

Análise e Visualização de Dados

MP40122-123 - Técnicas De Data Mining e Machine Learning Aplicadas às Ciências Biomédicas

Prof. Mateus Grellert



MESTRADO
ENGENHARIA ELETRÔNICA
E COMPUTAÇÃO

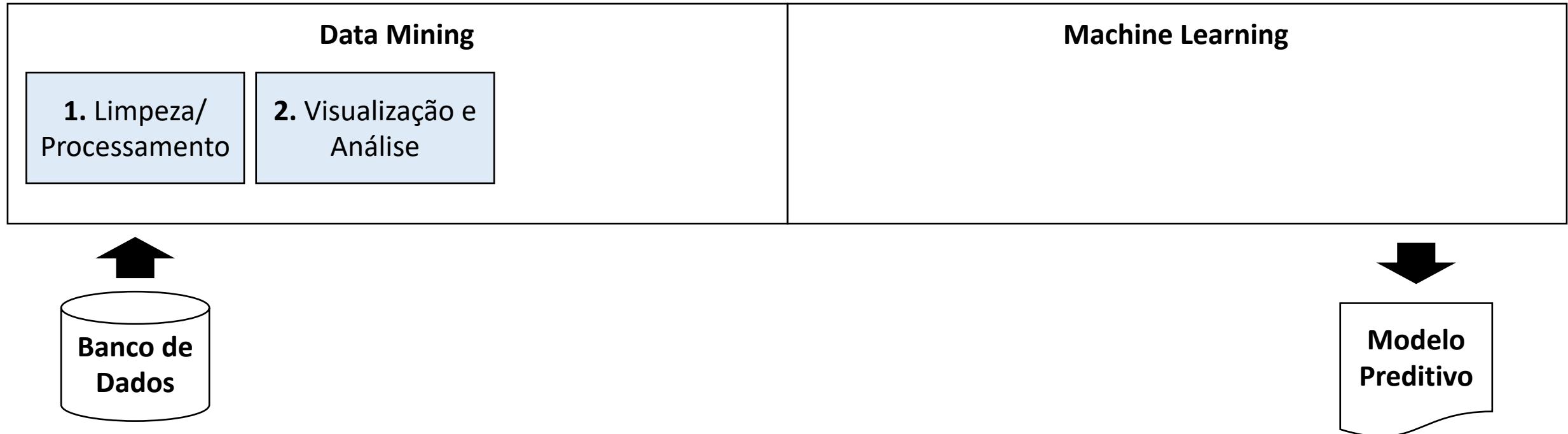
Na aula anterior

- Python icon Fluxograma do desenvolvimento de aplicações de Machine Learning
- Python icon Ferramentas para ML (Python)
- Python icon Importando bancos em Python
- Python icon Técnicas de pré-processamento:
 - Python icon Imputação
 - Python icon Remoção de outliers
 - Python icon Denoising



Desenvolvimento de Aplicações DM/ML

🐍 Fluxo de projeto integrando DM e ML:



Análise de Atributos

- Python Antes de iniciar o treinamento de modelos preditivos, podemos fazer uma análise de cada atributo
- Python Vantagens:
 - Python Detectar atributos ruidosos, inconsistentes
 - Python Conhecer a distribuição dos atributos
 - Python Aumentar o conhecimento sobre o problema
 - Python Desenvolver novos atributos com base no conjunto inicial

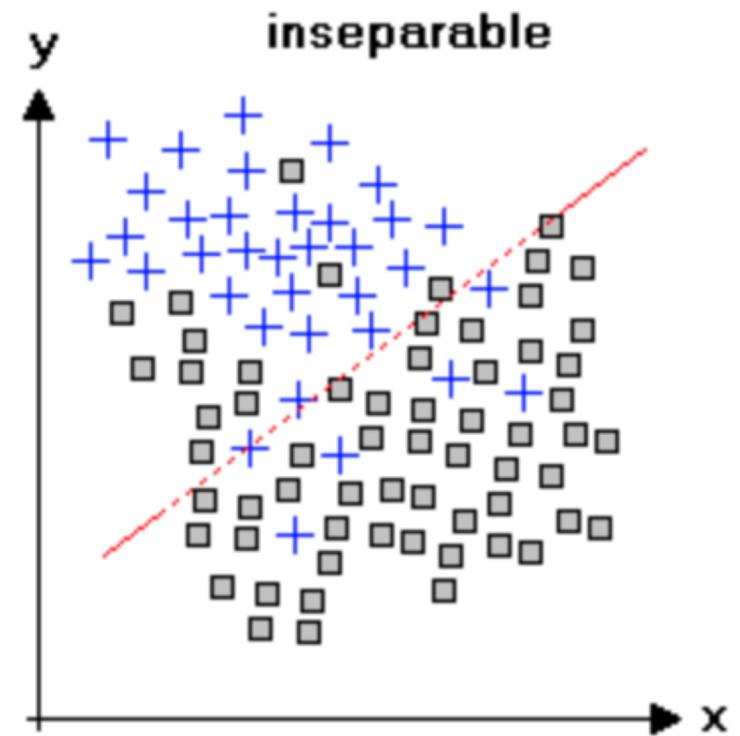
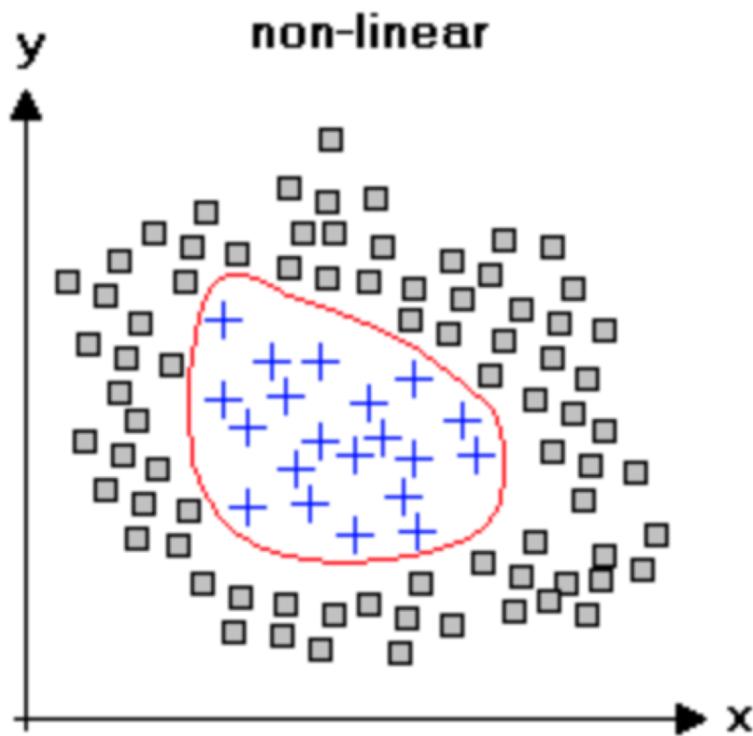
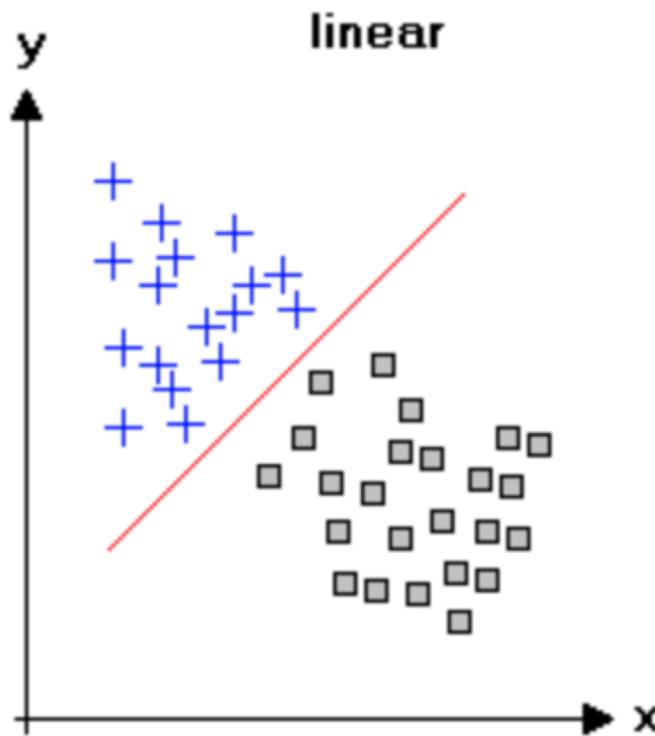


Análise de Atributos

- Python icon: O que define um atributo como **importante**?
- Python icon: Várias métricas
- Python icon: Uma forma comum de avaliar inicialmente se um atributo é importante se refere a **quão separável** são os grupos de desfecho sabendo o valor desse atributo



Análise de Atributos



Análise de Atributos – 5 Number Summary

- 🐍 Uma forma resumida e poderosa de descrever os atributos
- 🐍 Contém 5 valores: **mínimo, quartil inferior, mediana, quartil superior, máximo**

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Análise de Atributos – 5 Number Summary

- Python Com esses valores, conseguimos saber as seguintes características dos nossos atributos:
 - Python Localização (da mediana)
 - Python Espalhamento (com base nos quartis)
 - Python Limites (com mínimo e máximo)
 - Python Testes de normalidade
 - Python Detecção de outliers

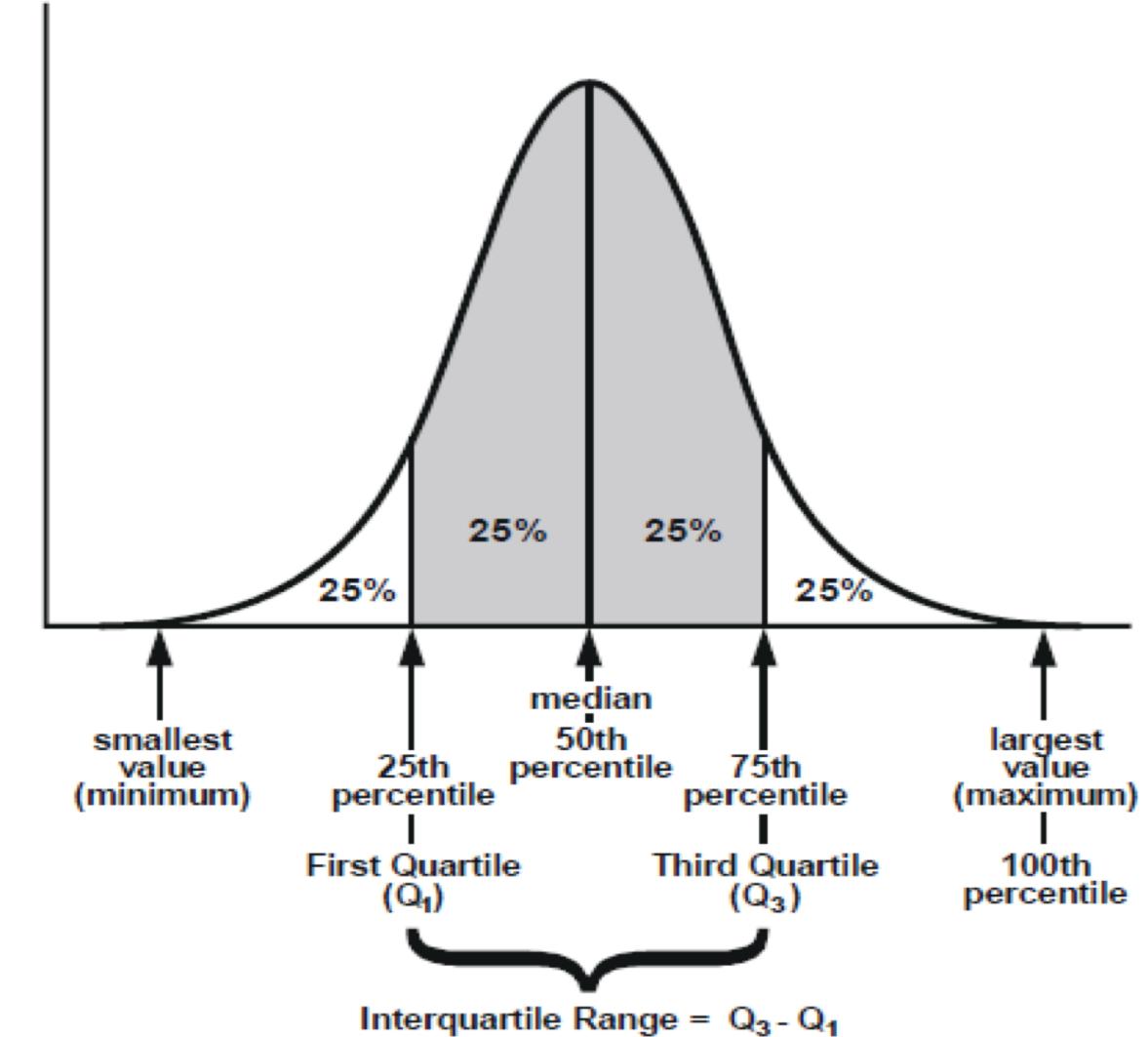


Análise de Atributos – 5 Number Summary

Python icon: Os 5 números de uma distribuição Gaussiana

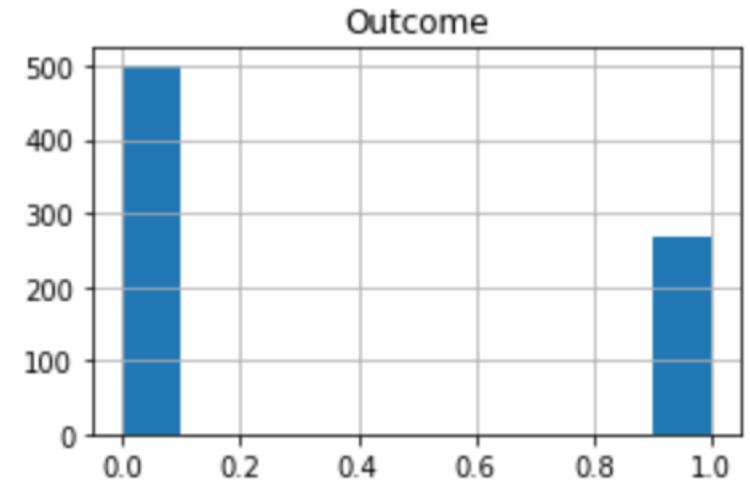
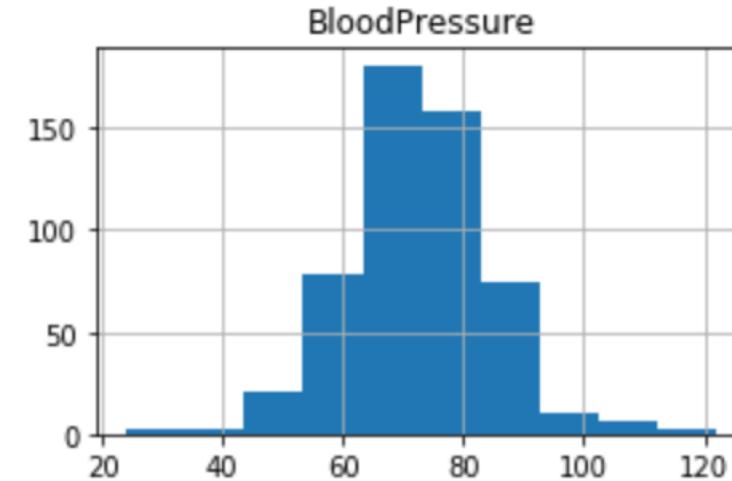
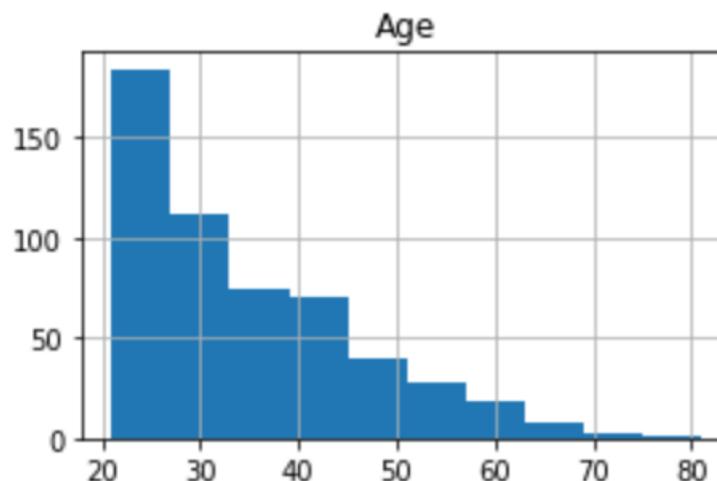
Python icon: $IQR = Q_3 - Q_1$

Python icon: 50% da distribuição se encontra no IQR



Análise de Atributos – Histograma

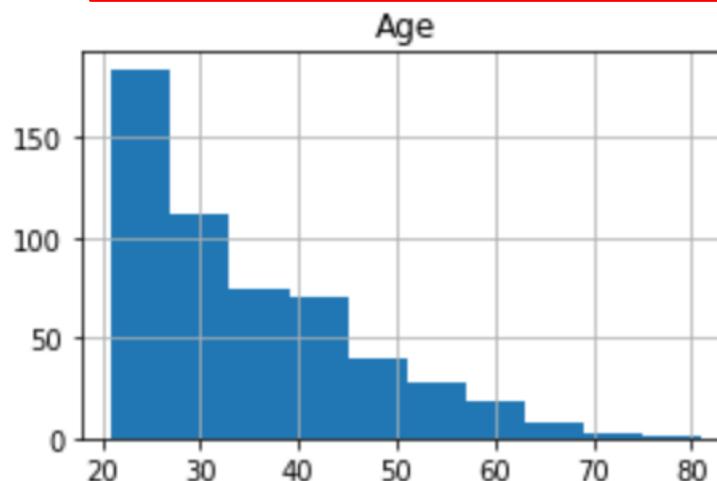
🐍 **Histogramas** representam a ocorrência para cada intervalo de valores do atributo



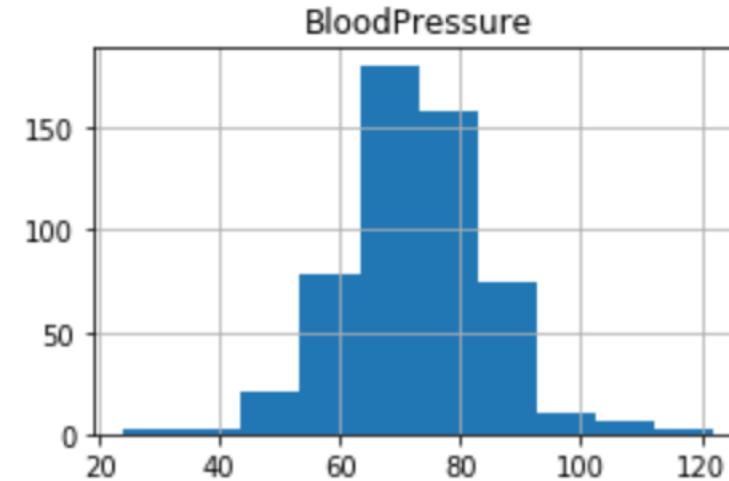
Análise de Atributos – Histograma

 **Histogramas** representam a ocorrência para cada intervalo de valores do atributo

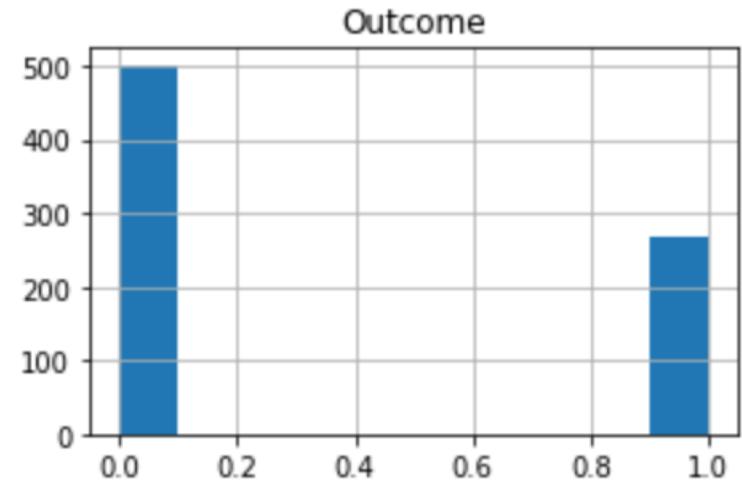
distribuição deslocada à esquerda



distribuição normal

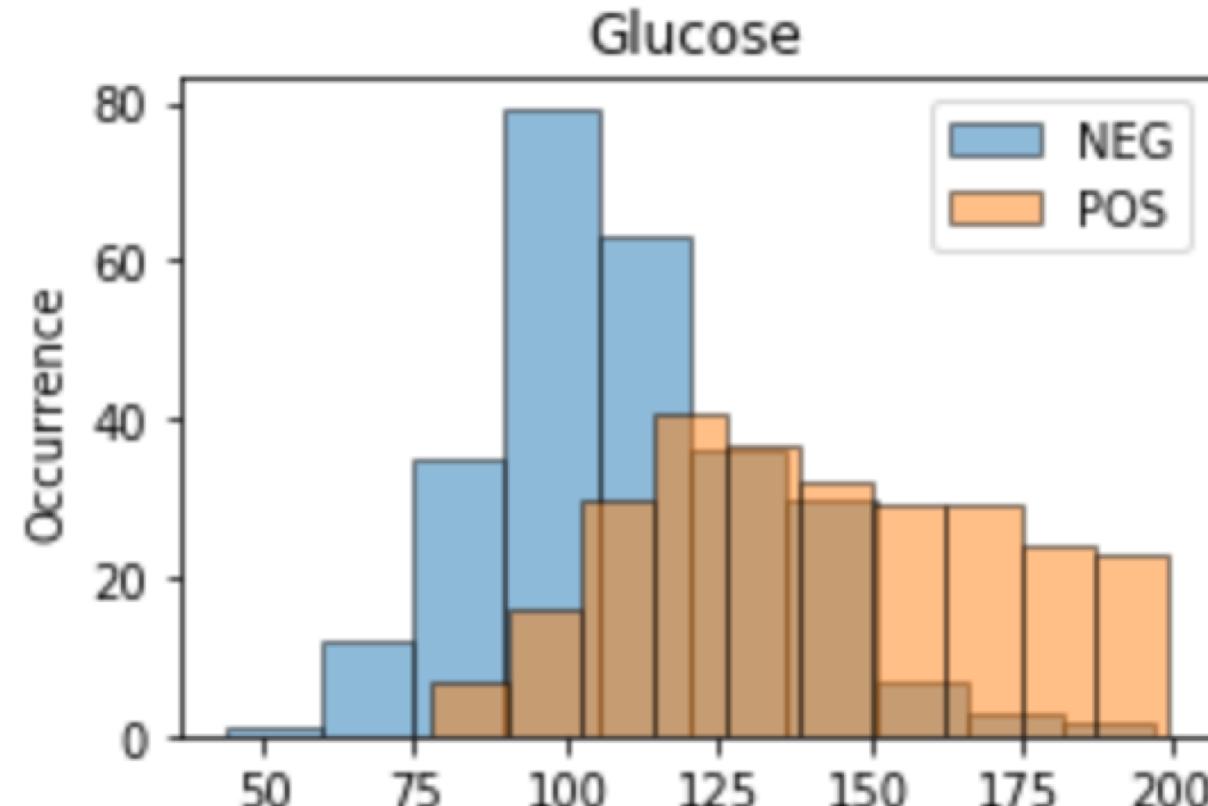


distribuição Binomial



Análise de Atributos – Histograma

🐍 Um **histograma agrupado** por desfecho é melhor para detectar a relação entre um atributo e o desfecho



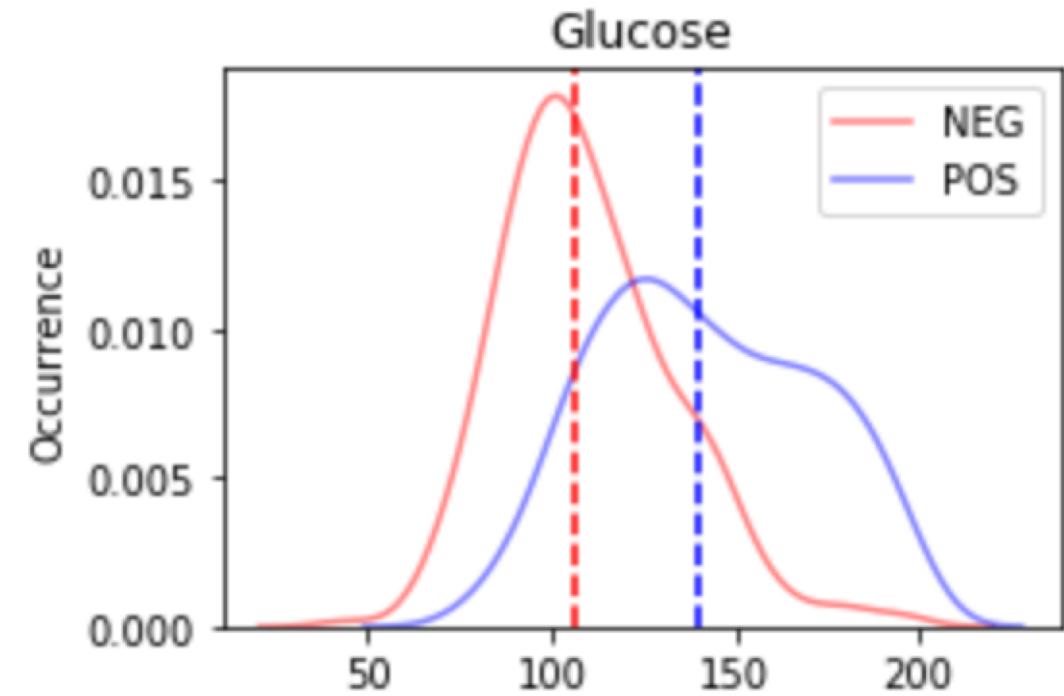
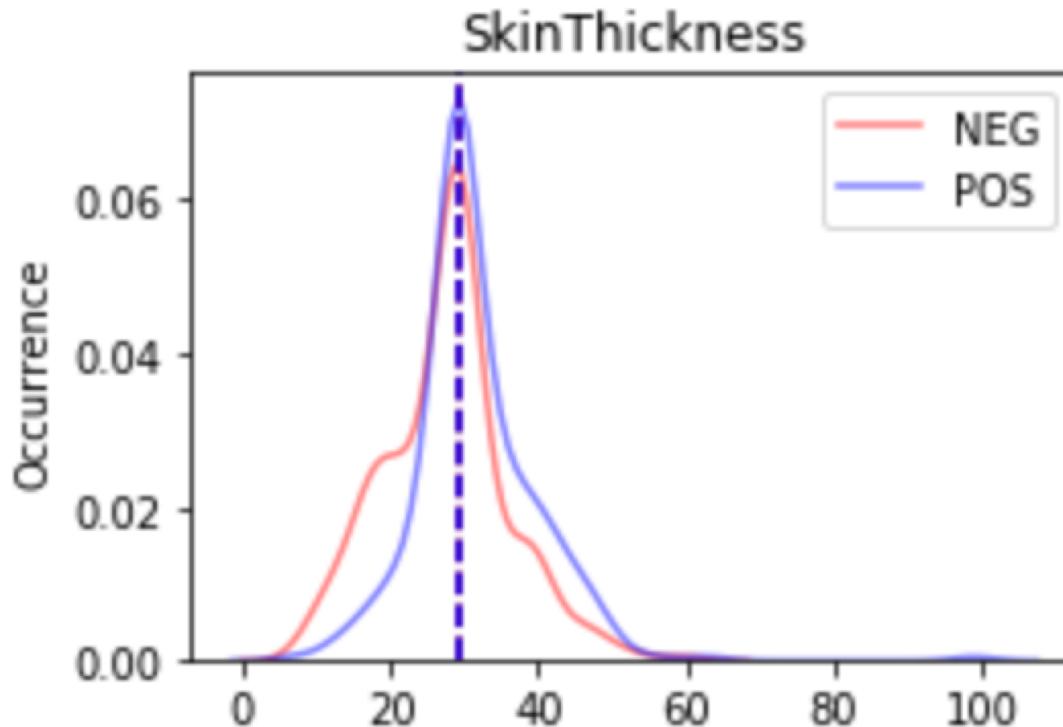
Análise de Atributos – Histograma

- Python icon: Histogramas podem ajudar a detectar **outliers** e testar a **normalidade** da distribuição
- Python icon: Por que é importante saber se uma distribuição é normal (ou paramétrica)?
 - Python icon: Isso vai determinar que tipo de teste estatístico podemos usar
- Python icon: Por exemplo, se quisermos saber a correlação entre duas variáveis, usamos:
 - Python icon: Pearson se a distribuição é **paramétrica**
 - Python icon: Spearman se for **não paramétrica**



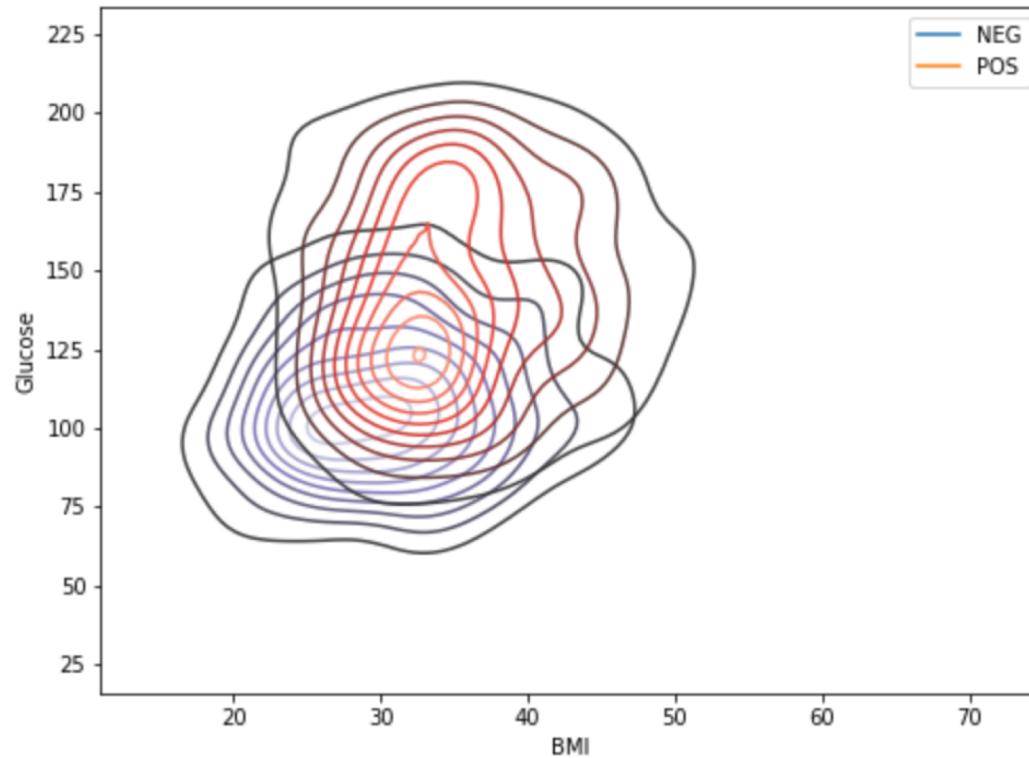
Análise de Atributos – PDF

🐍 Outra forma de avaliar a distribuição é de acordo com a PDF de cada grupo (POS/NEG)



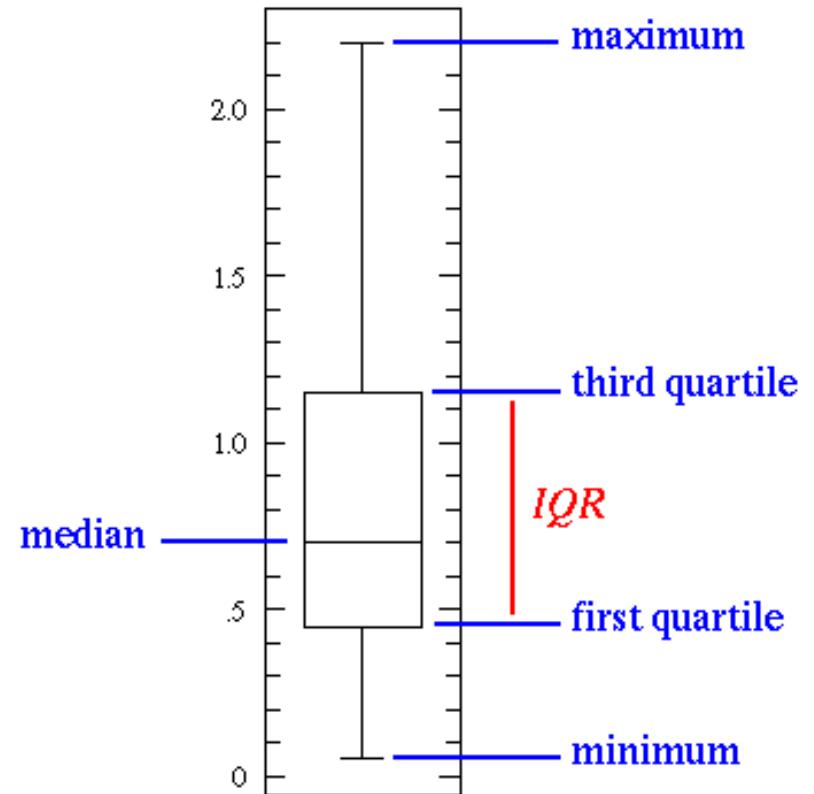
Análise de Atributos – PDF

🐍 Visualizar a PDF 2D pode ajudar a entender se duas variáveis ajudam a separar nossos dados!



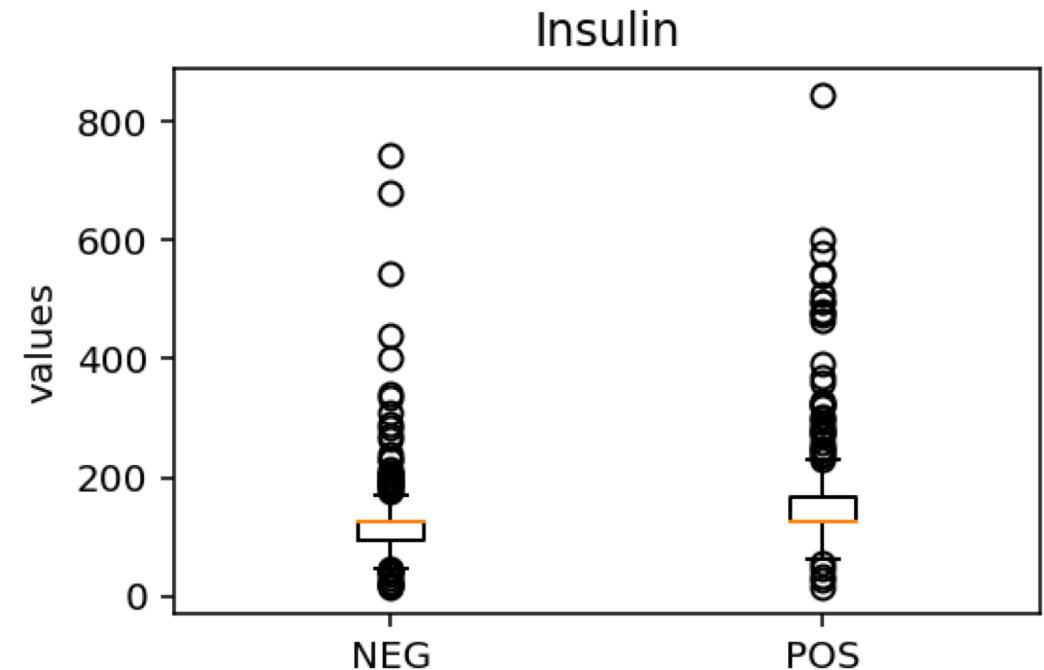
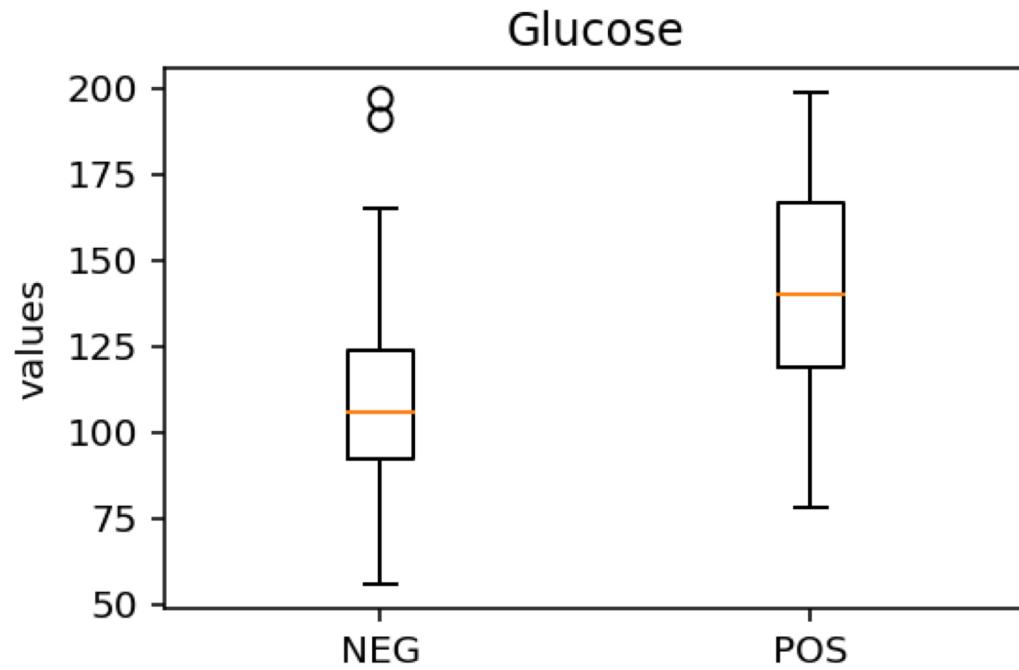
Análise de Atributos – PDF

- Python Outra ferramenta de visualização muito útil é gráfico de caixas (box plots)
- Python Um box plot contém a representação visual do **5-number summary**



Análise de Atributos – PDF

🐍 Exemplo de **box plots** pros atributos do PIMA



Análise de correlação

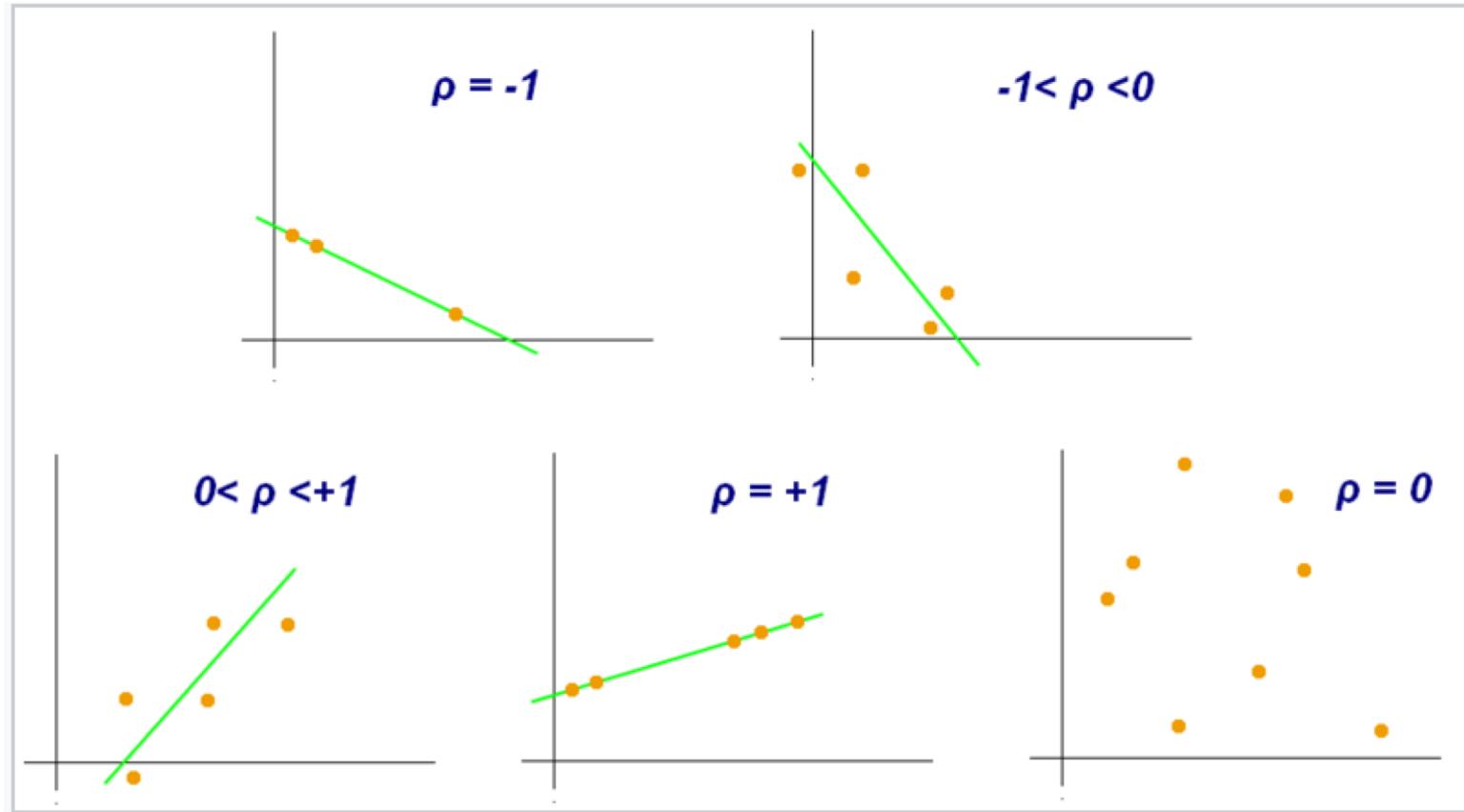
- Python Até agora não temos nenhuma métrica quantitativa para mensurar a relação entre duas
- Python A **correlação de Pearson** (ρ) é uma possível forma de avaliar isso

$$\rho_{X,Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$



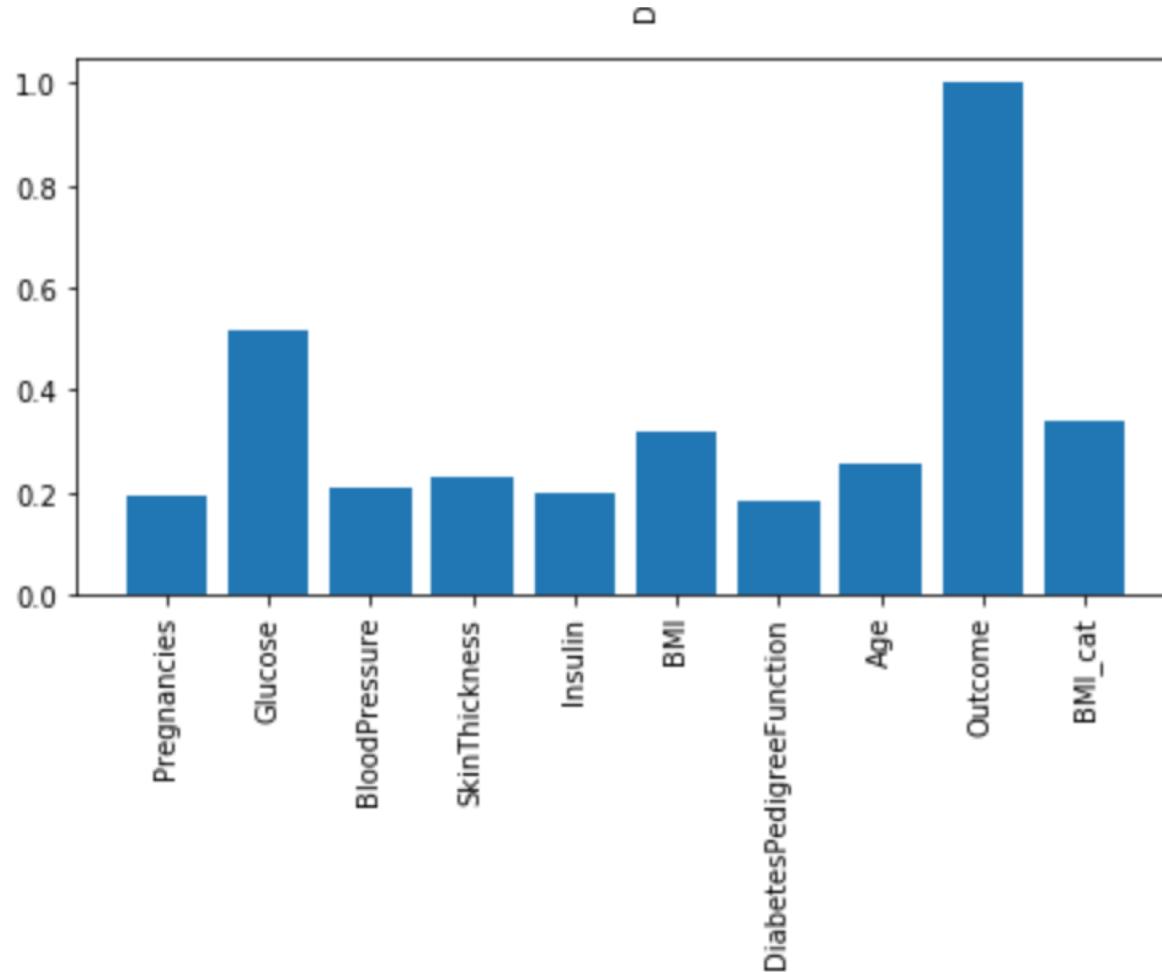
Análise de correlação

Python Exemplos de correlação para diferentes tipos de distribuição



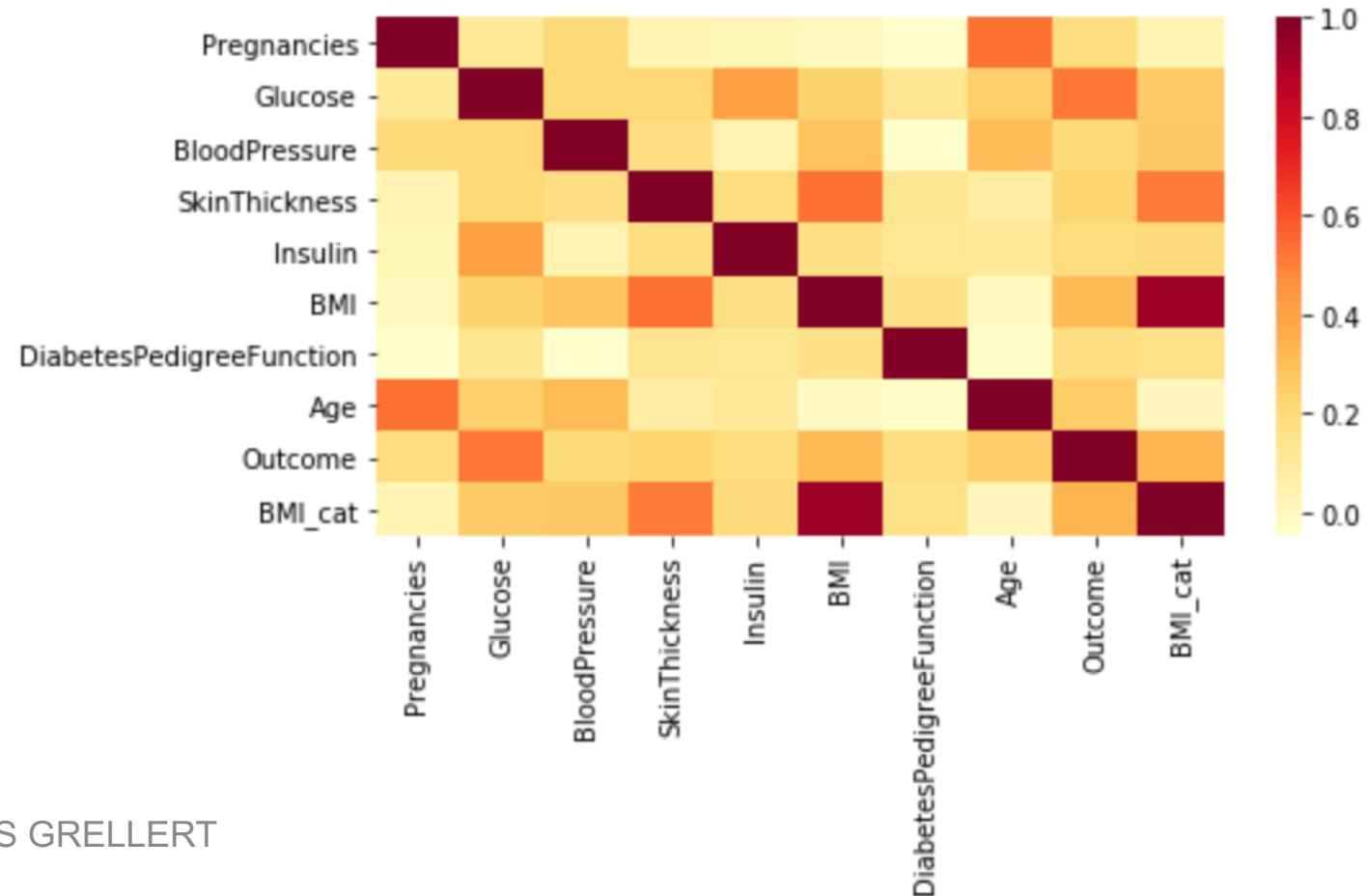
Análise de correlação

🐍 Exemplo para nossos conhecidos atributos: ABS (ρ)



Análise de correlação

🐍 A **matriz de correlação** contém as relações entre todos os atributos do banco (incluindo o desfecho)(



Análise de correlação

Python Interpretação da correlação (regra de dedão)

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Fonte: Mukaka, M. (2012). A guide to appropriate use of Correlation coefficient in medical research. *Malawi Medical Journal : The Journal of Medical Association of Malawi*, 24(3), 69–71.



Testes Estatísticos

- Python A pesquisa estatística na saúde envolve muitas vezes um simples teste de hipótese
- Python Para provar, por exemplo que Glicose tem um efeito significativo na Diabetes, podemos usar um teste de **regressão logística**
- Python O teste certo depende do tipo da variável dependente e da saída



Testes Estatísticos

		Outcome variable					
		Nominal	Categorical (>2 Categories)	Ordinal	Quantitative Discrete	Quantitative Non-Normal	Quantitative Normal
Input Variable	Nominal	χ^2 or Fisher's	χ^2	χ^2 -trend or Mann-Whitney	Mann-Whitney	Mann-Whitney or log-rank ^a	Student's <i>t</i> test
	Categorical (2>categories)	χ^2	χ^2	Kruskal-Wallis ^b	Kruskal-Wallis ^b	Kruskal-Wallis ^b	Analysis of variance ^c
	Ordinal (Ordered categories)	χ^2 -trend or Mann-Whitney	e	Spearman rank	Spearman rank	Spearman rank	Spearman rank or linear regression ^d
	Quantitative Discrete	Logistic regression	e	e	Spearman rank	Spearman rank	Spearman rank or linear regression ^d
	Quantitative non-Normal	Logistic regression	e	e	e	Plot data and Pearson or Spearman rank	Plot data and Pearson or Spearman rank and linear regression
	Quantitative Normal	Logistic regression	e	e	e	Linear regression ^d	Pearson and linear regression

Fonte: “Parametric and Non-parametric tests for comparing two or more groups”, [link](#)



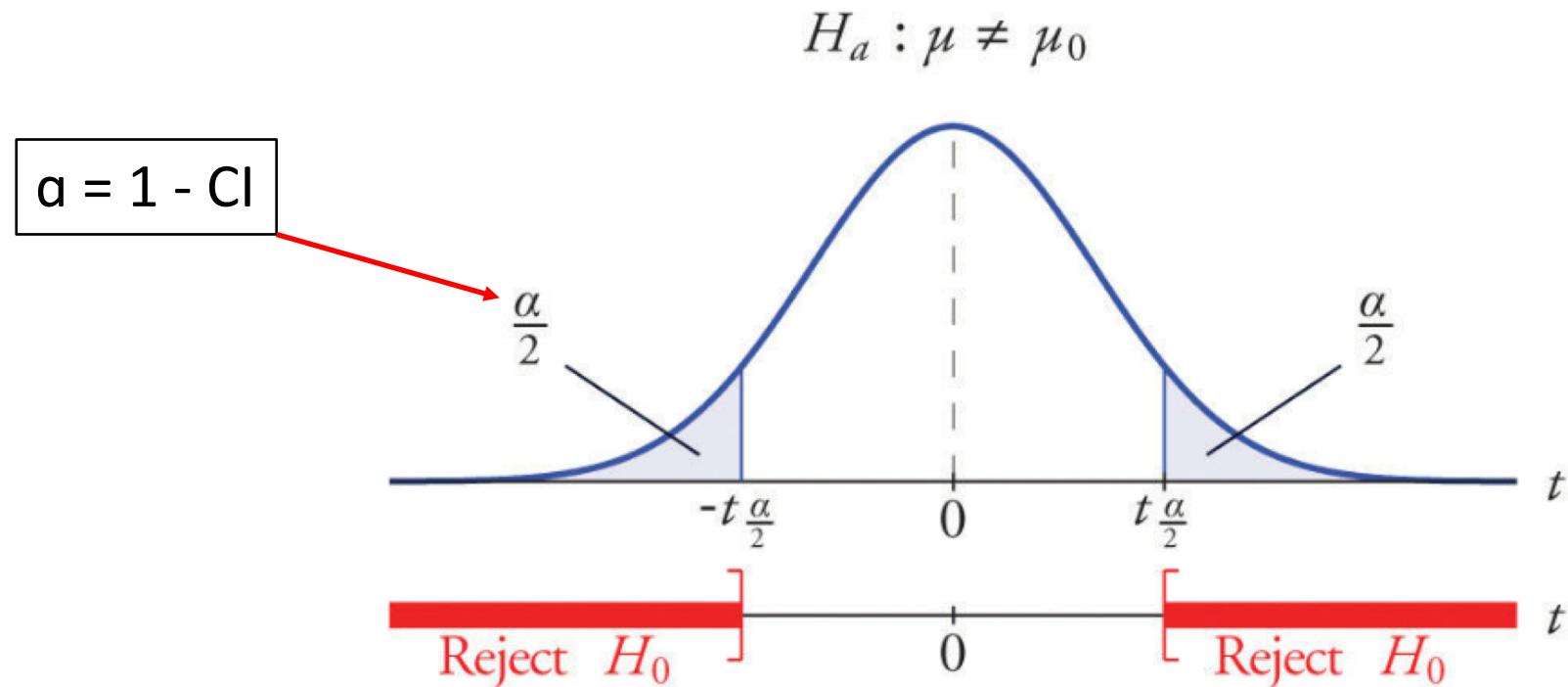
Testes Estatísticos

- Python Os testes seguem a forma: hipótese nula (H_0) contra uma hipótese alternativa (H_a)
- Python Normalmente $H_0 = \text{o atributo não é significativo para o desfecho}$
- Python Os testes computam um **valor estatístico** e compararam com os valores críticos para decidir se a H_0 deve ou não ser rejeitada
- Python Os valores críticos dependem do intervalo de confiança (a porcentagem da população que queremos explicar)
- Python Normalmente o CI é **95%**



Testes Estatísticos

Python Exemplo de um teste de hipótese



Testes Estatísticos

- Python Exemplo:
 - Queremos saber se o atributo **Glicose** tem efeito sobre nosso desfecho de **diabetes**
 - Como Glicose é um atributo numérico e nosso desfecho é categórico, vamos usar **Regressão Logística**

Testes Estatísticos

Python Resultado de uma regressão logística

Optimization terminated successfully.

Current function value: 0.549856

Iterations 6

Logit Regression Results

Dep. Variable:	Outcome	No. Observations:	536
Model:	Logit	Df Residuals:	534
Method:	MLE	Df Model:	1
Date:	Fri, 26 Oct 2018	Pseudo R-squ.:	0.2067
Time:	19:40:07	Log-Likelihood:	-294.72
converged:	True	LL-Null:	-371.53
		LLR p-value:	2.820e-35

	coef	std err	z	P> z	[0.025	0.975]
const	-5.0963	0.500	-10.187	0.000	-6.077	-4.116
Glucose	0.0408	0.004	10.256	0.000	0.033	0.049

Efeito da variável Glucose no Outcome

P < 0.05 → resultado significativo



Testes Estatísticos

Python Odds Ratio (OR): medida comum para detectar o efeito da variável

$$OR(V) = e^{Coeff_V}$$

- Python Interpretação é a seguinte: cada aumento de unidade na variável V aumenta o desfecho em OR(V) vezes
- Python No nosso caso: cada aumento unitário na Glicose aumenta as chances de um diagnóstico positivo em 1.041%



Conclusão

- 🐍 As técnicas de hoje só ajudam a fazer análises estatísticas com poucos parâmetros
- 🐍 Vamos aprender nas próximas aulas como **combinar parâmetros** para gerar modelos preditivos poderosos **mesmo que os atributos sejam fracamente correlacionados** com o desfecho



Próxima Aula



Treinamento de Modelos Preditivos



PROF. MATEUS GRELLERT