

Ciência dos Dados

Aula 04

Explorando Variáveis Quantitativas

Aula ANTERIOR (Aula03)

Ao final desta aula, o aluno deve ser capaz de:

- Construir tabelas de frequências e interpretar resultados considerando uma variável quantitativa e de forma cruzada com uma outra variável qualitativa.
- Explicar vantagens e desvantagens sobre o uso da frequência ou da densidade na construção de um histograma e saber interpretá-lo.
- Desenvolver contas para obtenção de média, mediana e moda e associar ordenação dessas medidas de acordo com assimetria dos dados.

Objetivos de Aprendizagem

Os alunos devem ser capazes de:

- Desenvolver medidas de dispersão que gerem informações para interpretação de variáveis quantitativas.
- Interpretar o comportamento de uma variável quantitativa a partir dos formatos de um box-plot.

Acompanhe, previamente, o PLANO DE AULA
no BLACKBOARD!

Medidas de dispersão

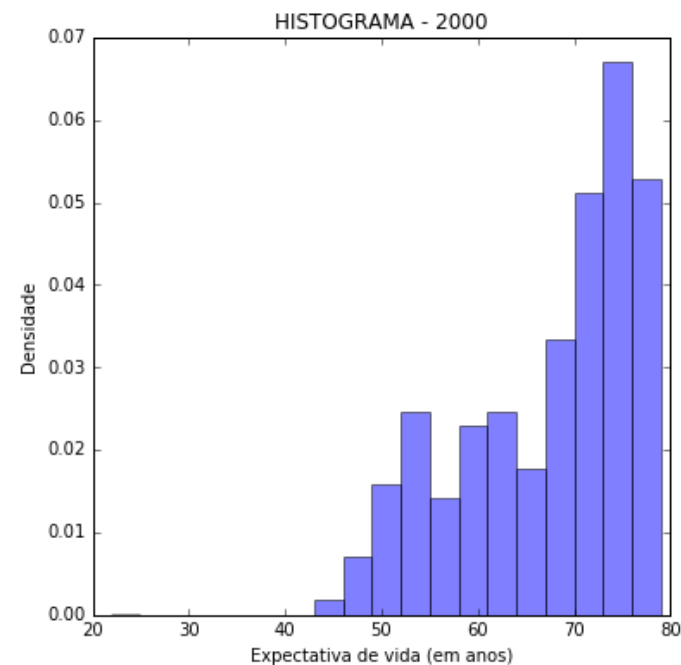
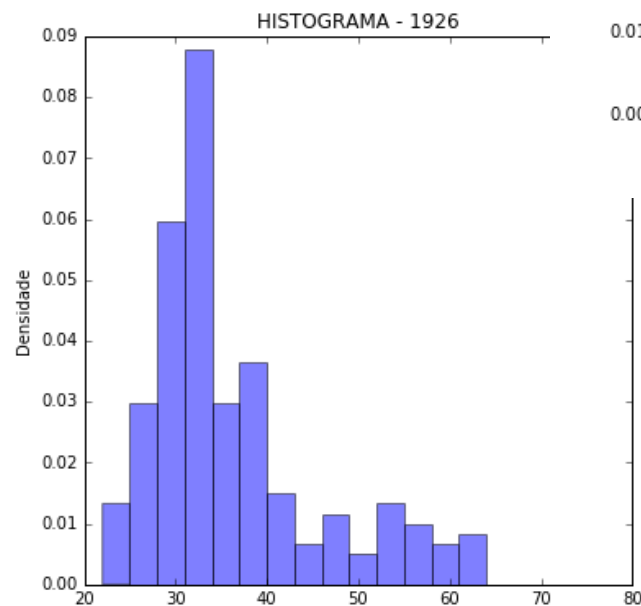
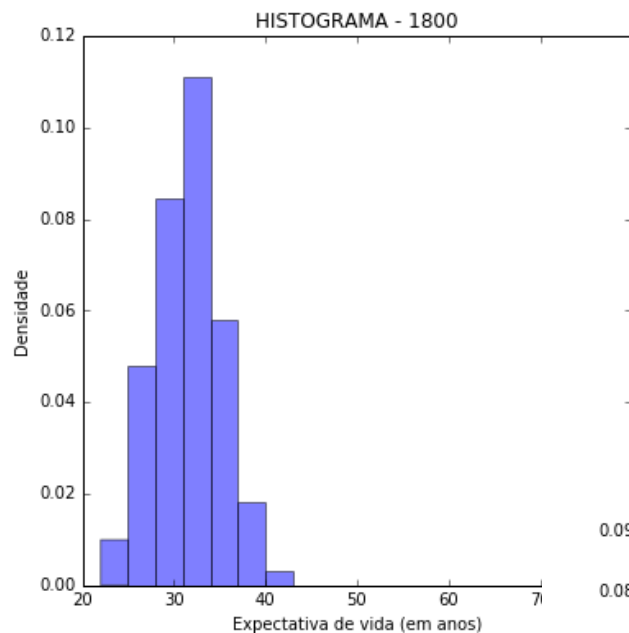
Expectativa de vida

Expectativa de vida (em anos) dos anos 1800, 1926 e 2000.

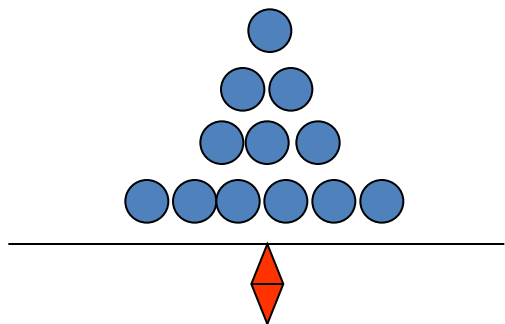
Medidas Resumo	1800	1926	2000
Média	31,5	36,3	68,0
Mediana	31,8	32,8	71,2
Desvio Padrão	3,76	9,59	9,21
Variância	14,16	91,99	84,86
Desvio Médio Absoluto	2,98	7,40	7,79
Tamanho amostral	201	201	201

Expectativa de vida

Expectativa de vida (em anos) dos anos 1800, 1926 e 2000.

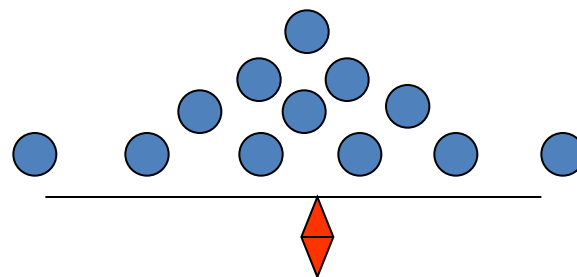


Medidas de Dispersão baseadas em distâncias a uma medida de tendência central



Baixa variabilidade

**As observações
estão próximas à
medida de tendência
central**



Alta variabilidade

**As observações
estão mais distantes
da medida de
tendência central**

Medidas de dispersão

Variância: a variância da amostra é a média das diferenças ao quadrado entre cada uma das observações e a média do conjunto.

$$\text{var}(X) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Características:

- Um dos problemas de usar a variância como medida de dispersão é o fato de sua unidade não ser a mesma unidade em que a variável foi medida (os valores dos dados estão elevados ao quadrado). A solução é extrair a raiz quadrada positiva da variância, já que, com isso, se volta à unidade original da variável.

Medidas de dispersão

Desvio Padrão: o desvio padrão de um conjunto de valores amostrais é uma medida da variação dos valores em torno da média. É uma espécie de desvio médio dos valores em relação à média aritmética.

$$dp(X) = \sqrt{\text{var}(X)} = \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 / n}$$

Características:

- É positivo. É zero apenas quando todos os valores dos dados são o mesmo número.
- O valor do desvio padrão pode crescer dramaticamente com a inclusão de um ou mais *outliers*.
- A unidade de medida é a mesma da variável X.

Desvio Médio Absoluto

Desvio Médio Absoluto: é a distância média dos dados até a média (aritmética).

$$dm(X) = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

Características:

Na prática, não é uma medida muito utilizada para resumir a variabilidade de um conjunto de dados. Isso ocorre porque a função módulo não é diferenciável o que cria dificuldades para alguns métodos de inferência estatística.

Percentis e Boxplot

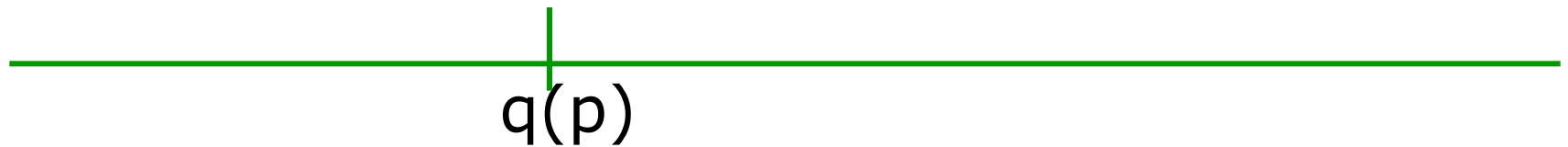
Magalhães e Pedroso de Lima, Cap 1

Percentil ou Quantil

Amostra ordenada

$p\%$ menores
observações

$(100-p)\%$ maiores
observações



Quantil ou Percentil de ordem p ($0 < p < 100$): é o valor que divide o conjunto de dados ordenado em 2 partes: uma delas com $p\%$ dos menores valores e a outra com $(100-p)\%$ dos maiores valores.

Expectativa de vida - Percentis

Expectativa de vida (em anos) dos anos 1800, 1926 e 2000.

Ordem	1800	1926	2000
Mínimo	23	23	46
10%	26	27	53
20%	29	30	60
30%	30	31	64
40%	31	32	69
50%	32	33	71
60%	32	35	73
70%	33	38	74
80%	35	41	76
90%	36	53	78
Máximo	43	63	83

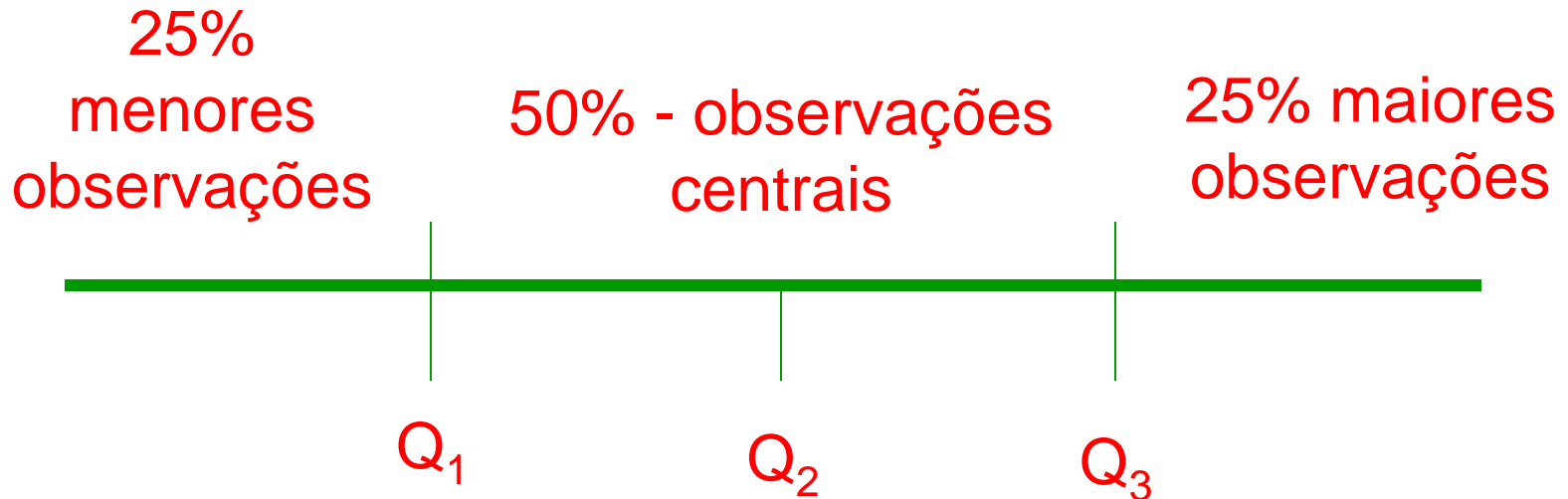
Life expectancy	1800	1926	2000
Brazil	32	31,99072	71,9

Comando Python:

`np.percentile (variável, ordem entre 0 e 100)` ou `variável.quantile(ordem em 0 e 1)`

Quartis

Amostra ordenada



Quartil: são valores que dividem o conjunto de dados ordenados em quatro partes iguais.

Cada parte contendo 25% dos dados.

Intervalo Interquartil: $IQ = Q_3 - Q_1$

Exercício: Dirigindo Bêbado

Excedendo o limite de 0,1 grama de álcool por litro de sangue:
paga multa, perde a carteira e tem carro apreendido.

Excedendo o limite de 0,6 grama de álcool por litro de sangue:
pode ser preso.

Os 14 motoristas abaixo foram parados em uma blitz policial.

0,32	0,39	0,02	0,18	0,13	0,49	0,63
0,08	0,08	0,16	0,08	0,26	0,16	0,25

Dado que a atual lei proíbe dirigir com níveis acima de 0,1, parece que esses níveis estão acima do permitido?

Encontre os quartis, calcule o IQ e discuta a assimetria.

Exercício: Dirigindo Bêbado

Conjunto de Dados:	0,32	0,39	0,02	0,18	0,13	0,49	0,73
	0,08	0,08	0,16	0,08	0,26	0,16	0,25

Dados	0,02	0,08	0,08	0,08	0,13	0,16	0,16
Ordenados:	0,18	0,25	0,26	0,32	0,39	0,49	0,73

$$Q1 = 0,08$$

$$Q2 = 0,17$$

$$Q3 = 0,32$$

$$IQ = Q3 - Q1 = 0,32 - 0,08 = 0,24$$

Boxplot

O **boxplot** é uma figura que possibilita visualizar várias características de um conjunto de dados como:

- as de tendência central (mediana)
- de posição (primeiro quartil e terceiro quartil)
- de dispersão (intervalo entre quartis)
- de assimetria
- pode **identificar os valores considerados como possíveis extremos.**

Identificação de possíveis valores aberrantes

Um ponto (w) será considerado suspeito de ser aberrante se:

$$w > LS = Q_3 + 1,5 \text{ IQ (Limite Superior), ou}$$

$$w < LI = Q_1 - 1,5 \text{ IQ (Limite Inferior),}$$

sendo $\text{IQ} = Q_3 - Q_1$ (intervalo interquartílico)

Propriedade: num modelo Normal, apenas 0,7% dos dados estão fora desses limites.

Limitação: em distribuições assimétricas a regra tende a identificar um número excessivo de valores suspeitos.

Boxplot

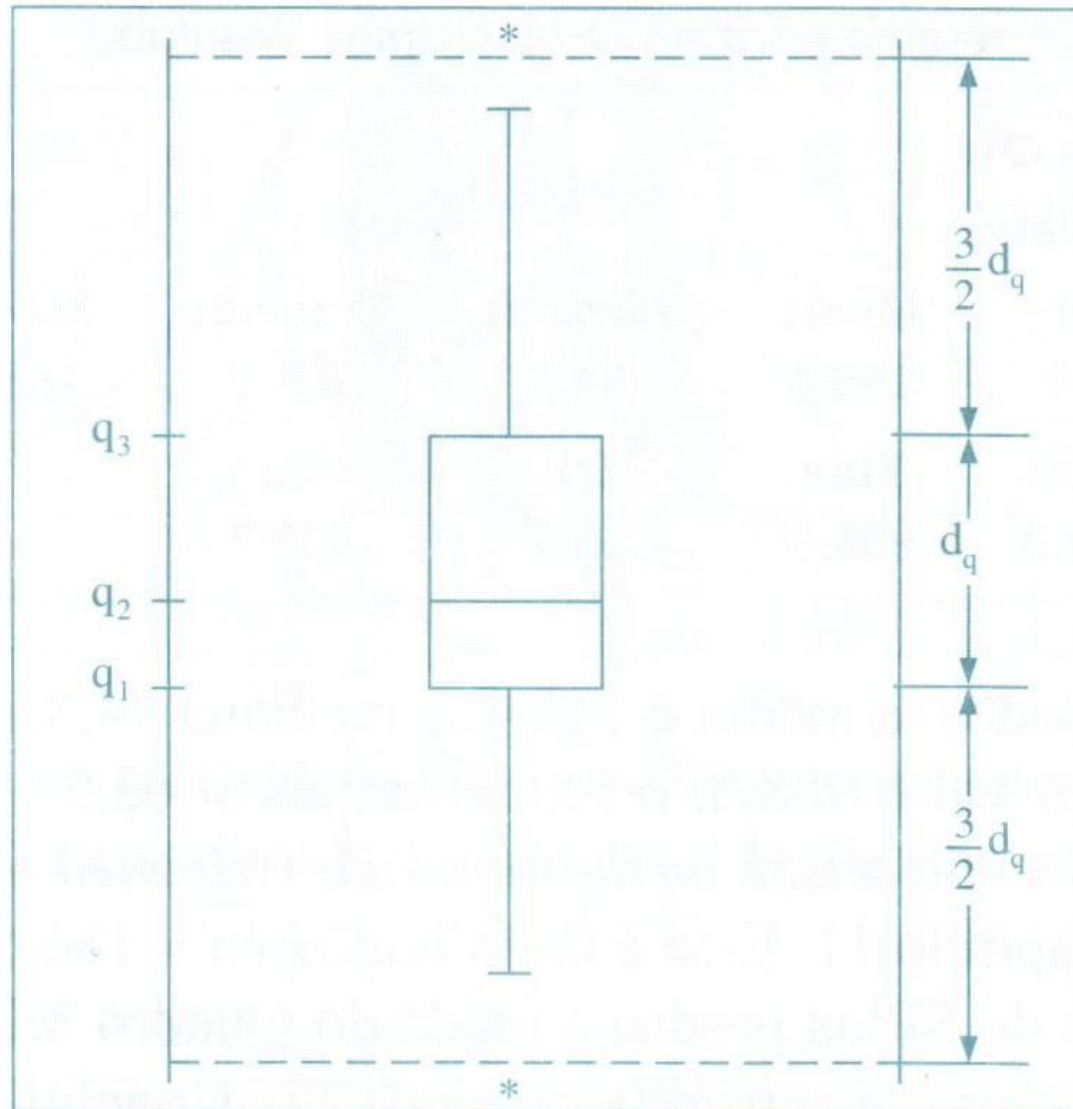
Caixa que contém 50% das observações centradas. Parte superior é Q3 e o inferior é Q1. Mediana Q2 está dentro da caixa. ($IQ = Q3 - Q1$)

Linha acima da caixa segue até o ponto mais remoto que não exceda o limite superior ($LS = Q3 + 1,5 IQ$)

Linha abaixo da caixa segue até o ponto mais remoto mas maior do que o limite inferior ($LI = Q1 - 1,5 IQ$)

Possível ponto extremo ou aberrante ("outlier") – aqueles acima do LS ou abaixo do LI

Boxplot



Fonte: Bussab e Morettin

Exercício: Dirigindo Bêbado

Conjunto de Dados:	0,32	0,39	0,02	0,18	0,13	0,49	0,73
	0,08	0,08	0,16	0,08	0,26	0,16	0,25

Dados	0,02	0,08	0,08	0,08	0,13	0,16	0,16
Ordenados:	0,18	0,25	0,26	0,32	0,39	0,49	0,73

$$Q1 = 0,08$$

$$LI = 0,08 - 1,5 (0,32 - 0,08) = -0,28$$

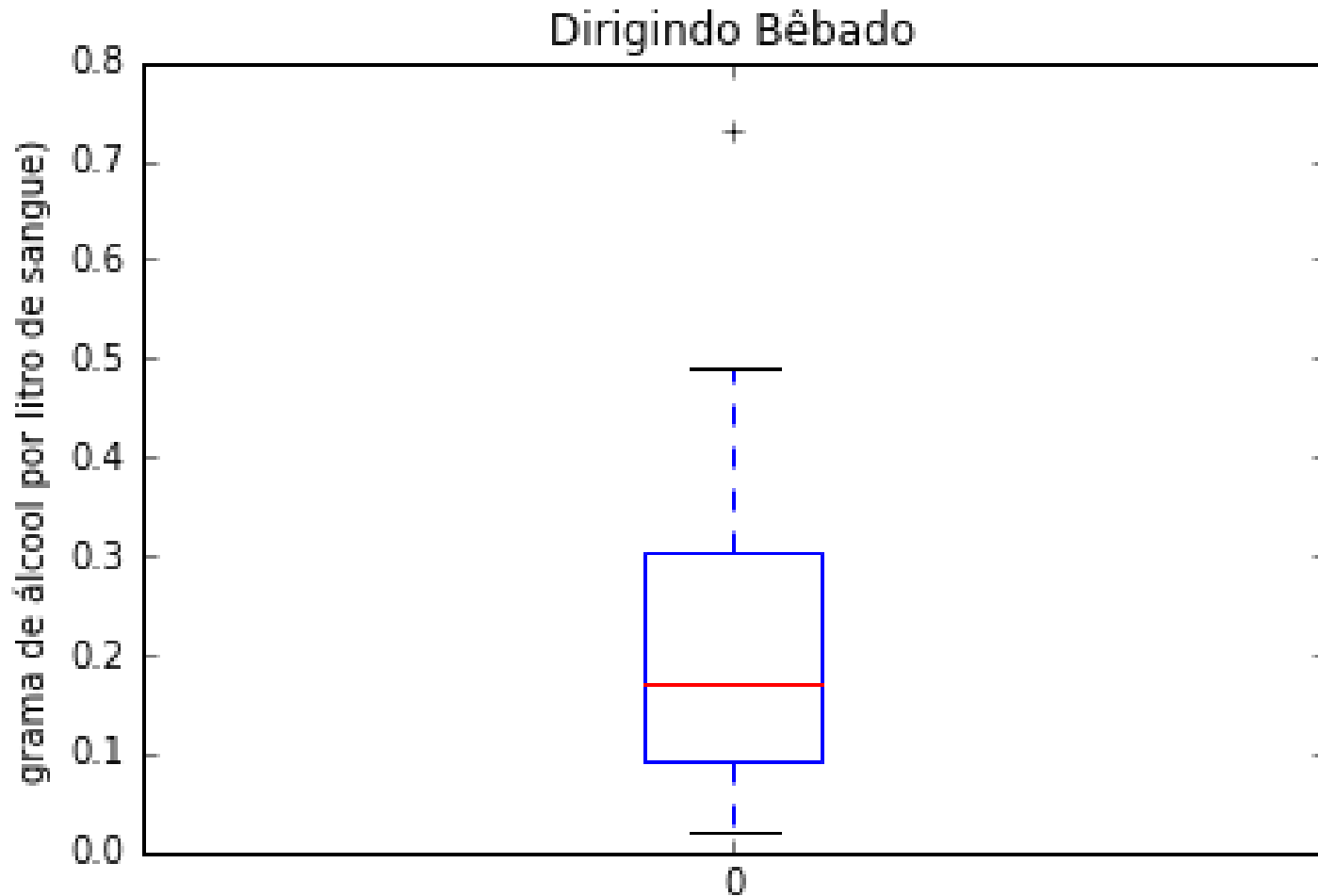
$$Q2 = 0,17$$

$$LS = 0,32 + 1,5 (0,32 - 0,08) = 0,68$$

$$Q3 = 0,32$$

$$IQ = Q3 - Q1 = 0,32 - 0,08 = 0,24$$

Boxplot



Comando Python:

`variaveis.plot(kind='box')`

ou

`variáveis.plot.box()`

Atividade com LIFE... (variável quantitativa)

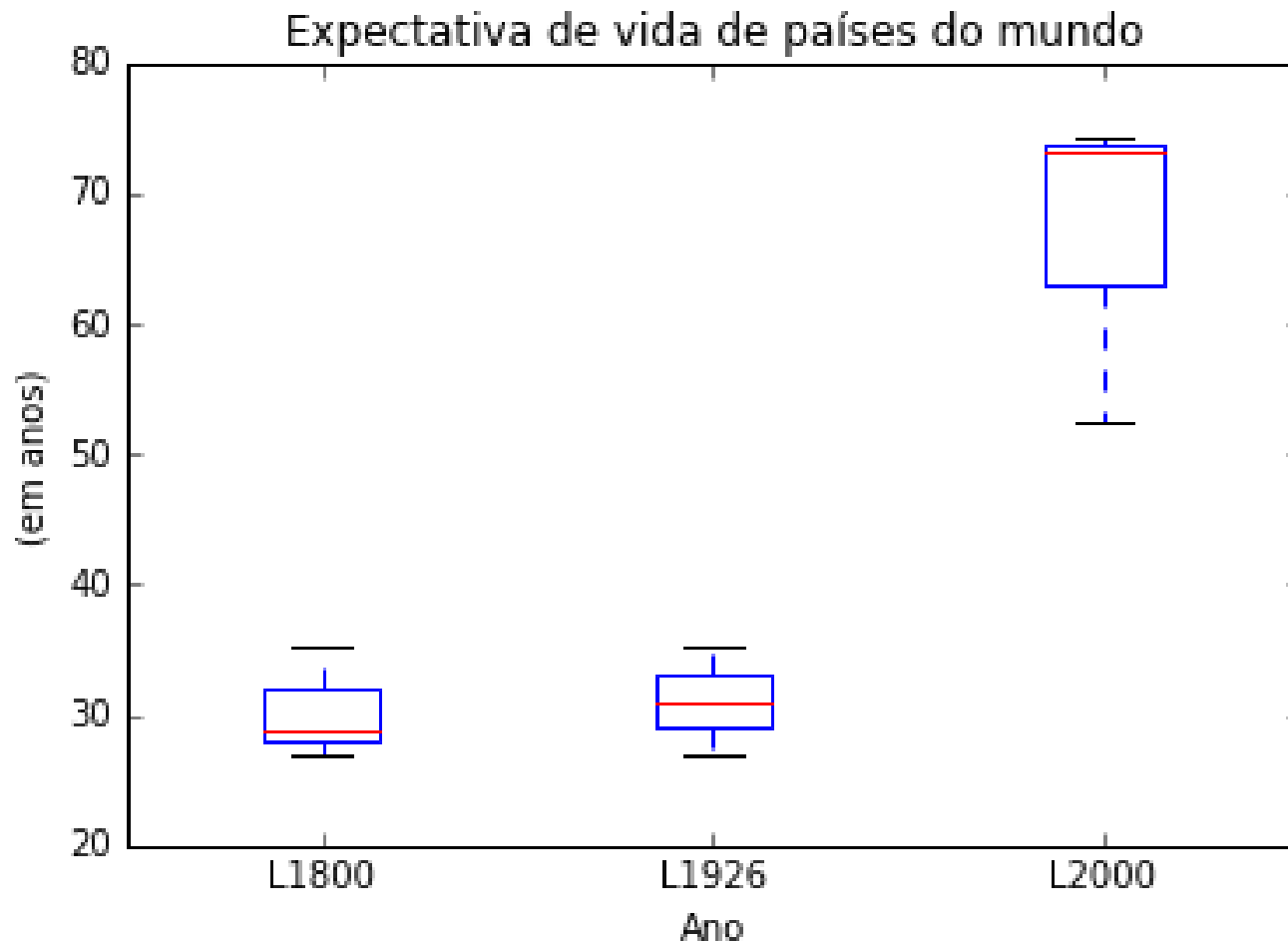
10 minutos:

Análise descritiva nas expectativas de vida de países em três anos diferentes.

Arquivo:

```
Aula04_Atividade1_VariáveisQuantitativas_LIFE.ipynb
```

Boxplot



Comando Python:

`variaveis.plot(kind='box')`

ou

`variáveis.plot.box()`

Atividade com Salários...

(variável quantitativa)

30 minutos:

Análise descritiva de salários de quatro profissões diferentes.

Arquivo:

Aula04_Atividade2_VariáveisQuantitativas_SALARIOS.ipynb

APS 2 – Check durante próxima semana.

Devem apresentar aos ninjas:

1. DataFrame filtrado e suficiente. Filtrado significa ter pessoas com o perfil decorrente ao objetivo estipulado ao PROJETO1 (pode haver mais filtros). Suficiente significa que contem dados suficientes para uma análise após a filtragem. CUIDADO COM DROPNA!!
2. Algumas análises e gráficos univariados de pelo menos uma variável quantitativa e uma variável qualitativa.
3. Formulação da pergunta (ou perguntas) que o projeto faz ou responde, demonstrando que foi encontrado um foco dentro do problema escolhido. PERGUNTA NÃO PRECISA SER TÍTULO DO PROJETO (alinhado ao objetivo).
4. Ler a rubrica e explicar para o Ninja em que ponto o seu projeto está no dia do check e o que ainda precisa fazer para finalizar o projeto.

Os horários para CHECK ainda serão liberados, mas acontecerão durante a próxima semana! Enviaremos em breve!
Deverão chegar ATÉ 20 minutos após início do horário de Check.

Preparo para próxima aula

Os alunos devem se preparar com:

AULA ESTUDIO