Análise sistemática acerca de um *raw dataset*: Dinâmica do Mercado de Dados: 2021 vs 2022

Erick Coutinho
Universidade Federal do Ceará
Campus Jardins de Anita
Curso de Ciência de Dados
Itapajé, CE-Brazil
erickcoutinho@alu.ufc.br

Shelda Souza
Universidade Federal do Ceará
Campus Jardins de Anita
Curso de Ciência de Dados
Itapajé, CE-Brazil
sheldasouza@alu.ufc.br

Maverick Alekyne
Universidade Federal do Ceará
Campus Jardins de Anita
Curso de Ciência de Dados
Itapajé, CE-Brazil
alekyne10@alu.ufc.br

***Abstract*—O artigo examina um conjunto de dados sobre o mercado de trabalho na área de dados, fornecido pela Data Hackers, destacando sua importância na compreensão das dinâmicas e tendências em constante evolução nesse setor. A análise oferece insights valiosos sobre a demografia dos profissionais, desafios enfrentados, preferências de ferramentas e tecnologias, e expectativas em relação ao ambiente de trabalho. Ao explorar os dados, identificamos padrões que podem influenciar estratégias de recrutamento, desenvolvimento de habilidades e retenção de talentos. Entre as 2.645 instâncias analisadas, observamos que o cargo de arquiteto de dados possui o maior salário. Além disso, constatamos que a maioria dos profissionais na área é do sexo masculino (1.924), com uma média salarial de 9.659, enquanto as profissionais do sexo feminino (436) apresentam uma média salarial de 8.036. Essas análises iniciais proporcionam insights valiosos para compreender a satisfação profissional, fatores motivacionais e as principais demandas do mercado. A partir dessa exploração, é possível contribuir de diversas formas para o aprimoramento desse setor em crescimento.**

***Index Terms*—Análise de Dados. Insights valiosos. Desafios e preferências.**

I. Introdução

A área de dados atualmente cresce rapidamente no Brasil e no mundo, com a digitalização massiva de processos e a explosão de dados gerados a cada momento criaram uma demanda sem precedentes por profissionais qualificados em lidar com essas informações. Empresas perceberam que a capacidade de extrair insights valiosos dos dados pode conferir uma vantagem competitiva significativa. Além disso, o avanço das tecnologias de armazenamento, processamento e análise de dados, aliado ao aumento da disponibilidade de dados em tempo real, abriu novas possibilidades para a utilização estratégica das informações. O surgimento de áreas como inteligência artificial, machine learning e análise preditiva também desempenhou um papel crucial, ampliando o escopo e a sofisticação das aplicações da área de dados.

A análise é justificada pela sua relevância em proporcionar uma compreensão mais profunda das dinâmicas e tendências específicas desse setor em constante evolução. Ao coletar e analisar esses dados, é possível extrair insights valiosos sobre a demografia dos profissionais, os desafios enfrentados, as preferências tecnológicas e as perspectivas em relação ao ambiente de trabalho.

Do ponto de vista científico, a análise desse dataset é fundamental para informar e embasar estratégias educacionais, decisões empresariais e políticas públicas relacionadas à área de dados no Brasil. A constante evolução desse campo exige uma compreensão detalhada das necessidades do mercado de trabalho, e os dados coletados proporcionam uma base sólida para pesquisas empíricas que impactam positivamente diversas esferas da sociedade.

II. Pré-processamento dos dados

A. Descrição da mineração dos dados

O primeiro passo é a renomeação das colunas para tornar a nomenclatura mais clara. Em seguida, são realizados procedimentos de pré-processamento, como a substituição de valores ausentes por ”Não informado” e a padronização dos valores da coluna ”Cargo Atual” para letras minúsculas.

B. Exploração e Visualização dos dados

Inicialmente, são exploradas visualizações que destacam a distribuição dos níveis de ensino em todo o conjunto de dados, oferecendo insights sobre a educação dos funcionários. Além disso, a análise se estende à comparação desses níveis entre gestores, revelando diferenças educacionais potenciais entre grupos profissionais.

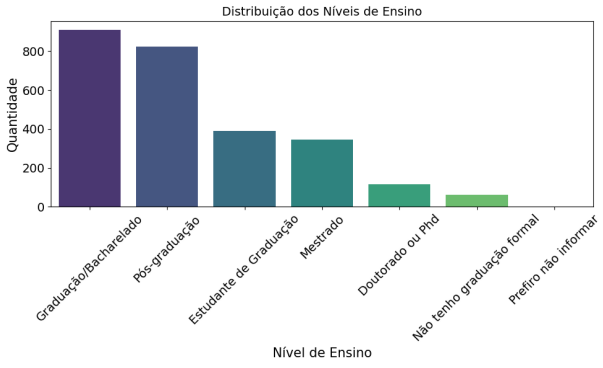

Fig. 1: Nivel de ensino dos entrevistados.

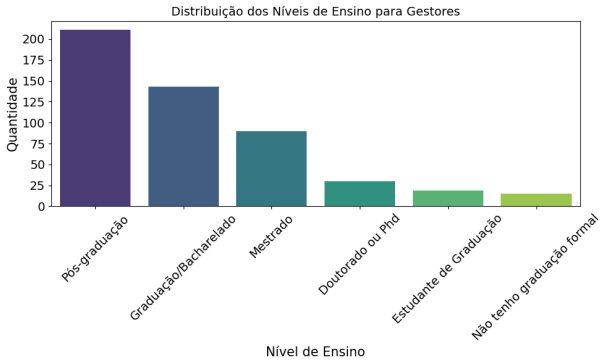

Fig. 2: Nivel de ensino dos gestores.

A análise também se estende à identificação dos principais critérios considerados ao decidir mudar de emprego, apresentando uma contagem detalhada e um gráfico que destaca visualmente esses critérios.

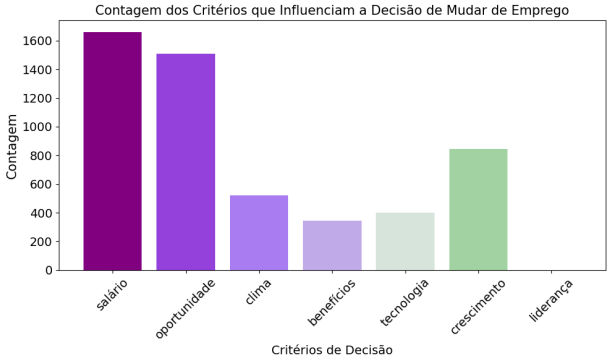

Fig. 3: Criterios de mudança de emprego

Categoria	Frequência
Salário	1659 vezes
Oportunidade	1509 vezes
Clima	522 vezes
Benefícios	345 vezes
Tecnologia	399 vezes
Crescimento	845 vezes
Liderança	0 vezes

TABLE I: Contagem de motivos

A análise de disparidade salarial de gênero é abordada de maneira detalhada, destacando médias salariais e comparando salários entre homens e mulheres nos mesmos cargos. A visualização final destaca a contagem de gênero, proporcionando uma compreensão rápida da distribuição de homens e mulheres na amostra.

Gênero	Média de Salário
Feminino	8036.048165
Masculino	9659.544699

TABLE II: Tabela de Gênero e Média de Salário.

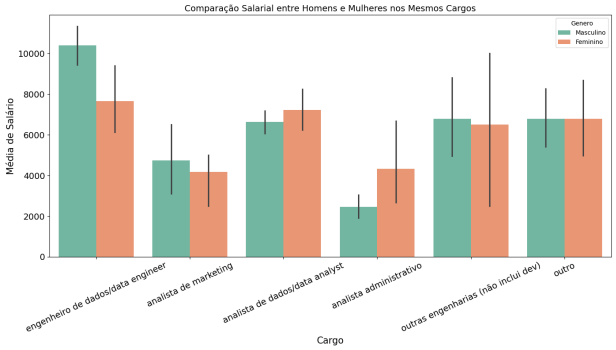

Fig. 4: Comparação salarial - Cargos

C. Principais técnicas usadas na primeira etapa da disciplina: pré-processamento

- Renomeação das colunas do conjunto de dados para tornar os nomes mais descritivos e padronizados.
- Os valores ausenteres (NA) foram substituidos pelo texto ”Não Informado”, em contraste com a abordagem convencional de preenchimento utilizando métricas estatísticas como média, mediana ou moda, essa abordagem foi realizada devido a natureza das variáveis qualitativas, a qual não se prestam bem a análises estatísticas tradicionais, uma vez que não apresentam uma estrutura numérica.
- Para visualização, usamos a linguagem Python atilizando mais as bibliotecas Seaborn e Matplotlib.

III. Avaliando um segundo *raw dataset*

Para comparação com a etapa anterior, iremos analisar um conjunto de dados com o mesmo propósito e proveniente da mesma empresa, porem, de 2022. Ao examinarmos ambos os conjuntos de dados, procuramos identificar padrões, mudanças ao longo do tempo e nuances adicionais que possam aprimorar nossa compreensão da temática em foco. Essa análise comparativa não apenas nos permitirá observar possíveis evoluções, mas também nos ajudará a identificar áreas de interesse e pontos de foco para as próximas etapas da pesquisa.

IV. Resultados

A primeira comparação que conduziremos será sobre o nível de ensino dos entrevistados, visando determinar se houve alterações ao longo do período analisado.

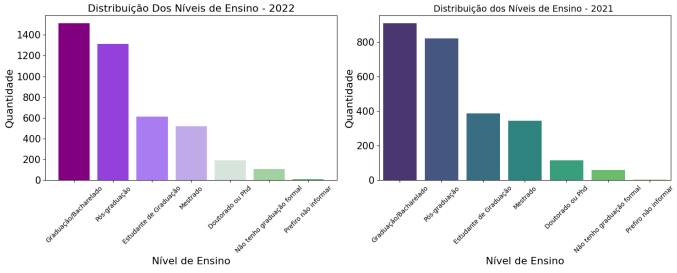

Fig. 5: Comparação de Ensino - 2021/2022

Como observado, as variações entre os anos de 2021 e 2022 foram praticamente insignificantes, sugerindo uma tendência à estabilidade nesse período. A alteração mais expressiva ocorreu na quantidade de entrevistados, registrando 4271 em 2022 em comparação com 2645 em 2021.

A segunda análise realizada consistiu na avaliação do uso de linguagens de programação no ambiente de trabalho, com o objetivo de identificar quais linguagens são mais frequentemente utilizadas pelos entrevistados, e a mudança de linguagem com o tempo.

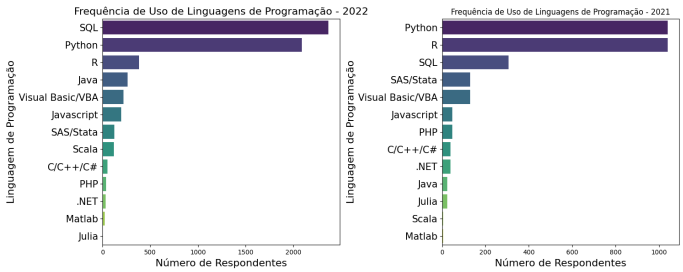

Fig. 6: Linguagens utilizadas - 2021/2022

Ao analisar o gráfico, nota-se que a linguagem R era a mais predominante na área de dados, juntamente com o Python. No entanto, em 2022, um ano posterior, ela sofreu uma redução no número de usuários, caindo para a terceira posição. Por outro lado, o SQL, que ocupava a terceira posição em 2021, ascendeu para o primeiro lugar em 2022. É importante destacar que os entrevistados nas análises não são necessariamente os mesmos, o que pode resultar em disparidades e variações significativas nos resultados.

Na última análise, sob uma perspectiva social, buscamos contrastar a média salarial e a presença de homens e mulheres na área de dados. Essa abordagem visa evidenciar diferenças existentes entre os gêneros no contexto profissional, historicamente, a presença feminina na área de Tecnologia da Informação (TI) tem sido reduzida, influenciada por estereótipos e desafios culturais. Contudo, nos últimos anos, observamos uma mudança significativa impulsionada por iniciativas e movimentos dedicados à promoção da diversidade de gênero na TI.

Gênero	**Média de Salário**
Feminino	6958.750473
Masculino	8896.650908

TABLE III: Média de Salário por Gênero - 2022

Gênero	**Contagem**
Masculino	3194
Feminino	1056

TABLE IV: Contagem de Entrevistados por Gênero - 2022

As tabelas apresentadas fornecem a média salarial por gênero dos entrevistados em 2022, revelando uma disparidade significativa. Os homens têm uma média salarial aproximadamente 27,8% superior à das mulheres. Além disso, em termos de cargos ocupados, os homens superam as mulheres em incríveis 202%.

Gênero	Média de Salário
Feminino	8036.048165
Masculino	9659.544699

TABLE V: Média de Salário por Gênero - 2021

Genero	Quantidade
Masculino	2144
Feminino	493

TABLE VI: Contagem de Entrevistados por Gênero - 2021

Na análise dos dados de 2021, nota-se que a disparidade salarial entre homens e mulheres é de cerca de 20,2

CONCLUSÕES

O mercado de dados tem experimentado um crescimento exponencial ao longo dos anos, tornando-se um elemento fundamental em diversas indústrias e setores. Esse crescimento é impulsionado pela crescente importância de insights baseados em dados para tomadas de decisão estratégicas. Empresas estão cada vez mais reconhecendo o valor intrínseco dos dados e a capacidade de transformá-los em conhecimento acionável [3]. A ascensão de tecnologias como inteligência artificial, machine learning e análise de dados contribuiu significativamente para a evolução do mercado de dados. Essas tecnologias capacitam organizações a extrair padrões, identificar tendências e tomar decisões mais informadas, impulsionando a inovação e a eficiência operacional [5].

Em Resumo, a análise comparativa entre os anos de 2021 e 2022 no mercado de dados proporcionou insights valiosos sobre as tendências e dinâmicas dessa área em constante evolução. Observou-se notáveis disparidades de gênero, tanto em questões salariais quanto na distribuição de cargos. Além disso, foram identificadas mudanças significativas no cenário das linguagens de programação predominantes, variações no nível de ensino dos profissionais atuantes na área e os motivos que levam indivíduos a permanecerem ou buscarem oportunidades em outras empresas. Esses elementos revelam a complexidade e a diversidade de fatores que influenciam o panorama do mercado de dados, fornecendo uma compreensão abrangente das dinâmicas em jogo. [5].

Dessa vez, observando os dados referentes a 2021, nota-se que a disparidade salarial entre homens e mulheres é de cerca de 20,2%, uma cifra menor em comparação com o ano de 2022, com diferença percentual de 7,6% Quanto aos cargos ocupados, os homens superam as mulheres em 334%, uma porcentagem ainda mais expressiva do que a observada no ano subsequente, indicando um aumento da participação feminina no mercado de trabalho [4] [2].

Observamos uma alteração significativa na preferência por

linguagens de programação, com o SQL ascendendo para a liderança em 2022. Essa mudança reflete a constante adaptação dos profissionais às demandas do mercado e a evolução das ferramentas tecnológicas [4].

Contudo, é essencial ressaltar que as conclusões extraídas baseiam-se nos dados disponíveis e na análise específica realizada. A variação na composição dos entrevistados entre os anos pode impactar as conclusões, indicando a importância de uma abordagem cautelosa na interpretação dos resultados. A tendência geral aponta para a estabilidade em diversos aspectos, indicando uma consistência nas escolhas e nas características dos profissionais de dados ao longo do período analisado. Essa estabilidade sugere uma maturidade no mercado, mas também ressalta a necessidade contínua de monitorar e abordar questões críticas, como a disparidade de gênero. Em suma, esta análise proporciona insights valiosos para compreender as nuances do mercado de dados, enfatizando a importância de estratégias inclusivas e a constante adaptação às mudanças no cenário tecnológico. O monitoramento contínuo dessas tendências é crucial para promover um ambiente mais equitativo e sustentável para todos os profissionais envolvidos na área de dados.

REFERENCES

[1] REDATOR. Profissões na área de dados: entenda as possíveis carreiras. Disponível em: ¡https://industriall.ai/blog/profissoes-na-area-de-dados-entenda-as-possiveis-carreiras¿. Acesso em: 29 nov. 2023.

[2] Data Hackers 2021 Great Expectations. Disponível em: ¡https://www.kaggle.com/code/mpwolke/data-hackers-2021-great-expectations¿. Acesso em: 29 nov. 2023.

[3] MELO, C. Qual o Cenário de Data Science no Brasil hoje? Disponível em: ¡https://sigmoidal.ai/qual-o-cenario-de-data-science-no-brasil-hoje/¿. Acesso em: 15 nov. 2023.

[4] State of Data 2022 Um raio-x dos profissionais de dados do Brasil. [s.l: s.n.]. Disponível em: ¡https://abrir.link/QCxds¿. Acesso em: 29 nov. 2023.

[5] Ciência de Dados Brasil: Descubra o Cenário de Ciência de Dados no Brasil. Disponível em: ¡https://abrir.link/g90Kql¿. Acesso em: 21 nov. 2023.