



0,0 aignosi

Análise de Planta de Flotação

Resultados da Análise Exploratória de Dados

PEDRO HENRIQUE DE MENEZES COSME

03/11/2025

Contexto

- **Objetivo:** Investigar os dados dos 21 sensores da planta para entender a dinâmica do processo e avaliar a viabilidade de se prever a % de Sílica (impureza) no futuro.
- **O "Porquê":** A medição de laboratório atual leva pelo menos 1 hora para retornar um valor.
- **O Problema:** Quando descobrimos um problema de qualidade, já é tarde demais. Os engenheiros de processo operam no escuro.
- **O foco da EDA:** Antes de propor um modelo, foi necessário caracterizar a dinâmica fundamental do processo. A análise focou em três hipóteses centrais:
 - Linearidade: Avaliar se a saída (% Silica Concentrate) possui correlação linear direta com as variáveis de entrada.
 - Temporalidade: Quantificar os atrasos de processo (lags) entre as ações de controle e o resultado final.
 - Estados Operacionais: Determinar se a planta opera em regimes distintos e se esses regimes impactam a qualidade da sílica.

Base de dados

- **Fonte:** Dados da planta de flotação, de Março a Setembro de 2017.
- **Volume:** 737.453 medições, que foram agrupadas em intervalos de 1 hora para análise.
- **Variáveis:** 24 colunas no total, incluindo:
 - **Entradas/Perturbações:** Qualidade do minério que entra (Ex: % Sílica na Alimentação).
 - **Controles do Processo:** Ações dos operadores (Ex: Fluxo de Amido, Fluxo de Amina).
 - **Variáveis de Estado:** Condições da planta (Ex: Nível e Fluxo de Ar em 7 colunas).
 - **Nosso Alvo (Saída):** % Sílica no Concentrado (a medida do laboratório).

Após agrupar os dados por hora, foram encontradas 318 horas de dados faltantes.

ANÁLISE

Isso não foi uma falha aleatória de sensor. Foi um único evento contínuo: uma parada total da planta (ou da coleta de dados) por 13 dias e 6 horas em Março de 2017.

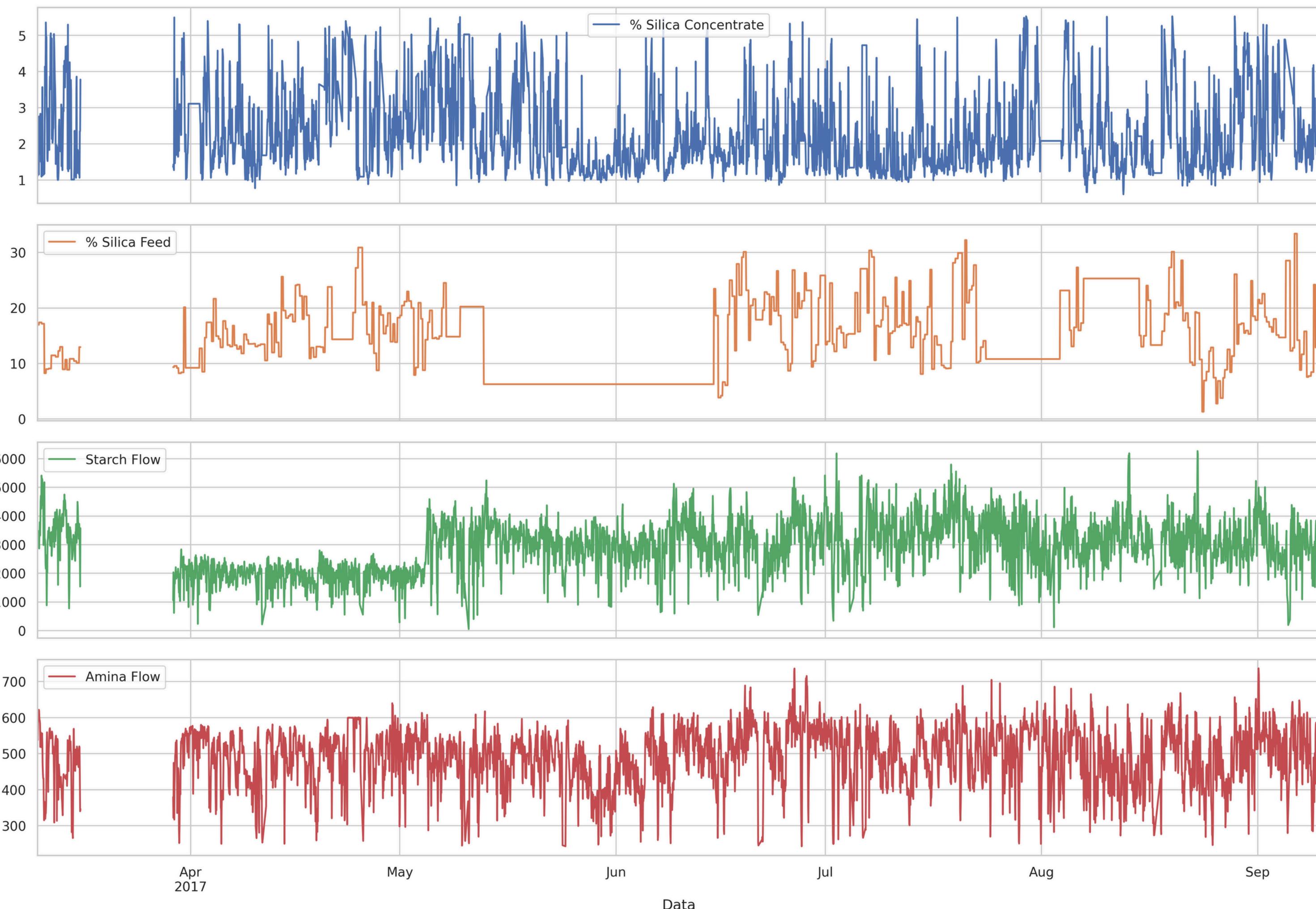
A AÇÃO

Para garantir a integridade estatística, dividimos os dados em "Pré-Parada" e "Pós-Parada"

FOCO DA ANÁLISE

Concentramos nossa análise no bloco "Pós-Parada" (aprox. 5,5 meses de dados contínuos), que representa o comportamento mais recente da planta.

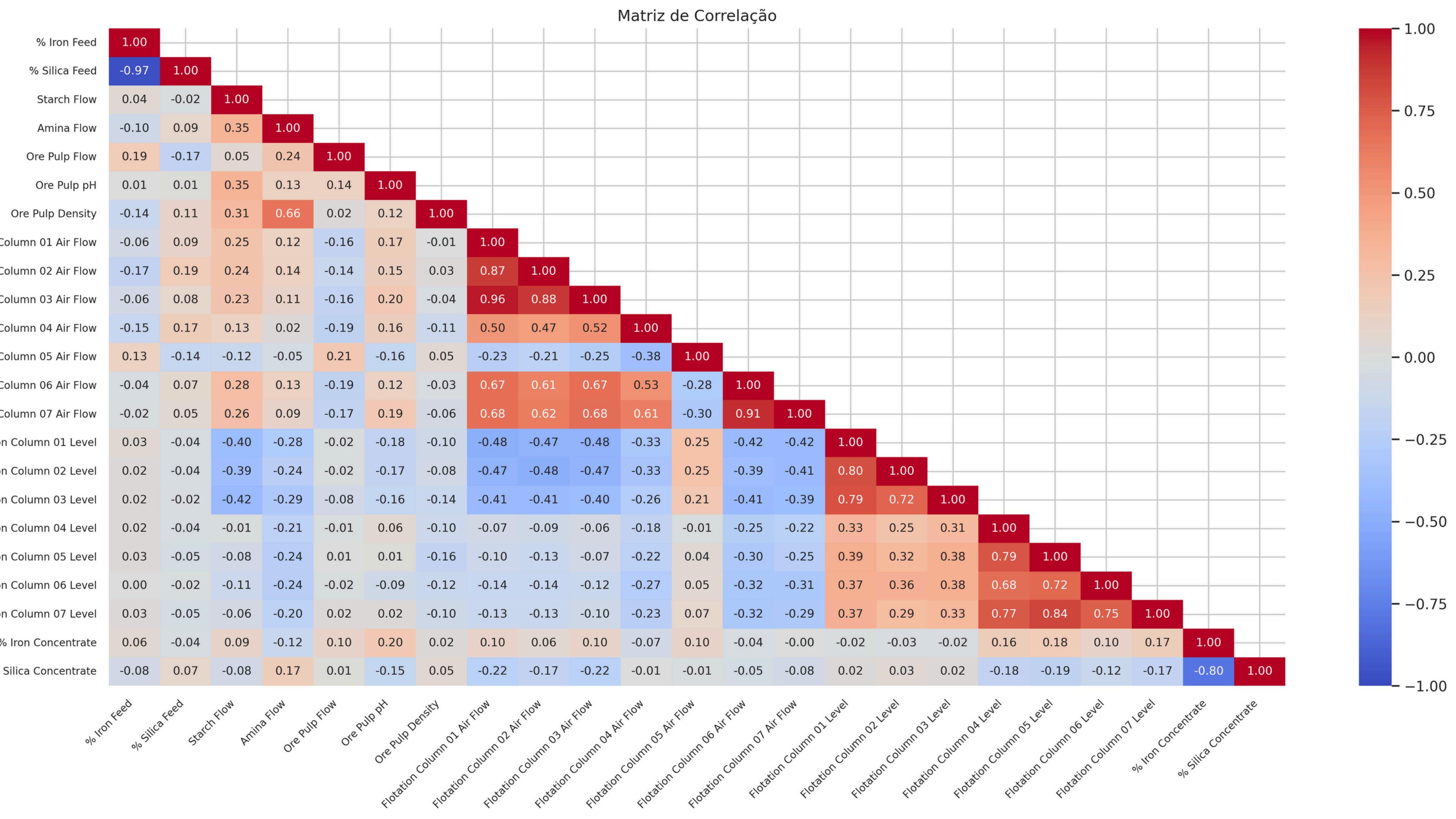
Análise de Série Temporal (Março a Setembro 2017)



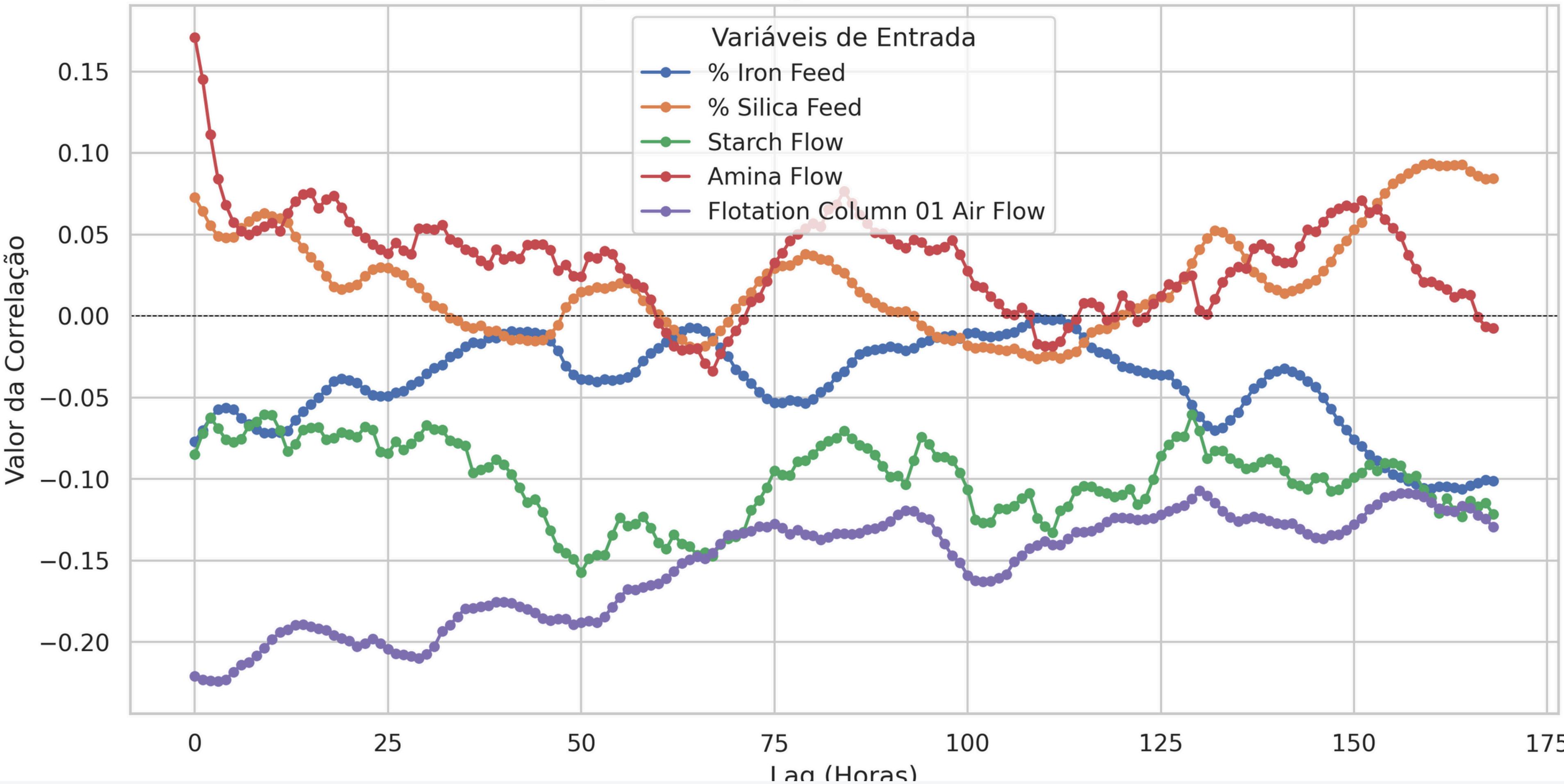
O processo é complexo e não-linear

A primeira hipótese foi buscar uma correlação linear simples.

- Pergunta: A Sílica que entra na planta prevê a Sílica que sai?
- Resposta: Não. A correlação instantânea é quase zero (0.07).
- Segunda Hipótese: "A sílica que entra às 10h impacta a saída das 14h?"
- Resposta: Também não. A análise de lag falhou em encontrar qualquer correlação linear significativa.
- Conclusão Principal: Um modelo linear simples irá falhar. A qualidade final não é prevista por uma única variável.



Análise de Correlação com Lag (Atraso) vs. "% Silica Concentrate"



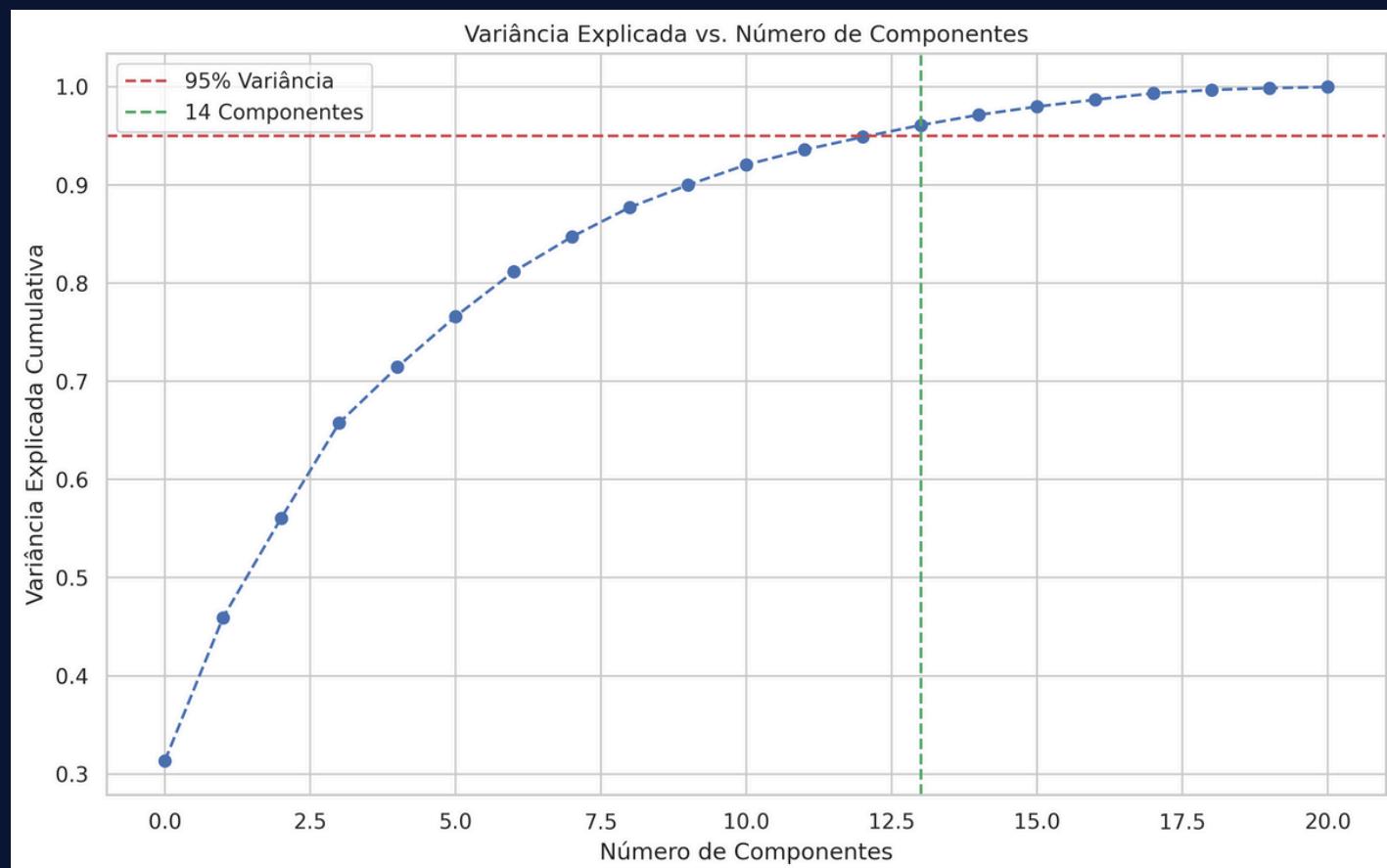
Análise PCA: Os Regimes Operacionais da Planta

- **O Problema (Por que PCA?):**

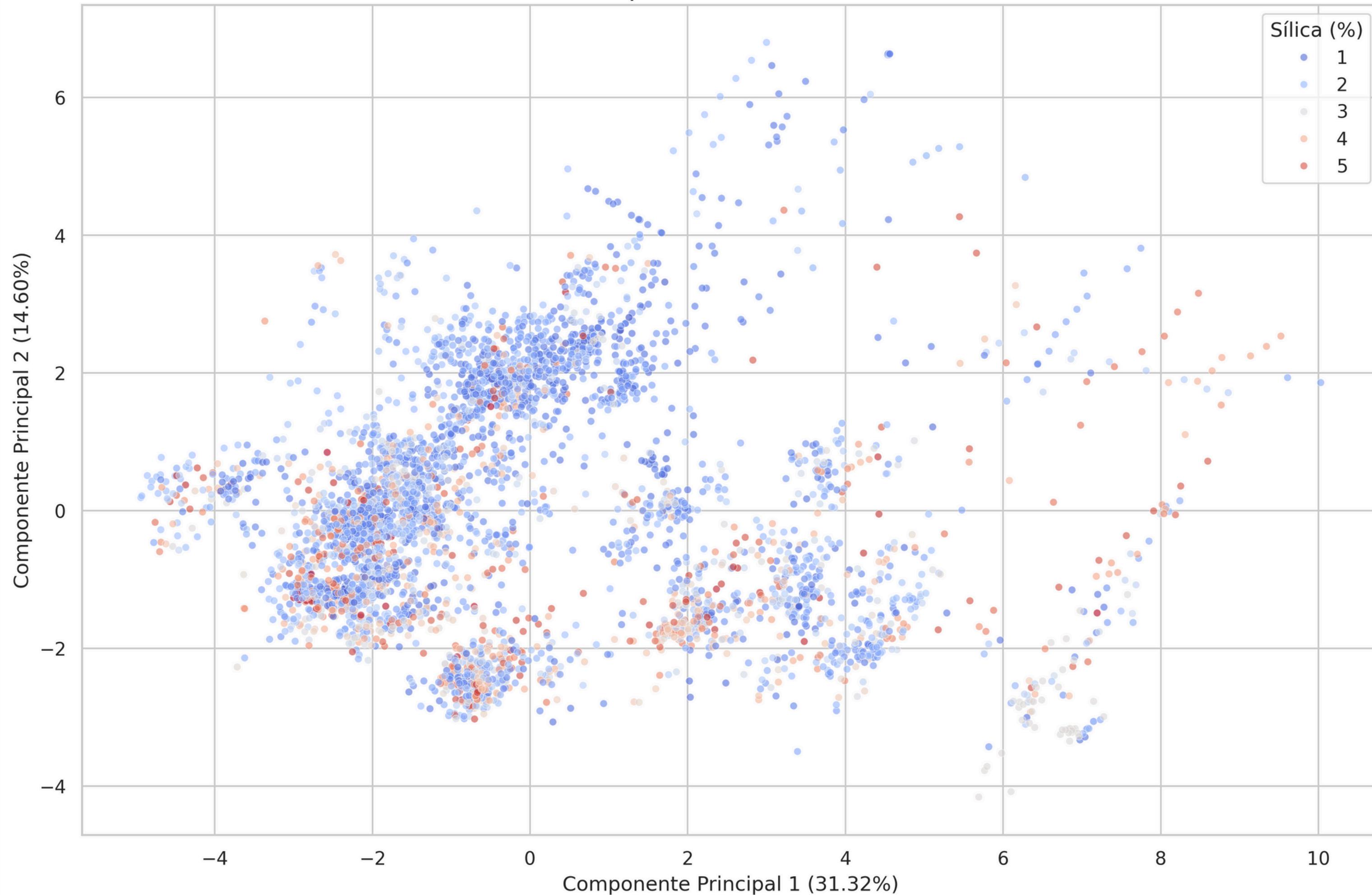
- A análise de correlação simples falhou em encontrar preditores lineares.
- A análise de lag também falhou, indicando que o processo não é linear-simples. Silica Feed (sozinha) não prevê Silica Concentrate (sozinha).
- A matriz de correlação também mostrou alta redundância (ex: as 7 colunas de "Air Flow" se movem juntas).

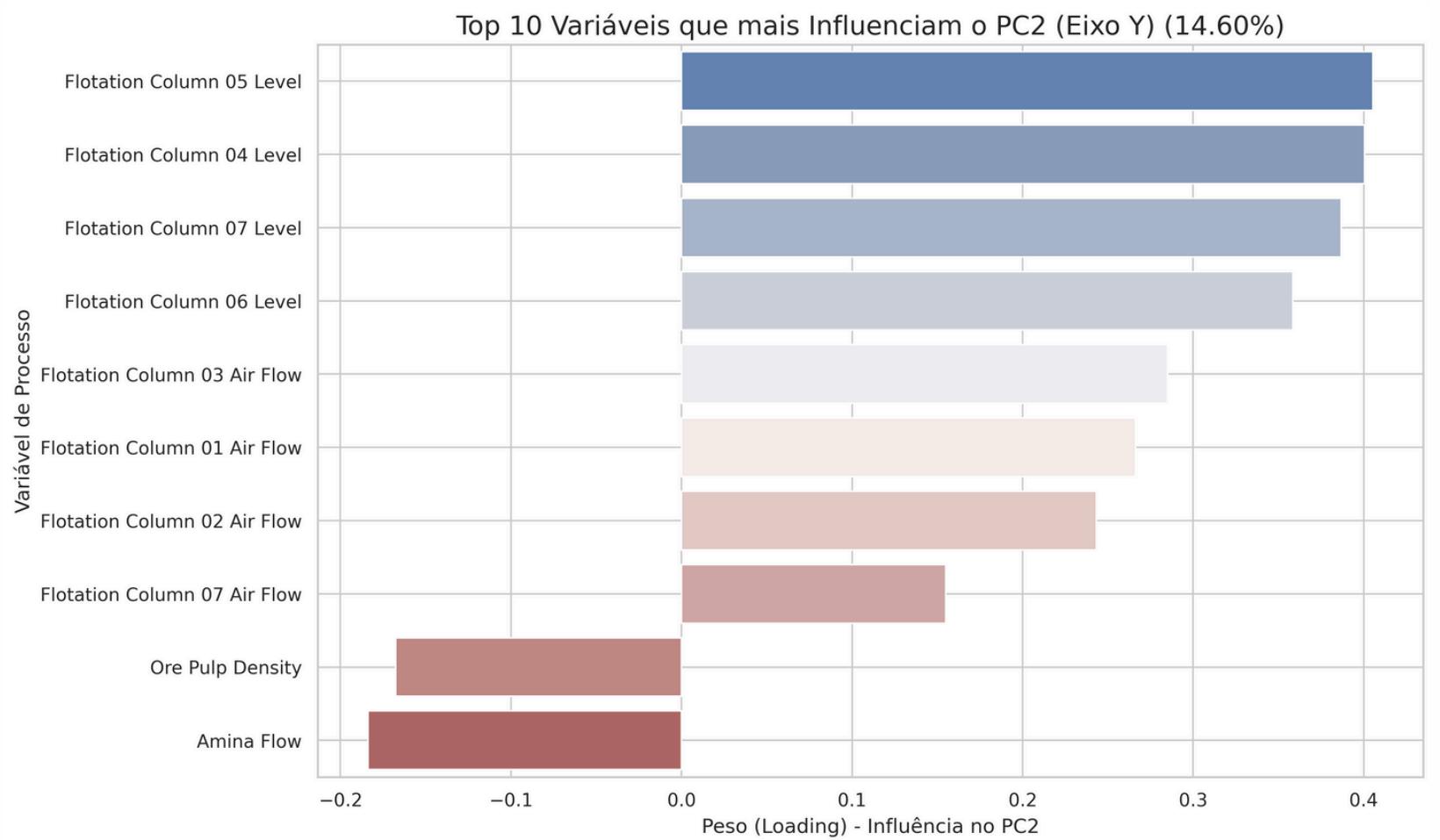
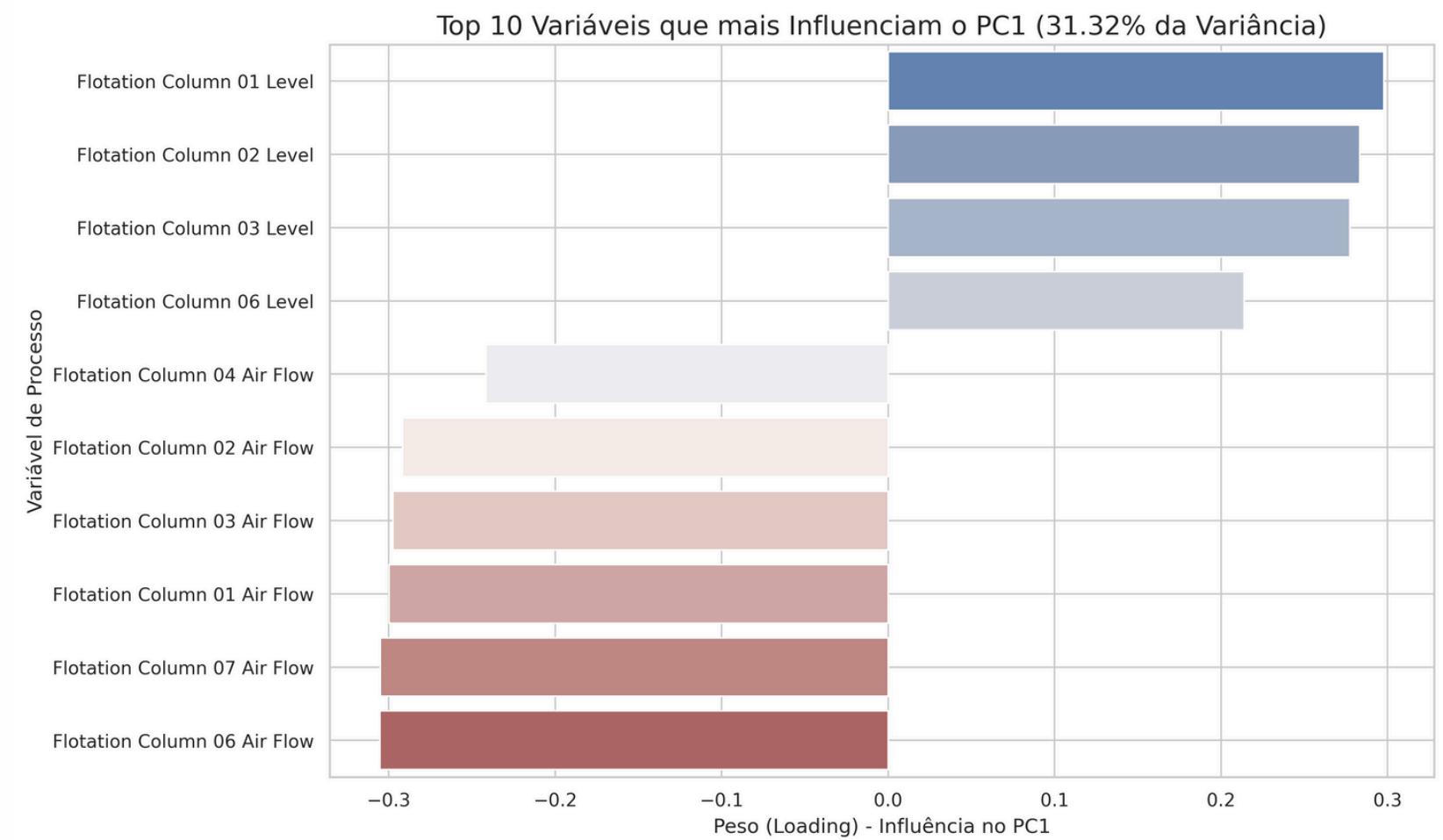
- **A hipótese:**

- Se o processo não é linear, a qualidade final deve depender de uma combinação complexa de todas as 21 variáveis (um "estado operacional").
- Com o PCA é possível "comprimir" essas 21 variáveis para um conjunto de variáveis que expliquem 95% da variância das próprias variáveis de processo.



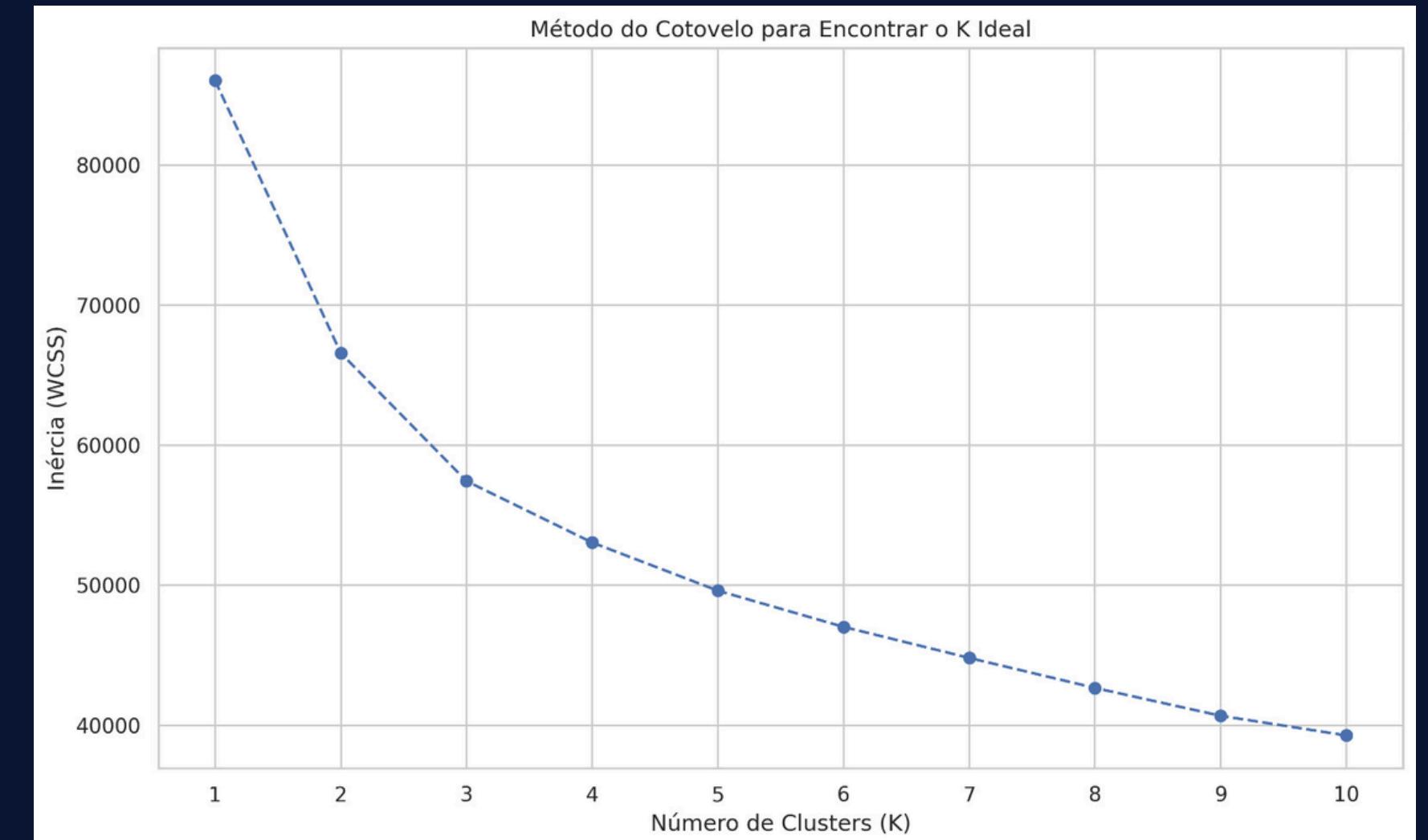
PCA (PC1 vs PC2) Colorido pela Concentração de Sílica (Alvo)
PC1+PC2 explicam 45.92% da variância





Os 3 Regimes de Operação da Planta

- A análise PCA sugeriu visualmente que a planta opera em estado distintos.
- Foi usado o K-Means com a finalidade de agrupar as horas de operação com base apenas nas 21 variáveis de processo . O objetivo é testar se os grupos encontrados correspondiam a diferentes níveis de % Silica Concentrate.
- A operação da planta foi divida em três regimes:
 - Regime 0
 - Regime 1
 - Regime 2



K-Means Clusters (K=3) no Espaço PCA



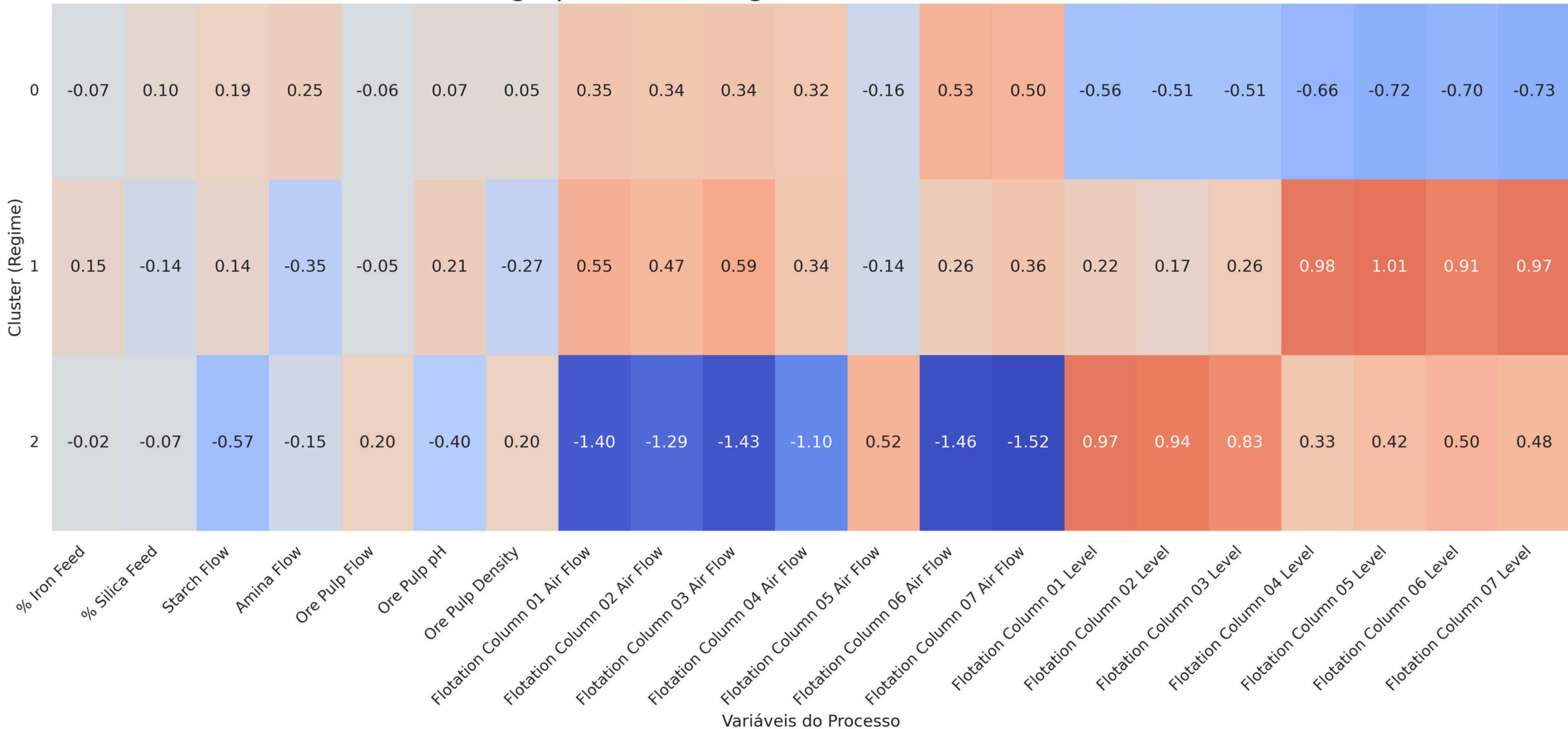
O Desempenho de Cada Regime

Média de % de Sílica para cada regime:

Regime	Nome do Regime	% de Sílica (Impureza)	Frequência
Regime 1	"Estado Ótimo"	1.85% (O Melhor)	26.4% do tempo
Regime 0	"Estado Padrão"	2.46% (Na média)	50.6% do tempo
Regime 2	"Estado Indesejável"	2.59% (O Pior)	23.0% do tempo

Conclusão: A planta opera em 3 estados distintos, cada um com um resultado de qualidade explicitamente distinto.

Fingerprint dos 3 Regimes (Valores Padronizados)



A Anatomia do "Estado Indesejável" (Regime 2)

Causa do pior resultado de qualidade. As médias deste regime:

- **Causa Principal:** Este regime é definido por Fluxos de Ar significativamente baixos.
 - Média do Fluxo de Ar (Col 1): 239 (vs. 290+ nos outros regimes)
- **Sintoma:** Os Níveis das Colunas (1-3) transbordam, atingindo os valores mais altos de todos os regimes.
- **Ação de Controle:** O sistema usa a menor quantidade de Amido (Starch Flow).
- **Resultado:** É um estado de falha ou controle ruim. A "agitação", causada pelo fluxo de ar, pode ser muito baixa, a flotação é ineficiente e as colunas enchem, resultando no pior produto final.



A Anatomia do "Estado Ótimo" (Regime 1)

- **Condição:** Este regime acontece quando a planta recebe o melhor minério (menor % de Sílica na entrada).
- **Causa Principal:** O sistema opera em um estado de "alta agitação".
 - Média do Fluxo de Ar (Col 1): 296.5 (O mais alto)
- **Ação de Controle:** Opera com Níveis de Coluna altos (mas controlados) e usa menos reagentes (Amina e Amido).
- **Interpretação:** Quando o minério é bom, o sistema opera de forma agressiva (alto fluxo de ar) e eficiente (menos reagentes) para produzir o melhor resultado.



O Processo tem Alta Inércia (Memória)

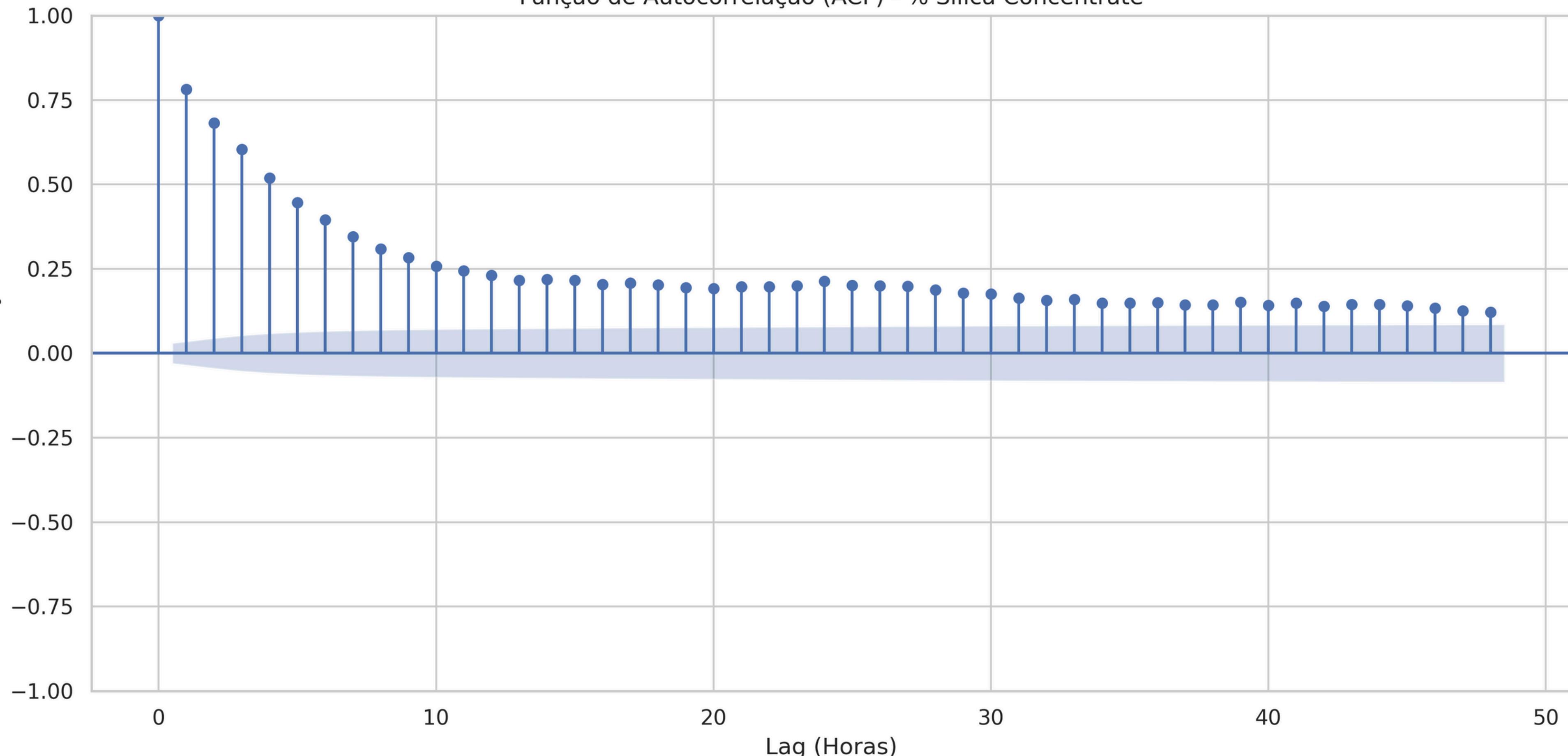
Inércia do Processo (Análise ACF):

- O processo tem uma "memória" muito forte. Se a sílica está alta agora, ela tende a continuar alta por muitas horas.

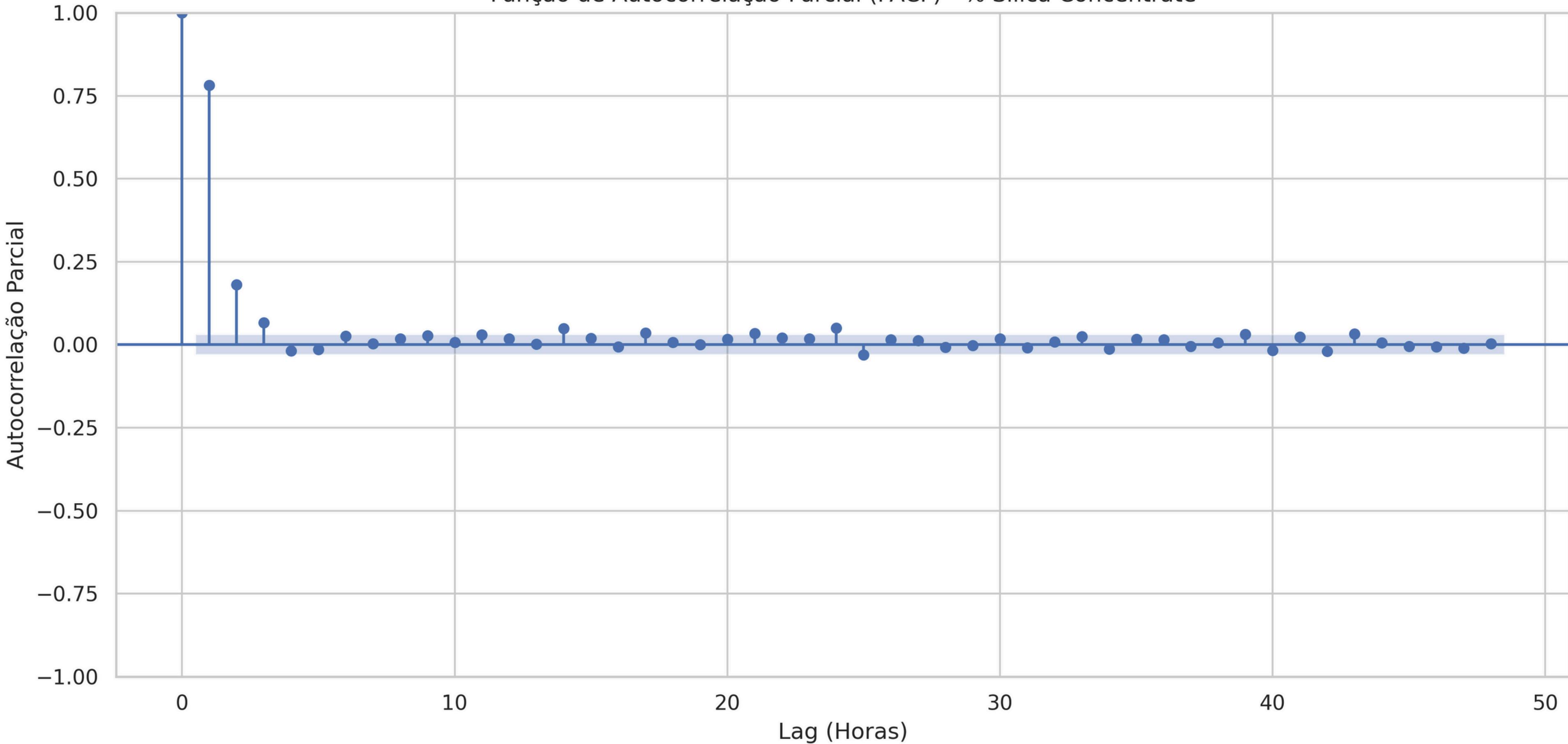
Conclusão

- A planta não muda de regime facilmente. Ela "trava" em um estado (bom ou ruim) por horas, tornando a previsão ainda mais crucial.

Função de Autocorrelação (ACF) - % Silica Concentrate



Função de Autocorrelação Parcial (PACF) - % Silica Concentrate



Descobertas

O PROCESSO É COMPLEXO E NÃO-LINEAR:

A qualidade final não pode ser prevista por uma única variável de entrada.

Nossas análises de correlação simples e com lag (atraso) falharam. Isso prova que modelos lineares simples (Entrada → Saída) não funcionam.

A QUALIDADE DEPENDE DE REGIMES DE OPERAÇÃO

A qualidade final depende do "Estado Operacional" da planta, que é uma combinação de todas as 21 variáveis.

EXISTEM 3 REGIMES

Provou-se a existência de 3 regimes operacionais e seus resultados: "Ótimo" (1.85% de sílica), "Padrão" (2.46%) e "Indesejável" (2.59%).

TEMOS AS "RECEITAS"

Sabemos a "impressão digital" exata de sensores (fluxos de ar, níveis, reagentes) que define cada um desses 3 estados.

O PROCESSO É AUTORREGRESSIVO (TEM "MEMÓRIA")

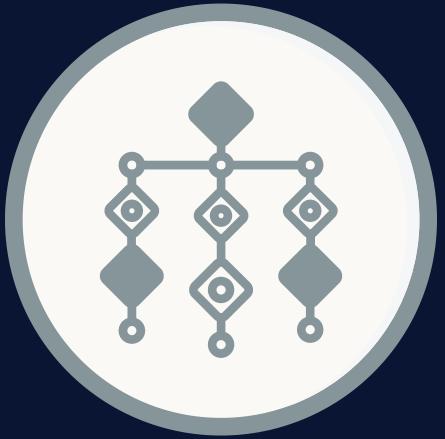
A análise ACF/PACF provou que o processo tem inércia.

O nível de sílica de agora (t) é o melhor preditor do nível de sílica da próxima hora ($t+1$). Isso significa que a planta "trava" nesses regimes por horas.

Próximos Passos

- A Análise Exploratória de Dados está completa. Foi possível descobrir a dinâmica da planta.
- Recomendação: Focar em um modelo de Classificação (prever um estado).
- O Novo Objetivo: Em vez de prever o valor exato da sílica (um problema de regressão difícil), prever:
- "Com base nos sensores, em qual dos 3 regimes (Ótimo, Padrão, Indesejável) a planta entrará na próxima hora?"
-





Possível modelo de ML

- **Random Forest:**
 - Objetivo: Prever o próximo Regime Operacional.
 - Saída do Modelo: A probabilidade de a planta entrar em cada um dos 3 regimes na próxima hora
- **Por que Classificação é Melhor que Regressão?**
 - A EDA provou que o processo é não-linear. A correlação entre qualquer entrada (como % Silica Feed) e o valor exato da saída é muito baixo, mesmo com atrasos.
 - É Mais Acionável: É mais útil para um engenheiro saber "a planta está entrando no estado de falha" do que "a sílica será 2.59%".
- **Por que um modelo de árvore?**
 - Modelos baseados em árvore (como Random Forest) são excelentes para encontrar as relações não-lineares.
 - Uso da Memória do Processo: A análise ACF/PACF provou que o processo é Autorregressivo. O modelo usará features de lag (ex: Nível(t-1), Nível(t-2)) como preditores.

★OBRIGADO