DADOS, PYTHON E ESTATÍSTICA: CAMINHOS PARA STEM

Ana Júlia Amaro Pereira Rocha

SEMANA 2

DIA 4

O QUANTO OS DADOS VARIAM

Como já comentei anteriormente, a **média nem sempre resume bem o conjunto de dados.** Duas turmas podem ter a mesma média de notas, por exemplo, mas serem bem diferentes. Uma pode ter notas muito parecidas entre si e a outra pode ter notas muito altas e muito baixas.

Por isso, precisamos das **medidas de dispersão.** Porque elas mostram o quão espalhados os dados estão em relação ao valor central. Vamos focar em duas medidas principais: **variância** e **desvio padrão.**

MEDIDAS DE DISPERSÃO

Variância: mede o quanto, em média, os valores se afastam da média. Quanto maior a variância, "mais diferentes são os valores entre si".

Para População (σ^2):

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- x_i: cada valor do conjunto de dados.
- μ: média populacional.
- N: número total de elementos na população.

Para Amostra (s^2):

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

- x_i: cada valor da amostra.
- $\bar{x}$: média amostral.
- n: número total de elementos na amostra.

MEDIDAS DE DISPERSÃO

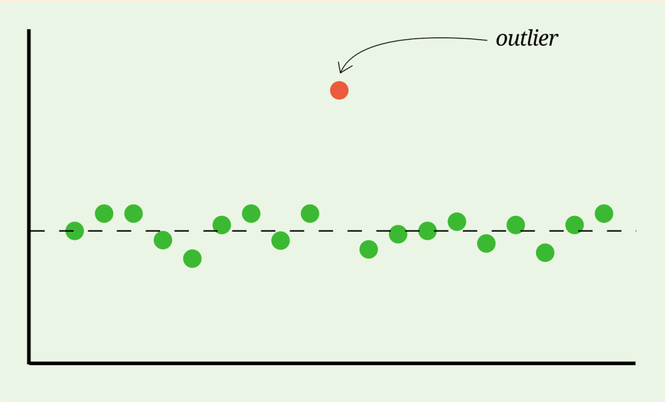
Desvio Padrão: indica, aproximadamente, quanto os dados costumam variar em relação à média. É uma forma mais intuitiva de ver a variância.

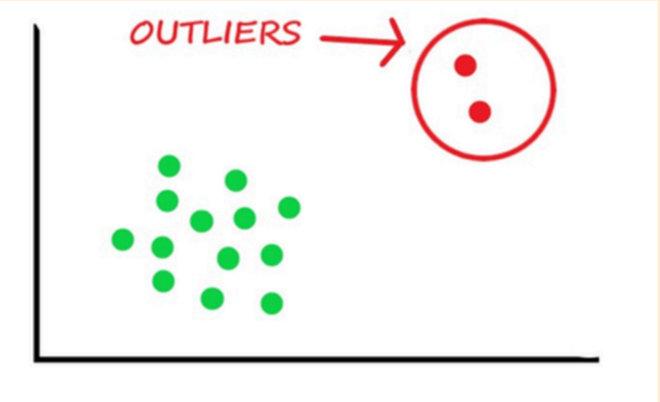
Um **desvio padrão pequeno** indica **dados mais consistentes.**
Um **desvio padrão grande** indica **dados mais irregulares.**

$$\sigma = \sqrt{\sigma^2} \quad \text{ou} \quad s = \sqrt{s^2}$$

OUTLIERS

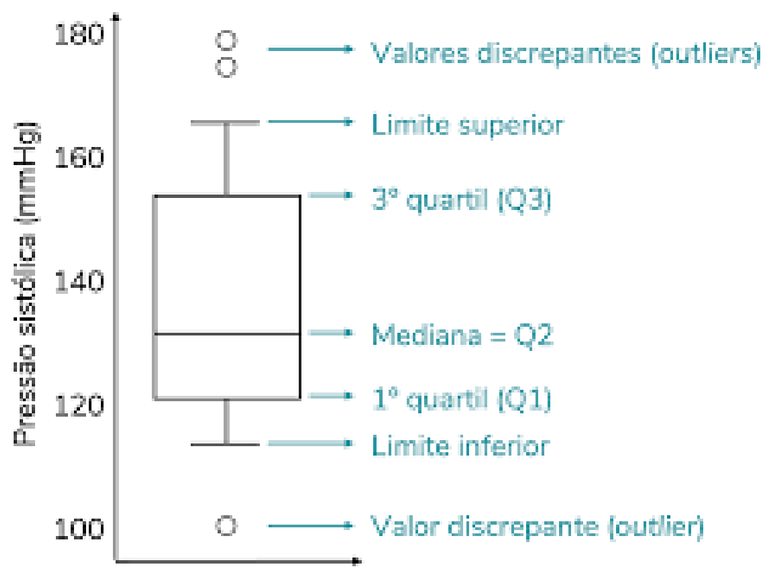
Como já comentei em sala, os **outliers** são valores muito distantes do restante dos dados, podendo acontecer por erros de digitação, erro de coleta ou em situações reais, mas raras.
O "problema" deles é que **podem distorcer médias e análises.** Então, precisamos investigá-los e, assim, até podemos apagá-los em alguns casos.

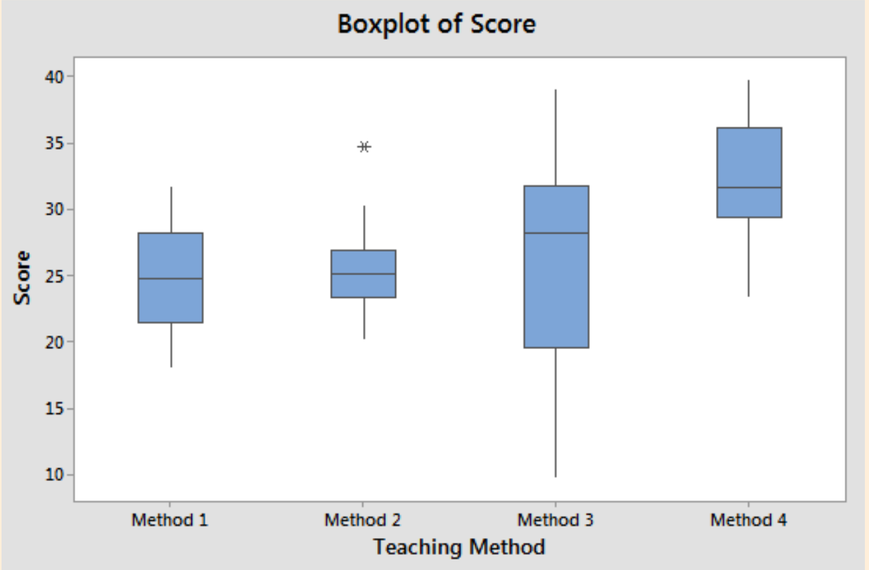




BOXPLOT

É um gráfico que mostra mediana, quartis, dispersão e possíveis outliers. O boxplot **ajuda a visualizar rapidamente como os dados estão distribuídos.**





CORRELAÇÃO VS CAUSALIDADE

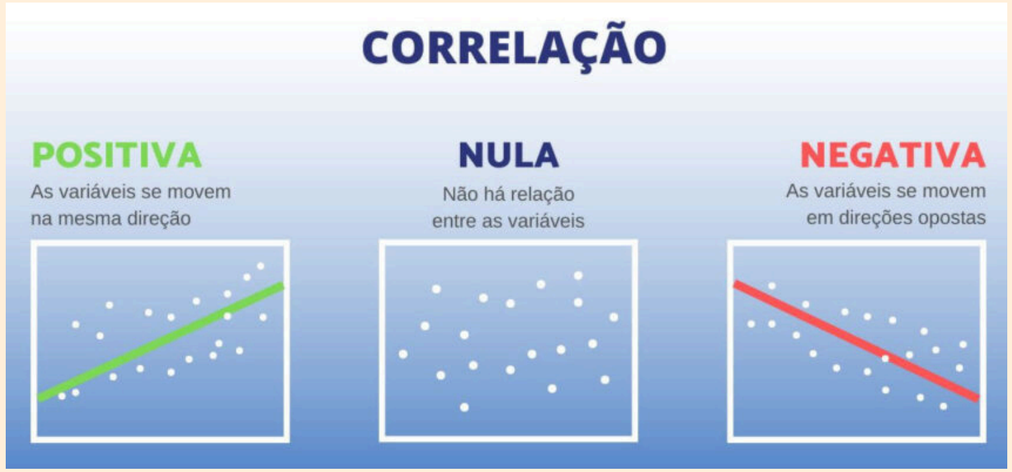
Correlação indica que duas variáveis mudam juntas, podendo acontecer de forma positiva quando ambas aumentam ou de forma negativa quando uma aumenta e a outra diminui. Mas, **correlação não implica causalidade.**

Duas coisas podem ser relacionadas sem que uma cause a outra. Às vezes, existe uma terceira variável “escondida” ou é apenas uma coincidência.

Um exemplo clássico é a relação entre vendas de sorvete e ataques de tubarão. Ambos aumentam no verão, mas sorvete não causa ataques, a terceira variável (calor/verão) causa ambos, levando mais pessoas a tomar sorvete e ir ao mar.

COEFICIENTE DE PEARSON

O coeficiente de correlação de Pearson mede o grau de relação linear entre duas variáveis numéricas, variando de **-1 (relação negativa forte)**, **0 (sem relação)** e **1 (relação positiva forte).**
Não prova causalidade, apenas indica associação.



ESTUDO DE CASO

A seguir temos 10 situações as quais vamos analisar juntos quando se trata de correlação e quando é causalidade.

1 - Vendas de sorvete e casos de insolação aumentam no verão.
Correlação. O calor influencia em ambos.

2 - Quanto mais horas de estudo, maior a nota na prova.
Causalidade, ainda que dependa de outros fatores como qualidade do estudo e do sono também.

3 - Uso de óculos aumenta com o nível de escolaridade.
Correlação. A idade pode ser a “variável escondida”.

ESTUDO DE CASO

4 - Número de ciclistas e acidentes de bicicleta crescem juntos.
Correlação. Mais ciclistas implica mais exposição ao risco.

5 - Quanto mais policiamento em um bairro, maior o número de crimes registrados.
Correlação. O policiamento responde ao crime, não o causa.

6 - Crianças que leem mais têm melhor vocabulário.
Causalidade, embora o ambiente familiar também influencie.

7 - Pessoas que dormem menos usam mais café.
Correlação. Fatores como rotina de trabalho e estresse causam ambos.

ESTUDO DE CASO

8 - Cidades com mais hospitais têm mais mortes registradas.
Correlação. O tamanho da população é a variável escondida.

9 - Alunos que participam de aulas de reforço melhoram as notas.
Causalidade. Embora fatores como motivação e apoio familiar influenciem.

10 - Quanto mais pessoas usam redes sociais, maior o número de diagnósticos de ansiedade.
Correlação. Pode haver influência mútua, mas é difícil estabelecer uma relação de causa, outros fatores estão envolvidos.