We rate dogs. Data Wranging

This has been a challenging project, because there are a lot of data from 3 different sources:

- Original twitter archive provided by Udacity.
- Predictions data provided by Udacity from an url.
- Additional twitter data obtained from Twitter API.

The last source has been the most difficult at the beginning because it has been the first time I have used an API.
Then, it is so important the assess part, to see the characteristic of each source. There are 3 different data frames:

- Twitter archive has 2356 entries and 17 columns.
- Image predictions has 1075 entries and 12 columns
- Twitter info has 3999 entries and 3 columns.

The main data frame is twitter archive because is where almost all important information appears, and where I have found many quality issues.
After assessing the data, I have started with the cleaning part. First at all, it is important to join all the different data frames in order to work with only one. A problem has been that there were lot of columns and some of them, like the dog stage ones, were in four different columns. So, I have extracted the dog stage from text column and remove the no needed columns.
Regarding to quality issues, I have decided to face 8 I have found.
It has been important to deal with all the retweets in the data frame. So, the first thing I have done is to remove it, because they are not necessary. Also, there were tweets with no images and, as they are important in this account, I have removed these rows.
Some issues were related to how the information is shown. For example, text column did not appear complete, some sources were difficult to read and some ratings were incorrectly shown.
The other issues I have found were related to dog's names, some of them were incorrect and it makes no sense that the missing values appear as 'None' and no as 'NaN'.
Finally, some data types were not the most adequate.

After cleaning the info, I have saved it in a new file and see some insights, like the most and least days of activity, the favorite tweet, the most

common names. Also, I have done two graphs that shows the Retweets and favorites over the time, and the ratings over time.

In general, the part of cleaning has been the most difficult to me. Because it is difficult to discover the issues and how to deal with it.
I have had to look for information in the documentation and in stackoverflow.