data type to datetime using the datetime library, this was done to check if there were any tweets beyond August 1st 2017 that filtered through but none did. The second issue to be cleaned was to drop the columns that had missing data since some of this data could not be autofilled with the average or mean. Columns from twitter-archive-enhanced.csv, and that had missing data were dropped using the *pandas.drop* function. The name column in image predictions was not dropped but the columns of the confidence levels were dropped and merged into one column ‘breed’ to make it easier for readers to assess. When the cleaning was done all three datasets were merged into one dataset called 'twitter_archive_master.csv'.