Rural Wildfires – Statistical analysis and inferences

2022/2023 – Data Science Engineering Fundaments – Statistics

Authors:

- João Fragão

- Jorge Vieira

- Tomás Rodrigues

Structure of the project

Methodology for this project: 6 Steps Statistical Investigation

Step 1: *Which factors can have an impact on a wider burned area and number of fire occurrences?*

Variables of this study:

Type of cause – Qualitative/categorical

Region: Littoral/Interior - Qualitative/categorical
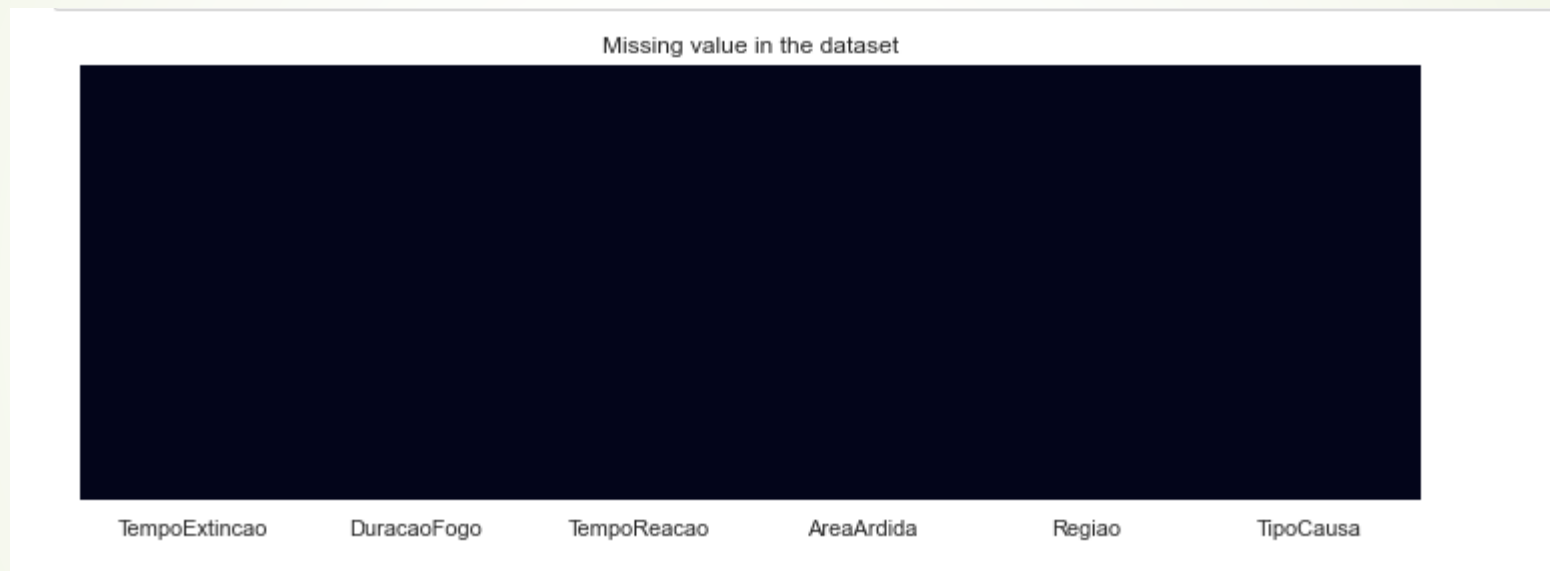
Fire duration - Quantitative

Burned area - Quantitative

Extinction time - Quantitative

Reaction time - Quantitative

Step 2: Getting the data - provided by the Professor

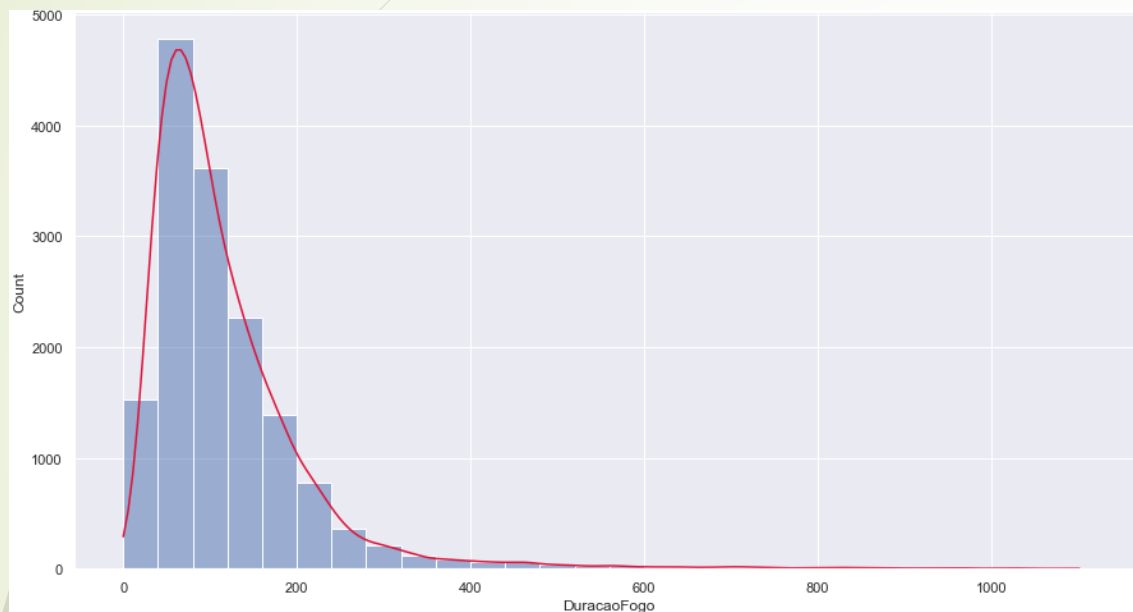## **Step 3** – Exploring the data
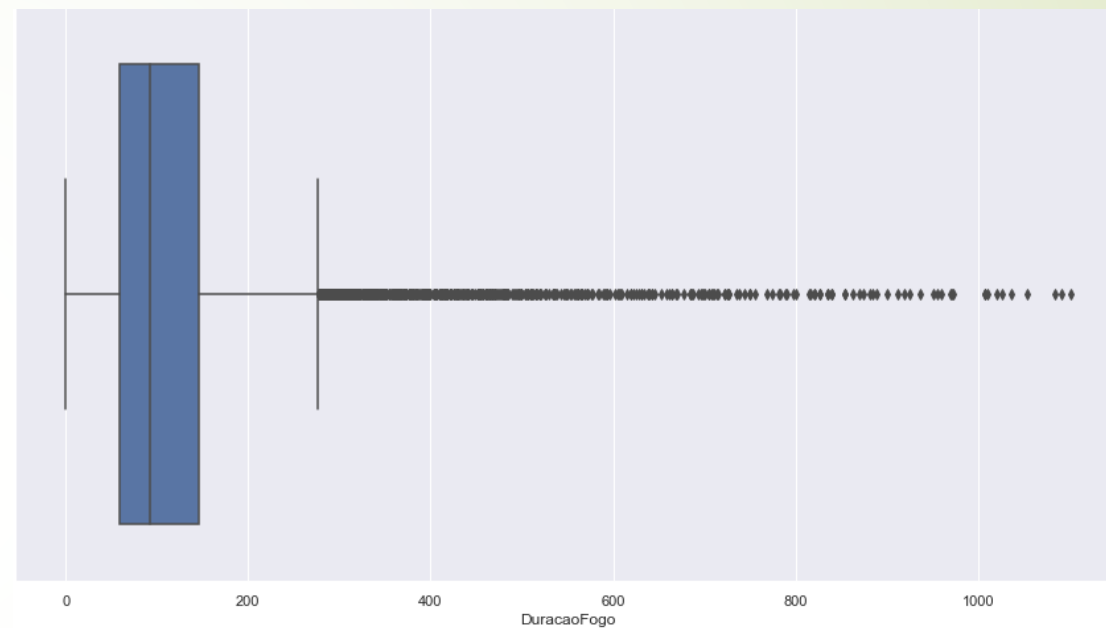## Data cleaning – Null and Error values

After cleaning the data there were no missing values or null values!

Step 3 – Exploring the data
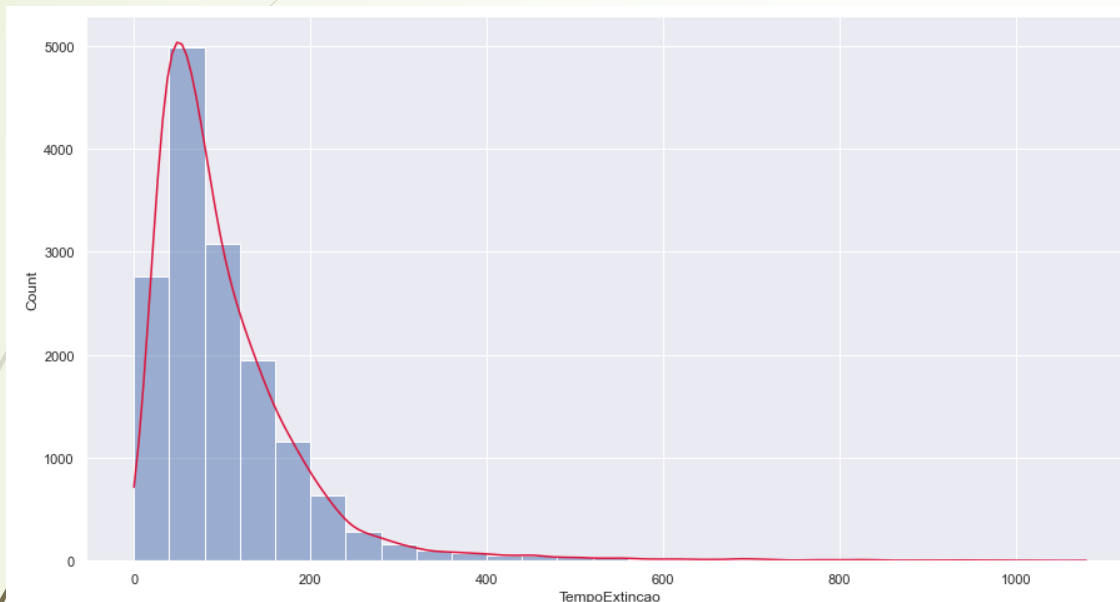## Univariate Analysis – **Fire duration**

- **Histogram** (Relationship between fire duration and number of occurrences)
- **Mode fire duration**: 80min (over 4500 times)
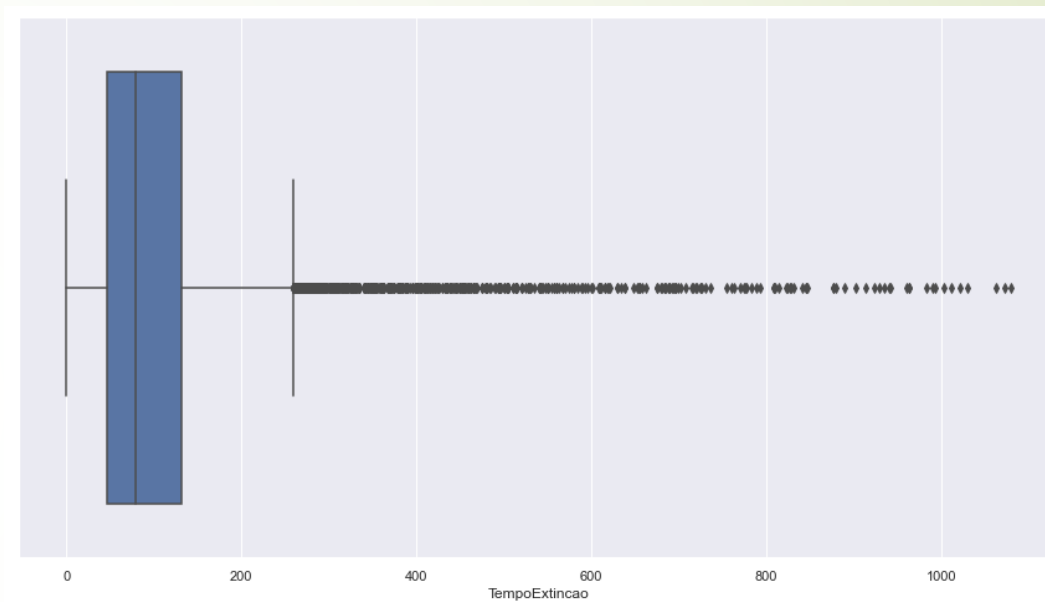- Follows a right skewed distribution (Mode <Median<Mean)

- **Boxplot** for the Fire duration
- **Note**: Outliers removed
- 25%: 59
- Median(50%): 92
- 75%: 146
- Max: 1102
- Count: 15453
- Mean: 117,4

Step 3 – Exploring the data
# Univariate Analysis – **Extinguishing time**

- **Histogram** (Relationship between the extinguishing time and number of occurrences)
- Mode for the extinguishing time: 80min (over 5000 times)
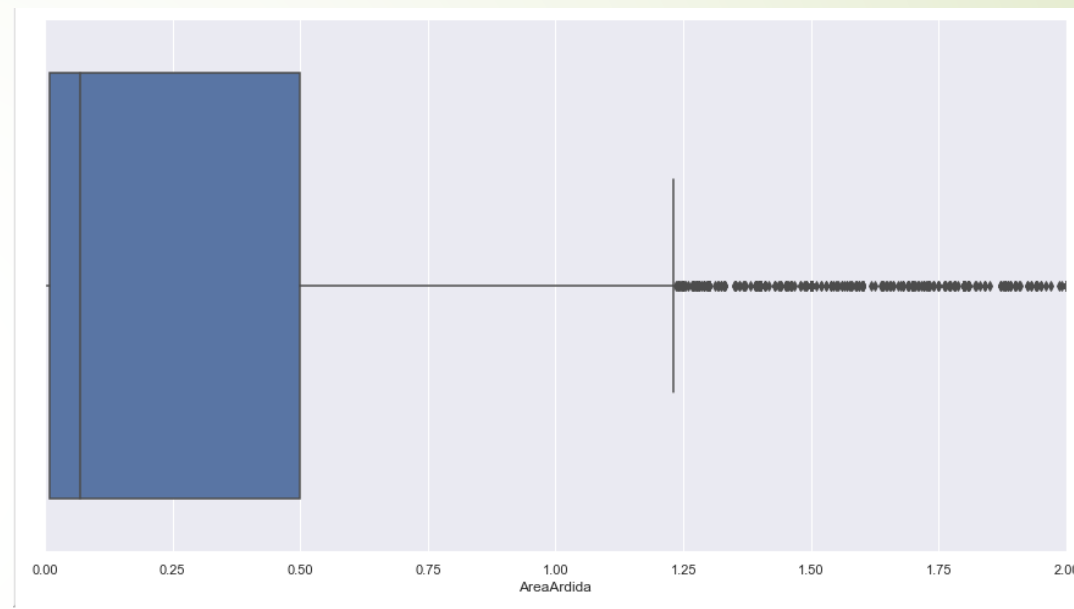- Follows a right skewed distribution (Mode <Median<Average)

- **Boxplot** for the extinguishing time
- **Note**: Outliers removed
- 25%: 47
- Median(50%): 92
- 75%: 132
- Max: 1080
- Count: 15453
- Mean:104,4

Step 3 – Exploring the data
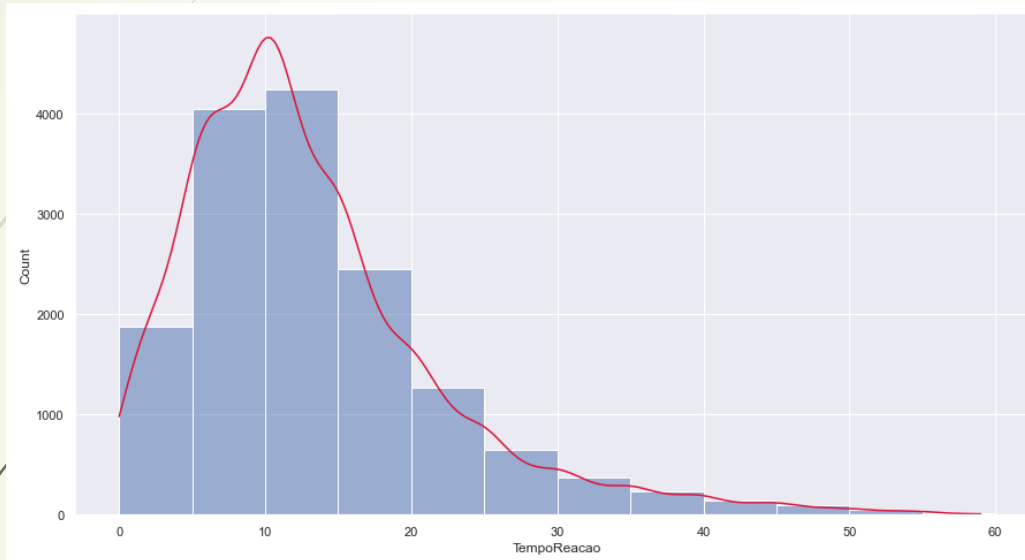# Univariate Analysis – **Burned area**
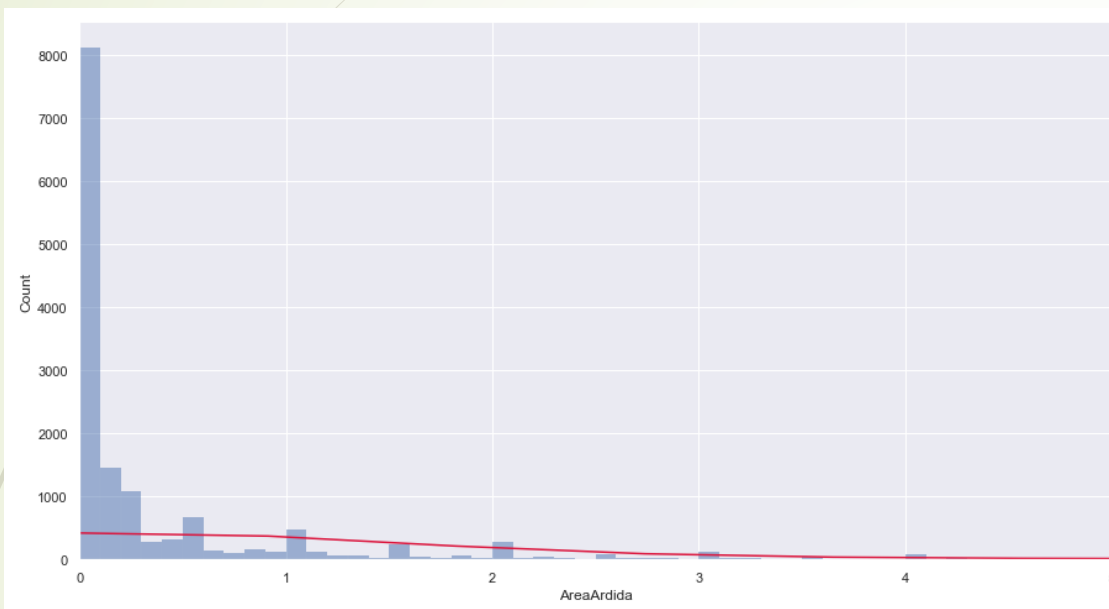
- **Histogram** (Relationship between the burned area and occurrences)
- **Mode** for the burned area: 0,1 acres
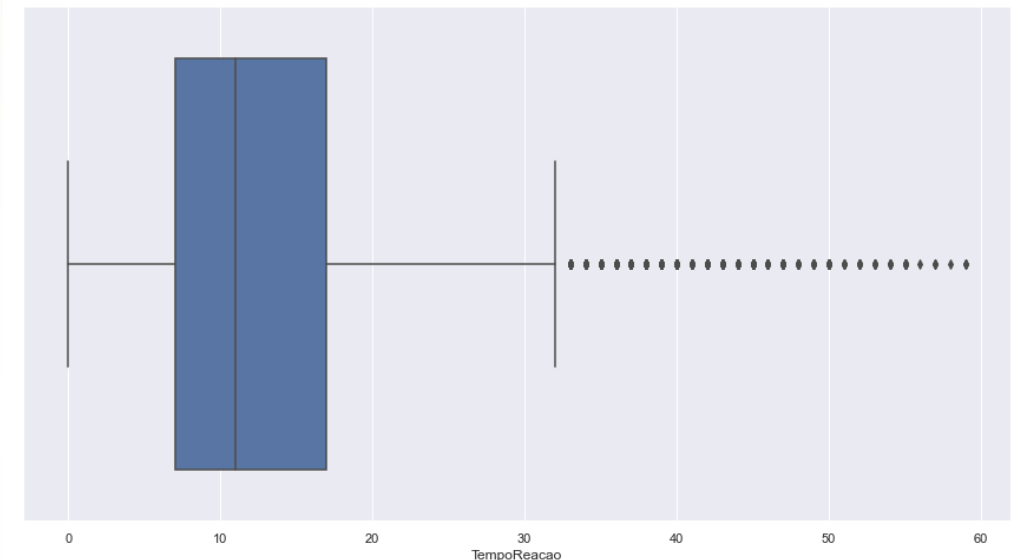- Follows a right skewed distribution (Mode <Median<Average)

- **Boxplot** for the burned area;
- **Note**: Outliers removed
- 25%: 0,01
- Median(50%): 0,07
- 75%: 0,5
- Max: 181,7
- Count: 15453
- Mean: 1,7

Step 3 – Exploring the data
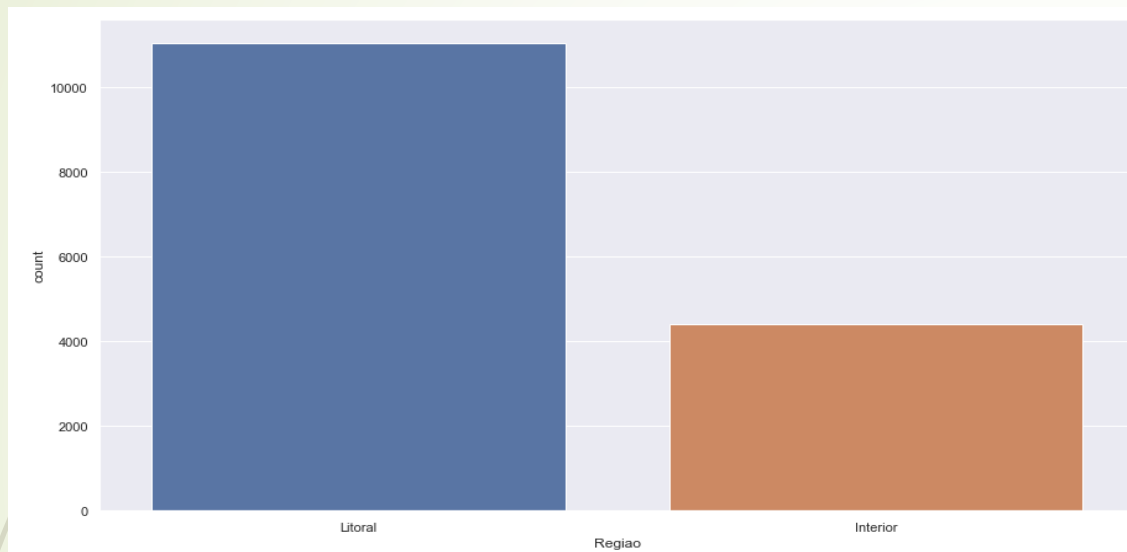# Univariate Analysis – **Reaction time**

- **Histogram** for the Reaction time;
- **Mode**: 15min;
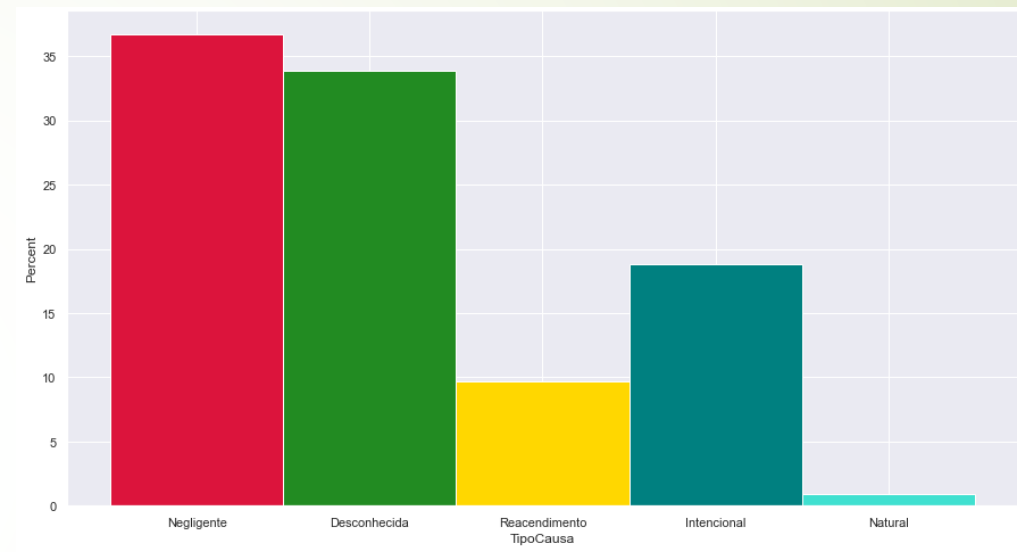- Follows a right skewed distribution (Mode <Median<Average)

- **Boxplot** for the Reaction time;
- **Note**: Outliers removed
- 25%: 7min
- Median(50%): 11min
- 75%: 17min
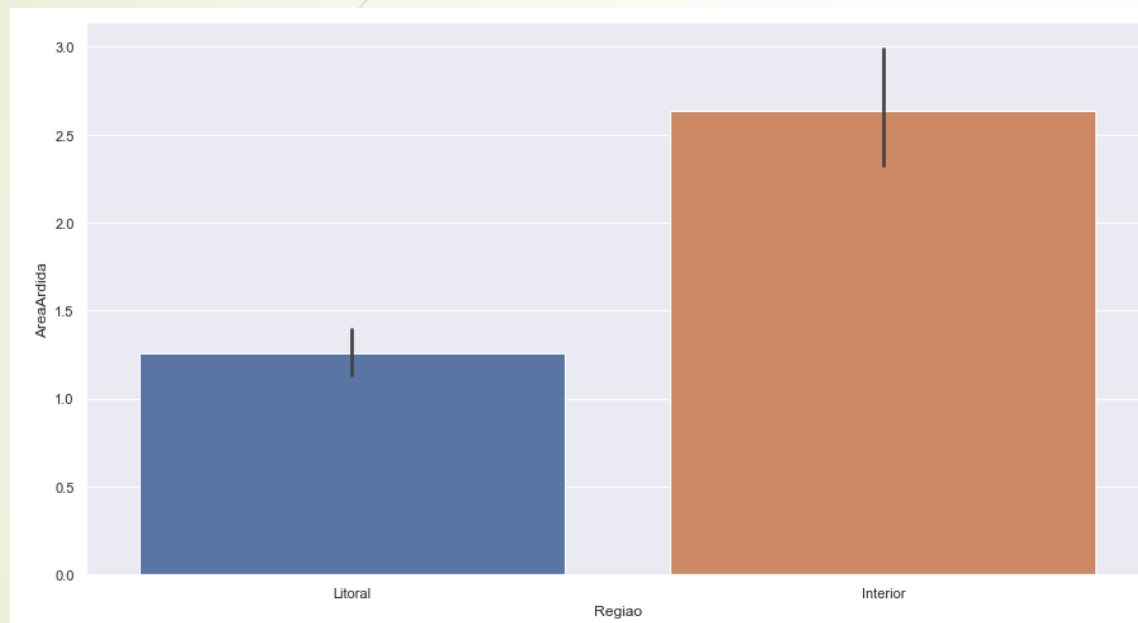- Max: 59min
- Count: 15453
- Mean: 13min

- **Bar chart** for the number of occurrences per region;
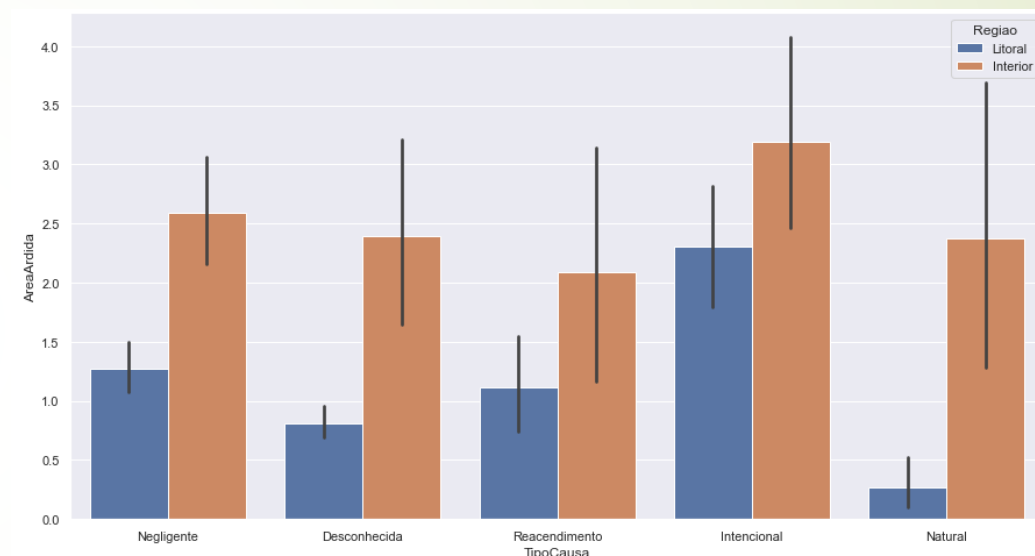- What **stands out**? Littoral with many more occurrences (70%/30% Ratio - approx.)

- **Bar chart** for the causes;
- **Negligence** and **Unknown** causes stand out (70%aprox);
- **Intentional** cause "in between" (almost 20%);
- **Natural** and **Reignition** are the lowest(10%aprox)
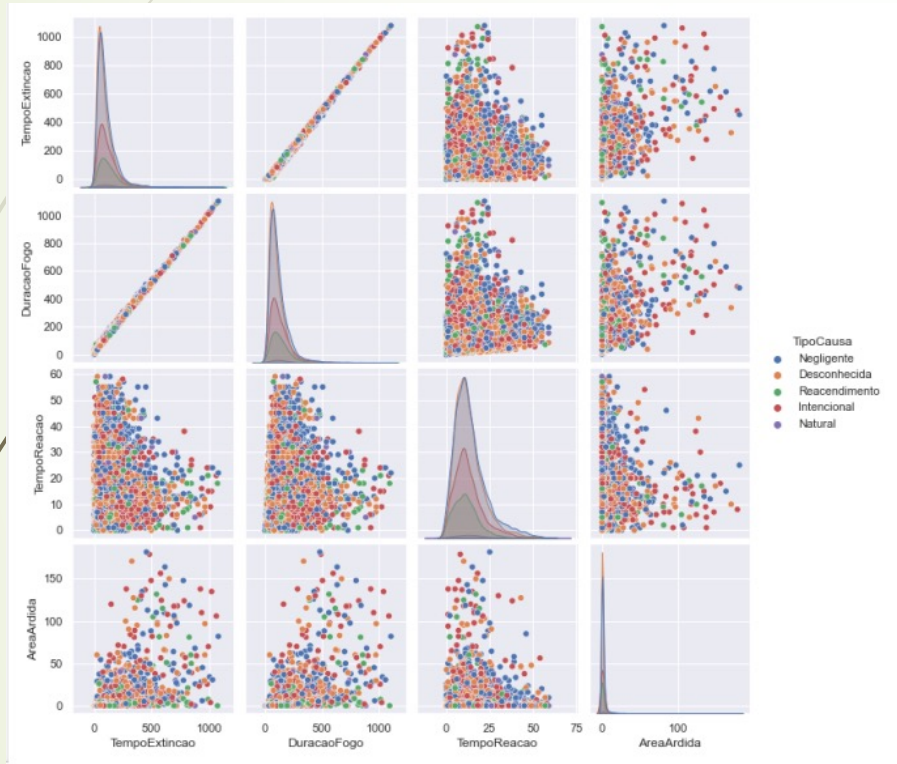
- **Bar chart** for the average burned area per region;
- What **stands out**? Interior with a wider burned area on average.
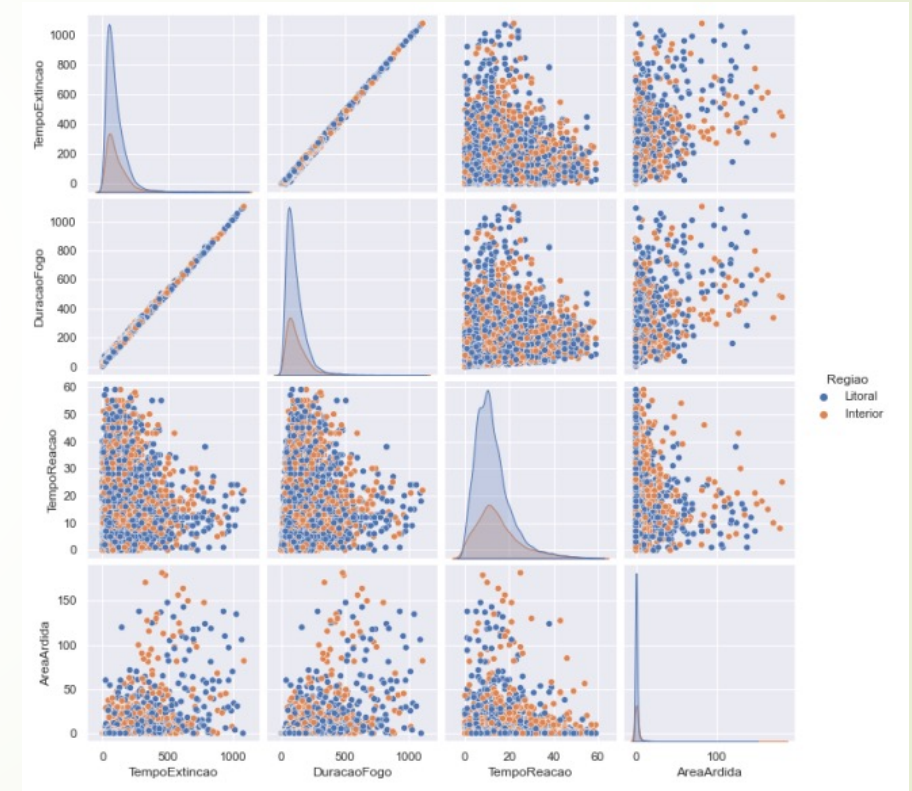
- **Bar chart** for the causes and burned area, split into littoral and interior;
- What **stands out?** The average burned area is higher in the interior than it is on the littoral;

Step 3 – Exploring the data
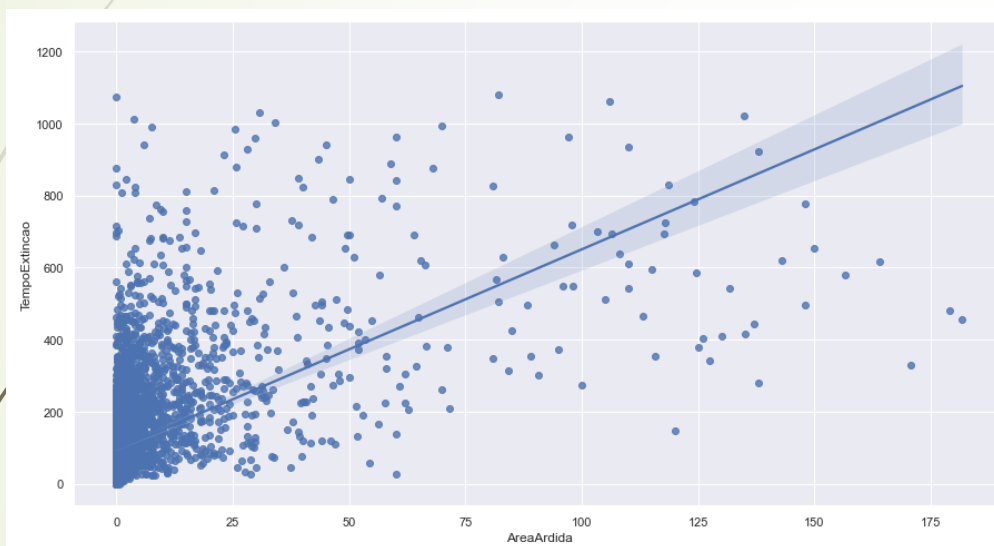# Multivariate Analysis - Corrrelations
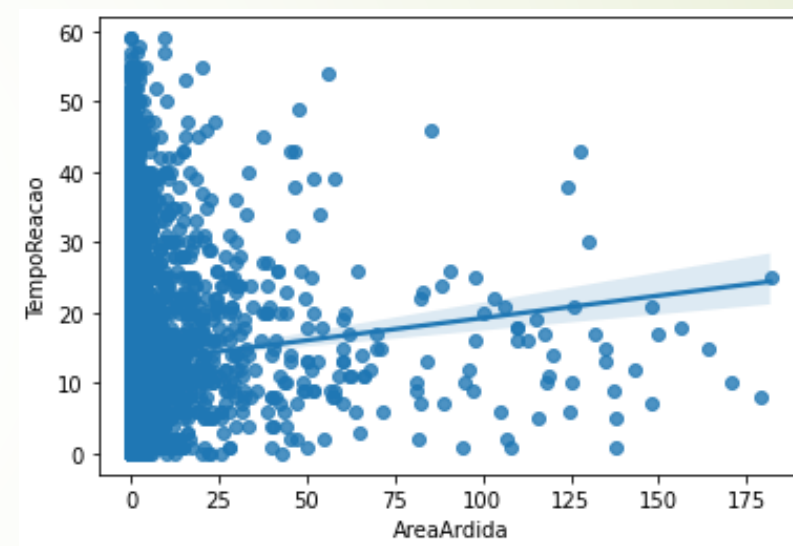

Correlations with **type of cause**

Correlations with **the Regions**

Step 3 – Exploring the data
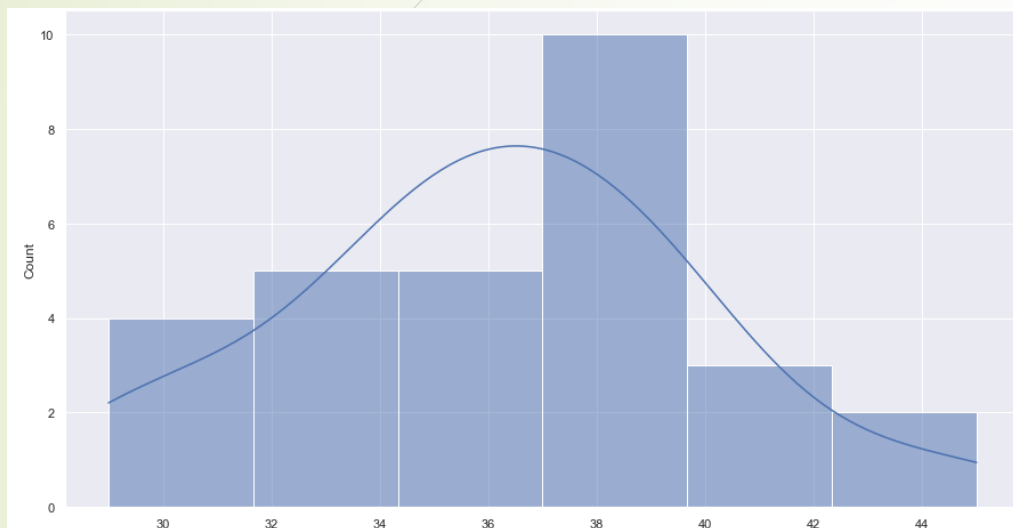# Multivariate Analysis - Correlations

Scatter plot that correlates the burned area and the extinguishing time

Scatter plot that correlates the burned area and the Reaction time

Step 4 – Draw inferences

1st Inference - Can we claim with 99% certainty that **more than 71% of the fire outbreaks** with an area below **10ha occur on the littoral**?

- Histogram for the fire outbreak occurrences on the littoral ;
- 29 samples, 50 rows each sample;
- Ho: ʋ ≥ 0,71 vs. H1: ʋ < 0,71;
- Confidence interval = ((70,74;74,48)

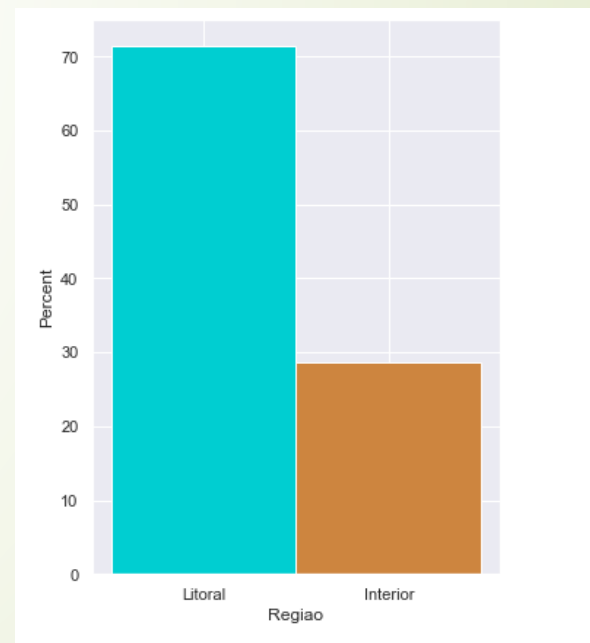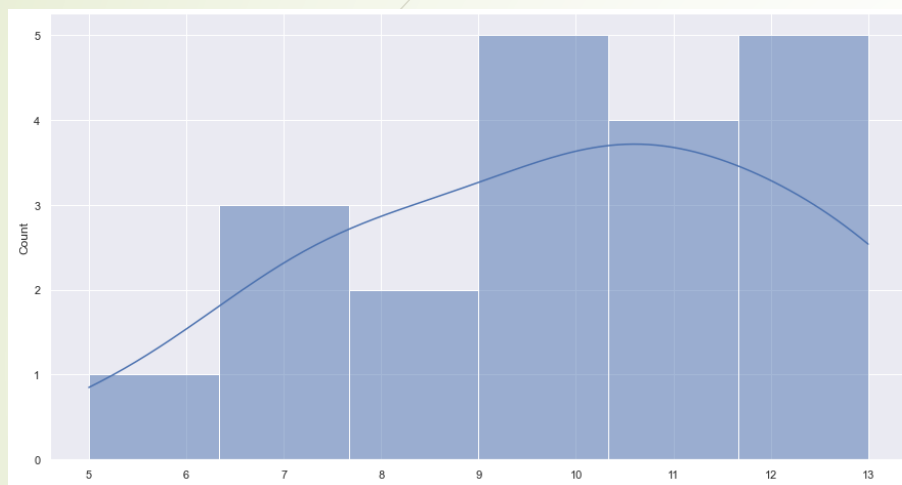

Conclusion: Since 71% is within the confidence interval, we can claim with 99% certainty that **more than 71% of the fire outbreaks** with a **burned area below 10 ha** occur on the **littoral**.

Step 4 – Draw inferences

2nd Inference - Can we claim with 99% certainty that **more than 35% of the fires** with a burned area above **100ha** have an **intentional cause behind**?

- ➡ Histogram for the number of occurrences intentionally out of 30 rows 20 samples;

- ➡ Ho: υ ≥ 0,35 vs. H1: υ < 0,35

- ➡ Confidence interval = (34,8;37,2)

Conclusion: Since 35% is within the confidence interval, we have evidence to state that more than 35% of the fires with a burned area above 100ha have an intentional cause behind.
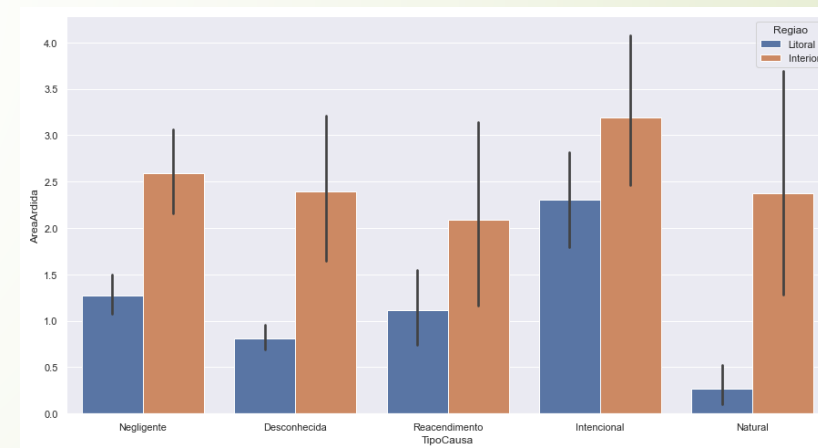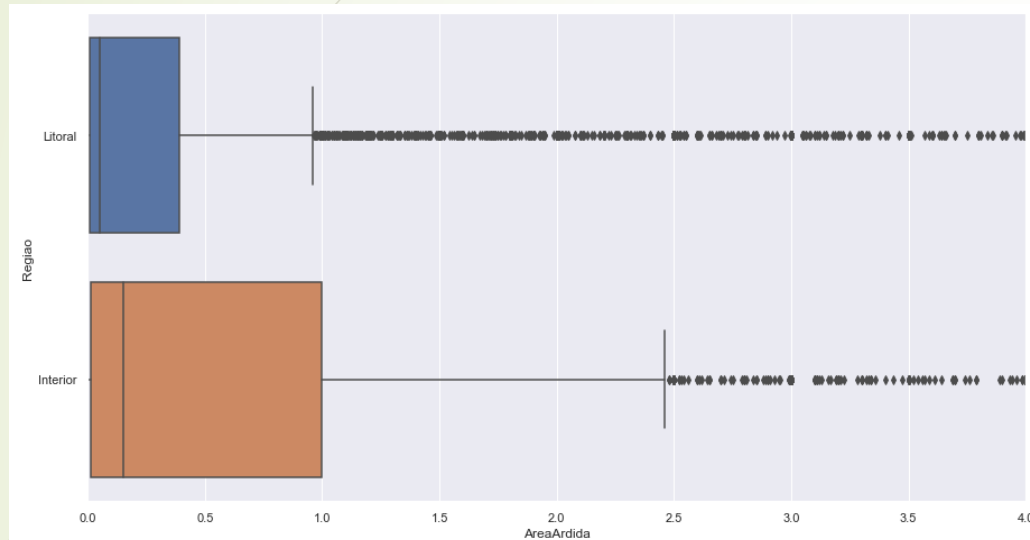
Step 4 – Draw inferences

3rd Inference - Can we claim with 99% certainty that **the mean difference of the burned area** in the **interior** is superior to the burned area on the **coast** ?

- ➡ **Boxplot** for the burned area, both on the littoral and in the interior;

- ➡ $H0: \mu1 \geq \mu2 \quad H0: \mu1 - \mu2 \geq 0$

- ➡ $H1: \mu1 < \mu2 \quad H1: \mu1 - \mu2 < 0$

significance Level (a)	0,01
Pooled Variance	25771,06358
T value	2,588238405
Margin of error	0,182310038
99% lower bound	2,537689962
99% upper bound	2,902310038

Conclusion: The results suggest that there is strong evidence that **the Mean difference** of the burned area in the **interior** will be **2,54 to 2,9 (ha)** higher than the burned area on the coast.

Step 4 – Draw inferences

4th Inference - Can we claim with 99% certainty that **the mean difference of the burned area** that has a **natural cause** behind is superior to the one with **an intentional cause behind**?
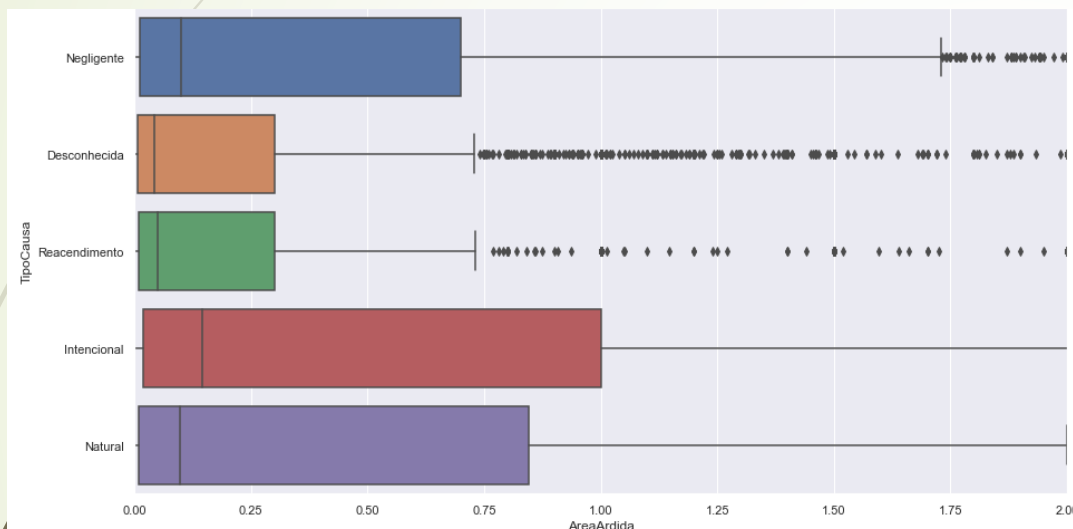
➡ **Boxplot** for the burned area, and the causes behind;

➡ $H0: \mu1 \geq \mu2 \quad H0: \mu1 - \mu2 \geq 0$

➡ $H1: \mu1 < \mu2 \quad H1: \mu1 - \mu2 < 0$

|	
confidence Level	99%
significance Level (a)	0,01
Pooled Variance	4222,61
T value	2,600887278
Margin of error	2,664139735
99% lower bound	-2,124139735
99% upper bound	3,204139735

❌ **Conclusion**: We cannot claim with 99% certainty that **the mean difference** of the burned area caused **naturally** will be higher than the one caused **intentionally,** since the mean difference can vary from (-2,12; 3,2).
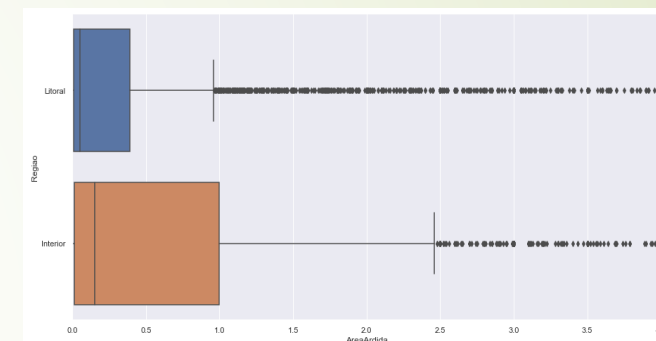
Step 5 – Conclusions (all in all)
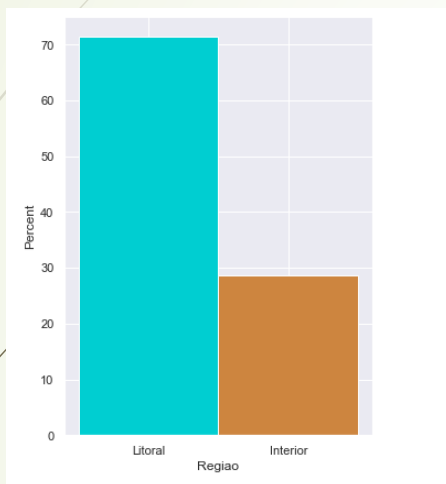
More than 70% of the **fire outbreaks** with a burned area below 10ha occur on the **littoral**
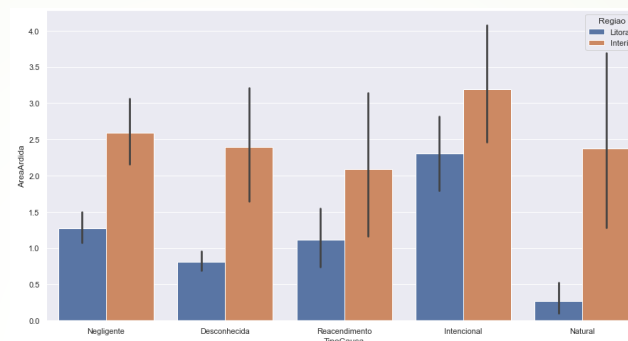
More than 35% of the fires with a burned area above **100ha** have an **intentional** cause behind

The **interior** has wider burned areas than the **littoral**, **on average**. 2,54 to 2,9 ha more

Step 6 – Look back and ahead

- Project had a small scale, couldn't study all variables

- Results could be different if we considered outliers

- Other interesting **variables** that could be **studied**?

    Humidity, wind, temperature, vegetation density, terrain slope, proximity to fire department,…

- Other interesting studies? **Evolution** through the years (2022-2025)

END