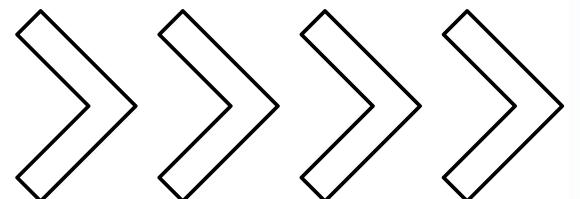


ANÁLISE PREDITIVA DE VENDAS COM BIG DATA USANDO APACHE SPARK E DOCKER SWARM



Grupo:

Anna Flávia Lopes Ferreira
João Marcos Marmontelo
Matheus Resende Furtado

Índice

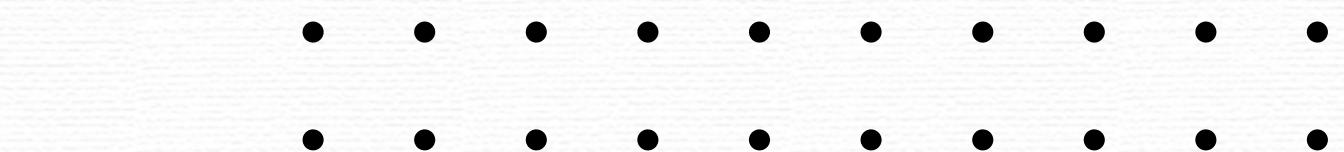
- | | | | |
|-----------|------------------------|-----------|-------------------------------------|
| 03 | Motivação | 09 | Metodologia |
| 04 | Objetivo | 10 | Resultados e Execução
do Projeto |
| 05 | Fonte de Dados | | |
| 06 | Por que é Big Data? | | |
| 07 | Fluxo de Arquitetura | | |
| 08 | Tecnologias Utilizadas | | |

-
-
-
-
-
-
-
-
-
-
-
-
-
-
-
-

Motivação

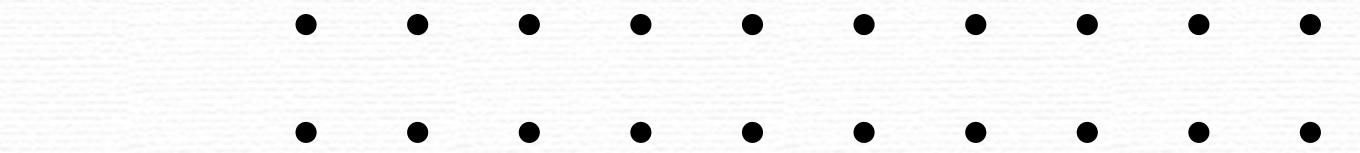
- Dados de vendas são fonte rica para análise e previsão.
- Pequenos negócios nem sempre têm ferramentas para isso.
- Decisões comerciais baseadas em dados melhoram a competitividade.
- Ferramentas de Big Data tornam isso possível de forma escalável.





Objetivo

- Construir um pipeline preditivo escalável.
- Prever valores de vendas com base em dados históricos.
- Utilizar ferramentas de Big Data modernas:
 - Apache Spark
 - PySpark
 - Spark MLlib
 - Docker Swarm

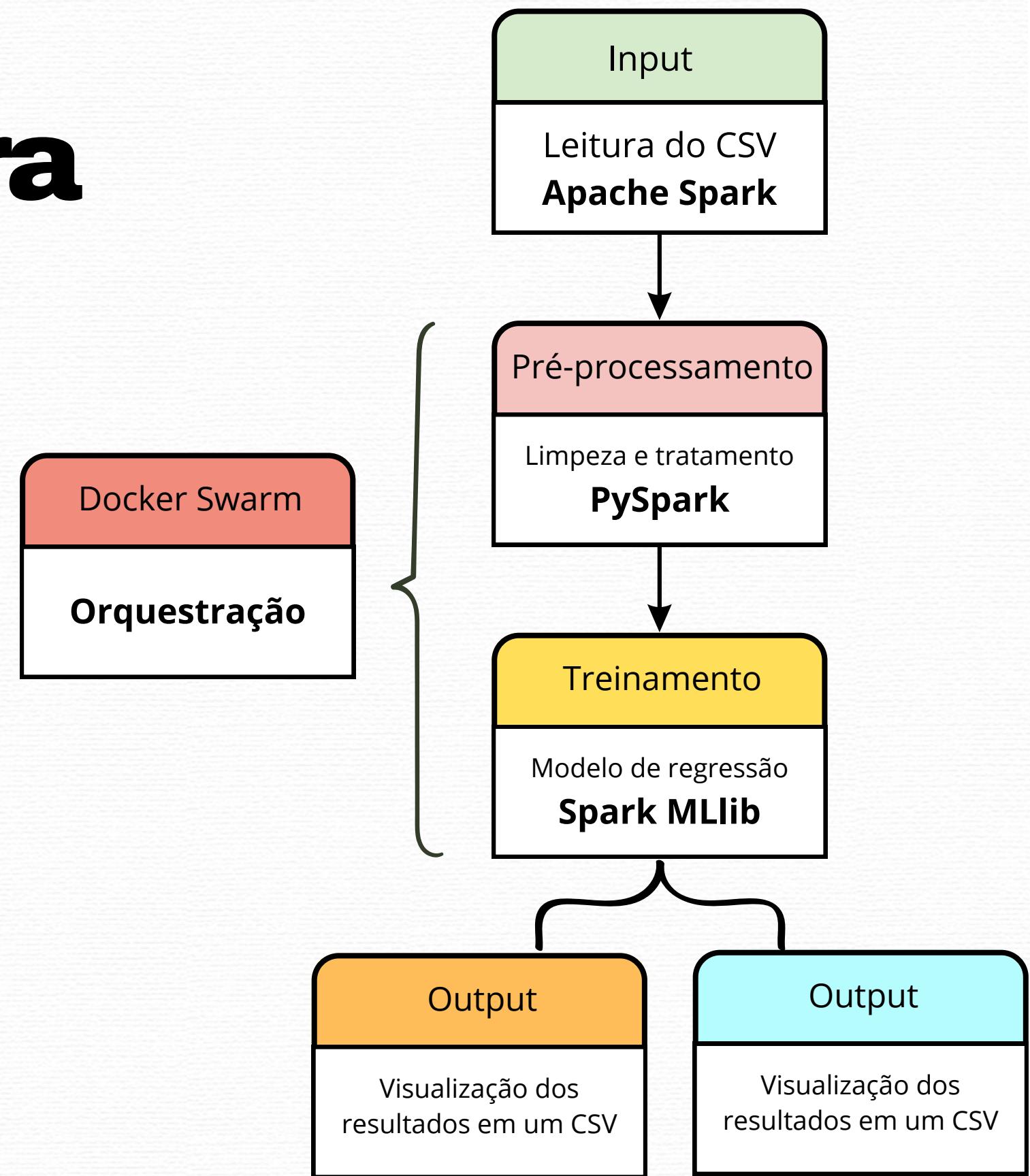


Fonte de Dados

[Big Sales Data - Kaggle](#)

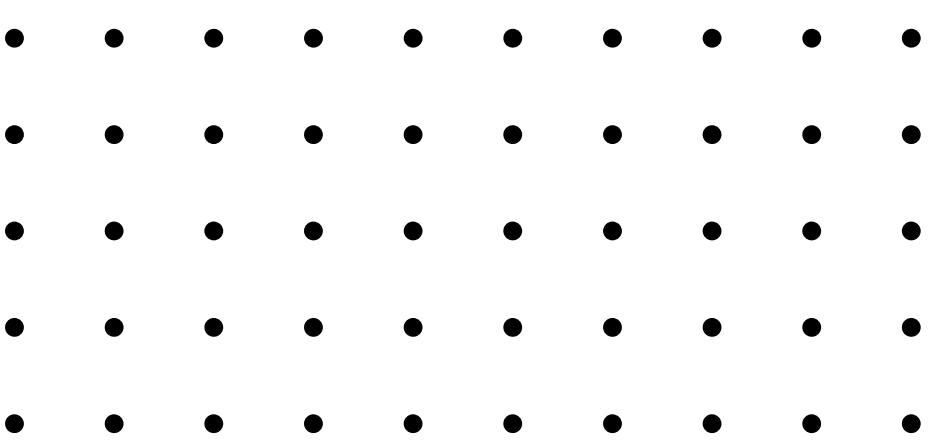
- **Tamanho:** 4,7GB
- **Linhas:** 19.666.763
- **Colunas:** 24
- **Colunas utilizadas:**
 - Data da venda
 - Loja
 - Fornecedor
 - Volume, custo, preço
 - Quantidade e valor da venda

Fluxo da arquitetura



Metodologia

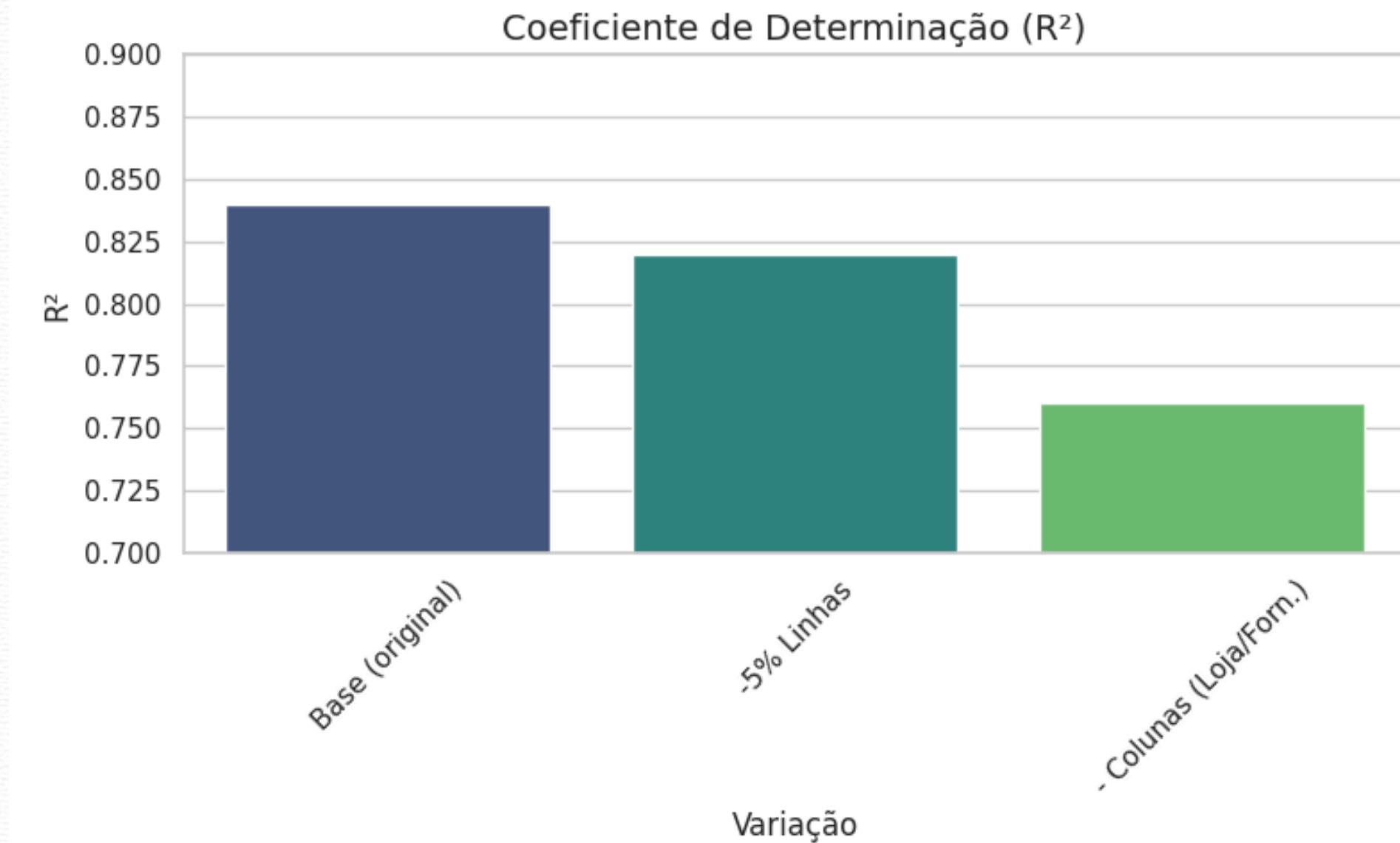
- Leitura e limpeza de dados com Spark
- Conversão de tipos, remoção de nulos
- Vetorização dos atributos (volume, custo, preço, etc.)
- Treinamento com 70% dos dados
- Teste com os 30% restantes
- Geração de arquivo com previsões



Discussão dos Resultados

Coeficiente de Determinação

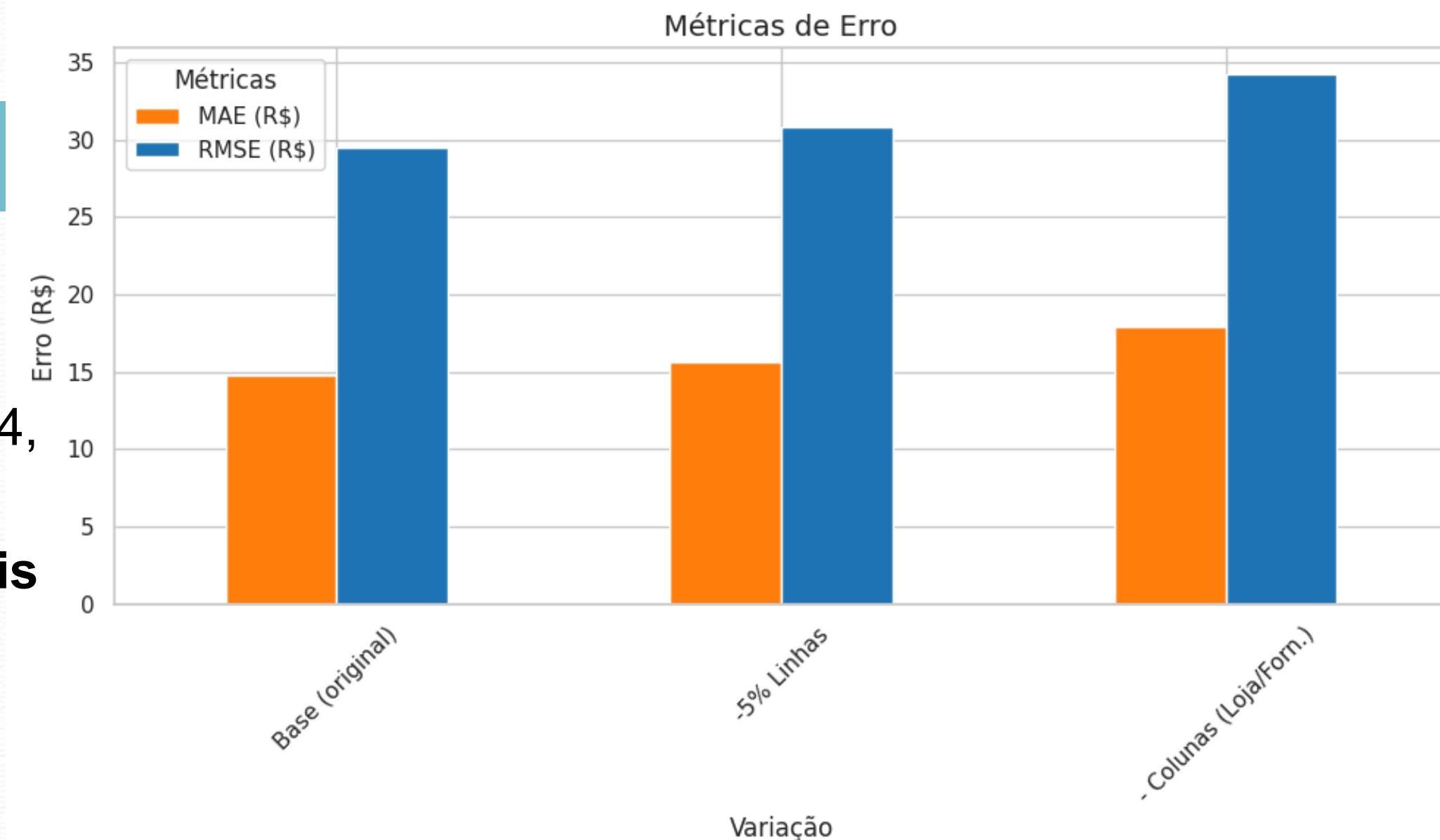
- **Base original:** $R^2 \approx 0,84$
- **Base com 5% das linhas removidas:** $R^2 \approx 0,82$
- **Base sem colunas informativas:** $R^2 \approx 0,75$
- **Conclusão:** Variáveis categóricas têm alto valor preditivo



Discussão dos Resultados

Métricas de Erro

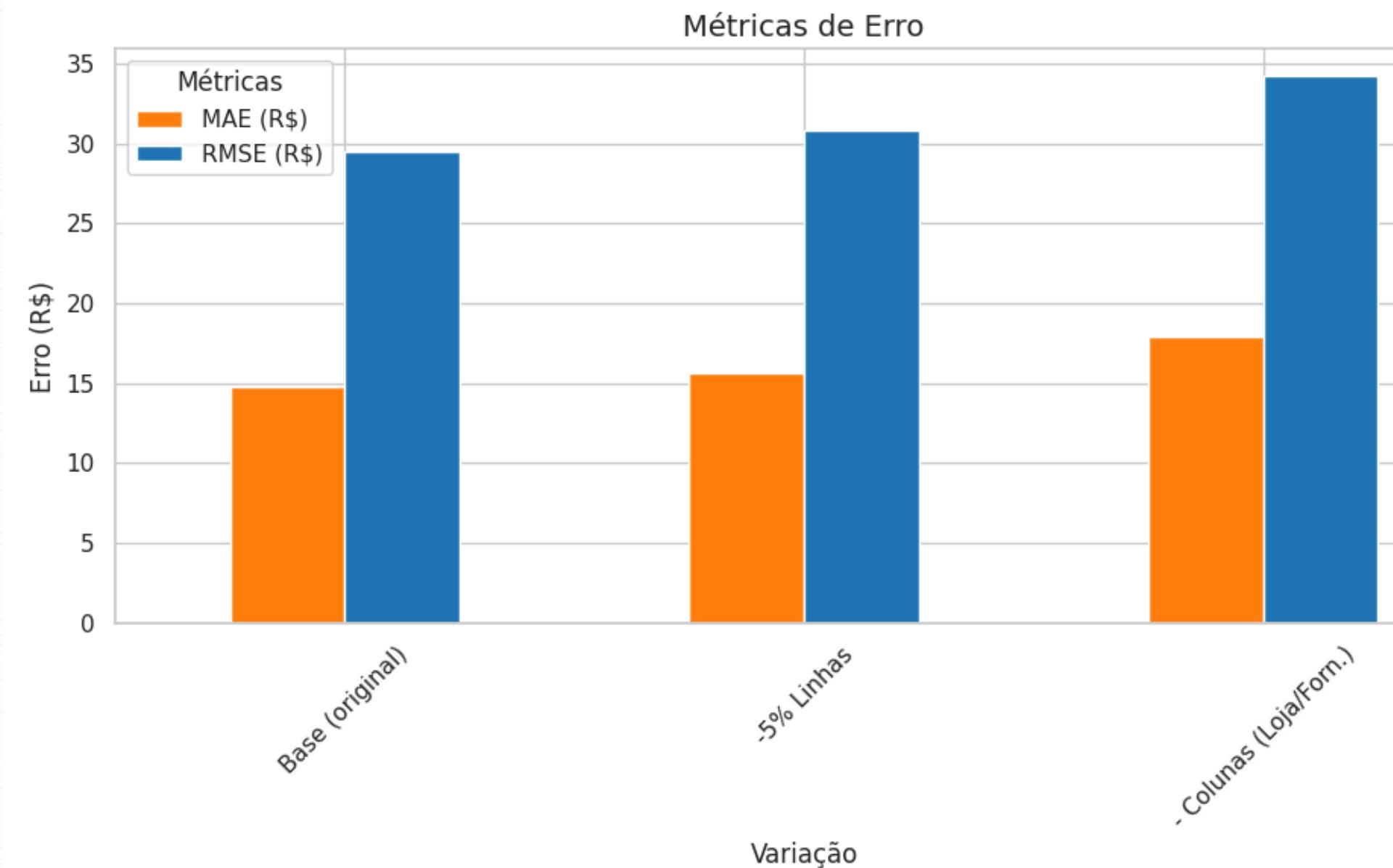
- **Base original:** MAE \approx R\$15, RMSE \approx R\$29
- **Sem 5% das linhas:** leve aumento nos erros
- **Sem colunas informativas:** RMSE $>$ R\$34, MAE \approx R\$18
- **Conclusão:** Qualidade dos dados tem mais impacto que volume



Discussão dos Resultados

Benchmark de Tempo

- Etapas medidas: **pré-processamento, treinamento, predição**
- Execuções rápidas: **todas abaixo de 9 minutos**
- Base sem colunas → **mais leve**
- Spark + Docker Swarm mostrou boa escalabilidade



Referências

- JUNIOR, Edilson Athayde. Apache Spark e Docker Swarm na prática. Medium, 22 jan. 2021. Disponível em: <http://medium.com/@edilsonathaydejunior/apache-spark-e-docker-swarm-na-pr%C3%A1tica-6795e0846f1c>.
- IBM Big Data – What is Big Data – United States. (s.d.). Disponível em <http://www.ibm.com/big-data/us/en/>;
- UNIVERSIDADE FEDERAL DO RIO DE JANEIRO. Big Data e os 5 V's. Grupo de Teleinformática e Automação – GTA/UFRJ, [s.d.]. Disponível em: JUNIOR, Edilson Athayde. Apache Spark e Docker Swarm na prática. Medium, 3 set. 2020. Disponível em: <https://medium.com/@edilsonathaydejunior/apache-spark-e-docker-swarm-na-pr%C3%A1tica-6795e0846f1c>. Acesso em: 9 jun. 2025.
- PIGMENT. Big Sales Data. Kaggle, [s.d.]. Disponível em: <https://www.kaggle.com/datasets/pigment/big-sales-data>. Acesso em: 9 jun. 2025.

Obrigado!