# WeRateDogs Data Wrangling Project :

This project was part of the data wrangling section of the Udacity Data Analyst Nanodegree program. It is dedicated to data wrangling concepts. In addition to deploying skills like web-scraping, for gathering data from different sources. Assessing the quality of the retrieved data and cleaning it
WeRateDogs Twitter account tweets was analyzed using Python, documented in a Jupyter Notebook (wrangle_act.ipynb).

This project included:

- Data wrangling, which consists of:
    - Gathering data
    - Assessing data
    - Cleaning data
- Storing, analyzing, and visualizing our wrangled data
- Reporting on 1) Data wrangling and 2) Data analyses and visualizations

**Gathering Data for the Project :**

- The WeRateDogs ‘twitter_archive_enhanced.csv’ datafile was manually downloaded from Udacity project details.
- The second used file of image predictions (image_predictions.tsv) hosted on Udacity's servers was downloading with BeautifulSoup and Requests library
- I’ve had issues getting developer’s account so I contacted Udacity instructor and received the final json file and the code. I proceeded then to read each tweet into a list then store it in a dataframe called ‘tweet_info’.

**Assessing the Gathered Data :**

After gathering the data and reading it into dataframes, assessment comes next to check data quality and tidiness. I detected and documented the following quality issues :

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be strings instead of float, because we don't want any operations on them, they also have many NaN values.

- retweeted_status_timestamp, timestamp should be datetime instead of object (string).
- name, doggo, floofer, pupper, puppo columns include None values instead of NaN, but Pandas treats None and NaN as essentially interchangeable for indicating missing or null values so it won't be an issue.
- rating_numerator, rating_denominator are integers but prefered to be float.
- Some of the elements of the name of the dog are not real name like "a","None" "an", should be replacedby null-value.
- The timestamp column is an object. It has to be a datetime object.
- There are 2075 rows in the images dataframe and 2356 rows in the archive dataframe.
- In several columns, null values are not treated as null values.
- Missing values in 'name' and dog stages showing as 'None' instead of NaN.
- tweet_archive contains retweets
- Some tweets don't include images

**Tidiness Issues：**

- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo instead of one column filled with their values.
- Join 'tweet_info' and 'image_predictions' to 'twitter_archive'
- We can see that all three columns share tweet_id column, all are the same type so we can merge these dataframes on that column.

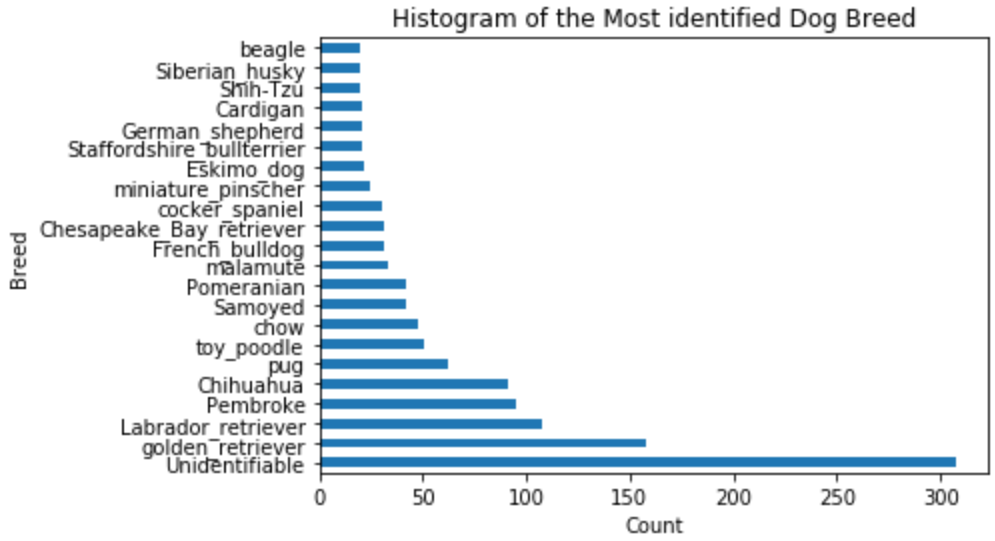
**Data Cleaning :**

Cleaning data is the third step in data wrangling process. Where I fixed the quality and tidiness issues that were identified in the assess step. It’s a very important step that ensures quality analysis with real and truthful insight.

It should be considered that maybe not all of the quality and tidiness issues were spotted and addressed in the assessment section but most of it had been done in the wrangle_act jupyter notebook.

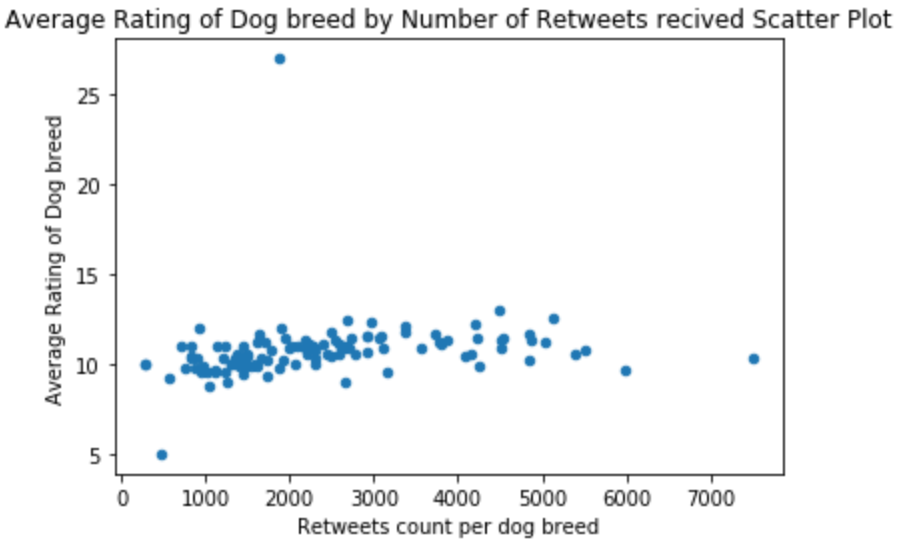
Data wrangling is a core concept that every data scientist should be familiar with to draw truthful insights from the different datasets available in the world. Analyzing, visualize, or modeling unclean and messy datasets could lead to missing out on important insights or drawing wrong conclusions.

**Analysis and Visualizations :**




From graph we can see that there are more than 300 Unidentifiable dog breeds, which could mean that the algorithm used needs more training and more data to draw predictions from. Golden retrievers, Labrador retrievers, Pembrokes  and Chihuahuas are by far the most identified breed, maybe the most owned by users.

On other hand, clumber, EntleBucher, Scotch_terrier, Japanese_spaniel, standard_schnauzer, Bouvier_des_Flandres and Irish_wolfhound are least identified dogs for the algorithm. We can't conclude why these breeds are the top breeds and others aren’t. It could be because they are commonly owned or they are the easiest to identify by the AI algorithm.




In this scatterplot, the data points line up nicely, however, the fact that the line would be horizontal means that the input values (that is, the x-values) are irrelevant to the output values (that is, the y-values).

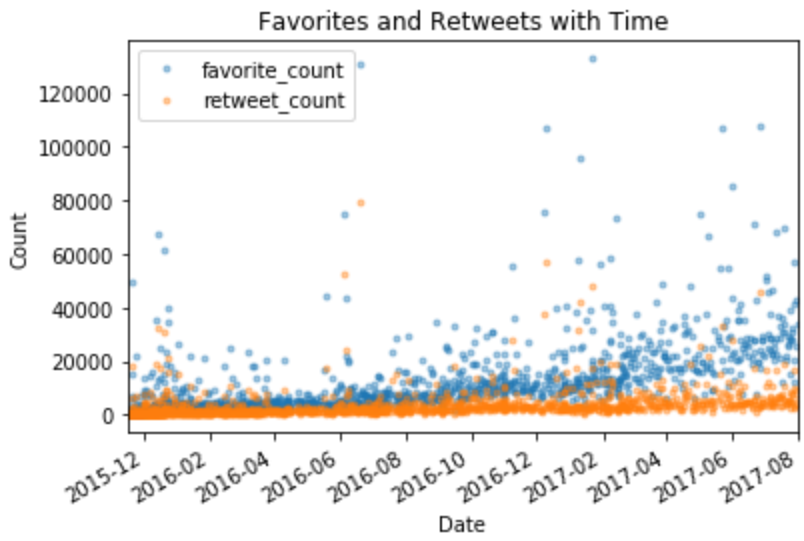
So there is a definite trend to the data, and there is an excellent good-fit line for it, but that line only says that the input values are irrelevant. If the inputs are irrelevant, then there can't possibly be a correlation between inputs and outputs. That is, the amount of retweets received by a dog breed doesn’t correlate in anyway with the average rating received,.

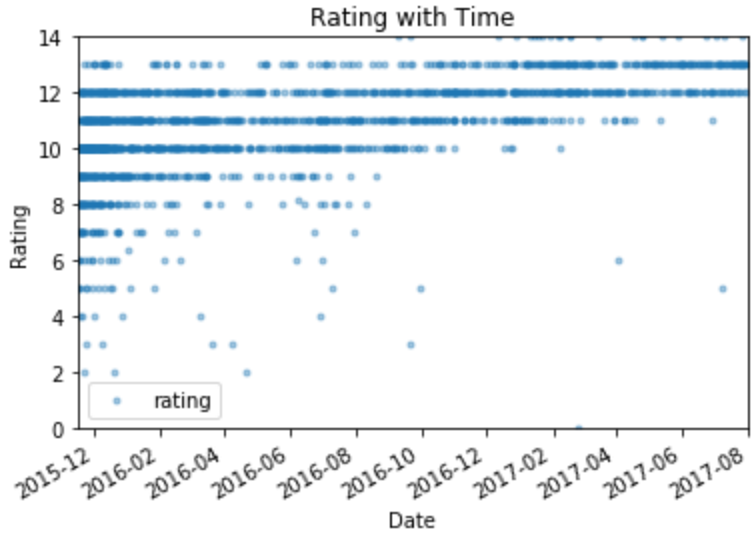
A similar scatterplot was done for Favorite counts and a pattern similar to this was achieved.

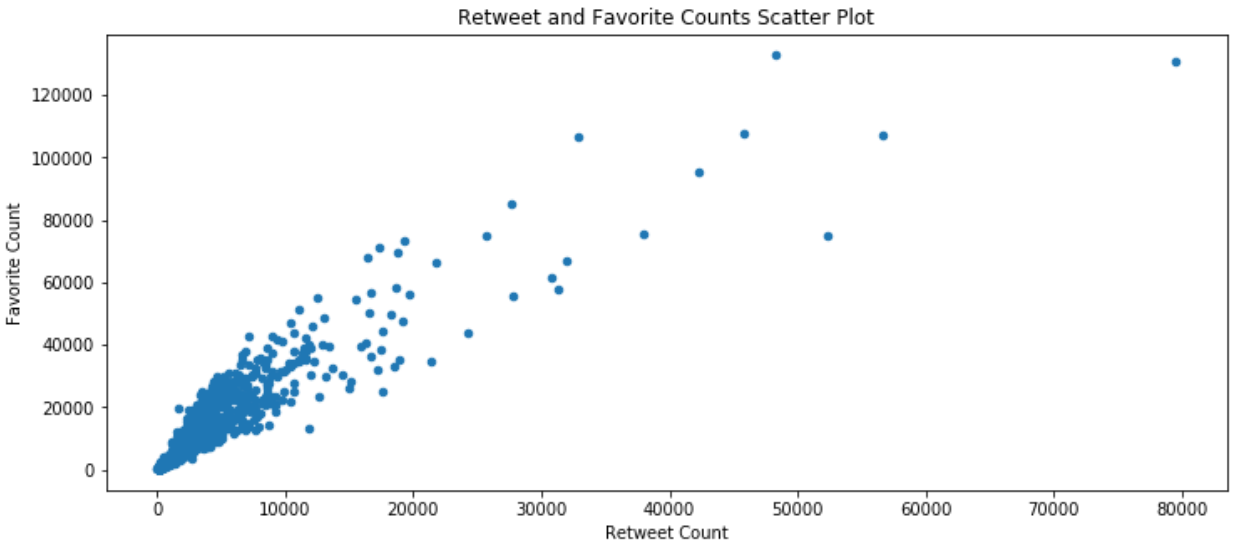
In addition to a statical analysis was done regarding dogs age Floofers minimum rating is above 10. We can assume that Floofer stage is the most favorable by users and floofers are consistently good dogs.

**Effect of Time :**

From these two plots, we can see the gradual increase of both favorites and retweets over time, and ratings which could mean more users have started following the account, making it more popular and rating more dogs.


Favorites and Retweets with Time
favorite_count
retweet_count
Count
Date
0
20000
40000
60000
80000
100000
120000
2015-12
2016-02
2016-04
2016-06
2016-08
2016-10
2016-12
2017-02
2017-04
2017-06
2017-08


Rating with Time
rating
Rating
Date
0
2
4
6
8
10
12
14
2015-12
2016-02
2016-04
2016-06
2016-08
2016-10
2016-12
2017-02
2017-04
2017-06
2017-08




From this scatter plot we can that there is almost a proportional relationship between the retweets and favorite counts, it could be said that the amount of favorites are almost as twice the retweets.

**Conclusion :**

Data wrangling is a core concept that every data scientist should be familiar with to draw truthful insights from the different datasets available in the world. Analyzing, visualize, or modeling unclean and messy datasets could lead to missing out on important insights or drawing wrong conclusions.