wrangle_report

August 10, 2020

1 Wrangle Report

The data wrangling process for this project encompassed three steps:

1.1 1. Data Gathering:

Gathering the data was done from three different sources via different methods. - The first file `WeRateDogs Twitter archive` was uploaded and read directly. - The second file `Image Prediction` was downloaded programmatically from the internet and read in the notebook. - Additional data `tweet counts` and `tweet favorites` were scraped using Twitter API for the tweets existing in the Twitter Archive file.

1.2 2. Data Assessing:

The three data sources were assessed both *visually* and *programmatically*. The assessment resulted in discovering 8 key quality issues and 3 tidiness issues.

1.2.1 The *quality issues* discovered were:

In the `Twitter Archive` dataframe:
(1) The existence of non-original tweets in the form of replies (`in_reply_to_status_id`, `in_reply_to_user_id`);
(2) The existence of non-original tweets in the form of retweets (`retweeted_status_id`,`retweeted_status_user_id`, `retweeted_status_timestamp`);
(3) Tweets that had no urls;
(4) A column that had an incorrect object data type (`timestamp`);
(5) Some rating denominators were invalid in the `rating_denominator` column;
(6) Some dog names were invalid in the `name` column;
(7) Redundant tags in the `source` column.
In the `Image Prediction` dataframe:
(8) There was a consistency problem in the formatting of predictions columns (not all predictions in `p1`, `p2` and `p3` began with an uppercase).

1.2.2 The *tidiness issues* were:

(1) Three columns of the `Twitter Archive` dataframe reflected the same variable and should have been recorded as a single column (`doggo`, `floofer`, `pupper`, `puppo`);

(2) The `Image Prediction` dataframe was in a wide format when it would be more efficient to have it in a long format;

(3) The three dataframes gathered represented the same observational unit (dog tweets) and better be in one dataframe.

1.3 3. Data Cleaning:

The data cleaning process was done for the quality issues first then for the tidiness issues.

1.3.1 Cleaning the *quality issues*:

In the `Twitter Archive` dataframe:
(1) Non-original tweets in form of replies were deleted;
(2) Non-original tweets in form of retweets were deleted;
(3) Tweets with missing urls were deleted;
(4) The `timestamp` column was converted to a datetime data type;
(5) The out-of-range denominators and corresponding numerators were rescaled to adhere to the common denominator for all ratings;
(6) Dogs with invalid names were renamed as having an unavailable entry of `None`;
(7) Redundant tags were deleted keeping only relevant content;
In the `Image Prediction` dataframe:
(8) All prediction columns (`p1`, `p2` and `p3`)were converted to a title-format.

1.3.2 Cleaning the *tidiness issues*:

(1) The three columns of the `Twitter Archive` dataframe were melted into one column `dog_stage`;

(2) The `Image Prediction` dataframe was converted to a long-format;

(3) The three dataframes were merged together, deleting missing columns and selecting only a subset of the `Image Prediction` dataframe that is of interest to the analysis.

Each step of the wranglin process is documented in detail inside the `wrangle_act` Jupyter notebook. The cleaning process is also documented with relevant definitions, codes and testing procedures.