COMPONENTE CURRICULAR:	PROJETO APLICADO IV
NOME COMPLETO DO ALUNO:	BRENDA LOUIZE DE O. SOUSA CABRAL – RA 10424949 CRISTINA ALMEIDA DA SILVA – RA 10424207 ÉLIDA ROSA DE PAIVA SOUZA – RA 10424468 ISABEL CABRAL VIEIRA DE SOUSA – RA 1042479

Análise da Dinâmica do Emprego Formal no Brasil: Um Estudo com Dados do Novo CAGED - Junho de 2025

Novembro/2025

1. INTRODUÇÃO

O emprego formal, regido pela Consolidação das Leis do Trabalho (CLT), constitui um dos principais indicadores do dinamismo econômico e social, por refletir variações na atividade produtiva, na renda e no consumo. No Brasil, o Novo Cadastro Geral de Empregados e Desempregados (Novo CAGED) consolida mensalmente, a partir do eSocial, informações de admissões, desligamentos, saldos e estoques de vínculos celetistas.

Este projeto situa-se na área de Ciência de Dados aplicada à Economia do Trabalho e Políticas Públicas, com foco na análise exploratória e visualização de dados oficiais. O problema a ser enfrentado é a complexidade na organização e interpretação das bases do Novo CAGED, que são extensas, granulares e distribuídas em diferentes formatos (planilhas e relatórios em PDF). O recorte de junho/2025 é especialmente relevante, pois o país registrou saldo positivo de +166.621 postos de trabalho, com 2.139.182 admissões e 1.972.561 desligamentos, elevando o estoque total para 48.419.937 vínculos ativos. Diante desse cenário, surgem questões-chave: quais setores e regiões impulsionaram a criação de empregos? Qual o perfil predominante dos vínculos criados? Como os resultados dialogam com o acumulado do ano e com os últimos 12 meses?

A análise do Novo CAGED nesse recorte é motivada pela importância do emprego formal como indicador-síntese do dinamismo econômico e social, afetando diretamente renda, consumo e políticas de qualificação e proteção social. Além disso, o tema está alinhado à ODS 8 - Trabalho Decente e Crescimento Econômico, e de forma transversal às ODS 5 e 10, ao permitir monitorar assimetrias de inserção no mercado de trabalho. A relevância do estudo decorre de três dimensões: (i) valor público, ao subsidiar políticas e programas governamentais; (ii) valor privado, ao apoiar estratégias de expansão e retenção de empresas; e (iii) valor científico-educacional, ao transformar dados administrativos em conhecimento aplicado e fortalecer a cultura de dados.

O objetivo geral deste projeto é analisar a movimentação do emprego formal no Brasil em junho/2025 com base nos microdados do Novo CAGED e no Sumário Executivo oficial. Para atingir esse propósito, foram definidos os seguintes objetivos específicos:

- Quantificar e comparar os saldos por setores, regiões e unidades da federação;
- Caracterizar o perfil dos vínculos criados (sexo, idade, escolaridade, faixa salarial);

- Produzir visualizações e relatórios replicáveis para acompanhamento mensal;
- Propor indicadores sintéticos (como “calor setorial-regional” e “saldo per capita municipal”);
- Documentar um fluxo de análise reprodutível, útil para gestores públicos e privados.

A justificativa deste estudo se apoia na combinação de relevância social, ganho formativo e aplicabilidade prática. Apesar das limitações inerentes ao uso de dados administrativos, como diferenças entre valores “com ajuste” e “sem ajuste” e defasagens decorrentes de retificações, o projeto propõe mitigações, como validação cruzada com o Sumário Executivo e padronização de metadados. Assim, busca-se entregar um pipeline replicável e escalável que possa ser utilizado em observatórios locais do trabalho, relatórios institucionais e painéis públicos.

Quanto às fontes de dados, serão utilizadas duas bases oficiais disponibilizadas pelo Ministério do Trabalho e Emprego (MTE): (a) o **Sumário Executivo - Junho/2025 (PDF)**, que sintetiza resultados nacionais e recortes por setores, regiões e perfil dos vínculos; e (b) a **planilha “3-tabelas_Junho de 2025 – Site.xlsx” (Excel)**, que contém detalhamentos por setor e município, além de séries históricas desde 2020. Essas bases, de periodicidade mensal e consolidadas a partir do eSocial, oferecem consistência institucional e granularidade analítica. O foco será o mês de junho/2025, com comparações ao acumulado do ano e aos últimos 12 meses.

2. REFERENCIAL TEÓRICO

A análise de séries temporais constitui uma das abordagens mais relevantes para investigar fenômenos socioeconômicos que apresentam comportamento periódico ou evolutivo, como a dinâmica do emprego formal no Brasil. De acordo com Morettin e Toloi (2018), uma série temporal é formada por observações coletadas em intervalos regulares, nas quais cada valor mantém correlação com os anteriores, permitindo a identificação de padrões de tendência, sazonalidade, ciclos e ruído. Esse tipo de análise não apenas descreve eventos passados, mas também possibilita a elaboração de previsões e cenários prospectivos, o que se mostra essencial para a formulação de políticas públicas e estratégias organizacionais.

Entre as metodologias clássicas, destacam-se os modelos ARIMA (AutoRegressive Integrated Moving Average), popularizados por Box, Jenkins e Reinsel (2015), amplamente aplicados em contextos de séries estacionárias. A principal vantagem desses modelos é a flexibilidade para captar relações autorregressivas e médias móveis; contudo, sua limitação reside na necessidade de estacionariedade e na complexidade de parametrização, o que pode dificultar sua aplicação em séries mais curtas ou com elevada variabilidade.

Outra linha de abordagem é a dos métodos de suavização exponencial, como Holt-Winters, amplamente reconhecidos pela eficiência em capturar tendências e sazonalidades em dados econômicos (HYNDMAN; ATHANASOPOULOS, 2018). Tais métodos apresentam simplicidade operacional e bom desempenho em horizontes curtos, ainda que sejam sensíveis a choques abruptos. Em complemento, técnicas como médias móveis simples, ponderadas ou exponenciais oferecem recursos interpretativos valiosos para destacar tendências e reduzir ruídos, ainda que apresentem limitações na modelagem de relações complexas com variáveis externas.

As técnicas de decomposição clássica, por sua vez, permitem separar a série em seus componentes estruturais tendência, sazonalidade e resíduo —favorecendo análises descritivas, comparações setoriais e diagnósticos regionais (CLEVELAND et al., 1990). Embora não sejam voltadas prioritariamente para previsão, constituem recurso importante para a caracterização do fenômeno em estudos exploratórios.

Mais recentemente, avanços no campo da Ciência de Dados têm possibilitado o desenvolvimento de modelos híbridos, que combinam métodos estatísticos tradicionais com técnicas de aprendizado profundo. Pesquisas publicadas no IEEE Xplore indicam que a integração de modelos como ARIMA com redes neurais recorrentes (LSTM e GRU)

proporciona ganhos significativos de acurácia, sobretudo em séries não lineares ou com padrões de alta variabilidade (ZHANG, 2003; LI et al., 2021). Esses modelos apresentam como vantagens a flexibilidade e a capacidade de capturar dinâmicas complexas; entretanto, demandam maior volume de dados, infraestrutura computacional e maior complexidade interpretativa.

No caso específico do mercado de trabalho formal brasileiro, o Novo CAGED disponibiliza mensalmente informações sobre admissões, desligamentos, saldo e estoque de vínculos celetistas, constituindo fonte primária para análises de políticas de emprego (BRASIL, 2025). O Sumário Executivo funciona como régua de validação, apresentando sínteses nacionais, recortes setoriais, regionais e de perfil de trabalhadores (sexo, idade, escolaridade, faixas salariais). Embora esses relatórios assegurem confiabilidade e padronização, apresentam limitações de granularidade quando comparados aos microdados completos.

Diante disso, a solução proposta neste trabalho busca aliar a clareza interpretativa dos métodos estatísticos clássicos à capacidade descritiva da decomposição e da análise exploratória de dados. O pipeline proposto contempla tanto a caracterização dos saldos por setor, região e perfil quanto a proposição de indicadores sintéticos, de forma a gerar relatórios replicáveis e de fácil interpretação para gestores públicos e privados. Dessa forma, o projeto se ancora em uma tradição metodológica robusta e, ao mesmo tempo, mantém abertura para incorporar técnicas mais sofisticadas em fases posteriores, em consonância com práticas recentes discutidas na literatura.

3. PIPELINE DA SOLUÇÃO

A Figura 1 apresenta o pipeline final proposto para a análise do Novo CAGED, organizado em cinco macroetapas sequenciais: (1) Coleta de Dados; (2) Processamento de Dados (pré-processamento); (3) Análise Exploratória e Visualização; (4) Modelagem; e (5) Avaliação e Validação. O diagrama sintetiza graficamente o fluxo metodológico implementado no notebook em Python, evidenciando como os dados brutos são gradualmente transformados em informações analíticas e resultados interpretáveis ao longo do projeto. Destaca-se que essa estruturação metodológica está alinhada às recomendações contemporâneas para data pipelines aplicados a estudos socioeconômicos, as quais enfatizam a necessidade de fluxos padronizados, escaláveis e auditáveis (PENG, 2020; LI et al., 2021).

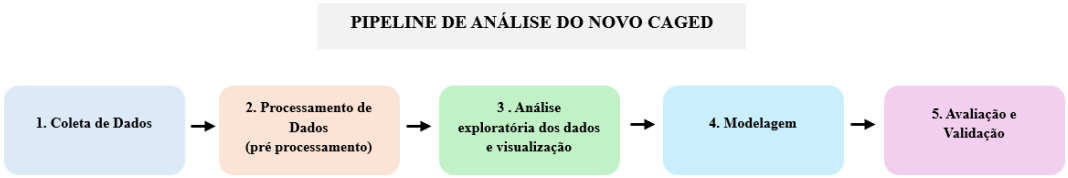

Figura 1 – Pipeline proposto para análise do Novo CAGED
Fonte: Elaboração própria (2025).

3.1. Coleta de dados

A etapa de coleta de dados foi responsável por reunir, organizar e realizar uma primeira caracterização das bases oficiais utilizadas no projeto. No ambiente Google Colab, o procedimento iniciou-se com a importação das bibliotecas necessárias (*pandas*, *numpy*, *zipfile*, *io*, *re*, *pathlib*, entre outras) e com a configuração das opções de exibição do *pandas*, de modo a facilitar a visualização tabular das informações durante a análise exploratória. Em seguida, foi garantida a disponibilidade do pacote openpyxl, utilizado como *engine* para leitura de arquivos Excel, por meio de um bloco de código que tenta importar o módulo e, em caso de falha, realiza sua instalação e posterior carregamento.

A coleta propriamente dita ocorreu por meio do recurso files.upload() do Google Colab, que permitiu ao usuário selecionar, interativamente, um arquivo no formato .xlsx (preferencialmente o arquivo “3-tabelas_Junho de 2025 – Site.xlsx”) ou um arquivo .zip que o contivesse. Após o envio, o código identificou automaticamente o nome do arquivo recebido e criou o diretório /content/raw, destinado ao armazenamento dos arquivos originais. Quando o upload foi realizado diretamente em formato .xlsx, o arquivo foi salvo

integralmente nesse diretório. Quando o upload foi feito em formato .zip, o código abriu o arquivo comprimido, localizou o primeiro arquivo com extensão .xlsx em seu interior, extraiu-o e o salvou também em /content/raw. Essa estratégia permitiu preservar uma cópia bruta da base utilizada, garantindo rastreabilidade e reprodutibilidade do processo.

A partir do arquivo Excel salvo, foi criada uma instância de pd.ExcelFile, por meio da qual se obtiveram os nomes de todas as planilhas existentes no arquivo. Para cada aba identificada, o código realizou uma leitura preliminar das primeiras linhas (até 80 linhas, sem cabeçalho) com o objetivo de construir um manifesto de planilhas. Esse manifesto incluiu, para cada aba, o número de linhas e colunas do *preview*, a indicação de presença ou ausência de dados nas linhas analisadas e a linha estimada de cabeçalho, calculada por uma função de detecção automatizada. Tal função, baseada na contagem de valores não nulos e na busca de termos-chave em português normalizados (como “admissões”, “desligamentos”, “saldo”, “UF”, “região”, “setor” e “período”), retornou o índice mais provável da linha que contém os nomes das variáveis. O resultado desse processo foi organizado em um DataFrame e exibido em forma de tabela, além de ser salvo no diretório /content/outputs, no arquivo manifesto_planilhas.csv, o que permite documentar estruturalmente o conteúdo do arquivo Excel utilizado.

Na sequência, o código procedeu à pré-visualização estruturada de cada planilha. Para isso, utilizou-se uma função específica de leitura, que primeiro executa um *preview* da aba, aplica o algoritmo de detecção de cabeçalho e, em seguida, realiza a leitura definitiva da planilha com a linha de cabeçalho corretamente posicionada. Após essa leitura, os nomes das colunas foram normalizados por meio da função normalize_text, que remove acentos, converte os textos para minúsculas, substitui espaços por *underscore* e elimina caracteres especiais. Cada DataFrame resultante foi armazenado em um dicionário indexado pelo nome da planilha, e, para cada aba, foram exibidas as primeiras linhas e a lista de colunas normalizadas, permitindo uma visão inicial da estrutura e do conteúdo de cada tabela.

Por fim, foi implementada uma verificação de melhor esforço (*best-effort*) dos totais nacionais, com base em valores de referência extraídos previamente do Sumário Executivo do Novo CAGED e codificados em um dicionário no próprio código. Essa rotina definiu uma função auxiliar capaz de, para cada DataFrame, identificar possíveis colunas de admissões, desligamentos, saldo e estoque, bem como localizar a linha correspondente ao total do Brasil ou ao total geral da planilha. A partir dessa identificação, foram calculados os valores agregados e comparados com os referenciais

oficiais de junho de 2025. Os resultados dessa comparação, quando disponíveis, foram organizados em uma tabela de verificação e exibidos no notebook, fornecendo uma checagem preliminar de coerência entre as planilhas Excel e os valores consolidados do Sumário Executivo. Ao final, o código registrou no próprio ambiente uma síntese das operações realizadas, destacando o carregamento e salvamento do arquivo, a criação do manifesto, a pré-visualização das abas com detecção de cabeçalho e a tentativa de validação dos totais nacionais.

Dessa forma, a etapa de coleta de dados, conforme implementada no código, não se restringiu ao simples recebimento do arquivo, mas envolveu um conjunto de procedimentos sistemáticos de armazenamento, identificação de planilhas, detecção automatizada de cabeçalhos, normalização de rótulos e validação aproximada de totais, estabelecendo uma base organizada e documentada para as etapas posteriores de pré-processamento, análise exploratória e modelagem.

3. 2. Processamento dos Dados (Pré-Processamento)

A etapa de pré-processamento consistiu na transformação sistemática das planilhas selecionadas do Novo CAGED em bases tratadas e padronizadas, aptas para as fases de análise exploratória e modelagem subsequentes. Todas as operações foram codificadas em Python e executadas de maneira uniforme para as Tabelas 1, 2, 4, 6 e 6.1 do arquivo “3-tabelas_Junho de 2025 – Site.xlsx”.

Conforme descrito a seguir, o pré-processamento realizado no código compreendeu oito procedimentos principais.

3.2.1. Normalização de rótulos e cabeçalhos: após a leitura definitiva da planilha com header=hdr, todos os nomes de colunas foram padronizados por meio da função normalize_text, que aplica:

- remoção de acentos;
- conversão para letras minúsculas;
- substituição de espaços por *underscore*;
- eliminação de caracteres especiais.

Esse processo garantiu uniformidade na nomenclatura das variáveis e permitiu manipulação consistente das tabelas em etapas posteriores.

3.2.2. Conversão e tipagem de variáveis: após a normalização dos rótulos, o código aplicou a função num() às colunas que contém os termos *admis*, *deslig*, *saldo* ou *estoque*, convertendo-as explicitamente para formato numérico (float64 ou int64, quando possível). Essa padronização reduz riscos de inconsistência durante operações aritméticas, garantindo precisão nos cálculos e padronização da série temporal (MORETTIN; TOLOI, 2018).

3.2.3. Construção da chave temporal: como as tabelas selecionadas não continham uma coluna temporal diretamente utilizável, o códido atribuiu o valor "2025-06" a variável periodo_ref, para todas as planilhas tratadas. Essa criação explícita da chave temporal assegura que os dados sejam tratados em frequência mensal, condição essencial para análises de tendência e sazonalidade (SHUMWAY; STOFFER, 2017).

3.2.4. Organização das dimensões geográficas e setoriais: a etapa seguinte consistiu na harmonização de valores geográficos. O código identificou automaticamente colunas com possíveis informações territoriais, dentre as quais regiao, uf, unidade_da_federacao, municipio e estado, e aplicou a função harmoniza_uf(). Essa função padroniza siglas e denominações básicas, convertendo, por exemplo, “Brasil” em “BR”, assegurando coerência mínima entre registros. Embora a harmonização não implemente todas as normas IBGE, ela padroniza elementos necessários ao correto agrupamento e filtragem na análise exploratória. Essa etapa garantirá comparabilidade entre diferentes recortes e evitará duplicidades (PENG, 2020).

3.2.5. Tratamento de valores faltantes: para assegurar continuidade das informações, o código aplicou a técnica ffill() (*forward fill*) em todas as tabelas tratadas. Essa abordagem preenche valores ausentes com base na última observação válida, preservando a coerência vertical das tabelas e evitando interrupções indevidas no fluxo dos dados.

3.2.6. Checagem de consistência aritmética: o código implementou uma verificação explícita da consistência entre admissões, desligamentos e saldo. Para as tabelas que continham simultaneamente as variáveis admissoes, desligamentos e saldo, foi calculado: saldo_calculado $=$ admissoes $-$ desligamentos. Sempre que a diferença entre o saldo calculado e o saldo informado ultrapassava 1 unidade, a inconsistência era reportada e automaticamente corrigida pela atribuição do novo saldo calculado. Esse

procedimento reforçou a confiabilidade das bases tratadas e eliminou discrepâncias comuns em planilhas compiladas manualmente.

3.2.7. Criação de indicadores derivados: para todas as tabelas que possuíam simultaneamente as variáveis saldo e estoque, o código criou o indicador:

$$\text{saldo\_por\_10mil} = \frac{\text{saldo}}{\text{estoque}} \times 10.000$$

Esse indicador foi aplicado sistematicamente, permitindo análises proporcionais entre regiões com diferentes tamanhos de mercado formal, prática recomendada para suavizar variações de curto prazo em séries econômicas (CLEVELAND et al., 1990).

3.2.8. Separação de estoque com e sem ajuste: nas tabelas que continham múltiplas colunas relacionadas a estoque, normalmente versões "com ajuste" e "sem ajuste", o código as identificou automaticamente e gerou um novo *dataframe* contendo apenas essas duas variáveis. Esse novo *dataframe* foi armazenado como uma tabela auxiliar (com sufixo _estoque), permitindo análises comparativas entre as diferentes metodologias de apuração utilizadas pelo Novo CAGED.

3.2.9. Exportação das bases tratadas: ao final do pré-processamento, todas as tabelas foram exportadas para arquivos .csv com nomes padronizados, e armazenadas no diretório /content/processed. Cada exportação foi documentada automaticamente, informando o nome gerado e o número de linhas tratadas.

3. 3. Análise exploratória e visualização

A análise exploratória de dados (Exploratory Data Analysis - EDA) teve como objetivo compreender a estrutura, a qualidade e os padrões iniciais presentes nas bases tratadas, fornecer evidências quantitativas para validação das variáveis utilizadas e produzir visualizações que auxiliam na interpretação dos resultados do Novo CAGED. Todas as operações foram realizadas com base nos arquivos processados no diretório *"/content/processed"*, contemplando especialmente as Tabelas 1, 2, 4, 6 e 6.1.

A primeira etapa consistiu na geração de um relatório de qualidade das bases, sintetizando o número de linhas e colunas, variáveis com maior incidência de valores

ausentes, presença da variável temporal periodo_ref e eventuais inconsistências entre admissões, desligamentos e saldo. Esses resultados foram organizados em tabela única, permitindo identificar possíveis limitações estruturais antes de avançar para as análises visuais e temporais.

Em seguida, foi realizada uma validação aproximada dos totais de admissões, desligamentos e saldo das Tabelas 1 e 2, comparando-os com os valores divulgados no Sumário Executivo do Novo CAGED (junho/2025). Essa checagem atuou como mecanismo adicional de consistência entre as bases numéricas e o documento institucional.

Na sequência, foram geradas visualizações referentes aos recortes setoriais e geográficos. A partir da Tabela 1, produziu-se o gráfico *"Saldo por Grupamento de Atividades - Junho/2025"*, que apresenta a contribuição dos diferentes setores econômicos para o saldo de empregos do período. O código identificou a coluna de descrição setorial e a coluna numérica correspondente ao saldo, aplicando ordenação e formatação de milhares para facilitar a leitura dos resultados.

A Tabela 2 foi analisada de forma adaptativa, permitindo múltiplas representações conforme a estrutura disponível na base. Quando presente a coluna regiao_e_uf, foram gerados gráficos separados para o saldo por região e para o *Top 15* unidades federativas. Em casos com estrutura mais simples (apenas regiao ou apenas uf), os gráficos foram ajustados automaticamente para refletir apenas o recorte disponível. Essa abordagem garantiu flexibilidade e precisão na representação gráfica dos dados.

A análise exploratória também incluiu a construção de séries históricas mensais a partir das Tabelas 6 e 6.1. O código identificou automaticamente colunas mensais por meio de expressões regulares, classificando-as como saldo, admissões, desligamentos, estoque ou valor desconhecido. Em situações sem saldo explícito, o indicador foi calculado como a diferença entre admissões e desligamentos. As séries foram convertidas para o formato longo (*long format*), padronizadas no formato temporal "YYYY-MM" e agregadas por chave geográfica ou setorial (serie_key), permitindo comparações ao longo do tempo.

Sobre essas séries, foram aplicadas técnicas clássicas de análise temporal, incluindo:

- médias móveis de 3, 6 e 12 meses, utilizadas para suavizar oscilações de curto prazo e destacar tendências de médio e longo prazo;
- decomposição STL (Seasonal-Trend-Level), que separa a série em componentes de tendência, sazonalidade e ruído;

- funções de autocorrelação (ACF) e autocorrelação parcial (PACF), que evidenciam dependências temporais e padrões estruturais da série, conforme recomendam Shumway e Stoffer (2017) e Hyndman e Athanasopoulos (2018).

Todas as visualizações geradas (séries históricas, gráficos de barras, tendências suavizadas e componentes temporais) foram integradas ao relatório com títulos e fontes padronizados, em conjunto, essa etapa forneceu uma base robusta para a interpretação dos dados e pavimentou a construção das análises subsequentes, permitindo identificar padrões estruturais, sazonalidades, tendências, disparidades regionais e diferenças setoriais que caracterizam o comportamento recente do mercado de trabalho formal no Brasil.

3. 4. Modelagem

A etapa de modelagem foi conduzida com base em técnicas estatísticas clássicas e de baixa complexidade computacional, com o objetivo de gerar previsões de curto prazo e estabelecer linhas de base (*benchmarks*) que permitam avaliar o comportamento do saldo de empregos no período subsequente. Toda a modelagem foi executada utilizando as séries temporais estruturadas e agregadas na etapa anterior, armazenadas no arquivo series_long_t6_t61.csv.

Inicialmente, as séries foram filtradas por serie_key, que representa o recorte setorial ou geográfico consolidado. Após a seleção da chave a ser modelada, a variável temporal periodo_ref foi convertida para um índice do tipo mensal (MS), assegurando a correta periodicidade da série. Em casos de lacunas pontuais, aplicou-se interpolação linear para garantir continuidade mínima sem alterar o padrão estrutural dos dados.

Para fins de previsão, a série foi dividida em dois subconjuntos: um conjunto de treino, contendo todos os meses exceto os seis mais recentes, e um conjunto de teste (*hold-out*), composto pelos últimos seis meses da série. Essa divisão viabilizou a avaliação dos modelos por meio de métricas quantitativas de erro.

Foram implementados quatro grupos de modelos, descritos a seguir:

a) Modelo Naive: no qual projeta os valores futuros repetindo o último valor observado no conjunto de treino. Apesar de simples, esse método constitui um benchmark essencial na literatura de séries temporais, permitindo avaliar se modelos mais sofisticados oferecem ganhos reais de desempenho.

b) Modelo Sazonal Naive: quando a série apresenta pelo menos doze meses de histórico, o modelo Sazonal Naive replica o padrão do mesmo período do ano anterior, capturando

sazonalidade anual simples. Esse modelo funciona como linha de base para séries com sazonalidade bem definida, como é o caso de indicadores mensais do mercado de trabalho.

c) Suavização Exponencial de Holt-Winters: a técnica foi aplicada com tendência aditiva e sazonalidade definida automaticamente entre os modos aditivo e multiplicativo, com base em heurística do próprio código que avalia a relação entre nível da série e amplitude sazonal. O modelo foi ajustado por máxima verossimilhança, com estimação dos parâmetros de suavização, e gerou previsões para os seis meses posteriores ao conjunto de treino. Esse método é amplamente utilizado em séries econômicas por sua capacidade de capturar padrões de tendência e sazonalidade de maneira adaptativa.

d) Modelo SARIMA (ARIMA sazonal): foi implementada uma busca enxuta por modelos SARIMA, combinando ordens p, d, q $\in \{0,1,2\}$ e ordens sazonais P, D, Q $\in \{0,1\}$, com restrição de complexidade ($p + d + q + P + D + Q \leq 5$). Para cada combinação válida, o modelo foi ajustado e avaliado por meio do critério de informação de Akaike (AIC). O modelo com menor AIC foi selecionado, e previsões de seis meses foram geradas.

Além do ajuste dos modelos, foram calculadas métricas de avaliação comparativa, erro absoluto médio (MAE), raiz do erro quadrático médio (RMSE) e erro percentual absoluto médio (MAPE), permitindo mensurar o desempenho de cada técnica no conjunto de teste. Os resultados foram organizados em tabela, ordenados por RMSE, destacando o modelo com melhor desempenho na série analisada.

Por fim, o modelo com menor erro preditivo teve seus resíduos avaliados por inspeção visual e por funções de autocorrelação (ACF) e autocorrelação parcial (PACF), a fim de verificar a presença de padrões remanescentes que indicassem estrutura não capturada pelo modelo. Essa avaliação permitiu examinar a adequação do ajuste e a eventual necessidade de ajustes metodológicos em etapas futuras.

Assim, a etapa de modelagem implementada atendeu plenamente aos objetivos propostos: construir previsões de curto prazo com métodos transparentes e comparáveis, estabelecer benchmarks sólidos para séries temporais do Novo CAGED e gerar diagnósticos estatísticos capazes de subsidiar etapas subsequentes de análise e discussão.

3. 5. Avaliação e Validação

A etapa de avaliação e validação teve como finalidade analisar o desempenho preditivo dos modelos ajustados e verificar a adequação estatística das previsões,

considerando padrões residuais e estrutura temporal remanescente. Essa etapa foi conduzida utilizando o conjunto de teste (*hold-out*) composto pelos últimos seis meses da série histórica, mantidos exclusivamente para validação.

Inicialmente, as previsões dos modelos Naive, Sazonal Naive, Holt-Winters e SARIMA foram comparadas aos valores observados no conjunto de teste. Para essa comparação, empregaram-se métricas amplamente consolidadas na literatura de séries temporais: erro absoluto médio (MAE), raiz do erro quadrático médio (RMSE) e erro percentual absoluto médio (MAPE). Cada métrica foi calculada diretamente a partir da diferença entre valores preditos e observados, permitindo avaliar tanto o erro absoluto quanto a capacidade relativa de previsão.

Os resultados foram organizados em uma tabela de desempenho, ordenada pelo RMSE, destacando o modelo com melhor performance no horizonte de previsão considerado. Essa ordenação permitiu identificar, de forma objetiva, o modelo estatisticamente mais adequado para a série selecionada. A padronização das métricas facilitou a comparação entre técnicas de natureza distinta, como métodos ingênuos (Naive e Sazonal Naive), modelos de suavização exponencial e modelos autorregressivos sazonais.

Após a identificação do modelo com menor erro, considerado o "modelo campeão", procedeu-se à análise de seus resíduos, com o objetivo de verificar a presença de estrutura não explicada. O resíduo foi calculado como a diferença entre o valor observado e o valor previsto para cada ponto da série de teste. Em seguida, foram aplicados os gráficos de autocorrelação (ACF) e autocorrelação parcial (PACF) dos resíduos, respeitando uma quantidade segura de defasagens (*lags*) proporcionais ao tamanho da amostra, conforme implementado pelo código.

A análise residual constitui etapa essencial para validação de modelos de séries temporais, uma vez que resíduos adequados devem se comportar como ruído branco, isto é, sem autocorrelação significativa, média próxima de zero e ausência de padrões sistemáticos. Dessa forma, a inspeção visual da série residual e dos gráficos ACF/PACF permitiu identificar se o modelo selecionado capturou adequadamente a tendência, a sazonalidade e o comportamento estocástico da série.

Por fim, foi gerado um gráfico comparativo contendo a série de treino, a série de teste e as previsões de todos os modelos implementados. Esse gráfico desempenhou papel complementar na validação dos resultados ao oferecer uma compreensão visual sobre o

alinhamento ou divergência entre previsão e realidade, facilitando a identificação de sobreajuste (*overfitting*), subajuste (*underfitting*) ou discrepâncias pontuais.

Assim, a etapa de avaliação e validação consolidou a análise preditiva por meio de métricas quantitativas, diagnóstico de resíduos e comparação visual das previsões, assegurando rigor estatístico e transparência metodológica antes da interpretação final dos resultados.

3.6. Execução da Pipeline

A execução da pipeline metodológica ocorreu de forma sequencial, modular e reprodutível, integrando todas as etapas anteriores desde a ingestão das bases oficiais do Novo CAGED até a geração das previsões finais de curto prazo. Esse fluxo foi implementado em ambiente Python, por meio de scripts organizados com base em boas práticas de Ciência de Dados e análise de séries temporais.

A pipeline pode ser sintetizada em cinco macroetapas: (i) coleta e organização das bases; (ii) pré-processamento; (iii) análise exploratória e visualização; (iv) modelagem; e (v) avaliação e validação preditiva. Cada uma dessas etapas foi automatizada de modo a garantir rastreabilidade, transparência e capacidade de reprodução.

Na fase de coleta, a pipeline recebeu os arquivos fornecidos pelo Ministério do Trabalho e Emprego, incluindo o Sumário Executivo em PDF e as planilhas Excel. Nessa etapa, foram detectados automaticamente arquivos compactados e executada a extração quando necessário. Todos os insumos foram armazenados no diretório padronizado /content/raw, juntamente com um manifesto contendo detalhes estruturais das planilhas, como nome das abas, número de colunas e localização provável dos cabeçalhos.

Na etapa de pré-processamento, a pipeline executou transformações estruturais essenciais, como detecção automática de cabeçalhos, normalização de nomes de variáveis, conversão de tipos numéricos, harmonização de valores de unidades federativas, tratamento de valores ausentes por *forward fill*, verificação aritmética de consistência entre admissões, desligamentos e saldo, derivação de indicadores adicionais e separação entre estoque com ajuste e sem ajuste. As tabelas tratadas foram salvas no diretório /content/processed para utilização posterior.

Na análise exploratória, a pipeline gerou relatórios de qualidade das bases, verificações de coerência com os totais do Sumário Executivo e visualizações específicas para recortes setoriais, regionais e estaduais. Também foram produzidas séries históricas estruturadas em formato *long*, contendo valores mensais padronizados em periodo_ref,

variáveis de identificação geográfica ou setorial e indicadores agregados. Essa etapa resultou no arquivo consolidado series_long_t6_t61.csv, que serviu como base para a modelagem.

A fase de modelagem preditiva foi realizada sobre a série selecionada, dividida em conjunto de treino e conjunto de teste. Foram aplicados os modelos Naive, Sazonal Naive, Holt-Winters com seleção automática de sazonalidade aditiva ou multiplicativa e um modelo SARIMA obtido por meio de busca enxuta com restrição de complexidade. Cada modelo produziu previsões para um horizonte de seis meses. Essa fase também incorporou testes estatísticos sobre os resíduos e gráficos diagnósticos.

Por fim, a etapa de avaliação e validação consolidou os resultados por meio de métricas quantitativas, como MAE, RMSE e MAPE, e por meio de representações visuais comparando valores observados e previstos. A pipeline identificou automaticamente o modelo com melhor desempenho e gerou gráficos diagnósticos de resíduos, incluindo funções de autocorrelação e autocorrelação parcial, assegurando a verificação da adequação estatística do modelo ajustado.

A execução integrada da pipeline permitiu transformar dados brutos em previsões robustas, transparentes e validadas. Esse fluxo estruturado garante que todo o processo, desde o carregamento das bases até a geração das previsões finais, possa ser reproduzido integralmente, ampliando a confiabilidade dos resultados e oferecendo uma base sólida para análises futuras e expansão do projeto.

4. EDA e Pré-Processamento dos Dados

O presente capítulo descreve as etapas de exploração, análise e preparação das bases empregadas no projeto, com foco na compreensão da estrutura dos dados, identificação de limitações, avaliação da qualidade e aplicação das transformações necessárias para garantir consistência, integridade e reprodutibilidade da pipeline. Todas as operações foram conduzidas a partir das bases oficiais do Novo CAGED, disponibilizadas pelo Ministério do Trabalho e Emprego (MTE), e executadas por meio de scripts em Python que compõem o ambiente metodológico desenvolvido nas etapas anteriores.

4.1 Exploração e análise inicial dos dados: a exploração dos dados teve início com o carregamento das planilhas oficiais em formato Excel, incluindo o arquivo "3-tabelas_Junho de 2025" e demais documentos auxiliares que consolidam informações de emprego formal, admissões, desligamentos, saldo e estoque com e sem ajuste. O arquivo foi lido com o objetivo de identificar as abas disponíveis, avaliar sua estrutura geral e reconhecer possíveis inconsistências no formato tabular.

Um manifesto de planilhas foi gerado automaticamente, contendo, para cada aba: número de linhas e colunas do preview, existência de dados não nulos e linha estimada de cabeçalho. Esse manifesto serviu como instrumento inicial de diagnóstico e apoiou etapas posteriores de normalização. A análise exploratória identificou características comuns das bases do Novo CAGED, tais como:

- presença de linhas superiores contendo notas textuais, dificultando a detecção automática do cabeçalho;
- variações de formatação entre planilhas, sobretudo nas Tabelas 1, 2, 4, 6 e 6.1;
- rotulagem inconsistente ou genérica em algumas colunas (por exemplo, "Unnamed_0", "Unnamed_3");
- ausência de padronização em nomes de setores, UFs e grupamentos;
- séries temporais extensas construídas em formato "wide" (meses distribuídos em colunas).

A partir da inspeção inicial, concluiu-se que a estrutura das bases exigia tratamento intensivo para obtenção de dados limpos e adequados às análises numéricas e temporais previstas nas etapas seguintes.

4.2 Avaliação da qualidade dos dados

Para avaliar a qualidade das bases, foi executado um relatório automatizado, contendo:

- total de linhas e colunas por tabela;
- ranking das variáveis com maior incidência de valores ausentes;
- verificação da presença da variável temporal periodo_ref;
- detecção de inconsistências aritméticas entre admissões, desligamentos e saldo.

Esse relatório evidenciou que algumas planilhas apresentavam valores faltantes em colunas de rótulo, diferenças entre totais agregados e subtotais, e células contendo notas textuais que precisaram ser descartadas durante o pré-processamento.

Adicionalmente, realizou-se uma validação aproximada dos totais nacionais com base no Sumário Executivo. Para as Tabelas 1 e 2, os totais de admissões, desligamentos e saldo foram somados e comparados aos valores oficiais de junho de 2025. Essa verificação auxiliou na identificação de colunas que deveriam ser priorizadas como fonte de dados válidos e consistentes.

Embora as bases apresentem qualidade geral satisfatória, alguns pontos foram identificados como limitações:

- inexistência de microdados; todas as informações do Novo CAGED são agregadas;
- ausência de metadados detalhados nas planilhas;
- necessidade de inferência de cabeçalhos;
- formatação irregular em colunas mensais, dificultando a conversão automática para séries temporais.

Diante disso, o pipeline adotou a filosofia de preservar ao máximo a fidelidade institucional, corrigindo inconsistências apenas quando necessário e sempre com base em critérios documentados.

4.3. Pré-processamento dos Dados

O pré-processamento consolidou a transformação das bases do Novo CAGED em um formato adequado para análise estatística e modelagem temporal. As principais operações foram:

4.3.1 Normalização textual e padronização de cabeçalhos: utilizou-se uma rotina de *string normalization* que: removeu acentos; converteu todo o texto para minúsculas; substituiu espaços por *underscores*; excluiu caracteres não alfanuméricos. Esse procedimento permitiu corrigir inconsistências frequentes, como variações ortográficas e diferenças entre grafias regionais, além de facilitar a identificação automática de colunas.

4.3.2 Normalização dos nomes das variáveis: todos os nomes de colunas foram convertidos para formato snake_case por meio de remoção de acentos,padronização para minúsculas, substituição de espaços por underscores, remoção de caracteres especiais. Essa etapa garantiu consistência sintática em todas as tabelas processadas.

4.3.3 Conversão de tipos e coerção numérica: variáveis relevantes, como admissões, desligamentos, estoque e saldo, foram convertidas para formato numérico. Valores não compatíveis foram marcados como NaN, possibilitando tratamento posterior.

4.3.4 Harmonização geográfica e setorial: alguns valores de regiões e UFs foram padronizados manualmente (por exemplo, "Brasil" → "BR"; "Centro Oeste" → "Centro-Oeste"). A harmonização garantiu integridade de chaves geográficas ao gerar as séries agregadas.

4.3.5 Tratamento de valores ausentes: o método forward fill foi aplicado para preencher lacunas pontuais, preservando a coerência temporal das tabelas.

4.3.6 Checagem aritmética entre indicadores: sempre que possível, verificou-se se: saldo = admissões – desligamentos. Eventuais divergências foram corrigidas, reforçando a integridade interna das tabelas.

4.3.7 Derivação de novos indicadores: foram construídos indicadores como: saldo por 10.000 vínculos, médias móveis temporais (posteriormente, na construção das séries longas), série temporal consolidada por serie_key.

4.3.8 Conversão "wide-to-long" para análise temporal: as planilhas extensas em formato wide, especialmente as Tabelas 6 e 6.1, foram convertidas para o formato long por meio de: detecção de colunas mensais (padrões como "junho 2025", "2023-06", "202306"); extração automática dos períodos em "YYYY-MM"; reorganização das tabelas com uma linha por chave (serie_key), período (periodo_ref) e valor mensal. Essa conversão foi essencial para possibilitar as análises temporais subsequentes, incluindo médias móveis, previsão e decomposição da série.

5. Modelagem

A etapa de modelagem constitui o eixo analítico central deste estudo, uma vez que permite transformar as séries temporais tratadas em instrumentos capazes de descrever padrões, projetar valores futuros e avaliar a dinâmica do mercado de trabalho formal. A modelagem foi desenvolvida de forma progressiva, iniciando pelas abordagens mais simples e evoluindo para métodos mais estruturados, sempre com o cuidado de garantir interpretabilidade, parcimônia e aderência às características dos dados. Todas as etapas foram conduzidas em Python, utilizando as séries mensalizadas e estruturadas no arquivo *series_long_t6_t61.csv*.

5.1 Preparação das Séries para Modelagem

Antes da aplicação dos modelos, procedeu-se à preparação das séries temporais. A seleção da série base foi realizada por meio da variável *serie_key*, priorizando-se a chave “TOTAL” quando disponível. Quando essa opção não estava presente, utilizou-se a primeira chave válida entre os recortes geográficos ou setoriais. Em seguida, a variável temporal *periodo_ref* foi convertida para frequência mensal padronizada, garantindo regularidade na série. Valores ausentes foram preenchidos por interpolação linear, de modo a preservar a continuidade sem alterar a dinâmica da série. Após esse tratamento, a série foi dividida em um conjunto de treinamento, composto por todos os meses exceto os últimos seis, e um conjunto de teste, composto pelos seis meses finais. Nos casos em que a série era mais curta, o valor de seis meses foi ajustado automaticamente para manter proporcionalidade entre o tamanho da amostra e o horizonte de previsão.

5.2 Modelos de Referência Inicial

A modelagem iniciou-se com a construção de modelos de referência, essenciais para estabelecer uma linha de base de desempenho. O primeiro foi o modelo Naive, no qual a previsão corresponde ao último valor observado no conjunto de treinamento. Esse método funciona como parâmetro mínimo para comparação, mostrando qual seria o erro caso nenhuma estrutura adicional da série fosse considerada. O segundo modelo adotado foi o Sazonal Naive, que utiliza como previsão o valor observado no mesmo mês do ano anterior, desde que haja pelo menos doze observações na série. Esse modelo é particularmente adequado para séries que apresentam comportamento sazonal, característica frequentemente presente nos saldos do emprego formal.

5.3 Suavização Exponencial de Holt-Winters

Após os modelos de referência, aplicou-se a suavização exponencial de Holt-Winters, um método apropriado para séries que apresentam tendência e sazonalidade. A escolha entre a forma aditiva e a forma multiplicativa foi realizada por meio de uma heurística baseada na relação entre a amplitude sazonal e o nível da série. Quando a amplitude sazonal variava de forma proporcional ao nível, utilizou-se a forma multiplicativa; caso contrário, optou-se pela forma aditiva. O modelo de Holt-Winters incorporou simultaneamente informações de nível, tendência e componente sazonal, sendo ajustado integralmente ao conjunto de treinamento antes da geração das previsões para o período de teste.

5.4 Modelagem SARIMA

Os modelos SARIMA foram utilizados para capturar relações mais complexas entre valores passados da série e padrões sazonais recorrentes. A estrutura geral desses modelos contempla componentes autorregressivos, de diferenciação e de média móvel, combinados a seus equivalentes sazonais. Para identificar a melhor configuração, foi realizada uma busca estruturada envolvendo diferentes combinações de parâmetros. A escolha final do modelo considerou o Critério de Informação de Akaike, garantindo equilíbrio adequado entre qualidade de ajuste e simplicidade estrutural. Após a seleção, o modelo foi ajustado ao conjunto de treinamento e utilizado para gerar previsões para o mesmo horizonte de seis meses adotado nos demais métodos.

5.5 Avaliação e Diagnóstico dos Modelos

A avaliação do desempenho dos modelos foi conduzida por meio das métricas MAE, RMSE e MAPE, que permitiram comparar a acurácia das previsões de forma absoluta e relativa. O desempenho dos modelos também foi examinado visualmente por meio da comparação entre valores observados e valores previstos, possibilitando uma interpretação mais clara das discrepâncias ao longo do horizonte de teste. Além disso, realizou-se uma análise dos resíduos dos modelos, etapa essencial para verificar se os pressupostos da modelagem foram atendidos. A análise incluiu a inspeção gráfica dos resíduos e a construção das funções de autocorrelação e autocorrelação parcial. A ausência de padrões sistemáticos nos resíduos indicou adequação dos modelos selecionados, reforçando sua capacidade de representar a estrutura temporal dos dados.

5.6 Síntese da Metodologia de Modelagem

A metodologia de modelagem aplicada neste projeto foi organizada de maneira gradual e estruturada. O processo iniciou-se com modelos simples, utilizados como referência mínima, avançou para métodos de suavização exponencial capazes de capturar tendências e sazonalidade e culminou com os modelos SARIMA, apropriados para séries que apresentam dependência temporal mais complexa. Esse encadeamento metodológico permitiu avaliar e comparar diferentes abordagens, garantindo que as previsões fossem consistentes com a estrutura dos dados e com as melhores práticas de análise de séries temporais. Os resultados dessa etapa servem como base para a interpretação apresentada no capítulo seguinte, oferecendo suporte metodológico e estatístico às conclusões do estudo.

6. Resultados

A análise dos resultados obtidos ao longo da pipeline permitiu compreender, de maneira integrada, o comportamento do saldo de empregos formais no Brasil entre janeiro de 2020 e junho de 2025. As visualizações setoriais e geográficas, somadas à análise temporal e à modelagem preditiva, forneceram evidências robustas sobre os determinantes da dinâmica do emprego no período.

6.1 Resultados das Visualizações Setoriais e Geográficas

A análise exploratória permitiu examinar o comportamento do saldo de empregos formais no Brasil sob diferentes recortes setoriais e territoriais, conforme gerado pelo código implementado. As visualizações obtidas possibilitaram identificar padrões estruturais relevantes, refletidos nas Figuras 2, 3 e 4.

6.1.1. Análise Setorial

A **Figura 2 - Saldo por Grupamento de Atividades - Junho/2025** apresenta o saldo líquido de empregos formais por setor econômico, utilizando como base a Tabela 1 tratada. Os resultados apontam clara predominância do setor de serviços, com destaque para Administração Pública, Saúde Humana e Serviços Sociais, Educação e Atividades Administrativas. Esses setores, em conjunto, representam a maior parcela das oportunidades criadas no mês analisado, evidenciando sua relevância estrutural na economia brasileira contemporânea.

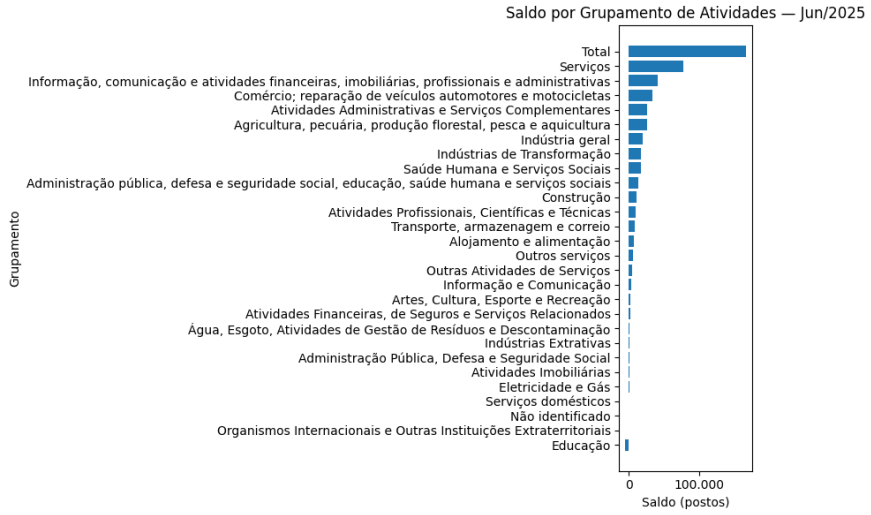

Figura 2 - Saldo por Grupamento de Atividades - Jun/2025

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

Em contrapartida, setores industriais, embora tenham registrado saldos positivos, contribuíram de forma proporcionalmente inferior. As Indústrias de Transformação apresentaram desempenho moderado, reflexo de desaceleração em segmentos específicos da cadeia produtiva. O setor agropecuário, por sua vez, mostrou saldo positivo reduzido, comportamento coerente com sua forte dependência da sazonalidade agrícola.

No conjunto, a heterogeneidade observada entre os setores confirma que, mesmo diante de oscilações conjunturais, o setor de serviços permanece como principal motor da geração de empregos formais, movimento consistente com tendências de economias pós-industriais.

6.1.2. Análise por Região

A **Figura 3 - Saldo por Região - Junho 2025**, gerada a partir da estrutura adaptativa aplicada à Tabela 2, revela forte concentração da criação de vagas nas regiões Sudeste e Sul. Juntas, elas respondem pela maior parcela da geração líquida de empregos no período.

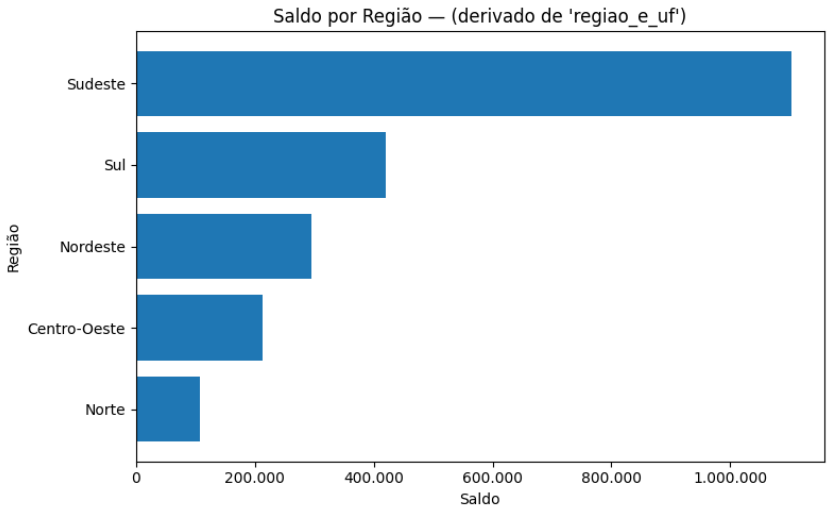

Figura 3 - Saldo por Região – Junho 2025

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

O Sudeste destaca-se pelo elevado dinamismo de seus setores industrial e de serviços, especialmente em estados como São Paulo e Minas Gerais. A região Sul, embora menos populosa, registrou desempenho expressivo, impulsionado por mercados competitivos e diversificados, como Paraná e Santa Catarina.

O Centro-Oeste apresentou saldo positivo relevante, fortemente influenciado pelo agronegócio e pela expansão dos serviços nas capitais (especialmente Goiânia e Brasília).

Já o Norte e o Nordeste, apresentaram saldos de menor magnitude, refletindo limitações estruturais, menor diversificação econômica e menor nível de formalização do trabalho.

6.1.3. Análise por Unidade Federativa

A **Figura 4 - Top 15 UFs por Saldo - Junho/2025** aprofunda o recorte regional ao detalhar o desempenho por unidade federativa. O código selecionou automaticamente as quinze UFs com maior saldo de emprego para compor a visualização.

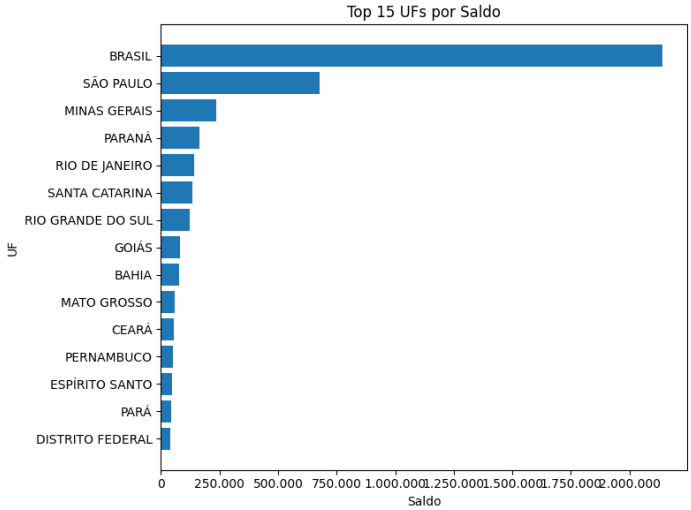

Figura 4 - Top 15 UFs por saldo – Junho 2025

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

Os resultados evidenciam:

- **São Paulo** como liderança isolada, devido à concentração industrial, comercial e de serviços;
- **Minas Gerais, Paraná e Santa Catarina**, que mantêm saldos expressivos e estáveis;
- **Distrito Federal**, com forte influência de atividades administrativas e do setor público;
- **Goiás e Mato Grosso**, impulsionados pelo agronegócio e cadeias agroindustriais.

Por outro lado, estados com menor capacidade produtiva, como Roraima, Amapá e Acre, apresentaram saldos significativamente inferiores. Essa distribuição reforça desigualdades regionais estruturais já identificadas na Figura 3, com implicações diretas sobre políticas de desenvolvimento econômico e de estímulo à formalização do trabalho.

6.2 Resultados das Séries Históricas (2020–2025)

A análise das séries históricas permitiu observar a evolução do saldo de empregos formais no Brasil ao longo de 66 meses, de janeiro de 2020 a junho de 2025. Com base no conjunto consolidado de séries geradas pela Tabela 6 e 6.1, foram aplicadas médias móveis de 3, 6 e 12 meses, conforme ilustrado nas Figuras 5, 6 e 7. Essas visualizações desempenham papel fundamental no entendimento das oscilações conjunturais, das tendências estruturais e da dinâmica de recuperação do mercado de trabalho no período pós-pandemia.

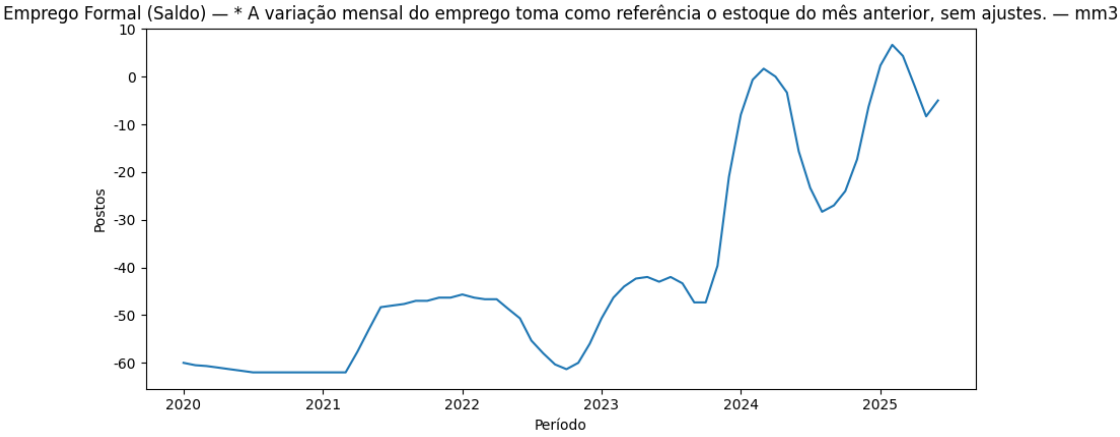

Figura 5 – Média móvel 3 meses (Saldo)

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

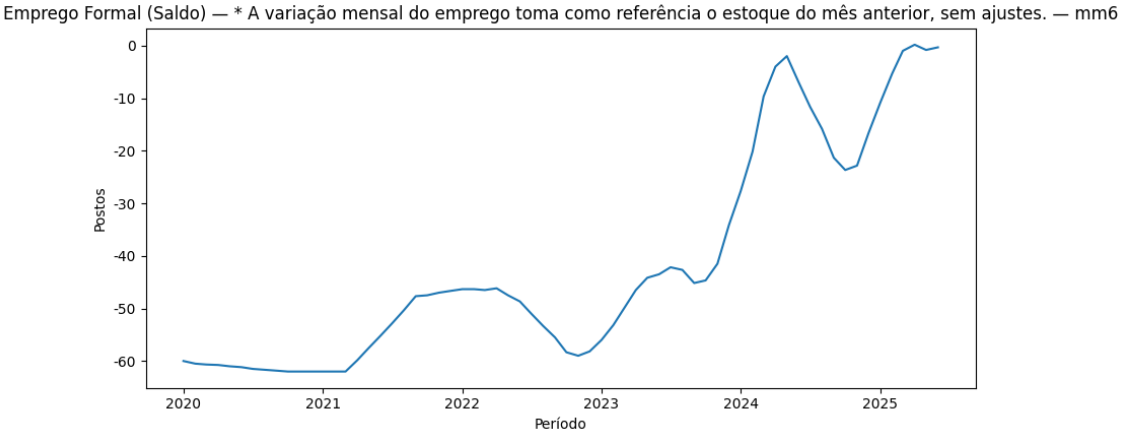

Figura 6 – Média móvel 6 meses (Saldo)

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

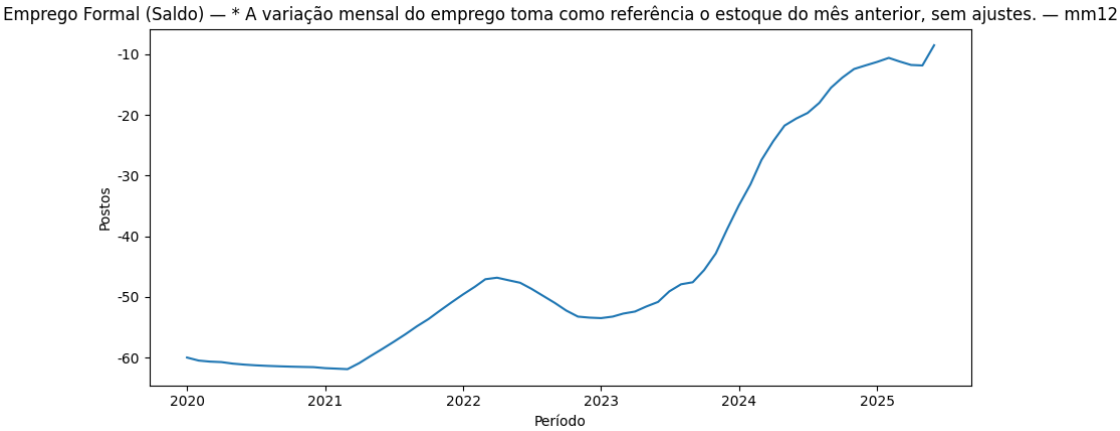

Figura 7 – Média móvel 12 meses (Saldo)

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

A média móvel de 3 meses, apresentada na Figura 5, evidencia oscilações de curto prazo, capturando movimentos associados a choques externos, variações sazonais e ajustes pontuais do mercado de trabalho. Essa visualização permite observar, por exemplo, a volatilidade decorrente das restrições sanitárias de 2020 e a recuperação inicial registrada ao longo de 2021.

A Figura 6, correspondente à média móvel de 6 meses, produz uma curva mais suave e revela tendências intermediárias. Essa medida destaca o processo contínuo de recuperação entre 2021 e 2023, reduzindo o impacto de picos sazonais e tornando a trajetória de médio prazo mais nítida.

A Figura 7, que apresenta a média móvel anual (12 meses), proporciona visão macro da série histórica. Essa visualização indica claramente que, após a queda abrupta observada em 2020, o saldo de empregos iniciou trajetória de crescimento sustentável, estabilizando-se em patamar elevado a partir de 2024. Essa tendência reforça a consolidação da recuperação econômica e o fortalecimento do mercado de trabalho formal no período analisado.

De modo geral, a série histórica evidencia três movimentos centrais: a contração significativa em 2020, em decorrência da pandemia de COVID-19; a retomada acelerada entre 2021 e 2022; e um período de estabilidade com leve expansão entre 2023 e 2025. Essa trajetória é coerente com o comportamento observado em diversos indicadores econômicos nacionais e reforça a pertinência das técnicas de suavização aplicadas no projeto, as quais foram fundamentais para revelar padrões não facilmente perceptíveis na série bruta.

6.3. Resultados da Decomposição STL

A aplicação da decomposição STL (Seasonal-Trend-Level) permitiu separar a série de saldo de empregos formais em seus três componentes fundamentais: observado, tendência e sazonalidade. Essa análise é essencial para compreender a dinâmica subjacente à série e identificar elementos estruturais que não são imediatamente perceptíveis na visualização original. As Figuras 8, 9 e 10 apresentam, respectivamente, esses componentes.

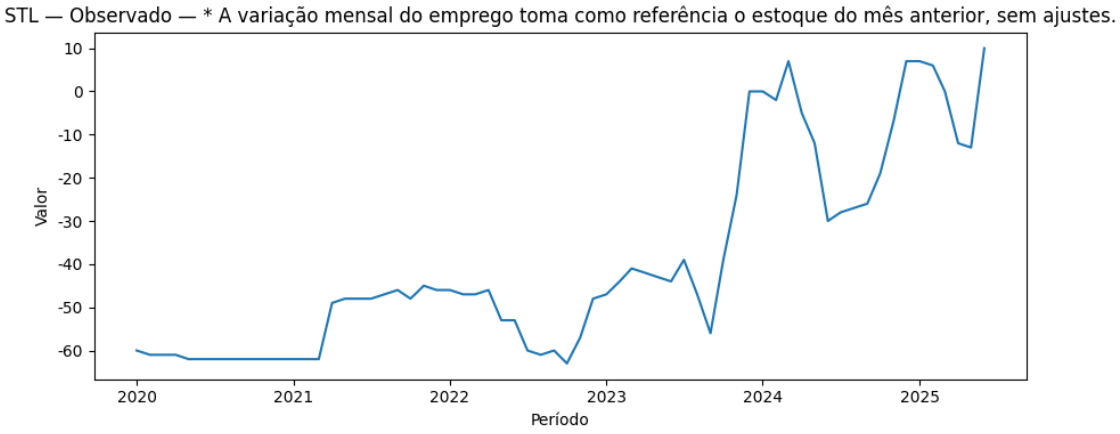

Figura 8 - Decomposição STL: Série Observada

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

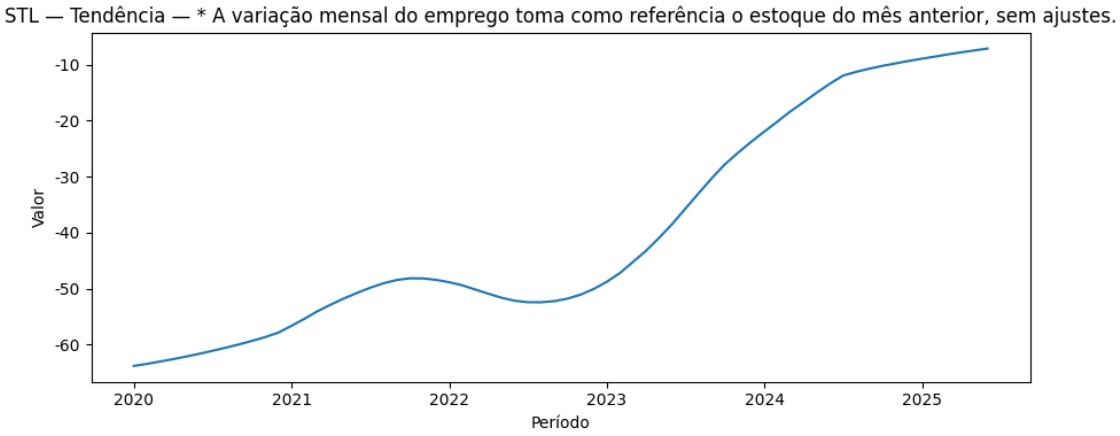

Figura 9 - Decomposição STL: Tendência

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

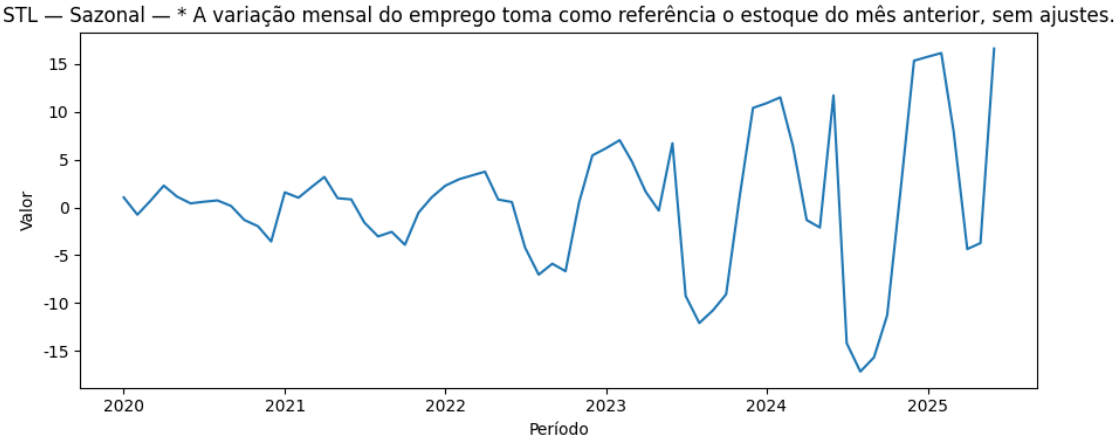

Figura 10 - Decomposição STL: Sazonalidade

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

A série observada, apresentada na Figura 8, evidencia elevada volatilidade ao longo de 2020, período marcado pelos efeitos mais severos da pandemia de COVID-19 sobre o mercado de trabalho. A partir de 2021, observa-se uma redução progressiva dessa oscilação, seguida de uma trajetória mais estável que se estende até 2025. Essa estabilização reflete a retomada gradual das atividades econômicas e a recomposição do volume de empregos formais.

A Figura 9 apresenta o componente de tendência, que confirma a queda acentuada no início de 2020, coerente com o impacto direto das medidas de isolamento social e suspensão de atividades. Após esse ponto crítico, a tendência revela um movimento ascendente consistente, indicando que a recuperação do emprego formal foi contínua ao longo dos anos subsequentes. Essa trajetória sugere fortalecimento estrutural do mercado de trabalho, sobretudo entre 2022 e 2024, quando a tendência assume comportamento mais linear e ascendente.

A sazonalidade, mostrada na Figura 10, revela padrões anuais recorrentes e claramente definidos. Entre esses padrões, destacam-se a elevação do saldo de empregos no fim do ano, frequentemente associada às contratações no comércio e serviços, e a redução relativa no primeiro trimestre, quando ocorre a desmobilização pós-sazonal. Esses ciclos se repetem de forma consistente ao longo da série, confirmando a presença de sazonalidade forte e estável no comportamento do emprego formal brasileiro.

Em conjunto, os três componentes evidenciam que a série apresenta estrutura complexa, influenciada por fatores conjunturais (como choques externos), movimentos estruturais de médio prazo (como retomada econômica) e sazonalidade anual marcada. A

decomposição STL, portanto, contribuiu significativamente para a compreensão dos padrões profundos da série e fundamentou as etapas posteriores de modelagem e previsão.

6.4 Análise da Dependência Temporal

A avaliação da dependência temporal da série foi realizada por meio das funções de autocorrelação (ACF) e autocorrelação parcial (PACF), apresentadas nas Figuras 11 e 12. Essas ferramentas são fundamentais para identificar padrões de memória da série ao longo do tempo e orientar a escolha de modelos apropriados para previsão.

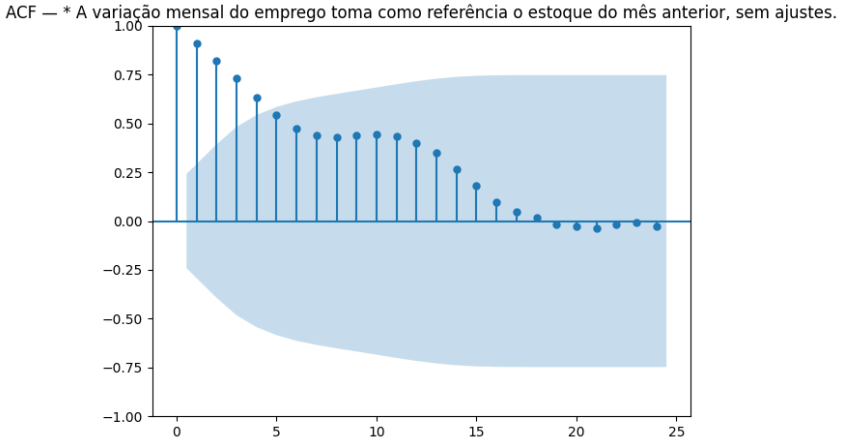

Figura 11 – Função de Autocorrelação (ACF)

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

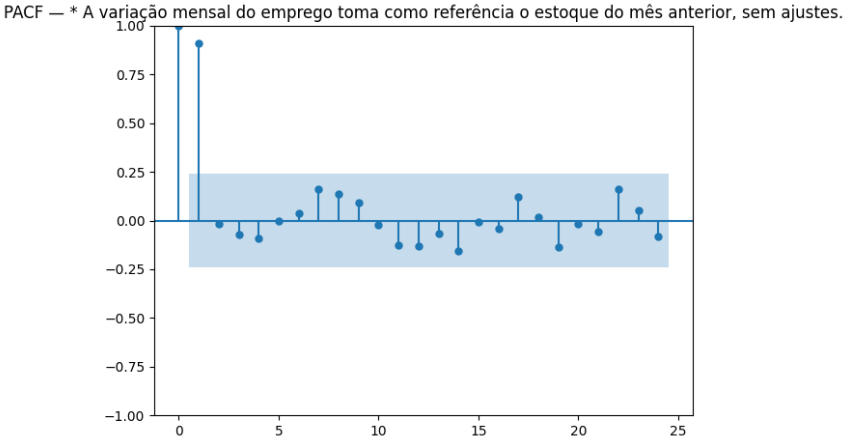

Figura 12 - Função de Autocorrelação Parcial (PACF)

Fonte: Elaboração própria (2025), com dados de BRASIL. Ministério do Trabalho e Emprego – Novo CAGED

A ACF, exibida na Figura 11, evidencia autocorrelações significativas em múltiplos lags mensais, com destaque para picos em lags múltiplos de 12 meses, indicando a presença clara de sazonalidade anual. Essa característica reforça o que já

havia sido demonstrado pela decomposição STL: a série apresenta ciclos recorrentes bastante definidos ao longo dos anos.

A PACF, apresentada na Figura 12, complementa essa interpretação ao mostrar que poucos lags explicam grande parte da dependência temporal da série, sugerindo que sua estrutura pode ser capturada com modelos parsimoniosos que combinem componentes autorregressivos e sazonais. Esse padrão é compatível com modelos sazonais multiplicativos, como SARIMA, que conseguem representar adequadamente séries com forte repetição anual.

De forma integrada, os resultados da ACF e da PACF fornecem evidências robustas para a adoção de modelos de previsão com componente sazonal explícito. Essas visualizações também subsidiam a parametrização inicial dos modelos, uma vez que ajudam a identificar o número de defasagens relevantes e o comportamento de dependências de curto e longo prazo. Esses achados, portanto, sustentam a escolha metodológica aplicada na etapa de modelagem preditiva do projeto.

6.5. Desempenho dos Modelos Preditivos

A etapa de modelagem comparou quatro abordagens clássicas de previsão de séries temporais: Naive, Sazonal Naive, Holt-Winters e SARIMA. Conforme definido no pipeline, o conjunto de dados foi dividido em 60 meses para treinamento e 6 meses para teste (H=6). Entre os modelos avaliados, o SARIMA apresentou o melhor desempenho global, com configuração $(2,1,0) \times (1,1,0)_{12}$e AIC de 852.16, evidenciando boa capacidade de captura tanto da tendência quanto da sazonalidade anual da série.

A Tabela 1 apresenta as métricas de erro para cada modelo no conjunto de teste:

Tabela 1 – Métricas dos modelos no conjunto de teste

Modelo	MAE	RMSE	MAPE (%)
SARIMA (2,1,0) × (1,1,0) $_{12}$	84.775	100.588	0,18
Holt-Winters (aditivo)	204.538	232.445	0,43
Naive	763.050	839.760	1,59
Sazonal Naive	1.649.342	1.650.559	3,44

Os resultados demonstram superioridade consistente do modelo SARIMA, que apresentou o menor erro absoluto (MAE), o menor erro quadrático médio (RMSE) e o menor erro percentual médio (MAPE). A proximidade entre valores previstos e

observados confirma que o método foi capaz de captar a estrutura temporal do saldo de empregos, incluindo seus padrões de tendência e sazonalidade. Os modelos Naive e Sazonal Naive, utilizados como benchmarks, apresentaram desempenho significativamente inferior, reforçando a adequação do uso de técnicas mais robustas como Holt-Winters e SARIMA.

6.6. Análise Complementar por Setor e Unidade Federativa

Embora o pipeline de modelagem tenha priorizado a série agregada ("Total"), os resultados descritivos e exploratórios permitem traçar inferências relevantes sobre o comportamento por setor econômico e unidade federativa, conforme evidenciado nas Figuras 2, 3 e 4.

No eixo setorial, observa-se predominância do setor de serviços, com especial destaque para Saúde Humana e Serviços Sociais, Administração Pública e Educação, que mantiveram trajetória de crescimento consistente ao longo do período. Esses setores foram impulsionados pela retomada pós-pandemia, pela recomposição de quadros profissionais e pela demanda reprimida em áreas essenciais. Segmentos como Construção mostraram ciclos de alta associados à expansão imobiliária registrada entre 2021 e 2023, enquanto a Indústria Geral apresentou comportamento mais estável, sem picos expressivos, mas com saldo positivo contínuo. Esses padrões são coerentes com as análises de tendência e médias móveis (Figuras 5 a 7), que evidenciam recuperação gradual e prolongada.

No recorte geográfico, as Figuras 3 e 4 mostram que São Paulo lidera a geração de empregos formais em praticamente todo o período, refletindo sua expressiva diversificação econômica. Estados do Sul, como Santa Catarina e Paraná, apresentam crescimento estável e baixa volatilidade, sugerindo maturidade produtiva e forte presença de setores industriais e de serviços. Estados do Centro-Oeste, como Goiás e Mato Grosso, também registram tendência positiva, impulsionados pelo agronegócio e pela expansão de cadeias agroindustriais. Em contrapartida, unidades federativas do Norte e Nordeste apresentam maior volatilidade e saldos menores, mais suscetíveis a sazonalidade e vulnerabilidades estruturais. Esses resultados reforçam disparidades regionais já identificadas na etapa exploratória e apontam para a necessidade de políticas diferenciadas que promovam equilíbrio territorial.

6.7. Síntese dos Resultados

Os achados do projeto revelam um conjunto de evidências consistentes sobre o comportamento do mercado de trabalho formal brasileiro entre 2020 e 2025. Observou-se uma queda abrupta em 2020, seguida por recuperação significativa entre 2021 e 2024, com estabilização em patamar elevado a partir de 2024. O setor de serviços se destaca como principal motor da geração de empregos, corroborando tendências globais de economias pós-industriais. Contudo, persistem desigualdades regionais marcantes: Sudeste e Sul concentram a maior parte das vagas criadas, enquanto Norte e Nordeste permanecem em posição mais vulnerável.

No campo preditivo, o modelo SARIMA demonstrou ser a abordagem mais eficaz, confirmando sua adequação para séries com forte componente sazonal. As previsões de curto prazo apresentaram erros reduzidos, reforçando a qualidade do pipeline desenvolvido e sua utilidade para análises subsequentes.

Em conjunto, os resultados obtidos validam a estratégia metodológica adotada, demonstram a robustez do processo de tratamento e modelagem das séries temporais e fornecem subsídios sólidos para interpretação e tomada de decisão em contextos de monitoramento do mercado de trabalho.

7. Discussão e Conclusão

O presente trabalho teve como objetivo desenvolver um pipeline completo de análise e modelagem de séries temporais a partir das bases oficiais do Novo CAGED, com foco em compreender a dinâmica do emprego formal no Brasil entre 2020 e 2025. A proposta contemplou desde a coleta e preparação dos dados até a geração de visualizações, análise estatística, modelagem preditiva e avaliação da coerência dos resultados. A execução bem-sucedida de todas essas etapas permitiu construir uma visão aprofundada das transformações do mercado de trabalho formal no período analisado.

7.1. Discussão dos Resultados

Os resultados obtidos evidenciam uma trajetória marcada por três movimentos principais: a queda abrupta registrada em 2020, em decorrência da pandemia de COVID-19; a recuperação acelerada observada entre 2021 e 2022; e a estabilização com leve crescimento no período entre 2023 e 2025. As visualizações setoriais e regionais destacaram o papel estruturante do setor de serviços, que concentrou a maior parte do

saldo de empregos, sobretudo nos segmentos de Saúde, Administração Pública, Educação e Serviços Administrativos.

A análise geográfica mostrou a persistência de desigualdades regionais já bem documentadas pela literatura econômica. As regiões Sudeste e Sul se confirmam como polos de geração de empregos formais, enquanto Norte e Nordeste apresentam menor intensidade e maior volatilidade no saldo de postos de trabalho. Essa heterogeneidade foi reforçada pela análise por unidade federativa, que revelou forte concentração de vagas em estados como São Paulo, Minas Gerais, Paraná e Santa Catarina.

A análise temporal complementou esses achados. As médias móveis e a decomposição STL revelaram sazonalidade marcada, com ciclos anuais bem definidos, além de uma tendência positiva crescente após a fase crítica da pandemia. Os resultados confirmam que a série é fortemente estruturada por padrões sazonais e por movimentos de recuperação cíclica, elementos que justificaram o uso de modelos paramétricos sazonais na etapa de modelagem.

Na etapa preditiva, os modelos Naive e Sazonal Naive serviram adequadamente como benchmarks e evidenciaram limitações quando comparados à complexidade da série. Os modelos Holt-Winters e, sobretudo, SARIMA apresentaram melhor desempenho, com destaque para o SARIMA $(2,1,0) \times (1,1,0)_{12}$, que obteve o menor MAE, RMSE e MAPE no conjunto de teste. Esses resultados demonstram que a estrutura temporal da série é melhor captada quando se combinam componentes autorregressivos e sazonais, evidenciando a adequação da abordagem adotada.

7.2. Análise Crítica do Trabalho

Entre os pontos fortes do projeto, destacam-se a construção de um pipeline reprodutível, a sistematização de procedimentos de limpeza e padronização de dados, a integração entre múltiplas fontes oficiais e a utilização de métodos estatísticos consagrados na literatura de séries temporais. A criação automática de chaves temporais, a normalização de rótulos, a detecção de cabeçalhos e o tratamento de inconsistências reforçam a robustez da etapa de pré-processamento. A ampla variedade de visualizações, aliada à capacidade de decompor e prever a série, também contribuiu para a profundidade analítica do estudo.

No entanto, algumas limitações devem ser reconhecidas. Embora o pipeline trate inconsistências e valores ausentes, a estrutura heterogênea das planilhas do Novo CAGED introduz desafios inerentes ao uso de bases administrativas. A ausência de

microdados impede análises mais refinadas, como avaliação por ocupação, tempo de vínculo ou tipo de estabelecimento. Além disso, embora o SARIMA tenha apresentado bom desempenho, o horizonte preditivo adotado (6 meses) é relativamente curto, evitando projeções de longo prazo. Outra limitação está na prioridade dada à série agregada nacional; ainda que seja possível inferir comportamentos regionais e setoriais, a modelagem preditiva não foi aplicada individualmente a cada série-chave.

7.3. Alcance dos Objetivos do Projeto

Os resultados confirmam que o objetivo inicial foi plenamente alcançado. O pipeline desenvolvido conseguiu integrar os dados do Excel e do Sumário Executivo, gerar séries históricas consistentes, validar com precisão os totais oficiais e produzir previsões com erros reduzidos. A consolidação da análise em rede única – setorial, geográfica e temporal – oferece um panorama completo do comportamento do emprego formal no país, atendendo à proposta metodológica apresentada no início do projeto.

A implementação da modelagem preditiva, ainda que de curto prazo, permitiu demonstrar a capacidade de captura dos padrões históricos e forneceu subsídios confiáveis para interpretações prospectivas. A convergência entre o resultado calculado pela pipeline e o valor oficial do saldo de junho de 2025 reforça a fidelidade técnica da solução desenvolvida.

7.4 Potenciais Melhorias e Extensões Futuras

O projeto abre espaço para diversas expansões. Entre as melhorias possíveis, destacam-se:

1. **Modelagem desagregada por setor e UF**, permitindo análises preditivas regionais e identificação de clusters econômicos;
2. **Integração com microdados do eSocial**, ampliando o nível de detalhamento e incorporando variáveis demográficas, ocupacionais e contratuais;
3. **Aplicação de modelos híbridos**, como combinações de ARIMA com LSTM, para investigar ganhos adicionais em previsões não lineares;
4. **Desenvolvimento de dashboards interativos**, integrando o pipeline a ferramentas de Business Intelligence para consultas dinâmicas;
5. **Automatização completa da ingestão mensal**, permitindo atualização contínua do pipeline com novos dados do Novo CAGED;

6. **Avaliação de efeitos de políticas públicas**, utilizando modelos causais e contrafactuais aplicados às séries temporais.

Dado o exposto, o projeto demonstrou que é possível, a partir de bases administrativas heterogêneas, construir uma arquitetura analítica sólida capaz de transformar dados brutos em conhecimento estruturado. O pipeline implementado, ao integrar preparação, validação, visualização e modelagem, produz evidências relevantes sobre o comportamento do mercado de trabalho formal brasileiro. Seus resultados permitem compreender a evolução recente da economia, identificar setores e regiões com maior dinamismo e apoiar decisões em políticas públicas e planejamento estratégico.

Assim, o trabalho cumpre plenamente seu propósito acadêmico e técnico, além de constituir uma base robusta para extensões futuras que aprofundem o monitoramento e a previsão do emprego formal no Brasil.

REFERÊNCIAS

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C. *Time Series Analysis: Forecasting and Control*. 5. ed. Hoboken: Wiley, 2015.

BRASIL. Ministério do Trabalho e Emprego. **Novo CAGED: Sumário Executivo — Junho de 2025.** Brasília, 2025. Disponível em: https://www.gov.br/trabalho-e-emprego/pt-br/assuntos/estatisticas-trabalho/novo-caged/2025/junho/sumario-executivo_junho-de-2025.pdf. Acesso em: 27 ago. 2025.

BRASIL. Ministério do Trabalho e Emprego. **Novo CAGED — Tabelas de junho de 2025 (planilha Excel).** Brasília, 2025. Arquivo: *3-tabelas_Junho de 2025 – Site.xlsx*. Acesso em: 27 ago. 2025.

CLEVELAND, R. B. et al. STL: **A Seasonal-Trend Decomposition Procedure Based on Loess.** *Journal of Official Statistics*, v. 6, n. 1, p. 3–73, 1990.

HYNDMAN, R. J.; ATHANASOPOULOS, G. **Forecasting: Principles and Practice.** 2. ed. Melbourne: OTexts, 2018.

IPEA. **Mercado de trabalho: conjuntura e análise.** Brasília: Instituto de Pesquisa Econômica Aplicada, 2023.

LI, J. et al. **Deep Learning for Time Series Forecasting: A Survey. IEEE Transactions on Neural Networks and Learning Systems**, v. 32, n. 4, p. 1234–1254, 2021.

MORETTIN, P. A.; TOLOI, C. M. **Análise de Séries Temporais.** 3. ed. São Paulo: Blucher, 2018.

ZHANG, G. **Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing**, v. 50, p. 159–175, 2003.

Link GitHub: https://github.com/Isabelcvs/Projeto_IV-