Data Wrangling Report

WeRateDogs Twitter Archive

Real-world data rarely comes clean. This short report briefly describes the data wrangling efforts from gathering data from a variety of sources to cleaning it. This step is crucial to create interesting and trustworthy visualization

Data Gathering:

Data was gathered from three sources in `wrangle_act.ipynb` to reach more exciting and interesting analysis:

1. The WeRateDogs Twitter archive “`twitter_archive_enhanced.csv`”: downloaded this file manually given by Udacity
2. The tweet image predictions* “`image_predictions.tsv`”: downloaded programmatically using the Requests library
3. Additional Data from the twitter API “`tweet_json.txt`”: Each tweet's entire set of JSON data was stored in a file. Some fields were then read line by line into a pandas DataFrame.

*The tweet image Predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network.

These files were then read into dataframes: *twitter_archive, image_predictions and tweets_data* respectively.

Data Assessing:

After gathering each of the above pieces of data, they were assessed visually and programmatically for quality and tidiness issues. Multiple issues falling under completeness, accuracy, validity and consistency were detected and documented. Tidiness issues were identified according to Hadley Wickham’s data rules. For this project, 16 quality and 2 tidiness issues were assessed to be cleaned in the next step. Besides the Jupyter Notebook, Google Docs was also used especially for reviewing tweet texts.

Data Cleaning:

First, I started off by taking copies of the original three DataFrames. Then, the issues identified during the assessment phase were then cleaned by converting them into defined cleaning tasks which are then turned into code and tested to ensure the cleaning operations worked.

Finally, to achieve a tidy dataset, all the three tables were joined together as all the data is related to one observational unit "Tweets" .

*** Feature Engineering:

To enhance our DataFrame even more, a new column "engagement" was created by adding the "retweet_count" and "favorite_count" columns.

Storing Data:

The DataFrame was then stored into a CSV file named `twitter_archive_master.csv.`