

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 two datasets from Kaggle the first is of 122 Brazilian weather stations between the years 2000-2017 and the is split into three files which are from 1961-2019. [1] [10]

The data set in consists of a four comma separate value (CSV) files, 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. Amalgamating the two of the three files to compose one full dataset was the first big task that needed to be handled. Unneeded columns were dropped and then merged on the station id and observation time columns before being merged with the rest of the data files. Two variables were chosen to be the merge point in order to avoid duplicating other rows.

The feather format was chosen because of its compatibility with R and Python. It offers extremely fast 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. 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 ease of access. Missing data can be handled by dropping missing rows, interpolation or imputation which fills missing data with the mean or median value.

Another reason for dropping said rows is that there are 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 fewer 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 analyzing 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 make an assertion or infer future weather conditions in Sao Goncalo 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. Not all-weather stations have been in operation since 2008 creating a slight discrepancy in having less data for some weather stations.

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 stabilized 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 unearthed that the original analysis had anomalies which skewed the results. Another approach to handle this would be to normalize the dataset with a log transformation which is similar to what was done or box-cox transformation, however data needs to be positively skewed in this case it isn’t.

Humidity was observed to make sure it’s 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 and Sao Paulo.

The average temperature in brazil is 22 degrees Celsius with an average humidity rating of 83.7 out of 100 as gathered by the 122 weather stations across the country. Temperature ranges from 22 to 25.8 degrees in Sao Goncalo, unfortunately there is a lot of data missing for Sao Paulo, the assumption is because of the drought and potential power issues much of it was not recorded, from what was recorded is the heights of rainfall activity. These ranged from 1.6 mm as far as 3.2 mm with the lowest recordings being in 2012 and 2016 at 1.6 mm and 1.6 mm respectively.

Other less important graphics were rendered in the code file as they wouldn’t fit in the written report, these included humidity trends, rainfall in other years etc. Other cities were assessed and picked based on the amount of data that wasn’t missing.

Data analysis and Visualisation

Figure 2 is a quick mockup 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, its ~450 Km away from Sao Paulo. [7] [8] Approximately 70% of electricity in Brazil is generated via hydropower plans, because of this, droughts have a propound impact on Brazil. [7] [8] [9]

Given more time to explore other forecasting methods and experience analyzing times series data a better solution could have been paired. In Figure 2 is a depicting the temperature fluctuations in 2008. It’s 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 it’s 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, after this observation was made the dataset was updated accordingly as mentioned above.

Rainfall in Sao Goncalo had a sharp jump from 1.0 ml to ~2.5 ml between 2010 and 2012 but entered a downward trend thereafter as seen in Figure 7. This isn’t particularly alarming as rainfall levels are still higher than they were in 2008, weather conditions are difficult to predict and this analysis doesn’t have all the data needed to make assertions about the outcome of this change along with its long-term effects.

However, temperature upper bounds have been on an upward trend along with average temperatures recorded in Sao Goncalo, which can be the effect of a rising population in said area however this information isn’t enough to claim the region will go into a drought more so make an observation based on the data available that the rainfall fluctuations and rising temperature should be inspected closer.

Since there is little data from Sao Paulo and other weather stations that are in the red/drought region, its difficult to make assertions about the state of Sao Goncalo, however because of its proximity to Sao Paulo, downward temperature and rainfall trends it is definitely on the radar for problems further down the line.

**Closing statement**

Sao Goncalo definitely in the potential red zone for future droughts, however more data, time and analysis is needed to make a more educated prediction.

The sort falls of this analysis are that, much of the needed data is missing. Interpreting and analyzing meteorological results is not easy due to the difficult to predict nature of weather conditions. A lack of expertise or an expert resulted in consulting online resources that vary in quality putting much weight on the data available or lack thereof to make judgements on the analysis.

Appendix

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

Different ways to impute in python https://www.youtube.com/watch?v=m\_qKhnaYZlc&ab\_channel=DataSchool

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

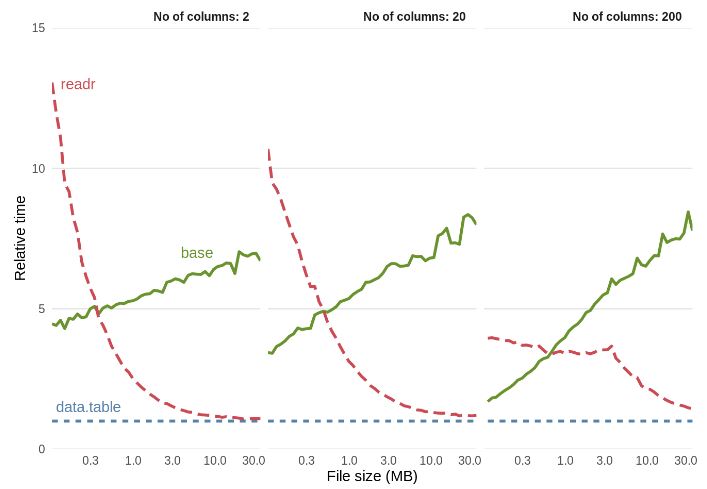
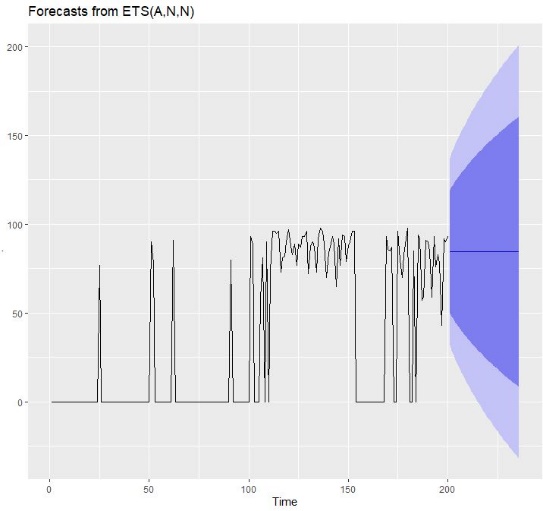
 

Figure 1: Binary file format performance benchmark [2]

Figure 2: Forecast of humidity using historical data

l

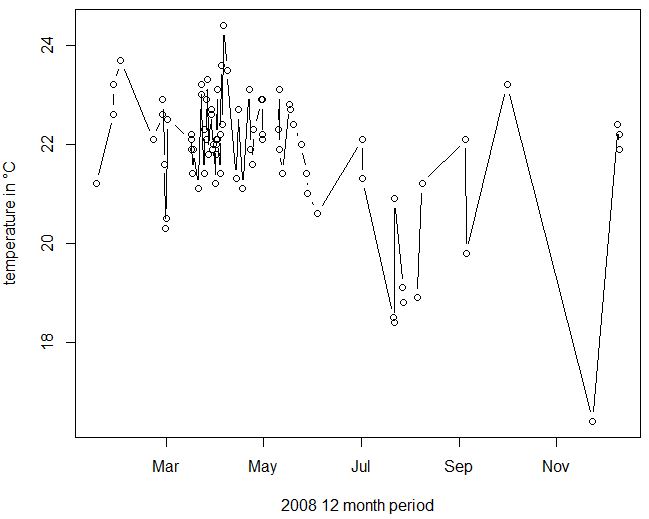


Figure 3: Sao Goncalo temperature 2008

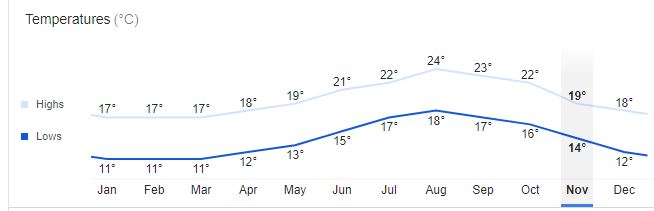


Figure 4: Current Temperature Fluctuations in Sao Goncalo [3]

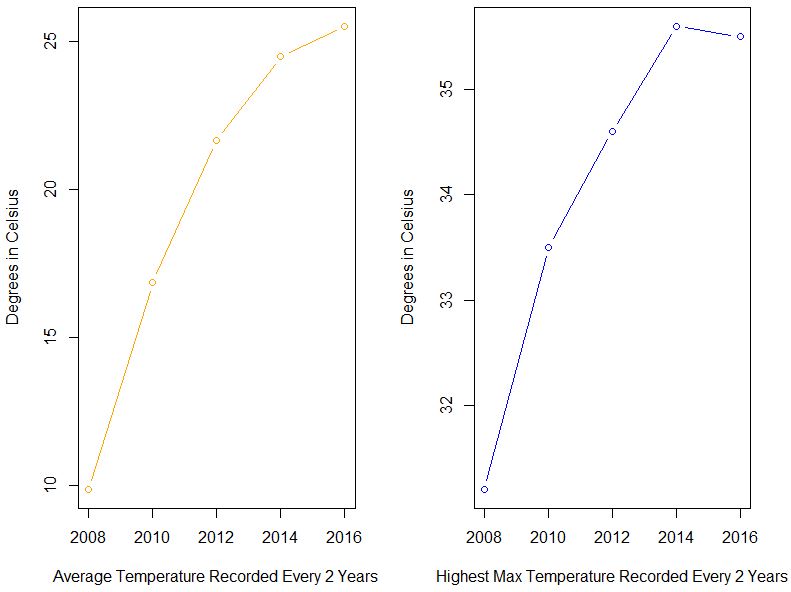


Figure 5: Graphs with anomaly comparing upper bound and average temperatures

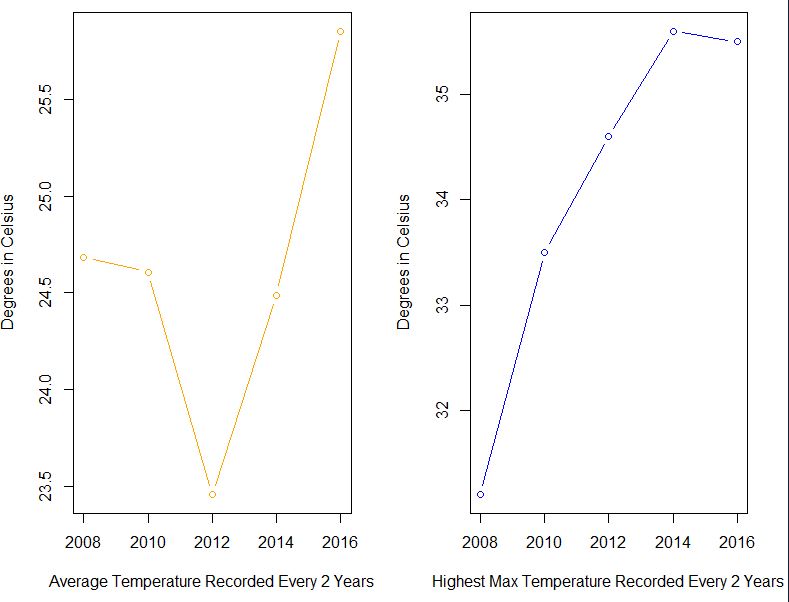


Figure 6: Updated comparison between temperature averages and upper bound max values

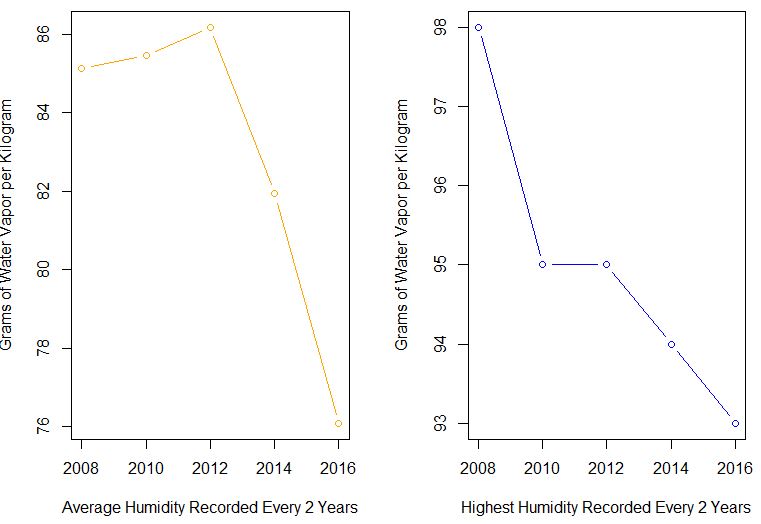
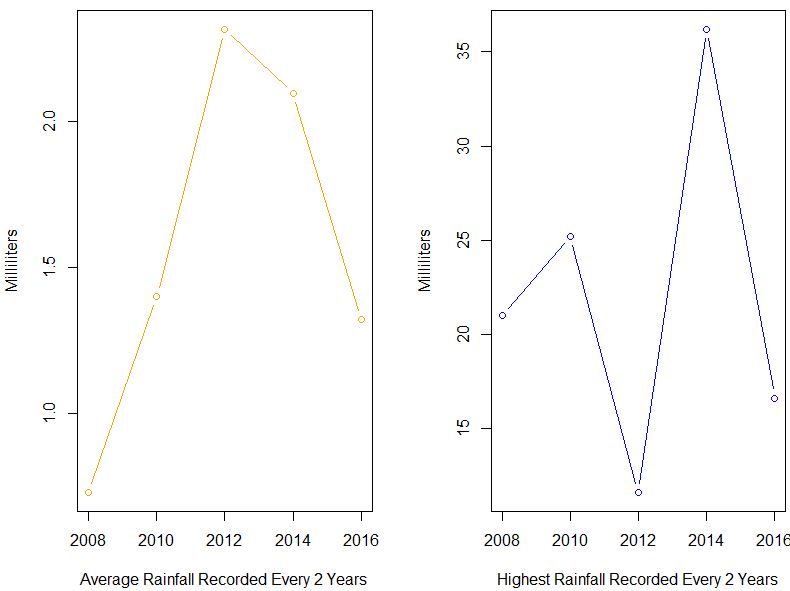


Figure 7: Rainfall levels compared with upper bounds max values

Figure 8: Humidity levels compared with upper bounds max values

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]**

**Brazilian drought, 2014-2017,** https://en.wikipedia.org/wiki/2014%E2%80%9317\_Brazilian\_drought#:~:text=The%202014%E2%80%9317%20Brazilian%20drought,worst%20drought%20in%20100%20years.&text=In%20these%20areas%20the%20rains,the%20drought%20worsened%20from%202015. [7]

**Drought and Water Crises Series Editor: Donald A. Wilhite, unknow year, Page 4** <https://openknowledge.worldbank.org/bitstream/handle/10986/28559/9781498765664.pdf?sequence=1&isAllowed=y> [8]

**Drought and Water Crises Series Editor: Donald A. Wilhite, unknow year** <https://openknowledge.worldbank.org/bitstream/handle/10986/28559/9781498765664.pdf?sequence=1&isAllowed=y> [9]

**Marciano Saraiva, 2020**, <https://www.kaggle.com/saraivaufc/conventional-weather-stations-brazil?select=weather_stations_codes.csv> [10]