Introduction

The dataset we have used for wrangling 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 are given out of 10. After scrapping the data, tidiness and quality issues were assessed and then cleaned.

The following steps are involved in the data wrangling process:

- Gathering Data
- Assessing Data
- Cleaning Data

Gathering Data:

Data was gathered mainly from three sources:

1) The enhanced twitter archive file was downloaded manually and consists of variables such as tweet id, source, text, name, numerator rating etc.
2) The tweet image prediction file was hosted on udacity's server and was downloaded programmatically and consists of variables such as tweet_id, img_num, jpg_url etc.
3) Twitter API and JSON file was gathered using twitter's API and consisted of twitter_id, retweets and favourites.

Assessing Data:

All the three files were gathered and assessed in the jupyter notebook using various methods such as .head(), .sample(), .info() etc
After which all the tidiness and quality issues were resolved:

Quality Issues:

1) The timestamp should be DateTime instead of an object.
2) Removing columns which are of no use.
3) Only original tweets are required, retweets are to be removed.
4) Removing rows with column_rating other than 10.
5) Replacing irrelevant names starting with lowercase.
6) Relacing None in the names column with nan.

7) Combining doggo, floofer, pupper and puppo columns into a single column.
8) All the columns in df_tweet should be integers.

Tidiness issues:
1) Merging all the columns on tweet_id
2) Combining numerator and denominator column into a single column

Cleaning Data:

This part of the wrangling process was divided into 3 steps i.e. Define, code ad test.
Initially, the copy of each data frame was created and the cleaning was done on the copies of the data, not on the original data. So that if any error occurs original copy of dataframe is not ruptured.
Every cleaning step involved a different process for it to be cleaned and the complete process was stored in wrangle_act.ipynb.

Conclusion:

Data wrangling is a very and a core skill which is used by every Data Analyst and Data scientist. It is very rare that the data we want to use can be assessed from a single source and is clean. The data has to be collected from sources and with different processes and then can be put to use.