Data Wrangling report

by Japheth Rutoh

Data: WeRateDogs dataset

This report describes briefly the efforts made in wrangling the data from WeRateDogs twitter.

The WeRateDogs dataset was found in three separate sources. These are the following:

- A provided flat file - `twitter-archive-enhanced.csv`
- A link provided to get data hosted privately by Udacity - `image_predictions.tsv`
- Data retrieved using the Twitter Api - `tweet_json.txt`

For the first data the `pandas` library was required to read the flat file; with the use of pd.read_csv().

The hosted link required the use of Python's inbuilt library `requests` and `BeautifulSoup` to parse through the response, in order to get the required data.

The data on twitter was retrieved using `tweepy` , a twitter API. The data received was in JSON format and this required Python's `json` library.

The next step of the wrangling process used was **Assessment**. Visual and programmatic assessments were used in this step.

For visual assessment, scrolling through the datasets in tabular form. This helped find the most issues in the data.

Programmatic assessment required use of pandas `pd.DataFrame.info()` , `pd.DataFrame.head()` , `pd.DataFrame.describe()` functions and many more.

The next step was to briefly document each issue as we found it. This helped to make it easy later in the cleaning process.

This section is divided into two:

- Detecting **Quality** issues

  > Issues that prevent us from doing useful analysis on the data.

- Detecting **Tidiness** issues

  > Issues that affect the structure of the data.

The next step was going through each of the issues and fixing them. This is the **cleaning** process. For each issue, we first described how we would solve the issue in the **define** step, then we fixed the issue in the **code** step. After coding, came the **test** step where we wrote code to check if the issue was fixed.

> **NB**: While fixing some issues we would find newer issues that needed fixing. This proved a point in the Wrangling process that this process is *iterative*.

In this whole process we found and fixed 13 issues found in the dataset.

After finishing up the wrangling process, we needed to save the final data in a format for later use. We saved the gathered, assessed and cleaned data in `twitter_archive_master.csv`.

Next we provided insights we found while wrangling the data in a few sentences.

The final step was to provide a few visualizations to back the insights found in the data wrangling process.