Trabalho Final RDI

Alice Moreira Marques-22306521
Eduardo Galvão de Aquino Cavalheiro-22303433
Eduardo Sousa Hirle de Freitas-22303593
Gabriel Almeida Poppi Durante-22302431
Isadora Almeida Poppi Durante-22302370
João Victor Ferreira Marques-22303180

2024-12-09

Contents

0.1 Seção 1: Web Scraping

0.2 Justificativa da Escolha do Mercado Livre

O Mercado Livre foi escolhido como alvo de raspagem de dados devido à sua relevância e popularidade na América Latina, especialmente no Brasil, onde se destaca como uma das maiores plataformas de e-commerce. A plataforma oferece uma grande variedade de produtos e informações, como preços, descrições e links de produtos, o que torna a raspagem de dados uma excelente fonte de informações para análises de mercado, comparações de preços e insights sobre o comportamento dos consumidores. Além disso, a coleta de dados de e-commerce pode ser útil para pesquisas sobre tendências de consumo e estratégias de marketing.

0.3 Ferramentas Utilizadas

0.3.1 Linguagem de Programação

A linguagem de programação escolhida para realizar a raspagem de dados foi o Python, uma das linguagens mais populares no campo da ciência de dados. Python é amplamente utilizado para automação de processos e análise de grandes volumes de dados, com várias bibliotecas dedicadas ao processamento e coleta de informações de sites.

0.3.2 Bibliotecas Utilizadas

- requests: Utilizada para enviar requisições HTTP e obter o conteúdo das páginas da web.

- BeautifulSoup (bs4): Responsável pelo parsing do HTML das páginas, permitindo a extração de informações relevantes.
- time: Usada para introduzir atrasos entre as requisições, evitando sobrecarga nos servidores e prevenindo o bloqueio por parte do site.
- pandas: Utilizada para organizar e manipular os dados coletados em um formato estruturado (DataFrame), facilitando a análise posterior.

0.4 Passo a Passo do Processo de Coleta

A coleta de dados seguiu um processo estruturado, que pode ser descrito nas etapas a seguir:

1.Definição do Produto e URL: O produto alvo da coleta foi o iPhone. Com isso, foi criada a URL de busca para os produtos relacionados a este item no Mercado Livre. 2. Envio de Requisições HTTP: Utilizou-se a biblioteca `requests` para enviar requisições HTTP para as páginas do Mercado Livre, coletando o conteúdo HTML das páginas de produtos. 3. Parse do Conteúdo HTML: Usando o BeautifulSoup, o conteúdo HTML retornado foi analisado para identificar e extrair as informações relevantes, como descrições dos produtos, preços e links. 4. Armazenamento dos Dados: As informações extraídas foram armazenadas em uma lista de dicionários, com cada item contendo o nome do produto, preço e link para o anúncio. 5. Criação do DataFrame: Usou-se a biblioteca pandas para organizar os dados coletados em um DataFrame, que possibilita a análise de forma estruturada e facilita a exportação para formatos como CSV ou Excel. 6. Implementação de Delay: Para evitar o bloqueio da aplicação devido a requisições excessivas, foi adicionado um intervalo de 2 segundos entre as requisições utilizando a função `time.sleep(2)`.

0.5 Comentários sobre a Qualidade e Características dos Dados Coletados

Os dados coletados incluem informações como descrições de produtos, preços e links para os anúncios no Mercado Livre. A qualidade dos dados depende de vários fatores:

- Descrição do Produto: As descrições podem variar de acordo com o vendedor e a categoria do produto. Isso pode gerar inconsistências nas informações.
- Preço: O preço é geralmente consistente, mas pode variar devido a promoções ou descontos temporários, o que pode afetar a comparabilidade entre os preços.
- Link para o Produto: O link fornecido direciona para a página do produto, permitindo uma verificação mais detalhada das informações.

Esses dados são bastante úteis para análise de mercado e comparação de preços, mas devem ser analisados com cautela devido às possíveis variações nas descrições dos produtos.

0.6 Possíveis Limitações ou Desafios Enfrentados

Durante o processo de raspagem, alguns desafios e limitações podem ser identificados:

- Mudanças na Estrutura do Site: O Mercado Livre pode alterar sua estrutura HTML, o que pode causar erros no scraping. Essas mudanças exigem ajustes no código.
- Limitações de Acesso: O site pode bloquear o acesso após várias requisições seguidas, o que pode ser contornado com o uso de proxies ou aumentando o tempo entre as requisições.
- Qualidade dos Dados: A descrição dos produtos pode ser inconsistente, com variações nas palavras-chave ou erros de digitação, afetando a uniformidade dos dados.
- Proteção Contra Scraping: O Mercado Livre pode usar métodos para impedir scraping excessivo, como CAPTCHA ou bloqueio de IPs, dificultando a coleta de dados em larga escala.

0.7 Aspectos Éticos da Coleta de Dados

A raspagem de dados deve ser realizada de maneira ética para garantir que não haja violação da privacidade dos usuários ou danos à plataforma. Alguns princípios importantes incluem:

- Transparência: Deve-se ser transparente sobre os objetivos da coleta de dados. A coleta deve ser feita de forma que respeite as expectativas dos usuários em relação à privacidade.
- Respeito ao Trabalho de Outros: O scraping não deve sobrecarregar os servidores do site ou prejudicar a experiência dos usuários. Isso pode ser mitigado ao usar delays nas requisições.
- Uso Responsável dos Dados: Os dados coletados devem ser utilizados para fins legítimos, como análise de mercado, e não para práticas comerciais desleais ou prejudiciais.

0.8 Aspectos Legais da Coleta de Dados

A coleta de dados deve ser feita de acordo com a legislação vigente, especialmente no que diz respeito à LGPD (Lei Geral de Proteção de Dados) e aos Termos de Uso da plataforma. Alguns pontos importantes incluem:

- Termos de Serviço: O Mercado Livre pode ter restrições sobre scraping em seus termos de uso, sendo fundamental verificar essas diretrizes antes de realizar a coleta.
- Proteção de Dados Pessoais: A coleta de dados deve respeitar a privacidade dos usuários e não violar as leis de proteção de dados pessoais.
- Propriedade Intelectual: O conteúdo coletado, como descrições e imagens de produtos, pode ser protegido por direitos autorais, e seu uso indevido pode resultar em infrações legais.

0.9 Boas Práticas e Recomendações

Para realizar a raspagem de dados de maneira ética e legal, recomenda-se:

- Limitar a Taxa de Requisições: Evitar sobrecarregar o servidor do Mercado Livre implementando atrasos nas requisições.
- Respeitar o `robots.txt`: Verificar as diretrizes de acesso do site para garantir que a raspagem não viole as regras definidas pela plataforma.
- Obter Permissão: Quando possível, solicitar permissão explícita para realizar a coleta de dados, especialmente em larga escala.

0.10 Demonstração da análize dos produtos via tabela

```
##     Produto                 Preço (R$)  Link
##  0  Apple iPhone 14 P...         4.97   https://www.mercadolivre.co...
##  1  Apple iPhone 14 (...         4.099  https://www.mercadolivre.co...
##  2  Apple iPhone 15 (...         4.554  https://www.mercadolivre.co...
##  3  Apple iPhone 15 (...       455      https://www.mercadolivre.co...
##  4  Apple iPhone 15 (...         5.337  https://www.mercadolivre.co...
##  5  Apple iPhone 15 (...         4.803  https://www.mercadolivre.co...
##  6  Apple iPhone 16 (...         5.337  https://www.mercadolivre.co...
##  7  Apple iPhone 16 (...       533      https://www.mercadolivre.co...
##  8  Apple iPhone 13 (...         5.199  https://www.mercadolivre.co...
##  9  Apple iPhone 14 (...         4.679  https://www.mercadolivre.co...
## 10  Apple iPhone 14 (...         5.199  https://www.mercadolivre.co...
## 11  Apple iPhone 13 (...       519      https://www.mercadolivre.co...
## 12  iPhone 12 Pro 512...         5.199  https://www.mercadolivre.co...
## 13  Apple iPhone 14 P...         4.679  https://www.mercadolivre.co...
## 14  Apple iPhone 14 (...         5.199  https://www.mercadolivre.co...
## 15  Apple iPhone 16 P...       519      https://www.mercadolivre.co...
## 16  iPhone XR 128 GB ...         5.199  https://www.mercadolivre.co...
## 17  iPhone 15 Pro Max...         4.679  https://www.mercadolivre.co...
## 18  Apple iPhone 16 (...         5.199  https://www.mercadolivre.co...
## 19  Apple iPhone 12 (...       519      https://www.mercadolivre.co...
```

0.11 Seção 2: ETL e EDA com o Dataset "State of Data 2023"

O presente trabalho analisa o conjunto de dados "State of Data 2023", que apresenta uma visão abrangente do mercado de trabalho na área de dados no Brasil. A pesquisa foi conduzida pela Data Hackers, a maior comunidade de dados do Brasil, em parceria com a Bain & Company, uma consultoria global renomada. Este estudo tem como objetivos principais:

1. Preparar os dados para análise através de um pipeline de ETL (Extract, Transform, Load).
2. Realizar uma Análise Exploratória de Dados (EDA) para identificar padrões, tendências e práticas relevantes no setor de dados.

O resultado esperado é uma síntese detalhada sobre o panorama da área de dados no Brasil, servindo como base para tomada de decisão por empresas e profissionais.

0.12 Descrição do Dataset

0.12.1 Sobre a Pesquisa

- Fonte: Kaggle - Data Hackers (State of Data Brazil 2023).
- Período de Coleta: 16 de outubro a 6 de dezembro de 2023.
- Número de Participantes: 5.293 respondentes (aumento de 24% em relação à edição anterior).
- Objetivo da Pesquisa:
  - Mapear o mercado de trabalho em dados no Brasil.
  - Explorar ferramentas utilizadas, práticas adotadas e desafios enfrentados.
  - Avaliar o impacto de tecnologias emergentes, como IA Generativa e LLMs.

0.12.2 Estrutura do Dataset

- Número de Variáveis: 399 colunas organizadas em 8 categorias principais:
  1. Dados Demográficos: Informações como localização, gênero e outros dados de perfil.
  2. Dados sobre Carreira: Cargos, salários, níveis de experiência e rotatividade.
  3. Desafios dos Gestores: Principais dificuldades enfrentadas no setor.
  4. Conhecimentos na Área de Dados: Ferramentas, linguagens de programação e práticas adotadas.
  5. Objetivos de Carreira: Metas e aspirações profissionais.
  6. Engenharia de Dados (DE): Ferramentas e conhecimentos específicos da área.
  7. Análise de Dados (DA): Habilidades relacionadas à visualização e análise de dados.
  8. Ciência de Dados (DS): Modelagem, aprendizado de máquina e práticas avançadas.
- Formato dos Dados:

- Predominância de variáveis categóricas.
- Algumas perguntas possuem respostas multivaloradas, armazenadas em colunas adicionais com identificação no formato `P<Parte><Pergunta>_<Opção>` (exemplo: `P4_d` representa ferramentas utilizadas no trabalho).

0.12.3 Processamento e Anonimização

- Privacidade: Dados sensíveis foram transformados ou excluídos para proteger a identidade dos participantes.
- Agrupamentos: Estados com baixa representatividade foram agrupados por região.
- Outliers: Dados que poderiam identificar participantes foram tratados ou removidos.
- Limitações: A anonimização reduziu a granularidade de alguns dados, como a divisão por estados, que foram agrupados em regiões.

0.13 Pipeline de ETL

0.13.1 Extração

- Fonte: Dataset baixado do Kaggle no formato CSV.
- Formato de Codificação: UTF-8 para evitar problemas de caracteres especiais.
- Objetivo: Garantir a integridade dos dados para posterior processamento e análise.

0.13.2 Transformação

0.13.2.1 Seleção de Colunas Relevantes As seguintes colunas foram selecionadas por sua relevância para o objetivo do trabalho: - Ferramentas Utilizadas: (`P4_d`) Linguagens de programação e ferramentas usadas no trabalho. - Áreas de Atuação: (`P4_a`, `P4_a_1`) Cargos e atividades desempenhadas. - Fontes de Dados Processadas: (`P4_b`) Tipos de fontes analisadas pelos profissionais. - Uso de ChatGPT/LLMs: (`P4_m`) Utilização de modelos de linguagem no trabalho. - IA Generativa: (`P3_f`) Casos de uso de IA generativa nas empresas. - Tamanho das

Equipes: (`P3_a`) Número de profissionais de dados por empresa. - Papéis no Time de Dados: (`P3_b`) Funções desempenhadas nas equipes.

0.13.2.2 Limpeza e Padronização

1. Nomes de Colunas:
    - Remoção de caracteres especiais, como parênteses e aspas.
    - Renomeação para termos mais intuitivos, como "Ferramentas" e "Fontes de Dados".
    - Exemplo: Antes: `P4_d , Quais das linguagens listadas abaixo você utiliza no trabalho?`. Depois: `"ferramentas"`.
2. Tratamento de Valores Ausentes:
    - Substituição por `"Não informado"`, garantindo a completude do dataset.
3. Separação de Respostas Multivaloradas:
    - Divisão de respostas em valores individuais (exemplo: ferramentas separadas por vírgulas).
4. Padronização de Textos:
    - Conversão de textos para letras minúsculas e remoção de espaços extras.

0.13.2.3 Transformações Adicionais

- Agrupamento de Categorias:
    - Respostas menos frequentes foram agrupadas como `"Outros"` para facilitar a visualização.
- Conversão de Tipos de Dados:
    - Garantia de que variáveis categóricas e numéricas estivessem corretamente formatadas.

0.13.3 Carregamento

- O dataset transformado foi armazenado em um DataFrame em memória, preparado para a próxima etapa (EDA).
- As colunas foram renomeadas para facilitar o entendimento, incluindo nomes como:
  - `"ferramentas"`, `"áreas de atuação"`, `"fontes de dados"`, `"chat_llm"`, `"ia generativa"`, `"tamanho da empresa"`.

0.14 Análise Exploratória de Dados (AED)

A Análise Exploratória de Dados (AED) foi conduzida para compreender as características do dataset "State of Data 2023" e identificar padrões nas respostas. Essa etapa se concentrou nas variáveis mais relevantes, analisando distribuições, frequências e características gerais dos dados.

0.14.1 Ferramentas Utilizadas

- Coluna Analisada: `P4_d` (Ferramentas).
- Procedimentos:
  - Divisão dos valores multivalorados para contabilizar cada ferramenta individualmente.
  - Padronização dos nomes para evitar duplicidade (ex.: "Python" e " python").
  - Gráfico de barras horizontais exibiu as ferramentas mais mencionadas.
- Top 5 Ferramentas Mais Utilizadas:
  1. Python
  2. SQL
  3. Excel
  4. Power BI
  5. Tableau

0.14.2 Fontes de Dados Processadas

- Coluna Analisada: `P4_b` (Fontes de Dados).
- Procedimentos:
  - Separação de valores multivalorados.
  - Contagem das fontes mais utilizadas e visualização gráfica.
- Top 5 Fontes de Dados:
  1. Bancos Relacionais (SQL)
  2. Planilhas (Excel/Google Sheets)
  3. APIs
  4. Data Lakes
  5. Dados Web (Web Scraping)

0.14.3 Uso de ChatGPT e LLMs

- Coluna Analisada: `P4_m`.
- Procedimentos:
  - Frequências calculadas para as categorias de resposta ("Sim" e "Não").
  - Gráfico de barras apresentou a distribuição.
- Distribuição:
  - 70% dos profissionais utilizam LLMs no trabalho.

0.14.4 Áreas de Atuação

- Colunas Analisadas: `P4_a`, `P4_a_1`.
- Procedimentos:
  - Análise de frequências por área.
  - Gráfico de barras exibiu as distribuições.
- Top 3 Áreas de Atuação:
  1. Cientista de Dados
  2. Engenheiro de Dados
  3. Analista de Dados

0.14.5 Tamanho das Equipes

- Coluna Analisada: `P3_a`.
- Procedimentos:
  - Frequências calculadas para as faixas de tamanho.
  - Gráfico de barras categorizou por número de integrantes.
- Distribuição:
  - Pequenas equipes (<10 pessoas) predominam.

0.14.6 Aplicações de IA Generativa

- Coluna Analisada: `P3_f`.
- Procedimentos:
  - Agrupamento das respostas pouco frequentes como "Outros".
  - Gráficos de barras separados por categoria.
- Principais Aplicações
  - Automação de Processos
  - Geração de Relatórios
  - Geração de Conteúdo

Gráfico: "Top 10 Ferramentas Utilizadas no Trabalho"

0.14.6.1 Descrição Este gráfico apresenta as 10 principais ferramentas de tecnologia utilizadas pelos profissionais de dados no Brasil, conforme os dados coletados na pesquisa State of Data 2023. A análise reflete a frequência de menções de cada ferramenta pelos participantes.

- Eixo X (Frequência): Quantidade de profissionais que mencionaram cada ferramenta.

- Eixo Y (Ferramentas): Nomes das ferramentas ou linguagens de programação mais citadas.
- Top Ferramentas:

    1. SQL: A ferramenta mais mencionada, com 3.156 menções.
    2. Python: Segunda mais citada, amplamente usada em análise de dados e machine learning.
    3. R: Ferramenta popular entre estatísticos.
    4. Outras ferramentas incluem Visual Basic/VBA, JavaScript, e linguagens mais específicas como SAS/Stata e Scala.
    5. Um número significativo de participantes respondeu "Não informado" ou "Não utilizo nenhuma das linguagens listadas".

0.14.6.2 Insight Obtido

1. SQL como Padrão de Mercado:

    - SQL é uma linguagem fundamental para acesso, manipulação e gestão de dados em bancos relacionais. A alta frequência reflete sua relevância transversal, sendo adotada por analistas, engenheiros e cientistas de dados.

2. Python na Segunda Posição:

    - Python é amplamente utilizado por sua versatilidade e pelas bibliotecas especializadas para análise de dados, aprendizado de máquina e automação de processos.

3. Prevalência de R:

    - A presença do R destaca seu uso em contextos mais acadêmicos ou em estatísticas avançadas.

4. Outras Ferramentas:

    - Linguagens como JavaScript e Visual Basic/VBA indicam um uso mais nichado para automação e desenvolvimento de aplicações interativas.
    - Ferramentas específicas, como SAS/Stata e Scala, são menos usadas, mas continuam sendo importantes para empresas com necessidades mais específicas.

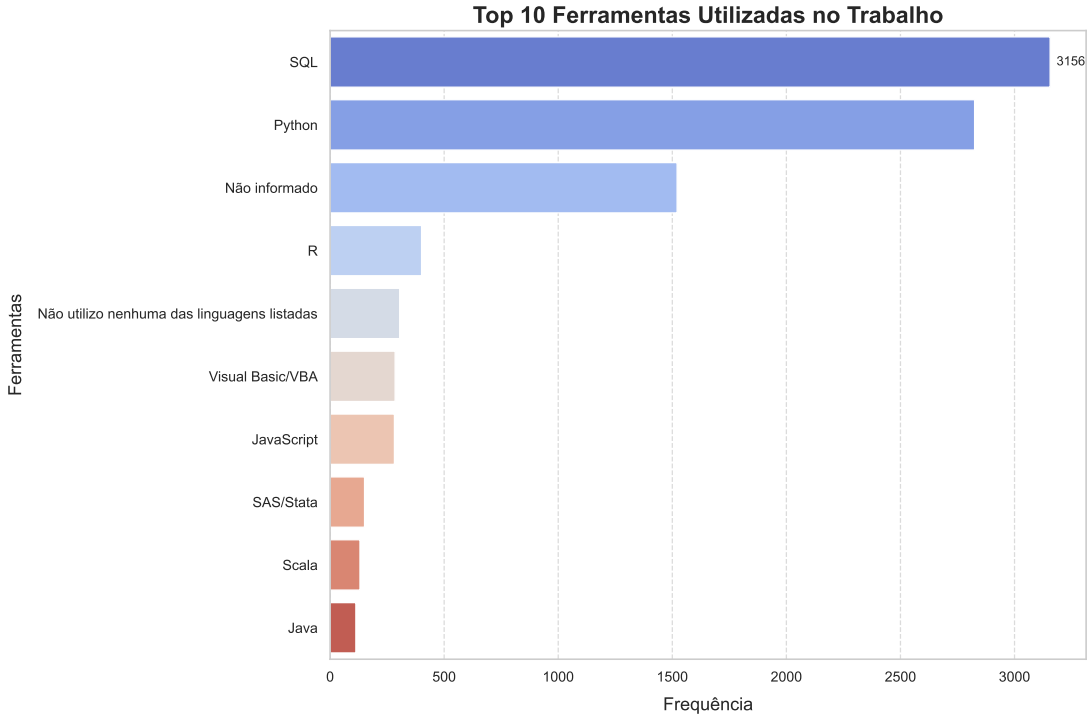
0.14.6.3 Importância

1. Desenvolvimento de Habilidades Prioritárias:
    - SQL e Python são as principais ferramentas que profissionais da área de dados devem dominar para atender às demandas do mercado de trabalho.
    - Essa priorização ajuda na formação de novos profissionais e na estruturação de cursos de capacitação.
2. Adaptação às Necessidades do Mercado:
    - A diversidade de ferramentas demonstra que o mercado brasileiro está amadurecendo, com a adoção de soluções específicas para diferentes etapas do ciclo de vida dos dados.

0.14.6.4 Ensinamentos Práticos

1. Para Profissionais:
    - Focar no aprendizado de SQL e Python como pilares essenciais para entrar ou crescer no mercado de dados.
    - Considerar a aprendizagem de ferramentas de nicho (como R, SAS/Stata) para ganhar vantagem em mercados especializados.
2. Para Empresas:
    - Investir na capacitação de suas equipes em SQL e Python para garantir eficiência nas operações de dados.
    - Avaliar o uso de ferramentas mais avançadas (R, Scala) quando houver necessidades específicas de análise ou processamento.
3. Para Instituições de Ensino:
    - Incorporar o ensino de SQL e Python como componentes essenciais de programas de formação em ciência de dados e tecnologia.

Este gráfico oferece uma visão estratégica sobre as competências mais demandadas no setor de dados, ajudando profissionais e empresas a alinharem seus objetivos com as exigências do mercado.



” Gráfico: “Top 10 Fontes de Dados Utilizadas no Trabalho”

0.14.6.5 Descrição O gráfico apresenta as 10 principais fontes de dados utilizadas pelos profissionais de dados no Brasil, conforme identificado na pesquisa State of Data 2023.

- Eixo X (Frequência): Número de vezes que cada fonte foi mencionada.
- Eixo Y (Fontes de Dados): Categorias das fontes de dados mais utilizadas.
- Top Fontes de Dados:

    1. Planilhas (3.429 menções): Amplamente usadas, principalmente em análises básicas.
    2. Dados relacionais (bancos SQL): Utilizados para armazenamento e análise de dados estruturados.
    3. Textos/Documentos: Dados não estruturados em grande destaque.
    4. Bancos NoSQL: Para armazenamento de dados não estruturados e escaláveis.
    5. Dados georreferenciados: Utilizados em análises espaciais.

6. Outras fontes incluem imagens, áudios, vídeos, e APIs.

0.14.6.6 Insight Obtido

1. Domínio de Planilhas e Bancos Relacionais:
   - Planilhas permanecem como a principal ferramenta para análises rápidas e simples, especialmente em pequenas empresas ou equipes.
   - Bancos SQL são indispensáveis para análises mais estruturadas e complexas, indicando sua ampla aceitação no mercado.
2. Crescimento do Uso de Dados Não Estruturados:
   - A presença significativa de textos/documentos e dados georreferenciados evidencia a expansão de aplicações para dados mais diversificados.
   - Bancos NoSQL ganham espaço em ambientes com necessidade de escalabilidade e flexibilidade.
3. Uso de APIs e Mídia:
   - O uso de APIs para extração de dados externos reflete a crescente interconexão entre sistemas.
   - Fontes como imagens, áudios e vídeos indicam a adoção de tecnologias de IA para análise multimodal.

0.14.6.7 Importância

1. Adaptação às Necessidades do Mercado:
   - O domínio de dados relacionais e não relacionais é essencial para atender às demandas atuais do mercado.
   - Empresas devem equilibrar o uso de fontes estruturadas (SQL) e não estruturadas (NoSQL, APIs, textos).
2. Diferenciação Competitiva:
   - Profissionais que conseguem trabalhar com dados não estruturados (textos, imagens, vídeos) estão mais preparados para lidar com aplicações modernas, como IA e big data.

0.14.6.8 Ensinamentos Práticos

1. Para Profissionais:
    - Fortalecer habilidades em planilhas e bancos relacionais (SQL).
    - Expandir conhecimento em fontes não estruturadas (textos, bancos NoSQL) para atender à crescente demanda por análise avançada.
2. Para Empresas:
    - Priorizar a modernização da infraestrutura de dados, incorporando bancos NoSQL e APIs para maior flexibilidade e escalabilidade.
    - Investir em ferramentas para análise de dados não estruturados, como sistemas de processamento de linguagem natural (NLP) ou análise de imagens.
3. Para Instituições de Ensino:
    - Reforçar a base em bancos SQL enquanto introduz fundamentos de bancos NoSQL e fontes de dados multimodais, preparando os alunos para desafios atuais e futuros.

Este gráfico destaca a importância de dominar fontes estruturadas e a transição para a análise de dados não estruturados, refletindo a evolução das necessidades de mercado e tecnologia.

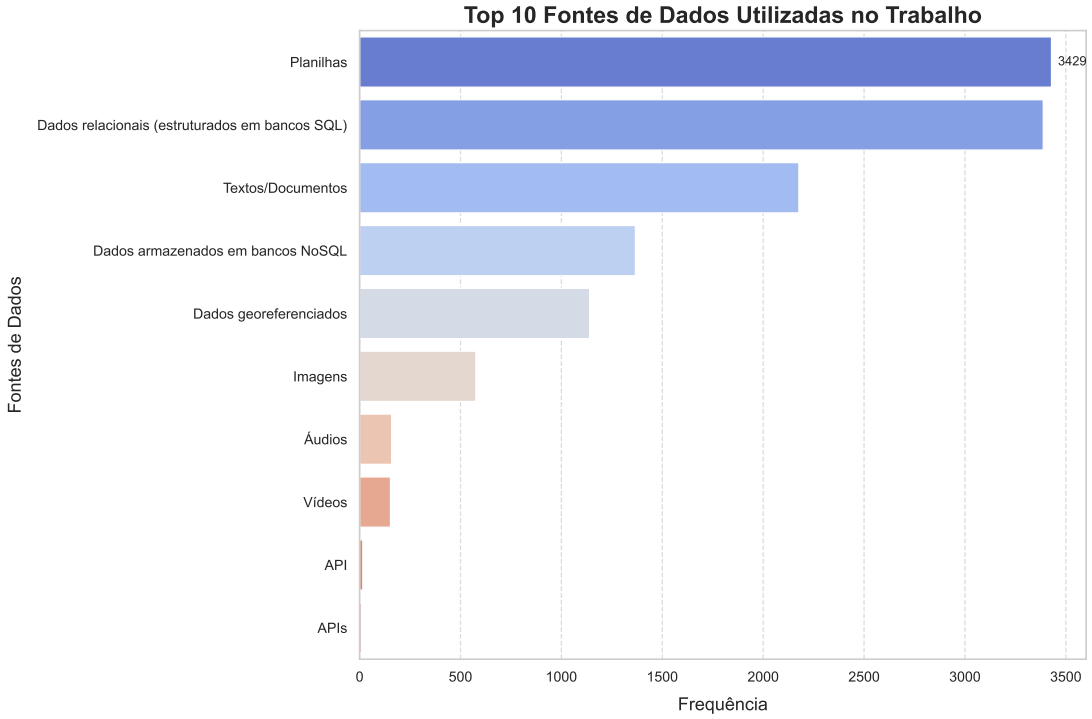


Gráfico: “Uso de ChatGPT/LLMs no Trabalho”

0.14.6.9 Descrição O gráfico apresenta a frequência de respostas relacionadas ao uso de ChatGPT/LLMs (Large Language Models) no ambiente de trabalho, conforme identificado na pesquisa State of Data 2023. Cada barra representa uma categoria de uso ou não uso das ferramentas de IA generativa.

- Eixo X (Frequência): Número de vezes que cada categoria foi mencionada pelos participantes.
- Eixo Y (Uso de ChatGPT/LLMs): Descrição do tipo de uso ou da ausência de uso de ferramentas de IA generativa.

0.14.6.10 Categorias Principais

1. Utilizo apenas soluções gratuitas (como ChatGPT free):

- Representa a maior frequência (2.219 menções).
- Profissionais que adotam versões gratuitas para tarefas diárias.

2. Não utilizo nenhum tipo de solução de IA Generativa:
    - A segunda maior categoria, com participantes que não empregam IA generativa em suas atividades.
3. Utilizo soluções no estilo "Copilot" (ex.: GitHub Copilot, ChatGPT Plus):
    - Usadas para maior produtividade em desenvolvimento e automação.
4. Utilizo soluções pagas de IA Generativa:
    - Inclui modelos avançados como ChatGPT Plus e MidJourney, com pagamento pelo próprio profissional ou pela empresa.

0.14.6.11 Insight Obtido

1. Alta Adoção de Ferramentas Gratuitas:
    - A predominância de soluções gratuitas reflete o impacto democratizante de ferramentas como ChatGPT, permitindo acesso a IA generativa mesmo em empresas de pequeno porte ou para profissionais autônomos.
2. Barreira para Soluções Pagas:
    - Apesar de amplamente reconhecidas, soluções pagas ainda possuem menor adoção, seja por questões financeiras ou por não serem vistas como indispensáveis.
3. Significativo Número de Não Usuários:
    - A presença de muitos profissionais que não utilizam IA generativa indica que há barreiras culturais, técnicas ou de infraestrutura que limitam a adoção.
4. Crescimento de Soluções Estilo "Copilot":
    - Ferramentas como GitHub Copilot mostram o interesse crescente em soluções que integram IA generativa diretamente em fluxos de trabalho específicos.

0.14.6.12 Importância

1. Evolução Tecnológica no Trabalho:
    - A adoção crescente de IA generativa aponta para uma mudança significativa na forma como tarefas analíticas e criativas são realizadas.
2. Acessibilidade como Diferencial:
    - Soluções gratuitas têm um papel crucial na popularização da IA generativa, nivelando o campo de atuação entre empresas de diferentes tamanhos.
3. Necessidade de Treinamento:
    - Profissionais precisam ser capacitados para explorar melhor as vantagens dessas ferramentas, especialmente para tirar maior proveito de soluções pagas e avançadas.

0.14.6.13 Ensinamentos Práticos

1. Para Profissionais:
    - Familiarizar-se com ferramentas gratuitas como ChatGPT e avaliar o impacto delas em produtividade e eficiência.
    - Experimentar soluções pagas para entender os benefícios adicionais em projetos mais complexos.
2. Para Empresas:
    - Investir em treinamento para aumentar a adoção de ferramentas de IA generativa e fomentar a cultura de inovação.
    - Avaliar o custo-benefício de soluções pagas, como o GitHub Copilot, para tarefas específicas.
3. Para Instituições de Ensino:
    - Incluir o uso de ferramentas como ChatGPT nos currículos, capacitando novos profissionais para o mercado em transformação.

Este gráfico evidencia a democratização da IA generativa e os desafios de adoção em diferentes contextos, destacando o papel das ferramentas gratuitas e as oportunidades de expansão com soluções mais avançadas.

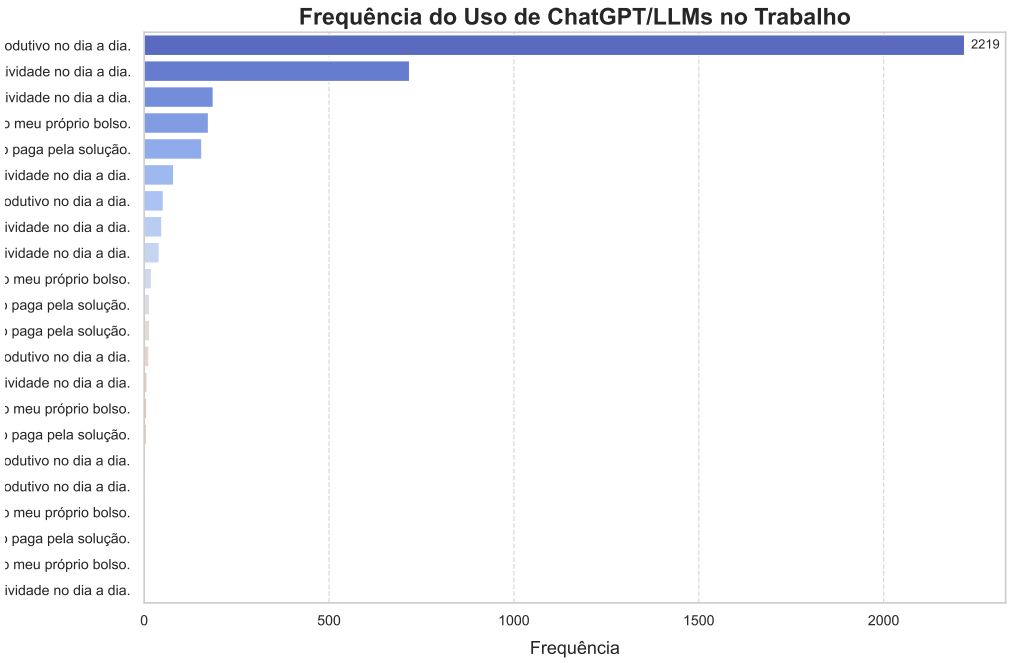


Gráfico: “Distribuição de Áreas de Atuação”

0.14.6.14 Descrição Este gráfico de barras apresenta a distribuição das áreas de atuação dos participantes da pesquisa State of Data 2023. Cada barra representa uma categoria de atuação no mercado de dados, acompanhada da respectiva frequência de profissionais.

- Eixo X (Frequência): Número de profissionais que atuam em cada área específica.
- Eixo Y (Área de Atuação): Principais categorias de atuação no setor de dados.

0.14.6.15 Categorias Principais

1. Análise de Dados:
    - Área mais representativa, com 1.795 respondentes.
    - Inclui atividades de coleta, análise e visualização de dados para apoio na tomada de decisão.
2. Engenharia de Dados:
    - Com grande representatividade, envolve a construção e manutenção de pipelines e infraestruturas de dados.
3. Gestores:
    - Envolve liderança e estratégia para equipes de dados, com significativa presença.
4. Ciência de Dados:
    - Profissionais focados em modelagem preditiva, aprendizado de máquina e análises avançadas.
5. Outras Atividades e Buscando Oportunidade:
    - Inclui profissionais que não se encaixam nas categorias acima ou estão em transição de carreira.

0.14.6.16 Insight Obtido

1. Análise de Dados como Ponto de Entrada:
    - A predominância da análise de dados destaca essa área como a mais acessível para iniciantes e essencial para empresas que estão estruturando times de dados.
2. Demanda Crescente por Engenharia de Dados:
    - A alta frequência de engenheiros de dados reflete a crescente demanda por infraestrutura robusta para lidar com volumes maiores e mais complexos de dados.
3. Presença Significativa de Gestores:

- A representatividade de gestores sugere que as empresas estão amadurecendo na organização de suas equipes de dados, criando lideranças dedicadas.

4. Ciência de Dados em Crescimento:
    - Apesar de ser uma área avançada, a ciência de dados mostra boa representatividade, indicando que as empresas estão investindo em modelos preditivos e inteligência artificial.
5. Transição e Busca por Oportunidades:
    - A presença de respondentes buscando oportunidades reforça a atratividade da área de dados e a necessidade de capacitação contínua.

0.14.6.17 Importância

1. Para Profissionais:
    - Identificar áreas com maior representatividade pode ajudar na escolha de uma trajetória de carreira.
    - Reconhecer a importância da análise de dados como porta de entrada e da engenharia de dados para a sustentação de projetos maiores.
2. Para Empresas:
    - Investir em equipes multidisciplinares que combinem analistas, engenheiros e cientistas de dados para alcançar uma estratégia de dados completa.
    - Desenvolver lideranças para gerenciar times em crescimento e garantir alinhamento com os objetivos estratégicos.
3. Para Instituições de Ensino e Formação:
    - Focar em programas que atendam às demandas de habilidades em análise e engenharia de dados.
    - Criar currículos que reflitam a relevância dessas áreas no mercado.

0.14.6.18 Ensinamentos Práticos

1. Iniciantes devem começar por análise de dados, uma área fundamental e de fácil entrada no mercado.
2. Empresas em expansão precisam priorizar a contratação de engenheiros de dados para garantir a escalabilidade e a integridade dos sistemas.
3. Equipes de dados devem ter gestores experientes para assegurar a integração entre áreas e maximizar o retorno dos investimentos em dados.
4. Profissionais em transição devem buscar capacitação em análise ou engenharia de dados, áreas com maior demanda e que oferecem amplo espaço para crescimento.

Este gráfico reflete um mercado robusto e em expansão, destacando tanto áreas maduras quanto oportunidades de entrada para novos profissionais.

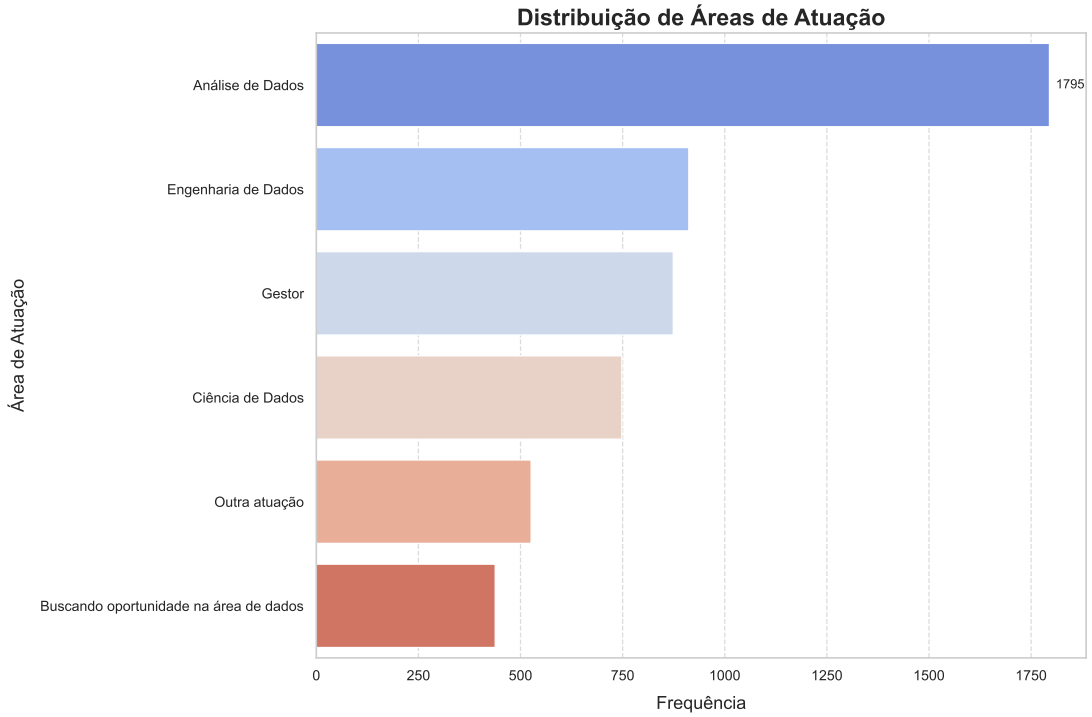


Gráfico: “Distribuição de Atuação Refletida no Dia a Dia”

0.14.6.19 Descrição O gráfico acima representa a distribuição das áreas de atuação que os profissionais identificaram como refletidas em suas atividades do dia a dia. Ele foi gerado a partir das respostas à pesquisa State of Data 2023. Cada barra mostra uma categoria de atuação e a frequência com que ela foi mencionada.

- Eixo X (Frequência): Quantidade de participantes que escolheram cada tipo de atuação.
- Eixo Y (Atuação Refletida no Dia a Dia): As categorias de atuação reconhecidas pelos profissionais como representativas de seu trabalho diário.

0.14.6.20 Categorias Representadas

1. Análise de Dados/BI:
    - Atividade mais mencionada, com 1.795 ocorrências.
    - Envolve cruzamento de dados, identificação de padrões, geração de insights e criação de dashboards e relatórios para suporte à tomada de decisão.
2. Engenharia de Dados:
    - Representa papéis técnicos ligados ao design de arquitetura de dados, desenvolvimento de pipelines, e implementação de soluções como data lakes e data warehouses.
3. Ciência de Dados/Machine Learning/AI:
    - Profissionais que aplicam modelos preditivos e algoritmos para resolver problemas do negócio e otimizar processos.
4. Nenhuma das Frentes Citadas:
    - Pequeno grupo que não se identifica com as categorias principais, indicando funções menos tradicionais ou generalistas.

0.14.6.21 Insights Obtidos

1. Predominância de Análise de Dados:
    - A análise de dados é a atividade mais refletida no dia a dia dos profissionais, reforçando sua importância como base para a tomada de decisões nas empresas.
2. Demanda Técnica por Engenharia de Dados:
    - A significativa presença de engenheiros de dados destaca a necessidade de infraestruturas robustas e escaláveis para suportar operações analíticas e científicas.
3. Crescimento de Ciência de Dados e AI:
    - Apesar de menor em relação às outras áreas, o número de profissionais trabalhando com ciência de dados e AI sugere que as empresas estão adotando modelos avançados para otimizar suas operações.
4. Especialização do Mercado:
    - A segmentação entre essas áreas demonstra que o mercado de dados está se estruturando em funções específicas, indicando maior maturidade no setor.

0.14.6.22 Importância

1. Para Profissionais:
    - Entender as principais frentes de atuação ajuda na escolha e no direcionamento de carreira.
    - A análise de dados pode ser uma entrada natural para o mercado, enquanto engenharia e ciência de dados requerem maior especialização técnica.
2. Para Empresas:
    - A alocação adequada de recursos em análise, engenharia e ciência de dados é essencial para obter insights estratégicos e escalar operações.
    - A presença de diferentes frentes reflete a necessidade de equipes multidisciplinares para cobrir toda a jornada de dados.

3. Para Instituições de Ensino:
    - Cursos devem priorizar habilidades em análise e engenharia de dados, alinhando-se às demandas mais frequentes do mercado.
    - Programas avançados podem incluir ciência de dados e inteligência artificial para profissionais mais experientes.

0.14.6.23 Ensinamentos Práticos

1. Análise de Dados como Base:
    - A maioria dos profissionais reflete essa atividade em seu dia a dia, destacando sua importância como base de qualquer estratégia de dados.
2. Engenharia e Ciência de Dados como Diferenciais:
    - Enquanto a análise é predominante, as áreas de engenharia e ciência de dados são diferenciais para empresas e profissionais buscando excelência técnica.
3. Foco no Desenvolvimento de Carreira:
    - Profissionais podem começar em análise de dados e se especializar posteriormente em engenharia ou ciência de dados, dependendo de suas metas e habilidades.

O gráfico evidencia a diversidade e segmentação das frentes de atuação no mercado de dados, oferecendo insights valiosos para planejamento de carreira, estruturação de equipes e capacitação profissional.

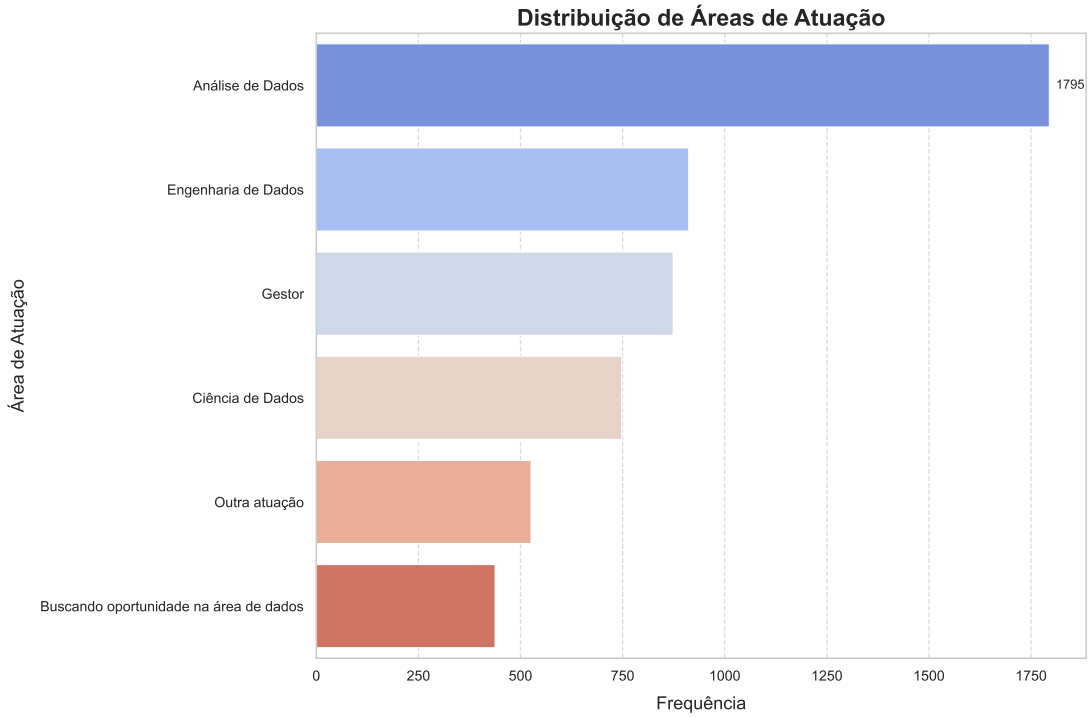


Gráfico: “Distribuição do Tamanho da Empresa”

0.14.6.24 Descrição O gráfico acima representa a distribuição das empresas participantes da pesquisa State of Data 2023, categorizadas de acordo com o tamanho de suas equipes de dados. Ele foi gerado com base na frequência de respostas dos profissionais, que indicaram o número de pessoas atuando diretamente com dados em suas organizações.

- Eixo X (Frequência): Quantidade de empresas em cada categoria.
- Eixo Y (Tamanho da Empresa): Faixas de tamanho baseadas no número de profissionais atuando na área de dados.

0.14.6.25 Categorias Representadas

1. Acima de 300 pessoas:

- Representa grandes empresas, indicando organizações com alto grau de maturidade em dados.

2. 1 - 3 pessoas:
    - Pequenos times ou empresas iniciando esforços relacionados à área de dados.
3. 4 - 10 pessoas:
    - Times intermediários, geralmente encontrados em empresas de médio porte.
4. 11 - 20, 21 - 50 pessoas:
    - Times em expansão, frequentemente alinhados a empresas em crescimento.
5. 101 - 300, 51 - 100 pessoas:
    - Times robustos, indicativos de empresas estruturadas ou grandes corporações.
6. Sem equipe de dados:
    - Empresas que ainda não possuem profissionais dedicados à área de dados.

0.14.6.26 Insights Obtidos

1. Predominância de Equipes Pequenas:
    - A maioria das empresas possui equipes pequenas, especialmente entre 1 a 10 profissionais, sugerindo que muitas empresas ainda estão no estágio inicial de desenvolvimento da área de dados.
2. Presença de Grandes Times:
    - O número significativo de empresas com mais de 300 pessoas em suas equipes indica que grandes corporações já possuem alta maturidade na área de dados.
3. Ausência de Profissionais em Algumas Empresas:

- A presença de empresas que ainda não possuem equipes de dados reflete que o mercado ainda tem espaço para adoção inicial de práticas de análise e engenharia de dados.

4. Segmentação do Mercado:

   - O mercado é diversificado, com uma ampla distribuição entre pequenas, médias e grandes equipes, indicando diferentes níveis de maturidade e demanda por profissionais da área.

0.14.6.27 Importância

1. Para Profissionais:

   - Profissionais podem identificar oportunidades tanto em empresas grandes com equipes maduras quanto em empresas menores que estão começando na área de dados, apresentando grande potencial de crescimento.

2. Para Empresas:

   - Empresas podem usar essa informação para benchmarking, avaliando o tamanho de suas equipes em relação à média do mercado.
   - Organizações com equipes pequenas podem considerar a expansão para acompanhar as tendências do setor.

3. Para o Mercado:

   - O gráfico evidencia a necessidade de formação e capacitação de mais profissionais para atender à demanda crescente, principalmente em pequenas e médias empresas.

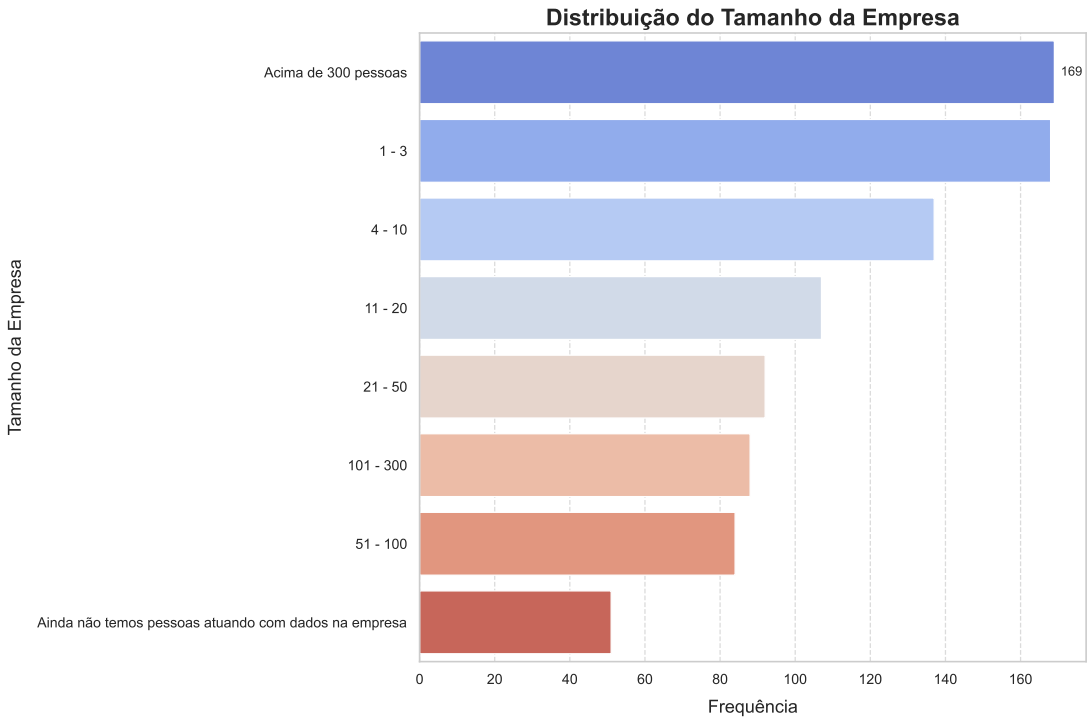
0.14.6.28 Ensinamentos Práticos

1. Oportunidade de Crescimento para Profissionais:

   - Equipes pequenas sugerem que há espaço para os profissionais se destacarem e liderarem iniciativas de dados.

2. Demanda por Profissionais de Dados em Expansão:
    - Empresas de médio porte, que estão ampliando suas equipes, precisam de profissionais experientes para estruturar a área.
3. Desafios em Grandes Corporações:
    - Empresas com grandes equipes de dados necessitam de profissionais altamente especializados para lidar com infraestruturas complexas e projetos avançados.

Este gráfico reforça a importância de adaptar estratégias de contratação e treinamento de acordo com o estágio de maturidade da empresa na área de dados. Ele também demonstra que o mercado de dados no Brasil está em franca expansão e diversificação.



Gráficos: Relação entre Papéis em Times de Dados e o Tamanho da Empresa

0.14.6.29 Descrição dos Gráficos Os gráficos apresentados ilustram a distribuição de diferentes papéis desempenhados em times de dados de empresas de variados tamanhos. Cada gráfico corresponde a um papel específico, incluindo: 1. Engenheiro de Dados 2. Cientista de Dados 3. Analista de Dados 4. Analista de Business Intelligence (BI) 5. Não Informado (papéis não especificados)

O eixo horizontal representa as faixas de tamanho das empresas, enquanto o eixo vertical apresenta a frequência de empresas com profissionais desempenhando esses papéis.

0.14.6.30 Insights Obtidos

1. Engenheiro de Dados:
    - Concentração de engenheiros de dados em empresas com mais de 300 funcionários.
    - Empresas menores (1-10 pessoas) apresentam menor número de engenheiros, sugerindo que essa função é mais crítica em corporações de grande porte, com infraestrutura de dados complexa.
2. Cientista de Dados:
    - Semelhante aos engenheiros, cientistas de dados estão amplamente presentes em empresas grandes (acima de 300 funcionários).
    - O papel também é encontrado em empresas médias (101-300 pessoas), indicando a relevância da modelagem e análise avançada para empresas com recursos intermediários.
3. Analista de Dados:
    - Uma das funções mais amplamente distribuídas, presente em empresas de todos os tamanhos.
    - Destaque para empresas grandes, mas uma distribuição mais uniforme sugere que essa função é essencial independentemente do porte da organização.
4. Analista de BI:
    - Amplamente presente em empresas maiores, mas também com boa representação em empresas médias e pequenas (11-50 funcionários).
    - Indica que a visualização e interpretação de dados são necessidades universais para suporte estratégico.

5. Papéis Não Informados:
    - A frequência elevada no grupo "Não Informado" reflete a falta de clareza na definição de papéis em muitas empresas, sugerindo oportunidade para melhor categorização e alinhamento.

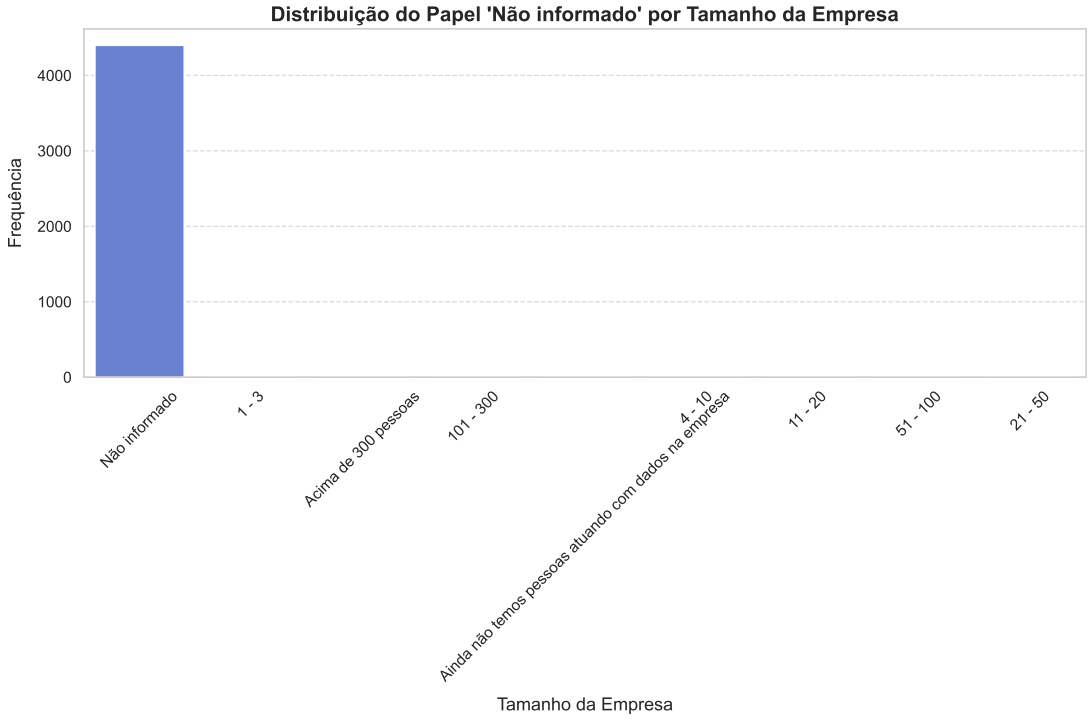
0.14.6.31 Importância e Ensinamentos

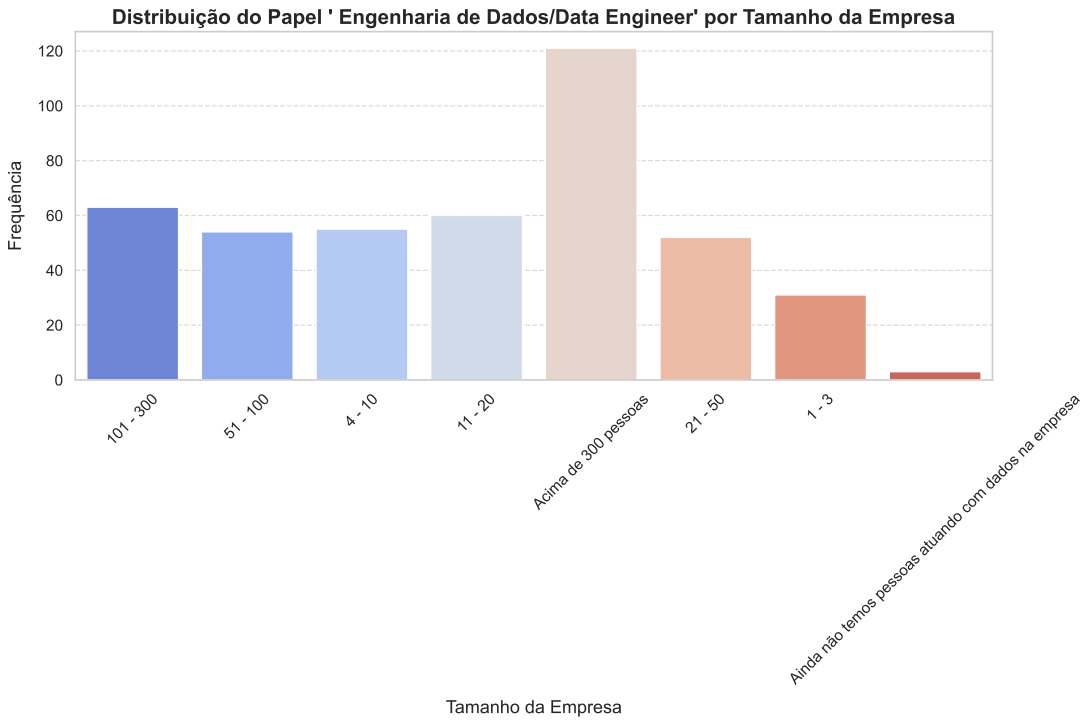
1. Evolução dos Times de Dados:
    - Empresas maiores investem em times diversificados, incluindo papéis especializados como Engenheiro e Cientista de Dados, enquanto empresas menores priorizam funções generalistas como Analistas de Dados.
2. Adaptação ao Contexto Organizacional:
    - Empresas com equipes menores tendem a priorizar profissionais com habilidades amplas, enquanto organizações maiores alocam especialistas para lidar com problemas específicos.
3. Potencial de Estruturação:
    - A alta frequência de "Não Informado" aponta para a necessidade de padronizar descrições de cargos, o que pode beneficiar tanto empregadores quanto profissionais no alinhamento de expectativas e na definição de carreiras.

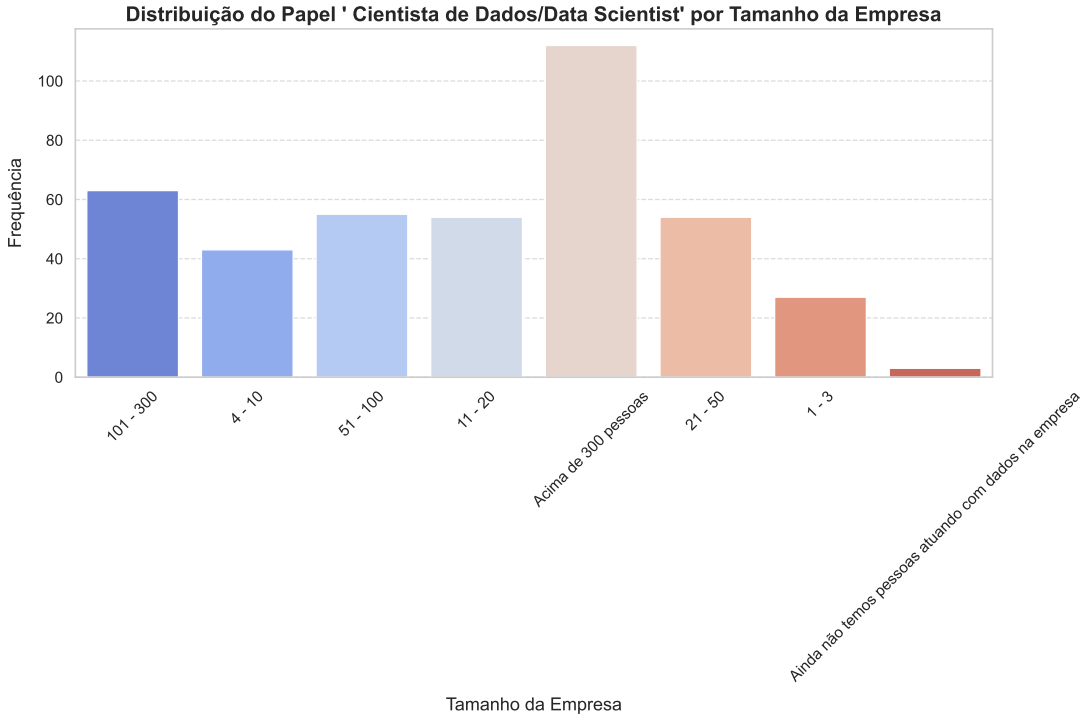
0.14.6.32 Aplicabilidade Prática

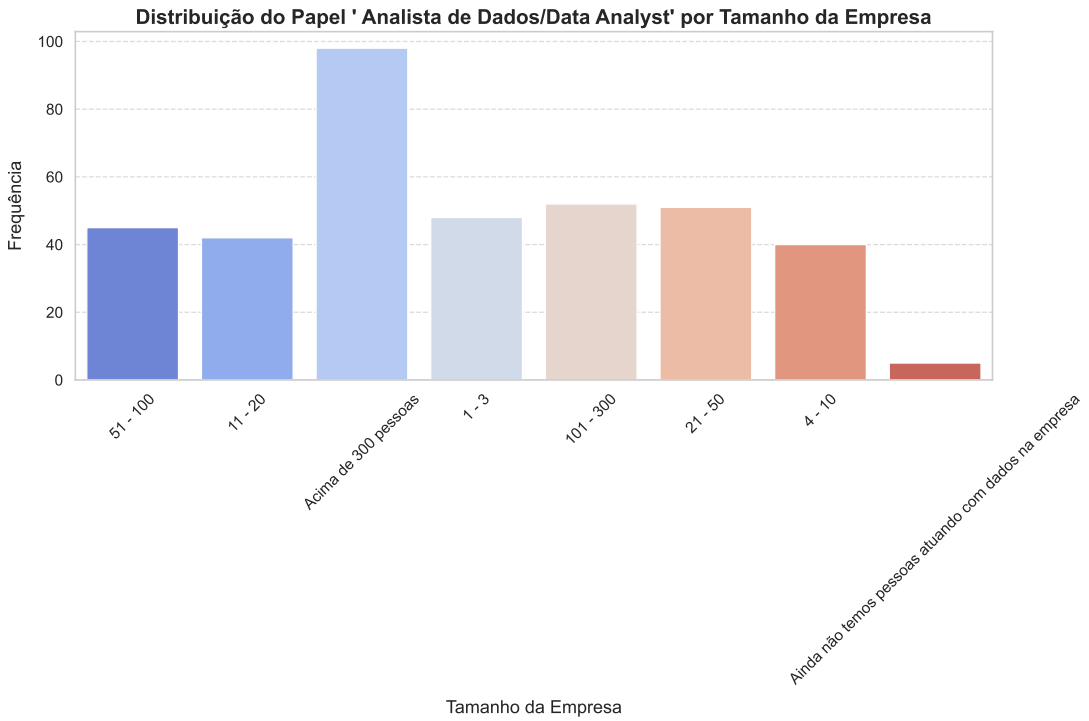
- Empresas de Pequeno Porte: Devem considerar o treinamento de analistas para lidar com demandas múltiplas em ambientes de recursos limitados.
- Empresas de Grande Porte: Indicativo de que investimentos em especialistas, como engenheiros de dados, são cruciais para lidar com arquiteturas complexas e grandes volumes de dados.
- Mercado de Trabalho: Profissionais devem alinhar suas especializações às demandas organizacionais baseadas no porte das empresas em que desejam atuar.

```
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='Tamanho Empresa', ylabel='count'>
## Text(0.5, 1.0, "Distribuição do Papel 'Não informado' por Tamanho da Empresa")
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2, 3, 4, 5, 6, 7, 8], [Text(0, 0, 'Não informado'), Text(1, 0, '1 - 3')
## (array([   0., 1000., 2000., 3000., 4000., 5000.]), [Text(0, 0.0, '0'), Text(0,
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='Tamanho Empresa', ylabel='count'>
## Text(0.5, 1.0, "Distribuição do Papel ' Engenharia de Dados/Data Engineer' por
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2, 3, 4, 5, 6, 7], [Text(0, 0, '101 - 300'), Text(1, 0, '51 - 100'), Tex
## (array([  0.,  20.,  40.,  60.,  80., 100., 120., 140.]), [Text(0, 0.0, '0'), Te
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='Tamanho Empresa', ylabel='count'>
## Text(0.5, 1.0, "Distribuição do Papel ' Cientista de Dados/Data Scientist' por
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2, 3, 4, 5, 6, 7], [Text(0, 0, '101 - 300'), Text(1, 0, '4 - 10'), Text
## (array([  0.,  20.,  40.,  60.,  80., 100., 120.]), [Text(0, 0.0, '0'), Text(0,
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='Tamanho Empresa', ylabel='count'>
## Text(0.5, 1.0, "Distribuição do Papel ' Analista de Dados/Data Analyst' por Tama
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2, 3, 4, 5, 6, 7], [Text(0, 0, '51 - 100'), Text(1, 0, '11 - 20'), Text
## (array([  0.,  20.,  40.,  60.,  80., 100., 120.]), [Text(0, 0.0, '0'), Text(0,
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='Tamanho Empresa', ylabel='count'>
## Text(0.5, 1.0, "Distribuição do Papel ' Analista de Business Intelligence/BI' po
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2, 3, 4, 5, 6, 7], [Text(0, 0, '101 - 300'), Text(1, 0, '51 - 100'), Tex
## (array([ 0., 10., 20., 30., 40., 50., 60., 70., 80., 90.]), [Text(0, 0.0, '0'),
```


Distribuição do Papel 'Não informado' por Tamanho da Empresa
Frequência
4000
3000
2000
1000
0
Não informado
1 - 3
Acima de 300 pessoas
101 - 300
4 - 10
Ainda não temos pessoas atuando com dados na empresa
11 - 20
51 - 100
21 - 50
Tamanho da Empresa


Distribuição do Papel ' Engenharia de Dados/Data Engineer' por Tamanho da Empresa
Frequência
120
100
80
60
40
20
0
101 - 300
51 - 100
4 - 10
11 - 20
Acima de 300 pessoas
21 - 50
1 - 3
Ainda não temos pessoas atuando com dados na empresa
Tamanho da Empresa


Distribuição do Papel ' Cientista de Dados/Data Scientist' por Tamanho da Empresa
Frequência
0
20
40
60
80
100
101 - 300
4 - 10
51 - 100
11 - 20
Acima de 300 pessoas
21 - 50
1 - 3
Ainda não temos pessoas atuando com dados na empresa
Tamanho da Empresa


Distribuição do Papel ' Analista de Dados/Data Analyst' por Tamanho da Empresa
Frequência
0
20
40
60
80
100
51 - 100
11 - 20
Acima de 300 pessoas
1 - 3
101 - 300
21 - 50
4 - 10
Ainda não temos pessoas atuando com dados na empresa
Tamanho da Empresa

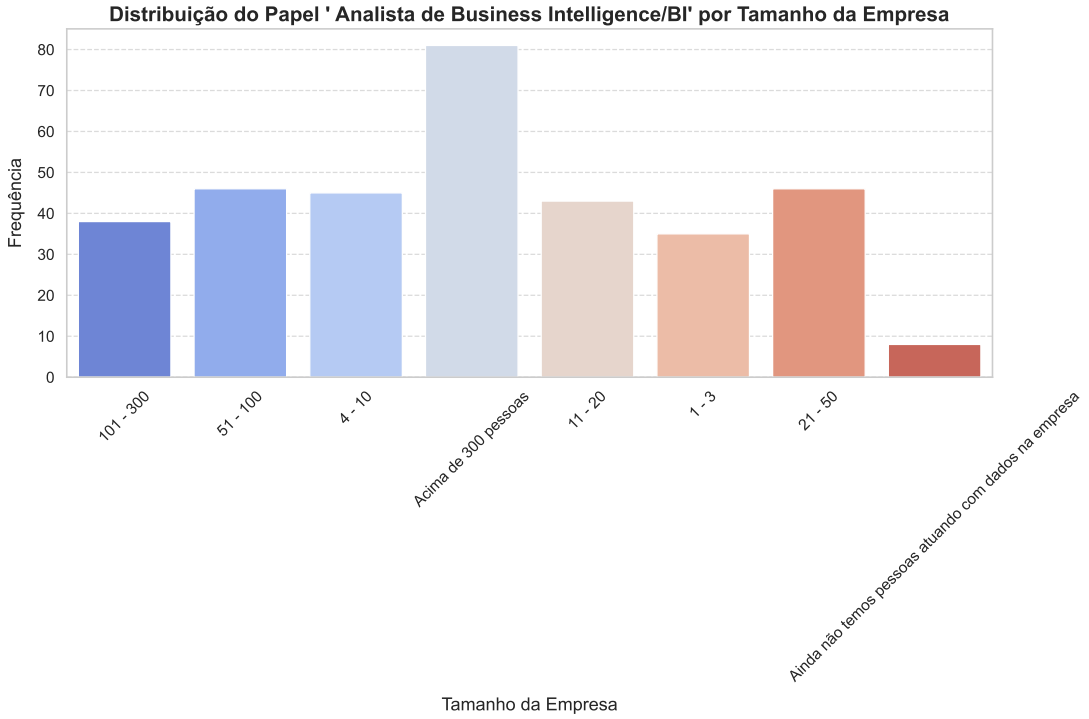


Gráfico: Uso de IA Generativa por Tamanho da Empresa

0.14.6.33 Descrição do Gráfico Este conjunto de gráficos apresenta a distribuição do uso de ferramentas de IA generativa (como ChatGPT e LLMs) em diferentes tamanhos de empresas, categorizadas por faixas numéricas de colaboradores. Cada gráfico detalha um cenário específico de utilização, como uso independente, direcionamento centralizado ou exploração por equipes internas.

0.14.6.34 Insights Obtidos

1. Distribuição do Uso Independente:
    - Empresas com equipes menores (1-3 colaboradores) lideram no uso independente de IA generativa, com o objetivo de aumentar a produtividade.
    - Isso pode ser atribuído à flexibilidade e à necessidade de otimizar recursos em equipes menores.

2. Direcionamento Centralizado:
    - Empresas maiores (acima de 300 colaboradores) destacam-se na implementação de IA generativa com direcionamento centralizado e apoio financeiro.
    - Este comportamento reflete uma maior maturidade e investimento em soluções tecnológicas.
3. Indefinição no Uso de IA Generativa:
    - Em empresas menores (1-3 colaboradores), há um maior índice de respostas "Não sei opinar", indicando uma possível falta de clareza ou conhecimento sobre as iniciativas de IA.
4. Exploração Inicial:
    - Empresas de médio porte (11-50 colaboradores) frequentemente utilizam IA generativa em fases experimentais, visando tanto eficiência operacional quanto diferenciação de produtos.

0.14.6.35 Importância dos Resultados

- O uso de IA generativa varia significativamente com o tamanho da empresa, refletindo recursos e prioridades estratégicas diferentes.
- Empresas pequenas tendem a explorar IA de forma mais independente, enquanto empresas maiores adotam uma abordagem estruturada e centralizada.
- Estes padrões são indicativos do estágio de maturidade tecnológica e da capacidade de investimento de cada segmento.

0.14.6.36 Ensino e Aplicações

- Para Empresas:
    - Empresas menores podem se inspirar nas estratégias centralizadas de empresas maiores para maximizar o impacto da IA generativa.
    - Empresas médias podem priorizar iniciativas experimentais para identificar rapidamente aplicações com alto ROI.
- Para Profissionais de Dados:
    - Compreender os padrões de uso de IA generativa ajuda a identificar oportunidades de mercado.

  - Profissionais podem focar em criar soluções customizadas para empresas menores e em integrar tecnologias em ambientes mais complexos de empresas grandes.

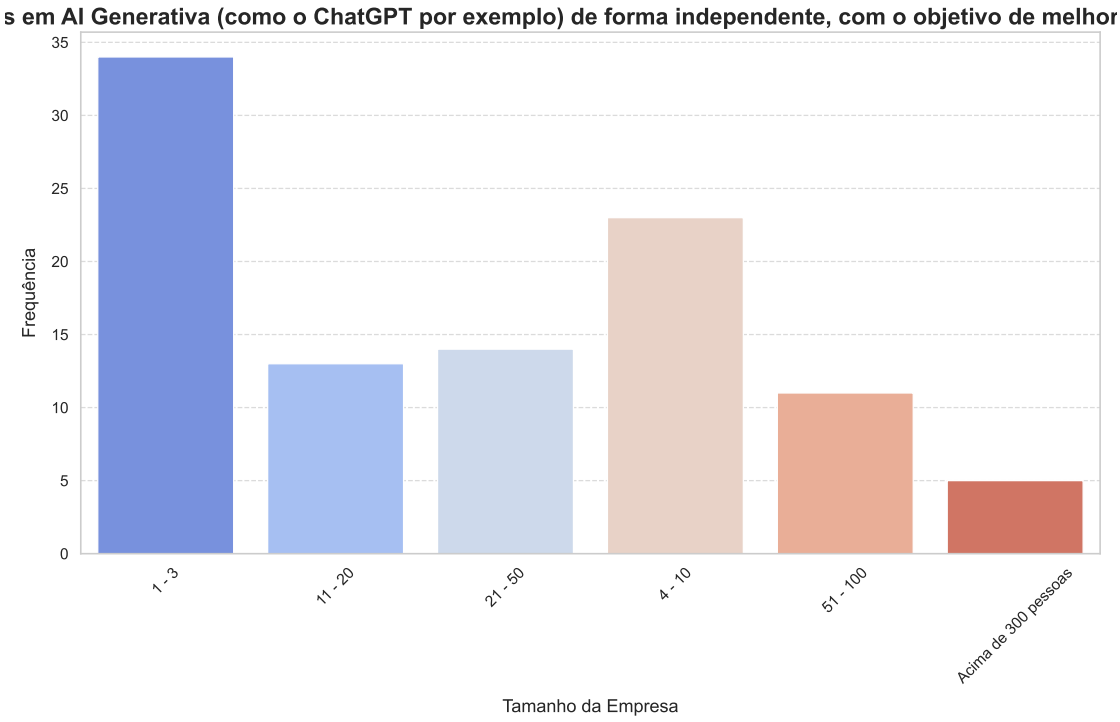
- Para Pesquisadores e Estudantes:
  - Este estudo reforça a necessidade de adaptar ferramentas e práticas ao contexto organizacional.
  - A análise comparativa entre tamanhos de empresas serve como base para futuras pesquisas sobre adoção tecnológica.

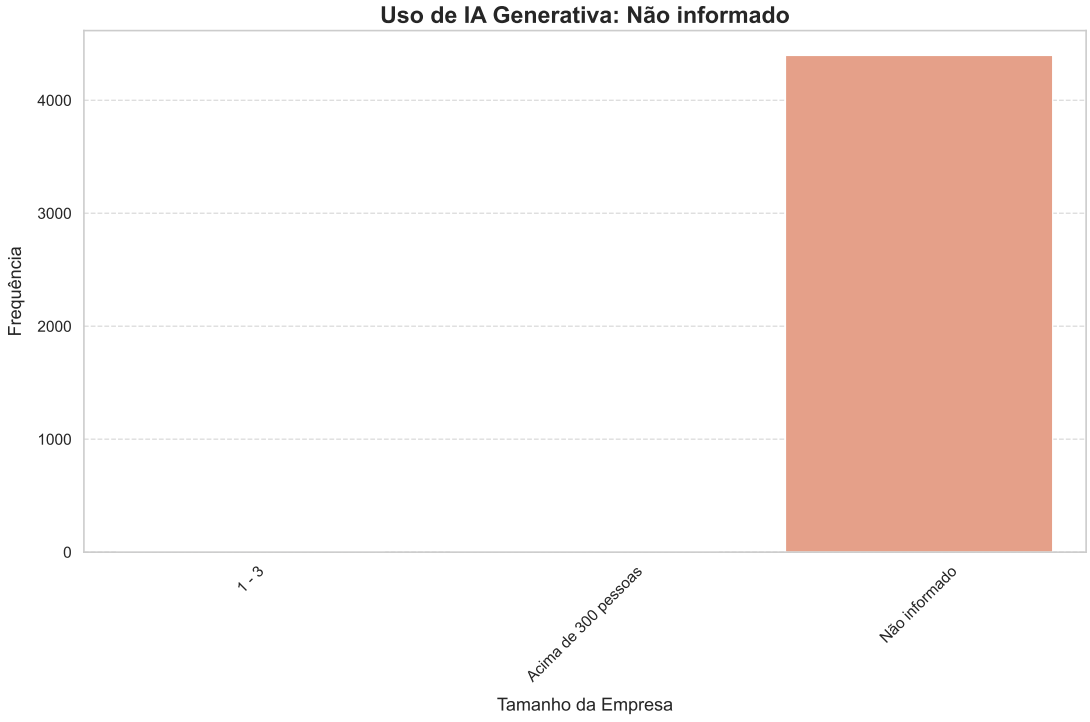
Esses gráficos destacam como diferentes tamanhos de empresas lidam com a inovação tecnológica, fornecendo um panorama útil para decisões estratégicas e estudos de mercado.

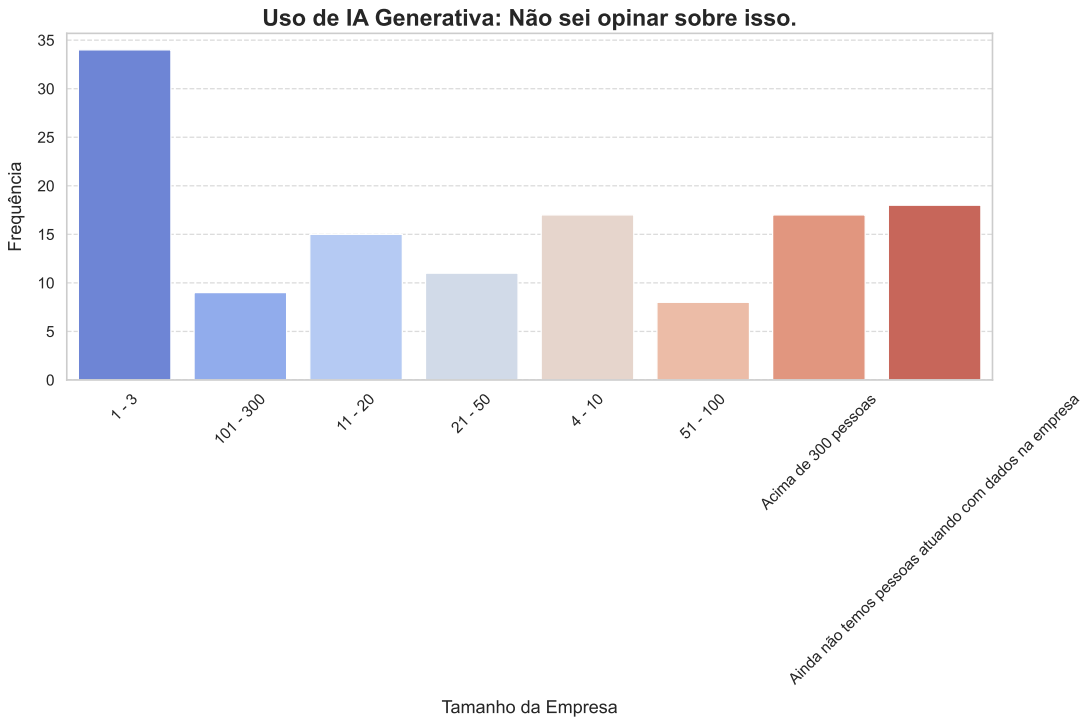
```
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Colaboradores utilizando soluções baseadas
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2, 3, 4, 5], [Text(0, 0, '1 - 3'), Text(1, 0, '11 - 20'), Text(2, 0, '2
## (array([ 0.,  5., 10., 15., 20., 25., 30., 35., 40.]), [Text(0, 0.0, '0'), Text
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Não informado')
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2], [Text(0, 0, '1 - 3'), Text(1, 0, 'Acima de 300 pessoas'), Text(2, 0
## (array([   0., 1000., 2000., 3000., 4000., 5000.]), [Text(0, 0.0, '0'), Text(0,
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Não sei opinar sobre isso.')
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2, 3, 4, 5, 6, 7], [Text(0, 0, '1 - 3'), Text(1, 0, '101 - 300'), Text(2
## (array([ 0.,  5., 10., 15., 20., 25., 30., 35., 40.]), [Text(0, 0.0, '0'), Text
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Não tenho visto soluções de IA Generatica
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
```

```
## ([0, 1, 2, 3, 4, 5, 6, 7], [Text(0, 0, '1 - 3'), Text(1, 0, '101 - 300'), Text(
## (array([ 0.,  5., 10., 15., 20., 25., 30.]), [Text(0, 0.0, '0'), Text(0, 5.0, '5
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Outros')
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2, 3, 4, 5, 6, 7], [Text(0, 0, '1 - 3'), Text(1, 0, '101 - 300'), Text(
## (array([ 0., 10., 20., 30., 40., 50., 60., 70., 80., 90.]), [Text(0, 0.0, '0'),
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Uma ou mais equipes testando e aplicando s
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2, 3, 4, 5], [Text(0, 0, '1 - 3'), Text(1, 0, '101 - 300'), Text(2, 0,
## (array([ 0.,  2.,  4.,  6.,  8., 10., 12.]), [Text(0, 0.0, '0'), Text(0, 2.0, '2
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Uma ou mais equipes testando e aplicando s
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0, 1, 2, 3, 4], [Text(0, 0, '101 - 300'), Text(1, 0, '11 - 20'), Text(2, 0, '4
## (array([ 0.,  2.,  4.,  6.,  8., 10., 12., 14., 16.]), [Text(0, 0.0, '0'), Text
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Equipes de desenvolvimento utilizando solu
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0], [Text(0, 0, 'Acima de 300 pessoas')])
## (array([0., 1., 2., 3., 4., 5., 6.]), [Text(0, 0.0, '0'), Text(0, 1.0, '1'), Tex
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Equipes de desenvolvimento utilizando solu
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0], [Text(0, 0, 'Acima de 300 pessoas')])
## (array([0., 1., 2., 3., 4., 5., 6., 7., 8.]), [Text(0, 0.0, '0'), Text(0, 1.0,
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Existe um direcionamento centralizado para
```

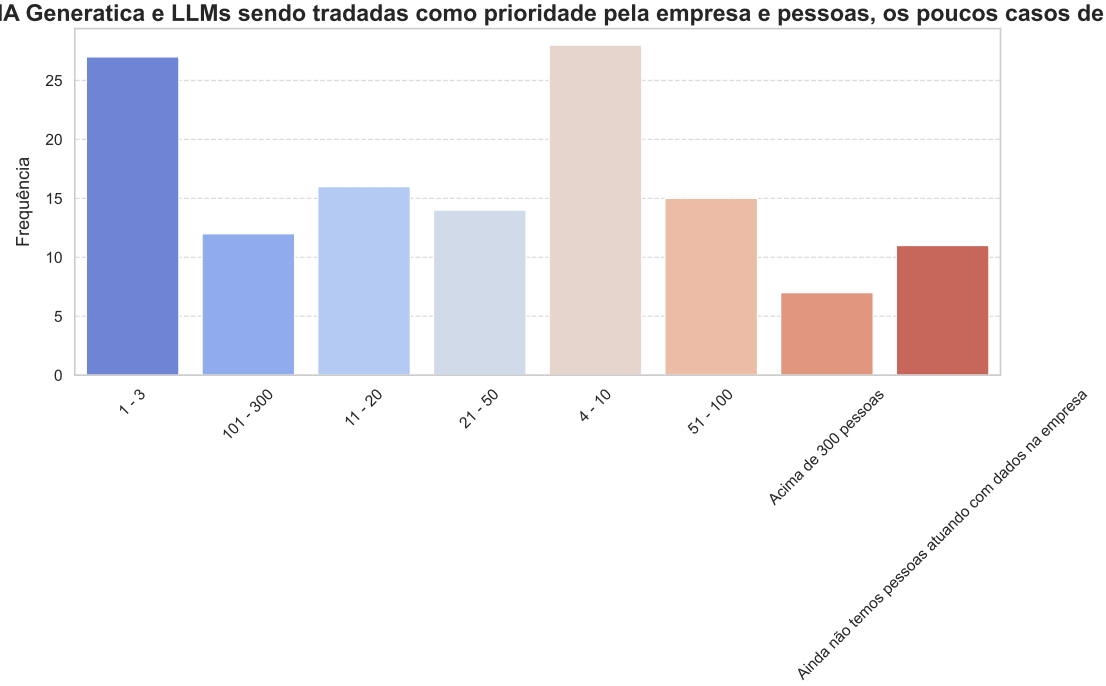
```
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0], [Text(0, 0, 'Acima de 300 pessoas')])
## (array([0., 1., 2., 3., 4., 5., 6., 7., 8., 9.]), [Text(0, 0.0, '0'), Text(0, 1
## <Figure size 1200x800 with 0 Axes>
## <Axes: xlabel='tamanho_empresa', ylabel='Frequência'>
## Text(0.5, 1.0, 'Uso de IA Generativa: Uma ou mais equipes testando e aplicando s
## Text(0.5, 0, 'Tamanho da Empresa')
## Text(0, 0.5, 'Frequência')
## ([0], [Text(0, 0, 'Acima de 300 pessoas')])
## (array([0., 1., 2., 3., 4., 5., 6.]), [Text(0, 0.0, '0'), Text(0, 1.0, '1'), Tex
```



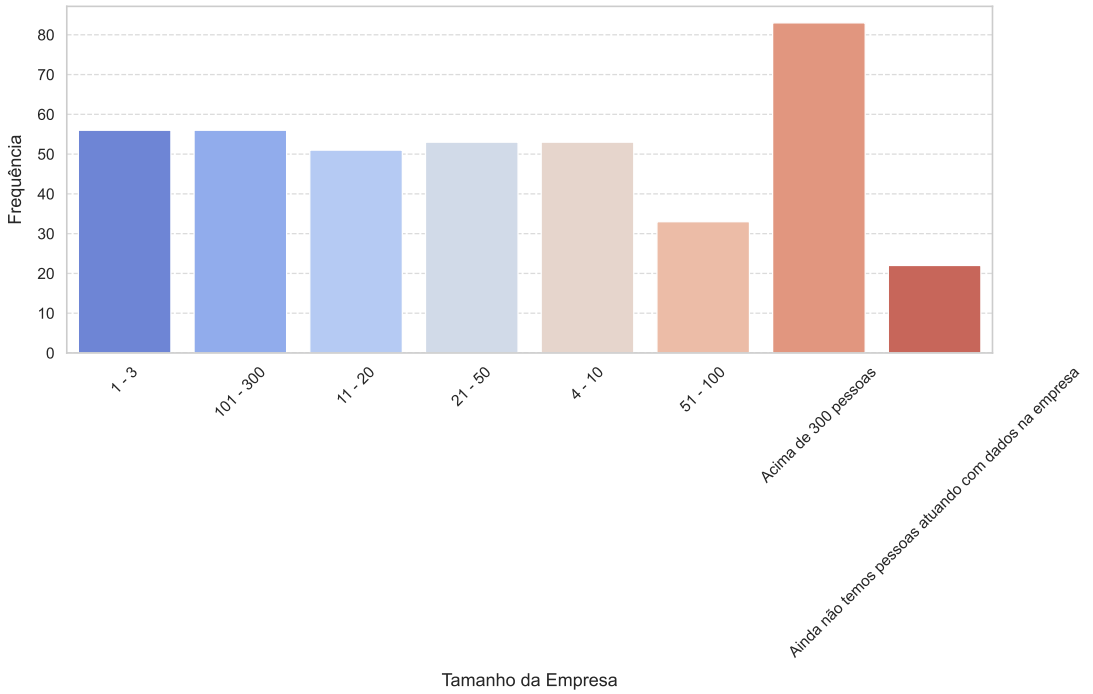

Uso de IA Generativa: Não informado
Frequência
4000
3000
2000
1000
0
1 - 3
Acima de 300 pessoas
Não informado
Tamanho da Empresa


Uso de IA Generativa: Não sei opinar sobre isso.
Frequência
35
30
25
20
15
10
5
0
1 - 3
101 - 300
11 - 20
21 - 50
4 - 10
51 - 100
Acima de 300 pessoas
Ainda não temos pessoas atuando com dados na empresa
Tamanho da Empresa

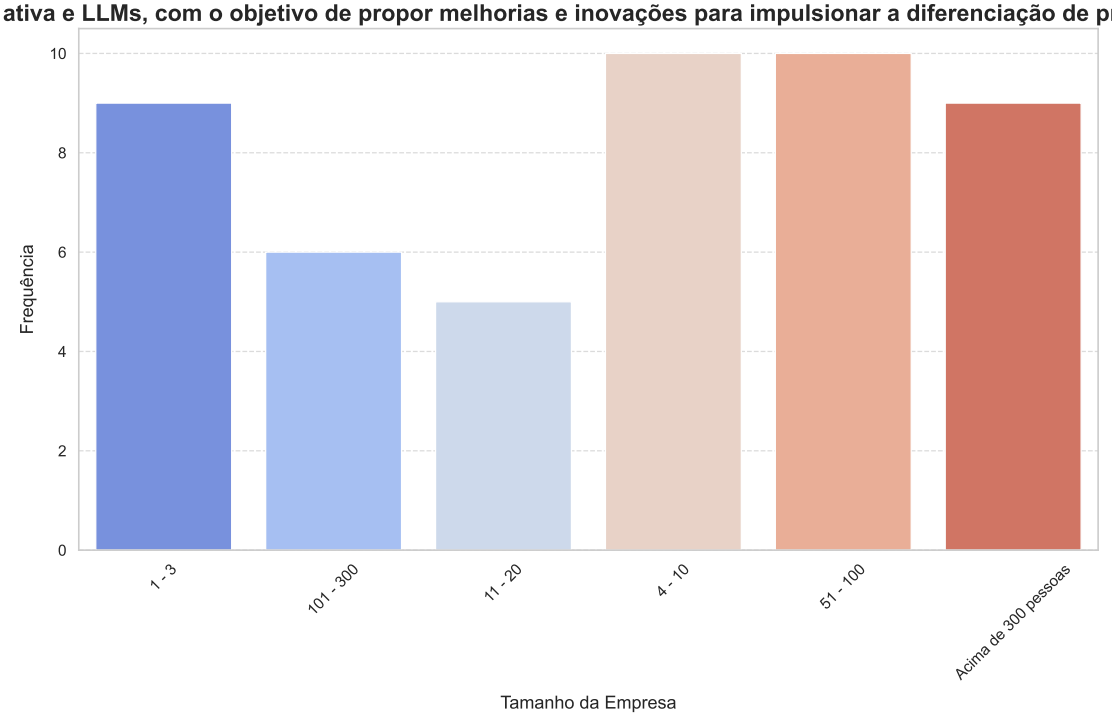
IA Generatica e LLMs sendo tradadas como prioridade pela empresa e pessoas, os poucos casos de



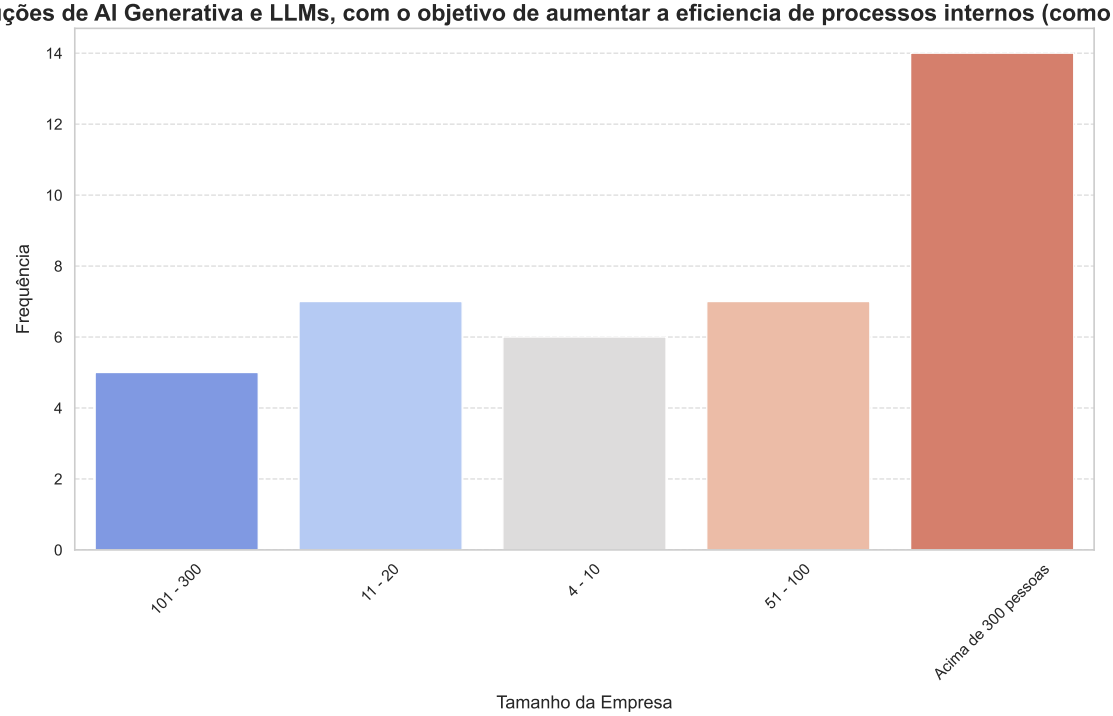
Uso de IA Generativa: Outros



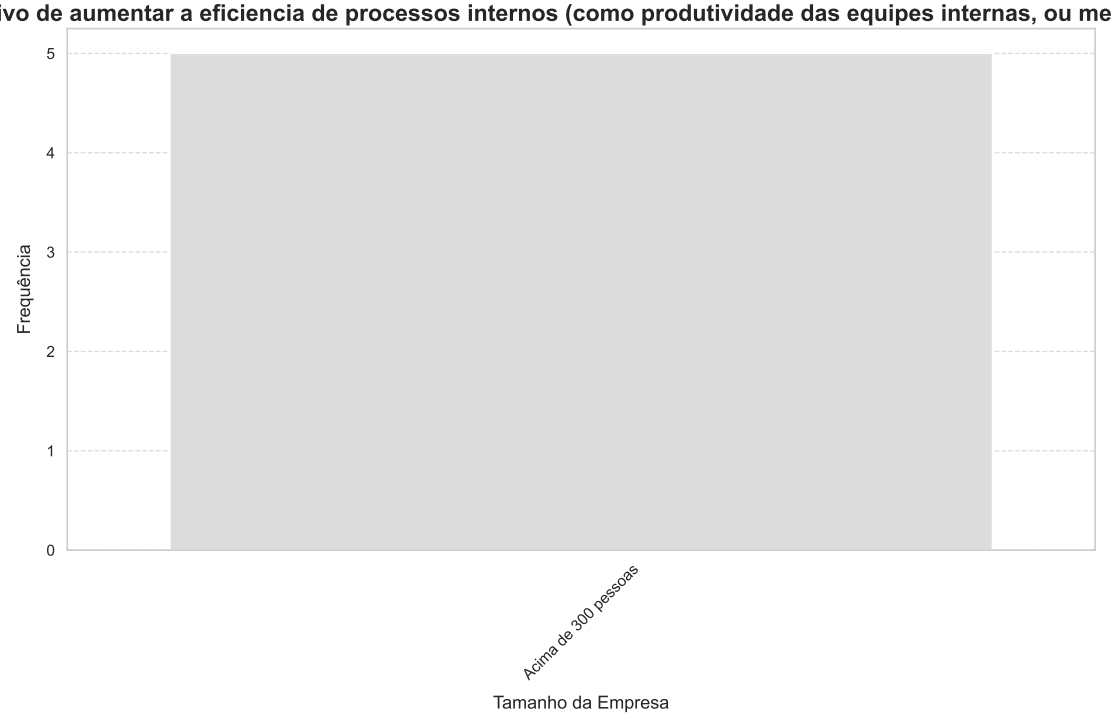
ativa e LLMs, com o objetivo de propor melhorias e inovações para impulsionar a diferenciação de pı



ıções de AI Generativa e LLMs, com o objetivo de aumentar a eficiencia de processos internos (como

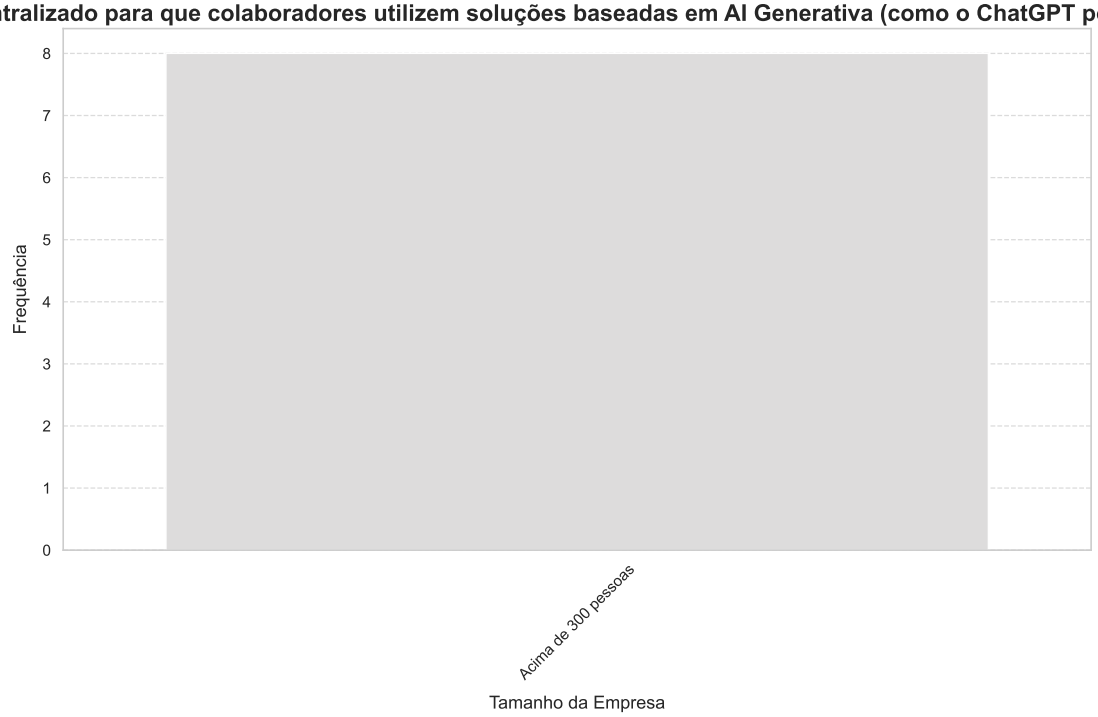


ivo de aumentar a eficiencia de processos internos (como produtividade das equipes internas, ou me

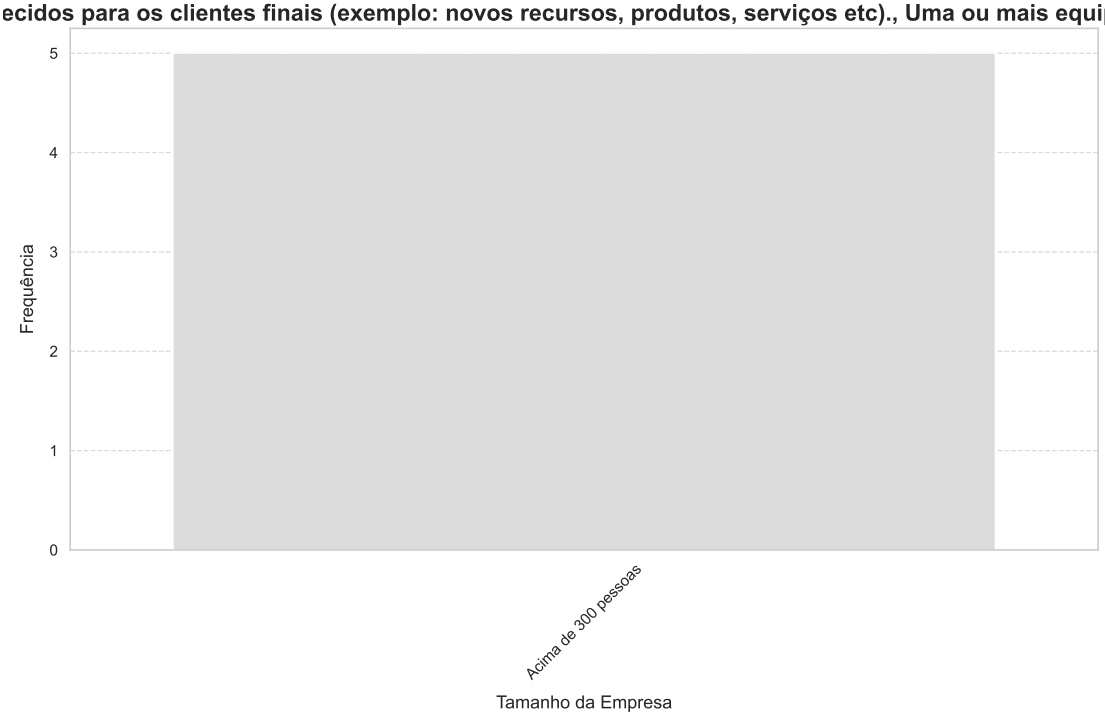


ivo de propor melhorias e inovações para impulsionar a diferenciação de produtos oferecidos para os

ıtralizado para que colaboradores utilizem soluções baseadas em AI Generativa (como o ChatGPT po



ecidos para os clientes finais (exemplo: novos recursos, produtos, serviços etc)., Uma ou mais equip



0.15 Conclusão

O trabalho realizado abordou duas frentes principais de análise:

0.15.1 Web Scraping

Foi implementado um script de scraping para o site Mercado Livre, com foco na coleta de dados relacionados a produtos da linha iPhone. Utilizando ferramentas como Python, Requests e BeautifulSoup, foram extraídos nome do produto, preço e link para compra. A análise foi limitada a 10 páginas, enfrentando desafios como duplicatas, preços incompletos e bloqueios pelo site. Os dados coletados foram apresentados em gráficos que permitiram interpretar a distribuição e comparação de preços, oferecendo uma visão clara do mercado analisado.

0.15.2 ETL e Análise Exploratória de Dados (EDA)

Com o dataset "State of Data 2023", obtido da Data Hackers em parceria com a Bain & Company, foi implementado um pipeline de ETL para transformar os dados e garantir sua integridade. A Análise Exploratória de Dados (EDA) destacou ferramentas mais utilizadas no mercado, fontes de dados preferenciais, aplicação de IA generativa e áreas de atuação dos profissionais. Esses insights forneceram uma visão abrangente sobre o mercado de dados no Brasil.

0.15.3 Considerações Finais

O trabalho demonstrou domínio técnico em ferramentas de manipulação e análise de dados, bem como capacidade de extrair insights práticos. Na parte de web scraping, foram identificadas limitações e sugeridas melhorias para futuras implementações. A análise do dataset "State of Data 2023" trouxe insights valiosos sobre o mercado de dados, sendo útil para profissionais, empresas e instituições de ensino. Este projeto reflete o uso estratégico de ciência de dados para resolver problemas reais e agregar valor a partir de dados disponíveis.

0.16 Bibliografia

1. Mercado Livre. Disponível em: https://www.mercadolivre.com.br. Acesso em: [data de acesso].
2. Data Hackers, Bain & Company. "State of Data 2023". Pesquisa conduzida entre outubro e dezembro de 2023. Disponível em: https://www.kaggle.com.

3. Hunter, J.D. “Matplotlib: A 2D Graphics Environment.” *Computing in Science & Engineering*, vol. 9, no. 3, 2007, pp. 90–95. DOI: 10.1109/MCSE.2007.55.
4. Pedregosa, F. et al. “Scikit-learn: Machine Learning in Python.” *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–2830.
5. Richardson, L. “Beautiful Soup Documentation.” Disponível em: https://www.crummy.com/software/BeautifulSoup/. Acesso em: [data de acesso].