# wrangle_report

March 26, 2019

## 1 Wrangle Report

### 1.1 Gathering Data

**Enhanced Twitter Archive** I was provided this data as a .csv file, so I initially read the file. This file contained information on tweeter's ID as well as rating information and dog stage.

**Tweet Image Predictions** I downloaded this data as a .tsv from a URL with the Requests library. This file contained information on the image predictions of the breed of dog (or other objects) in each tweet.

**WeRateDogs Twitter Archive** I created a new data frame by, firstly, querying the Twitter API for each tweet in the Twitter archive and saving JSON in a .txt file and then, using pandas, creating the data frame by extracting tweet_id, retweets, and favorites data from the .txt file.

### 1.2 Assessing Data

First, I visually assessed the three data frames listed above for any issues. I then used some common programmatic assessments in pandas to find any further issues.

#### 1.2.1 Quality Issues

The Enhanced Twitter Archive file contained missing values for dog stages, contained multiple values for dog stage, contained incorrect rating numerators and rating denominators, contained unnecessary entries related to retweets, needed 'None' values to be replaced with 'NaN' for the dog stage entries, and contained unnecessary columns (retweets and expanded URLs).

The Tweet Image Predictions file contained entries where p1_dog, p2_dog, and p3_dog were all "False". Also, entries in p1, p2, and p3 were not consistently lower case.

The WeRateDogs Twitter Archive file contained columns that were the wrong data type (objects instead of integers) and contained missing 'favorites' data.

#### 1.2.2 Tidiness Issues

The Enhanced Twitter Archive file contained four different rows for the dog stages instead of one, rating_numerator and rating denominator were in different columns instead of one. The three data frames also needed to be combined.

## 1.3 Cleaning Data

**Enhanced Twitter Archive**

1. I created a new data frame omitting rows with missing values for dog stages by using the Pandas `.query()` function.
2. I created a new data frame omitting rows with multiple values for dog stages by using the Pandas `.loc()` function.
3. I corrected the rating_numerator and rating_denominator columns by using the Pandas `.text.str.extract()` function and then converting rating_numerator and rating_denominator to floats by using Pandas `astype(str).astype(float)` function.
4. I created a new data frame omitting retweets by using `.query()`.
5. I replaced 'None' with 'NaN' for all dog stages by using the NumPy `.where()` function (so that the columns could then be merged).
6. I dropped columns pertaining to retweets and expanded URLs by using Pandas `.drop()` function.

**Tweet Image Predictions**

1. I created a new data frame omitting entries where p1_dog, p2_dog, and p3_dog were all "False" by using the Pandas `.query()` function.
2. I converted strings in p1, p2, and p3 to lowercase by using the Pandas `.str.lower()` function.

**WeRateDogs Twitter Archive**

1. I converted all columns in the dataset from an object to an interger by using the Pandas `.astype(str).astype(int)` function.
2. I created a new dataframe without any missing values for the favorites column by using the Pandas `.query()` function.