Anterior
Desafio

Próximo
Conteúdo do Livro

Infográfico

Para análises de dados usando correlações, o analista precisa do conhecimento claro do comportamento dos cálculos de correlação perante o conjunto de dados estudado. Além disso, é importante o uso de bibliotecas matemáticas pré-implementadas. Isso facilita na redução da *escritabilidade* do desenvolvedor (ou seja, o quanto a biblioteca pode ser usada para a análise de dados de um domínio específico) e permite a interpretação dos resultados de forma mais clara.

No Infográfico a seguir, são apresentadas definições sucintas de correlações com dicas de código usando Pandas, Numpy e Seaborn.

Correlação é uma medida de quão bem duas variáveis estão relacionadas entre si. Análises de correlações estão presentes no dia a dia em diversas áreas da ciência.

Essa medida pode ser aplicada na Medicina: você pode, por exemplo, correlacionar a idade de uma pessoa com seus níveis de açúcar no sangue. Aqui, as unidades são completamente diferentes; a idade é medida em anos e o nível de açúcar no sangue é medido em mmol/L (uma medida de concentração).

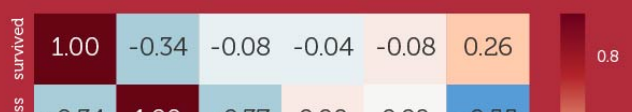
Na vida pessoal: medir seu salario pessoal com gastos pessoais. Na medida que seu salário aumenta em seu emprego, seus gastos pessoais também aumentam.

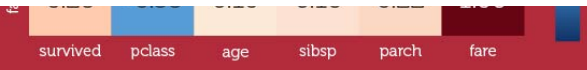
Em institutos educacionais: o tempo de estudo dos alunos com o desempenho em testes avaliativos ou o tempo de prova que o aluno fez o teste, com sua respectiva nota.

Esse documento apresenta o conceito de correlação de Pearson e funções usando Pandas, Numpy e Seaborn, Inicialmente na correlação de Pearson, para variáveis numéricas. Após, é explicada a correlação para variáveis categóricas utilizando tabelas cruzadas. Por fim, é apresentado o conceito de *multicolinearidade*, cujo conceito é aplicado em situações onde duas ou mais variáveis independentes apresentam uma forte correlação entre si.

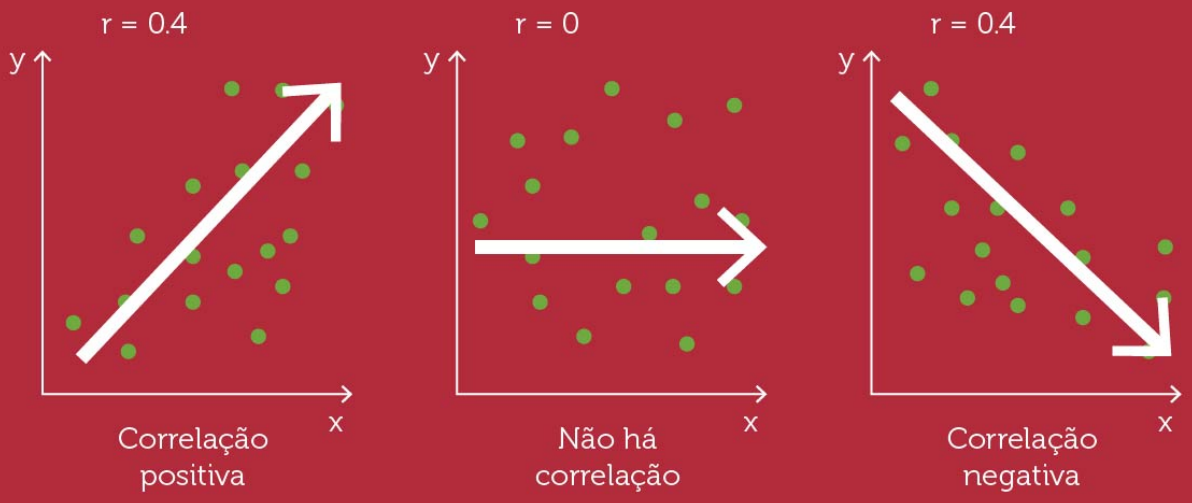
CORRELAÇÃO DE PEARSON

Cálculo da correlação de Pearson (para variáveis numéricas).





- Um coeficiente de correlação de 1 significa que, para cada crescimento positivo em uma variável, como x, há um crescimento positivo na mesma proporção de y. Um maior consumo de energia, por exemplo, aumenta em quase perfeita correlação com o preço na fatura do mês da conta de luz.
- Um coeficiente de correlação de -1 significa que, para cada aumento positivo em uma variável, há uma decrescimento negativa de uma proporção fixa na outra. Por exemplo, quanto maior a velocidade do automóvel, diminui-se em (quase) perfeita correlação o tempo de chegada ao destino.
- Um coeficiente de correlação zero significa que, para cada aumento, não há um aumento positivo ou negativo da correlação. As variáveis x e y não estão relacionadas.



CORRELAÇÃO PARA VARIÁVEIS CATEGÓRICAS E TABULAÇÃO CRUZADA

Para esse cenário, é necessária a manipulação de *dataframes* Pandas com a função *crosstab*. Essa função permite o relacionamento entra variáveis categóricas utilizando alguma métrica matemática (soma, media, etc.). Além disso, outra técnica muito comum para esse cenário é a medida de V de Cramer. Essa medida é de correlação simétrica entre variáveis categóricas. Sua fórmula é:

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}}$$

φc representa V de Cramer*.
χ2 é a estatística de teste independente qui-quadrado.
N é o tamanho da amostra envolvida no teste.
k é o menor número de categorias de cada variável.

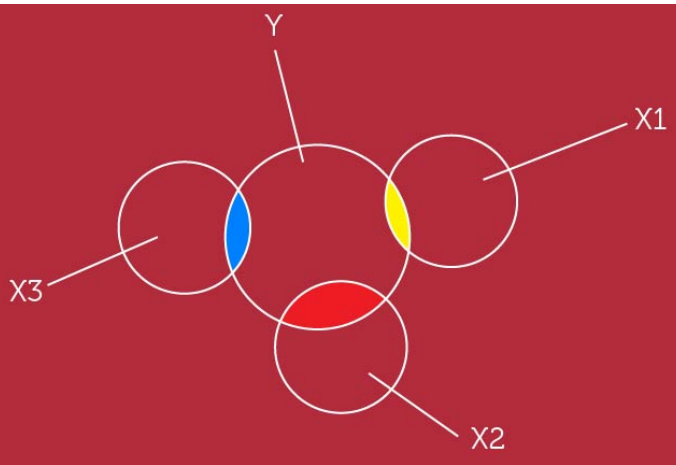
	Base	Age					
		Under 18	18-24	25-34	35-44	45-54	55+
	204	59 29%	43 21%	38 19%	36 18%	20 10%	8 4%
Frequency of visit							
Daily	18 9%	9 4%	5 2%	4 2%	- -	- -	- -
Twice a week	35 17%	11 5%	8 4%	8 4%	7 3%	- -	1 0%
Weekly	64 31%	16 8%	8 4%	16 8%	16 8%	4 2%	4 2%
Monthly	87 43%	23 11%	22 11%	10 5%	13 6%	16 8%	3 1%

MULTICOLINEARIDADE

Resume-se como alta correlação de duas ou mais variáveis independentes.
O fator de inflação da variância (VIF) é a principal métrica para a identificação dessas variáveis.

$$VIF = \frac{1}{1 - R^2}$$

VIF $\leq$ 1 — as variáveis não são correlacionadas.
1< VIF $\leq$ 5 — as variáveis são moderadamente correlacionadas.
VIF > 5 — as variáveis são altamente correlacionadas.



Anterior
Desafio

Próximo
Conteúdo do Livro

Implementar tabelas cruzadas em Python
Ex.:
pd.crosstab(df.coluna1,df.coluna2, margins=True, margins_name="Total").

Identificação de multicolinearidade em Python
statsmodels.stats.outliers_influence.variance_inflation_factor.