DAND Term 2, Project 3: Wrangle and Analyse Data - report

Data wrangling is about obtaining and preparing data for analysis, through a process of gathering, assessing and cleaning. Here, I gathered data from 3 different sources: a twitter archive made available for downloading, a related set of image predictions that I downloaded programmatically, and the WeRateDogs data as available through the Twitter API. The latter was the most complicated as it requires to obtain and set up a dev account, as well as the somewhat unwieldy Twitter API. Data records from these 3 sources can be connected using a tweet's unique identifier.

Assessment showed that in terms of data quality, the data were not as dirty as the medical data used in the Udacity lesson, perhaps because human intervention was only inolved in few steps of data handling: mainly the programmatic extraction from tweet text was not very sophisticated, the naming of columns, the inconsistent capitalisation of dog breeds.

However the structure of the data was relatively complex: there are different kinds of tweets (normal tweets, retweets, replies, quote tweets), of which the normal tweets usually have a standard format consisting of a dog picture and a rating. This format often also includes naming the dog, and sometimes adding a “dog stage”. While the same information is available about standard tweets (e.g. timestamp, favourite/retweet counts), different additional information can only exist for other types of tweets, such as replies or retweets.

This complexity made structuring the data according to the principles of data tidyness less straighforward. Data tidying included restructuring the available data into two data frames linked by the unique tweet_id, one organised around tweets as observational unit, the other organised around dog predictions as the observational unit. This restructuring into two data frames involved moving certain variables such as dog stage, name, rating from one dataframe to the other, as well as dropping some redundant columns in a final tidying step. Also, dog stage and tweet type wer encoded as single categorical variables.

The data cleaning steps can be summarised as changes to data type (unique identifiers, timestamps) ; correcting string extraction from tweet texts ; simplifying the columns, names and structure of the dog predictions table ; re-stucturing the data according to tidy data principles – dog stage and tweet type as a single variable, re-structuring records to fit into one of two tables representing a single observational unit.

The final analysis and visualisation was relatively straightforward, as the tidiness meant data had been structured to facilitate analysis: it was simple to answer questions relating to tweets (evolution of twitter account popularity, distribution and relationship between tweet favourites/retweets ; factors affecting tweet popularity) using one dataframe, and questions about dogs using the other dataframe (top dog breeds and names, the meaningfulness of the dog rating system).

In reflection, the process of imposing a consistent structure also sometimes meant removing some complexity to make data more uniform. For example, encoding “dog stage” as a variable taking on one of 6 different categories meant that those representing several dog stages were summarised as “several”, simplifying the complexity of all the permutations that are possible with 4 different dog stages. Another example was removing a small number of dog predictions originating from retweeting extraneous news sources, replies and self-retweets (which always have a 0 favourite count) – this made the data set more homogenous and easier to analyse as dog predictions corresponded to the standard

format used by WeRateDogs, and e.g. the complexity of favourite counts having a different meaning on retweets didn't need to be handled.

A second reflection is that it would have been very useful to define some clear criteria for data cleaning and restructuring at the outset of the process. However this was not really possible because there were no clear questions or overall problem set out that the data needed to answer, from which such criteria could have been deduced. As a result, it became an iterative process of data cleaning and becoming aware of questions that were of possible interest and that could be answered.