Data wrangling is a practice in Data Science that involves gathering, cleaning, analyzing and storing data for analysis and presentation as the case may be.

In this case study, we used data from WeRateDogs Twitter handle.

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The rating is done on a scale of 10, however exceptional ratings may exceed 10, so the data contains ratings like $\frac{12}{10}$, $\frac{13}{10}$, $\frac{15}{10}$, etc.

Python and relevant libraries are used for the process of data wrangling, visualization and presentations. The core ones are:

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Json
- Tweepy

This article does not cover the process of the data wrangling but focus more of presenting the insight of what data WeRateDogs have.

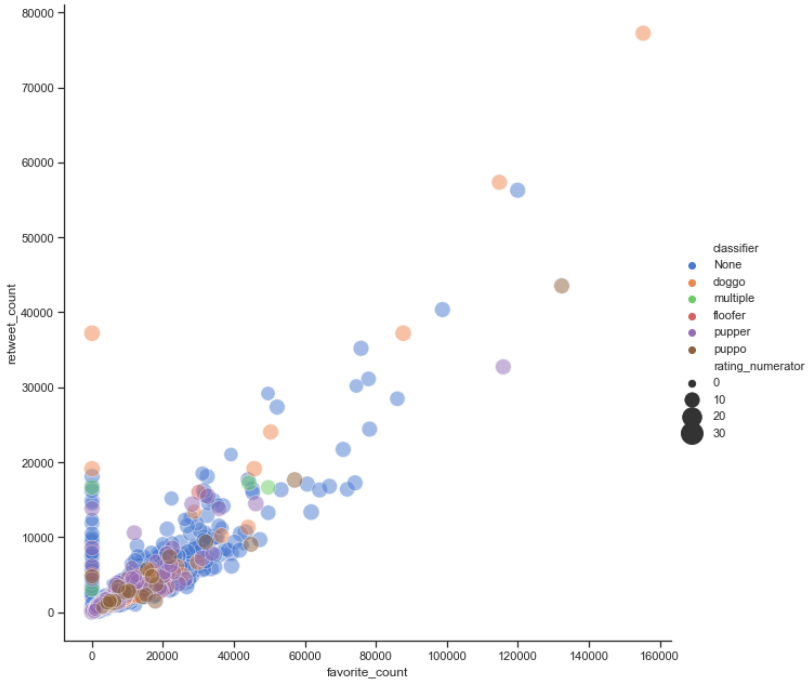
All the relevant data are merge into one after the wrangling process. A summary of the structure is shown below. The wrangling is mainly for educational purpose so the scope covers identifying 8 cleanliness issues and fixing them.

```
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   tweet_id            2040 non-null   object
 1   timestamp           2040 non-null   datetime64[ns, UTC]
 2   source              2040 non-null   object
 3   text                2040 non-null   object
 4   expanded_urls       2040 non-null   object
 5   rating_numerator    2040 non-null   int64
 6   rating_denominator  2040 non-null   int64
 7   name                2040 non-null   object
 8   doggo               2040 non-null   object
 9   floofer             2040 non-null   object
 10  pupper              2040 non-null   object
 11  puppo               2040 non-null   object
 12  retweet_count       2040 non-null   int64
 13  favorite_count      2040 non-null   int64
 14  jpg_url             2040 non-null   object
 15  img_num             2040 non-null   int64
 16  p1                  2040 non-null   object
 17  p1_conf             2040 non-null   float64
 18  p1_dog              2040 non-null   bool
 19  p2                  2040 non-null   object
 20  p2_conf             2040 non-null   float64
 21  p2_dog              2040 non-null   bool
 22  p3                  2040 non-null   object
 23  p3_conf             2040 non-null   float64
 24  p3_dog              2040 non-null   bool
dtypes: bool(3), datetime64[ns, UTC](1), float64(3), int64(5), object(13)
memory usage: 372.5+ KB
```

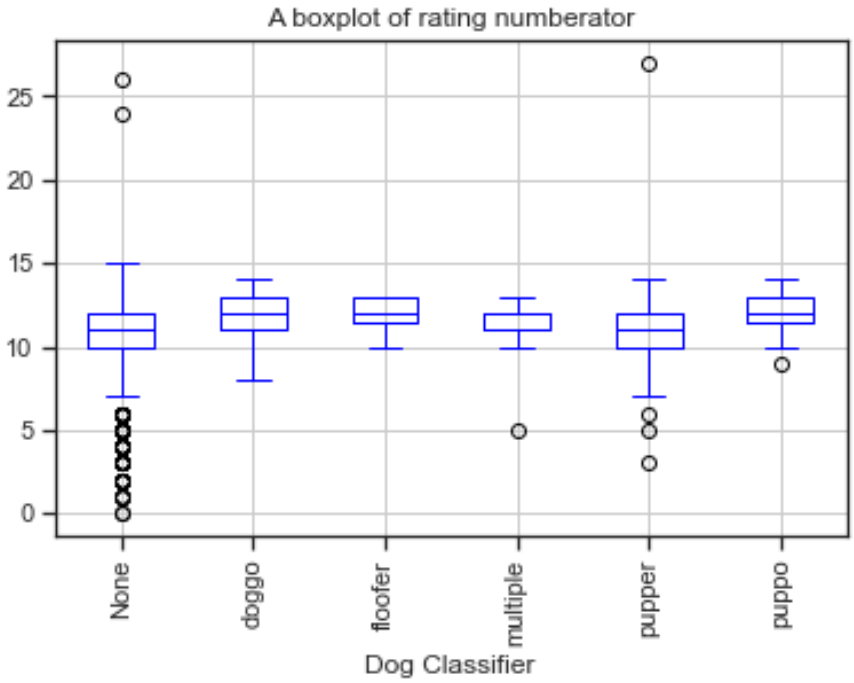
Table 1- WeRateDogs Dataframe after Wrangling

Insights from the Data

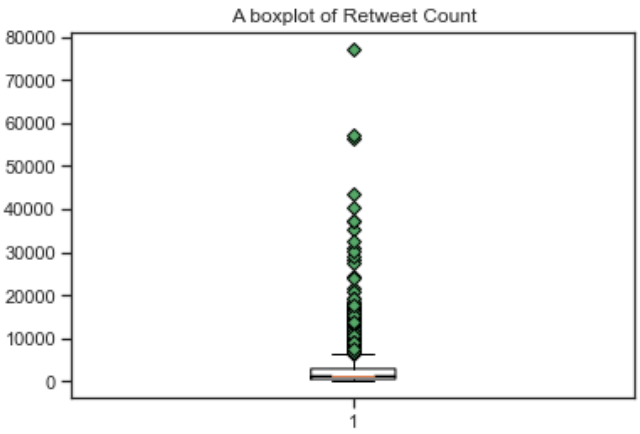
1. The favorites and retweets the tweets get are closely correlated and most of the ratings in our data for the classes of dog stages are 10 above as seen in the scatter plot below.

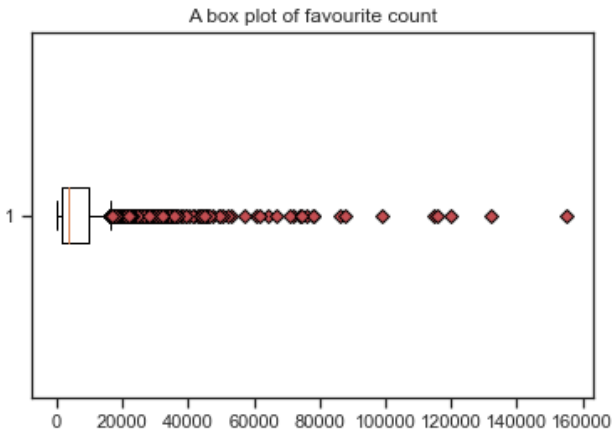


2. We can check for outliers in our data using a box plot. The figure below depicts the outliers in the dogs' stage classification against the rating numerator.

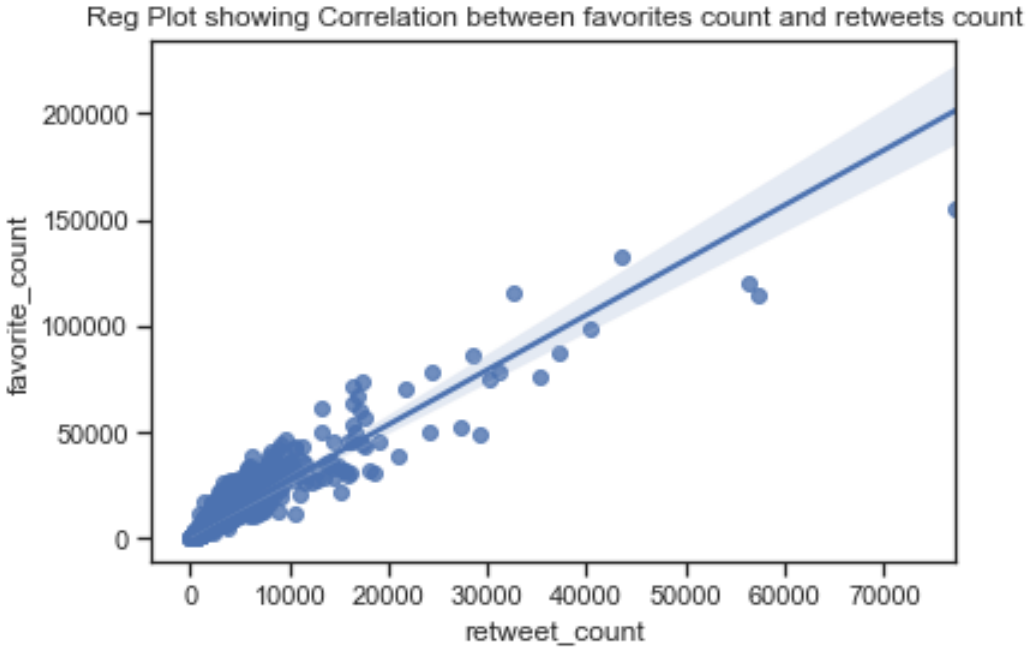


3. We could also do the same for retweet counts and favorite counts. See below respectively.

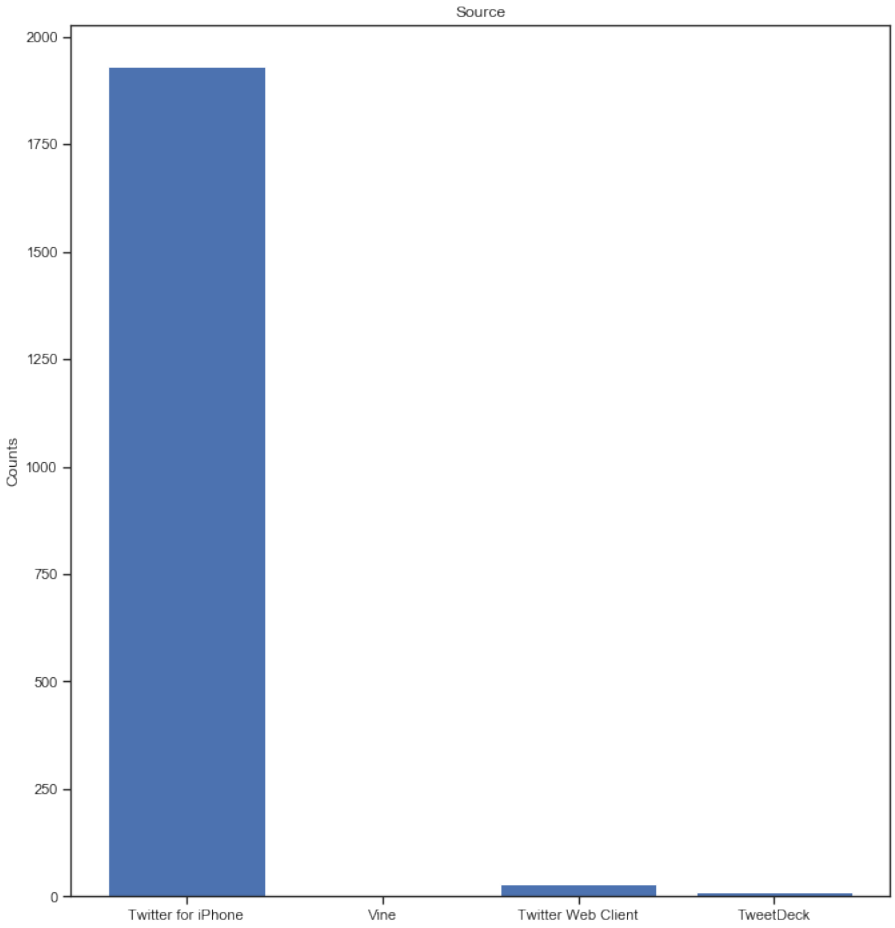




4. The correlation between retweets and favorite can be well shown using a Regression Plot. The figure below shows that there is a strong correlation between the two variables.



5. We want to give account of the devices used to interact with WeRateDogs account. A bar plot will come in handy here. The diagram below shows the information about devices used to retweet and favorite the tweets



More could be put to use from the cleaned data, especially in generating ML algorithms or Neural networks but this is beyond this article.