Wrangle Report

This report illustrates how the data was wrangled.

The dataset wrangled within this project is the Twitter user @dog_rates, also known as we WeRateDogs. WeRateDogs is a twitter account that rates people's dogs with a humorous comment about the dog. The ratings have a majority of its ratings with a denominator of 10.

The numerators at times were greater than 10. e.g 11/10, 12/10, 13/10, etc. This is so because the dogs were very good dog bents. The followers of WeRateDogs stood at over 4 millions. The account has received international media coverage

This report was completed with a Jupyter notebook and exported to several document extensions. The Wrangling process or steps of the project was structured into three (3) Steps:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

They were further explained below:

1. *Gathering Data*

The master data was a combination of three (3) different data sources connected with the tweet_did

1. Enhance Tweet Archive:

This contains data extracted programmatically from tweet data sent by WeRateDogs to Udacity through email which is intended to be used solely for this project. Ratings, dog name, and dog stage and some related information were some of the content of the data.

The file was downloaded from a link and imported with panda's library (pd.read_csv) into the main project for analysis. Here is the link for the data https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv which was downloaded, uploaded and read into a pandas DataFrame.

2. The tweet image predictions

This file is present in each tweet according to a neural network. It is produced by running every image in the WeRateDogs Twitter Archive by classifying the breeds of dogs. Along side each tweet ids were the image url, and the image predictions number which corresponds to the most confident prediction ranging from 1 to 4 because tweets can have up to 4 images. The dataset is hosted on Udacity server and it was downloaded programmatically using the Requests Library from the following url:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Additional data from the Twitter API

The purpose here was to gather each tweet's retweet count and favorite ("like") count at the minimum. Hence, it was attained by using querying twitter's API and tweepy library which later stored in a txt file called tweet_json. A twitter developer account was used to get the API keys and tokens.

After querying, the necessary fields pulled from the API for this analysis were the tweet id, retweet count and favorite count. The data was then saved to a 'tweet_data.csv' file for future purposes. This was done without the index column to avoid wrong field name.

2. Accessing Data

After gathering the 3 datasets, they were individually assessed virtually and programmatically for both quality and tidiness issues:

The following issues were noticed and resolved:

- Tidiness:
  - Dog stage data is separated into 4 columns
  - There is a relationship amongst the data but they have been separated
- Quality:
  i. There are 181 retweets based off the retweeted_status_ids
  ii. Some dog names are invalid (e.g, a, an, & the, instead of a normal name)
  iii. Invalid tweet_id data type (e.g integer instead of string)
  iv. Invalid timestamp data type (string not datetime)
  v. 440 rating numerators less than 10 (e.g 1998)
  vi. Row 313 has 0 denominator
  vii. 23 rating denominators not equal to 10
  viii. Underscores were used as a peparator of columns p1, p2 and p3 instead of spaces

3. Cleaning Data

The above issues were later cleaned and stored in a master dataset.