WeRateDogs Twitter Archive – Wrangling Efforts

In this project, I (Prem Rajamohan) performed the Data Wrangling steps (Gather, Assess, and Clean) on three different datasets related to the WeRateDogs twitter feed. In this report, I am outlining my thought process and the efforts taken to wrangle the different data pieces.

Gathering

For this project, I gathered data from three different sources

1. Enhanced twitter archive feed – this is considered as the 'main' dataset and this was provided directly to me as a 'csv' file
2. Image predictions feed – this data set contains the predictions of a neural network algorithm that attempts to classify the pictures of the dogs in the tweets. This was hosted by Udacity and downloaded by me using the 'requests' library
3. Retweet and Favorite count – I obtained this dataset from the Twitter API using the tweepy library. There are multiple other columns in the Twitter APT but the project's focus is around retweets and favorites

Assessing and Cleaning

I assessed each data set individually and noted the quality/tidy issues first. Only after cleaning each of the issues, I went about merging the three different data pieces together to form a final clean master table.

In the enhanced twitter archive feed, I did the following –

1. The timestamp column was of string data type and I converted this to a datetime format
2. There were retweets and replies included in the data feed. I removed them since this project focused strictly on actual tweets
3. WeRateDogs is known for its unique rating system where the numerator is almost always < 10 but I noticed few cases where the denominator was <> 10 and the numerator was extreme values. I removed such records
4. The name column contained invalid values such as 'a' etc. but I did not clean them since that column is not going to be particularly useful
5. I combined the 4 columns (doggo, puppo, pupper, and flooder) into one column since very few records had multiple categorizations and the data is much more tidy this way
6. I removed the records where expanded_urls was NULL

In the image predictions feed, I did the following –

1. I got rid of the columns p2 and p3 since p1 is the most confident answer and hence most likely to be true
2. Once I did that, there were few records where the p1_dog column was False and I deleted such records since we cannot trust those values

3. I also removed the jpg column since that is not relevant and is already available in the base feed
4. I also made a mental note of the final number of tweets here since the tweets that do not join to the enhanced twitter feed will be considered missing

In the twitter APT feed – there were no real issues but just a few missing records when compared to the enhanced twitter archive feed and they will have missing values.

Finally, I combined the three different datasets together and created a clean table with ~2K records and only the relevant columns.