

# Wrangle\_report

## 0.1 Reporting: wrangle\_report

This report details the process of wrangling a tweet archive from the Twitter user @dog rates, commonly known as WeRateDogs.

The dataset was separated into three sections:

- The first upgraded Twitter archive of 2356 tweets, including basic tweet data, rating, dog name, and dog "stage."
- Next, more data from the Twitter API, containing interesting columns such as retweet count and favorite count.
- Image Predictions Database

After acquiring the data, the next step was to analyse it, beginning by reading it into a pandas dataframe then, utilizing programmatic approach methods such as `.head()`, `.info()`, `.describe()`, various characteristics, tidiness, and quality concerns were checked out. Some of the basic observations made during the evaluation stage included: incorrect data type for some characteristics, extreme values in a few columns (commonly referred to outliers), a number of unnecessary columns, and so on.

Following the assessment stage, the data was cleaned based on the observations made.

- I began by making copies of the original pieces of data.
- Other cleaning processes included removing retweet and reply rows as the analysis process concerns just the tweets, extracting numerator rating from the text column (the initial extracted numerator rating was not properly extracted), dealing with inconsistency in the rating denominator column by dropping affected entries, creating new column from existing columns (rating was removing the extreme values from the rating (%) column, converting certain columns to their appropriate datatype and cleaning invalid names from the name column.
- One intriguing wrestling process was removing the extreme values from the rating (%) column (the rating (%) column was created by dividing rating numerator by rating denominator and multiplying the result by 100).
- The interquartile range approach is used to remove the outlier, specifically by selecting a higher threshold. In order to properly separate out the extreme numbers, I noticed the various quantiles beginning with the lowest, 25th percentiles, 50th percentiles (median), 75th percentile, 99th percentile, and 100th percentile (maximum). A closer examination of the 99th and 100th percentiles revealed the presence of an outlier. To deal with this, I set my upper limit using the code below:



```
#Finding outliers
import numpy as np
from scipy.stats import iqr
iqr = iqr(rate_dogs_clean['rating(%)'])
upper_threshold = np.quantile(rate_dogs_clean['rating(%)'], 0.99) + 1.5 *
iqr
```

Another type of cleaning that involved data structure (tidiness) was done specifically with Twitter archive data. The dog stage column, which was anticipated to be a single column, was split into four independent columns, which were combined using the `str.cat()` method columns into a single column called dog stage.

Finally, to facilitate analysis and minimize redundancy, I consolidated the three different tables into one master dataset and saved the dataframe in csv format.