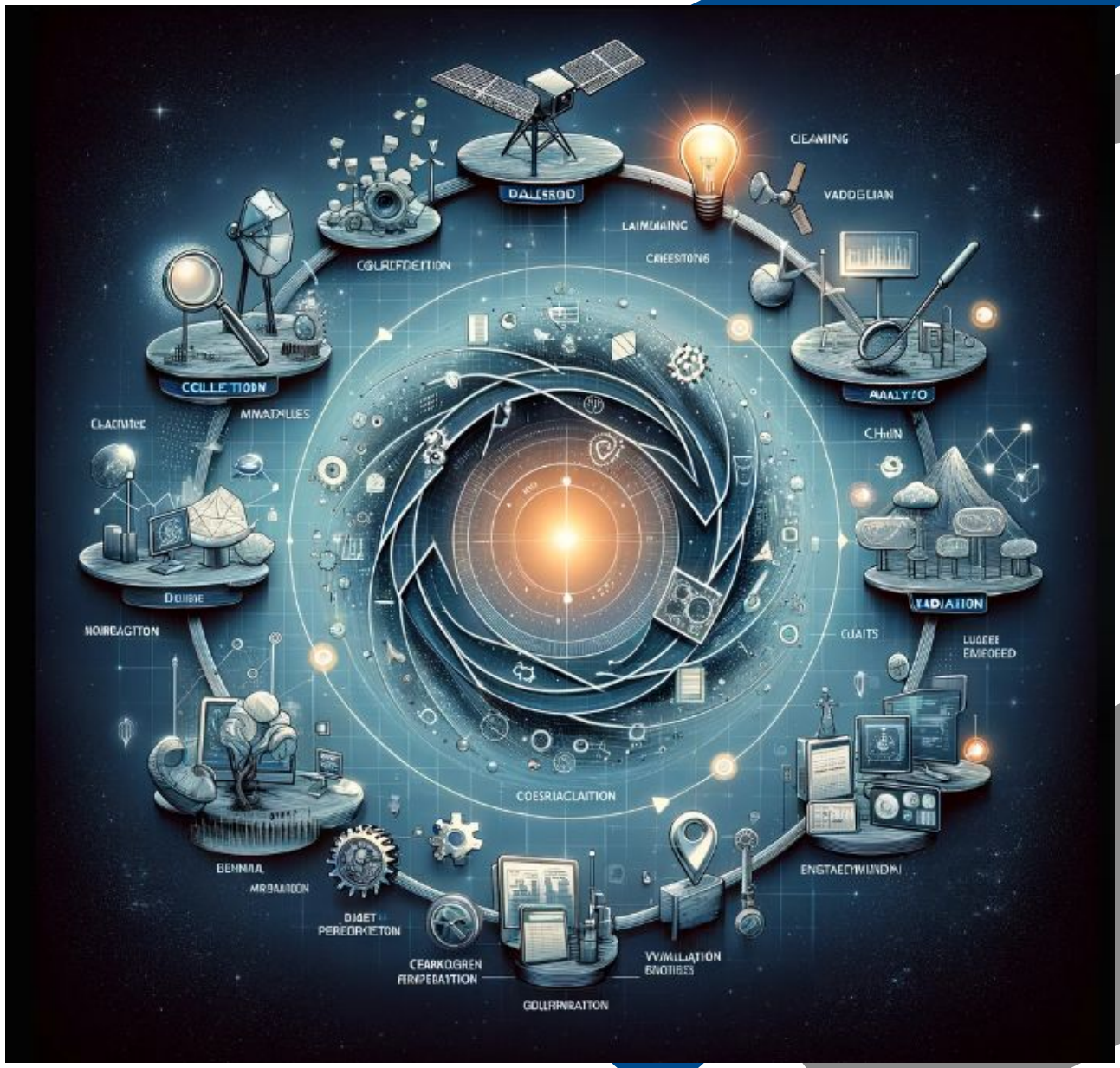
Senac

Fecomércio Sesc

Data Science – Princípios e Técnicas



Abril

2024

Data Science

Onde me encontrar:

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

e

https://github.com/MatmJr

Vimos na aula passada

Data Science

Estamos trabalhando com 891 observações de 12 variáveis. Para tornar as coisas um pouco mais explícitas, já que alguns dos nomes das variáveis não são totalmente claros, aqui está o que temos que lidar:

Data Science

Nome da Variável	Descrição
Survived	Sobreviveu (1) ou morreu (0)
Pclass	Classe do passageiro (1 = Upper Class, 2 = Middle Class, 3 = Lower Class)
Name	Nome do passageiro
Sex	Sexo do passageiro
Age	Idade do passageiro
SibSp	Número de irmãos/cônjuges a bordo
Parch	Número de pais/filhos a bordo
Ticket	Número do bilhete
Fare	Tarifa
Cabin	Cabine
Embarked	Porto de embarque C = Cherbourg, Q = Queenstown, S = Southampton

Data Science

Na etapa de pré-processamento:

```
data.isnull().sum()

Survived      0
Pclass        0
Name          0
Sex           0
Age         177
SibSp         0
Parch         0
Ticket        0
Fare          0
Cabin       687
Embarked      2
Title         0
dtype: int64
```

```
data.isna().sum()

Survived    0
Pclass      0
Name        0
Sex         0
Age         0
SibSp       0
Parch       0
Ticket      0
Fare        0
Cabin       0
Embarked    0
Title       0
dtype: int64
```

Data Science

Montamos um script que acessa o arquivo e faz o processamento dos dados.

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.

Data Science

Exemplo 1: Determine a variância das idades dos passageiros do Titanic:

```
data["Age"].var()
```

Data Science

```
data.groupby(["Sex","Pclass"])["Age"].var()
```

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 idades dos passageiros do Titanic.

```
data["Age"].std()
```

Data Science

```
data.groupby(["Sex","Pclass"])["Age"].std()
```

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.

Data Science

```
data["Age"].max() - data["Age"].min()
```

Distribuição de Frequência

A distribuição de frequência é uma ferramenta estatística essencial para analisar e visualizar como os dados estão distribuídos ao longo de diferentes categorias ou intervalos contínuos.

Data Science

Ela permite observar a frequência com que determinados valores ou grupos de valores ocorrem dentro de um conjunto de dados, facilitando a identificação de padrões e a compreensão da estrutura dos dados.

Data Science

A construção de uma distribuição de frequência pode ser realizada através da tabulação dos dados em classes, que podem ser definidas por intervalos de valores. Cada classe contém uma contagem de quantas vezes valores dentro daquele intervalo aparecem no conjunto de dados, conhecida como frequência.

Data Science

A representação visual dessas frequências, geralmente por meio de histogramas (gráficos de barras), proporciona uma compreensão clara e imediata das características principais dos dados, como a concentração de observações em determinados intervalos.

Data Science

Exemplo 3: Construa um histograma para visualizar a distribuição de frequência das idades dos passageiros do Titanic.

Vamos precisar de uma biblioteca nova: matplotlib.

Data Science

```
import matplotlib.pyplot as plt

plt.hist(data["Age"])
plt.title('Distribuição de Idades dos Passageiros do Titanic')
plt.xlabel('Idades')
plt.ylabel('Frequência')
plt.show()
```

Data Science

Box-Plot

O box-plot é uma ferramenta gráfica essencial na estatística descritiva, fornecendo uma visão clara da distribuição dos dados, bem como de sua concentração e dispersão. Este gráfico destaca quão distantes os valores extremos estão em relação ao corpo principal dos dados, facilitando a identificação de outliers.

Data Science

Um box-plot é estruturado a partir de cinco valores chave: o mínimo, o primeiro quartil (Q1), a mediana (Q2), o terceiro quartil (Q3) e o máximo. Esses pontos são utilizados para construir o gráfico e comparar a distribuição de diferentes conjuntos de dados.

Data Science

Para desenhar um box-plot, posiciona-se uma caixa retangular sobre um eixo numérico, que pode ser horizontal ou vertical. Os extremos da caixa são definidos pelos primeiro e terceiro quartis, encapsulando aproximadamente 50% dos dados entre eles.

Data Science

O box-plot é uma ferramenta poderosa para uma análise rápida e eficaz, permitindo comparações diretas entre diferentes grupos de dados e ajudando a destacar características como simetria e dispersão de maneira intuitiva e imediata.

Data Science

Vamos precisar de mais uma biblioteca para plotar gráficos, a seaborn

```
import seaborn as sns
sns.boxplot(x= data.Age)
```

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 idades dos passageiros do Titanic.

```
data["Age"].skew()
```

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.

Data Science

Exemplo 5: Determine a curtose das idades dos passageiros do Titanic.

```
data["Age"].kurtosis()
```

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.

Data Science

Essa curva suaviza os picos e vales dos histogramas ao passar pelas diferentes classes de dados, oferecendo uma representação contínua e detalhada da distribuição.

```
sns.displot(data.Age, bins = [0,10,20,30,40,50,60,70,80,90], kde= True)
```

Dúvidas?

Marco Mialaret, MSc

Telefone:
81 98160 7018

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