Data Wrangling Report

The first step was to prepare a python script in spyder editor. This script reads twitter_archive dataset supplied and stores it in a dataframe. After authentication the script iterates over the tweet ids and for each id we extract the status object and store append it to a list. Some of the tweets ids returned an error that indicated no status with such id exist. A try and catch method solved this. That is why the extracted data has fewer entries than the tweet ids supplied.

We iterate over the status lists to extract retweet count, favorite count and tweed id then store them in a 2 dimension array. The 2D array was used to populate the tweets_data.csv file.

Most of the assessment was done visually on excel. The 3 files were read into the wrangle_act.ipynb and assessment done programmatically. On the name column of tweet_archive, it was noted that sensible dog names started with a capital letter. Therefore, any dog name starting with small letter was dropped. The doggo, pupper, puppo and floofer columns were supposed to represent dog size/category, at least based on slang dictionary definition of the four names. These was need to rename them and merge them to new column called dog group.

The expanded_url column in twitter_archive has duplicated links in same cell. These were split and the first link selected. Some predicted breed names started with small letter and had an underscore. All predicted breed names were capitalized and underscore replaced with a blank space.

The ‘we rate dogs’ channel has information on rating that suggest the denominator should be 10. Therefore all values in the rating_denominator were set to 10. The source links in the supplied twitter_archive dataset were not properly extracted as they contained anchor tags. The anchor tags had to go.

My analysis took interest in variables from the tweets_clean and archive_clean data. The archive_clean set had few observations so it was merged with tweets_clean using a left join, tweets_clean being the left table. This is to avoid losing valuable data in the tweets_clean table. The resulting dataframe was merged with images_clean data using a left join; with images_clean being the left table. This ensured we only select tweet_ids that are on the images_clean dataset. The cleaned datasets were stored in csv files.