wrangle_report

June 15, 2020

1 Data Wrangling Report

1.1 Introduction

The task in the project is to gather data connected with the Twitter account "WeRateDogs" from different sources, assess the data, clean it and do some basic analysis.

1.2 Data Gathering

The data comes from different sources. First there is the data from the account @dog_rates (WeRateDogs). This was provided by Udacity to be downloaded manually from their website. This data is the tweet archive from the account itself. The data was kept in the dataframe `twitter_archive`.

Second there is data from a neural network from Udacity that classifies the breeds of the dogs within each tweet of the twitter archive. This data was to be downloaded programmatically from a Udacity webserver. For the wrangling process it was loaded into the dataframe `img_pred`.

The last data was additional data for the tweets. This data was to be downloaded from Twitter itself using an API. Afterwards the data had to be converted from JSON to a Pandas dataframe, named `add_twitter_data`.

1.3 Assessing

The data was not clean and it showed several tidiness issues. Some problems for the individual tables were:

1.3.1 `twitter_archive`

For several columns the dataype was not correct. For example the id columns were all int, but should have been str, as they serve as names like zip codes. The timestamps were were objects (strings) and had to be converted to datetime.

The table contains a lot of 'None' values which are are simple strings, but no NumPy NaNs.

The table contains also data that is not needed for the analysis, e.g. data of retweets. A number of rows also contain no photos, what makes this data unnecessary as well.

It is noticeable that in the column `name` a number of entries are not dog names, but articles like 'a' or 'an', and a lot of rows contain 'None'.

A few rows did not have the standard denominator for the rating of 10. While the numerator is supposed to be larger than the denominator, a smaller number exceeds 10 by far.

A tidiness problem was, that the table contained four columns for the dogs stages, i.e. that one variable is distributed over four columns which violates the principle that each variable forms a

column. ### `img_pred` The names in this table are not very descriptive, what makes it difficult to understand the variables. Also the datatype for the tweet id is an integer instead of a string.

To comply with the definition of a tidy dataset, the table has to be merged with the table `twitter_archive` so that each observation is a row.

1.3.2 `add_twitter_data`

Same as for `img_pred`, this table has to be merged with the table `twitter_archive` so that each observation is a row.

1.4 Cleaning

I started the cleaning process by merging the tables tables on the `tweet_id` column into the table `twitter_master` so that I can do some of the cleaning in one step, e.g. converting datatypes.

Basically I addressed all of the above mentioned issues. I replaced 'None' values in the table with NumPy NaNs, so that aggregations would not be distorted. I also replaced some not so descriptive column names with more descriptive ones. I dropped a number of columns, which were not needed, and a number of rows with missing or erroneous data, watching out that I do not lose too much of the data. And I have combined the four columns of the dog stages into one.

1.5 Analysing

Im my analysis I took a look at: * Distribution of the ratings * Distributionof the dog stages * The scatter matrix to see if there are any correlations * Distribution of the top 10 dog breeds