Wrangle Report

Introduction

This project involves gathering data from WeRateDogs twitter account. They rate dogs based on the images uploaded by dog owners who would like their dogs to be rated. For this project we are using three datasets.

Data Wrangling Steps

The three datasets used are the twitter-archive-enhanced.csv, tweet-json.txt and, image-predictions.tsv. The following are the steps undertaken in my wrangling process.

Step 1: I loaded the three datasets and created a list to store the data in the twitter_json.txt and read the contents one by one before appending the retrieved data in another and then creating a data frame from the data.

Step 2: Created a copy for each of the datasets so as to be able to get back the original data in case I messed up. I also converted the timestamp to datetime format.

Step 3: I dropped some columns that were not necessary for my analysis in twitter_archive_clean. I also renamed the id column in the the data frame created from the twitter_json.txt data to tweet_id so that it could match the other two data frames. From the dogtionary, the breed floofer is known as floof and this was renamed also. This step also involved dropping the columns with missing values in the twitter-archive-enhanced.csv file.

Step 4: Converted the tweet_id columns to string and merged the three datasets together using the common column tweet_id.