Wrangle Report

Introduction

WeRateDogs is a Twitter account with over 4 million followers that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators are almost always greater than 10. 11/10, 12/10, 13/10, etc. This project is mainly focused on the data wrangling process of the WeRateDogs Data Analysis Project. The wrangling process was done in three stages.

Data Gathering

In the first stage, the datsets needed for the analysis was gathered. We use three datasets for this project. The first dataset (WeRateDogs Twitter archive dataset) was downloaded as available on the udacity website. The second dataset (Image prediction) which contains a table full of image prediction was downloaded programmatically using the Requests library while the third dataset which contains the numbers of likes and retweets for each tweet had to be queried from the Twitter website using the Twitter API. The data queried was stored as a json data which was programatically read line by line to extract the tweetID, likes and retweets and finally converted to a dataframe.

Assessing

In the Second stage, the datasets was first assessed visually using a spreadsheet application(Microsoft Excel) and then programmatically using some functions. While assessing the dataset, some few tidyness and qualities were noticed. Some quality issues included lots of missing values in reply to status information columns, values in ratings denominator greater than 10 in Twitter archive table, Underscore in the dog names column in image prediction table, False results in predictions, meaning some of the pictures are not images of dogs. Also some of the extracted columns (e.g dog ratings and dog names) had a few errors. All these issues were well documented in the analysis.

Cleaning

After assessing and documentation of these issues, the next stage involved taking actions to clean and handle these issues. Different python functions were programatically used to perform cleaning operations and afterwards tests were carried out on the dataset to ensure that they were properly cleaned. The datasets were merged after all issues had been handled and stored