# Estudo sobre Consumo de Eletricidade e Água

**Sidclay da Silva**

**Agosto 2020**

# *Electricity and Water Consumption Study*

***Sidclay da Silva***

***August 2020***

## Introdução

Este documento contém um estudo autônomo que inclui a análise do consumo de água e eletricidade na residência do autor, e a relação deste consumo com as variações de temperatura e umidade relativa do local.

## Descrição do problema

O consumo de água e eletricidade para a residência do autor se apresentam de forma irregular ao longo dos meses, mesmo sendo o perfil familiar regular, hábitos de consumo aparentemente regulares, aparelhos elétricos regulares e instalações elétrica e hidráulica inferior a sete anos. Este estudo visa analisar o consumo e verificar se existe alguma relação entre o consumo e as variações de temperatura e umidade relativa que justifiquem as irregularidades apresentadas nos relatórios de consumo das companhias de água e energia.

## Informações de referência

A residência está localizada na cidade de Juiz de Fora, estado de Minas Gerais (BR). A cidade possui 1.435,7 km$^2$ de extensão e população de 568.873 habitantes, segundo Censo 2010 realizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE), veja seção Referências item 1. Localizada sob as coordenadas -21.7642 latitude e -43.3496 longitude, está a 698 metros de altitude e a cerca de 180 km do mar, com clima tropical de altitude.

Como na maior parte das residências do país, não há sistema de calefação ou ar-condicionado no ambiente, não há sistema de aquecimento de água, os chuveiros usam energia elétrica e não há água aquecida nas torneiras, não há sistema de geração próprio de energia e também não há sistema de coleta de água de chuva ou poço.

## *Introduction*

*This document contains a autonomy study which includes analysis of water and electricity consumption in the author residence, and the relation between the consumption and the variation of local temperature and relative humidity.*

## *Problem description*

*Water and electricity consumption are irregular through the months for the author residence, even being stable the family profile, apparently stable the consumption habits, regular electric appliances and electric and hydraulic installation below seven years. This study is meant to analyse the water and electricity consumption and verify if there is any relation the consumption and the temperature and relative humidity variation that would justify the irregularities presented by the water and electricity companies reports.*

## *Reference information*

*The residence is located in Juiz de Fora city, state of Minas Gerais (BR). The city has an area of 1,453.7 km$^2$ e its population is 568.873 inhabitants, according to the Census 2010 by* Instituto Brasileiro de Geografia e Estatística *(IBGE), the Brazilian institute for geography and statistics, see References section item 1. Its geographical coordinates are -21.7642 latitude and -43.3496 longitude, 698 meters of altitude and approximately 180 km distant from the sea, presenting high-altitude tropical climate.*

*Like most of the residencies in the country, there is no heating or air conditioning systems, no water heating systems as well, shower make use of electricity and there is no warm tap water, none kind of own power generator is installed, no rainwater is collected and there is no well.*

## Ferramentas utilizadas

Para este estudo foi utilizado um computador MacBook Pro mid 2010, processador 2,66 GHz Intel Core 2 Duo com 4 GB de memória RAM, rodando com o sistema operacional macOS High Sierra 10.13. Os códigos foram desenvolvidos em ambiente Jupyter Lab em Notebooks com núcleo Python 3.6, documentos Markdown também foram elaborados em ambiente Jupyter Lab. Este relatório foi elaborado no Apple Pages 8.1 e exportado em formato PDF.

## *Tools used*

*For this study a equipment MacBook Pro mid 2010 has been used, 2,66 GHz Intel Core 2 Duo processor with 4 GB RAM memory, running with macOS High Sierra 10.13. Codes have been developed in Jupiter Lab environment within Notebooks with Python 3.6 kernel, Markdown documents have also been created in Jupiter Lab environment. This report has been created in Apple Pages 8.1 and exported to PDF format.*

## Aquisição e preparação de dados

Os dados de consumo de água e eletricidade foram foram coletados pelo autor a partir dos relatórios de consumo apresentados junto com as faturas mensais das companhias. Dados de variação de temperatura e umidade relativa foram selecionados e extraídos do banco de dados público do Instituto Nacional de Meteorologia (INMET), veja seção Referências item 2. Seguem detalhes para cada conjunto de dados utilizados.

### Dados sobre consumo de eletricidade

Os dados coletados compreendem o período entre 2014-10 e 2020-06 e foram salvos em um arquivo no formato CSV, conforme estrutura na Fig.01, e carregados em um objeto *Pandas dataframe*, conforme Fig.02.

## *Data acquisition and preparation*

*Water and electricity consumption data have been collected by the author from the consumption reports provided with the monthly invoices by each of the respective companies. Data on temperature and relative humidity have been selected and downloaded from* Instituto Nacional de Meteorologia *(INMET), the Brazilian national institute for meteorology, see References section item 2. Follow details for each of the data sets used.*

### *Electricity consumption data*

*The collected data covers the period between 2014-10 and 2020-06 and have been stored in a CSV, according to the structure in Fig.01, and loaded into a* Pandas dataframe *object, shown in Fig.02.*

| | Column | Data type | Description |
|---|---|---|---|
| 1 | **period** | `string` in format *YYYY-MM* | Year and month of the billing reference |
| 2 | **previous_check_date** | `string` in format *DD/MM/YYYY* | Date of the previous consumption reading |
| 3 | **current_check_date** | `string` in format *DD/MM/YYYY* | Date of the current consumption reading |
| 4 | **previous_mark** | `integer` | Electricity meter indicator mark of previous consumption reading |
| 5 | **current_mark** | `integer` | Electricity meter indicator mark of current consumption reading |
| 6 | **consumption_kwh** | `integer` | Electricity consumption in kilowatt-hour, difference between *current_mark* and *previous_mark* |

Fig.01 - Eletricidade, estrutura do arquivo CSV
*Fig.01 - Electricity, CSV file structure*

| | period | previous_check_date | current_check_date | previous_mark | current_mark | consumption_kwh |
|---|---|---|---|---|---|---|
| **0** | 2020-06 | 27/05/2020 | 26/06/2020 | 16.642 | 17.002 | 360 |
| **1** | 2020-05 | 29/04/2020 | 27/05/2020 | 16.298 | 16.642 | 344 |
| **2** | 2020-04 | 27/03/2020 | 29/04/2020 | 15.996 | 16.298 | 302 |
| **3** | 2020-03 | 27/02/2020 | 27/03/2020 | 15.726 | 15.996 | 270 |
| **4** | 2020-02 | 27/01/2020 | 27/02/2020 | 15.469 | 15.726 | 257 |
| **...** | ... | ... | ... | ... | ... | ... |
| **64** | 2015-02 | 26/01/2015 | 19/02/2015 | 601.000 | 641.000 | 37 |
| **65** | 2015-01 | 26/12/2014 | 26/01/2015 | 411.000 | 601.000 | 193 |
| **66** | 2014-12 | 26/11/2014 | 26/12/2014 | 223.000 | 411.000 | 188 |
| **67** | 2014-11 | 24/10/2014 | 26/11/2014 | 73.000 | 223.000 | 150 |
| **68** | 2014-10 | 24/09/2014 | 24/10/2014 | 71.000 | 73.000 | 2 |

69 rows × 6 columns

Fig.02 - Eletricidade, *Pandas dataframe*
*Fig.02 - Electricity,* Pandas *dataframe*

Como preparação dos dados, o registro referente ao período 2014-10 foi removido, pois a residência estive desocupada durante praticamente todo o mês.
Dados nas colunas *previous_check* e *current_check* foram convertidos do padrão brasileiro de data (DD/MM/AAAA) para o formato de data *Pandas* (AAAA-MM-DD) e o tipo de dado destas colunas convertido de *object* para *datetime*.
Duas novas colunas foram criadas a partir da coluna *period* (AAAA-MM), *year* com a informação do ano (AAAA) e *month* com a informação do mês (MM).
O dia da leitura pela companhia de energia é irregular, causando faturamentos com variados número de dias, para o período coletado a variação está entre 24 e 35 dias, veja Fig.03.

*During data preparation, the observation for period 2014-10 has been removed, for that period the residence was unoccupied for almost the whole month.*
*Data on columns* previous_check *and* current_check *have been converted from Brazilian date standard (DD/MM/YYYY) to* Pandas *date format (YYYY-MM-DD) and columns data type have have been converted from* object *to* datetime.
*Two new columns have been created based on column* period *(YYYY-MM),* year *(YYYY) and* month*(MM).*
*The reading date by the electricity company is irregular, due to this, the invoicing has different quantity of days, for the collected period the variation is between 24 and 35 days, see Fig.03*

```
# check differencies in days between checks
(df_electr['current_check'] - df_electr['previous_check']).dt.days.unique()
```

```
array([33, 30, 31, 24, 35, 28, 29, 32])
```

Fig.03 - Eletricidade, variação no período de leitura
*Fig.03 - Electricity, variation on reading period*

Para estimar uma medida de consumo balanceada, três novas colunas foram criadas.
- ***consumption_days***: tempo em dias entre leituras $(current\_check - previous\_check)$
- ***consumption_day_kwh***: consumo calculado por dia $\left(\frac{consumption\_kwh}{consumption\_days}\right)$
- ***consumption_30d_kwh***: consumo estimado em 30 dias $(consumption\_day\_kwh \times 30)$

*To estimate a balanced consumption measure, three new columns have been created.*
- **consumption_days***: time in days between readings* $(current\_check - previous\_check)$
- **consumption_day_kwh***: calculated consumption per day* $\left(\frac{consumption\_kwh}{consumption\_days}\right)$
- **consumption_30d_kwh***: estimated consumption in 30 days* $(consumption\_day\_kwh \times 30)$

Não havia havia problema de ausência de dados, finalmente o *dataframe* foi reduzido somente às colunas relevantes para análise dos dados, veja Fig.04.
- ***period***: ano e mês
- ***year***: ano extraído de *period*
- ***month*** - mês extraído de *period*
- ***consumption_kwh***: consumo de eletricidade em kWh
- ***consumption_days***: tempo em dias da medição
- ***consumption_day_kwh***: consumo calculado por dia em kWh
- ***consumption_30d_kwh***: consumo estimado para 30 dias em kWh

*The was no missing data issue, finally the* dataframe *has been reduced to the columns relevant to data analysis, see Fig.04.*
- **period***: year and month*
- **year***: year extracted from* period
- **month***: month extracted from* period
- **consumption_kwh***: electricity consumption in kWh*
- **consumption_days***: measured time in days*
- **consumption_day_kwh***: calculated consumption per day in kWh*
- **consumption_30d_kwh***: estimated consumption in 30 days in kWh*

| | period | year | month | consumption_kwh | consumption_days | consumption_day_kwh | consumption_30d_kwh |
|---|---|---|---|---|---|---|---|
| **0** | 2014-11 | 2014 | 11 | 150 | 33 | 4.545 | 136.35 |
| **1** | 2014-12 | 2014 | 12 | 188 | 30 | 6.267 | 188.01 |
| **2** | 2015-01 | 2015 | 01 | 193 | 31 | 6.226 | 186.78 |
| **3** | 2015-02 | 2015 | 02 | 37 | 24 | 1.542 | 46.26 |
| **4** | 2015-03 | 2015 | 03 | 306 | 35 | 8.743 | 262.29 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **63** | 2020-02 | 2020 | 02 | 257 | 31 | 8.290 | 248.70 |
| **64** | 2020-03 | 2020 | 03 | 270 | 29 | 9.310 | 279.30 |
| **65** | 2020-04 | 2020 | 04 | 302 | 33 | 9.152 | 274.56 |
| **66** | 2020-05 | 2020 | 05 | 344 | 28 | 12.286 | 368.58 |
| **67** | 2020-06 | 2020 | 06 | 360 | 30 | 12.000 | 360.00 |

68 rows × 7 columns

Fig.04 - Eletricidade, após preparação de dados
*Fig.04 - Electricity, after data preparation*

## Dados sobre consumo de água

Os dados coletados compreendem o período entre 2014-10 e 2020-06 e foram salvos em um arquivo no formato CSV, conforme estrutura na Fig.05, e carregados em um objeto *Pandas dataframe*, conforme Fig.06.

## *Water consumption data*

*The collected data covers the period between 2014-10 and 2020-06 and have been stored in a CSV, according to the structure in Fig.05, and loaded into a* Pandas *dataframe object, shown in Fig.06.*

| | Column | Data type | Description |
|---|---|---|---|
| 1 | **period** | `string` in format *YYYY-MM* | Year and month of the billing reference |
| 2 | **previous_check_date** | `string` in format *DD/MM/YYYY* | Date of the previous consumption reading |
| 3 | **current_check_date** | `string` in format *DD/MM/YYYY* | Date of the current consumption reading |
| 4 | **previous_mark** | `integer` | Hydrometer indicator mark of previous consumption reading |
| 5 | **current_mark** | `integer` | Hydrometer indicator mark of current consumption reading |
| 6 | **consumption_kwh** | `integer` | Water consumption in cubic meter ($m^3$), difference between *current_mark* and *previous_mark* |

*Fig.05 - Água, estrutura do arquivo CSV*
*Fig.05 - Water, CSV file structure*

| | period | previous_check_date | current_check_date | previous_mark | current_mark | consumption_m3 |
|---|---|---|---|---|---|---|
| **0** | 2020-06 | 29/05/2020 | 30/06/2020 | 258 | 270 | 12 |
| **1** | 2020-05 | 30/04/2020 | 29/05/2020 | 247 | 258 | 11 |
| **2** | 2020-04 | 30/03/2020 | 30/04/2020 | 235 | 247 | 12 |
| **3** | 2020-03 | 28/02/2020 | 30/03/2020 | 221 | 235 | 14 |
| **4** | 2020-02 | 30/01/2020 | 28/02/2020 | 215 | 221 | 6 |
| **...** | ... | ... | ... | ... | ... | ... |
| **64** | 2015-02 | 31/01/2015 | 04/03/2015 | 69 | 81 | 12 |
| **65** | 2015-01 | 30/12/2014 | 31/01/2015 | 51 | 69 | 18 |
| **66** | 2014-12 | 02/12/2014 | 30/12/2014 | 35 | 51 | 16 |
| **67** | 2014-11 | 30/10/2014 | 02/12/2014 | 21 | 35 | 14 |
| **68** | 2014-10 | 30/09/2014 | 30/10/2014 | 20 | 21 | 1 |

69 rows × 6 columns

Fig.06 - Água, *Pandas dataframe*
*Fig.06 - Water,* Pandas *dataframe*

Como preparação dos dados, o registro referente ao período 2014-10 foi removido, pois a residência estive desocupada durante praticamente todo o mês.

*During data preparation, the observation for period 2014-10 has been removed, for that period the residence was unoccupied for almost the whole month.*

Dados nas colunas *previous_check* e *current_check* foram convertidos do padrão brasileiro de data (DD/MM/AAAA) para o formato de data *Pandas* (AAAA-MM-DD) e o tipo de dado destas colunas convertido de *object* para *datetime*.
Duas novas colunas foram criadas a partir da coluna *period* (AAAA-MM), *year* com a informação do ano (AAAA) e *month* com a informação do mês (MM).
O dia da leitura pela companhia de água é irregular, causando faturamentos com variados número de dias, para o período coletado a variação está entre 28 e 34 dias, veja Fig.07.

*Data on columns* previous_check *and* current_check *have been converted from Brazilian date standard (DD/MM/YYYY) to* Pandas *date format (YYYY-MM-DD) and columns data type have have been converted from* object *to* datetime.
*Two new columns have been created based on column* period *(YYYY-MM),* year *(YYYY) and* month*(MM).*
*The reading date by the water company is irregular, due to this, the invoicing has different quantity of days, for the collected period the variation is between 28 and 34 days, see Fig.07*

```
# check differencies in days between checks
(df_water['current_check'] - df_water['previous_check']).dt.days.unique()
```

```
array([33, 28, 32, 29, 30, 31, 34])
```

Fig.07 - Água, variação no período de leitura
*Fig.07 - Water, variation on reading period*

Para estimar uma medida de consumo balanceada, três novas colunas foram criadas.
- ***consumption_days***: tempo em dias entre leituras $(current\_check - previous\_check)$
- ***consumption_day_m3***: consumo calculado por dia $\left(\frac{consumption\_m3}{consumption\_days}\right)$
- ***consumption_30d_m3***: consumo estimado em 30 dias $(consumption\_day\_m3 \times 30)$

*To estimate a balanced consumption measure, three new columns have been created.*
- **consumption_days***: time in days between readings* $(current\_check - previous\_check)$
- **consumption_day_m3***: calculated consumption per day* $\left(\frac{consumption\_m3}{consumption\_days}\right)$
- **consumption_30d_m3***: estimated consumption in 30 days* $(consumption\_day\_m3 \times 30)$

Devido a redução de precisão na medição, a cada cinco anos os hidrômetros são substituídos. Verificando-se as entradas que apresentavam a coluna *current_mark* menor que *previous_mark* foi possível apurar que o hidrômetro foi substituído em 2018-02, veja Fig.08.

*Due to the measurement accuracy reduction, the hydrometers are replaced every fifth year. Checking observations having column* current_mark *lower than* previous_mark *it was possible to find out that the hydrometer was replaced in 2018-02, see Fig.08.*

```
# check when current_mark is lower than previous_mark
df_water[df_water['current_mark'] < df_water['previous_mark']]
```

| | period | previous_check | current_check | previous_mark | current_mark | consumption_m3 | year | month |
|---|---|---|---|---|---|---|---|---|
| **39** | 2018-02 | 2018-02-02 | 2018-03-07 | 434 | 3 | 10 | 2018 | 02 |

Fig.08 - Água, troca do hidrômetro
*Fig.08 - Water, hydrometer replacement*

Não havia havia problema de ausência de dados, finalmente o *dataframe* foi reduzido somente às colunas relevantes para análise dos dados, veja Fig.09.

*The was no missing data issue, finally the* dataframe *has been reduced to the columns relevant to data analysis, see Fig.09.*

- ***period***: ano e mês
- ***year***: ano extraído de *period*
- ***month***: mês extraído de *period*
- ***consumption_m3***: consumo de água em m³
- ***consumption_days***: tempo em dias da medição
- ***consumption_day_m3***: consumo calculado por dia em m³
- ***consumption_30d_m3***: consumo estimado para 30 dias em m³

- **period***: year and month*
- **year***: year extracted from* period
- **month***: month extracted from* period
- **consumption_m3***: water consumption in m³*
- **consumption_days***: measured time in days*
- **consumption_day_m3***: calculated consumption per day in m³*
- **consumption_30d_m3***: estimated consumption in 30 days in m³*

| | period | year | month | consumption_m3 | consumption_days | consumption_day_m3 | consumption_30d_m3 |
|---|---|---|---|---|---|---|---|
| **0** | 2014-11 | 2014 | 11 | 14 | 33 | 0.424 | 12.72 |
| **1** | 2014-12 | 2014 | 12 | 16 | 28 | 0.571 | 17.13 |
| **2** | 2015-01 | 2015 | 01 | 18 | 32 | 0.562 | 16.86 |
| **3** | 2015-02 | 2015 | 02 | 12 | 32 | 0.375 | 11.25 |
| **4** | 2015-03 | 2015 | 03 | 16 | 29 | 0.552 | 16.56 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **63** | 2020-02 | 2020 | 02 | 6 | 29 | 0.207 | 6.21 |
| **64** | 2020-03 | 2020 | 03 | 14 | 31 | 0.452 | 13.56 |
| **65** | 2020-04 | 2020 | 04 | 12 | 31 | 0.387 | 11.61 |
| **66** | 2020-05 | 2020 | 05 | 11 | 29 | 0.379 | 11.37 |
| **67** | 2020-06 | 2020 | 06 | 12 | 32 | 0.375 | 11.25 |

68 rows × 7 columns

Fig.09 - Água, após preparação de dados
*Fig.09 - Water, after data preparation*

## Dados meteorológicos

Os dados meteorológicos compreendem o período de 2014-01 a 2020-06 e foram extraídos do banco de dados público do INMET. Este instituto mantém estações meteorológicas convencionais e automáticas espalhadas pelo país, nas estações convencionais os dados são registrados três vezes ao dia, às 00:00, 12:00 e 18:00 horas, nas estações automáticas os dados são registrados de hora em hora. Em Juiz de Fora estão instaladas uma estação convencional e uma automática, para este estudo foram usados os dados da estação automática A518.

Os dados de cada estação estão disponíveis para consulta no portal do INMET na *internet*, ver seção Referências item 2. Na opção salvar dados, estes são agrupados por ano, em arquivos no formato ZIP contento os dados de todas as estações. Para este estudo foram salvos os arquivos dos anos entre 2014 e 2020, este último até 30 de junho, e extraídos somente os arquivos da estação A518, sete no total. Os arquivos CSV possuem uma primeira parte de identificação da estação entre as linhas 1 e 8, conforme Fig.10, e uma segunda parte com o registro dos dados a partir da linha 9, conforme Fig.11.

## *Meteorological data*

*The meteorological data covers the period between 2014-01 and 2020-06 and have been extracted from the INMET public data bank. The institute keeps conventional and automatic meteorological stations across the country, for conventional stations the data are recorded three times a day, at 00:00, 12:00 and 18:00 o'clock, for automatic stations the data are recorded hourly. In Juiz de Fora one conventional and one automatic station are installed, for this study the data from automatic station A518 have been used.*

*Data of all stations are available from the INMET web portal, see Reference section item 2. In its save data option, the data are grouped by year in a ZIP file format, containing data of all stations. For this study files from years between 2014 and 2020 have been downloaded, and only files of station A518 extracted, seven of them in total. The CSV files contain a first portion to identify the station from row 1 to 8, according to Fig.10, and a second portion with the recorded data staring from row 9, according to the Fig.11.*

| | Station identification | | | |
|---|---|---|---|---|
| **Row** | **Name** | **Data until 2018** | **Data after 2018** | **Description** |
| 1 | REGIÃO | SE | SE | Geographical region |
| 2 | UF | MG | MG | Federation state |
| **3** | **ESTAÇÃO** | **JUIZ DE FORA** | **JUIZ DE FORA** | **Station name / location** |
| **4** | **CODIGO (WMO)** | **A518** | **A518** | **Station id** |
| 5 | LATITUDE | -21,76999999 | -21,769965 | Station latitude |
| 6 | LONGITUDE | -43,36416666 | -43,364329 | Station longitude |
| 7 | ALTITUDE | 950 | 936,88 | Station altitude (m) |
| 8 | DATA DE FUNDAÇÃO | 2007-05-26 | 26/05/07 | Station foundation date |

Fig.10 - Meteorologia, identificação da estação no arquivo CSV

*Fig.10 - Meteorology, station identification in CSV file*

| | Data description | | |
|---|---|---|---|
| **Column** | **Name** | **Data type** | **Description** |
| **0** | **DATA** | `string` | **Date in format *YYYY-MM-DD* until 2018, and *YYYY/MM/DD* after 2018** |
| **1** | **HORA UTC** | `string` | **Time in format *HH:MM* until 2018, and *HHMM UTC* after 2018** |
| 2 | PRECIPITAÇÃO TOTAL HORARIA | `float` | Total rain (mm) |
| 3 | PRESSAO ATMOSFERICA HORARIA | `float` | Atmospheric pressure (mB) |
| 4 | PRESSÃO ATMOSFERICA MAX HORA ANT | `float` | Maximal atmospheric pressure during previous hour (mB) |
| 5 | PRESSÃO ATMOSFERICA MIN HORA ANT | `float` | Minimal atmospheric pressure during previous hour (mB) |
| 6 | RADIACAO GLOBAL | `float` | Global radiation ($W/m^2$) |
| **7** | **TEMPERATURA HORARIA** | `float` | **Temperature (°C)** |
| 8 | TEMPERATURA PONTO DE ORVALHO | `float` | Dew point temperature (°C) |
| **9** | **TEMPERATURA MAX HORA ANT** | `float` | **Maximal temperature during previous hour (°C)** |
| **10** | **TEMPERATURA MIN HORA ANT** | `float` | **Minimal temperature during previous hour (°C)** |
| 11 | TEMPERATURA PONTO DE ORVALHO MAX HORA ANT | `float` | Maximal dew point temperature during prevous hour(°C) |
| 12 | TEMPERATURA PONTO DE ORVALHO MIN HORA ANT | `float` | Minimal dew point temperature during prevous hour(°C) |
| **13** | **UMIDADE REL MAX HORA ANT** | `integer` | **Maximal relative humidity during previous hour (%)** |
| **14** | **UMIDADE REL MIN HORA ANT** | `integer` | **Minimal relative humidity during previous hour (%)** |
| **15** | **UMIDADE REL HORARIA** | `integer` | **Relative humidity (%)** |
| 16 | VENTO DIREÇÃO HORARIA | `integer` | Wind direction (°) |
| 17 | VENTO RAJADA MAXIMA | `float` | Maximal wind gust (m/s) |
| 18 | VENTO VELOCIDADE HORARIA | `float` | Wind speed (m/s) |

Fig.11 - Meteorologia, estrutura de dados no arquivo CSV

*Fig.11 - Meteorology, data structure in CSV file*

Sendo 24 registros por dia, os arquivos CSV para os anos de 2014 a 2019 possuíam 18 colunas e entre 8.760 e 8,784 linhas, para o ano de 2020 eram 4.344 linhas. Para carregar os dados dos arquivos CSV em um único *Pandas dataframe*, foi executado um laço que carregava os dados de um arquivo por vez em um *dataframe* temporário e somente as colunas relevantes para análise eram armazenadas do *dataframe* definitivo, conforme Fig.12.

*Being 24 records per day, the CSV files for years between 2014 and 2019 contained 18 columns and between 8,760 and 8,784 rows, for year 2020 were 4,344 rows. To load the data from CSV files into a unique Pandas dataframe, a loop has been ran, which loaded one file each time into a temporary dataframe, and only the columns relevant for data analysis stored into a final dataframe, according to Fig.12.*

- ***date***: data do registro
- ***time_utc***: horário do registro no fuso horário UTC
- ***temp*** - temperatura instantanea em ºC
- ***temp_min***: temperatura mínima na a hora anterior em ºC
- ***temp_max***: temperatura máxima na a hora anterior em ºC
- ***rel_humidity***: umidade relativa instantanea em %
- ***rel_humidity_min***: umidade relativa mínima na hora anterior em %
- ***rel_humidity_max***: umidade relativa máxima na hora anterior em %

No total, 56.952 registros foram carregados, e, após a execução do laço, o *dataframe* temporário foi removido da memória.

- **date***: recording date*
- **time_utc***: recording time UTC*
- **temp** *- instant temperature in ºC*
- **temp_min***: minimal temperature during previous hour in ºC*
- **temp_max***: maximal temperature during previous hour in ºC*
- **rel_humidity***: instant relative humidity in %*
- **rel_humidity_min***: minimal relative humidity during previous hour in %*
- **rel_humidity_max***: maximal relative humidity during previous hour in %*

In total, 56,952 records have been loaded, and after the loop ran, the temporary dataframe has been deleted from memory.

| | date | time_utc | temp | temp_min | temp_max | rel_humidity | rel_humidity_min | rel_humidity_max |
|---|---|---|---|---|---|---|---|---|
| **0** | 2014-01-01 | 00:00 | 22,9 | 22,7 | 23,1 | 78.0 | 78.0 | 81.0 |
| **1** | 2014-01-01 | 01:00 | 23 | 22,8 | 23,5 | 76.0 | 72.0 | 78.0 |
| **2** | 2014-01-01 | 02:00 | 22,6 | 22,6 | 23,1 | 80.0 | 75.0 | 80.0 |
| **3** | 2014-01-01 | 03:00 | 22,4 | 22,2 | 22,6 | 82.0 | 80.0 | 84.0 |
| **4** | 2014-01-01 | 04:00 | 22,1 | 22 | 22,7 | 84.0 | 80.0 | 85.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **56947** | 2020-06-30 | 1900 UTC | 19,4 | 19,2 | 20 | 74.0 | 72.0 | 75.0 |
| **56948** | 2020-06-30 | 2000 UTC | 19 | 19 | 19,5 | 76.0 | 73.0 | 76.0 |
| **56949** | 2020-06-30 | 2100 UTC | 18,4 | 18,3 | 19 | 78.0 | 76.0 | 78.0 |
| **56950** | 2020-06-30 | 2200 UTC | 17,7 | 17,7 | 18,4 | 80.0 | 78.0 | 80.0 |
| **56951** | 2020-06-30 | 2300 UTC | 17,1 | 17 | 17,7 | 84.0 | 80.0 | 84.0 |

56952 rows × 8 columns

Fig.12 - Meteorologia, *Pandas dataframe*
*Fig.12 - Meteorology,* Pandas *dataframe*

A coluna *date* foi carregada do arquivo CSV no formato *datetime*, nenhum ajuste foi necessário.
A coluna *time_utc* foi carregada no formato *object*. Somente as duas primeiras posições à esquerda, referente à hora são relevantes, as demais posições foram removidas e o tipo de dados da coluna convertido para *int32*.
As colunas *temp*, *temp_min* e *temp_max* foram carregadas no formato *object* e os dados no formato brasileiro de números, com a vírgula como separador de decimais. A vírgula foi substituída pelo ponto e o tipo de dados da coluna convertido para *float64*.
As colunas *rel_humidity*, *rel_humidity_min* e *rel_humidity_max* foram carregadas no formato *float64*, nenhum ajuste foi necessário.
Veja Fig.13, dataframe após os ajustes nos dados.

*Column* date *has been loaded from CSV file in* datetime *format, no adjustment required.*
*Column* time_utc *has been loaded in* object *format. Only its first two left positions, which refer to the hour, are relevant, the remaining positions have been removed and the column data type converted to* int32*.*
*Columns* temp*,* temp_min *and* temp_max *have been loaded in* object *format and the data in Brazilian numbers standard, comma as decimal separator. The comma has been replaced by decimal separator (.) and the column data type converted to* float64*.*
*Columns* rel_humidity*,* rel_humidity_min*,* rel_humidity_max *have been loaded in* flot64 *format, no adjustments required.*
*See Fig.13, dataframe after data adjustments.*

| | date | time_utc | temp | temp_min | temp_max | rel_humidity | rel_humidity_min | rel_humidity_max |
|---|---|---|---|---|---|---|---|---|
| **0** | 2014-01-01 | 0 | 22.9 | 22.7 | 23.1 | 78.0 | 78.0 | 81.0 |
| **1** | 2014-01-01 | 1 | 23.0 | 22.8 | 23.5 | 76.0 | 72.0 | 78.0 |
| **2** | 2014-01-01 | 2 | 22.6 | 22.6 | 23.1 | 80.0 | 75.0 | 80.0 |
| **3** | 2014-01-01 | 3 | 22.4 | 22.2 | 22.6 | 82.0 | 80.0 | 84.0 |
| **4** | 2014-01-01 | 4 | 22.1 | 22.0 | 22.7 | 84.0 | 80.0 | 85.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **56947** | 2020-06-30 | 19 | 19.4 | 19.2 | 20.0 | 74.0 | 72.0 | 75.0 |
| **56948** | 2020-06-30 | 20 | 19.0 | 19.0 | 19.5 | 76.0 | 73.0 | 76.0 |
| **56949** | 2020-06-30 | 21 | 18.4 | 18.3 | 19.0 | 78.0 | 76.0 | 78.0 |
| **56950** | 2020-06-30 | 22 | 17.7 | 17.7 | 18.4 | 80.0 | 78.0 | 80.0 |
| **56951** | 2020-06-30 | 23 | 17.1 | 17.0 | 17.7 | 84.0 | 80.0 | 84.0 |

56952 rows × 8 columns

Fig.13 - Meteorologia, *dataframe* ajuste no tipo de dados
*Fig.13 - Meteorology, dateframe data type adjustments*

Foram encontrados registros com dados ausentes nas colunas *temp*, *temp_min*, *temp_max*, *rel_humidity*, *rel_humidity_min* e *rel_humidity_max*. Haviam entradas com o valor *-9999* e sem valor, identificadas como *NaN*, veja exemplo da coluna *temp* na Fig.14. O volume de dados ausentes representava entre 1,31% e 1,43% dos dados cada coluna, uma taxa muito baixa que não implicou risco à integridade dos dados, veja resumo na Fig.15.

*Missing data has been found in columns* temp, temp_min, temp_max, rel_humidity, rel_humidity_min *and* rel_humidity_max. *They have been identified with the value* -9999 *and* NaN, *see example from column* temp *in Fig.14. The missing data volume represented between 1.31% and 1.43% of the data in each column, a very low rate which was not an issue to the data integrity, see summary in Fig.15.*

```
# check for missing values
df_weather[(df_weather['temp']==-9999) | df_weather['temp'].isnull()]
```

| | date | time_utc | temp | temp_min | temp_max | rel_humidity | rel_humidity_min | rel_humidity_max |
|---|---|---|---|---|---|---|---|---|
| **9281** | 2015-01-22 | 17 | -9999.0 | -9999.0 | -9999.0 | -9999.0 | -9999.0 | -9999.0 |
| **42923** | 2018-11-24 | 11 | -9999.0 | -9999.0 | -9999.0 | -9999.0 | -9999.0 | -9999.0 |
| **42924** | 2018-11-24 | 12 | -9999.0 | -9999.0 | -9999.0 | -9999.0 | -9999.0 | -9999.0 |
| **42925** | 2018-11-24 | 13 | -9999.0 | -9999.0 | -9999.0 | -9999.0 | -9999.0 | -9999.0 |
| **42926** | 2018-11-24 | 14 | -9999.0 | -9999.0 | -9999.0 | -9999.0 | -9999.0 | -9999.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **46300** | 2019-04-14 | 4 | NaN | NaN | NaN | NaN | NaN | NaN |
| **46304** | 2019-04-14 | 8 | NaN | NaN | NaN | NaN | NaN | NaN |
| **46306** | 2019-04-14 | 10 | NaN | NaN | NaN | NaN | NaN | NaN |
| **46322** | 2019-04-15 | 2 | NaN | NaN | NaN | NaN | NaN | NaN |
| **54259** | 2020-03-10 | 19 | NaN | NaN | NaN | NaN | NaN | NaN |

748 rows × 8 columns

Fig.14 - Meteorologia, dados ausentes na coluna *temp*
*Fig.14 - Meteorologia, missing data in column* temp

```
# print the number of missing values for each column (2 to -1)
for i in range(2, df_weather.shape[1]):
    vnum = df_weather[(df_weather.iloc[:,i]==-9999) | (df_weather.iloc[:,i].isnull())].shape[0]
    print('{} missing values in column {} ({:.2%})'.format(vnum, df_weather.columns[i], vnum/df_weather.shape[0]))
```

```
748 missing values in column temp (1.31%)
770 missing values in column temp_min (1.35%)
770 missing values in column temp_max (1.35%)
794 missing values in column rel_humidity (1.39%)
817 missing values in column rel_humidity_min (1.43%)
817 missing values in column rel_humidity_max (1.43%)
```

Fig.15 - Meteorologia, resumo dados ausentes
*Fig.15 - Meteorology, missing data summary*

Os dados ausentes foram substituídos pela média calculada a partir do mesmo dia e horário nos demais anos, para todas as colunas onde foram encontrados dados ausentes. Como no exemplo da Fig.14, a coluna *temp* na data 2015-01-22 horário 17 foi substituída pela temperatura média calculada do dia 22 de janeiro às 17:00 horas dos anos 2104 e de 2016 a 2020.

## Análise dos dados

Após a preparação dos dados foi realizada a análise de cada tópico separadamente, consumo de eletricidade, consumo de água, variação de temperatura e variação de umidade relativa.

### Análise do consumo de eletricidade

A primeira análise foi baseada nas estatísticas do *dataframe*. Identificou-se uma grande diferença entre os valores mínimo e máximo, na coluna *consumption_kwh* estes valores são *37,0* e *360,0 kWh* respectivamente. O mesmo comportamento foi observado na coluna *consumption_day_kwh*, *1,542* e *12,286* respectivamente, o que mostra alta irregularidade no consumo. O intervalo interquartil também é bastante grande, na coluna *consumption_kwh* vai de *217,0* a *285,5*, o mesmo comportamento na coluna *consumption_day_kwh*, de *7,246* a *9,156*. A irregularidade é confirmada pelo também elevado desvio padrão, *57,059* para *consumption_kwh* e *1,803* para *consumption_day_kwh*.
A medição irregular pela companhia de energia pode criar desbalanceamento nas faturas, o que pode ser balanceado pelo cálculo do consumo diário, mas este também mostra alta irregularidade no consumo no período avaliado. Veja Fig.16.

*The missing data have been replaced by the mean calculated from the same day and time for the remaining years, for all the columns which missing data have been found. As in Fig.14 example, the column* temp *for date 2015-01-22 time 17 has been replaced by the mean temperature calculated for January 22 at 17:00 o'clock from years 2014 and from 2016 until 2020.*

## *Data analysis*

*After data preparation, a data analysis has been conducted for each topic separately, electricity consumption, water consumption, temperature variation and relative humidity variation.*

### *Electricity consumption analysis*

*The first analysis has been done on dataframe statistics. It has been identified a large range between minimum and maximum values, for column* consumption_kwh *they are* 37.0 *and* 360.0 kWh *respectively. The same behaviour has been observed in column* consumption_day_kwh, 1.542 *and* 12.286 *respectively, what shows high level of irregularity in consumption. The interquartile interval is also large, for column* consumption_kwh *it goes from* 217.0 *to* 285.5*, the same behaviour for column* consumption_day_kwh*, from* 7.246 *to* 9.156*. The irregularity is confirmed by the also high standard deviation,* 57.059 *for* consumption_kwh *and* 1.803 *for* consumption_day_kwh*.*
*Irregular readings by the electricity company could create some unbalanced invoices, what could be balanced by the calculated consumption per day, but it also shows high irregularity in consumption for the evaluated period. See Fig.16.*

| | consumption_kwh | consumption_days | consumption_day_kwh | consumption_30d_kwh |
|---|---|---|---|---|
| count | 68.000000 | 68.000000 | 68.000000 | 68.000000 |
| mean | 248.955882 | 30.470588 | 8.149162 | 244.474853 |
| std | 57.058868 | 1.888461 | 1.802550 | 54.076489 |
| min | 37.000000 | 24.000000 | 1.542000 | 46.260000 |
| 25% | 217.000000 | 29.000000 | 7.245750 | 217.372500 |
| 50% | 247.500000 | 30.000000 | 8.233000 | 246.990000 |
| 75% | 285.500000 | 32.000000 | 9.155750 | 274.672500 |
| max | 360.000000 | 35.000000 | 12.286000 | 368.580000 |

Fig.16 - Eletricidade, estatísticas do *dataframe*
*Fig.16 - Electricity, dataframe statistics*

Histogramas para comparação das frequências nas colunas *consumption_kwh* e *consumption_30d_kwh* mostraram maior concentração de consumo na faixa entre *200* e *300 kWh*, e um provável mês fora do perfil de consumo abaixo de *100 kWh*. Comparando-se os dois histogramas também é visível que o consumo estimado para 30 dias tem maior concentração em torno de *250 kWh*, veja Fig.17.

*Histograms comparing the frequencies in columns* consumption_khw *and* consumption_30d_kwh *showed higher consumption concentration in range between* 200 *and* 300 kWh*, and one probably outlier month with consumption below* 100 kWh*. Comparing the two histograms, it is visible that the estimated 30 days consumption has higher concentration around* 250 kWh*, see Fig.17.*

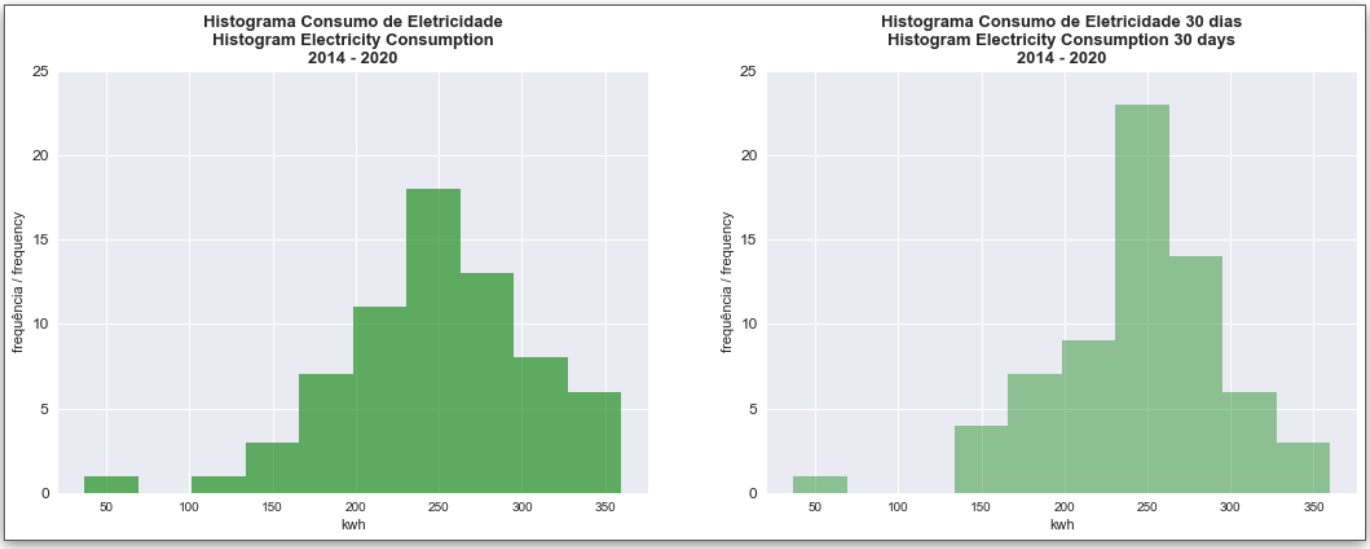


Fig.17 - Eletricidade, histogramas

*Fig.17 - Electricity, histograms*

O gráfico de linhas na Fig.18 mostra o aumento no consumo ao longo do período avaliado, com regular redução entre os meses de janeiro e fevereiro, e aumento entre os meses de maio e agosto.

A média móvel de 12 meses mostra crescente aumento de consumo entre 2014 e meados de 2017, e então certa estabilidade até final de 2019, mas um pico no consumo em 2020 inicia nova subida na média.

O gráfico também mostra a curva do consumo estimado para 30 dias, que tende ser mais atenuada, em relação à curva do consumo.

*The line plot in Fig.18 shows consumption increase along the evaluated period, with a regular reduction between January and February, and increase between May and August.*

*The 12 months mean shows continuous increasing in consumption between 2014 and middle of 2017, and then some stability until end of 2019, but a consumption peak in 2020 starts a new mean increasing.*

*The plot also shows the estimated 30 days consumption curve, which tends to be smoother than the consumption curve.*

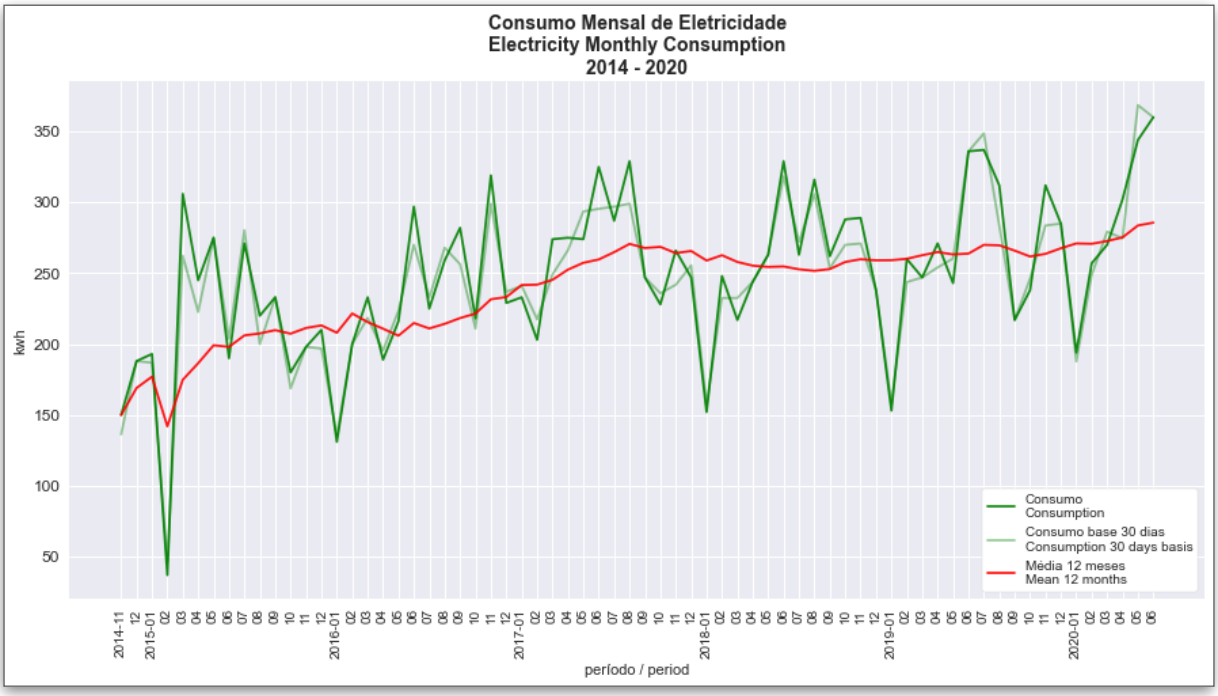


Fig.18 - Eletricidade, gráfico consumo mensal

*Fig.18 - Electricity, monthly consumption plot*

O gráfico de barras na Fig.19 mostra a média calculada por mês para o consumo e o consumo estimado 30 dias para o período avaliado, é visível o aumento de consumo nos meses de maio a agosto, sendo a média de julho bem superior à média geral do ano. Inversamente uma redução acentuada nos meses de janeiro e fevereiro, bem inferior à média geral. É possível também notar que a média para o consumo estimado 30 dias é inferior em quase todos os meses, exceto para os meses de janeiro, maio e julho quando ficou superior, e dezembro quando está no mesmo nível.

*The bar plot in Fig.19 shows the mean calculated by month for consumption and estimated consumption 30 days for the evaluated period, it is visible the consumption increase between May and August, being the mean for July greatly higher than the yearly overall mean. On the other hand, there is a acute reduction in January and February, greatly lower than overall mean. Is is also noticeable that the mean for the estimated 30 days is lower than consumption in almost all of the months, except from January, May and July, when it was higher, and December, when is was on the same level.*

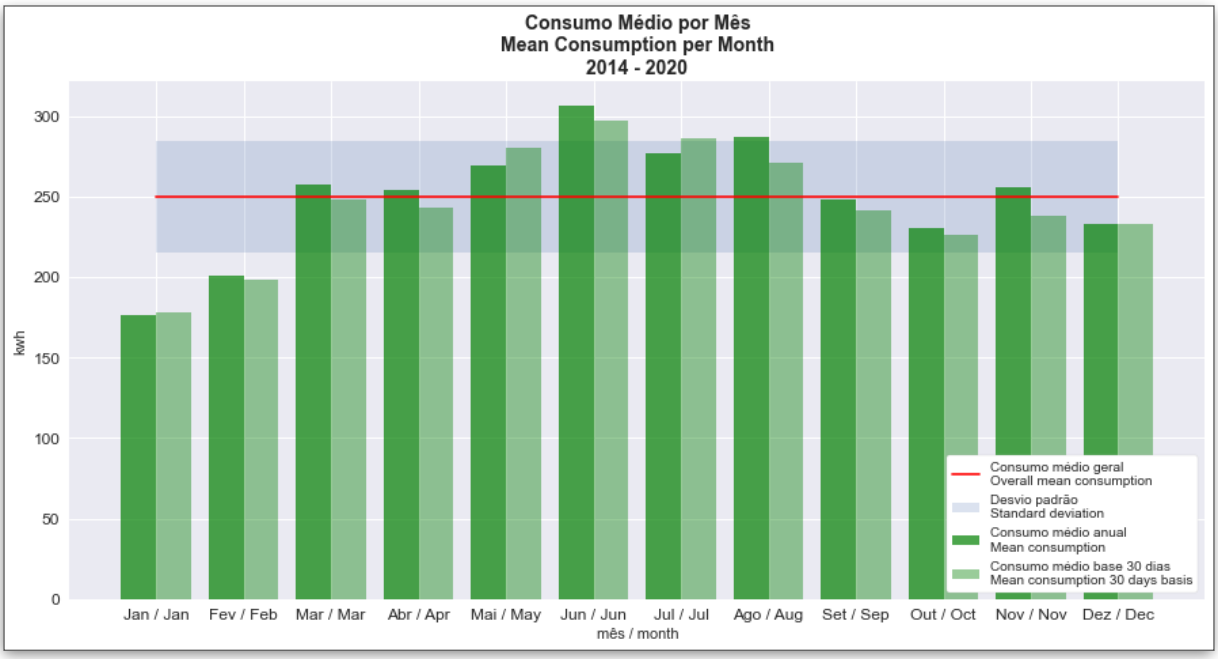


Fig.19: Eletricidade, gráfico média consumo por mês

*Fig.19 - Electricity, mean consumption by month*

A análise isolada do consumo de eletricidade mostra indícios de padrões sazonais, com mais consumo no período que coincide com o período de inverno e menos consumo com o período que coincide com o verão, e também com o período mais longo de férias.

*The electricity consumption analysis itself shows some signs of seasonal patterns, with higher consumption during the period which corresponds to the Winter and lower lower consumption during the period which corresponds to the Summer, and also the long vacation time.*

### Análise do consumo de água

Na análise das estatísticas do *dataframe*, identificou-se também uma grande diferença entre os valores mínimo e máximo, na coluna *consumption_m3* estes valores são 4,0 e 18,0 m³ respectivamente. O mesmo comportamento foi observado na coluna *consumption_day_m3*, 0,125 e 0,571 respectivamente. Ao contrário do consumo de eletricidade, o intervalo interquartil é curto, na coluna *consumption_m3* vai de 9,0 a 11,0, e na coluna *consumption_day_m3* de 0,298 a 0,375. Neste caso o desvio padrão é baixo, 2,582 para *consumption_m3* e 0,082 para *consumption_day_m3*. Aparentemente há certa regularidade no consumo de água, a grande diferença entre os valores mínimo e máximo podem ser meses de consumo fora do padrão.

### *Water consumption analysis*

*Within dataframe statistics analysis it also noticed a huge difference between minimum and maximum values, for column consumption_m3 they are 4.0 and 18.0 m³ respectively. The same behaviour has been observed for column consumption_day_m3, 0.125 and 0.571 respectively. Contrary to the electricity consumption, the interquartile intervale is short, for column consumption_m3 it goes from 9.0 to 11.0, and for column consumption_day_m3 from 0.298 to 0.375. In this case the standard deviation is low, 2,582 for consumption_m3 and 0,082 for consumption_day_m3. Apparently, there is some regularity on water consumption, the huge difference between minimum and maximum values may are outliers.*

| | consumption_m3 | consumption_days | consumption_day_m3 | consumption_30d_m3 |
|---|---|---|---|---|
| count | 68.000000 | 68.000000 | 68.000000 | 68.000000 |
| mean | 10.147059 | 30.441176 | 0.333029 | 9.990882 |
| std | 2.581592 | 1.605795 | 0.082238 | 2.467126 |
| min | 4.000000 | 28.000000 | 0.125000 | 3.750000 |
| 25% | 9.000000 | 29.000000 | 0.297500 | 8.925000 |
| 50% | 10.000000 | 30.000000 | 0.323000 | 9.690000 |
| 75% | 11.000000 | 32.000000 | 0.375000 | 11.250000 |
| max | 18.000000 | 34.000000 | 0.571000 | 17.130000 |

Fig.20 - Água, estatísticas do *dataframe*
*Fig.20 - Water, dataframe statistics*

A medição irregular pela companhia de água também pode criar desbalanceamento nas faturas, o que pode ser balanceado pelo cálculo do consumo diário. Mas neste caso é particularmente importante notar que a irregularidade nas leituras pode criar impactos nos valores cobrados, uma vez que a tarifa de água é segmentada em faixas de consumo, para cobrar mais de consumidores que gastam mais. Faixas superiores de consumo têm tarifas mais caras, veja Fig.21 as faixas de consumo.

*Irregular readings by the water company could also create some unbalanced invoices, what could be balanced by the calculated consumption per day. But in these case it is particularly important to notice that the irregular readings could impact charges, due to the fact that the water consumption tariff is segmented in consumption ranges, allowing to charge more consumers who expend more water. Higher consumption ranges have higher tariffs, see Fig.21 the consumption ranges.*

| TARIFF | RANGE ($m^3$) |
|---|---|
| T1 | >0 and <=5 |
| T2 | >5 and <=10 |
| T3 | >10 and <=15 |
| T4 | >15 and <=20 |
| T5 | >20 and <=40 |
| T6 | >40 |

Fig.21 - Água, faixas de consumo
*Fig.21 - Water, consumption ranges*

Histograma para comparação das frequências na coluna *consumption_m3* mostrou maior concentração na faixa entre *7* e *12 m³*, e prováveis meses fora do perfil de consumo abaixo de *7* e acima de *15 m³*. Para a coluna *consumption_30d_m3,* o histograma mostrou maior concentração na faixa entre *8* e *12 m³*, e prováveis meses fora do perfil de consumo abaixo de *5* e acima de *15 m³*. A diferença de frequências entre as colunas *consumption_m3* e *consumption_30d_m3* é evidente, mas não há evidencias de que poderia haver impactos nas tarifas cobradas, veja Fig.22.

*A histogram comparing the frequencies for column consumption_m3 showed higher concentration in the range between 7 and* 12 m³*, and probably some outliers below 7 and above* 15 m³*. For column consumption_30d_m3, the histogram showed higher concentration in the range between* 8 *and* 12 m³*, and probably some outliers below* 5 *and* 15 m3*. It is clear the difference between frequencies for columns consumption_m3 and consumption_30d_m3, but there is no evidence that would have impact in the tariff charged, see Fig.22.*

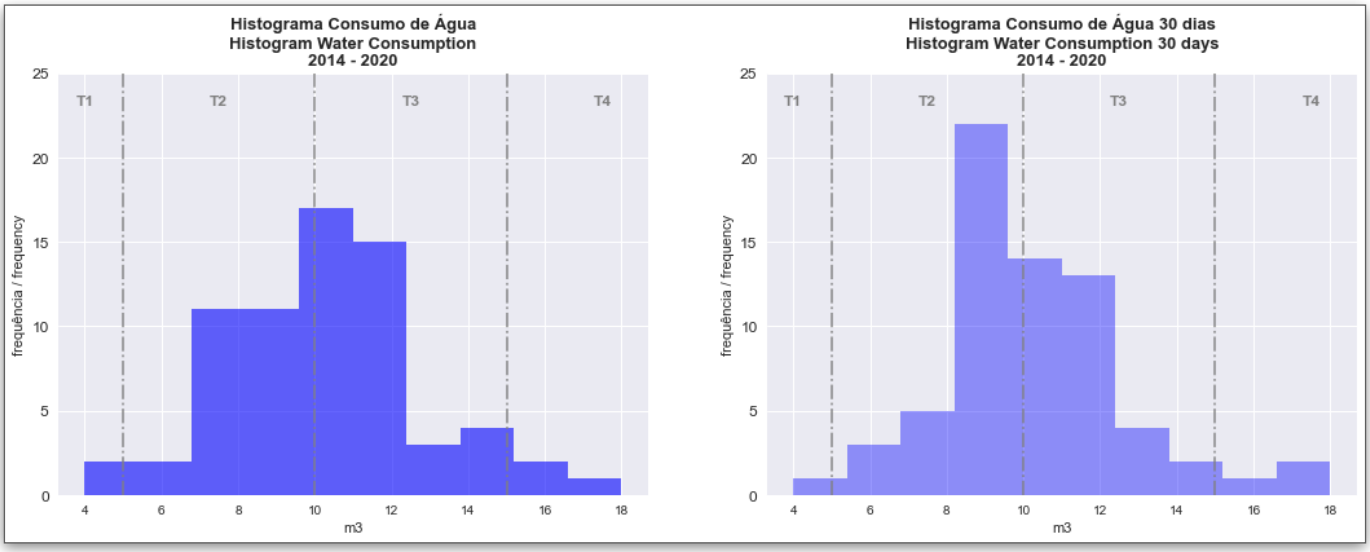


Fig.22 - Água, histogramas

*Fig.22 - Water, histograms*

O gráfico de linhas na Fig.23 mostra contínua redução no consumo entre 2014 e o primeiro trimestre de 2018, então um leve aumento até meados de 2019 e novamente redução. Aparecem diversos meses com leituras fora do padrão com valores abaixo de *7 m³* e picos acima de *12 m³*, uma variação anormal é observada a partir de meados de 2019. A troca do higrômetro em fevereiro de 2018 não parece ter impactado na leitura do consumo.

O gráfico também mostra a curva do consumo estimado para 30 dias, que tende ser mais atenuada, em relação à curva do consumo. Em comparação ao consumo medido pela companhia de água, o consumo estimado 30 dias ficou situado em diferentes faixas de consumo, e consequente tarifa, em quatro pontos, dois acima, 2017-02 e 2018-03, e dois abaixo, 2017-09 e 2017-10.

*The line plot in Fig.23 shows continuous consumption reduction between 2014 and middle of 2018, then there is a light increase unit middle 2019 and again reduction. There some outliers with values below 7 m³ and above 12 m³, there is also abnormal variation observed from middle 2019 on. The hydrometer replacement in February 2018 does not seem impacted in consumption reading.*

*The plot also shows the estimated 30 days consumption curve, which tends to be smoother than the consumption curve. In comparison to the consumption by the water company, the estimated 30 days consumption lays in different consumption ranges, and consequently different tariff, in four points, two above, 2017-02 and 2018-03, and two below, 2017-09 and 2019-10.*

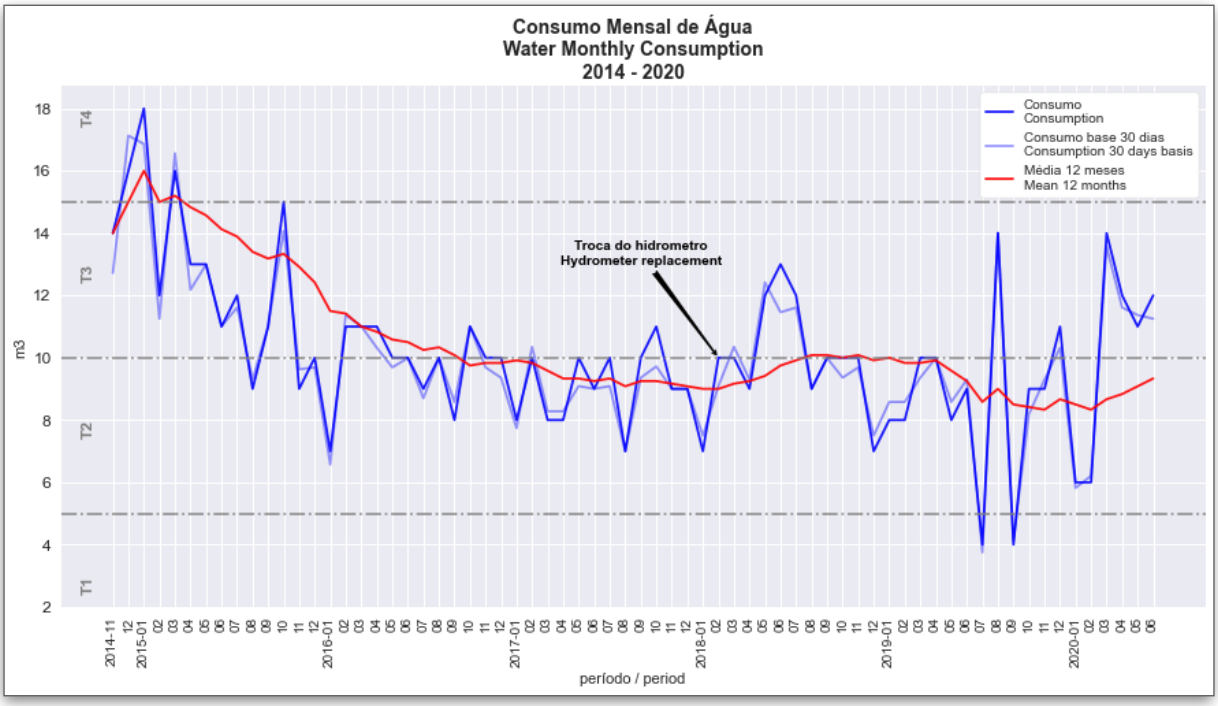


Fig.23 - Água, gráfico consumo mensal

*Fig.23 - Water, monthly consumption plot*

Na Fig.24, o gráfico de barras mostra a média calculada por mês para o consumo e o consumo estimado 30 dias para o período avaliado, é visível o aumento de consumo nos meses de março e outubro, bem superior à média geral do ano. Também é visível redução nos meses de janeiro, fevereiro, julho e setembro bem inferior à média geral. A média para o consumo estimado 30 dias é nitidamente inferior nos meses de junho e outubro.
À excessão da redução no mês de janeiro, talvez associada ao período mais longo de férias, não há aparentemente padrão sazonal para o consumo de água.

*In Fig.24, the bar plot shows the mean calculated by month for consumption and estimated consumption 30 days for the evaluated period, it is visible the consumption increase in March and October, their means are greatly higher than the yearly overall mean. It is also visible reduction in January, February, July and September, their means are greatly lower than overall mean. The mean for the estimated 30 days is visible lower than consumption in June and October. Except from reduction in January, maybe associated to the long vacation period, apparently there is no sazonal pattern in water consumption.*

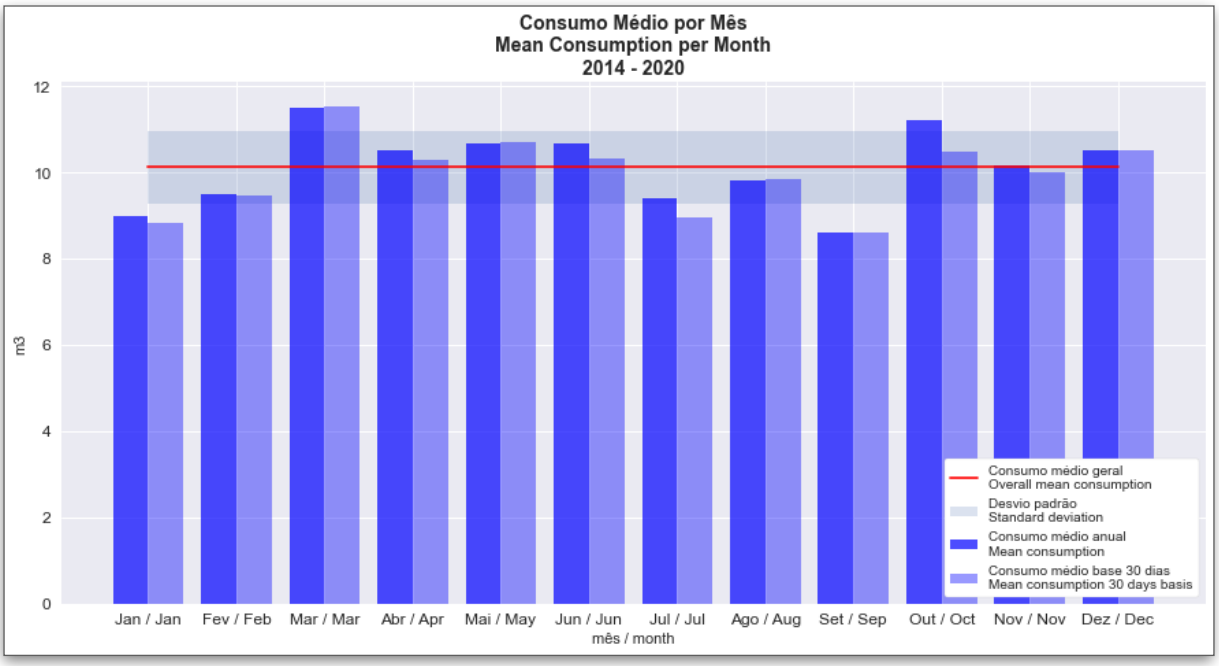


Fig.24: Água, gráfico média consumo por mês
*Fig.24 - Water, mean consumption by month*

## Análise da variação da temperatura

As estatísticas do *dataframe* mostraram uma grande variação entre as temperaturas mínima e máxima, o menor valor na coluna *temp_min* é *7,20ºC* e o maior valor na coluna *temp_max* é *35,00ºC*. Por outro lado, a diferença entre a média e a mediana da coluna *temp* é mínima, *19,48ºC* média e *19,20ºC mediana*, com desvio padrão *3,93*, e intervalo interquartil entre *16,70ºC* e *22,00ºC*, o que pode indicar extremos de temperatura mínima e máxima fora dos padrões locais, veja Fig.25.

## *Temperature variation analysis*

*Dataframe statistics showed huge variation between minimal and maximal temperatures, the lowest value in column* tem_min *is* 7.20ºC *and the maximal value in column* temp_max *is* 35.00ºC*. On the other hand, the difference between mean and median of column* temp *is minimal,* 19.48ºC *mean and* 19.20ºC *median, standard deviation* 3,93 *and interquartile interval between* 16.70ºC *and* 22.00ºC*, what could indicate that extreme low and high temperatures are outlier, for the local patterns, see Fig.25.*

| | time_utc | temp | temp_min | temp_max | rel_humidity | rel_humidity_min | rel_humidity_max |
|---|---|---|---|---|---|---|---|
| **count** | 56952.000000 | 56952.000000 | 56952.000000 | 56952.000000 | 56952.000000 | 56952.000000 | 56952.000000 |
| **mean** | 11.500000 | 19.482771 | 19.001689 | 20.017200 | 79.192069 | 76.765764 | 81.431679 |
| **std** | 6.922247 | 3.934341 | 3.747882 | 4.131531 | 15.725555 | 16.628299 | 14.736559 |
| **min** | 0.000000 | 7.300000 | 7.200000 | 7.400000 | 19.000000 | 17.000000 | 22.000000 |
| **25%** | 5.750000 | 16.700000 | 16.400000 | 17.100000 | 68.000000 | 65.000000 | 72.000000 |
| **50%** | 11.500000 | 19.200000 | 18.900000 | 19.700000 | 83.000000 | 80.000000 | 86.000000 |
| **75%** | 17.250000 | 22.000000 | 21.400000 | 22.700000 | 93.000000 | 91.000000 | 94.000000 |
| **max** | 23.000000 | 34.700000 | 33.200000 | 35.000000 | 100.000000 | 99.000000 | 100.000000 |

Fig.25 - Temperatura, estatísticas do *dataframe*
*Fig.25 - Temperature, dataframe statistics*

Para possibilitar a análise da variação de temperatura ao longo do dia por mês, os dados foram agrupados em quatro períodos conforme o horário, da seguinte forma;

- **Período 0**: $[0,6[$ (00:00 - 05:59)
- **Período 1**: $[6,12[$ (06:00 - 11:59)
- **Período 2**: $[12,18[$ (12:00 - 17:59)
- **Período 3**: $[18,0[$ (18:00 - 23:59)

Para a coluna *temp* calculou-se a média por período, para coluna *temp_min* selecionou-se o menor valor por período e para a coluna *temp_max* selecionou-se o maior valor por período.
O gráfico na Fig.26 foi elaborado a partir dos dados agrupados, ele exibe nas barras a temperatura média por período do dia para cada mês, para os dados avaliados, e exibe também a faixa entre a média da temperatura mínima e média da temperatura máxima para cada período do dia. Neste gráfico é possível notar que os períodos 0 e 1, entre 00:00 e 11:59 h, apresentam temperaturas inferiores aos períodos 2 e 3, entre 12:00 e 23:59 h, e também menor variação entre mínimo e máximo, em todos os meses para os dados avaliados.

*To allow the analysis of temperature variation along the day by month, the data has been grouped in four periods according to the time as following;*

- **Period 0**: $[0,6[$ *(00:00 - 05:59)*
- **Period 1**: $[6,12[$ *(06:00 - 11:59)*
- **Period 2**: $[12,18[$ *(12:00 - 17:59)*
- **Period 3**: $[18,0[$ *(18:00 - 23:59)*

*For column* temp *the period mean has been calculated, for column* temp_min *the minimum period value has been selected, and for column* temp_max *the maximum period value has been selected.*
*The plot in Fig.26 has been designed based on the grouped data, on its bars it shows the mean temperature for each period by month, for the evaluated data, and it also shows the range between mean minimum temperature and mean maximum temperature for each period. From this plot is noticeable that periods 0 and 1, from 00:00 to 11:59 o'clock, have lower temperature than periods 2 and 3, from 12:00 to 23:59 o'clock, and lower variation between minimum and maximum as well, for all months for the evaluated data.*

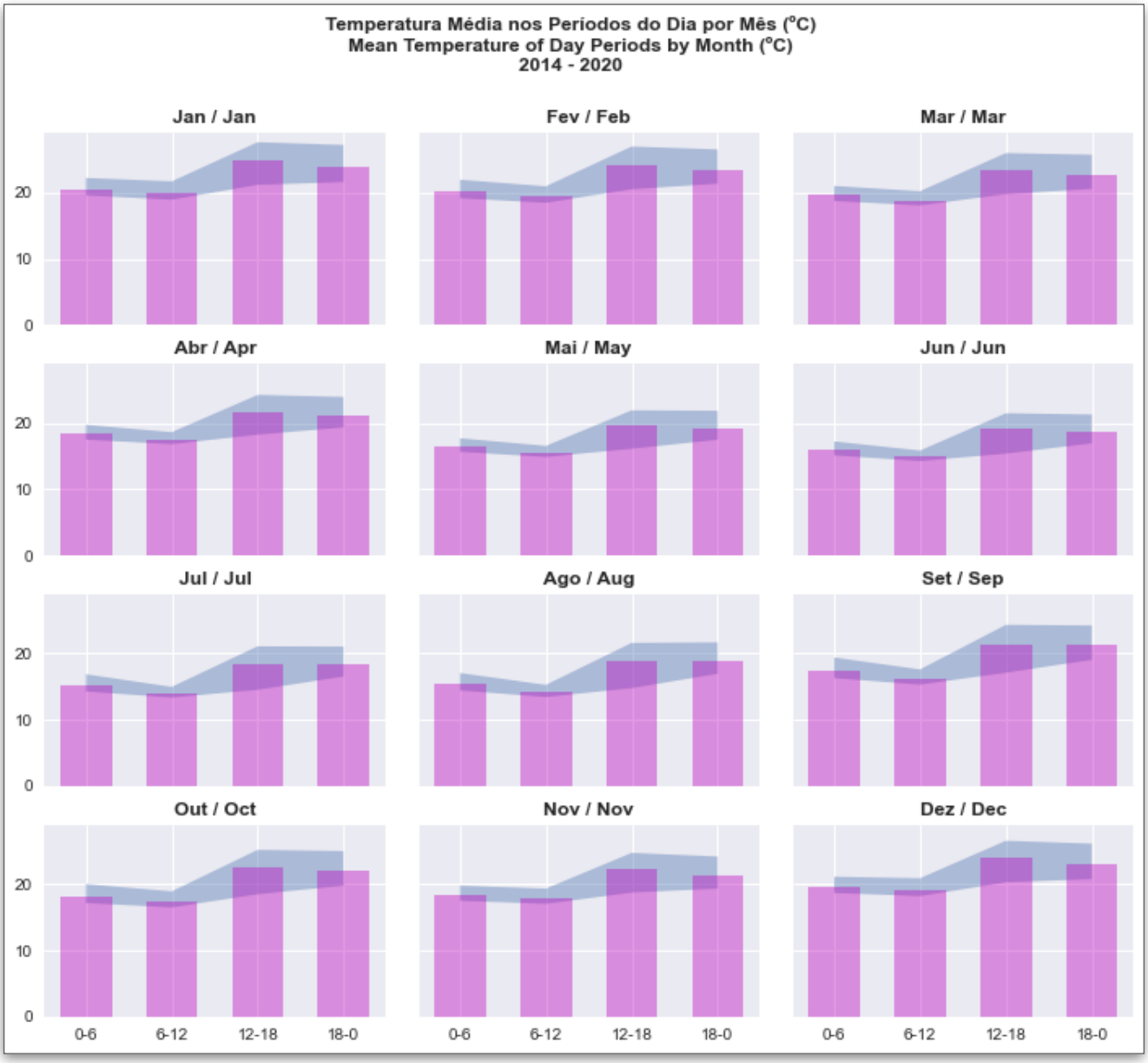


Fig.26 - Temperatura, média por período do dia

*Fig.26 - Temperature, mean per day period*

Também foi verificada a variação de temperatura por dia, os dados foram agrupados por mês e dia, calculou-se a média por dia para a coluna *temp*, selecionou-se o menor valor para coluna *temp_min* e o maior valor coluna *temp_max*.

O gráfico na Fig.27 exibe nas barras a temperatura média e a faixa entre a média das mínimas e média das máximas por dia para cada mês, para os dados avaliados. Neste gráfico é possível notar acentuada queda na média ao final de abril e grande variação na média diária entre os meses de julho a outubro. A variação entre as temperaturas mínima e máxima parece menor nos meses de maio e junho.

*The temperature variation by day has also been verified, the data has been grouped by month and da, the mean for column* temp *has been calculated, the lower value for column* temp_min *and higher value for column* temp_max *selected.*

*The plot in Fig.27 shows the mean temperature on bars and the range between mean minimum and mean maximum per day by month, for the evaluated data. From this plot is possible to notice a acute drop in temperature by end of April and high variation on daily mean between June and October. The minimum and maximum variation seems to be lower in May and June.*

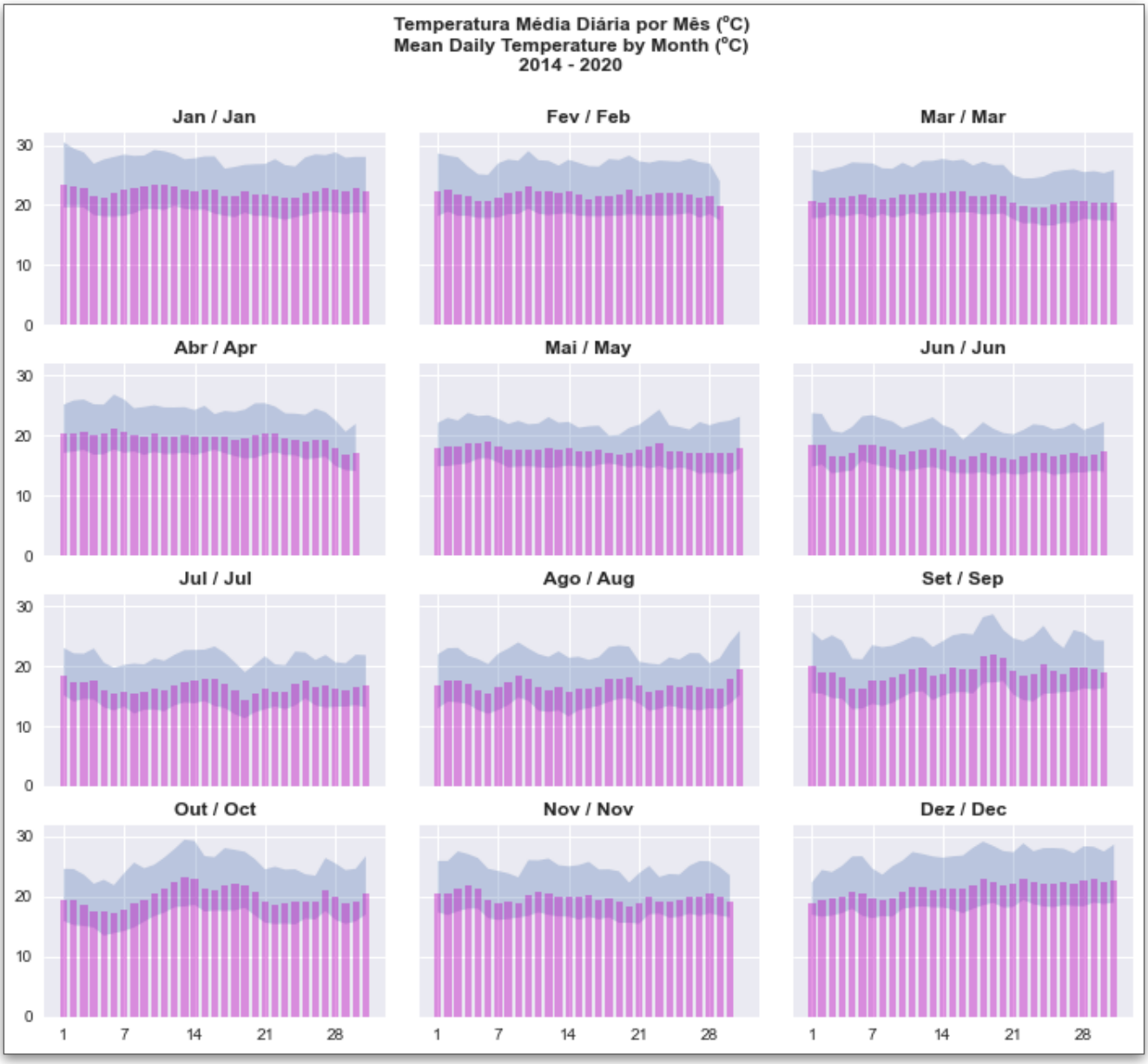


Fig.27 - Temperatura, média por dia

*Fig.27 - Temperature, mean per day*

O gráfico da Fig.28 mostra a variação da temperatura ao longo do ano com a faixa entre a média das mínimas e a média das máximas, para o período avaliado. O período entre os meses de maio e agosto é o mais frio do ano, entre setembro e novembro as temperaturas são mais amenas, entre dezembro e março é o período mais quente e em abril a temperatura também é amena. A média da temperatura máxima é inferior a *30ºC* e a média da temperatura mínima está em torno de *13ºC*, o que pode confirmar que extremos de na temperatura mínima e máxima não são o padrão, ou não são frequentes, para a região.

*The plot in Fig.28 shows the temperature variation along the year with the mean minimum and mean maximum range, for the evaluated period. Between May and August is the coldest year period, between September and November the temperature is mild, between December and March is the hottest period and in April temperature is again mild.*

*The mean of maximum temperature is lower than 30ºC and the mean of minimum temperature is approximately 13ºC, what could confirm that extremely high and low temperature are outlier, or not frequent, for the region.*

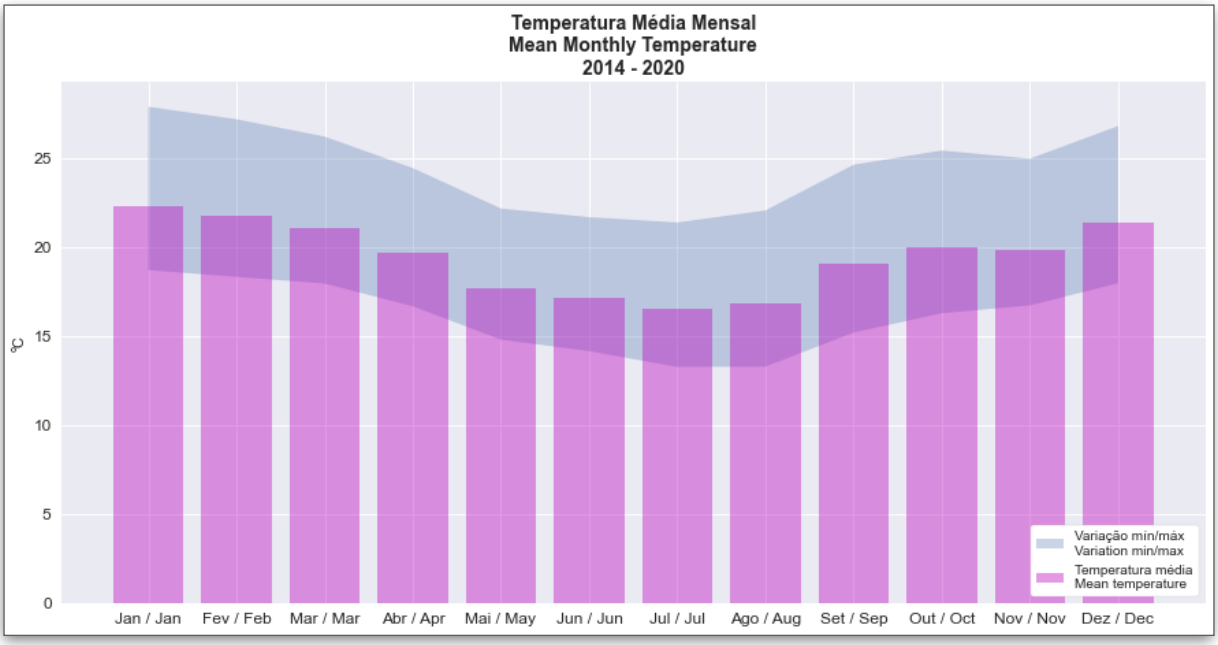


Fig.28 - Temperatura, média por mês

*Fig.28 - Temperature, mean per month*

O gráfico de linhas na Fig.29 mostra a variação da temperatura com a faixa entre a média das mínimas e a média das máximas por mês, para o período entre 2014-11 e 2020-06. Pelo gráfico é possível ver que a temperatura média mais baixa foi no inverno de 2017 *15ºC*, com a média mínima atingindo o menor valor *12ºC*, a temperatura média mais alta foi no verão de 2014-2015 *24ºC*, quando a média da máxima também atingiu o pico *30ºC*. Também e visível que a média da temperatura máxima neste período caiu de *30ºC* em 2014-2015 para aproximadamente *26ºC* em 2019-2020, embora aparentemente o padrão da temperatura média não apresente variação significativa neste intervalo.

*The line plot in Fig.29 shows the temperature variation with the mean minimum and mean maximum range per month, for the period between 2014-11 and 2020-06. From the plot is visible that the lowest mean temperature was in Winter 2017* 15ºC*, having the mean of minimum also at its lowest value* 12ºC*, the highest mean temperature was in Summer 2014-2015* 24ºC*, when the mean of maximum also achieved its peak* 30ºC*. It is also visible that the mean of maximum temperature has dropped for the period from* 30ºC *in 2014-2015 to approximately* 26ºC *in 2019-2020, even though apparently the mean temperature pattern does not seem to have significant variation in this period.*

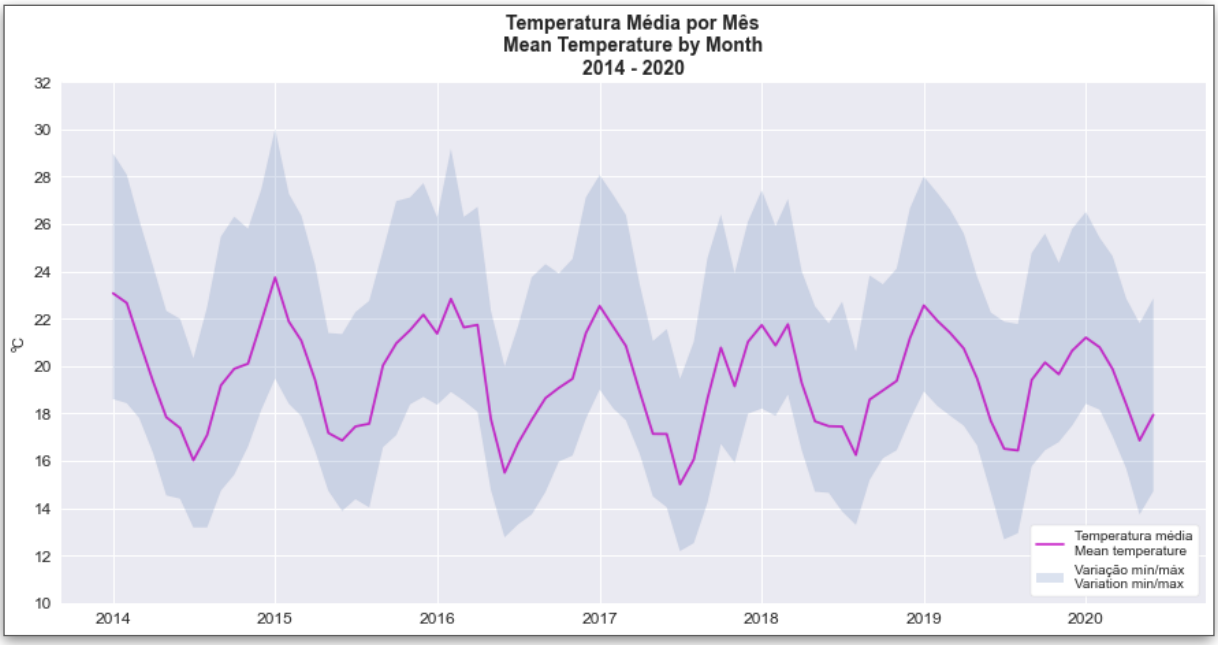


Fig.29 - Temperatura, média por mês para o período avaliado

*Fig.29 - Temperature, men per month for the evaluated period*

Selecionando-se no *dataframe* os registros com a mais alta temperatura máxima duas datas muito próximas são retornadas, 2014-10-14 e 2014-10-19, ambas tendo registrado *35,0ºC* de temperatura máxima, o que está bastante acima das médias da temperatura máxima e que está próximo do período de maiores médias, veja Fig.30.

*When selecting from the dataframe the observations with the highest maximum temperature, two close dates are returned, 2014-10-14 and 2014-10-19, both having recorded* 35.0ºC *maximum temperature, what is much higher than the mean of maximum temperature and it is also close to the period with the highest means, see Fig.30.*

```
# chech day with highest temperature
df_weatherday[df_weatherday['temp_max'] == max(df_weatherday['temp_max'])]
```

| | date | temp_mean | temp_min | temp_max | rel_humidity_mean | rel_humidity_min | rel_humidity_max | year_month |
|---|---|---|---|---|---|---|---|---|
| **286** | 2014-10-14 | 25.7 | 21.1 | 35.0 | 45.2 | 18.0 | 69.0 | 2014-10 |
| **291** | 2014-10-19 | 25.6 | 20.0 | 35.0 | 57.8 | 25.0 | 83.0 | 2014-10 |

Fig.30 - Temperature, dia com mais alta máxima
*Fig.30 - Temperature, day with the highest maximum*

Selecionando-se os registros com a mais baixa temperatura mínima uma data é retornada, 2013-06-13, tendo registrado temperatura mínima de *7,2ºC*, bem abaixo das médias da temperatura mínima, veja Fig.31.

*When selecting the observations with the lowest minimum temperature one day is returned, 2013-06-13, having recorded* 7.2ºC *minimum temperature, much lower than the mean of minimum temperature, see Fig.31.*

```
# chech day with lowest temperature
df_weatherday[df_weatherday['temp_min'] == min(df_weatherday['temp_min'])]
```

| | date | temp_mean | temp_min | temp_max | rel_humidity_mean | rel_humidity_min | rel_humidity_max | year_month |
|---|---|---|---|---|---|---|---|---|
| **894** | 2016-06-13 | 10.5 | 7.2 | 15.6 | 75.0 | 41.0 | 95.0 | 2016-06 |

Fig.31 - Temperature, dia com mais baixa mínima
*Fig.31 - Temperature, day with the lowest minimum*

### Análise da variação da umidade relativa

As estatísticas do *dataframe* mostraram a umidade relativa média alta *79,19%*, típico para o clima tropical de altitude da região, o desvio padrão é também alto *15,73%*, e o intervalo interquartil vai de *68,00* a *93,00%*, aparentemente a uma grande variação na umidade relativa.

A mínima registrada para o período avaliado é de *17.00%*, muito baixo para a região. Normalmente o INMET emite alerta quando a umidade relativa atinge 30% e é considerada situação crítica em 20%. Veja Fig. 32.

### *Relative humidity variation analysis*

*Dataframe statistics showed that mean relative humidity is high* 79.19%, *typical for the region high-altitude tropical climate, the standard deviation is also high* 15.73%, *and the interquartile interval ranges from* 68.00 *to* 93.00%, *apparently there is a huge variation in relative humidity.*

*The minimum recorded for the evaluated period is* 17.00%, *extremely low for the region. Normally the INMET issues an alert when relative humidity drops to 30% and it is considered critical situation at 20%. See Fig.32.*

| | time_utc | temp | temp_min | temp_max | rel_humidity | rel_humidity_min | rel_humidity_max |
|---|---|---|---|---|---|---|---|
| **count** | 56952.000000 | 56952.000000 | 56952.000000 | 56952.000000 | 56952.000000 | 56952.000000 | 56952.000000 |
| **mean** | 11.500000 | 19.482771 | 19.001689 | 20.017200 | 79.192069 | 76.765764 | 81.431679 |
| **std** | 6.922247 | 3.934341 | 3.747882 | 4.131531 | 15.725555 | 16.628299 | 14.736559 |
| **min** | 0.000000 | 7.300000 | 7.200000 | 7.400000 | 19.000000 | 17.000000 | 22.000000 |
| **25%** | 5.750000 | 16.700000 | 16.400000 | 17.100000 | 68.000000 | 65.000000 | 72.000000 |
| **50%** | 11.500000 | 19.200000 | 18.900000 | 19.700000 | 83.000000 | 80.000000 | 86.000000 |
| **75%** | 17.250000 | 22.000000 | 21.400000 | 22.700000 | 93.000000 | 91.000000 | 94.000000 |
| **max** | 23.000000 | 34.700000 | 33.200000 | 35.000000 | 100.000000 | 99.000000 | 100.000000 |

Fig.32 - Umidade relativa, estatísticas do *dataframe*
*Fig.32 - Relative humidity, dataframe statistics*

Da mesma forma que para a análise da variação de temperatura, para a análise da umidade relativa ao longo do dia por mês, os dados também foram agrupados em quatro períodos conforme o horário;

- **Período 0**: $[0,6[$ (00:00 - 05:59)
- **Período 1**: $[6,12[$ (06:00 - 11:59)
- **Período 2**: $[12,18[$ (12:00 - 17:59)
- **Período 3**: $[18,0[$ (18:00 - 23:59)

Para a coluna *rel_humidity* foi calculada a média por período, para coluna *rel_humidity_min* foi selecionado o menor valor por período e para a coluna *rel_humidity_max* foi selecionado o maior valor por período.
O gráfico na Fig.33 exibe nas barras a umidade relativa média por período do dia para cada mês e exibe também a faixa entre a média das mínima e a média das máxima para cada período do dia. Neste gráfico é possível notar que os períodos 0 e 1, entre 00:00 e 11:59 h, apresentam humidade relativa mais alta que os períodos 2 e 3, entre 12:00 e 23:59 h, e também menor variação entre mínimo e máximo, em todos os meses para os dados avaliados. Interessante notar também que o comportamento inverso foi observado na análise da temperatura, Fig.26.

*Like for the temperature variation analysis, for the relative humidity analysis along the day by month, the data has been grouped in four periods according to the time;*

- **Period 0**: $[0,6[$ *(00:00 - 05:59)*
- **Period 1**: $[6,12[$ *(06:00 - 11:59)*
- **Period 2**: $[12,18[$ *(12:00 - 17:59)*
- **Period 3**: $[18,0[$ *(18:00 - 23:59)*

*For column* rel_humidity *the period mean has been calculated, for column* rel_humidity_min *the minimum period value has been selected, and for column* rel_humidity_max *the maximum period value has been selected.*
*The plot in Fig.33 shows on its bars the mean relative humidity per period of the day by month and it also shows the range between the mean minimum and mean maximum for each period of the day. From this plot is noticeable that periods 0 and 1, from 00:00 to 11:59 o'clock, have higher relative humidity than periods 2 and 3, from 12:00 to 23:59 o'clock, and lower variation between minimum and maximum as well, for all months for the evaluated data. It is interesting to also notice that the opposite behaviour has been observed when analysing the temperature, Fig.26.*


Fig.33 - Umidade relativa, média por período do dia

*Fig.33 - Relative humidity, mean by day period*

Para a análise da variação da umidade relativa por dia, os dados foram agrupados por mês e dia, calculou-se a média por dia para a coluna *rel_humidity*, selecionou-se o menor valor para coluna *rel_humidity_min* e o maior valor coluna *rel_humidity_max*.

O gráfico na Fig.34 exibe nas barras a umidade relativa média e a faixa entre a média das mínima e média das máximas por dia para cada mês, para os dados avaliados. É possível notar elevada média, acima de 75% em todas as barras entre março e junho, e novamente entre a segunda semana de novembro e segunda de dezembro, em setembro e outubro aparentemente está o período com as menores médias, com um queda acentuada em meados de outubro.

*For the analysis of relative humidity per day the data has been grouped by month and da, the mean for column* rel_humidity *has been calculated, the lower value for column* rel_humidity_min *and higher value for column* rel_humidity_max *selected.*

*The plot in Fig.34 shows the mean relative humidity on bars and the range between mean minimum and mean maximum per day by month, for the evaluated data. It is possible to notice high mean, above 75%, for bars between March and June, and again between second week November and second week December, apparently in September and October is the period with lowest means, with a acute drop in middle October.*


Fig.34 - Umidade relativa, média por dia

*Fig.34 - Relative humidity, mean per day*

O gráfico da Fig.35 mostra a variação da umidade relativa ao longo do ano com a faixa entre as média das mínimas e média das máximas, para o período avaliado. O período de março a junho e novembro realmente apresentam as médias mais altas e setembro a média mais baixa, mas a variação da umidade relativa ao longo ano não é alta, a média permanece acima de 70% em todos os meses. Interessante notar com é larga a faixa entre as médias das mínimas e as médias das máximas.

*The plot in Fig.35 shows the relative humidity variation along the year with the mean minimum and mean maximum range, for the evaluated period. The period from March to June and November actually present the highest means and September the lowest mean, but the relative humidity variation is not high, the mean stays above 70% in all months. Interesting to notice how large is the range between mean minimum and mean maximum.*

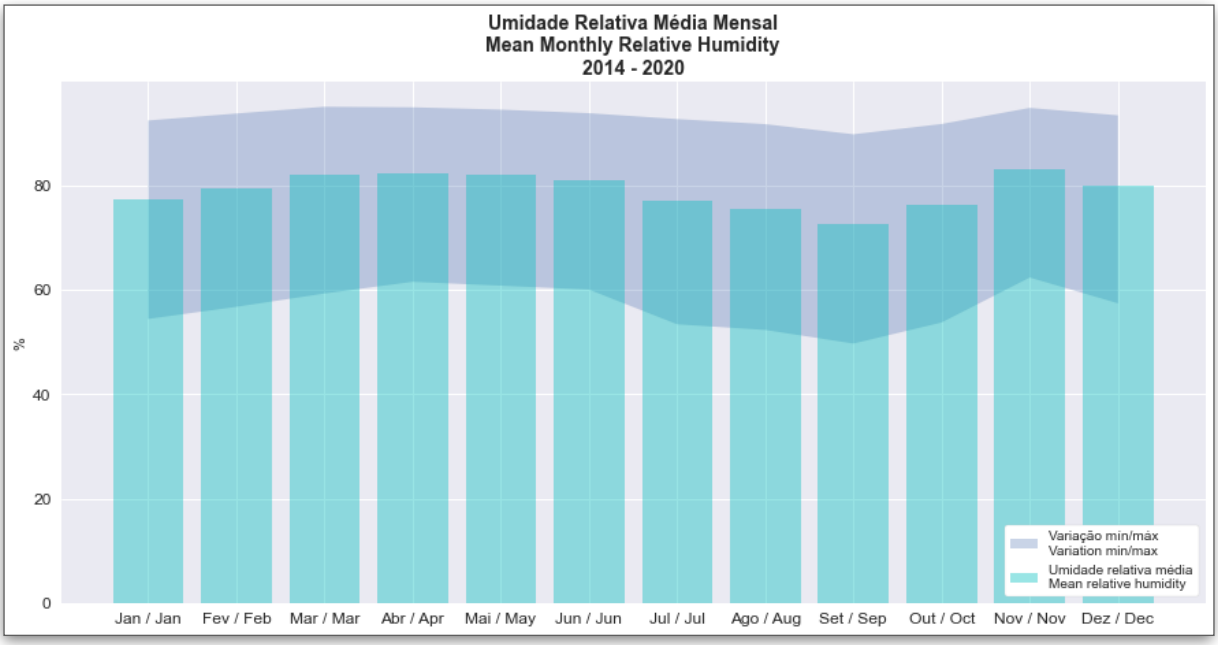


Fig.35 - Umidade relativa, média por mês

*Fig.35 - Relative humidity, mean per month*

O gráfico de linhas na Fig.36 mostra a variação da umidade relativa com a faixa entre a média das mínimas e a média das máximas por mês, para o período entre 2014-11 e 2020-06. Pelo gráfico é possível ver que a umidade relativa média fica entre 70 e 87%, exceto para o três ocorrências no final de 2014 e uma em 2017, quando ficou abaixo de 70%. A médias das mínimas tem variação alta, entre 40 e 65%, com os valores mais baixos também em 2014 e 2017. Interessante notar que o mesmo período ao final de 2014 também apresentou a temperatura média e média da máxima mais altas.

*The line plot in Fig.29 shows the relative humidity variation with the mean minimum and mean maximum range per month, for the period between 2014-11 and 2020-06. From the plot is visible that mean relative humidity lays between 70 and 87%, except from three occurrences at the end of 2014 and one in 2017, when it was below 70%. The mean minimum has high variation, between 40 and 65%, with the lowest values in 2014 and 2017 as well. It is interesting to notice that the same period at the end of 2014 also recorded the highest mean temperature and mean maximum temperature.*

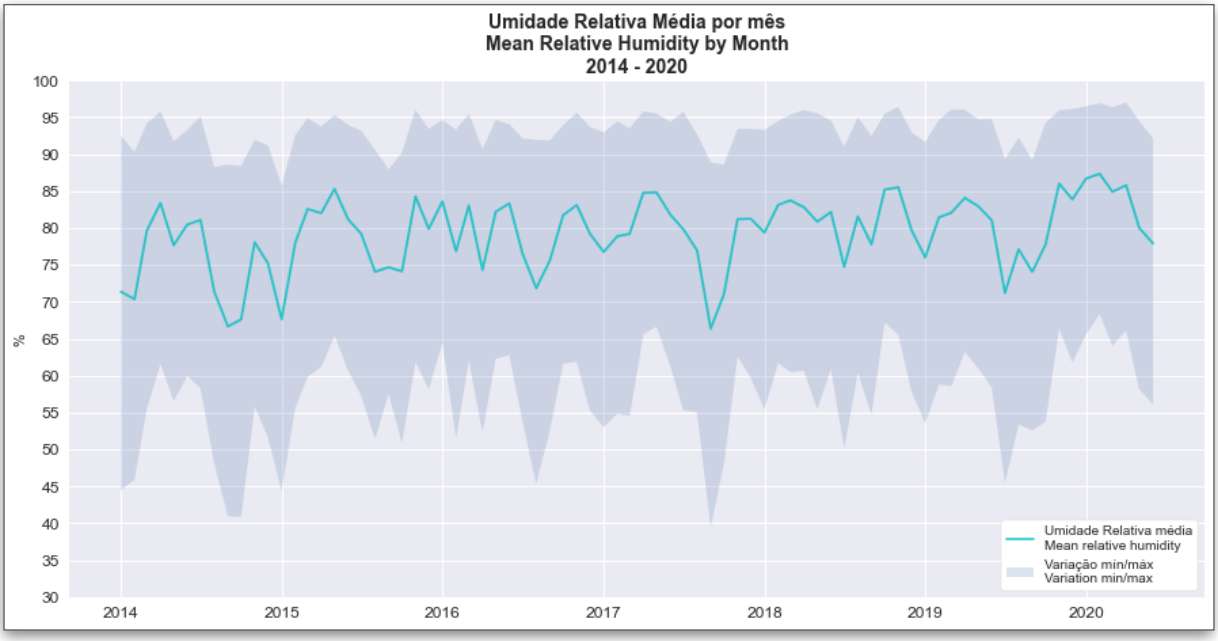


Fig.36 - Umidade relativa, média por mês para o período avaliado

*Fig.36 - Relative humidity, men per month for the evaluated period*

Selecionando-se os registros com a mais baixa umidade relativa mínima uma data é retornada, 2017-10-14, tendo registrado umidade relativa mínima de *17%*, bem abaixo da média das mínima, veja Fig.37.

*When selecting the observations with the lowest minimum relative humidity one day is returned, 2017-10-14, having recorded* 17% *minimum relative humidity, much lower than the mean of minimum, see Fig.37.*

```
# chech day with lowest relative humidity
df_weatherday[df_weatherday['rel_humidity_min'] == min(df_weatherday['rel_humidity_min'])]
```

| | date | temp_mean | temp_min | temp_max | rel_humidity_mean | rel_humidity_min | rel_humidity_max | year_month |
|---|---|---|---|---|---|---|---|---|
| **1382** | 2017-10-14 | 25.3 | 20.3 | 32.1 | 41.8 | 17.0 | 70.0 | 2017-10 |

Fig.37 - Umidade relativa, dia com mais baixa mínima

*Fig.37 - Relative humidity, day with the lowest minimum*

## Análise de correlação

Após a análise de cada conjunto de dados separadamente, consumo de eletricidade, consumo de água, variação de temperatura e variação de umidade relativa, foi analisada a correlação, consumo de água versus variação de umidade relativa e variação de temperatura, e consumo de eletricidade versus variação de umidade relativa e variação temperatura.

Para possibilitar a análise de correlação os dados foram inseridos em um único objeto *dataframe* tendo como referência o período, ano e mês convertidos para o formato *Pandas datetime*. As seguinte colunas foram selecionadas;

- ***period***: coluna comum
- ***electricity***: médio do consumo de eletricidade por dia
- ***water***: média do consumo de água por dia
- ***temperature***: temperatura média
- ***humidity***: umidade relativa média

Os dados foram restritos ao período comum a todos os conjuntos de dados, entre 2014-11 e 2020-06. Veja Fig38.

## *Correlation analysis*

*After data analysis of each dataset separately, electricity consumption, water consumption, temperature variation and relative humidity variation, the correlation analysis has been conducted, water consumption versus temperature variation and relative humidity variation, and electricity consumption versus temperature variation and relative humidity variation.*

*To allow the correlation analysis the data have been inserted into a unique dataframe object having as reference the period, year and month converted to* Pandas datetime *format. The following columns have been selected;*

- **period***: common column*
- **electricity***: mean of electricity consumption per day*
- **water***: mean of water consumption per day*
- **temperature***: mean temperature*
- **humidity***: mean relative humidity*

*The data have been restricted to the common period for all the datasets, between 2014-11 and 2020-06. See Fig38.*

| | period | electricity | water | temperature | humidity |
|---|---|---|---|---|---|
| **0** | 2014-11-01 | 4.545 | 0.424 | 20.1 | 78.1 |
| **1** | 2014-12-01 | 6.267 | 0.571 | 21.9 | 75.2 |
| **2** | 2015-01-01 | 6.226 | 0.562 | 23.7 | 67.7 |
| **3** | 2015-02-01 | 1.542 | 0.375 | 21.9 | 78.0 |
| **4** | 2015-03-01 | 8.743 | 0.552 | 21.1 | 82.6 |
| **...** | ... | ... | ... | ... | ... |
| **63** | 2020-02-01 | 8.290 | 0.207 | 20.8 | 87.4 |
| **64** | 2020-03-01 | 9.310 | 0.452 | 19.9 | 84.9 |
| **65** | 2020-04-01 | 9.152 | 0.387 | 18.4 | 85.8 |
| **66** | 2020-05-01 | 12.286 | 0.379 | 16.9 | 80.0 |
| **67** | 2020-06-01 | 12.000 | 0.375 | 17.9 | 77.9 |

68 rows × 5 columns

Fig.38 - Correlação, *Pandas dataframe*

*Fig.38 - Correlation, Pandas dataframe*

**Análise de correlação no consumo de água**

Inicialmente a correlação linear entre consumo de água e variação de humidade relativa e variação de temperatura foi verificada utilizando-se um *heatmap,* elaborado a partir do Coeficiente de Correlação de Pearson (*r*). O coeficiente resultante é um número real entre *-1* e *1* com a seguinte interpretação;

- ***-1***: perfeita correlação linear negativa entre as variáveis
- ***0***: não há correlação linear entre variáveis, independentes linearmente
- ***1***: perfeita correlação linear positiva entre as variáveis

A diagonal representa a correlação linear da variável com ela mesma, *r = 1,00*. O *heatmap* mostra que a correlação linear entre consumo de água e umidade relativa é fraca, *r = -0,10*, e a correlação linear entre consumo de água e temperatura também é fraca, *r = 0,10*, veja Fig.39.

***Water consumption correlation analysis***

*Initially the linear correlation between water consumption and relative humidity variation and temperature variation has been verified with a heatmat, based on the Pearson Correlation Coefficient (*r*). The resulting coefficient is a real number between* -1 *and* 1 *with the following interpretation;*

- **-1***: perfect negative linear correlation between variables*
- **0***: no linear correlation between variables, linear independent*
- **1***: perfect positive linear correlation between variables*

*The diagonal represents the linear correlation between the variable and itself,* r = 1.00*. The heatmap shows that linear correlation between water consumption and relative humidity is weak,* r = -0.10*, and the linear correlation between water consumption and temperature is also weak,* r = 0.10*, see Fig.39.*

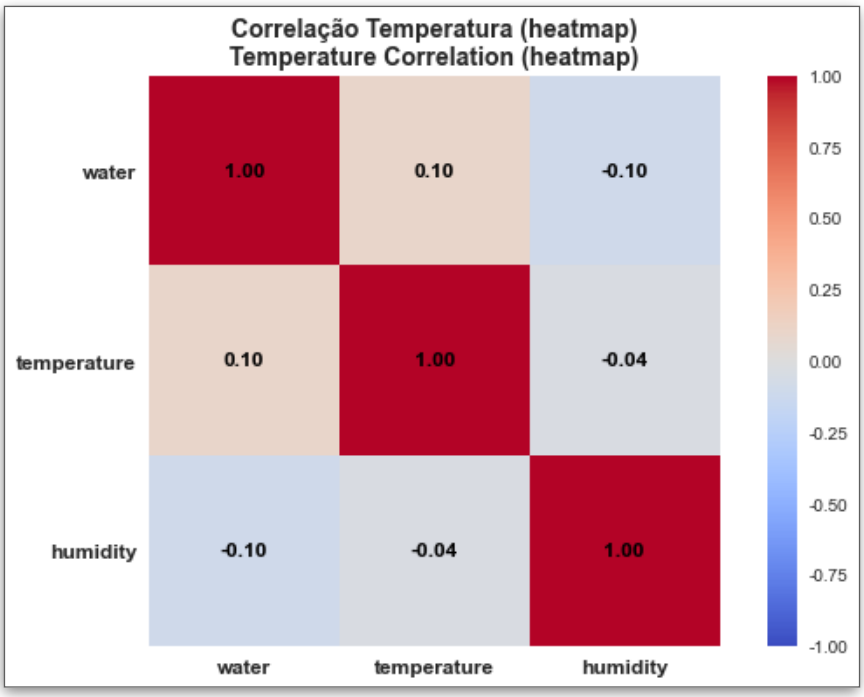


Fig.39 - Correlação, *heatmap* consumo de água

*Fig.39 - Correlation, water consumption heatmap*

O gráfico na Fig.40 mostra a relação entre o consumo de água e umidade relativa, com a linha de regressão. É visível no gráfico que a linha está quase na posição horizontal e não há correlação linear entre os pontos.

*The plot in Fig.40 shows the relation between water consumption and relative humidity, with the regression line. It is visible from the plot that the line lays almost in horizontal position and there is no linear correlation between the dots.*

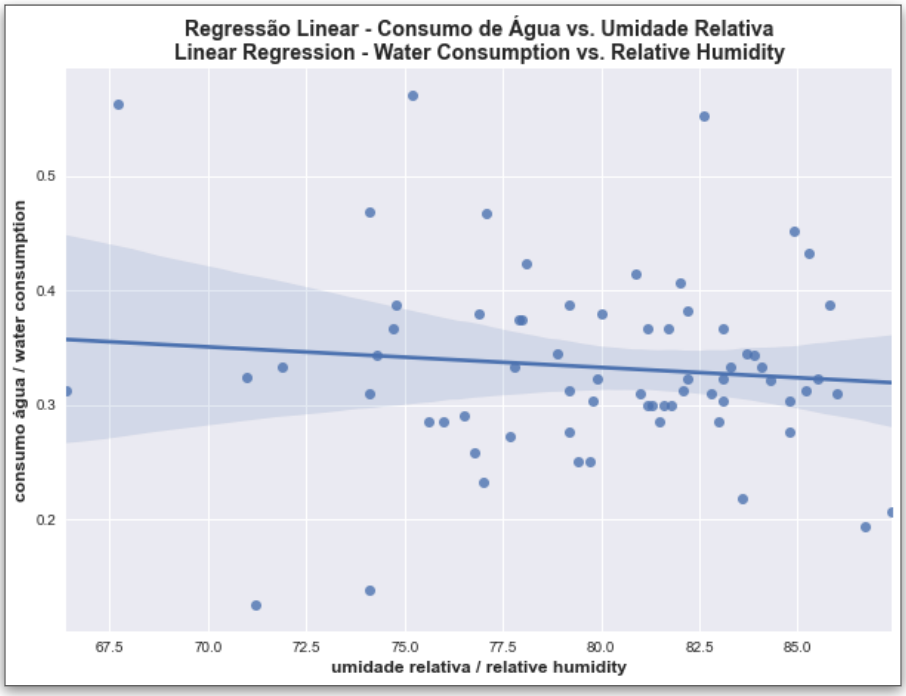


Fig.40 - Correlação, regressão linear consumo de água vs. umidade relativa

*Fig.40 - Correlation, linear regression water consumption vs. relative humidity*

A Fig.41 mostra novamente o Coeficiente de Correlação de Pearson (*r*) e a significância da correlação (*valor-P*) entre consumo de água e variação de umidade relativa. O *valor-P* é um número real entre *0* e *1*, ele pode ser interpretado da seguinte maneira;

- ***> 0,100***: não há evidência de que a correlação seja significante
- ***< 0,100***: há fraca evidência de que a correlação seja significante
- ***< 0,050***: há moderada evidência de que a correlação seja significante
- ***< 0,001***: há forte evidência de que a correlação seja significante

Para a correlação linear entre o consumo de água e variação de umidade relativa, *r = -0,10* indica fraca correlação e *valor-P = 0,421* indica que não há evidência de que a correlação seja significante, ou seja, para o período avaliado não há indícios de que o consumo de água esteja relacionado à variação na umidade relativa.

*The Fig.41 shows once more the Pearson Correlation Coefficient (*r*) and the correlation significance (*P-value*) between water consumption and relative humidity variation. The* P-value *is a real number between* 0 *and* 1*, it can be interpreted as following;*

- **> 0.100***: there is no evidence that the correlation is significant*
- **< 0.100***: there is weak evidence that the correlation is significant*
- **< 0.050***: there is moderate evidence that the correlation is significant*
- **< 0.001***: there is strong evidence that the correlation is significant*

*For the linear correlation between water consumption and relative humidity,* r = -0.10 *indicates weak correlation and* P-value = 0.421 *indicates that there is no evidence that the correlation is significant, what means that for the evaluated period there is no indication that the water consumption is related to the relative humidity variation.*

```
# check Pearson r and P-value
r, p = stats.pearsonr(df_features['water'], df_features['humidity'])

# print results
print('Correlation: Water consumption vs. Relative Humidity\n\
 Pearson r = {:.2f}\n P-value = {:.9f}'.format(r,p))
```

```
Correlation: Water consumption vs. Relative Humidity
 Pearson r = -0.10
 P-value = 0.421329772

For Pearson r = -0.10 there is a weak correlation.

For P-value > 0.1 there is no evidence of statistically significant correlation.
```

Fig.41 - Correlação, significância consumo de água vs. umidade relativa

*Fig.41 - Correlation, significance water consumption vs. relative humidity*

O gráfico na Fig.42 mostra a relação entre o consumo de água e temperatura, com a linha de regressão. É visível no gráfico que a linha está também quase na posição horizontal e não há correlação linear entre os pontos.

*The plot in Fig.42 shows the relation between water consumption and temperature, with the regression line. It is visible from the plot that the line also lays almost in horizontal position and there is no linear correlation between the dots.*

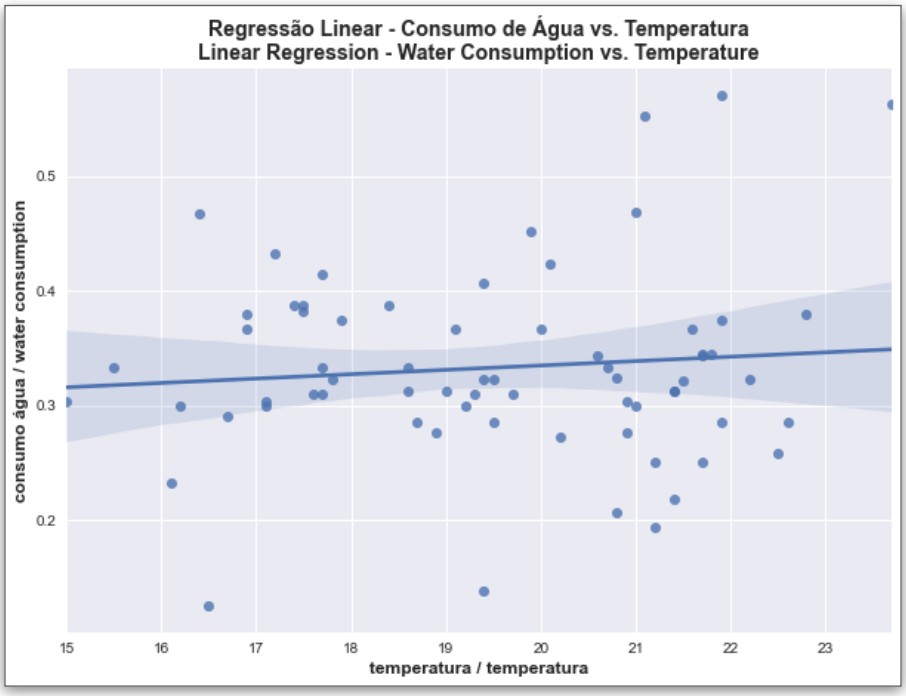


Fig.42 - Correlação, regressão linear consumo de água vs. temperatura

*Fig.42 - Correlation, linear regression water consumption vs. temperature*

A Fig.43 mostra o Coeficiente de Correlação de Pearson (*r*) e a significância da correlação (*valor-P*) entre consumo de água e variação de temperatura, *r = 0,10* indica fraca correlação e *valor-P = 0,438* indica que não há evidência de que a correlação seja significante, ou seja, para o período avaliado não há indícios de que o consumo de água esteja relacionado à variação na temperatura.

*The Fig.43 shows the Pearson Correlation Coefficient (*r*) and the correlation significance (*P-value*) between water consumption and temperature variation,* r = 0.10 *indicates weak correlation and* P-value = 0.438 *indicates that there is no evidence that the correlation is significant, what means that for the evaluated period there is no indication that the water consumption is related to the temperature variation.*

```
# check Pearson r and P-value
r, p = stats.pearsonr(df_features['water'], df_features['temperature'])

# print results
print('Correlation: Water consumption vs. Temperature\n\
 Pearson r = {:.2f}\n P-value = {:.9f}'.format(r,p))
```

```
Correlation: Water consumption vs. Temperature
 Pearson r = 0.10
 P-value = 0.437977141
```

For **Pearson r = -0.10** there is a weak correlation.

For **P-value > 0.1** there is no evidence of statistically significant correlation.

Fig.43 - Correlação, significância consumo de água vs. temperatura

*Fig.43 - Correlation, significance water consumption vs. temperature*

**Análise de correlação no consumo de eletricidade**

A correlação linear entre consumo de eletricidade e variação de humidade relativa e variação de temperatura também foi verificada utilizando-se um *heatmap,* elaborado a partir do Coeficiente de Correlação de Pearson (*r*).

***Electricity consumption correlation analysis***

*The linear correlation between electricity consumption and relative humidity variation and temperature variation has also been verified with a heatmat, based on the Pearson Correlation Coefficient (*r*).*

O *heatmap* mostra que a correlação linear entre consumo de eletricidade e umidade relativa é fraca, *r = 0,16*, e a correlação linear entre consumo de eletricidade e temperatura é moderadamente e negativa, *r = -0,62*, veja Fig.44.

*The heatmap shows that linear correlation between electricity consumption and relative humidity is weak,* r = 0.16*, and the linear correlation between electricity consumption and temperature is moderate and negative,* r = -0.62*, see Fig.44.*

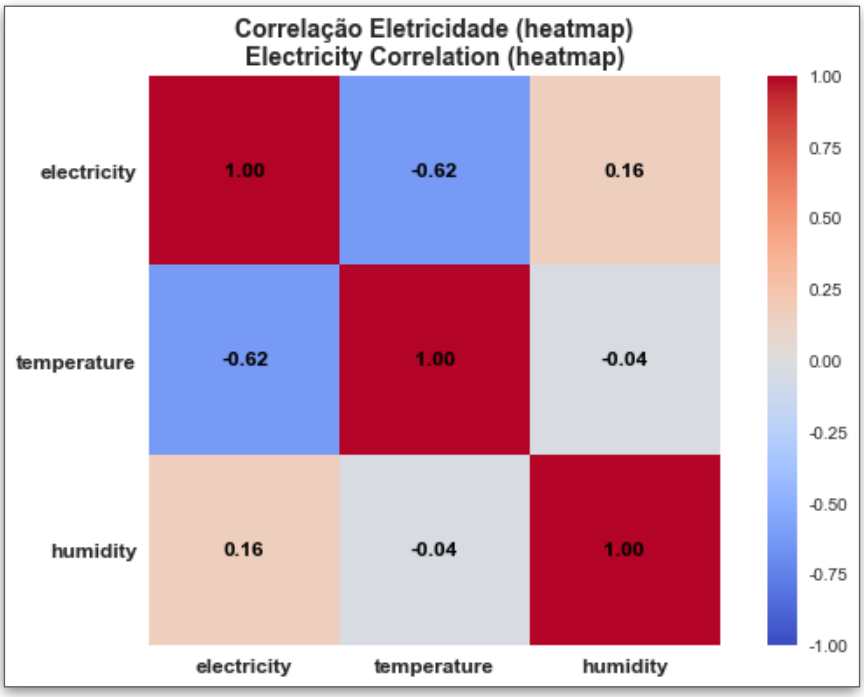


Fig.44 - Correlação, *heatmap* consumo de eletricidade

*Fig.44 - Correlation, electricity consumption heatmap*

O gráfico na Fig.45 mostra a relação entre o consumo de eletricidade e umidade relativa, com a linha de regressão. É visível no gráfico, assim como para o consumo de água, que a linha está quase na posição horizontal e não há correlação linear entre os pontos.

*The plot in Fig.45 shows the relation between electricity consumption and relative humidity, with the regression line. It is visible from the plot that the line, like for water consumption, lays almost in horizontal position and there is no linear correlation between the dots.*

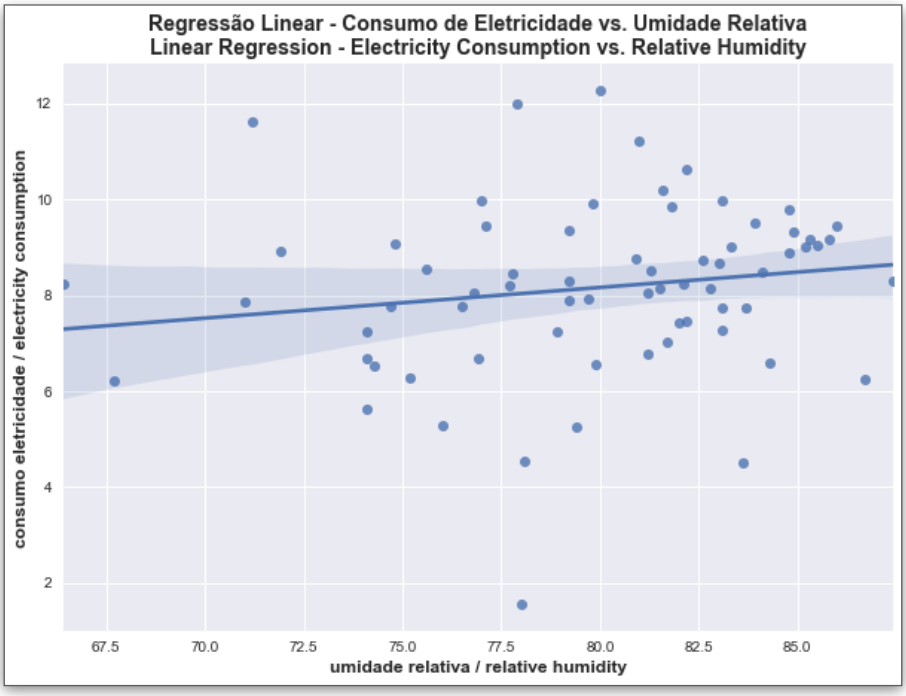


Fig.45 - Correlação, regressão linear consumo de eletricidade vs. umidade relativa

*Fig.45 - Correlation, linear regression electricity consumption vs. relative humidity*

A Fig.46 mostra o Coeficiente de Correlação de Pearson (*r*) e a significância da correlação (*valor-P*) entre consumo de eletricidade e variação de umidade relativa, *r = 0,16* indica fraca correlação e *valor-P = 0,189* indica que não há evidência de que a correlação seja significante, ou seja, para o período avaliado não há indícios de que o consumo de eletricidade esteja relacionado à variação na umidade relativa.

*The Fig.46 shows the Pearson Correlation Coefficient (*r*) and the correlation significance (*P-value*) between electricity consumption and relative humidity variation,* r = 0.16 *indicates weak correlation and* P-value = 0.189 *indicates that there is no evidence that the correlation is significant, what means that for the evaluated period there is no indication that the electricity consumption is related to the relative humidity variation.*

```
# check Pearson r and P-value
r, p = stats.pearsonr(df_features['electricity'], df_features['humidity'])

# print results
print('Correlation: Electricity consumption vs. Relative Humidity\n\
 Pearson r = {:.2f}\n P-value = {:.9f}'.format(r,p))
```

```
Correlation: Electricity consumption vs. Relative Humidity
 Pearson r = 0.16
 P-value = 0.188638451
```

For **Pearson r = 0.16** there is a weak correlation.

For **P-value > 0.1** there is no evidence of statistically significant correlation.

Fig.46 - Correlação, significância consumo de eletricidade vs. umidade relativa

*Fig.46 - Correlation, significance electricity consumption vs. relative humidity*

O gráfico na Fig.47 mostra a relação entre o consumo de eletricidade e temperatura, com a linha de regressão. É possível perceber no gráfico a inclinação na linha e certa aproximação dos pontos à linha de regressão. Como indicou o *heatmap* a correlação é moderada, assim não há uma coincidência exata com a linha de regressão. A correlação negativa indica que à medida que a temperatura aumenta os valores no consumo de eletricidade diminuem.

*The plot in Fig.47 shows the relation between electricity consumption and temperature, with the regression line. It is possible to notice from the plot that there is a line coefficient and some proximity of the dots to the regression line. As indicated by the heatmap, the correlation is moderate, there is no exact coincidence to the regression line. The negative correlation indicates that when temperature rises the values on electricity consumption drop.*

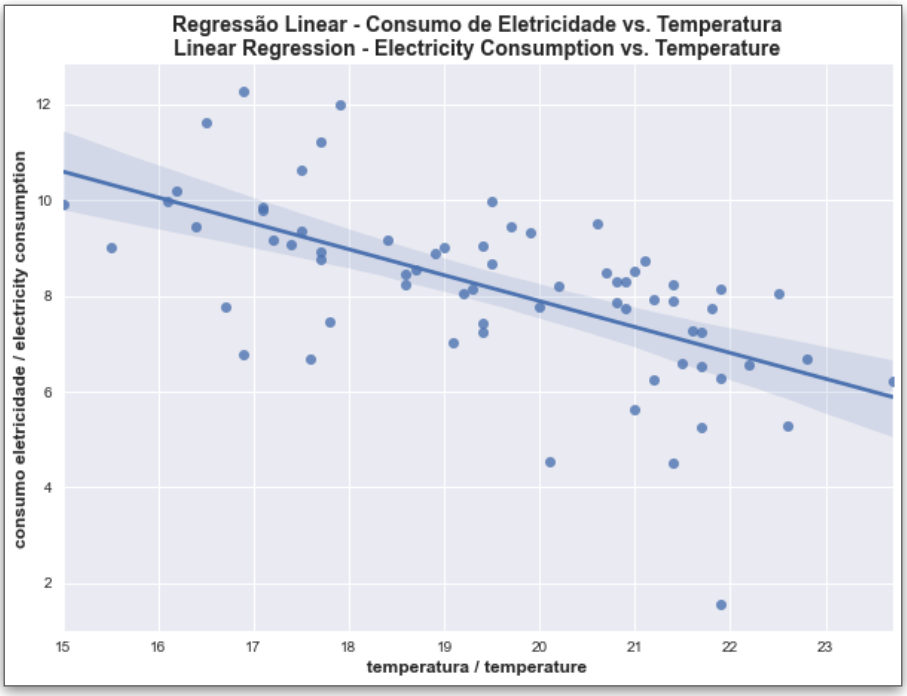


Fig.47 - Correlação, regressão linear consumo de eletricidade vs. temperatura

*Fig.47 - Correlation, linear regression electricity consumption vs. temperature*

A Fig.48 mostra o Coeficiente de Correlação de Pearson (*r*) e a significância da correlação (*valor-P*) entre consumo de eletricidade e variação de temperatura, *r = -0,62* indica moderada correlação e *valor-P = 0,000000017* indica que há forte evidência de que a correlação seja significante, ou seja, para o período avaliado há indícios de que o consumo de eletricidade esteja inversamente relacionado à variação na temperatura.

*The Fig.48 shows the Pearson Correlation Coefficient (*r*) and the correlation significance (*P-value*) between electricity consumption and temperature variation,*
r = -0.62 *indicates moderate correlation and*
P-value = 0.000000017 *indicates that there is strong evidence that the correlation is significant, what means that for the evaluated period there is indication that the electricity consumption is inversely related to the temperature variation.*

```
# check Pearson r and P-value
r, p = stats.pearsonr(df_features['electricity'], df_features['temperature'])

# print results
print('Correlation: Electricity consumption vs. Temperature\n\
 Pearson r = {:.2f}\n P-value = {:.9f}'.format(r,p))
```

```
Correlation: Electricity consumption vs. Temperature
 Pearson r = -0.62
 P-value = 0.000000017
```

For **Pearson r = -0.62** there is a moderate negative correlation.

For **P-value < 0.001** there is evidence of statistically significant correlation.

Fig.48 - Correlação, significância consumo de eletricidade vs. temperatura
*Fig.48 - Correlation, significance electricity consumption vs. temperature*

## Conclusão

O objetivo deste estudo autônomo foi analisar o consumo de água e eletricidade na residência do autor e verificar se existe relação entre a variação no consumo e a variação de temperatura e variação de umidade relativa.
Para a análise foram coletados os dados de consumo das faturas mensais das companhias de água e eletricidade, os dados meteorológicos foram selecionados a partir do banco de dados público do Instituto Nacional de Meteorologia (INMET) em seu portal na *internet*, veja seção Referências item 2. Os dados analisados compreenderam o período de 2014-11 a 2020-06.
Para o período, a análise mostrou que o consumo de água não segue nenhum padrão aparente de sazonalidade, somente redução no período de férias mais longas no verão. Para o consumo de eletricidade, o consumo foi mais alto nos meses que coincidem com o inverno e mais baixo nos meses que coincidem com o verão, também há redução de consumo no período prolongado de férias no verão.
Quanto à correlação, a análise mostrou que não há indícios de que o consumo de água esteja relacionado com a variação de umidade relativa e nem com a variação de temperatura, e que também não há indícios de que o consumo de eletricidade esteja relacionado à variação de umidade relativa, mas a análise mostrou que existe moderada correlação entre consumo de eletricidade e variação de temperatura, com Coeficiente de Correlação de Pearson igual a *-0,62*, e que há forte evidência de que a correlação seja significante, com valor-P igual a *0,000000017*.

## *Conclusion*

*The objective of this autonomy study was to analyse water and electricity consumption in author's residence and to verify if there any relation between consumption and temperature variation and relative humidity variation.*
*For this analysis consumption data has been collected from invoices provided by water and electricity companies, the meteorological data has been selected from the public data base of* Instituto Nacional de Meteorologia *(*INMET*), the Brazilian institute for meteorology, on its internet portal, see References section item 2. The analysed data covers the period from 2014-11 to 2020-06.*
*For this period, the analysis showed that water consumption apparently does not follow any seasonal pattern, there is only consumption reduction during the long vacation period in Summer. For electricity consumption, the consumption was higher for months during Winter time e lower for months during Summer time, the electricity consumption was reduced during long vacation period in Summer as well.*
*Regarding the correlation, the analysis showed that there is no indication that the water consumption is related to neither relative humidity variation nor temperature variation, it also showed that there is no indication that electricity consumption is related to relative humidity variation, but the analysis showed moderate correlation between electricity consumption and temperature variation, with Pearson Correlation Coefficient equals to* -0,62*, and there is strong indication that the correlation is significant, with P-value equals to* 0,000000017*.*

O estudo atingiu seu objetivo proporcionando uma visão do consumo e também respondendo à questão da correlação.

## Caminhos futuros

A partir deste ponto poderia ser desenvolvido um modelo baseado na relação entre o consumo de eletricidade e variação de temperatura para que se entenda melhor esta relação, definindo-se uma equação matemática para se representar a relação também seria possível conhecer seus limites e até que ponto a variação na temperatura influencia no consumo de energia. Infelizmente o conjunto de dados não contém dados suficientes para elaborar um modelo de *machine learning*, treiná-lo e realizar testes de verificação, para isso o ideal seria uma série temporal mais longa com dados de mais residências e com grupos de perfis definidos.

## Referencias


1. **IBGE - Instituto Brasileiro de Geografia e Estatística**
   https://www.ibge.gov.br

2. **INMET - Instituto Nacional de Meteorologia**
   https://portal.inmet.gov.br


*The study achieved its goal providing some view regarding consumption and answering the correlation question as well.*

## *Future directions*

*From this point a model could be developed based on relation between electricity consumption and temperature variation to understand better this relation, having a defined mathematical equation to represent the relation it would also be possible to know its limits and up to which point the temperature variation influences the electricity consumption. Unfortunately the dataset does not contain sufficient data to define a machine learning model, train it and run verification tests, this would require longer time series with more residences data and with profile groups defined.*

## *References*