Minicurso de Ciência de Dados

Aula 6 - Tratamento de Dados

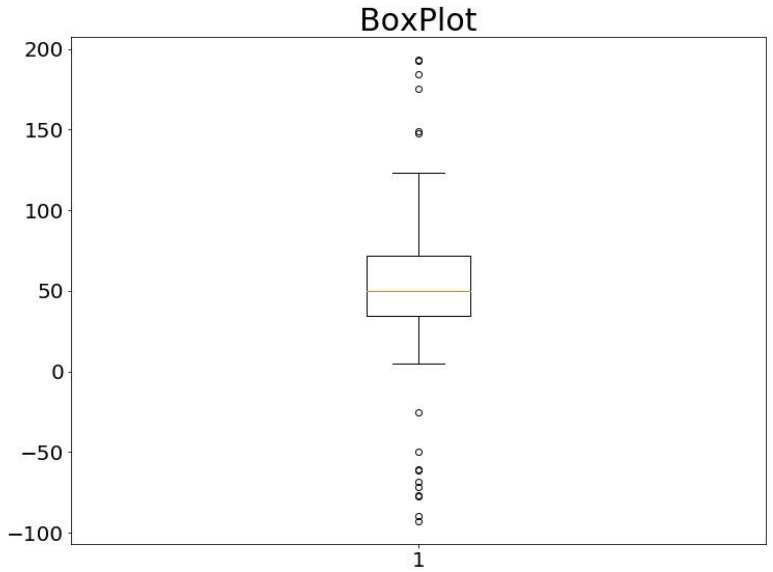
Kally Chung

5 de Fevereiro de 2020

I CiDWeek

Outlier

- São dados que se diferenciam dos outros e podem ser encontrados por estarem distante dos demais dados. Uma das formas da identificação dos outliers pode ocorrer através da visualização de dados (boxplot, gráfico de dispersão).



Outlier

- O que fazer com dados atípicos?
  - Opção 1: Interpretar o outlier perguntando para a fonte.

    Exemplo: Quando se trabalha com *Fonte de Renda*

  - Opção 2: Se livrar dos outliers

    Exemplo: Quando se trabalha com *Ano*

- Como encontrar outliers?
  - Conheça seus dados.

Em qual ano você concluiu sua graduação?
2014
2016
2013
2018
2015
2017
2016
2017
2008
204
2008
2012
2018
2000
2015

Conhecendo seus dados

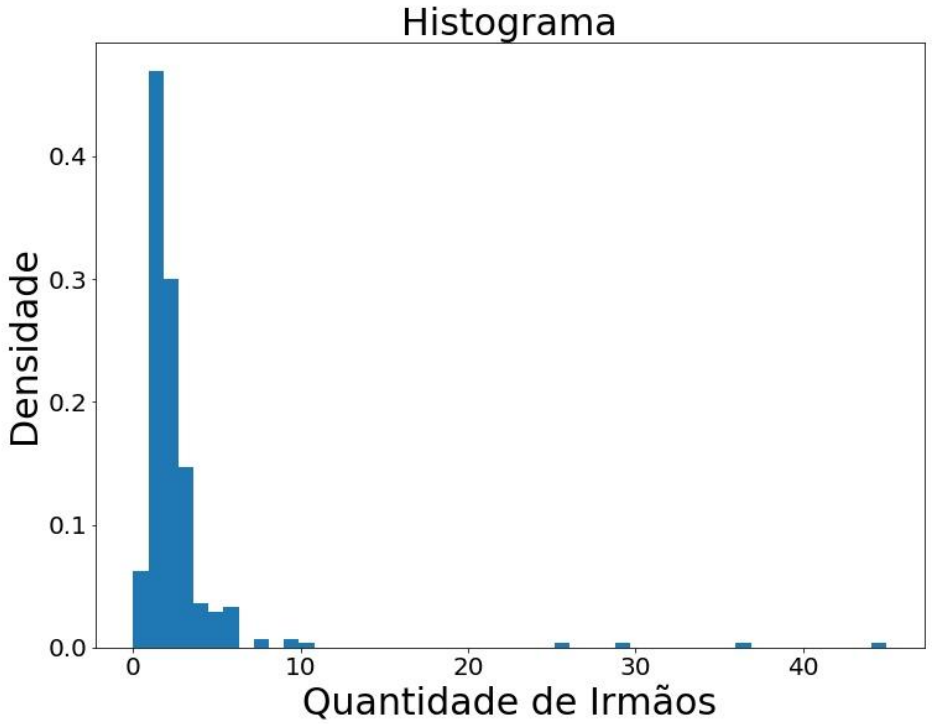
Dados Unidimensionais

- Faça a descritiva dos dados, ou seja
    - Quantas observações existem?
    - Qual o menor/maior dado?
    - Qual a média e o desvio padrão?

- Plote o histograma dos dados

- Exemplo: Quantidade de irmãos
  - 303 observações
  - Nenhum dado faltante

```
Média =  2.396039603960396
D.Padrão =  4.064250233276887
Mediana =  2
Moda =  1
Min =  0
Max =  45
```



Conhecendo seus dados

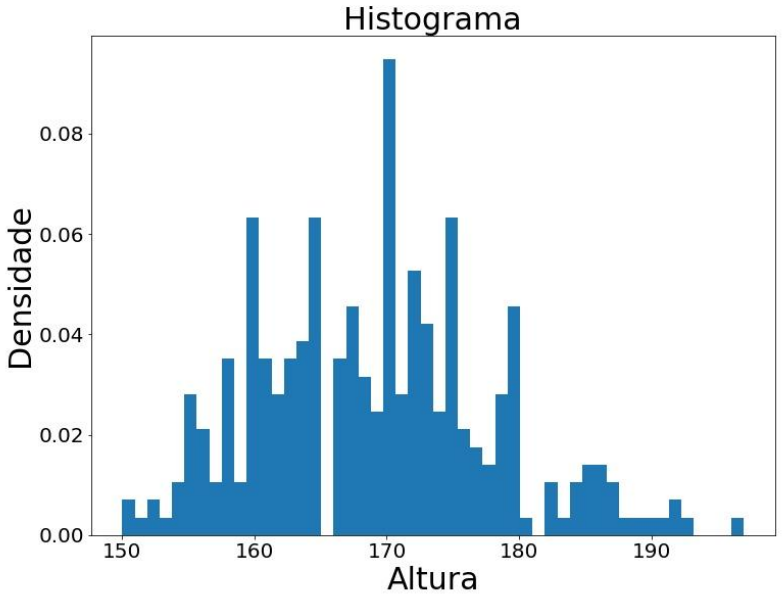
Dados Bidimensionais

- Faça a descritiva de cada dado, ou seja
    - Quantas observações existem?
    - Qual o menor/maior dado?
    - Qual a média e o desvio padrão?

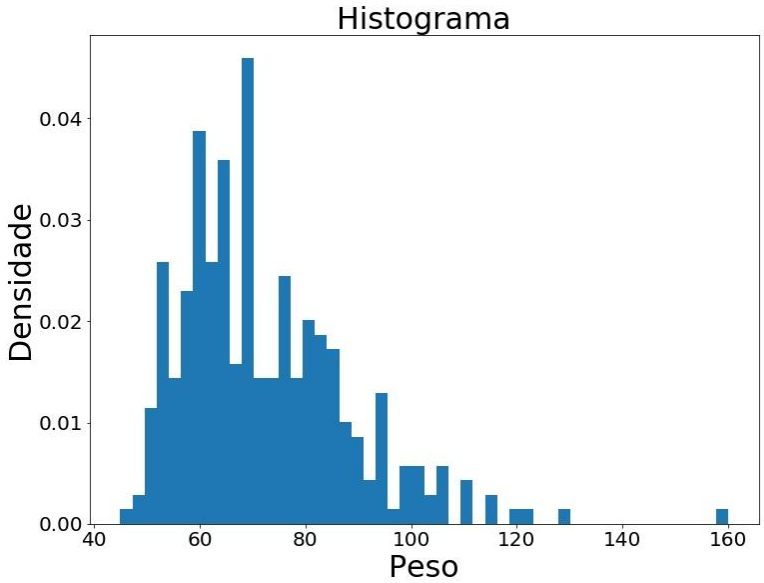
- Plote o gráfico de dispersão

- Exemplo: Peso (cm) e altura (kg)

```
Altura
------
Média =  169.0957095709571
D.Padrão =  8.903125506411854
Mediana =  169
Moda =  170
Min =  150
Max =  197
```




```
Peso
----
Média =  72.3003300330033
D.Padrão =  16.050383210852658
Mediana =  69
Moda =  60
Min =  45
Max =  160
```



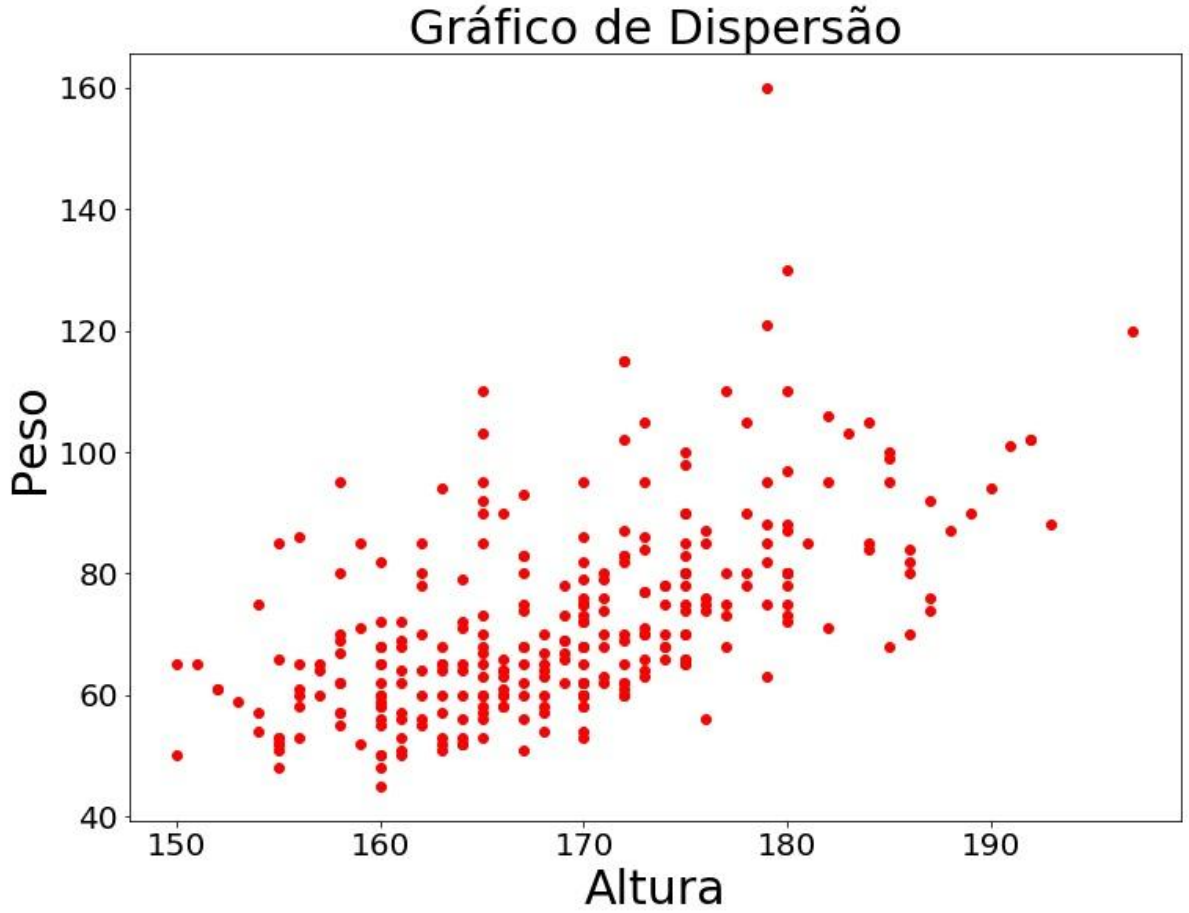


Gráfico de Dispersão
Peso
Altura
160
140
120
100
80
60
40
150
160
170
180
190

CiDAMO

Conhecendo seus dados

Correlação entre duas variáveis

- O coeficiente de correlação entre X e Y é uma medida para o grau de associação da relação linear entre as variáveis X e Y.
- O valor está sempre -1 e 1, em que r = 0 indica a ausência de associação.
- Quando r > 0, tem-se a correlação positiva, que significa que à medida que a variável X cresce, variável Y também cresce.
- Quando r < 0, tem-se a correlação negativa, que significa que à medida que X cresce, Y decresce.

Conhecendo seus dados

Correlação entre duas variáveis

Valor de r (+ ou -)	Interpretação
0 a 0,19	Correlação bem fraca
0,20 a 0,39	Correlação fraca
0,40 a 0,69	Correlação moderada
0,70 a 0,89	Correlação forte
0,90 a 1,00	Correlação muito forte

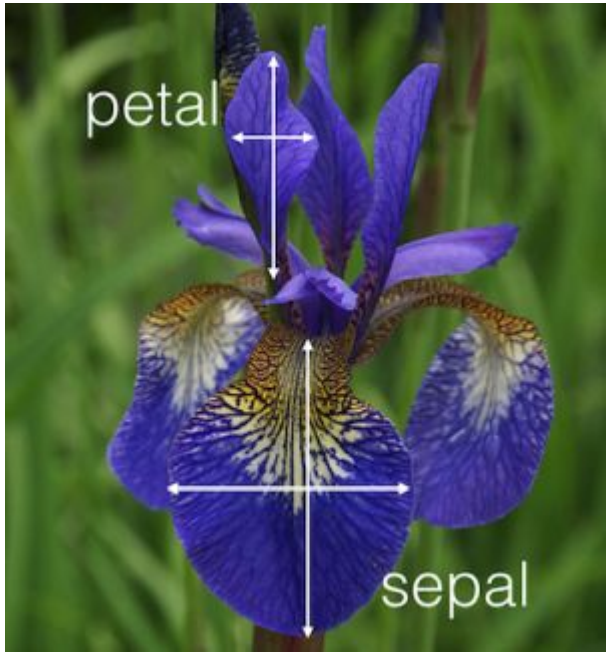
http://leg.ufpr.br/~paulojus/CE003/ce003/node8.html

Conhecendo seus dados

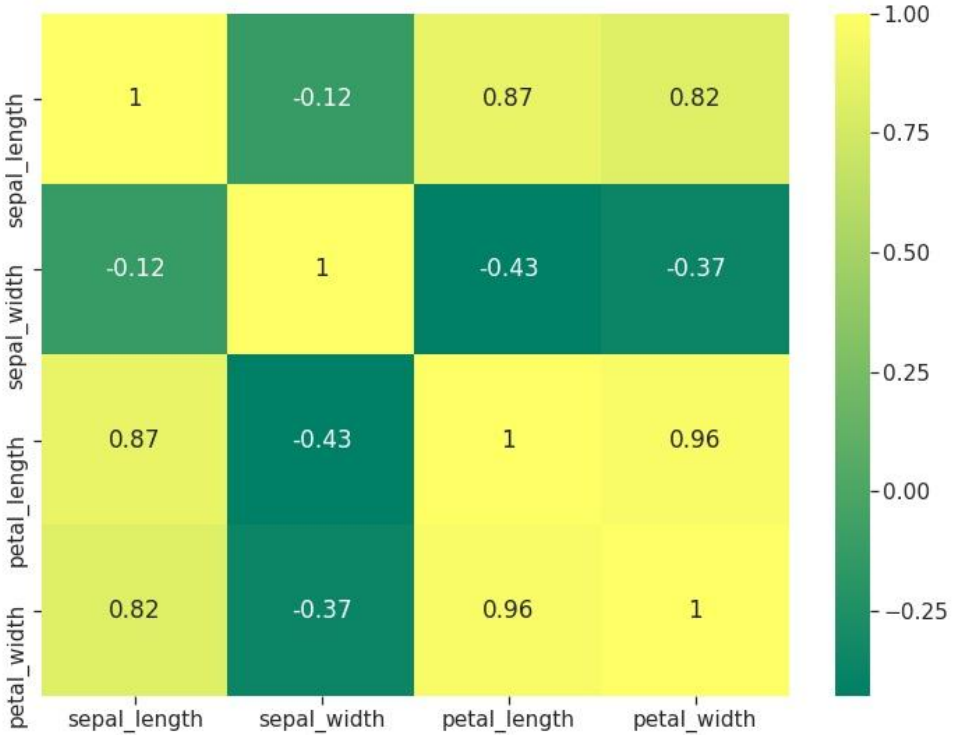
Dados Multidimensionais

- Analize a matriz de correlação
- Plote a matriz de gráfico de dispersão

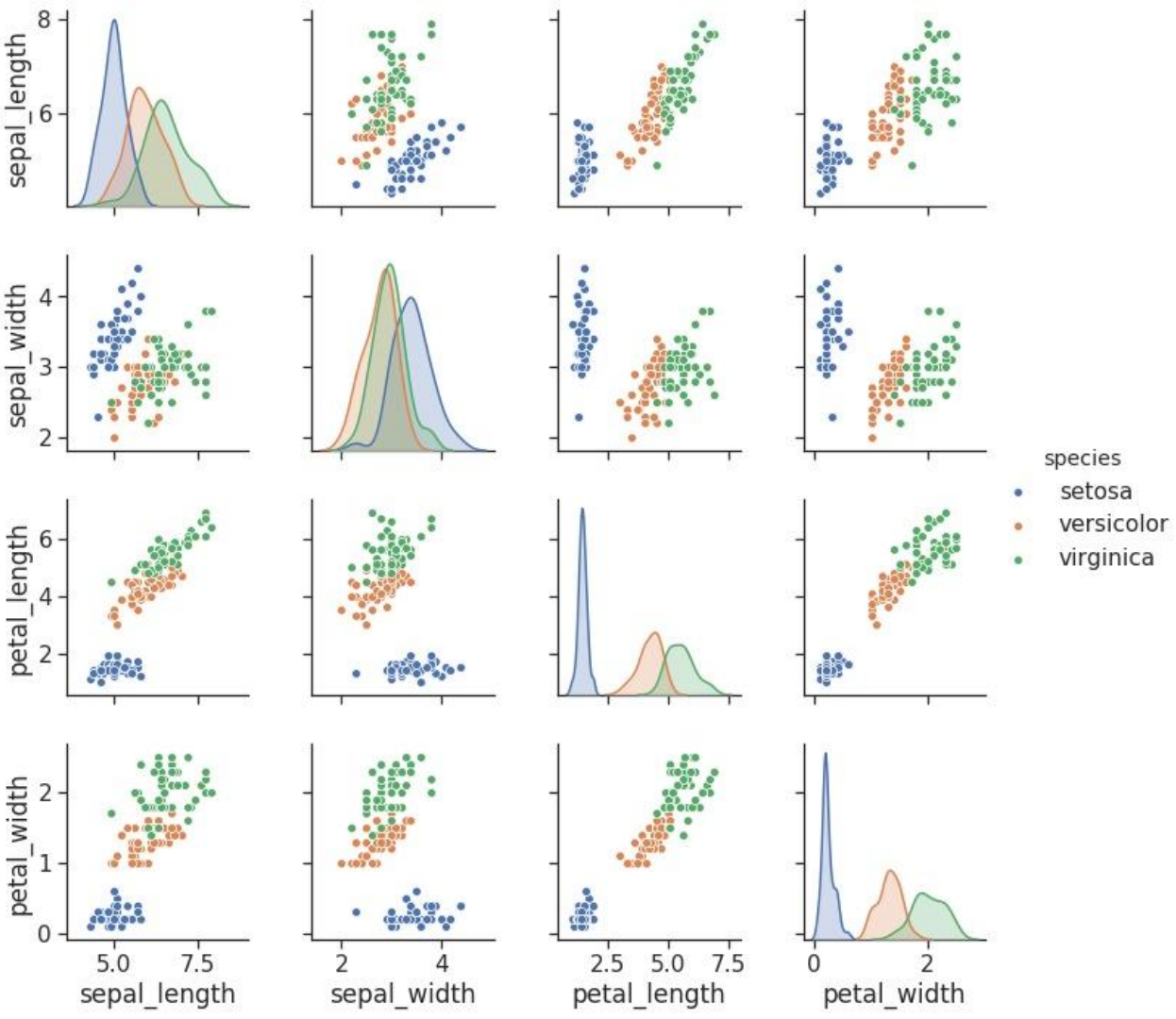
- Exemplo: conjunto de dados Iris
  - 50 observações de 3 espécies das flores: Iris setosa, Iris virginica e Iris versicolor
  - Cada observação contém a largura e o comprimento da pétala e sépala



CiDAMO



Matriz de correlação dos variáveis do conjunto de dados Iris.



CiDAMO

Matriz de gráfico de dispersão do conjunto de dados do Iris

Dados faltantes

CiDAMO

- Delete linhas
- Substitua pela média, mediana e/ou moda
- Substitua por uma amostra aleatória
- Defina uma nova categoria
- Faça previsão dos valores faltantes
- Use métodos que funcionam com missing data

Dados faltantes

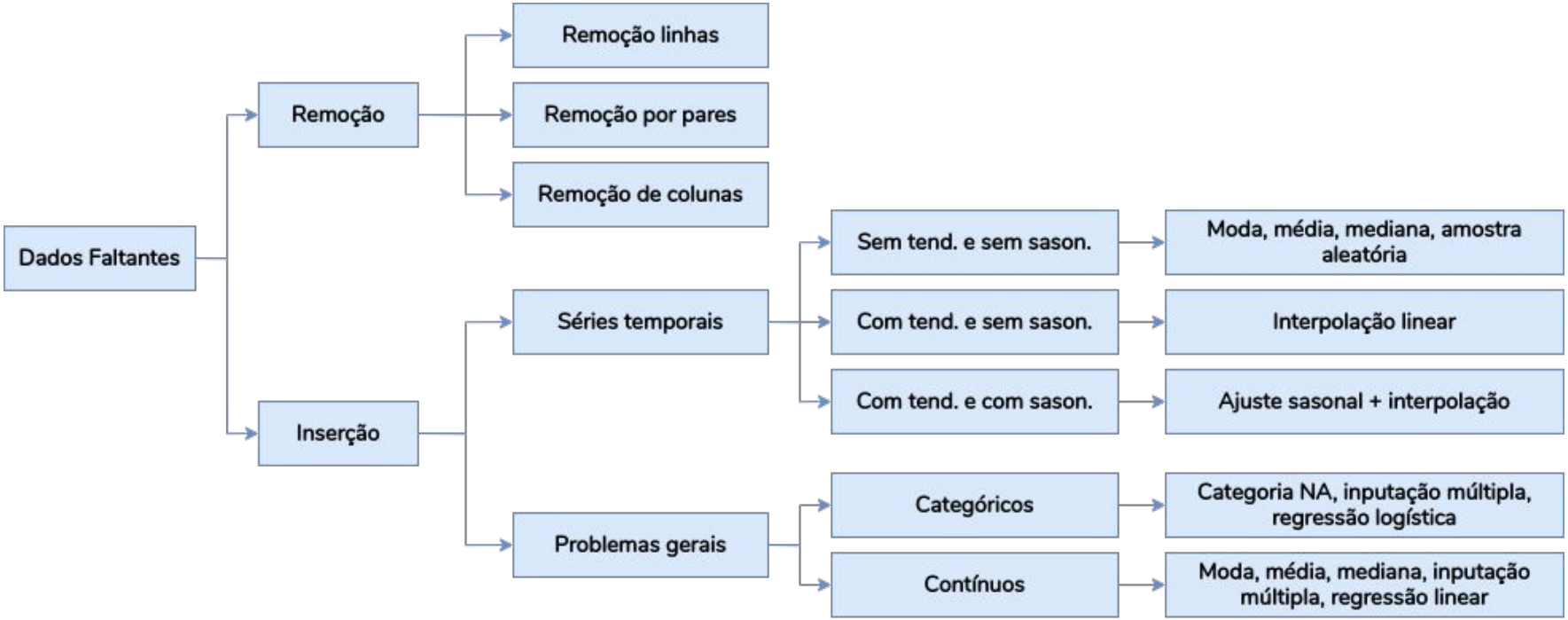


Diagrama traduzido de https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

Pré-processamento

CiDAMO

- Normalização (Standardization, Scaling, Normalization)
- Binarização (Binarization)
- Codificação One-Hot (One-hot encoding)
- Codificação por categoria (Label Encoding)

- Usar antes da modelagem

Normalização Tipo Standardization

CiDAMO

- Standardization:

$$X_{esc} = \frac{X - média(X)}{dp(X)}$$

- Rescala a distribuição da variável X para a distribuição Normal de média 0(zero) e desvio padrão 1.
- Ajuda a remover o viés das variáveis

Normalização Tipo Scaling

CiDAMO

Min-Max Method

- Scaling na escala [0, 1]:

$$X_{esc} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- Scaling na escala [a, b], a < b:

$$X_{esc} = a + \frac{(b - a)(X - X_{min})}{X_{max} - X_{min}}$$

Normalização Tipo *Normalization*

- Norma tipo 1:

$$Z = \sum_{i=1}^{N} |X_i|$$

- Norma tipo 2:

$$Z = \sqrt{\sum_{i=1}^{N} X_i^2}$$

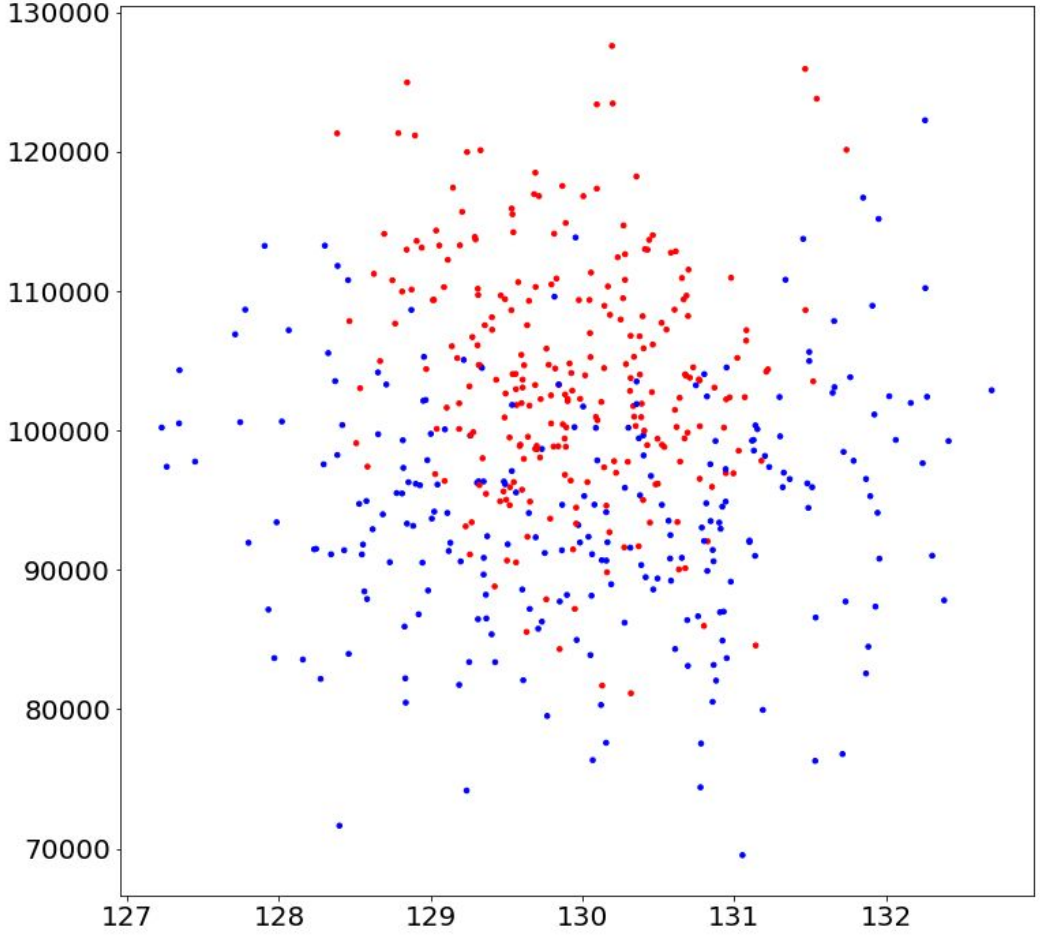
- Norma tipo inf:

$$Z = max(X)$$

- Normalization:

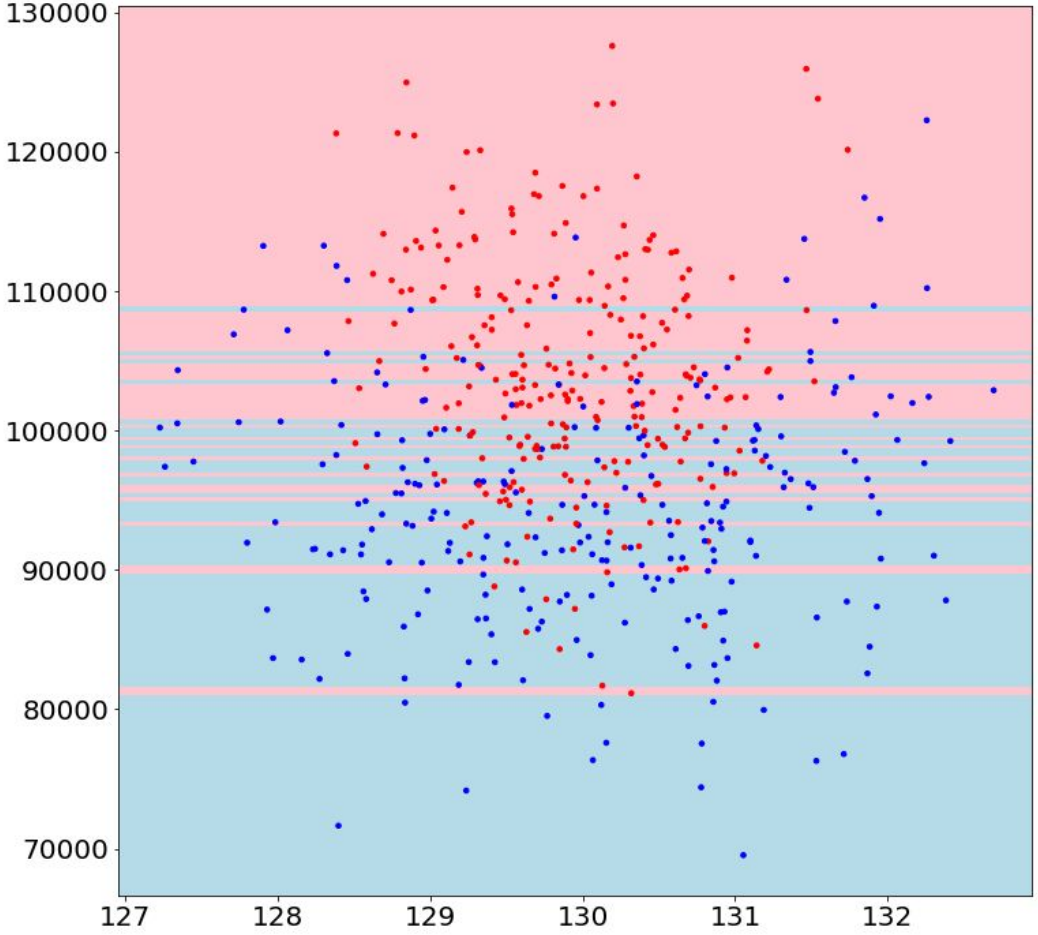
$$X_{esc} = \frac{X}{Z}$$

- Usado para ajustar os valores numa escala comum.
- A soma dos valores escalados é 1.

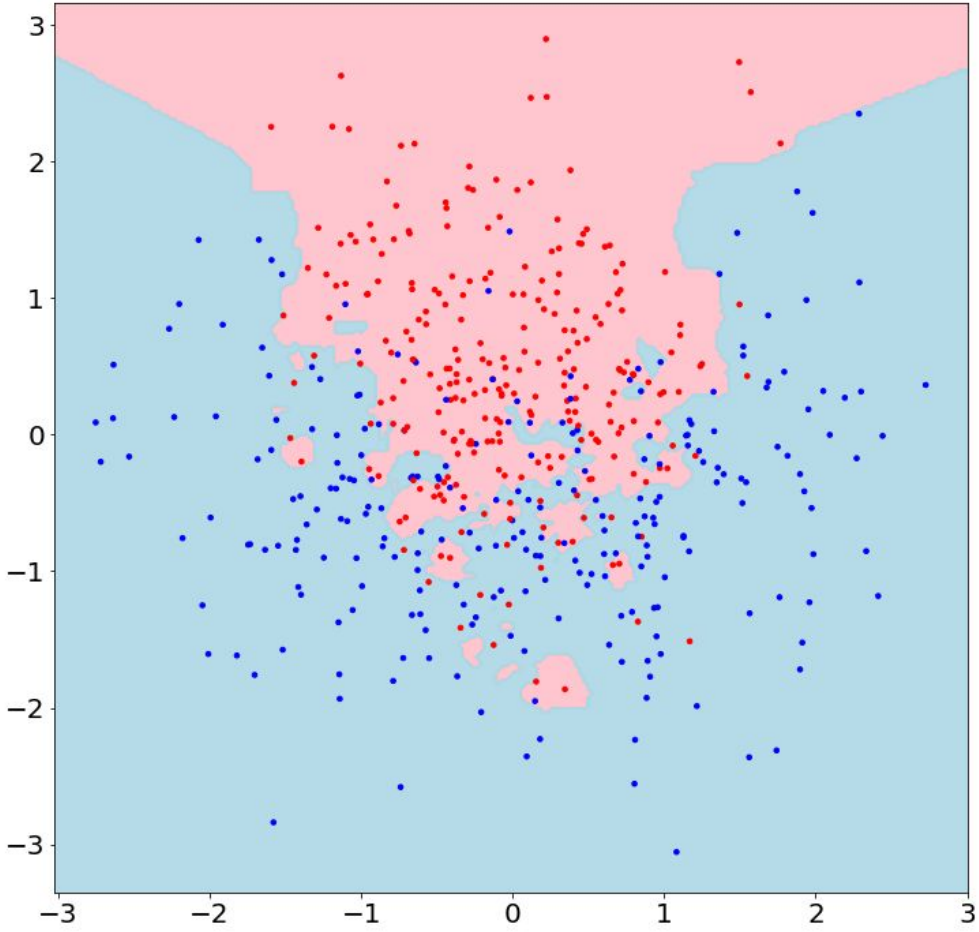


CiDAMO

Dados simulados



KNN aplicado em cada ponto do espaço



CiDAMO

Dados normalizados e KNN aplicado em cada ponto do espaço

Binarização (*Binarization*)

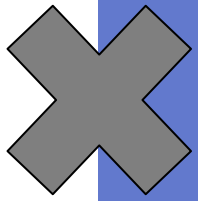
CiDAMO

- Usado para converter uma variável quantitativa em variável binária

- Ao invés de considerar a quantidade, considera-se a presença ou ausência de uma característica.

One-Hot Encoding

- Converte variável qualitativa em uma matriz de variáveis binárias.
- Nessa conversão, as variáveis são binárias, no entanto, dependendo da cardinalidade, as dimensões da matriz podem comprometer eficiência (tempo) do método.



Label/Ordinal Encoding

- Converte variável qualitativa em variável quantitativa, mantendo a mesma estrutura vetorial.
- A conversão das variáveis qualitativas para quantitativas pode gerar problemas com operações matemáticas.

In [1]:
```
import numpy as np
import pandas as pd

data = pd.read_csv('diamante-treino.csv')
data.head()
```

Out[1]:

	Price	Carat	Color	Clarity	Cut
0	6877	1.40	H	SI2	Very Good
1	4416	1.01	J	VS1	Very Good
2	4866	0.80	F	VVS1	Ideal
3	3522	0.92	E	SI1	Very Good
4	1102	0.40	E	VS1	Excellent

In [2]:
```
from sklearn.preprocessing import OneHotEncoder, LabelEncoder, OrdinalEncoder

lbl = LabelEncoder()
colors = lbl.fit_transform(data[['Cut']])
print(lbl.classes_)
print(data.Cut[0:5])
print(colors[0:5])
```

```
['Excellent' 'Good' 'Ideal' 'Very Good']
0    Very Good
1    Very Good
2        Ideal
3    Very Good
4    Excellent
Name: Cut, dtype: object
[3 3 2 3 0]
```

In [3]:
```
ohe = OneHotEncoder()
colors = ohe.fit_transform(data[['Cut']])
print(ohe.categories_)
print(data.Cut[0:5])
print(colors[0:5,:].todense())
```

```
[array(['Excellent', 'Good', 'Ideal', 'Very Good'], dtype=object)]
0    Very Good
1    Very Good
2        Ideal
3    Very Good
4    Excellent
Name: Cut, dtype: object
[[0. 0. 0. 1.]
 [0. 0. 0. 1.]
 [0. 0. 1. 0.]
 [0. 0. 0. 1.]
 [1. 0. 0. 0.]]
```

In [4]:
```
ohe = OneHotEncoder(drop='first')
colors = ohe.fit_transform(data[['Cut']])
print(ohe.categories_)
print(data.Cut[0:5])
print(colors[0:5,:].todense())
```

```
[array(['Excellent', 'Good', 'Ideal', 'Very Good'], dtype=object)]
0    Very Good
1    Very Good
2        Ideal
3    Very Good
4    Excellent
Name: Cut, dtype: object
[[0. 0. 1.]
 [0. 0. 1.]
 [0. 1. 0.]
 [0. 0. 1.]
 [0. 0. 0.]]
```

In [5]:
```
ord = OrdinalEncoder(categories=[['Good', 'Very Good', 'Excellent', 'Ideal']])
cuts = ord.fit_transform(data[['Cut']])
print(ord.categories)
print(data.Cut[0:5])
print(cuts[0:5])
```

```
[['Good', 'Very Good', 'Excellent', 'Ideal']]
0    Very Good
1    Very Good
2        Ideal
3    Very Good
4    Excellent
Name: Cut, dtype: object
[[1.]
 [1.]
 [3.]
 [1.]
 [2.]]
```

Referências

- Outlier
  - Livro: Data Science from Scratch, Joel Grus
- Correlação
  - http://leg.ufpr.br/~paulojus/CE003/ce003/node8.html
- Missing data
  - https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/
  - https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4
- Pré-processamento
  - Python Machine Learning Cookbook, Giuseppe Ciaburro & Prateek Joshi

Obrigada

Estes slides e as imagens aqui presente são propriedade intelectual de seus autores, exceto quando explicitado o contrário.

Distribuição pública dentro da licença CC-BY-SA 4.0