



Escola Politécnica de Pernambuco

Especialização em Ciência de Dados e Analytics

Introdução à Ciência de Dados

Aula 5


Prof. Dr. Alexandre Maciel
alexandre.maciel@upe.br

ANÁLISE DESCRITIVA DE DADOS





- Não é Mineração de Dados, apenas descreve os dados.
- Resulta na descoberta de padrões consistentes.
- Univariadas ou bivariadas.

BASE DE DADOS EXEMPLO

**UC Irvine**
Machine Learning
Repository

Datasets Contribute Dataset ▾ About

 [Login](#)





Mammographic Mass


Donated on 10/28/2007


Discrimination of benign and malignant mammographic masses based on BI-RADS attributes and the patient's age.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate	Life	Classification
Attribute Type	# Instances	# Attributes
Integer	961	6


 **DOWNLOAD**

 **CITE**

 3 citations

 1877 views

Creators

 Matthias Elter


DOI

[10.24432/C53K6Z](https://doi.org/10.24432/C53K6Z)

License

This dataset is licensed under a [Creative Commons Attribution](#)

Information



Additional Information

Mammography is the most effective method for breast cancer screening available today. However, the low positive predictive value of breast biopsy resulting from mammogram interpretation leads to approximately...

[SHOW MORE](#) ▾

DESCRIÇÃO DA BASE

Number of Instances: 961

Number of Attributes: 6 (1 goal field, 1 non-predictive, 4 predictive attributes)

Attribute Information:

1. BI-RADS assessment: 1 to 5 (ordinal, non-predictive)
2. Age: patient's age in years (integer)
3. Shape: mass shape: round=1 oval=2 lobular=3 irregular=4 (nominal)
4. Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5 (nominal)
5. Density: mass density high=1 iso=2 low=3 fat-containing=4 (ordinal)
6. Severity: benign=0 or malignant=1 (binominal)

Missing Attribute Values: Yes

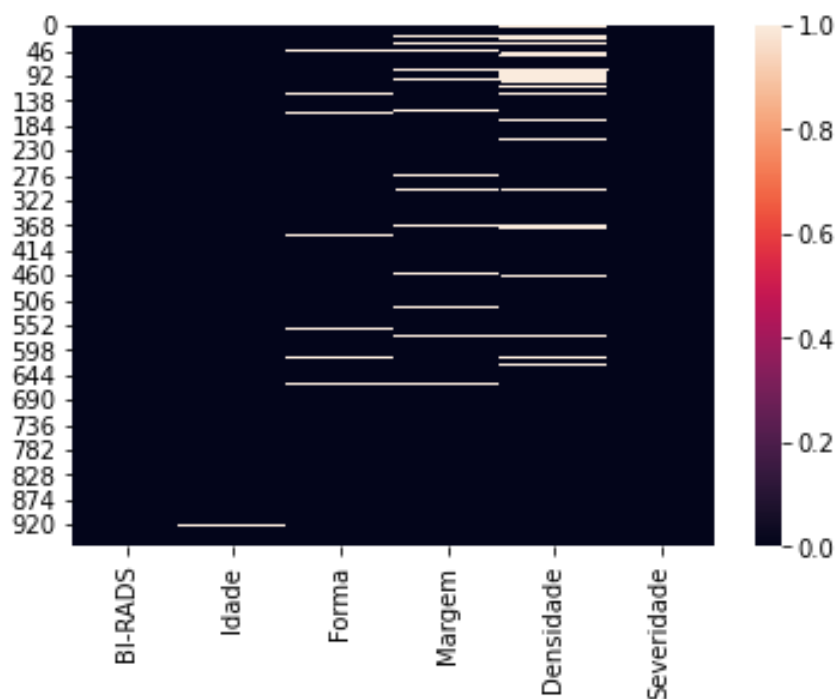
- BI-RADS assessment: 2
- Age: 5
- Shape: 31
- Margin: 48
- Density: 76
- Severity: 0

Class Distribution: benign: 516; malignant: 445

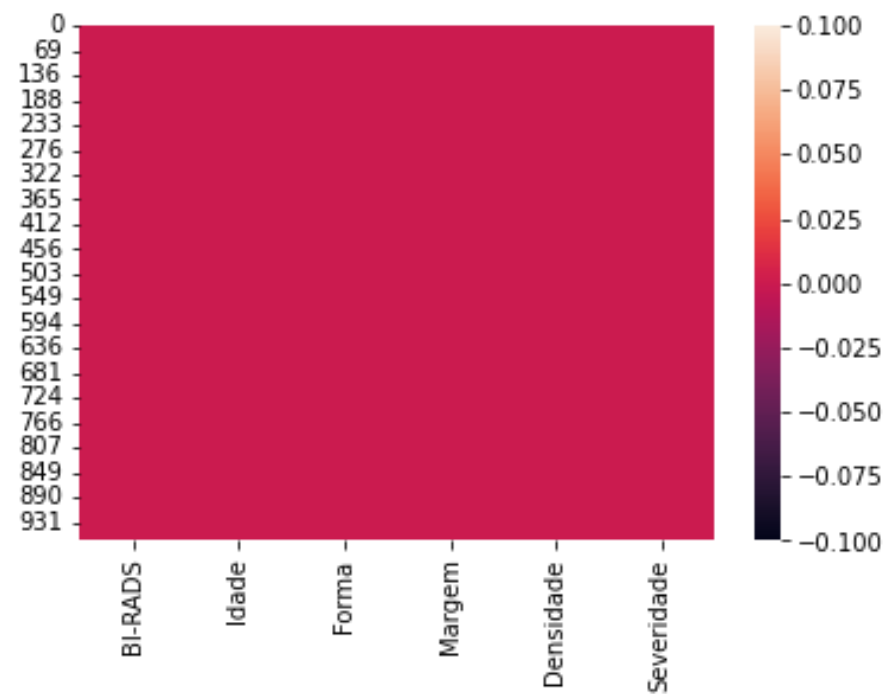
HEAD DA BASE DE DADOS

	BI-RADS	Idade	Forma	Margem	Densidade	Severidade
0	5.0	67.0	3.0	5.0	3.0	1
1	4.0	43.0	1.0	1.0	NaN	1
2	5.0	58.0	4.0	5.0	3.0	1
3	4.0	28.0	1.0	1.0	3.0	0
4	5.0	74.0	1.0	5.0	NaN	1
5	4.0	65.0	1.0	NaN	3.0	0
6	4.0	70.0	NaN	NaN	3.0	0
7	5.0	42.0	1.0	NaN	3.0	0
8	5.0	57.0	1.0	5.0	3.0	1
9	5.0	60.0	NaN	5.0	1.0	1

DADOS AUSENTES



961 instâncias



830 instâncias

MATRIZ DE CORRELAÇÃO

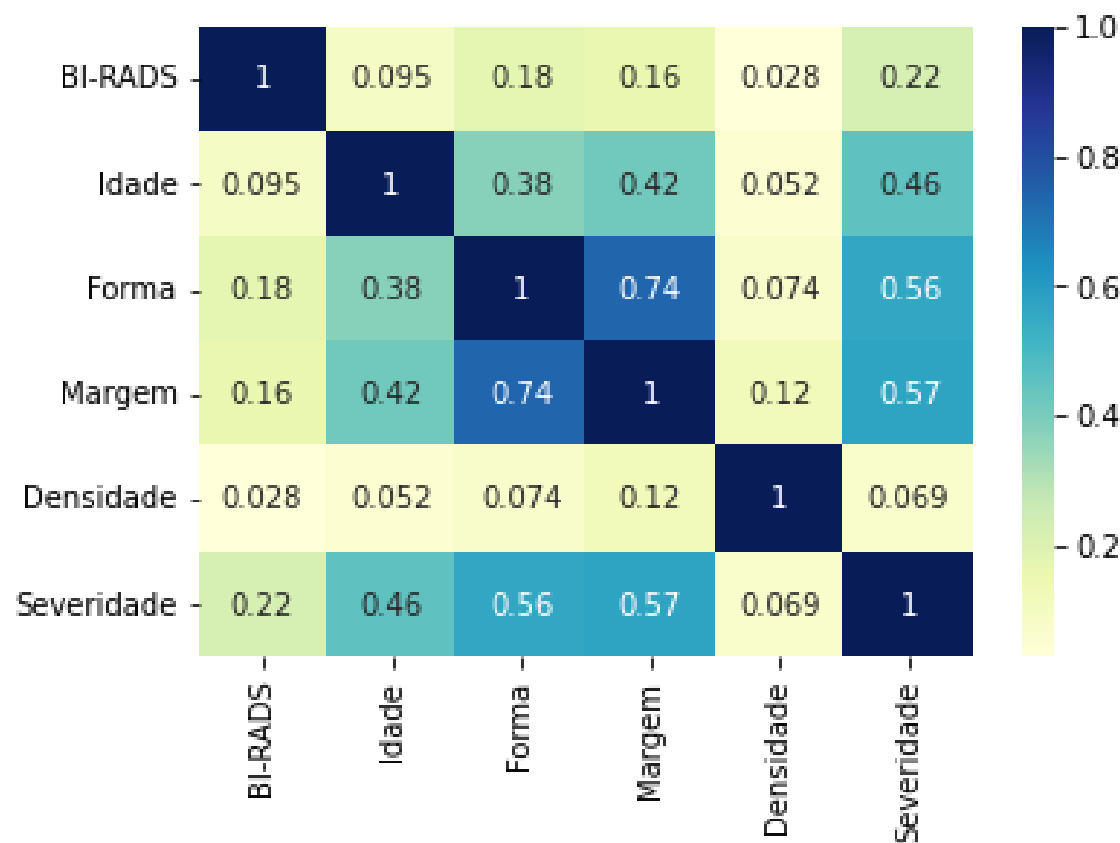
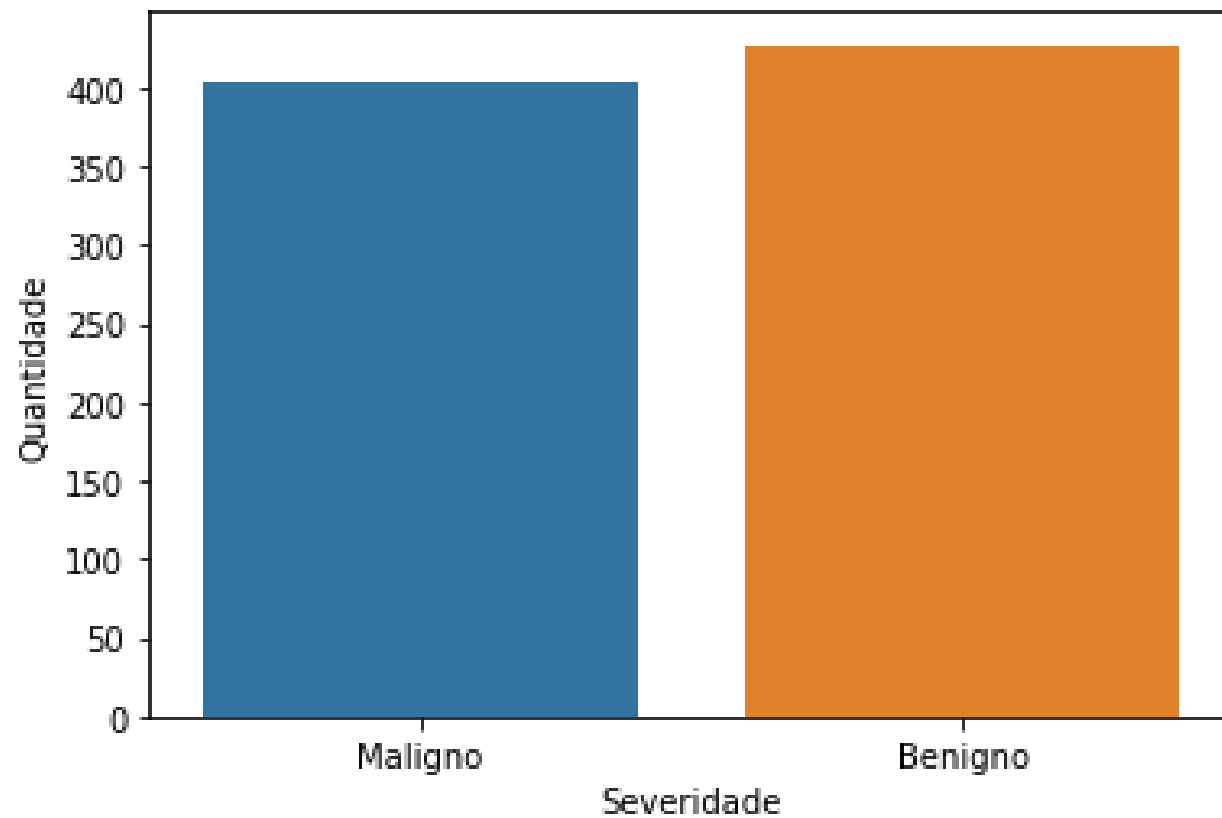
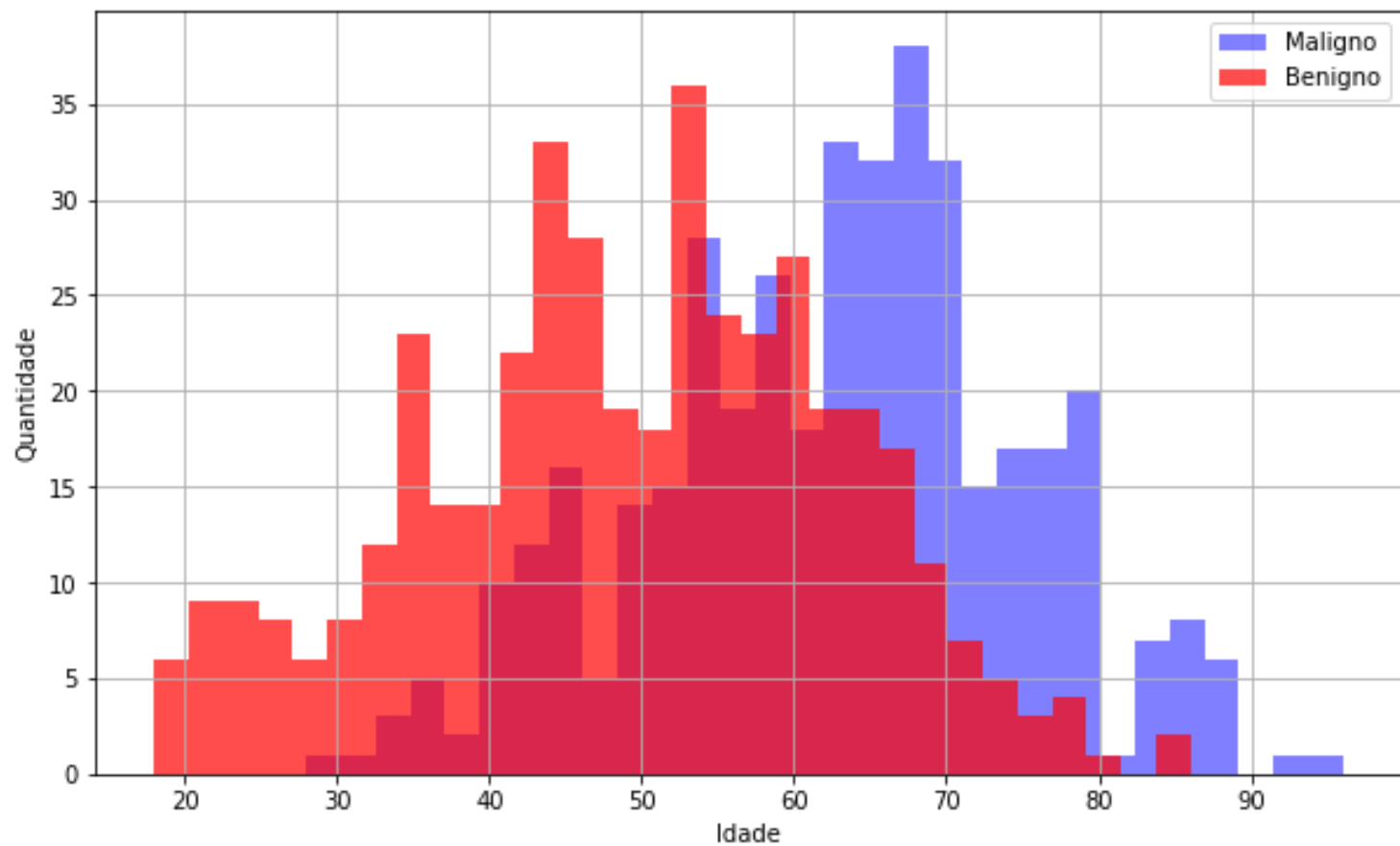


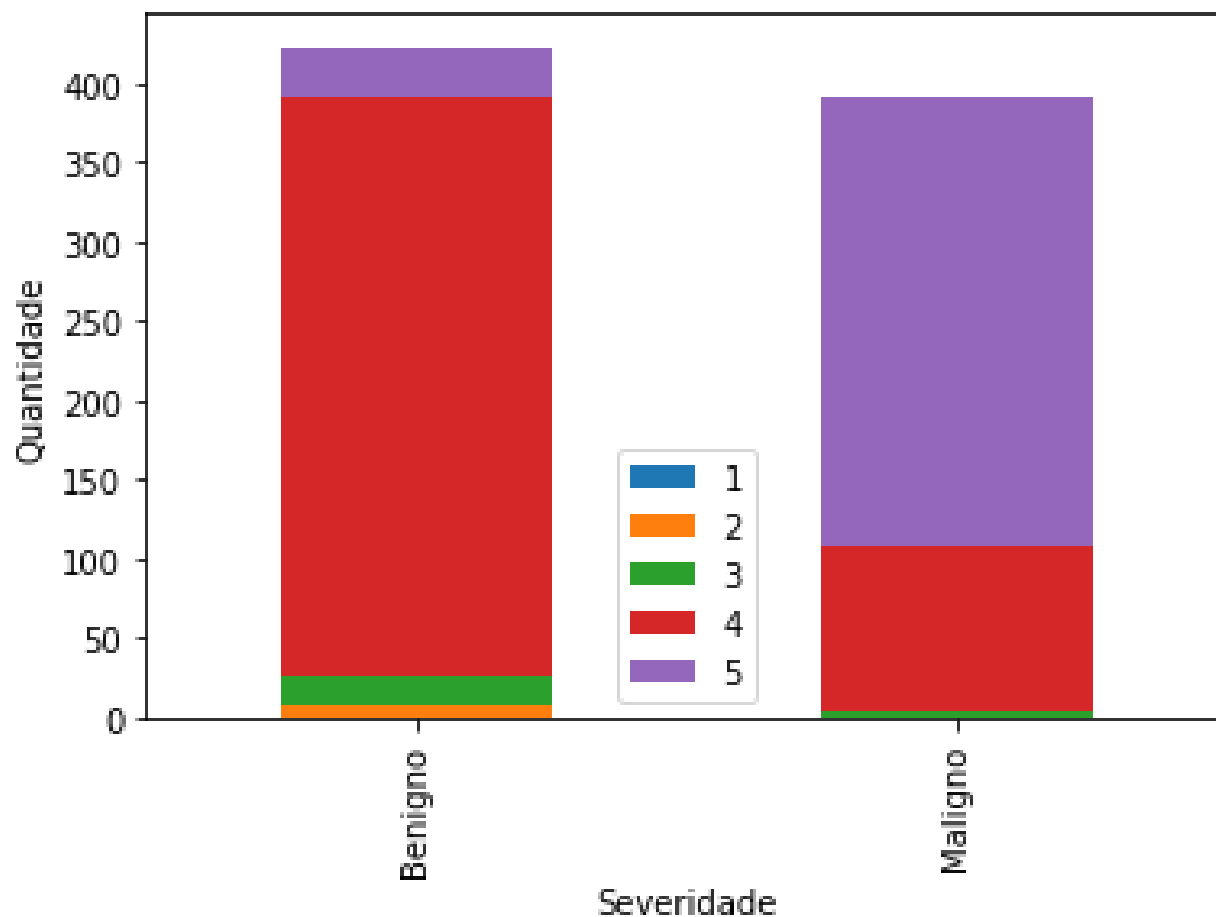
GRÁFICO DE COLUNAS



HISTOGRAMA



COLUNAS EMPILHADA



DISTRIBUIÇÃO DE FREQUÊNCIA

Limite inferior	Limite superior	Frequência absoluta	Frequência relativa	Frequência acumulada	
18	34	68	8,19%	68	8,19%
35	46	164	19,75%	232	27,95%
47	58	223	26,86%	455	54,81%
59	79	337	40,60%	792	95,42%
80	92	38	4,33%	830	100%

GRÁFICO DE PARETO

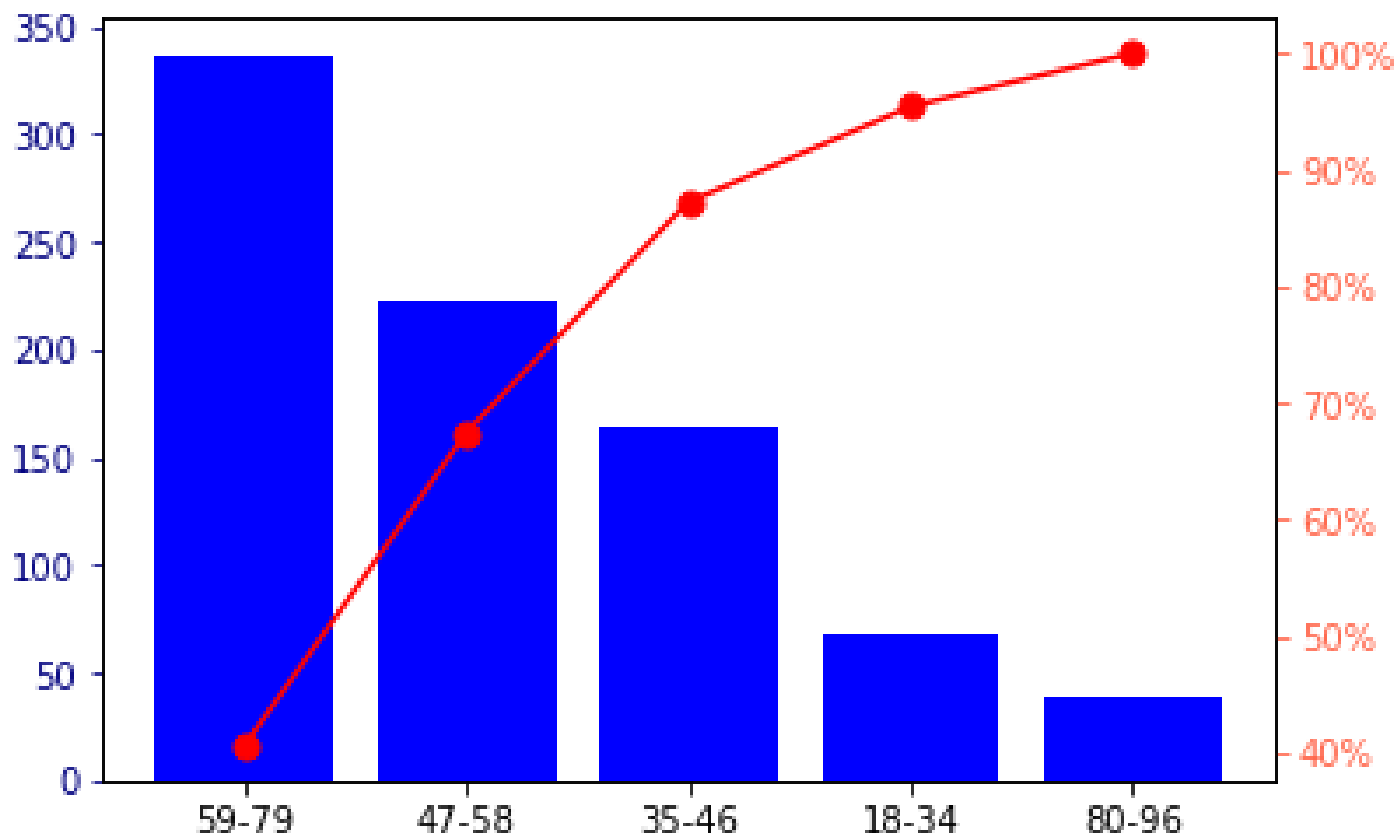


GRÁFICO DE LINHA

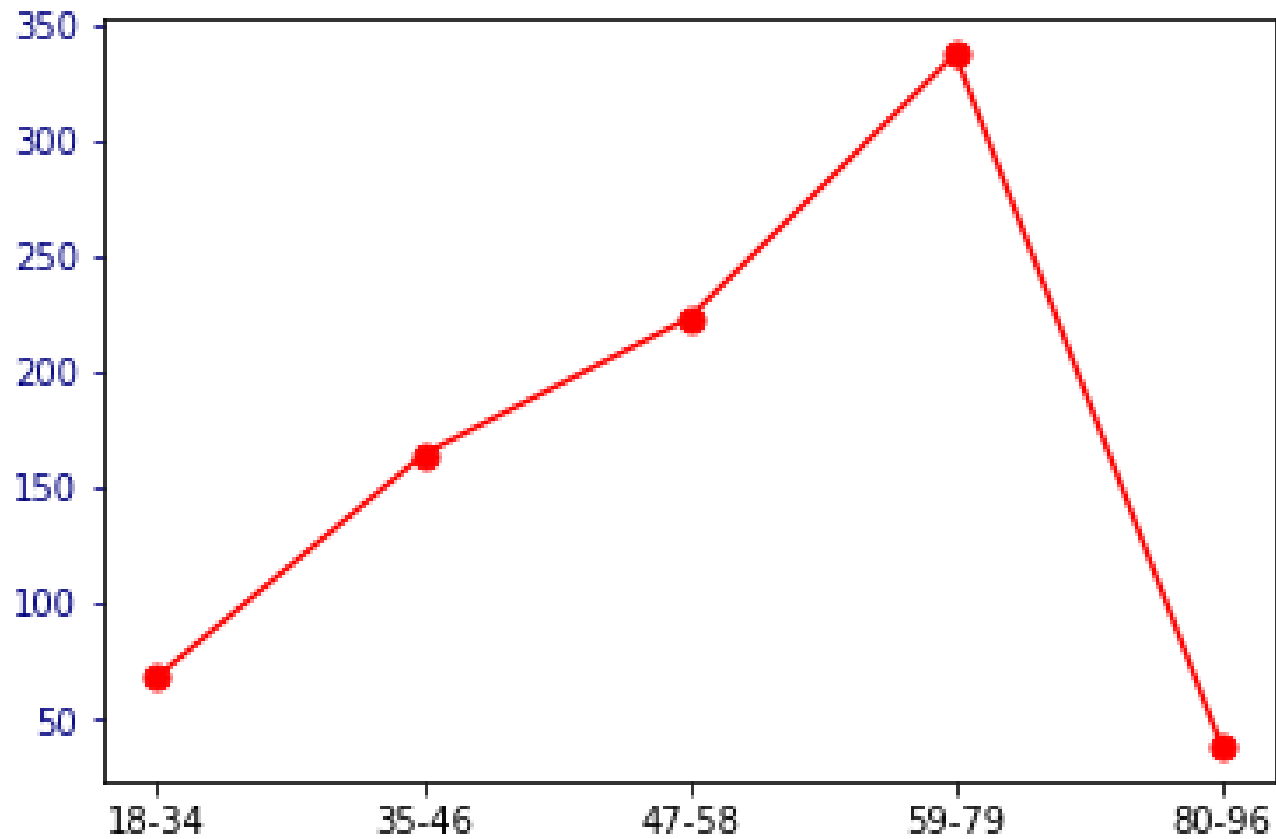


GRÁFICO DE SETORES

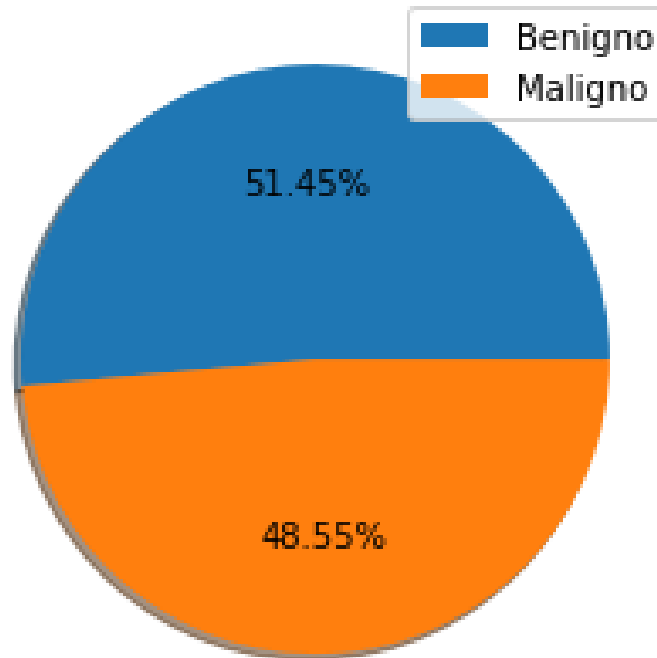
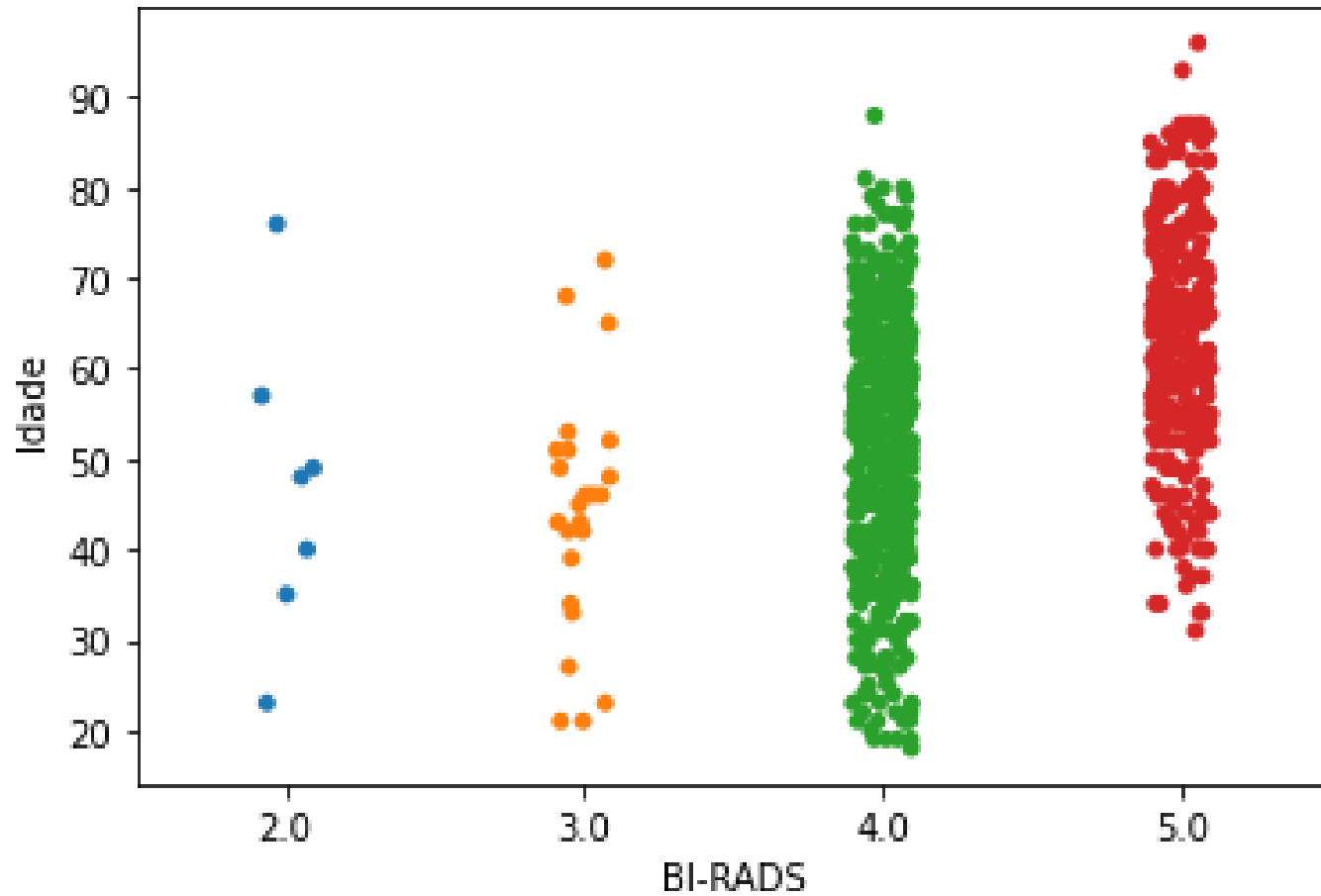


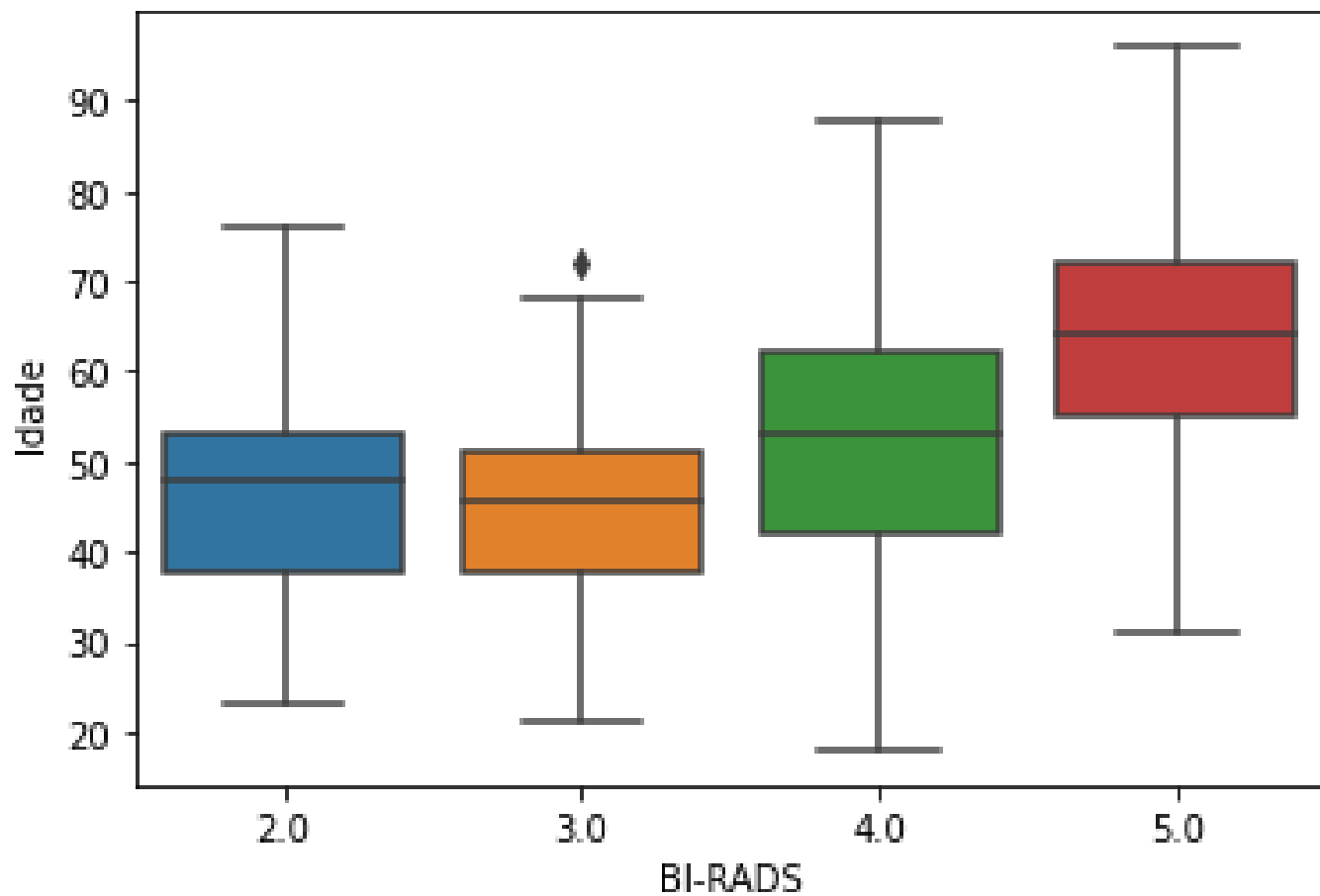
GRÁFICO DE DISPERSÃO



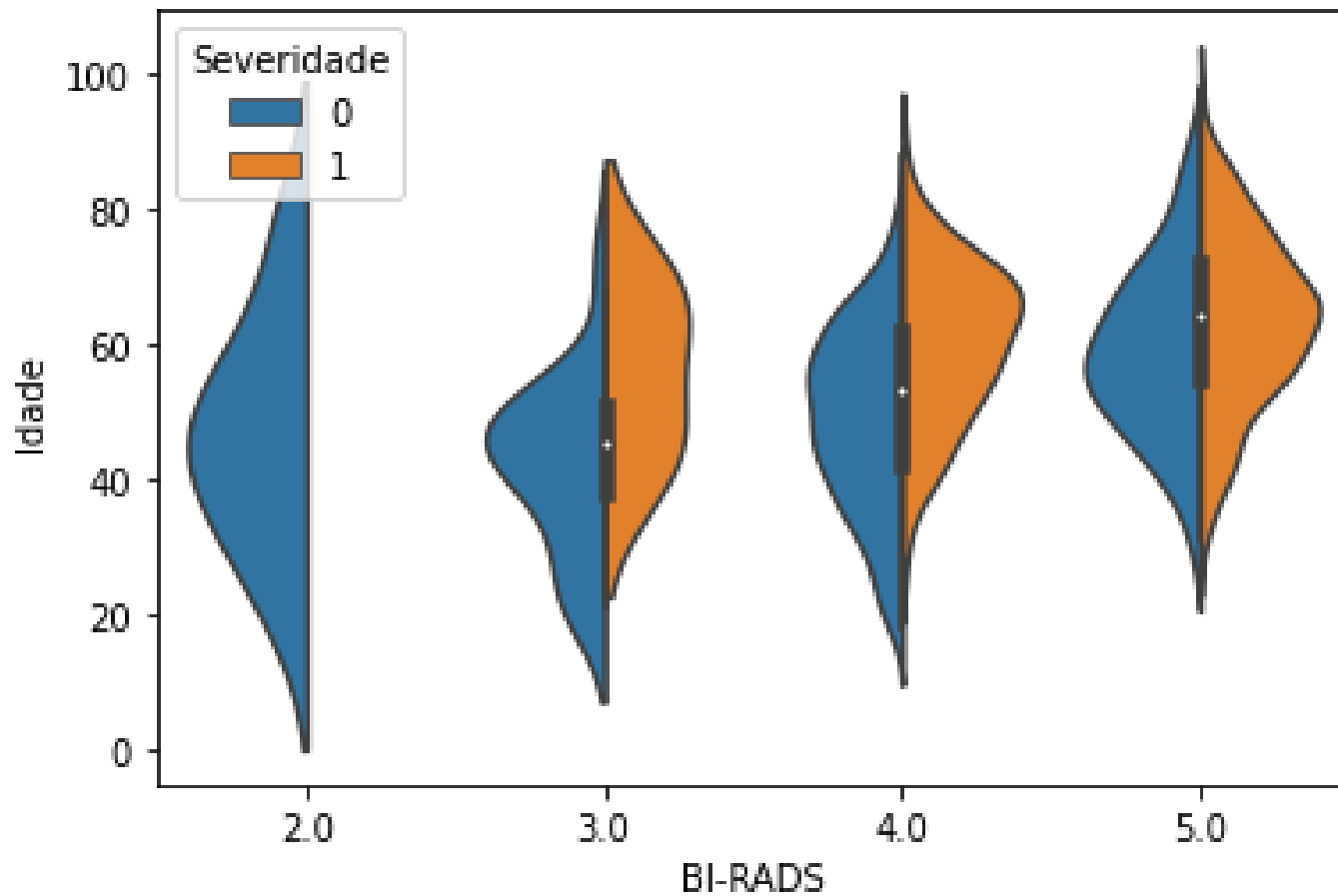
MEDIDAS DE RESUMO

	BI-RADS	Idade	Forma	Margem	Densidade	Severidade
count	830.000000	830.000000	830.000000	830.000000	830.000000	830.000000
mean	4.393976	55.781928	2.781928	2.813253	2.915663	0.485542
std	1.888371	14.671782	1.242361	1.567175	0.350936	0.500092
min	0.000000	18.000000	1.000000	1.000000	1.000000	0.000000
25%	4.000000	46.000000	2.000000	1.000000	3.000000	0.000000
50%	4.000000	57.000000	3.000000	3.000000	3.000000	0.000000
75%	5.000000	66.000000	4.000000	4.000000	3.000000	1.000000
max	55.000000	96.000000	4.000000	5.000000	4.000000	1.000000

BOX PLOT



VIOLIN PLOT



ATIVIDADE 5

1. Escolha um atributo numérico e um nominal (classe?) da base de dados do seu projeto.
2. Elabore:
Histograma, gráfico de dispersão e boxplot.
3. Entregue os slides (+3) de sua apresentação.
4. Faça o upload no Google Classroom.