Estratégia Concursos

Aula 00

TCE-PI (Cargos de TI) Passo Estratégico de Análise de Dados - 2024 (Pós-Edital)

Autor:
Fernando Pedrosa Lopes

16 de Agosto de 2024

ANÁLISE EXPLORATÓRIA E REPRESENTAÇÃO DE DADOS

Sumário

CONTEÚDO

Visualização e Análise Exploratória de Dados. Visualizações para cada tipo de dado. Tipos de gráficos. Representação de Dados.

ANÁLISE ESTATÍSTICA

Inicialmente, convém destacar o percentual de incidência do assunto, dentro da disciplina **Ciência de Dados e Estatística** em concursos/cargos similares. Quanto maior o percentual de cobrança de um dado assunto, maior sua importância.

Obs.: *um mesmo assunto pode ser classificado em mais de um tópico devido à multidisciplinaridade de conteúdo*.

Assunto	Relevância na disciplina em concursos similares
Estatística descritiva (análise exploratória de dados)	23.7 %
Cálculo de Probabilidades	8.3 %
Conhecimentos de estatística	7.7 %
Principais distribuições de probabilidade	4.7 %
Amostragem	2.4 %
Modelos lineares	2.4 %
Inferência estatística	1.8 %
Estatística não paramétrica	0.6 %

GLOSSÁRIO DE TERMOS

Faremos uma lista de termos que são relevantes ao entendimento do assunto desta aula. Caso tenha alguma dúvida durante a leitura, esta seção pode lhe ajudar a esclarecer.

AED - AED, ou Análise Exploratória de Dados, é uma técnica de estatística utilizada para analisar e investigar conjuntos de dados, a fim de descobrir padrões, anomalias, testar hipóteses e verificar suposições com a ajuda de gráficos estatísticos e ferramentas de resumo.

Observação - Em análise de dados, uma observação refere-se a um único ponto de dados ou registro em um conjunto de dados. Por exemplo, em um conjunto de dados contendo informações sobre pessoas, cada pessoa seria uma observação.

Feature - Uma feature, ou variável, em um conjunto de dados é uma característica individual que é medida para cada observação. Por exemplo, em um conjunto de dados contendo informações sobre casas, as features poderiam incluir o número de quartos, a área total e o ano de construção.

Rótulo - Um rótulo é o valor que queremos prever ou classificar em aprendizado de máquina supervisionado. Por exemplo, em um problema de classificação de spam, o rótulo seria se cada e-mail é 'spam' ou 'não spam'.

Metadado - Metadados são dados sobre os dados. Eles fornecem informações como quando e por quem os dados foram coletados, como os dados estão formatados, a fonte dos dados, etc.

Gráfico de Linha - Um gráfico de linha é um tipo de gráfico que exibe informações como uma série de pontos de dados chamados 'marcadores' conectados por segmentos de linha reta. Ele é usado principalmente para visualizar uma tendência nos dados ao longo do tempo, conhecida como série temporal.

Gráfico de Barras - Um gráfico de barras é um gráfico que usa barras retangulares para representar comparações entre categorias. Uma dimensão das barras representa a categoria sendo comparada, enquanto a outra dimensão representa o valor para essa categoria.

Gráfico de Pizza - Um gráfico de pizza é um gráfico circular que é dividido em fatias para ilustrar proporções numéricas. Em um gráfico de pizza, o arco de cada fatia (e, portanto, seu ângulo e área central), é proporcional à quantidade que representa.

Gráfico de Dispersão - Um gráfico de dispersão usa pontos gráficos para representar os valores de duas variáveis diferentes de um conjunto de dados. A posição de um ponto depende de seus valores de duas variáveis, uma no eixo x e outra no eixo y.

Gráfico de Contorno - Um gráfico de contorno é uma representação gráfica tridimensional em duas dimensões, onde três variáveis são representadas por três eixos. Os dois eixos x e y representam as variáveis independentes, enquanto a cor ou o nível de cinza representam a terceira variável.

Gráfico de Área - Um gráfico de área representa a magnitude de uma tendência ao longo do tempo, conectando vários pontos de dados em uma linha e preenchendo a área abaixo dela.

Gráfico de Rede - Um gráfico de rede é um tipo de gráfico que é usado para visualizar conexões ou relacionamentos entre entidades. Cada entidade é representada como um nó e cada relacionamento é representada como uma aresta que conecta os nós.

Histograma - Um histograma é um gráfico que mostra a distribuição de frequência de um conjunto de dados contínuos ou discretos. Ele é útil para visualizar a distribuição dos dados.

Box Plot - Um box plot, também conhecido como diagrama de caixa, é um gráfico que representa estatísticas resumidas de um conjunto de dados, incluindo os quartis, a mediana, e possíveis outliers. Ele é usado para visualizar a distribuição e a dispersão dos dados.

ROTEIRO DE REVISÃO

A ideia desta seção é apresentar um roteiro para que você realize uma revisão completa do assunto e, ao mesmo tempo, destacar aspectos do conteúdo que merecem atenção.

Introdução

Análise Exploratória de Dados, ou AED, é um método usado para analisar e resumir conjuntos de dados. Foi introduzida por John Tukey em 1977, e desde então, tem se tornado uma etapa essencial em qualquer análise de dados.

A ideia por trás da AED é que antes de podermos fazer suposições ou modelos, precisamos saber o que nossos dados estão nos dizendo. AED permite a exploração, a compreensão da estrutura, das relações e das principais características dos dados.

Seus principais objetivos são:

- **Maximizar a percepção dos dados:** Visa a facilitar o entendimento dos dados. Ela ajuda a identificar as principais variáveis, as relações entre elas e as variáveis que proporcionam a maior informação.
- **Identificar variáveis importantes:** Para fazer previsões ou para analisar a variância, a AED nos ajuda a identificar as variáveis mais importantes que contribuiriam para as previsões.
- **Detectar outliers e anomalias:** AED pode nos ajudar a detectar possíveis outliers e anomalias que podem alterar o resultado da análise de dados.
- **Testar suposições:** Podemos usar a AED para testar se nossas suposições ou a hipótese de que os dados foram gerados por um determinado processo são verdadeiras.
- **Desenvolver modelos:** Com base na AED, podemos criar modelos que nos ajudam a entender os dados e fazer previsões.

Podemos dizer que AED é uma combinação de técnicas gráficas e quantitativas. As técnicas gráficas envolvem a criação de gráficos de dados, como histogramas, gráficos de dispersão e gráficos de caixa para visualizar os dados. As técnicas quantitativas envolvem o cálculo de estatísticas descritivas, como a média, mediana, variância e desvio padrão.

Observações de dados e seus componentes

Conjuntos de dados, especialmente em um contexto de análise de dados, podem ser bastante complexos e variados, mas geralmente contêm alguns componentes principais. Veja os principais:

- **Variáveis:** Uma variável é qualquer característica, número ou quantidade que pode ser medida ou contada. Elas são divididas em dois tipos principais - variáveis quantitativas e variáveis qualitativas. As variáveis quantitativas são numéricas e representam uma medida (como altura, peso, idade). As variáveis qualitativas, também chamadas de categóricas, são não numéricas e representam categorias ou grupos (como gênero, cor dos olhos).
- **Observações:** Uma observação, também conhecida como registro, é uma única instância em um conjunto de dados. Em um conjunto de dados tabular, uma observação geralmente corresponde a uma linha.
- **Features:** Em Machine Learning, as features são variáveis individuais independentes que atuam como entrada em um modelo. Por exemplo, em um conjunto de dados sobre preços de casas, as features podem incluir o número de quartos, o tamanho do lote, o ano de construção, entre outros.
- **Rótulos:** Em contextos de aprendizado supervisionado, os rótulos representam a "resposta" ou o resultado que o modelo está tentando prever ou classificar. No exemplo do conjunto de dados de preços de casas, o preço de venda da casa seria o rótulo.

- **Metadados:** Metadados são "dados sobre os dados". Eles fornecem informações adicionais sobre o conjunto de dados, como quando e como os dados foram coletados, quem os coletou, como estão estruturados, etc.

Claro que cada conjunto de dados pode ter diferentes componentes dependendo de sua natureza e do contexto em que é usado.

Visualização de Dados

A visualização de dados é uma técnica que envolve o uso de elementos visuais - como gráficos, mapas e gráficos - para representar e entender complexos conjuntos de dados. É uma maneira eficaz de transmitir informações de forma rápida e intuitiva, ajudando a entender padrões, tendências, correlações e outliers nos dados.

Veja algumas das principais razões para entendermos melhor a visualização de dados:

- **Compreensão rápida:** Podemos processar informações visuais muito mais rápido do que informações textuais. A visualização dos dados permite uma compreensão mais rápida dos dados em comparação com a leitura de tabelas ou relatórios.
- **Descoberta de padrões, tendências e correlações:** Os gráficos e as representações visuais ajudam a identificar padrões, tendências e correlações nos dados que podem não ser evidentes apenas olhando para os números.
- **Simplificação de dados complexos:** A visualização de dados pode tornar os dados complexos mais acessíveis, compreensíveis e utilizáveis. Ela permite que pessoas de diferentes formações, que não são especialistas em análise de dados, entendam as complexidades dos dados.
- **Comunicação eficaz:** É uma maneira eficaz de comunicar informações, seja em apresentações, relatórios ou publicações online. Ela pode captar a atenção do público e transmitir uma mensagem de forma clara e convincente.

Existem muitos tipos diferentes de visualizações de dados, incluindo **gráficos de barras, gráficos de linha, gráficos de pizza, histogramas, gráficos de dispersão, gráficos de área, gráficos de rede e muitos mais**. Cada tipo de gráfico tem seus próprios usos e é mais adequado para diferentes tipos de dados e diferentes questões de pesquisa.

Escolher o tipo de gráfico correto para visualizar seus dados é um passo importante para comunicar seus resultados de forma eficaz, portanto, é necessário considerar algumas questões na escolha do gráfico, tais como:

- **Seus dados e seus objetivos:** Antes de escolher um gráfico, você precisa entender o que seus dados representam e o que você deseja comunicar. Seu objetivo é mostrar tendências ao longo do tempo? Comparar categorias? Mostrar distribuição ou correlação entre duas variáveis? O tipo de gráfico que você escolhe deve se alinhar com seus objetivos.
- **Considere o tipo de dados que você tem:** Diferentes gráficos são apropriados para diferentes tipos de dados. Por exemplo, os gráficos de barras e de pizza são ótimos para dados categóricos, enquanto os gráficos de linha são melhores para dados temporais. Gráficos de dispersão são úteis para mostrar a relação entre duas variáveis contínuas.
- **Simplicidade é fundamental:** Como regra geral, as visualizações de dados devem ser o mais simples possível, mas ainda assim eficazes. Evite gráficos excessivamente complicados ou confusos. Se o gráfico não é fácil de entender, provavelmente não é a escolha certa.
- **Não engane com gráficos:** Certifique-se de que o gráfico não está distorcendo os dados de uma forma que possa ser enganosa. Por exemplo, começar um gráfico de barras em um número diferente de zero pode fazer com que as diferenças pareçam mais dramáticas do que realmente são.

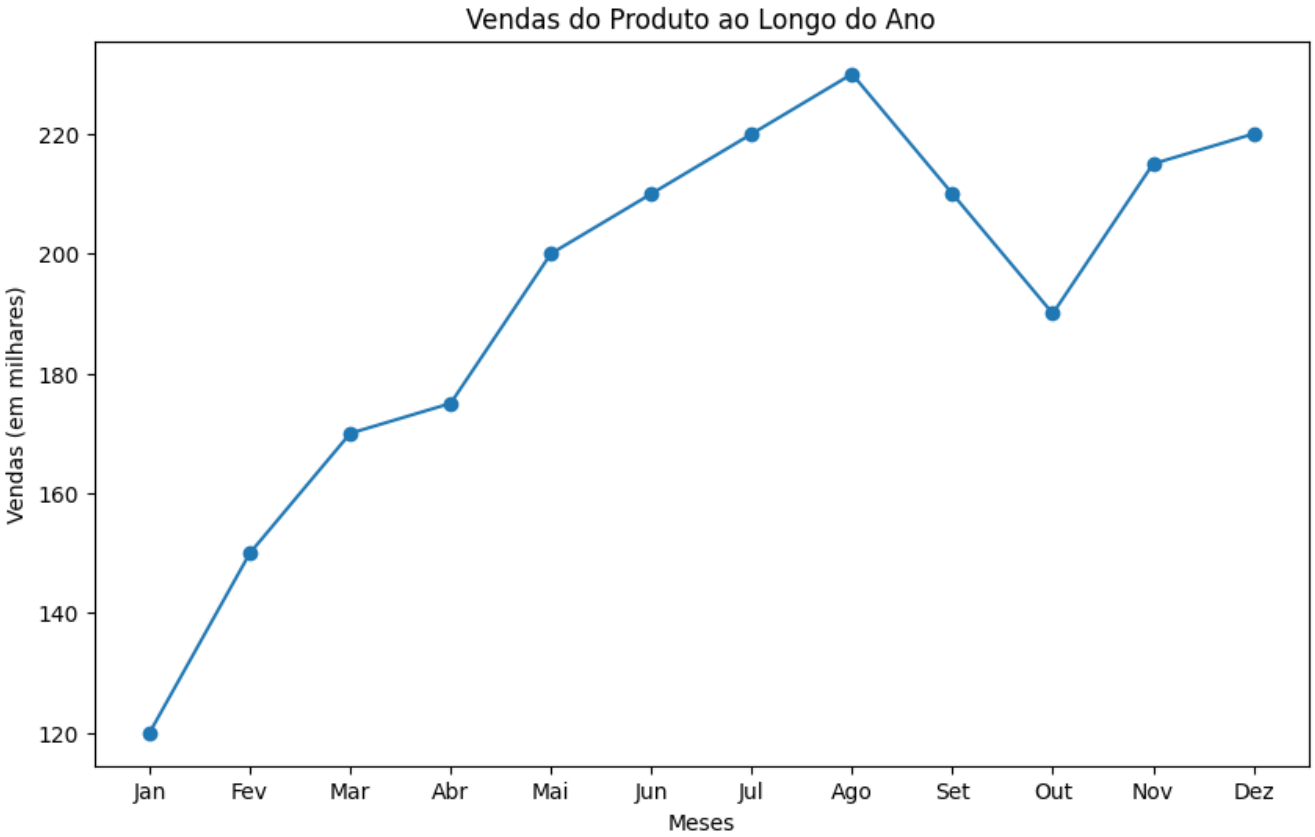
Nas próximas seções, vamos estudar os diferentes tipos de visualizações e gráficos em maiores detalhes.

Gráfico de Linha

Quando usar

Gráficos de linha são melhores para visualizar tendências ou mudanças em dados contínuos ao longo do tempo. Eles são especialmente úteis para a análise de séries temporais - quando temos pontos de dados em intervalos sucessivos. Por exemplo, você pode usar um gráfico de linha para visualizar as mudanças nas temperaturas médias ao longo do ano, os preços das ações em uma semana, as vendas de um produto ao longo do tempo, entre outros.

Exemplo



Suponha que você é o gerente de vendas de uma empresa de varejo e está interessado em entender como as vendas de um determinado produto mudaram ao longo do ano passado. Aqui estão os dados mensais de vendas que você tem (em milhares de unidades):

- Janeiro: 120
- Fevereiro: 150
- Março: 170
- Abril: 175
- Maio: 200
- Junho: 210
- Julho: 220
- Agosto: 230
- Setembro: 210
- Outubro: 190
- Novembro: 215
- Dezembro: 220

Ao visualizar esses dados em um gráfico de linha, você pode ver claramente a tendência de aumento das vendas ao longo do ano, com um pequeno declínio em outubro. Isso pode ser devido a várias razões - talvez haja um aumento sazonal na demanda durante os meses de verão,

ou talvez uma campanha de marketing tenha sido particularmente eficaz durante esse período. Da mesma forma, o declínio em outubro pode indicar uma diminuição sazonal na demanda ou talvez tenha havido problemas de estoque que limitaram a quantidade de produto disponível para venda.

Em qualquer caso, o gráfico de linha torna essas tendências muito mais fáceis de ver e entender do que se você estivesse apenas olhando para os números brutos. E essa visibilidade melhorada pode ajudá-lo a tomar decisões mais informadas sobre como gerenciar as vendas e o estoque desse produto no futuro.

Gráfico de Barras

Quando usar

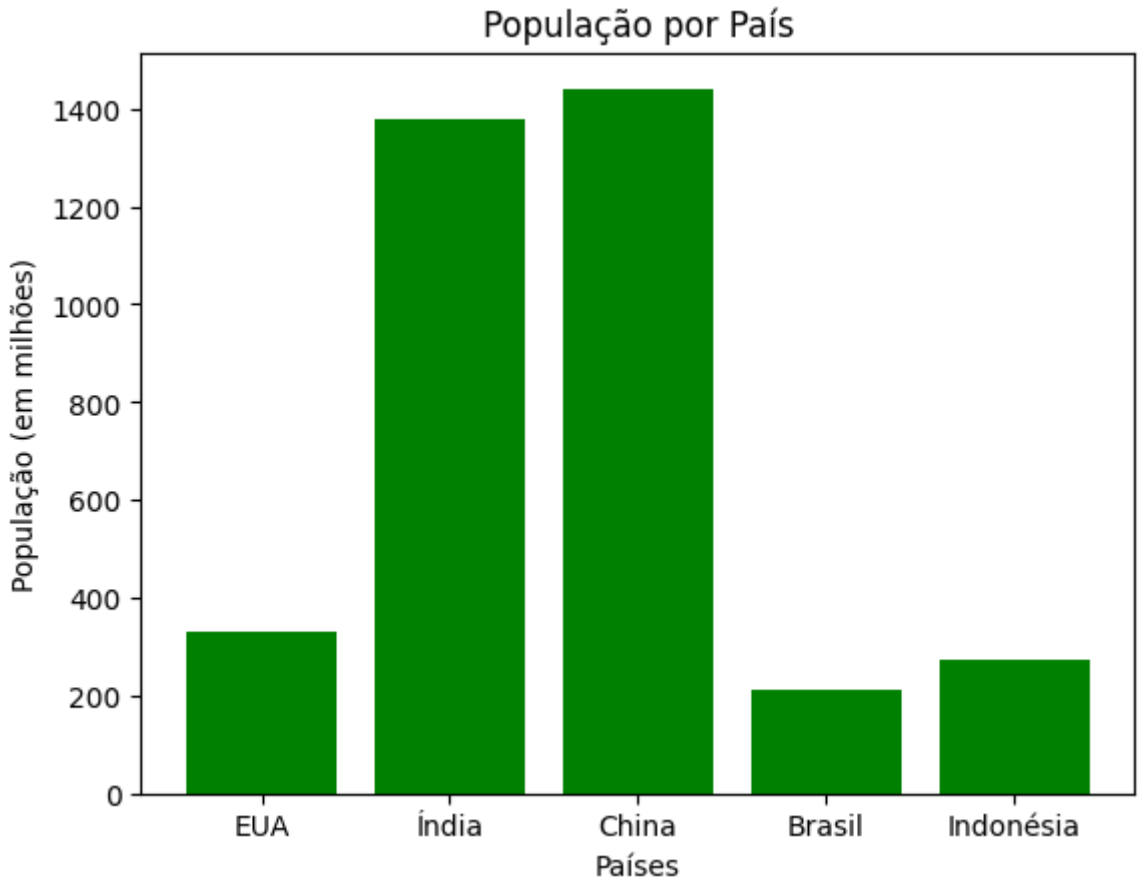
Gráficos de barras são muito versáteis e úteis para comparar quantidades de diferentes categorias. Eles podem ser usados para representar dados categóricos ou numéricos. Também são úteis quando você deseja comparar uma única categoria de dados entre subgrupos individuais.

Exemplo

Vamos presumir que temos dados sobre a população de diferentes países e queremos visualizar esses dados. Aqui estão os dados:

- Estados Unidos: 331 milhões
- Índia: 1380 milhões
- China: 1441 milhões
- Brasil: 213 milhões
- Indonésia: 273 milhões

Podemos usar um gráfico de barras para representar esses dados da seguinte maneira:



Neste exemplo, estamos visualizando a população de diferentes países. O gráfico de barras fornece uma representação visual clara das diferenças de população entre os países. Podemos ver imediatamente que a China e a Índia têm populações muito maiores em comparação com os outros países listados. Por outro lado, o Brasil e os Estados Unidos têm populações menores em comparação. Isso poderia ser útil, por exemplo, para uma empresa que está tentando decidir em que países lançar um novo produto, pois o tamanho da população poderia potencialmente indicar o tamanho do mercado.

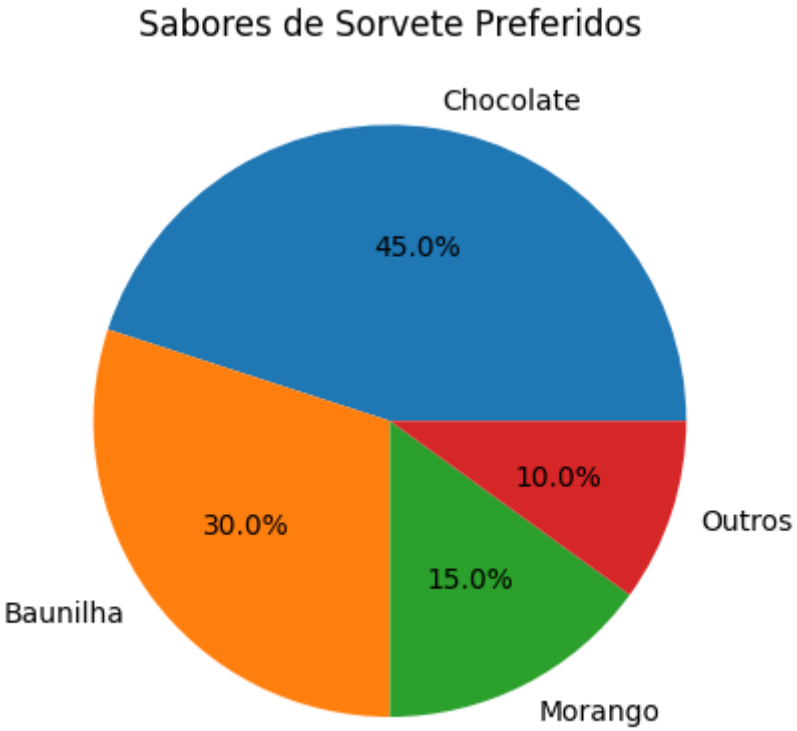
Gráfico de Pizza

Quando usar

Gráficos de pizza são úteis para visualizar proporções ou porcentagens de um todo. Eles representam categorias de dados como fatias de um círculo, onde o tamanho de cada fatia é proporcional à porcentagem que ela representa do todo. Portanto, eles são ideais quando você quer comparar partes de um todo. No entanto, eles podem se tornar confusos se tiverem muitas

fatias. Para dados com muitas categorias, outros tipos de gráficos, como gráficos de barras, podem ser mais adequados.

Exemplo



Vamos usar o exemplo de uma pesquisa feita em uma escola para descobrir qual é o sabor preferido de sorvete entre os estudantes. Suponha que os dados sejam assim:

- Chocolate: 45%
- Baunilha: 30%
- Morango: 15%
- Outros: 10%

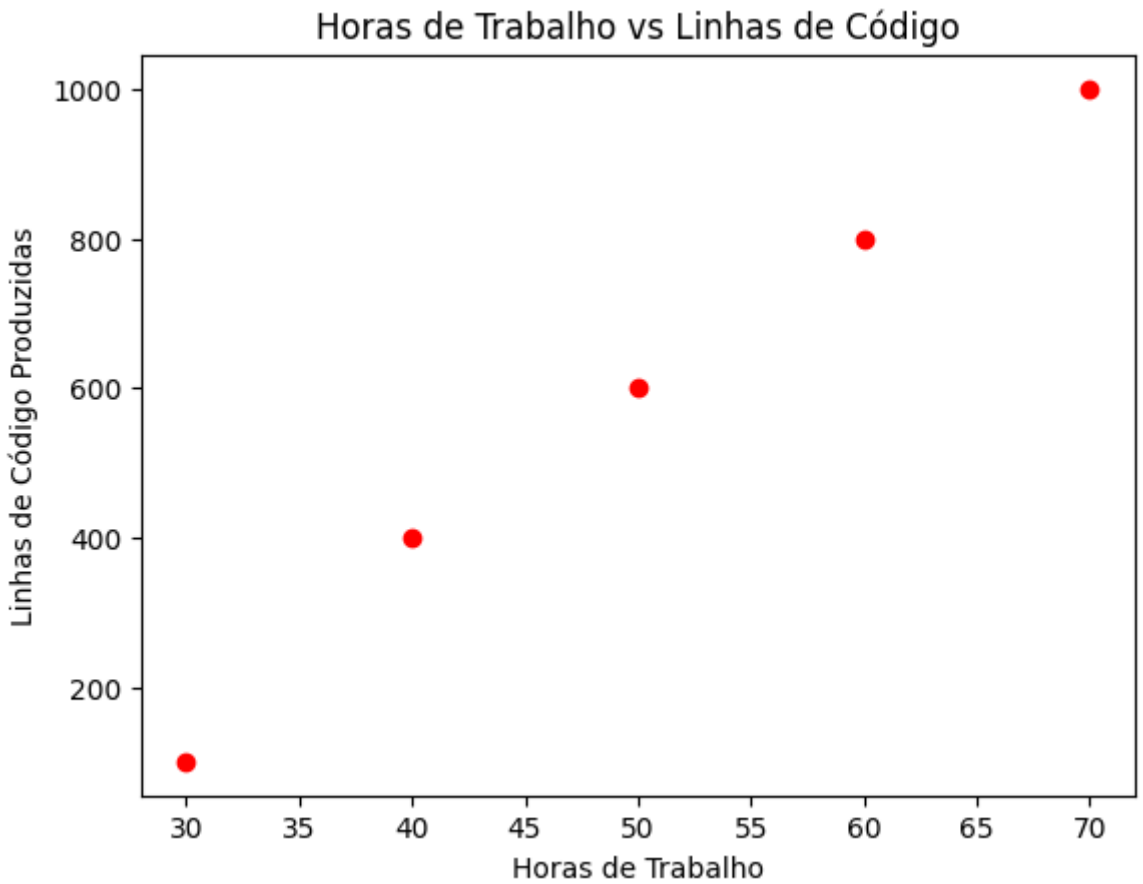
Neste exemplo, a escola conduziu uma pesquisa para descobrir qual sabor de sorvete é o favorito dos alunos. Os dados mostram a porcentagem de alunos que preferem cada sabor. A partir do gráfico de pizza, podemos ver claramente que o chocolate é o sabor mais popular, seguido pela baunilha. Os sabores Morango e outros sabores combinados representam 25% das preferências dos alunos. Portanto, se a escola estivesse planejando servir sorvete em um evento, poderia usar essas informações para garantir que tenha bastante sorvete de chocolate e baunilha disponível.

Gráfico de Dispersão

Quando usar

Gráficos de dispersão são úteis para visualizar a relação entre duas variáveis quantitativas. Eles permitem que você veja se existe uma correlação entre as duas variáveis - por exemplo, se uma variável aumenta, a outra variável também aumenta (correlação positiva), se uma variável aumenta enquanto a outra diminui (correlação negativa), ou se não há correlação entre as variáveis.

Exemplo



Suponha que temos dados de uma empresa de tecnologia sobre o número de horas que seus funcionários trabalham por semana e a quantidade de código que produzem. Aqui estão os dados para 5 funcionários:

- Funcionário 1: 30 horas, 100 linhas de código
- Funcionário 2: 40 horas, 400 linhas de código

- Funcionário 3: 50 horas, 600 linhas de código
- Funcionário 4: 60 horas, 800 linhas de código
- Funcionário 5: 70 horas, 1000 linhas de código

Neste exemplo, estamos visualizando a relação entre o número de horas que os funcionários trabalham e a quantidade de código que eles produzem. A partir do gráfico de dispersão, podemos ver que parece haver uma correlação positiva entre as duas variáveis: quanto mais horas um funcionário trabalha, mais linhas de código eles parecem produzir. Isto é apenas um exemplo simplificado. Na prática, a quantidade de código produzido por um programador não é necessariamente correlacionada com as horas trabalhadas, pois a eficiência e a experiência do programador também desempenham um papel importante.

Gráfico de Contorno

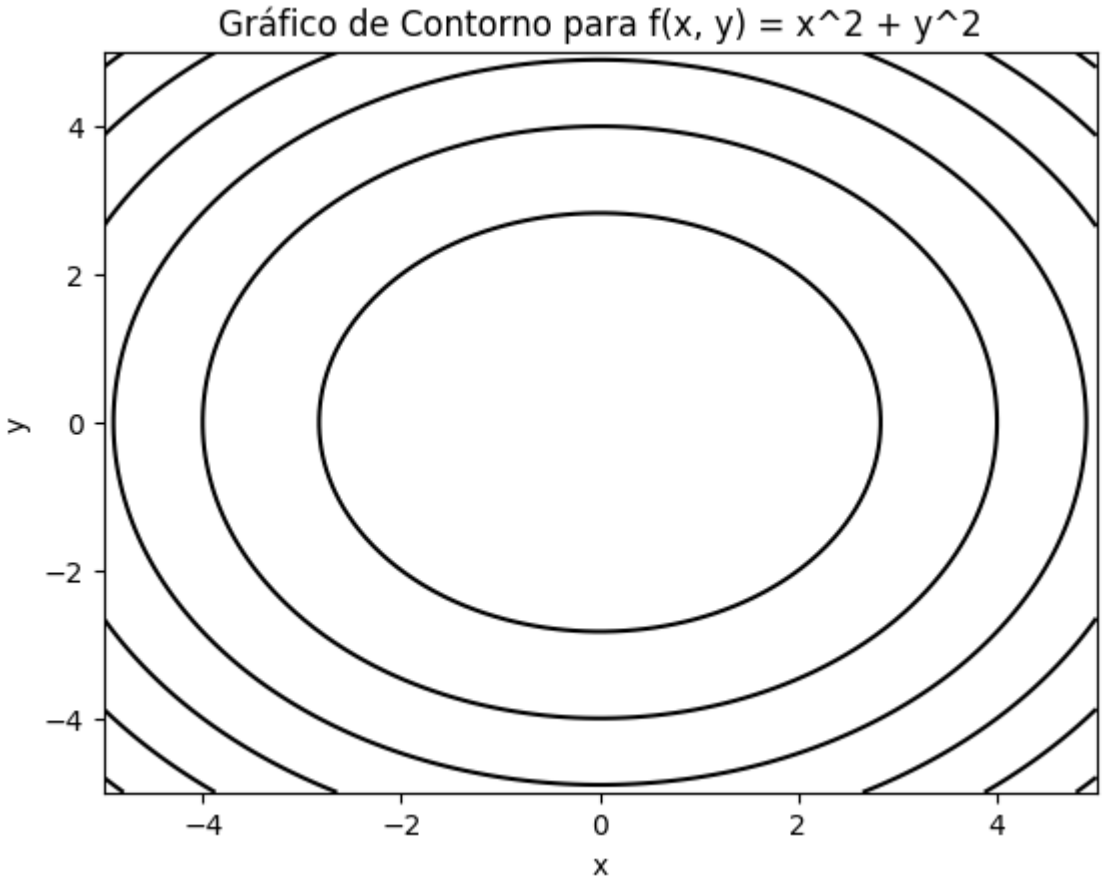
Quando usar

Gráficos de contorno, também conhecidos como gráficos de nível ou gráficos de superfície 2D, são diagramas tridimensionais representados em duas dimensões. Eles são geralmente usados em ciências físicas e engenharia para representar variáveis que dependem de duas outras. Cada curva (contorno) em um gráfico de contorno representa uma linha de igual valor (isolinha) para a função de três dimensões.

Esses gráficos são úteis quando você precisa visualizar uma superfície tridimensional em duas dimensões. Comumente, eles são usados em campos como meteorologia (para visualizar campos de temperatura ou pressão) ou em aprendizado de máquina para visualizar funções de decisão em classificadores.

Exemplo

Por exemplo, considere que temos uma função f(x, y) = x^2 + y^2 e queremos visualizar esta função usando um gráfico de contorno. Poderíamos plotar o gráfico da seguinte maneira:



Neste exemplo, estamos visualizando a função f(x, y) = x^2 + y^2. Cada contorno no gráfico representa um conjunto de pontos (x, y) que têm o mesmo valor para a função f(x, y). A partir deste gráfico, podemos ver que o valor de f(x, y) aumenta à medida que nos afastamos do centro (0,0).

Em um gráfico de contorno mais complexo, as linhas de contorno podem não ser apenas círculos, mas podem ter várias formas diferentes. Isso permite que você visualize superfícies complexas em três dimensões usando apenas um gráfico bidimensional.

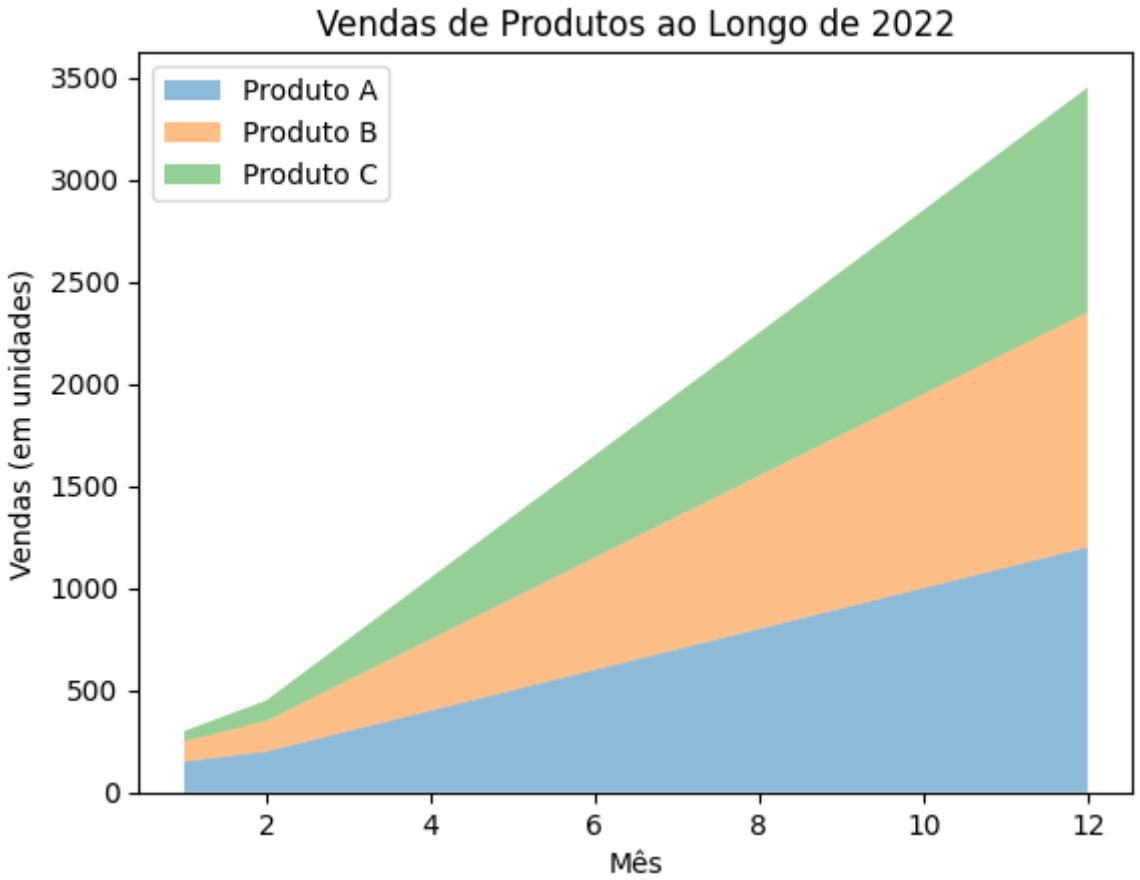
Gráfico de Área

Quando usar

Gráficos de área são essencialmente gráficos de linha, exceto que a área abaixo da linha é preenchida com cor ou textura. Eles são comumente usados para representar quantidades acumuladas ao longo do tempo, ou seja, para mostrar tendências de dados ao longo do tempo

entre diferentes grupos. Eles também são úteis quando você quer enfatizar a magnitude da mudança ao longo do tempo.

Exemplo



Por exemplo, considere que temos dados sobre as vendas de três diferentes produtos em uma loja durante o ano de 2022. Aqui estão os dados:

- Produto A: [150, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100, 1200]
- Produto B: [100, 150, 250, 350, 450, 550, 650, 750, 850, 950, 1050, 1150]
- Produto C: [50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1100]

Neste exemplo, estamos visualizando as vendas de três produtos diferentes ao longo de um ano. Cada cor no gráfico representa um produto diferente. A partir do gráfico de área, podemos ver como as vendas de cada produto mudaram ao longo do tempo. Podemos ver que as vendas de todos os produtos aumentaram ao longo do ano, com o Produto A sempre vendendo mais do que o Produto B, que por sua vez sempre vendeu mais do que o Produto C. Portanto, se a loja estivesse planejando o estoque para o próximo ano, ela poderia usar essas informações para garantir que tem mais do Produto A em estoque do que dos outros produtos.

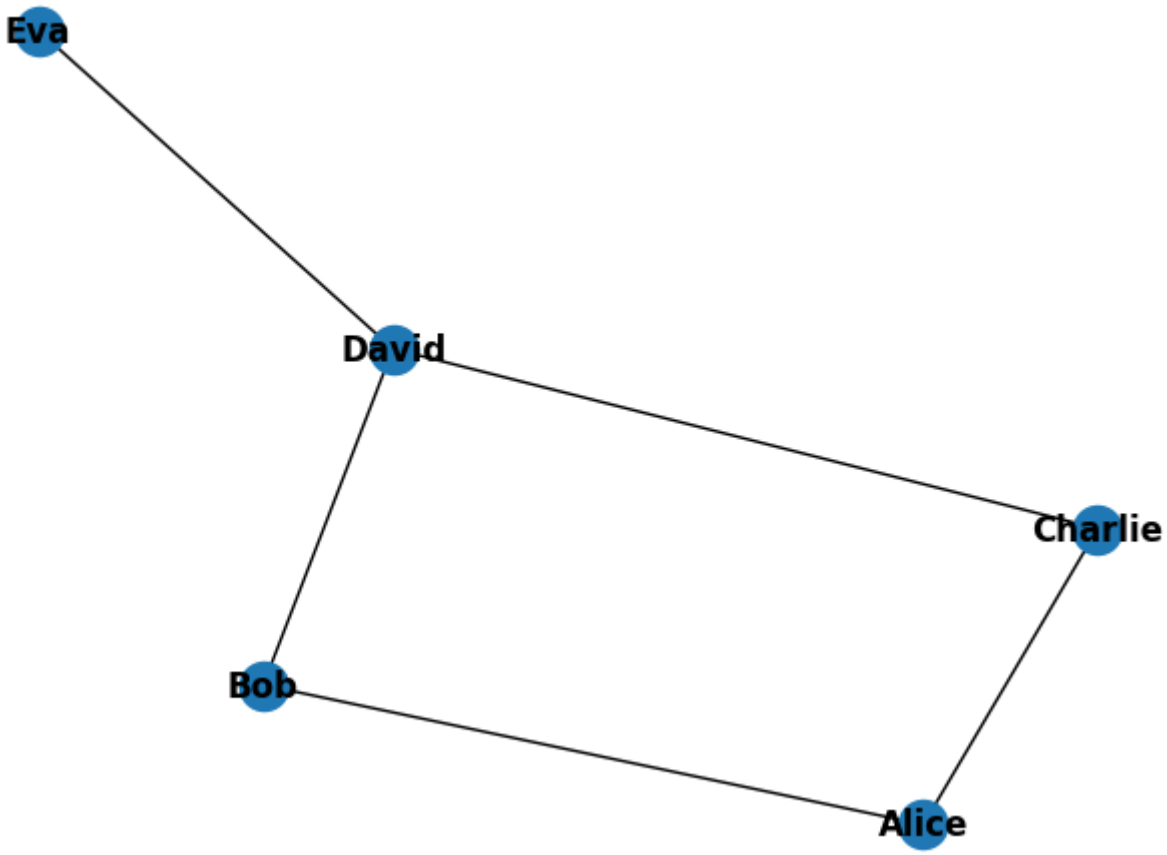
Gráfico de Rede

Quando usar

Gráficos de rede, também conhecidos como gráficos de grafos, são usados para visualizar relações ou conexões entre nós. Cada nó representa uma entidade e cada aresta (linha) representa uma conexão entre duas entidades. Esses gráficos são úteis quando se está trabalhando com relacionamentos complexos, como redes sociais, sistemas de recomendação, diagramas de fluxo, entre outros.

Exemplo

Vamos considerar um exemplo onde temos uma pequena rede social de 5 pessoas e queremos visualizar as conexões entre elas. Podemos representar isso usando um gráfico de rede da seguinte maneira:



Neste exemplo, estamos visualizando uma pequena rede social. Cada nó representa uma pessoa e cada aresta representa uma conexão entre duas pessoas. Podemos ver imediatamente que Alice está conectada a Bob e Charlie, Bob está conectado a Alice e David, Charlie está conectado a Alice e David, David está conectado a Bob, Charlie e Eva, e Eva está conectada apenas a David.

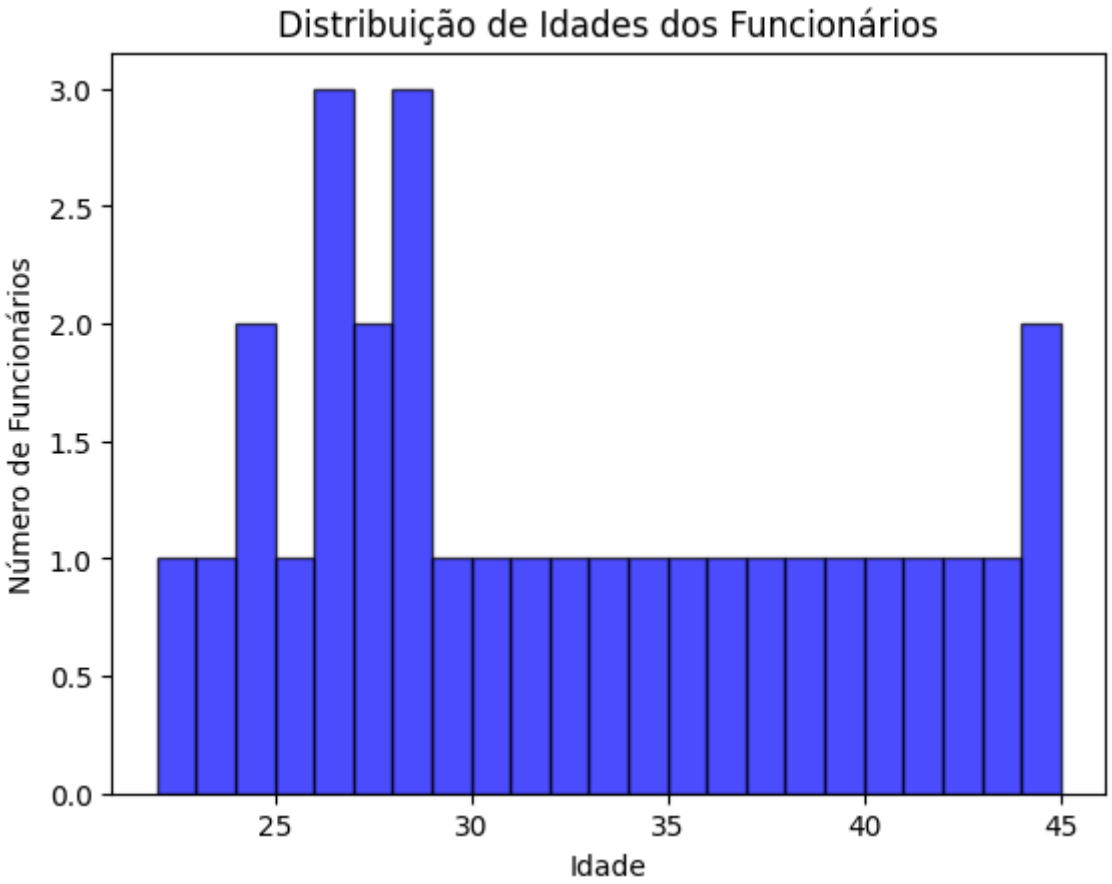
Portanto, se essa fosse uma rede social e quiséssemos enviar uma mensagem a todos o mais rápido possível, começaríamos com David, pois ele está conectado à maioria das pessoas. Este é um exemplo simples, mas os gráficos de rede podem ser usados para visualizar conjuntos de dados muito mais complexos com milhares ou mesmo milhões de nós e arestas.

Histograma

Quando usar

Um histograma é um gráfico que mostra a distribuição de frequência de um conjunto de dados contínuos ou discretos. É útil quando se quer ter uma ideia da densidade e da distribuição central dos dados. Histogramas são comumente usados em estatística e análise de dados para visualizar a distribuição dos dados.

Exemplo



Suponha que temos dados sobre as idades dos funcionários de uma empresa e queremos visualizar a distribuição dessas idades. Aqui estão as idades de 30 funcionários:

idades = [22, 23, 24, 24, 25, 26, 26, 26, 27, 27, 28, 28, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45]

Neste exemplo, estamos visualizando a distribuição de idades dos funcionários de uma empresa. Cada barra do histograma representa um intervalo de idades, e a altura da barra representa o número de funcionários que têm uma idade dentro daquele intervalo.

A partir do histograma, podemos ver que a maioria dos funcionários tem entre 26 e 28 anos, com um número decrescente de funcionários à medida que a idade aumenta. Isso poderia sugerir que a empresa tem uma força de trabalho relativamente jovem. Se a empresa estivesse considerando implementar políticas ou benefícios destinados a funcionários de certas idades, essas informações poderiam ser úteis.

Box Plot

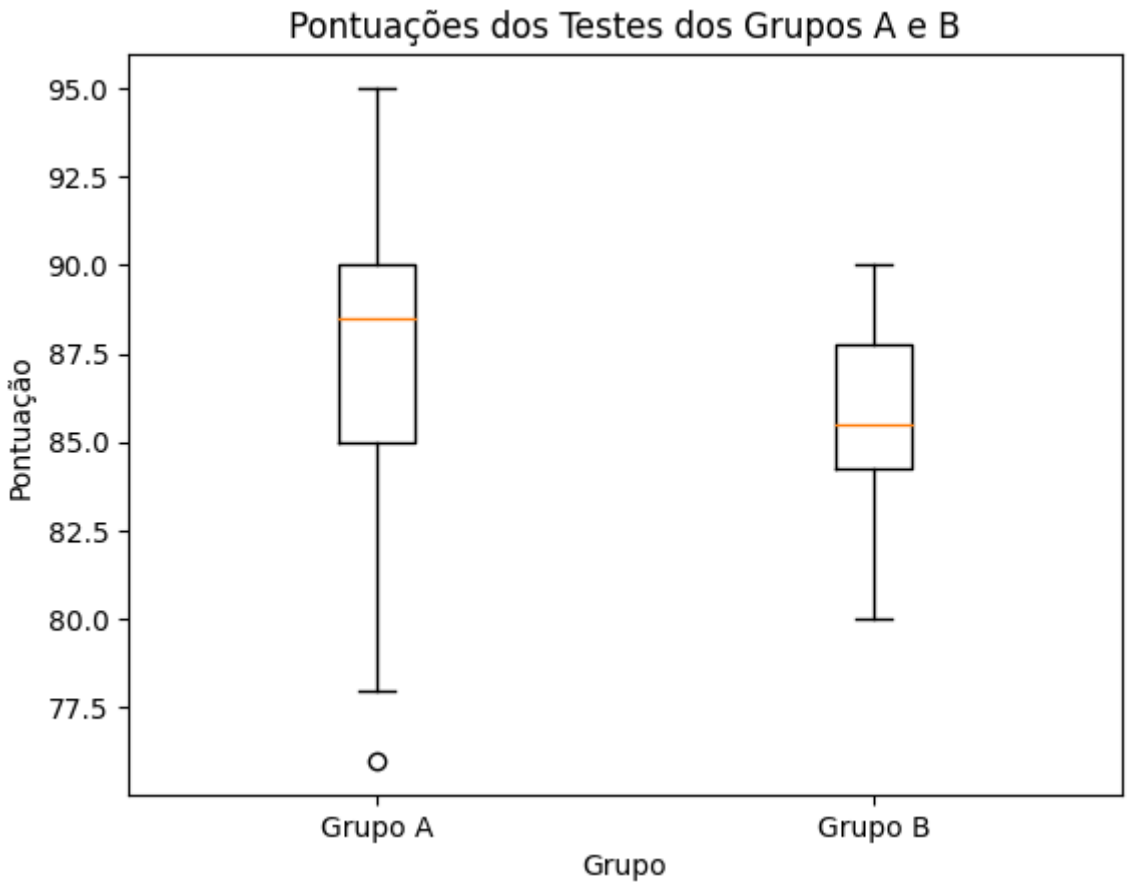
Quando usar

Um box plot, também conhecido como diagrama de caixa, é um gráfico usado para representar estatísticas resumidas de dados variáveis, como média, mediana, quartis, máximo e mínimo. Ele é útil para visualizar a distribuição e a dispersão dos dados, bem como para identificar possíveis outliers. É um gráfico comum em análise estatística e é útil para comparar distribuições entre diferentes conjuntos de dados.

Exemplo

Vamos supor que temos dados sobre as pontuações de dois grupos de estudantes em um teste. Aqui estão as pontuações:

- Grupo A: [85, 90, 78, 92, 88, 76, 95, 89, 90, 85]
- Grupo B: [80, 85, 88, 84, 82, 90, 88, 85, 87, 86]



Neste exemplo, estamos visualizando as pontuações dos testes dos Grupos A e B. Cada box plot representa um grupo de estudantes. A linha no meio da caixa representa a mediana das pontuações do grupo, enquanto a parte superior e inferior da caixa representam o terceiro quartil (Q3) e o primeiro quartil (Q1), respectivamente. A linha superior (bigode superior) representa o máximo (exceto possíveis outliers) e a linha inferior (bigode inferior) representa o mínimo (exceto possíveis outliers).

A partir deste gráfico, podemos ver que a mediana das pontuações do Grupo A é ligeiramente superior à do Grupo B. No entanto, a distribuição das pontuações é mais ampla no Grupo A do que no Grupo B, indicando uma maior variação nas pontuações do Grupo A. Isso pode sugerir que, enquanto o Grupo A teve uma pontuação média mais alta, o desempenho dos alunos neste grupo foi mais variado. Por outro lado, os alunos do Grupo B foram mais consistentes em suas pontuações.

Representação de Dados

A representação de dados é um conceito fundamental na ciência da computação e tecnologia da informação, referindo-se à maneira como as informações são codificadas, armazenadas e manipuladas dentro de um sistema computacional.

Vamos estudar a representação de dados numéricos e dados textuais (no final das contas, é tudo transformado para dados binários no computador). Antes de adentrarmos nas representações, é importante entendermos os **sistemas de numeração**.

Sistemas de Numeração

Sistemas de numeração são métodos para representar números utilizando um conjunto específico de símbolos. Vejamos os quatro principais sistemas de numeração: Decimal, Binário, Octal e Hexadecimal.

Sistema Decimal (Base 10)

- **Base**: 10
- **Símbolos**: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9
- **Utilização**: É o sistema padrão utilizado na vida cotidiana para representar números.
- **Exemplo**: O número 572 em decimal é representado da mesma forma: 572.

Sistema Binário (Base 2)

- **Base**: 2
- **Símbolos**: 0, 1
- **Utilização**: Utilizado internamente pelos computadores, já que os circuitos digitais têm dois estados: ligado (1) e desligado (0).
- **Conversão:**

  Para converter um número binário para decimal, você pode utilizar a seguinte fórmula:

$$\text{Decimal} = b_n \times 2^n + b_{n-1} \times 2^{n-1} + \ldots + b_1 \times 2^1 + b_0 \times 2^0$$

  onde b_i é o dígito na posição i, e n é a posição do dígito mais significativo (começando do 0).

  Exemplo: Convertendo 1011 para decimal:

$$1 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 = 8 + 0 + 2 + 1 = 11$$

Sistema Octal (Base 8)

- **Base**: 8
- **Símbolos**: 0, 1, 2, 3, 4, 5, 6, 7
- **Utilização**: Menos comum hoje em dia, mas já foi usado em sistemas como o Unix para representar permissões de arquivo.
- **Conversão**:

  A conversão de um número octal para decimal utiliza um princípio semelhante ao binário, mas com base 8:

$$\text{Decimal} = o_n \times 8^n + o_{n-1} \times 8^{n-1} + \ldots + o_1 \times 8^1 + o_0 \times 8^0$$

  Exemplo: Convertendo 157 para decimal:

$$1 \times 8^2 + 5 \times 8^1 + 7 \times 8^0 = 64 + 40 + 7 = 111$$

Sistema Hexadecimal (Base 16)

- **Base**: 16
- **Símbolos**: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, A (10), B (11), C (12), D (13), E (14), F (15)
- **Utilização**: Utilizado em programação para representar valores binários de forma mais concisa, especialmente em cores e endereços de memória.
- **Conversão**:

A conversão de um número hexadecimal para decimal segue a mesma lógica, mas com base 16:

$$\text{Decimal} = h_n \times 16^n + h_{n-1} \times 16^{n-1} + \ldots + h_1 \times 16^1 + h_0 \times 16^0$$

Lembre-se de que os dígitos A, B, C, D, E e F representam 10, 11, 12, 13, 14 e 15, respectivamente.

Exemplo: Convertendo 1A3 para decimal:

$$1 \times 16^2 + 10 \times 16^1 + 3 \times 16^0 = 256 + 160 + 3 = 419$$

A tabela abaixo mostra como alguns números são representados em cada base:

BASE	DECIMAL	BINÁRIO	OCTAL	HEXADECIMAL
REPRESENTAÇÃO	0	0	0	0
	1	1	1	1
	2	10	2	2
	3	11	3	3
	4	100	4	4
	5	101	5	5
	6	110	6	6
	7	111	7	7
	8	1000	10	8
	9	1001	11	9
	10	1010	12	A
	11	1011	13	B
	12	1100	14	C
	13	1101	15	D
	14	1110	16	E
	15	1111	17	F
	16	10000	20	10
	20	10100	24	14
	30	11110	36	1E

	40	101000	50	28
	50	110010	62	32
	60	111100	74	3C
	70	1000110	106	46
	80	1010000	120	50
	90	1011010	132	5A
	100	11001000	144	64
	1000	1111101000	1750	3E8
	2989	101110101101	5655	BAD

Representação de números inteiros

Vamos ver os principais métodos e considerações envolvidos na representação de números inteiros.

Representação de Números Inteiros Sem Sinal (Unsigned Integers)

Representação Binária

- Números inteiros sem sinal são representados na memória usando o sistema binário.
- Por exemplo, o número 5 seria representado como 00000101 em um byte (8 bits).

Limitações

- A representação sem sinal só pode representar números não negativos.
- O valor máximo depende do número de bits disponíveis. Com 8 bits, o valor máximo é 255.

Representação de Números Inteiros com Sinal (Signed Integers)

Complemento a Dois

- A forma mais comum de representar números inteiros com sinal é usando o complemento a dois.
- O bit mais significativo (à esquerda) é usado como bit de sinal: 0 para positivo, 1 para negativo.
- Os outros bits representam o valor absoluto do número.

Veja como você representaria -5 em complemento de dois:

Primeiro, represente o valor absoluto do número (neste caso, 5) em binário:

5=00000101

Em seguida, inverta todos os bits do número. Isso é conhecido como o complemento de um:

00000101→11111010

Finalmente, adicione 1 ao complemento de um. Isso nos dá o complemento de dois:

11111010+1=11111011

Agora, 11111011 representa -5 em complemento de dois usando um byte.

Por Que Usar o Complemento de Dois?

O complemento de dois é útil porque permite que os circuitos do computador usem as mesmas operações de adição e subtração para números tanto positivos quanto negativos. Por exemplo, se você somar 11111011 (representando -5) com 00000101 (representando +5), obterá 00000000, que é 0 em binário, exatamente como esperado.

Representação de Largura Fixa vs. Largura Variável

Largura Fixa (como int, short, long)

- Utiliza um número fixo de bits, geralmente 8, 16, 32 ou 64.
- Mais eficiente em termos de velocidade, mas pode desperdiçar espaço se a faixa de valores for limitada.

Largura Variável (como BigInt em algumas linguagens)

- Utiliza uma quantidade variável de bits, dependendo do valor.
- Mais eficiente em termos de espaço, mas geralmente mais lento para operar.

Considerações de Plataforma e Linguagem

- Diferentes sistemas operacionais e linguagens de programação podem ter suas próprias convenções para representar inteiros.
- Considerações como a ordem dos bytes podem afetar como os números inteiros são armazenados na memória.

Representação de números reais

A representação de números reais na memória do computador é uma tarefa complexa, pois esses números podem ter partes fracionárias que variam em magnitude e precisão. Os números reais são comumente representados usando o padrão de ponto flutuante, e o padrão IEEE 754 é normalmente utilizado. Aqui está uma visão geral de como os números reais são representados na memória:

Padrão IEEE 754

Este é o padrão mais comum para representação de ponto flutuante. Define como números reais são armazenados em formatos de precisão simples e dupla.

Componentes

- **Bit de Sinal**: Indica o sinal do número (0 para positivo, 1 para negativo).
- **Expoente**: Representa o poder de 2 que o número é multiplicado.
- **Mantissa (ou Fração)**: Representa a parte fracionária do número.

Precisão Simples (32 bits)

Na precisão simples, o número é representado em 32 bits, divididos da seguinte forma:

- 1 bit para o sinal
- 8 bits para o expoente
- 23 bits para a mantissa

Precisão Dupla (64 bits)

Na precisão dupla, o número é representado em 64 bits:

- 1 bit para o sinal
- 11 bits para o expoente
- 52 bits para a mantissa

Representação Normalizada

Os números são normalmente representados de forma que o dígito mais significativo à esquerda da parte fracionária seja 1. Isso permite maior precisão.

Valores Especiais

O padrão IEEE 754 inclui representações para zero, infinito, e "não é um número" (NaN), o que ajuda no tratamento de casos excepcionais.

Exemplos

Positivo Normalizado

- Número Real: +6.75
- Precisão Simples (32 bits): Sinal: 0, Expoente: 10000001, Mantissa: 10110000000000000000000
- Precisão Dupla (64 bits): Sinal: 0, Expoente: 10000000001, Mantissa: 1011000000000000000000000000000000000000000000000000

Negativo Normalizado

- Número Real: -6.75
- Precisão Simples (32 bits): Sinal: 1, Expoente: 10000001, Mantissa: 10110000000000000000000
- Precisão Dupla (64 bits): Sinal: 1, Expoente: 10000000001, Mantissa: 1011000000000000000000000000000000000000000000000000

Zero Positivo

- Número Real: +0
- Precisão Simples (32 bits): Sinal: 0, Expoente: 00000000, Mantissa: 00000000000000000000000
- Precisão Dupla (64 bits): Sinal: 0, Expoente: 00000000000, Mantissa: 0000000000000000000000000000000000000000000000000000

Zero Negativo

- Número Real: -0
- Precisão Simples (32 bits): Sinal: 1, Expoente: 00000000, Mantissa: 00000000000000000000000
- Precisão Dupla (64 bits): Sinal: 1, Expoente: 00000000000, Mantissa: 0000000000000000000000000000000000000000000000000000

Infinito Positivo

- Número Real: +∞
- Precisão Simples (32 bits): Sinal: 0, Expoente: 11111111, Mantissa: 00000000000000000000000
- Precisão Dupla (64 bits): Sinal: 0, Expoente: 11111111111, Mantissa: 0000000000000000000000000000000000000000000000000000

Infinito Negativo

- Número Real: -∞
- Precisão Simples (32 bits): Sinal: 1, Expoente: 11111111, Mantissa: 00000000000000000000000
- Precisão Dupla (64 bits): Sinal: 1, Expoente: 11111111111, Mantissa: 0000000000000000000000000000000000000000000000000000
- Nota: Infinito

Não é um Número (NaN)

- Número Real: NaN
- Precisão Simples (32 bits): Sinal: -, Expoente: 11111111, Mantissa: xxxxxxxxxxxxxxxxxxxxxxx
- Precisão Dupla (64 bits): Sinal: -, Expoente: 11111111111, Mantissa: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
- Nota: NaN (a mantissa não é toda zero)

Limitações e Desafios

- **Precisão**: A precisão é limitada pela quantidade de bits disponíveis. Isso pode levar a erros de arredondamento.
- **Alcance**: A representação de ponto flutuante tem limites para números muito grandes ou muito pequenos.
- **Velocidade**: Operações de ponto flutuante podem ser mais lentas do que operações inteiras.

Representação de dados textuais

A representação textual em sistemas de computadores é um aspecto importante, pois facilita a comunicação e o processamento de informações em diversos idiomas e scripts. Vejamos mais detalhes sobre as representações textuais ASCII, UNICODE e UTF-8:

ASCII (American Standard Code for Information Interchange)

- **Definição**: ASCII é um padrão de codificação de caracteres que usa 7 bits para representar cada caractere.
- **Características**:
  - Contém 128 caracteres, incluindo letras do alfabeto inglês, dígitos, pontuação e caracteres de controle.
  - Fácil de implementar e amplamente suportado.
- **Limitações**:

  - Restrito ao alfabeto inglês e símbolos comuns; não suporta caracteres de outros idiomas ou scripts.
- **Uso**: Amplamente utilizado em sistemas e protocolos antigos.

Veja os caracteres ASCII nas tabelas abaixo:

Hexa	Nome	Significado	Hexa	Nome	Significado
0	NUL	Null	10	DLE	Data Link Escape
1	SOH	Start Of Heading	11	DC1	Device Control 1
2	STX	Start Of TeXt	12	DC2	Device Control 2
3	ETX	End Of TeXt	13	DC3	Device Control 3
4	EOT	End Of Transmission	14	DC4	Device Control 4
5	ENQ	Enquiry	15	NAK	Negative AcKnowledgement
6	ACK	ACKnowledgement	16	SYN	SYNchronous idle
7	BEL	BELl	17	ETB	End of Transmission Block
8	BS	BackSpace	18	CAN	CANcel
9	HT	Horizontal Tab	19	EM	End of Medium
A	LF	Line Feed	1A	SUB	SUBstitute
B	VT	Vertical Tab	1B	ESC	ESCape
C	FF	Form Feed	1C	FS	File Separator
D	CR	Carriage Return	1D	GS	Group Separator
E	SO	Shift Out	1E	RS	Record Separator
F	SI	Shift In	1F	US	Unit Separator

Já a segunda parte da tabela é uma continuação da primeira, mas representando caracteres imprimíveis:

Hexa	Car	Hexa	Car	Hexa	Car	Hexa	Car	Hexa	Car	Hexa	Car
20	Espaço	30	0	40	@	50	P	60	'	70	p
21	!	31	1	41	A	51	Q	61	a	71	q
22	"	32	2	42	B	52	R	62	b	72	r
23	#	33	3	43	C	53	S	63	c	73	s
24	$	34	4	44	D	54	T	64	d	74	t
25	%	35	5	45	E	55	U	65	e	75	u
26	&	36	6	46	F	56	V	66	f	76	v
27	'	37	7	47	G	57	W	67	g	77	w
28	(	38	8	48	H	58	X	68	h	78	x
29	)	39	9	49	I	59	Y	69	i	79	y
2A	*	3A	:	4A	J	5A	Z	6A	j	7A	z
2B	+	3B	;	4B	K	5B	[	6B	k	7B	{
2C	,	3C	<	4C	L	5C	\	6C	l	7C	|
2D	-	3D	=	4D	M	5D	]	6D	m	7D	}
2E	.	3E	>	4E	N	5E	^	6E	n	7E	~
2F	/	3F	?	4F	O	5F	_	6F	o	7F	DEL

UNICODE

- **Definição**: UNICODE é um padrão de codificação que fornece um código único para cada caractere, independentemente da plataforma, programa ou idioma.

- **Características**:
  - Suporta mais de 1 milhão de caracteres, incluindo quase todos os scripts escritos do mundo.
  - Pode ser implementado através de diferentes codificações, como UTF-8, UTF-16 e UTF-32.
- **Vantagens**:
  - Facilita a internacionalização, permitindo a representação de muitos idiomas em um único documento.
- **Desafios**:
  - A variedade de codificações pode criar complexidade na interpretação e conversão de textos.

UTF-8 (8-bit Unicode Transformation Format)

- **Definição**: UTF-8 é uma codificação de caracteres que representa UNICODE em 8-bit. É uma forma específica de implementar o padrão UNICODE.
- **Características**:
  - Utiliza um número variável de bytes (1 a 4) para cada caractere.
  - Compatível com ASCII, significando que os primeiros 128 caracteres são os mesmos que em ASCII.
  - Capaz de representar qualquer caractere no padrão UNICODE.
- **Vantagens**:
  - Eficiente em termos de espaço para textos em inglês e outros scripts que usam caracteres ASCII.
  - Amplamente adotado na web e em diversos sistemas operacionais.
- **Desafios**:
  - A representação de caracteres que exigem mais bytes pode ser menos eficiente em termos de espaço.

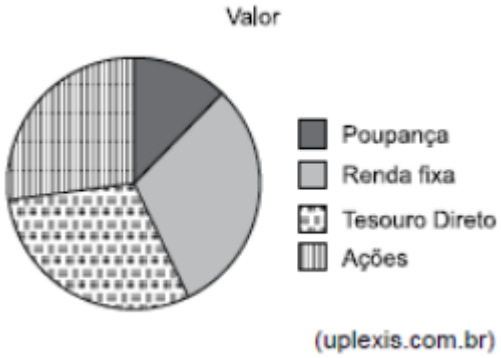
QUESTÕES ESTRATÉGICAS

Nesta seção, apresentamos e comentamos uma amostra de questões objetivas selecionadas estrategicamente: são questões com nível de dificuldade semelhante ao que você deve esperar para a sua prova e que, em conjunto, abordam os principais pontos do assunto.

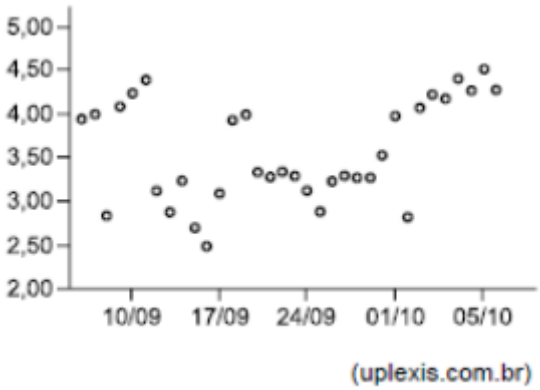
A ideia, aqui, não é que você fixe o conteúdo por meio de uma bateria extensa de questões, mas que você faça uma boa revisão global do assunto a partir de, relativamente, poucas questões.

1. **(VUNESP / UNICAMP - 2019)** Assinale dentre os exemplos a seguir, o gráfico de dispersão.

a)

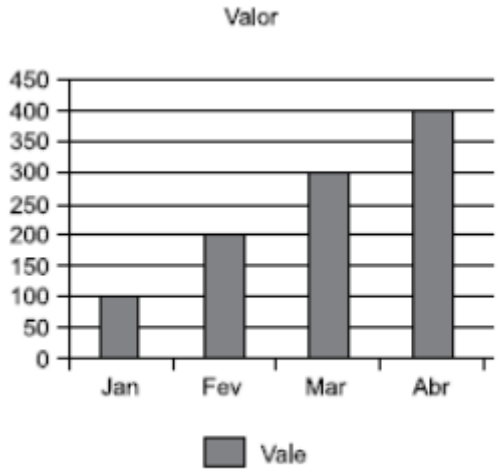


b)



(uplexis.com.br)

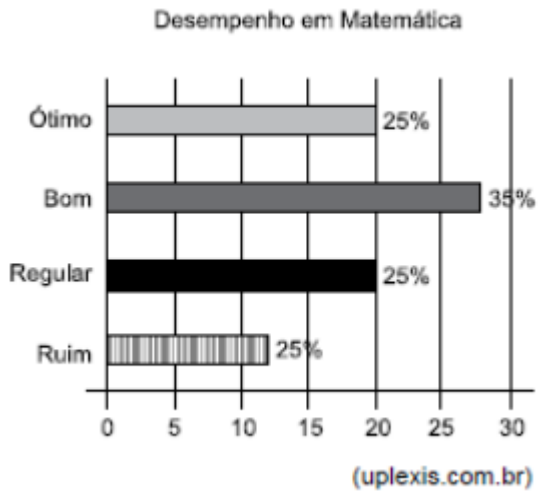
c)



(uplexis.com.br)

d)

e)



Comentários:

(a) Errado, trata-se de um Gráfico de Pizza; (b) Correto; (c) Errado, trata-se de um Gráfico de Linhas; (d) Errado, trata-se de um Gráfico de Barras Vertical (Gráfico de Colunas); (e) Errado, trata-se de um Gráfico de Barras Horizontal.

Gabarito: Letra B

2. **(CESPE / APEX - 2022)** O gráfico por meio do qual é possível representar localização, dispersão, assimetria, comprimento da cauda e outliers, mediante o mínimo, o primeiro quartil, a mediana, o terceiro quartil e o máximo, é denominado:

a) gráfico de linha.

b) gráfico de setor.

c) box plot.

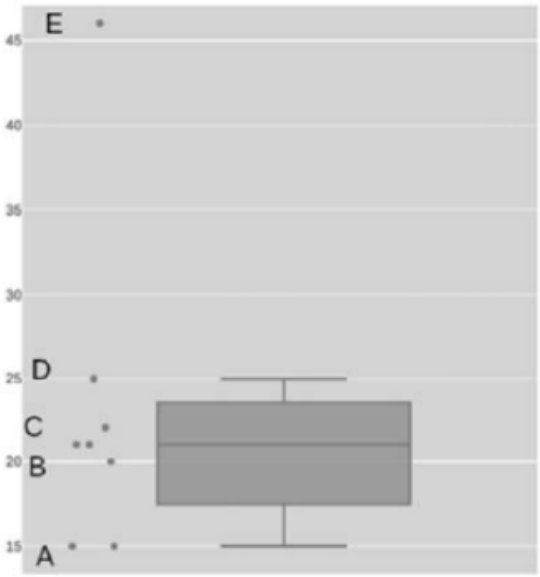
d) scatter plot.

Comentários:

O gráfico que permite representar tudo isso é o BoxPlot. Ele mostra a média dos dados, assim como a variância, bem como os quartis (25%, 50% e 75%), os pontos fora da curva (outliers), localização, dispersão, assimetria, comprimento de cauda, etc.

Gabarito: Letra C

3. **(CESPE / TELEBRAS - 2022)** No gráfico boxplot anteriormente apresentado, o outlier do conjunto de dados é representado pelo ponto:



a) A.

b) E.

c) B.

d) C.

e) D.

Comentários:

O *outlier* é o valor fora da curva, discrepante, destoante! *Onde vocês veem um valor bem distante dos usuais?* No ponto E!

Gabarito: Letra B

4. **(CESGRANRIO / Banco do Brasil – 2021)** Após a coleta de dados em um determinado contexto (variáveis A, B, C, … X), uma das formas mais simples e iniciais de análise é a geração e a avaliação de um histograma para uma variável selecionada (ex: X), como por exemplo, em um estudo

climático, em que os dados coletados poderiam incluir a temperatura máxima observada em toda a Terra ao longo de dez anos.

Nesse caso, o histograma adequado é um gráfico em que são apresentadas as:

a) últimas dez médias móveis da variável X

b) somas das médias dos quadrados de cada valor de uma variável X

c) variações de uma variável X ao longo do tempo

d) médias históricas da variável X nos últimos sete dias

e) frequências de uma variável X em intervalos de valores

Comentários:

Um histograma é um gráfico que mostra a distribuição de frequência de dados, com os dados organizados em classes ou intervalos e contados em barras verticais. Ele é utilizado para mostrar como a frequência dos dados se distribui ao longo de um intervalo de valores.

Cada retângulo do histograma representa uma classe, a altura do retângulo representa a frequência e a largura do retângulo representa a amplitude ou intervalo de valores. Logo, o histograma adequado é um gráfico em que são apresentadas as frequências de uma variável X em intervalos de valores.

Gabarito: Letra E

5. **(COPESE-UFT / Prefeitura de Palmas - 2014)** O Box-Plot (gráfico de caixa) é ferramenta útil na análise exploratória de dados. O propósito do gráfico é fornecer ao analista uma primeira ideia da distribuição dos dados. Sobre o gráfico em questão, analise as afirmativas.

I. Quando a linha que representa a mediana estiver equidistante dos outros quartis a distribuição será simétrica.

II. Quando a linha que representa a mediana estiver mais próxima do 1º quartil que do 3º quartil a distribuição será assimétrica à direita.

III. Quando a linha que representa a mediana estiver mais próxima do 3º quartil que do 1º quartil a distribuição será assimétrica à esquerda.

Marque a alternativa CORRETA.

a) Apenas a afirmativa I está correta.

b) Apenas as afirmativas I e II estão corretas.

c) Todas as afirmativas estão corretas.

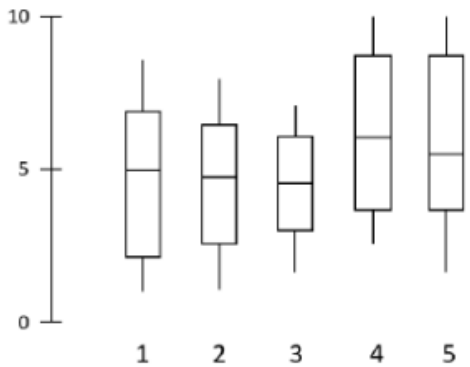
d) Todas as afirmativas estão incorretas.

Comentários:

(I) Correto. Uma distribuição simétrica significa que os dados têm uma simetria em relação à mediana. Ademais, a quantidade de dados acima da mediana é aproximadamente igual à quantidade de dados abaixo da mediana; (II) Correto. Uma distribuição assimétrica à direita significa que a maioria dos dados está concentrada à direita da mediana, ou seja, temos uma maior concentração de valores altos na distribuição. Isto geralmente indica que a distribuição tem uma cauda longa à direita, logo existem alguns valores muito elevados que estão muito distantes da mediana. (III) Correto. Uma distribuição assimétrica à esquerda significa que existe um lado mais curto da distribuição, que é o lado esquerdo. Isso indica que a maioria dos dados está concentrada na direita, e que há poucos dados em relação à esquerda.

Gabarito: Letra D

9. **(FGV / TRT13 - 2022)** Os diagramas a seguir são Box-Plots de notas de cinco turmas de alunos de um mesmo colégio numa prova de matemática:



A maior mediana das notas foi obtida pela turma:

a) 1.
b) 2.
c) 3.

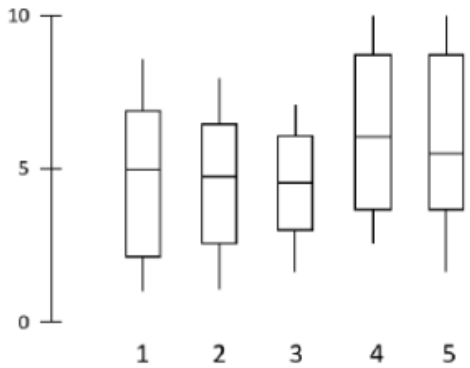
d) 4.
e) 5.

Comentários:

Mediana é o Q2! A maior mediana está representada na Turma 4, dado que ela tem a linha central mais alta.

Gabarito: Letra D

10. (FGV / TRT13 - 2022) Os diagramas a seguir são Box-Plots de notas de cinco turmas de alunos de um mesmo colégio numa prova de matemática:



A turma com notas mais homogêneas nessa prova foi a:

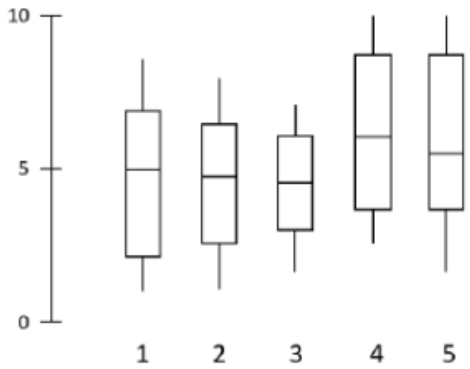
a) 1.
b) 2.
c) 3.
d) 4.
e) 5.

Comentários:

A turma com notas mais homogêneas é aquela em que tem o menor intervalo interquartílico. Logo, é possível ver pela imagem que a turma representada pela menor caixa é a Turma 3.

Gabarito: Letra C

11. **(FGV / TRT13 - 2022)** Os diagramas a seguir são Box-Plots de notas de cinco turmas de alunos de um mesmo colégio numa prova de matemática:



A turma com notas mais homogêneas nessa prova foi a:

a) 1.
b) 2.
c) 3.
d) 4.
e) 5.

Comentários:

A turma com notas mais homogêneas é aquela em que tem o menor intervalo interquartílico. Logo, é possível ver pela imagem que a turma representada pela menor caixa é a Turma 3.

Gabarito: Letra C

12. **(CESGRANRIO / Banco do Brasil – 2021)** Após a coleta de dados em um determinado contexto (variáveis A, B, C, … X), uma das formas mais simples e iniciais de análise é a geração e a avaliação de um histograma para uma variável selecionada (ex: X), como por exemplo, em um estudo climático, em que os dados coletados poderiam incluir a temperatura máxima observada em toda a Terra ao longo de dez anos.

Nesse caso, o histograma adequado é um gráfico em que são apresentadas as:

a) últimas dez médias móveis da variável X
b) somas das médias dos quadrados de cada valor de uma variável X
c) variações de uma variável X ao longo do tempo
d) médias históricas da variável X nos últimos sete dias
e) frequências de uma variável X em intervalos de valores

Comentários:

Um histograma é um gráfico que mostra a distribuição de frequência de dados, com os dados organizados em classes ou intervalos e contados em barras verticais. Ele é utilizado para mostrar como a frequência dos dados se distribui ao longo de um intervalo de valores.

Cada retângulo do histograma representa uma classe, a altura do retângulo representa a frequência e a largura do retângulo representa a amplitude ou intervalo de valores. Logo, o histograma adequado é um gráfico em que são apresentadas as frequências de uma variável X em intervalos de valores.

Gabarito: Letra E

13. **(CESGRANRIO / BB – 2021)** Um funcionário de um banco foi incumbido de acompanhar o perfil dos clientes de um determinado produto por meio da Análise de Dados, de forma a aprimorar as atividades de marketing relativas a esse produto. Para isso, ele utilizou a variável classe social desses clientes, coletada pelo banco, que tem os valores A, B, C, D e E, sem referência a valores contínuos.

Sabendo-se que essa é uma escala ordinal, qual é a medida de tendência central adequada para analisar essa variável?

a) média aritmética

b) média geométrica

c) mediana

d) quartis

e) variância

Comentários:

VARIÁVEL	MÉDIA	MEDIANA	MODA
QUANTITATIVA DISCRETA	OK	OK	OK
QUANTITATIVA CONTÍNUA	OK	OK	OK
QUALITATIVA ORDINAL		OK	OK
QUALITATIVA NOMINAL			OK

Basta lembrar dessa tabelinha. Por meio de uma escala ou variável ordinal, é possível avaliar a mediana ou a moda.

Gabarito: C

QUESTIONÁRIO DE REVISÃO E APERFEIÇOAMENTO

A ideia do questionário é elevar o nível da sua compreensão no assunto e, ao mesmo tempo, proporcionar uma outra forma de revisão de pontos importantes do conteúdo, a partir de perguntas que exigem respostas subjetivas.

São questões um pouco mais desafiadoras, porque a redação de seu enunciado não ajuda na sua resolução, como ocorre nas clássicas questões objetivas.

O objetivo é que você realize uma auto explicação mental de alguns pontos do conteúdo, para consolidar melhor o que aprendeu ;)

Além disso, as questões objetivas, em regra, abordam pontos isolados de um dado assunto. Assim, ao resolver várias questões objetivas, o candidato acaba memorizando pontos isolados do conteúdo, mas muitas vezes acaba não entendendo como esses pontos se conectam.

Assim, no questionário, buscaremos trazer também situações que ajudem você a conectar melhor os diversos pontos do conteúdo, na medida do possível.

É importante frisar que não estamos adentrando em um nível de profundidade maior que o exigido na sua prova, mas apenas permitindo que você compreenda melhor o assunto de modo a facilitar a resolução de questões objetivas típicas de concursos, ok?

Nosso compromisso é proporcionar a você uma revisão de alto nível!

Vamos ao nosso questionário:

Perguntas

1. O que é Análise Exploratória de Dados (AED) e quais são seus principais objetivos?
2. O que é uma 'observação' em um conjunto de dados?
3. O que é uma 'feature' em um conjunto de dados?
4. Qual é a principal utilidade de um gráfico de linha e quando deve ser usado?
5. O que é um rótulo em aprendizado de máquina supervisionado?
6. Como um gráfico de barras pode ser útil na análise de dados?
7. Quando é apropriado usar um gráfico de pizza?
8. Por que um gráfico de dispersão é útil e quando deve ser usado?
9. Como um gráfico de contorno é útil e quando deve ser usado?
10. O que é um gráfico de área e quando deve ser usado?
11. Qual é a principal utilidade de um gráfico de rede?
12. O que é um histograma e quando deve ser usado?
13. O que é um box plot e quando deve ser usado?
14. O que são metadados?
15. Por que é importante escolher o tipo de gráfico correto ao visualizar dados?
16. Qual é a diferença entre um gráfico de barras e um histograma?
17. Por que poderíamos querer usar um gráfico de dispersão em vez de um gráfico de linha para representar dados?
18. O que é um outlier e como um box plot pode ajudar a identificá-lo?
19. Quando é preferível usar um gráfico de área em vez de um gráfico de linha?
20. Como um gráfico de rede pode ser útil na análise de dados?
21. Explique o papel do bit de sinal, expoente e mantissa na representação de números reais no padrão IEEE 754.
22. Como é representado o número -5 em complemento de dois usando um byte?
23. Quais são as principais diferenças entre as codificações ASCII, UNICODE e UTF-8?
24. Descreva o processo de converter um número binário para decimal.
25. O que é o padrão de ponto flutuante IEEE 754 e por que é importante?

26. Como os números inteiros negativos são representados na memória do computador?
27. Qual é a diferença entre precisão simples e dupla na representação de números de ponto flutuante?
28. Explique o conceito de UNICODE e sua importância na representação textual global.
29. Como você converteria o número 1A3 de hexadecimal para decimal?
30. Quais são as limitações do padrão ASCII e como elas foram superadas com a introdução do UNICODE?

Perguntas e Respostas

1. O que é Análise Exploratória de Dados (AED) e quais são seus principais objetivos?
   Resposta: AED é uma abordagem para analisar conjuntos de dados a fim de resumir suas características principais, geralmente com métodos visuais. Seus objetivos principais incluem a identificação de tendências e padrões, a detecção de outliers e anomalias, o teste de hipóteses e a preparação de dados para modelagem preditiva.

2. O que é uma 'observação' em um conjunto de dados?
   Resposta: Uma observação em um conjunto de dados refere-se a um único ponto de dados ou registro.

3. O que é uma 'feature' em um conjunto de dados?
   Resposta: Uma feature, ou variável, em um conjunto de dados é uma característica individual que é medida para cada observação.

4. Qual é a principal utilidade de um gráfico de linha e quando deve ser usado?
   Resposta: Um gráfico de linha é usado principalmente para visualizar uma tendência nos dados ao longo do tempo, sendo particularmente útil para séries temporais.

5. O que é um rótulo em aprendizado de máquina supervisionado?
   Resposta: Um rótulo é o valor que queremos prever ou classificar em aprendizado de máquina supervisionado.

6. Como um gráfico de barras pode ser útil na análise de dados?
   Resposta: Um gráfico de barras é útil para comparar quantidades de diferentes categorias ou grupos.

7. Quando é apropriado usar um gráfico de pizza?
   Resposta: Um gráfico de pizza é apropriado quando se quer ilustrar proporções numéricas entre categorias diferentes de um todo.

8. Por que um gráfico de dispersão é útil e quando deve ser usado?
   Resposta: Um gráfico de dispersão é útil para visualizar a relação entre duas variáveis numéricas e é comumente usado quando se quer entender como uma variável é afetada por outra.

9. Como um gráfico de contorno é útil e quando deve ser usado?
   Resposta: Um gráfico de contorno é útil para visualizar uma relação tridimensional em duas dimensões. É usado quando há três variáveis numéricas e se quer ver como as duas variáveis independentes afetam a variável dependente.

10. O que é um gráfico de área e quando deve ser usado?
    Resposta: Um gráfico de área representa a magnitude de uma tendência ao longo do tempo. Ele deve ser usado quando se quer mostrar como uma quantidade total é composta ao longo do tempo, para várias categorias.

11. Qual é a principal utilidade de um gráfico de rede?
    Resposta: Um gráfico de rede é útil para visualizar conexões ou relacionamentos entre entidades.

12. O que é um histograma e quando deve ser usado?
    Resposta: Um histograma é um gráfico que mostra a distribuição de frequência de um conjunto de dados. É útil para visualizar a distribuição dos dados e deve ser usado quando se quer ter uma ideia da densidade e da distribuição central dos dados.

13. O que é um box plot e quando deve ser usado?
    Resposta: Um box plot é um gráfico que representa estatísticas resumidas de um conjunto de dados, incluindo os quartis, a mediana, e possíveis outliers. Ele é usado para visualizar a distribuição e a dispersão dos dados.

14. O que são metadados?
    Resposta: Metadados são dados sobre os dados. Eles fornecem informações como quando e por quem os dados foram coletados, como os dados estão formatados, a fonte dos dados, etc.

15. Por que é importante escolher o tipo de gráfico correto ao visualizar dados?
    Resposta: A escolha do gráfico correto é importante porque diferentes gráficos são apropriados para diferentes tipos de dados e objetivos de análise. O gráfico correto pode tornar os dados mais fáceis de entender e pode revelar insights que de outra forma poderiam ser perdidos.

16. Qual é a diferença entre um gráfico de barras e um histograma?
    Resposta: Embora ambos usem barras para representar dados, um gráfico de barras compara

diferentes grupos ou categorias, enquanto um histograma mostra a distribuição de uma única variável contínua.

17. Por que poderíamos querer usar um gráfico de dispersão em vez de um gráfico de linha para representar dados?
    Resposta: Um gráfico de dispersão é preferido quando queremos entender a relação entre duas variáveis numéricas, independentemente do tempo. Um gráfico de linha, por outro lado, é usado principalmente para mostrar uma tendência ao longo do tempo.

18. O que é um outlier e como um box plot pode ajudar a identificá-lo?
    Resposta: Um outlier é um valor que é significativamente diferente da maioria dos outros valores em um conjunto de dados. Um box plot pode ajudar a identificar outliers representando-os como pontos individuais que estão distantes do corpo principal do gráfico (a 'caixa').

19. Quando é preferível usar um gráfico de área em vez de um gráfico de linha?
    Resposta: Um gráfico de área é preferível quando se quer mostrar como uma quantidade total é composta ao longo do tempo, especialmente quando há várias categorias a serem comparadas.

20. Como um gráfico de rede pode ser útil na análise de dados?
    Resposta: Um gráfico de rede é útil para visualizar complexas interações e relações entre diferentes entidades. Ele pode ajudar a identificar grupos ou comunidades dentro dos dados, entender a influência entre os nós e analisar a estrutura geral das conexões.

21. Explique o papel do bit de sinal, expoente e mantissa na representação de números reais no padrão IEEE 754.
    Resposta: O bit de sinal indica se o número é positivo ou negativo (0 para positivo, 1 para negativo). O expoente determina a potência de 2 pela qual a parte fracionária é multiplicada. A mantissa, ou fração, representa a parte fracionária do número em notação científica. Juntos, esses componentes permitem a representação precisa de uma vasta gama de números reais.

22. Como é representado o número -5 em complemento de dois usando um byte?
    Resposta: O número -5 é representado em complemento de dois como 11111011. Esse processo envolve pegar a representação binária de 5, inverter todos os bits e adicionar 1 ao resultado.

23. Quais são as principais diferenças entre as codificações ASCII, UNICODE e UTF-8?
    Resposta: ASCII é uma codificação de 7 bits que suporta 128 caracteres, principalmente para o inglês. UNICODE é um padrão global que pode representar mais de 1 milhão de caracteres,

incluindo scripts de diversos idiomas. UTF-8 é uma implementação específica do UNICODE que usa 1 a 4 bytes por caractere e é compatível com ASCII.

24. Descreva o processo de converter um número binário para decimal.
    Resposta: Para converter um número binário para decimal, multiplicamos cada dígito pelo valor da potência de 2 correspondente à sua posição e somamos os resultados. Por exemplo, 1011 em binário é 1 * 2^3 + 0 * 2^2 + 1 * 2^1 + 1 * 2^0 = 11 em decimal.

25. O que é o padrão de ponto flutuante IEEE 754 e por que é importante?
    Resposta: O padrão IEEE 754 define a representação e a aritmética de números de ponto flutuante em computadores. É crucial para garantir consistência e precisão em cálculos científicos, gráficos e financeiros em diferentes plataformas e linguagens de programação.

26. Como os números inteiros negativos são representados na memória do computador?
    Resposta: Os números inteiros negativos são comumente representados usando o complemento de dois, onde o bit mais significativo é usado como bit de sinal e os outros bits representam o valor absoluto do número em forma complementar.

27. Qual é a diferença entre precisão simples e dupla na representação de números de ponto flutuante?
    Resposta: A precisão simples usa 32 bits (1 bit de sinal, 8 bits de expoente, 23 bits de mantissa), enquanto a precisão dupla usa 64 bits (1 bit de sinal, 11 bits de expoente, 52 bits de mantissa). A precisão dupla oferece maior alcance e precisão, mas consome mais espaço de memória.

28. Explique o conceito de UNICODE e sua importância na representação textual global.
    Resposta: UNICODE é um padrão de codificação que fornece um código único para cada caractere, abrangendo quase todos os scripts escritos do mundo. É essencial para a internacionalização, permitindo a representação e troca de textos em diversos idiomas e culturas.

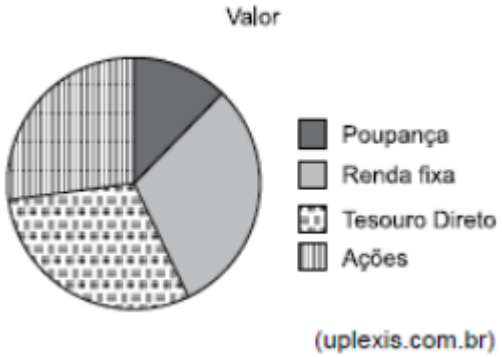
29. Como você converteria o número 1A3 de hexadecimal para decimal?
    Resposta: A conversão do número 1A3 em hexadecimal para decimal é feita como 1 * 16^2 + 10 * 16^1 + 3 * 16^0 = 256 + 160 + 3 = 419.

30. Quais são as limitações do padrão ASCII e como elas foram superadas com a introdução do UNICODE?
    Resposta: O padrão ASCII está limitado a 128 caracteres, adequado principalmente para o alfabeto inglês e símbolos comuns. Não suporta caracteres de outros idiomas ou scripts. O UNICODE supera essas limitações, fornecendo um padrão global que pode representar caracteres de quase todos os scripts escritos do mundo.

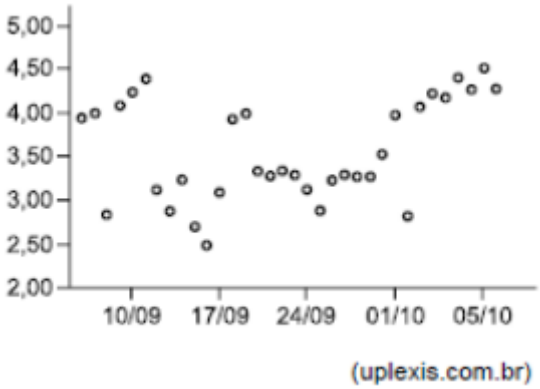
LISTA DE QUESTÕES ESTRATÉGICAS

1. **(VUNESP / UNICAMP - 2019)** Assinale dentre os exemplos a seguir, o gráfico de dispersão.

a)

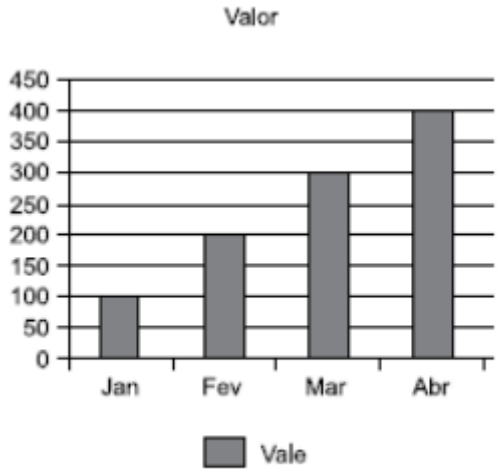


b)



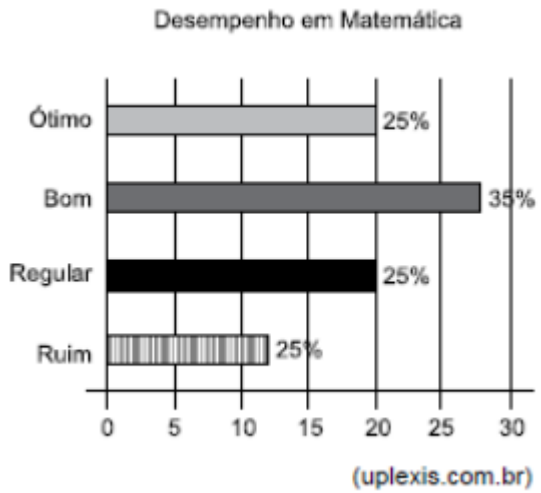
(uplexis.com.br)

c)



(uplexis.com.br)

d)



e)

2. **(CESPE / APEX - 2022)** O gráfico por meio do qual é possível representar localização, dispersão, assimetria, comprimento da cauda e outliers, mediante o mínimo, o primeiro quartil, a mediana, o terceiro quartil e o máximo, é denominado:

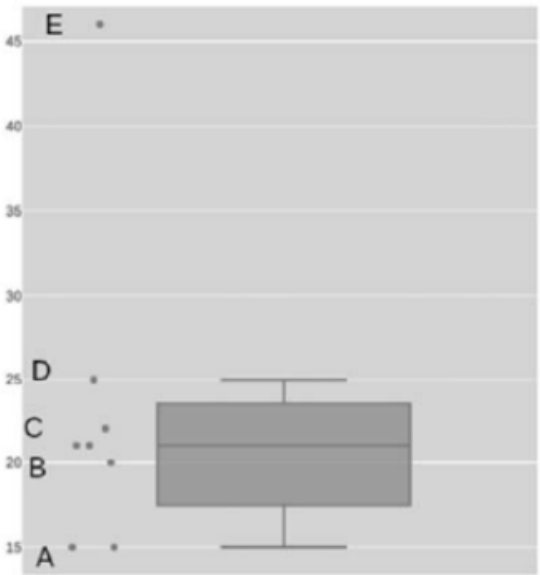
a) gráfico de linha.

b) gráfico de setor.

c) box plot.

d) scatter plot.

3. **(CESPE / TELEBRAS - 2022)** No gráfico boxplot anteriormente apresentado, o outlier do conjunto de dados é representado pelo ponto:



a) A.

b) E.

c) B.

d) C.

e) D.

4. **(CESGRANRIO / Banco do Brasil – 2021)** Após a coleta de dados em um determinado contexto (variáveis A, B, C, … X), uma das formas mais simples e iniciais de análise é a geração e a avaliação de um histograma para uma variável selecionada (ex: X), como por exemplo, em um estudo climático, em que os dados coletados poderiam incluir a temperatura máxima observada em toda a Terra ao longo de dez anos.

   Nesse caso, o histograma adequado é um gráfico em que são apresentadas as:

   a) últimas dez médias móveis da variável X

   b) somas das médias dos quadrados de cada valor de uma variável X

   c) variações de uma variável X ao longo do tempo

   d) médias históricas da variável X nos últimos sete dias

   e) frequências de uma variável X em intervalos de valores

5. **(COPESE-UFT / Prefeitura de Palmas - 2014)** O Box-Plot (gráfico de caixa) é ferramenta útil na análise exploratória de dados. O propósito do gráfico é fornecer ao analista uma primeira ideia da distribuição dos dados. Sobre o gráfico em questão, analise as afirmativas.

   I. Quando a linha que representa a mediana estiver equidistante dos outros quartis a distribuição será simétrica.

   II. Quando a linha que representa a mediana estiver mais próxima do 1º quartil que do 3º quartil a distribuição será assimétrica à direita.

   III. Quando a linha que representa a mediana estiver mais próxima do 3º quartil que do 1º quartil a distribuição será assimétrica à esquerda.

Marque a alternativa CORRETA.

a) Apenas a afirmativa I está correta.

b) Apenas as afirmativas I e II estão corretas.

c) Todas as afirmativas estão corretas.

d) Todas as afirmativas estão incorretas.

6. **(CESPE / ANAC – 2009)** Geralmente, números inteiros são representados em ponto fixo e números fracionários, em ponto flutuante.

Comentários:

Perfeito! Na prática, cálculos com números inteiros e com números em ponto fixo são idênticos, apenas se assume que a vírgula do número inteiro está localizada na posição mais à direita do valor (o valor inteiro 10 pode ser representado em ponto fixo como 10,0). Logo, para permitir esse tipo de cálculo, os números inteiros já são representados como números em ponto fixo – é claro que tudo isso também é válido para números binários. Já os números fracionários realmente são representados em ponto flutuante porque oferecem uma precisão muito maior utilizando menos bits.

Gabarito: Correto

7. **(CESPE / Polícia Federal – 2004)** Para aplicações científicas, é comum a utilização de números de ponto flutuante em vez de números inteiros. Os processadores atuais suportam a norma IEEE, na qual um número de ponto flutuante com precisão simples é representado em 32 bits, utilizando a notação científica com um bit para o sinal, 8 bits para o expoente e 23 bits para a mantissa.

Comentários:

Perfeito! Aplicações científicas realmente utilizam números de ponto flutuante em vez de números inteiros por conta da precisão. De fato, processadores suportam ponto flutuante formado por 32 bits, sendo 1 bit para o sinal, 8 bits para o expoente e 23 bits para mantissa.

Gabarito: Correto

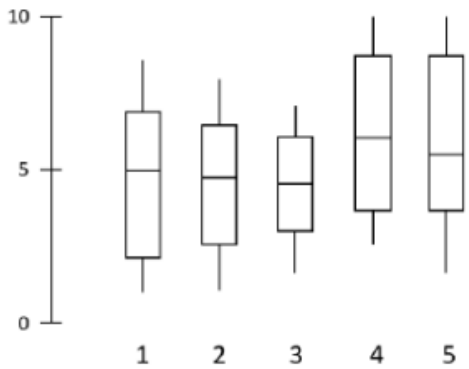
8. **(CESPE / TRT17 – 2013)** Um dado do tipo ponto-flutuante, cujo valor é definido em termos de precisão e faixa de valores, pode pertencer ao conjunto dos números reais, racionais ou irracionais.

Comentários:

Perfeito! Ele define números reais, logo também define números racionais e irracionais porque ambos fazem parte dos números reais.

Gabarito: Correto

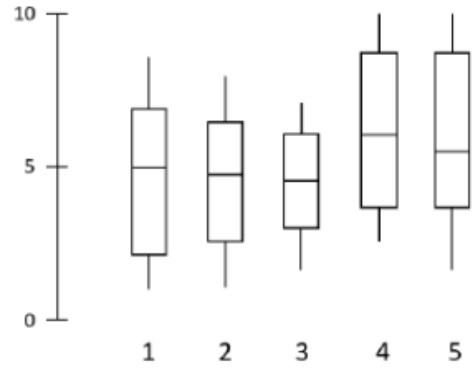
9. **(FGV / TRT13 - 2022)** Os diagramas a seguir são Box-Plots de notas de cinco turmas de alunos de um mesmo colégio numa prova de matemática:



A maior mediana das notas foi obtida pela turma:

a) 1.
b) 2.
c) 3.
d) 4.
e) 5.

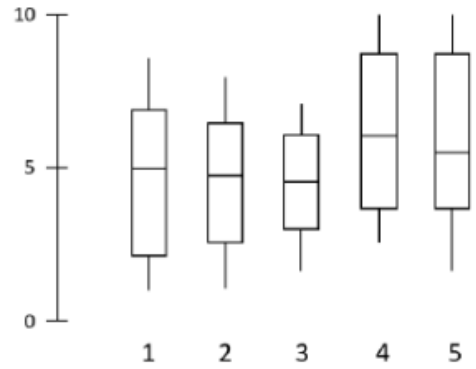
10. **(FGV / TRT13 - 2022)** Os diagramas a seguir são Box-Plots de notas de cinco turmas de alunos de um mesmo colégio numa prova de matemática:



A turma com notas mais homogêneas nessa prova foi a:

a) 1.
b) 2.
c) 3.
d) 4.
e) 5.

11. (FGV / TRT13 - 2022) Os diagramas a seguir são Box-Plots de notas de cinco turmas de alunos de um mesmo colégio numa prova de matemática:



A turma com notas mais homogêneas nessa prova foi a:

a) 1.
b) 2.
c) 3.
d) 4.

e) 5.

12. **(CESGRANRIO / Banco do Brasil – 2021)** Após a coleta de dados em um determinado contexto (variáveis A, B, C, … X), uma das formas mais simples e iniciais de análise é a geração e a avaliação de um histograma para uma variável selecionada (ex: X), como por exemplo, em um estudo climático, em que os dados coletados poderiam incluir a temperatura máxima observada em toda a Terra ao longo de dez anos.

Nesse caso, o histograma adequado é um gráfico em que são apresentadas as:

a) últimas dez médias móveis da variável X
b) somas das médias dos quadrados de cada valor de uma variável X
c) variações de uma variável X ao longo do tempo
d) médias históricas da variável X nos últimos sete dias
e) frequências de uma variável X em intervalos de valores

13. **(CESGRANRIO / BB – 2021)** Um funcionário de um banco foi incumbido de acompanhar o perfil dos clientes de um determinado produto por meio da Análise de Dados, de forma a aprimorar as atividades de marketing relativas a esse produto. Para isso, ele utilizou a variável classe social desses clientes, coletada pelo banco, que tem os valores A, B, C, D e E, sem referência a valores contínuos.

Sabendo-se que essa é uma escala ordinal, qual é a medida de tendência central adequada para analisar essa variável?

a) média aritmética

b) média geométrica

c) mediana

d) quartis

e) variância

Gabaritos

1. B
2. C
3. B
4. E
5. D
6. C
7. C
8. C
9. D
10. C
11. C
12. E
13. C

ESSA LEI TODO MUNDO CONHECE:
PIRATARIA É CRIME.

Mas é sempre bom revisar o porquê e como você pode ser prejudicado com essa prática.

1. 

Professor investe seu tempo para elaborar os cursos e o site os coloca à venda.

2. 

Pirata divulga ilicitamente (grupos de rateio), utilizando-se do anonimato, nomes falsos ou laranjas (geralmente o pirata se anuncia como formador de "grupos solidários" de rateio que não visam lucro).

3. 

Pirata cria alunos fake praticando falsidade ideológica, comprando cursos do site em nome de pessoas aleatórias (usando nome, CPF, endereço e telefone de terceiros sem autorização).

4. 

Pirata compra, muitas vezes, clonando cartões de crédito (por vezes o sistema anti-fraude não consegue identificar o golpe a tempo).

5. 

Pirata fere os Termos de Uso, adultera as aulas e retira a identificação dos arquivos PDF (justamente porque a atividade é ilegal e ele não quer que seus fakes sejam identificados).

6. 

Pirata revende as aulas protegidas por direitos autorais, praticando concorrência desleal e em flagrante desrespeito à Lei de Direitos Autorais (Lei 9.610/98).

7. 

Concurseiro(a) desinformado participa de rateio, achando que nada disso está acontecendo e esperando se tornar servidor público para exigir o cumprimento das leis.

8. 

O professor que elaborou o curso não ganha nada, o site não recebe nada, e a pessoa que praticou todos os ilícitos anteriores (pirata) fica com o lucro.

Deixando de lado esse mar de sujeira, aproveitamos para agradecer a todos que adquirem os cursos honestamente e permitem que o site continue existindo.