wrangle_report

June 1, 2022

0.1 WeRateDogs - wrangle_report

- The dataset wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRate-Dogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

0.1.1 Data gathering

Data was gathered from 3 sources:

- Directly downloaded the WeRateDogs Twitter archive data (twitter_archive_enhanced.csv) from udacity
- Download the tweet image prediction from https://d17h27t6h515a5.cloudfront.net/topher/2017/August/predictions/image-predictions.tsv
- Additional query data gathered via Twitter API

0.2 Assesment

0.2.1 The following quality data issues were found:

1. Twitter archive: Dog Name contains incorect values (a, the etc)
2. There are tweets that are retweets.
3. Dropping unneded (missing too much values) columns using drop function.
4. Images without image url
5. Convert timestamp object variable to datetime.
6. Source links can be converted to a short category string
7. Image predictions: Make name from p1,p2,p3 title case
8. Ratings are objects.

0.2.2 The following data tidiness issues were found:

1. Dog stages should be in one column
2. Merge all data into one data frame

0.3 Cleaning

Cleaning process can be analyzed in depth in the wrangle_act notebook . Each step was documented using the template : * Define: Explaining the problem and the approach. * Code: The complete code that run to fix the data. * Test: Assess the data again to make sure the code works and fixed the issue

After the wrangling phase was terminated the result was saved into a csv file and can be used for future analysis.

0.4 Conclusions

When gathering data user data from external parties it comes with some „garbage" attached to it: some of it because of the way the apis are designed and some if it because of the way twitter was used the users. The performed data wrangling is not by all means an exhaustive one , for other vizualisations/insights additional cleaning may be needed.