

Ciência de Dados Aplicada ao Mercado de Trabalho de Pernambuco: processamento e visualização dos microdados da PNAD Contínua

Data Science Applied to the Labor Market in Pernambuco: Processing and Visualization of PNAD Contínua Microdata

Karen Evellyn Vieira Ribeiro

kevr@discente.ifpe.edu.br

Rodrigo Sena Rodrigues

rsr12@discente.ifpe.edu.br

Lutemberg Francisco de Andrade Santana

lutemberg.santana@paulista.ifpe.edu.br

RESUMO

O presente trabalho descreve o desenvolvimento de uma solução baseada em Ciência de Dados aplicada à aquisição, tratamento, integração, análise e visualização de microdados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) do Instituto Brasileiro de Geografia e Estatística (IBGE), focando na exploração das características e desigualdades estruturais do mercado de trabalho de Pernambuco no período de 2019 a 2025. O percurso metodológico envolveu a coleta automatizada de microdados do portal do IBGE através de um *pipeline* implementado na linguagem *Python*, extração e gerenciamento de arquivos compactados locais e tratamento de dados de largura fixa para matrizes estruturadas via *pandas* e *numpy*. Os resultados obtidos seguiram uma linha analítica e uma linha tecnológica: na analítica, através da análise exploratória dos dados, notou-se uma desigualdade sociodemográfica e de rendimento no estado; como entrega na linha tecnológica, desenvolveu-se uma aplicação *web* interativa utilizando o *framework Streamlit* e a biblioteca *Plotly*, viabilizando a exploração dinâmica de indicadores e o acesso a informações socioeconômicas do estado. O estudo comprova a eficácia do uso da tecnologia como um instrumento ágil de diagnóstico social para apoiar a investigação das características e desigualdades do mercado de trabalho.

Palavras-chave: Ciência de Dados; PNAD Contínua; Mercado de trabalho; Desigualdade socioeconômica; Painel interativo.

ABSTRACT

This study describes the development of a Data Science-based solution applied to the acquisition, processing, integration, analysis, and visualization of microdata from the Continuous National Household Sample Survey (PNAD Contínua), conducted by the Brazilian Institute of Geography and Statistics (IBGE). The study focuses on exploring the characteristics and structural inequalities of the labor market in the state of Pernambuco, Brazil, from 2019 to 2025. The methodological approach involved the automated collection of microdata from the IBGE portal through a pipeline implemented in Python, as well as the local extraction and management of compressed files and the processing of fixed-width data into structured datasets using the pandas and numpy libraries. The results followed both an analytical and a technological approach. From the analytical perspective, exploratory data analysis revealed sociodemographic and income inequalities within the state. From the technological perspective, an interactive web application was developed using the Streamlit framework and the Plotly library, enabling the dynamic exploration of labor market indicators and access to socioeconomic information about the state. The study demonstrates the effectiveness of technology as an agile social diagnostic tool to support the investigation of the characteristics and inequalities present in the labor market.

Keywords: Data Science; PNAD Contínua; Labor market; Socioeconomic inequality; Interactive dashboard.

1 INTRODUÇÃO

O mercado de trabalho brasileiro é marcado por transformações econômicas, tecnológicas e sociais que afetam as formas de inserção ocupacional, os níveis de rendimento e as condições de trabalho da população. Essas transformações ocorrem em um contexto caracterizado por desigualdades persistentes, expressas em diferenças de renda, escolaridade, formalização dos vínculos e oportunidades de acesso às ocupações. Nesse contexto, Saboia *et al.* (2021) destacam a heterogeneidade do mercado de trabalho brasileiro, evidenciada pela coexistência de distintas formas de inserção ocupacional, níveis de formalização e condições de rendimento. A compreensão dessas dinâmicas exige, portanto, a análise conjunta de diferentes dimensões capazes de caracterizar a estrutura e as desigualdades presentes no mercado de trabalho.

Entre essas dimensões, a escolaridade assume papel relevante na compreensão das diferenças de inserção e rendimento. Entretanto, a relação entre

educação e mercado de trabalho não ocorre de maneira isolada ou automática. Conforme discutido por Frigotto (2007, 2018), a formação deve ser compreendida em articulação com as condições econômicas e sociais que estruturam o mundo do trabalho. De maneira complementar, Bourdieu (1998) demonstra que os efeitos associados ao capital educacional são mediados pelas condições sociais nas quais os indivíduos estão inseridos, de modo que a ampliação da escolaridade, embora relevante, não elimina, por si só, as desigualdades existentes.

Além da escolaridade, fatores como posição na ocupação, formalização do vínculo, jornada de trabalho, rendimento e características sociodemográficas contribuem para compreender as diferentes formas de participação no mercado de trabalho. Conforme Saboia *et al.* (2021), a análise das condições de inserção laboral requer considerar a heterogeneidade das ocupações e dos vínculos, uma vez que diferentes formas de participação no mercado de trabalho estão associadas a condições distintas de proteção e remuneração. A heterogeneidade dessas condições evidencia a necessidade de análises capazes de considerar simultaneamente múltiplas variáveis, evitando interpretações baseadas em uma única dimensão da inserção ocupacional.

No estado de Pernambuco, a investigação dessas características assume particular relevância diante das diferenças observadas nas formas de inserção laboral e nos níveis de rendimento. Para analisar essas questões, a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), realizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE), constitui uma importante fonte de informações, por disponibilizar microdados relacionados às características demográficas, educacionais e laborais da população.

Entretanto, a utilização dos microdados da PNAD Contínua envolve desafios relacionados ao volume de informações, à organização dos arquivos, à seleção e interpretação das variáveis e à necessidade de tratamento computacional para sua utilização analítica. Diante dessas características, técnicas e ferramentas de Ciência de Dados apresentam potencial para sistematizar as etapas de aquisição, tratamento, integração, análise e visualização das informações, transformando dados brutos em estruturas organizadas e passíveis de exploração.

A partir dessa perspectiva, estabelece-se a seguinte pergunta de pesquisa: como técnicas e ferramentas de Ciência de Dados podem ser empregadas na

aquisição, no tratamento, na integração, na análise e na visualização dos microdados da PNAD Contínua para apoiar a investigação das características e desigualdades do mercado de trabalho de Pernambuco?

Em consonância com essa questão, o objetivo geral deste trabalho consiste em desenvolver uma solução baseada em Ciência de Dados para automatizar a aquisição, o tratamento, a integração, a análise e a visualização dos microdados da PNAD Contínua, possibilitando a exploração das características e desigualdades do mercado de trabalho de Pernambuco no período de 2019 a 2025.

Para o alcance desse objetivo, definem-se como objetivos específicos: automatizar a aquisição e o processamento dos microdados trimestrais referentes ao estado; estruturar, tratar e integrar os dados selecionados, considerando as características das variáveis e suas condições de aplicabilidade; analisar aspectos sociodemográficos e laborais relacionados à inserção ocupacional, à formalização, à escolaridade, à jornada de trabalho e ao rendimento; identificar padrões e associações entre as variáveis selecionadas; e desenvolver uma aplicação *web* interativa destinada à visualização e à exploração dos indicadores produzidos.

A articulação entre esses objetivos permite integrar a dimensão computacional do trabalho à investigação empírica do mercado de trabalho pernambucano. Para tanto, são empregados *scripts* desenvolvidos em *Python* para automatizar a aquisição, a manipulação e o tratamento dos arquivos, associados a ferramentas de análise e visualização interativa. Esse fluxo possibilita explorar, de forma integrada, diferentes dimensões do mercado de trabalho, contemplando informações relativas à posição ocupacional, à formalização dos vínculos, à jornada de trabalho, à escolaridade, ao rendimento e às características sociodemográficas.

Em razão da natureza aplicada e do caráter exploratório-descritivo da pesquisa, o estudo não se orienta pelo teste de uma hipótese previamente estabelecida, mas pela investigação da questão de pesquisa a partir do desenvolvimento e da aplicação da solução computacional proposta. Nesse sentido, os resultados são examinados com o propósito de identificar e descrever padrões, distribuições e associações entre as variáveis selecionadas, sem pretensão de estabelecer relações causais.

A escolha do período compreendido entre 2019 e 2025 permite contemplar diferentes momentos recentes do mercado de trabalho pernambucano, abrangendo

alterações ocorridas ao longo desse intervalo. A delimitação temporal possibilita examinar as variáveis selecionadas em diferentes períodos, preservando, contudo, o caráter exploratório e descritivo da investigação e as particularidades metodológicas da PNAD Contínua.

O desenvolvimento da solução também se justifica pela possibilidade de ampliar as condições de acesso e exploração de dados públicos por meio de ferramentas computacionais. A automatização do processamento dos microdados e sua integração a uma aplicação *web* interativa permitem que informações originalmente distribuídas em arquivos de maior complexidade sejam organizadas e apresentadas de maneira estruturada, favorecendo a análise dos indicadores e demonstrando a aplicação prática de conhecimentos de Ciência de Dados e desenvolvimento de sistemas na investigação de questões socioeconômicas.

Dessa forma, o trabalho articula conhecimentos de Análise e Desenvolvimento de Sistemas e Ciência de Dados à investigação empírica do mercado de trabalho, tendo Pernambuco como domínio de aplicação. Sua contribuição compreende tanto a construção de um fluxo computacional para aquisição, processamento, análise e visualização dos microdados quanto a exploração das informações produzidas, permitindo identificar características, padrões e desigualdades relacionadas à inserção ocupacional, à escolaridade, à jornada de trabalho e aos rendimentos.

Além desta introdução, o trabalho está estruturado em quatro seções. A Seção 2 apresenta o Referencial Teórico, abordando as principais discussões sobre mercado de trabalho, desigualdades e inserção ocupacional, além dos fundamentos de Ciência de Dados que sustentam o estudo. Na sequência, a Seção 3 descreve a Metodologia, contemplando a base de dados, o recorte da pesquisa, as ferramentas computacionais e os procedimentos de tratamento e análise. Já a Seção 4 apresenta e discute os Resultados, reunindo os padrões identificados nos dados e a aplicação *web* desenvolvida para visualização e exploração interativa dos indicadores. Por fim, a Seção 5 apresenta as Considerações Finais, destacando as principais contribuições e limitações do estudo, bem como as possibilidades de continuidade do trabalho.

2 REFERENCIAL TEÓRICO

Esta Seção apresenta os fundamentos teóricos que sustentam a pesquisa, contemplando as discussões sobre mercado de trabalho, desigualdades e inserção ocupacional, bem como os conceitos relacionados à Ciência de Dados e à Mineração de Dados e os recursos computacionais empregados na coleta, tratamento, análise e visualização de grandes volumes de informações.

2.1 Mercado de Trabalho, Desigualdades e Inserção Ocupacional

O cenário econômico e social contemporâneo impõe desafios significativos para a inserção e a permanência dos indivíduos no mercado de trabalho. No Brasil, essas dinâmicas são marcadas pela heterogeneidade das formas de ocupação e por desigualdades nas condições de trabalho e nos níveis de rendimento, refletindo características estruturais da economia e da sociedade. Nesse contexto, a qualificação e a escolaridade constituem dimensões relevantes, uma vez que podem ampliar as possibilidades de inserção profissional e de acesso a oportunidades de emprego e renda (Rego *et al.*, 2021).

A relação entre educação e trabalho, contudo, é complexa. Conforme apontado por Rego *et al.* (2021), a educação, isoladamente, não é capaz de gerar emprego, uma vez que a inserção laboral depende de mecanismos econômicos e estruturais mais amplos. Ainda assim, a qualificação profissional pode atuar como condicionante da qualidade dessa inserção. Essa perspectiva também encontra respaldo em Frigotto (2007, 2018), ao compreender as relações entre educação e trabalho no interior das estruturas econômicas e sociais, afastando interpretações que atribuam exclusivamente às características individuais dos trabalhadores a responsabilidade por suas condições de inserção profissional.

De maneira complementar, Bourdieu (1998) destaca que a escolaridade e o capital educacional se articulam a outros recursos econômicos, sociais e culturais. Dessa forma, indivíduos com níveis semelhantes de formação podem apresentar condições distintas de inserção e rendimento, uma vez que as oportunidades disponíveis no mercado de trabalho também são influenciadas pelas desigualdades existentes na estrutura social. A escolaridade constitui, portanto, uma dimensão relevante para a compreensão das diferenças observadas no mercado de trabalho, mas seus efeitos devem ser analisados em interação com outros atributos

individuais, sociais e econômicos.

A heterogeneidade das formas de inserção ocupacional constitui uma característica importante do mercado de trabalho brasileiro. Saboia *et al.* (2021), ao analisarem sua evolução com base na PNAD e na PNAD Contínua, demonstram que as condições de ocupação, desemprego, informalidade e rendimento estão relacionadas às transformações da conjuntura econômica. Os autores destacam que o aprofundamento da crise econômica brasileira a partir de 2014 provocou reversões em indicadores que anteriormente apresentavam comportamento favorável, com crescimento da desocupação, da subutilização da força de trabalho e da informalidade.

Nesse sentido, a posição na ocupação constitui uma dimensão relevante para a compreensão das diferentes formas de participação dos trabalhadores na atividade econômica. Empregados com carteira assinada, trabalhadores sem carteira, servidores públicos, trabalhadores domésticos, empregadores e trabalhadores por conta própria apresentam condições distintas quanto à natureza do vínculo, à estabilidade e ao acesso aos mecanismos de proteção social. Saboia *et al.* (2021) ressaltam, entretanto, que a informalidade pode ser definida a partir de diferentes critérios, entre os quais se encontram a posição na ocupação, a contribuição previdenciária e a regularização da atividade econômica.

Essa distinção é particularmente importante para a interpretação do trabalho por conta própria. A categoria reúne situações ocupacionais heterogêneas e, portanto, não deve ser compreendida automaticamente como sinônimo de empreendedorismo, informalidade ou precariedade. A natureza dessa inserção depende de aspectos relacionados às condições de exercício da atividade, aos rendimentos, à proteção social e às características do próprio trabalhador. Dessa maneira, a análise da posição ocupacional deve ser articulada a outras dimensões do mercado de trabalho.

As desigualdades também se manifestam de forma expressiva na distribuição dos rendimentos. A renda proveniente do trabalho representa um componente relevante da renda das famílias brasileiras, fazendo com que alterações nas condições do mercado de trabalho produzam repercussões sobre a própria distribuição de renda. Saboia *et al.* (2021) demonstram que a evolução dos rendimentos está associada tanto à conjuntura econômica quanto a fatores

institucionais, destacando que períodos de expansão das contratações formais, crescimento do rendimento real e valorização do salário mínimo estiveram relacionados à melhoria de indicadores distributivos.

A desigualdade de rendimentos, entretanto, possui caráter multidimensional. Ferreira, Firpo e Messina (2017) identificam que as disparidades salariais no mercado de trabalho brasileiro estão relacionadas a dimensões como gênero, raça, informalidade, localização urbano-rural e escolaridade. Nesse sentido, diferenças de remuneração não devem ser atribuídas a uma única característica, mas compreendidas a partir da interação entre atributos individuais, posição ocupacional, condições institucionais e características estruturais do mercado de trabalho.

Entre essas dimensões, a escolaridade ocupa posição relevante na literatura sobre diferenciais de rendimento. Níveis mais elevados de formação podem estar associados ao desenvolvimento de conhecimentos e competências e ao acesso a determinadas ocupações e posições profissionais. Contudo, conforme discutem Frigotto (2007, 2018) e Bourdieu (1998), a relação entre formação e resultados econômicos não ocorre de maneira automática, pois as possibilidades de inserção profissional são mediadas pelas condições sociais, pelas características da estrutura produtiva e pelas oportunidades disponíveis aos diferentes grupos.

A jornada de trabalho constitui outra dimensão importante para a caracterização da inserção ocupacional. O número de horas trabalhadas está relacionado às diferentes formas de organização do trabalho, aos regimes contratuais e às características das ocupações. Urani (1995), ao analisar as relações entre jornada de trabalho, salários e absorção da mão de obra no mercado de trabalho brasileiro, destaca que a compreensão dessas diferenças requer considerar conjuntamente aspectos como escolaridade, posição na ocupação e remuneração. Dessa forma, jornada e rendimento devem ser analisados conjuntamente com as demais dimensões que caracterizam a inserção do trabalhador.

Nesse contexto, a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) constitui uma importante fonte para o acompanhamento das transformações do mercado de trabalho brasileiro. Sua periodicidade trimestral permite observar alterações conjunturais e sazonais relacionadas à ocupação e aos rendimentos, ampliando as possibilidades de investigação das condições de

inserção laboral. Saboia *et al.* (2021) destacam que essa característica possibilita análises mais detalhadas das mudanças ocorridas ao longo do tempo.

Dessa forma, a análise do mercado de trabalho requer considerar conjuntamente inserção ocupacional, formalização, jornada, escolaridade, rendimento e características sociodemográficas. A utilização dos microdados da PNAD Contínua amplia as possibilidades de investigação dessas dimensões e demanda ferramentas capazes de organizar, tratar e analisar grandes volumes de informações, estabelecendo sua articulação com a Ciência de Dados e a Mineração de Dados.

2.2 Ciência de Dados e Mineração de Dados

A crescente digitalização da sociedade e os avanços das tecnologias da informação têm ampliado significativamente a produção e a disponibilidade de dados, caracterizados por elevado volume, variedade e velocidade, fenômeno frequentemente associado ao conceito de *Big Data*. Nesse contexto, a Ciência de Dados e a Mineração de Dados (*Data Mining*) assumem relevância ao fornecer métodos e técnicas capazes de organizar, explorar e analisar grandes conjuntos de informações, possibilitando a identificação de padrões e a geração de conhecimento a partir dos dados (Lampropoulos, 2021).

A Ciência de Dados possui caráter interdisciplinar, articulando conhecimentos provenientes da computação, da estatística e de diferentes áreas de aplicação. Sua utilização permite investigar fenômenos complexos a partir de dados estruturados e não estruturados, transformando registros brutos em informações passíveis de interpretação. Nesse sentido, sua aplicação mostra-se pertinente à análise do mercado de trabalho, cuja compreensão envolve múltiplas dimensões, como inserção ocupacional, formalização, jornada, escolaridade, rendimento e características sociodemográficas.

A Mineração de Dados, por sua vez, está relacionada ao processo de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases – KDD*), que busca identificar padrões, associações e tendências relevantes em grandes conjuntos de informações. Lampropoulos (2021) destaca que esse processo pode incorporar métodos estatísticos e técnicas de aprendizado de máquina (*machine learning*), possibilitando a exploração sistemática dos dados e a

obtenção de informações que não seriam facilmente identificadas por procedimentos convencionais.

Embora Lampropoulos (2021) desenvolva parte dessa discussão no âmbito da Mineração de Dados Educacionais (*Educational Data Mining* – EDM), os fundamentos da Mineração de Dados possuem aplicação mais ampla e podem ser empregados em diferentes campos do conhecimento. No presente estudo, essa abordagem é direcionada à análise de dados socioeconômicos e do mercado de trabalho, tendo como fonte os microdados da PNAD Contínua. A amplitude dessa base permite estabelecer diferentes recortes e relações entre variáveis, favorecendo a investigação das desigualdades e das formas de inserção ocupacional discutidas na Seção anterior.

A utilização de grandes bases de dados, contudo, não se resume ao acesso às informações. É necessário estabelecer procedimentos que possibilitem transformar os registros disponíveis em dados adequados à análise. Gorard, See e Siddiqui (2020) ressaltam que a simples disponibilidade de dados ou evidências não é suficiente para produzir conhecimento, sendo necessário organizá-los e transformá-los em formatos que favoreçam sua interpretação e utilização. De maneira convergente, Schildkamp (2019) compreende o uso de dados como um processo que envolve coleta, interpretação e produção de sentido, destacando a importância de converter dados brutos em informações capazes de subsidiar análises e decisões.

Essa perspectiva é particularmente relevante na utilização de microdados provenientes de pesquisas oficiais, cuja estrutura pode envolver grande quantidade de registros, diferentes formatos de arquivos e múltiplas variáveis. A transformação dessas informações em uma base analítica exige, portanto, procedimentos computacionais capazes de assegurar sua aquisição, organização e preparação, estabelecendo um fluxo entre a disponibilidade dos dados e sua posterior análise.

Nesse sentido, a Ciência de Dados e a Mineração de Dados fornecem o suporte conceitual e tecnológico para a análise das diferentes dimensões do mercado de trabalho. Sua aplicação envolve etapas sucessivas de obtenção, organização, tratamento e análise das informações. Diante do volume e da diversidade dos dados, a automatização desses processos torna-se relevante para garantir maior eficiência e sistematização, especialmente nas etapas de coleta e

organização dos arquivos, abordadas na Seção seguinte.

2.3 Coleta Automatizada de Dados e Manipulação de Sistemas de Arquivos na Extração de Dados

A etapa inicial de estruturação de grandes bases de dados frequentemente esbarra na dispersão das informações e na lentidão da coleta manual. Para transpor essa barreira, o *Web Scraping* (ou extração de dados da *web*) emerge como uma alternativa para automatização da coleta, permitindo a extração de informações diretamente de páginas online e portais governamentais. Conforme ressalta Farias *et al.* (2024), essa técnica é amplamente adotada para obter dados de forma eficiente, reduzindo a necessidade de intervenção humana constante, o que se torna relevante quando o volume de dados é extenso ou exige atualizações regulares.

No ecossistema da linguagem *Python*, o processo de *web scraping* é orquestrado por bibliotecas nativas e de terceiros. A biblioteca *requests*, por exemplo, é responsável por enviar as requisições *HTTP* aos servidores *web* e capturar as respostas brutas. Em seguida, ferramentas como o *BeautifulSoup* realizam a varredura da estrutura *HTML* ou *XML* recebida, transformando os dados não estruturados em objetos navegáveis e permitindo a extração precisa de *links*, textos e arquivos com base em *tags* e identificadores específicos (Farias *et al.*, 2024).

Após a coleta automatizada, a gestão eficiente do fluxo da informação requer práticas de engenharia de dados, que englobam a aquisição, o armazenamento e a recuperação de arquivos. Na pesquisa científica, os projetos podem enfrentar problemas relacionados à desorganização de arquivos e à existência de diferentes formatos de dados, o que pode dificultar a interpretação, a reutilização e o próprio uso de técnicas de mineração de dados.

Dessa forma, a manipulação via código dos sistemas de arquivos é importante para a criação de rotinas automatizadas que descompacte e estruture os dados brutos. Como demonstrado por Le Piane (2023), o uso de métodos computacionais orientados a dados permite converter grandes volumes de arquivos, como os microdados brutos em formato *Fixed-Width File* (FWF), para formatos padronizados e estruturados, como arquivos *JSON* ou *CSV*. A automação dessa conversão favorece a integração dos dados e sua utilização nas etapas posteriores

de análise computacional (Le Piane, 2023).

2.4 Tratamento de Dados (*Data Wrangling*) e Vetorização

A adequação dos dados extraídos para o consumo analítico é uma das fases mais críticas de projetos em Ciência de Dados. Estudos indicam que o pré-processamento consiste em uma tarefa demorada, reforçando a necessidade de ferramentas eficientes (Kabir *et al.*, 2024). Nesse sentido, Gholizadeh (2022) destaca a biblioteca *pandas* como uma ferramenta rápida, flexível e poderosa desenvolvida em *Python* especificamente para a análise e o tratamento de dados (*data wrangling*).

O *pandas* simplifica o processo de tratamento ao utilizar estruturas bidimensionais chamadas *DataFrames*, que permitem lidar com inconsistências, dados ausentes (*NaN*), filtragens e transformações de colunas com eficiência em grandes conjuntos de dados. Como apontam Kabir *et al.* (2024), a limpeza adequada contribui para maior consistência dos resultados das estatísticas descritivas e dos modelos analíticos, reduzindo problemas decorrentes de inconsistências presentes nos dados originais.

Complementarmente, a biblioteca *NumPy* introduz cálculos numéricos de alto desempenho, permitindo a vetorização de operações matemáticas (Gholizadeh, 2022). Em vez de percorrer grandes conjuntos de dados iterativamente, o que pode apresentar maior custo computacional, a vetorização permite aplicar operações lógicas e aritméticas simultaneamente a matrizes numéricas. A combinação de *pandas* para estruturação dos dados e *NumPy* para cálculo matricial é descrita na literatura como uma base importante para a análise de dados na linguagem *Python* (McKinney, 2011; Kabir *et al.*, 2024).

2.5 Desenvolvimento Rápido de Aplicações Web Orientadas a Dados

O ciclo de vida dos dados também envolve a etapa de comunicação dos resultados. A visualização de dados transcende a estética, atuando como mecanismo para tornar um grande volume de informações compreensível para usuários técnicos e leigos, permitindo a identificação de padrões e o suporte à tomada de decisão (Kabir *et al.*, 2024). A literatura enfatiza que o ecossistema *Python* conta com bibliotecas robustas para a criação de representações visuais,

desde as mais estáticas até gráficos interativos avançados (Saabith *et al.*, 2021).

Historicamente, construir painéis para apresentar essas visualizações exigia equipes multidisciplinares com conhecimento em desenvolvimento *web*. Atualmente, o uso de ambientes e bibliotecas orientadas a interfaces gráficas do usuário (*GUI*) favorece o desenvolvimento de Aplicações *Web* Orientadas a Dados (*Data Apps*) (Saabith *et al.*, 2021).

2.6 Visualização de Dados e Ferramentas de Monitoramento

Para que a mineração e o monitoramento de dados contribuam para a compreensão das características e desigualdades do mercado de trabalho, a visualização de dados desempenha papel relevante ao facilitar a identificação e a interpretação de padrões, tendências e relações presentes nas informações analisadas.

Ali e Vidya Sagar (2023) definem a visualização de dados como a representação gráfica de informações utilizando elementos visuais como gráficos e mapas. Essa prática é essencial para revelar padrões, tendências e correlações que seriam difíceis de detectar em planilhas ou textos brutos, facilitando a comunicação de *insights* complexos para diferentes públicos.

Segundo a revisão realizada por Ali e Vidya Sagar (2023), existem diversas ferramentas com características distintas que podem ser aplicadas à análise e visualização de dados:

- *Plotly*: para o desenvolvimento de ferramentas de monitoramento modernas, a interatividade é fundamental. Ali e Vidya Sagar (2023) destacam o *Plotly* como uma biblioteca poderosa para criar visualizações interativas baseadas na *web*. Ela permite que os usuários explorem os dados por meio de recursos como *zoom* e seleção em tempo real.
- *Streamlit*: associado a essas ferramentas, o *Streamlit*, mencionado no contexto de *frameworks web* para *Python*, permite a criação de *dashboards* analíticos completos sem a necessidade de conhecimento avançado em *JavaScript*, entregando o que está sendo criado diretamente na rede, através da captura de mudanças realizadas no repositório, de maneira automática, em tempo real.

O uso de *Plotly* e *Streamlit* mostra-se adequado a projetos que exigem o

compartilhamento de gráficos interativos e painéis de controle para monitoramento de indicadores (Ali e Vidya Sagar, 2023).

A aplicação dessas ferramentas na análise de dados socioeconômicos permite apresentar indicadores de maneira intuitiva e interativa, favorecendo a identificação de tendências e padrões e facilitando a exploração das informações por diferentes usuários. No contexto do mercado de trabalho, esses recursos podem contribuir para a visualização e o monitoramento de indicadores relacionados à ocupação, escolaridade, rendimento e demais características analisadas (Ali e Vidya Sagar, 2023).

2.7 Trabalhos relacionados

A revisão da literatura e do estado da arte é fundamental para situar a presente pesquisa em relação às soluções tecnológicas e investigações socioeconômicas já desenvolvidas. Para esta Seção, foram selecionados trabalhos que convergem com este estudo em dois eixos principais: o uso da Ciência de Dados e painéis interativos (eixo tecnológico) e a análise de bases governamentais aplicadas ao mercado de trabalho e às desigualdades socioeconômicas (eixo socioeconômico), conforme demonstrado na Tabela 1.

Tabela 1: Comparação entre os trabalhos relacionados e a proposta desta pesquisa

Trabalho	Base de dados PNAD	Engenharia de dados	Dashboard/ Painel interativo	Foco em desigualdade e renda
Bueno (2023)	X	✓	✓	✓
Santos (2023)	X	X	✓	X
Vieira e Souza (2025)	✓	X	✓	X
O presente trabalho (2026)	✓	✓	✓	✓

Fonte: Elaborado pelos autores, 2026.

No contexto do uso de dados governamentais para a tomada de decisão, Vieira e Souza (2025) analisaram a aplicação de um *dashboard* construído com dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua) pela Secretaria de Planejamento do Estado de Pernambuco (SEPLAG). Embora o

trabalho compartilhe com esta pesquisa a mesma base de dados (PNAD) e o mesmo recorte territorial (Pernambuco), o estudo de Vieira e Souza (2025) concentrou-se estritamente na avaliação heurística de *design* da informação e usabilidade da interface desenvolvida em *Power BI*. Diferentemente, o presente trabalho foca na Engenharia de Dados *end-to-end*, propondo a construção de um *pipeline* em *Python* para o tratamento automatizado dos microdados e a construção de um painel interativo nativo na *web*.

Abordando a intersecção entre tecnologia e disparidades de renda, Bueno (2023) desenvolveu um sistema automatizado para extração e processamento de dados com o intuito de investigar as desigualdades salariais de gênero no mercado de trabalho brasileiro entre 2012 e 2021. O trabalho assemelha-se a este estudo pela adoção da linguagem *Python* no processo de *Extract, Transform, Load* (ETL). Entretanto, a pesquisa de Bueno (2023) baseou-se nos registros administrativos da RAIS (Relação Anual de Informações Sociais) e utilizou o *software Power BI* para a visualização, enquanto este trabalho processa os microdados da PNAD Contínua, que possibilitam analisar diferentes formas de inserção no mercado de trabalho, e desenvolve uma Aplicação *Web* Orientada a Dados (*Data App*).

Por fim, estreitando o foco para a gestão educacional local, Santos (2023) desenvolveu painéis de Inteligência de Negócios utilizando dados do próprio Instituto Federal de Pernambuco (IFPE) – Campus Paulista. A autora empregou técnicas de automação para extrair informações de planilhas de desempenho, consolidando-as em *dashboards* para apoiar a diretoria de ensino. Embora compartilhe o mesmo ambiente institucional de aplicação, o trabalho de Santos (2023) restringiu-se ao tratamento de dados internos de desempenho acadêmico, enquanto o presente estudo direciona a aplicação das técnicas de Ciência de Dados para uma base governamental de abrangência populacional voltada à investigação do mercado de trabalho.

A análise dos trabalhos relacionados evidencia a utilização de diferentes estratégias de Ciência de Dados para aquisição, tratamento e visualização de informações provenientes de bases governamentais, ao mesmo tempo em que permite identificar as particularidades da abordagem adotada nesta pesquisa. Sendo assim, a partir das informações supracitadas, a Seção seguinte descreve os procedimentos metodológicos empregados no desenvolvimento do estudo,

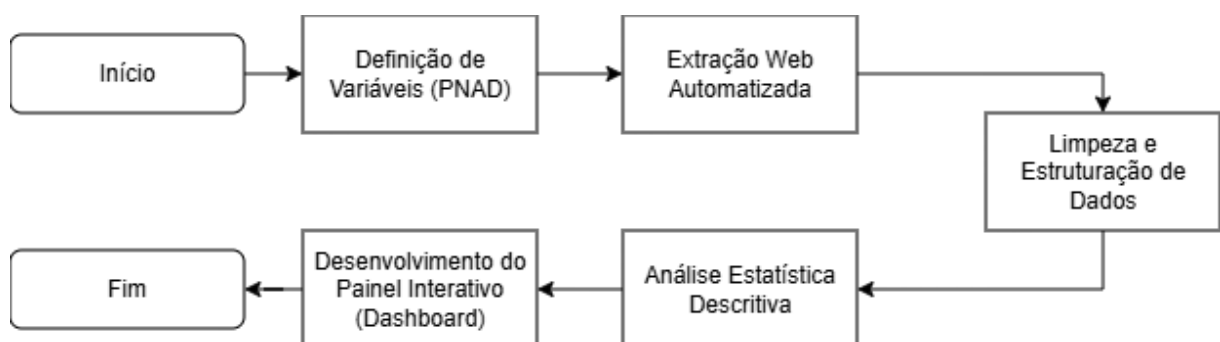
contemplando a fonte dos dados, as ferramentas utilizadas e as etapas de aquisição, tratamento, análise e visualização dos microdados da PNAD Contínua.

3 METODOLOGIA

A pesquisa adotou uma abordagem quantitativa, de natureza aplicada e caráter exploratório-descritivo, estruturada na análise de dados socioeconômicos e laborais, com foco nas características e desigualdades do mercado de trabalho do estado de Pernambuco. A natureza aplicada decorre do desenvolvimento de uma solução computacional destinada à aquisição, ao tratamento, à análise e à visualização de dados, enquanto o caráter exploratório-descritivo está relacionado à identificação e à caracterização de padrões, distribuições e associações entre as variáveis investigadas.

O procedimento metodológico foi estruturado em um fluxo contínuo de processamento de dados (*pipeline*), compreendendo desde a obtenção dos microdados até a disponibilização das informações em uma aplicação *web* interativa. A Figura 1 sintetiza esse procedimento e seus respectivos insumos e produtos, organizados em cinco etapas: (1) definição das variáveis e aquisição dos dados; (2) coleta automatizada e processamento bruto; (3) tratamento e integração dos dados; (4) análise exploratória; e (5) visualização dos dados e desenvolvimento do painel interativo.

Figura 1: Fluxograma da metodologia utilizada no estudo.



Fonte: Elaborado pelos autores.

Na primeira etapa, foram realizadas a definição das variáveis e a aquisição da base de dados. O estudo utilizou exclusivamente os microdados trimestrais da

Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), disponibilizados pelo Instituto Brasileiro de Geografia e Estatística (IBGE). A escolha dessa fonte decorreu de sua abrangência e da disponibilidade de informações relacionadas às características sociodemográficas e às condições de inserção no mercado de trabalho, possibilitando analisar dimensões como ocupação, posição ocupacional, jornada de trabalho, escolaridade e rendimento.

Por se tratar de uma pesquisa amostral domiciliar, os registros presentes nos microdados da PNAD Contínua correspondem a observações da amostra e não devem ser interpretados diretamente como quantitativos da população. Essa característica foi considerada tanto na estruturação da base quanto na interpretação dos resultados, especialmente na distinção entre as frequências de registros observadas nos microdados e os indicadores para os quais foram aplicados os pesos amostrais disponibilizados pela pesquisa.

A delimitação territorial foi realizada mediante a seleção dos registros cuja variável correspondente à Unidade da Federação (UF) identifica Pernambuco (código 26). O dicionário de variáveis e a documentação técnica disponibilizados juntamente com os microdados orientaram a seleção e a interpretação das variáveis utilizadas. Foram contempladas características sociodemográficas, como idade, sexo, cor/raça e nível de instrução, e características laborais, como posição na ocupação, jornada de trabalho e rendimento, em consonância com as dimensões discutidas no referencial teórico.

Na segunda etapa, foi realizada a coleta automatizada e o processamento bruto dos dados. A extração e a organização dos arquivos ocorreram em ambiente computacional utilizando a linguagem *Python*. Para automatizar a aquisição dos arquivos, foram utilizadas as bibliotecas *requests* e *BeautifulSoup*, empregadas na comunicação *HTTP* e na identificação dos recursos disponibilizados no ambiente *web* do IBGE. Esse procedimento permitiu sistematizar a obtenção dos arquivos correspondentes aos períodos selecionados e de seus respectivos documentos auxiliares.

Na sequência, o gerenciamento dos diretórios locais e a extração dos arquivos compactados foram realizados por meio das bibliotecas nativas *os* e *zipfile*. A automatização dessas tarefas buscou reduzir procedimentos manuais repetitivos e manter um fluxo padronizado para o processamento dos diferentes arquivos

trimestrais utilizados na pesquisa.

A terceira etapa consistiu no tratamento e na integração dos dados. Os arquivos brutos da PNAD Contínua, disponibilizados em estrutura de largura fixa (FWF), foram lidos e convertidos em estruturas tabulares organizadas em *DataFrames*, utilizando principalmente as bibliotecas *pandas* e *NumPy*. A interpretação desses arquivos considerou a estrutura e as especificações das variáveis estabelecidas na documentação da pesquisa.

O tratamento contemplou procedimentos de seleção das variáveis de interesse, conversão de tipagens, padronização das estruturas, filtragem territorial para Pernambuco e organização dos dados para utilização nas etapas analíticas subsequentes. Os valores ausentes foram considerados de acordo com a natureza e a aplicabilidade de cada variável, uma vez que, em pesquisas estruturadas por questionários, a ausência de preenchimento não representa necessariamente erro ou perda de informação, podendo decorrer dos próprios critérios de aplicação das perguntas.

Após o tratamento dos arquivos correspondentes aos diferentes períodos, os registros foram consolidados em uma base estruturada para fins de análise. Essa consolidação reuniu as observações disponíveis nos diferentes trimestres compreendidos entre 2019 e 2025 e não foi interpretada como acompanhamento longitudinal dos mesmos indivíduos ao longo do período. Dessa forma, os quantitativos apresentados como registros ou observações da base representam ocorrências nos microdados processados, e não necessariamente indivíduos únicos acompanhados durante todo o intervalo temporal. Ao final dessa etapa, a base processada foi exportada em formato .csv, possibilitando sua utilização nas análises e na aplicação desenvolvida.

A quarta etapa consistiu na análise exploratória dos dados. Sobre a base consolidada foram aplicadas técnicas de estatística descritiva e exploratória, incluindo distribuições de frequência, medidas de tendência central e dispersão, cruzamentos entre variáveis e medidas de associação, de acordo com as características das informações investigadas. Essas análises tiveram como finalidade identificar e descrever padrões relacionados às características sociodemográficas, à inserção ocupacional, à escolaridade, à jornada de trabalho e aos rendimentos observados no mercado de trabalho pernambucano.

Considerando a natureza amostral da PNAD Contínua, foi utilizada a variável V1028, correspondente ao peso amostral disponibilizado nos microdados, nos cálculos dos indicadores da aplicação que requerem ponderação. A utilização desse peso permitiu considerar a representatividade atribuída a cada observação da amostra na produção das estimativas ponderadas.

Desse modo, foram diferenciadas as frequências de registros da base, que correspondem às observações efetivamente processadas, das estimativas ponderadas, calculadas mediante a aplicação da V1028. Assim, essa distinção foi considerada na apresentação e na interpretação dos resultados, evitando que as contagens observadas na amostra fossem diretamente interpretadas como quantitativos da população pernambucana.

Cabe destacar que as análises desenvolvidas possuem natureza predominantemente descritiva e exploratória. Dessa forma, diferenças, padrões e associações identificados entre escolaridade, idade, cor/raça, posição ocupacional, jornada de trabalho e rendimento não foram interpretados como evidências de relações de causa e efeito.

A quinta e última etapa consistiu na visualização dos dados e no desenvolvimento de um painel interativo (*Data App*). Os dados processados e os indicadores gerados nas etapas anteriores foram integrados a uma aplicação *web* desenvolvida em *Python*, utilizando o *framework Streamlit* para construção da interface e a biblioteca *Plotly* para elaboração das representações gráficas interativas.

A aplicação foi estruturada de modo a estabelecer uma interface entre os dados processados e o usuário final, possibilitando a exploração dos indicadores por meio de elementos visuais e filtros definidos no sistema. A integração entre as rotinas de processamento e a interface permitiu que as informações fossem apresentadas de acordo com os parâmetros selecionados, sem a necessidade de manipulação direta dos arquivos de microdados pelo usuário. O *Data App* foi disponibilizado *on-line* por meio da plataforma do *Streamlit*¹.

A sistematização das cinco etapas possibilitou integrar aquisição, organização, tratamento, análise e visualização em um mesmo fluxo computacional. A documentação das rotinas desenvolvidas e o versionamento dos *scripts* foram

¹ <https://pnad-dashboard.streamlit.app>

mantidos em repositório no *GitHub*², contribuindo para a organização e a rastreabilidade dos procedimentos computacionais empregados no desenvolvimento da solução.

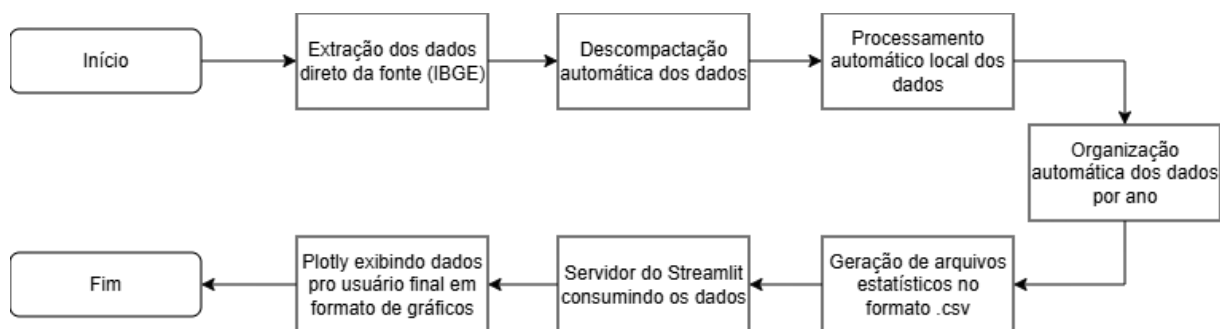
O recorte temporal compreendeu o período de 2019 a 2025, contemplando diferentes momentos da dinâmica do mercado de trabalho pernambucano. Essa delimitação permitiu explorar as características e associações existentes entre variáveis como escolaridade, cor/raça, faixa etária, posição ocupacional, jornada de trabalho e rendimento no período analisado, respeitando o caráter descritivo e exploratório da investigação.

A Figura 1 apresenta, portanto, a visão geral do percurso metodológico adotado, evidenciando a articulação entre as etapas de obtenção dos microdados, processamento computacional, análise e disponibilização das informações. Essa estrutura permite compreender não apenas as técnicas empregadas, mas também a sequência lógica entre os dados de entrada, os procedimentos executados e os produtos resultantes de cada etapa.

Diferentemente, a Figura 2 detalha o fluxo de processamento do sistema desenvolvido, evidenciando o percurso computacional realizado pelos dados. O diagrama representa o fluxo iniciado com a obtenção dos microdados disponibilizados pelo IBGE, seguido pela descompactação dos arquivos e pelo tratamento automatizado por meio de *scripts* desenvolvidos em *Python*, com utilização de *pandas* e *NumPy*. Após o processamento, os dados estruturados são armazenados e disponibilizados para consumo pela aplicação *web* desenvolvida em *Streamlit*, responsável pela apresentação dos indicadores e das visualizações interativas ao usuário final.

Figura 2: Fluxograma do funcionamento do sistema.

² <https://github.com/rodrigosenal7/tcc>



Fonte: Elaborado pelos autores.

Essa distinção entre as duas representações é relevante: a Figura 1 sintetiza o método da pesquisa e suas etapas, enquanto a Figura 2 representa a arquitetura e o fluxo operacional da solução computacional desenvolvida. Dessa forma, as figuras se complementam ao apresentar, respectivamente, a perspectiva metodológica da investigação e a perspectiva tecnológica do sistema implementado.

Dessa forma, o percurso metodológico integrou as etapas de aquisição, tratamento, integração, análise e visualização dos microdados da PNAD Contínua em um fluxo computacional sistematizado. A aplicação dos procedimentos de Ciência de Dados e de análise exploratória possibilitou estruturar as informações necessárias à investigação das características e desigualdades do mercado de trabalho pernambucano. Com base nesse percurso, a Seção seguinte apresenta e discute os resultados obtidos a partir dos dados processados e das variáveis analisadas.

4 RESULTADOS E DISCUSSÕES

4.1 Consolidação e Limpeza da Base de Dados

O processamento automatizado dos microdados trimestrais da PNAD Contínua, delimitado para o estado de Pernambuco no período de 2019 a 2025, resultou na consolidação de uma base de dados estruturada. A extração e o agrupamento dos arquivos geraram um banco final composto por 530.602 registros amostrais e 19 variáveis analíticas, abrangendo informações demográficas, educacionais e relacionadas ao mercado de trabalho.

Durante a etapa de inspeção da estrutura dos dados, observou-se a incidência de valores ausentes, representados computacionalmente como *NaN* (*Not a Number*), em diversas colunas. Variáveis relacionadas à dinâmica laboral, a

exemplo da condição de servidor público, posse de carteira assinada, rendimento mensal e jornada de trabalho, apresentaram quantidade expressiva de valores não preenchidos. Comportamento semelhante foi observado em variáveis educacionais, como a conclusão do curso frequentado.

A consulta ao dicionário de variáveis disponibilizado pelo IBGE indicou, contudo, que parte significativa dessas ocorrências não correspondia necessariamente a falhas de coleta ou perdas informacionais, mas às próprias condições de aplicabilidade das perguntas previstas no questionário. Como exemplo, informações sobre rendimento e horas trabalhadas não se aplicam a determinados indivíduos fora da condição ocupacional correspondente. As variáveis estruturais e demográficas básicas, como ano, trimestre, peso amostral, sexo, idade e cor/raça, apresentaram preenchimento completo para os registros processados.

A identificação desse padrão orientou os procedimentos de tratamento dos dados (*Data Cleaning*). Nas variáveis categóricas, os valores cuja ausência correspondia à não aplicabilidade foram identificados de forma a evitar a exclusão indevida dos registros. Nas variáveis contínuas, como rendimento bruto e horas semanais trabalhadas, os valores não aplicáveis foram mantidos como *NaN*, evitando sua incorporação indevida aos cálculos das medidas descritivas. Dessa forma, os 530.602 registros foram preservados na base consolidada, respeitando as condições específicas de preenchimento das variáveis analisadas.

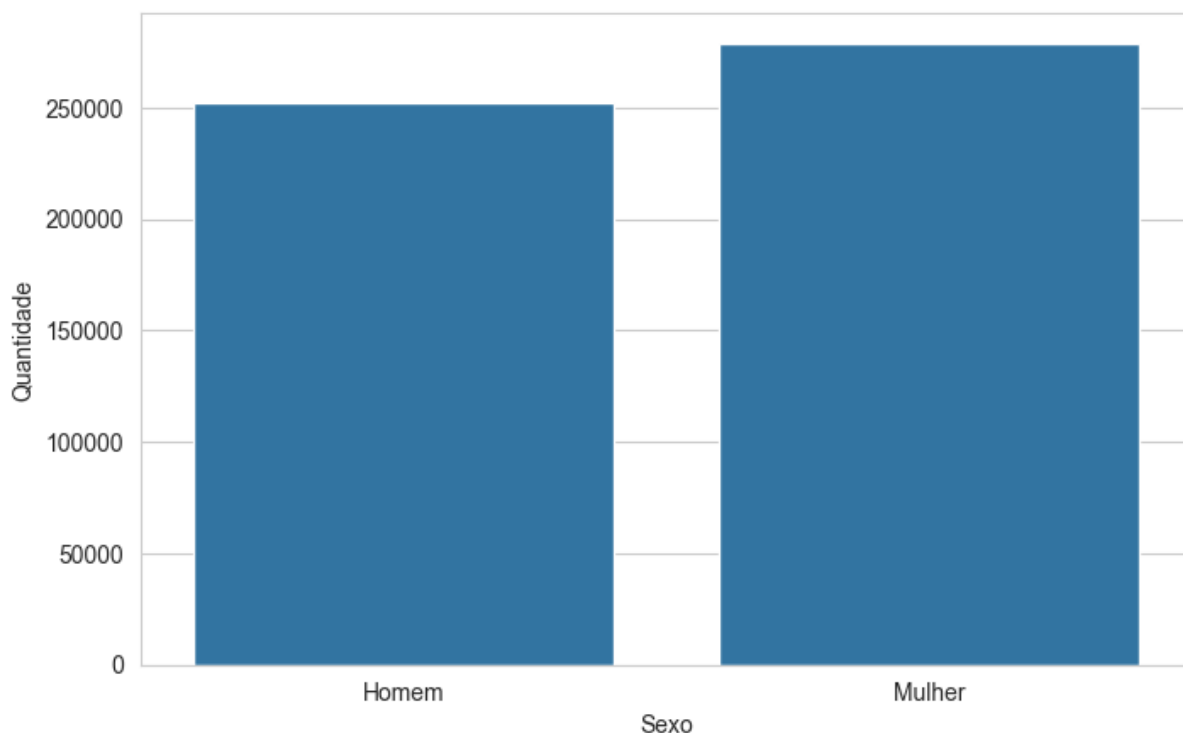
A consolidação dos diferentes arquivos trimestrais permitiu, portanto, construir uma base integrada para a execução das análises subsequentes. Conforme estabelecido na metodologia, esses registros representam observações provenientes dos diferentes períodos compreendidos entre 2019 e 2025, não sendo interpretados como indivíduos únicos acompanhados longitudinalmente.

4.2 Perfil Sociodemográfico e Educacional

A caracterização sociodemográfica dos registros analisados evidencia um leve predomínio de mulheres, que correspondem a 52,5% (278.686) da base, ante 47,5% (251.916) de homens, conforme ilustrado na Figura 3. Esses valores representam a distribuição dos registros processados no conjunto consolidado e devem ser diferenciados das estimativas populacionais obtidas mediante

ponderação.

Figura 3: Gráfico de distribuição do número de pessoas da amostra segundo o sexo, Pernambuco, 2019–2025.

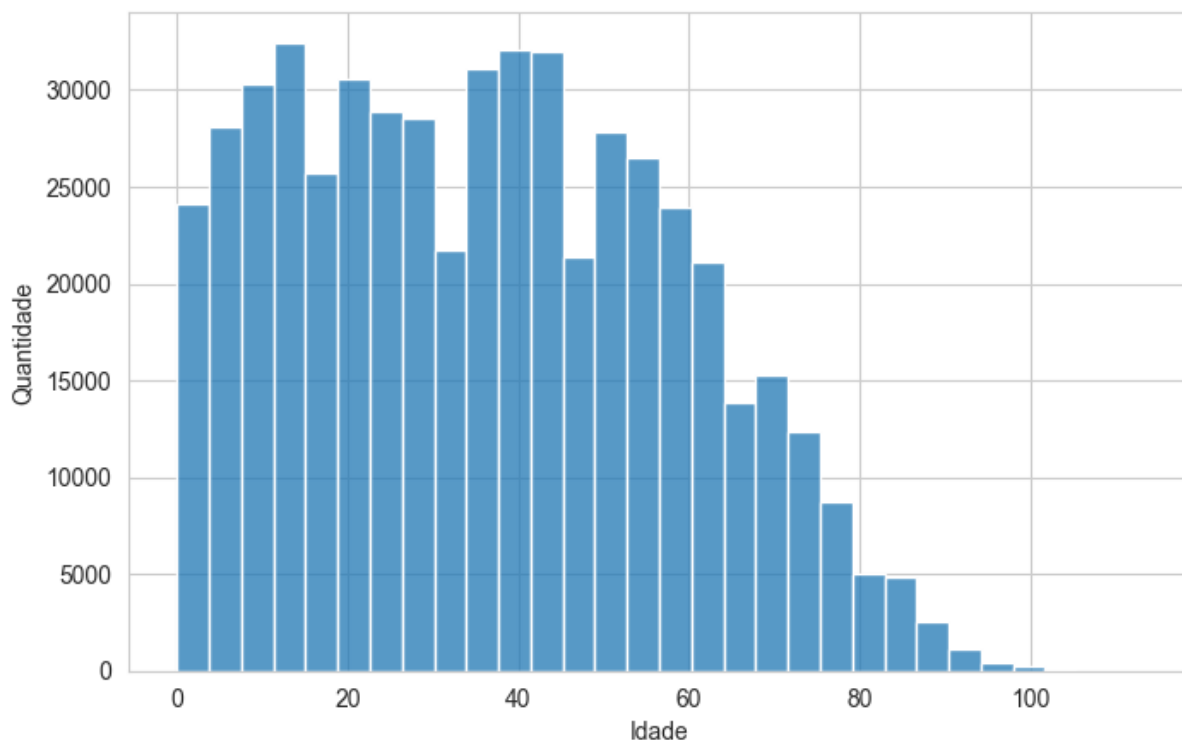


Fonte: Elaborado pelos autores.

Em relação à faixa etária, a base abrange registros de indivíduos de 0 a 113 anos, apresentando idade média de 36 anos e mediana de 35 anos. Conforme demonstrado na Figura 4, a análise quartílica indica que 75% dos registros apresentam idade de até 53 anos, evidenciando maior concentração da distribuição nas faixas etárias mais jovens e intermediárias.

Figura 4: Gráfico de distribuição do número de pessoas da amostra segundo a

idade, Pernambuco, 2019–2025.

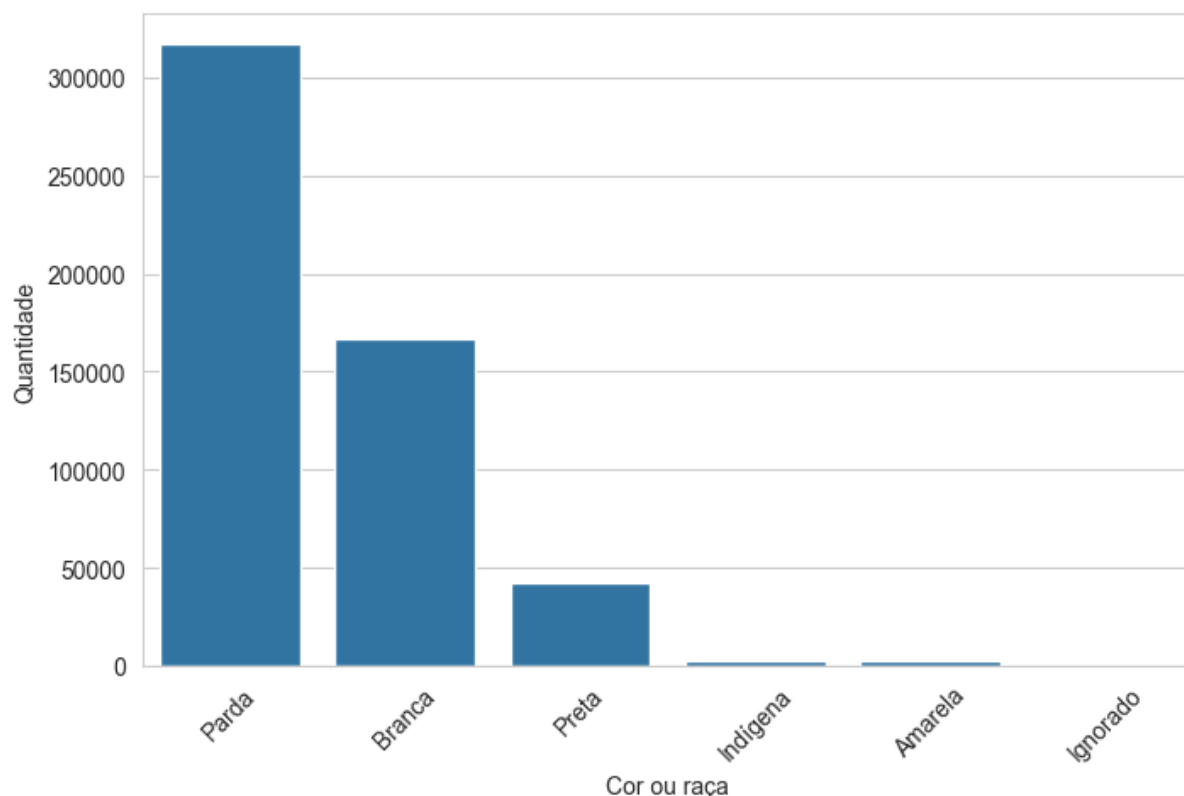


Fonte: Elaborado pelos autores.

Quanto à composição étnico-racial, observa-se maior frequência de registros de indivíduos autodeclarados pardos, totalizando 316.789 observações. Em seguida, encontram-se os registros de indivíduos brancos (166.498) e pretos (42.552). As categorias indígena (2.487) e amarela (2.224) apresentaram menores frequências na base analisada, enquanto 52 registros não possuíam informação de cor/raça. A Figura 5 apresenta visualmente essa distribuição.

Figura 5: Gráfico de distribuição do número de pessoas da amostra segundo a

categoria de cor ou raça declarada, Pernambuco, 2019–2025.

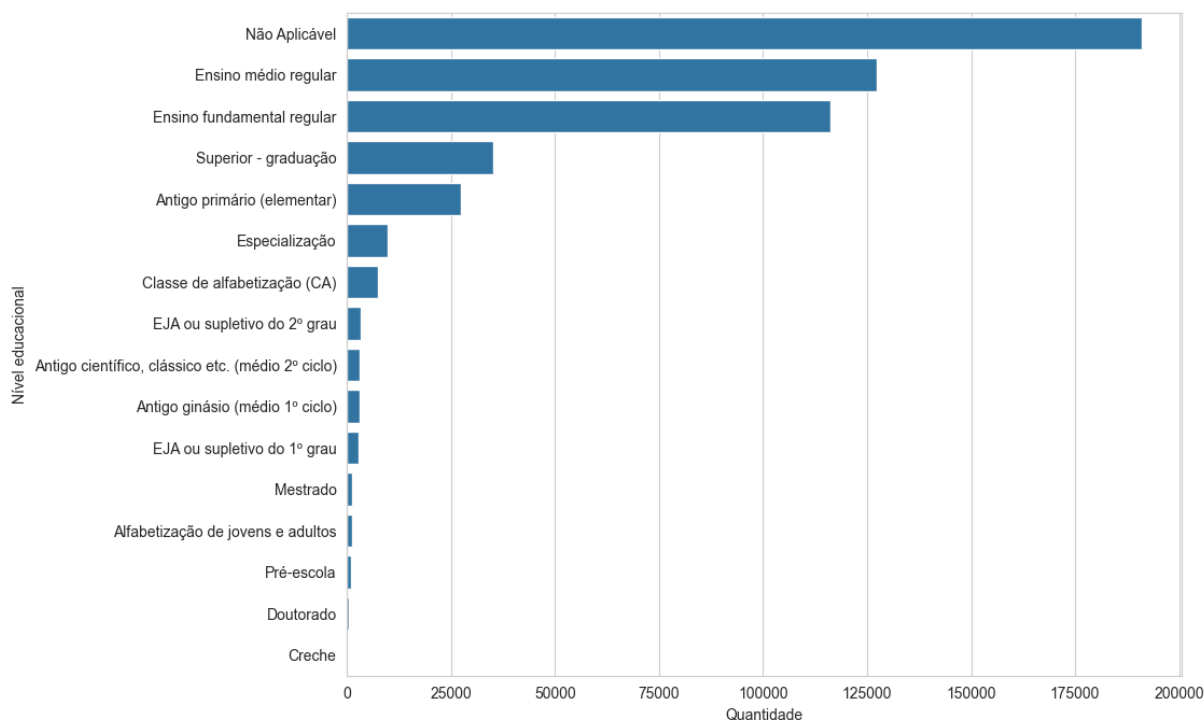


Fonte: Elaborado pelos autores.

No aspecto educacional, os registros apresentam maior concentração nos níveis correspondentes à educação básica. A Figura 6 demonstra que o ensino médio regular agrupa o maior número de observações, com 127.351 registros, seguido pelo ensino fundamental regular, com 116.177. A base contabilizou ainda 35.028 registros associados à graduação, além de 9.833 relacionados à especialização, 1.263 ao mestrado e 546 ao doutorado. Esses dados evidenciam redução progressiva das frequências observadas à medida que se avançam os níveis de escolaridade.

Figura 6: Gráfico de distribuição do número de pessoas da amostra segundo o nível

de escolaridade, Pernambuco, 2019–2025.



Fonte: Elaborado pelos autores.

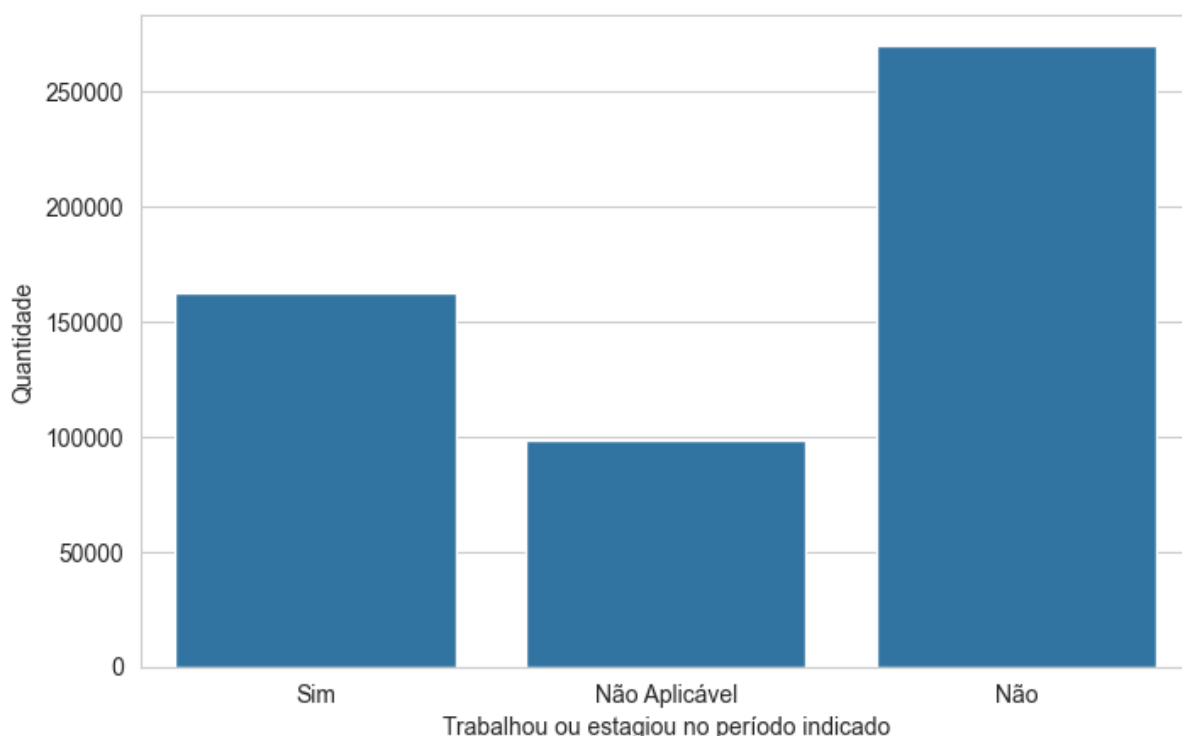
Os resultados sociodemográficos e educacionais fornecem, assim, uma caracterização inicial da base consolidada e permitem contextualizar as análises sobre as diferentes formas de inserção ocupacional apresentadas na Seção seguinte.

4.3 Estrutura do Mercado de Trabalho em Pernambuco

A análise da condição de ocupação da amostra revela a dinâmica de participação da população pernambucana na força de trabalho. Conforme ilustrado pela distribuição na Figura 7, os dados demonstram que, no período avaliado, 162.235 indivíduos declararam ter exercido alguma atividade ou estágio remunerado por pelo menos uma hora na semana de referência. Em contrapartida, a maior parcela da população da base (269.829 registros) foi classificada como não ocupada, o que engloba tanto pessoas desempregadas em busca de oportunidades quanto indivíduos fora da força de trabalho (como estudantes em dedicação exclusiva, aposentados e donas de casa).

Figura 7: Gráfico de distribuição do número de pessoas da amostra segundo a

condição de ocupação no período de referência, Pernambuco, 2019–2025.



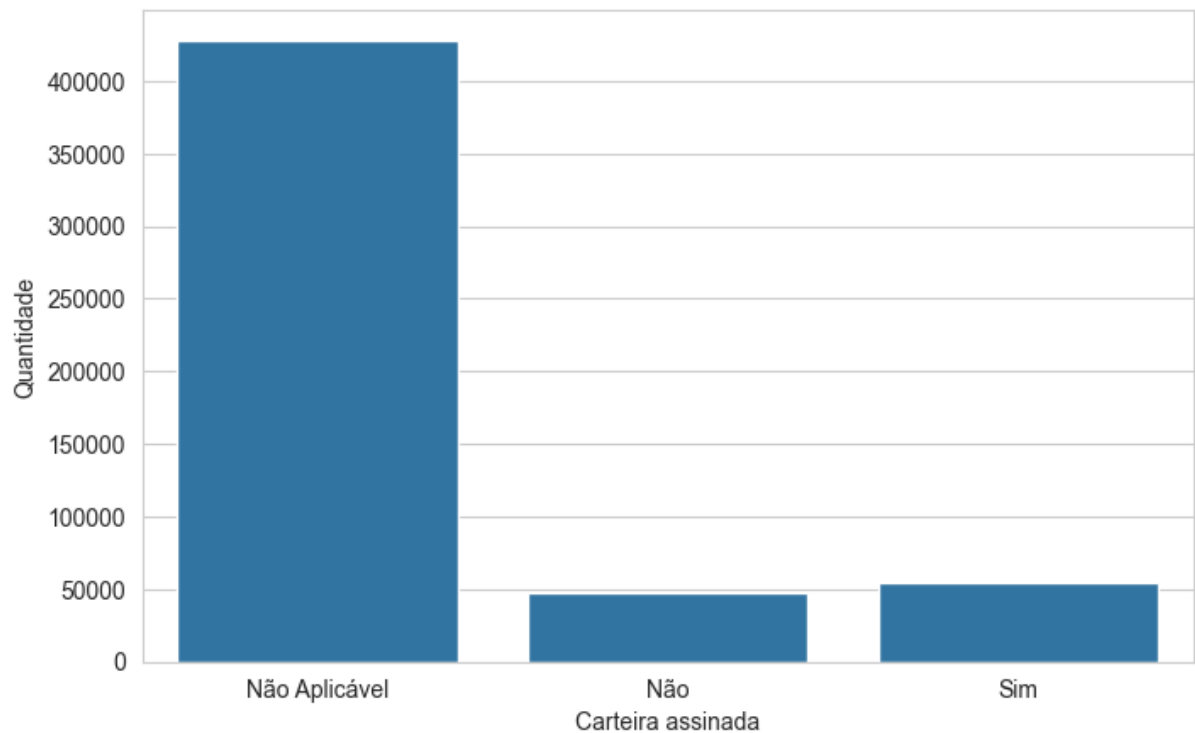
Fonte: Elaborado pelos autores.

A expressiva presença do trabalho por conta própria deve, entretanto, ser interpretada com cautela. Conforme discutido no referencial teórico, essa categoria reúne formas heterogêneas de inserção e não deve ser automaticamente considerada sinônimo de empreendedorismo, informalidade ou precariedade. Sua interpretação requer considerar conjuntamente aspectos relacionados à proteção social, aos rendimentos e às demais condições de exercício da atividade, conforme destacado por Saboia *et al.* (2021).

A dimensão da formalização pode ser complementada pela análise da posse de carteira de trabalho. Considerando o subconjunto de registros para o qual essa variável é aplicável, os resultados apresentados na Figura 8 mostram 54.740 registros com carteira assinada e 48.013 registros sem formalização do vínculo por meio de carteira.

Figura 8: Gráfico de distribuição do número de trabalhadores segundo a condição

de posse de carteira de trabalho assinada, Pernambuco, 2019–2025.

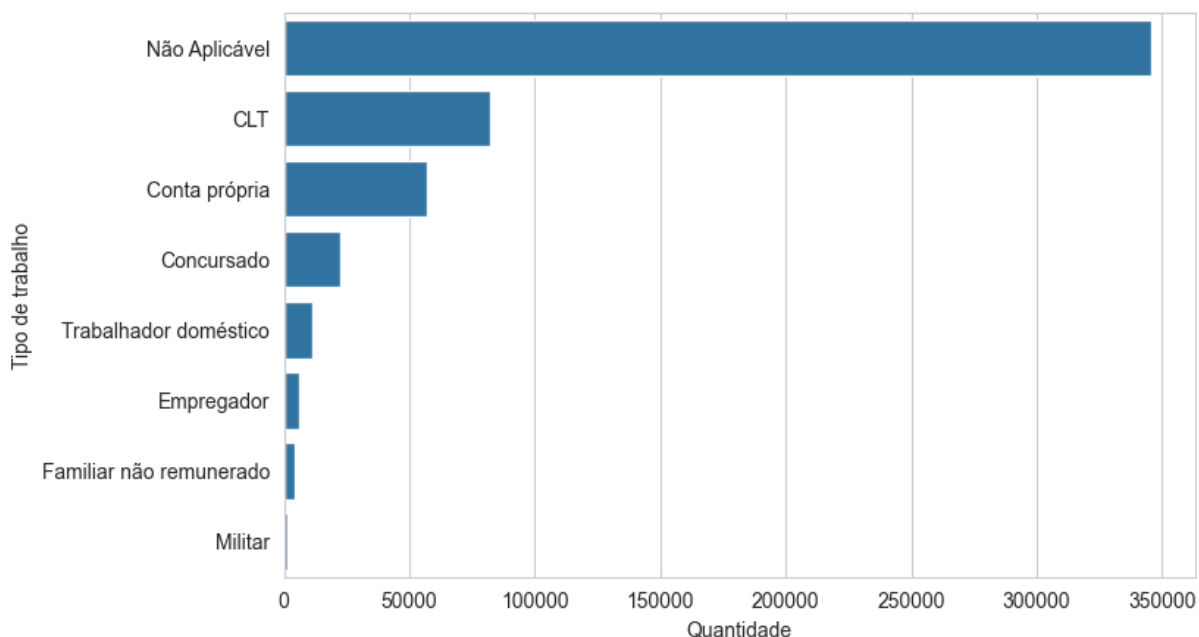


Fonte: Elaborado pelos autores.

A proximidade entre esses quantitativos evidencia a presença expressiva de vínculos sem carteira assinada no subconjunto analisado, aspecto relevante para a compreensão das diferentes condições de formalização do mercado de trabalho pernambucano, mostrado na Figura 9.

Figura 9: Gráfico de distribuição do número de pessoas ocupadas segundo a

posição na ocupação no trabalho principal, Pernambuco, 2019–2025.

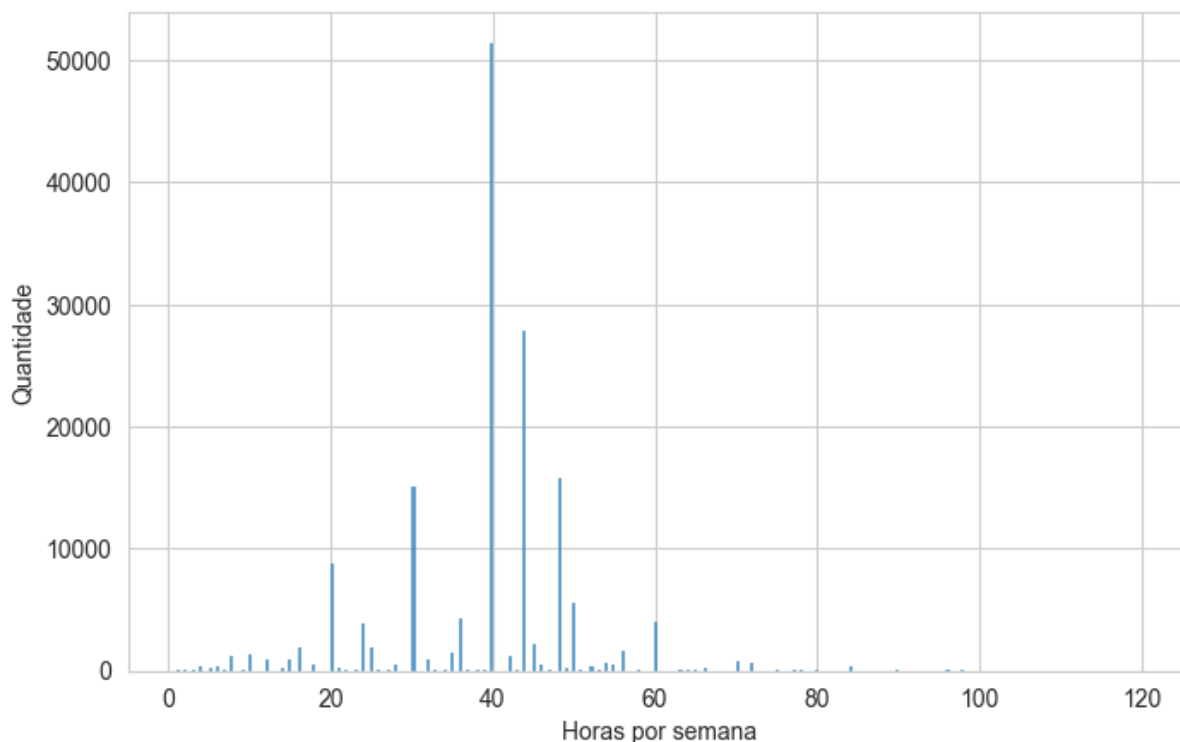


Fonte: Elaborado pelos autores.

No que se refere à jornada semanal de trabalho, a média observada foi de aproximadamente 39 horas, enquanto a mediana correspondeu a 40 horas. A análise da distribuição aponta ainda que 75% dos registros válidos para essa variável apresentam jornadas de até 44 horas semanais. O histograma apresentado na Figura 10 evidencia concentrações nas faixas de 40 e 44 horas. Também são observadas concentrações secundárias em 20 e 30 horas, indicando a ocorrência de jornadas inferiores ao padrão predominante, enquanto registros de 48 horas ou mais revelam a presença de jornadas mais extensas.

Figura 10: Gráfico de distribuição do número de pessoas ocupadas segundo as

horas habitualmente trabalhadas por semana, Pernambuco, 2019–2025.



Fonte: Elaborado pelos autores.

A heterogeneidade observada nas formas de ocupação, na formalização dos vínculos e nas jornadas de trabalho reforça a necessidade de analisar conjuntamente diferentes dimensões da inserção ocupacional. Conforme discutido por Saboia *et al.* (2021), as condições de participação no mercado de trabalho não se restringem à situação de ocupação, envolvendo também características relacionadas ao vínculo, à proteção social e aos rendimentos.

4.4 Desigualdades de Renda e o Retorno da Escolaridade

O ponto analítico principal deste estudo reside na investigação da distribuição de renda e em seus principais condicionantes. A análise estatística da variável de rendimento bruto mensal revelou uma profunda assimetria no mercado de trabalho pernambucano, conforme mostrado na Tabela 2. Enquanto a renda média estimada foi de R\$1.878,45, a mediana situou-se em apenas R\$1.212,00. Essa grande disparidade (onde 75% da população ganha até R\$1.900,00) indica que a maior parcela da população ocupada aufer remunerações muito próximas ao salário

mínimo, mas a média estadual é puxada para cima por *outliers*, uma minoria de indivíduos que concentra ganhos extremos, alcançando picos de até R\$200.000,00 mensais.

Tabela 2: Estatísticas descritivas da variável renda.

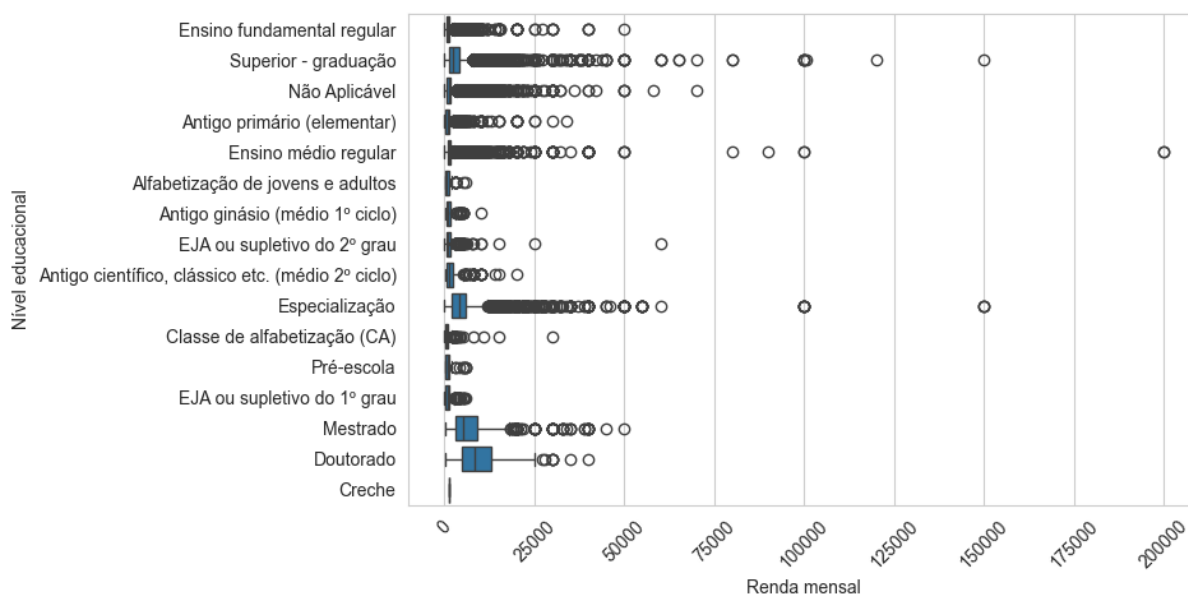
Métrica	Valor
Média	1878.44
Desvio-padrão	3138.83
Mínimo	4.00
25%	700.00
50%	1212.00
75%	1900.00
Máximo	200000.00

Fonte: Elaborado pelos autores.

A escolaridade apresentou-se como uma dimensão relevante na diferenciação dos rendimentos dos trabalhadores analisados. A visualização da renda segundo os níveis de instrução, apresentada no *boxplot* da Figura 11, evidencia uma associação positiva entre escolaridade e rendimento. Trabalhadores com ensino fundamental ou inseridos em modalidades de alfabetização de jovens e adultos apresentam remunerações médias entre R\$ 700,00 e R\$ 1.000,00, enquanto aqueles com ensino médio alcançam aproximadamente R\$ 1.500,00. As diferenças tornam-se mais expressivas entre os trabalhadores com ensino superior, cujos rendimentos ultrapassam R\$ 3.500,00. Ressalta-se que esses resultados indicam que níveis mais elevados de escolaridade estão associados a maiores rendimentos, embora essa relação deva ser interpretada em conjunto com outros fatores econômicos, sociais e ocupacionais que também influenciam a remuneração.

Figura 11: *Boxplot* comparativo da renda mensal por nível de escolaridade, com

indicação da mediana, intervalo interquartil, dispersão e valores atípicos,
Pernambuco, 2019–2025.



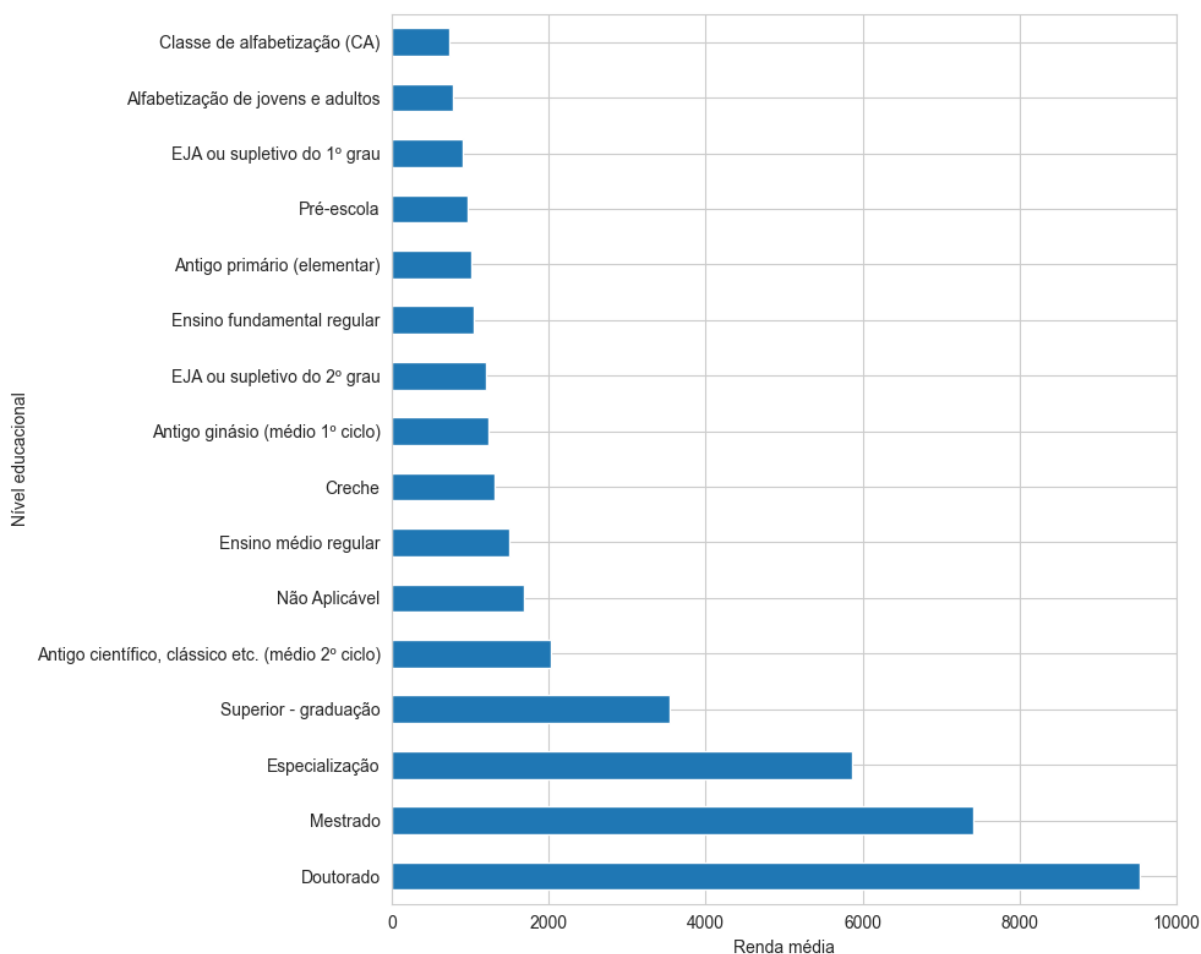
Fonte: Elaborado pelos autores.

A diferença de rendimentos também se mantém nos níveis de pós-graduação. Conforme a distribuição da renda média apresentada na Figura 12, registram-se rendimentos médios de R\$ 5.865,00 para especialistas, R\$ 7.409,00 para mestres e R\$ 9.530,00 para concluintes de doutorado. Esses resultados reforçam a associação observada entre níveis mais elevados de escolaridade e maiores rendimentos no conjunto analisado, sem implicar, isoladamente, relação causal entre as variáveis.

Cabe destacar que essa associação encontra respaldo na discussão desenvolvida no referencial teórico sobre diferenciais de rendimento. Entretanto, conforme Bourdieu (1998), Frigotto (2007, 2018) e Ferreira, Firpo e Messina (2017), a escolaridade não deve ser considerada isoladamente, uma vez que as condições de remuneração e inserção ocupacional também se relacionam a características sociais, econômicas e institucionais.

Figura 12: Gráfico de distribuição da renda mensal média das pessoas ocupadas

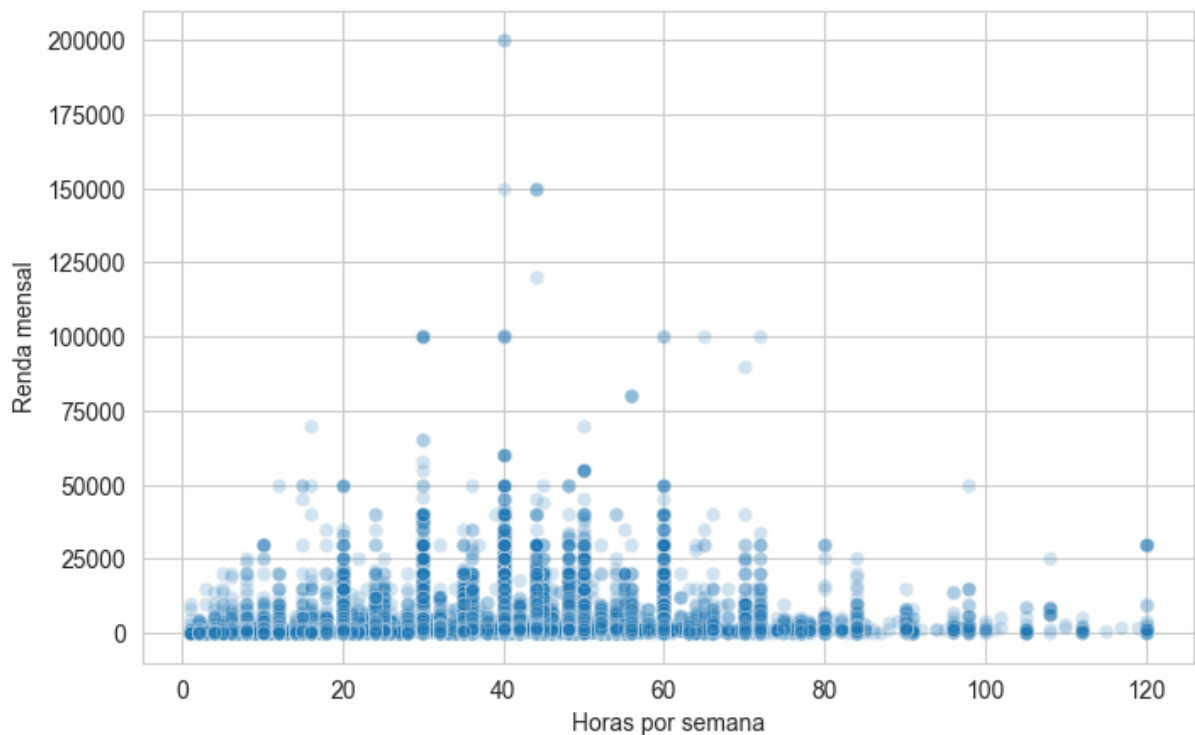
segundo o nível de escolaridade, Pernambuco, 2019–2025



Fonte: Elaborado pelos autores.

O gráfico de dispersão apresentado na Figura 13 evidencia elevada dispersão dos rendimentos para diferentes jornadas de trabalho, inclusive entre registros com jornadas semanais elevadas. Conforme discutido no referencial teórico, a desigualdade de rendimentos possui caráter multidimensional, envolvendo também posição ocupacional, formalização e características sociodemográficas (Ferreira *et al.*, 2017). Dessa forma, os resultados devem ser interpretados como associações identificadas no conjunto analisado, sem atribuição de relações causais entre as variáveis.

Figura 13: Gráfico de dispersão entre a renda mensal e as horas habitualmente trabalhadas por semana entre as pessoas ocupadas, Pernambuco, 2019–2025.



Fonte: Elaborado pelos autores.

4.5 Automação e Painel Interativo

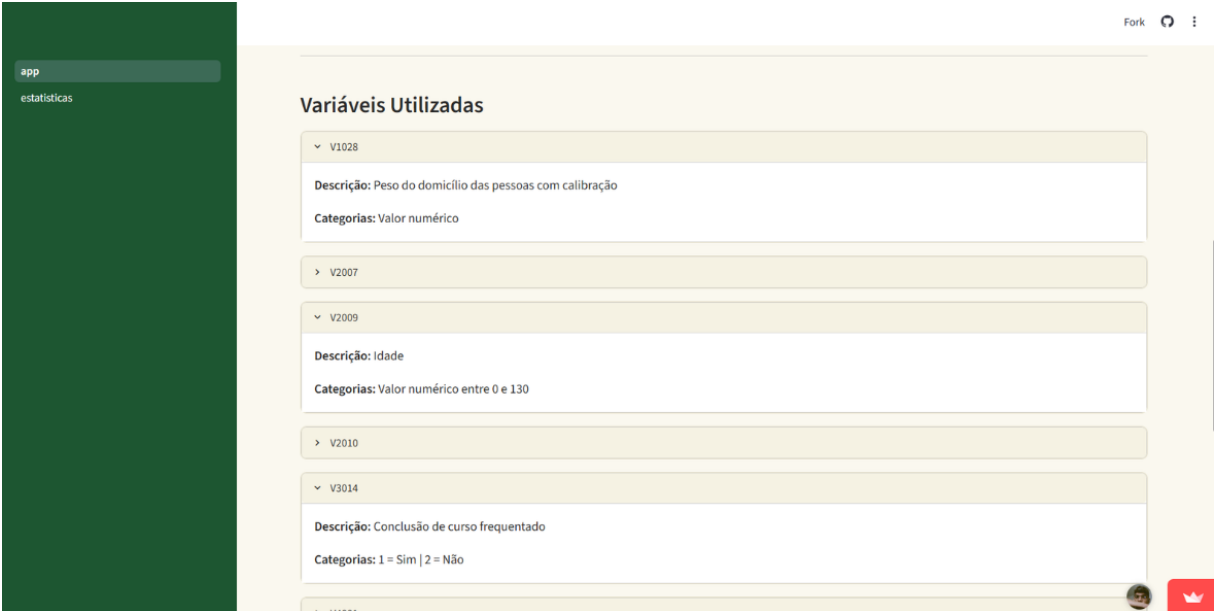
A etapa final deste trabalho consistiu na tradução do conhecimento empírico extraído dos microdados em uma ferramenta de *software* acessível, culminando no desenvolvimento de um painel interativo (*dashboard*) servido na plataforma do *Streamlit*, conforme ilustrado nas Figura 14, Figura 15 e Figura 16. Para garantir a reprodutibilidade e a modularidade da aplicação, o sistema foi estruturado em *scripts* específicos para cálculos, filtros e geração visual.

Figura 14: Página inicial do painel interativo da PNAD Contínua.



Fonte: Elaborado pelos autores.

Figura 15: Dicionário de variáveis utilizadas com sua descrição e categorias envolvidas



Fonte: Elaborado pelos autores.

Figura 16: Página de estatísticas do painel interativo da PNAD Contínua.



Fonte: Elaborado pelos autores.

No processamento dos indicadores que requerem ponderação, o módulo estatístico incorporou a variável V1028, correspondente ao peso amostral disponibilizado nos microdados da PNAD Contínua. A função implementada permitiu calcular médias ponderadas, relacionando, entre outras variáveis, o rendimento financeiro (V403312) ao respectivo peso amostral. Esse procedimento possibilitou produzir estimativas ponderadas para os indicadores implementados no painel, respeitando a distinção entre as frequências dos registros processados e as estimativas obtidas mediante ponderação.

Para a interface destinada ao usuário final, a aplicação utilizou o *framework* *Streamlit*, integrado à biblioteca *Plotly* para a geração das visualizações. Essa estrutura possibilitou apresentar os indicadores por meio de gráficos de linhas, barras, dispersão e outros elementos interativos, favorecendo a exploração dos resultados produzidos pelo sistema, conforme mostrado na Figura 17.

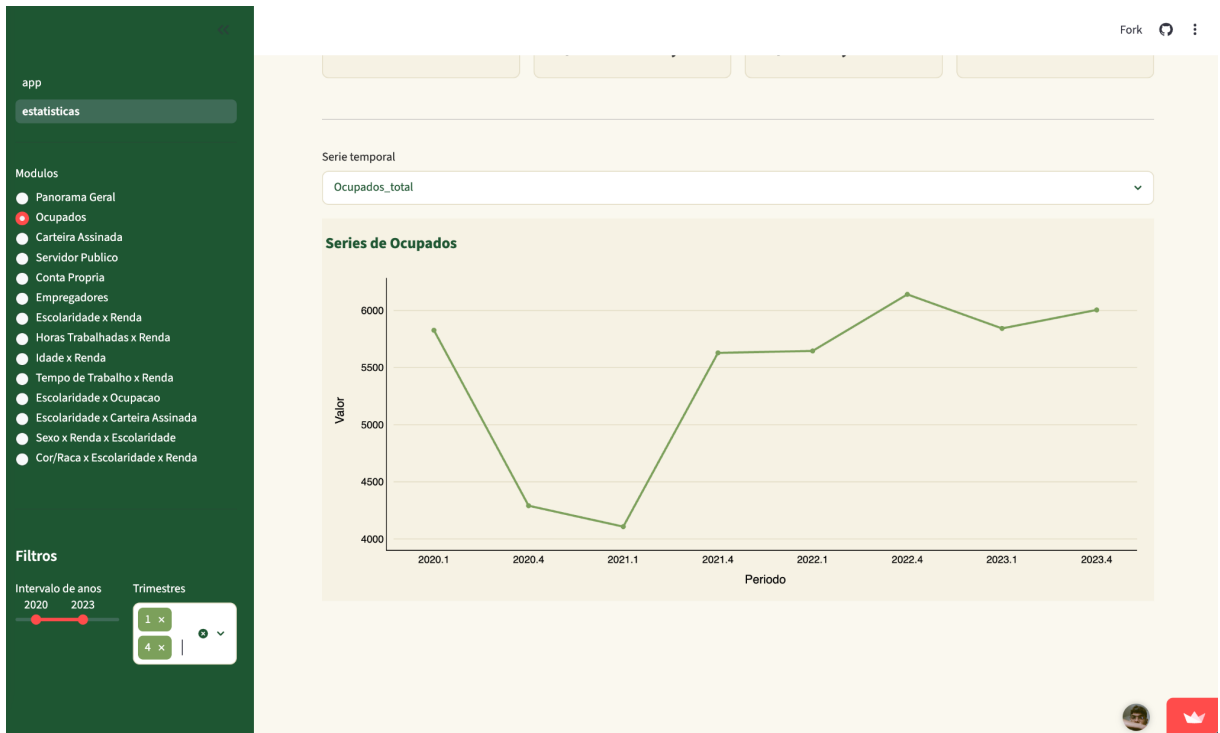
Figura 17: Gráfico de Total de pessoas no painel interativo da PNAD Contínua.



Fonte: Elaborado pelos autores.

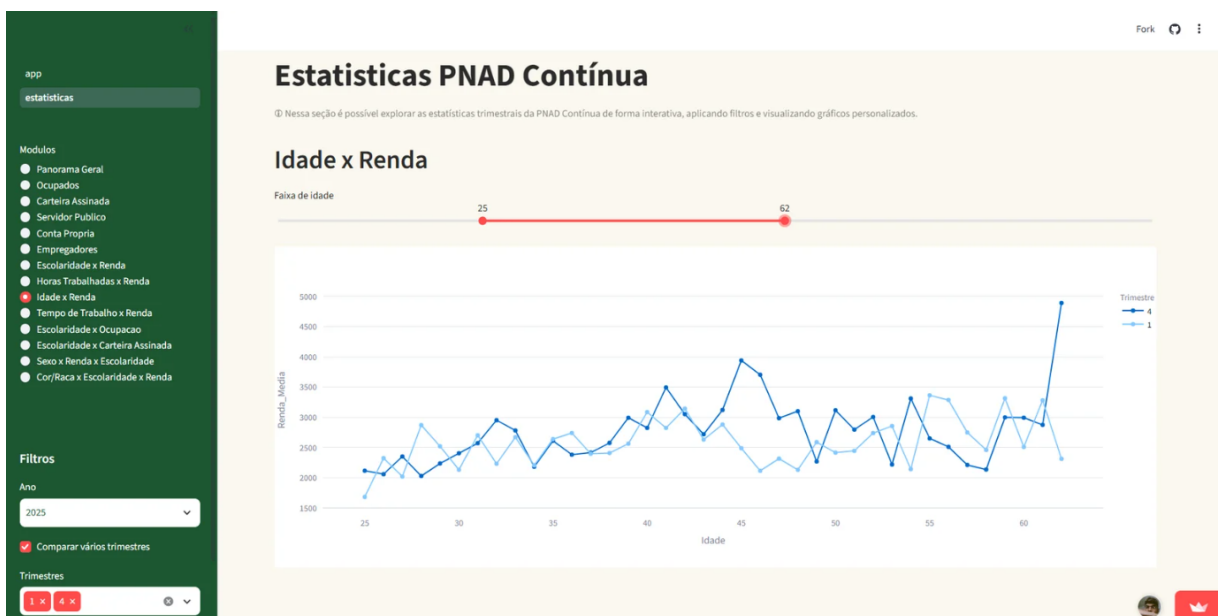
A interatividade da aplicação foi complementada pelo módulo de filtros. Por meio de componentes do *Streamlit*, como barras de seleção e caixas de seleção múltipla, o usuário pode explorar diferentes recortes da base, selecionando intervalos de anos entre 2019 e 2025 e os trimestres de interesse, conforme demonstrado na Figura 18. Além disso, é possível fazer as mesmas manipulações em módulos que oferecem dados cruzados, como mostrado na Figura 19. Dessa forma, a aplicação permite transformar uma base composta por mais de 530 mil registros em visualizações dinâmicas dos indicadores selecionados, sem que o usuário precise manipular diretamente os microdados originais.

Figura 18: Gráfico de Total de ocupados com filtro no painel interativo da PNAD Contínua.



Fonte: Elaborado pelos autores.

Figura 19: Gráfico de Total de ocupados com filtro no painel interativo da PNAD Contínua.



Fonte: Elaborado pelos autores.

A construção do painel materializa, assim, o fluxo de Ciência de Dados proposto neste trabalho, integrando aquisição, tratamento, análise e visualização em uma única solução computacional. Além de permitir a exploração dos indicadores relacionados ao mercado de trabalho pernambucano, a aplicação evidencia a viabilidade do uso de ferramentas de desenvolvimento em *Python* para tornar grandes bases de dados governamentais mais acessíveis à análise e à interpretação.

De forma integrada, os resultados demonstram tanto a viabilidade do fluxo computacional desenvolvido quanto às possibilidades analíticas proporcionadas pelos microdados da PNAD Contínua. As análises permitiram identificar padrões relacionados à inserção ocupacional, à escolaridade, à jornada de trabalho e aos rendimentos, respeitando o caráter descritivo e exploratório do estudo. Sendo assim, esses achados, associados ao desenvolvimento do painel interativo, fornecem os elementos necessários para a retomada do objetivo da pesquisa e a apresentação das considerações finais na Seção seguinte.

5 CONSIDERAÇÕES FINAIS

Este estudo teve como objetivo desenvolver uma solução baseada em Ciência de Dados para automatizar a aquisição, o tratamento, a integração, a análise e a visualização dos microdados da PNAD Contínua, possibilitando a exploração das características e desigualdades do mercado de trabalho de Pernambuco entre 2019 e 2025. A proposta foi operacionalizada por meio de um fluxo de processamento de dados em *Python*, integrado a uma aplicação *web* interativa desenvolvida com *Streamlit* e *Plotly*.

Os resultados obtidos evidenciam características relevantes do mercado de trabalho pernambucano, especialmente no que se refere à distribuição dos rendimentos, às formas de inserção ocupacional e às diferenças associadas à escolaridade. A concentração de trabalhadores em faixas de menor remuneração e a heterogeneidade das formas de ocupação dialogam com Saboia *et al.* (2021), que destacam a informalidade, a posição ocupacional e os rendimentos como dimensões importantes para a compreensão das condições de inserção no mercado de trabalho brasileiro.

No que se refere à escolaridade, os resultados indicam uma associação positiva entre níveis mais elevados de instrução e maiores rendimentos. Trabalhadores com ensino superior e pós-graduação apresentaram rendimentos médios superiores aos observados entre aqueles com menor nível de escolaridade. Essa relação, entretanto, não deve ser interpretada de forma isolada ou causal. Conforme discutido por Bourdieu (1998) e Frigotto (2007, 2018), a escolaridade se articula a outras condições econômicas e sociais que também influenciam as oportunidades de inserção e remuneração. Nesse sentido, os resultados reforçam a necessidade de compreender as desigualdades de rendimento a partir de uma perspectiva multidimensional, considerando fatores individuais, sociais e ocupacionais (Ferreira *et al.*, 2017).

No âmbito tecnológico, o trabalho demonstrou a viabilidade da aplicação integrada de técnicas de Ciência de Dados no tratamento e na exploração de microdados governamentais. O desenvolvimento do painel interativo possibilitou transformar uma base extensa e complexa em uma estrutura mais acessível para consulta, filtragem e visualização de indicadores. Nesse sentido, a solução desenvolvida aproxima-se da perspectiva apresentada por Gorard, See e Siddiqui (2020), segundo a qual a disponibilidade dos dados, por si só, não é suficiente, sendo necessário transformá-los em formatos que favoreçam sua interpretação e utilização. Da mesma forma, os recursos de visualização empregados contribuem para a identificação de padrões e tendências em grandes conjuntos de informações, conforme discutido por Ali e Vidya Sagar (2023).

A utilização do peso amostral da PNAD Contínua no módulo estatístico possibilitou a obtenção de estimativas ponderadas para os indicadores analisados, contribuindo para uma interpretação mais adequada das informações provenientes da pesquisa. Dessa forma, considera-se que o objetivo proposto foi alcançado, uma vez que o estudo resultou na construção de um fluxo automatizado de aquisição, tratamento, análise e visualização dos microdados da PNAD Contínua, associado a uma aplicação interativa voltada à exploração das características e desigualdades do mercado de trabalho de Pernambuco.

Como limitações, destaca-se o caráter amostral da PNAD Contínua e a natureza predominantemente descritiva das análises realizadas, o que não permite estabelecer relações causais entre escolaridade, rendimento e formas de inserção

ocupacional. Além disso, a consolidação de registros provenientes de diferentes períodos exige cautela na interpretação dos resultados, especialmente em análises temporais e na comparação entre grupos populacionais. Essas limitações não reduzem a contribuição da solução desenvolvida, mas delimitam o alcance das inferências que podem ser realizadas a partir dos resultados apresentados.

Para pesquisas futuras, recomenda-se o aprofundamento das análises por meio de técnicas estatísticas multivariadas, modelos preditivos e comparações temporais mais detalhadas. Também podem ser consideradas integrações com outras bases de dados do mercado de trabalho, como RAIS e CAGED, ampliando as possibilidades de investigação das desigualdades ocupacionais e de rendimento. O aprimoramento do painel interativo e a incorporação de novos indicadores também constituem possibilidades de continuidade e expansão da solução desenvolvida, envolvendo, por exemplo, a disponibilização em infraestrutura de nuvem.

REFERÊNCIAS

- ALI, A.; VIDYA SAGAR, S. D. Assessing the Performance of Python Data Visualization Libraries: A Review. *International Journal of Computer Engineering in Research Trends*, v. 10, n. 1, p. 28-39, jan. 2023. Disponível em: https://www.researchgate.net/profile/Lavanya-Addepalli/publication/369533034_Assessing_the_Performance_of_Python_Data_Visualization_Libraries_A_Review/links/6420a158315dfb4cceaee503/Assessing-the-Performance-of-Python-Data-Visualization-Libraries-A-Review.pdf. Acesso em: 19 ago. 2026
- BOURDIEU, Pierre. *A reprodução: elementos para uma teoria do sistema de ensino*. 6. ed. Rio de Janeiro: Francisco Alves, 1998. Disponível em: https://www.academia.edu/23478968/A_Reprodu%C3%A7%C3%A3o_Elementos_para_uma_Teoria_do_Sistema_de_Ensino_BORDIEU_Pierre_PASSERON. Acesso em: 19 ago. 2026
- BUENO, Douglas Marcelo Constantini. *Automação e engenharia de dados para a análise das disparidades salariais de gênero no mercado de trabalho brasileiro: um estudo de 2012 a 2021*. 2023. 53 f. Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Controle e Automação) – Instituto de Ciência e Tecnologia, Universidade Estadual Paulista "Júlio de Mesquita Filho", Sorocaba, 2023. Disponível em: <https://repositorio.unesp.br/entities/publication/cb6a2763-f678-40aa-8eb7-706af879fe75>. Acesso em: 22 jun. 2026.
- FARIAS, W. A. S. et al. Web Scraping as a scientific tool for theoretical reference. *Research Square*, [S. l.], v. 1, p. 1-22, jan. 2024. DOI: 10.21203/rs.3.rs-3854342/v1. Disponível em: <https://www.researchsquare.com/article/rs-3854342/v1>. Acesso em: 20 jun. 2026

FERREIRA, Francisco H. G.; FIRPO, Sergio P.; MESSINA, Julián. *Ageing Poorly? Accounting for the Decline in Earnings Inequality in Brazil, 1995-2012*. Washington, DC: World Bank, 2017. Policy Research Working Paper, n. 8018. Disponível em: <https://openknowledge.worldbank.org/entities/publication/2bb2a740-f502-537d-9567-847b18254422>. Acesso em: 15 jul. 2026

FRIGOTTO, Gaudêncio. *A produtividade da escola improdutiva: um (re)exame das relações entre educação e estrutura econômico-social capitalista*. 6. ed. São Paulo: Cortez, 2007. Disponível em: https://konektacommerce.nyc3.cdn.digitaloceanspaces.com/TEXT_SAMPLE_CONTENT/produtividade-da-escola-improdutiva-a-89230-1.pdf. Acesso em: 19 ago. 2026

_____. *Educação e a crise do capitalismo real*. 7. ed. São Paulo: Cortez, 2018. Disponível em: <https://pt.scribd.com/doc/282997998/FRIGOTTO-Educacao-e-Crise-Do-Capitalismo-Real>. Acesso em: 19 ago. 2026

GHOLIZADEH, S. Top Popular Python Libraries in Research. *Journal of Robotics and Automation Research*, Cape Town, v. 3, n. 2, p. 142-145, 2022. Disponível em: https://www.researchgate.net/publication/361306436_Top_Popular_Python_Libraries_in_Research. Acesso em: 20 jun. 2026

GORARD, S.; SEE, B. H.; SIDDIQUI, N. What is the evidence on the best way to get evidence into use in education?, *Review of Education*, v. 8, 2020. Disponível em: <https://bera-journals.onlinelibrary.wiley.com/doi/10.1002/rev3.3200>. Acesso em: 15 dez. 2025

KABIR, M. A. et al. Python for Data Analytics: A Systematic Literature Review of Tools, Techniques, and Applications. *Academic Journal on Science, Technology, Engineering & Mathematics Education*, Ohio, v. 4, n. 4, p. 134-154, 2024. DOI: 10.69593/ajsteme.v4i04.146. Disponível em: https://www.researchgate.net/publication/385805476_Python_For_Data_Analytics_A_Systematic_Literature_Review_Of_Tools_Techniques_And_Applications. Acesso em: 20 jun. 2026

LAMPROPOULOS, G. *Educational Data Mining and Learning Analytics in the 21st Century*. In: *Encyclopedia of Data Science and Machine Learning*. IGI Global, 2021. p. 1642-1651. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0736585318304234>. Acesso em: 19 ago. 2026

LE PIANE, F. *Data-driven and data-oriented methods for materials science and technologies*. 2023. Tese (Doutorado em Ciência da Computação e Engenharia) – Alma Mater Studiorum Università di Bologna, Bolonha, 2023. Disponível em: https://amsdottorato.unibo.it/id/eprint/10976/4/Thesis_final_LePiane.pdf. Acesso em: 20 jun. 2026

REGO, F. A.; ROSAS, I. R. C.; PRADOS, R. M. N. Educação Profissional e Tecnológica como alternativa de acesso ao mercado de trabalho. *Brazilian Journal of Development*, Curitiba, v. 7, n. 2, p. 14585-14596, fev. 2021. Disponível em:

<https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/24536>. Acesso em: 19 ago. 2026

SAABITH, A. L. S.; VINOTHRAJ, T.; FAREEZ, M. M. M. A Review on Python Libraries and Ides for Data Science. *International Journal of Research in Engineering and Science (IJRES)*, [S. l.], v. 9, n. 11, p. 36-53, nov. 2021. Disponível em: https://www.researchgate.net/publication/357898994_A_Review_on_Python_Libraries_and_IDEs_for_Data_Science. Acesso em: 20 jun. 2026

SABOIA, João; HALLAK NETO, João; SIMÕES, André; DICK, Paulo C. Mercado de trabalho, salário-mínimo e distribuição de renda no Brasil no passado recente. *Revista de Economia Contemporânea*, Rio de Janeiro, v. 25, n. 2, p. 1-30, 2021. Disponível em: <https://www.scielo.br/j/rec/a/PfsVrtZPhMvPFv8hsD8R4fk/abstract/?lang=pt>. Acesso em: 16 jul. 2026.

SANTOS, Elisama Sabrina dos. *Construção de painéis de inteligência de negócios com indicadores educacionais no IFPE Campus Paulista*. 2023. Trabalho de Conclusão de Curso (Tecnologia em Análise e Desenvolvimento de Sistemas) – Instituto Federal de Pernambuco, Paulista, 2023. Disponível em: <https://repositorio.ifpe.edu.br/xmlui/handle/123456789/1034>. Acesso em: 15 jun. 2026.

SCHILDKAMP, K. Data-based decision-making for school improvement: Research insights and gaps. *Educational Research*, v. 61, n. 3, p. 257-273, 2019. Disponível em: <https://www.tandfonline.com/doi/epdf/10.1080/00131881.2019.1625716?needAccess=true>. Acesso em: 19 ago. 2026

URANI, André. *Jornada de trabalho, salários e absorção de mão-de-obra na indústria de transformação brasileira: 1981-1990*. Rio de Janeiro: Banco Nacional de Desenvolvimento Econômico e Social, 1995. 32 p. (Textos para Discussão, n. 37). Disponível em: <https://web.bndes.gov.br/bib/jspui/handle/1408/19143>. Acesso em: 16 jul. 2026.

VIEIRA, Gislayne Pinheiro Malagueta; SOUZA, Eduardo A. B. M. *Design de dashboards na Gestão Pública: Análise de caso com dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD)*. In: CONGRESSO INTERNACIONAL DE DESIGN DA INFORMAÇÃO, 12., 2025, São Paulo. Anais [...]. São Paulo: Sociedade Brasileira de Design da Informação, 2025. p. 1-14. Disponível em: <https://pdf.blucher.com.br/designproceedings/cidi2025/1091406.pdf>. Acesso em: 22 jun. 2026.