Project “Wrangle and Analyze Data” wrangle report

There where 3 files and every file recived different treatment because they were very different.

First, the “archive” file.

This was our main file, which all the project build upon it. We added to this file the data we need from the other and didn’t use them anymore.

This file has a lot of problems – some of them we treated well, some of them we didn’t because we don’t need all the data for our investigation. We don’t need the text for example, or the Url’s (because we had the analyzed Url’s on the other file).

Some of the problems in this file:

quality:

- Erroneous datatypes (timestamp)
- in the text column, sometimes there isn't a text, just link, sometimes there is spoiled link like id 878404777348136000 or 873337748698140000 or 863471782782697000
- in some columns there are 0 values in favorite_count and thousands in retweet_count, it's not logical!
- the normal rating_numerator is like 10-20 , but there are nominators of hundreds ...and the max value in thousands.
- seems like we haven't all the urls - there are "2295 non null" from 2354 rows.
- the mean value of the rating_denominator isn't 10 which means that there are denominator different from 10.
- There is a None value as a string in some columns and not NaN

tidiness

- There are many urls in a big mass - inside the text, and in the expanded_urls, some are 2 or 3 urls in one columns, some are of image and some not.
- i was prefer all the "doggo,floofer,pupper,puppo" in one column, so i could do value counts easily.

Than we are going to the “image predictions” file.

In this file there wasn’t problems like in the first, there wasn’t things unorganized or null, but there were problems in the content of the file.

quality:

- seems that there are images which are not dogs at all ! i saw a turtle and elephant.
- seems like the system sometimes recognizes the dog just in third probability like in this url : https://pbs.twimg.com/media/CT5PY90WoAAQGLo.jpg there is a dog in a shopping cart. the system recognize it as a "shopping cart" and shoping basket in p1 and p2.

I refer to this problems as quality issues, because it is spoiled data which could damage our analysis.

On the third file, the “tweets-json.txt” there was a problem of a lot information we don’t need, or we have already.

The cleaning process

First we took from the json file only two columns we interested in “re-tweets” and “favorites” we don’t need all the other columns there.

We did it by “merge” function in pandas which gives us the ability to preform inner join.

Than we began with the “image predictions” file.

Why ? Because in this way if something isn’t there, it would automatically removed from the archive file. So we shouldn’t work twice.

We took that strategy:

1. Remove every line that there isn’t a dog recognition at all.
2. If there is a dog recognition , we will took the one with the highest probability.

So we did a dataframe with only the dog breeds and id’s , and merge it to the archive file.

On the archive file we removed all the columns we don’t need, and cleaned the columns we need. That's means when there are non logical values, we replaced them with the mean or the median values, we put maximum value to the “numerator” of 15, all the “denominator” we changed to 10, we replaced “None” with NaN, and put all the “staging” in one column.

By the end we finished with one clean file with dog breeds, staging, favorites and rating.

The columns in the new file (‘twitter_archive_master.csv’)are:

'tweet_id', 'timestamp', 'rating_numerator', 'rating_denominator', 'retweet_count', 'favorite_count', 'dog_breed', 'stage'