# wrangle_report

November 22, 2021

## 1 Wrangling Report for WeRateDogs

### 1.1 Table of Contents



### 1.2 Introduction

This report is to document the the data wrangling I undertook as part of the Wrangling and Analyze Data project for the Udacity Data Analysis Nanodegree. The data we wrangled was from the a twitter archive called WeRateDogs.

### 1.3 Step 1: Gathering Data

I used a number of data sources in this first step, some supplied as part of the project while the rest I queried.

- Twitter Archive Data: A csv file supplied as part of the project and downloaded (from: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv).
- Tweet Image Predictions: a tsv file programatically downloaded from https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv which includes the dog breed predictions along with links to the jpg.
- API & JSON: After creating a twitter account, and gaining developer access (which was a painful experience when dealing with the analyst processing the request!) I was able to use Twitter API. Using the provided key, token & secrets I extracted tweet_id, retweet_count & favorite_count the latter two as integers. This pulled data used the Tweet IDs from the Archive Data.

### 1.4 Step 2: Assessing Data

At this stage I did both manual assessments & programatic assessments of the data. The manual assessments I did both within the workspace & within excel for both quality & tidiness issues.

Below I will go over the Quality & Tidiness I uncovered for each of the Data Sources we used.

**Twitter Archive Data**

- We can see the NaN value (indicating blanks) repeatedly show up in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded urls
- Some of the rating denominators are not 10. After manually going through these and filtering out replies and retweets we can see 5 that need fixed and 1 that needs deleted for lacking a rating at all.
- Fix: After so many requests, this is Bretagne. She was the last surviving 9/11 search dog, and our second ever 14/10. RIP https://t.co/XAVDNDaVgQ
- Fix: Happy 4/20 from the squad! 13/10 for all https://t.co/eV1diwds8a
- Fix: This is Bluebert. He just saw that both #FinalFur match ups are split 50/50. Amazed af. 11/10 https://t.co/Kky1DPG4iq
- Fix: This is Darrel. He just robbed a 7/11 and is in a high speed police chase. Was just spotted by the helicopter 10/10 https://t.co/7EsP8LmSp5
- Fix: This is an Albanian 3 1/2 legged Episcopalian. Loves well-polished hardwood flooring. Penis on the collar. 9/10 https://t.co/d9NcXFKwLv
- Delete: Meet Sam. She smiles 24/7 & secretly aspires to be a reindeer.
- A lot of the entries under the name column are not correct, the most predominate "None" & "a".
- Tweet_id should be a string but is currently an integer.
- The Timestamp is in object (string) format and should be in datetime format.
- 181 lines of data need to be removed as they have a valid retweeted_status_user_id showing they're retweets.
- 78 lines also need to be removed as they have a valid in_reply_to_status_id showing they're replies.
- Multiple dog "stage" (doggo, floffer, pupper, puppo) columns aren't needed.

**Predictions**

- Again Tweet_id should be a string but is currently an integer.
- Multiple predictions for dog type/breed are in each tweet, we don't need more than 1 prediction.
- Some of the predictions aren't valid dog breeds.

**Tweet JSON**

- No issues, we already have our counts as integers due to the earlier mentioned part in Step 1.

**Common Issue**

- The three datasets need merged into a cleaned master file.

### 1.5 Step 3: Cleaning Data

I undertook the following steps to Clean the Data:

**All**

- Made copies of each of the datasets.
- Merged into a Masterfile.

**Twitter Archive Data**

- Removed the 181 retweets.
- Removed the 78 replies.
- Converted the Timestamp from object to datetime format.
- Converted Tweet_id from an integer to a string.
- Manually fixed the 5 with incorrect ratings based on their text.
- Deleted the entry with no score.
- Created a new rating column based on the rating_numenator & rating_denominator. This should tidy up the non-10 valid denominators.

**Predictions**

- Converted Tweet_id from an integer to a string.
- Droped extra dog predictions, we only need p1 not p2 & p3.
- Droped non-dog predictions.

**Tweet Json**

- None.

**All**

- Made copies of each of the datasets.
- Merged into a Masterfile.

### 1.6 Further Steps

From here on we are no longer doing Data Wrangling but working on the twitter_archive_master.csv dataset to draw insights and create useful visualisations for the WeRateDogs Twitter Data.

```
In [ ]: #### Convert to PDF
        !jupyter nbconvert --to pdf wrangle_report.ipynb
```