UNIVERSIDADE PRESBITERIANA MACKENZIE

PROJETO APLICADO II – CURSO CIÊNCIA DE DADOS

TURMA 201825166.000.03A – GRUPO 19

GUILHERME AUGUSTO LEAL OLIVEIRA

GUILHERME ROCHA DE SOUZA DUARTE GUILHERME SANTOS OLIVEIRA RICARDO ZULIAN DE SOUZA AMARAL

ESTUDO DE ATAS DO COPOM COMO FERRAMENTA PREDITORA

São Paulo 2025

TURMA 201825166.000.03A – GRUPO 19

GUILHERME AUGUSTO LEAL OLIVEIRA

GUILHERME ROCHA DE SOUZA DUARTE

GUILHERME SANTOS OLIVEIRA

RICARDO ZULIAN DE SOUZA MARAL

ESTUDO DE ATAS DO COPOM COMO FERRAMENTA PREDITORA

Projeto aplicado apresentado à Universidade Presbiteriana Mackenzie como requisito parcial para conclusão da disciplina Projeto Aplicado II.

Orientador: Professor Felipe Albino dos Santos

São Paulo 2025

# - SUMÁRIO

[1 - SUMÁRIO 3](#_Toc210300458)

[2 - TABELAS, QUADROS E FIGURAS 3](#_Toc210300459)

[2.1- QUADROS 3](#_Toc210300460)

[2.2– FIGURAS 4](#_Toc210300461)

[3 - TERMOS CHAVE 4](#_Toc210300462)

[4 - GLOSSÁRIO 4](#_Toc210300463)

[5- RECURSOS EXTERNOS 4](#_Toc210300464)

[6- INTRODUÇÃO 4](#_Toc210300465)

[7- A EMPRESA 4](#_Toc210300466)

[8 – O COPOM 5](#_Toc210300467)

[9 – OBJETIVO 5](#_Toc210300468)

[9.1 PRODUÇÃO DE RESUMOS OU SUMÁRIOS 5](#_Toc210300469)

[9.2 PRODUÇÃO DE INDICADORES 5](#_Toc210300470)

[10 - A BASE DE DADOS 6](#_Toc210300471)

[11 - ANÁLISE EXPLORATÓRIA 8](#_Toc210300472)

[11.1 - AQUISIÇÃO DE DADOS E ESTRUTURA INICIAL 8](#_Toc210300473)

[11.2 - ENGENHARIA DE VARIÁVEIS E ESTATÍSTICAS DESCRITIVAS 9](#_Toc210300474)

[11.3 - ANÁLISE GRÁFICA E DEFINIÇÃO DO PERÍODO DE ESTUDO 9](#_Toc210300475)

[11.4 - DELIMITAÇÃO E FILTRAGEM DO CONJUNTO DE DADOS 10](#_Toc210300476)

[11.5 - ANÁLISE GRÁFICA DO SUBCONJUNTO DELIMITADO 11](#_Toc210300477)

[11.6 - FORMAÇÃO DA BASE DE DADOS E AQUISIÇÃO DO TEXTO INTEGRAL 12](#_Toc210300478)

[11.7 - ANÁLISE ESTATÍSTICA DA EXTENSÃO DOS TEXTO 13](#_Toc210300479)

[11.8 - AQUISIÇÃO E ANÁLISE DA SÉRIE HISTÓRICA DO IPCA 14](#_Toc210300480)

[11.9 - AQUISIÇÃO E ANÁLISE DA SÉRIE HISTÓRICA DA TAXA SELIC 15](#_Toc210300481)

[11.10 - LIMPEZA E NORMALIZAÇÃO DO CONTEÚDO TEXTUAL (PLN) 17](#_Toc210300482)

[11.11 - CONSOLIDAÇÃO FINAL DOS DADOS E ENGENHARIA DA VARIÁVEL ALVO 18](#_Toc210300483)

[12 - EMBASAMENTO 19](#_Toc210300471)

[12.1 – DEFINIÇÃO DA LINGUAGEM DE PROGRAMAÇÃO 19](#_Toc210300485)

[12.2 – BASE TEÓRICA E MÉTODOS 20](#_Toc210300486)

[12.3 – CÁLCULO DE ACURÁCIA 20](#_Toc210300487)

[13 - MODELAGEM E RESULTADOS 20](#_Toc210300488)

[14 – STORYTELLING 20](#_Toc210300489)

[15 - REFERÊNCIAS 20](#_Toc210300490)

# - TABELAS, QUADROS E FIGURAS

## 2.1- QUADROS

Quadros 1 a 6 – Estrutura da Base de Dados 6,7

## 2.2– FIGURAS

Figura 1 – Acesso às APIs 8

Figura 2 – Prazo de Publicação 9

Figura 3 – Gráfico do Prazo de Publicação 9

Figura 4 – Boxplot do Prazo de Publicação 10

Figura 5 – Filtragem dos Dados 11

Figura 6 – Gráfico do Prazo de Publicação Filtrado 12

Figura 6 – Boxplot do Prazo de Publicação Filtrado 12

# 3 - TERMOS CHAVE

BACEN, COPOM, SELIC, INFLAÇÃO.

# - GLOSSÁRIO

**BACEN** – Abreviatura utilizada no mercado financeiro pra Banco Central do Brasil, órgão federal responsável pela política monetária nacional.

**COPOM** – Comitê de política monetária. Formado pelos diretores do BACEN, se reúne regularmente para definir os parâmetros da política monetária nacional.

**SELIC** – A taxa de juros básica da economia nacional.

**INFLAÇÃO** – Medida estatística da variação de preços de produtos e serviços no mercado nacional.

# 5- RECURSOS EXTERNOS

Os documentos e o código desenvolvidos para a realização deste estudo podem ser encontrados no Github.

Segue o repositório: <https://github.com/guilhermersduarte/Projeto-Aplicado2-Grupo19>

# 6- INTRODUÇÃO

Este projeto que tem por objetivo compilar e processar os dados produzidos pelas reuniões periódicas do COPOM e através de técnicas de processamento de linguagem natural criar indicadores para provisão de comportamento de juros e inflação.

A empresa escolhida como usuária da informação é o Banco do Brasil. Antecipar corretamente o comportamento da inflação e das taxas de juros tem grande valor nas operações no mercado financeiro e no gerenciamento de carteiras de crédito.

Ferramentas em Python e bibliotecas de processamento de linguagem natural serão utilizadas para capturar, processar e analisar os produtos do COPOM e dados históricos de taxa de juros e inflação.

# 7- A EMPRESA

O Banco do Brasil (BB) foi fundado em 1808 e é o primeiro banco a operar no Brasil e um dos primeiros da américa latina. Criado por Dom João VI, incialmente operava como banco emissor e financiador do governo. Hoje é um dos maiores bancos no país, sendo uma empresa de economia mista, com o governo federal sendo seu maior acionista.

O Banco do Brasil é reconhecido pela excelência na gestão de recursos financeiros e oferta de produtos de investimento. A tesouraria atua na administração de liquidez, estão de riscos financeiros e negociação de ativos no mercado financeiro. Sua carteira de crédito é uma das maiores no país, com destaque ao financiamento do agronegócio. O BB também to uma gestora de recursos de terceiros, a BB Asset Managemente, com fundos de renda fixa, ações e multimercado.

# 8 – O COPOM

O COPOM se reúne a cada 45 dias para definir a taxa básica de juros vigente no Brasil pelos próximos 45 dias.

Na reunião se define a taxa SELIC, que a taxa de juros média praticada nos títulos da dívida federal de um dia útil. Influencia todas as outras taxas de juros praticadas no Brasil, de Títulos Federais de longo prazo ao rotativo do cartão de crédito.

A taxa de juros é a ferramenta utilizada no controle da inflação. Uma inflação elevada indica uma economia aquecida e excesso de dinheiro em circulação. Nesse cenário o COPOM eleva a taxa de juros, tornando o crédito mais caro e a poupança mais interessante, reduzindo assim a quantidade de dinheiro em circulação e consequentemente a evolução dos preços. Uma inflação em queda acentuada ou até uma deflação (queda de preços) indicam economia estagnada e falta de dinheiro. Uma redução da taxa de juros busca aumentar a quantidade de dinheiro em circulação e assim aquecer a economia.

Cada reunião tem três produtos:

-A **taxa** propriamente dita, expressada em percentual (na data da produção deste documento a taxa selic era de 15% ao ano);

-Um **comunicado**, publicado ao final da reunião. O comunicado tem um título informativo – já apresenta a decisão no próprio título – e uma justificativa curta, de uma lauda, justificando essa decisão. Apresenta também o resultado da votação do comitê, destacando os membros favoráveis e contrários à decisão.

-Uma **ata**, publicada uma semana depois da decisão. Documento mais extenso, detalhando os fatores o processo observado na tomada da decisão.

# 9 – OBJETIVO

As atas e comunicados tem um formato estrito e a linguagem utilizada é recorrente. Pequenas nuances em terminologia indicam intenções e expectativas. Durante o desenvolvimento dos sistemas da Biruta serão exploradas duas possibilidades envolvendo o processamento de linguagem natural:

## 9.1 Produção de resumos ou sumários

Utilizar técnicas de processamento de linguagem natural para agilmente apresentar versões concisas e com terminologia acessível dos documentos produzidos pelo COPOM.

## 9.2 Produção de indicadores

Rotulando cada decisão com o comportamento da economia (inflação) e do próprio COPOM (taxa de juros) em períodos subsequentes desenvolver um modelo de classificação e previsão dessas variáveis em prazos a serem estipulados.

# 10 - A BASE DE DADOS

As bases primárias de dados são obtidas diretamente dos sistemas do BACEN via API. Aqui os 4 pontos de consulta e a informação fornecida:

- Lista de atas do COPOM

[https://www.bcb.gov.br/api/servico/sitebcb/copom**/atas**](https://www.bcb.gov.br/api/servico/sitebcb/copom/atas)

Tem registros a partir de 28/01/1998

|  |  |  |  |
| --- | --- | --- | --- |
| Nome do Campo | Tipo do Campo | Descrição | Exemplo |
| nroReuniao | integer | 255 | 255 |
| dataReferencia | string<date> | Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21) | 2023-06-21 |
| dataPublicacao | string<date> | Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21) | 2023-06-21 |
| titulo | string | Título da publicação | 255ª Reunião - 20-21 junho, 2023 |

- Detalhes sobre uma ata do COPOM

[https://www.bcb.gov.br/api/servico/sitebcb/copom**/atas\_detalhes**](https://www.bcb.gov.br/api/servico/sitebcb/copom/atas_detalhes)

Tem registros a partir de 28/01/1998

|  |  |  |  |
| --- | --- | --- | --- |
| Nome do Campo | Tipo do Campo | Descrição | Exemplo |
| nroReuniao | integer | 255 | 255 |
| dataReferencia | string<date> | Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21) | 2023-06-21 |
| dataPublicacao | string<date> | Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21) | 2023-06-21 |
| titulo | string | Título da publicação | 255ª Reunião - 20-21 junho, 2023 |
| urlPdfAta | string | Caminho completo para acesso ao arquivo PDF do relatório. Disponível só a partir da reunião de número 200 | https://www.bcb.gov.br/content/copom/atascopom/ Copom255-not20230621255.pdf |
| textoAta | string | Conteúdo da ata, em formato texto com tags html | <div id=\"atacompleta\"><div id=\"ataconteudo\"><h3 class=\"secao\">A) Atualização da conjuntura econômica e do cenário do ... |

- **Lista de comunicados do COPOM** [https://www.bcb.gov.br/api/servico/sitebcb/copom**/comunicados**](https://www.bcb.gov.br/api/servico/sitebcb/copom/comunicados)

Tem registros a partir de 19/04/2000.

|  |  |  |  |
| --- | --- | --- | --- |
| Nome do Campo | Tipo do Campo | Descrição | Exemplo |
| nroReuniao | integer | 255 | 255 |
| dataReferencia | string<date> | Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21) | 2023-06-21 |
| titulo | string | Título da publicação | 255ª Reunião - 20-21 junho, 2023 |

- Detalhes sobre um comunicado do COPOM

[https://www.bcb.gov.br/api/servico/sitebcb/copom**/atas\_detalhes**](https://www.bcb.gov.br/api/servico/sitebcb/copom/atas_detalhes)

Tem registros a partir de 19/04/2000

|  |  |  |  |
| --- | --- | --- | --- |
| Nome do Campo | Tipo do Campo | Descrição | Exemplo |
| nroReuniao | integer | 255 | 255 |
| dataReferencia | string<date> | Data do último dia da reunião do Copom, com formato conforme especificação RFC-3339 (ex: 2023-06-21) | 2023-06-21 |
| titulo | string | Título do comunicado | Copom mantém a taxa Selic em 13,75% a.a. |
| textoComunicado | string | Conteúdo do comunicado, em formato texto com tags html | <p style=\"text-align&#58;justify;\">Em sua 252ª reunião, o Comitê de Política Monetária (Copom) decidiu manter a taxa Selic ... |

- Histórico de taxas de juros SELIC

<https://api.bcb.gov.br/dados/serie/bcdata.sgs.432/dados?formato=json&dataInicial=01/01/2016&dataFinal=31/12/2025>

Tem registros diários a partir de 05/03/1999.

|  |  |  |  |
| --- | --- | --- | --- |
| Nome do Campo | Tipo do Campo | Descrição | Exemplo |
| data | string<date> | Apresentação da taxa anual diariamente. | 21/06/203 |
| valor | float | Taxa anual no dia | 15.00 |

- Histórico de taxas de inflação IPCA

<https://api.bcb.gov.br/dados/serie/bcdata.sgs.433/dados?formato=json>

Tem registros mensais a partir de

|  |  |  |  |
| --- | --- | --- | --- |
| Nome do Campo | Tipo do Campo | Descrição | Exemplo |
| data | string<date> | Apresentação da taxa mensal mensalmente | 21/06/203 |
| valor | float | Taxa no mês anterior | 15.00 |

# 11 - ANÁLISE EXPLORATÓRIA

A base de dados selecionada para o projeto é composta pelas Atas das Reuniões do Comitê de Política Monetária (COPOM), obtidas diretamente através da API pública do Banco Central do Brasil (BACEN).

## 11.1 - AQUISIÇÃO DE DADOS E ESTRUTURA INICIAL

A análise exploratória (AED) iniciou-se com a fase de aquisição de dados, utilizando a biblioteca requests para realizar uma requisição HTTP GET à *URL* da API do COPOM. A resposta em formato JSON foi então processada e estruturada em um **Pandas DataFrame**.

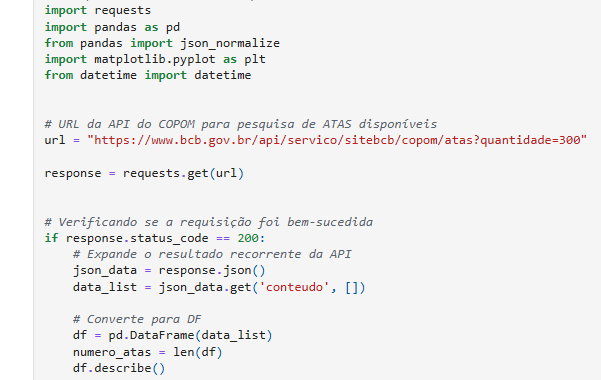


Figura 1

import requests e import pandas: Importação das bibliotecas para requisições *web* e manipulação de dados.

* **url:** Define o *endpoint* da API para as atas, solicitando até 300 registros.
* **response.status\_code == 200:** Garante que a requisição foi bem-sucedida.
* **pd.DataFrame(data\_list):** Converte a lista de dicionários (data\_list) extraída do JSON da API em uma estrutura de dados tabular (df).

O *dataset* inicial contém **253 registros**, hoje a lista tem 253 registros, com a primeira reunião em 28 de janeiro de 1998 e a última em 17 de setembro de 2025, a primeira publicação ocorreu em 22 de maio de 1998 e a última em 23 de setembro de 2025.

As colunas chave incluem nroReuniao, dataReferencia, dataPublicacao e título

(que implicitamente remete ao conteúdo da Ata).

## 11.2 - ENGENHARIA DE VARIÁVEIS E ESTATÍSTICAS DESCRITIVAS

Para aprofundar a AED, as colunas de datas foram convertidas para o tipo datetime e foi calculada uma *feature* temporal: o prazo de publicação.

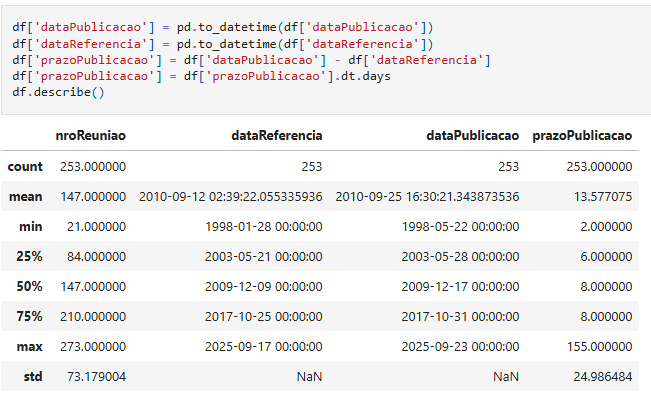


Figura 2

## 11.3 - ANÁLISE GRÁFICA E DEFINIÇÃO DO PERÍODO DE ESTUDO

A análise da *feature* prazoPublicacao foi aprofundada por meio de visualizações e estatísticas específicas para identificar a distribuição e o comportamento temporal do prazo de divulgação.

Gráfico, Gráfico de linhas

O conteúdo gerado por IA pode estar incorreto.

Figura 3

Gráfico, Gráfico de dispersão

O conteúdo gerado por IA pode estar incorreto.

Figura 4

Prazo mínimo: 2

Prazo médio: 13.577075098814229

Prazo mediano: 8.0

Prazo máximo: 155

## 11.4 - DELIMITAÇÃO E FILTRAGEM DO CONJUNTO DE DADOS

Após a constatação de uma quebra estrutural na variável prazoPublicacao coincidente com a adoção do **Regime de Metas de Inflação** em junho de 1999, o *dataset* foi filtrado para garantir que a análise se concentre apenas no período sob o regime monetário atual.

Código Explicado (Filtragem do DataFrame):

O código a seguir cria o novo *DataFrame* (df1999), retendo apenas os registros cuja dataReferencia é posterior a **1º de junho de 1999**. Em seguida, são geradas novas estatísticas descritivas para o subconjunto filtrado.

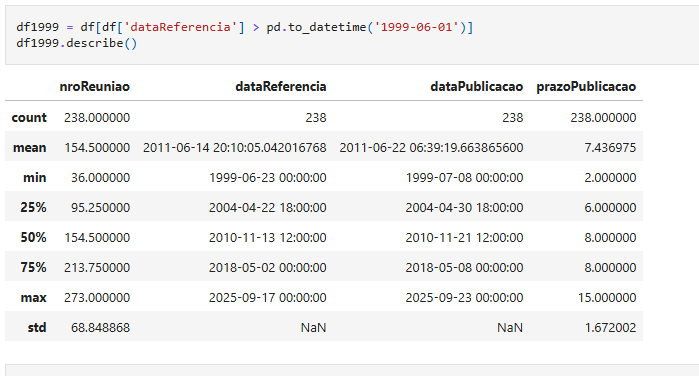


Figura 5

**Validade do Estudo:** Ao focar nos **238 registros** posteriores a junho de 1999, o estudo garante que o conteúdo textual das Atas (a ser usado como variável preditora) está contextualizado sob o **regime de política monetária moderna** (Metas de Inflação), aumentando a validade e a relevância das futuras análises.

## 11.5 - ANÁLISE GRÁFICA DO SUBCONJUNTO DELIMITADO

Para confirmar a homogeneidade do período selecionado e validar a eliminação da quebra estrutural, foram geradas novas visualizações e estatísticas focadas apenas no *DataFrame* pós-1999 (df1999).

**Gráfico de Linha:** Diferentemente da série original, o novo gráfico demonstra que a variável prazoPublicacao opera dentro de uma faixa estreita de valores (entre 2 e 15 dias). A alta volatilidade e os picos de longa duração (acima de 100 dias) foram eliminados, indicando um **comportamento estável e previsível**.

Gráfico, Gráfico de caixa estreita

O conteúdo gerado por IA pode estar incorreto.

Figura 6

**Boxplot:** O *boxplot* confirma a **baixa dispersão** dos dados. O *range* interquartil (IQR) é pequeno, e os *outliers* remanescentes (até 15 dias) são significativamente menos extremos em comparação ao *dataset* completo.

Gráfico, Gráfico de caixa estreita

O conteúdo gerado por IA pode estar incorreto.

Figura 7

Prazo mínimo: 2

Prazo médio: 7.436974789915967

Prazo mediano: 8.0

Prazo máximo: 15

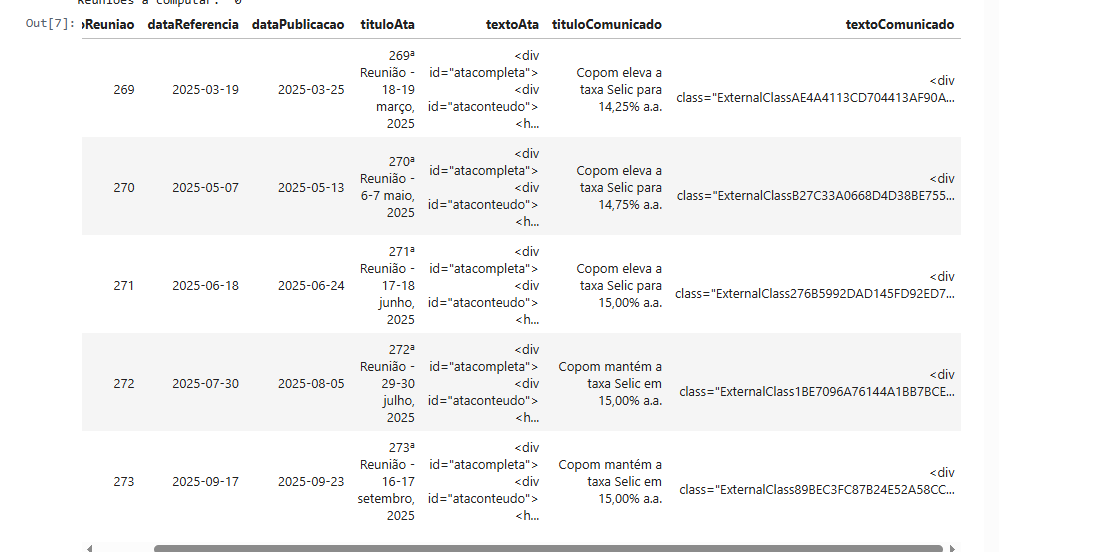
**Estatísticas:** A **Média (7.44 dias)** e a **Mediana (8.0 dias)** são muito próximas, sugerindo uma distribuição mais simétrica e menos distorcida pelos valores extremos (já removidos).

**Conclusão da AED:** O *dataset* df1999 está agora **homogêneo** e **pronto** para a próxima fase do **Tratamento de Dados**, que é a extração de *features* textuais e a preparação do conjunto de treinamento e teste.

## 11.6 - FORMAÇÃO DA BASE DE DADOS CONSOLIDADA E AQUISIÇÃO DO TEXTO INTEGRAL

A base filtrada (df1999) contém 238 registros válidos, iniciados em junho de 1999, cobrindo todas as reuniões sujeitas à regulamentação atual do Regime de Metas de Inflação. Nesta etapa, foi realizada a **aquisição programática** dos textos completos das Atas e dos Comunicados para cada reunião, utilizando os *endpoints* específicos da API do Banco Central.

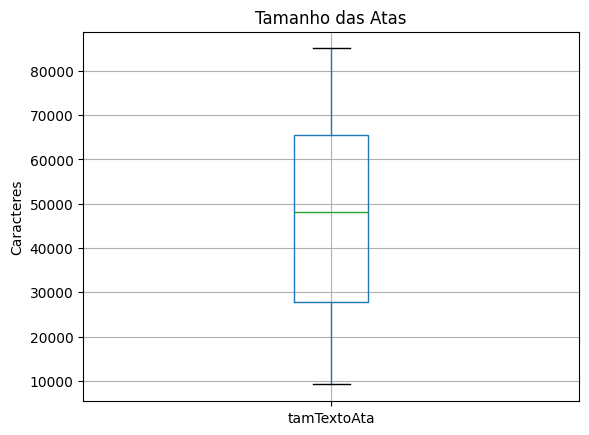
O processo iterativo abaixo utiliza o número de cada reunião (nroReuniao) para fazer duas requisições HTTP distintas: uma para obter o texto integral da Ata e outra para o Comunicado. Os dados textuais (textoAta e textoComunicado) são então consolidados em um novo *DataFrame* (df\_copom), que será a base primária para o Processamento de Linguagem Natural (PLN).



**Análise:** O sucesso na aquisição dos dados é confirmado pelo preenchimento das colunas textoAta e textoComunicado. É notável, contudo, que o texto bruto contém tags HTML (<div id="atacompleta">..., <div class="ExternalClass...">...). Este é um problema de ruído que será tratado na próxima fase: o **Processamento de Linguagem Natural (PLN)**.

## 11.7 - ANÁLISE ESTATÍSTICA DA EXTENSÃO DOS TEXTO

Com o *DataFrame* df\_copom consolidado e contendo o texto integral das Atas e Comunicados, a próxima fase da Análise Exploratória focou na inspeção das *features* textuais. Esta análise preliminar da extensão (número de caracteres) é vital para planejar as etapas subsequentes de limpeza, tokenização e vetorização no Processamento de Linguagem Natural (PLN).

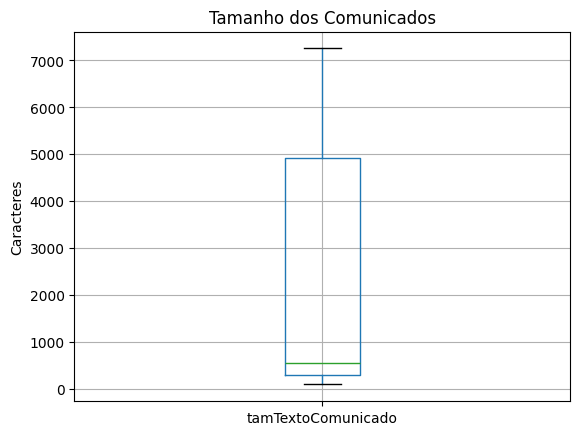


Tamanho mínimo das atas: 9177.0

Tamanho máximo das atas: 85056.0

Tamanho médio das atas: 46994.7572815534

Tamanho mediano das atas: 48215.0



Tamanho mínimo dos comunicados: 99.0

Tamanho máximo dos comunicados: 7254.0

Tamanho médio dos comunicados: 2098.5526315789475

Tamanho mediano dos comunicados: 560.5

## 11.8 - AQUISIÇÃO E ANÁLISE DA SÉRIE HISTÓRICA DO IPCA

A fase de tratamento de dados exige a integração de variáveis macroeconômicas cruciais ao lado das *features* textuais. Esta subseção detalha a aquisição e a análise da série histórica do **IPCA (Índice Nacional de Preços ao Consumidor Amplo)**, que será utilizada como uma **variável preditora numérica**.

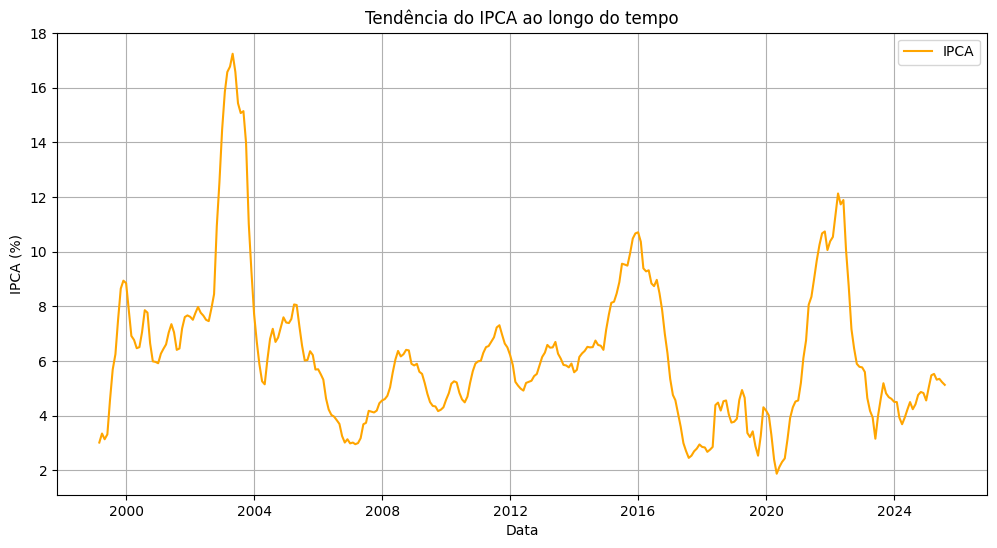
Contexto e Necessidade de Limpeza Textual

Conforme a análise de extensão textual (Seção 11.7), o conteúdo das atas e comunicados possui **formatação HTML** (<div id="atacompleta">...), necessitando de um pré-processamento para **remoção de *tags*** e outros caracteres especiais antes da vetorização. A média de caracteres de aproximadamente **47.000 (Atas)** e **2.098 (Comunicados)** confirma a riqueza da informação textual a ser explorada.

*Nota Institucional:* Observa-se que os comunicados começaram a ser publicados a partir da reunião 46, em 26 de abril de 2000. Para reuniões anteriores (dentro da base filtrada pós-1999), somente o texto das Atas estará disponível, o que será tratado na modelagem como ausência de *feature* textual do comunicado.

Aquisição e Tratamento do IPCA

A série histórica do IPCA foi obtida através da **API de Séries Temporais (SGS)** do Banco Central do Brasil, utilizando o código **13522** (IPCA Acumulado 12 Meses) para garantir a relevância da informação de inflação de longo prazo para as decisões do COPOM.



Estatísticas descritivas do IPCA:

count 318.000000

mean 6.228553

std 2.697288

min 1.880000

25% 4.500000

50% 5.875000

75% 7.262500

max 17.240000

Name: valor, dtype: float64

Código e Gráfico Explicados (Tendência):

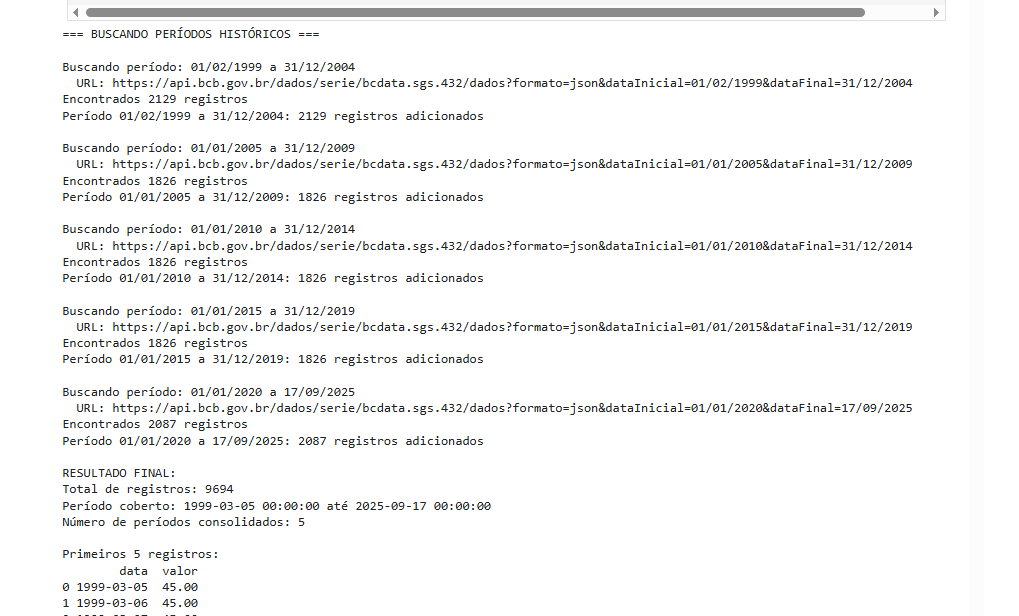
O gráfico de linha confirma a natureza de série temporal volátil do IPCA, mostrando picos (como o máximo de 17.24%) e períodos de estabilidade.

## 11.9 - AQUISIÇÃO E ANÁLISE DA SÉRIE HISTÓRICA DA TAXA SELIC

A segunda variável macroeconômica fundamental para a previsão é a **Taxa Selic Diária (SGS 432)**, que representa o *target* das decisões do COPOM. Devido à sua natureza de indicador diário e às limitações impostas pela API do Banco Central para o intervalo de datas, foi necessária uma abordagem iterativa para garantir a integridade da série histórica completa.

Estratégia de Aquisição Segmentada

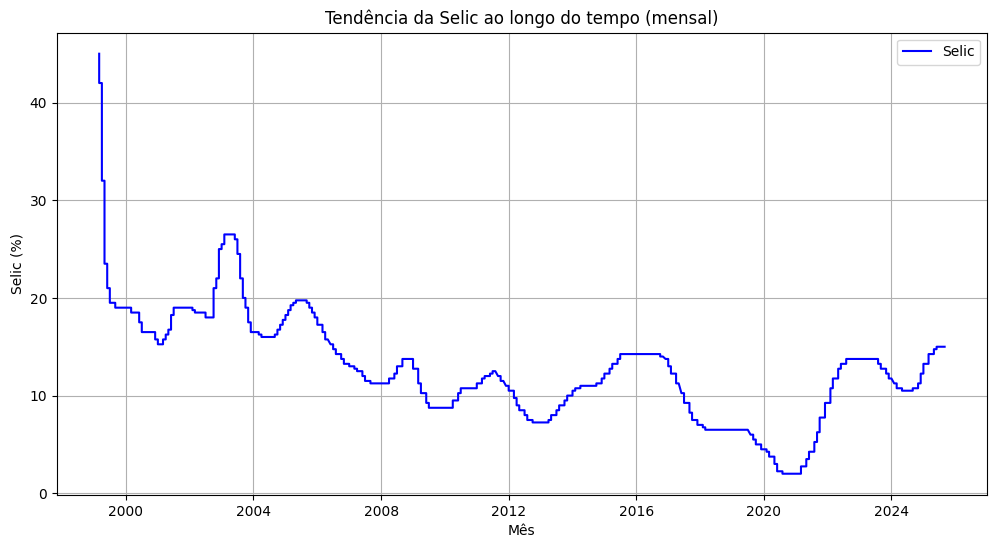
Para contornar as restrições de *range* da API, a busca pela série diária da Selic foi segmentada em blocos de tempo menores. O código implementa uma função (buscar\_selic\_periodo) que é executada sequencialmente por cinco períodos, desde fevereiro de 1999 até a data mais recente disponível na base de Atas.



A execução desta estratégia foi bem-sucedida, resultando na consolidação de **9.694 registros** diários da Taxa Selic, cobrindo o período contínuo de **1999-03-05 a 2025-09-17**.

Análise Estatística e Gráfica

Para fins de visualização e *feature engineering*, a série diária foi agrupada por mês.



Estatísticas descritivas da Selic:

count 9694.000000

mean 12.727151

std 5.480852

min 2.000000

25% 9.250000

50% 12.250000

75% 16.000000

max 45.000000

Name: valor, dtype: float64

Análise: O gráfico de tendência revela a **alta volatilidade** histórica da taxa Selic, uma característica esperada para a política monetária brasileira, que utiliza a taxa de juros como principal ferramenta de controle inflacionário. O **Desvio Padrão (≈6.27)** é elevado, demonstrando grandes oscilações ao longo do período de estudo. A série exibe picos elevados no início (máximo de 45.00%), seguidos de uma tendência decrescente de longo prazo, intercalada por ciclos de alta e baixa.

## 11.10 - LIMPEZA E NORMALIZAÇÃO DO CONTEÚDO TEXTUAL (PLN)

A Análise Exploratória (Seções 11.6 e 11.7) confirmou que as colunas textoAta e textoComunicado contêm **formatação HTML**, que deve ser removida antes de qualquer técnica de Processamento de Linguagem Natural (PLN). A remoção desse ruído é um passo fundamental para garantir que a vetorização textual capture apenas o significado semântico das palavras.

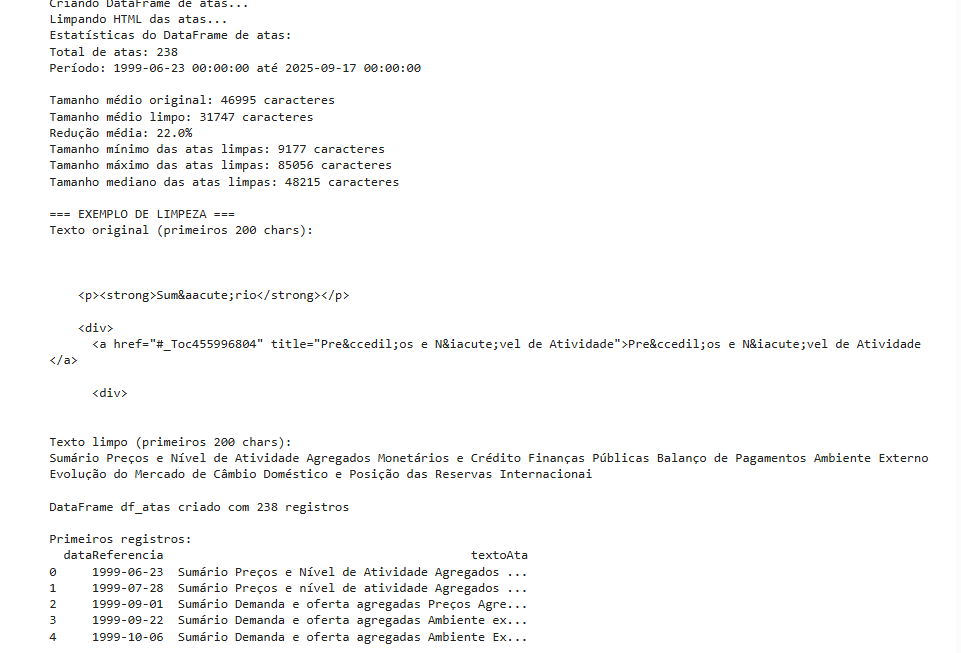
Esta subseção detalha a aplicação da função de limpeza de HTML e a normalização do texto da Ata.

Função de Limpeza e Aplicação em df\_atas

A função limpar\_html foi implementada para remover *tags* e normalizar o texto:

**BeautifulSoup:** Utilizada para fazer a raspagem do conteúdo e extrair eficientemente o texto puro, descartando todas as *tags* HTML.

**re (Expressões Regulares):** Aplicada para normalizar o texto, substituindo múltiplos espaços, quebras de linha (\n) e tabs (\t) por um único espaço, e removendo caracteres especiais desnecessários.



A redução média de **22.0%** no tamanho das Atas após a limpeza (de 46.995 para 31.747 caracteres) é significativa e confirma que uma grande quantidade de *tags* HTML, metadados e outros elementos não textuais foram removidos com sucesso.

Gráfico, Gráfico de caixa estreita

O conteúdo gerado por IA pode estar incorreto.

O processo resultou na criação do **DataFrame df\_atas** com **238 registros** válidos e a coluna textoAta perfeitamente limpa e normalizada. O período coberto é de **1999-06-23 a 2025-09-17**, consistente com o filtro temporal estabelecido. A etapa de pré-processamento textual está, portanto, concluída para as Atas.

## 11.11 - CONSOLIDAÇÃO FINAL DOS DADOS E ENGENHARIA DA VARIÁVEL ALVO

A etapa final do pré-processamento consiste na consolidação dos *datasets* filtrados e tratados: o texto limpo das Atas (df\_atas), a série diária da Selic (df\_selic) e a série mensal do IPCA (df\_ipca). Além disso, é criada a **Variável Alvo (y)**, que define o objetivo de regressão do modelo.

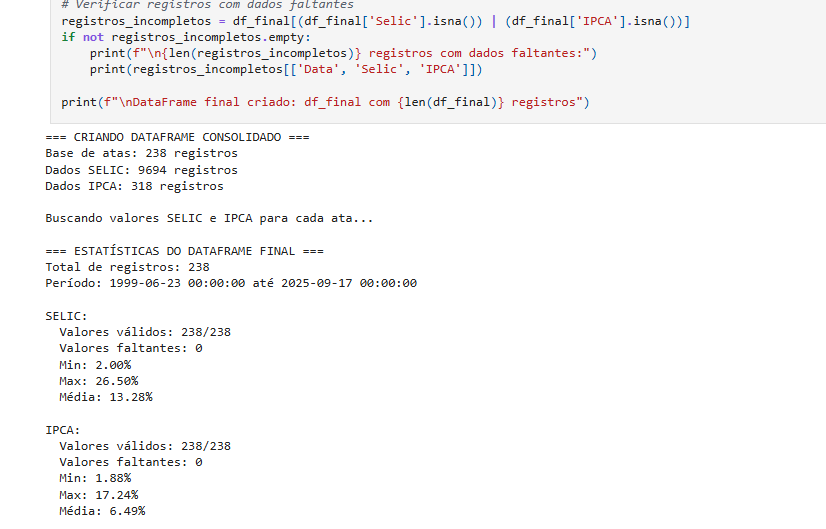
Estratégia de Junção e Mapeamento de Variáveis

Para garantir a coerência temporal, a junção dos *datasets* foi realizada através de funções customizadas que mapeiam os valores mais relevantes para a data de referência de cada Ata:

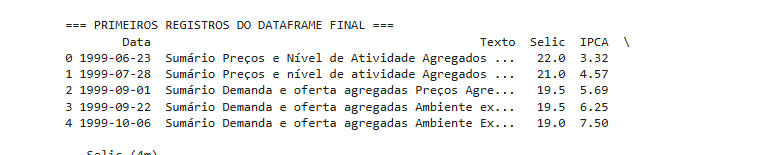
* **Selic (Selic):** Utiliza-se a função buscar\_selic para encontrar o valor da Selic em vigor na Data exata da Ata. Caso a data não exista (feriado ou fim de semana), busca-se o valor imediatamente anterior, garantindo 100% de preenchimento.
* **IPCA (IPCA):** A função buscar\_ipca\_anterior mapeia o valor do IPCA (acumulado 12 meses) **referente ao mês anterior** à data da Ata, uma vez que a inflação é divulgada com *lag*.

**Análise e Integridade do Dataset Final**

O processo de *feature engineering* e *merge* resultou em um *DataFrame* final (df\_final) com **238 registros**. A estratégia de mapeamento foi bem-sucedida, eliminando valores nulos (NA) nas variáveis preditoras numéricas.



Abaixo estão os primeiros e últimos registros do *dataset* final, confirmando a correta integração dos dados



**Criação do rótulo**

Foi criada uma nova coluna, ‘Selic (6m)’, com o valor da taxa seis meses após a data da reunião. Comparando o valor da taxa no futuro com a atual, foi criada uma nova coluna, de sentimento, categórica com 3 valores:

‘hawkish’: se a taxa subiu após a reunião;

‘neutral’: se se manteve;

‘dovish’: se a taxa caiu.

**Definição do Dataset de Modelagem**

O *DataFrame* **df\_final** está totalmente preparado, sem dados faltantes nas colunas essenciais, e define claramente as variáveis para o modelo:

Variáveis Preditora (X): Texto (limpo)

Variável Alvo (y): Sentimento

# 12 - EMBASAMENTO

## 12.1 – DEFINIÇÃO DA LINGUAGEM DE PROGRAMAÇÃO

Foi escolhida PYTHON como a linguagem de desenvolvimento para o projeto. A ampla disponibilidade de literatura e bibliotecas nas disciplinas de Ciências de Dados é o fator principal. A familiaridades de todos os integrantes do grupo com a mesma também é decisiva.

## 12.2 – BASE TEÓRICA E MÉTODOS

Foi tomada a decisão de abordar o problema inicialmente como um problema de classificação. Os objetos do estudo, inflação e taxa de juros, são numéricos e poderiam ser alvos de regressão. Contudo, para uma primeira abordagem foi decidido que um estudo de sentimento, avaliando se estas variáveis devem subir, cair ou se manter em prazos pré-determinados é mais prático e atende aos objetivos do estudo, que é indicar o posicionamento ideal da instituição considerando a direção de variação dos índices, e não a intensidade.

As atas são textos extensos, com repetições de terminologia e uso ambíguo de palavras. Isto torna o uso de regressão logística ou ‘bag of words’ menos indicados. O estudo aqui proposto é uma análise de sentimento, uma das várias possibilidades dentro do que comummente se chama de ‘Estudo de Linguagem Natural, ou natural linguage processing – NLP - em inglês. O processamento de NLP pode envolver transformação ou interpretação de texto, capacidade que exploraremos aqui. Redes neurais recorrente ou convolucionais (CNN/RNN) capturam contextos sequenciais de palavras ou frases. Modelos avançados com uso de transformadores (BERT) utilizam mecanismos de atenção para priorizar elementos relevantes do texto.

Para a implementação do modelo de NLP selecionamos o TensorFlow, desenvolvido e mantido pela Google. A biblioteca permite a implementação de redes neurais com grande flexibilidade e inclui ferramentas de suporte na preparação dos dados, definição de pipelines e avaliação dos modelos.

Um fator muito decisivo na escolha da ferramenta é a existência do TensorFlow HUB e a possibilidade de se integrar camadas personalizadas pré-treinadas. As atas a serem estudadas são extensas, mas não são muitas. Desta forma a qualidade do modelo que será obtido a partir das mesmas é incerto no início do processo e a posterior adição de camadas pré-trinadas virá suprindo a deficiência na massa de dados do projeto.

## 12.3 – CÁLCULO DE ACURÁCIA

Para a avaliação da qualidade do modelo foi definido a utilização das seguintes ferramentas:

* **Matriz de confusão:** Permite a visualização da qualidade dos resultados gerados pelo modelo
* **F1-Score:** Oferece um valor mensurável para a qualidade dos resultados obtidos, considerando tanto a *Precisão* quanto o *Recall*.

# 13 - MODELAGEM E RESULTADOS

Aqui será apresentado o processo de modelagem e resultados.

# 14 – STORYTELLING

Aqui será feita a proposta de apresentação do projeto

# 15 - REFERÊNCIAS

1. COPOM - <https://www.bcb.gov.br/conteudo/dadosabertos/BCBDeinf/elements_copom.html#/>
2. BACEN TIME SERIES - <https://www3.bcb.gov.br/sgspub/localizarseries/localizarSeries.do?method=prepararTelaLocalizarSeries>
3. GitHub - Projeto Aplicado 2: <https://github.com/guilhermersduarte/Projeto-Aplicado2-Grupo19>
4. Taxa SELIC - <https://www.bcb.gov.br/controleinflacao/taxaselic%20?modalAberto=administracao-da-taxa-selic>
5. GÉRON, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. 2. ed. Sebastopol, CA: O'Reilly Media, 2019.