Wrangle Report

Introduction

This dataset is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. The objective of this project is to wrangle the messy Twitter data to create interesting and trustworthy analysis and visualizations.

The following steps were taken to wrangle the data

- Gathering data
- Assessing data
- Cleaning data

Step 1 : Gathering Data

1. The WeRateDogs Twitter archive. Downloaded the file manually from the following link: twitter_ archive_enhanced.csv
2. The tweet image predictions, i.e., what breed of dog is present in each tweet. This file(image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Using the tweet IDs from the Twitter archive, the Twitter API was queried for each tweet's JSON data using Python's Tweepy library and stored the entire set of tweet data in a file called tweet_json.txt file.

Step 2 : Assessing Data

As a second step of the Data wrangling process, assessing helped to figure out several Quality issues (content related) and some tidiness issues (structural types).

Quality

- Timestamp column is a string
- There were many Missing/ Misspelled dog names
- Some text lines contain links
- Exclude columns that are not required for analysis
- Rating denominator higher than 10 and some lower than 10
- Retweets present in the file: texts start with "RT @"
- There were "&amp" characters present in text
- Some breed names have the first letter lowercase in p1, p2, p3 columns
- id column needs to be renamed to tweet_id to make it similar to other data frames

Tidiness

- Dog stages split into 4 different columns in twitter dataframe
- Combine the 3 separate data frames into 1

Step 3: Cleaning Data

After Assessing the data the data it was time to clean the data, which means addressing all the quality and tidiness issues listed above. Copies of each of the dataset were made before the Cleaning step. Cleaning also involved removing duplicates and missing data. The dataset was then stored as twitter_archive_master.csv.