**Para saber mais: desvio padrão**

O desvio padrão é uma medida estatística que indica a dispersão ou variabilidade dos dados em relação à média. Ele representa o quão distantes os valores individuais estão da média do conjunto de dados. Quanto maior o desvio padrão, maior é a dispersão dos dados em torno da média; e quanto menor, menor é a dispersão.

O desvio padrão auxilia na compreensão da variabilidade de um conjunto de dados e avalia o quão representativa é a média. Ele nos fornece uma medida da consistência ou inconsistência dos dados em relação à média e é comumente utilizado para medir a confiança e a margem de erro em cálculos estatísticos, que você aprenderá um pouco mais nos próximos cursos da formação.

#### Vamos para um exemplo?

Podemos notar o comportamento do desvio padrão observando o código abaixo que cria um DataFrame que contém a quantidade de visitas em um website por dia da semana em duas semanas:

**import** pandas **as** pd

**import** numpy **as** np

# Criando DataFrame de exemplo

dados = np.array([[100, 120, 90, 110, 130], [80, 150, 70, 140, 60]])

df = pd.DataFrame(dados, index = ["Semana 1", "Semana 2"],

columns= ["Segunda", "Terça", "Quarta", "Quinta", "Sexta"])

# Calculando o desvio padrão das visitas por semana

desvio\_padrao\_semana1 = df.loc["Semana 1"].std()

desvio\_padrao\_semana2 = df.loc["Semana 2"].std()

print(df)

print("\nDesvio Padrão da Semana 1:", desvio\_padrao\_semana1.round(2))

print("Desvio Padrão da Semana 2:", desvio\_padrao\_semana2.round(2))

Observe abaixo a saída desse código:

Segunda Terça Quarta Quinta Sexta

Semana 1 100 120 90 110 130

Semana 2 80 150 70 140 60

Desvio Padrão da Semana 1: 15.81

Desvio Padrão da Semana 2: 41.83

Lendo a resposta gerada pelo código, percebemos que o desvio padrão da **Semana 2** foi maior que o da **Semana 1**. Por quê? Sabendo que estamos trabalhando com dados de uma mesma natureza (quantidade de visitas), isso indica que possuímos uma maior dispersão dos dados de visitas na **Semana 2** em relação à média semanal (100 visitas), enquanto que na **Semana 1** possuímos uma menor dispersão da quantidade de visitas de acordo com a mesma semana (110 visitas).

Trazendo para ciência de dados, podemos interpretar que houve maior instabilidade ou flutuação no número de visitas em diferentes dias da semana na **Semana 2**, o que pode aumentar o interesse em analisar o que pode ter ocorrido e como planejar um maior engajamento para a semana inteira.

**Para saber mais: Kaggle**

Uma das ferramentas que podemos utilizar para explorar os dados em diversos contextos da ciência de dados é o **[Kaggle](https://www.kaggle.com/" \t "_blank)**. Ele é uma plataforma online que conecta entusiastas de ciência de dados e aprendizado de máquina, oferecendo um vasto repositório de competições, conjuntos de dados, notebooks, e uma comunidade ativa para troca de conhecimento e experiências. Este ambiente permite que as pessoas explorem e participem de desafios propostos por empresas e organizações, os quais vão desde previsão de vendas até diagnósticos médicos avançados, oferecendo prêmios para as melhores soluções.

Para estudantes e profissionais da área de ciência de dados, o Kaggle atua como uma ferramenta de aprendizado de mão na massa, permitindo aplicar conceitos aprendidos na área por meio de problemas reais. Podemos acessar uma série de notebooks disponíveis na plataforma compartilhados pela comunidade partindo da análise exploratória com o storytelling dos dados aos modelos de machine learning documentados passo a passo.

Além disso, os desafios do Kaggle possibilitam às pessoas estudantes testar e aplicar o que vem aprendendo em cenários competitivos, com feedback sobre a performance de seus modelos através dos rankings. Essa interação direta com problemas do mundo real e a possibilidade de ver como outros participantes abordaram o mesmo problema são inestimáveis para o desenvolvimento de habilidades práticas em ciência de dados.