Licap

Formação Cientista de Dados

PUC Minas

Licap

Formação do Cientista de Dados

Exploração e pré-processamento da base de dados

Luis Enrique Zárate

Conteúdo do Curso

Licap

Exploração dos dados

1. Remoção de variáveis pela frequência de valores
2. Esparcidade de variáveis
3. Monotonicidade de variáveis
4. Incremento dimensional de uma variável
5. Outliers de uma variável
6. Realce com polarização
7. Coerência nas instâncias
8. Análise da Causalidade
9. Erros de medição
10. Análise da variabilidade

Descrição estatística de variáveis

Exploração de variáveis

Licap

Remoção de Variáveis pela frequência de valores

A informação básica de uma variável compreende o número de valores distintos e a frequência de cada valor.

Valor da Variável	Frequencia de ocorrência
A	1
B	2
C	15
D	2
E	1

A variável deve ser uma constante ou a amostra está polarizada.

Ações:

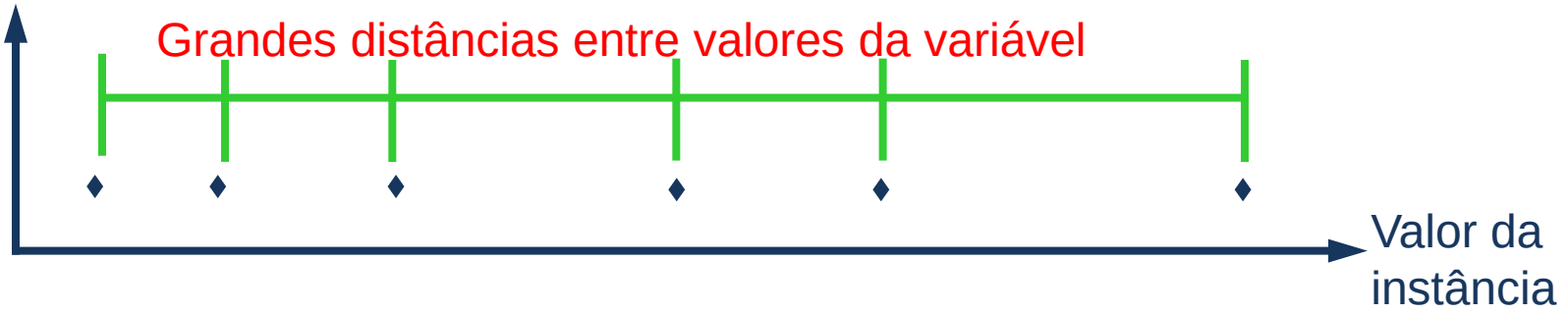
Obtem-se uma nova amostra, remove-se a variável, segmenta-se o conjunto de dados, ou coloca-se restrições ao conhecimento extraído.

Exploração de variáveis

Licap

Esparcidade de Variáveis

- As variáveis esparças podem ter pouco relevância, portanto poderiam ser eliminadas.
- Porém poderiam ser transformadas em variáveis categóricas ordinales (“Colapso de variável”).
- O minerador deverá analisar para tomar essa decisão



Exploração de variáveis

Licap

Esparcidade de Variáveis

- Exemplo: consideremos a variável valor do empréstimo x R$1000,00 de uma base contendo 5000 registros:

Empréstimos = {0, 0, 10, 100, 0, 5, 20, 150, 2, 0, 0, 1250, 0,....., 35, 2, 10, 1500}

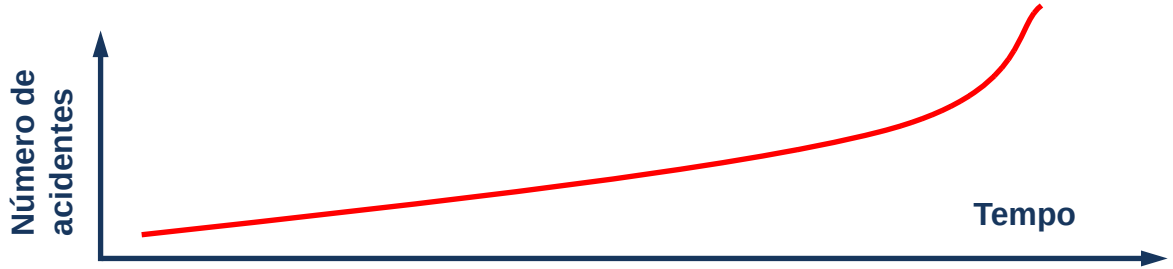
- Existe grande esparcidade no conjunto de valores.
- A esparcidade é resultado de uma amostra inadequada ou retrata a realidade?
- Se a esparcidade retrata a realidade seria conveniente converter a variável em variável categórica ordinal: {**Baixo empréstimo, Médio Empréstimo e Alto emprestimo}** de acordo a faixas a serem ajustadas adequadamente.

Exploração de variáveis

Monotonicidade de Variáveis

Uma variável que se incrementa sem limites é chamada de monotónica.
Exemplos: número de acidentes, número de compras pela internet, número de fraudes, número de acessos a websites, etc.

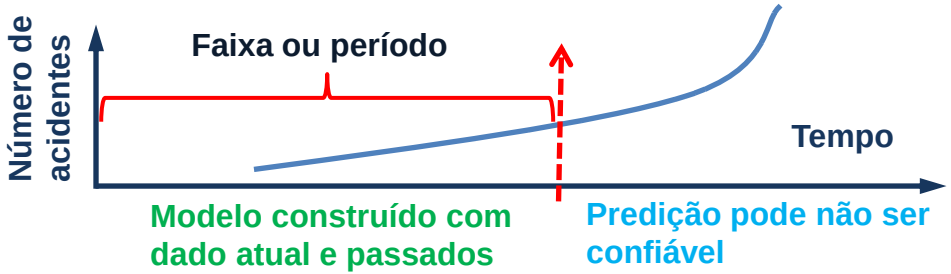
As variáveis monotónicas têm que ser transformadas em variáveis no-monotónicas para serem utilizadas na mineração e deverão ser limitados a uma faixa.



Exploração de variáveis

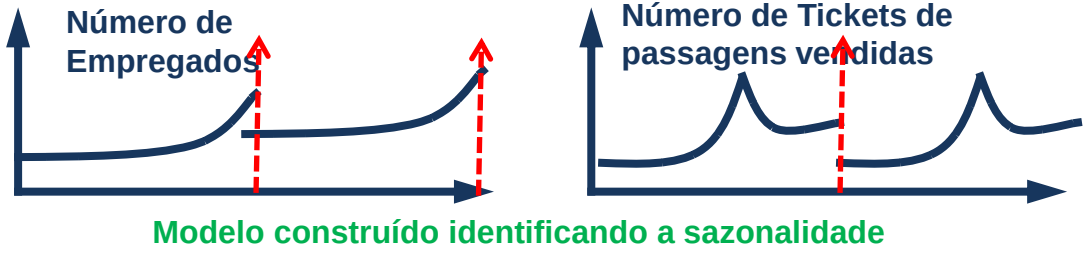
Monotonicidade de Variáveis

Quando é considerada uma faixa para a variável monotónica, o modelo minerado pode não ser confiável na predição.



Transformações aplicáveis:

Identificar a estacionaridade da variável.



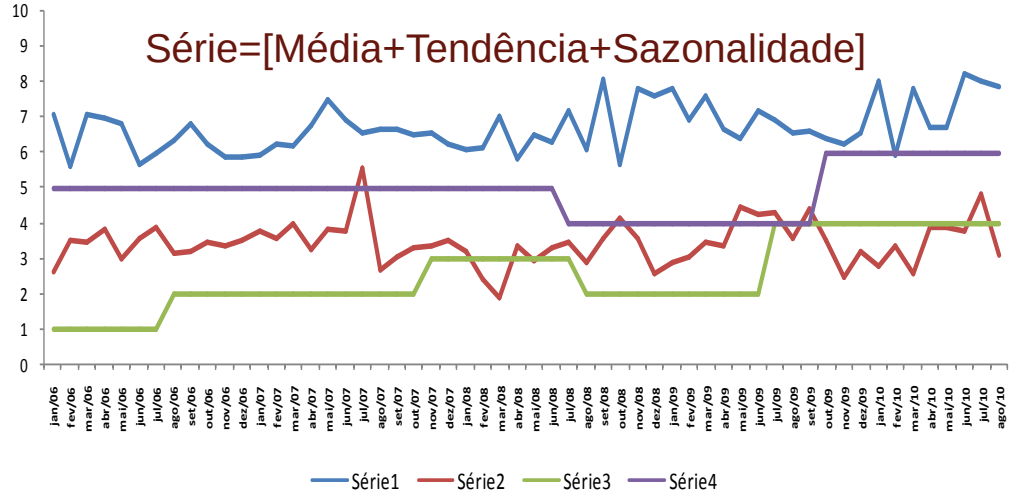
Exploração de variáveis

Monotonicidade de Variáveis

Transformações aplicáveis:

Tratar a informação dos dados como uma Série Temporal, extraíndo médias, tendências e sazonalidades.

Cada variável monotónica requer uma transformação específica.

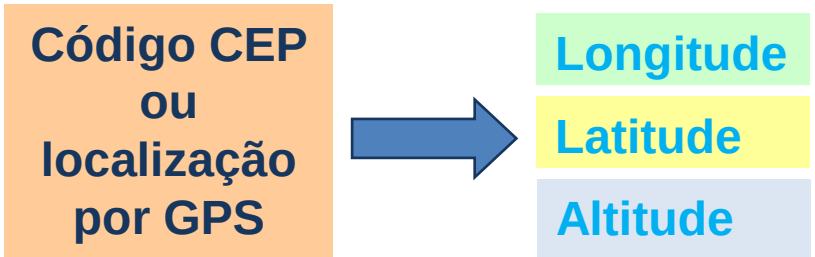


Exploração de variáveis

Licap

Incremento Dimensional de uma Variável

Existem algumas circunstâncias onde a dimensão de uma variável requer ser incrementada.



Exploração de variáveis

Licap

Outliers de uma Variável

É uma valor (ou valores) com baixa frequencia de ocorrência localizado longe das maiores concentrações dos valores da variável.

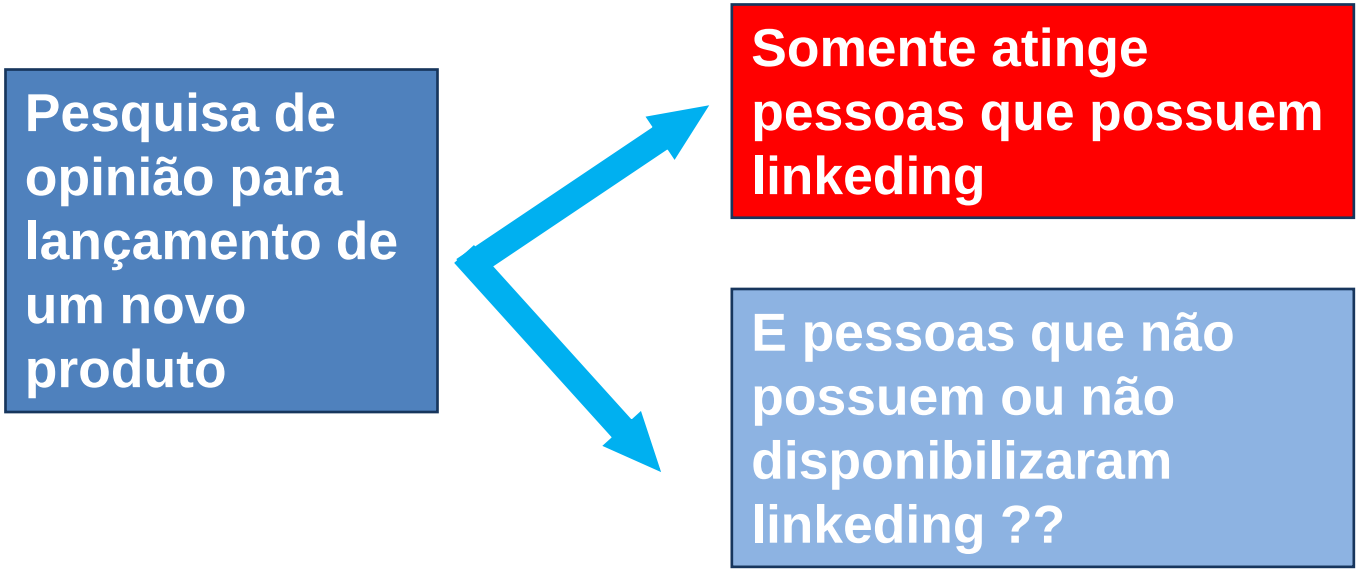
A grande questão é saber se os *“outliers”* são um erro ou não. Pois estes podem distorcer a resposta de um modelo de ML.

Exploração de variáveis

Licap

Realce com Polarização

A polarização deve ser conhecida e controlada. A polarização pode impor restrições ao modelo.

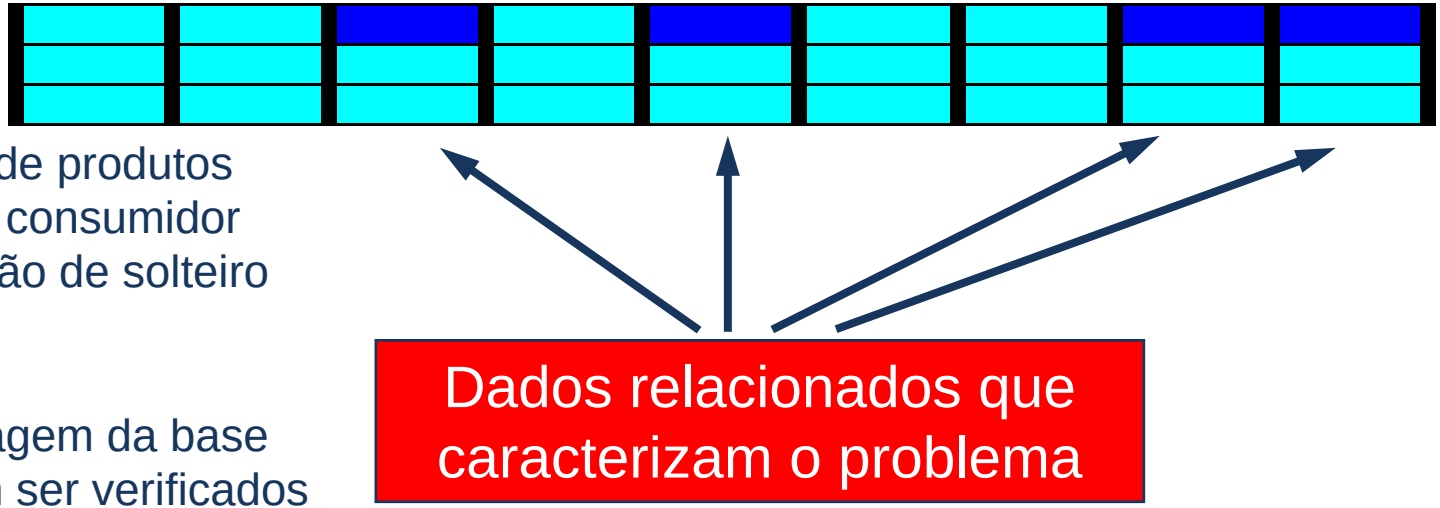


Exploração de variáveis

Licap

Coerência nas instâncias

O Minerador deve explorar e observar a consistência dos conjunto de dados relacionados ao problema.



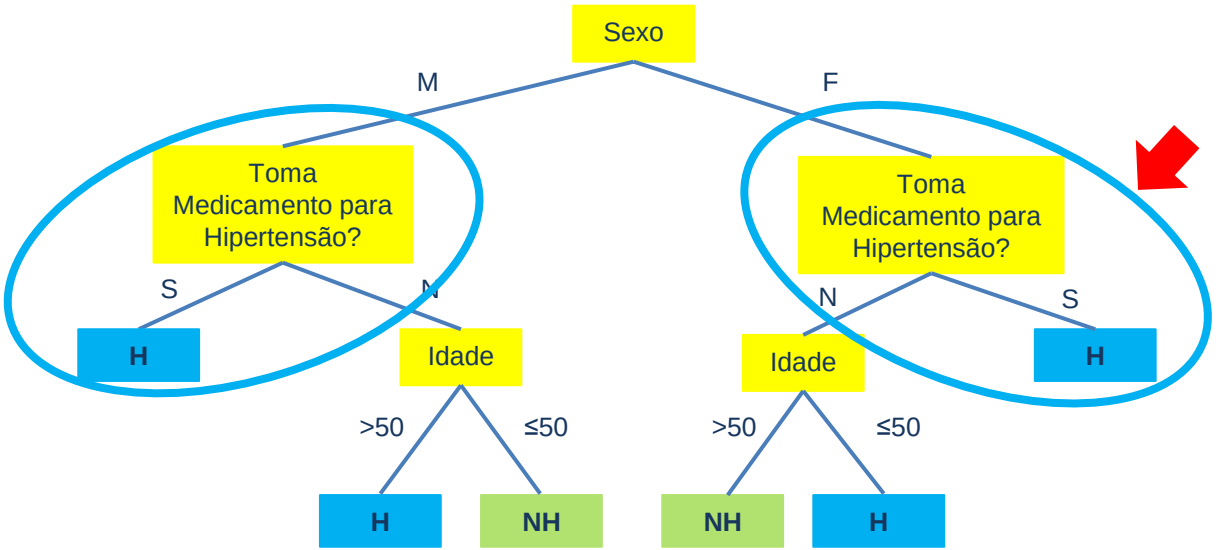
Ex. O consumo de produtos pode variar se o consumidor passa da condição de solteiro para casado.

Durante a montagem da base de dados devem ser verificados as mudanças de perfis.

Exploração de variáveis

Consequência vs. Causalidade

O Minerador deve selecionar variáveis causadoras e não variáveis consequência associadas a CLASSE num problema de predição.



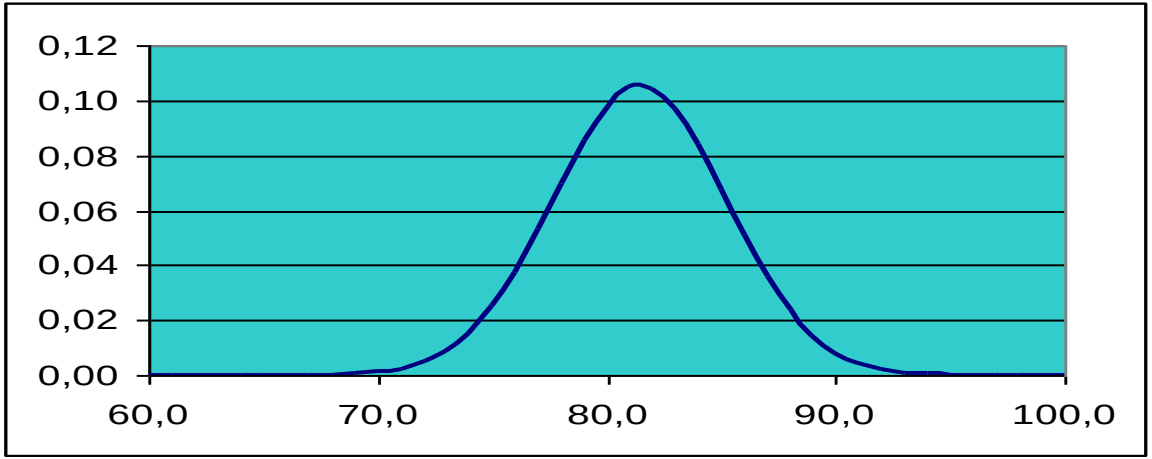
É uma variável de Consequência e não de Causalidade. Isto leva a modelos “Naive”

Exploração de variáveis

Licap

Erros de Medição:

Medida incorreta: sempre é aleatória com distribuição gaussiana

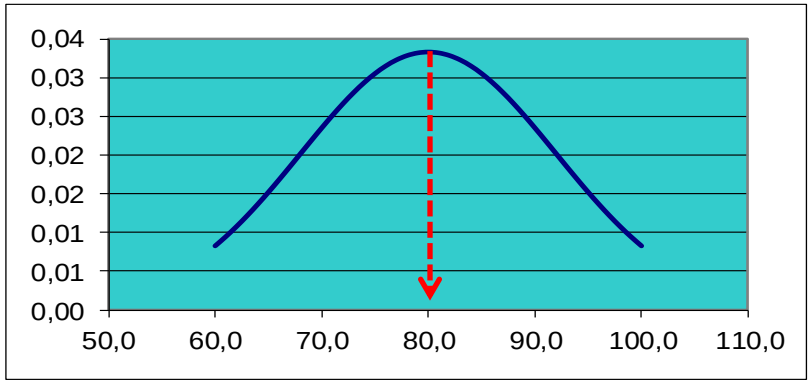


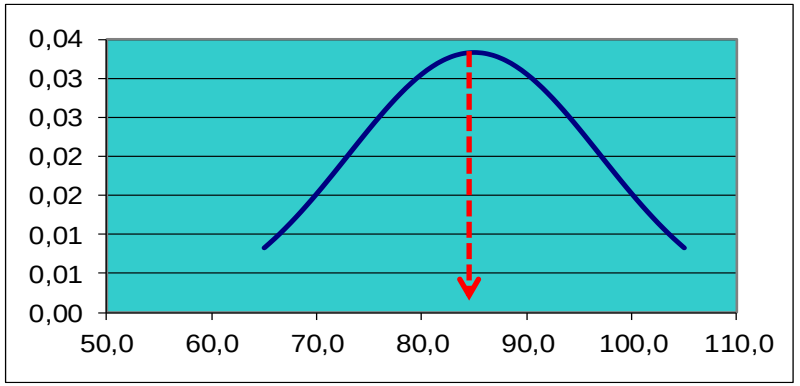
Exploração de variáveis

Erros de Medição:

Erros de calibração: polarização constante. Falha de calibragem

Exemplo: Média deslocada e Desvio padrão constante.



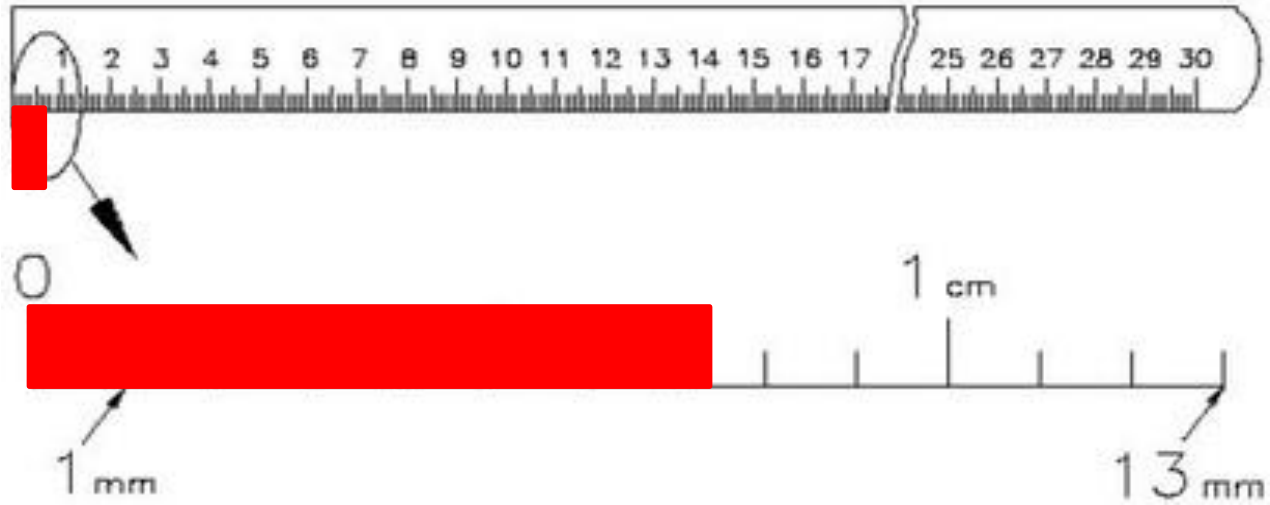


Exploração de variáveis

Licap

Erros de Medição:

Erros de precisão: trunca o valor medido.



Exploração de variáveis

Licap

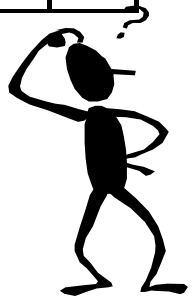
Análise da Variabilidade

Está relacionado à frequencia de valores que uma variável pode adotar

Considere a seguinte amostra representativa de uma população:

49	63	44	25	16	34	62	55	40	31	44	37	48	65	83	53	39	15	25	52
68	35	64	71	43	76	39	61	51	30	32	74	28	64	46	31	79	69	38	69
53	32	69	39	32	67	17	52	64	64	25	28	64	65	70	44	43	72	37	31
67	69	64	74	32	25	65	39	75	36	26	59	28	23	40	56	77	68	46	48

Existe algum padrão evidente??

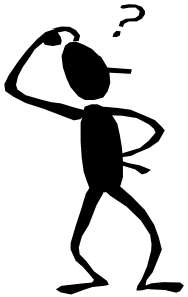


Exploração de variáveis

Licap

Considere que a amostra está ordenada

15	16	17	23	25	25	25	25	26	28	28	28	30	31	31	31	32	32	32	32
34	35	36	37	37	38	39	39	39	39	40	40	43	43	44	44	44	46	46	48
48	49	51	52	52	53	53	55	56	59	61	62	63	64	64	64	64	64	64	65
65	65	67	67	68	68	69	69	69	69	70	71	72	74	74	75	76	77	79	83

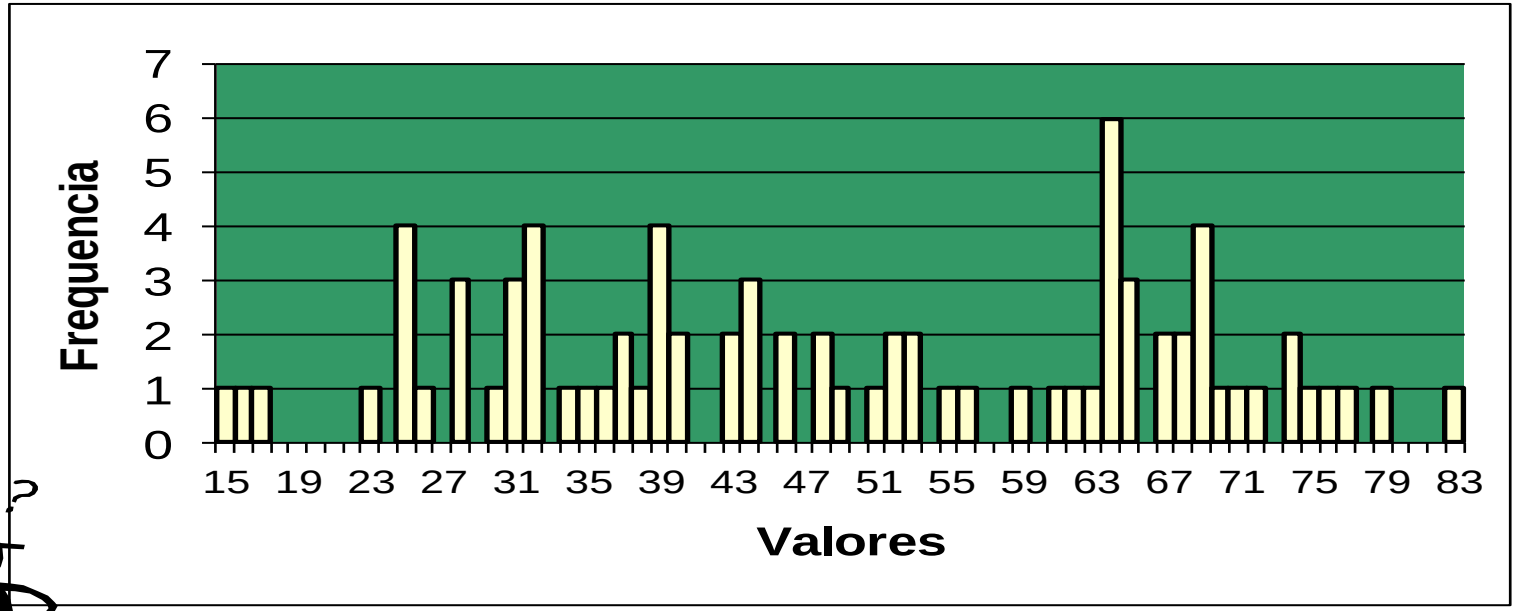


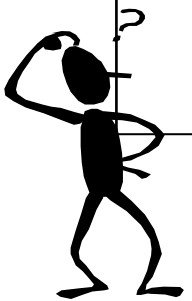
Pode existir algum padrão embora é difícil identificá-o !!

Exploração de variáveis

Licap

Considerando o histograma de frequências





O padrão é ainda difícil de ser detectado !!!

Exploração de variáveis

Licap

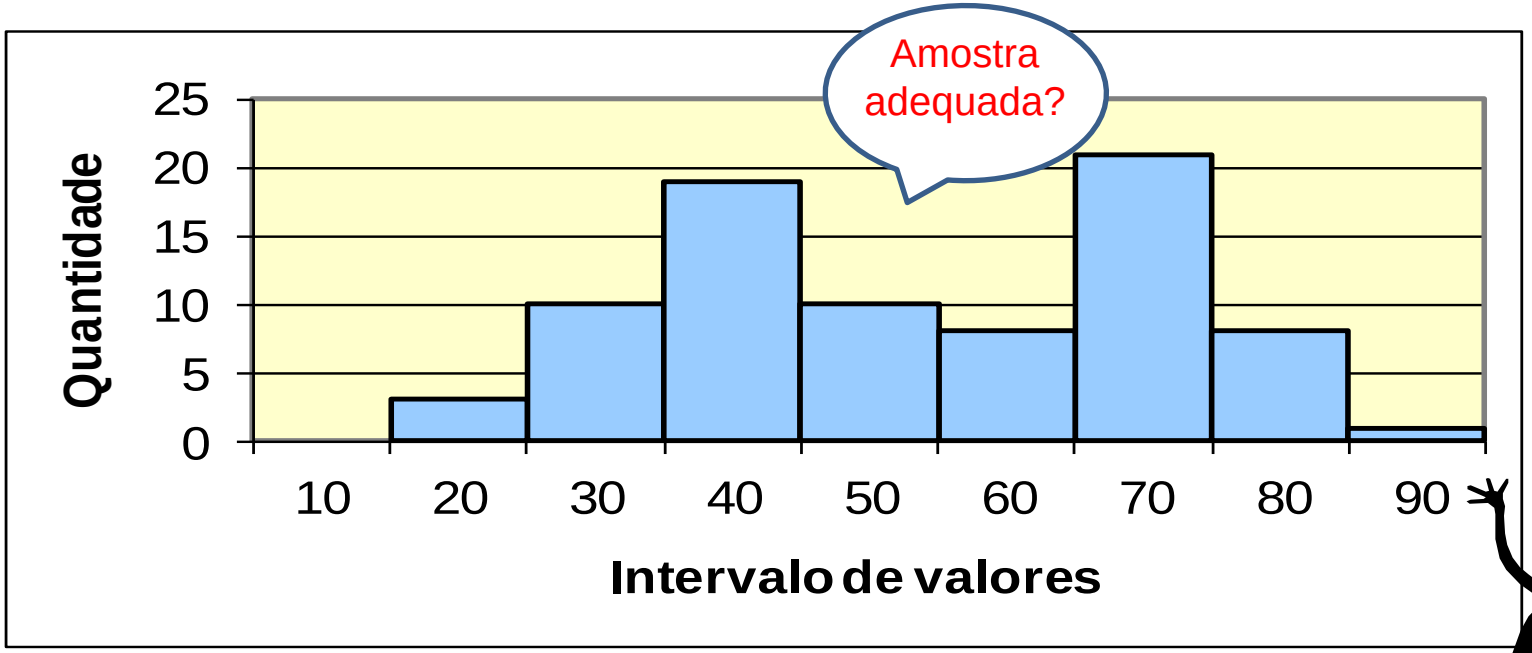
Considerando o suavizamento de curvas

É possível observar algum padrão !!!

Exploração de variáveis

Licap

Considerando o histograma por intervalos



Agora sim é possível observar um padrão !!!

Descrição estatística de variáveis

Licap

Seja o conjunto de dados relativos à variável peso de 80 pessoas:

81,80	87,10	82,70	79,80	81,30	79,50	88,50	75,90
81,60	73,90	84,50	87,10	82,00	79,30	82,50	87,10
83,00	87,30	79,70	82,00	83,60	84,50	80,40	78,10
86,40	76,70	83,70	78,40	76,00	80,90	80,20	78,90
77,40	78,50	82,90	81,90	80,70	78,40	78,00	81,40
84,60	79,50	82,30	80,50	80,70	79,00	90,00	79,90
86,80	80,10	83,20	78,20	80,40	85,50	85,50	79,30
83,00	78,10	83,40	83,60	85,70	86,80	86,50	83,80
86,80	83,50	79,90	76,60	84,30	78,50	74,40	71,80
79,10	82,10	84,50	78,40	80,70	70,70	78,50	85,20

Media: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\bar{x}$

Mediana: $\tilde{x} = x_{([n+1]/2)}$

para "n" impar

$$\tilde{x} = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

para "n" par

Desvio Padrão:

$$S(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

n-1: para amostra; n: para população

Variância:

$$Var(x) = S^2(x)$$

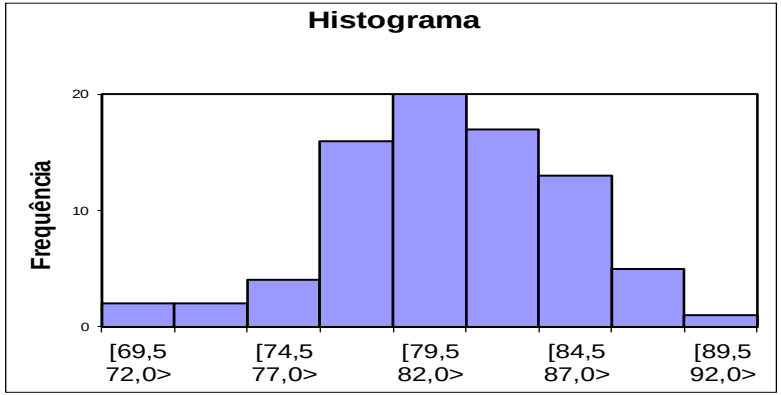
Medidas de Têndencia Central	
Média	81,44
Mediana	81,35
D. Pad	3,79
Variância	14,36

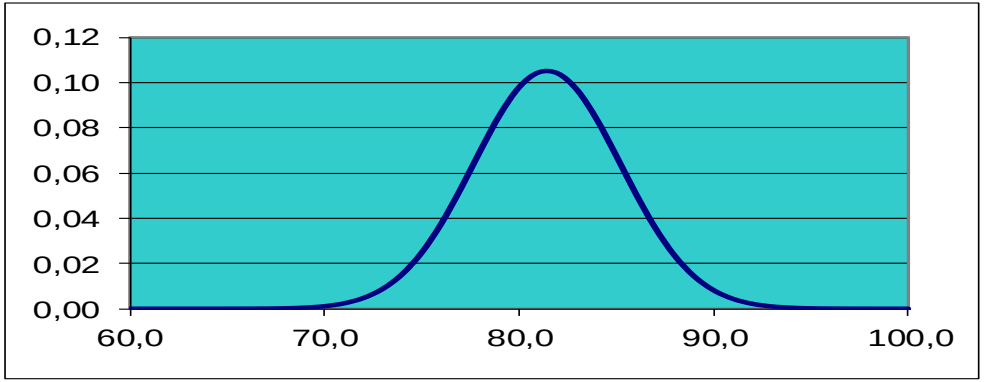
Moda: Valor mais frequente de uma variável. Se aplica para dados discretos e categóricos.

Descrição estatística de variáveis

Licap

Função densidade de probabilidade - Distribuição Normal





Medidas de Têndencia Central	
Média	81,44
Mediana	81,35
D. Pad	3,79
Variância	14,36

Intervalo	Probabilidade	
	Interna	Externa
$\mu \pm 1\sigma$	68,2%	31,74%
$\mu \pm 2\sigma$	95,46%	4,54%
$\mu \pm 3\sigma$	99,73%	0,27%

$N(\mu,\sigma)$

Descrição tabular da base de dados

Licap

Total de registros na base = 200

Atributo	Relevância	Tipo	Valores	Valor mínimo	Valor máximo	Média	Desvio Padrão	Moda	Distribuição	Dados ausentes
Idade	Fato	Quant. (discreto)	.....	21	60	30	10	42	Normal	5
Peso	Fato	Quant. (contínuo)	.....	60,5	90,8	74,2	5,1	...	Normal	3
Sexo	Julga.	Qual. (dico)	{M,F}	.....	.....	M: 70% F: 30%	.....	M: 150 F: 50		0
...		...		...	...	...	...	...	...	...

Licap

Formação Cientista de Dados

Obrigado!

PUC Minas