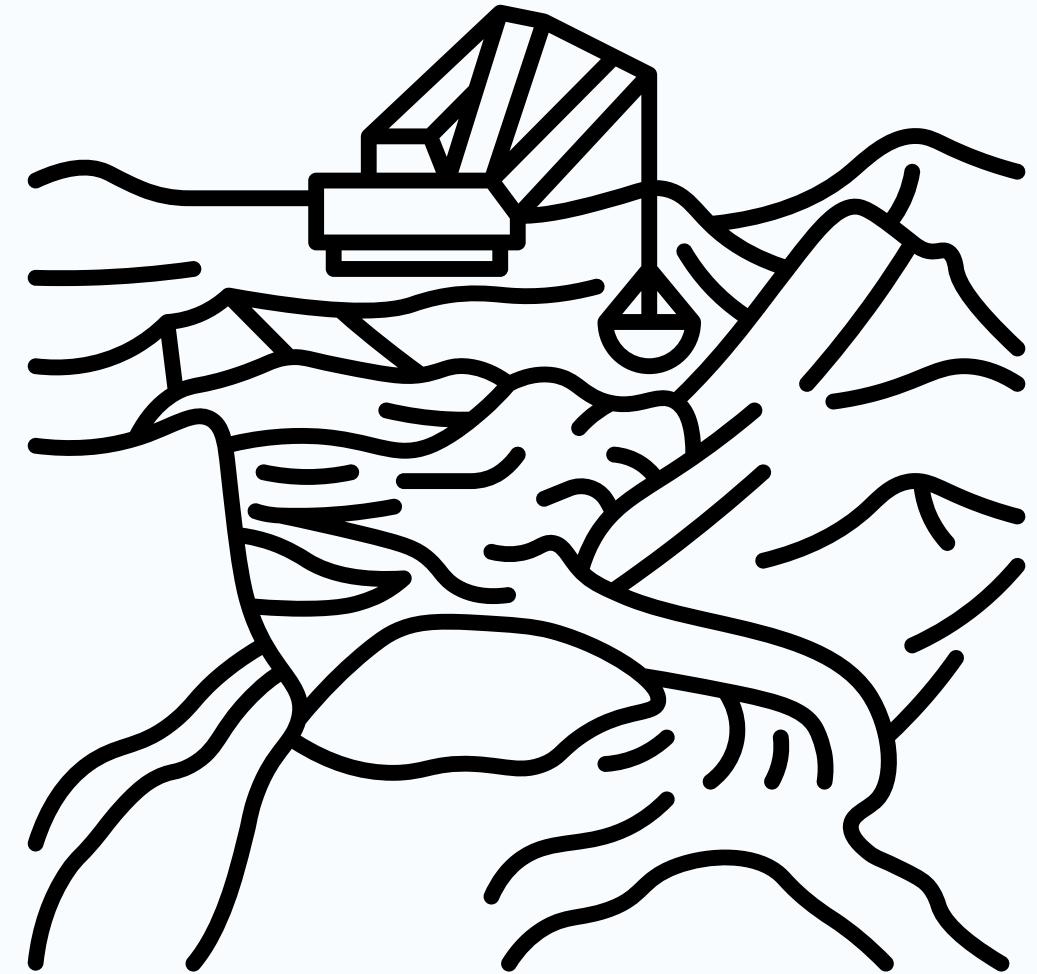


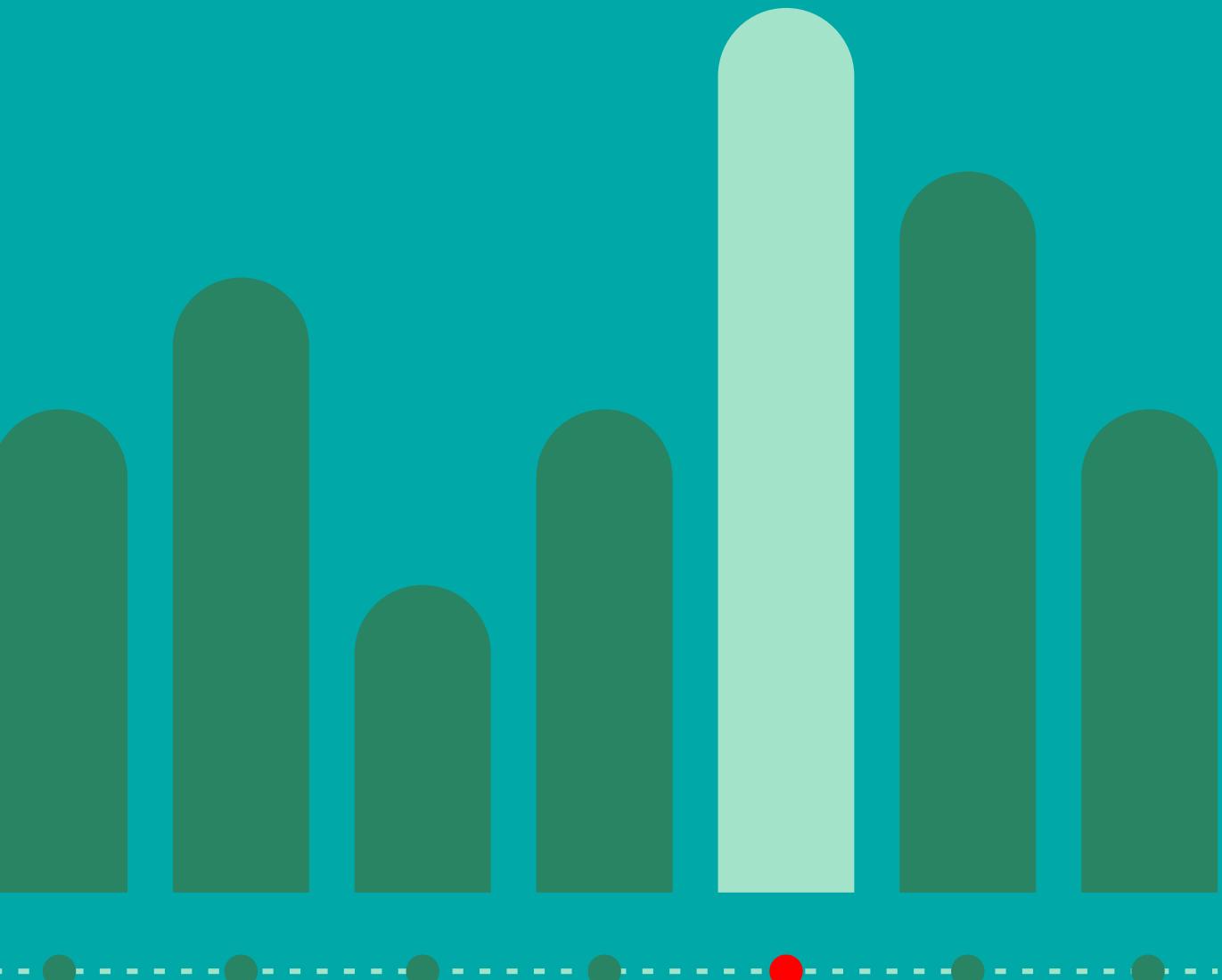
ANÁLISE E PREVISÃO DA QUALIDADE DO CONCENTRADO DE MINÉRIO DE FERRO: EXPLORANDO SÉRIES TEMPORAIS E IMPACTOS DE PROCESSO



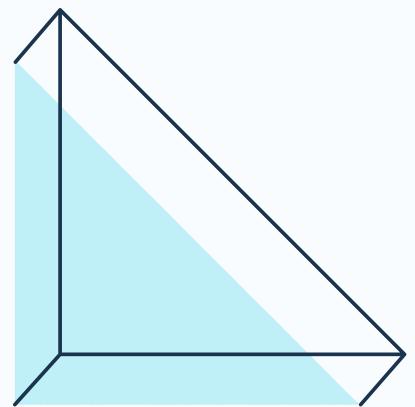
Apresentado por Enio Moreira

SUMÁRIO

- Exploração de dados (EDA) e Hipóteses
 - Amostragem Temporal
 - Distribuição de Variáveis
 - Outliers
 - Dependência Temporal
 - Correlação
 - PCA
- Construção dos Modelos
 - Métricas de comparação: R^2 e RMSE
- MLPRegressor
 - Um pouco sobre
 - Métricas de treinamento e teste: R^2 , RMSE e Função de perda
- Série predita x Série Real

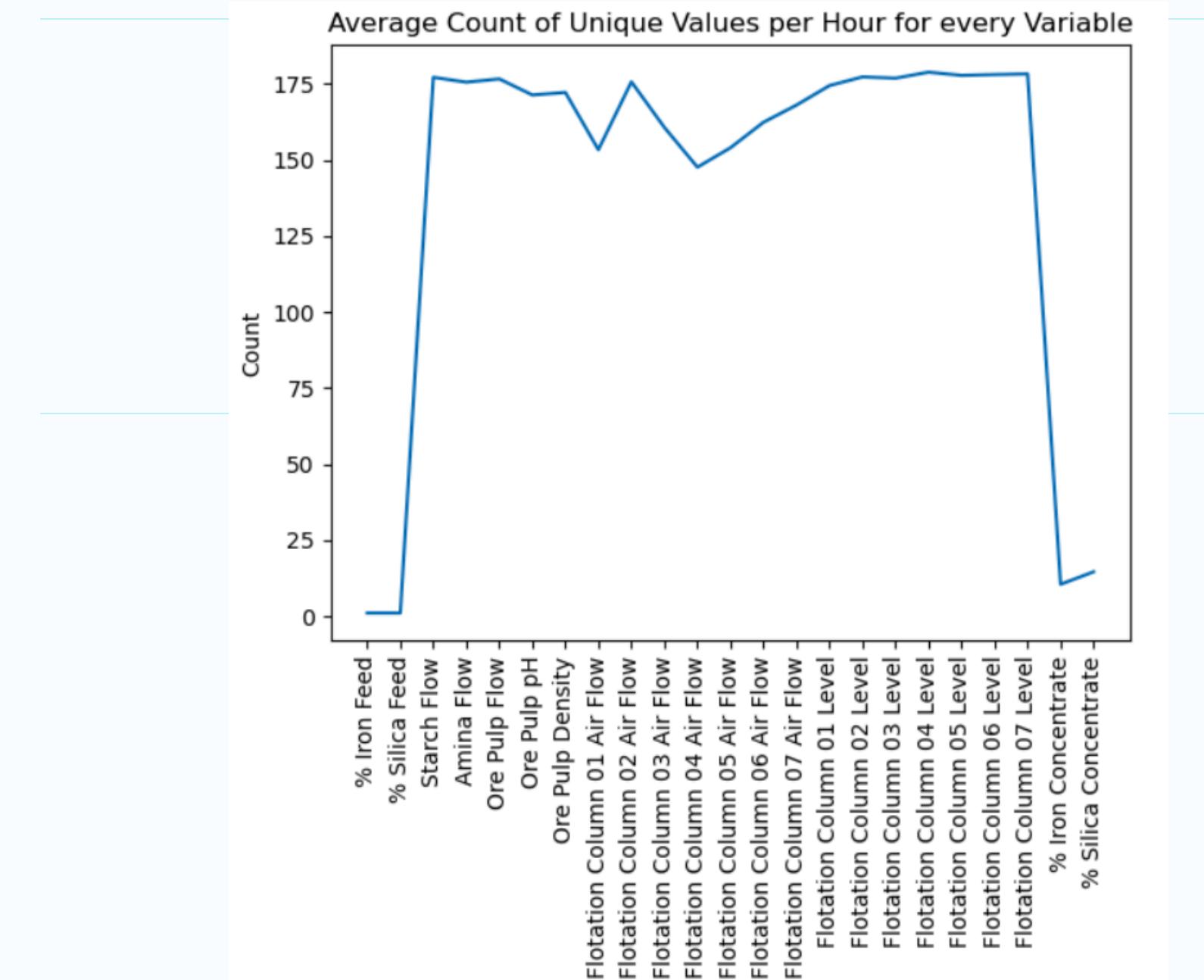


EDA e Hipóteses

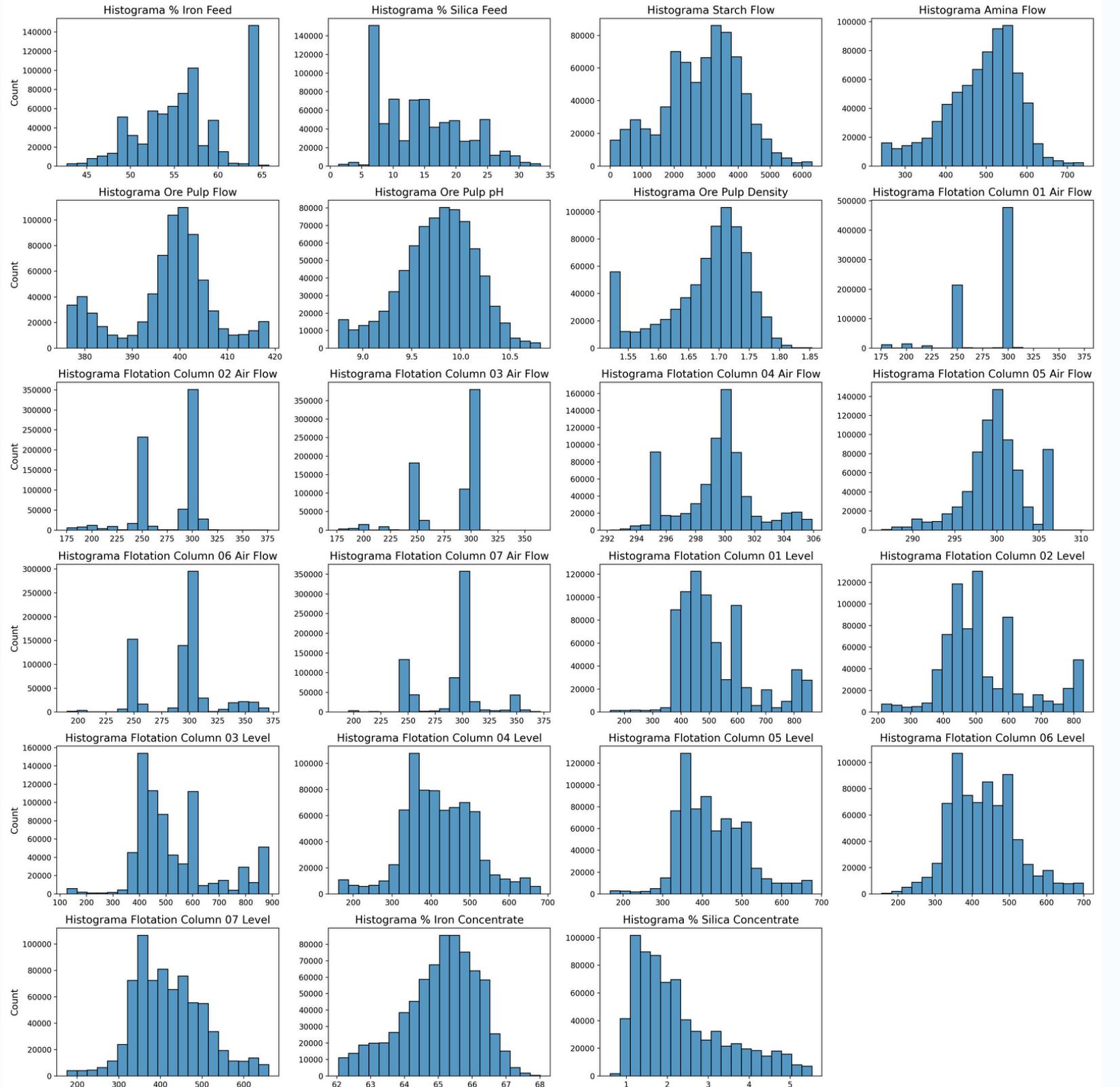


Amostragem temporal

Todas as colunas foram reamostradas para um tempo de amostragem de 20 segundos. Nesse processo, foram excluídas do estudo as datas com menos de 180 registros. Para verificar a distribuição da amostragem, visualizei todas as características, identificando quais estavam registradas por hora e quais seguiam a amostragem de 20 segundos.

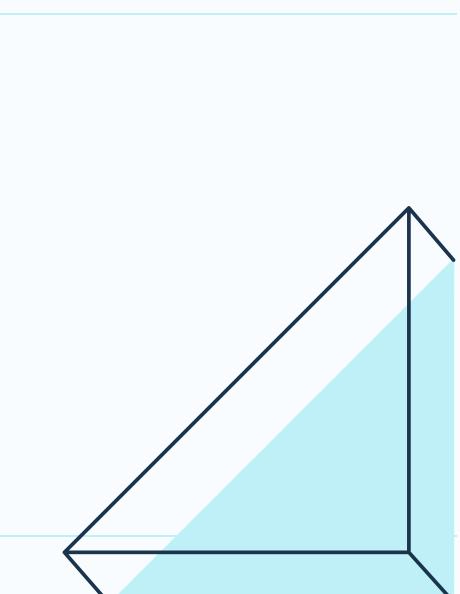


EDA e Hipóteses

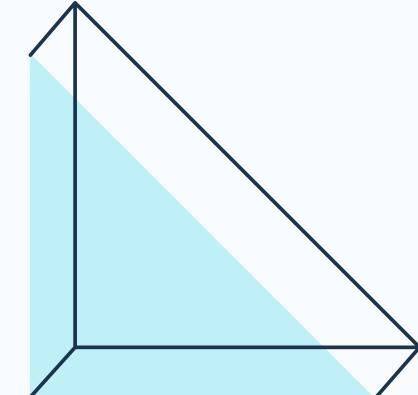


Distribuição de variáveis

Analisando a distribuição de cada coluna (feature), observa-se que várias variáveis apresentam uma distribuição aproximadamente normal, com exceção das sete colunas de Air Flow, cujas distribuições são bastante irregulares. Além disso, nota-se um comportamento quase inversamente simétrico entre as distribuições das variáveis de alimentação (% Iron Feed e % Silica Feed). Esse comportamento faz sentido, pois um maior percentual de ferro na alimentação reduz o espaço disponível para outros elementos, como impurezas de sílica, e vice-versa.

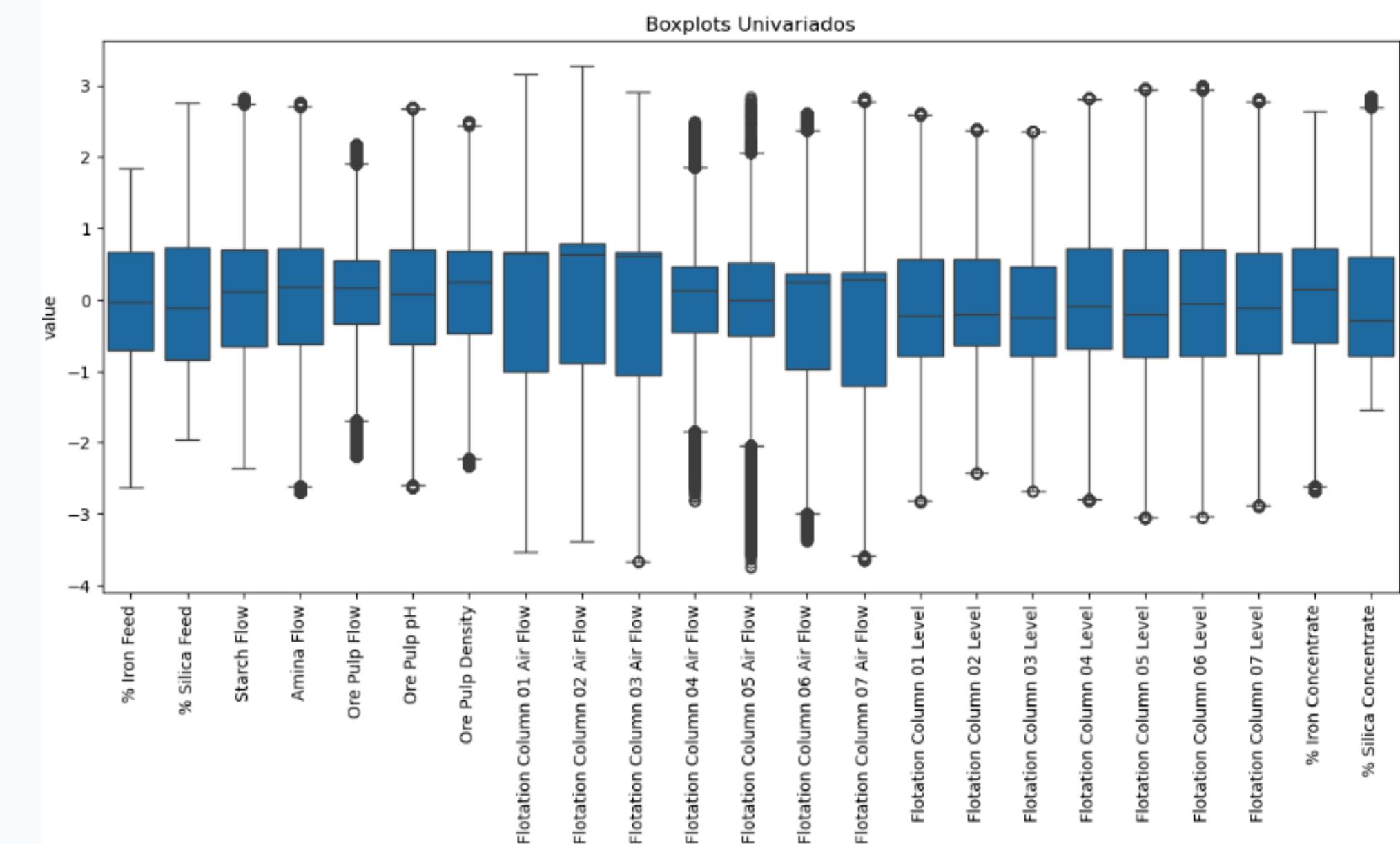


EDA e Hipóteses

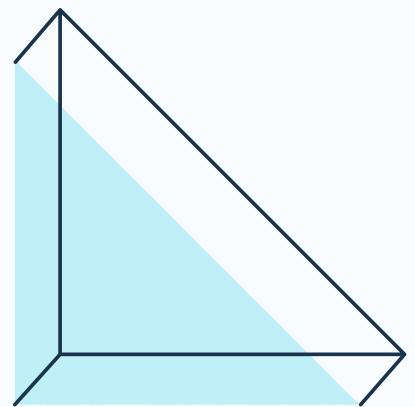


Identificação de Outliers

Para detectar valores discrepantes, foram gerados boxplots para cada feature. Observou-se que algumas colunas de Flotation Column Air Flow apresentam um número elevado de outliers extremos, enquanto Ore Pulp Flow também exibe pontos discrepantes significativos.

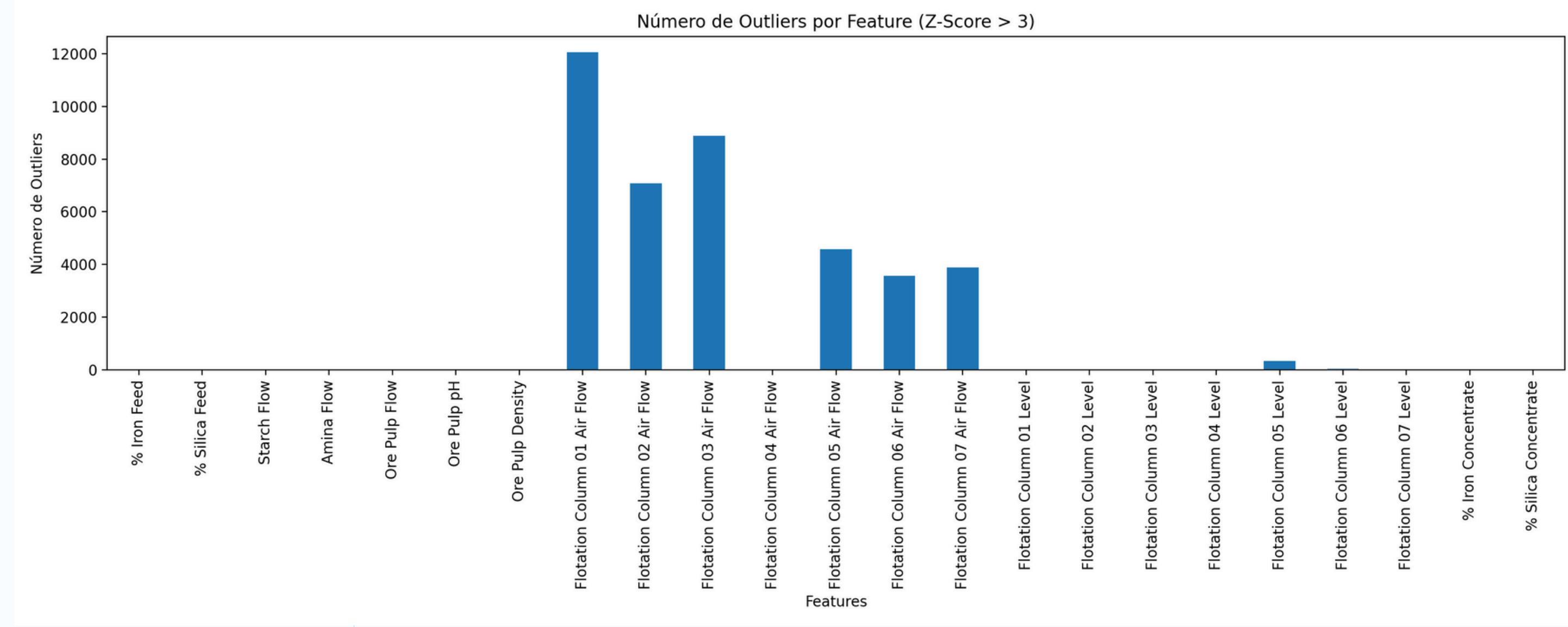


EDA e Hipóteses



Cálculo do Z-Score

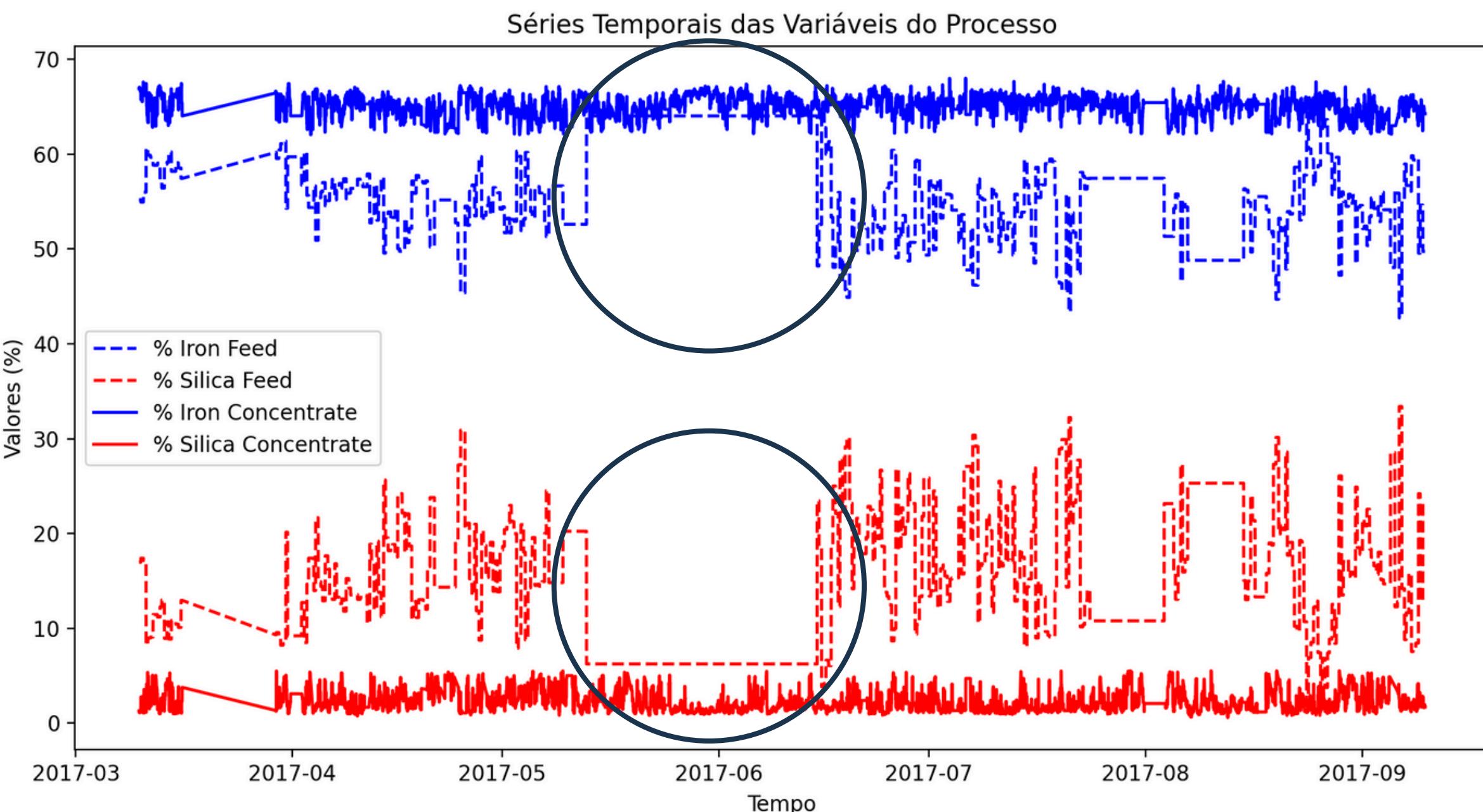
Para complementar a análise visual e reforçar a identificação dos outliers, foi calculado o Z-Score de cada variável.



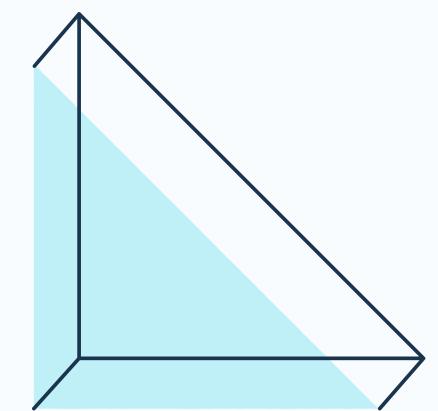
EDA e Hipóteses

Dependência Temporal

A pureza do concentrado de minério de ferro (% Iron Concentrate e % Silica Concentrate) apresenta dependência temporal, sugerindo que os valores passados influenciam os valores futuros devido a efeitos de amortecimento no processo.

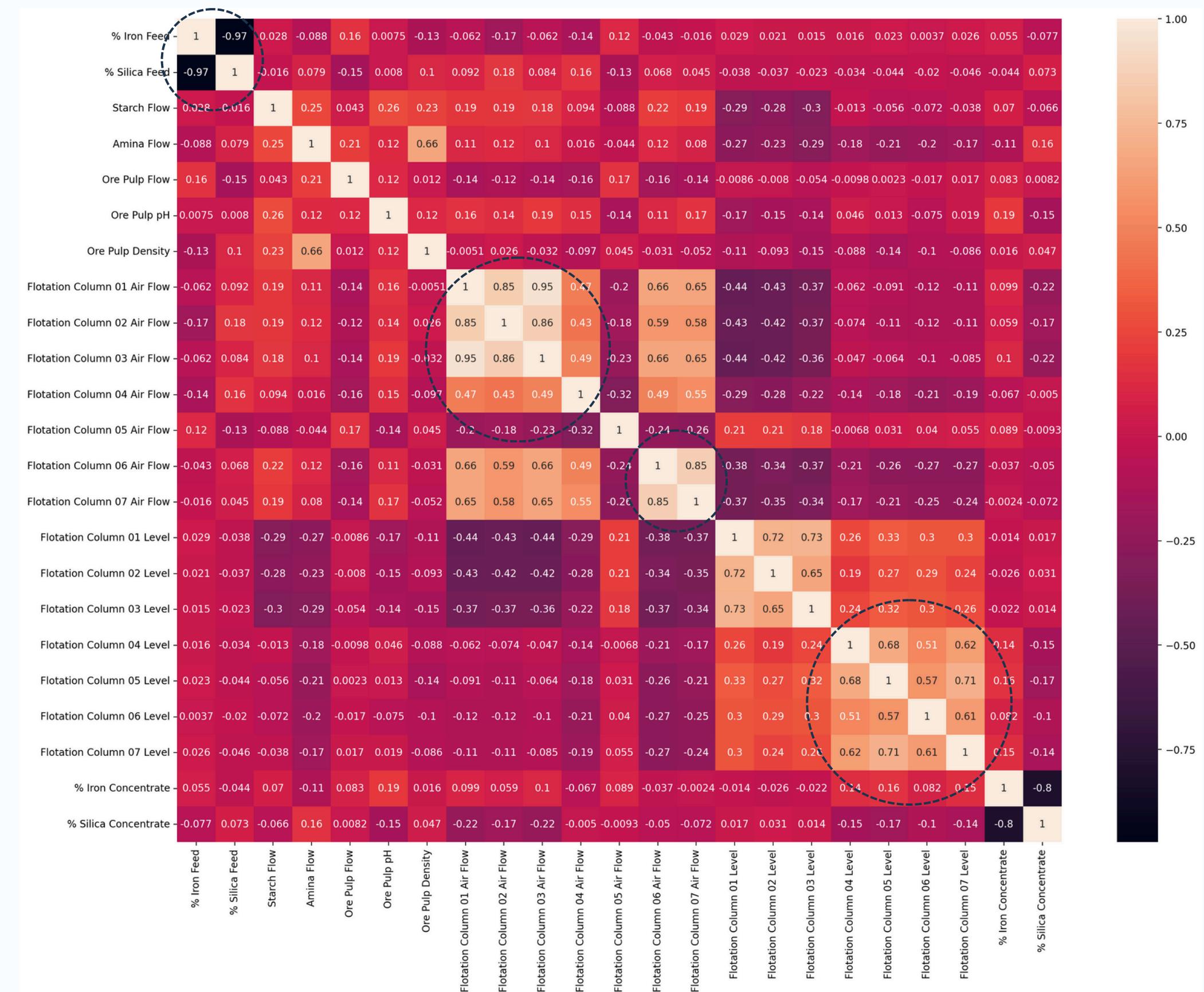


EDA e Hipóteses



Correlação

Apesar de a maioria das variáveis não apresentar correlação forte com a variável alvo, a variável % Iron Concentrate tem uma correlação significativa.

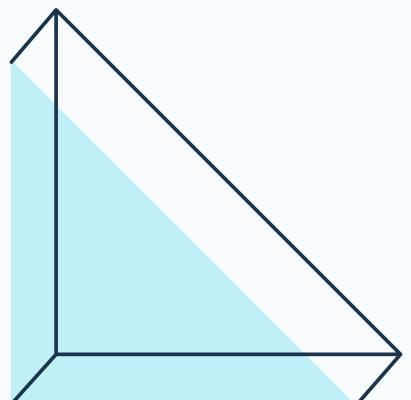


EDA e Hipóteses

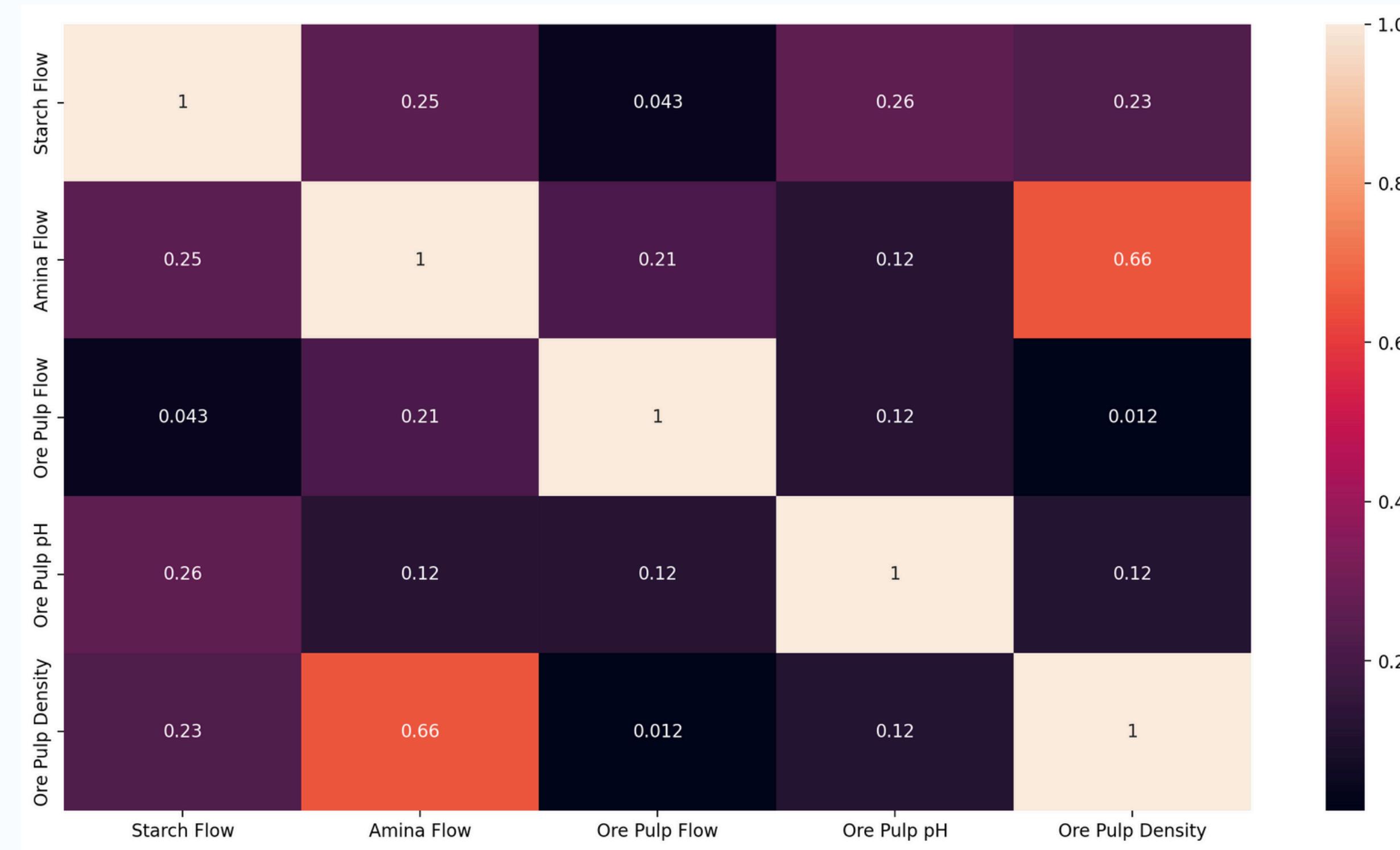
Divisão dos Dados em DataFrames Correlacionados

Embora algumas colunas não apresentem uma correlação forte com a variável alvo, identificou-se uma correlação significativa entre elas. Dessa forma, os dados foram agrupados em data frames com base nessas relações.

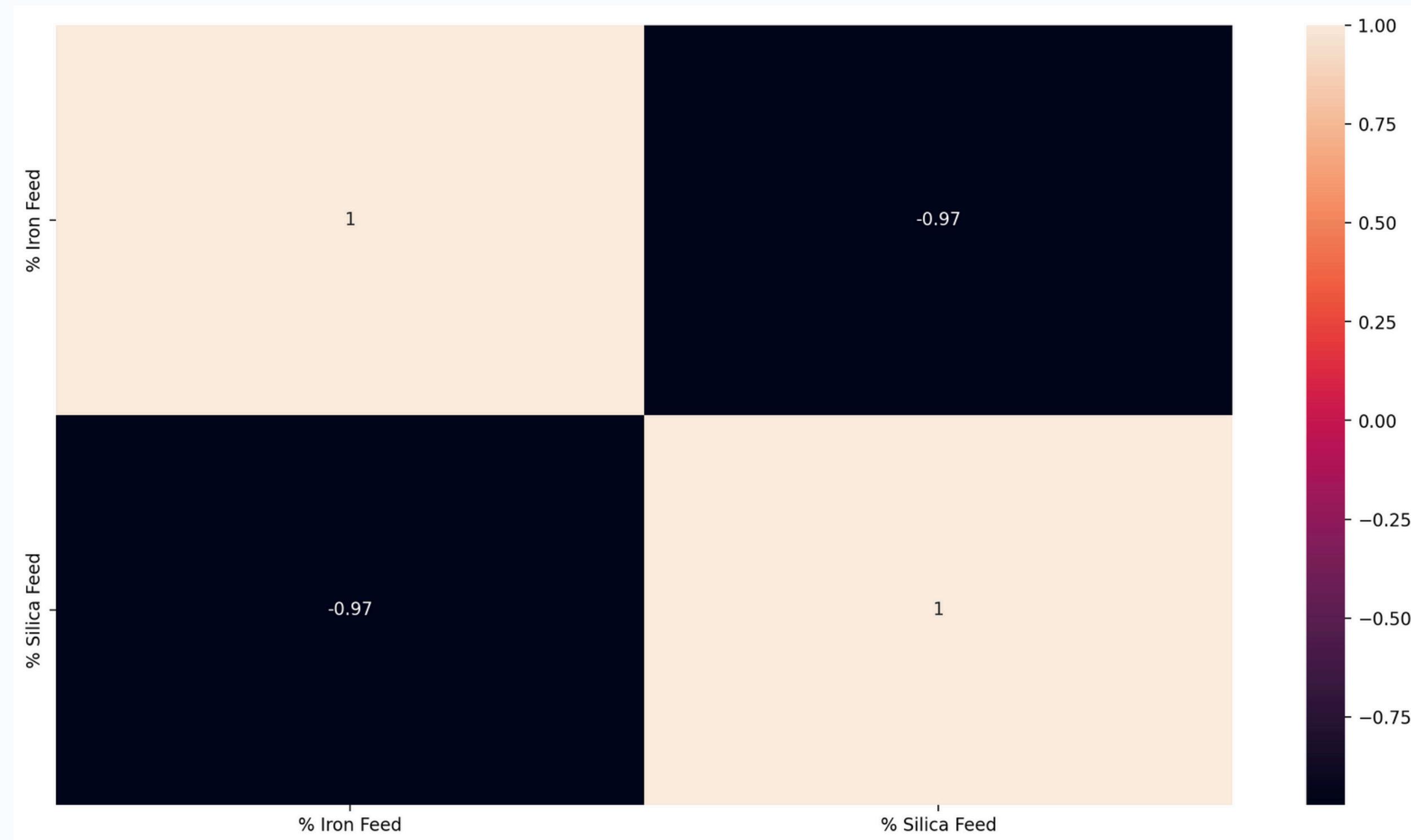
- Importantes: Informada na descrição do data set
- Feed: Variáveis de entada
- Air Flow: Colunas de flotação
- Level: Colunas de flotação



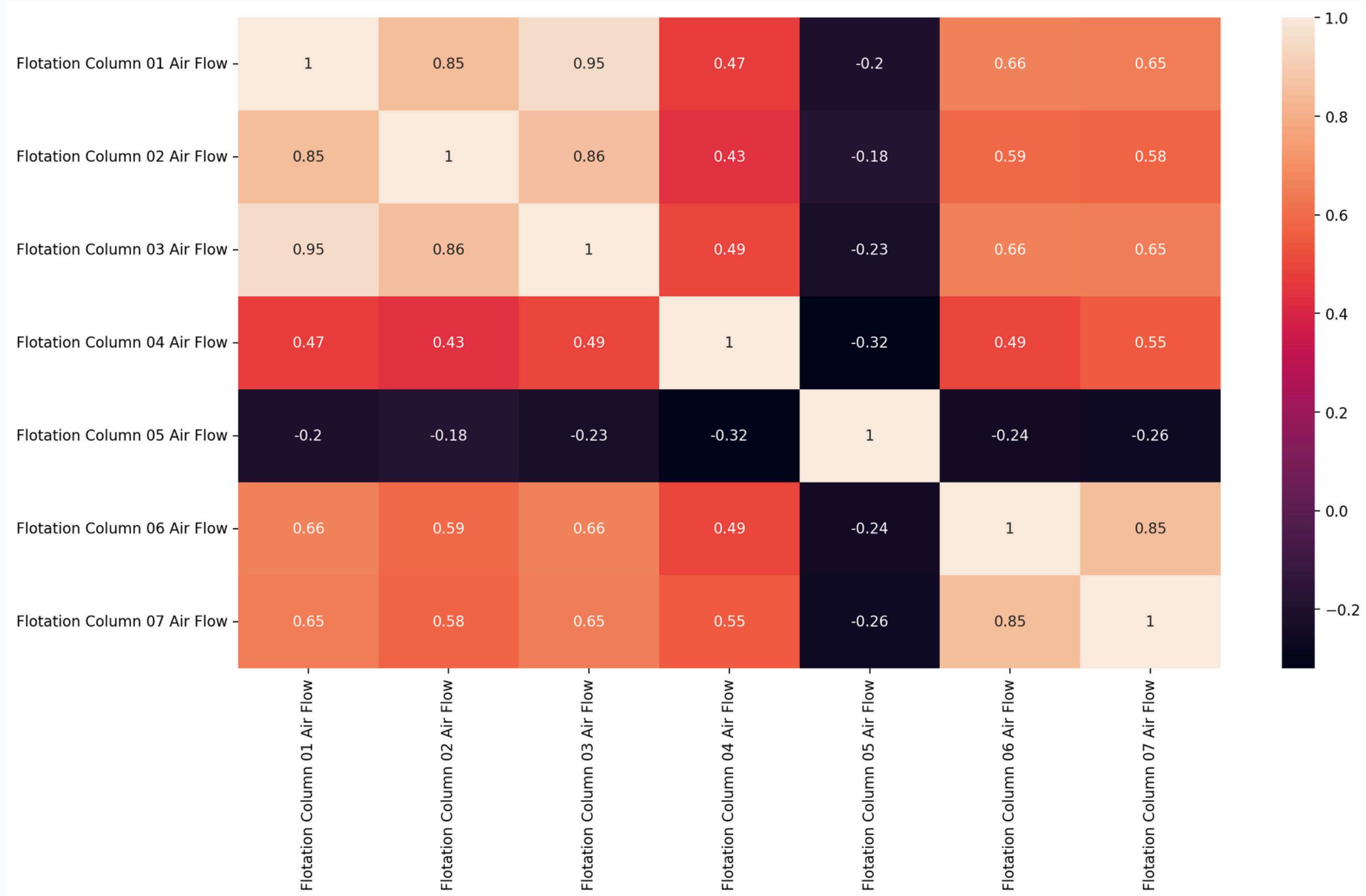
Importantes



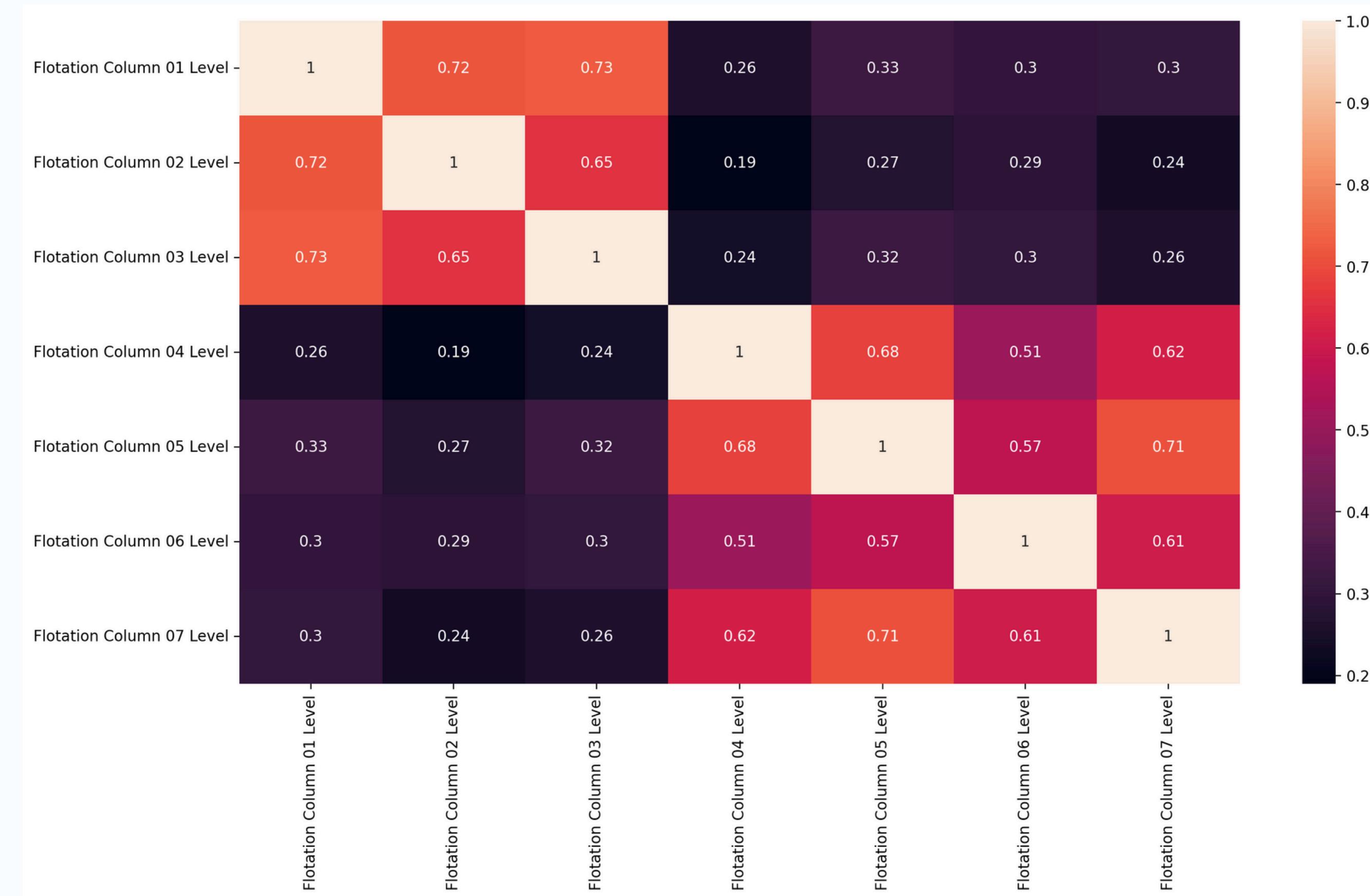
Feed



Air Flow



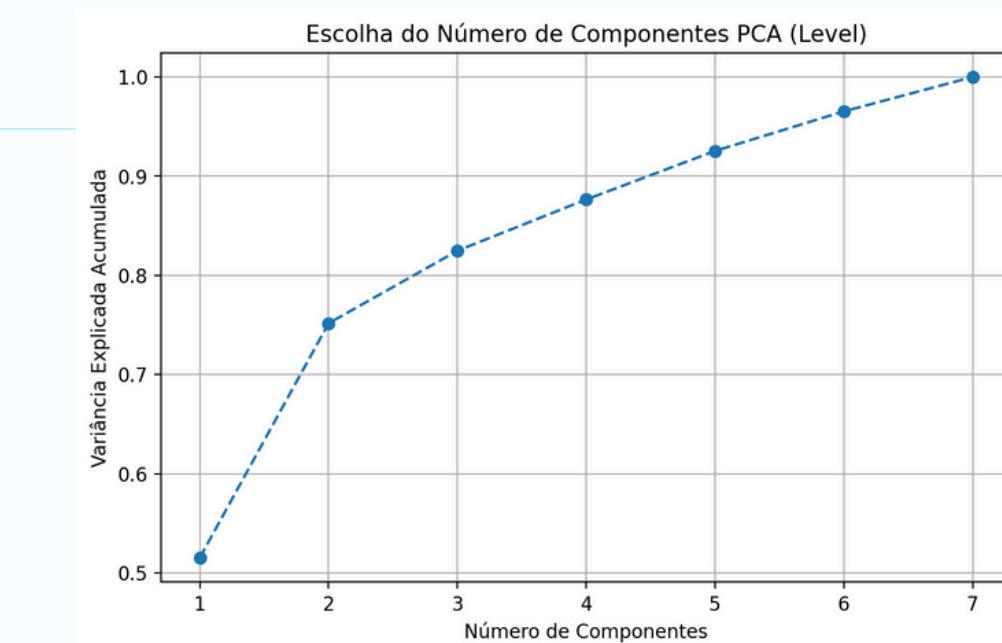
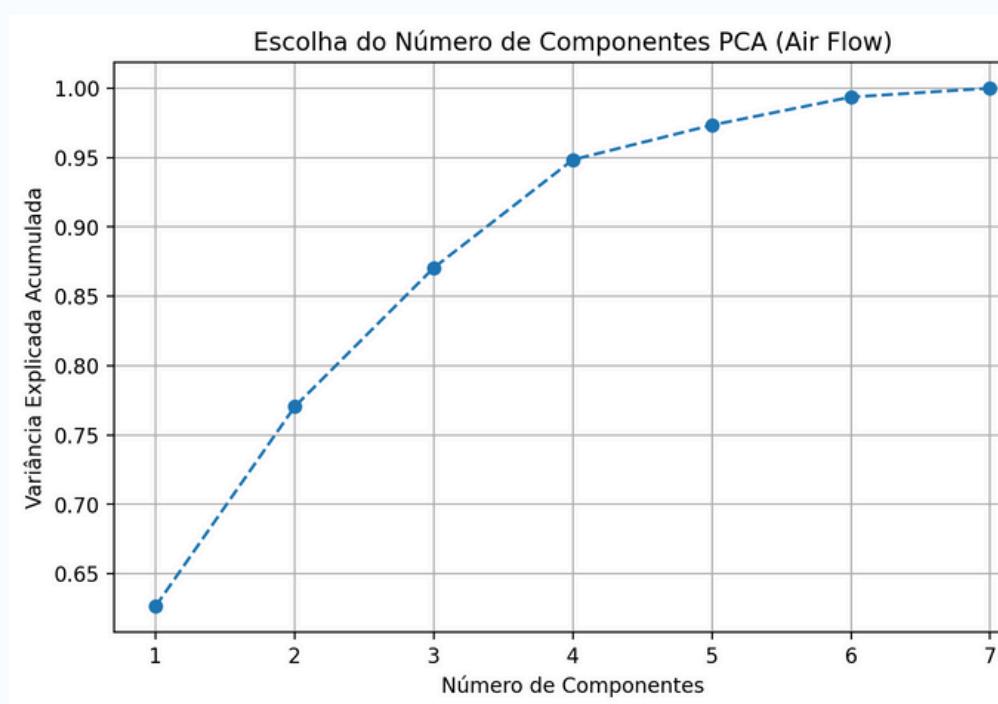
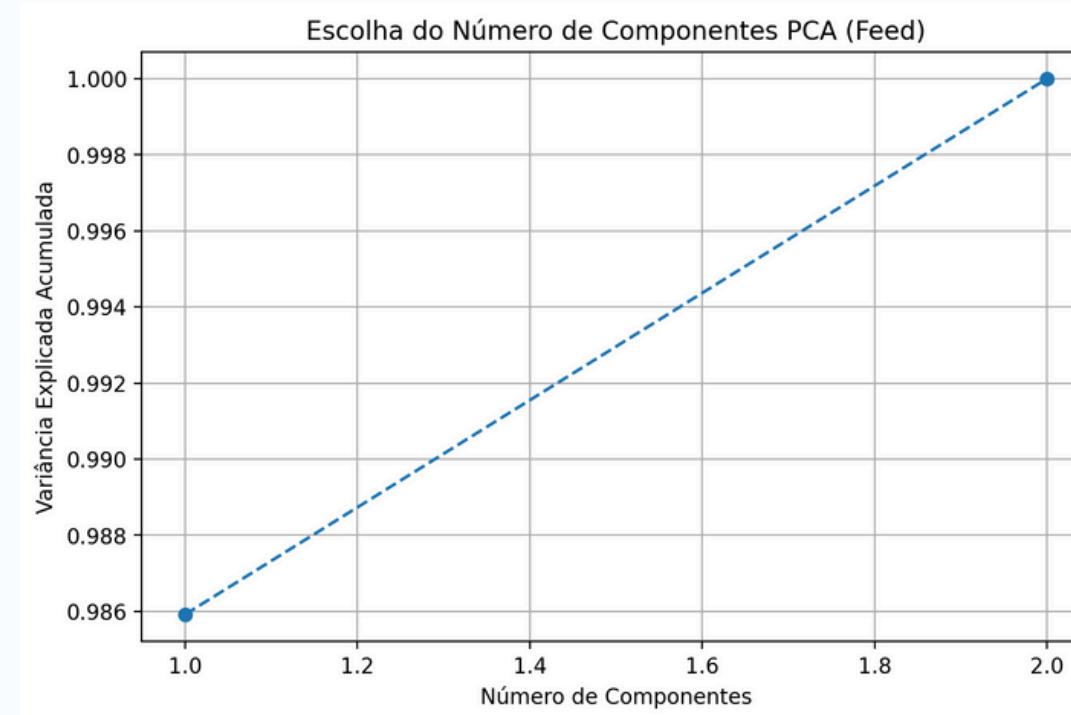
Level



EDA e Hipóteses

Aplicação do PCA

O PCA foi aplicado separadamente em cada novo dataframe, visando determinar o número ideal de componentes para capturar entre 93% e 95% da variância explicada.



Construção dos modelos

	R ² (médio)	RMSE (médio)
Lasso	0.678	0.639
Ridge	0.678	0.639
Random Forest	0.874	0.399
MLPRegressor	0.943	0.269

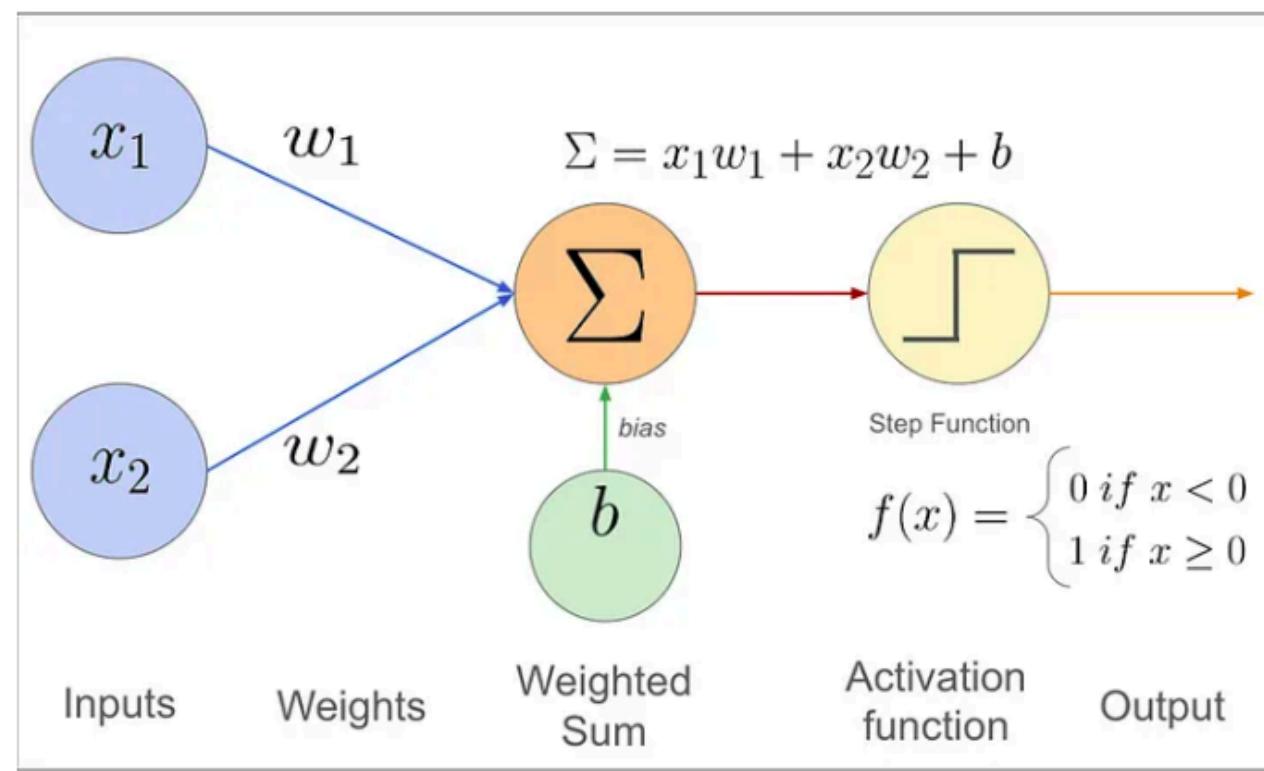
Modelos testados

- Lasso Regression: Usa regularização L1 (Linear)
- Ridge Regression: Usa regularização L2 (Linear)
- Random Forest Regressor: Baseado em múltiplas árvores de decisão (Não linear)
- MLPRegressor: Rede Neural

Métricas

- R² (Coeficiente de Determinação): Indica a proporção da variância dos dados explicada pelo modelo. Varia de 0 a 1, onde valores próximos de 1 indicam um melhor ajuste aos dados.
- RMSE (Root Mean Squared Error): Mede a diferença média entre os valores previstos e os reais, penalizando erros maiores. Quanto menor o RMSE, melhor o desempenho do modelo.

MLPRegressor

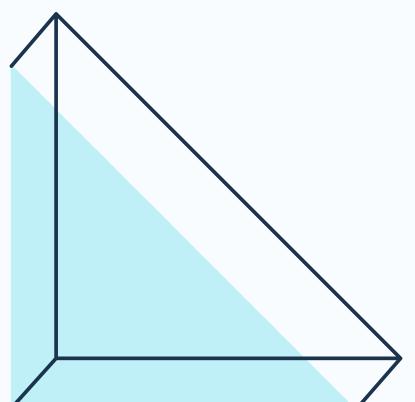


Fonte: <https://muneesba.medium.com/deep-learning-101-lesson-7-perceptron-f6a698d81be8>

Um pouco sobre

O Perceptron foi introduzido por McCulloch e Pitts como uma versão computacional de um neurônio biológico, podendo ser entendido como uma soma ponderada das entradas utilizadas em uma função de ativação para calcular uma saída.

Com base nessa abstração, Minsky e Papert propuseram perceptrons organizados em camadas **interconectadas** e confirmaram as Redes Neurais com mais de uma camada oculta como uma ferramenta poderosa, capaz de resolver problemas de **separação não linear**.

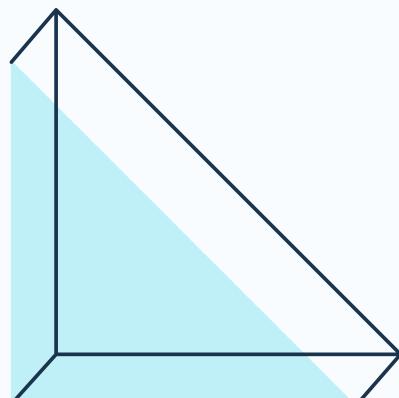


MLPRegressor

Procura pelos melhores hiper parâmetros

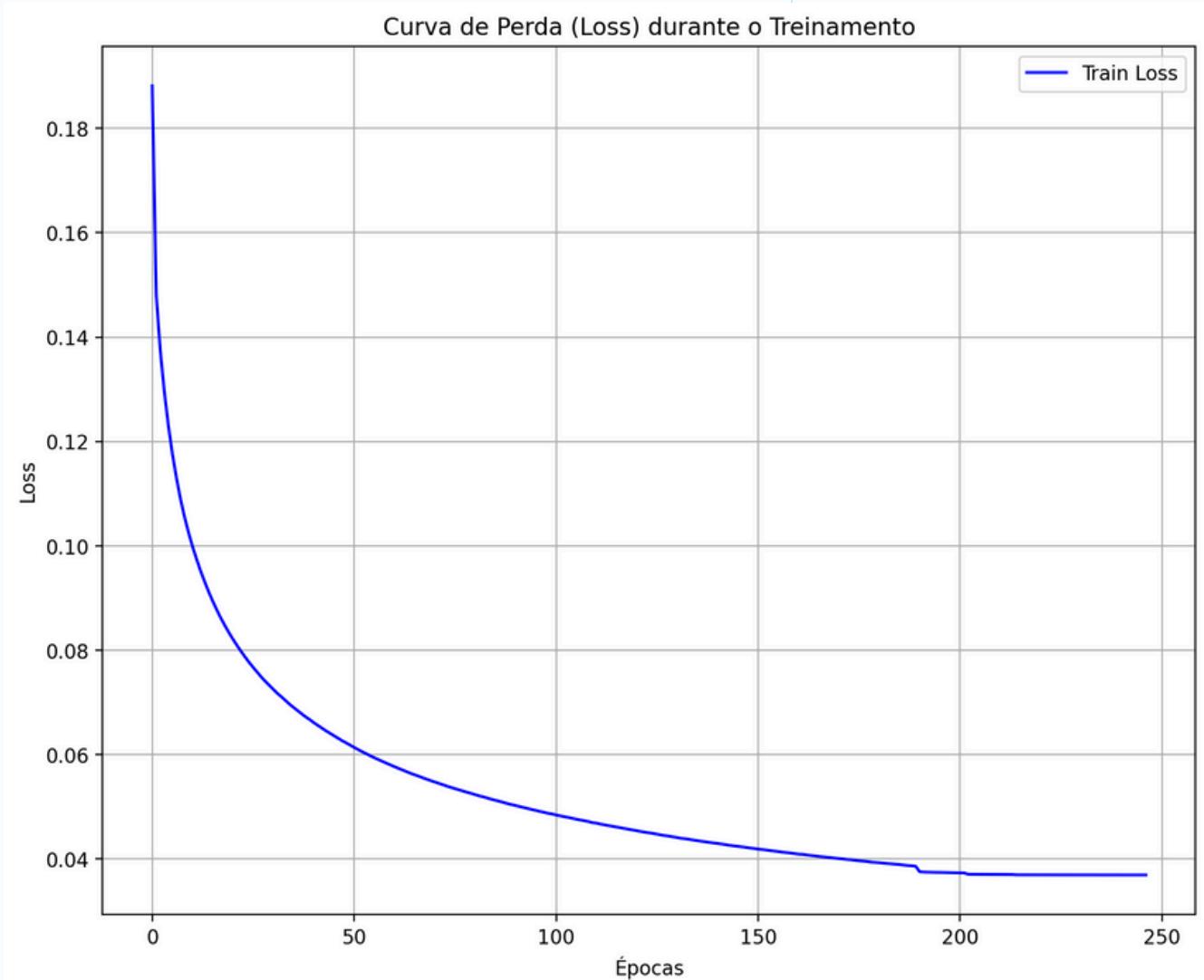
RandomizedSearchCV: É possível explorar diferentes combinações de hiper parâmetros, como o número de camadas, número de neurônios, taxa de aprendizado e outras variáveis importantes para otimizar a rede neural e melhorar seus resultados.

solver	learning_rate	hidden_layer_sizes	alpha	activation
'sgd'	'adaptive'	(200, 100)	0.001	'relu'

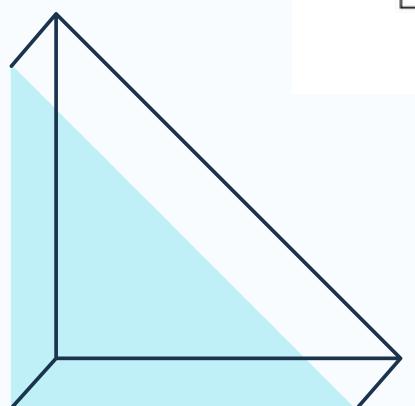


MLPRegressor

Métrica: Treino X Teste



	R ²	RMSE
Treino	0.943	0.269
Teste	0.939	0.278



Série predita X

Série real

