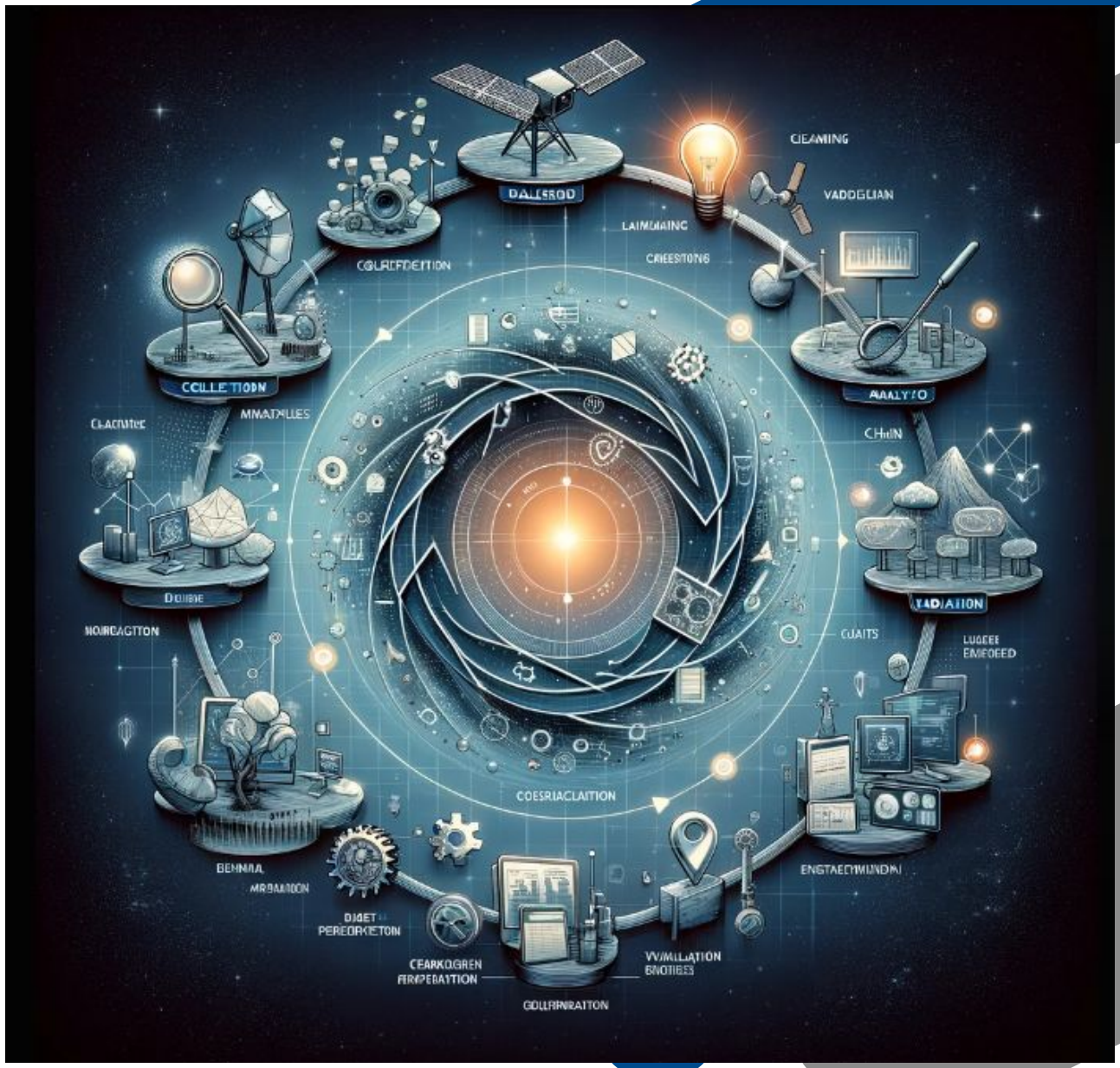
Senac

Fecomércio Sesc

Data Science – Princípios e Técnicas



Abril

2025

Data Science

Onde me encontrar:

https://www.linkedin.com/in/marco-mialaret-junior/

e

https://github.com/MatmJr

Nas aulas passadas

Data Science

A **estatística descritiva** é uma ferramenta essencial nesse processo, utilizando duas abordagens principais:

Data Science

- A abordagem quantitativa, que descreve e resume os dados numericamente.

- A abordagem visual, que ilustra os dados por meio de gráficos e visualizações.

Data Science

Vimos na aula passada:

- A tendência central informa sobre os centros dos dados. Medidas úteis incluem a média, mediana e moda.

Data Science

Veremos na aula de hoje:

- A variabilidade informa sobre a dispersão dos dados. Medidas úteis incluem variância e desvio padrão.

- A correlação (ou variabilidade conjunta) informa sobre a relação entre um par de variáveis em um conjunto de dados. Medidas úteis incluem a covariância e o coeficiente de correlação.

As medidas de Dispersão

Data Science

As medidas de tendência central, por si só, não são suficientes para fornecer uma descrição completa dos dados. É essencial considerar também as medidas de variabilidade, que quantificam a dispersão dos pontos de dados em relação à média. Nesta seção, você aprenderá a identificar e calcular as principais medidas de variabilidade:

Data Science

- Variância
- Desvio padrão
- Amplitude
- Assimetria
- Curtose

Variância

A **variância** é uma medida estatística que indica a dispersão dos valores em um conjunto de dados. Ela calcula o quão distante cada ponto de dados está da média do conjunto, proporcionando uma visão da variabilidade geral. A variância é geralmente representada pelo símbolo σ^2.

Exemplo 1: Determine a variância das transações Pix:

Desvio Padrão

O **desvio padrão** da amostra é uma importante medida de dispersão dos dados, intimamente relacionada à variância da amostra. O desvio padrão, representado por s, é obtido ao se calcular a raiz quadrada positiva da variância da amostra.

Data Science

Essa medida é frequentemente preferida à variância, pois é expressa na mesma unidade que os dados originais, facilitando a interpretação. Após calcular a variância, o desvio padrão pode ser facilmente determinado usando Python.

Data Science

Exemplo 2: Calcule o desvio padrão das transações pix:

Amplitude

Amplitude é uma medida de dispersão estatística que indica a diferença entre o maior e o menor valor em um conjunto de dados. Essa métrica é particularmente útil para entender a escala total dentro da qual os dados variam, oferecendo uma visão rápida da extensão dos valores observados.

Data Science

Para calcular a amplitude, simplesmente subtraímos o valor mínimo do valor máximo encontrado no conjunto de dados. Apesar de sua simplicidade, a amplitude tem limitações significativas, principalmente porque é extremamente sensível a valores atípicos (outliers). Um único valor extremamente alto ou extremamente baixo pode distorcer a percepção da variação geral dos dados.

Assimetria

A **assimetria** é uma medida que indica o grau de desvio na simetria de uma distribuição de dados. Em termos simples, a assimetria quantifica o quanto a distribuição de uma característica se afasta de uma distribuição perfeitamente simétrica.

Data Science

Observação: Valores negativos de assimetria indicam a presença de uma cauda dominante à esquerda da distribuição, como ilustrado no primeiro conjunto de dados. Por outro lado, valores positivos indicam uma cauda mais longa ou mais espessa à direita, conforme demonstrado no segundo conjunto. Uma assimetria próxima de zero (por exemplo, entre -0,5 e 0,5) sugere que a distribuição é relativamente simétrica.

Data Science

Exemplo 4: Determine o coeficiente de assimetria para as transações Pix:

Curtose

Curtose é uma medida estatística que descreve o "achatamento" ou "pico" de uma distribuição em relação à distribuição normal. Outliers em uma amostra impactam a curtose de forma mais significativa do que afetam a assimetria, pois, numa distribuição simétrica, caudas mais pesadas em ambos os extremos elevam a curtose.

Data Science

Em contraste com a assimetria, onde as caudas opostas podem se neutralizar, na curtose ambas contribuem para seu aumento. Diferentemente da média e do desvio padrão, que são expressos nas mesmas unidades dos dados, e da variância, que é expressa no quadrado dessas unidades, a curtose é uma medida adimensional. Ela representa um coeficiente que indica o grau de achatamento da distribuição dos dados.

Exemplo 5: Determine a curtose das transações Pix:

Data Science

O padrão de referência para a curtose é a distribuição normal, que tem uma curtose de 3. Por esse motivo, frequentemente utiliza-se o termo excesso de curtose, que é calculado como curtose menos 3.

Data Science

- Mesocúrtica: Uma distribuição é classificada como mesocúrtica quando sua curtose é aproximadamente igual a 3 (ou um excesso de curtose próximo de 0, usando a medida do pandas). Essa distribuição tem uma forma semelhante à distribuição normal em termos de "pico".

Data Science

- Platicúrtica: Uma distribuição com curtose menor que 3 (ou excesso de curtose negativo no pandas) é chamada de platicúrtica. Comparada à distribuição normal, ela tende a ter caudas mais curtas e finas e um pico central mais baixo e largo.

Data Science

- Leptocúrtica: Quando a curtose é maior que 3 (ou excesso de curtose positivo no pandas), a distribuição é classificada como leptocúrtica. Esse tipo de distribuição possui caudas mais longas e gordas, com um pico central mais alto e agudo em comparação com uma distribuição normal.

Data Science

Ao analisar a assimetria e a curtose das idades, observamos que os dados não estão distribuídos de maneira centralizada e apresentam uma forma verticalmente alongada. Para obter uma visualização mais precisa da distribuição dos dados, é útil empregar uma curva conhecida como função densidade de probabilidade.

Dúvidas?

Marco Mialaret, MSc

Telefone:
81 98160 7018

E-mail:
marco.junior@pe.senac.br