Documentation of Data Wrangling Steps for Analysis of WeRateDogs Twitter Account

Data was gathered from different sources: a local csv-file, a programmatically downloaded file and data downloaded via the twitter API. After assessing the data visually and programmatically some issues were found.
Regarding the data loaded from the csv-file there were found some quality and tidiness issues. First of all, the tweets which could be identified as retweets or replies were dropped from the data frame, as they do not contain a valid rating.
Furthermore, some of the remaining tweets had missing values in the numerator and denominator column or had invalid values. The ratings were extracted from the text column with regular expressions and merged with the available values in the numerator and denominator columns. Additionally, the numerator and denominator were calculated to rating, so that there are not two columns for one variable. After that outliers were removed, any value higher than 1.4 or lower 1.0 was removed.
There were four columns: doggo, floofer, pupper, puppo, which describe the dog stage. These columns were merged to one column to ensure to have only one variable dog stage. Invalid names like the 55 entries 'a' in the column name were removed and the info completed by extracting names from the text column.
Furthermore, the datetime column was converted to datatype datetime.
After the cleaning process of 2356 tweets there are only 1191 tweets left, for 231 there is a value in stage and 837 values in name,

The programmatically downloaded dataframe contained three object predictions per tweet image. In the cleaning process only one prediction was kept, the predicted object with the highest confidence level while being a dog.

The columns of retweet count and favorite count of the data downloaded via the twitter API were converted to integer.

Finally a master dataframe was created by merging the data loaded from the csv-file with the programmatically downloaded dataframe and the data downloaded via the twitter API.