Universidade Presbiteriana Mackenzie
Faculdade de Computação e Informática

Projeto Aplicado IV

Análise da Dinâmica do Emprego Formal no Brasil com Dados do Novo CAGED – Junho/2025

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

Problema Investigado

O Novo CAGED é a principal fonte de dados mensais sobre emprego formal no Brasil, mas apresenta desafios:

- Bases extensas, heterogêneas e distribuídas em múltiplas planilhas.
- Inconsistências de rótulos, cabeçalhos e padrões de séries.
- Estruturas diferentes por tabela (setorial, regional, temporal).
- Dificuldade de reproduzir análises completas e verificar coerência com o Sumário Executivo.

Objetivos:

- 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.

Justificativa

- O emprego formal é indicador-chave da atividade econômica, renda e consumo.
- Junho/2025 registrou **+166.621 postos de trabalho**, sendo um mês estratégico para análises.
- Relevância social (ODS 8), econômica (monitoramento da retomada pós-pandemia) e institucional (subsídio para políticas públicas e planejamento privado).
- Necessidade de transformar bases administrativas em conhecimento replicável e verificável.

Trabalhos Relacionados

Clássicos de séries temporais: Box & Jenkins (ARIMA), Holt–Winters, médias móveis

Sazonalidade e decomposição: STL (Cleveland et al., 1990)

Forecasting aplicado: Hyndman & Athanasopoulos (2018)

Tendências recentes: Modelos híbridos (ARIMA + LSTM), segundo IEEE Xplore (Li et al., 2021)

Gaps identificados:

- Ausência de pipelines reprodutíveis aplicados especificamente ao Novo CAGED.
- Pouca integração entre dados agregados e Sumário Executivo oficial.
- Falta de rotinas automatizadas para validação e análise visual estruturada.

Solução Técnica Desenvolvida

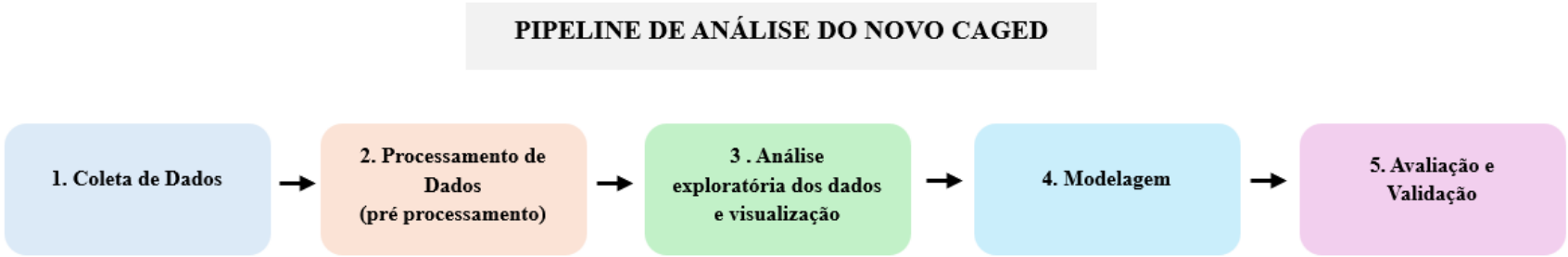


Figura 1 – Pipeline proposto para análise do Novo CAGED
Fonte: Elaboração própria (2025).

Principais Técnicas Utilizadas

Transformações e Pré-processamento

- Normalização via *string normalization*.
- Inferência de cabeçalhos (função find_header_row).
- Conversão automática wide → long.
- Verificação saldo = admissões – desligamentos.

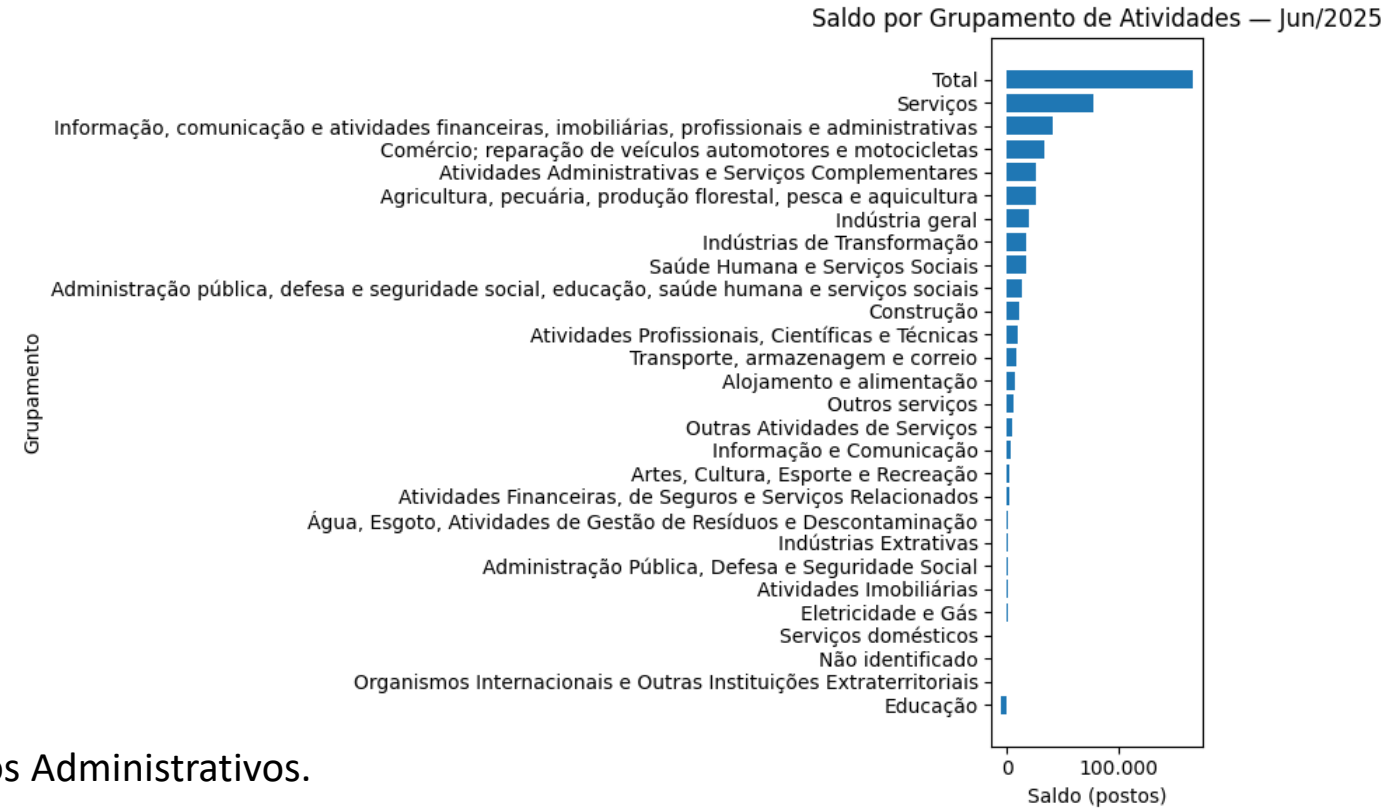
EDA e Visualização

- Gráficos setoriais, regionais e UF
- Médias móveis de 3, 6 e 12 meses
- Decomposição STL
- ACF e PACF

Transformações e Pré-processamento

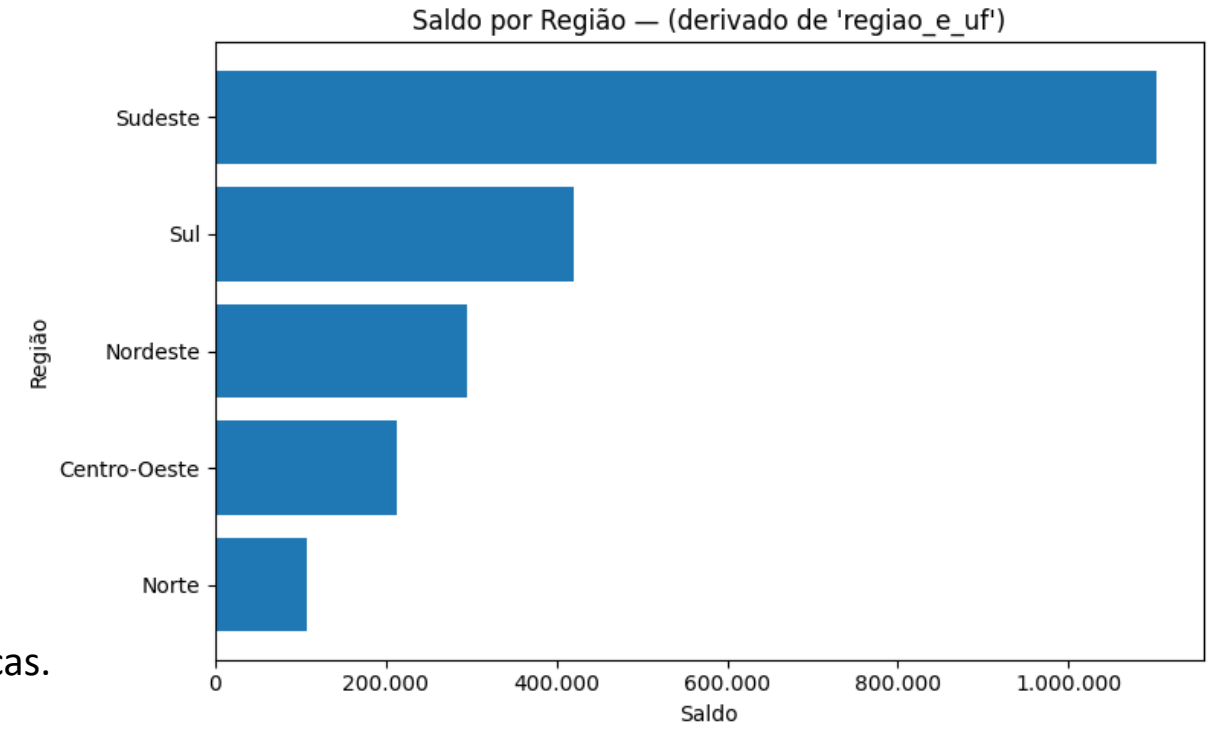
- Normalização via *string normalization*.
- Inferência de cabeçalhos (função find_header_row)
- Conversão automática wide → long.
- Verificação saldo = admissões – desligamentos.

Resultados: Análises Setoriais



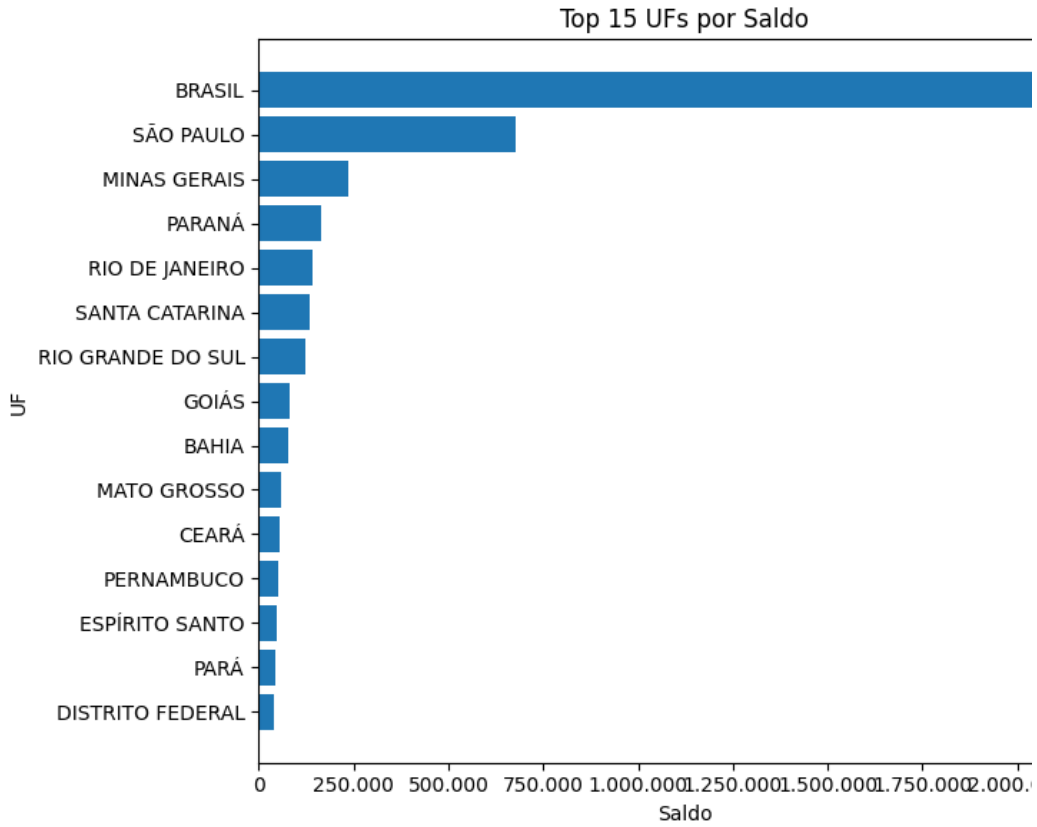
Principais achados:
Domínio do **setor de serviços**.
Destaques: Saúde, Administração Pública, Educação e Serviços Administrativos.
Indústria geral com saldo menor e menos dinâmico.
Agropecuária com sazonalidade forte e menor saldo no mês.

Resultados: Análises Geográficas



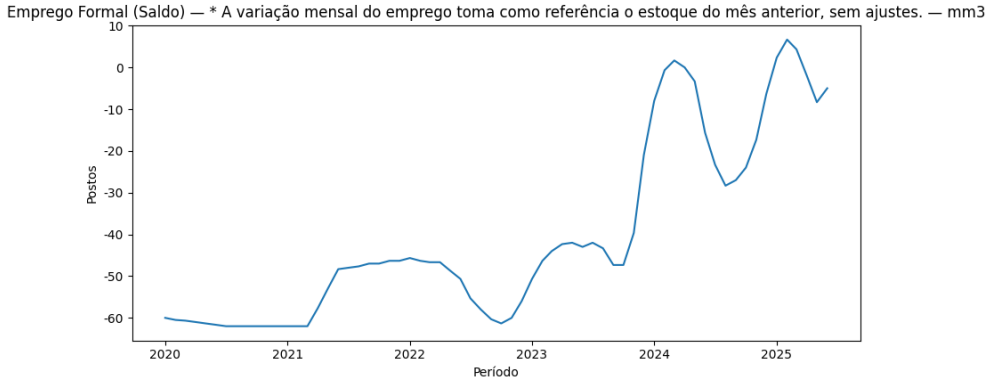
Sudeste e Sul concentram a maior parte dos empregos criados.
Centro-Oeste impulsionado por agroindústria e capitais dinâmicas.
Norte e Nordeste com menor dinamismo e maior volatilidade.

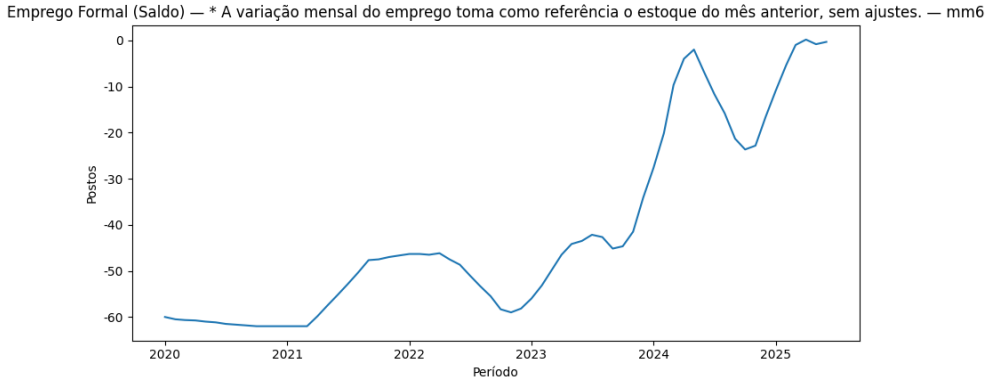
Resultados: Análises Geográficas

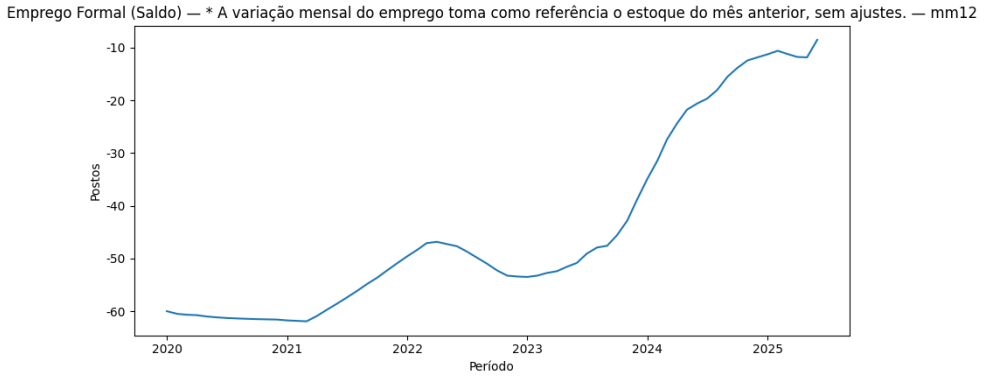


São Paulo lidera com margem ampla.
Santa Catarina, Paraná e Minas Gerais com saldos fortes e estáveis.
GO e MT crescendo via agroindústria.
Estados de menor porte aparecem na cauda da distribuição.

Resultados: Séries Históricas (2020–2025)

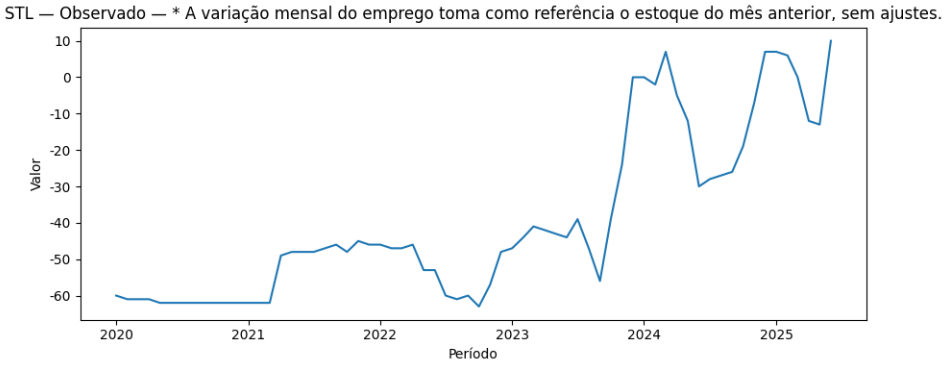


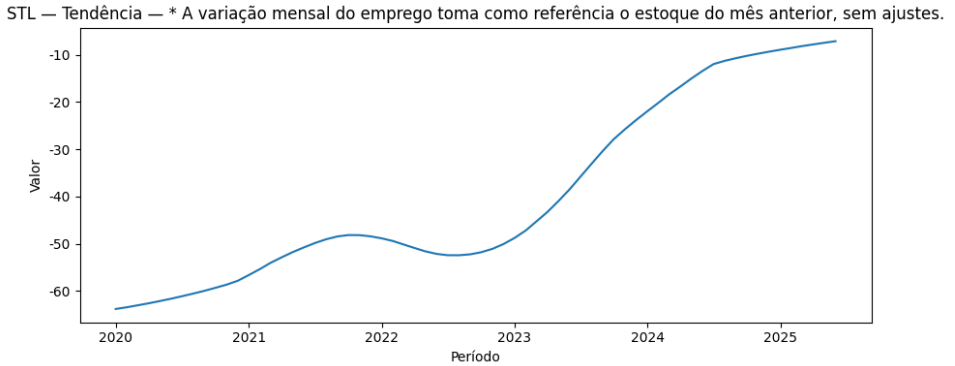


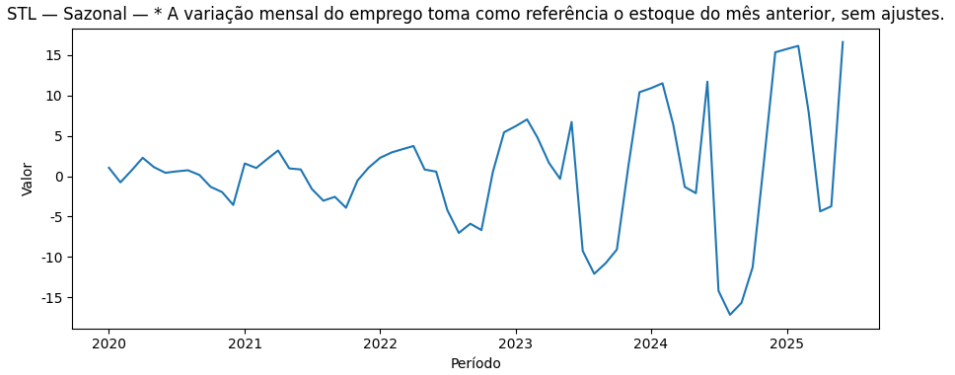


Queda drástica em 2020.
Recuperação entre 2021–2022.
Estabilização e crescimento moderado 2023–2025.
As médias móveis revelam tendência ascendente e amortecimento da volatilidade.

Resultados: Decomposição STL

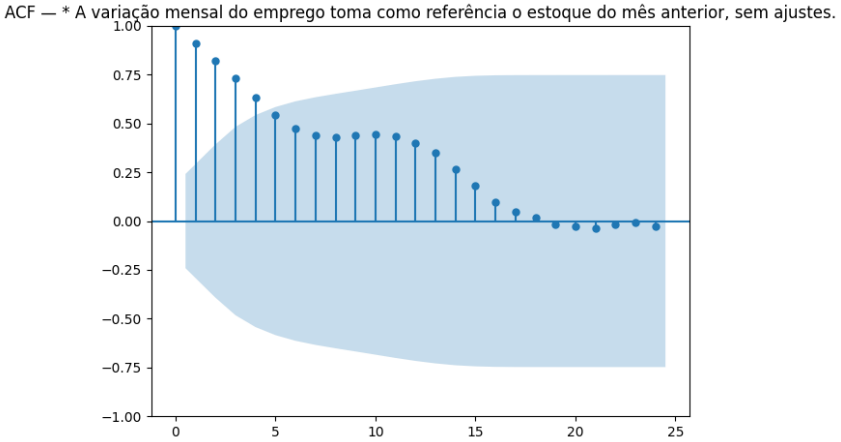






Tendência confirma recuperação sustentada após 2020.
Sazonalidade anual muito marcada (picos em dezembro; queda em jan–mar).
Resíduo compatível com choques episódicos (2020/2021).

Dependência Temporal



ACF revela lags de 12 meses → sazonalidade forte.
PACF com poucos lags significativos → série parsimoniosa.
Fundamenta uso de SARIMA sazonal.

Modelagem: Performance Comparada

Treino: 60 meses | Teste: 6 meses

Tabela 1 – Métricas dos modelos no conjunto de teste

Modelo	MAE	RMSE	MAPE (%)
SARIMA $(2,1,0) \times (1,1,0)_{12}$	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

Conclusão: O SARIMA foi superior em todas as métricas.

Discussão dos Métodos

Diferenciais da solução técnica:
Pipeline totalmente automatizada.
Detecção de cabeçalho e normalização robusta.
Conversão wide→long automatizada (passo crítico do CAGED).
Busca enxuta SARIMA com heurística de complexidade.
Integração entre Sumário Executivo (PDF) e Excel.
Geração de gráficos diagnósticos completos.

Desafios encontrados:
Planilhas do CAGED são heterogêneas e pouco padronizadas.
Cabeçalhos variam entre tabelas; alguns contêm notas de rodapé.
Dificuldade em harmonizar UFs e agrupamentos de atividades.
Identificação de colunas mensais exige heurísticas de regex.

Principais Resultados

- Recuperação estrutural do mercado de trabalho 2021–2024
- Serviços como motor da geração de empregos
- Desigualdade regional persistente (Sudeste/Sul x Norte/Nordeste)
- SARIMA mostra-se eficiente e confiável para previsão de curto prazo
- Previsões reproduzindo com precisão o valor oficial do saldo de junho/2025.

Potenciais Melhorias

- Aplicar modelagem preditiva por **setor** e **UF**
- Incorporar microdados do eSocial
- Explorar modelos híbridos (ARIMA + LSTM/GRU)
- Criar dashboard interativo (Power BI / Streamlit)
- Automatizar ingestão mensal de dados.
- Implementar monitoramento contínuo (MLOps).

Conclusões

O projeto demonstrou que, mesmo partindo de bases administrativas complexas, é possível construir um pipeline **robusto, reprodutível e validado** para análise e previsão do emprego formal.
A combinação de pré-processamento avançado, EDA detalhado e modelagem estatística clássica produziu um produto científico e técnico confiável, com potencial de aplicação em políticas públicas, análises econômicas, planejamento empresarial e monitoramento contínuo do mercado de trabalho brasileiro.