**1) Mencione qual é a responsabilidade de um analista de dados?**

A responsabilidade de um analista de dados inclui,

* Fornecer suporte a todas as análises de dados e coordenar com clientes e equipes
* Resolver problemas associados aos negócios para clientes e desempenho [auditor](https://career.guru99.com/pt/top-10-internal-audit-interview-questions/) em dados
* Analisar resultados e interpretar dados usando técnicas estatísticas e fornecer relatórios contínuos
* Priorizar as necessidades do negócio e trabalhar em estreita colaboração com as necessidades de gestão e informação
* Identificar novos processos ou áreas para oportunidades de melhoria
* Analise, identifique e interprete tendências ou padrões em conjuntos de dados complexos
* Adquira dados de fontes de dados primárias ou secundárias e mantenha bancos de dados / sistemas de dados
* Filtre e “limpe” dados e revise relatórios de computador
* Determine indicadores de desempenho para localizar e corrigir problemas de código
* Protegendo o banco de dados desenvolvendo um sistema de acesso determinando o nível de acesso do usuário

**2) O que é necessário para se tornar um analista de dados?**

Para se tornar um analista de dados,

* Conhecimento robusto em pacotes de relatórios (Business Objects), linguagem de programação (frameworks XML, Javascript ou ETL), bancos de dados ([SQL](https://www.guru99.com/sql-server-questions.html), [SQL](https://www.guru99.com/sql-server-questions.html" \o "SQL)ite, etc.)
* Fortes habilidades com capacidade de analisar, organizar, coletar e disseminar big data com precisão
* Conhecimento técnico em design de banco de dados, modelos de dados, mineração de dados e técnicas de segmentação
* Forte conhecimento em pacotes estatísticos para análise de grandes conjuntos de dados ([SAS](https://career.guru99.com/pt/top-17-sas-interview-questions/), **Excel**, SPSS, etc.)

**3) Mencione quais são as várias etapas de um projeto analítico?**

Várias etapas em um projeto de análise incluem

* Definição de problema
* Exploração de dados
* Preparação de dados
* Modelagem
* Validação de dados
* Implementação e rastreamento

**4) Mencione o que é limpeza de dados?**

A limpeza de dados, também conhecida como limpeza de dados, trata da identificação e remoção de erros e inconsistências dos dados, a fim de melhorar a qualidade dos dados.

**5) Liste algumas das melhores práticas para limpeza de dados?**

Algumas das melhores práticas para limpeza de dados incluem,

* Classifique os dados por atributos diferentes
* Para grandes conjuntos de dados, limpe-os passo a passo e melhore os dados a cada etapa até obter uma boa qualidade de dados
* Para grandes conjuntos de dados, divida-os em pequenos dados. Trabalhar com menos dados aumentará sua velocidade de iteração
* Para lidar com tarefas de limpeza comuns, crie um conjunto de funções/ferramentas/scripts utilitários. Pode incluir o remapeamento de valores com base em um arquivo CSV ou banco de dados SQL ou pesquisa e substituição de regex, apagando todos os valores que não correspondem a um regex
* Se você tiver problemas com a limpeza dos dados, organize-os por frequência estimada e ataque os problemas mais comuns
* Analise as estatísticas resumidas de cada coluna (desvio padrão, média, número de valores ausentes)
* Acompanhe todas as operações de limpeza de datas, para que você possa alterar alterações ou remover operações, se necessário

**6) Explique o que é**[**logística**](https://career.guru99.com/pt/top-20-logistician-interview-questions/)**regressão?**

A regressão logística é um método estatístico para examinar um conjunto de dados no qual existem uma ou mais variáveis ​​independentes que definem um resultado.

**7) Lista das melhores ferramentas que podem ser úteis para análise de dados?**

A seguir estão as melhores ferramentas de análise de dados

* [Quadro](https://career.guru99.com/pt/top-10-tableau-interview-questions/)
* RapidMiner
* OpenRefine
* KNIME
* Operadores de pesquisa do Google
* Solver
* NodeXL
* io
* Wolfram Alpha's
* Tabelas do Google Fusion

**8) Mencione qual é a diferença entre mineração de dados e criação de perfil de dados?**

A diferença entre mineração de dados e criação de perfil de dados é que

**Perfil de dados:** Tem como alvo a análise de instância de atributos individuais. Fornece informações sobre vários atributos como faixa de valores, valor discreto e sua frequência, ocorrência de valores nulos, tipo de dados, comprimento, etc.

**Mineração de dados:** Centra-se na análise de cluster, detecção de registros incomuns, dependências, descoberta de sequências, manutenção de relações entre vários atributos, etc.

**9) Liste alguns problemas comuns enfrentados pelo analista de dados?**

Alguns dos problemas comuns enfrentados pelo analista de dados são

* Erro de ortografia comum
* Entradas duplicadas
* Valores faltantes
* Valores ilegais
* Representações de valores variáveis
* Identificando dados sobrepostos

**10) Menciona o nome do framework desenvolvido pela Apache para processamento de grande conjunto de dados para uma aplicação em ambiente de computação distribuída?**

[Hadoop](https://career.guru99.com/pt/top-25-hadoop-admin-interview-questions-and-answers/) e MapReduce é a estrutura de programação desenvolvida pela Apache para processar grandes conjuntos de dados para uma aplicação em um ambiente de computação distribuída.

**11) Mencione quais são os padrões ausentes que geralmente são observados?**

Os padrões ausentes que geralmente são observados são

* Faltando completamente ao acaso
* faltando ao acaso
* Faltando isso depende do próprio valor faltante
* Faltando isso depende da variável de entrada não observada

**12) Explique o que é o método de imputação KNN?**

Na imputação KNN, os valores dos atributos ausentes são imputados usando o valor dos atributos mais semelhante ao atributo cujos valores estão ausentes. Usando uma função de distância, a similaridade de dois atributos é determinada.

**3) Mencione quais são os métodos de validação de dados usados ​​pelo analista de dados?**

Normalmente, os métodos usados ​​pelo analista de dados para validação de dados são

* Triagem de dados
* Verificação de dados

**14) Explique o que deve ser feito com dados suspeitos ou faltantes?**

* Prepare um relatório de validação que forneça informações de todos os dados suspeitos. Deve fornecer informações como critérios de validação que falharam e a data e hora da ocorrência
* Pessoal experiente deve examinar os dados suspeitos para determinar sua aceitabilidade
* Dados inválidos devem ser atribuídos e substituídos por um código de validação
* Para trabalhar com dados ausentes, use a melhor estratégia de análise, como método de exclusão, métodos de imputação única, métodos baseados em modelo, etc.

**15) Mencionou como lidar com os problemas de múltiplas fontes?**

Para lidar com os problemas de múltiplas fontes,

* Reestruturação de esquemas para realizar uma integração de esquema
* Identifique registros semelhantes e mescle-os em um único registro contendo todos os atributos relevantes sem redundância

**16) Explique o que é um Outlier?**

O outlier é um termo comumente usado por analistas para se referir a um valor que parece distante e diverge de um padrão geral em uma amostra. Existem dois tipos de outliers

* Univariada
* Multivariada

**17) Explique o que é algoritmo de cluster hierárquico?**

O algoritmo de agrupamento hierárquico combina e divide grupos existentes, criando uma estrutura hierárquica que mostra a ordem em que os grupos são divididos ou mesclados.

**18) Explique o que é algoritmo K-mean?**

K significa é um método de particionamento famoso. Os objetos são classificados como pertencentes a um dos K ​​grupos, k escolhidos a priori.

No algoritmo K-mean,

* Os clusters são esféricos: os pontos de dados em um cluster estão centralizados em torno desse cluster
* A variação/distribuição dos clusters é semelhante: cada ponto de dados pertence ao cluster mais próximo

**19) Mencione quais são as principais habilidades exigidas para um analista de dados?**

Um cientista de dados deve ter as seguintes habilidades

* **Conhecimento de banco de dados**
* gerenciamento de banco de dados
* Combinação de dados
* Consultando
* Manipulação de dados
* **Análise Preditiva**
* Estatísticas descritivas básicas
* Modelagem preditiva
* Análise avançada
* **Conhecimento de Big Data**
* Análise de Big Data
* Análise de dados não estruturados
* Aprendizado de máquinas
* **Habilidade de apresentação**
* Visualização de dados
* Apresentação de insights
* Projeto de relatório

**20) Explique o que é filtragem colaborativa?**

A filtragem colaborativa é um algoritmo simples para criar um sistema de recomendação baseado em dados comportamentais do usuário. Os componentes mais importantes da filtragem colaborativa são **usuários- itens- interesse**.

Um bom exemplo de filtragem colaborativa é quando você vê uma declaração como “recomendado para você” em sites de compras online que aparece com base no seu histórico de navegação.

**21) Explique quais são as ferramentas utilizadas em Big Data?**

As ferramentas usadas em Big Data incluem

* Hadoop
* [Colméia](https://career.guru99.com/pt/top-30-hive-interview-questions/)
* Porco
* Calha
* cornaca
* Sqoop

**22) Explique o que é KPI, desenho de experimentos e regra 80/20?**

**KPI**: Significa Key Performance Indicator, é uma métrica que consiste em qualquer combinação de planilhas, relatórios ou gráficos sobre processos de negócios

**Projeto de experimentos**: É o processo inicial usado para dividir seus dados, amostrar e configurar dados para análise estatística

**regras 80/20**: Isso significa que 80% de sua receita vem de 20% de seus clientes

**23) Explique o que é Map Reduce?**

Map-reduce é uma estrutura para processar grandes conjuntos de dados, dividindo-os em subconjuntos, processando cada subconjunto em um servidor diferente e então combinando os resultados obtidos em cada um.

**24) Explique o que é Clustering? Quais são as propriedades dos algoritmos de cluster?**

Clustering é um método de classificação aplicado aos dados. O algoritmo de clustering divide um conjunto de dados em grupos ou clusters naturais.

As propriedades do algoritmo de agrupamento são

* Hierárquico ou plano
* Iterativo
* Duro e macio
* Disjuntivo

**25) Quais são alguns dos métodos estatísticos úteis para analistas de dados?**

Os métodos estatísticos úteis para cientistas de dados são

* método bayesiano
* processo de Markov
* Processos espaciais e de cluster
* Estatísticas de classificação, percentil, detecção de valores discrepantes
* Técnicas de imputação, etc.
* Algoritmo simplex
* Otimização matemática

**26) O que é análise de série temporal?**

A análise de série temporal pode ser feita em dois domínios, domínio da frequência e domínio do tempo. Na análise de série temporal, a saída de um processo específico pode ser prevista analisando os dados anteriores com a ajuda de vários métodos, como suavização exponencial, método de regressão log-linear, etc.

**27) Explique o que é análise de correlograma?**

Uma análise de correlograma é a forma comum de análise espacial em geografia. Consiste em uma série de coeficientes de autocorrelação estimados calculados para uma relação espacial diferente. Pode ser usado para construir um correlograma para dados baseados em distância, quando os dados brutos são expressos como distância em vez de valores em pontos individuais.

**28) O que é uma tabela hash?**

Na computação, uma tabela hash é um mapa de chaves para valores. É um [estrutura de dados](https://career.guru99.com/pt/top-50-data-structure-interview-questions/) usado para implementar uma matriz associativa. Ele usa uma função hash para calcular um índice em um [ordem](https://career.guru99.com/pt/top-50-array-interview-questions-answers/) de slots, dos quais o valor desejado pode ser obtido.

**29) O que são colisões de tabelas hash? Como isso é evitado?**

Uma colisão de tabela hash acontece quando duas chaves diferentes fazem hash com o mesmo valor. Dois dados não podem ser armazenados no mesmo slot do array.

Para evitar a colisão de tabelas hash, existem muitas técnicas, aqui listamos duas

* **Encadeamento separado**:

Ele usa a estrutura de dados para armazenar vários itens que fazem hash no mesmo slot.

* **Endereçamento aberto**:

Ele procura outros slots usando uma segunda função e armazena o item no primeiro slot vazio encontrado

**29) Explique o que é imputação? Liste diferentes tipos de técnicas de imputação?**

Durante a imputação, substituímos os dados ausentes por valores substituídos. Os tipos de técnicas de imputação envolvidas são

* **Imputação Única**
* Imputação de hot-deck: um valor ausente é imputado de um registro semelhante selecionado aleatoriamente com a ajuda de um cartão perfurado
* Imputação de plataforma fria: funciona da mesma forma que a imputação de plataforma quente, mas é mais avançada e seleciona doadores de outros conjuntos de dados
* Imputação de média: envolve a substituição do valor ausente pela média daquela variável para todos os outros casos
* Imputação de regressão: envolve a substituição do valor ausente pelos valores previstos de uma variável com base em outras variáveis
* Regressão estocástica: é igual à imputação de regressão, mas adiciona a variância média da regressão à imputação de regressão
* **Imputação múltipla**
* Ao contrário da imputação única, a imputação múltipla estima os valores várias vezes

**30) Qual método de imputação é mais favorável?**

Embora a imputação única seja amplamente utilizada, ela não reflete a incerteza criada pela falta de dados aleatórios. Portanto, a imputação múltipla é mais favorável do que a imputação única no caso de dados faltantes aleatoriamente.

**31) Explique o que é n-grama?**

**N-grama:**

Um n-grama é uma sequência contígua de n itens de uma determinada sequência de texto ou fala. É um tipo de modelo de linguagem probabilística para prever o próximo item dessa sequência na forma de (n-1).

**32) Explique quais são os critérios para um bom modelo de dados?**

Os critérios para um bom modelo de dados incluem

* Pode ser facilmente consumido
* Grandes alterações de dados em um bom modelo devem ser escalonáveis
* Deve fornecer desempenho previsível
* Um bom modelo pode se adaptar às mudanças nos requisitos