Report on Data Wrangling process

The following steps were taking in the data wrangling process;

Data Gathering

The data gathered from three sources;

- The Enhanced Twitter Archive derive from The WeRateDogs Twitter page, it contains basic tweet data for 2356 tweets. The Twitter-archive-enhanced.csv was read into a pandas dataframe.
- The Tweet image prediction data file was generated from images in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs. The image_prediction.tsv was downloaded using python's Request library for the URL given and stored in a dataframe.
- Using the Twitter API by generating Twitter API keys, secrets, and token for an app, the Tweepy Python library and the tweet ids from the Enhanced Twitter Archive file, data consisting of the number of retweets, favourites, etc. the data was extracted in JSON data format and stored in a file tweet_json.txt and then read into a pandas dataframe.

Data assessment

A visual and programmatic assessment for the data was carried out on the data sets for quality and tidiness issues, below are the result of the assessment;

Enhanced Twitter Archive

Quality

i. Rows with values in retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp are retweets and not the actual tweets.
ii. Rows with values in in_reply_to_status_id and in_reply_to_user_id are replies to an original tweet.
iii. The source, expanded_urls have duplicate column values or not need for the analysis.
iv. The columns doggo,floofer,pupper,puppo are string data types.
v. Simplify the prediction result columns p1_dog,p2_dog,p3_dog into one column.
vi. Incorrect values in the rating_denominator column.
vii. Incorrect values in the rating_numerator column
viii. Lower case string values are not Dog names in the name column.
ix. Source of device used placed in 'a' HTML tags in the source columns.

x. The timestamp data type is (string)object.
xi. The tweet_id column data type is an integer.
xii. The expanded_urls column is not needed for the analysis.

Tidiness

1.retweets, favourites, text, created, jpg_url, confidence level,verified_dog_prediction from the tweet_api data and image_prediction data tables should be part of the df_twitter_archive data table.

2.Combine the dog stage columns doggo,floofer,pupper,puppo into one column.

Tweet image prediction data file

Quality

1. Fewer records than the enhanced Twitter archive,2075 instead of 2356.
2. The img_num column is not needed for the analysis.

Tidiness

1.All the columns should be part of the Enhanced Twitter Archive table using the tweet_id columns.

2.Columns p1_dog, p2_dog and p3_dog should be combined.

Tweeter API data

Quality

1. Days of the week need for analysis.
2. Fewer or missing records in the tweet API data table, compared to the enhanced Twitter archive,2331 instead of 2356 due to failed API retrieval.

Tidiness

The retweet, favourites, text and created_at columns should be part of the enhanced Twitter archive.

Data Cleansing

- The rows representing retweets replies and not need for the final analysis were dropped.

- The tables image_predict and twitter_api were merged to the enhanced Twitter archive table using the tweet_id column, which dealt with the issue of differences in the records as only records with matching records are left.
- The Dog stage columns were combined, with a data type of category.
- The lower case string values for the name column were converted to None as they are not dog names.
- The Beautiful soup library was used to remove the text between HTML tags, to determine source platforms.
- The days of the Week column was derived from the created column.
- The two rating columns were re-extracted from the text columns to compare and replace the existing values.
- The prediction result columns were simplified and graded.
- The column with wrong data types was converted to types for analysis and storage.