Introdução à Ciência de Dados para Administração

Fundamentos, Técnicas e Aplicações

Washington Santos da Silva

19 de janeiro de 2026

Índice

Prefácio **4**

Introdução **6**

I Parte 1 **7**

Fundamentos da Ciência de Dados **8**

1 Visão Geral de Ciência de Dados **9**
1.1 Introdução . . . 9
1.2 A Economia de Dados: Um Breve Histórico . . . 9
1.3 Big Data e a Explosão do Volume de Dados . . . 10
1.4 O que é Ciência de Dados? . . . 14
1.5 Metodologias e Processos: CRISP-DM . . . 14
1.6 Resumo e Próximos Passos . . . 16

2 Papéis profissionais na área de dados **17**
Analista de Dados . . . 17
Cientista de Dados (Iniciante) . . . 18
Engenheiro de Dados . . . 19
2.1 Áreas de aplicação . . . 19
2.2 Habilidades interpessoais e analíticas . . . 19
2.3 Resumo e Próximos Passos . . . 20

3 A Metodologia CRISP-DM **22**
3.1 O problema do "código antes do plano" . . . 22
3.2 O papel do administrador em projetos de dados . . . 22
3.3 Ferramentas são meios, não fins . . . 22
3.4 CRISP-DM como estrutura orientadora . . . 23
3.5 Fase 1: Compreensão do Negócio (*Business Understanding*) . . . 23
3.5.1 Checklist da Fase 1 . . . 24
3.6 Fase 2: Compreensão dos Dados (*Data Understanding*) . . . 24
3.6.1 Checklist da Fase 2 . . . 24
3.7 Fase 3: Preparação dos Dados (*Data Preparation*) . . . 25
3.7.1 Checklist da Fase 3 . . . 25

3.8 Fase 4: Modelagem (*Modeling*) . . . 25
3.8.1 Checklist da Fase 4 . . . 25
3.9 Fase 5: Avaliação (*Evaluation*) . . . 26
3.9.1 Checklist da Fase 5 . . . 26
3.10 Fase 6: Implementação (*Deployment*) . . . 26
3.10.1 Checklist da Fase 6 . . . 26
3.11 Resumo e Próximos Passos . . . 27

4 Tipos de Análise de Dados em CRISP-DM **28**
4.1 Visão geral dos tipos de análise no CRISP-DM . . . 28
4.2 Análise descritiva . . . 28
4.3 Análise diagnóstica . . . 29
4.4 Análise preditiva . . . 29
4.5 Análise prescritiva . . . 30
4.6 Resumo e Próximos Passos . . . 31

5 Um Estudo de Caso Introdutório **32**
5.1 Estrutura do projeto e organização dos arquivos . . . 32
5.2 O Caso Junglivet Whisky Company . . . 33
5.3 Fase 1: Compreensão do Negócio . . . 34
5.4 Fase 2: Compreensão dos Dados . . . 34
5.4.1 Dicionário de dados . . . 34
5.4.2 Importação dos dados . . . 35
5.4.3 Inspeção inicial dos dados . . . 36
5.5 Fase 3: Preparação dos Dados . . . 38
5.6 Análise exploratória de dados . . . 39
5.6.1 Relação entre fornecedor e qualidade . . . 39
5.6.2 Relação entre mestre responsável e qualidade . . . 43
5.6.3 Relação entre cor e qualidade . . . 45
5.6.4 Conclusão da Análise Exploratória . . . 46
5.7 Próximas Fases . . . 47
5.8 Resumo e Próximos Passos . . . 48

Referências **49**

Prefácio

Este livro resulta da minha experiência em ministrar a disciplina *Introdução à Ciência de Dados* nos anos de 2024 e 2025 para o curso noturno de Bacharelado em Administração do IFMG – Campus Formiga. A intenção é que o livro organize e aprofunde os materiais utilizados em sala de aula, com o objetivo de oferecer aos estudantes uma referência aberta e, espero, com boa qualidade.

A estrutura do livro é inspirada no programa da disciplina e deve funcionar tanto como material de apoio às aulas quanto como um registro dos conteúdos trabalhados ao longo da disciplina. O leitor encontrará referências a materiais complementares e à recursos online, que podem ser utilizadas para estudo autônomo.

Embora pensado prioritariamente para os alunos do curso de Administração do IFMG – Campus Formiga, o texto foi escrito de modo a poder ser útil a estudantes de Administração e áreas afins, especialmente àqueles que estão iniciando o estudo de disciplinas que envolvam análise de dados e estatística.

O leitor a quem este livro se dirige costuma ter algum contato prévio com programação, geralmente por meio de disciplinas introdutórias, mas frequentemente apresenta dificuldades com lógica, matemática básica e conceitos estatísticos. Além disso, é comum a dependência quase exclusiva de planilhas eletrônicas, o que limita a reprodutibilidade, a transparência e a escalabilidade das análises. Este livro parte dessa realidade e não pressupõe fluência prévia em programação ou estatística.

A proposta é deliberada: as dificuldades inerentes à Ciência de Dados não são evitadas, pelo contrário, mas introduzidas de forma gradual e contextualizada. Conceitos, técnicas e ferramentas são apresentados passo a passo, sempre que possível ancorados em exemplos ligados à área de Administração. O objetivo não é ensinar apenas comandos, mas contribuir para a formação de um modo de pensar analítico e crítico, necessários para a proposição de soluções basedas em dados.

A linguagem de programação adotada é *R*, escolhida pelo fato de considerá-la a melhor linguagem de programação para a análise e visualização de dados. O livro também introduz noções básicas de *SQL*, reconhecendo sua importância no trabalho com bases de dados reais. Ferramentas como *RStudio*, *Git* e *GitHub*, bem como o uso de um *terminal*, são apresentadas como parte do ambiente computacional utilizado, voltado à organização, documentação e controle de versões. O sistema *Quarto* é utilizado extensivamente, enfatizando a importância da

reprodutibilidade, **transparência** e a importância da **documentação** dos procedimentos de análise.

– TODO: tópicos de probabilidade, estatística (aprendizagem de máquinda?) que serão abordados.

Minha intenção é que este projeto, em estágio inicial de desenvolvimento, forneça uma contribuição, ainda que modesta, para minimizar um grave problema estrutural: a escassez de referências atualizadas em português na área de Ciência de Dados, a dificuldade de acesso a livros importados por instituições públicas, dada a forte restrição orçamentária à que estas instituições estão submetidas há muitos anos. Nesse contexto, considero que a produção de materiais abertos, de alto nível e adaptados ao contexto local torna-se muito necessária.

Washington Santos da Silva
Professor do IFMG – Campus Formiga

Introdução

TODO.

Parte I

Parte 1

Fundamentos da Ciência de Dados

Esta parte do livro tem como objetivo apresentar os fundamentos históricos, conceituais e metodológicos da chamada economia de dados e da Ciência de Dados. Busca-se fornecer uma base para a compreensão geral do campo e introduzir alguns métodos e princípios que considero relevantes.

Ao longo dos capítulos que compõem esta parte, discutem-se as transformações históricas que levaram à consolidação da economia de dados, os principais papéis profissionais envolvidos na área, metodologias de organização do trabalho e os diferentes tipos de análise empregados.

A Parte I também introduz, de maneira gradual, princípios fundamentais de boas práticas em projetos de análise de dados, como organização de arquivos, reprodutibilidade e documentação. Esses princípios serão retomados e aprofundados ao longo do livro.

O último capítulo desta parte contém um estudo de caso que ilustra a metodologia CRISP-DM e apresenta, de forma introdutória, a organização de um projeto de análise de dados, o uso da linguagem R e uma análise descritiva simples.

Ao final desta parte, espera-se que o leitor desenvolva uma visão geral da Ciência de Dados, por que ela se tornou central na economia de dados contemporânea, como projetos de dados são estruturados e de que forma conceitos, ferramentas e processos se articulam. Essa visão geral servirá como base para as partes e capítulos seguintes, nos quais veremos em mais detalhes as ferramentas computacionais, técnicas e aplicações.

1 Visão Geral de Ciência de Dados

1.1 Introdução

Nas últimas décadas, dados passaram a ocupar um papel central nas decisões econômicas, organizacionais e governamentais. Atividades cotidianas como compras online, interações em redes sociais ou o uso do GPS em smartphones geram continuamente grandes volumes de dados, que alimentam o que hoje se convencionou chamar de *economia de dados*.

Esse fenômeno não surgiu de forma repentina. Ele é resultado de um processo histórico no qual a coleta, o armazenamento e a análise de dados foram progressivamente incorporados às práticas de gestão, pesquisa e tomada de decisão. Compreender essa trajetória é fundamental para entender por que a Ciência de Dados se tornou uma área central no mundo contemporâneo.

1.2 A Economia de Dados: Um Breve Histórico

As raízes da economia de dados remontam ao século XIX, quando jornais norte-americanos passaram a coletar informações sistemáticas sobre seus leitores e a realizar levantamentos para antecipar resultados eleitorais. Já nesse período inicial, dados eram utilizados como instrumento para reduzir incertezas e orientar decisões editoriais e comerciais, ainda que de forma incipiente e pouco padronizada (Harkness, 2021a).

No início do século XX, com a consolidação do marketing como área organizacional, empresas e pesquisadores passaram a estruturar departamentos dedicados ao estudo sistemático do comportamento do consumidor. A coleta de dados deixou de ser episódica e passou a integrar processos contínuos de análise, voltados à compreensão de mercados, preferências e padrões de consumo (Harkness, 2021a).

A partir da década de 1930, a introdução de métodos estatísticos de amostragem, notadamente os trabalhos associados a George Gallup, marcou uma mudança qualitativa importante. Previsões baseadas em dados passaram a apoiar-se em fundamentos estatísticos mais sólidos, substituindo abordagens baseadas em grandes volumes de respostas não controladas por técnicas cientificamente mais rigorosas (Harkness, 2021b).

Entre as décadas de 1950 e 1980, empresas como a Nielsen consolidaram sistemas de observação contínua de hábitos de consumo e audiência. Esse período reforçou uma distinção central na economia de dados: observar comportamentos reais, de forma sistemática, frequentemente

produz informações mais confiáveis do que simplesmente perguntar aos indivíduos sobre suas intenções ou opiniões Harkness (2021c).

Com a digitalização da economia a partir dos anos 1990, o uso de códigos de barras, programas de fidelidade e, posteriormente, plataformas digitais e redes sociais transformou os dados em um ativo estratégico de escala global. A capacidade de coletar, armazenar e analisar grandes volumes de informações passou a redefinir modelos de negócio e estruturas competitivas em diversos setores (Harkness, 2021c).

Mais recentemente, o avanço de modelos de inteligência artificial, como os grandes modelos de linguagem, representa o estágio mais sofisticado dessa trajetória. Esses sistemas dependem fortemente de grandes volumes de dados para seu treinamento e funcionamento, evidenciando que, na economia contemporânea, dados não são apenas um subproduto das atividades organizacionais, mas um recurso central que condiciona inovação, eficiência e poder econômico (Harkness, 2021d).

1.3 Big Data e a Explosão do Volume de Dados

O relatório especial *The Data Deluge*, publicado pela *The Economist* (The Economist, 2010), marcou um ponto de inflexão na forma como economistas, gestores e formuladores de políticas passaram a enxergar os dados na economia contemporânea. O texto parte da constatação de que o mundo entrou em uma era caracterizada por uma produção de dados sem precedentes, impulsionada pela digitalização de processos, pela popularização da internet, pelo uso massivo de dispositivos móveis e pela automação de atividades econômicas e sociais.

O artigo destaca que o crescimento do volume de dados não é apenas quantitativo, mas também qualitativo: empresas e governos passaram a registrar transações, comportamentos e interações em níveis de detalhe antes inimagináveis. Esse fenômeno transforma dados em um ativo econômico estratégico, comparável a recursos tradicionais como capital e trabalho, mas com características próprias, como alta reutilização e forte dependência de capacidades analíticas.

A *The Economist* enfatiza que o verdadeiro desafio não reside na coleta dos dados, mas em sua **organização, análise e interpretação**. Grandes volumes de dados, por si só, não geram conhecimento nem vantagem competitiva. Somente quando combinados com métodos estatísticos, computacionais e analíticos adequados é que os dados podem apoiar decisões gerenciais mais informadas, melhorar previsões e revelar padrões econômicos relevantes.

Por fim, a matéria alerta para os riscos associados à "ilusão dos dados": quanto maior o volume de informações disponíveis, maior também a probabilidade de erros, correlações espúrias e interpretações equivocadas. Nesse contexto, competências analíticas sólidas tornam-se essenciais para administradores, economistas e profissionais das ciências sociais aplicadas, que passam a lidar com um ambiente decisório cada vez mais orientado por dados.

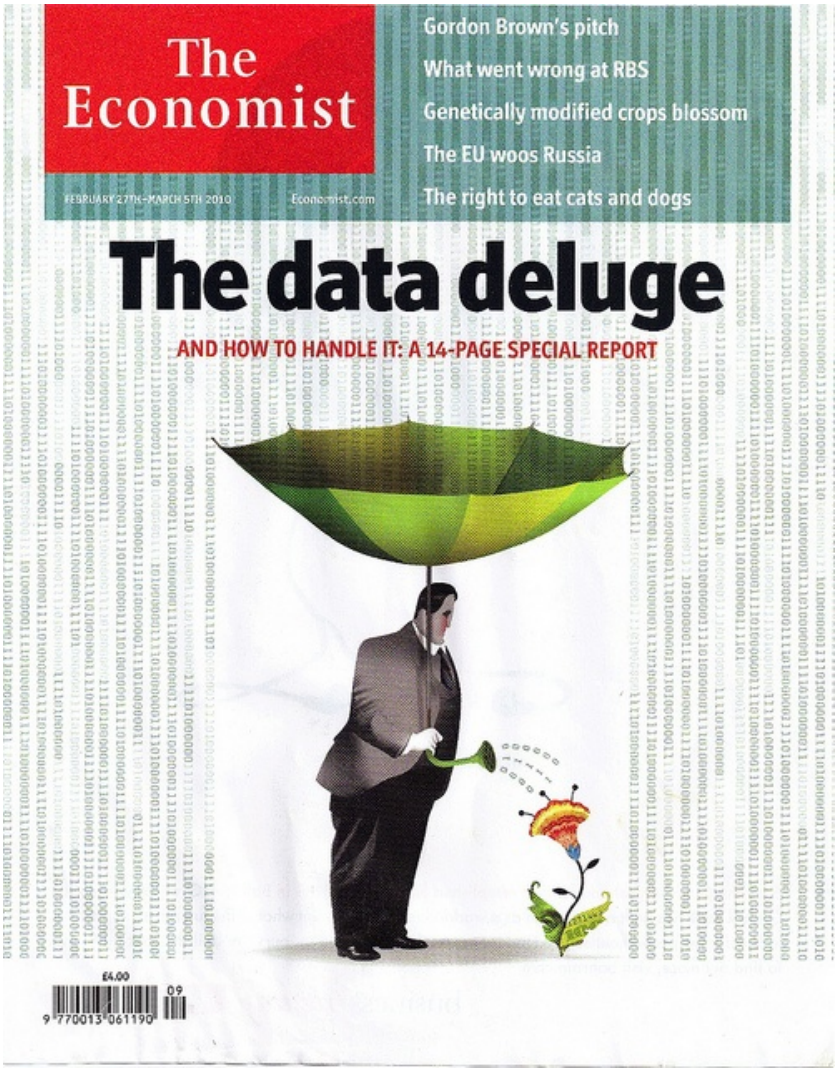

Figura 1.1: Capa da revista The Economist (2010) sobre o dilúvio de dados.

Este crescimento exponencial da produção de dados levou ao surgimento do conceito de *Big Data*, frequentemente caracterizado pelos chamados *cinco Vs*: volume, velocidade, variedade, veracidade e valor. Esses atributos ajudam a compreender não apenas a quantidade de dados gerados, mas também os desafios associados ao seu processamento e uso.

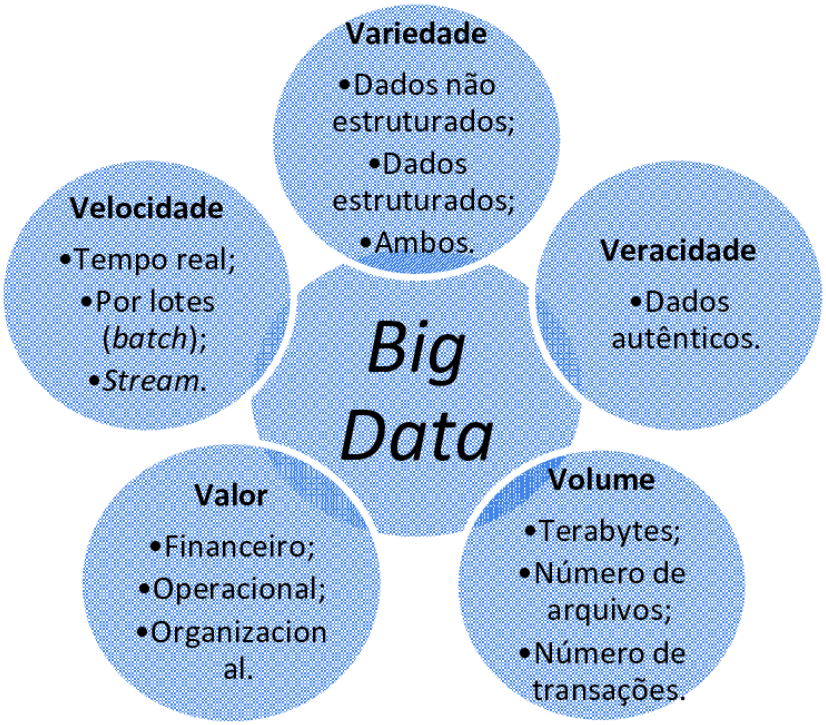

Figura 1.2: Os cinco Vs do Big Data.

Esse aumento no volume e na complexidade dos dados tornou insuficiente o uso exclusivo de ferramentas tradicionais, como planilhas eletrônicas, e criou a necessidade de métodos, linguagens e infraestruturas mais robustas para análise. Não por acaso, publicações de referência na área de negócios passaram a destacar dados como um dos recursos mais valiosos da economia contemporânea.

Outro marco simbólico importante na consolidação da chamada economia de dados foi a capa da revista *The Economist*, publicada em 2017, cujo título era *The World's Most Valuable Resource* ("O recurso mais valioso do mundo") e cujo subtítulo anunciava *Data and the new rules of competition* ("Dados e as novas regras da competição"). A mensagem central não era apenas a comparação entre dados e petróleo, mas a ideia de que os dados haviam se tornado o principal recurso estratégico capaz de redefinir a dinâmica competitiva entre empresas e setores (The Economist, 2017).

Diferentemente de recursos tradicionais, os dados não geram valor de forma automática. Seu valor emerge da capacidade de coletá-los, organizá-los, analisá-los e, sobretudo, utilizá-los de

maneira sistemática na tomada de decisão. Nesse sentido, a capa da *The Economist* aponta para uma mudança mais profunda: empresas competitivas passam a ser aquelas capazes de transformar dados em conhecimento operacional e vantagem estratégica contínua.

Essa mudança implica novas regras de competição. Escala, velocidade de análise, capacidade de experimentação e aprendizado contínuo tornam-se fatores centrais. Organizações que dominam esses elementos conseguem adaptar produtos, processos e estratégias com maior rapidez, enquanto aquelas que tratam dados apenas como subprodutos operacionais tendem a perder relevância.

Ao destacar os dados como o recurso mais valioso da economia contemporânea, a revista reforça a necessidade de métodos, ferramentas e competências voltadas não apenas à análise técnica, mas à integração entre dados, estratégia e decisão. É nesse contexto que a Ciência de Dados se consolida como área essencial para a Administração, indo além do uso de tecnologias específicas e passando a influenciar diretamente a forma como as organizações competem.

Figura 1.3: Capa da revista The Economist (2017) sobre o recurso mais valioso do mundo.

1.4 O que é Ciência de Dados?

Diante desse contexto, surge a Ciência de Dados como uma abordagem estruturada para extrair significado e valor de grandes volumes de dados. Embora o termo possa parecer intimidador, sua ideia central é relativamente simples: usar métodos analíticos e computacionais para transformar dados brutos em informações úteis para a tomada de decisão.

De forma sintética, a Ciência de Dados integra conhecimentos de estatística, computação e domínio do negócio, combinando técnicas dessas áreas para lidar com problemas reais.

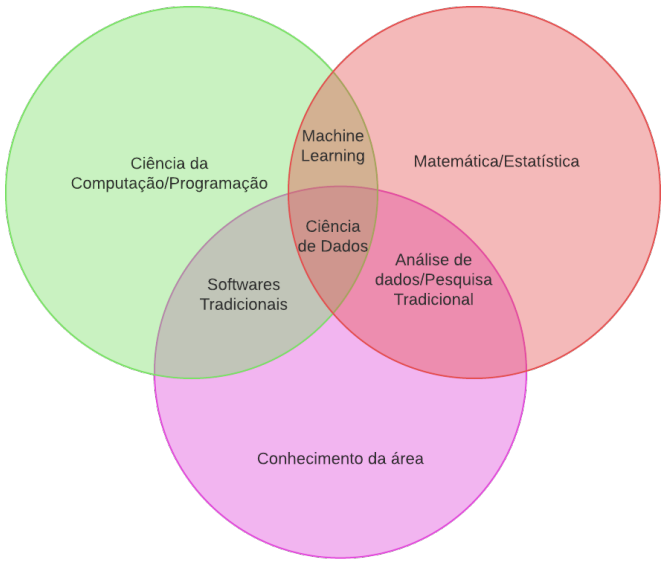

Figura 1.4: A Ciência de Dados como interseção entre estatística, computação e conhecimento do domínio.

Essa característica interdisciplinar explica tanto o potencial da área quanto a diversidade de formações presentes em equipes de dados.

1.5 Metodologias e Processos: CRISP-DM

A prática da Ciência de Dados costuma ser organizada por meio de metodologias de projeto. Uma das mais difundidas é o CRISP-DM (*Cross Industry Standard Process for Data Mining*) (Chapman *et al.*, 2000), que estrutura o trabalho em etapas iterativas, desde o entendimento do problema de negócio até a implantação de soluções baseadas em dados.

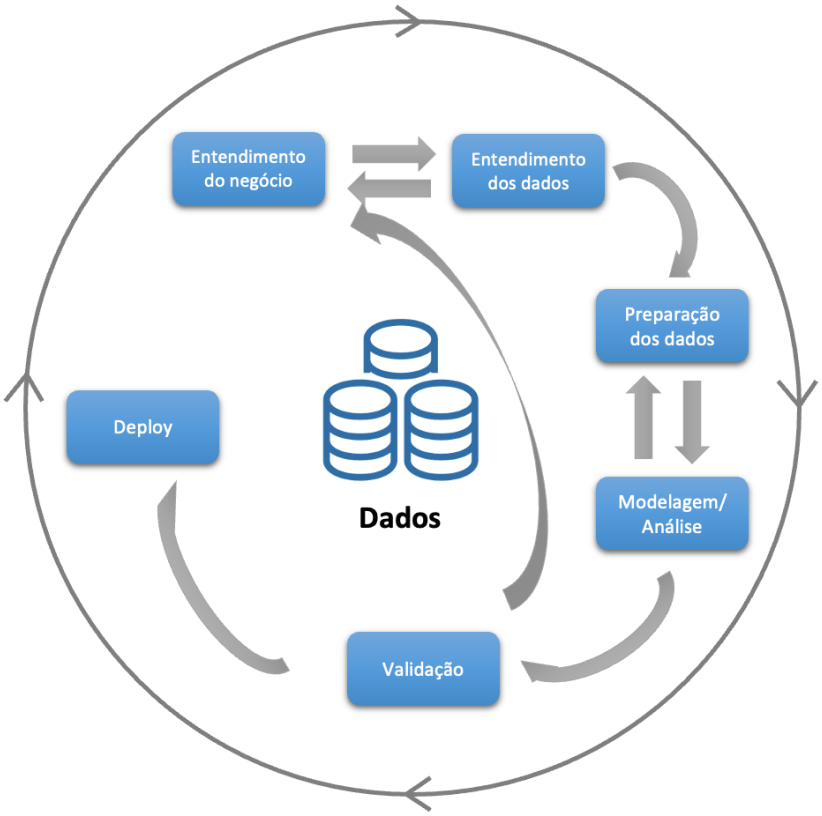

Figura 1.5: Etapas do processo CRISP-DM aplicadas a projetos de Ciência de Dados.

Esse modelo destaca que projetos de dados raramente seguem um caminho linear. É comum que análises retornem a etapas anteriores à medida que novos insights surgem ou que hipóteses iniciais precisem ser revistas.

1.6 Resumo e Próximos Passos

Ao longo deste capítulo, vimos como a chamada *economia de dados* se consolidou historicamente e por que a explosão no volume e na complexidade das informações tornou necessária uma abordagem estruturada, hoje conhecida como Ciência de Dados. Também discutimos que gerar valor a partir de dados exige mais do que tecnologia: requer métodos, organização e capacidade de transformar informação em decisão.

Na prática, esse trabalho não é realizado por um único tipo de profissional. Projetos orientados por dados envolvem funções distintas, com responsabilidades e competências complementares. No próximo capítulo, apresentamos os principais papéis profissionais na área de dados — Analista de Dados, Cientista de Dados e Engenheiro de Dados — e discutimos como eles se relacionam com problemas reais de Administração e áreas afins.

2 Papéis profissionais na área de dados

O crescimento da área de dados deu origem a diferentes papéis profissionais, que variam conforme o foco técnico, o grau de especialização e a posição no fluxo de produção de valor a partir dos dados. Entre os papéis mais comuns estão o **Analista de Dados**, o **Cientista de Dados** e o **Engenheiro de Dados**.

Embora essas funções sejam conceitualmente distintas, na prática, especialmente em organizações de pequeno e médio porte, é comum que um mesmo profissional acumule responsabilidades associadas a mais de um papel. Essa sobreposição é particularmente relevante no contexto brasileiro e deve ser levada em conta ao interpretar descrições formais de cargos.

Nas subseções a seguir, apresentam-se esses papéis de forma sintética, destacando habilidades técnicas, competências analíticas e exemplos típicos de aplicação.

Analista de Dados

O Analista de Dados atua principalmente na exploração, organização e interpretação de dados, com foco em apoiar decisões operacionais e táticas por meio de análises descritivas e diagnósticas.

Habilidades técnicas

- Domínio de ferramentas de visualização de dados, como Power BI ou Tableau, e conhecimento avançado em planilhas eletrônicas, como o MS Excel.
- Proficiência em linguagens de programação voltadas à análise de dados, especialmente R e/ou Python.
- Experiência prática, ainda que em nível introdutório, com bancos de dados e linguagem SQL.

Habilidades analíticas

- Capacidade de realizar análises estatísticas básicas e interpretar grandes volumes de dados para identificar padrões, tendências e anomalias.

Exemplos de aplicações práticas

- Análise de dados em setores como finanças, saúde e turismo, com o objetivo de gerar insights para melhoria de processos, gestão de recursos e atendimento ao cliente.
- Elaboração de análises **descritivas** e **diagnósticas**: a análise descritiva busca compreender *o que* ocorreu, enquanto a análise diagnóstica procura explicar *por que* ocorreu.

Cientista de Dados (Iniciante)

O Cientista de Dados atua de forma mais aprofundada na modelagem e na construção de soluções analíticas, combinando estatística, programação e conhecimento do problema de negócio.

Habilidades técnicas

- Proficiência em linguagens de programação como R e/ou Python, com uso de bibliotecas especializadas (por exemplo, tidyverse, pandas, scikit-learn, tidymodels).
- Conhecimento de SQL e bancos de dados relacionais e não relacionais.
- Familiaridade com ferramentas de versionamento de código (Git) e ambientes de desenvolvimento.

Habilidades analíticas

- Conhecimento sólido de estatística aplicada e aprendizagem de máquina.
- Capacidade de preparar, transformar e organizar conjuntos de dados para análise e modelagem.
- Capacidade de implementar algoritmos básicos de aprendizagem de máquina sob supervisão ou em projetos de escopo limitado.

Exemplos de aplicações práticas

- Desenvolvimento de modelos de classificação e regressão para problemas como previsão, segmentação de clientes e detecção de anomalias.
- Realização de análises **preditivas** e **prescritivas**, utilizando dados históricos para antecipar comportamentos e apoiar recomendações de ação.
- Criação de provas de conceito (POCs) para validação de hipóteses de negócio baseadas em dados.
- Comunicação de resultados técnicos em formato acessível a públicos não técnicos.

Engenheiro de Dados

O Engenheiro de Dados é o profissional responsável pela construção e manutenção da infraestrutura que permite o armazenamento, o processamento e o acesso eficiente aos dados utilizados pelas equipes analíticas.

- Atua no projeto, desenvolvimento e otimização de *pipelines* de dados, *data warehouses* e *data lakes*.
- Seu foco principal é garantir que os dados sejam confiáveis, consistentes, acessíveis e escaláveis, servindo como base para o trabalho de analistas e cientistas de dados.
- Em geral, esse papel envolve menor ênfase em análises estatísticas e maior concentração em aspectos de arquitetura, desempenho e integração de sistemas.

2.1 Áreas de aplicação

As aplicações da Ciência de Dados são amplas e afetam diretamente o cotidiano das organizações. Em finanças, destacam-se análises de risco de crédito, detecção de fraudes e gestão de investimentos. Em marketing, técnicas de segmentação de clientes, análise de sentimentos e monitoramento de mídias sociais são amplamente utilizadas.

Esses exemplos ilustram como dados podem ser usados para compreender o passado, explicar causas e antecipar cenários futuros, correspondendo às análises descritivas, diagnósticas e preditivas.

2.2 Habilidades interpessoais e analíticas

Além de competências técnicas, profissionais de dados precisam desenvolver habilidades interpessoais e analíticas que permitam transformar resultados quantitativos em decisões organizacionais concretas.

Essas habilidades são essenciais para conectar análises de dados a problemas reais de negócio e para comunicar resultados de forma eficaz a diferentes públicos.

- **Pensamento analítico**: Abordar problemas de forma estruturada, formular perguntas relevantes, selecionar informações apropriadas e buscar soluções baseadas em evidências.
- **Conhecimento do negócio**: Compreender os objetivos e estrutura da organização, o contexto do mercado e a forma como análises de dados se relacionam com metas estratégicas e operacionais.

- **Comunicação oral e escrita**: Capacidade de explicar resultados técnicos a pessoas sem formação técnica, utilizando linguagem clara, exemplos práticos e evitando jargões desnecessários.
- **Pensamento crítico**: Questionar suposições implícitas, avaliar a qualidade e as limitações dos dados disponíveis e considerar interpretações alternativas antes de chegar a conclusões.
- **Contar histórias com dados**: Organizar resultados e indicadores de modo a construir uma narrativa coerente, capaz de destacar os principais insights e apoiar processos de tomada de decisão.
- **Trabalho em equipe**: Colaborar com profissionais de diferentes áreas, compreender necessidades diversas e integrar perspectivas técnicas e organizacionais.
- **Gerenciamento de projetos**: Planejar etapas de trabalho, definir prioridades, estabelecer prazos realistas e comunicar o andamento das atividades às partes interessadas.
- **Adaptabilidade**: Lidar com mudanças de requisitos, ferramentas e tecnologias em um campo caracterizado por rápida evolução.
- **Curiosidade intelectual**: Demonstrar interesse contínuo em aprender, formular novas perguntas e explorar dados de forma sistemática e responsável.

> Dica
>
> Na prática profissional em Ciência de Dados, a capacidade de **estruturar problemas**, **compreender o contexto do negócio** e **comunicar resultados** de forma clara tende a ser mais determinante do que o domínio isolado de ferramentas. As habilidades técnicas adquirem valor quando integradas a essas competências centrais.

2.3 Resumo e Próximos Passos

Neste capítulo, discutimos os principais papéis profissionais na área de dados, bem como as competências técnicas, analíticas e interpessoais associadas a cada função. Vimos que Analistas de Dados, Cientistas de Dados e Engenheiros de Dados contribuem de forma complementar para transformar dados em informação útil e apoiar decisões organizacionais.

Na prática, porém, a atuação desses profissionais não ocorre de maneira isolada ou desordenada. Projetos de Ciência de Dados exigem coordenação, alinhamento estratégico e definição clara de objetivos, etapas e entregáveis. É nesse ponto que metodologias de projeto tornam-se essenciais.

No próximo capítulo, apresentamos a metodologia CRISP-DM, amplamente adotada em projetos de Ciência de Dados, como uma estrutura orientadora para organizar o trabalho das equipes,

alinhar análises aos objetivos do negócio e reduzir riscos associados a abordagens improvisadas ou excessivamente centradas em ferramentas.

3 A Metodologia CRISP-DM

Este capítulo apresenta a metodologia CRISP-DM (*Cross-Industry Standard Process for Data Mining*) (Chapman *et al.*, 2000), amplamente utilizada para estruturar projetos de Ciência de Dados. O objetivo é oferecer uma visão clara e operacional do método, com ênfase no papel do administrador na condução, coordenação e alinhamento estratégico desses projetos.

3.1 O problema do "código antes do plano"

Projetos de dados frequentemente falham não por limitações técnicas, mas pela ausência de uma metodologia clara. É comum que equipes iniciem o desenvolvimento de modelos ou scripts antes mesmo de compreender o problema de negócio a ser resolvido.

Esse tipo de abordagem tende a gerar desalinhamento entre soluções técnicas e necessidades organizacionais, ciclos recorrentes de retrabalho, desperdício de tempo e recursos, além de dificuldades para escalar projetos piloto para soluções corporativas.

3.2 O papel do administrador em projetos de dados

Nesse contexto, o administrador desempenha papel central. Cabe a ele assegurar que projetos de dados tenham início a partir de objetivos de negócio claramente definidos, com critérios mensuráveis de sucesso estabelecidos antes da implementação técnica.

Além disso, o administrador atua como elo entre equipes técnicas e as partes interessadas no projeto, articulando expectativas, restrições e prioridades organizacionais, e garantindo que os resultados produzidos sejam relevantes para a tomada de decisão.

3.3 Ferramentas são meios, não fins

Ferramentas como R, Python, SQL, Quarto e Git são essenciais para projetos de Ciência de Dados, mas não constituem um fim em si mesmas. Sem uma metodologia orientadora, mesmo códigos tecnicamente sofisticados podem resolver o problema errado.

O CRISP-DM fornece contexto e direção para o uso dessas ferramentas, enfatizando que a modelagem é apenas uma das etapas de um processo mais amplo, orientado por objetivos organizacionais.

3.4 CRISP-DM como estrutura orientadora

O CRISP-DM organiza projetos de dados de forma sistemática e iterativa. Diferentemente de abordagens centradas em ferramentas, o processo se inicia na compreensão do negócio e se encerra com a avaliação e implementação dos resultados no contexto organizacional.

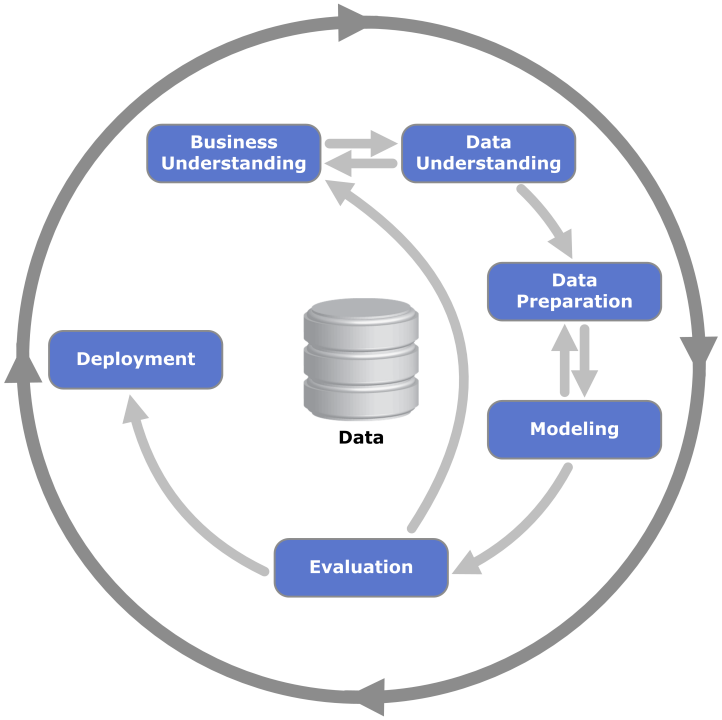

Figura 3.1: Fases da metodologia CRISP-DM e sua natureza iterativa.

As fases funcionam como pontos de verificação, permitindo avaliar o progresso, alinhar expectativas e decidir se o projeto deve avançar, ser ajustado ou interrompido.

3.5 Fase 1: Compreensão do Negócio (*Business Understanding*)

A primeira fase do CRISP-DM tem como foco alinhar a análise de dados aos objetivos empresariais. Antes de qualquer coleta, modelagem ou análise, é necessário compreender claramente qual problema se deseja resolver, por que ele é relevante e para quem.

Nessa etapa, objetivos de negócio são definidos, a situação organizacional é avaliada, as partes interessadas são identificadas, metas analíticas são estabelecidas e um plano de projeto é elaborado de forma realista.

3.5.1 Checklist da Fase 1

#	Tarefa	Resultados esperados
1.1	Determinar objetivos de negócio	Contexto e critérios de sucesso
1.2	Avaliar situação	Recursos, restrições e riscos
1.3	Alinhar partes interessadas	Expectativas e responsabilidades
1.4	Determinar objetivos de análise	Metas analíticas e critérios técnicos
1.5	Produzir plano do projeto	Cronograma e responsabilidades

3.6 Fase 2: Compreensão dos Dados (*Data Understanding*)

A compreensão dos dados envolve conhecer profundamente as informações que serão utilizadas no projeto, independentemente de elas já existirem ou precisarem ser produzidas.

Em muitos projetos, os dados não estão imediatamente disponíveis, sendo necessário coletá-los a partir de fontes externas, bases públicas, sistemas de terceiros ou por meio de instrumentos como surveys, experimentos ou registros operacionais. Nessa fase, também são avaliadas viabilidade, custos, limitações éticas e qualidade dos dados.

As atividades incluem coleta ou produção inicial, descrição, exploração e verificação de qualidade, fornecendo base empírica sólida para as etapas seguintes.

3.6.1 Checklist da Fase 2

#	Tarefa	Resultados esperados
2.1	Coletar ou produzir dados	Relatório ou desenho de dados
2.2	Descrever dados	Relatório de descrição
2.3	Explorar dados	Relatório de exploração
2.4	Verificar qualidade dos dados	Relatório de qualidade

3.7 Fase 3: Preparação dos Dados (*Data Preparation*)

A preparação dos dados é frequentemente a fase mais trabalhosa do processo. Seu objetivo é transformar dados brutos ou heterogêneos em um conjunto adequado para análise e modelagem.

Inclui seleção de registros e variáveis, tratamento de inconsistências, criação de atributos derivados, integração de múltiplas fontes e padronização de formatos, com documentação das decisões adotadas.

3.7.1 Checklist da Fase 3

#	Tarefa	Resultados esperados
3.1	Selecionar dados	Critérios de inclusão e exclusão
3.2	Limpar dados	Relatório de limpeza
3.3	Construir dados	Atributos derivados
3.4	Integrar dados	Conjuntos integrados
3.5	Formatar dados	Conjunto final documentado

3.8 Fase 4: Modelagem (*Modeling*)

Na fase de modelagem, técnicas estatísticas e de aprendagem de máquisa são aplicadas aosdados preparados. São escolhidos algoritmos compatíveis com os objetivos do projeto, definidos procedimentos de validação e ajustados parâmetros dos modelos.

Decisões tomadas nessa fase dependem fortemente das escolhas realizadas nas etapas anteriores e, com frequência, exigem iteração com fases anteriores.

3.8.1 Checklist da Fase 4

#	Tarefa	Resultados esperados
4.1	Selecionar técnicas	Técnica e pressupostos
4.2	Gerar design de teste	Estratégia de validação
4.3	Construir modelo	Modelos e parâmetros
4.4	Avaliar modelo	Métricas e ajustes

3.9 Fase 5: Avaliação (*Evaluation*)

A avaliação verifica se os resultados obtidos atendem aos objetivos de negócio definidos inicialmente. Não se trata apenas de desempenho técnico, mas de utilidade prática, viabilidade de implementação e impacto organizacional.

Nesta etapa, decide-se se o modelo está pronto para uso, se ajustes adicionais são necessários ou se novas análises devem ser conduzidas.

3.9.1 Checklist da Fase 5

#	Tarefa	Resultados esperados
5.1	Avaliar resultados	Comparação com critérios de sucesso
5.2	Revisar processo	Lições aprendidas
5.3	Determinar próximos passos	Decisões e ações

3.10 Fase 6: Implementação (*Deployment*)

A implementação transforma análises e modelos em instrumentos efetivos de decisão. Envolve integração com sistemas organizacionais, definição de métricas de monitoramento, manutenção e comunicação dos resultados aos gestores.

Essa fase também inclui a documentação do projeto e a consolidação de aprendizados, contribuindo para a maturidade analítica da organização.

3.10.1 Checklist da Fase 6

#	Tarefa	Resultados esperados
6.1	Planejar implantação	Plano de implantação
6.2	Planejar monitoramento e manutenção	Plano de monitoramento
6.3	Produzir relatório final	Relatório e apresentação
6.4	Revisar projeto	Documentação de experiência

3.11 Resumo e Próximos Passos

Ao longo deste capítulo, discutimos a metodologia CRISP-DM como uma estrutura orientadora para projetos de Ciência de Dados, destacando a importância de iniciar análises a partir de objetivos de negócio claramente definidos e de organizar o trabalho de forma iterativa e sistemática. Vimos que projetos bem sucedidos dependem menos de ferramentas isoladas e mais da articulação entre problemas, dados, métodos e decisões.

No entanto, estruturar um projeto é apenas parte do desafio. Em cada fase do CRISP-DM, diferentes tipos de análise podem ser empregados, cada um respondendo a perguntas específicas e oferecendo níveis distintos de apoio à tomada de decisão. Nem toda análise tem o mesmo propósito, nem exige o mesmo grau de sofisticação técnica.

No próximo capítulo, apresentamos os principais tipos de análise de dados — descritiva, diagnóstica, preditiva e prescritiva — e discutimos como eles se relacionam com as etapas do CRISP-DM e com as necessidades práticas da Administração.

4 Tipos de Análise de Dados em CRISP-DM

Ao longo do processo CRISP-DM, diferentes tipos de análise podem ser empregados, variando em complexidade técnica, grau de formalização e valor estratégico. Cada tipo de análise responde a uma pergunta distinta de negócio e está associado a decisões tomadas em fases específicas do processo analítico.

As análises descritiva, diagnóstica, preditiva e prescritiva não são excludentes. Pelo contrário, costumam ser adotadas de forma progressiva e complementar, à medida que a organização desenvolve suas capacidades analíticas e passa a integrar dados de maneira mais sistemática em seus processos decisórios.

4.1 Visão geral dos tipos de análise no CRISP-DM

O termo *analytics* refere-se a um espectro contínuo de técnicas analíticas que evoluem desde a organização e compreensão do passado até a recomendação de ações orientadas para o futuro.

No contexto do CRISP-DM, esses tipos de análise não correspondem a fases isoladas, mas atravessam o processo como um todo. Análises descritivas e diagnósticas são predominantes nas fases de Compreensão do Negócio e dos Dados, enquanto análises preditivas e prescritivas tornam-se centrais nas fases de Modelagem, Avaliação e Implementação.

A maturidade analítica de uma organização pode ser avaliada pelo equilíbrio e pela profundidade com que esses quatro tipos de análise são utilizados de forma integrada, e não apenas pela adoção de técnicas mais sofisticadas.

4.2 Análise descritiva

A análise descritiva representa o nível inicial do uso sistemático de dados. Seu objetivo é organizar, resumir e comunicar informações históricas de modo a tornar o passado compreensível para gestores e tomadores de decisão.

O que aconteceu?

- **Objetivo**: Descrever e sintetizar dados históricos, identificando padrões, tendências e comportamentos recorrentes.
- **Técnicas**: Estatísticas descritivas, tabelas, visualizações e dashboards.
- **Complexidade**:

No contexto da Administração, a análise descritiva está fortemente associada a relatórios gerenciais e sistemas de acompanhamento operacional. Ela fornece a base informacional sobre a qual análises mais avançadas podem ser construídas.

Exemplos incluem relatórios de vendas por canal em períodos promocionais, painéis de monitoramento de indicadores operacionais e análises de distribuição de clientes por região ou perfil de consumo.

4.3 Análise diagnóstica

A análise diagnóstica aprofunda a análise descritiva ao buscar explicações para os padrões observados. Enquanto a análise descritiva responde ao *o que* aconteceu, a diagnóstica procura compreender *por que* esses resultados ocorreram.

Por que aconteceu?

- **Objetivo**: Investigar causas, relações e fatores associados aos resultados observados.
- **Técnicas**: Análise de correlação, segmentação, comparações entre grupos, *drill-down* e análise de fatores.
- **Complexidade**:

Esse tipo de análise é fundamental para apoiar decisões corretivas e ajustes de estratégia, como identificar fatores associados à queda nas vendas após um reajuste de preços ou compreender as causas do aumento do turnover em determinadas unidades organizacionais.

No CRISP-DM, análises diagnósticas são recorrentes nas fases de Compreensão do Negócio e dos Dados, orientando decisões sobre coleta, preparação e seleção de variáveis relevantes.

4.4 Análise preditiva

A análise preditiva utiliza dados históricos para estimar comportamentos futuros ou resultados prováveis. Nesse nível, modelos estatísticos e de aprendizagem de máquina passam a

desempenhar papel central.

O que provavelmente acontecerá?

- **Objetivo**: Estimar tendências futuras e resultados prováveis com base em padrões observados nos dados.
- **Técnicas**: Modelos de regressão, séries temporais e algoritmos de classificação.
- **Complexidade**:

Exemplos típicos em Administração incluem previsão de demanda para produtos sazonais, modelos de propensão à inadimplência em instituições financeiras e estimativas de giro de estoque para apoiar decisões de compras e logística.

No CRISP-DM, a análise preditiva está fortemente associada à fase de Modelagem, mas depende diretamente das decisões tomadas nas etapas anteriores de compreensão e preparação dos dados.

4.5 Análise prescritiva

A análise prescritiva representa o nível mais avançado da jornada analítica. Seu foco não está apenas em prever resultados, mas em recomendar ações que maximizem objetivos organizacionais, considerando restrições, custos e trade-offs.

O que devemos fazer?

- **Objetivo**: Recomendar decisões e ações otimizadas com base em análises descritivas, diagnósticas e preditivas.
- **Técnicas**: Otimização, simulação, algoritmos de decisão e sistemas de recomendação.
- **Complexidade**:

No contexto empresarial, a análise prescritiva é utilizada em problemas como otimização do mix de produtos por loja, recomendação personalizada em plataformas de e-commerce e definição automática de rotas logísticas em ambientes urbanos complexos.

Esse tipo de análise exige não apenas maturidade técnica, mas também processos decisórios bem estruturados, integração com sistemas operacionais e clareza quanto aos objetivos estratégicos da organização.

4.6 Resumo e Próximos Passos

Neste capítulo, discutimos quatro tipos de análise de dados — descritiva, diagnóstica, preditiva e prescritiva — destacando que cada uma responde a perguntas distintas e oferece diferentes formas de apoio à tomada de decisão. Também vimos que esses tipos de análise atravessam as etapas de um projeto e se articulam com a metodologia CRISP-DM.

No próximo capítulo, apresentamos um estudo de caso introdutório que integra, em um único exemplo, etapas iniciais do CRISP-DM, organização de projeto e análise exploratória usando a linguagem R. O objetivo é oferecer uma visão da "floresta": um panorama do fluxo completo de trabalho que será aprofundado, com mais detalhes e novas ferramentas, ao longo dos capítulos seguintes.

5 Um Estudo de Caso Introdutório

Este capítulo apresenta um estudo de caso introdutório com o objetivo de integrar, em um único exemplo, conceitos, ferramentas e práticas que serão aprofundados ao longo dos capítulos seguintes. A intenção, certamente, não é esgotar os tópicos abordados, mas oferecer uma visão geral do processo de análise de dados no contexto da metodologia CRISP-DM. Este estudo de caso é baseado em um exemplo fictício adaptado do livro de Jung (2024).

Como ler este capítulo: o objetivo aqui é oferecer uma visão geral do fluxo de trabalho de um projeto de Ciência de Dados, do problema de negócio até uma análise exploratória inicial. Você não precisa dominar a linguagem R, RStudio, Quarto ou Git e GitHub neste momento.

Se esta é sua primeira leitura da Parte 1, concentre-se em entender o que está sendo feito e por que cada decisão é tomada. Na Parte 2, as ferramentas serão apresentadas sistematicamente; depois disso, recomenda-se retornar a este capítulo e reler os trechos de código com maior segurança.

5.1 Estrutura do projeto e organização dos arquivos

Antes de iniciar qualquer análise, é fundamental definir uma estrutura mínima de projeto. A organização adequada dos arquivos facilita a reprodutibilidade, a colaboração e a manutenção do trabalho ao longo do tempo, além de tornar mais claras as etapas do processo analítico.

Ao longo da disciplina, os projetos são organizados de forma padronizada, separando dados brutos, dados limpos (ou processados), scripts e relatórios quarto. Essa mesma lógica será adotada neste livro, servindo como referência para os exemplos e códigos apresentados nos próximos capítulos.

Uma estrutura básica de projeto pode ser representada da seguinte forma:

```
projeto_junglivet/
  dados/                      # arquivos de dados utilizados no projeto
     brutos/                  # dados originais, sem modificações
     limpos/                  # dados após limpeza e transformações
  scripts/                    # scripts R organizados por etapa da análise
     01-importacao.R          # leitura e inspeção inicial dos dados
     02-preparacao.R          # limpeza, transformação e criação de variáveis
```

```
    03-analise_exploratoria.R  # gráficos e estatísticas descritivas
relatorios/                    # relatórios reprodutíveis do projeto
    01_relatorio.qmd           # relatório principal em Quarto
README.md                      # descrição do projeto e instruções gerais
.gitignore                     # arquivos e pastas ignorados pelo Git
projeto_junglivet.Rproj        # arquivo do projeto RStudio
```

Essa separação explícita evita que dados originais sejam corrompidos, torna as etapas do processo analítico mais transparentes e facilita a verificação e a reprodução das análises realizadas. Além disso, diversas instituições e empresas com alta maturidade analítica possuem uma estrutura padronizada para projetos envolvendo código e dados.

> Organização do Projeto
>
> A *organização das pastas e arquivos do projeto* não é um detalhe técnico, mas parte central da reprodutibilidade e do trabalho em equipe. Um projeto bem estruturado permite que análises sejam compreendidas, verificadas e reproduzidas por outras pessoas — ou pelo próprio autor em um momento posterior.

Na prática, recomendamos **fortemente** criar inicialmente um repositório vazio no GitHub, cloná-lo localmente e, na pasta clonada do repositório, criar um projeto RStudio. A partir desse projeto, os arquivos e diretórios podem ser organizados de forma estruturada, com controle de versão via Git e GitHub.

Nota sobre reprodutibilidade: ao longo deste estudo de caso, assumimos que o trabalho é conduzido em um projeto RStudio, com os arquivos organizados em pastas e com o histórico de alterações registrado por controle de versão (Git) e sincronizado em um repositório no GitHub. O relatório é produzido em Quarto, permitindo que resultados, figuras e tabelas sejam reproduzidos a partir do código e dos dados.

5.2 O Caso Junglivet Whisky Company

Neste estudo de caso, aplicamos as três primeiras fases da metodologia CRISP-DM:

1. **Compreensão do Negócio** (*Business Understanding*)
2. **Compreensão dos Dados** (*Data Understanding*)
3. **Preparação dos Dados** (*Data Preparation*)

Você acaba de ser contratado como analista de dados na *Junglivet Whisky Company*. A empresa enfrenta reclamações recorrentes sobre a qualidade do whisky produzido, e a direção busca identificar possíveis causas para o problema.

Os dados fornecidos correspondem ao registro da linha de produção das últimas duas semanas.

5.3 Fase 1: Compreensão do Negócio

O primeiro passo do CRISP-DM consiste em entender claramente o problema de negócio, antes de qualquer decisão técnica.

- **Problema de negócio**: queda na qualidade do whisky produzido.
- **Objetivo**: identificar fatores associados à redução da qualidade.
- **Critério de sucesso**: evidenciar fatores operacionais que influenciam negativamente o indicador de qualidade.

Nesta fase, ainda não buscamos respostas nos dados, mas sim formular as perguntas corretas.

5.4 Fase 2: Compreensão dos Dados

Nesta etapa, analisamos os dados disponíveis, sua estrutura e limitações, antes de qualquer transformação ou modelagem.

5.4.1 Dicionário de dados

A documentação das variáveis é uma prática fundamental em projetos de análise de dados. Ela garante interpretação consistente e reduz ambiguidades ao longo do processo analítico.

Um dicionário de dados é uma documentação estruturada que descreve o significado, formato, uso e relacionamentos de cada variável em um conjunto de dados.

Ele funciona como um guia essencial para compreender corretamente as informações disponíveis, garantindo que todos os usuários interpretem os dados de maneira consistente.

O arquivo de dados fornecido contém as seguintes colunas (ou variáveis):

- **DAY**: dia da produção.
- **MONTH**: mês da produção.
- **MANUFACTURER**: mestre responsável pela produção.
- **PRODUCT**: tipo de produto.
- **SHIFT**: turno de produção.
- **COLOR**: indicador de cor (0 a 1).
- **MALTING**: fornecedor do malte.
- **TASTING**: indicador de qualidade (0 a 1000).

5.4.2 Importação dos dados

Antes de iniciar a análise, é necessário garantir que os pacotes R utilizados no projeto estejam instalados e carregados. Neste livro, utilizaremos o pacote **pacman** para facilitar esse processo, pois ele permite instalar e carregar pacotes de forma automática quando necessário.

A instalação de pacotes deve ser feita apenas uma vez em cada ambiente. Após isso, basta carregá-los normalmente. O uso do **pacman** reduz problemas comuns enfrentados por iniciantes, como erros relacionados a pacotes ausentes.

```
# Verifica se o pacote 'pacman' está instalado.
# Caso não esteja, realiza a instalação a partir do CRAN.
if (!requireNamespace("pacman", quietly = TRUE)) {
  install.packages("pacman")
}

# Carrega o pacote 'pacman' na sessão atual
library(pacman)

# A função p_load():
# - instala automaticamente pacotes ausentes
# - carrega todos os pacotes listados
pacman::p_load(
  here,      # define caminhos relativos ao diretório raiz do projeto
  readr,     # leitura eficiente de arquivos CSV
  dplyr,     # manipulação e transformação de dados
  tidyr,     # função drop_na()
  ggplot2    # visualização de dados
)
```

O processo de importação de dados é um passo fundamental em qualquer análise. Neste caso, utilizamos duas ferramentas importantes:

- O pacote `here` permite definir caminhos relativos ao diretório raiz do projeto, o que torna o código mais portátil e facilita o compartilhamento. Independentemente de onde o projeto esteja armazenado em diferentes computadores, o pacote `here` encontrará automaticamente os arquivos a partir da raiz do projeto.
- O pacote `readr`, parte do tidyverse, oferece funções otimizadas para leitura de arquivos, como a `read_csv()`, que é mais rápida que a função base do R e oferece tratamento mais consistente dos tipos de dados. Além disso, ela converte automaticamente strings vazias para NA, indica o tipo de cada coluna importada e preserva os nomes das variáveis originais.

```r
# Define o caminho relativo do arquivo de dados
# a partir da raiz do projeto.
caminho <- here::here("dados/brutos/productionlog_sample.csv")

# Importa o arquivo com a função read_csv
dados_destilaria <- readr::read_csv(caminho)
```

A partir desse ponto, os dados estão disponíveis na ambiente de R e podem ser explorados, verificados e preparados para as etapas seguintes da análise.

5.4.3 Inspeção inicial dos dados

Após importar os dados, é essencial verificar sua estrutura para entender o que temos disponível. A função `glimpse()` do pacote `dplyr` nos oferece uma visão concisa e informativa sobre:

- Quais variáveis (colunas) estão presentes no conjunto de dados.
- Qual o tipo ou classe de cada variável.
- Os primeiros valores de cada variável.
- O número total de observações (linhas).

```r
# fornece visão geral da estrutura dos dados
dplyr::glimpse(dados_destilaria)
```

```
Rows: 21
Columns: 8
$ DAY          <dbl> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, NA, 6, 6, 7, 7, 8, 8, 9, 9,~
$ MONTH        <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, NA, 4, 4, 4, 4, 4, 4, 4, 4,~
$ MANUFACTURER <chr> "Leonard", "Carlson", "Leonard", "Carlson", "Leonard", "C~
$ PRODUCT      <chr> "Junglivet", "Junglivet Premium", "Junglivet", "Junglivet~
$ SHIFT        <chr> "Morning", "Evening", "Morning", "Evening", "Morning", "E~
$ COLOR        <dbl> 0.27, 0.27, 0.28, 0.32, 0.32, 0.28, 0.29, 0.29, 0.33, 0.2~
$ MALTING      <chr> "Inhouse", "Burns Best Ltd.", "Inhouse", "Inhouse", "Matr~
$ TASTING      <dbl> 895, 879, 938, 900, 917, 900, 934, 951, 852, 850, NA, 991~
```

Antes de avançar para estatísticas resumidas, é útil observar algumas linhas do conjunto de dados. Isso ajuda a localizar problemas comuns, como valores ausentes, categorias inesperadas e registros inconsistentes.

```r
# Exibe as primeiras linhas do conjunto de dados
dplyr::slice_head(dados_destilaria, n = 12)
```

```
# A tibble: 12 x 8
     DAY MONTH MANUFACTURER PRODUCT           SHIFT   COLOR MALTING      TASTING
   <dbl> <dbl> <chr>        <chr>             <chr>   <dbl> <chr>          <dbl>
 1     1     4 Leonard      Junglivet         Morning  0.27 Inhouse          895
 2     1     4 Carlson      Junglivet Premium Evening  0.27 Burns Best ~     879
 3     2     4 Leonard      Junglivet         Morning  0.28 Inhouse          938
 4     2     4 Carlson      Junglivet         Evening  0.32 Inhouse          900
 5     3     4 Leonard      Junglivet         Morning  0.32 Matro Ltd.       917
 6     3     4 Carlson      Junglivet         Evening  0.28 Inhouse          900
 7     4     4 Leonard      Junglivet         Morning  0.29 Inhouse          934
 8     4     4 Gumble       Junglivet Premium Evening  0.29 Matro Ltd.       951
 9     5     4 Leonard      Junglivet         Morning  0.33 Matro Ltd.       852
10     5     4 Carlson      Junglivet         Evening  0.27 Inhouse          850
11    NA    NA <NA>         <NA>              <NA>    NA    <NA>             NA
12     6     4 Carlson      Junglivet         Morning  0.3  Inhouse          991
```

Além disso, estatísticas descritivas básicas ajudam a identificar valores ausentes, escalas e possíveis inconsistências.

A função `summary()` resume as principais estatísticas das variáveis e indica automaticamente a presença de valores ausentes (NA)

```
# fornece estatísticas descritivas dos dados
summary(dados_destilaria)
```

```
      DAY            MONTH   MANUFACTURER         PRODUCT
 Min.   : 1.0   Min.   :4   Length:21          Length:21
 1st Qu.: 3.0   1st Qu.:4   Class :character   Class :character
 Median : 5.5   Median :4   Mode  :character   Mode  :character
 Mean   : 5.5   Mean   :4
 3rd Qu.: 8.0   3rd Qu.:4
 Max.   :10.0   Max.   :4
 NA's   :1      NA's   :1
    SHIFT               COLOR          MALTING             TASTING
 Length:21          Min.   :0.2600   Length:21          Min.   :822.0
 Class :character   1st Qu.:0.2775   Class :character   1st Qu.:875.0
 Mode  :character   Median :0.3000   Mode  :character   Median :925.5
                    Mean   :0.2955                      Mean   :918.5
                    3rd Qu.:0.3100                      3rd Qu.:957.8
                    Max.   :0.3500                      Max.   :999.0
                    NA's   :1                           NA's   :1
```

A saída da função mostra que há pelo menos um valor faltante (NA). Com a inspeção de algumas linhas, é possível localizar rapidamente esse tipo de problema e decidir como tratá-lo na preparação dos dados (por exemplo, remover a observação ou investigar sua origem).

5.5 Fase 3: Preparação dos Dados

Nesta fase, preparamos os dados para análise, renomeando variáveis, convertendo cada variável para um tipo ou classe de dados adequado, tratando valores ausentes, removendo colunas irrelevantes e garantindo que temos dados de qualidade para trabalhar.

O código a seguir executa as seguintes operações para limpar os dados:

1. **Remove** a coluna `MONTH` que é desnecessária para a análise.
2. **Renomeia** todas as colunas para nomes mais descritivos em português, facilitando a interpretação.
3. **Converte** as variáveis para seus tipos/classes de dados apropriados: `numeric` para valores quantitativos (dia, cor, indicador_qualidade) e `factor` para variáveis categóricas (fabricante, tipo_produto, turno, fornecedor_malte).
4. Remove linhas com valores ausentes para garantir a integridade dos dados nas análises subsequentes.

Vamos utilizar o operador pipe (`%>%`) do tidyverse para encadear as operações de limpeza e transformação de dados de forma mais legível. Cada operação recebe o resultado da anterior e aplica uma nova transformação.

Observe como organizamos o código com identação consistente e comentários explicativos para cada operação. Esta é uma boa prática que torna o código mais legível e facilita sua manutenção.

```
# Pipeline para criar uma nova data frame
# contendo os dados limpos.

# define o objeto que armazenará os dados limpos
# criado a partir de dados_destilaria
dados_destilaria_limpos <- dados_destilaria %>%
  # Remove a variável MONTH
  select(-MONTH) %>%
  # Renomeia as variáveis para nomes mais descritivos
  rename(
    dia = DAY,
    mestre_responsavel = MANUFACTURER,
```

```
    tipo_produto = PRODUCT,
    turno = SHIFT,
    cor = COLOR,
    fornecedor_malte = MALTING,
    indicador_qualidade = TASTING
  ) %>%
  # Converte explicitamente os tipos das variáveis
  mutate(
    dia = as.numeric(dia),
    mestre_responsavel = as.factor(mestre_responsavel),
    tipo_produto = as.factor(tipo_produto),
    turno = as.factor(turno),
    cor = as.numeric(cor),
    fornecedor_malte = as.factor(fornecedor_malte),
    indicador_qualidade = as.numeric(indicador_qualidade)
  ) %>%
  # Remove observações com valores ausentes
  drop_na()
```

5.6 Análise exploratória de dados

A Análise Exploratória de Dados (AED) é uma abordagem fundamental que nos permite investigar e compreender as características principais de um conjunto de dados.

Utilizamos técnicas visuais e estatísticas para:

- Identificar padrões, tendências e relações entre variáveis.
- Detectar valores atípicos (outliers) e anomalias.
- Verificar hipóteses preliminares sobre possíveis causas do problema.
- Orientar análises mais detalhadas e modelagens futuras.

Com os dados devidamente preparados, vamos explorar graficamente
relações entre algumas variáveis e o indicador de qualidade do whisky para identificar potenciais fatores que explicam os problemas enfrentados pela destilaria.

5.6.1 Relação entre fornecedor e qualidade

O boxplot ou diagrama de caixa é uma ferramenta útil para visualizar a distribuição de variáveis numéricas agrupadas por categorias.

Neste gráfico:

- A linha horizontal dentro da caixa representa a **mediana** (percentil 50).
- Os limites inferior e superior da caixa representam o **primeiro quartil** (percentil 25) e o **terceiro quartil** (percentil 75), respectivamente.
- As "hastes" (whiskers) se estendem até 1,5 vezes o intervalo interquartil (IQR) que é a diferença entre o terceiro e o primeiro quartil.
- Pontos individuais além das hastes representam **outliers** (valores atípicos)

Essa visualização permite comparar as distribuições do indicador de qualidade entre os diferentes fornecedores de malte, oferecendo indícios iniciais sobre possíveis relações sistemáticas entre o insumo utilizado e a qualidade final do produto.

```
# Boxplot comparativa da qualidade por fornecedor de malte
ggplot(dados_destilaria_limpos, aes(x = fornecedor_malte, y = indicador_qualidade)) +
  # Cria boxplots para representar a distribuição dos dados
  geom_boxplot() +
  # Aplica um tema minimalista para melhor visualização
  theme_minimal() +
  # Define títulos e rótulos dos eixos
  labs(
    title = "Qualidade do Whisky por Fornecedor de Malte",
    x = "Fornecedor",
    y = "Pontuação de Qualidade"
  )
```

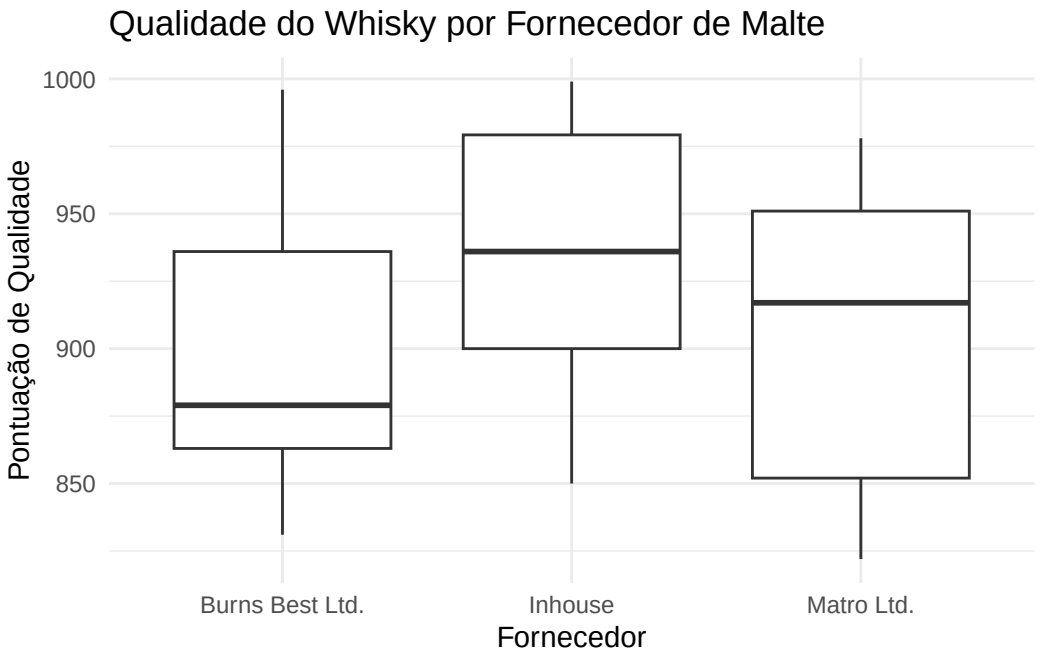

Figura 5.1: Boxplot comparativo de qualidade do whisky por fornecedor de malte.

O boxplot sugere diferenças sistemáticas na distribuição do indicador de qualidade entre as diferentes origens do malte. No conjunto observado, o malte produzido internamente pela própria destilaria (Inhouse) está associado à maior mediana do indicador de qualidade, o que sugere melhor desempenho típico quando o insumo não é adquirido de fornecedores externos.

O fornecedor *Matro Ltd.* aparece com mediana intermediária, mas com maior dispersão (IQR mais alto), sugerindo maior variabilidade no padrão de qualidade associado a esse insumo. Já o fornecedor *Burns Best Ltd.* apresenta a menor mediana do indicador de qualidade, com a maior parte das observações concentrada em níveis relativamente mais baixos.

É importante destacar que essas evidências são exploratórias: o gráfico não prova causalidade e, sem testes estatísticos, não é possível afirmar se as diferenças observadas são estatisticamente significativas. Ainda assim, o fornecedor de malte emerge como um candidato natural para investigação nas próximas fases do CRISP-DM.

Para complementar a leitura visual do boxplot, calculamos medidas numéricas por fornecedor. Especificamente, reportamos:

- a mediana do indicador de qualidade (medida de posição),
- o IQR (intervalo interquartil, medida de dispersão coerente com o boxplot), e
- o tamanho da amostra (n) em cada grupo.

Essa tabela ajuda a confirmar, com números, as diferenças sugeridas pelo gráfico e também permite avaliar o quanto os resultados podem ser afetados por tamanhos de amostra pequenos.

> Por que utilizamos mediana e IQR?
>
> Na análise exploratória, optamos por resumir os dados utilizando a **mediana** e o **intervalo interquartil (IQR)**, em vez da média e do desvio-padrão. Essa escolha não é arbitrária. Distribuições assimétricas ou com valores extremos (outliers) são comuns em dados reais de processos produtivos. Nessas situações, a média pode ser fortemente influenciada por poucos valores atípicos, enquanto a mediana representa melhor o comportamento típico dos dados. De forma análoga, o IQR fornece uma medida de dispersão mais robusta, coerente com a informação visual apresentada no boxplot.
> Esses conceitos — assimetria, medidas de posição e medidas de dispersão — serão estudados de forma sistemática na parte do livro dedicada à análise estatística. Neste momento, o objetivo é apenas compreender como essas medidas ajudam a interpretar padrões observados nos dados.

```
dados_destilaria_limpos %>%
  group_by(fornecedor_malte) %>%
  summarise(
    qualidade_mediana = median(indicador_qualidade),
    iqr_qualidade = IQR(indicador_qualidade),
    n = n()
  ) %>%
  arrange(desc(qualidade_mediana))
```

```
# A tibble: 3 x 4
  fornecedor_malte qualidade_mediana iqr_qualidade     n
  <fct>                        <dbl>         <dbl> <int>
1 Inhouse                        936          79.2    10
2 Matro Ltd.                     917          99       5
3 Burns Best Ltd.                879          73       5
```

As estatísticas descritivas reforçam a leitura do boxplot. A origem interna do malte (*Inhouse*) está associada à maior mediana do indicador de qualidade, seguida pelos fornecedores externos *Matro Ltd.* e *Burns Best Ltd.*

Quanto à dispersão, o IQR sugere maior variabilidade para o fornecedor *Matro Ltd.* (IQR = 99) do que para o malte produzido internamente pela destilaria (Inhouse, IQR = 79,2) e para o fornecedor *Burns Best Ltd.* (IQR = 73). Em termos operacionais, esse resultado pode indicar menor consistência na qualidade do insumo fornecido por *Matro Ltd.*

Por fim, é necessária cautela na interpretação destes resultados, pois os tamanhos de amostra são pequenos (n = 5 para dois fornecedores) e diferentes em relação ao malte próprio da destilaria. Em etapas posteriores, métodos estatísticos formais poderão avaliar se as diferenças observadas persistem e se são compatíveis com variações aleatórias.

5.6.2 Relação entre mestre responsável e qualidade

Os mestres responsáveis pela produção podem influenciar significativamente a qualidade do produto final devido às suas técnicas, experiência e atenção aos detalhes. Novamente, um boxplot comparativo é uma visualização útil para ilustrar a relação entre o indicador numérico de qualidade e o mestre responsável.

```
# Boxplot comparativo entre indicador de qualidade e mestre destilador
ggplot(dados_destilaria_limpos, aes(x = mestre_responsavel, y = indicador_qualidade)) +
  geom_boxplot() +
  theme_minimal() +
  labs(
    title = "Qualidade do Whisky por Mestre Responsável",
    x = "Mestre Responsável",
    y = "Indicador de Qualidade"
  )
```

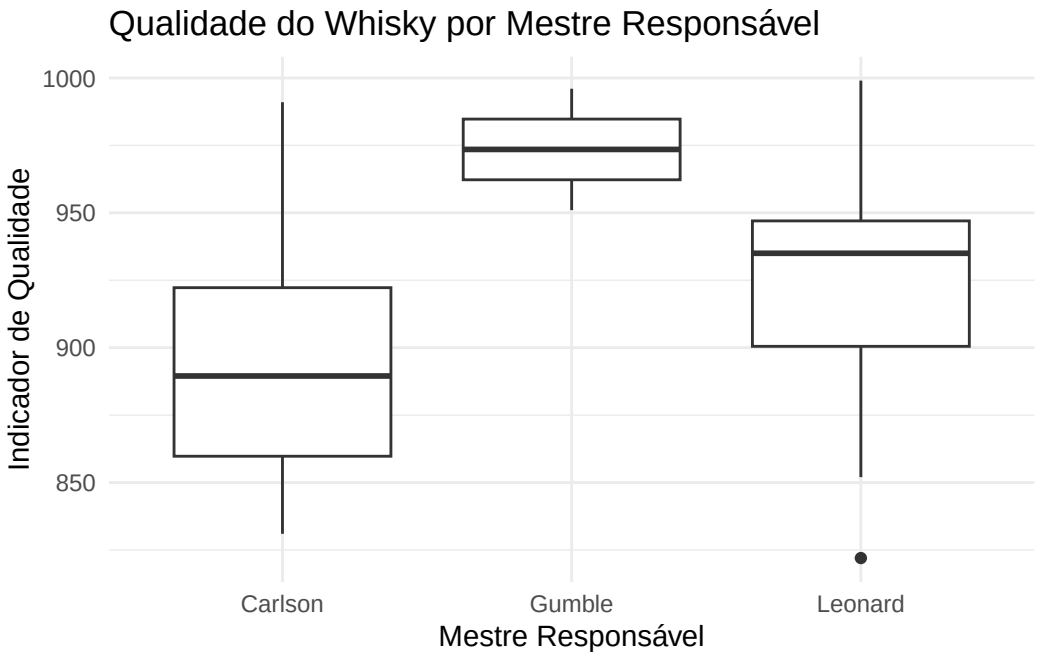

Figura 5.2: Boxplot comparativo entre qualidade do whisky e mestre destilador.

O boxplot comparativo sugere diferenças na distribuição do indicador de qualidade entre os mestres responsáveis pela produção. No conjunto observado, os whiskies produzidos pelo mestre Gumble apresentam a maior mediana do indicador de qualidade e menor dispersão aparente em relação aos demais.

Os whiskies produzidos pelo mestre Leonard apresentam mediana intermediária do indicador de qualidade, enquanto aqueles produzidos pelo mestre Carlson exibem a menor mediana e maior dispersão visual, indicando maior variabilidade nos resultados observados.

Para complementar a análise gráfica, calculamos estatísticas descritivas por mestre responsável, incluindo uma medida de posição (mediana), uma medida de dispersão (IQR) e o tamanho da amostra (n) em cada grupo.

```
dados_destilaria_limpos %>%
  group_by(mestre_responsavel) %>%
  summarise(
    qualidade_mediana = median(indicador_qualidade),
    iqr_qualidade = IQR(indicador_qualidade),
    n = n()
  ) %>%
  arrange(desc(qualidade_mediana))
```

```
# A tibble: 3 x 4
  mestre_responsavel qualidade_mediana iqr_qualidade     n
  <fct>                           <dbl>         <dbl> <int>
1 Gumble                           974.          22.5     2
2 Leonard                          935           46.5    10
3 Carlson                          890.          62.5     8
```

As estatísticas descritivas corroboram a hierarquia observada no boxplot: os whiskies produzidos pelo mestre Gumble apresentam a maior mediana do indicador de qualidade, seguidos por Leonard e Carlson.

Em termos de dispersão, o IQR sugere menor variabilidade para Gumble e maior variabilidade para Carlson. Contudo, essa comparação deve ser interpretada com extrema cautela, dado o número muito reduzido de observações associadas ao mestre Gumble (n = 2).

Esse desbalanceamento amostral limita os possíveis insights e ilustra uma situação comum em análises reais: padrões visuais ou estatísticos podem ser fortemente influenciados pela quantidade de dados disponíveis, reforçando a importância de análises mais robustas nas fases posteriores do CRISP-DM.

5.6.3 Relação entre cor e qualidade

O gráfico de dispersão, ou *scatter plot*, é uma ferramenta adequada para explorar a relação entre duas variáveis numéricas. Ao analisar a relação entre o indicador de cor do whisky e seu indicador de qualidade:

- cada ponto representa uma amostra produzida;
- o eixo horizontal mostra o valor do indicador de cor;
- o eixo vertical indica a pontuação de qualidade;
- a curva de suavização (*LOESS*) auxilia na visualização do padrão médio dos dados, sem assumir uma relação linear pré-definida.

Essa visualização permite investigar se determinados **intervalos de valores** do indicador de cor tendem a estar associados a níveis mais elevados ou mais baixos de qualidade, oferecendo indícios iniciais sobre possíveis padrões no processo produtivo.

```
ggplot(dados_destilaria_limpos, aes(x = cor, y = indicador_qualidade)) +
  geom_point() +
  geom_smooth(method = "loess", se = FALSE) +
  theme_minimal() +
  labs(
    title = "Relação entre Cor e Qualidade do Whisky",
    x = "Indicador de Cor",
    y = "Indicador de Qualidade"
  )
```

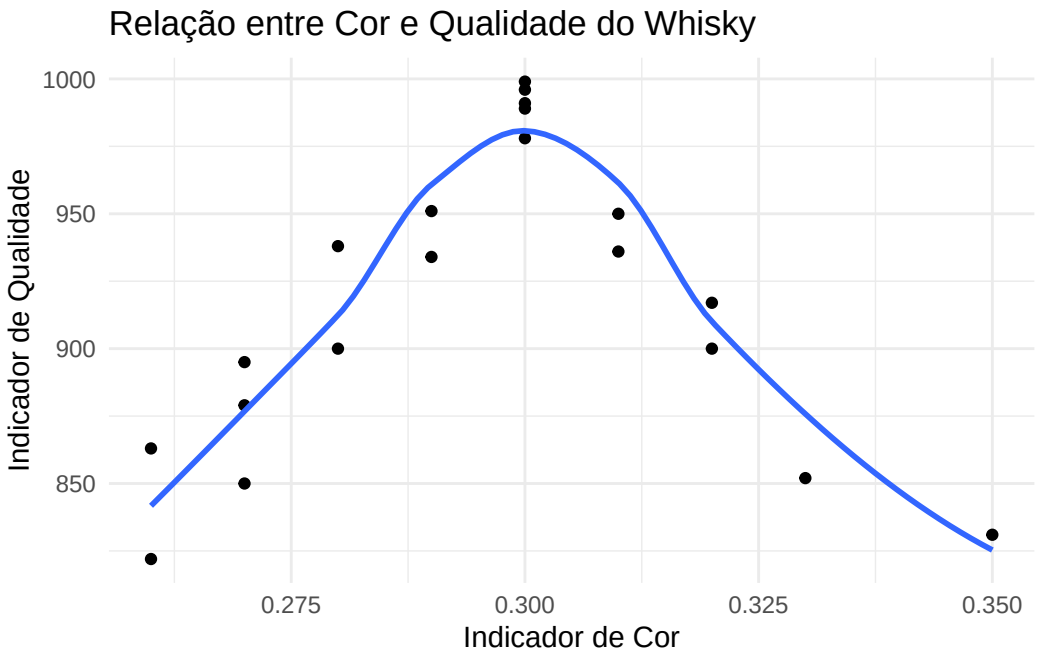

Figura 5.3: Gráfico de dispersão entre qualidade e cor do whisky.

O gráfico de dispersão sugere uma relação não linear entre o indicador de cor e a qualidade do whisky. A curva de suavização indica que valores de cor em torno de 0,3 estão associados, em média, a níveis mais elevados de qualidade no conjunto observado.

Por outro lado, valores muito baixos ou muito elevados do indicador de cor tendem a estar associados a desempenhos inferiores no indicador de qualidade. Esse padrão visual sugere que a cor pode refletir condições do processo produtivo que afetam a qualidade final do produto.

Do ponto de vista operacional, esse resultado é particularmente relevante, pois o indicador de cor pode ser monitorado durante a produção, antes das etapas finais de avaliação sensorial. No entanto, essa evidência deve ser interpretada como exploratória, servindo como base para a formulação de hipóteses que poderão ser avaliadas de forma mais rigorosa nas fases posteriores do CRISP-DM.

5.6.4 Conclusão da Análise Exploratória

A análise exploratória dos dados da linha de produção da Junglivet Whisky Company indica que a variação observada na qualidade do produto não ocorre de forma puramente aleatória. Pelo contrário, os resultados sugerem associações entre o indicador de qualidade e diferentes dimensões do processo produtivo.

Em particular, a origem do malte — distinguindo insumos produzidos internamente daqueles adquiridos de fornecedores externos — e o indicador de cor do whisky emergem como variáveis potencialmente relevantes, apresentando padrões consistentes de associação com o desempenho em qualidade. Além disso, a análise por mestre responsável sugere diferenças na distribuição do indicador de qualidade, embora essas evidências sejam fortemente limitadas pelo desbalanceamento dos tamanhos de amostra entre os mestres.

É fundamental destacar que esses achados têm caráter estritamente exploratório. As associações observadas não constituem provas de causalidade, nem permitem inferências conclusivas sobre o efeito isolado de cada fator. Ainda assim, elas cumprem o papel central da Análise Exploratória de Dados ao fornecer hipóteses analíticas bem fundamentadas e ao orientar as próximas fases do processo CRISP-DM, nas quais métodos estatísticos mais formais poderão ser empregados para avaliar a robustez e a relevância prática desses padrões.

5.7 Próximas Fases

A Análise Exploratória de Dados cumpriu o papel de identificar padrões, associações e possíveis fontes de variação na qualidade do whisky. A partir dessas evidências iniciais, o projeto avança para as próximas fases da metodologia CRISP-DM, nas quais essas hipóteses poderão ser avaliadas de forma mais sistemática.

- **Modelagem**: nesta fase, serão aplicados métodos estatísticos apropriados para avaliar se as diferenças observadas na análise exploratória — associadas à origem do malte, ao mestre responsável e a intervalos do indicador de cor — são consistentes e não atribuíveis apenas a variações aleatórias. Exemplos de abordagens incluem testes de comparação entre grupos e modelos simples de explicação da variabilidade observada.
- **Avaliação**: os resultados obtidos na etapa de modelagem serão interpretados à luz dos objetivos de negócio da empresa. Além da significância estatística, serão considerados a magnitude dos efeitos estimados, a robustez dos resultados e suas implicações práticas para o processo produtivo.
- **Implantação**: com base nas evidências avaliadas, poderão ser propostas ações operacionais concretas, como a revisão de fornecedores externos, ajustes em etapas do processo produtivo e a definição de indicadores de monitoramento contínuo da qualidade.

> Modelagem
>
> No contexto da Ciência de Dados e CRISP-DM, modelagem **não se confunde** com aprendizagem de máquina. Modelar pode significar aplicar métodos estatísticos tradicionais para comparar grupos, explicar relações e apoiar decisões, sem a necessidade de construir modelos preditivos complexos.

Checklist rápido deste estudo de caso (CRISP-DM)

Ao final deste capítulo, verifique se você consegue responder "sim" às questões abaixo:

1. Eu consigo descrever o problema de negócio, o objetivo e o critério de sucesso do projeto?
2. Eu identifiquei as variáveis do arquivo e compreendi seus significados (dicionário de dados)?
3. Eu inspecionei a estrutura do conjunto de dados e reconheci problemas potenciais (por exemplo, valores ausentes e colunas irrelevantes)?
4. Eu gerei uma versão limpa e documentada do conjunto de dados?
5. Eu combinei gráficos e estatísticas descritivas para levantar hipóteses iniciais (por exemplo, fornecedor, mestre responsável e cor)?
6. Eu consigo explicar por que os achados aqui são exploratórios e não representam prova de causalidade?

5.8 Resumo e Próximos Passos

Ao longo dos capítulos da Parte 1 do livro, vimos como conceitos, papéis, metodologias e tipos de análise se articulam em um projeto de Ciência de Dados simplificado. Essa visão geral fornece o contexto necessário para compreender não apenas *o que* deve ser feito, mas, sobretudo, *por que* determinadas decisões são tomadas ao longo do processo analítico.

Na Parte 2, o foco se desloca para as ferramentas que tornam possível a execução prática desses projetos, aprofundando os princípios e as boas práticas introduzidas até aqui.

Referências

CHAPMAN, P. *et al.* **CRISP-DM 1.0: Step-by-step data mining guide**. [s.l.] CRISP-DM Consortium, 2000.

HARKNESS, T. The history of the data economy: Part I: The birth of customer insight. **Significance**, v. 18, n. 2, p. 12–15, a2021.

_____. The history of the data economy: Part II: Analytics arrives. **Significance**, v. 18, n. 4, p. 16–19, b2021.

_____. The history of the data economy: Part III: The new kings and queens of data. **Significance**, v. 18, n. 5, p. 16–19, c2021.

_____. The history of the data economy: Part IV: The future. **Significance**, v. 18, n. 6, p. 12–15, d2021.

JUNG, D. **The Modern Business Data Analyst: A Case Study Introduction into Business Data Analytics with CRISP-DM and R**. Cham, Switzerland: Springer Nature Switzerland, 2024.

THE ECONOMIST. **The Data Deluge**, 27 fev. 2010. Disponível em: <https://www.economist.com/weeklyedition/2010-02-27>. Acesso em: 21 jan. 2026

_____. **The World's Most Valuable Resource**, 2017. Disponível em: <https://www.economist.com/weeklyedition/2017-05-06>. Acesso em: 20 jan. 2026