Internal Report on the wrangling data project

Brief review of the data wrantling process

The whole process of wrangling the data of WeRateDogs consists of gathering data of various types from different sources, assessing data in terms of quality and tidiness issues, and cleaning and merging the data.

Among various stages of the wrangling process, I hereby focus on my efforts into the data cleansing. The data set of tweets include various types of attributes. Some of them, such as the dogs' names and rating scores, are given in the text data. I paid attention to whether those attributes were consistent with the texts. Indeed, it turns out that somw rows had wrong names: "a", "very", and "the", for example. I looked into the text data in detail and tried imputing the correct names of dogs. Likewise, I closely looked into the text data with an emphasis on whether the rating scores in the text data were correctly extracted and input into the "rating_numerator" and "rating_denominator" columns.

Secondly, I made an effort into making the datasets compact without losing the essential quality of data. For example, I programmatically simplified four columns into a new one "dog_stage". A similar simplification was applied to the dog breeds as well. Furthermore, I created a new column which computes the rating scores. Moreover, I truncated the data predictions only to the most likely prediction. As such the data sets were simplified while redundant information was removed, so that we merged different datasets into a new one in terms of tweet IDs.

Admittedly, there are some limitations to our data wrangling procedure. For example, the text data do not always include all the information necessary for correcting the other columns. Indeed, there are still rows missing the information on dog stages and dog names. Moreover, further considerations should lead to another data quality/tidiness issue, which we might deal with elsewhere.

(302 words)