


**Grupo 2**

- Álvaro Mota
- João Rodrigues
- Maria Lúcia
- Rui Gomes



# Implementação de um modelo preditivo para antecipação da procura de produtos e otimização da gestão de stock

# Agenda

## **01** BUSINESS UNDERSTANDING

## **02** DATA UNDERSTANDING

## **03** DATA PREPARATION

## **04** MODELAÇÃO

## **05** AVALIAÇÃO

## **06** VISUALIZAÇÃO (PowerBI; Streamlit)

## **07** PROJECT OVERVIEW

## **08** CONCLUSÕES E RECOMENDAÇÕES

# Business Understanding

## Contexto

Cadeia de diversas lojas de venda a retalho distribuídas por várias cidades turcas.

Objetivo: Implementação de um modelo preditivo para antecipação da procura de produtos e otimização da gestão de stock



O maior problema da Gestão de Stock é o equilíbrio da disponibilidade de produtos sem acumulação excessiva de inventário ou ruturas de stock.

## Impacto

- Excesso
  - Custos elevados
  - Imobilização de capital
  - Risco de obsolescência
- Ruturas
  - Perdas de vendas
  - Insatisfação do cliente

# Business Understanding

## Oportunidade do Negócio

- **Valor da Abordagem Preditiva**

- Antecipar a procura com maior precisão
- Minimizar desperdícios e melhorar serviço ao cliente
- Apoiar decisões logísticas de forma fundamentada
- Otimizar capital de rotação através de melhor gestão de inventário

- **Impacto Esperado**

- Aumento de vendas e de disponibilidade
- Diminuição de excesso de inventário
- Melhoria na satisfação do cliente
- Otimização de custos operacionais



**Prever a demanda é um desafio complexo**

# Agenda

**01** BUSINESS UNDERSTANDING

**02** DATA UNDERSTANDING

**03** DATA PREPARATION

**04** MODELAÇÃO

**05** AVALIAÇÃO

**06** VISUALIZAÇÃO (PowerBI; Streamlit)

**07** PROJECT OVERVIEW

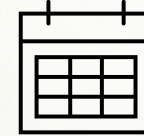
**08** CONCLUSÕES E RECOMENDAÇÕES



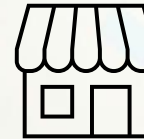
# Data Understanding



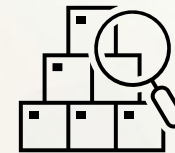
Registo diário com **8.9M** registos



02-01-2017 a 31-10-2019 = **1033 dias**

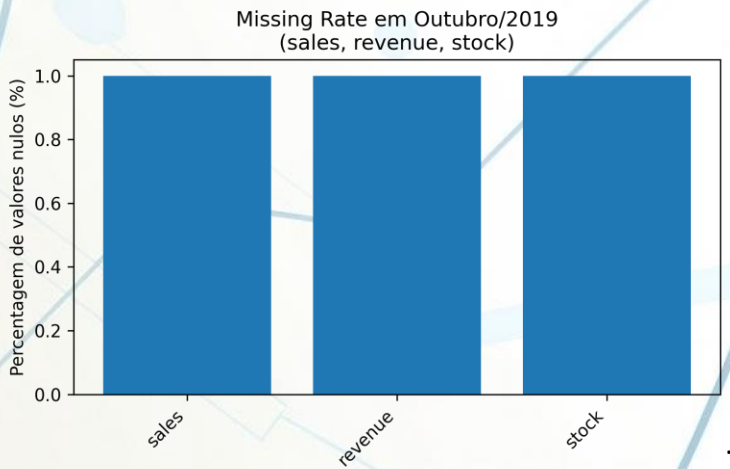
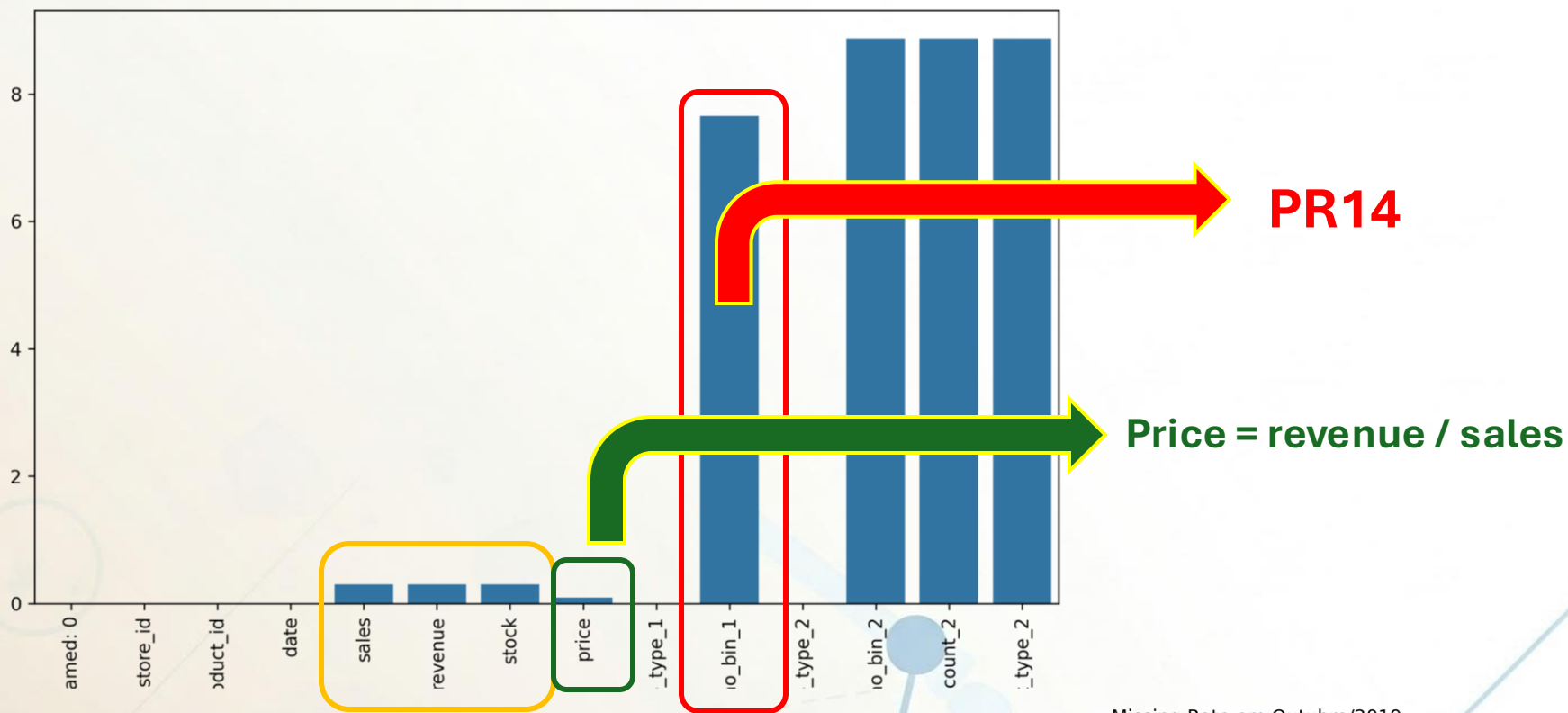
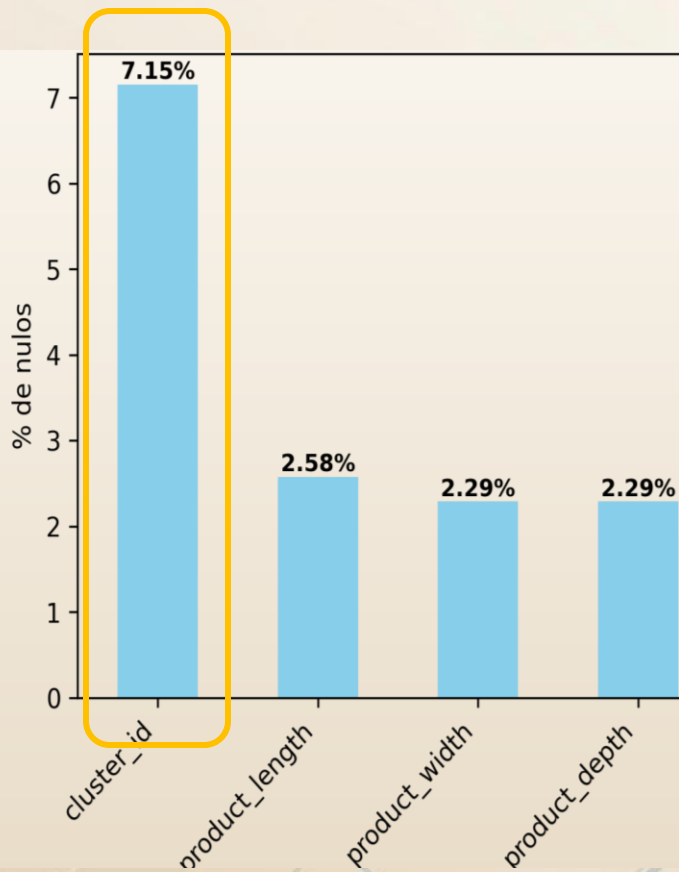


**63 lojas** em 19 cidades Turcas



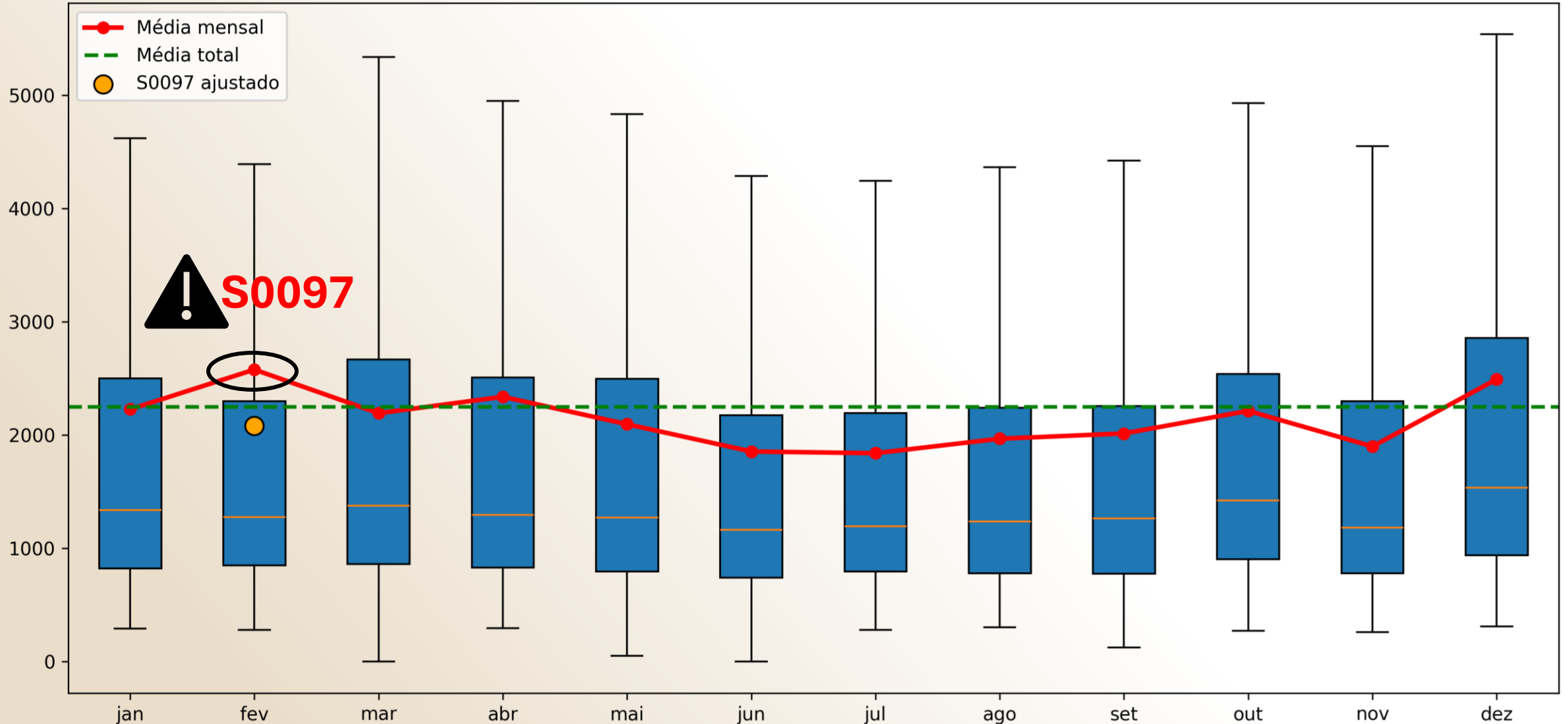
Info de 699 produtos, mas apenas **600** na presentes na tabela de factos.

# Data Understanding - Nulos



# Data Understanding - Outliers

Distribuição Mensal de Quantidades Vendidas

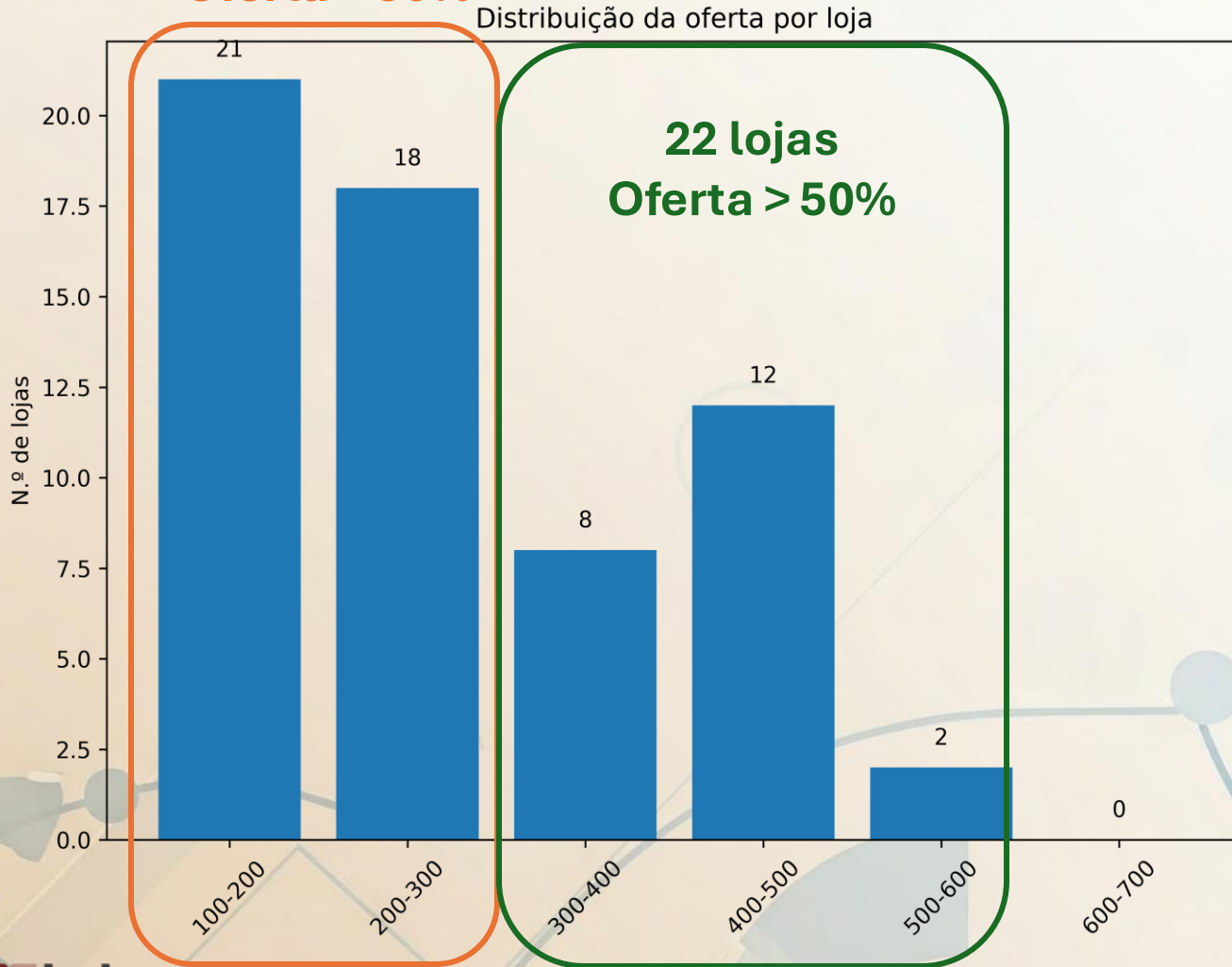




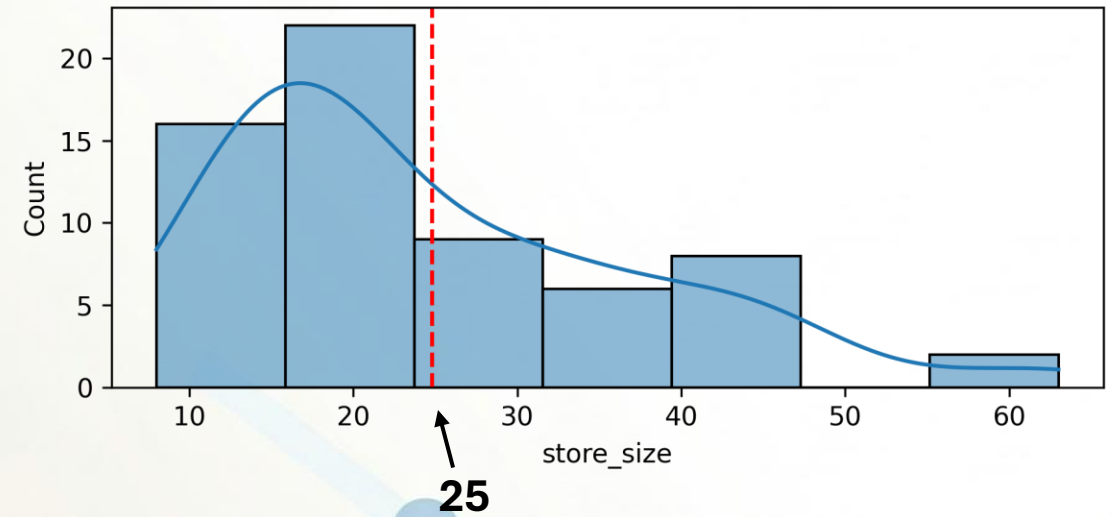
# Data Understanding – EDA

39 lojas

Oferta < 50%

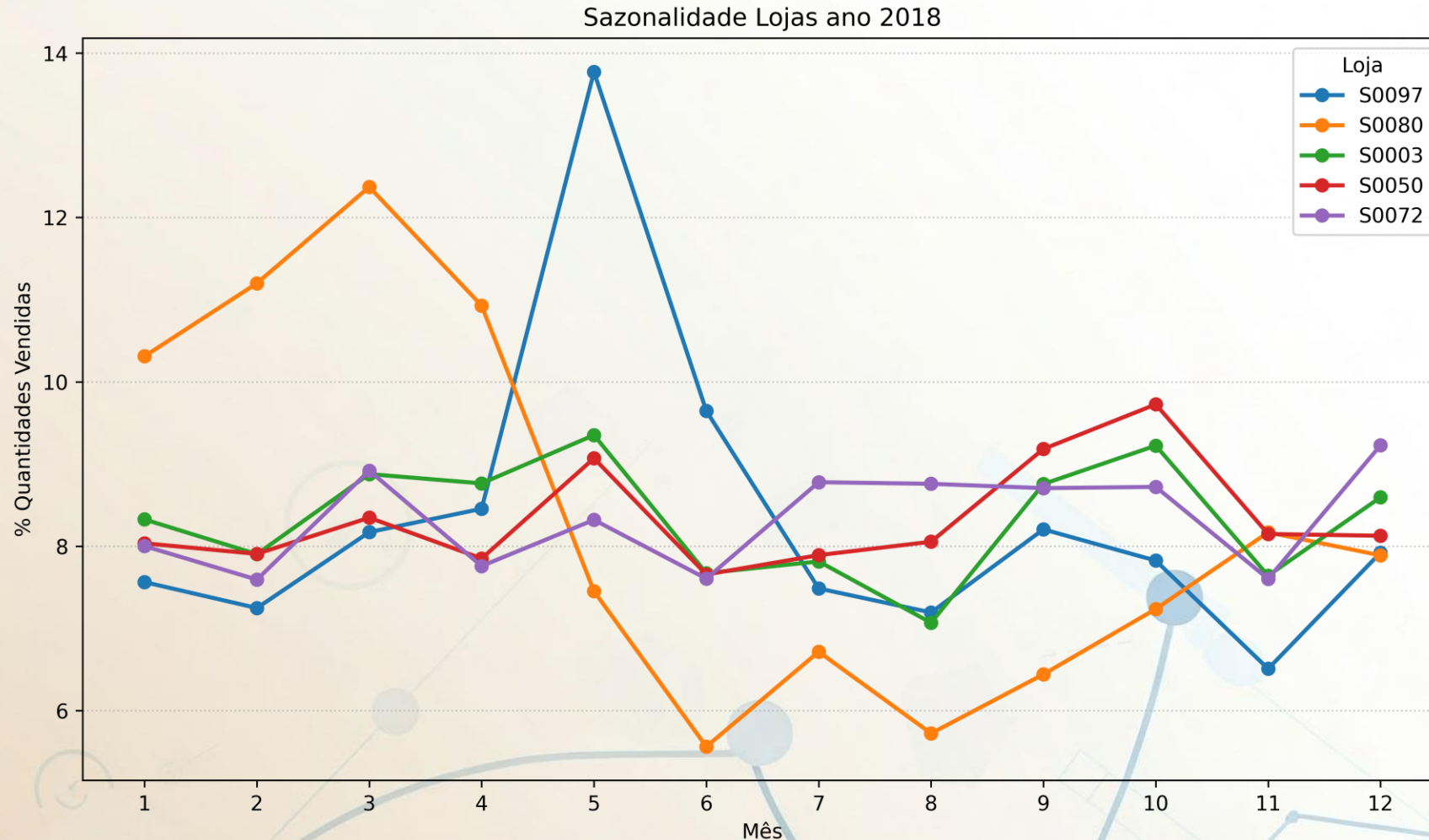


Distribuição de store\_size



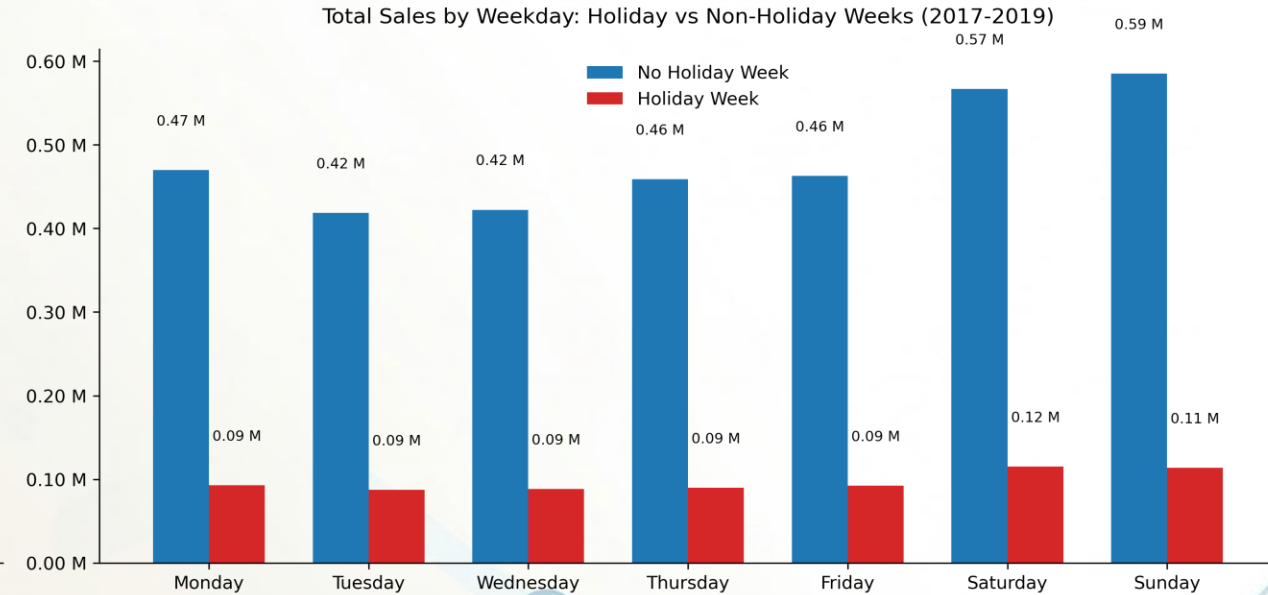
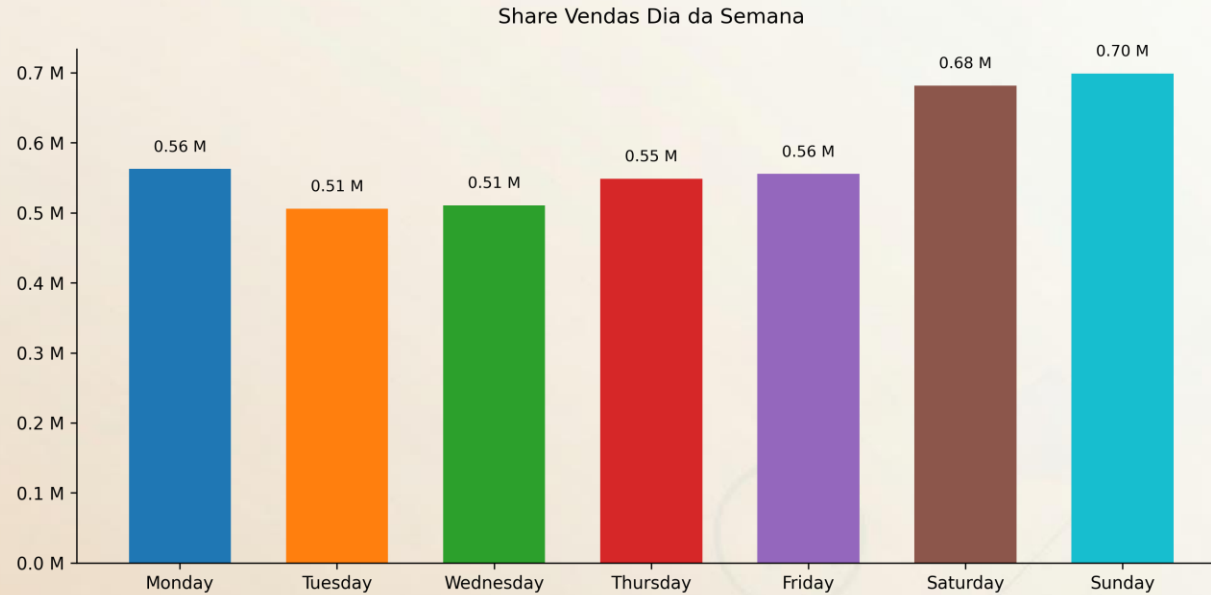
16704 combinações produto-loja distintos

# Data Understanding – EDA



**Lojas podem assumir comportamentos distintos de vendas ao longo do ano**

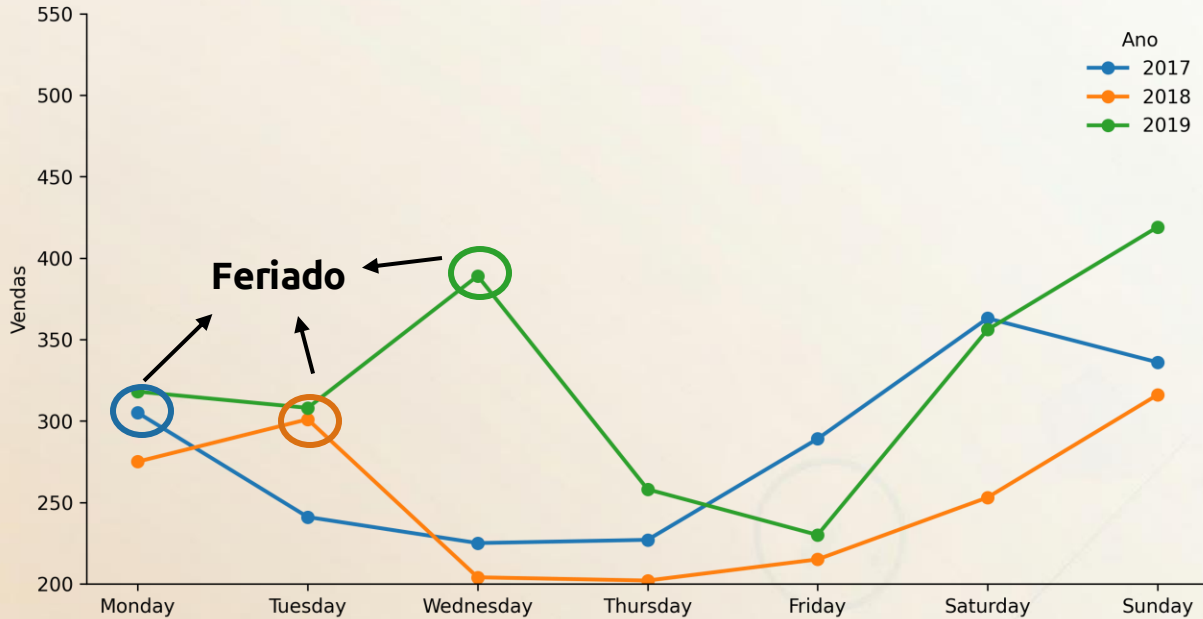
# Data Understanding – EDA



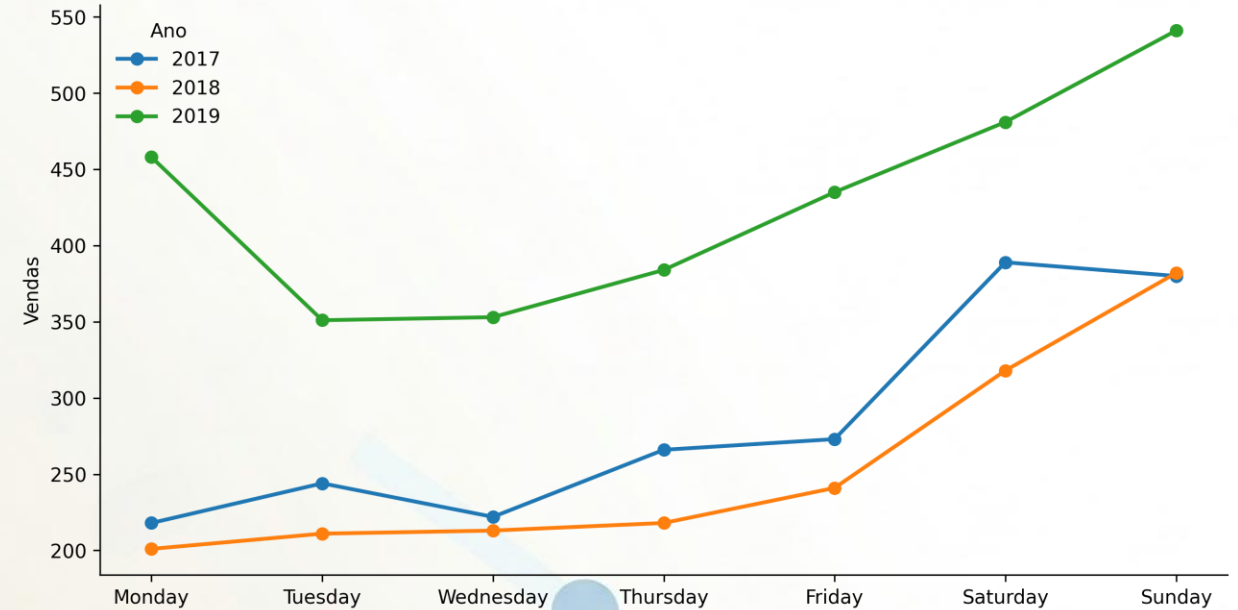
**Feriados durante a semana poderão distribuir as vendas pela semana se não forem ao FDS**

# Data Understanding – EDA

Product P0103 - Vendas em semana com feriado



Product P0103 - Vendas em semana sem feriado

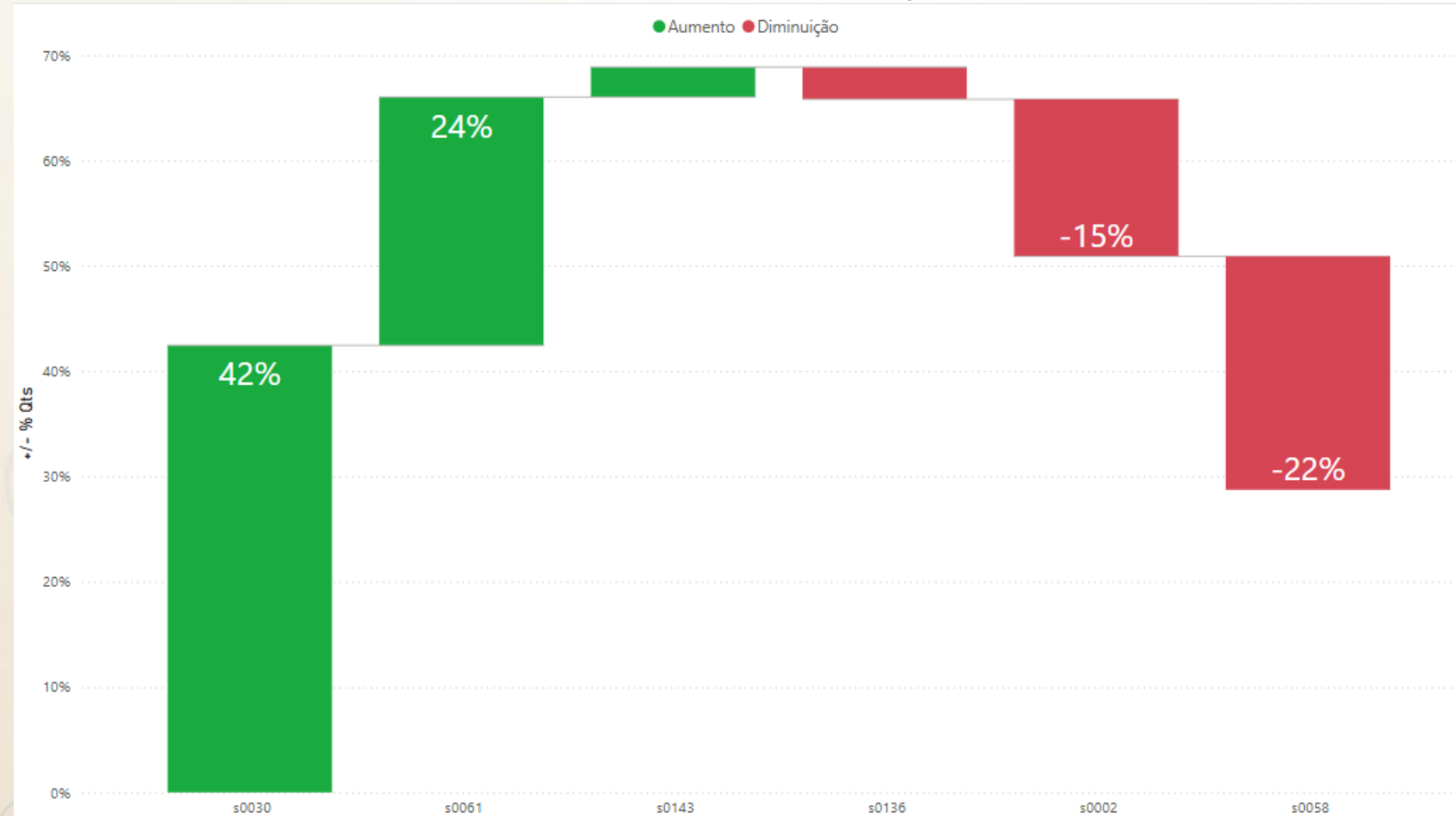


**Feriados poderão influenciar a procura e as vendas na restante semana**

# Data Understanding – EDA

Semana:							
1	8	15	22	29	36	43	50
2	9	16	23	30	37	44	51
3	10	17	24	31	38	45	52
4	11	18	25	32	39	46	
5	12	19	26	33	40	47	
6	13	20	27	34	41	48	
7	14	21	28	35	42	49	

Ritmo a 4 semanas de 6 lojas distintas



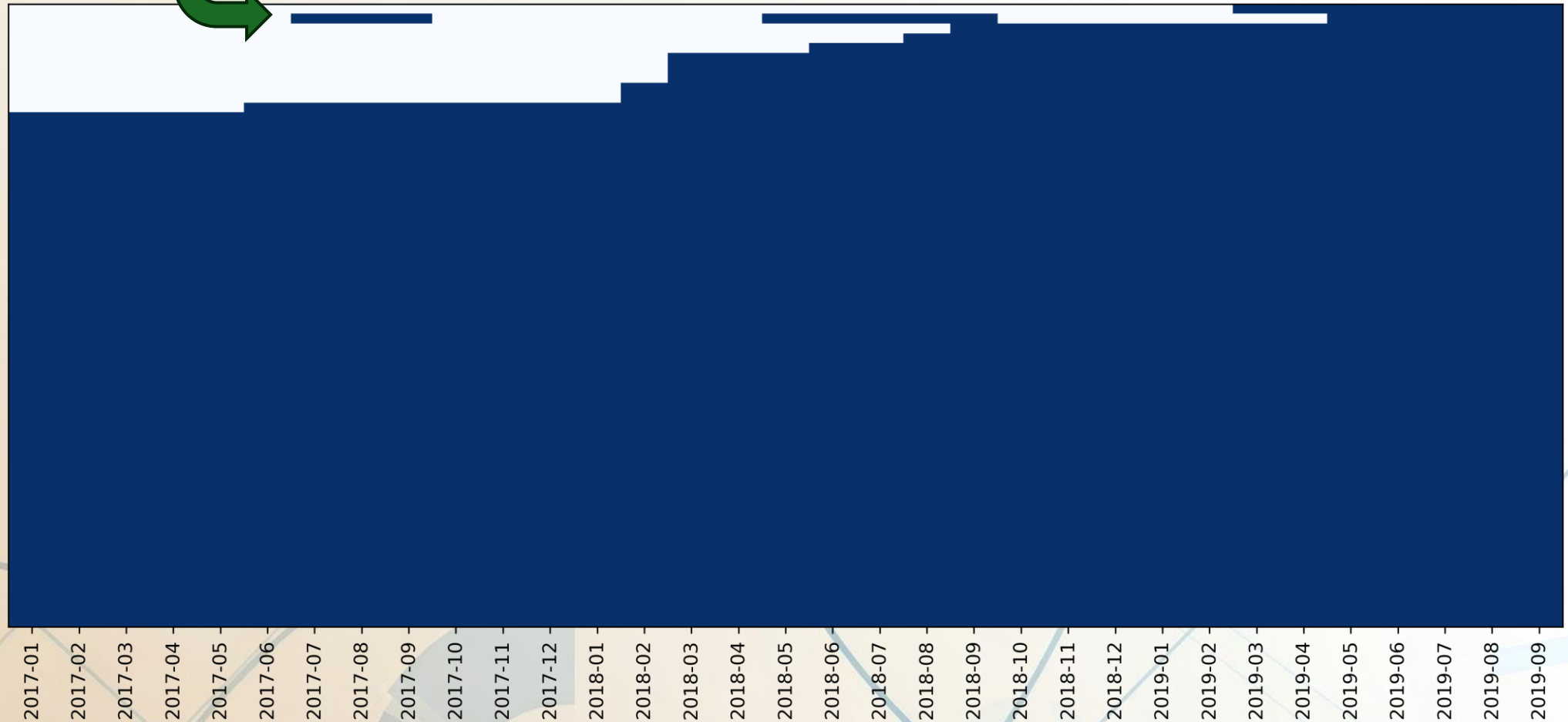
Ritmo a 4 semanas e última semana poderão ajudar a prever procura



# Data Understanding – EDA

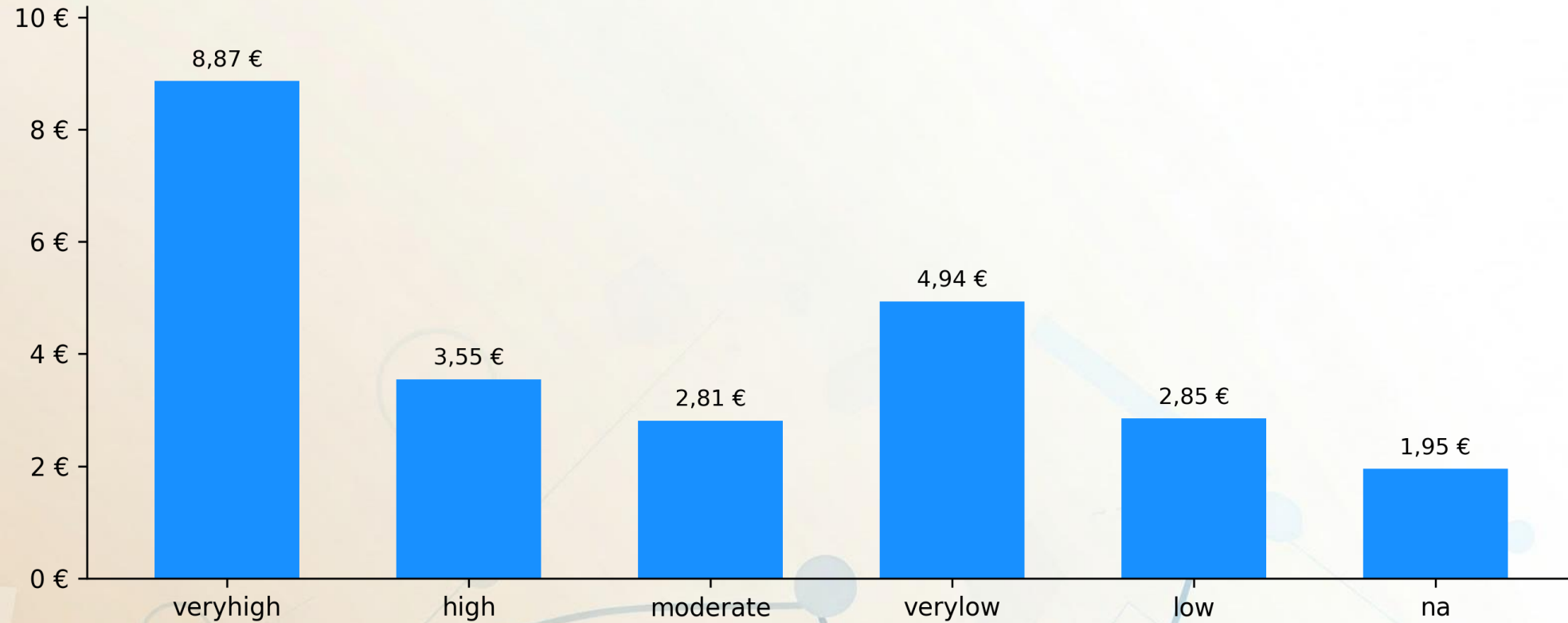
S0136, única loja sazonal

Mapa de calor: atividade de vendas por loja e mês  
(ordenado)



# Data Understanding – EDA

Cesto Médio por Nível de Promoção (promo\_bin\_1)



**Promoções poderão ter impacto na procura.**

# Agenda

**01** BUSINESS UNDERSTANDING

**02** DATA UNDERSTANDING

**03** DATA PREPARATION

**04** MODELAÇÃO

**05** AVALIAÇÃO

**06** VISUALIZAÇÃO (PowerBI; Streamlit)

**07** PROJECT OVERVIEW

**08** CONCLUSÕES E RECOMENDAÇÕES

# Data Preparation

## Arquitetura Medallion



Bronze

- **Ingestão** diária das 3 fontes: sales, cities e product na Bronze LakeHouse do Fabric

Silver

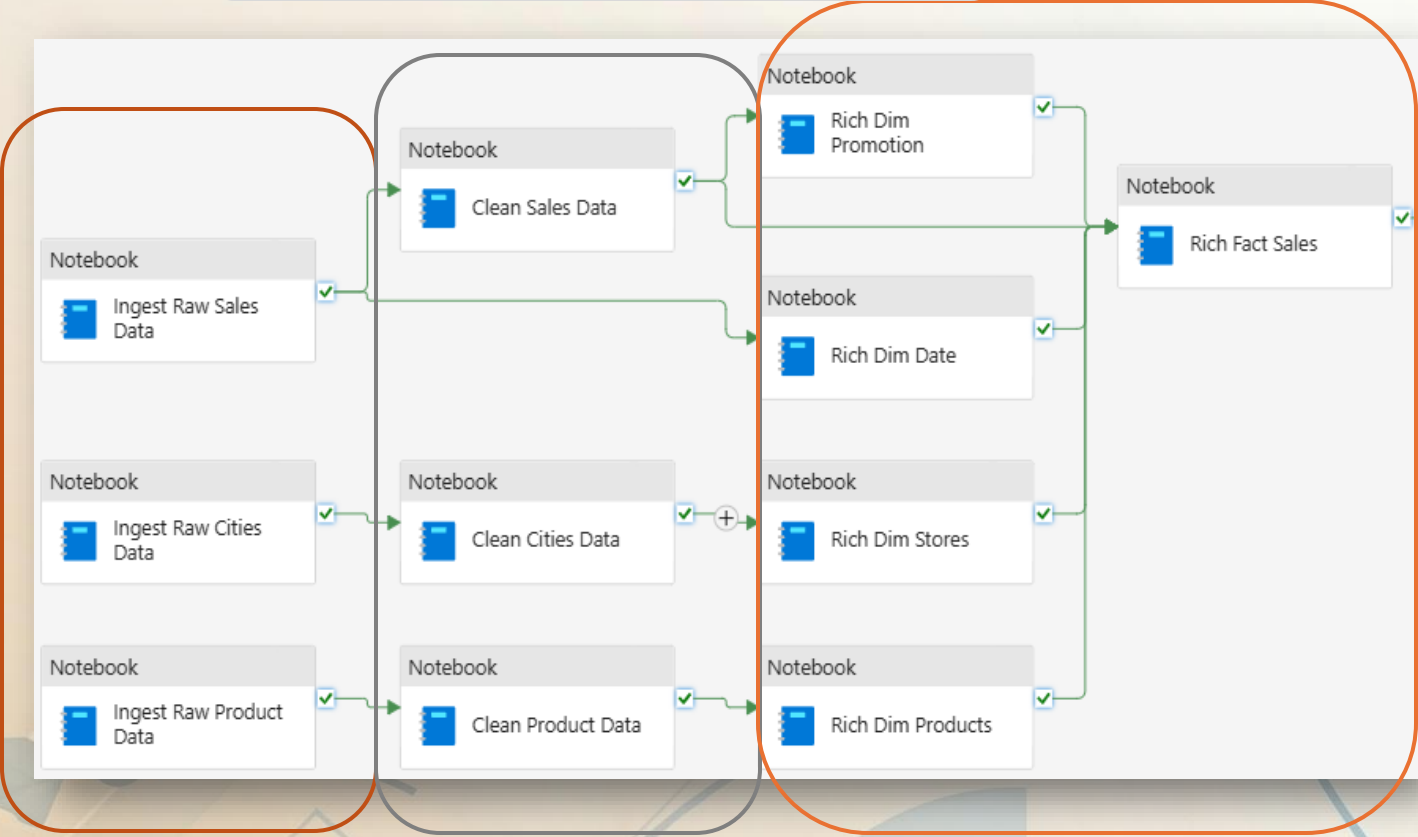
- **Validação da unicidade das tabelas**
- **Validação hierárquica das categorias:** confirmar que só existe relação com 1 hierarquia anterior
- **Eliminação de linhas:** Outubro 2019 – sem sales, revenue e stock
- **Correção de tipos** – datas para datatype, p.e
- **Normalização textual** – id's , nomes e categorias padronizadas para minúsculas e sem espaços
- **Correção textual das cidades** – tratar caracteres especiais nos nomes das cidades
- **Duplicados nas chaves primárias** - manter última ocorrência
- **Imputação de Nulos** - cluster\_id, price , promo\_bin\_1

Gold

- **Reorganização e reforço das tabelas com criação de variáveis relevantes**
  - Problema / negócio: 5\_business\_days\_offset , week\_start\_date, week\_end\_date
  - EDA : is\_weekend, is\_holiday, is\_promotion\_active
- **Criação do StarSchema**
  - Tabelas dimensão : date, product, store, promotion
  - Tabela de factos: sales

# Data Preparation

## Arquitetura Medallion



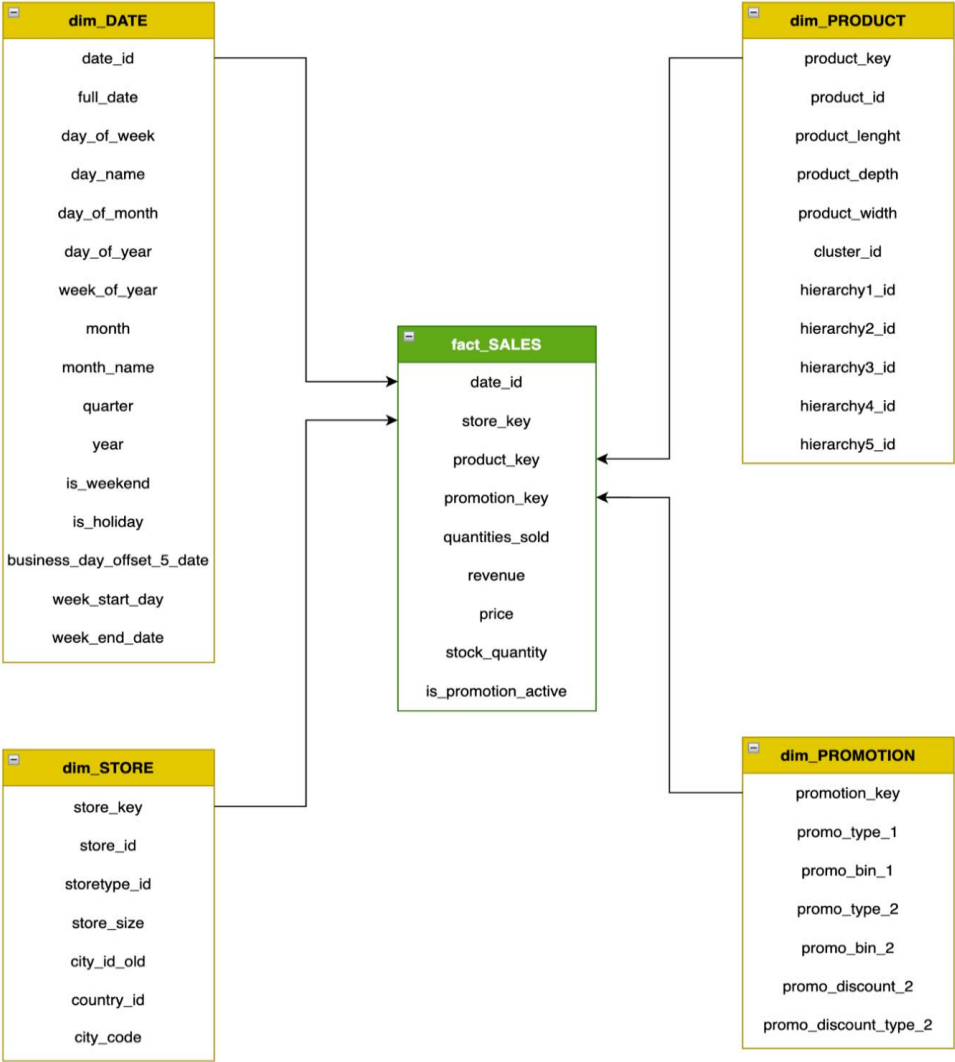
BRONZE LH

SILVER LH

GOLD LH

BI & Analytics

Machine Learning





# Agenda

**01** BUSINESS UNDERSTANDING

**02** DATA UNDERSTANDING

**03** DATA PREPARATION

**04** MODELAÇÃO

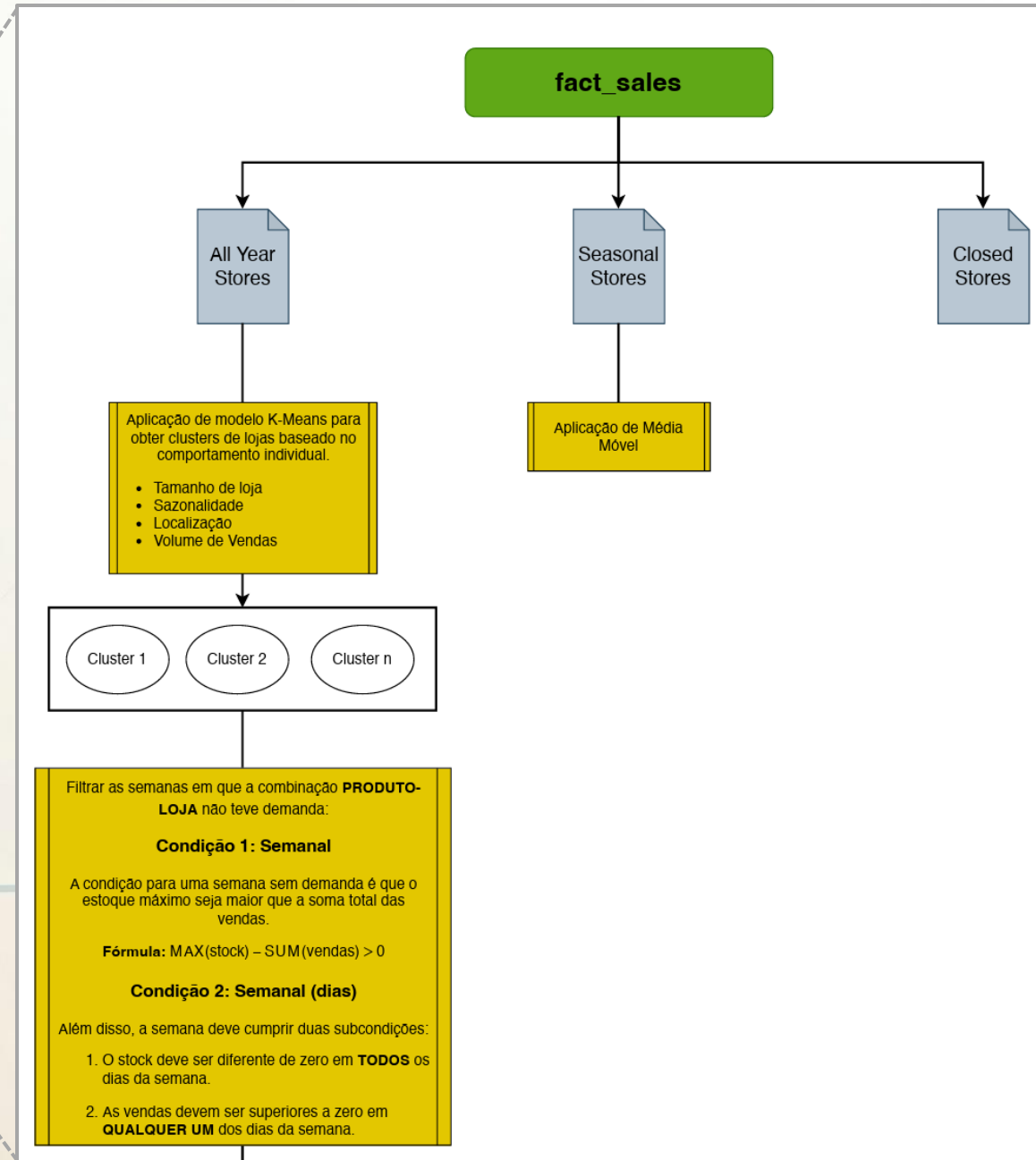
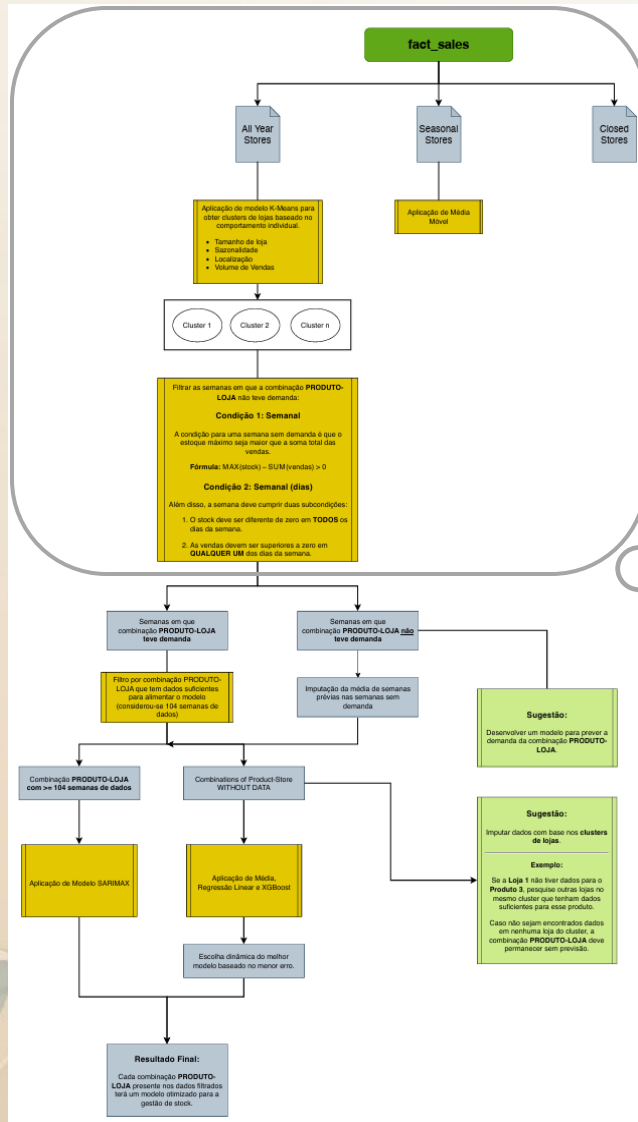
**05** AVALIAÇÃO

**06** VISUALIZAÇÃO (PowerBI; Streamlit)

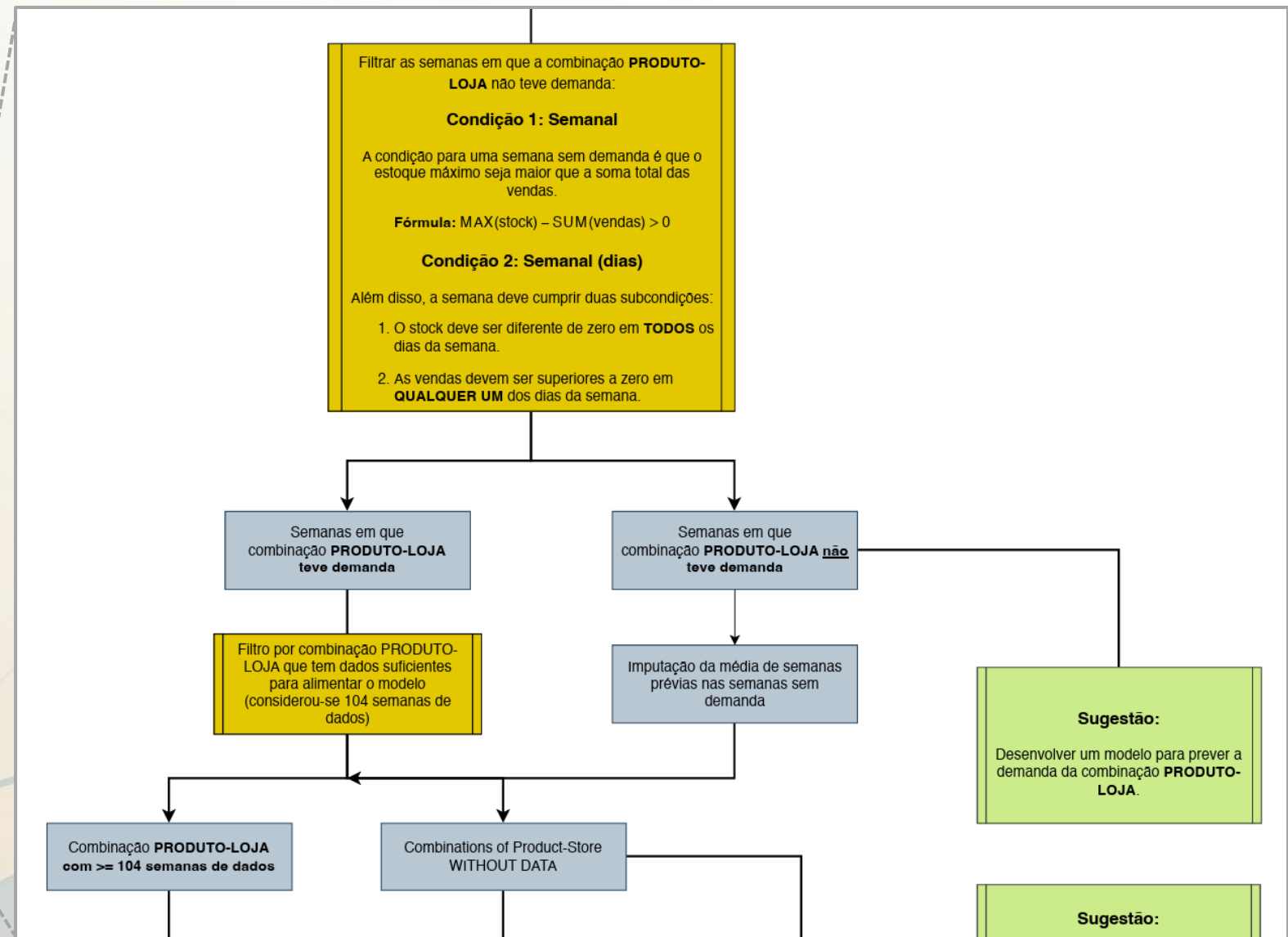
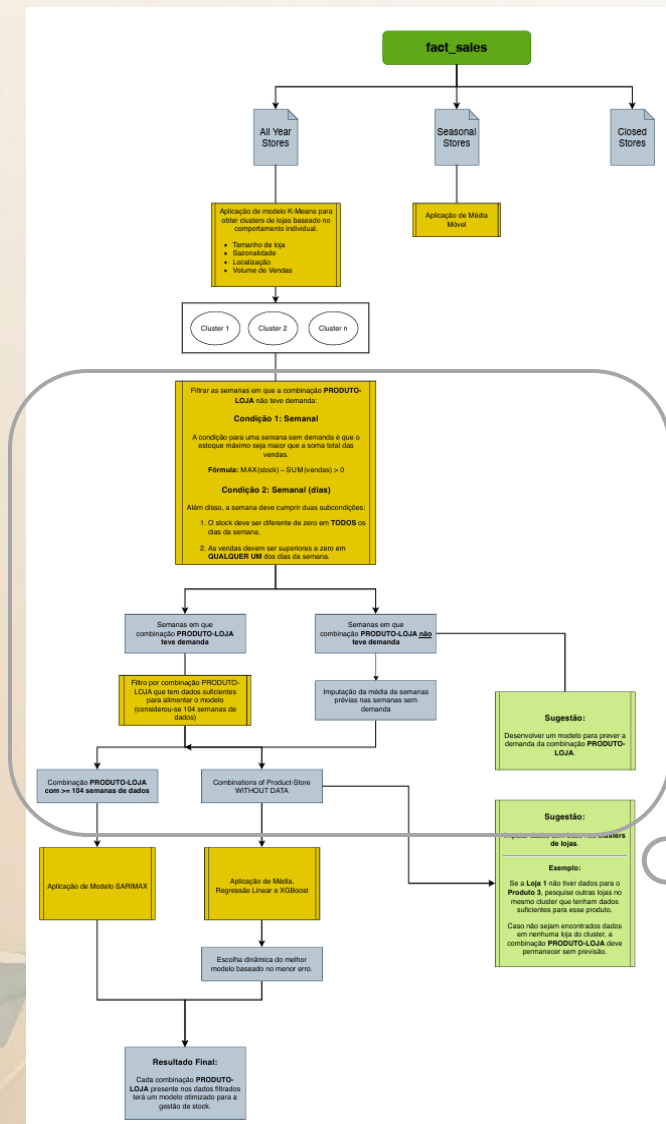
**07** PROJECT OVERVIEW

**08** CONCLUSÕES E RECOMENDAÇÕES

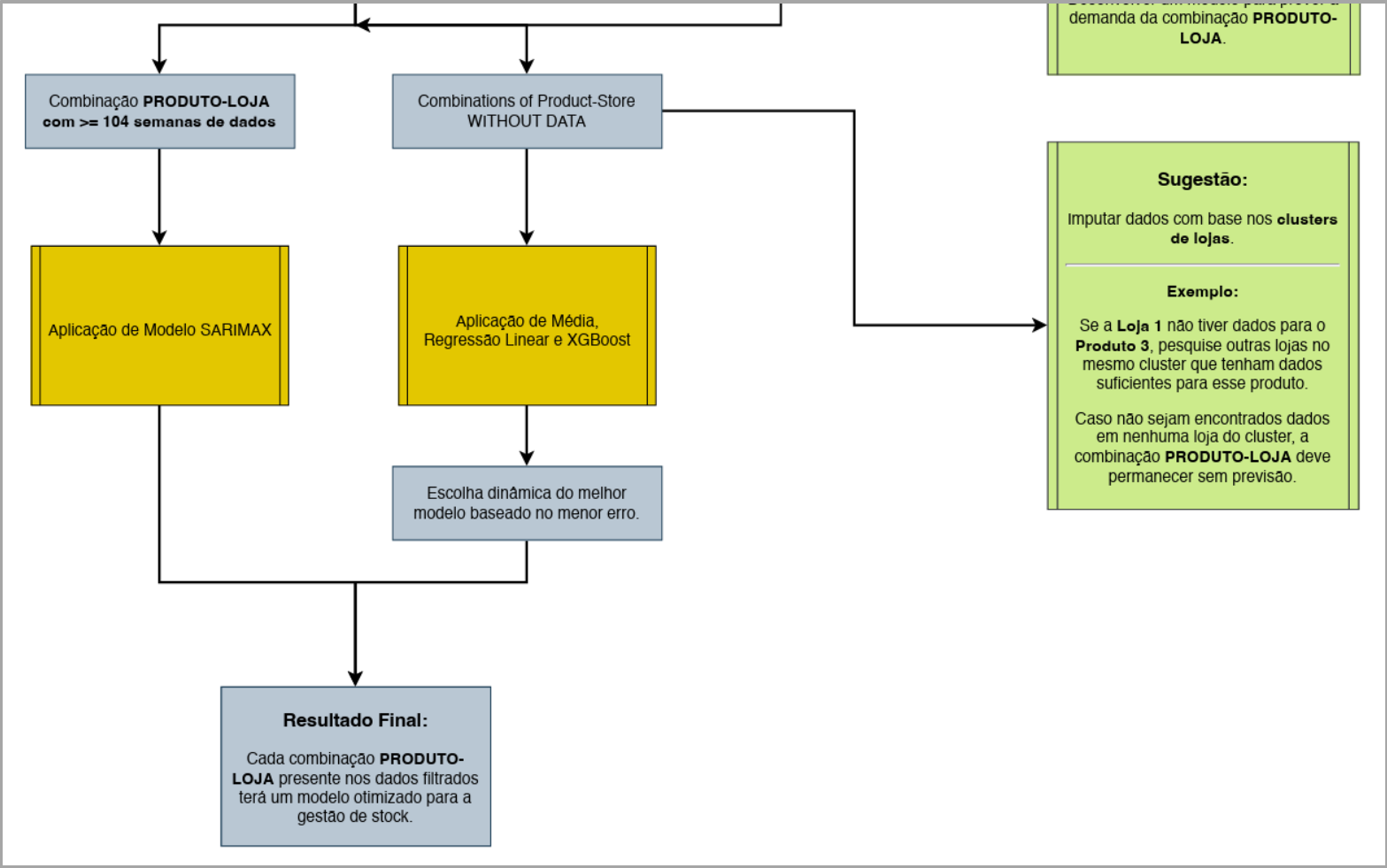
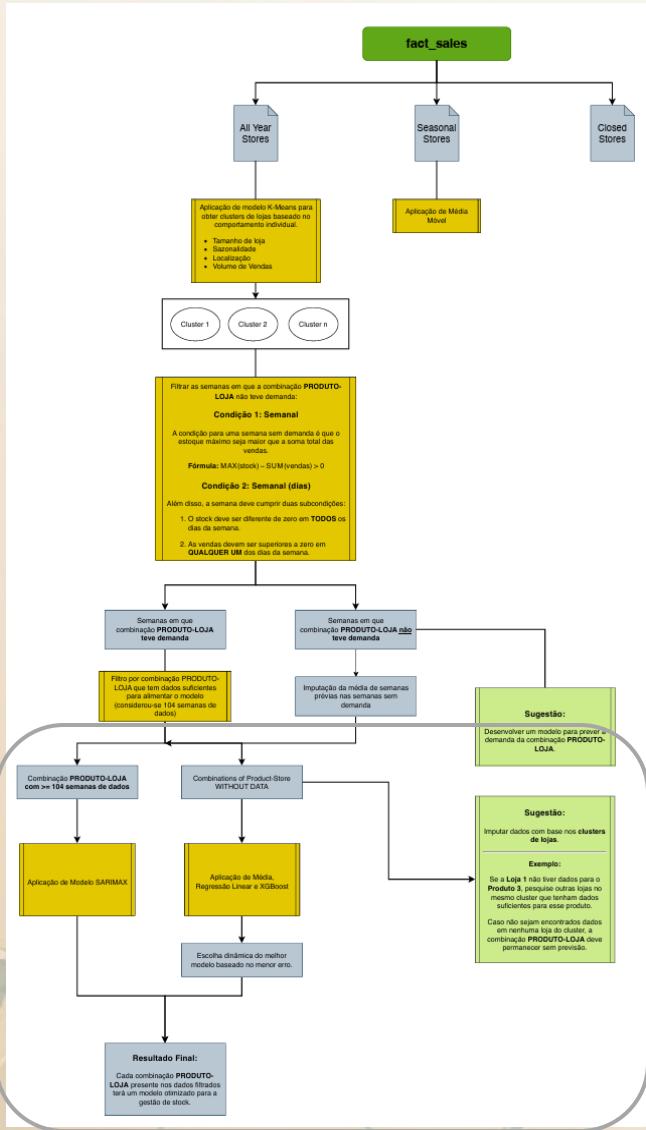
# Modelação



# Modelação



# Modelação



# Modelação - Feature Engineering

Grupo de Features	Justificação
Lag Features	<i>lag_1, lag_4</i> capturam padrões recentes de procura. Ajudam os modelos a aprender dependências temporais através do uso de valores passados como preditores.
Médias Móveis	<i>rolling_mean_4</i> suaviza flutuações de curto prazo e capta tendências locais. Evita reações exageradas a picos ou ruídos nos dados.
Transformação de Preço	<i>log_price</i> reduz a assimetria nos dados de preço, especialmente quando o preço varia exponencialmente.
Métricas de Inventário	<i>inv_days_cover</i> expressa quantos dias/semanas de stock estão disponíveis com base nas vendas recentes ( <i>rolling_mean_4</i> ).
Sazonalidade	<i>month_sin, month_cos, fourier_sin_52_k1, fourier_cos_52_k1</i> codificam padrões cíclicos (mensais e anuais). Features padrão em modelos de séries temporais para representar sazonalidade sem codificar eventos no calendário.
Índice Temporal	<i>t</i> acompanha a progressão temporal por combinação produto-loja.



# Agenda

**01** BUSINESS UNDERSTANDING

**02** DATA UNDERSTANDING

**03** DATA PREPARATION

**04** MODELAÇÃO

**05** AVALIAÇÃO

**06** VISUALIZAÇÃO (PowerBI; Streamlit)

**07** PROJECT OVERVIEW

**08** CONCLUSÕES E RECOMENDAÇÕES

# Segmentação de lojas

## Cluster 0

### Lojas de pequena dimensão e de performance básica

- Pequena superfície
- Baixo volume por transação, mas com frequência elevada
- Perfil tipo: lojas de conveniência, ou pontos de venda rápida

## Cluster 2

### Lojas de media dimensão e de performance eficiente

- Média superfície
- Excelente performance diária
- Equilíbrio entre volume e frequência de vendas
- Perfil tipo: lojas bem localizadas em zonas de alto tráfego

## Cluster 1

### Lojas de grande dimensão e de performance excecional

- Grande superfície
- Vendas massivas mas esporádicas
- Alto volume de receita e quantidade por transação
- Perfil tipo: loja flagship, centro de distribuição ou loja corporativa B2B

## Cluster 3

### Lojas premium com vendas pontuais

- Média a grande superfície
- Vendas de alto valor
- Baixa frequência diária mas alto ticket médio
- Perfil tipo: lojas especializadas ou de produtos premium

# Modelos para Previsão

## Produtos com dados suficientes

### ARIMA

- Erro médio absoluto (MAE) foi de **0.53**.
- MAE representa **12.19%** das vendas médias e **4.49%** do desvio padrão.
- **R<sup>2</sup> foi 0** - indica que o modelo tem um desempenho igual a uma previsão baseada na média.

**Erros médios baixos, mas nenhuma explicabilidade da variabilidade (R<sup>2</sup>).**



## Observações Críticas e Próximos Passos

- Desenvolvimento de modelo de previsão de demanda.
- **Desafio de Dados:** A alta frequência de vendas a zero (demanda intermitente) limita a capacidade dos modelos de explicar a variabilidade (R<sup>2</sup>).
- Explorar **modelos mais adaptados** a este tipo de dados para capturar melhor a dinâmica da demanda.

## Produtos sem dados suficientes

### XGBoost

- Selecionado como o melhor modelo para **5.049 combinações**.
- Erro médio absoluto (MAE) foi de **0.061**.
- MAE representa **1.77%** das vendas médias e **0.48%** do desvio padrão.
- R<sup>2</sup> foi **-0.0017** - indica que o modelo tem um desempenho pior do que uma previsão baseada na média.

### Baseline (Média)

- Selecionada como o melhor modelo para **3.864 combinações**.
- Erro médio absoluto (MAE) foi de **0.012**.
- MAE corresponde a **0.36%** das vendas médias e **0.10%** do desvio padrão.
- R<sup>2</sup> foi **0.113**.

### Regressão Linear

- Selecionado como o melhor modelo para **3.554 combinações**.
- Erro médio absoluto (MAE) foi de **0.19**.
- MAE representa **5.67%** das vendas médias e **1.55%** do desvio padrão.
- R<sup>2</sup> foi **0.033**.

**Erros médios baixos, mas pouca explicabilidade da variabilidade (R<sup>2</sup>).**

# Agenda

**01** BUSINESS UNDERSTANDING

**02** DATA UNDERSTANDING

**03** DATA PREPARATION

**04** MODELAÇÃO

**05** AVALIAÇÃO

**06** VISUALIZAÇÃO (PowerBI; Streamlit)

**07** PROJECT OVERVIEW

**08** CONCLUSÕES E RECOMENDAÇÕES

# Visualização e Experimentação

PowerBI e Streamlit

Deploy

## Stock Forecasting App - Model Experiments

Products with Data Products with No Data

### Forecasting for Products with Enough Data (ARIMAX)

This section uses an ARIMAX model, which incorporates exogenous variables.

Select Product and Store Combination:

Product: 2, Store: 17

Select Features (Exogenous Variables):

is\_weekend ×

is\_holiday ×

store\_size ×

promo\_flag ×

lag\_1 ×

lag\_4 ×

rolling\_mean\_4 ×

log\_price ×

#### ARIMAX Parameters (p, d, q)

ARIMA 'p' (AR order)

1-+

ARIMA 'd' (I order)

1-+

ARIMA 'q' (MA order)

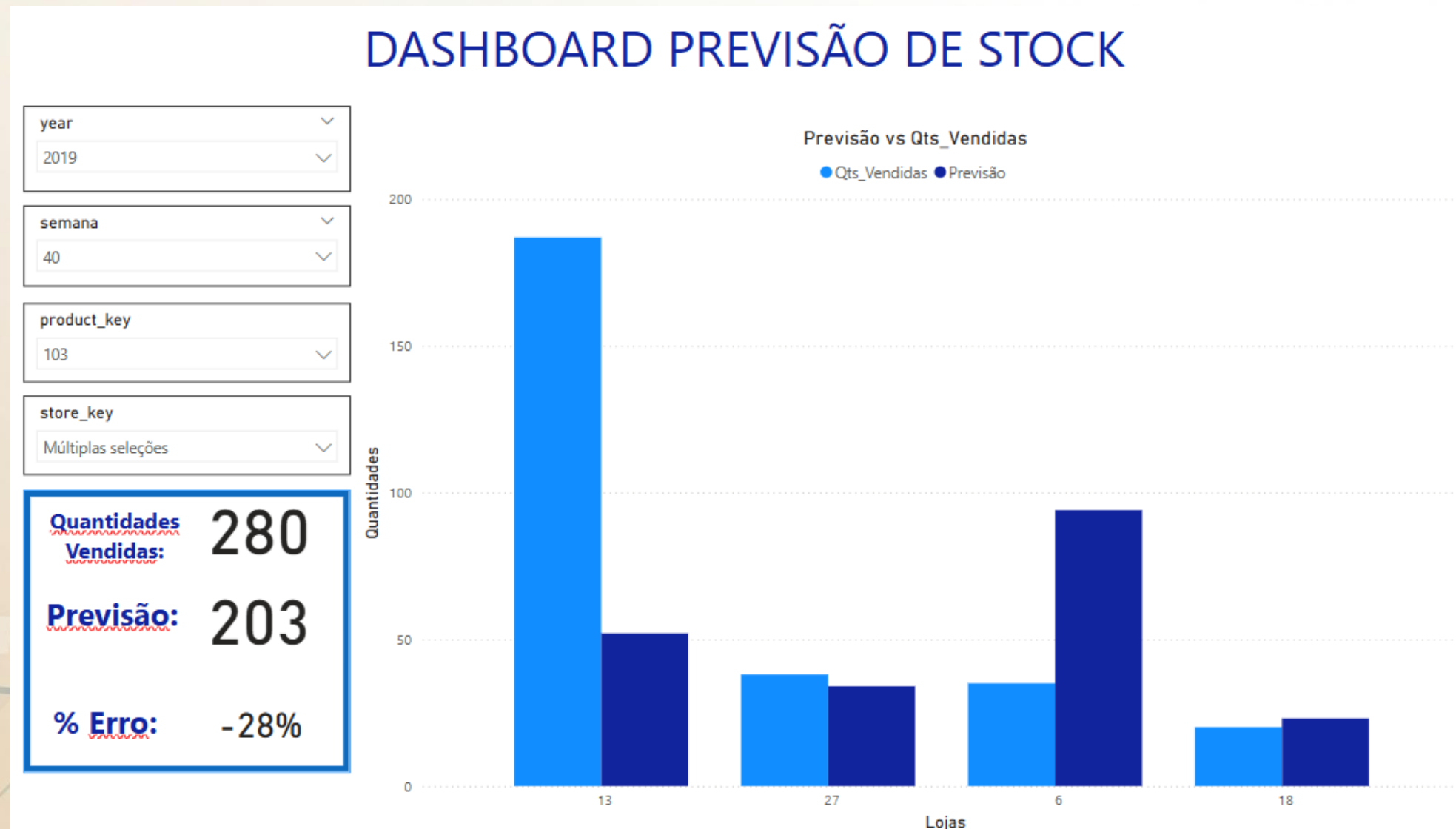
1-+

Run ARIMAX Model



# Visualização e Experimentação

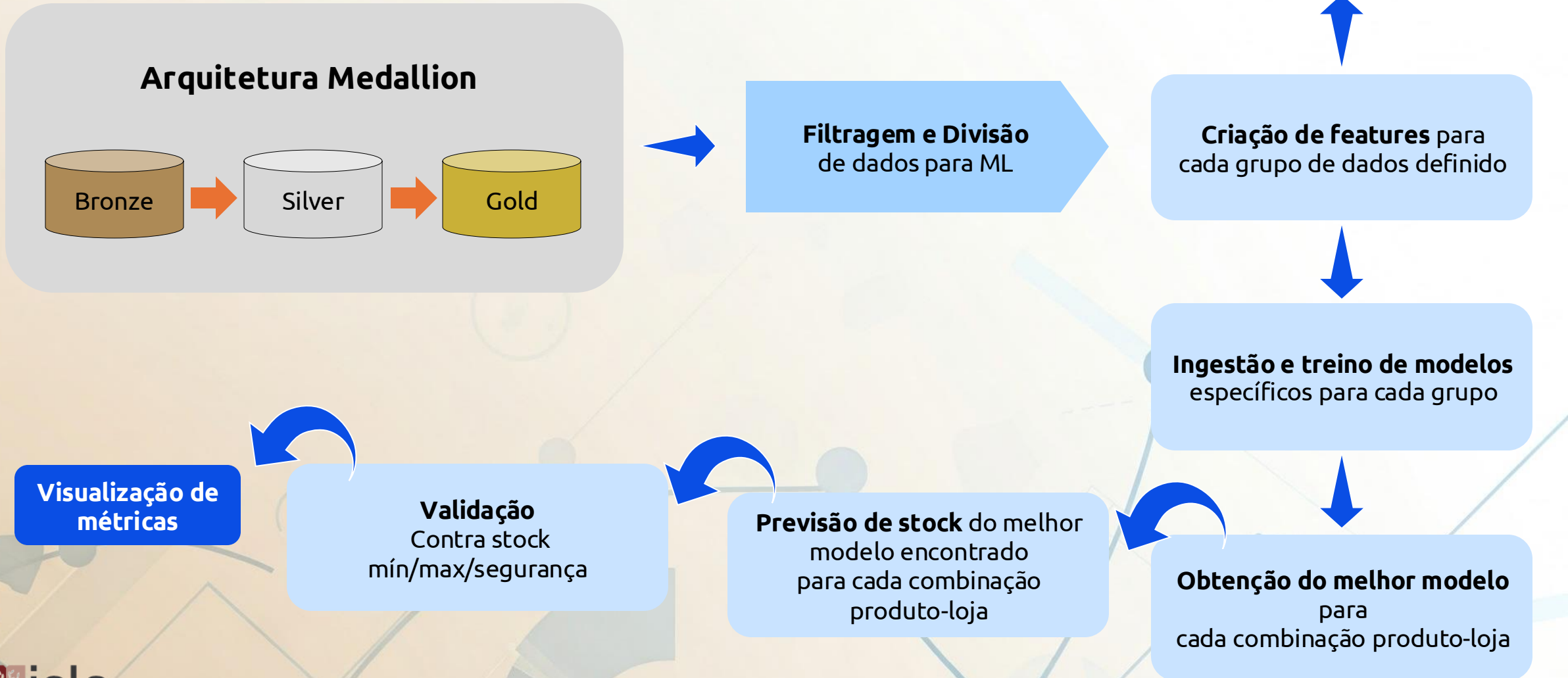
PowerBI e Streamlit



# Agenda

- 01** BUSINESS UNDERSTANDING
- 02** DATA UNDERSTANDING
- 03** DATA PREPARATION
- 04** MODELAÇÃO
- 05** AVALIAÇÃO
- 06** VISUALIZAÇÃO (PowerBI; Streamlit)
- 07** PROJECT OVERVIEW
- 08** CONCLUSÕES E RECOMENDAÇÕES

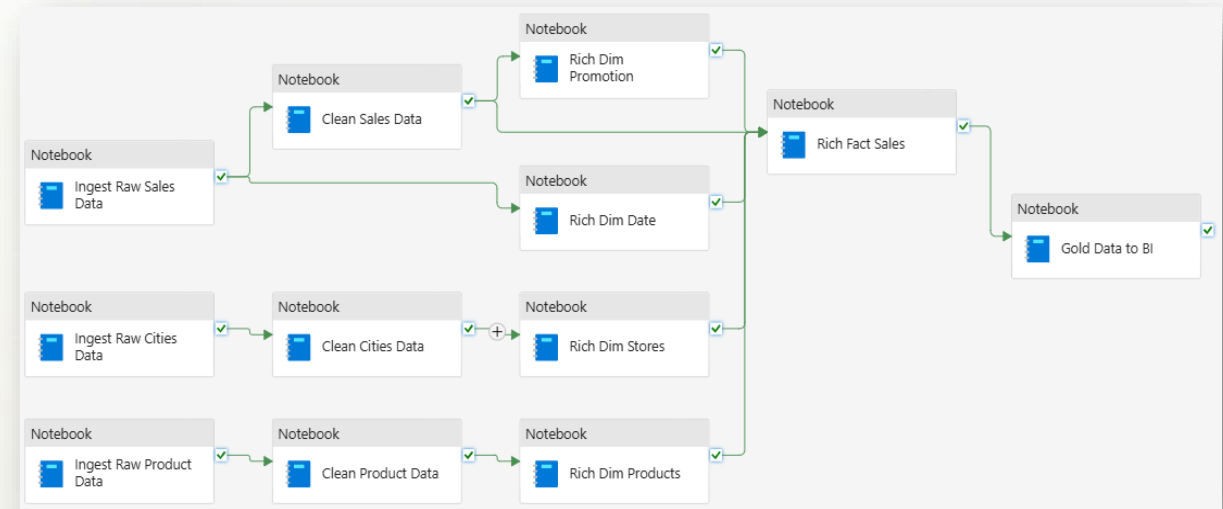
# Overview do Projeto



# Overview do Projeto

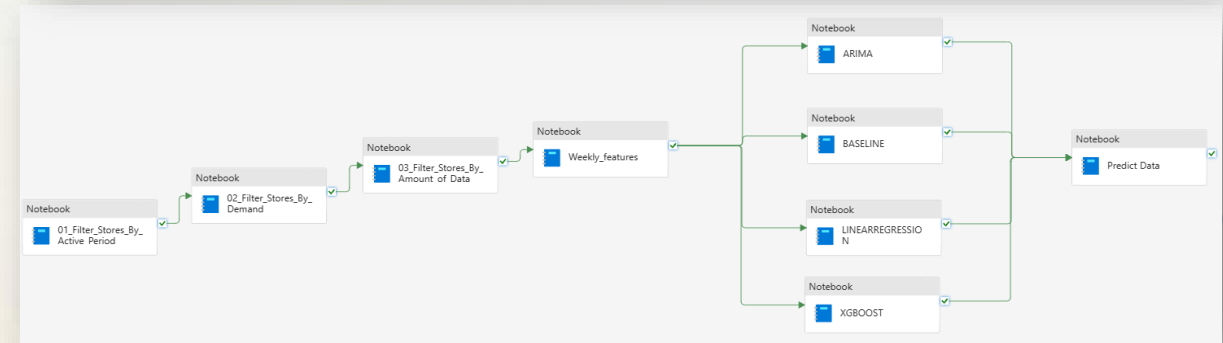
1

**DATA FOUNDATION (Medallion)**  
Desenvolvimento inicial da infraestrutura  
*Janela de execução: Diariamente*



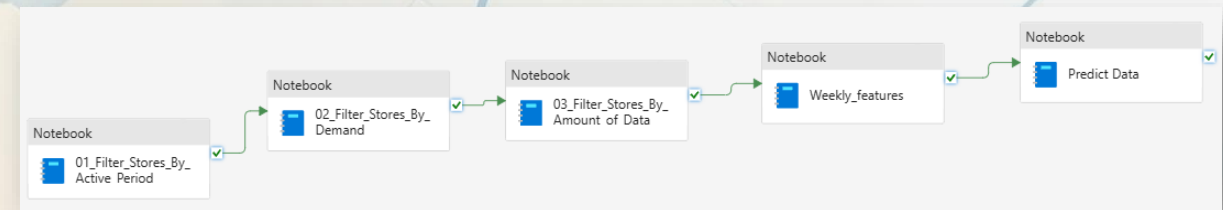
2

**MODEL SELECTION**  
Criação de *Features* dinamicamente.  
Seleção do melhor modelo baseada nas métricas de erro  
*Janela de execução: Semanalmente (Segunda-feira)*



3

**PREDICTION**  
Previsão de stock  
*Janela de execução: Semanalmente (Quinta-feira)*



# Agenda

- 01** BUSINESS UNDERSTANDING
- 02** DATA UNDERSTANDING
- 03** DATA PREPARATION
- 04** MODELAÇÃO
- 05** AVALIAÇÃO
- 06** VISUALIZAÇÃO (PowerBI; Streamlit)
- 07** PROJECT OVERVIEW
- 08** CONCLUSÕES E RECOMENDAÇÕES

# Next steps

**Desenvolvimento de um modelo preditivo de demanda para cada combinação loja-produto**

Obtenção de dados de demanda passíveis de imputação nos dados atuais.

**Estudo de novas features para serem introduzidas no modelo de SARIMAX, Regressão Linear e XGBoost**

Identificar e introduzir novas variáveis nos modelos existentes:

- Dados de marketing e promoções
- Indicadores económicos
- Dados climáticos

**Estudo da aplicação de outros modelos nos dados modelados**

Investigação de técnicas de cutting-edge para melhorar a precisão:

- **Redes Neurais:** Para padrões complexos
- **LSTM:** Para sequências temporais longas
- **Ensemble Methods:** Combinação de múltiplos modelos

## **Benefícios Diretos:**

- Redução do stock parado – menos capital imobilizado
- Diminuição de ruturas de stock – melhor disponibilidade de produto
- Otimização de compras – timing e quantidades ideais
- Melhoria da imagem - menos descontos por excesso de stock





# Obrigado pela atenção!



Pós-Graduação em Analytics e Data Science Empresarial

U.C.: Projeto 2

05 de Julho de 2025

## Grupo 2

- Álvaro Mota
- João Rodrigues
- Maria Lúcia
- Rui Gomes