

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 centred 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.

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, as illustrated in Figure 1. 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 column before being dropped as they’re no longer needed, shaving off 300 Mb. The aim was to also fill out any missing values using the columns mentioned above.

The extracted data frames are then concatenated back into a single feather file for easy of access. Missing data can be handled by dropping missing rows, interpolation or imputation which fills missing data with the mean or median value.

The reasoning behind dropping these rows is that after exploring the years 2007-2010 for the Sao Gancalo station there are large chunks missing which possibly infer that said weather station was not in operation in that year/period and imputing data for a whole year is not a good solution as it won’t be a good representation. Those years will be omitted from further exploration.

Another reason for dropping said rows is that there is almost one million rows of data, finding a more intact chunk shouldn’t be much of an issue with so many weather stations to choose from. Interpolation in this case produces a straight line between the gaps since there’s not much to infer from as most gaps are in ranges of a few months to a year.

The air temperature column had to be revisited as it introduced anomalies into the analysis, upper and lower bound temperature averages weren’t matching up to other sources. There was a discrepancy between when a weather station was established and the period after. Data collected further down the line of a weather station did not have such gaps in values and had much less missing values.

This was tackled by replacing 0 values in the air temperature column with NA followed by using na omit as an easy way clean it up, this levelled out the averages collected in earlier years with very little change in the later years.

Data exploration

This section focuses on analysing changes in temperature, rain fall and humidity over a few years all of which are variables of interest, among others, when inspecting and trying to forecast droughts. The final goal is to predict/forecast using historical data or identify patterns in the above variables of interest. All graphs unless stated otherwise have been tailored using data specifically from the Sao Goncalo weather station.

An interesting pattern emerged as the project went on. During the analysis of temperature data it was found that the higher and lower bounds of min and max temperatures were very high as compared to the average which in the early years was extremely low (10°C) and far apart from other years. After some background reading and validating data through other online sources.

Such as weather-spark [4] where the kaggle dataset values were compared with weather-sparks data and come to a consensus what can be trusted and what can or should be dismissed as an equipment fault etc. The hypothesis is that as the population in San Goncalo went up in parallel with temperature as can be seen on population.city [5] and as the population stabilised so has the temperature rise. This can be observed in Figure 6.

In Figure 5 it was very suspicious that the average temperature had such a perfect correlation with the upper bound temperatures, as aesthetically pleasing as it was this raised an alarm that unearth that the original analysis had anomalies which skewed the averages.

Humidity was observed to make sure its consistent with rainfall as there should be a connection between the two. Following on from that, similar metrics were observed for the whole country and compare it to Sao Goncalo.

Data analysis and Visualisation

Figure 2 is a quick mock up of a forecasting attempt of humidity levels in Sao Goncalo over the year two thousand and eight. There isn’t much that can be concluded from this graph as the forecast creates a straight line. The reason Sao Goncalo was chosen is because it borders the parts of Brazil most prone to droughts. [7]

Given more time to explore other forecasting methods of times series data maybe a better solution could have been paired. In Figure 2 is a depicting the temperature fluctuations in two thousand and eight. Its difficult to determine the causes of such steep drops. It seems unrealistic that temperatures would have such high drops all of a sudden. Potential causes could be faulty or poor quality equipment and potential outages etc.

Figure 4 is a quick screen grab from google of temperature fluctuations in Sao Goncalo. Figure 3 and 4 have a twelve years difference but its a fair estimation that temperatures don’t drop that low in Sao Goncalo supporting that there must have been some form of outages in said weather station, from this the dataset was then cleaned further and any zero values were also dropped from the temperature column.

Appendix

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

**Sao Goncalo Monthly Climate Averages,** https://www.worldweatheronline.com/sao-goncalo-weather-averages/madeira/pt.aspx

Figure 1: Binary file format performance benchmark [2]

Figure 2: Forecast of humidity using historical data

l

Figure 3: Sao Goncalo temperature 2008

Figure 4: Current Temperature Fluctuations in Sao Goncalo [3]

Figure 5: Comparison between temperature averages and upper bound max values

Figure 6: Updated comparison between temperature averages and upper bound max values

Figure 7: Rainfall levels compared with upper bounds max values

Figure 8: Humidity levels compared with upper bounds max values

Figure 9: Rainfall levels from all of Brazil compared with upper bounds max values

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]

**Average Weather in São Gonçalo Brazil**, https://weatherspark.com/y/30562/Average-Weather-in-S%C3%A3o-Gon%C3%A7alo-Brazil-Year-Round [4]

**What is the population of São Gonçalo?** Answer: São Gonçalo, Brazil (Administrative unit: Rio de Janeiro)http://population.city/brazil/sao-goncalo/ [5]

**São Gonçalo, State of Rio de Janeiro, Brazil, 2020, Weather averages,** https://www.google.com/search?rlz=1C1CHBD\_enIE857IE857&sxsrf=ALeKk03\_1IAilNMrNltFeZwAYHHgKI4AEg%3A1605381806982&ei=ri6wX-HKO7Wf1fAPocyRoAI&q=s%C3%A3o%20gon%C3%A7alo%20brazil%20temperatures&oq=s%C3%A3o%20gon%C3%A7alo%20brazil%20temperatures&gs\_lcp=CgZwc3ktYWIQAzoECCMQJzoHCCMQsAIQJ1CXkAZY4ZkGYJKcBmgAcAB4AIABVYgBpQSSAQE4mAEAoAEBqgEHZ3dzLXdpesABAQ&sclient=psy-ab&ved=0ahUKEwjh2dWi4YLtAhW1TxUIHSFmBCQQ4dUDCA0&uact=5 **[6]**

[7]