wrangle_report

May 22, 2021

Verena Dietrich

1 Data sources

There are three data sources for information about the twitter account "WeRateDogs": 1. An on hand file for downloading provided from Udacity with a database of tweets (twitter_archive_enhanced.csv) 2. A file for downloading from a URL with predictions of what is shown on the images of the tweets (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv) 3. The twitter API that provides information about the tweets given in the first file

2 Gathering data

I downloaded the twitter_archive_enhanced.csv and named the data frame *twitter_archive*. I requested the file image-predictions.tsv from the URL, saved it as a tsv-file and imported it as a data frame with the name *image_predictions*. I generated a developer account to use API to collect further information about each tweet in the twitter_archive_enhanced.csv and stored the information in separated lines in a json-file with the Jason and the tweepy library. I imported the file as a data frame with the name *tweet_info*.

3 Assessing data

I analysed the data manually with the methods like .info(), .describe() .head(), .tail(), .duplicated(), isnull() and .plot() provided by the library pandas. That way I located several quality and tidiness issues:

3.1 Qulity Issues in the twitter_archive data frame:

1. The column tweet_id is of datatype integer (int64) and not a string.
2. The column timestamp is not datetime data type.
3. Some tweets are retweets and not the original tweet.
4. To ratings are not comparable if the denominators vary.
5. Most of the entries in the columns retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp are empty.

6. The entries in the source column are surrounded by html tags, but only the content between '>' and '<' is relevant.
7. The name column contains strings like "a", "the", "an", "my"....

3.2 Quality issues in the image_predictions data frame:

8. The column tweet_id is of datatype integer (int64) and not a string.
9. Some of the fist letters in the column predictions are written in capital letters, some not.

3.3 Tidiness issues:

9. In twitter_archive the categories "doggo", "floofer", "pupper" and "puppo" are in 4 seperated columns.
10. The tables tweet_info, twitter_archive and image_prediction are in sperate tabels. There are more issues that can be find, but in my opinion that are the most imported issues to fix to perform a detailed analysis that can lead in interesting and reliable insights.

4 Cleaning data

I started with the tidiness issues to prepare a single data frame with the inner merge to get only tweets with full information. That made the cleaning steps easier and with less repetition. After that I went through all quality issues by describing the issue, defining the tasks that are to do to eliminate the issue and testing the result.