Data Wrangling Udacity Project

The project is composed of the 4 standard wrangling sections: Gather, Assess, Clean and Store, plus two other sections, namely a small cleaning iteration section and an Analysis section.

Gather:
I dealt with 3 different files:
- *twitter-archive-enhanced.csv*, which I downloaded from the Udacity website
- *image-predictions.tsv*, which I downloaded programmatically using the Requests library
- *tweet_json.txt*, which I named as such after downloading the json formatted data with the Twitter API.
This was the only real 'Gathering' challenge.

Assess:
I used the key points of the project motivation as guidelines. there were many issues with the data I did not address. However, here is the list of those I did:

Quality:
1. Some columns were incomplete, with several empty cells.
2. Not all the tweets contained images (photos).
3. The tweet IDs were in the int64 format in all the three files.
4. There were many retweets in the the first file, which I cleaned as per key points request.
5. The denominator of the rating was often different from 10.
6. Some of the prediction entries did not represent dog breeds in any of the p1, p2, p3 columns.

Tidiness:
1. The columns 'doggo', 'floofer', 'pupper', 'puppo' were structured in such a way that if one of the words 'doggo', 'floofer', 'pupper', 'puppo' were in the tweet, then the value would turn to the respective column name (i.e. 'doggo', 'floofer', etc). It is more tidy to have one column say, "dog_phase", where we then specified each phase with numbers from 1 to 4. However, I noticed
that the same tweet could contain more than one of these nicknames/"dog-stages". Therefore, as a second option I could keep all columns. I chose the first option and described the choice in the wrangle act.
2. All the DFs had to be merged in a single tabular piece of data.

Clean:
In the cleaning phase I addressed programmatically all the issues assessed. As per missing data, I did not use imputation, rather I got rid of many entries for the sake of being sure about the quality of the data.
Final Cleaning Iteration:
In this phase I addressed a couple of data quality issues that remained after the three pieces were joined. Namely I had many rows with incomplete data as the result of the join and some data type issue. A third, tidiness issue attained the dog breed predictions. I fixed it merely because of the analysis I conducted later.

Store:
I stored the DF in a .csv file.

Analysis:
• Bottom 10 ratings including outliers
• Top 10 most frequently predicted breeds
• Top 10 most liked breeds on average
• Showing a correlation between Favorite count and Retweet count
This were 4 insights despite the 3 required. Amongst those 4, the first is worth addressing here as it could prompt a new wrangling iteration: it revealed that some of the low ratings referred to tweets whose pictures did not contain dogs in the first place.