WRANGLE REPORT

This is the report on the data wrangling process of the WeRateDogs twitter feed.

The wrangling procedure was divided into 3 stages.

- Data Gathering
- Data Assessing
- Data Cleaning

Data Gathering –

To gather all the required data, I started with manually downloading the provided Twitter Archive data. This dataset contains information on Tweets, time of the tweets, dog ratings (both numerator and denominator), dog stage name, reply status and of course twee id which is the primary key. This dataset was read into panda dataframe.

The twitter archive data has some notable omissions like retweets and favorite counts. To get this data, another source of data was needed. And for this, there were two ways to do it, pull the data from WeRateDogs twitter feed using Twitter API and the tweet_id that is a primary key from the twitter archive dataset. Another way of getting the retweet and favourite count is by writing the already provided retweet and favorite file into an empty file and loading this file into panda data frame.

Finally, additional data was provided on the classification of the breeds of dogs. This dataset contains the predicted breeds of dogs and the confidence rating of the prediction. This data was downloaded programmatically using requests.get() method. The downloaded dataset content was writing into panda data frame.

Data Assessing –

The assessment of the gathered data was done in two ways, visual assessment and programmatic assessment. Some noticeable issues with the three dataset at the end of assessment are as follows

Quality Issues

- There were unusual names in the name column of Twitter archive dataset.
- The tweet_id datatype should be string
- There were some missing values in 'in_reply_to_status_id', 'in_reply_to_user_id', 'retweeted_status_id', and 'retweeted_status_user_id' column.

- There was an unusual value as the maximum value of the denominator count
- The timestamp had a wrong datatype; string.
- Some dog stages names are ‘unknown’
- The image prediction dataset has lesser number of rows than the twitter_archive data.
- The choice of naming the column of prediction as p1, p2 is not exactly descriptive.

Tidiness Issues

- Dog names should be in one column not 3
- All the dataset should be merged for better analysis
- Some columns are not necessary for the analysis

Data Cleaning

The first step of the cleaning process was to make a copy of the whole dataset so as to preserve the data in case I want to reverse some decision or I need something in the future. The data quality issues were addressed. Null vaues were removed and redundant columns were dropped. Image Prediction dataset was cleaned by selecting the highest confidence rating and dropping the rest of the rating. The column names were renamed from p1, p2 to dog_ breed and confidence rating. The timestamp datatype was properly converted to datetime. The denominators that were lower than 10 and higher were eliminated and the numerator count was limited to 20 as maximum. The unusual name in the name column was cleaned by dropping the names that did not start with capital letter. Finally, all the three dataset were merge into one data file for easier visualization and analysis.