

School of Computer Science

Data Wrangling in Fulfilment of

DATA9910

Maksymilian Drzezdzon

C15311966

Degree: TU060/1

Module Coordinator: David Leonard

Declaration of Ownership: I declare that the attached work is entirely my own and that all sources have been acknowledged: 🗹

**Date: 2020/10/25**

Data Cleaning

This assignment is centered around a Kaggle data set from 122 Brazilian weather stations. [1]

The data set in question consists of a single comma separate value (CSV) file, with a variety of missing values, different data formats and redundant columns which this section aims to tackle. Any points made in this section can be inspected further in the attached code file.

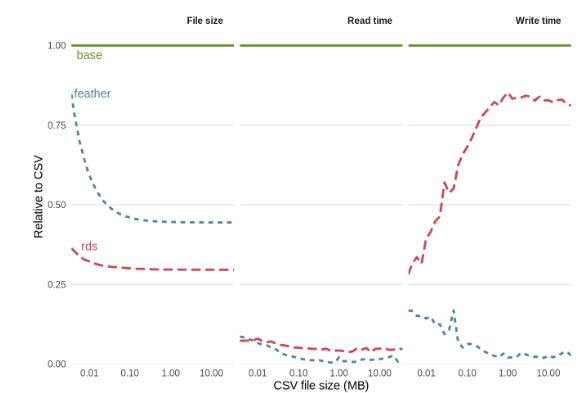
Renaming columns to be more descriptive was a priority, as it makes deciphering what is being observed clearer. Selecting an appropriate data format for ease of loading followed suit.

Figure 1: Binary file format performance benchmark [2]

The feather format was chosen because of its compatibility with R and Python. It offers superior load and write speed, which is ideal for constantly updating and reloading data. What the feather format is not good for is storing data long term, as frequent updates to the package do not guarantee compatibility with previously compiled feather files. [3]

The year, month, day and hour columns where converged and formatted into an existing but unformatted date and time columns before being dropped as they’re no longer needed, shaving off 300 Mb.

The extracted data frames are then concatenated back into a single separate file as a quick reference

Data exploration

This section focuses on further nuances exhibited throughout the data set and researching libraires that could help with further data exploration. Examining how missing values are dispersed guiding later decisions such as which columns one should focus on or drop.

Such libraries or methods are glimpse and str both preforming similar functions of printing contents and column names, summary which provides general statistical information on each column if one is not explicitly specified. A powerful function called skim provides a comprehensive statistical overview of each column if one is not specified such as standard of deviation, quartile values and a rough but useful histogram visualizing each columns data skew.

Interpolation can be used to handle empty values if there is a good enough reason for missing data or if it can be easily inferred.

Separating a few randomly selected cities and analyzing each providing deeper insights and potentially providing context on some missing data. Merging final data frames into a single dataset to support analysis conducted in this report.

Data analysis and visualization

This section builds on findings from the previous section and expresses them. comparing a few selected cities comparing and contrasting, generating questions that are further expressed in the code

Data analysis and visualization

Supplemental material

Garrett Grolemund and Hadley Wickham, R for Data Science, 2016

References

**PROPPG, 2016** <https://www.kaggle.com/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region> [1]

**Figure 1: Section 5.4.3** <https://csgillespie.github.io/efficientR/input-output.html> [2]

**Hadley Wickham, 2016, What should you not use feather for** https://blog.rstudio.com/2016/03/29/feather/ [3]