Wrangle report – Georgios Pallas

I started this project by gathering data related to the WeRateDogs twitter account. First, I read the archive file containing tweets and information about the dogs. Secondly, I gathered the prediction data about the dog breed from a url. Lastly, I used the twitter API to gather additional information (retweet and favorite count) for the tweets that are contained in the archive file.

After I have gathered all my data, I assessed them in order to identify quality and tidiness issues. Examples of quality issues are with regard to the data:

- Completeness
- Validity
- Accuracy
- Consistency

Furthermore, tidy data need to follow the following format:

- Each variable forms a column
- Each observation forms a row
- Each type of observational unit forms a table

I applied both a visual and a programmatic assessment to the data. In the visual assessment, I could observe issues like tidiness, for example, the dog types, although they form one variable, they were split to different columns and needed to be merged in one column. I could also observe quality issues like names that are not correct or missing values. With the programmatic assessment I was able to identify erroneous data types, values that might have been inaccurate, for example, the numerator and denominator values, duplicated values and others. Furthermore, by applying programmatic assessment I could delve into the details of the data. For example, some values that looked erroneous but they were correct, or some missing values, for example about dog types, that I was able to identify them in the text of the tweets.

After assessing the data, I cleaned them, for some of the quality and tidiness issues that I had identified. In the cleaning process, I followed the format:

- Define
- Code
- Test

And in some cases, in between the cleaning steps, I re-assessed some of my previous observations to see if they are still an issue (because some of the cleaning steps might had fix at the same time other issues).

After cleaning my data, I visualized and analyzed them. I analyzed issues, for example, with regard to the popularity of the tweets over time (using the retweet and favorite counts), or what is the number of tweets of the account over time, or how the confidence of the image predictions is distributed.

There were also issues that I didn’t clean. For example, in the text of the tweets in the archive dataset, there is short link that is not useful and could be cleaned. Or I could keep the short link instead of the expanded url column. Other issues that I didn't explore are with regard to the names of the dogs. I have

cleaned the erroneous names, however there might be in the text names of the dogs that have not been recorded in the dog column. Also, I could have treated different some of the values. For example, I decided to keep the numerator values that are between 8 and 14, due to the large number of tweets with these scores. However, I could have kept all the available numbers – usually the low rating was not about dogs, but it could have been useful to keep these scores if I was going to focus more on the prediction table and how accurate the image prediction was –.

I could conclude, that depending on the analysis I would like to perform and specific questions I would like to answer, then I would clean my data in a slightly different way. So, I can always re-assess and iterate on any of the steps, if I see that this would be helpful in my analysis. Finally, the analysis was based on the tweets between November 2015 and August 2017, to which image predictions were available.