Introduction:

This Wangle and Analyze Data Project is part of the Udacity Data Analyst Nanodegree program, the data set is from the twitter account WeRateDogs, as know as: @dog_rates, which provides information about dogs. The data set contains 2300+ pieces of data, which were stored in 3 different sources. After done gathering, assessing and cleaning the data.

Real world data sets rarely come in a clean form, and so does this one. Before we can do further analyzing job, we need to do some wrangling work to make it as suitable as much for the upcoming analyzing session. The wrangling work are separated into 3 parts: gathering, assessing and cleaning.

Gathering:

Since the data sets are stored in different forms, multiple libraries and tools are implemented. They are pandas, numpy, os, re, json, tweepy, matplotlib and particularly twitter API is used to gather favorite count and retweet count, and Request library are used to download the image_predictions.tsv file from the Udacity server, and so forth. The data sets are gathered from the files named:

twitter-archive-enhanced.csv

tweet_json.txt

image_predictions.tsv

Assessing:

The following methods are employed to assess the data:

.info()

.value_counts()

.head()

.sample()

These are the methods of checking the quality of the data sets both programmatically and visually.

Cleaning:

The following are tidiness issues and quality issues found in the data set, 2 for the tidiness session and 8 for the quality one, which are shown down below.

Tidiness issue:

1. Mergering data sets;

2. Simplifying dog stages.

Quality issue:

1. Removing retweets;

2. setting 'tweet_id' as string and;

3. 'timestamp' as datetime form;

4. & 5. Lower case for names and misspelling of names;

6. Inaccuracy of ratings.

7. Renaming some columns and;

8. Deleting some irrelevant ones.

Conclusion:

After done all of the 3 parts, the data set is ready for further analyzing.
However, some outliers in the data set are not removed on purpose.